Computing multiple solutions of topology optimization problems

Ioannis P. A. Papadopoulos

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

There are a number of people without whom this thesis would have not been possible. Firstly, thank you to Prof. Patrick Farrell whose guidance, charisma, and energy has been indispensable. I could not have asked for a better supervisor for the past three years and I will remember all our meetings and trips (in-person and virtual) very fondly. A huge thank you to my other supervisor Prof. Endre Süli whose rigor and wealth of knowledge has been inspiring. I am deeply grateful for the research freedom he has given me and his support with seeing the results to the end. I would like to thank Prof. Thomas Surowiec, for not only his mathematical input, but his kind words throughout my studies and hosting a wonderful trip to Marburg. I am also thankful to Prof. Anton Schiela for funding and hosting a trip to Bayreuth. Moreover, I am grateful to Prof. Coralia Cartis, Prof. Ricardo Ruiz Baier, Prof. Andrew Wathen, and Prof. Sarah Waters for their comments during the transfer and confirmation vivas.

I am deeply grateful to numerous others, who were not directly involved with the mathematics, but were nevertheless just as important in the process. To my parents for their unwavering support. To my girlfriend, Maaike van Swieten, for being there for me in the ups as well as the lows. I am more grateful than you could ever know. To Thomas Pak for his support and company. To Réka Kovács, Bétina Frinault, Alun Vaughan-Jackson, and Adam Fraser for their kindness and banter. To my office-mates Jonah Duncan, James Kohout, and Asad Chaudhary with whom I had the pleasure of sharing the experience of the research roller-coaster. To the MMSC graduating class of 2017; never have I ever met such a talented group of individuals. I am sure we will all remain friends for life and (almost surely) colleagues. To Patrick's other PhD students for their constant source of mathematical inspiration, in particular, Francis Aznaran, Alexei Gazca, and Jingmin Xia. To the numerical analysis group in Oxford for helping me explore my numerical interests and to the PDE CDT Cohort 4 for helping me cultivate my PDE interests. To the SIAM Student Chapter at Oxford as well as my fellow committee members Ambrose Yim and Joe Field. I would also like to give my thanks to the Oxford University Mountaineering Club. Thank you to Nadav Gropper for being an excellent team coach and to all my other team mates for their relaxed, yet focused, approach to training.

Last but not least, I would like to thank The MathWorks, Inc. for their financial support and for hosting me in two very enjoyable summer internships.

Abstract

Topology optimization finds the optimal material distribution of a continuum in a domain, subject to PDE and volume constraints. Density-based models often result in a PDE, volume and inequality constrained, nonconvex, infinite-dimensional optimization problem. These problems can exhibit many local minima. In practice, heuristics are used to obtain the global minimum, but these can fail even in the simplest of cases.

In this thesis we address two core issues related to the nonconvexity of topology optimization problems: the convergence of the discretization and the computation of the solutions. First, we consider the convergence of a finite element discretization of a fluid topology optimization problem. Results available in literature show that there exists a sequence of finite element solutions that weakly(-*) converges to a solution of the analytical problem. We improve on these classical results. In particular, by fixing any isolated minimizer, we show that there exists a sequence of finite element solutions that *strongly* converges to that minimizer. Moreover, these results hold for both traditional conforming finite element methods and more sophisticated divergence-free discontinuous Galerkin finite element methods.

We then focus on developing a solver that can systematically compute multiple minimizers of a general density-based topology optimization problem. This leads to the successful computation of 42 distinct solutions of a two-dimensional fluid topology optimization problem. Finally, by developing preconditioners for the linear systems that arise during the optimization process, we are able to apply the solver to three-dimensional fluid topology optimization problems. This culminates in an example where we compute 11 distinct three-dimensional solutions.

Contents

Any single optimization formulation . . . will be inherently nonconvex.

—Joakim Petersson & Ole Sigmund, 1998

1 Introduction

1.1 Topology optimization

The design of optimal structures is a ubiquitous task faced in engineering. An important mathematical technique extensively used in the initial stages of the design is known as *topology optimization*. The goal of topology optimization is to find the optimal design of a structure or device that minimizes an objective functional. The resulting algorithms are flexible and allow for initial guesses that are substantially different to the final solution; in particular, prior knowledge of the optimal shape or topology of the solution is not required. Topology optimization is a more general technique than its cousin shape optimization, which requires knowledge of the topology of the solution. The differences are highlighted in Fig. 1.1. Due to its flexibility, topology optimization has found uses in a number of industrial applications including airplane wings [1], semiconductor laser designs [6], sophisticated pumps [15], and orthopaedic implants [172] to name but a few. However, this flexibility tends to come at the cost of an infinite-dimensional, nonconvex, nonsmooth, and constrained optimization problem. As we will discuss throughout this thesis, such models have a rich complexity in their analysis and pose a distinct difficulty in the computation of the solutions.

A typical topology optimization problem consists of three components: the formulation, the discretization, and the optimization strategy. The formulation defines an objective functional to be minimized and states the necessary constraints. The material properties are normally described by functions that are determined by

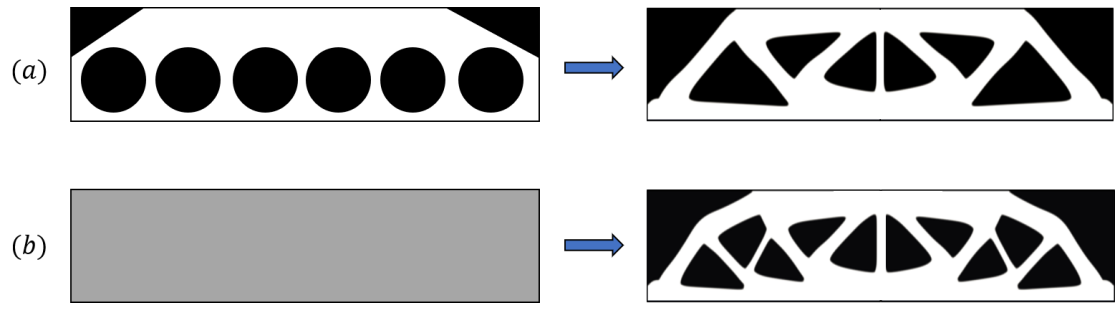


Figure 1.1: Shape (a) vs. topology (b) optimization approaches for minimizing the compliance of a structure undergoing a force. The initial designs are shown on the left and the resulting optimal solutions on the right. Shape optimization keeps the topology of the initial guess during the optimization process. Topology optimization does not have this requirement and finds a solution with four additional holes, minimizing the compliance by a further 6%.

partial differential equations (PDEs). Irrespective of the complexity of the PDEs, analytical solutions to topology optimization problems are very rarely known in practice, even for baseline problems. Hence, we must discretize the optimization problem in order to numerically compute approximations to the analytical solutions. The resulting finite-dimensional problem then requires an optimization strategy to compute the corresponding finite-dimensional solutions. The point at which the problem is discretized differs between different optimization strategies. We will opt for an optimize-then-discretize approach, where the first-order optimality conditions of the infinite-dimensional problem are first derived and then discretized. This is in contrast to discretizing the optimization problem at the level of the objective functional which often results in mesh dependence [144]. The choice of formulation, discretization, and optimization strategy tend to be linked.

In this thesis, we focus entirely on the *density* approach to topology optimization. First introduced independently by Bendsøe [28] and Zhou and Rozvany [188], in the density approach, the topology of the solution is parameterized by a function $\rho : \Omega \to \{0, 1\}$. Optimal regions are defined by the set $\{\rho = 1 \text{ a.e.}\}$, whereas the subdomain $\{\rho = 0 \text{ a.e.}\}$ is interpreted as void or holes in the design domain. Hence, the problem is reduced to finding the optimal function ρ. In general, restricting the range of ρ to $\{0, 1\}$ results in ill-posed or numerically intractable problems. Hence, the problem is often relaxed to finding functions $\rho : \Omega \to [0, 1]$ and an interpolation

scheme is constructed to penalize intermediate values of ρ. The density approach can be simple to implement. For example, Sigmund and coworkers have designed code in MATLAB, for certain problems, that are under 100 lines [18, 74, 150].

The construction of the interpolation scheme gives rise to different submodels of the density approach. Bendsøe [28] and Zhou and Rozvany's [188] approach was coined as the solid isotropic material with penalization (SIMP) approach. SIMP is the most popular interpolation scheme when considering elastic materials [29] and heat transfer [79]. Other interpolation schemes are known as the rational approximation of material properties (RAMP) [160] and SINH [44]. An advantage of RAMP is that there is nonzero sensitivity even if the material distribution value is zero, which can help remedy certain numerical instabilities. Since the topology is represented as a function, any number of classical discretization techniques can be applied, including finite difference discretizations [29], finite element discretizations [14, 36, 64, 65], and finite volume discretizations [57, 58, 79, 98, 101, 111, 133, 161, 169]. The resulting optimization problem is a PDE-constrained optimization problem. Therefore, many popular optimization strategies are based on calculating the sensitivities or adjoints of the optimization problem [90, Ch. 1] or by solving the first-order optimality conditions. By far, the most popular optimization strategy is a nested sensitivity approach. Here, an update for the material distribution function is computed via solving the adjoint equation and then the updates for the state variables are found by solving the forward problem [140, 162, 163, 188]. The box constraints on the material distribution complicate the computation of the updates. A popular algorithm to handle the box constraints is the method of moving asymptotes (MMA) of Svanberg [162].

An alternative to a nested approach is the use of simultaneous analysis and design (SAND) methods [65, 92]. In a SAND approach, the material distribution and state variables are simultaneously updated. The advantage of a SAND method is that it can be reformulated as solving the first-order optimality conditions, which can be handled by Newton-like methods [90, Ch. 2]. Cost per iteration of Newton-like methods can be higher but the convergence is often superlinear [90, Ch. 2.4.2].

Most importantly for the work in this thesis, Newton-like methods can be coupled with *deflation*; an algorithm for computing multiple solutions of nonlinear PDEs [66]. Rojas-Labanda and Stolpe [139] benchmarked various sensitivity and SAND based optimizers. The conclusion was that SAND methods (in particular those based on barrier methods like IPOPT [170]) tended to find better minimizers than their sensitivity-based competitors.

Although they are beyond the scope of this thesis, we briefly mention two other highly successful parameterizations for topology optimization problems; hard-kill methods and boundary variation methods [53, 152]. In hard-kill methods, the domain is discretized into (normally square) elements, and the elements are either tagged with zero or one. If they are tagged as one, they are part of the optimal design and if not, they are void. In some ways, this can be interpreted as a density-approach where the material distribution is discretized with a piecewise constant discretization and the range is not relaxed from $\{0, 1\}$. The removal (or addition) of elements is based on heuristics which do not necessarily involve gradient information. The most famous hard-kill algorithm is known as the evolutionary structural optimization method (ESO) [179, 180].

Instead of tracking whole regions, boundary variation methods instead track the boundaries between the structure and the void. An advantage of these methods is that they result in clear boundaries, a common criticism of the density approach. Two commonly used choices of boundary variation methods are level-set methods [12, 146] and phase-field methods [38]. In level-set methods, the boundary is tracked by a scalar level-set function which is evolved by implicitly solving the Hamilton–Jacobi equation. Topological derivatives are sometimes incorporated in the optimization process in order to aid the introduction of holes away from the free boundary [11, 47, 173]. In phase-field methods, the boundary is represented by a scalar phase-field function. Here, the boundary is not tracked throughout the optimization process and the mechanism of the different "phases" are typically described by the Allen–Cahn or Cahn–Hilliard equations.

The first mathematical derivation of a topology optimization problem is often attributed to a paper of Bendsøe and Kikuchi from 1988 [30]. Their goal was to find the topology of a two-dimensional linearly elastic material, restricted to occupying up to a half of a rectangular domain, that minimizes the displacement caused by a force. This particular problem is known as compliance minimization. Then, in 2003, the first topology optimization problem for fluids was proposed by Borrvall and Petersson [36]. They derived a density-based model with an interpolation scheme similar to RAMP. The setup of an example of a Borrvall–Petersson problem, called the double-pipe [36, Sec. 4.5], is featured in Fig. 1.2. Here, a Stokes fluid enters a rectangular domain from two inlets on the left-hand side and exits through two outlets on the opposing side. The problem enforces a restriction on the volume that the fluid can occupy: namely, 1/3 of the total area of the rectangle. The goal is to find the optimal channels, carrying the fluid, that minimize the total power dissipation of the flow. The velocity profiles of two solutions are given in Fig. 1.3.

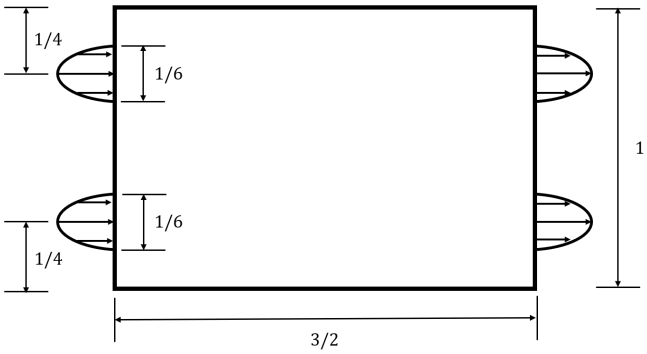


Figure 1.2: The setup of the Borrvall–Petersson double-pipe problem, an example of a topology optimization problem for fluids in Stokes flow. The goal is to find the channels, restricted to occupying up to 1/3 of the area of the domain, carrying the fluid from the inlets to the outlets that minimize the power dissipation of the fluid. The setup of the double-pipe problem is further discussed in Section 4.6.1.

The double-pipe example highlights a key feature of topology optimization problems: in general they are nonconvex and can support multiple local minima. This is irrespective of whether a density approach, hard-kill method, or boundary variation method is used [153, Sec. 5]. As topology optimization is typically used in the discovery phase of design, the ability to compute solutions different to those

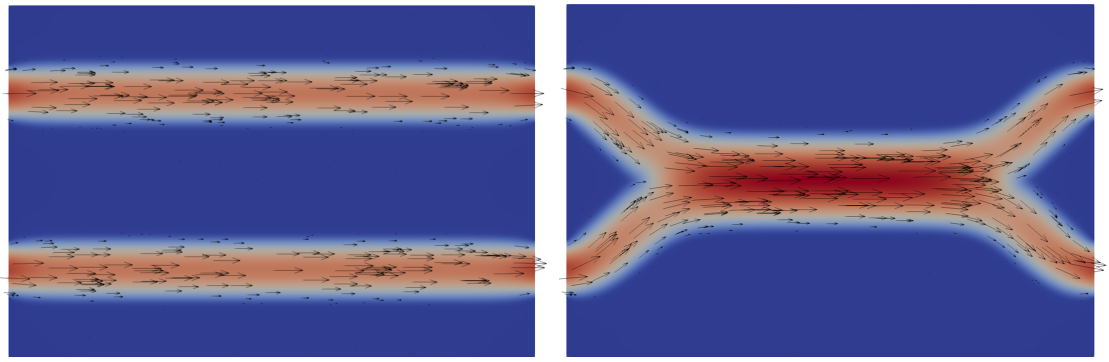


Figure 1.3: The velocity of two (locally) optimal solutions for the double-pipe problem. The arrows indicate the direction and magnitude of the flow. The two solutions have different topologies and the solution on the right results in a power dissipation value that is roughly 3/4 of the power dissipation of the solution on the left.

anticipated by the designer is extremely valuable. By finding multiple solutions, the designer is able to choose the best available in a postprocessing step. In particular, designs that are undesirable due to manufacturing or aesthetic reasons can be discarded. For this reason, many industrial applications can benefit from having a choice of multiple (locally) optimal configurations [59].

1.2 Nonconvex optimization

Nonconvex optimization problems can exhibit complicated solution landscapes consisting of many local and global minima. To this end, we introduce the following definitions of a global and local minimizer.

Definition 1.1 (Global minimizer)**.** *Consider the set X and a functional $J : X \to \mathbb{R}$. We say that $x_* \in X$ is a global minimizer of J, in the feasible set X, if $J(x_*) \leq J(x)$ for all $x \in X$.*

Definition 1.2 (Local minimizer)**.** *Consider the set X and a functional $J : X \to \mathbb{R}$. We say that $x_* \in X$ is a local minimizer of J, in the feasible set X, if there exists an open neighborhood $N \subset X$ around x_* such that $J(x_*) \leq J(x)$ for all $x \in N$.*

Many efficient optimization strategies, particularly for PDE-constrained optimization problems, are local and utilize gradient information of the problem to recover a minimum. Assuming sufficient regularity, these algorithms aim to

find a stationary solution, i.e. a solution to the first-order optimality conditions. For convex problems, this is equivalent to finding the global minimum. However, this is not the case for nonconvex problems. Although many algorithms ensure the computation of a descent direction, there is often no guarantee whether they converge to a solution that is local or global.

The complications that arise due to the nonconvexity of the infinite-dimensional optimization problem are two-fold. Firstly, the nonconvexity complicates the numerical analysis of the convergence of the discretization. In the context of topology optimization, there is alarmingly scarce analysis on whether the finite element method converges in a suitable sense to the analytical minimizers. Results on finite element convergence for density-based topology optimization problems were pioneered by Petersson and coworkers [35, 36, 124–127]. Other finite element convergence results can also be found in the works of Bourdin [37], Greifenstein and Stingl [84], Haslinger and Mäkinen [86], and Talischi and Paulino [164]. Often, the proven convergence is weak in the material distribution, which can allow oscillations in the approximation. These oscillations are normally called checkerboarding in the context of topology optimization (see Section 2.3). In some problems, the convergence of the material distribution is improved to strong convergence. However, in none of the cited papers does the analysis thoroughly discuss the nonconvexity of the problem. It is not clear if there exist finite element sequences that converge to every analytical minimizer of the nonconvex problems.

Secondly, even if we assume that the discretizations are well-behaved, the challenge of actually computing the discretized minimizers still remains [153]. As highlighted by Sigmund and Petersson [153, Sec. 4], the most common trick in topology optimization literature to globally optimize the problem is via continuation of model parameters. This was first utilized in the context of topology optimization by Allaire and Francfort [10]. At its best, continuation is heuristic, and at its worst it can completely fail. Stolpe and Svanberg [159] have provided elementary examples where this occurs. For example, a SIMP formulation [29] of the compliance

minimization of a six-bar truss can be reduced to the discretized optimization problem [159, Sec. 3.1],

$$\min_{(x_1,x_2)\in\mathbb{R}^2} \left(\max\{ \frac{8\beta_t}{x_1^{p_s}+5x_2^{p_s}} + \frac{2\beta_t}{5x_1^{p_s}+x_2^{p_s}}, \frac{8}{x_1^{p_s}+5x_2^{p_s}} + \frac{18}{5x_1^{p_s}+x_2^{p_s}} \} \right)$$
$$\text{such that } \; x_1 + x_2 = 1, \quad 0 \leq x_1, x_2 \leq 1.$$

Here p_s denotes the SIMP continuation parameter and $\beta_t = 2(1-\nu_t^2)/E$, where ν_t is the Poisson ratio and E is the modulus of elasticity. A typical strategy is to find a minimizer to the optimization problem at $p_s = 1$, and then, at each continuation step, use the previous solution as an initial guess for the next value of p_s. In this case, suppose we fix $\beta_t = 2.6$. A poor starting guess for $p_s = 1$ can result in convergence to the local minimizer $\mathbf{x} = (0.5, 0.5)$. Then, even as $p_s \to \infty$, the continuation method will always return $\mathbf{x} = (0.5, 0.5)$ and will not converge to the true global solution, $\mathbf{x} = (0, 1)$.

There have been other approaches utilized in the topology optimization literature. One such technique is a multistart approach, e.g. as applied by Rezayat et al. [137]. Here, the optimization strategy is initialized at various initial guesses in the hope that the optimization strategy will converge to different minimizers. However, in general such an approach has a number of limitations:

- Even after discretization, most topology optimization involve computing solutions that are very high-dimensional. Hence, selecting appropriate initial guesses is nontrivial;
- Without additional mechanisms, this approach can (and often does) converge to previously found solutions;
- There is no natural termination criterion for stopping the solution landscape exploration.

A different approach is the use of global search techniques. These algorithms do not necessarily rely on gradient information and use heuristic and stochastic methods to update the design. These optimization strategies include genetic algorithms [22, 23, 94, 110, 171, 187], simulated annealing [148], and differential evolution schemes

[176]. However, as documented by Sigmund [151], such approaches do not scale well as the problem size increases. Moreover, they do not necessarily perform any better than their gradient-based counterparts [151, Sec. 2]. Other global search methods include branch-and-bound strategies introduced by Stolpe and others [4, 5, 134, 184]. However, these have only been applied to truss topology optimization problems and do not have a clear extension to general density-based models.

Another strategy, as investigated by Zhang and Narato [185], is to apply the tunneling method [106] to these problems, adapting the MMA algorithm. Tunneling proceeds by finding a single minimum, then looking for other controls that yield the same functional value (attempting to tunnel into other basins) by solving an auxiliary equation. Deflation is used in the tunneling phase to ensure that the Gauss–Newton procedure applied to the tunneling functional does not converge to the current state. An advantage of this method is that it builds upon current state-of-the-art strategies already used in the community. However, the tunneling phase has an additional cost to the algorithm and, for each application, the tunneling parameters require careful tuning throughout the optimization process [185, Sec. 2]; failing to tune correctly might result in convergence to solutions that have already been found.

1.3 Structure and aims of this thesis

The goal of this thesis is a full numerical treatment of the nonconvexity of topology optimization problems. This can be split broadly into two parts. The first step is to prove that appropriate discretizations can suitably approximate all the minimizers of the analytical problem. The second step is to develop a solver for systematically computing the multiple discretized solutions.

1.3.1 Analysis

Density-approach models in topology optimization are carefully constructed to enforce the necessary properties of the continuum in question, whilst also penalizing intermediate values of the material distribution function. Typically, the material distribution is coupled to the state variables that are present in the standard

continuum equations. This coupling means that the analytical properties of the state variables, e.g. higher regularity properties, do not immediately translate to the solutions of the topology optimization problem. Nevertheless, the solutions tend to exhibit interesting properties that are not strictly enforced in the model. For instance, in some models, numerical experiments tend to reveal transitions in the material distribution that are not sharp, even if the model allows for jumps.

In Chapter 2, we introduce topology optimization models for minimizing the compliance of elastic structures and minimizing the power dissipation of fluids. In the latter model, we prove that the volume constraint on the material distribution is binding, the support of the material distribution is contained within the support of the velocity, regularity results for the velocity and pressure, and a surprising regularity result for the material distribution. These analytical results form the first half of a manuscript that has been submitted for publication [123].

Since a topology optimization problem can feature design domains with complicated geometries and the assumed regularity of the solutions is low, we opt for a finite element discretization in this work. The finite element method is an umbrella term for a large class of different discretizations; we refer to Brenner and Scott [39] for an introduction. At its core, the finite element method triangulates the design domain into simple cells and approximates analytical functions by gluing together piecewise polynomials defined on each cell. After posing the problem in variational form, a finite element method is defined by the choice of the finite element. The piecewise polynomial can be represented by finite-dimensional vectors and, hence, can be used for the numerical purposes. A useful property for any finite element discretization is that as the mesh size tends to zero (i.e. the maximum diameter of the cells in the triangulation decreases in size), the finite element solution converges to the analytical solution it is approximating. For (linear) problems with unique solutions, this is normally shown by deriving approximation estimates [39, Sec. 2.8]. These estimates become increasingly difficult to prove (if they hold) for nonlinear problems. Hence, convergence is normally proven indirectly via compactness theorems that can be used to show that subsequences of bounded sequences of finite element solutions converge.

In the case where there are multiple solutions, one needs to be careful when taking such subsequences as different subsequences may converge to different solutions.

As previously mentioned, the literature on the convergence of finite element methods for topology optimization problems is largely underdeveloped. In Chapter 3, we focus on the Borrvall–Petersson topology optimization problem. We consider different finite element discretizations, from classical conforming methods to more modern divergence-free discontinuous Galerkin discretizations. The results for these different families are similar: for any given isolated minimizer of the problem, there exists a sequence of finite element solutions, satisfying first-order optimality conditions, that strongly converges to that minimizer. These are the first results for the strong convergence of any finite element discretization for the Borrvall–Petersson problem. We also derive the first error bounds found in literature and show that the convergence rates of a conforming finite element method can be bounded above by the convergence of the solutions in weaker norms. The convergence results for conforming finite element discretizations form the second half of a manuscript that has been submitted for publication [123]. The convergence results for divergence-free discontinuous Galerkin discretizations form a separate manuscript that has also been submitted for publication [120].

1.3.2 Solvers for the computation of multiple solutions

Given a model and suitable finite element method that can approximate all the analytical isolated minimizers, the next challenge is to develop an algorithm that can systematically compute these different minimizers. In Chapter 4, we develop a solver called the *deflated barrier method*. Although the analytical and numerical results mainly focus on the Borrvall–Petersson model, the deflated barrier method is flexible and can be applied to a variety of density-based topology optimization problems. The deflated barrier method is constructed so that we retain the property that no prior knowledge of the solutions is required. Hence, in all examples computed in this thesis, the initial guess for the material distribution is a constant value in the design domain. The deflated barrier method can be split into three components:

- deflation, a mechanism for computing multiple solutions of nonlinear problems;
- a primal-dual active set strategy, a Newton-like method that can enforce box constraints on the solutions;
- barrier terms that aid the global nonlinear convergence of the solver.

The deflated barrier method can be seen as a SAND approach. When applied to the Borrvall–Petersson problem, the setup of the solver is similar to Evgrafov's state space Newton's method [65]. The use of a primal-dual active set strategy and barrier terms aid the robustness of the algorithm. The greatest novelty that the algorithm offers is the application of deflation to topology optimization problems. Deflation is a technique originally developed for computing all the roots of a polynomial [175] that has been extended to computing multiple solutions of nonlinear PDEs [66] and nonlinear variational inequalities [68]. Given a problem with multiple solutions, the first solution is computed as normal. Deflation is then used to quotient out the discovered solution, whilst keeping all the other solutions intact. The nonlinear solver can no longer converge to the deflated solution and, hence, will converge to a different solution when reapplied to the problem. Numerically, deflation is cheap, requiring the evaluation of one inner product after each iteration of a Newton-like solver and can be incorporated as a post-processing step with minimal overhead. In Section 4.6 we apply the deflated barrier method to a variety of compliance and Borrvall–Petersson topology optimization problems. The flagship example is the computation of 42 distinct solutions to a Borrvall–Petersson problem with a Navier–Stokes constraint and five holes in the design domain. The deflated barrier method exhibits superlinear convergence for all problems. For conforming discretizations of Borrvall–Petersson problems, we find that the iteration counts are mesh independent and we utilize the solutions to confirm the convergence results from Chapter 3. The iteration counts are not mesh independent for compliance problems, however, we use a grid-sequencing strategy that keeps iteration counts reasonable even on fine meshes. The work in this chapter has been published in the SIAM Journal on Scientific Computing [122].

A critique of SAND algorithms is that each iteration is more expensive to compute than their nested approach counterparts. Although the overall iteration counts may be much lower due to the superlinear convergence, the computational expense to compute one iteration may become prohibitive when these algorithms are applied to three-dimensional topology optimization problems. In Chapter 5, we develop preconditioners for the linear systems that arise when the deflated barrier method is applied to a Borrvall–Petersson problem with a divergence-free discontinuous Galerkin finite element discretization. By one application of block preconditioning, the computational work is immediately reduced to that of a nested approach whilst still retaining the superlinear convergence properties. However, we then apply further block preconditioning and control the innermost Schur complement with an augmented Lagrangian term. As a final step, we develop a robust multigrid cycle for the augmented momentum block. The multigrid cycle features a specialized relaxation method, that can handle the semi-definite terms arising in the augmented Lagrangian, and a representation of the active set (as defined by the primal-dual active set strategy) on the coarser levels. This preconditioner allows us to apply the deflated barrier method to compute multiple solutions of three-dimensional Borrvall–Petersson problems. The main example in this chapter is the computation of eleven distinct solutions of a quadruple-pipe problem with five internal holes in the design domain. The work in this chapter has been included in a manuscript that is in preparation [121].

In Chapter 6, we summarize the work and discuss directions for future extensions.

The art of structure is where to put the holes.

—Robert Le Ricolais, 1894–1977

2 Topology optimization

In this chapter we discuss the formulation of a density approach to topology optimization problems. Once the general problem is defined, we develop two particular models of topology optimization problems: the topology optimization of the minimization of the displacement of a linearly elastic material, and the topology optimization of the power dissipation of fluids. We discuss the fluid model in detail, proving a number of novel results. We show that isolated minimizers of the fluid problem are binding at the volume constraint, satisfy first-order optimality conditions, and possess higher regularity.

2.1 Functional analysis

Before we construct the models, we first define appropriate spaces for the solutions. Throughout this work $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, denotes an open and bounded set. Let $C^k(\Omega)$ denote the set of real-valued k-times continuously differentiable functions in Ω and let $C_c^\infty(\Omega)$ denote the set of smooth functions with compact support in Ω. By considering the closure of Ω, denoted $\bar{\Omega}$, we define the norm $\|\cdot\|_{C^k(\bar{\Omega})}$, for any $u \in C^k(\bar{\Omega})$ by

$$\|u\|_{C^k(\bar{\Omega})} = \sum_{i=1}^{k} \sup_{x \in \bar{\Omega}} |\nabla^i u(x)|. \tag{2.1}$$

We use the standard notation for the Lebesgue and Sobolev spaces equipped with their standard norms, e.g. $(L^r(\Omega), \|\cdot\|_{L^r(\Omega)})$ and $(W^{k,r}(\Omega), \|\cdot\|_{W^{k,r}(\Omega)})$. The Banach space $(L^\infty(\Omega), \|\cdot\|_{L^\infty(\Omega)})$ denotes the vector space of essentially bounded measurable

functions equipped with the essential supremum norm. We denote dual spaces with a star *. The space of traces on the boundary of a function in $W^{1,r}(\Omega)$ is denoted by $W^{1/r',r}(\partial\Omega)$ where r' is the Hölder conjugate of r, i.e. $1/r + 1/r' = 1$. When it exists, the boundary trace operator is denoted by $|_{\partial\Omega} : W^{1,r}(\Omega) \to W^{1/r',r}(\partial\Omega)$. When $k = 1$ and $r = 2$, we define the Hilbert space $H^1(\Omega) := W^{1,2}(\Omega)$ equipped with the inner product $(u, v)_{H^1(\Omega)} := \int_\Omega uv + \nabla u \cdot \nabla v \,\mathrm{d}x$ where ∇ denotes the weak gradient [63, Ch. 5.2]. Moreover, the spaces $W^{k,r}(\Omega)^d$, $d \in \{2, 3\}$, are the set of vector-valued functions $\boldsymbol{u} = (u_1, \ldots, u_d)$ such that each component $u_i \in W^{k,r}(\Omega)$, $i = 1, \ldots, d$. We define the following subspaces as:

$$L^2_0(\Omega) := \left\{ q \in L^2(\Omega) : \int_\Omega q \,\mathrm{d}x = 0 \right\}, \tag{2.2}$$

$$H^1_{\mathrm{div}}(\Omega)^d := \{\boldsymbol{v} \in H^1(\Omega)^d : \mathrm{div}(\boldsymbol{v}) = 0 \text{ a.e. in } \Omega\}, \tag{2.3}$$

$$H^1_0(\Omega)^d := \{\boldsymbol{v} \in H^1(\Omega)^d : \boldsymbol{v}|_{\partial\Omega} = \boldsymbol{0}\}. \tag{2.4}$$

Let $\Gamma \subseteq \partial\Omega$ be a subset of the boundary with nonzero Hausdorff measure $\mathcal{H}^{d-1}(\Gamma) > 0$. If $\boldsymbol{g} \in H^{1/2}(\Gamma)^d$, then, the following subspaces are defined as:

$$H^1_{|\Gamma,\boldsymbol{g}}(\Omega)^d := \{\boldsymbol{v} \in H^1(\Omega)^d : \boldsymbol{v}|_\Gamma = \boldsymbol{g}\}, \tag{2.5}$$

$$H^1_{|\Gamma,\boldsymbol{g},\mathrm{div}}(\Omega)^d := H^1_{|\Gamma,\boldsymbol{g}}(\Omega)^d \cap H^1_{\mathrm{div}}(\Omega)^d. \tag{2.6}$$

We also define the following function spaces which will be utilized when considering nonconforming velocity finite element spaces in Chapter 3:

$$\boldsymbol{H}(\mathrm{div};\Omega) := \left\{ \boldsymbol{v} \in L^2(\Omega)^d : \mathrm{div}(\boldsymbol{v}) \in L^2(\Omega) \right\}, \tag{2.7}$$

$$\boldsymbol{H}_0(\mathrm{div};\Omega) := \left\{ \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega) : (\boldsymbol{v} \cdot \boldsymbol{n})|_{\partial\Omega} = 0 \right\}, \tag{2.8}$$

$$\boldsymbol{H}_{\boldsymbol{g}}(\mathrm{div};\Omega) := \left\{ \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega) : (\boldsymbol{v} \cdot \boldsymbol{n})|_{\partial\Omega} - \boldsymbol{g} \cdot \boldsymbol{n} = 0 \right\}, \tag{2.9}$$

$$\boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega) := \left\{ \boldsymbol{v} \in \boldsymbol{H}_{\boldsymbol{g}}(\mathrm{div};\Omega) : \mathrm{div}(\boldsymbol{v}) = 0 \text{ a.e. in } \Omega \right\}. \tag{2.10}$$

We note that $(\boldsymbol{v}\cdot\boldsymbol{n})|_{\partial\Omega}$ is well defined for all $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega)$ [19, Th. 3.12]. Moreover, $\boldsymbol{H}(\mathrm{div};\Omega)$ is a Hilbert space when equipped with the inner product

$$(\boldsymbol{u}, \boldsymbol{v})_{\boldsymbol{H}(\mathrm{div};\Omega)} := \int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} + \mathrm{div}(\boldsymbol{u})\,\mathrm{div}(\boldsymbol{v}) \,\mathrm{d}x. \tag{2.11}$$

Throughout this work, we invoke a number of results from functional analysis. For the convenience of the reader, we quote the main results that we use below.

Theorem 2.1 (Hölder's inequality, App. B.2 in [63]). *Assume that* $1 \leq q,\, q' \leq \infty$, *with* $1/q + 1/q' = 1$. *Then, if* $u \in L^q(\Omega)$ *and* $v \in L^{q'}(\Omega)$, *we have that*

$$\int_\Omega |uv| \, \mathrm{d}x \leq \|u\|_{L^q(\Omega)} \|v\|_{L^{q'}(\Omega)}. \tag{2.12}$$

Theorem 2.2 (Sobolev embedding theorem, Th. 5.4 in [8]). *Let* Ω *be a bounded and open subset of* $\mathbb{R}^d$ *with a Lipschitz boundary. Assume that* $u \in W^{k,q}(\Omega)$, $1 \leq q < \infty$.

(Case A). *If* $k < d/q$ *and* $q \leq s \leq \frac{dq}{d-kq}$, *we have that* $u \in L^s(\Omega)$. *Moreover,*

$$\|u\|_{L^s(\Omega)} \leq C(k, q, d, \Omega) \|u\|_{W^{k,q}(\Omega)}. \tag{2.13}$$

(Case B). *If* $k = d/q$ *and* $q \leq s < \infty$, *we have that* $u \in L^s(\Omega)$. *Moreover,*

$$\|u\|_{L^s(\Omega)} \leq C(k, q, d, \Omega) \|u\|_{W^{k,q}(\Omega)}. \tag{2.14}$$

(Case C). *If* $k = 1$, $d < q \leq \infty$ *and* $r = 1 - d/q$, *we have that* $u \in C^{0,r}(\bar{\Omega})$. *Moreover,*

$$\|u\|_{C^{0,r}(\bar{\Omega})} \leq C(k, q, d, \Omega) \|u\|_{W^{k,q}(\Omega)}. \tag{2.15}$$

Theorem 2.3 (Rellich–Kondrachov theorem, Th. 6.2 in [8]). *Let* Ω *be a bounded and open subset of* $\mathbb{R}^d$ *with a Lipschitz boundary. Suppose that* $1 \leq q < d$, *then for each* $1 \leq s < q'$, *where* q' *is the Hölder conjugate of* q, $W^{1,q}(\Omega)$ *is compactly embedded in* $L^s(\Omega)$. *This is denoted by* $W^{1,q}(\Omega) \subset\subset L^s(\Omega)$.

Theorem 2.4 (Heine–Borel theorem, Th. 11.18 in [75]). *A subset of* $\mathbb{R}^n$, $1 \leq n < \infty$, *is sequentially compact if and only if it is bounded and closed in* $\mathbb{R}^n$.

Theorem 2.5 (Mazur's theorem, App. D.4 in [63]). *Let* X *be a reflexive Banach space. Then, every convex and norm-closed subset of* X *is weakly closed.*

Theorem 2.6 (Eberlein–Šmulian theorem, Th. A.62 in [76]). *Let* E *be a subset of a Banach space* X. *Then the weak closure of* E *is weakly compact if and only if for any sequence* $(x_n) \subset E$ *there exists a subsequence weakly convergent to some element of* X, *i.e. if and only if the weak closure of* E *is weakly sequentially compact.*

Theorem 2.7 (Kakutani's theorem, Th. A.65 in [76]). *A Banach space is reflexive if and only if the closed unit ball* $\{x \in X : \|x\|_X \leq 1\}$ *is weakly compact.*

Corollary 2.1 (Corollary of Kakutani's theorem). *If a Banach space* X *is reflexive, then every norm-closed, bounded, and convex subset of* X *is weakly compact.*

Proof. By Kakutani's theorem, we have that (by rescaling if necessary) every norm-closed ball of X is weakly compact. Let $K \subseteq X$ be norm-closed, bounded, and convex. By Mazur's theorem, since K is norm-closed, it is weakly closed. Since K is bounded, there exists a ball B such that $K \subseteq B$. Hence, since B is weakly compact, then K is weakly compact. □

Theorem 2.8 (Banach–Alaoglu theorem, Th. A.52 in [76]). *If* E *is a neighborhood of* 0 *in a locally convex topological vector space* X*, then*

$$K := \{L \in X' : |L(x)| \leq 1 \ \textit{for every} \ x \in E\}$$

is weakly- compact.*

Theorem 2.9 (Banach's closed range theorem, [25]). *Suppose that* X *and* Y *are Banach spaces, and* $K : Z \to Y$ *is a closed linear operator, where* Z *is dense in* X*. Let* $\ker(K) := \{x \in Z : Kx = 0\}$ *denote the kernel of* K *and let* $K^* : Y^* \to X^*$ *be the transpose of* K*, defined by* $\langle K^* y^*, x\rangle = \langle y^*, Kx\rangle$*, where* X^* *and* Y^* *denote the dual spaces of* X *and* Y*, respectively, and* $\langle \cdot, \cdot \rangle$ *is the duality pairing between* Y^* *and* Y*, or* X^* *and* X*. Then, the following properties are equivalent:*

- $\mathrm{im}(K)$*, the range of* K*, is closed in* Y*;*
- $\mathrm{im}(K^*)$*, the range of* K^**, is closed in* X^**;*
- $\mathrm{im}(K) = [\ker(K^*)]^\circ := \{y \in Y : \langle y^*, y\rangle = 0 \textit{ for all } y^* \in \ker(K^*)\}$*;*
- $\mathrm{im}(K^*) = [\ker(K)]^\circ := \{x^* \in X^* : \langle x^*, x\rangle = 0 \textit{ for all } x \in \ker(K)\}$.

Theorem 2.10 (Implicit Function Theorem, App. C.7 in [63]). *Consider the open set* $Z \subset \mathbb{R}^{n+m}$ *and a function* $\boldsymbol{f} \in C^1(Z)^m$*,* $\boldsymbol{f} : Z \to \mathbb{R}^m$*. Consider a point* $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^{n+m}$*, where* $\mathbf{x}_0 \in \mathbb{R}^n$ *and* $\mathbf{y}_0 \in \mathbb{R}^m$*. Suppose that* $|\det \nabla_{\mathbf{y}} \boldsymbol{f}(\mathbf{x}_0, \mathbf{y}_0)| \neq 0$.

Let $\mathbf{z}_0 = \boldsymbol{f}(\mathbf{x}_0, \mathbf{y}_0)$. *Then there exists an open set* $D \subset Z$, *with* $(\mathbf{x}_0, \mathbf{y}_0) \in D$, *an open set* $X \subset \mathbb{R}^n$, *with* $\mathbf{x}_0 \in X$, *and a* C^1 *mapping* $\boldsymbol{g} : X \to \mathbb{R}^m$, *such that*

1. $\boldsymbol{g}(\mathbf{x}_0) = \mathbf{y}_0$;
2. $\boldsymbol{f}(\mathbf{x}, \boldsymbol{g}(\mathbf{x})) = \mathbf{z}_0$, $\mathbf{x} \in X$,

and

3. *if* $(\mathbf{x}, \mathbf{y}) \in D$ *and* $\boldsymbol{f}(\mathbf{x}, \mathbf{y}) = \mathbf{z}_0$, *then* $\mathbf{y} = \boldsymbol{g}(\mathbf{x})$;
4. *if* $\boldsymbol{f} \in C^k(Z)^m$, *then* $\boldsymbol{g} \in C^k(X)^m$ *for* $k \geq 2$.

2.2 General formulation

A density approach to modeling a topology optimization problem is formulated as follows:

$$\min_{\boldsymbol{u} \in \boldsymbol{U},\, \rho \in C_\gamma} J(\boldsymbol{u}, \rho) \tag{2.16}$$

$$\text{subject to} \begin{cases} F(\boldsymbol{u}, \rho) = 0 & \text{(PDE constraint)}, \\ \int_\Omega (\gamma - \rho)\, \mathrm{d}x \geq 0 & \text{(volume constraint)}, \\ 0 \leq \rho \leq 1 & \text{(box constraints)}, \\ G_k(\boldsymbol{u}, \rho) \geq 0, \quad k = 1, \ldots, m & \text{(additional constraints)}. \end{cases} \tag{2.17}$$

Typically $\boldsymbol{u}$ is a variable of physical interest, often called the *state* variable and $\boldsymbol{U}$ is the space of admissible state functions. For example, in incompressible fluid flow, $\boldsymbol{u}$ often represents the velocity and pressure of the flow and, in elasticity problems, $\boldsymbol{u}$ denotes the displacement. As mentioned in the previous chapter, the material distribution function ρ encodes the topology of the optimal design as a subset of the bounded (design) domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$. The box constraints on ρ ensure that we can interpret the topology of the solution. Namely, the set where $\{\rho = 1 \text{ a.e.}\} \subset \Omega$ is the optimal region that the continuum should occupy and the set $\{\rho = 0 \text{ a.e.}\}$ is interpreted as the holes in the optimal solution. The volume constraint fixes an upper limit for the fraction of the domain Ω that the optimal solution can occupy. We define the space of admissible material distributions C_γ as the following:

$$C_\gamma := \left\{ \eta \in L^\infty(\Omega) : 0 \leq \eta \leq 1 \text{ a.e. in } \Omega, \ \int_\Omega \eta \, \mathrm{d}x \leq \gamma |\Omega| \right\}, \tag{2.18}$$

where $\gamma \in (0,1)$ is the volume fraction. We see that C_γ incorporates the volume constraint and box constraints on ρ as part of its definition. Hence, when we define specific models later, we will not explicitly write out the volume and box constraints. The functional $J : \boldsymbol{U} \times C_\gamma \to \mathbb{R}$ is a functional that maps from the spaces of admissible state and material distribution functions to the real numbers. The objective of the topology optimization problem is to find the pair $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that minimizes J. The PDE constraint $F : \boldsymbol{U} \times C_\gamma \to X$, where X is an appropriate Banach space, enforces the relevant physical properties on the state $\boldsymbol{u}$. For fluid problems, these can be momentum and incompressibility conditions and for elasticity, these can be the equations of linear elasticity. Designing the functional J and PDE constraint F so that they represent the necessary physical requirements and constraints, whilst also penalizing intermediate values of ρ, is nontrivial and requires a different construction for each physical system. The additional constraints often model further desirable properties for the solutions. For example, these can include the bound formulation of buckling topology optimization problems [29, Ch. 2.1.2] or stress constraints [29, Ch. 2.3]. Neither of the models we derive below will have additional constraints.

2.3 Compliance of elastic structures

A significant portion of the topology optimization literature focuses on finding the optimal topology of an elastic material that minimizes its displacement, whilst experiencing a force, that can only occupy a fraction of the domain. This is commonly referred to as the topology optimization of the compliance of a structure.

For simplicity we consider structures that obey linear elasticity. The optimization problem is posed as follows: given the volume fraction $\gamma \in (0,1)$, find $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that minimizes

$$J(\boldsymbol{u}, \rho) = \int_{\Gamma_N} \boldsymbol{f} \cdot \boldsymbol{u} \, \mathrm{d}s \tag{C}$$

subject to the linear elasticity PDE constraint

$$F(\boldsymbol{u},\rho)=\begin{cases}-\operatorname{div}(\mathsf{S})=0 & \text{in } \Omega,\\ \mathsf{S}=k(\rho)\left[2\mu_l\mathsf{D}(\boldsymbol{u})+\lambda_l\operatorname{tr}(\mathsf{D}(\boldsymbol{u}))\mathsf{I}\right] & \text{in } \Omega,\\ \mathsf{S}\boldsymbol{n}=\boldsymbol{f} & \text{on } \Gamma_N.\end{cases} \tag{2.19}$$

Here the space of admissible state functions is $\boldsymbol{U}=H^1_{\Gamma_D,\mathbf{0}}(\Omega)^d$, i.e. H^1 vector-valued functions which are zero on the boundary $\Gamma_D\subset\partial\Omega$. The state $\boldsymbol{u}$ denotes the displacement of the structure and S denotes the stress tensor. The traction $\boldsymbol{f}\in H^{1/2}(\Gamma_N)^d$ is known, $\Gamma_N,\Gamma_D\subset\partial\Omega$ are known disjoint boundaries on $\partial\Omega$ such that $\Gamma_N\cup\Gamma_D=\partial\Omega$, μ_l and λ_l are the Lamé coefficients, $\operatorname{tr}(\cdot)$ is the matrix-trace operator, I is the $d\times d$ identity matrix, $\boldsymbol{n}$ is the unit outward normal and

$$\mathsf{D}(\boldsymbol{u})=\frac{1}{2}(\nabla\boldsymbol{u}+\nabla\boldsymbol{u}^\top),\quad k(\rho)=\epsilon_{\text{SIMP}}+(1-\epsilon_{\text{SIMP}})\rho^{p_s},$$

where $0<\epsilon_{\text{SIMP}}\ll 1$ and $p_s\geq 1$. Unless stated otherwise, we choose $\epsilon_{\text{SIMP}}=10^{-5}$ and $p_s=3$. The use of $k(\rho)$ is known as the Solid Isotropic Material with Penalization (SIMP) model. Bendsøe and Sigmund [29, Ch. 1] provide a concise physical interpretation of the SIMP model. In essence, for ρ close to one, $k(\rho)$ is close to one, indicating the presence of material, whereas where ρ is close to zero, $k(\rho)$ approaches ϵ_{SIMP}, indicating void. It is typical to raise ρ to the power of $p_s>1$ in order to penalize intermediate values of ρ.

We now introduce a Lagrange multiplier $\boldsymbol{v}\in\boldsymbol{U}$ and reformulate (C) as finding the stationary points $(\boldsymbol{u},\rho,\boldsymbol{v})\in\boldsymbol{U}\times C_\gamma\times\boldsymbol{U}$ of

$$\begin{aligned}&\int_{\Gamma_N}\boldsymbol{f}\cdot\boldsymbol{u}\,\mathrm{d}s\\ &\quad+\int_\Omega k(\rho)\left[2\mu_l\mathsf{D}(\boldsymbol{u}):\mathsf{D}(\boldsymbol{v})+\lambda_l\operatorname{tr}(\mathsf{D}(\boldsymbol{u}))\cdot\operatorname{tr}(\mathsf{D}(\boldsymbol{v}))\right]\mathrm{d}x-\int_{\Gamma_N}\boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}s,\end{aligned} \tag{2.20}$$

which follows after an integration by parts of the PDE constraint.

We notice that (2.20) has symmetry that can be exploited to reduce the size of the problem. By deriving the Euler–Lagrange equations of (2.20), we see that the linear elasticity PDE constraint (2.19) on $\boldsymbol{u}$ must be satisfied. However, if we consider the adjoint equation involving $\boldsymbol{v}$, it can be verified that $\boldsymbol{v}=-\boldsymbol{u}$.

Substituting this relation into (2.20), we see that (2.20) is equivalent to finding the stationary points $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ of

$$2\int_{\Gamma_N} \boldsymbol{f} \cdot \boldsymbol{u}\, \mathrm{d}s - \int_\Omega k(\rho)\left[2\mu_l \mathbf{D}(\boldsymbol{u}) : \mathbf{D}(\boldsymbol{u}) + \lambda_l \mathrm{tr}(\mathbf{D}(\boldsymbol{u})) \cdot \mathrm{tr}(\mathbf{D}(\boldsymbol{u}))\right] \mathrm{d}x. \tag{2.21}$$

The substitution is useful as it greatly reduces the size of the problem after discretization.

Unfortunately, the optimization problem (C) is ill-posed in general and does not have minimizers in the continuous setting. Naïve attempts at finding minimizers often yield checkerboard patterns of ρ. Checkerboarding is a phenomenon where the discretized material distribution oscillates between the values zero and one between neighboring elements. An example of checkboarding is given in Fig. 2.1. Such solutions cannot be manufactured. Although a different choice of finite element

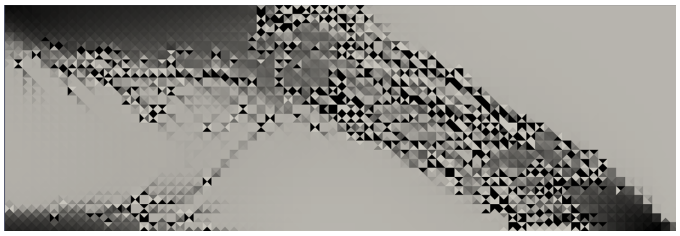


Figure 2.1: Checkerboarding behavior in ρ whilst attempting to find the minimizer of a Messerschmitt–Bölkow–Blohm (MBB) beam, an example of (C), without a restriction method. The values of ρ are wildly oscillating between elements.

spaces may avoid the checkerboarding, the solutions will still be mesh dependent. As the mesh is refined, the beams of the solutions will become ever thinner, leading to nonphysical solutions in the limit. There are several schemes employed by the topology optimization community to obtain physically reasonable solutions for ρ and they are known as *restriction methods* [29]. We opt for the addition of a Ginzburg–Landau energy term,

$$J_{\mathrm{GL}}(\boldsymbol{u}, \rho) := J(\boldsymbol{u}, \rho) + \frac{\beta\epsilon}{2}\int_\Omega |\nabla\rho|^2\, \mathrm{d}x + \frac{\beta}{2\epsilon}\int_\Omega \rho(1-\rho)\mathrm{d}x, \tag{C$_{\mathrm{GL}}$}$$

with $0 < \beta \ll 1$, $0 < \epsilon \ll 1$, to the objective function. J_{GL} requires ρ to be weakly differentiable. Hence we now seek a solution $\rho \in C_\gamma \cap H^1(\Omega)$. Physically, the Ginzburg–Landau term corresponds to penalizing fluctuations in the values of ρ. As

$\epsilon \to 0$, it was shown by Modica [113] that the Ginzburg–Landau energy Γ-converges to the perimeter functional associated with restricting $\rho(x) \in \{0, 1\}$, providing rigorous mathematical grounding for this choice of regularization. For sufficiently large values of β, this introduces minima and removes the checkerboarding effect. Other restriction methods used by the topology optimization community include gradient control [34], perimeter constraints [34], sensitivity filtering [37, 149], design filtering [45, 105] and regularized penalty [34].

2.4 Power dissipation of fluid flow

The first model for the topology optimization of a fluid was proposed by Borrvall and Petersson [36]. Their goal was to find the subdomain that minimizes the power dissipation of a fluid, subject to the Stokes equations and a volume constraint restricting the proportion of the domain that the fluid can occupy. In their paper, they derive *generalized Stokes equations*, which involve the classical velocity and pressure terms but also incorporate the material distribution, ρ, via an inverse permeability term α. The presence of fluid is indicated by a value of one in the material distribution whereas absence of fluid is represented by a value of zero. The inverse permeability term is constructed in order to favour solutions where ρ is close to zero or one. From the generalized Stokes equations, Borrvall and Petersson formulate an infinite-dimensional nonconvex optimization problem with inequality and box constraints. The derived optimization problem requires no further regularization for well-posedness, in contrast to the structural topology optimization in the previous section. The optimization problem supports (not necessarily unique) local minima.

Since Borrvall and Petersson's seminal work, there have been numerous extensions. Evgrafov [64], Olesen et al. [118], and Gersborg-Hansen et al. [80] extended the model to fluids satisfying stationary Navier–Stokes flow. Aage et al. [3] solved the first three-dimensional problem. Kreissl et al. [103] and Deng et al. [55] were the first to consider unsteady Navier–Stokes flow and Deng et al. [54] later included body forces. Alonso et al. considered rotating bodies in cylindrical coordinates [14,

16]. For a detailed review on the literature of the topology optimization of fluids, we refer to the work of Alexandersen and Andreasen [9].

2.4.1 The Borrvall–Petersson model

Given a volume constraint on a fluid in a fixed Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, the Borrvall–Petersson model attempts to minimize the energy lost by the flow due to viscous dissipation, whilst maximizing the flow velocities at the applied body force. More precisely, given the volume fraction $\gamma \in (0, 1)$, the objective is to find a velocity-material distribution pair $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that minimizes

$$J(\boldsymbol{u}, \rho) := \frac{1}{2} \int_\Omega \left(\alpha(\rho)|\boldsymbol{u}|^2 + \nu|\nabla \boldsymbol{u}|^2 - 2\boldsymbol{f} \cdot \boldsymbol{u} \right) \mathrm{d}x, \tag{BP}$$

where $\boldsymbol{U} = H^1_{|\partial\Omega, \boldsymbol{g}, \mathrm{div}}(\Omega)^d$ as defined in (2.6).

Here, $\boldsymbol{f} \in L^2(\Omega)^d$ is a body force and $\nu > 0$ is the (constant) viscosity. Moreover, the (possibly inhomogeneous) boundary data $\boldsymbol{g} \in H^{1/2}(\partial\Omega)^d$ and $\boldsymbol{g} = \boldsymbol{0}$ on a subset of the boundary $\Gamma \subset \partial\Omega$, with $\mathcal{H}^{d-1}(\Gamma) > 0$, i.e. Γ has nonzero Hausdorff measure on the boundary. Here, α is the inverse permeability, modeling the influence of the material distribution on the flow. For values of ρ close to one, $\alpha(\rho)$ is small, permitting fluid flow; for small values of ρ, $\alpha(\rho)$ is very large, restricting fluid flow. The function α satisfies the following properties:

(A1) $\alpha : [0, 1] \to [\underline{\alpha}, \overline{\alpha}]$ with $0 \leq \underline{\alpha} < \overline{\alpha} < \infty$;

(A2) α is convex and monotonically decreasing;

(A3) $\alpha(0) = \overline{\alpha}$ and $\alpha(1) = \underline{\alpha}$;

(A4) α is twice continuously differentiable,

generating an operator also denoted $\alpha : C_\gamma \to L^\infty(\Omega; [\underline{\alpha}, \overline{\alpha}])$. Typically, in the literature α takes the form [36, 65]

$$\alpha(\rho) = \bar{\alpha} \left(1 - \frac{\rho(q+1)}{\rho + q} \right), \tag{2.22}$$

where $q > 0$ is a penalty parameter, so that $\lim_{q \to \infty} \alpha(\rho) = \bar{\alpha}(1 - \rho)$.

Remark 2.1. *The inverse permeability term α can be interpreted as a fictitious permeability term as it appears in the Stokes–Brinkman equations. The Stokes–Brinkman equations model a slowly moving Newtonian fluid flowing through porous media and are given by*

$$\nu \mathbf{K}^{-1}\boldsymbol{u} - \nu\Delta\boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } \Omega, \tag{2.23}$$

$$\operatorname{div}(\boldsymbol{u}) = 0 \text{ in } \Omega, \tag{2.24}$$

$$\boldsymbol{u}|_{\partial\Omega} = \boldsymbol{g} \text{ on } \partial\Omega, \tag{2.25}$$

where $\boldsymbol{u}$ and p are the fluid velocity and pressure, respectively, and $\mathbf{K}$ is a permeability tensor allowed to vary over the spatial domain, Ω. In the Borrvall–Petersson case, we can equate $\nu\mathbf{K}^{-1} = \alpha(\rho)\mathbf{I}$. A rough argument (assuming that $\boldsymbol{u}$, ρ, and p are continuous and $\bar{\alpha} \gg 1$), gives the following: for $x \in \Omega$ where

$$\rho(x) = 1,\ (2.23) \approx -\nu\Delta\boldsymbol{u}(x) + \nabla p(x) = \boldsymbol{f}(x) \implies \text{Stokes momentum equation},$$

$$\rho(x) = 0,\ (2.23) \approx \bar{\alpha}\boldsymbol{u}(x) = \boldsymbol{f}(x) \implies \boldsymbol{u}(x) \approx \mathbf{0}.$$

To summarize, in regions where $\rho = 1$, (2.23) reduces to the standard Stokes momentum equation and where $\rho = 0$, the velocity is approximately zero.

Remark 2.2. *At first glance, the optimization problem (BP) lacks a PDE constraint $F : \boldsymbol{U} \times C_\gamma \to X$. Physically, the velocity $\boldsymbol{u}$ satisfies a generalized momentum equation and an incompressibility constraint. The incompressibility constraint is incorporated into the solution space $\boldsymbol{U}$ and, therefore, is not explicitly stated. In Proposition 2.4, we will show that isolated minimizers of (BP) automatically satisfy a weak form of the generalized Stokes momentum equation formulated by Borrvall and Petersson [36, Eq. 12]. Hence, it does not need to be enforced as a separate constraint. In the case where the fluid satisfies a different fluid momentum equation such as a Navier–Stokes momentum equation, the same power dissipation functional (BP) is minimized and an alternative momentum equation must be added as the PDE constraint.*

The objective functional (BP) can be interpreted as the total potential power of the flow. The first and second terms in the integral measure the energy lost by the flow through the porous medium and the energy lost due to viscous dissipation, respectively. The third term attempts to maximize the flow velocities at the applied body force. (BP) is discussed in further detail by Borrvall and Petersson [36].

Remark 2.3. *The integral in* (BP) *is well defined. Indeed, since* α *is assumed to be convex, it is Borel measurable; also since* $\rho \in C_\gamma$ *is Lebesgue measurable, the composition* $\alpha(\rho) : \Omega \to [\underline{\alpha}, \overline{\alpha}]$ *is Lebesgue measurable.*

The following existence theorem is due to Borrvall and Petersson [36, Th. 3.1].

Theorem 2.11. *Suppose that* $\Omega \subset \mathbb{R}^d$ *is a Lipschitz domain, with* $d \in \{2, 3\}$*, and* α *is continuously differentiable and satisfies properties (A1)–(A3). Then, there exists a pair* $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ *that minimizes* J *(as defined in* (BP)*).*

Due to the lack of strict convexity in (BP), a minimizing pair is not necessarily unique. The remainder of this subsection is concerned with the rigorous analysis of the minimizers of (BP).

In the results that follow, we are required to distinguish between different types of local and global minimizers.

Definition 2.1 (Isolated minimizer)**.** *Let* Z *be a Banach space and suppose that the function* $z_0 \in Z$ *is a local or global minimizer of the functional* $J : Z \to \mathbb{R}$*. We say that* z_0 *is isolated if there exists an open neighborhood* $E \subset Z$ *of* z_0 *such that there are no other minimizers contained in* E*.*

First we consider the relationship between an isolated minimizer $(\boldsymbol{u}, \rho)$ and the volume constraint. The volume constraint is typically modeled as an inequality constraint. However, as we show below, this constraint is active at an optimal solution. To the best of our knowledge, the following result is novel.

Proposition 2.1. *If the pair* $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ *is an isolated local or global minimizer of* J *as defined in* (BP) *and* $\gamma \in (0, 1)$*, then,* $\int_\Omega \rho \, \mathrm{d}x = \gamma |\Omega|$*.*

Proof by contradiction. Suppose there exists a pair $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that is an isolated local or global minimizer of $J(\boldsymbol{u}, \rho)$ such that $V := \int_\Omega \rho \, \mathrm{d}x < \gamma|\Omega|$. By the definition of an isolated minimizer, there exists an $r > 0$ such that for any $(\boldsymbol{w}, \eta)$ that satisfies,

$$\|\boldsymbol{u} - \boldsymbol{w}\|_{H^1(\Omega)} + \|\rho - \eta\|_{L^\infty(\Omega)} \leq r$$

then $J(\boldsymbol{u}, \rho) < J(\boldsymbol{w}, \eta)$. Then, for any function $\delta\rho \in C_\gamma$ such that

$$0 < \|\delta\rho\|_{L^1(\Omega)} \leq (\gamma|\Omega| - V), \tag{2.26}$$

$$0 < \|\delta\rho\|_{L^\infty(\Omega)} \leq r, \tag{2.27}$$

$$0 \leq \rho + \delta\rho \leq 1, \tag{2.28}$$

we have that $\rho + \delta\rho \in C_\gamma$ from (2.26) and (2.28) and $\rho + \delta\rho$ lies in the L^∞-r-neighborhood of ρ from (2.27). Such a $\delta\rho$ exists, for example

$$\delta\rho = c(1 - \rho), \quad \text{where } c = \min\left\{ \frac{r}{\|1 - \rho\|_{L^\infty(\Omega)}}, \frac{\gamma|\Omega| - V}{|\Omega| - V} \right\}.$$

We see that $c > 0$ since $r > 0$ and $V < \gamma|\Omega| < |\Omega|$. Furthermore $\delta\rho$ satisfies (2.26)–(2.28) since,

$$\begin{aligned} \|\delta\rho\|_{L^1(\Omega)} &= c \int_\Omega (1 - \rho) \mathrm{d}x \leq c(|\Omega| - V) \leq \gamma|\Omega| - V, \\ \|\delta\rho\|_{L^\infty(\Omega)} &\leq c\|1 - \rho\|_{L^\infty(\Omega)} \leq r, \\ 0 \leq \rho + \delta\rho &= \rho + c(1 - \rho) \leq \rho + 1 - \rho \leq 1. \end{aligned}$$

Since $\alpha(\cdot)$ is monotonically decreasing and ρ and $\delta\rho$ are non-negative and not equal to zero, then $\alpha(\rho + \delta\rho) \leq \alpha(\rho)$ a.e. and hence $J(\boldsymbol{u}, \rho + \delta\rho) \leq J(\boldsymbol{u}, \rho)$. As $\delta\rho \neq 0$, this contradicts the assumption that $(\boldsymbol{u}, \rho)$ is an isolated local or global minimizer. □

2.4.2 Support of ρ

The following lemma will be used in the proof of the next proposition.

Lemma 2.1. *Consider a nonzero function $\eta \in C_\gamma$ and the measurable non-empty set $E \subset\subset \mathrm{supp}(\eta)$, where* supp *denotes the support of a function, i.e. $\eta > 0$ a.e. in E. Then, there exists an $\epsilon' > 0$ such that, for all $\epsilon \in (0, \epsilon']$, there exists a set $E_\epsilon \subseteq E$, $|E_\epsilon| > 0$, where $\eta > \epsilon$ a.e. in E_ϵ.*

Proof. For a contradiction, suppose that there exists no such ϵ' such that $E_{\epsilon'}$ exists. This implies that

$$\text{for all } \; n \geq 0, \;\; |E \backslash \hat{E}_n| = 0, \tag{2.29}$$

where $\hat{E}_n := \{0 \leq \eta \leq 1/n \;\; \text{a.e. in} \;\; E\}$. We see that $\varnothing = E\backslash\hat{E}_1 \subseteq E\backslash\hat{E}_2 \subseteq \cdots \subseteq E\backslash\hat{E}_n \subseteq \cdots$, i.e. $E\backslash\hat{E}_n$ is ascending. By (2.29) we note that

$$\lim_{n\to\infty} |E\backslash\hat{E}_n| = 0. \tag{2.30}$$

Moreover,

$$\begin{aligned} \cup_{n=1}^{\infty} E\backslash\hat{E}_n &= \lim_{n\to\infty} E\backslash\{0 \leq \eta \leq 1/n \;\; \text{a.e. in} \;\; E\} \\ &= E\backslash\{\eta = 0 \;\; \text{a.e. in} \;\; E\} = E\backslash\varnothing = E. \end{aligned} \tag{2.31}$$

Now we see that

$$0 < |E| = |\cup_{n=1}^{\infty} E\backslash\hat{E}_n| = |\lim_{n\to\infty} E\backslash\hat{E}_n| = \lim_{n\to\infty} |E\backslash\hat{E}_n| = 0, \tag{2.32}$$

where the first equality follows from (2.31), the third equality follows from the continuity of the Lebesgue measure, and the fourth equality follows from (2.30). (2.32) is a contradiction and, therefore, such an $\epsilon' > 0$ must exist. By choosing $E_\epsilon = E_{\epsilon'}$ for all $0 < \epsilon \leq \epsilon'$, we conclude that the statement holds for all $\epsilon \in (0, \epsilon']$. □

The following result is novel. We show that if $(\boldsymbol{u}, \rho)$ is an isolated minimizer of (BP), then, the support of ρ is contained in the support of $\boldsymbol{u}$. This result will be useful for the numerical analysis of the finite element discretization in Chapter 3.

Proposition 2.2 (Support of ρ)**.** *Suppose that $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, with $d \in \{2, 3\}$, and α satisfies properties (A1)–(A4). Further assume that the minimizer $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ of (BP) is isolated. Then,* $\text{supp}(\rho) \subseteq U$*, where* $U := \text{supp}(\boldsymbol{u})$.

Proof. By definition of an isolated minimizer, there exists an $r > 0$ such that, for all $(\boldsymbol{w}, \eta) \in \boldsymbol{U} \times C_\gamma$, $(\boldsymbol{w}, \eta) \neq (\boldsymbol{u}, \rho)$ that satisfies

$$\|\boldsymbol{u} - \boldsymbol{w}\|_{H^1(\Omega)} + \|\rho - \eta\|_{L^\infty(\Omega)} \leq r,$$

we have that $J(\boldsymbol{u},\rho) < J(\boldsymbol{w},\eta)$. For a contradiction, suppose that there exists a set $E \subset \Omega$, $E \cap U = \varnothing$, of positive measure, where $\rho > 0$ a.e. in E. By Lemma 2.1, there exists an $\epsilon \in (0, r)$ such that there exists a set $E_\epsilon \subseteq E$, $|E_\epsilon| > 0$ where $\rho > \epsilon$ a.e. in E_ϵ. Define $\tilde{\rho}$ as

$$\tilde{\rho} := \begin{cases} \rho & \text{a.e. in } \Omega \backslash E_\epsilon, \\ \rho - \epsilon & \text{a.e. in } E_\epsilon. \end{cases} \tag{2.33}$$

As $\rho \in C_\gamma$, also $\tilde{\rho} \in C_\gamma$. We note that $\|\rho - \tilde{\rho}\|_{L^\infty(\Omega)} = \|\epsilon\|_{L^\infty(E_\epsilon)} < r$ and, therefore, $(\boldsymbol{u}, \tilde{\rho})$ lies inside the minimizing neighborhood of the $(\boldsymbol{u}, \rho)$. However, $J(\boldsymbol{u}, \tilde{\rho}) = J(\boldsymbol{u}, \rho)$ as ρ and $\tilde{\rho}$ only differ on the set E_ϵ, but $\boldsymbol{u} = \boldsymbol{0}$ a.e. in $E_\epsilon \subseteq E$ by assumption. This contradicts the assertion that $(\boldsymbol{u}, \rho)$ is an isolated minimizer. □

2.4.3 First-order optimality conditions

In this subsection, we show that isolated minimizers of (BP) also satisfy first-order optimality conditions. These conditions will be solved by our optimization strategy that we develop in later chapters. Moreover, they will be essential for our proofs of regularity in Section 2.4.4 and the proof of the convergence of a finite element discretization in Chapter 3.

Proposition 2.3 (Fréchet differentiability of J)**.** *Suppose that α satisfies (A1)–(A4). Then, $J : H^1(\Omega)^d \times L^s(\Omega) \to \mathbb{R}$ is Fréchet differentiable with respect to $\boldsymbol{u}$ and ρ, where $1 < s \leq \infty$ in two dimensions and $3 \leq s \leq \infty$ in three dimensions. Moreover, for all $\boldsymbol{v} \in H_0^1(\Omega)^d$ and $\eta \in C_\gamma$ we have that*

$$\langle J'_{\boldsymbol{u}}(\boldsymbol{u},\rho), \boldsymbol{v}\rangle = \int_\Omega \alpha(\rho)\boldsymbol{u}\cdot\boldsymbol{v} + \nu\nabla\boldsymbol{u} : \nabla\boldsymbol{v} - \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x, \tag{2.34}$$

$$\langle J'_{\rho}(\boldsymbol{u},\rho), \eta - \rho\rangle = \frac{1}{2}\int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2(\eta-\rho)\,\mathrm{d}x, \tag{2.35}$$

where $J'_{\boldsymbol{u}}(\boldsymbol{u},\rho)$ denotes the Fréchet deriviative of J with respect to $\boldsymbol{u}$ and $J'_{\rho}(\boldsymbol{u},\rho)$ denotes the Fréchet deriviative of J with respect to ρ.

Proof. Consider a variation $\boldsymbol{v} \in H_0^1(\Omega)^d$. Note that

$$\begin{aligned} J(\boldsymbol{u}+\boldsymbol{v},\rho) - J(\boldsymbol{u},\rho) &= \frac{1}{2}\int_\Omega \alpha(\rho)(|\boldsymbol{u}+\boldsymbol{v}|^2 - |\boldsymbol{u}|^2) + \nu(|\nabla(\boldsymbol{u}+\boldsymbol{v})|^2 - |\nabla\boldsymbol{u}|^2)\,\mathrm{d}x \\ &= \underbrace{\int_\Omega \alpha(\rho)\boldsymbol{u}\cdot\boldsymbol{v} + \nu\nabla\boldsymbol{u} : \nabla\boldsymbol{v}\,\mathrm{d}x}_{=:A\boldsymbol{v}} + \underbrace{\frac{1}{2}\int_\Omega \alpha(\rho)|\boldsymbol{v}|^2 + \nu|\nabla\boldsymbol{v}|^2\,\mathrm{d}x}_{=:R(\boldsymbol{v})}, \end{aligned}$$

where, for $(\boldsymbol{u}, \rho) \in H^1_{\boldsymbol{g},\mathrm{div}}(\Omega)^d \times C_\gamma$ fixed, A is a linear operator on $\boldsymbol{v}$ and R is a nonlinear operator on $\boldsymbol{v}$. Now we note that

$$R(\boldsymbol{v}) = |R(\boldsymbol{v})| \leq \max\{\bar{\alpha}, \nu\}\|\boldsymbol{v}\|^2_{H^1(\Omega)},$$

which implies that

$$\frac{|J(\boldsymbol{u}+\boldsymbol{v}, \rho) - J(\boldsymbol{u}, \rho) - A\boldsymbol{v}|}{\|\boldsymbol{v}\|_{H^1(\Omega)}} = \frac{|R(\boldsymbol{v})|}{\|\boldsymbol{v}\|_{H^1(\Omega)}} \to 0 \ \text{ as } \ \boldsymbol{v} \to \boldsymbol{0}.$$

This implies that J is Fréchet differentiable with respect to $\boldsymbol{u}$. Now let, for $(\boldsymbol{u}, \rho) \in H^1_{\boldsymbol{g},\mathrm{div}}(\Omega)^d \times C_\gamma$ fixed,

$$B\eta := \frac{1}{2}\int_\Omega \alpha'(\rho)\eta|\boldsymbol{u}|^2 \,\mathrm{d}x.$$

By assumption (A4), α is twice continuously differentiable and hence by two applications of the mean value theorem we see that

$$\begin{aligned} J(\boldsymbol{u}, \rho+\eta) - J(\boldsymbol{u}, \rho) - B\eta &= \frac{1}{2}\int_\Omega (\alpha(\rho+\eta) - \alpha(\rho) - \alpha'(\rho)\eta)|\boldsymbol{u}|^2 \,\mathrm{d}x \\ &= \frac{1}{2}\int_\Omega (\alpha'(\rho + c\eta) - \alpha'(\rho))\eta|\boldsymbol{u}|^2 \,\mathrm{d}x \\ &= \frac{1}{2}\int_\Omega \alpha''(\rho + cc'\eta)c\eta^2|\boldsymbol{u}|^2 \,\mathrm{d}x, \end{aligned}$$

for some $0 \leq c, c' \leq 1$. Now

$$\left|\frac{1}{2}\int_\Omega \alpha''(\rho + cc'\eta)c\eta^2|\boldsymbol{u}|^2 \,\mathrm{d}x\right| \leq \sup_{\zeta \in C_\gamma} |\alpha''(\zeta)|\|\boldsymbol{u}\|^2_{L^{2q'}(\Omega)}\|\eta\|^2_{L^{2q}(\Omega)},$$

where q' is the Hölder conjugate of q. By the Sobolev embedding theorem, the second term on the right-hand side is bounded if, in two dimensions, we have $2q' < \infty$ and, in three dimensions, we have $2q' \leq 6$. In turn this implies that $2q > 1$ in two dimensions and $2q \geq 3$ in three dimensions. By identifying $s := 2q$, we conclude that

$$\frac{|J(\boldsymbol{u}, \rho+\eta) - J(\boldsymbol{u}, \rho) - B\eta|}{\|\eta\|_{L^s(\Omega)}} \leq \sup_{\zeta \in C_\gamma} |\alpha''(\zeta)|\|\boldsymbol{u}\|^2_{L^{2q'}(\Omega)}\|\eta\|_{L^s(\Omega)} \to 0 \ \text{ as } \ \eta \to 0.$$

This implies that J is Fréchet differentiable with respect to ρ. □

Remark 2.4. *It can be checked that if α is $(n+1)$-times continuously differentiable, then J is n-times Fréchet differentiable with respect to $\boldsymbol{u}$ and ρ.*

The following result is novel. We show that if $(\boldsymbol{u}, \rho)$ is an isolated minimizer of the optimization problem (BP), then the minimizer also satisfies first-order optimality conditions consisting of two equations and a variational inequality.

Proposition 2.4 (First-order optimality conditions)**.** *Suppose that $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, with $d \in \{2, 3\}$, and α satisfies properties (A1)–(A4). By Theorem 2.11 there exists a local minimizer $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$. If the local minimizer is also an isolated minimizer, then there exists a unique Lagrange multiplier $p \in L^2_0(\Omega)$ such that the following necessary first-order optimality conditions hold:*

$$a_\rho(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) = l_{\boldsymbol{f}}(\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in H^1_0(\Omega)^d, \tag{FOC1}$$

$$b(\boldsymbol{u}, q) = 0 \qquad \text{for all } q \in L^2_0(\Omega), \tag{FOC2}$$

$$c_{\boldsymbol{u}}(\rho, \eta - \rho) \geq 0 \qquad \text{for all } \eta \in C_\gamma, \tag{FOC3a}$$

where

$$a_\rho(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega [\alpha(\rho)\boldsymbol{u} \cdot \boldsymbol{v} + \nu \nabla \boldsymbol{u} : \nabla \boldsymbol{v}] \, \mathrm{d}x, \quad l_{\boldsymbol{f}}(\boldsymbol{v}) := \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, \mathrm{d}x, \tag{2.36}$$

$$b(\boldsymbol{v}, q) := -\int_\Omega q \, \mathrm{div}(\boldsymbol{v}) \, \mathrm{d}x, \qquad c_{\boldsymbol{u}}(\rho, \eta) := \frac{1}{2} \int_\Omega \alpha'(\rho)\eta |\boldsymbol{u}|^2 \, \mathrm{d}x. \tag{2.37}$$

Proof. We will first show that (FOC1)–(FOC2) are satisfied by generalizing arguments, used for the Stokes system with a homogeneous Dirichlet boundary condition, found in [119]. For ease of notation we define $\boldsymbol{X}_{\boldsymbol{g}} := H^1_{|\partial\Omega, \boldsymbol{g}}(\Omega)^d$, $\boldsymbol{X}_0 := H^1_0(\Omega)^d$, $\boldsymbol{U}_0 := H^1_{\mathrm{div}}(\Omega)^d \cap \boldsymbol{X}_0$ and $M := L^2_0(\Omega)$. The respective dual spaces of $\boldsymbol{X}_0$, $\boldsymbol{U}_0$ and M are denoted with *. We also define the associated operators, $A \in \mathcal{L}(\boldsymbol{X}_{\boldsymbol{g}}, \boldsymbol{X}_0^*)$, $B \in \mathcal{L}(\boldsymbol{X}_{\boldsymbol{g}}, M)$ and $B_0 \in \mathcal{L}(\boldsymbol{X}_0, M)$ by

$$\langle A\boldsymbol{u}, \boldsymbol{v} \rangle := a_\rho(\boldsymbol{u}, \boldsymbol{v}), \quad \langle B\boldsymbol{w}, q \rangle := b(\boldsymbol{w}, q), \quad \text{and} \quad \langle B_0 \boldsymbol{v}, q \rangle := b(\boldsymbol{v}, q). \tag{2.38}$$

We note that $\ker(B_0) = \boldsymbol{U}_0$. From Theorem 2.11, we know that there exists a pair $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that is a local minimizer for (BP). For any given $\boldsymbol{v} \in \boldsymbol{U}_0$, we see that $\boldsymbol{u} + t\boldsymbol{v} \in \boldsymbol{U}$, $t \in \mathbb{R}$. If $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ is an isolated minimizer, then, by definition, there exists an $r > 0$ such that, for any $(\boldsymbol{w}, \eta) \in \boldsymbol{U} \times C_\gamma$, $(\boldsymbol{w}, \eta) \neq (\boldsymbol{u}, \rho)$ that satisfies

$$\|\boldsymbol{u} - \boldsymbol{w}\|_{H^1(\Omega)} + \|\rho - \eta\|_{L^\infty(\Omega)} \leq r \tag{2.39}$$

we have that $J(\boldsymbol{u},\rho) < J(\boldsymbol{w},\eta)$. Hence, for any given $\boldsymbol{v} \in \boldsymbol{U}_0$, if $0 < t \leq r/\|\boldsymbol{v}\|_{H^1(\Omega)}$, the following inequality holds

$$\frac{1}{t}(J(\boldsymbol{u}+t\boldsymbol{v},\rho) - J(\boldsymbol{u},\rho)) \geq 0. \tag{2.40}$$

By Proposition 2.3, J is Fréchet differentiable, and therefore also Gateaux differentiable, with respect to $\boldsymbol{u}$. Hence as $t \to 0_+$, we see that

$$\langle J'_{\boldsymbol{u}}(\boldsymbol{u},\rho), \boldsymbol{v}\rangle \geq 0 \ \text{ for all } \boldsymbol{v} \in \boldsymbol{U}_0. \tag{2.41}$$

By considering the same reasoning with $t < 0$, we deduce that

$$\langle J'_{\boldsymbol{u}}(\boldsymbol{u},\rho), \boldsymbol{v}\rangle = 0 \ \text{ for all } \boldsymbol{v} \in \boldsymbol{U}_0. \tag{2.42}$$

From Proposition 2.3, we know that $J'_{\boldsymbol{u}}(\boldsymbol{u},\rho) = A\boldsymbol{u} - \boldsymbol{f}$ and hence $A\boldsymbol{u} - \boldsymbol{f} \in \boldsymbol{U}_0^\circ$ where

$$\boldsymbol{U}_0^\circ := (\ker(B_0))^\circ = \{h \in \boldsymbol{X}_0^* : \langle h, \boldsymbol{v}\rangle = 0 \text{ for all } \boldsymbol{v} \in \boldsymbol{U}_0\}. \tag{2.43}$$

We know that the operator B_0 satisfies the following equivalent version of the inf-sup condition [82, Ch. 1, Sec. 4.1, Lem. 4.1]:

$$\text{there exists a } \beta > 0 \text{ such that, for all } q \in M,\ \|B_0^* q\|_{\boldsymbol{X}_0^*} \geq \beta \|q\|_M, \tag{2.44}$$

where B_0^* is the dual operator of B_0, defined by $\langle \boldsymbol{v}, B_0^* q\rangle = \langle B_0 \boldsymbol{v}, q\rangle$. This implies that B_0^* is injective (and therefore bijective) from M into $\mathrm{im}(B_0^*)$. Furthermore, it also implies that $(B_0^*)^{-1}$ is continuous. Consider $\boldsymbol{f} \in \mathrm{im}(B_0^*)$; then, there exists a $q \in M$ such that $\boldsymbol{f} = B_0^* q$ and

$$\|(B_0^*)^{-1}\boldsymbol{f}\|_M \leq \frac{1}{\beta}\|\boldsymbol{f}\|_{\boldsymbol{X}_0^*}. \tag{2.45}$$

Therefore, $\mathrm{im}(B_0^*)$ is closed.

Since $\mathrm{im}(B_0^*)$ is closed, by Banach's closed range theorem, we know that $\mathrm{im}(B_0^*) = (\ker(B_0))^\circ = \boldsymbol{U}_0^\circ$. Hence since, $A\boldsymbol{u} - \boldsymbol{f} \in \boldsymbol{U}_0^\circ$, there exists a $p \in M$ such that

$$A\boldsymbol{u} + B_0^* p = \boldsymbol{f}. \tag{2.46}$$

Since B_0^* is injective, p is also unique. Since $\boldsymbol{u} \in \boldsymbol{U}$, we have that $B\boldsymbol{u} = 0$. Hence (FOC1) and (FOC2) hold.

We will now show that (FOC3a) holds via a direct calculus of variations approach. We note that C_γ is a convex subset of a linear space. For any given $\zeta, \eta \in C_\gamma$ and $t \in [0,1]$, we therefore have that $\zeta + t(\eta - \zeta) \in C_\gamma$. Since $(\boldsymbol{u}, \rho)$ is a local minimizer, it follows that for each $\eta \in C_\gamma$, if $0 < t \leq r/\|\eta - \rho\|_{L^\infty(\Omega)}$, with r as in (2.39), then

$$\frac{1}{t}(J(\boldsymbol{u}, \rho + t(\eta - \rho)) - J(\boldsymbol{u}, \rho)) \geq 0. \tag{2.47}$$

From Proposition 2.3, we know that J is Fréchet differentiable, and therefore also Gateaux differentiable, with respect to ρ. Hence, by taking the limit as $t \to 0$, we see that

$$c_{\boldsymbol{u}}(\rho, \eta - \rho) = \langle J'_\rho(\boldsymbol{u}, \rho), \eta - \rho \rangle \geq 0 \ \text{ for all } \eta \in C_\gamma. \tag{2.48}$$

Therefore (FOC3a) holds. □

The variational inequality (FOC3a) only tests against functions that satisfy $\int_\Omega \eta \, \mathrm{d}x \leq \gamma|\Omega|$. This is a difficult restriction to enforce in practice. Hence, in the next proposition we show that we can expand the space of test functions at the cost of introducing a new unknown $\lambda \in \mathbb{R}$.

Proposition 2.5 (Relaxing the space of test functions)**.** *Suppose that $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, with $d \in \{2, 3\}$, α satisfies properties (A1)–(A4) and $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ is an isolated minimizer of (BP). Then, there exist unique Lagrange multipliers $p \in L^2_0(\Omega)$ and $\lambda \in \mathbb{R}$, such that, for all $(\eta, \boldsymbol{v}, q, \zeta) \in C_{[0,1]} \times H^1_0(\Omega)^d \times L^2_0(\Omega) \times \mathbb{R}$, the following necessary first-order optimality conditions are satisfied:*

$$a_\rho(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) = l_{\boldsymbol{f}}(\boldsymbol{v}), \tag{FOC1}$$

$$b(\boldsymbol{u}, q) = 0, \tag{FOC2}$$

$$c_{\boldsymbol{u},\lambda}(\rho, \eta - \rho) := \frac{1}{2}\int_\Omega (\alpha'(\rho)|\boldsymbol{u}|^2 + \lambda)(\eta - \rho)\mathrm{d}x \geq 0, \tag{FOC3b}$$

$$d_\rho(\lambda, \zeta) := -\zeta \int_\Omega (\gamma - \rho)\mathrm{d}x = 0. \tag{FOC4}$$

Here $C_{[0,1]} := \cup_{\gamma \in [0,1]} C_\gamma$, i.e. we relax the volume constraint on the variation in the material distribution.

Proof. The existence of $p \in L^2_0(\Omega)$ and the first two equations (FOC1) and (FOC2) follow from Proposition 2.4. The equation (FOC4) follows from Proposition 2.1. It remains to show that there exists a $\lambda \in \mathbb{R}$ such that the variational inequality (FOC3b) is satisfied. Consider the functional $I : C_{[0,1]} \to \mathbb{R}$ defined by

$$I(\eta) = \int_\Omega \gamma - \eta \, \mathrm{d}x. \tag{2.49}$$

Consider any two functions $\eta, \zeta \in C_{[0,1]}$ with $\|\zeta\|_{L^1(\Omega)} \neq \gamma|\Omega|$. Now for any $\tau, \sigma \in [0,1]$ consider the convex sum $\rho + \tau(1-\sigma)(\zeta - \rho) + \sigma(\eta - \rho) \in C_{[0,1]}$. Let $i(\tau, \sigma) = I(\rho + \tau(1-\sigma)(\zeta - \rho) + \sigma(\eta - \rho))$. Then,

$$\partial_\tau i(\tau, \sigma) = -\int_\Omega (1-\sigma)(\zeta - \rho) \mathrm{d}x, \tag{2.50}$$

$$\partial_\sigma i(\tau, \sigma) = -\int_\Omega (\eta - \rho) - \tau(\zeta - \rho) \mathrm{d}x. \tag{2.51}$$

The Implicit Function Theorem implies that there exists a $\phi \in C^1(\mathbb{R}; \mathbb{R})$ and $\sigma_0 > 0$ such that $\phi(0) = 0$ and $i(\phi(\sigma), \sigma) = 0$ for all $\sigma < \sigma_0$. Hence, by differentiating $i(\phi(\sigma), \sigma)$ with respect to σ, we see that

$$\partial_\tau i(\phi(\sigma), \sigma)\phi'(\sigma) + \partial_\sigma i(\phi(\sigma), \sigma) = 0 \quad \text{for all} \quad \sigma < \sigma_0. \tag{2.52}$$

Therefore, by considering $\sigma = 0$ and rearranging, we see that

$$\phi'(0) = -\frac{\int_\Omega \eta - \rho \, \mathrm{d}x}{\int_\Omega \zeta - \rho \, \mathrm{d}x}. \tag{2.53}$$

Since, by assumption $\|\zeta\|_{L^1(\Omega)} \neq \gamma|\Omega|$, then $\int_\Omega \zeta - \rho \, \mathrm{d}x \neq 0$ and, therefore, (2.53) is finite. Now let $w(\sigma) := \phi(\sigma)(1-\sigma)(\zeta - \rho) + \sigma(\eta - \rho)$. For all $\sigma < \sigma_0$, we have that $i(\phi(\sigma), \sigma) = 0$, which implies that $\rho + w(\sigma) \in C_\gamma$ for all $\sigma \in [0, \sigma_0)$. Let $j(\sigma) = J(\boldsymbol{u}, \rho + w(\sigma))$. Then, since $(\boldsymbol{u}, \rho)$ is an isolated minimizer, we have that

$$0 \leq \lim_{\sigma \to 0} \frac{1}{\sigma}(J(\boldsymbol{u}, \rho + w(\sigma)) - J(\boldsymbol{u}, \rho)) = j'(0). \tag{2.54}$$

By computing $j'(0)$ we see that

$$j'(0) = c_{\boldsymbol{u},0}(\rho, \eta - \rho + \phi'(0)(\zeta - \rho)) \geq 0 \tag{2.55}$$

By choosing

$$\lambda = -\frac{c_{\boldsymbol{u},0}(\rho, \zeta - \rho)}{\int_\Omega \zeta - \rho \, \mathrm{d}x}, \tag{2.56}$$

in (FOC3b), it follows from (2.55) that the variational inequality (FOC3b) holds. □

2.4.4 Regularity

In this subsection we prove novel regularity results for the velocity and pressure terms and a novel and surprising regularity result for the material distribution. We show that in a convex domain, and under sufficiently smooth data, if $(\boldsymbol{u}, \rho)$ is an isolated minimizer of (BP), then $\boldsymbol{u} \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$ where p is the Lagrange multiplier associated with $(\boldsymbol{u}, \rho)$. We also show that (without a convex domain assumption) if the inverse permeability α is strongly convex (which is the case for all choices found in literature), and under a homogeneous Dirichlet boundary condition on $\boldsymbol{u}$, we have that ρ is weakly differentiable inside any compact subset of the support of $\boldsymbol{u}$. Hence, although the aim of the topology optimization formulation is to recover a 0-1 material distribution function, we see that the transitions in the material distribution are necessarily not jumps as weakly differentiable functions cannot jump.

Proposition 2.6 (Regularity of $\boldsymbol{u}$ and p)**.** *Let the domain Ω be either a convex polygon in two dimensions or a convex polyhedron in three dimensions and consider the triple $(\boldsymbol{u}, \rho, p) \in \boldsymbol{U} \times C_\gamma \times L^2_0(\Omega)$ that satisfies (FOC1)–(FOC3a). Suppose that the forcing term $\boldsymbol{f} \in L^2(\Omega)^d$ and the boundary datum $\boldsymbol{g}$ is the boundary trace of a function $\hat{\boldsymbol{g}} \in H^2(\Omega)^d$ on the boundary $\partial\Omega$, where $\operatorname{div}(\hat{\boldsymbol{g}}) = 0$ a.e. in Ω. Then, $\boldsymbol{u} \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$.*

Proof. The idea of the proof is to reduce the system (FOC1)–(FOC2) to the standard Stokes system with a homogeneous Dirichlet boundary condition and then invoke the regularity results of Kellogg and Osborn [97] in two dimensions and Kozlov et al. [102] in three dimensions.

Let $\boldsymbol{w} := \boldsymbol{u} - \hat{\boldsymbol{g}}$. Since the trace operator is a linear operator, we see that $\boldsymbol{w}|_{\partial\Omega} = (\boldsymbol{u} - \hat{\boldsymbol{g}})|_{\partial\Omega} = \boldsymbol{g} - \boldsymbol{g} = \boldsymbol{0}$. Since $\boldsymbol{u} \in H^1(\Omega)^d$ and $\hat{\boldsymbol{g}} \in H^2(\Omega)^d$, then $\boldsymbol{w} \in H^1_0(\Omega)^d$.

By substituting $\boldsymbol{w}$ into (FOC1)–(FOC2), we see that (FOC1)–(FOC2) is equivalent to finding $(\boldsymbol{w}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ that satisfies for all $(\boldsymbol{v}, q) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$:

$$\int_\Omega \nabla \boldsymbol{w} : \nabla \boldsymbol{v} - p \operatorname{div}(\boldsymbol{v}) \, \mathrm{d}x = \int_\Omega (\boldsymbol{f} - \alpha(\rho)(\boldsymbol{w} + \hat{\boldsymbol{g}})) \cdot \boldsymbol{v} - \nabla \hat{\boldsymbol{g}} : \nabla \boldsymbol{v} \, \mathrm{d}x, \tag{2.57}$$

$$\int_\Omega q \operatorname{div}(\boldsymbol{w} + \hat{\boldsymbol{g}}) \, \mathrm{d}x = 0. \tag{2.58}$$

Define $\hat{\boldsymbol{f}}$ as $\hat{\boldsymbol{f}} := \boldsymbol{f} - \alpha(\rho)(\boldsymbol{w} + \hat{\boldsymbol{g}}) + \Delta \hat{\boldsymbol{g}}$. Since $\boldsymbol{f} \in L^2(\Omega)^d$, $\alpha(\rho) \in L^\infty(\Omega)$, $\boldsymbol{w} \in H_0^1(\Omega)^d$, and $\hat{\boldsymbol{g}} \in H^2(\Omega)^d$, then $\hat{\boldsymbol{f}} \in L^2(\Omega)^d$. By an application of integration by parts on the final term on the right-hand side of (2.57), and noting that $\operatorname{div}(\hat{\boldsymbol{g}}) = 0$ a.e. in Ω, we see that (2.57)–(2.58) is equivalent to finding $(\boldsymbol{w}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ that satisfies for all $(\boldsymbol{v}, q) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$:

$$\int_\Omega \nabla \boldsymbol{w} : \nabla \boldsymbol{v} - p \operatorname{div}(\boldsymbol{v}) \, \mathrm{d}x = \int_\Omega \hat{\boldsymbol{f}} \cdot \boldsymbol{v} \, \mathrm{d}x, \tag{2.59}$$

$$\int_\Omega q \operatorname{div}(\boldsymbol{w}) \, \mathrm{d}x = 0. \tag{2.60}$$

We note that (2.59)–(2.60) is the standard Stokes system with a homogeneous Dirichlet boundary condition and forcing term $\hat{\boldsymbol{f}} \in L^2(\Omega)^d$. Therefore, by the elliptic regularity of the Stokes system [97, 102], $\boldsymbol{w} \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$. Since $\boldsymbol{u} = \boldsymbol{w} + \hat{\boldsymbol{g}}$ and $\hat{\boldsymbol{g}} \in H^2(\Omega)^d$, we conclude that $\boldsymbol{u} \in H^2(\Omega)^d$. □

The following result concerning the regularity of the material distribution is novel.

Theorem 2.12 (Regularity of ρ)**.** *Suppose that the domain $\Omega \subset \mathbb{R}^d$ is bounded, the boundary is Lipschitz, and that the datum $\boldsymbol{g} = \boldsymbol{0}$ on $\partial\Omega$. Consider a local or global minimizer, $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$, of (BP) such that $\boldsymbol{u}$ is not the zero function and there exists a closed subset $U_\theta \subset \Omega$ with non-empty interior on which $|\boldsymbol{u}|^2$ is bounded below by a positive constant, $\theta > 0$. Suppose that (A1)–(A4) hold and that $\alpha \in C^2([0,1])$ is strongly convex, i.e.,*

(A5) There exists a constant $\alpha''_{\min} > 0$ such that $\alpha''(y) \geq \alpha''_{\min} > 0$ for all $y \in [0,1]$.

Then, $\nabla \rho$ exists in U_θ and $\rho \in C_\gamma \cap H^1(U_\theta)$.

Remark 2.5. *The statement $\rho \in H^1(U_\theta)$, for any $\theta > 0$, implies that the material distribution lives in H^1 in any compact subset of the support of the velocity.*

The assumption (A5) excludes the case where α is linear. This is consistent with previous theory, as Borrvall and Petersson [36, Sec. 3.2] showed that if α is linear then ρ is necessarily a 0-1 solution (a linear combination of Heaviside functions) and thus $\rho \notin H^1(\Omega)$ due to the jumps. However, the assumptions (A1)–(A5) do include the choice (2.22), where the lower bound in (A5) is $\alpha''_{\min} = 2\bar{\alpha}q/(q+1)^2$. We see that this lower bound degrades to zero as $q \to \infty$. As previously noted, the limit $q \to \infty$ coincides with $\alpha(\rho) \to \bar{\alpha}(1-\rho)$, which is a linear function.

Proof of Theorem 2.12. Let ∂_{x_k} denote the partial derivative with respect to x_k. If we can bound the L^2-norm of the difference quotients of ρ, in all coordinate directions in U_θ, above by constants independent of h, then, by taking the weak limit, we can deduce that $\partial_{x_k}\rho$ exists as an element of $L^2(U_\theta)$ for $1 \leq k \leq d$.

We define $U \subset \Omega$ as $U := \operatorname{supp}(\boldsymbol{u})$ and fix an open, bounded and connected domain $\hat{\Omega}$ such that $\hat{\Omega} = \Omega$ if $U \subset\subset \Omega$ and $\Omega \subset\subset \hat{\Omega}$ otherwise. In the case where U is not a compact subset of Ω, we extend $\boldsymbol{u}$ and ρ by zero to the whole of $\mathbb{R}^d$. Since the trace of $\boldsymbol{u}$ is zero on the boundary, the extension of $\boldsymbol{u}$ by zero lives in $H^1(\hat{\Omega})^d$. Let $0 < |h| < (1/2)\operatorname{dist}(U, \partial\hat{\Omega})$ and choose $k \in \{1, \dots, d\}$. We define ρ^h as

$$\rho^h(x) = \begin{cases} \rho(x + he_k) & \text{for } x \in \Omega - he_k, \\ 0 & \text{for } x \in \mathbb{R}^d \backslash (\Omega - he_k). \end{cases}$$

We define the difference quotient, D_k^h, in the k-th coordinate direction, as

$$D_k^h \rho(x) = \frac{\rho(x + he_k) - \rho(x)}{h}, \quad h \in \mathbb{R}\backslash\{0\},\ x \in \hat{\Omega}.$$

Let $\eta = (\rho^h + \rho^{-h})/2$. We note that $\eta \in C_\gamma$, since

$$0 \leq \frac{1}{2}\rho^h \leq \frac{1}{2} \text{ a.e. in } \Omega, \quad \text{and} \quad 0 \leq \frac{1}{2}\rho^{-h} \leq \frac{1}{2} \text{ a.e. in } \Omega,$$

which implies that $0 \leq \eta \leq 1$ a.e. in Ω and

$$\begin{aligned} \int_\Omega \eta \, \mathrm{d}x &= \frac{1}{2}\int_\Omega \rho^h + \rho^{-h} \, \mathrm{d}x \\ &= \frac{1}{2}\int_{\Omega - he_k \cap \Omega} \rho \, \mathrm{d}x + \frac{1}{2}\int_{\Omega + he_k \cap \Omega} \rho \, \mathrm{d}x \leq \int_\Omega \rho \, \mathrm{d}x \leq \gamma|\Omega|. \end{aligned}$$

If we multiply (FOC3a) through by 4 and divide by h^2 we see that

$$\frac{1}{h^2}\int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2(\rho^h+\rho^{-h}-2\rho)\,\mathrm{d}x \geq 0. \tag{2.61}$$

We note that,

$$D_k^{-h}(D_k^h\rho) = \frac{\frac{\rho-\rho^{-h}}{h} - \frac{\rho^h-\rho}{h}}{-h} = \frac{\rho^h+\rho^{-h}-2\rho}{h^2}.$$

Hence, because $\boldsymbol{u}$ is zero outside of Ω, (2.61) is equivalent to

$$\int_{\hat{\Omega}} \alpha'(\rho)|\boldsymbol{u}|^2(D_k^{-h}(D_k^h\rho))\,\mathrm{d}x \geq 0. \tag{2.62}$$

In order to obtain a first-order difference quotient, we will perform the finite difference analogue of integration by parts to shift the D_k^{-h} operator from $D_k^h\rho$ to $\alpha'(\rho)|\boldsymbol{u}|^2$. We note that, by definition, the left-hand side of (2.62) is equal to

$$-\frac{1}{h}\int_{\hat{\Omega}}(\alpha'(\rho)|\boldsymbol{u}|^2)(x)\left((D_k^h\rho)(x-he_k)-(D_k^h\rho)(x)\right)\mathrm{d}x, \tag{2.63}$$

which by a change of variables is equal to

$$-\frac{1}{h}\left(\int_{\hat{\Omega}-he_k}(\alpha'(\rho)|\boldsymbol{u}|^2)(x+he_k)(D_k^h\rho)(x)\mathrm{d}x - \int_{\hat{\Omega}}(\alpha'(\rho)|\boldsymbol{u}|^2)(x)(D_k^h\rho)(x)\mathrm{d}x\right).$$

We note that $U \subset\subset \hat{\Omega}$ and $|h| < (1/2)\mathrm{dist}(U,\partial\hat{\Omega})$, which implies that $U \subset\subset \hat{\Omega}-he_k$. Therefore,

$$\begin{aligned}&\int_{\hat{\Omega}-he_k}(\alpha'(\rho)|\boldsymbol{u}|^2)(x+he_k)(D_k^h\rho)(x)\mathrm{d}x\\ &\quad=\int_{U-he_k}(\alpha'(\rho)|\boldsymbol{u}|^2)(x+he_k)(D_k^h\rho)(x)\mathrm{d}x\\ &\quad=\int_{\hat{\Omega}}(\alpha'(\rho)|\boldsymbol{u}|^2)(x+he_k)(D_k^h\rho)(x)\mathrm{d}x.\end{aligned} \tag{2.64}$$

Therefore, from (2.62)–(2.64) we see that

$$\int_{\hat{\Omega}} D_k^h(\alpha'(\rho)|\boldsymbol{u}|^2)(D_k^h\rho)\,\mathrm{d}x \leq 0. \tag{2.65}$$

Now we wish to rewrite $D_k^h(\alpha'(\rho)|\boldsymbol{u}|^2)$ in a form that we can decouple from $D_k^h\rho$ in order to be able to bound (2.65) above and below. Now,

$$\begin{aligned}
D_k^h(\alpha'(\rho)|\boldsymbol{u}|^2)(x) &= \frac{1}{h}\left(\alpha'(\rho(x+he_k))|\boldsymbol{u}(x+he_k)|^2 - \alpha'(\rho(x))|\boldsymbol{u}(x)|^2\right)\\
&= \frac{1}{2h}\left(\alpha'(\rho(x+he_k))\left(|\boldsymbol{u}(x+he_k)|^2 - |\boldsymbol{u}(x)|^2\right)\right)\\
&\quad + \frac{1}{2h}\left(\alpha'(\rho(x))\left(|\boldsymbol{u}(x+he_k)|^2 - |\boldsymbol{u}(x)|^2\right)\right)\\
&\quad + \frac{1}{2h}\left(|\boldsymbol{u}(x+he_k)|^2\left(\alpha'(\rho(x+he_k)) - \alpha'(\rho(x))\right)\right)\\
&\quad + \frac{1}{2h}\left(|\boldsymbol{u}(x)|^2\left(\alpha'(\rho(x+he_k)) - \alpha'(\rho(x))\right)\right)\\
&= \frac{1}{2}\left(\alpha'(\rho^h) + \alpha'(\rho)\right)D_k^h(|\boldsymbol{u}|^2) + \frac{1}{2}\left(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2\right)D_k^h(\alpha'(\rho)).
\end{aligned}$$

Therefore, from (2.65) we see that

$$\begin{aligned}
\int_{\hat{\Omega}}\Big[&\frac{1}{2}\left(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2\right)D_k^h(\alpha'(\rho))\\
&+\frac{1}{2}\left(\alpha'(\rho^h) + \alpha'(\rho)\right)D_k^h|\boldsymbol{u}|^2\Big]D_k^h(\rho)\,\mathrm{d}x \le 0.
\end{aligned} \tag{2.66}$$

Now,

$$\begin{aligned}
&\frac{1}{2}\left(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2\right)D_k^h(\alpha'(\rho)) + \frac{1}{2}\left(\alpha'(\rho^h) + \alpha'(\rho)\right)D_k^h|\boldsymbol{u}|^2\\
&\quad= \frac{1}{h}\int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\Big[\alpha'\left(s\rho^h + (1-s)\rho\right)\frac{1}{2}\left(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2\right)\\
&\qquad+ \frac{1}{2}\left(\alpha'(\rho^h) + \alpha'(\rho)\right)\left|s\boldsymbol{u}^h + (1-s)\boldsymbol{u}\right|^2\Big]\,\mathrm{d}s\\
&\quad= \frac{1}{h}\underbrace{\int_0^1\left[\alpha''\left(s\rho^h + (1-s)\rho\right)\right]\mathrm{d}s}_{=:A}\frac{1}{2}\left(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2\right)(\rho^h - \rho)\\
&\qquad+ \frac{1}{2h}\left(\alpha'(\rho^h) + \alpha'(\rho)\right)\underbrace{\int_0^1\left[2\left(s\boldsymbol{u}^h + (1-s)\boldsymbol{u}\right)\right]\mathrm{d}s}_{=:\boldsymbol{B}}\cdot(\boldsymbol{u}^h - \boldsymbol{u}).
\end{aligned}$$

Hence, from (2.66) we find that

$$\frac{1}{2}\int_{\hat{\Omega}} A(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2)|D_k^h\rho|^2 + (\alpha'(\rho^h) + \alpha'(\rho))\boldsymbol{B}\cdot(D_k^h\boldsymbol{u})D_k^h\rho\,\mathrm{d}x \le 0. \tag{2.67}$$

Subtracting the second term on the left-hand side in (2.67) from both sides, taking absolute values on the right-hand side, using the Cauchy–Schwarz inequality and multiplying by 2, we see that

$$\int_{\hat{\Omega}} A(|\boldsymbol{u}^h|^2 + |\boldsymbol{u}|^2)|D_k^h\rho|^2\mathrm{d}x \le \int_{\hat{\Omega}}|\boldsymbol{B}||\alpha'(\rho^h) + \alpha'(\rho)||D_k^h\boldsymbol{u}||D_k^h\rho|\mathrm{d}x. \tag{2.68}$$

Furthermore we note that $A \geq \alpha''_{\min}$ and

$$\boldsymbol{B} = \int_0^1 \left[2\left(s\boldsymbol{u}^h + (1-s)\boldsymbol{u}\right)\right] \mathrm{d}s = 2\left[\frac{s^2}{2}\boldsymbol{u}^h + \left(s - \frac{s^2}{2}\right)\boldsymbol{u}\right]_0^1 = \boldsymbol{u}^h + \boldsymbol{u}.$$

Hence, using Cauchy's inequality and Young's inequality, we see that

$$\begin{aligned}
&\alpha''_{\min}\int_{U-he_k} |\boldsymbol{u}^h|^2 |D_k^h\rho|^2 \mathrm{d}x + \alpha''_{\min}\int_U |\boldsymbol{u}|^2 |D_k^h\rho|^2 \mathrm{d}x \\
&\leq \int_{\hat{\Omega}} |\boldsymbol{u}+\boldsymbol{u}^h||\alpha'(\rho^h)+\alpha'(\rho)||D_k^h\boldsymbol{u}||D_k^h\rho| \mathrm{d}x \\
&\leq \int_{U-he_k} |\boldsymbol{u}^h||\alpha'(\rho^h)+\alpha'(\rho)||D_k^h\boldsymbol{u}||D_k^h\rho| \mathrm{d}x \\
&\quad + \int_U |\boldsymbol{u}||\alpha'(\rho^h)+\alpha'(\rho)||D_k^h\boldsymbol{u}||D_k^h\rho| \mathrm{d}x \\
&\leq \frac{\epsilon}{2}\int_{U-he_k} |\boldsymbol{u}^h|^2 |D_k^h\rho|^2 \mathrm{d}x + \frac{\epsilon}{2}\int_U |\boldsymbol{u}|^2 |D_k^h\rho|^2 \mathrm{d}x \\
&\quad + \frac{1}{2\epsilon}\int_{U-he_k} |\alpha'(\rho^h)+\alpha'(\rho)|^2 |D_k^h\boldsymbol{u}|^2 \mathrm{d}x \\
&\quad + \frac{1}{2\epsilon}\int_U |\alpha'(\rho^h)+\alpha'(\rho)|^2 |D_k^h\boldsymbol{u}|^2 \mathrm{d}x.
\end{aligned} \tag{2.69}$$

By fixing $\epsilon = \alpha''_{\min}$, from (2.69) we see that,

$$\begin{aligned}
&\frac{\alpha''_{\min}}{2}\int_U |\boldsymbol{u}|^2 |D_k^h\rho|^2 \mathrm{d}x \\
&\leq \frac{\alpha''_{\min}}{2}\int_U |\boldsymbol{u}|^2 |D_k^h\rho|^2 \mathrm{d}x + \frac{\alpha''_{\min}}{2}\int_{U-he_k} |\boldsymbol{u}^h|^2 |D_k^h\rho|^2 \mathrm{d}x \\
&\leq \frac{1}{\alpha''_{\min}}\int_{\hat{\Omega}} |\alpha'(\rho^h)+\alpha'(\rho)|^2 |D_k^h\boldsymbol{u}|^2 \mathrm{d}x.
\end{aligned} \tag{2.70}$$

Now $|\alpha'(\rho^h)+\alpha'(\rho)|^2$ is bounded above by $4\sup_{\zeta\in C_\gamma}|\alpha'(\zeta)|^2$ which is independent of h. Consider a set $\tilde{\Omega} \subset \mathbb{R}^d$ such that $\hat{\Omega} \subset\subset \tilde{\Omega}$. We note that $\boldsymbol{u} \in H^1(\tilde{\Omega})^d$. By applying Theorem 3 in [63, pg. 294], we see that

$$\begin{aligned}
\int_U |\boldsymbol{u}|^2 |D_k^h\rho|^2 \, \mathrm{d}x &\leq \frac{\tilde{C}(\Omega)\sup_{\zeta\in C_\gamma}|\alpha'(\zeta)|^2}{(\alpha''_{\min})^2}\|\nabla\boldsymbol{u}\|^2_{L^2(\tilde{\Omega})} \\
&\leq \frac{\hat{C}(\Omega)\sup_{\zeta\in C_\gamma}|\alpha'(\zeta)|^2}{(\alpha''_{\min})^2}\|\nabla\boldsymbol{u}\|^2_{L^2(\Omega)} \leq C < \infty,
\end{aligned} \tag{2.71}$$

where $\tilde{C}$, $\hat{C}$ and C are constants. The bound is independent of h and k. Because, by hypothesis, there exists a subset $U_\theta \subset \Omega$ such that, $|\boldsymbol{u}|^2 \geq \theta > 0$ a.e. in U_θ, we see from (2.71) that, because $U_\theta \subset U = \mathrm{supp}(\boldsymbol{u})$, also

$$\theta\int_{U_\theta} |D_k^h\rho|^2 \mathrm{d}x \leq \int_{U_\theta} |\boldsymbol{u}|^2 |D_k^h\rho|^2 \, \mathrm{d}x \leq \int_U |\boldsymbol{u}|^2 |D_k^h\rho|^2 \, \mathrm{d}x \leq C. \tag{2.72}$$

Estimate (2.72) implies that

$$\sup_h \|D_k^h \rho\|_{L^2(U_\theta)} < \infty. \tag{2.73}$$

From (2.73) we see that there exists a function $\eta_k \in L^2(U_\theta)$ and a subsequence $h_i \to 0$ such that,

$$D_k^{h_i} \rho \rightharpoonup \eta_k \text{ weakly in } L^2(U_\theta).$$

Finally, we wish to identify η_k with $\partial_{x_k}\rho$. First choose any smooth and compactly supported function, $\phi \in C_c^\infty(U_\theta)$. We note that

$$\int_{U_\theta} \rho\, \partial_{x_k}\phi \,\mathrm{d}x \le C\|\rho\|_{L^\infty(U_\theta)}\|\phi\|_{W^{1,\infty}(U_\theta)} < \infty.$$

Since $\partial_{x_k}\phi$ is compactly supported in U_θ, it follows that

$$\int_{U_\theta} \rho\, \partial_{x_k}\phi \,\mathrm{d}x = \int_{\hat{\Omega}} \rho\, \partial_{x_k}\phi \,\mathrm{d}x.$$

Hence

$$\begin{aligned}\int_{U_\theta} \rho\, \partial_{x_k}\phi \,\mathrm{d}x &= \lim_{h_i\to 0}\int_{\hat{\Omega}} \rho\, D_k^{-h_i}\phi \,\mathrm{d}x \\ &= -\lim_{h_i\to 0}\int_{\hat{\Omega}} (D_k^{h_i}\rho)\phi \,\mathrm{d}x = -\lim_{h_i\to 0}\int_{U_\theta} (D_k^{h_i}\rho)\phi \,\mathrm{d}x = -\int_{U_\theta} \eta_k\phi \,\mathrm{d}x.\end{aligned}$$

Hence $\eta_k = \partial_{x_k}\rho$ a.e. in U_θ for $k = 1, \dots, d$. Therefore, from (2.71) we see by weak lower semicontinuity that

$$\int_{U_\theta} |\partial_{x_k}\rho|^2 \,\mathrm{d}x \le C(\Omega, \sup_{\zeta\in C_\gamma} |\alpha'(\zeta)|^2, \alpha''_{\min}), \tag{2.74}$$

for some constant C. We conclude that $\rho \in H^1(U_\theta) \cap C_\gamma$, $\theta > 0$. □

Remark 2.6. *The assumption that the boundary datum $\boldsymbol{g} = \mathbf{0}$ on $\partial\Omega$ is required to expand the domain of integration from Ω in (2.61) to $\hat{\Omega}$ in (2.62) in order to perform the finite difference analogue of integration by parts in (2.63). A homogeneous Dirichlet boundary condition on $\boldsymbol{u}$ is rarely imposed in practice. However, we observe during numerical experiments that ρ possesses additional regularity in the case of inhomogeneous Dirichlet boundary conditions, and we hypothesize that the results can be generalized to that case.*

A mathematical problem does not cease being mathematical just because we have discretized it.

— Arieh Iserles, 2009

3

Numerical analysis of the Borrvall–Petersson problem

The most common discretization for topology optimization problems is the finite element method. Despite this, there is little literature that investigates the convergence of such discretizations. In particular, it is rarely known if the finite element solutions strongly converge to their respective analytical solutions. Without the certainty of strong convergence, pathological behavior, such as checkerboarding, can occur. This problem is exacerbated by the nonconvexity of the models; finite element solutions might only converge to some minimizers and not others. In this chapter, we provide the first proofs of the strong convergence of both conforming and divergence-free discontinuous Galerkin (DG) finite element methods to all isolated minimizers of the Borrvall–Petersson problem.

Prior to this work, Borrvall and Petersson's original paper [36, Sec. 3.3] contained the only known results for the convergence of a finite element approximation to the Borrvall–Petersson problem. They considered a piecewise constant finite element approximation of the material distribution coupled with an inf-sup stable conforming quadrilateral finite element approximation of the velocity and the pressure. They showed that such approximations of the velocity and material distribution converge to an unspecified solution $(\boldsymbol{u}, \rho)$ of (BP) in the following sense [36, Th. 3.2]:

$$\begin{aligned}
&\boldsymbol{u}_h \rightharpoonup \boldsymbol{u} \ \text{ weakly in } \ H^1(\Omega)^d, \\
&\rho_h \overset{*}{\rightharpoonup} \rho \ \text{ weakly-* in } \ L^\infty(\Omega), \\
&\rho_h \to \rho \ \text{ strongly in } \ L^s(\Omega_b), \quad s \in [1, \infty),
\end{aligned}$$

where Ω_b is any measurable subset of Ω where ρ is equal to zero or one a.e. Their analysis suggests that a finite element method is a suitable discretization, but it left a number of open problems:

(P1) It is not clear which minimizer the sequence is converging to, as the nonconvexity of the problem provides multiple candidates for the limits;

(P2) The convergence is weak-* in the material distribution in regions where $\{0 < \rho < 1 \text{ a.e.}\} \subset \Omega$, which permits the presence of checkerboard patterns as $h \to 0$ [36, Sec. 3];

(P3) There are no convergence results for the finite element approximation of the pressure, p.

In general (P1) means that their result does not imply that there necessarily exists a sequence of finite element solutions that converges to the global minimizer.

In this chapter, we consider two finite element methods. The first is based on any conforming mixed finite element space such that the velocity and pressure spaces are inf-sup stable. Here, conforming means that the velocity, pressure, and material distribution finite element spaces are contained in $H^1(\Omega)^d$, $L^2_0(\Omega)$ and C_γ, respectively. We prove that, for *every* isolated minimizer of (BP), there exists a sequence of finite element solutions to the discretized first-order optimality conditions that strongly converges to the minimizer, as the mesh size tends to zero. More specifically, we show that, for each isolated analytical local minimizer, there exists a sequence of finite element solutions $(\boldsymbol{u}_h, \rho_h, p_h)$ that converges to it strongly in $H^1(\Omega)^d \times L^s(\Omega) \times L^2(\Omega)$ as $h \to 0$, where $s \in [1, \infty)$. We emphasize that the results hold in the case where the minima are isolated. This analysis resolves the open problems (P1)–(P3).

Our second finite element method involves divergence-free DG finite element spaces. Here, the pressure and material distribution finite element spaces remain conforming but the velocity finite element space is no longer contained in $H^1(\Omega)^d$ but rather in $\boldsymbol{H}(\mathrm{div}; \Omega)$. A divergence-free DG approximation of the velocity allows jumps in the tangential directions across faces of elements. Therefore, we employ

an interior penalty which increasingly penalizes these jumps as the mesh size tends to zero. The resulting theorem is similar to that of the conforming finite element method. Now, the velocity approximation strongly converges in a broken H^1-norm and the sequence of finite element solutions satisfy discretized first-order optimality conditions that involve interior penalty terms.

In many studies, conforming inf-sup stable mixed finite element methods are used to discretize the Borrvall–Petersson problem. The main advantage is the (relative) ease of implementation. In the past couple of decades, discontinuous Galerkin (DG) methods for fluid flow have become increasingly popular [51, 52, 78, 99, 100]. This is in part due to the existence of divergence-free DG finite element methods. Some stable finite element methods for fluid flow, such as the Taylor–Hood finite element pair, only weakly satisfy the incompressibility constraint $\operatorname{div}(\boldsymbol{u}) = 0$. This manifests as a dependence of the error in the velocity on the best approximation error in the pressure. In some problems, only weakly satisfying the incompressibility constraint has been observed to support instabilities that result in nonphysical solutions [96, 108]. In divergence-free finite element methods, the incompressibility constraint is satisfied pointwise, which is useful for ensuring pressure robustness [145] and deriving error bounds on the velocity that are independent of the error of the pressure.

In Borrvall–Petersson topology optimization problems, a natural mesh refinement to obtain sharper solutions is in regions where $0 < \rho < 1$ a.e. It can be empirically checked that mesh refinement in these regions does not guarantee improvements in the error of the pressure approximation. If the convergence for the velocity and material distribution rely heavily on the convergence of the pressure, then only doing mesh refinement in those areas caps the improvement in the errors for the velocity and material distribution. This motivates the need for discretizations that decouple the dependence of the errors of the velocity and material distribution from the approximation error of the pressure. Crucially, for the work described in Chapter 5, divergence-free finite element discretizations also allow for an easier characterization of the kernel of the discretized grad-div term. This characterization

is key to the development of our robust preconditioner and multigrid cycle for the systems that arise in our solver developed in Chapter 4 [72, 73, 91, 143].

H^1-conforming divergence-free finite element methods exist; for example the Scott–Vogelius finite element [145]. To ensure inf-sup stability for a general mesh in a k-th order Scott–Vogelius finite element method, the polynomial order for the velocity space must be $k \geq 2d$, $d \in \{2, 3\}$ [145, 186]. The expense of the high order method is normally justified by the accompanying high convergence rate. However, the material distribution ρ is often discretized with piecewise constant or continuous piecewise linear finite elements due to the box constraints on the material distribution. The box constraints not only cause algorithmic restrictions but also reduce the regularity of ρ. The relatively low order approximation of the material distribution then caps the order of convergence of the velocity and pressure (as discussed in Section 4.6) which negates the advantage of the high order method. Inf-sup stability can be achieved for $k \geq d$ if the mesh is barycentrically refined [132]. However, barycentrically refined meshes can be difficult to align with jumps in the analytical material distribution, which can lead to poorly resolved solutions. Moreover, barycentrically refined meshes complicate the generation of a mesh hierarchy for robust multigrid cycles [72]. In contrast, there exist low-order divergence-free DG finite element methods that are inf-sup stable on general meshes.

Throughout this chapter we denote the velocity, material distribution, and pressure finite element spaces by $\boldsymbol{X}_h$, $C_{\gamma,h} \subset C_\gamma$, and $M_h \subset L^2_0(\Omega)$. When we refer to a *conforming* finite element method, we assume that $\boldsymbol{X}_h \subset H^1(\Omega)^d$, whereas for a divergence-free DG method, we assume that $\boldsymbol{X}_h \subset \boldsymbol{H}(\mathrm{div}; \Omega)$. We denote a family of triangulations of the domain Ω by $(\mathcal{T}_h)$. The family is characterized by the mesh size $h := \max_{K \in \mathcal{T}_h} h_K$, where h_K is the diameter of the element $K \in \mathcal{T}_h$. We assume that every triangulation of the family $(\mathcal{T}_h)$ satisfies:

(M1) (Shape regularity). There exist constants $c_1, c_2 > 0$ such that

$$c_1 h_K^d \leq |K| \leq c_2 h_K^d \quad \text{for all } K \in \mathcal{T}_h.$$

as well as a submesh condition as found in the work of Buffa and Ortner [46, As. 2.1]. For a given h, let the set $\mathcal{F}_h$ denote the set of all facets of the triangulation $\mathcal{T}_h$ and h_F represent the diameter of each facet $F \in \mathcal{F}_h$. We also assume the following:

(M2) (Contact regularity). There exists a constant $c_1 > 0$ such that

$$c_1 h_K^{d-1} \leq \mathcal{H}^{d-1}(F) \text{ for all } F \in \mathcal{F}_h,\ K \in \mathcal{T}_h \text{ such that } F \subset \bar{K}.$$

3.1 Conforming finite element discretization

3.1.1 Assumptions and the first convergence theorem

In this subsection we assume that $\boldsymbol{X}_h \subset H^1(\Omega)^d$. We state our assumptions for the conforming finite element methods and the first main theorem of this chapter. We define $\boldsymbol{X}_{0,h}$ as

$$\boldsymbol{X}_{0,h} := \{\boldsymbol{v}_h \in X_h : \boldsymbol{v}_h|_{\partial\Omega} = \boldsymbol{0}\}. \tag{3.1}$$

In general, it will not be possible to represent the boundary data $\boldsymbol{g}$ exactly in the velocity finite element space. Hence, for each h, we instead consider boundary data $\boldsymbol{g}_h$ (which can be represented) and assume that

(A-C1) $\boldsymbol{g}_h \to \boldsymbol{g}$ strongly in $H^{1/2}(\partial\Omega)^d$.

We now define the space $\boldsymbol{X}_{\boldsymbol{g}_h,h} := \{\boldsymbol{v}_h \in \boldsymbol{X}_h : \boldsymbol{v}_h|_{\partial\Omega} = \boldsymbol{g}_h\}$. We will also assume that:

(A-C2) $\boldsymbol{X}_{0,h}$ and M_h satisfy the following inf-sup condition for some $c_b > 0$,

$$c_b \leq \inf_{q_h \in M_h \setminus \{0\}} \sup_{\boldsymbol{v}_h \in \boldsymbol{X}_{0,h} \setminus \{0\}} \frac{b(\boldsymbol{v}_h, q_h)}{\|\boldsymbol{v}_h\|_{H^1(\Omega)} \|q_h\|_{L^2(\Omega)}}. \tag{3.2}$$

(A-C3) The finite element spaces are dense in their respective function spaces, i.e., for any $(\boldsymbol{v}, \eta, q) \in H^1(\Omega)^d \times C_\gamma \times L^2_0(\Omega)$,

$$\begin{aligned}\lim_{h\to 0} \inf_{\boldsymbol{w}_h \in \boldsymbol{X}_h} \|\boldsymbol{v} - \boldsymbol{w}_h\|_{H^1(\Omega)} &= \lim_{h\to 0} \inf_{\zeta_h \in C_{\gamma,h}} \|\eta - \zeta_h\|_{L^2(\Omega)} \\ &= \lim_{h\to 0} \inf_{r_h \in M_h} \|q - r_h\|_{L^2(\Omega)} = 0.\end{aligned}$$

Theorem 3.1 (Convergence of the conforming finite element method)**.** *Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain in two dimensions or a polyhedral Lipschitz domain in three dimensions. Suppose that the inverse permeability α satisfies (A1)–(A5) and there exists an isolated local or global minimizer $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ of (BP). Moreover, assume that, for $\theta > 0$, U_θ is the subset of Ω where $|\boldsymbol{u}|^2 \geq \theta$ a.e. in U_θ and suppose that there exists a $\theta' > 0$ such that U_θ is closed and has non-empty interior for all $\theta \leq \theta'$. Let p denote the unique Lagrange multiplier associated with $(\boldsymbol{u}, \rho)$ such that $(\boldsymbol{u}, \rho, p)$ satisfy the first-order optimality conditions (FOC1)–(FOC3a).*

Consider the conforming finite element spaces $\boldsymbol{X}_h \subset H^1(\Omega)^d$, $C_{\gamma,h} \subset C_\gamma$, and $M_h \subset L^2_0(\Omega)$ and suppose that the assumptions (A-C1)–(A-C3) hold.

Then, there exists an $\bar{h} > 0$ such that, for $h \leq \bar{h}$, $h \to 0$, there is a sequence of solutions $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,h} \times C_{\gamma,h} \times M_h$ to the following discretized first-order optimality conditions

$$a_{\rho_h}(\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) = l_{\boldsymbol{f}}(\boldsymbol{v}_h) \qquad \text{for all } \boldsymbol{v}_h \in \boldsymbol{X}_{0,h}, \qquad (\text{FOC1}_h)$$

$$b(\boldsymbol{u}_h, q_h) = 0 \qquad \text{for all } q_h \in M_h, \qquad (\text{FOC2}_h)$$

$$c_{\boldsymbol{u}_h}(\rho_h, \eta_h - \rho_h) \geq 0 \qquad \text{for all } \eta_h \in C_{\gamma,h}, \qquad (\text{FOC3a}_h)$$

such that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$, $\rho_h \to \rho$ strongly in $L^s(\Omega)$, $s \in [1, \infty)$, and $p_h \to p$ strongly in $L^2(\Omega)$ as $h \to 0$.

In the next subsection we give the proof of Theorem 3.1. In Proposition 3.1, by fixing a ball around an isolated minimizer, we show that finite element minimizers of a *modified* optimization problem converge weakly in $H^1(\Omega)^d \times L^2(\Omega)$ to the analytical isolated minimizer. From this we deduce that there exists a subsequence of finite element minimizers $(\boldsymbol{u}_h)$ that converges strongly to the analytical isolated minimizer in $L^2(\Omega)^d$. We then strengthen the convergence of ρ_h to strong convergence in $L^s(\Omega)$, $s \in [1, \infty)$, in Proposition 3.2 and strengthen the convergence of $\boldsymbol{u}_h$ to strong convergence in $H^1(\Omega)^d$ in Proposition 3.3. In Proposition 3.4, we prove that there exists an $\bar{h} > 0$ such that there is a subsequence, $h < \bar{h}$, $h \to 0$, of strongly converging finite element minimizers that also satisfy the discretized first-order

optimality conditions of (BP). Finally, in Proposition 3.5, we show that the Lagrange multiplier, $p_h \in M_h$, that satisfies the discretized first-order optimality conditions, converges strongly in $L^2(\Omega)$ to the analytical Lagrange multiplier.

3.1.2 Proof of the convergence of a conforming finite element method

Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP) and an $r > 0$ such that $(\boldsymbol{u}, \rho)$ is the unique minimizer in $B_{r,H^1(\Omega)\times L^2(\Omega)}(\boldsymbol{u}, \rho) \cap (\boldsymbol{U} \times C_\gamma)$, where

$$\begin{aligned} &B_{r,H^1(\Omega)\times L^2(\Omega)}(\boldsymbol{u}, \rho) \\ &:= \{\boldsymbol{v} \in H^1(\Omega)^d,\ \eta \in C_\gamma : \|\boldsymbol{u} - \boldsymbol{v}\|_{H^1(\Omega)} + \|\rho - \eta\|_{L^2(\Omega)} \le r\}. \end{aligned} \tag{3.3}$$

Such an $r > 0$ exists by the definition of an isolated minimizer. We also define $B_{r,H^1(\Omega)}(\boldsymbol{u})$ and $B_{r,L^2(\Omega)}(\rho)$ by

$$B_{r,H^1(\Omega)}(\boldsymbol{u}) := \{\boldsymbol{v} \in H^1(\Omega)^d : \|\boldsymbol{u} - \boldsymbol{v}\|_{H^1(\Omega)} \le r\}, \tag{3.4}$$

$$B_{r,L^2(\Omega)}(\rho) := \{\eta \in C_\gamma : \|\rho - \eta\|_{L^2(\Omega)} \le r\}. \tag{3.5}$$

We note that

$$\begin{aligned} &(\boldsymbol{U} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})) \times (C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)) \\ &\qquad\qquad \subset B_{r,H^1(\Omega)\times L^2(\Omega)}(\boldsymbol{u}, \rho) \cap (\boldsymbol{U} \times C_\gamma) \end{aligned}$$

and hence $(\boldsymbol{u}, \rho)$ is also the unique minimizer in $(\boldsymbol{U} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})) \times (C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho))$.

Moreover, we define the spaces $\boldsymbol{U}_{\boldsymbol{g}_h,h}$ and $\boldsymbol{U}_{0,h}$ by

$$\begin{aligned} \boldsymbol{U}_{\boldsymbol{g}_h,h} &:= \{\boldsymbol{v}_h \in \boldsymbol{X}_{\boldsymbol{g}_h,h} : b(\boldsymbol{v}_h, q_h) = 0 \text{ for all } q_h \in M_h\}, \\ \boldsymbol{U}_{0,h} &:= \{\boldsymbol{v}_h \in \boldsymbol{X}_{0,h} : b(\boldsymbol{v}_h, q_h) = 0 \text{ for all } q_h \in M_h\}. \end{aligned}$$

Proposition 3.1 (Weak convergence of $(\boldsymbol{u}_h, \rho_h)$ in $H^1(\Omega)^d \times L^2(\Omega)$)**.** *Suppose that the conditions of Theorem 3.1 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP). Consider the finite-dimensional optimization problem: find $(\boldsymbol{u}_h, \rho_h)$ that minimizes*

$$\min_{(\boldsymbol{v}_h,\eta_h)\in(\boldsymbol{U}_{\boldsymbol{g}_h,h}\cap B_{r/2,H^1(\Omega)}(\boldsymbol{u}))\times(C_{\gamma,h}\cap B_{r/2,L^2(\Omega)}(\rho))} J(\boldsymbol{v}_h, \eta_h). \tag{BP$_h$}$$

Then, a global minimizer $(\boldsymbol{u}_h, \rho_h)$ *of* (BP_h) *exists and there exist subsequences (up to relabeling) such that*

$$\boldsymbol{u}_h \rightharpoonup \boldsymbol{u} \text{ weakly in } H^1(\Omega)^d, \tag{3.6}$$

$$\boldsymbol{u}_h \to \boldsymbol{u} \text{ strongly in } L^2(\Omega)^d, \tag{3.7}$$

$$\rho_h \overset{*}{\rightharpoonup} \rho \text{ weakly-* in } L^\infty(\Omega), \tag{3.8}$$

$$\rho_h \rightharpoonup \rho \text{ weakly in } L^s(\Omega),\ s \in [1, \infty). \tag{3.9}$$

Proof. The functional J is continuous and

$$(\boldsymbol{U}_{\boldsymbol{g}_h,h} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)) \tag{3.10}$$

is a finite-dimensional, closed and bounded set and, therefore, sequentially compact by the Heine–Borel theorem. Hence J obtains its infimum in $(\boldsymbol{U}_{\boldsymbol{g}_h,h} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho))$ and, therefore, a global minimizer $(\boldsymbol{u}_h, \rho_h)$ exists.

By a corollary of Kakutani's Theorem (Corollary 2.1), if a Banach space is reflexive then every norm-closed, bounded and convex subset of the Banach space is weakly compact and thus, by the Eberlein–Šmulian theorem, sequentially weakly compact. It can be checked that $H^1(\Omega)^d \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})$ and $C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)$ are norm-closed, bounded and convex subsets of the reflexive Banach spaces $H^1(\Omega)^d$ and $L^2(\Omega)$, respectively. Therefore, $H^1(\Omega)^d \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})$ is weakly sequentially compact in $H^1(\Omega)^d$ and $C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)$ is weakly sequentially compact in $L^2(\Omega)$.

Hence we can extract subsequences, $(\boldsymbol{u}_h)$ and (ρ_h) of the sequence generated by the global minimizers of (BP_h) such that

$$\boldsymbol{u}_h \rightharpoonup \hat{\boldsymbol{u}} \in H^1(\Omega)^d \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u}) \text{ weakly in } H^1(\Omega)^d, \tag{3.11}$$

$$\rho_h \rightharpoonup \hat{\rho} \in C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho) \text{ weakly in } L^2(\Omega). \tag{3.12}$$

By assumption (A-C3), there exists a sequence of finite element functions $\tilde{\rho}_h \in C_{\gamma,h}$ that strongly converges to ρ in $L^2(\Omega)$. Moreover let $\tilde{\boldsymbol{u}}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,h}$ be a finite element function taken from the sequence of finite element functions that satisfy $\tilde{\boldsymbol{u}}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$. Such a sequence is shown to exist in [36, Lem. 3.1].

We now wish to identify the limits $\hat{\boldsymbol{u}}$ and $\hat{\rho}$. Consider the following bound:

$$\begin{aligned}
&2|J(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h)-J(\boldsymbol{u},\rho)| \\
&\quad\le \int_\Omega |(\alpha(\rho)-\alpha(\tilde{\rho}_h))|\boldsymbol{u}|^2| + |\alpha(\tilde{\rho}_h)(|\boldsymbol{u}|^2-|\tilde{\boldsymbol{u}}_h|^2)|\,\mathrm{d}x \\
&\qquad + \int_\Omega \nu\left||\nabla\boldsymbol{u}|^2-|\nabla\tilde{\boldsymbol{u}}_h|^2\right| + 2|\boldsymbol{f}\cdot(\boldsymbol{u}-\tilde{\boldsymbol{u}}_h)|\,\mathrm{d}x \\
&\quad\le L_\alpha\|\boldsymbol{u}\|^2_{L^4(\Omega)}\|\tilde{\rho}_h-\rho\|_{L^2(\Omega)} \\
&\qquad + \bar{\alpha}\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{L^2(\Omega)}(\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{L^2(\Omega)}+2\|\boldsymbol{u}\|_{L^2(\Omega)}) \\
&\qquad + \nu\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{H^1(\Omega)}(\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{H^1(\Omega)}+2\|\boldsymbol{u}\|_{H^1(\Omega)}) \\
&\qquad + 2\|\boldsymbol{f}\|_{L^2(\Omega)}\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{L^2(\Omega)},
\end{aligned} \tag{3.13}$$

where L_α denotes the Lipschitz constant for α. From (3.13) we see that

$$J(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h)\to J(\boldsymbol{u},\rho) \ \text{ as } \ h\to 0.$$

Furthermore, for sufficiently small $h>0$, we note that

$$(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h)\in(\boldsymbol{U}_{\boldsymbol{g}_h,h}\cap B_{r/2,H^1(\Omega)}(\boldsymbol{u}))\times(C_{\gamma,h}\cap B_{r/2,L^2(\Omega)}(\rho)).$$

Therefore,

$$J(\boldsymbol{u}_h,\rho_h)\le J(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h). \tag{3.14}$$

By taking the limit as $h\to 0$ and utilizing the strong convergence of $\tilde{\boldsymbol{u}}_h$ and $\tilde{\rho}_h$ to $\boldsymbol{u}$ and ρ, respectively, we see that

$$\lim_{h\to 0} J(\boldsymbol{u}_h,\rho_h)\le J(\boldsymbol{u},\rho). \tag{3.15}$$

(A-C1) implies that

$$\boldsymbol{u}_h|_{\partial\Omega}=\boldsymbol{g}_h\to\boldsymbol{g} \ \text{ strongly in } \ H^{1/2}(\partial\Omega)^d. \tag{3.16}$$

By assumption (A-C3), for every $q\in L^2_0(\Omega)$, there exists a sequence of $\tilde{q}_h\in M_h$ such that $\tilde{q}_h\to q$ strongly in $L^2(\Omega)$. Since $\boldsymbol{u}_h\rightharpoonup\hat{\boldsymbol{u}}$ weakly in $H^1(\Omega)^d$ and $\boldsymbol{u}_h\in\boldsymbol{U}_{\boldsymbol{g}_h,h}$, we see that

$$b(\hat{\boldsymbol{u}},q)=\lim_{h\to 0} b(\boldsymbol{u}_h,\tilde{q}_h)+\lim_{h\to 0} b(\boldsymbol{u}_h,q-\tilde{q}_h)=0 \ \text{ for all } \ q\in L^2_0(\Omega). \tag{3.17}$$

Hence $\hat{\boldsymbol{u}}$ is pointwise divergence-free and together with (3.16), we deduce that $\hat{\boldsymbol{u}} \in \boldsymbol{U} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})$. By construction $\hat{\rho} \in C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)$.

Since $(\boldsymbol{u}, \rho)$ is the unique local minimizer of $(\boldsymbol{U} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})) \times (C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho))$, from (3.15), we see that

$$\lim_{h\to 0} J(\boldsymbol{u}_h, \rho_h) = J(\boldsymbol{u}, \rho). \tag{3.18}$$

Since $(\boldsymbol{u}, \rho)$ is the unique minimizer in the spaces we consider, we can identify the limits $\hat{\boldsymbol{u}}$ and $\hat{\rho}$ as $\boldsymbol{u}$ and ρ, respectively, and state that $\boldsymbol{u}_h \rightharpoonup \boldsymbol{u}$ weakly in $H^1(\Omega)^d$ and $\rho_h \rightharpoonup \rho$ weakly in $L^2(\Omega)$. By the Rellich–Kondrachov theorem, we can extract a further subsequence such that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $L^2(\Omega)^d$.

We note that by the Banach–Alaoglu theorem, the closed unit ball of the dual space of a normed vector space, (for example $L^1(\Omega)$), is compact in the weak-* topology. Hence, we can also find a subsequence such that $\rho_h \overset{*}{\rightharpoonup} \hat{\rho} \in C_\gamma \cap \{\eta : \|\rho - \eta\|_{L^\infty(\Omega)} \leq r/2\}$ weakly-* in $L^\infty(\Omega)$. By the uniqueness of the weak limit, we can identify $\hat{\rho} = \rho$ a.e. in Ω and, thus, we deduce that $\rho_h \overset{*}{\rightharpoonup} \rho$ weakly-* in $L^\infty(\Omega)$. Consequently, $\rho_h \rightharpoonup \rho$ weakly in $L^s(\Omega)$ for all $s \in [1, \infty)$. □

Corollary 3.1 (Strong convergence of ρ_h in $L^s(\Omega_b)$)**.** *Suppose that the conditions of Theorem 3.1 hold. Fix any isolated minimizer of* (BP) *and let Ω_b be any measurable subset of Ω of positive measure on which ρ is equal to zero or one a.e. (if such a set exists). Then, there exists a sequence of finite element minimizers, ρ_h, of* (BP_h) *that converge strongly in $L^s(\Omega_b)$ to the isolated local or global minimizer of* (BP)*, where $s \in [1, \infty)$.*

Proof. We have shown that there exists a sequence of finite element minimizers $(\boldsymbol{u}_h, \rho_h)$ of (BP_h) that converge to the isolated minimizer $(\boldsymbol{u}, \rho)$. In particular $\rho_h \overset{*}{\rightharpoonup} \rho$ weakly-* in $L^\infty(\Omega)$ and $\rho_h \rightharpoonup \rho$ weakly in $L^s(\Omega)$ for all $s \in [1, \infty)$. The result is then deduced by following the proof of Corollary 3.2 in [124]. □

Proposition 3.2 (Strong convergence of ρ_h in $L^s(\Omega)$, $s \in [1, \infty)$)**.** *Suppose that the conditions stated in Theorem 3.1 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of* (BP)*. Then, there exists a subsequence of minimizers, (ρ_h), of* (BP_h) *such that*

$$\rho_h \to \rho \text{ strongly in } L^s(\Omega), \quad s \in [1, \infty). \tag{3.19}$$

Proof. We note that $C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)$ is a convex set, and hence for any $\eta_h \in C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)$, $t \in [0,1]$, we have that $\rho_h + t(\eta_h - \rho_h) \in C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)$. Since $(\boldsymbol{u}_h, \rho_h)$ is a minimizer of (BP$_h$), by the arguments used in Proposition 2.4 we deduce that

$$\int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2(\eta_h - \rho_h)\mathrm{d}x \geq 0 \quad \text{for all} \quad \eta_h \in C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho). \tag{3.20}$$

Hence, (FOC3a) and (3.20) imply that for all $\eta \in C_\gamma$ and $\eta_h \in C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)$ we have that

$$\int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2\rho \,\mathrm{d}x \leq \int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2\eta \,\mathrm{d}x, \tag{3.21}$$

$$\int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2\rho_h \,\mathrm{d}x \leq \int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2\eta_h \,\mathrm{d}x. \tag{3.22}$$

By subtracting $\int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2\rho_h \mathrm{d}x$ from (3.21) and $\int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2\rho \,\mathrm{d}x$ from (3.22), we see that

$$\int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2(\rho - \rho_h) \,\mathrm{d}x \leq \int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2(\eta - \rho_h)\mathrm{d}x, \tag{3.23}$$

$$\int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2(\rho_h - \rho) \,\mathrm{d}x \leq \int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2(\eta_h - \rho) \,\mathrm{d}x. \tag{3.24}$$

Summing (3.23) and (3.24) and rearranging the left-hand side, we see that

$$\begin{aligned}\int_\Omega (\alpha'(\rho) - \alpha'(\rho_h))|\boldsymbol{u}|^2(\rho - \rho_h)\mathrm{d}x + \int_\Omega \alpha'(\rho_h)(|\boldsymbol{u}|^2 - |\boldsymbol{u}_h|^2)(\rho - \rho_h)\mathrm{d}x \\ \leq \int_\Omega \alpha'(\rho)|\boldsymbol{u}|^2(\eta - \rho_h)\mathrm{d}x + \int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2(\eta_h - \rho)\mathrm{d}x.\end{aligned} \tag{3.25}$$

By fixing $\eta = \rho_h \in C_\gamma$ and subtracting the second term on the left-hand side of (3.25) from both sides we deduce that

$$\begin{aligned}&\int_\Omega (\alpha'(\rho) - \alpha'(\rho_h))|\boldsymbol{u}|^2(\rho - \rho_h)\mathrm{d}x \\ &\quad \leq \int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2(\eta_h - \rho)\mathrm{d}x + \int_\Omega \alpha'(\rho_h)(|\boldsymbol{u}_h|^2 - |\boldsymbol{u}|^2)(\rho - \rho_h)\mathrm{d}x.\end{aligned} \tag{3.26}$$

By an application of the mean value theorem, we note that there exists a $c \in (0,1)$ such that

$$\begin{aligned}&\int_\Omega (\alpha'(\rho) - \alpha'(\rho_h))|\boldsymbol{u}|^2(\rho - \rho_h)\mathrm{d}x \\ &\quad = \int_\Omega \alpha''(\rho_h + c(\rho - \rho_h))|\boldsymbol{u}|^2(\rho - \rho_h)^2\mathrm{d}x.\end{aligned} \tag{3.27}$$

By (A5) and the definition of U_θ we bound (3.27) from below:

$$\begin{aligned}&\int_\Omega \alpha''(\rho_h + c(\rho-\rho_h))|\boldsymbol{u}|^2(\rho-\rho_h)^2\mathrm{d}x\\ &\quad\geq \int_{U_\theta} \alpha''(\rho_h + c(\rho-\rho_h))|\boldsymbol{u}|^2(\rho-\rho_h)^2\mathrm{d}x \geq \alpha''_{\min}\theta\|\rho-\rho_h\|^2_{L^2(U_\theta)}.\end{aligned} \tag{3.28}$$

Now we bound the right-hand side of (3.26) as follows,

$$\begin{aligned}&\int_\Omega \alpha'(\rho_h)|\boldsymbol{u}_h|^2(\eta_h-\rho)\mathrm{d}x + \int_\Omega \alpha'(\rho_h)(|\boldsymbol{u}_h|^2-|\boldsymbol{u}|^2)(\rho-\rho_h)\mathrm{d}x\\ &\quad\leq 2\alpha'_{\max}(\|\boldsymbol{u}\|^2_{L^4(\Omega)} + \|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{L^4(\Omega)})\|\rho-\eta_h\|_{L^2(\Omega)}\\ &\qquad+\alpha'_{\max}\|\rho-\rho_h\|_{L^q(\Omega)}\|\boldsymbol{u}+\boldsymbol{u}_h\|_{L^{q'}(\Omega)}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)},\end{aligned} \tag{3.29}$$

where $2 < q' < \infty$ in two dimensions, $2 < q' \leq 6$ in three dimensions, and $q = 2q'/(q'-2)$. We note that

$$\begin{aligned}\|\boldsymbol{u}+\boldsymbol{u}_h\|_{L^{q'}(\Omega)} &\leq \|\boldsymbol{u}\|_{L^{q'}(\Omega)} + \|\boldsymbol{u}_h\|_{L^{q'}(\Omega)}\\ &\leq \|\boldsymbol{u}\|_{H^1(\Omega)} + \|\boldsymbol{u}_h\|_{H^1(\Omega)} \leq \hat{C} < \infty,\end{aligned} \tag{3.30}$$

where the second inequality holds thanks to the Sobolev embedding theorem. Combining (3.26)–(3.30) we see that

$$\|\rho-\rho_h\|^2_{L^2(U_\theta)} \leq C\left(\|\rho-\eta_h\|_{L^2(\Omega)} + \|\rho-\rho_h\|_{L^q(\Omega)}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{L^2(\Omega)}\right), \tag{3.31}$$

where $C = C(\alpha'_{\max}, \alpha''_{\min}, \theta, \|\boldsymbol{u}\|_{L^4(\Omega)}, \hat{C})$. By assumption (A-C3), there exists a sequence of finite element functions $\tilde{\rho}_h \in C_{\gamma,h}$ such that $\tilde{\rho}_h \to \rho$ strongly in $L^2(\Omega)$. Thanks to the strong convergence, we note that for sufficiently small h, $\tilde{\rho}_h \in C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)$. Hence we can fix $\eta_h = \tilde{\rho}_h$. By Proposition 3.1, we know that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $L^2(\Omega)^d$ and since $\rho \in C_\gamma$, $\rho_h \in C_{\gamma,h} \subset C_\gamma$, then $\|\rho-\rho_h\|_{L^q(\Omega)} \leq |\Omega|^{1/q}\|\rho-\rho_h\|_{L^\infty(\Omega)} \leq |\Omega|^{1/q}$. Therefore, the right-hand side of (3.31) tends to zero as $h \to 0$. Hence, we deduce that

$$\rho_h \to \rho \text{ strongly in } L^2(U_\theta), \quad \theta > 0. \tag{3.32}$$

Now we note that

$$\|\rho-\rho_h\|_{L^2(\Omega)} = \|\rho-\rho_h\|_{L^2(U_\theta)} + \|\rho-\rho_h\|_{L^2(U\setminus U_\theta)} + \|\rho-\rho_h\|_{L^2(\Omega\setminus U)}. \tag{3.33}$$

If $U\setminus U_\theta$ or $\Omega\setminus U$ are empty, we neglect the corresponding term in (3.33) with no loss of generality. Suppose $\Omega\setminus U$ is non-empty. By definition of U, $\boldsymbol{u} = \boldsymbol{0}$ a.e. in $\Omega\setminus U$.

By Proposition 2.2, this implies that $\rho = 0$ a.e. in $\Omega\backslash U$. Therefore, Corollary 3.1 implies that

$$\rho_h \to \rho \ \text{ strongly in } \ L^2(\Omega\backslash U). \tag{3.34}$$

Suppose $U\backslash U_\theta$ is non-empty. Since, $\rho, \rho_h \in C_\gamma$ we see that

$$\|\rho - \rho_h\|_{L^2(U\backslash U_\theta)} \leq |U\backslash U_\theta|^{1/2} \to 0 \ \text{ as } \ \theta \to 0. \tag{3.35}$$

Therefore, by first taking the limit as $h \to 0$ and then by taking the limit as $\theta \to 0$, (3.32)–(3.35) imply that $\rho_h \to \rho$ strongly in $L^2(\Omega)$.

Since $\|\rho - \rho_h\|_{L^1(\Omega)} \leq |\Omega|^{1/2}\|\rho - \rho_h\|_{L^2(\Omega)}$, we see that $\rho_h \to \rho$ strongly in $L^1(\Omega)$. Hence, for any $s \in [1, \infty)$,

$$\int_\Omega |\rho - \rho_h|^s \mathrm{d}x = \int_\Omega |\rho - \rho_h|^{s-1}|\rho - \rho_h| \mathrm{d}x \leq 1^{s-1}\|\rho - \rho_h\|_{L^1(\Omega)}, \tag{3.36}$$

which implies that $\rho_h \to \rho$ strongly in $L^s(\Omega)$. □

Proposition 3.3 (Strong convergence of $\boldsymbol{u}_h$ in $H^1(\Omega)^d$)**.** *Suppose that the conditions stated in Theorem 3.1 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP). Then, there exists a subsequence of minimizers, $(\boldsymbol{u}_h)$, of (BP_h) such that*

$$\boldsymbol{u}_h \to \boldsymbol{u} \ \textit{strongly in} \ H^1(\Omega)^d. \tag{3.37}$$

Proof. We note that the set $\boldsymbol{W}_h := \boldsymbol{U}_{\boldsymbol{g}_h,h} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})$ is convex. Hence it can be shown that minimizers of (BP_h) satisfy the variational inequality

$$a_{\rho_h}(\boldsymbol{u}_h, \boldsymbol{w}_h - \boldsymbol{u}_h) - l_{\boldsymbol{f}}(\boldsymbol{w}_h - \boldsymbol{u}_h) \geq 0 \quad \text{for all} \quad \boldsymbol{w}_h \in \boldsymbol{W}_h. \tag{3.38}$$

From Proposition 2.4 we deduce that, for all $\boldsymbol{w}_h \in \boldsymbol{W}_h$,

$$a_\rho(\boldsymbol{u}, \boldsymbol{w}_h - \boldsymbol{u}_h) + b(\boldsymbol{w}_h - \boldsymbol{u}_h, p) = l_{\boldsymbol{f}}(\boldsymbol{w}_h - \boldsymbol{u}_h). \tag{3.39}$$

Hence

$$a_{\rho_h}(\boldsymbol{u}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) \leq a_\rho(\boldsymbol{u}, \boldsymbol{u}_h - \boldsymbol{w}_h) + b(\boldsymbol{u}_h - \boldsymbol{w}_h, p). \tag{3.40}$$

By subtracting $a_{\rho_h}(\boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h)$ from both sides we see that

$$\begin{aligned} &a_{\rho_h}(\boldsymbol{u}_h - \boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) \\ &\quad \leq a_\rho(\boldsymbol{u}, \boldsymbol{u}_h - \boldsymbol{w}_h) - a_{\rho_h}(\boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) + b(\boldsymbol{u}_h - \boldsymbol{w}_h, p - q_h). \end{aligned} \tag{3.41}$$

We note that a_{ρ_h} is coercive and bounded with constants,

$$c_a = \nu/(c_p^2 + 1), \quad C_a = \max\{\nu, \overline{\alpha}\},$$

and b is bounded with constant C_b. Hence,

$$\begin{aligned} \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\Omega)}^2 &\leq \frac{1}{c_a} a_{\rho_h}(\boldsymbol{u}_h - \boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) \\ &\leq \frac{1}{c_a} \left(a_\rho(\boldsymbol{u}, \boldsymbol{u}_h - \boldsymbol{w}_h) - a_{\rho_h}(\boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) + b(\boldsymbol{u}_h - \boldsymbol{w}_h, p - q_h)\right) \\ &= \frac{1}{c_a} \left(\int_\Omega \alpha(\rho_h)(\boldsymbol{u} - \boldsymbol{w}_h) \cdot (\boldsymbol{u}_h - \boldsymbol{w}_h) + (\alpha(\rho) - \alpha(\rho_h))\, \boldsymbol{u} \cdot (\boldsymbol{u}_h - \boldsymbol{w}_h) \mathrm{d}x \right. \\ &\quad \left. + \int_\Omega \nu \nabla(\boldsymbol{u} - \boldsymbol{w}_h) : \nabla(\boldsymbol{u}_h - \boldsymbol{w}_h) \mathrm{d}x + b(\boldsymbol{u}_h - \boldsymbol{w}_h, p - q_h)\right) \\ &\leq \frac{1}{c_a} \bar{\alpha} \|\boldsymbol{u} - \boldsymbol{w}_h\|_{L^2(\Omega)} \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{L^2(\Omega)} \\ &\quad + \frac{1}{c_a} \|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{L^2(\Omega)} \\ &\quad + \frac{\nu}{c_a} |\boldsymbol{u} - \boldsymbol{w}_h|_{H^1(\Omega)} |\boldsymbol{u}_h - \boldsymbol{w}_h|_{H^1(\Omega)} + \frac{C_b}{c_a} \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\Omega)} \|p - q_h\|_{L^2(\Omega)}. \end{aligned}$$

Hence,

$$\|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\Omega)} \leq C \left(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} + \|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + \|p - q_h\|_{L^2(\Omega)}\right),$$

where $C = C(\bar{\alpha}, \nu, C_a, c_a, C_b)$ is a constant. This implies that, for all $\boldsymbol{w}_h \in \boldsymbol{W}_h$,

$$\begin{aligned} &\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} \\ &\quad \leq C \left(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} + \|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + \|p - q_h\|_{L^2(\Omega)}\right). \end{aligned} \tag{3.42}$$

where $C' = C'(\bar{\alpha}, \nu, C_a, c_a, C_b, L_\alpha)$. For sufficiently small h, we note that $\tilde{\boldsymbol{u}}_h \in \boldsymbol{W}_h$ (where $\tilde{\boldsymbol{u}}_h$ is defined in the proof of Proposition 3.1) and $\tilde{\boldsymbol{u}}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$. Moreover, by assumption (A-C3), there exists a sequence of finite element functions $\tilde{p}_h \in M_h$ that converges to p strongly in $L^2(\Omega)$. Suppose $\boldsymbol{w}_h = \tilde{\boldsymbol{u}}_h$ and $q_h = \tilde{p}_h$. From Proposition 3.2, we know that there exists a subsequence (not indicated) such that $\rho_h \to \rho$ strongly in $L^4(\Omega)$. We now observe that

$$\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \leq L_\alpha \|\rho - \rho_h\|_{L^4(\Omega)} \|\boldsymbol{u}\|_{L^4(\Omega)} \tag{3.43}$$

where L_α is the Lipschitz constant for α. Hence by taking the limit as $h \to 0$, we deduce that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$. □

In the following proposition, we show that (up to a subsequence) minimizers of (BP_h) also satisfy the first-order optimality conditions, ($\mathrm{FOC1}_h$)–($\mathrm{FOC3a}_h$), that are the finite-dimensional analogue of the first-order optimality conditions (FOC1)–(FOC3a) associated with (BP). This allows us to consider the finite-dimensional optimization problem over the whole set $\boldsymbol{U}_{\boldsymbol{g}_h,h} \times C_{\gamma,h}$, rather than the restricted set $(\boldsymbol{U}_{\boldsymbol{g}_h,h} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho))$.

Proposition 3.4 (Discretized first-order optimality conditions)**.** *Suppose that the conditions stated in Theorem 3.1 hold. Then, there exists an $\bar{h} > 0$ such that for all $h < \bar{h}$, there exists a unique Lagrange multiplier $p_h \in M_h$ such that the functions $(\boldsymbol{u}_h, \rho_h)$ that locally minimize (BP_h) satisfy the first-order optimality conditions ($\mathrm{FOC1}_h$)–($\mathrm{FOC3a}_h$).*

Proof. From Proposition 3.3, we know that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$. Hence by definition of strong convergence, there exists an $\bar{h} > 0$ such that, for all $h \leq \bar{h}$, $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} \leq r/4$. Therefore, for each $\boldsymbol{v}_h \in \boldsymbol{U}_{0,h}$, if $|t| < r/(4\|\boldsymbol{v}_h\|_{H^1(\Omega)})$ then $\boldsymbol{u}_h + t\boldsymbol{v}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,h} \cap B_{r/2,H^1(\Omega)}(\boldsymbol{u})$. Now we can follow the reasoning of the proof of Proposition 2.4 (adding the subscript $_h$ where necessary) to deduce the existence of a unique $p_h \in M_h$ such that ($\mathrm{FOC1}_h$)–($\mathrm{FOC3a}_h$) hold. □

Proposition 3.5 (Strong convergence of p_h in $L^2(\Omega)$)**.** *Suppose that the conditions stated in Theorem 3.1 hold. Then, there is a subsequence of the unique $p_h \in M_h$ defined in Proposition 3.4 that converges strongly in $L^2(\Omega)$ to the $p \in L^2_0(\Omega)$ that solves (FOC1)–(FOC3a) for the given isolated minimizer $(\boldsymbol{u}, \rho)$.*

Proof. The inf-sup condition (A-C2) for M_h and $\boldsymbol{X}_{0,h}$ implies that, for any $q_h \in M_h$,

$$\begin{aligned} c_b\|q_h - p_h\|_{L^2(\Omega)} &\leq \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,h}\setminus\{0\}} \frac{b(\boldsymbol{w}_h, q_h - p_h)}{\|\boldsymbol{w}_h\|_{H^1(\Omega)}} \\ &= \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,h}\setminus\{0\}} \frac{b(\boldsymbol{w}_h, p - p_h) + b(\boldsymbol{w}_h, q_h - p)}{\|\boldsymbol{w}_h\|_{H^1(\Omega)}} \\ &\leq \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,h}\setminus\{0\}} \frac{|b(\boldsymbol{w}_h, p - p_h)| + |b(\boldsymbol{w}_h, q_h - p)|}{\|\boldsymbol{w}_h\|_{H^1(\Omega)}} \\ &= \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,h}\setminus\{0\}} \frac{|a_\rho(\boldsymbol{u}, \boldsymbol{w}_h) - a_{\rho_h}(\boldsymbol{u}_h, \boldsymbol{w}_h)| + |b(\boldsymbol{w}_h, q_h - p)|}{\|\boldsymbol{w}_h\|_{H^1(\Omega)}} \\ &\leq \|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + (\bar{\alpha} + 1)\,\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} \\ &\quad + C_b\|p - q_h\|_{L^2(\Omega)}. \end{aligned} \tag{3.44}$$

Hence,

$$\begin{aligned} \|p - p_h\|_{L^2(\Omega)} \leq C\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \\ + C\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} + C\|p - q_h\|_{L^2(\Omega)}, \end{aligned} \tag{3.45}$$

where $C = C(c_b, C_b, \bar{\alpha}, L_\alpha)$. By assumption (A-C3), there exists a sequence of finite element functions, $\tilde{p}_h \in M_h$ that satisfies $\tilde{p}_h \to p$ strongly in $L^2(\Omega)$. Let $q_h = \tilde{p}_h$. We have already shown that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$ in Proposition 3.3. Similarly, in the proof of Proposition 3.3 we also showed that $\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \to 0$. Hence we conclude that $p_h \to p$ strongly in $L^2(\Omega)$. □

We now have the necessary ingredients to prove Theorem 3.1.

Proof of Theorem 3.1. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP) and its unique associated Lagrange multiplier p. By the results of Propositions 3.1, 3.2, 3.3, and 3.4, there exists a mesh size $\bar{h}$ such that for, $h < \bar{h}$, there exists a sequence of finite element solutions $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{U}_{\boldsymbol{g}_h,h} \times C_{\gamma,h} \times M_h$ satisfying (FOC1$_h$)–(FOC3a$_h$) that converges to $(\boldsymbol{u}, \rho, p)$. By taking a subsequence if necessary (not indicated), Proposition 3.2 implies that $\rho_h \to \rho$ strongly in $L^s(\Omega)$, $s \in [1, \infty)$, Proposition 3.3 implies that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\Omega)^d$, and Proposition 3.5 implies that $p_h \to p$ strongly in $L^2(\Omega)$. □

Corollary 3.2 (Relaxing the volume constraint)**.** *Suppose that the conditions of Theorem 3.1 hold. Fix an isolated minimizer* $(\boldsymbol{u}, \rho)$ *of* (BP) *and its associated*

Lagrange multiplier p. Consider the finite element space $C_{[0,1],h} \subset C_{[0,1]}$ with $C_{[0,1]}$ as defined in Proposition 2.5. Then, there exists a sequence $(\lambda_h) \in \mathbb{R}$ such that the sequence $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,h} \times C_{\gamma,h} \times M_h$, satisfying ($\mathrm{FOC1}_h$)–($\mathrm{FOC3a}_h$), that converges to $(\boldsymbol{u}, \rho, p)$ strongly in $H^1(\Omega)^d \times L^s(\Omega) \times L^2(\Omega)$, $s \in [1, \infty)$, also satisfies, for all $(\eta_h, \zeta_h) \in C_{[0,1],h} \times \mathbb{R}$,

$$c_{\boldsymbol{u}_h,\lambda_h}(\rho_h, \eta_h - \rho_h) \geq 0, \tag{$\mathrm{FOC3b}_h$}$$

$$d_{\rho_h}(\lambda_h, \zeta_h) = 0, \tag{$\mathrm{FOC4}_h$}$$

with $c_{\boldsymbol{u}_h,\lambda_h}$ and d_{ρ_h} as defined in Proposition 2.5.

Proof. By construction $(\boldsymbol{u}_h, \rho_h)$ minimizes (BP_h). Hence, the result can be deduced in similar fashion to the proof of Proposition 2.5. □

3.2 Divergence-free DG finite element discretization

In this section we adapt the proof of Theorem 3.1 to a nonconforming divergence-free DG discretization of the velocity where $\boldsymbol{X}_h \not\subset H^1(\Omega)^d$ but $\boldsymbol{X}_h \subset \boldsymbol{H}(\mathrm{div}; \Omega)$. The two main difficulties are the following:

- $\nabla \boldsymbol{v}_h$ is not well-defined on Ω for a general function $\boldsymbol{v}_h \in \boldsymbol{X}_h$. Therefore, we need to introduce a new power dissipation functional J_h and form a_{h,η_h} so that $J_h(\boldsymbol{v}_h, \eta_h)$ and $a_{h,\eta_h}(\boldsymbol{v}_h, \cdot)$ are well-defined for functions $\boldsymbol{v}_h \in \boldsymbol{X}_h$ and $\eta_h \in C_{\gamma,h}$;
- A crucial result utilized in the proof of Theorem 3.1 was the extraction of a strongly converging subsequence $(\boldsymbol{u}_h)$ in $L^2(\Omega)^d$ from a weakly converging sequence in $H^1(\Omega)^d$ in Proposition 3.1. This compactness result can no longer be used as the minimizing velocity finite element functions do not generate a sequence that is bounded in $H^1(\Omega)^d$. Hence, a different compactness result is required to extract a strongly converging subsequence in $L^2(\Omega)^d$.

In order to define the modified functional J_h, we must first define notation for the elements and faces for a given mesh $\mathcal{T}_h$. We split the set of facets $\mathcal{F}_h$ into the union $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^\partial$ where $\mathcal{F}_h^i$ is the subset of interior facets and $\mathcal{F}_h^\partial$ collects all Dirichlet boundary facets $F \subset \partial\Omega$. If $F \in \mathcal{F}_h^i$, then $F = \overline{\partial K^+} \cap \overline{\partial K^-}$ for two elements $K^-, K^+ \in \mathcal{T}_h$. We write $\boldsymbol{n}_F^+$ and $\boldsymbol{n}_F^-$ to denote the outward normal unit vectors to the boundaries ∂K^+ and ∂K^-, respectively. If $F \in \mathcal{F}_h^\partial$, then $\boldsymbol{n}_F$ is the outer unit normal vector $\boldsymbol{n}$. We make the following assumption concerning the triangulation at the boundary:

(M3) (Boundary regularity). There exists a constant $c_1 > 0$ such that:

$$c_1 h \leq h_F \text{ for all } F \in \mathcal{F}_h^\partial.$$

We denote the space of discontinuous finite element functions with degree no higher than k by

$$X_{\mathrm{DG}_k} := \{v \in L^1(\Omega) : v|_K \in \mathcal{P}_k \text{ for all } K \in \mathcal{T}_h\}, \tag{3.46}$$

where $\mathcal{P}_k$ denotes the set of polynomials of order no higher than k. Let $\boldsymbol{\phi} \in (X_{\mathrm{DG}_k})^d$ and $\boldsymbol{\Phi} \in (X_{\mathrm{DG}_k})^{d\times d}$ be any piecewise vector or matrix-valued function, respectively with traces from within the interior of $K^\pm$ denoted by $\boldsymbol{\phi}^\pm$ and $\boldsymbol{\Phi}^\pm$, respectively. We define the jump $[\![\cdot]\!]_F$ and the average $\{\!\{\cdot\}\!\}_F$ operators across interior facets $F \in \mathcal{F}_h^i$ by

$$[\![\boldsymbol{\phi}]\!]_F = \boldsymbol{\phi}^+ \otimes \boldsymbol{n}_F^+ + \boldsymbol{\phi}^- \otimes \boldsymbol{n}_F^- \quad \text{and} \quad \{\!\{\boldsymbol{\Phi}\}\!\}_F = \frac{1}{2}\left(\boldsymbol{\Phi}^+ + \boldsymbol{\Phi}^-\right). \tag{3.47}$$

If $F \in \mathcal{F}_h^\partial$, we set $[\![\boldsymbol{\phi}]\!]_F = \boldsymbol{\phi} \otimes \boldsymbol{n}_F$ and $\{\!\{\boldsymbol{\Phi}\}\!\}_F = \boldsymbol{\Phi}$. We note that, for any $F \in \mathcal{F}_h^\partial$, $\int_F |[\![\boldsymbol{\phi}]\!]_F|^2 \,\mathrm{d}s = \int_F |\boldsymbol{\phi}|^2 \,\mathrm{d}s$.

A function, $v \in X_{\mathrm{DG}_k}$, only has a well-defined weak derivative on each element on the mesh. Hence, we define the broken Sobolev space $H^1(\mathcal{T}_h)$ as:

$$H^1(\mathcal{T}_h) := \{v \in L^1(\Omega) : v \in H^1(K) \text{ for all } K \in \mathcal{T}_h\}. \tag{3.48}$$

Moreover, for a function $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$, we define the broken H^1-seminorm and norms as:

$$|\boldsymbol{v}|^2_{H^1(\mathcal{T}_h)} := \sum_{K\in\mathcal{T}_h} \|\nabla \boldsymbol{v}\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^i} \int_F h_F^{-1} |[\![\boldsymbol{v}]\!]_F|^2 \mathrm{d}s, \tag{3.49}$$

$$\|\boldsymbol{v}\|^2_{H^1(\mathcal{T}_h)} := \|\boldsymbol{v}\|^2_{L^2(\Omega)} + |\boldsymbol{v}|^2_{H^1(\mathcal{T}_h)}, \tag{3.50}$$

$$\|\boldsymbol{v}\|^2_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} := \|\boldsymbol{v}\|^2_{H^1(\mathcal{T}_h)} + \sum_{F\in\mathcal{F}_h^\partial} \int_F h_F^{-1} |[\![\boldsymbol{v}-\boldsymbol{g}]\!]_F|^2 \mathrm{d}s. \tag{3.51}$$

The two families of DG finite elements of interest for the velocity are the Brezzi–Douglas–Marini (BDM) finite element [40, 41] and the Raviart–Thomas (RT) finite element [117, 136]. The k-th order BDM finite element is defined for $d = 2$ in [41, Sec. 2] and for $d = 3$ in [40, Sec. 2]. The k-th order RT finite element is defined in [136, Sec. 3] and [117, Sec. 2] for $d = 2$ and $d = 3$, respectively. The finite element spaces induced by the k-th order BDM and RT finite elements are denoted by $\boldsymbol{X}_{\mathrm{BDM}_k}$ and $\boldsymbol{X}_{\mathrm{RT}_k}$, respectively. We note that $\boldsymbol{X}_{\mathrm{RT}_k} \subset \boldsymbol{X}_{\mathrm{BDM}_k} \subset \boldsymbol{Z}_h$ where, for a given $k \geq 1$,

$$\boldsymbol{Z}_h := \{\boldsymbol{v} \in (X_{\mathrm{DG}_k})^d : \mathrm{div}(\boldsymbol{v}) \in X_{\mathrm{DG}_{k-1}} \cap L^2(\Omega)\}. \tag{3.52}$$

We note that $\boldsymbol{Z}_h \subset \boldsymbol{H}(\mathrm{div};\Omega)$ and $\boldsymbol{Z}_h \subset H^1(\mathcal{T}_h)^d$. We define the following subspaces of $\boldsymbol{X}_h$:

$$\boldsymbol{X}_{0,\boldsymbol{n},h} := \{\boldsymbol{v} \in \boldsymbol{X}_h : (\boldsymbol{v}\cdot\boldsymbol{n})|_{\partial\Omega} = 0\}, \tag{3.53}$$

$$\boldsymbol{X}_{\boldsymbol{g},\boldsymbol{n},h} := \{\boldsymbol{v} \in \boldsymbol{X}_h : (\boldsymbol{v}\cdot\boldsymbol{n})|_{\partial\Omega} - \boldsymbol{g}\cdot\boldsymbol{n} = 0\}. \tag{3.54}$$

We now define the discrete Borrvall–Petersson power dissipation functional for a DG finite element discretization. Consider the functions $\boldsymbol{u}_h \in \boldsymbol{X}_h \subset \boldsymbol{Z}_h$ and $\rho_h \in C_{\gamma,h}$. We note that, in general, $J(\boldsymbol{u}_h, \rho_h)$ is ill-defined as $\nabla \boldsymbol{u}_h$ is not defined on all of Ω. Hence, given a penalization parameter $\sigma > 0$, we define

the discrete analogue J_h as

$$\begin{aligned} J_h(\boldsymbol{u}_h, \rho_h) := \frac{1}{2}\int_\Omega \left(\alpha(\rho_h)|\boldsymbol{u}_h|^2 - 2\boldsymbol{f}\cdot\boldsymbol{u}_h\right)\mathrm{d}x + \frac{\nu}{2}\sum_{K\in\mathcal{T}_h}\int_K |\nabla\boldsymbol{u}_h|^2\mathrm{d}x \\ + \frac{1}{2}\sum_{F\in\mathcal{F}_h^i}\sigma h_F^{-1}\int_F |[\![\boldsymbol{u}_h]\!]_F|^2\mathrm{d}s - \sum_{F\in\mathcal{F}_h^i}\int_F \{\!\{\nabla\boldsymbol{u}_h\}\!\}_F : [\![\boldsymbol{u}_h]\!]_F\,\mathrm{d}s \\ + \frac{1}{2}\sum_{F\in\mathcal{F}_h^\partial}\sigma h_F^{-1}\int_F |[\![\boldsymbol{u}_h - \boldsymbol{g}_h]\!]_F|^2\mathrm{d}s \\ - \sum_{F\in\mathcal{F}_h^\partial}\int_F \{\!\{\nabla\boldsymbol{u}_h\}\!\}_F : [\![\boldsymbol{u}_h - \boldsymbol{g}_h]\!]_F\,\mathrm{d}s. \end{aligned} \tag{3.55}$$

Remark 3.1. *This particular choice for J_h as the discrete analogue of J is motivated by an interior penalty approach for DG formulations. In Proposition 3.12, we prove that the velocity minimizers of J_h satisfy a fluid momentum equation featuring terms that arise in the interior penalty DG discretization of the Stokes equations [52, 78].*

Remark 3.2. *For any $\boldsymbol{v}_h \in \boldsymbol{Z}_h$ the terms $\{\!\{\nabla\boldsymbol{v}_h\}\!\}_F$ and $[\![\boldsymbol{v}_h]\!]_F$ as they appear in J_h are well-defined [20, Sec. 3.1].*

Proposition 3.6 (Consistency of J_h)**.** *Consider any $(\boldsymbol{v}, \eta) \in H^1_{\boldsymbol{g}}(\Omega)^d \times C_\gamma$ such that $\boldsymbol{v} \in H^r(\Omega)^d$ for some $r > 3/2$. Then, J_h, $h > 0$ is consistent, i.e.*

$$J_h(\boldsymbol{v}, \eta) = J(\boldsymbol{v}, \eta). \tag{3.56}$$

Proof. Since $\boldsymbol{v} \in H^r(\Omega)^d$, for some $r > 3/2$, we note that $J_h(\boldsymbol{v}, \eta)$ is well-defined and there can be no jumps in $\boldsymbol{v}$ across elements. Hence, for all $F \in \mathcal{F}_h^i$, integrals involving $[\![\boldsymbol{v}]\!]_F$ are equal to zero. Moreover $\boldsymbol{v}|_{\partial\Omega} = \boldsymbol{g}$ and, therefore, $\int_F |[\![\boldsymbol{v} - \boldsymbol{g}]\!]_F|^2\,\mathrm{d}s = 0$ for all $F \in \mathcal{F}_h^\partial$. Hence,

$$J_h(\boldsymbol{v}, \eta) = \frac{1}{2}\int_\Omega \left(\alpha(\eta)|\boldsymbol{v}|^2 - 2\boldsymbol{f}\cdot\boldsymbol{v}\right)\mathrm{d}x + \frac{\nu}{2}\sum_{K\in\mathcal{T}_h}\int_K |\nabla\boldsymbol{v}|^2\mathrm{d}x = J(\boldsymbol{v}, \eta). \tag{3.57}$$

□

For a sufficiently large penalization parameter $\sigma > 0$, we define the broken form $a_{h,\rho}(\boldsymbol{u}, \boldsymbol{v})$ by

$$\begin{aligned} a_{h,\rho}(\boldsymbol{u}, \boldsymbol{v}) := \sum_{K\in\mathcal{T}_h}\int_K \alpha(\rho)\boldsymbol{u}\cdot\boldsymbol{v} + \nabla\boldsymbol{u} : \nabla\boldsymbol{v}\,\mathrm{d}x + \sum_{F\in\mathcal{F}_h}\sigma h_F^{-1}\int_F [\![\boldsymbol{u}]\!]_F : [\![\boldsymbol{v}]\!]_F\mathrm{d}s \\ - \sum_{F\in\mathcal{F}_h}\int_F \{\!\{\nabla\boldsymbol{u}\}\!\}_F : [\![\boldsymbol{v}]\!]_F\mathrm{d}s - \sum_{F\in\mathcal{F}_h}\int_F [\![\boldsymbol{u}]\!]_F : \{\!\{\nabla\boldsymbol{v}\}\!\}_F\mathrm{d}s, \end{aligned} \tag{3.58}$$

and the linear functional $l_{h,\boldsymbol{f},\boldsymbol{g}}$ as

$$\begin{aligned} l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{v}) := &\int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x \\ &+ \sum_{F\in\mathcal{F}_h^\partial} \sigma h_F^{-1}\int_F [\![\boldsymbol{g}]\!]_F : [\![\boldsymbol{v}]\!]_F\,\mathrm{d}s - \sum_{F\in\mathcal{F}_h^\partial}\int_F [\![\boldsymbol{g}]\!]_F : \{\!\{\nabla\boldsymbol{v}\}\!\}_F\,\mathrm{d}s. \end{aligned} \tag{3.59}$$

Proposition 3.7 (Consistency of $a_{h,\rho}$)**.** *Suppose that $(\boldsymbol{u},\rho)\in\boldsymbol{U}\times C_\gamma$ is an isolated minimizer of (BP) and let $p\in L^2_0(\Omega)$ denote the Lagrange multiplier such that $(\boldsymbol{u},\rho,p)$ satisfy (FOC1)–(FOC3a). Moreover, assume that $\boldsymbol{u}\in H^r(\Omega)^d$ for some $r>3/2$. Then, for all $\boldsymbol{v}_h\in H^1(\mathcal{T}_h)^d\cap\boldsymbol{H}_0(\mathrm{div};\Omega)$, we have that*

$$a_{h,\rho}(\boldsymbol{u},\boldsymbol{v}_h) + b(\boldsymbol{v}_h,p) = l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{v}_h). \tag{3.60}$$

Proof. By (FOC1), we have that, for all $\boldsymbol{w}_h\in H^1_0(\Omega)^d$,

$$a_\rho(\boldsymbol{u},\boldsymbol{w}_h) + b(\boldsymbol{w}_h,p) = l_{\boldsymbol{f}}(\boldsymbol{w}_h). \tag{3.61}$$

Since, by assumption, $\boldsymbol{f}\in L^2(\Omega)^d$, then, by an integration by parts, we see that

$$\alpha(\rho)\boldsymbol{u} - \nu\Delta\boldsymbol{u} + \nabla p = \boldsymbol{f} \ \text{ a.e. in } \Omega. \tag{3.62}$$

Therefore, as the set of smooth functions is dense in $L^2(\Omega)$, we can test (3.62) against any $\boldsymbol{v}_h\in H^1(\mathcal{T}_h)^d\cap\boldsymbol{H}_0(\mathrm{div};\Omega)\subset L^2(\Omega)^d$. Thus, by performing a second integration by parts, we have that

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h}\int_K \alpha(\rho)\boldsymbol{u}\cdot\boldsymbol{v}_h + \nu\nabla\boldsymbol{u} : \nabla\boldsymbol{v}_h - p\,\mathrm{div}(\boldsymbol{v}_h)\,\mathrm{d}x \\ &\quad - \sum_{F\in\mathcal{F}_h}\int_F \{\!\{\nabla\boldsymbol{u}\}\!\}_F : [\![\boldsymbol{v}_h]\!]_F\,\mathrm{d}s = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}_h\,\mathrm{d}x. \end{aligned} \tag{3.63}$$

The element-wise surface integrals arising by the integration by parts of the ∇p term drop out due to the continuity of $\boldsymbol{v}_h\cdot\boldsymbol{n}$ across faces of elements for all $\boldsymbol{v}_h\in\boldsymbol{H}_0(\mathrm{div};\Omega)$. Similarly the boundary surface integrals drop out since $\boldsymbol{v}_h\cdot\boldsymbol{n}=0$ on $\partial\Omega$. As $\boldsymbol{u}\in\boldsymbol{U}$, for all $F\in\mathcal{F}_h^i$, we have that $[\![\boldsymbol{u}]\!]_F=\boldsymbol{0}$ and for all $F\in\mathcal{F}_h^\partial$, $[\![\boldsymbol{u}]\!]_F=[\![\boldsymbol{g}]\!]_F$. As $\boldsymbol{u}\in H^r(\Omega)^d$, for some $r>3/2$, the traces of $\nabla\boldsymbol{u}$ on $F\in\mathcal{F}_h$ are well-defined. We conclude that (3.60) holds. □

Proposition 3.8 (Coercivity and boundedness of $a_{h,\rho}$)**.** *There exists a $\sigma_0 > 0$, such that for all $\sigma \geq \sigma_0$, $\boldsymbol{w}_h, \boldsymbol{u}_h \in \boldsymbol{X}_h$ and $\eta \in C_\gamma$, there exists constants $c_a, C_a > 0$ such that*

$$c_a \|\boldsymbol{w}_h\|^2_{H^1(\mathcal{T}_h)} \leq a_{h,\eta}(\boldsymbol{w}_h, \boldsymbol{w}_h), \tag{3.64}$$

$$a_{h,\eta}(\boldsymbol{w}_h, \boldsymbol{u}_h) \leq C_a \|\boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)} \|\boldsymbol{u}_h\|_{H^1(\mathcal{T}_h)}. \tag{3.65}$$

Proof. We note that, by assumption (A1), $0 \leq \alpha(\eta) \leq \bar{\alpha}$ for all $\eta \in C_\gamma$. Hence, the result follows from classical coercivity and boundedness results for DG discretizations for interior penalty methods [20, Sec. 4.1–4.2]. □

Definition 3.1. *For some $\boldsymbol{g}_h$ defined on $\partial\Omega$, that can represented exactly in $\boldsymbol{X}_h$, we define the spaces $\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h}$ and $\boldsymbol{U}_{0,\boldsymbol{n},h}$ as:*

$$\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} := \{\boldsymbol{u} \in \boldsymbol{X}_{\boldsymbol{g}_h,\boldsymbol{n},h} : b(\boldsymbol{u}_h, q_h) = 0 \text{ for all } q_h \in X_{\mathrm{DG}_{k-1}}\} \tag{3.66}$$

$$\boldsymbol{U}_{0,\boldsymbol{n},h} := \{\boldsymbol{u} \in \boldsymbol{X}_{0,\boldsymbol{n},h} : b(\boldsymbol{u}_h, q_h) = 0 \text{ for all } q_h \in X_{\mathrm{DG}_{k-1}}\}. \tag{3.67}$$

In the following lemma we provide the proof that functions $\boldsymbol{v}_h \in \boldsymbol{U}_{0,\boldsymbol{n},h}$ and $\boldsymbol{v}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h}$ are pointwise divergence-free.

Lemma 3.1 (Pointwise divergence-free)**.** *Suppose that $\boldsymbol{X}_h \subset \boldsymbol{Z}_h$. Consider a function $\boldsymbol{v}_h \in \boldsymbol{U}_{0,\boldsymbol{n},h}$ or $\boldsymbol{v}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h}$. Then,* $\operatorname{div}(\boldsymbol{v}_h) = 0$ *a.e. in* Ω.

Proof. Since $\boldsymbol{U}_{0,\boldsymbol{n},h}, \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \subset \boldsymbol{X}_h \subset \boldsymbol{Z}_h$ then, by definition, $\operatorname{div}(\boldsymbol{v}_h) \in X_{\mathrm{DG}_{k-1}}$. Hence, there exists a $q_h \in X_{\mathrm{DG}_{k-1}}$ such that $q_h = \operatorname{div}(\boldsymbol{v}_h)$. Therefore,

$$b(\boldsymbol{v}_h, q_h) = \|\operatorname{div}(\boldsymbol{v}_h)\|^2_{L^2(\Omega)} = 0, \tag{3.68}$$

which implies that $\operatorname{div}(\boldsymbol{v}_h) = 0$ a.e. in Ω. □

3.2.1 Assumptions and the second convergence theorem

As before, the boundary data $\boldsymbol{g}$ cannot be represented in the finite element space. Hence we instead approximate the boundary data with a finite element function $\boldsymbol{g}_h$ (which can be represented) and assume that

(A-DG1) $h^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)} \to 0$ as $h \to 0$.

We also assume that:

(A-DG2) $\boldsymbol{X}_{0,\boldsymbol{n},h}$ and M_h satisfy the following inf-sup condition for some $c_b > 0$,

$$c_b \leq \inf_{q_h \in M_h \setminus \{0\}} \sup_{\boldsymbol{v}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h} \setminus \{0\}} \frac{b(\boldsymbol{v}_h, q_h)}{\|\boldsymbol{v}_h\|_{H^1(\mathcal{T}_h)} \|q_h\|_{L^2(\Omega)}}. \tag{3.69}$$

(A-DG3) The finite element spaces are dense in their respective function spaces, i.e., for any $(\boldsymbol{v}, \eta, q) \in H^1(\mathcal{T}_h)^d \times C_\gamma \times L^2_0(\Omega)$,

$$\begin{aligned} \lim_{h \to 0} \inf_{\boldsymbol{w}_h \in \boldsymbol{X}_h} \|\boldsymbol{v} - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)} &= \lim_{h \to 0} \inf_{\zeta_h \in C_{\gamma,h}} \|\eta - \zeta_h\|_{L^2(\Omega)} \\ &= \lim_{h \to 0} \inf_{r_h \in M_h} \|q - r_h\|_{L^2(\Omega)} = 0. \end{aligned}$$

Remark 3.3. *The inf-sup (A-DG2) and density (A-DG3) conditions are satisfied by either $\boldsymbol{X}_h = \boldsymbol{X}_{\mathrm{BDM}_k}$ or $\boldsymbol{X}_h = \boldsymbol{X}_{\mathrm{RT}_k}$ with the choice of the pressure finite element space $M_h = X_{\mathrm{DG}_{k-1}}$ [51].*

We can now state our second main theorem of this chapter.

Theorem 3.2 (Convergence of the divergence-free DG finite element method)**.** *Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain in two dimensions or a polyhedral Lipschitz domain in three dimensions. Suppose that the inverse permeability α satisfies (A1)–(A5) and there exists an isolated local or global minimizer $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ of (BP) that has the additional regularity $\boldsymbol{u} \in H^r(\Omega)^d$ for some $r > 3/2$. Moreover, assume that, for $\theta > 0$, U_θ is the subset of Ω where $|\boldsymbol{u}|^2 \geq \theta$ a.e. in U_θ and suppose that there exists a $\theta' > 0$ such that U_θ is closed and has non-empty interior for all $\theta \leq \theta'$. Let p denote the unique Lagrange multiplier associated with $(\boldsymbol{u}, \rho)$ such that $(\boldsymbol{u}, \rho, p)$ satisfy the first-order optimality conditions (FOC1)–(FOC3a).*

Consider the finite element spaces $\boldsymbol{X}_h \subset \boldsymbol{Z}_h$, $C_{\gamma,h} \subset C_\gamma$, and $M_h \subset L^2_0(\Omega)$ and suppose that the assumptions (A-DG1)–(A-DG3) hold.

Then, there exists an $\bar{h} > 0$ such that, for $h \leq \bar{h}$, $h \to 0$, there is a sequence of solutions $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times C_{\gamma,h} \times M_h$ to the following discretized first-order

optimality conditions

$$a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) = l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{v}_h) \qquad \text{for all } \boldsymbol{v}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}, \qquad \text{(FOC1-DG}_h\text{)}$$

$$b(\boldsymbol{u}_h, q_h) = 0 \qquad \text{for all } q_h \in M_h, \qquad \text{(FOC2-DG}_h\text{)}$$

$$c_{\boldsymbol{u}_h}(\rho_h, \eta_h - \rho_h) \geq 0 \qquad \text{for all } \eta_h \in C_{\gamma,h}, \qquad \text{(FOC3a-DG}_h\text{)}$$

such that, $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0$, $\rho_h \to \rho$ *strongly in* $L^s(\Omega)$, $s \in [1, \infty)$, *and* $p_h \to p$ *strongly in* $L^2(\Omega)$ *as* $h \to 0$.

3.2.2 Proof of the convergence of a divergence-free DG finite element method

The proof follows a similar pattern to the proof of Theorem 3.1. The nonconvexity of the optimization problem is handled by fixing an isolated minimizer and introducing a modified optimization involving J_h that has the fixed isolated minimizer as the unique solution. We then show that there exists a sequence of discretized solutions that converges to the minimizer in the appropriate norms. The modified optimization problem is related back to the original optimization problem by showing that a subsequence of the strongly converging minimizers satisfy the first-order optimality conditions (FOC1-DG$_h$)–(FOC3a-DG$_h$), that can be solved numerically.

An important fact we utilize later is the existence of sequences in $\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h}$ that converge strongly to $\boldsymbol{u}$.

Lemma 3.2 (Strongly converging sequences)**.** *Suppose that (A-DG1)–(A-DG3) hold and* $\boldsymbol{X}_h \subset \boldsymbol{Z}_h$. *Consider any isolated minimizer* $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ *of* (BP). *Then, there exists a sequence of functions* $(\tilde{\boldsymbol{u}}_h, \tilde{p}_h) \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times M_h$ *such that* $\|\boldsymbol{u} - \tilde{\boldsymbol{u}}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0$ *and* $\|p - \tilde{p}_h\|_{L^2(\Omega)} \to 0$.

Proof. For sufficient large $\sigma > 0$, consider the problem, find $(\tilde{\boldsymbol{u}}_h, \tilde{p}_h) \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times M_h$ that satisfies

$$a_{h,\rho}(\tilde{\boldsymbol{u}}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, \tilde{p}_h) = l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{v}_h) \qquad \text{for all } \boldsymbol{v}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}, \tag{3.70}$$

$$b(\tilde{\boldsymbol{u}}_h, q_h) = 0 \qquad \text{for all } q_h \in M_h. \tag{3.71}$$

Then, under assumptions (A-DG1)–(A-DG3), by standard results for $\boldsymbol{H}(\mathrm{div};\Omega)$ finite element discretizations of the Stokes and Stokes–Brinkman equations with an interior penalty [52, 99, 100], the pair $(\tilde{\boldsymbol{u}}_h, \tilde{p}_h)$ exists, is unique, and $\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0$, $\|p - \tilde{p}_h\|_{L^2(\Omega)} \to 0$ as $h \to 0$. □

We now fix an isolated local or global minimizer $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ of (BP) and extend the definition of the functional J in (BP) to functions $\boldsymbol{v} \notin H^1_{\boldsymbol{g}}(\Omega)^d$ by

$$J(\boldsymbol{v}, \eta) = +\infty \quad \text{for all} \quad \boldsymbol{v} \notin H^1_{\boldsymbol{g}}(\Omega)^d,\ \eta \in C_\gamma. \tag{3.72}$$

We also fix an $r > 0$ such that $(\boldsymbol{u}, \rho)$ is the unique minimizer of (BP) in $B_{r,\boldsymbol{H}(\mathrm{div};\Omega)\times L^2(\Omega)}(\boldsymbol{u}, \rho) \cap (\boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega) \times C_\gamma)$, where

$$\begin{aligned} &B_{r,\boldsymbol{H}(\mathrm{div};\Omega)\times L^2(\Omega)}(\boldsymbol{u}, \rho) \\ &:= \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega),\ \eta \in C_\gamma : \|\boldsymbol{u} - \boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div};\Omega)} + \|\rho - \eta\|_{L^2(\Omega)} \le r\}. \end{aligned} \tag{3.73}$$

Such an $r > 0$ is guaranteed to exist by the definition of an isolated minimizer and the extension of J in (3.72) to functions $\boldsymbol{v} \in \boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega)$ such that $\boldsymbol{v} \notin H^1_{\boldsymbol{g}}(\Omega)^d$. We also define $B_{r,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$ by

$$B_{r,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u}) := \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div};\Omega) : \|\boldsymbol{u} - \boldsymbol{v}\|_{\boldsymbol{H}(\mathrm{div};\Omega)} \le r\}. \tag{3.74}$$

We note that

$$\begin{aligned} &(\boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega) \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)) \\ &\qquad\qquad \subset B_{r,\boldsymbol{H}(\mathrm{div};\Omega)\times L^2(\Omega)}(\boldsymbol{u}, \rho) \cap (\boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega) \times C_\gamma) \end{aligned}$$

and hence $(\boldsymbol{u}, \rho)$ is also the unique minimizer in $(\boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega) \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho))$.

Proposition 3.9 (Weak convergence of $(\boldsymbol{u}_h, \rho_h)$ in $\boldsymbol{H}(\mathrm{div};\Omega) \times L^2(\Omega)$)**.** *Suppose that the conditions of Theorem 3.2 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP). For a given $h > 0$, consider the finite-dimensional optimization problem: find $(\boldsymbol{u}_h, \rho_h) \in (\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho))$ that minimizes*

$$J_h(\boldsymbol{u}_h, \rho_h). \tag{BP-DG$_h$}$$

Then, a global minimizer $(\boldsymbol{u}_h, \rho_h)$ *of* (BP-DG$_h$) *exists and there exist subsequences (up to relabeling) such that as* $h \to 0$*:*

$$\boldsymbol{u}_h \rightharpoonup \boldsymbol{u} \text{ weakly in } \boldsymbol{H}(\mathrm{div};\Omega), \tag{3.75}$$

$$\boldsymbol{u}_h \to \boldsymbol{u} \text{ strongly in } L^q(\Omega)^d, \tag{3.76}$$

$$\boldsymbol{u}_h \to \boldsymbol{u} \text{ strongly in } L^r(\partial\Omega)^d, \tag{3.77}$$

$$\rho_h \overset{*}{\rightharpoonup} \rho \text{ weakly-* in } L^\infty(\Omega), \tag{3.78}$$

$$\rho_h \rightharpoonup \rho \text{ weakly in } L^s(\Omega),\ s \in [1,\infty), \tag{3.79}$$

where $1 \le q, r < \infty$ *in two dimensions and* $1 \le q < 6$, $1 \le r < 4$ *in three dimensions.*

Proof. The functional J_h is continuous and

$$(\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)) \tag{3.80}$$

is a finite-dimensional, closed and bounded set and, therefore, sequentially compact by the Heine–Borel theorem. Hence, J_h obtains its infimum in $(\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho))$ and, therefore, a global minimizer $(\boldsymbol{u}_h, \rho_h)$ exists for all $h > 0$.

It can be checked that $\boldsymbol{H}(\mathrm{div};\Omega) \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$ and $C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)$ are norm-closed, bounded and convex subsets of the reflexive Banach spaces $\boldsymbol{H}(\mathrm{div};\Omega)$ and $L^2(\Omega)$, respectively. Therefore, by the arguments in the proof of Proposition 3.1, $\boldsymbol{H}(\mathrm{div};\Omega) \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$ is weakly sequentially compact in $\boldsymbol{H}(\mathrm{div};\Omega)$ and $C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)$ is weakly sequentially compact in $L^2(\Omega)$.

Hence we extract subsequences (up to relabeling), $(\boldsymbol{u}_h)$ and (ρ_h) of the sequence generated by the global minimizers of (BP-DG$_h$) such that

$$\boldsymbol{u}_h \rightharpoonup \hat{\boldsymbol{u}} \in \boldsymbol{H}(\mathrm{div};\Omega) \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u}) \text{ weakly in } \boldsymbol{H}(\mathrm{div};\Omega), \tag{3.81}$$

$$\rho_h \rightharpoonup \hat{\rho} \in C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho) \text{ weakly in } L^2(\Omega). \tag{3.82}$$

By assumption (A-DG3), there exists a sequence of finite element functions $\tilde{\rho}_h \in C_{\gamma,h}$ that strongly converges to ρ in $L^2(\Omega)$. Moreover, Lemma 3.2 implies the existence of a sequence $(\tilde{\boldsymbol{u}}_h) \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h}$ that satisfies $\|\boldsymbol{u} - \tilde{\boldsymbol{u}}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0$.

We now wish to identify the limits $\hat{\boldsymbol{u}}$ and $\hat{\rho}$. Consider the following bound:

$$
\begin{aligned}
&2|J_h(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h)-J(\boldsymbol{u},\rho)|\\
&\quad\le\int_\Omega |(\alpha(\rho)-\alpha(\tilde{\rho}_h))|\boldsymbol{u}|^2|+|\alpha(\tilde{\rho}_h)(|\boldsymbol{u}|^2-|\tilde{\boldsymbol{u}}_h|^2)|+2|\boldsymbol{f}\cdot(\boldsymbol{u}-\tilde{\boldsymbol{u}}_h)|\,\mathrm{d}x\\
&\qquad+\nu\sum_{K\in\mathcal{T}_h}\int_K\left||\nabla\boldsymbol{u}|^2-|\nabla\tilde{\boldsymbol{u}}_h|^2|\right|\,\mathrm{d}x\\
&\qquad+\sum_{F\in\mathcal{F}_h^i}\int_F\sigma h_F^{-1}|[\![\tilde{\boldsymbol{u}}_h]\!]_F|^2\,\mathrm{d}s+\sum_{F\in\mathcal{F}_h^\partial}\int_F\sigma h_F^{-1}|[\![\tilde{\boldsymbol{u}}_h-\boldsymbol{g}_h]\!]_F|^2\,\mathrm{d}s\\
&\qquad+2\sum_{F\in\mathcal{F}_h^i}\int_F|\{\!\{\nabla\tilde{\boldsymbol{u}}_h\}\!\}_F:[\![\tilde{\boldsymbol{u}}_h]\!]_F|\,\mathrm{d}s\\
&\qquad+2\sum_{F\in\mathcal{F}_h^\partial}\int_F|\{\!\{\nabla\tilde{\boldsymbol{u}}_h\}\!\}_F:[\![\tilde{\boldsymbol{u}}_h-\boldsymbol{g}_h]\!]_F|\,\mathrm{d}s.
\end{aligned}
\tag{3.83}
$$

For all $\boldsymbol{v}\in H^1(\mathcal{T}_h)^d$, $\boldsymbol{\Phi}\in(X_{\mathrm{DG}_k})^{d\times d}$, $h>0$, the following inequality holds [46, Lem. 7]

$$
\begin{aligned}
&\sum_{F\in\mathcal{F}_h}\int_F|\{\!\{\boldsymbol{\Phi}\}\!\}_F:[\![\boldsymbol{v}]\!]_F|\,\mathrm{d}s\\
&\le C\left(\sum_{F\in\mathcal{F}_h}\int_F h_F^{-1}|[\![\boldsymbol{v}]\!]_F|^2\,\mathrm{d}s\right)^{1/2}\left(\sum_{K\in\mathcal{T}_h}\|\boldsymbol{\Phi}\|_{L^2(K)}^2\right)^{1/2},
\end{aligned}
\tag{3.84}
$$

for a constant C that only depends on the mesh quality. Hence, we see that

$$
\begin{aligned}
&2|J_h(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h)-J(\boldsymbol{u},\rho)|\\
&\quad\le L_\alpha\|\boldsymbol{u}\|_{L^4(\Omega)}^2\|\tilde{\rho}_h-\rho\|_{L^2(\Omega)}+2\|\boldsymbol{f}\|_{L^2(\Omega)}\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{L^2(\Omega)}\\
&\qquad+\bar{\alpha}\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{L^2(\Omega)}(\|\tilde{\boldsymbol{u}}_h-\boldsymbol{u}\|_{L^2(\Omega)}+2\|\boldsymbol{u}\|_{L^2(\Omega)})\\
&\qquad+\nu\sum_{K\in\mathcal{T}_h}\|\nabla\tilde{\boldsymbol{u}}_h-\nabla\boldsymbol{u}\|_{L^2(K)}(\|\nabla\tilde{\boldsymbol{u}}_h-\nabla\boldsymbol{u}\|_{L^2(K)}+2\|\nabla\boldsymbol{u}\|_{L^2(K)})\\
&\qquad+\sum_{F\in\mathcal{F}_h^i}\int_F\sigma h_F^{-1}|[\![\tilde{\boldsymbol{u}}_h]\!]_F|^2\mathrm{d}s+\sum_{F\in\mathcal{F}_h^\partial}\int_F\sigma h_F^{-1}|[\![\tilde{\boldsymbol{u}}_h-\boldsymbol{g}_h]\!]_F|^2\mathrm{d}s\\
&\qquad+C\left(\sum_{F\in\mathcal{F}_h^i}\int_F h_F^{-1}|[\![\tilde{\boldsymbol{u}}_h]\!]_F|^2\mathrm{d}s\right)^{1/2}\left(\sum_{K\in\mathcal{T}_h}\|\nabla\tilde{\boldsymbol{u}}_h\|_{L^2(K)}^2\right)^{1/2}\\
&\qquad+C\left(\sum_{F\in\mathcal{F}_h^\partial}\int_F h_F^{-1}|[\![\tilde{\boldsymbol{u}}_h-\boldsymbol{g}_h]\!]_F|^2\mathrm{d}s\right)^{1/2}\left(\sum_{K\in\mathcal{T}_h}\|\nabla\tilde{\boldsymbol{u}}_h\|_{L^2(K)}^2\right)^{1/2}.
\end{aligned}
\tag{3.85}
$$

Thanks to the strong convergence of $\tilde{\boldsymbol{u}}_h$ in the broken $H^1_{\boldsymbol{g}}$-norm to $\boldsymbol{u}$ and by assumption (A-DG1), from (3.85) we deduce that

$$
J_h(\tilde{\boldsymbol{u}}_h,\tilde{\rho}_h)\to J(\boldsymbol{u},\rho)\ \text{ as }\ h\to 0.
$$

Furthermore, for sufficiently small $h > 0$, we note that

$$(\tilde{\boldsymbol{u}}_h, \tilde{\rho}_h) \in (\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)).$$

Therefore, since $(\boldsymbol{u}_h, \rho_h)$ is a global minimizer of (BP-DG$_h$) in $B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho))$,

$$J_h(\boldsymbol{u}_h, \rho_h) \leq J_h(\tilde{\boldsymbol{u}}_h, \tilde{\rho}_h). \tag{3.86}$$

By taking the limit as $h \to 0$ and utilizing the strong convergence of $\tilde{\boldsymbol{u}}_h$ and $\tilde{\rho}_h$ to $\boldsymbol{u}$ and ρ, respectively, we see that

$$\lim_{h\to 0} J_h(\boldsymbol{u}_h, \rho_h) \leq J(\boldsymbol{u}, \rho). \tag{3.87}$$

By construction $\hat{\rho} \in C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)$. By assumption (A-DG3), for every $q \in L^2_0(\Omega)$, there exists a sequence of $\tilde{q}_h \in M_h$ such that $\tilde{q}_h \to q$ strongly in $L^2(\Omega)$. Since $\boldsymbol{u}_h \rightharpoonup \hat{\boldsymbol{u}}$ weakly in $\boldsymbol{H}(\mathrm{div};\Omega)$ and $\boldsymbol{u}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h}$, we see that

$$b(\hat{\boldsymbol{u}}, q) = \lim_{h\to 0} b(\boldsymbol{u}_h, \tilde{q}_h) + \lim_{h\to 0} b(\boldsymbol{u}_h, q - \tilde{q}_h) = 0 \ \text{ for all } \ q \in L^2_0(\Omega). \tag{3.88}$$

Hence, $\hat{\boldsymbol{u}}$ is pointwise divergence-free. The final step to identify $\hat{\boldsymbol{u}}$ as $\boldsymbol{u}$ is to show that $\hat{\boldsymbol{u}} \in H^1_{\boldsymbol{g}}(\Omega)^d$. Now, the sequence $(\boldsymbol{u}_h)$ also defines a bounded sequence in $H^1(\mathcal{T}_h)^d$ such that

$$\sup_{h>0} \left[\|\boldsymbol{u}_h\|_{L^1(\Omega)} + |\boldsymbol{u}_h|_{H^1(\mathcal{T}_h)} \right] < +\infty.$$

Hence, by the compact embedding lemma from Buffa and Ortner [46, Lem. 8], there exists a subsequence (up to relabeling) and a limit $\hat{\boldsymbol{w}} \in H^1(\Omega)^d$ such that

$$\boldsymbol{u}_h \to \hat{\boldsymbol{w}} \ \text{ strongly in } \ L^q(\Omega)^d, \tag{3.89}$$

where $1 \leq q < \infty$ in two dimensions and $1 \leq q < 6$ in three dimensions. By the uniqueness of limits $\hat{\boldsymbol{w}} = \hat{\boldsymbol{u}}$ a.e. in Ω and thus $\hat{\boldsymbol{u}} \in H^1(\Omega)^d$. Moreover, the same compact embedding lemma implies that

$$\boldsymbol{u}_h \to \hat{\boldsymbol{u}} \ \text{ strongly in } \ L^r(\partial\Omega)^d, \tag{3.90}$$

where $1 \leq r < \infty$ in two dimensions and $1 \leq r < 4$ in three dimensions. If $\|\boldsymbol{u}_h - \boldsymbol{g}\|_{L^2(\partial\Omega)} \not\to 0$, then $J_h(\boldsymbol{u}_h, \rho_h) \to +\infty$. Since $(\boldsymbol{u}_h)$ is a bounded sequence, we must have that $\|\boldsymbol{u}_h - \boldsymbol{g}\|_{L^2(\partial\Omega)} \to 0$. Hence,

$$\|\hat{\boldsymbol{u}} - \boldsymbol{g}\|_{L^2(\partial\Omega)} \leq \|\hat{\boldsymbol{u}} - \boldsymbol{u}_h\|_{L^2(\partial\Omega)} + \|\boldsymbol{u}_h - \boldsymbol{g}\|_{L^2(\partial\Omega)} \to 0. \tag{3.91}$$

Thus, (3.88), (3.89), and (3.91) imply that $\hat{\boldsymbol{u}} \in \boldsymbol{U} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$. Since $(\boldsymbol{u}, \rho)$ is the unique local minimizer of (BP) in

$$(\boldsymbol{H}_{\boldsymbol{g},\mathrm{div}}(\mathrm{div};\Omega) \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})) \times (C_\gamma \cap B_{r/2,L^2(\Omega)}(\rho)),$$

from (3.87), we see that

$$\lim_{h\to 0} J_h(\boldsymbol{u}_h, \rho_h) = J(\boldsymbol{u}, \rho). \tag{3.92}$$

Since $(\boldsymbol{u}, \rho)$ is the unique minimizer in the spaces we consider, we identify the limits $\hat{\boldsymbol{u}}$ and $\hat{\rho}$ as $\boldsymbol{u}$ and ρ, respectively, and state that $\boldsymbol{u}_h \rightharpoonup \boldsymbol{u}$ weakly in $\boldsymbol{H}(\mathrm{div};\Omega)$, $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $L^q(\Omega)^d$, $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $L^r(\partial\Omega)^d$ and $\rho_h \rightharpoonup \rho$ weakly in $L^2(\Omega)$, where $1 \leq q, r < \infty$ in two dimensions and $1 \leq q < 6$, $1 \leq r < 4$ in three dimensions.

By the same argument as in the proof of Proposition 3.1, we conclude that $\rho_h \overset{*}{\rightharpoonup} \rho$ weakly-* in $L^\infty(\Omega)$ and $\rho_h \rightharpoonup \rho$ weakly in $L^s(\Omega)$ for all $s \in [1, \infty)$. □

Proposition 3.10 (Strong convergence of ρ_h in $L^s(\Omega)$, $s \in [1,\infty)$)**.** *Suppose that the conditions stated in Theorem 3.2 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP). Then, there exists a subsequence of minimizers, (ρ_h), of (BP-DG$_h$) such that*

$$\rho_h \to \rho \text{ strongly in } L^s(\Omega), \quad s \in [1, \infty). \tag{3.93}$$

The proof of Proposition 3.10 follows from the proof of Proposition 3.2 with some small modifications.

Proposition 3.11 (Strong convergence of $\boldsymbol{u}_h$ in the $H^1_{\boldsymbol{g}}(\mathcal{T}_h)$-norm)**.** *Suppose that the conditions stated in Theorem 3.2 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP). Then, there exists a subsequence of minimizers, $(\boldsymbol{u}_h)$, of (BP-DG$_h$) such that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0. \tag{3.94}$$

Proof. We note that $\boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$ is a convex set, and hence for any $\boldsymbol{w}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$, $t \in [0,1]$, we have that $\boldsymbol{u}_h + t(\boldsymbol{w}_h - \boldsymbol{u}_h) \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$. Since $(\boldsymbol{u}_h, \rho_h)$ is a global minimizer of (BP-DG$_h$), we note that

$$\frac{1}{t}\left[J_h(\boldsymbol{u}_h + t(\boldsymbol{w}_h - \boldsymbol{u}_h), \rho_h) - J_h(\boldsymbol{u}_h, \rho_h)\right] \geq 0. \tag{3.95}$$

By taking the limit $t \to 0$, a calculation shows that, for all $\boldsymbol{w}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$,

$$a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{w}_h - \boldsymbol{u}_h) \geq l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{w}_h - \boldsymbol{u}_h). \tag{3.96}$$

We note that $\boldsymbol{w}_h - \boldsymbol{u}_h \in \boldsymbol{U}_{0,\boldsymbol{n},h}$. Hence, from Proposition 3.7 and Lemma 3.1, we deduce that

$$a_{h,\rho}(\boldsymbol{u}, \boldsymbol{w}_h - \boldsymbol{u}_h) = l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{w}_h - \boldsymbol{u}_h). \tag{3.97}$$

Therefore, from (3.96) and (3.97), we see that

$$\begin{aligned} a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) \leq\ & a_{h,\rho}(\boldsymbol{u}, \boldsymbol{u}_h - \boldsymbol{w}_h) \\ & + l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{u}_h - \boldsymbol{w}_h) - l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{u}_h - \boldsymbol{w}_h). \end{aligned} \tag{3.98}$$

Hence, by subtracting $a_{h,\rho_h}(\boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h)$ from both sides of (3.98), and utilizing the coercivity of $a_{h,\rho_h}(\cdot,\cdot)$ as stated in Proposition 3.8, we have that

$$\begin{aligned} c_a\|\boldsymbol{u}_h - \boldsymbol{w}_h\|^2_{H^1(\mathcal{T}_h)} \leq\ & a_{h,\rho}(\boldsymbol{u}, \boldsymbol{u}_h - \boldsymbol{w}_h) - a_{h,\rho_h}(\boldsymbol{w}_h, \boldsymbol{u}_h - \boldsymbol{w}_h) \\ & + l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{u}_h - \boldsymbol{w}_h) - l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{u}_h - \boldsymbol{w}_h). \end{aligned} \tag{3.99}$$

Now, by assumption (M3), for all $F \in \mathcal{F}_h^\partial$, there exists a $c > 0$ such that $h_F^{-1} \leq ch^{-1}$, where c depends on the mesh regularity. By taking the absolute value of the right-hand side of (3.99), collecting terms, utilizing the inequality (3.84), and the boundedness of a_{h,ρ_h} by C_a as stated in Proposition 3.8, we have that

$$\begin{aligned} c_a\|\boldsymbol{u}_h - \boldsymbol{w}_h\|^2_{H^1(\mathcal{T}_h)} \leq\ & \bar{\alpha}\|\boldsymbol{u} - \boldsymbol{w}_h\|_{L^2(\Omega)}\|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{L^2(\Omega)} \\ & + \|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)}\|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{L^2(\Omega)} \\ & + C_a\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}\|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)} \\ & + Ch^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)}\|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{L^2(\partial\Omega)} \\ & + Ch^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)}\|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}, \end{aligned} \tag{3.100}$$

for some constant $C = C(\sigma)$ that also depends on the mesh regularity.

We note that $\|\boldsymbol{u} - \boldsymbol{w}_h\|_{L^2(\Omega)} \leq \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}$ by definition. Moreover, by the broken trace theorem, as found in Buffa and Ortner [46, Th. 4.4], there exists a constant C_{BT} such that, for all $\boldsymbol{v} \in H^1(\mathcal{T}_h)^d$, $d \in \{2, 3\}$ we have

$$\|\boldsymbol{v}\|_{L^2(\partial\Omega)} \leq C_{\mathrm{BT}} \|\boldsymbol{v}\|_{H^1(\mathcal{T}_h)}. \tag{3.101}$$

Therefore, by bounding the $L^2(\Omega)$ and $L^2(\partial\Omega)$-norms of $\boldsymbol{u}_h - \boldsymbol{w}_h$ above by the broken H^1-norm, and dividing through by $c_a \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}$ we see that

$$\begin{aligned} \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)} \leq C\|\boldsymbol{u} - \boldsymbol{w}_h\|_{L^2(\Omega)} + C\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \\ + C\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)} + Ch^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)}, \end{aligned} \tag{3.102}$$

for some constant C that depends on $c_a, C_a, C_{\mathrm{BT}}, \bar{\alpha}, \sigma$ and the mesh regularity.

For sufficiently small h, we note that $\tilde{\boldsymbol{u}}_h \in \boldsymbol{U}_{\boldsymbol{g}_h, \boldsymbol{n}, h} \cap B_{r/2, \boldsymbol{H}(\mathrm{div};\Omega)}(\boldsymbol{u})$ (where $\tilde{\boldsymbol{u}}_h$ is defined in Lemma 3.2) and $\tilde{\boldsymbol{u}}_h \to \boldsymbol{u}$ strongly in $H^1(\mathcal{T}_h)^d$. Fix $\boldsymbol{w}_h = \tilde{\boldsymbol{u}}_h$.

From Proposition 3.10, we know that there exists a subsequence (not indicated) such that $\rho_h \to \rho$ strongly in $L^4(\Omega)$. We now observe that

$$\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \leq L_\alpha \|\rho - \rho_h\|_{L^4(\Omega)} \|\boldsymbol{u}\|_{L^4(\Omega)}. \tag{3.103}$$

$\|\boldsymbol{u}\|_{L^4(\Omega)}$ is bounded for $d \in \{2, 3\}$ thanks to the Sobolev embedding theorem. Hence, by taking the limit as $h \to 0$ in (3.102), from (A-DG1), Lemma 3.2, and (3.103), we deduce that $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\mathcal{T}_h)} \to 0$ as $h \to 0$. In Proposition 3.9, we showed that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $L^2(\partial\Omega)^d$. Hence, we conclude that $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0$ as $h \to 0$. □

Proposition 3.12 (Discretized first-order optimality conditions)**.** *Suppose that the conditions stated in Theorem 3.2 hold. Then, there exists an $\bar{h} > 0$ such that for all $h < \bar{h}$, there exists a unique Lagrange multiplier $p_h \in M_h$ such that the functions $(\boldsymbol{u}_h, \rho_h)$, that minimize (BP-DG$_h$), satisfy the first-order optimality conditions (FOC1-DG$_h$)–(FOC3a-DG$_h$).*

Proof. The goal is to utilize the strong convergence proven in Proposition 3.11 to obtain an equation in weak form for $\boldsymbol{u}_h$ tested against functions in $\boldsymbol{U}_{0,\boldsymbol{n},h}$. Then from the proof of Proposition 2.4, we can deduce the existence of p_h.

From Proposition 3.11, we know that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\mathcal{T}_h)^d$. Hence by definition of strong convergence, there exists an $\bar{h}_1 > 0$ such that, for all $h \leq \bar{h}_1$, $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\mathcal{T}_h)} \leq r/4$. Moreover, since $\boldsymbol{u} \in \boldsymbol{U}$, we have that $\operatorname{div}(\boldsymbol{u}) = 0$ a.e. in Ω and by Lemma 3.1, we have that $\operatorname{div}(\boldsymbol{u}_h) = 0$ a.e. in Ω. Therefore,

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{\boldsymbol{H}(\operatorname{div};\Omega)} &= \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{L^2(\Omega)} + \|\operatorname{div}(\boldsymbol{u} - \boldsymbol{u}_h)\|^2_{L^2(\Omega)} \\ &= \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{L^2(\Omega)} \leq \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{H^1(\mathcal{T}_h)} \leq r^2/16. \end{aligned} \tag{3.104}$$

Hence, for each $\boldsymbol{v}_h \in \boldsymbol{U}_{0,\boldsymbol{n},h}$, if $|t| < r/(4\|\boldsymbol{v}_h\|_{H^1(\mathcal{T}_h)})$ then $\boldsymbol{u}_h + t\boldsymbol{v}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,\boldsymbol{n},h} \cap B_{r/2,\boldsymbol{H}(\operatorname{div};\Omega)}(\boldsymbol{u})$. From Proposition 3.10 we have that $\rho_h \to \rho$ strongly in $L^2(\Omega)$. Hence, there exists an $\bar{h}_2 > 0$ such that, for all $h \leq \bar{h}_2$, $\|\rho - \rho_h\|_{L^2(\Omega)} \leq r/4$. Therefore, for each $\eta_h \in C_{\gamma,h}$, if $0 < t < r/(4\|\eta_h - \rho_h\|_{L^2(\Omega)})$ then $\rho_h + t(\eta_h - \rho_h) \in C_{\gamma,h} \cap B_{r/2,L^2(\Omega)}(\rho)$. Let $\bar{h} = \min\{\bar{h}_1, \bar{h}_2\}$ and consider $h \leq \bar{h}$.

Since $(\boldsymbol{u}_h, \rho_h)$ is a global minimizer of (BP-DG$_h$), then, for all $\boldsymbol{v}_h \in \boldsymbol{U}_{0,\boldsymbol{n},h}$, if $|t| < r/(4\|\boldsymbol{v}_h\|_{H^1(\mathcal{T}_h)})$ we have

$$\frac{1}{t}\left[J_h(\boldsymbol{u}_h + t\boldsymbol{v}_h, \rho_h) - J_h(\boldsymbol{u}_h, \rho_h)\right] \geq 0. \tag{3.105}$$

By considering the limits for $t \to 0_+$ and $t \to 0_-$, we have that, for all $\boldsymbol{v}_h \in \boldsymbol{U}_{0,\boldsymbol{n},h}$,

$$a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{v}_h) = l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{v}_h). \tag{3.106}$$

From (3.106), the existence of a unique $p_h \in M_h$ such that $(\boldsymbol{u}_h, \rho_h, p_h)$ satisfy (FOC1-DG$_h$)–(FOC2-DG$_h$) follows from the inf-sup condition (A-DG2) and the argument given in the proof of Proposition 2.4.

Similarly, since $(\boldsymbol{u}_h, \rho_h)$ is a global minimizer of (BP-DG$_h$), then, for all $\eta_h \in C_{\gamma,h}$, if $0 < t < r/(4\|\eta_h - \rho_h\|_{L^2(\Omega)})$ we have

$$\frac{1}{t}\left[J_h(\boldsymbol{u}_h, \rho_h + t(\eta_h - \rho_h)) - J_h(\boldsymbol{u}_h, \rho_h)\right] \geq 0. \tag{3.107}$$

By taking the limit as $t \to 0$, we deduce that (FOC3a-DG$_h$) holds. □

Proposition 3.13. *Suppose that the conditions stated in Theorem 3.2 hold. Then, there is a subsequence of the unique $p_h \in M_h$ defined in Proposition 3.12 that converges strongly in $L^2(\Omega)$ to the $p \in L^2_0(\Omega)$ that solves (FOC1)–(FOC3a) for the given isolated minimizer $(\boldsymbol{u}, \rho)$.*

Proof. The inf-sup condition (A-DG2) for M_h and $\boldsymbol{X}_{0,\boldsymbol{n},h}$ implies that, for any $q_h \in M_h$,

$$\begin{aligned} c_b\|q_h - p_h\|_{L^2(\Omega)} &\leq \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}\setminus\{0\}} \frac{b(\boldsymbol{w}_h, q_h - p_h)}{\|\boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}} \\ &= \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}\setminus\{0\}} \frac{b(\boldsymbol{w}_h, p - p_h) + b(\boldsymbol{w}_h, q_h - p)}{\|\boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}} \\ &\leq \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}\setminus\{0\}} \frac{|b(\boldsymbol{w}_h, p - p_h)| + |b(\boldsymbol{w}_h, q_h - p)|}{\|\boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}}. \end{aligned} \tag{3.108}$$

From Proposition 3.7 and Proposition 3.12, it follows that

$$b(\boldsymbol{w}_h, p - p_h) = a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{w}_h) - a_{h,\rho}(\boldsymbol{u}, \boldsymbol{w}_h) + l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{w}_h) - l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{w}_h). \tag{3.109}$$

Therefore,

$$\begin{aligned} &c_b\|q_h - p_h\|_{L^2(\Omega)} \\ &\leq \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}\setminus\{0\}} \frac{|a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{w}_h) - a_{h,\rho}(\boldsymbol{u}, \boldsymbol{w}_h)|}{\|\boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}} \\ &\quad + \sup_{\boldsymbol{w}_h \in \boldsymbol{X}_{0,\boldsymbol{n},h}\setminus\{0\}} \frac{|l_{h,\boldsymbol{f},\boldsymbol{g}}(\boldsymbol{w}_h) - l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{w}_h)| + |b(\boldsymbol{w}_h, q_h - p)|}{\|\boldsymbol{w}_h\|_{H^1(\mathcal{T}_h)}}. \end{aligned} \tag{3.110}$$

By using the same argument we used to bound (3.100) by (3.102), we see that (3.110) implies that

$$\begin{aligned} c_b\|q_h - p_h\|_{L^2(\Omega)} \leq C\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + C\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\mathcal{T}_h)} \\ + Ch^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)} + C_b\|p - q_h\|_{L^2(\Omega)}. \end{aligned} \tag{3.111}$$

where C_b is the boundedness constant for $b(\cdot,\cdot)$ and C is dependent on C_a, C_{BT}, $\bar{\alpha}$, L_α, ν, σ and the mesh regularity. Hence, by an application on the Cauchy–Schwarz inequality,

$$\begin{aligned} \|p - p_h\|_{L^2(\Omega)} \leq C\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + C\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\mathcal{T}_h)} \\ + Ch^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)} + C\|p - q_h\|_{L^2(\Omega)}, \end{aligned} \tag{3.112}$$

where C is dependent on c_b, C_b, C_a, C_{BT}, $\bar{\alpha}$, L_α, ν, σ and the mesh regularity.

By assumption (A-DG3), there exists a sequence of finite element functions, $\tilde{p}_h \in M_h$ that satisfies $\tilde{p}_h \to p$ strongly in $L^2(\Omega)$. Let $q_h = \tilde{p}_h$. We have already shown that $\boldsymbol{u}_h \to \boldsymbol{u}$ strongly in $H^1(\mathcal{T}_h)^d$ in Proposition 3.11. Similarly, in the proof of Proposition 3.11 we also showed that $\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \to 0$. By assumption (A-DG1), $h^{-1}\|\boldsymbol{g} - \boldsymbol{g}_h\|_{L^2(\partial\Omega)} \to 0$. Hence, we conclude that $p_h \to p$ strongly in $L^2(\Omega)$. □

We now have the required results to prove Theorem 3.2.

Proof of Theorem 3.2. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP) and its unique associated Lagrange multiplier p. By the results of Propositions 3.9, 3.10, 3.11, and 3.12, there exists a mesh size $\bar{h}$ such that for, $h < \bar{h}$, there exists a sequence of finite element solutions $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times C_{\gamma,h} \times M_h$ satisfying (FOC1-DG$_h$)–(FOC3a-DG$_h$) that converges to $(\boldsymbol{u}, \rho, p)$. By taking a subsequence if necessary (not indicated), Proposition 3.10, implies that $\rho_h \to \rho$ strongly in $L^s(\Omega)$, $s \in [1, \infty)$, Proposition 3.11 implies that $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1_{\boldsymbol{g}}(\mathcal{T}_h)} \to 0$, and Proposition 3.13 implies that $p_h \to p$ strongly in $L^2(\Omega)$. □

Corollary 3.3 (Relaxing the volume constraint)**.** *Suppose that the conditions of Theorem 3.2 hold. Fix an isolated minimizer $(\boldsymbol{u}, \rho)$ of (BP) and its associated Lagrange multiplier p. Consider the finite element space $C_{[0,1],h} \subset C_{[0,1]}$ with $C_{[0,1]}$ as defined in Proposition 2.5. Then, there exists a sequence $(\lambda_h) \in \mathbb{R}$ such that the sequence $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times C_{\gamma,h} \times M_h$, satisfying (FOC1-DG$_h$)–(FOC3a-DG$_h$), that converges to $(\boldsymbol{u}, \rho, p)$ strongly in $H^1_{\boldsymbol{g}}(\mathcal{T}_h)^d \times L^s(\Omega) \times L^2(\Omega)$, $s \in [1, \infty)$, also satisfies, for all $(\eta_h, \zeta_h) \in C_{[0,1],h} \times \mathbb{R}$,*

$$c_{\boldsymbol{u}_h,\lambda_h}(\rho_h, \eta_h - \rho_h) \geq 0, \qquad \text{(FOC3b-DG}_h\text{)}$$

$$d_{\rho_h}(\lambda_h, \zeta_h) = 0, \qquad \text{(FOC4-DG}_h\text{)}$$

with $c_{\boldsymbol{u}_h,\lambda_h}$ and d_{ρ_h} as defined in Proposition 2.5.

Proof. By construction $(\boldsymbol{u}_h, \rho_h)$ minimizes (BP-DG$_h$). Hence, the result can be deduced as in the proof of Proposition 2.5. □

For each isolated minimizer, Theorems 3.1 and 3.2 guarantee the existence of a sequence of mesh sizes $h_0 > h_1 > h_2 > \cdots$ such that the finite element solutions for that sequence, $(\boldsymbol{u}_{h_i}, \rho_{h_i}, p_{h_i})$, satisfy the discretized first-order optimality conditions and strongly converge to the analytical isolated minimizer in $H^1(\Omega)^d \times L^s(\Omega) \times L^2(\Omega)$, $s \in [1, \infty)$, for a conforming discretization and in $H^1_{\boldsymbol{g}}(\mathcal{T}_h)^d \times L^s(\Omega) \times L^2(\Omega)$, $s \in [1, \infty)$, for a divergence-free DG discretization. However, the uniqueness of the finite element solutions is not guaranteed. In practice, it is possible that for a given sequence of mesh sizes h_i, there exist two sequences of finite element solutions, $(\boldsymbol{u}_{h_i}, \rho_{h_i}, p_{h_i})$ and $(\hat{\boldsymbol{u}}_{h_i}, \hat{\rho}_{h_i}, \hat{p}_{h_i})$, that strongly converge to the same isolated minimizer. This possibility is discussed and numerically verified in Section 4.6.1.

3.3 Error bounds

In the following section we restrict ourselves to a conforming discretization and derive error bounds that depend on the L^2-norm error of the velocity. Hence, if we have an estimate of the convergence rate of the L^2-norm error of the velocity, we can deduce the convergence rates for the H^1-norm error of the velocity, the L^2-norm error of the material distribution, and the L^2-norm error of the pressure.

Theorem 3.3 (Error bounds)**.** *Suppose that the conditions stated in Theorem 3.1 hold,* $(\boldsymbol{u}, \rho)$ *is an isolated minimizer of* (BP) *and* $(\boldsymbol{u}, \rho, p)$ *satisfy the first-order optimality conditions* (FOC1)–(FOC3a)*. Moreover, let the domain be either a convex polygon in two dimensions or a convex polyhedron in three dimensions. Let* $(\boldsymbol{u}_h, \rho_h, p_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,h} \times C_{\gamma,h} \times M_h$ *denote finite element solutions satisfying* (FOC1_h)–(FOC3a_h) *that converge to* $(\boldsymbol{u}, \rho, p)$ *as* $h \to 0$*. Furthermore, suppose that the support of* ρ *is compactly contained in the support of* $\boldsymbol{u}$ *and there exists a mesh size* $h' > 0$ *such that, for all* $h < h'$*, the support of* ρ_h *is also compactly contained in the support of* $\boldsymbol{u}$*. Moreover, assume that the boundary data* $\boldsymbol{g}$ *is the restriction of a function* $\hat{\boldsymbol{g}} \in H^2(\Omega)^d$ *on the boundary* $\partial\Omega$*, where* $\operatorname{div}(\hat{\boldsymbol{g}}) = 0$ *a.e. in* Ω*.*

Then, for any $(\boldsymbol{w}_h, \eta_h, q_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,h} \times C_{\gamma,h} \times M_h$, $h < h'$, *the following error bounds hold:*

$$\begin{aligned}\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} \leq C \Big(&\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} + \|\rho - \eta_h\|^{1/2}_{L^2(\Omega)} \\ &+ \|p - q_h\|_{L^2(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|^r_{L^2(\Omega)} \Big),\end{aligned} \tag{3.113}$$

$$\|\rho - \rho_h\|_{L^2(\Omega)} \leq C \left(\|\rho - \eta_h\|^{1/2}_{L^2(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|^r_{L^2(\Omega)} \right), \tag{3.114}$$

$$\begin{aligned}\|p - p_h\|_{L^2(\Omega)} \leq C \Big(&\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} + \|\rho - \eta_h\|^{1/2}_{L^2(\Omega)} \\ &+ \|p - q_h\|_{L^2(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|^r_{L^2(\Omega)} \Big),\end{aligned} \tag{3.115}$$

where C *is a constant independent of* h, $r = 1 - \epsilon$, *for any* $0 < \epsilon < 1$, *in two dimensions, and* $r = 3/4$ *in three dimensions.*

Proof. (Material distribution error bound). From (3.31), we know that, for any $\theta > 0$,

$$\|\rho - \rho_h\|^2_{L^2(U_\theta)} \leq C \left(\|\rho - \eta_h\|_{L^2(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(\Omega)} \|\rho - \rho_h\|_{L^q(\Omega)} \right), \tag{3.116}$$

where C is a constant dependent on $\alpha''_{\min}, \alpha'_{\max}, \theta$, $\|\boldsymbol{u}\|_{L^4(\Omega)}$, and $\|\boldsymbol{u}\|_{L^q(\Omega)}$, $2 < q < \infty$ in two dimensions, and $3 \leq q < \infty$ in three dimensions. We note that since ρ, $\rho_h \in C_\gamma$,

$$\|\rho - \rho_h\|^q_{L^q(\Omega)} = \int_\Omega |\rho - \rho_h|^{q-2} |\rho - \rho_h|^2 \mathrm{d}x \leq \|\rho - \rho_h\|^2_{L^2(\Omega)}. \tag{3.117}$$

By assumption, the supports of ρ and ρ_h are compactly contained in the support of $\boldsymbol{u}$ and thus there exists a $\theta'(h') > 0$ such that $\operatorname{supp}(\rho)$, $\operatorname{supp}(\rho_h) \subset U_{\theta'}$ for all $h < h'$. Therefore, by an application of Young's inequality with ϵ, from (3.116) and (3.117) we deduce that

$$\begin{aligned}\|\rho - \rho_h\|^2_{L^2(\Omega)} &= \|\rho - \rho_h\|^2_{L^2(U_{\theta'})} \\ &\leq C \left(\|\rho - \eta_h\|_{L^2(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(\Omega)} \|\rho - \rho_h\|^{2/q}_{L^2(\Omega)} \right) \\ &\leq C \left(\|\rho - \eta_h\|_{L^2(\Omega)} + c(\epsilon) \|\boldsymbol{u} - \boldsymbol{u}_h\|^{2r}_{L^2(\Omega)} + \epsilon \|\rho - \rho_h\|^2_{L^2(\Omega)} \right),\end{aligned} \tag{3.118}$$

where $c(\epsilon) = (\epsilon q)^{-2r/q} (2r)^{-1}$ and r is half the Hölder conjugate of q, i.e. $0 < r < 1$ in two dimensions and $0 < r \leq 3/4$ in three dimensions. By fixing $\epsilon = C^{-1}/2$, moving the third term over to the left-hand side, taking the square root, and applying the Cauchy–Schwarz inequality, we deduce that (3.114) holds.

(Velocity error bound). Proposition 3.4 implies that

$$a_{\rho_h}(\boldsymbol{u}_h, \boldsymbol{w}_h - \boldsymbol{u}_h) - l_{\boldsymbol{f}}(\boldsymbol{w}_h - \boldsymbol{u}_h) = 0 \text{ for all } \boldsymbol{w}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,h}. \tag{3.119}$$

Hence by following the proof of Proposition 3.3, and replacing $\boldsymbol{W}_h$ by $\boldsymbol{U}_{\boldsymbol{g}_h,h}$ until (3.42), we conclude that, for all $\boldsymbol{w}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,h}$,

$$\begin{aligned} &\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} \\ &\quad \leq C \left(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} + \|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + \|p - q_h\|_{L^2(\Omega)} \right). \end{aligned} \tag{3.120}$$

Given the assumptions stated in this theorem, Proposition 2.6 implies that $\boldsymbol{u} \in H^2(\Omega)^d$. By the Sobolev embedding theorem we see that $\boldsymbol{u} \in L^\infty(\Omega)^d$. Thus,

$$\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} \leq L_\alpha \|\boldsymbol{u}\|_{L^\infty(\Omega)} \|\rho - \rho_h\|_{L^2(\Omega)}. \tag{3.121}$$

Hence, by (3.120) and (3.121), we deduce that

$$\begin{aligned} &\|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} \\ &\quad \leq C \left(\|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} + \|\rho - \rho_h\|_{L^2(\Omega)} + \|p - q_h\|_{L^2(\Omega)} \right). \end{aligned} \tag{3.122}$$

It is known that the inf-sup condition implies that $\inf_{\boldsymbol{w}_h \in \boldsymbol{U}_{\boldsymbol{g}_h,h}} \|\boldsymbol{u} - \boldsymbol{w}_h\|_{H^1(\Omega)} \leq C\|\boldsymbol{u} - \boldsymbol{v}_h\|_{H^1(\Omega)}$ for all $\boldsymbol{v}_h \in \boldsymbol{X}_{\boldsymbol{g}_h,h}$ [39, Th. 12.5.17]. By substituting (3.114) into (3.122), we deduce that (3.113) holds.

(Pressure error bound). In Proposition 3.5 we showed that

$$\begin{aligned} &\|p - p_h\|_{L^2(\Omega)} \\ &\quad \leq C \left(\|(\alpha(\rho) - \alpha(\rho_h))\boldsymbol{u}\|_{L^2(\Omega)} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{H^1(\Omega)} + \|p - q_h\|_{L^2(\Omega)} \right). \end{aligned} \tag{3.123}$$

By applying (3.121), and then by substituting (3.114) and (3.113) into (3.123), we conclude that (3.115) holds. □

Corollary 3.4. *Suppose that the conditions stated in Theorems 2.12 and 3.3 hold. Consider the Taylor–Hood finite element discretization,* $(\mathrm{CG}_2)^d \times \mathrm{CG}_1$*, for the velocity and pressure and any conforming finite element space for the material distribution* ρ*. Provided* $\rho \in H^1(\Omega)$*, suppose that the following best approximation error bound holds:*

$$\inf_{\eta_h \in C_{\gamma,h}} \|\rho - \eta_h\|_{L^2(\Omega)} \leq Ch\|\rho\|_{H^1(\Omega)}. \tag{3.124}$$

Then, the error bounds satisfy the following rate of convergence:

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{H^1(\Omega)}+\|\rho-\rho_h\|_{L^2(\Omega)}+\|p-p_h\|_{L^2(\Omega)}&\\ \leq \mathcal{O}(h^{1/2})+\mathcal{O}(\|\boldsymbol{u}-\boldsymbol{u}_h\|^r_{L^2(\Omega)}),&\end{aligned} \tag{3.125}$$

where $r = 1-\epsilon$, *for any* $0 < \epsilon < 1$ *in two dimensions and* $r = 3/4$ *in three dimensions.*

Proof. Proposition 2.6 implies that $\boldsymbol{u} \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$. Moreover, since the conditions stated in Theorem 2.12 hold and, by assumption, the support of ρ is compactly contained in the support of $\boldsymbol{u}$, then $\rho \in H^1(\Omega)$. Since $\boldsymbol{w}_h \in \boldsymbol{X}_{0,h}$ and $q_h \in M_h$ in (3.113)–(3.115) are arbitrary, by well-known approximation results (e.g. using the nodal interpolant for $\boldsymbol{u}$ and the Scott–Zhang interpolant for p) [39, Ch. 4]

$$\inf_{\boldsymbol{w}_h \in X_{0,h}} \|\boldsymbol{u}-\boldsymbol{w}_h\|_{H^1(\Omega)} \leq Ch\|\boldsymbol{u}\|_{H^2(\Omega)}, \tag{3.126}$$

$$\inf_{q_h \in M_h} \|p-q_h\|_{L^2(\Omega)} \leq Ch\|p\|_{H^1(\Omega)}. \tag{3.127}$$

By assumption an approximation result exists for ρ. Substituting (3.124), (3.126), and (3.127) into (3.113)–(3.115), we obtain (3.125). □

Corollary 3.5. *Suppose that the conditions of Corollary 3.4 hold and we fix the finite element space approximating the material distribution to* CG_1. *Moreover, suppose that the material distribution has additional regularity,* $\rho \in C_\gamma \cap H^2(\Omega)$ *and the following best approximation result holds*

$$\inf_{\eta_h \in C_{\gamma,h}} \|\rho-\eta_h\|_{L^2(\Omega)} \leq Ch^2\|\rho\|_{H^2(\Omega)}. \tag{3.128}$$

Then, the convergence rate (3.125) *can be sharpened to*

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{H^1(\Omega)}+\|\rho-\rho_h\|_{L^2(\Omega)}+\|p-p_h\|_{L^2(\Omega)}&\\ \leq \mathcal{O}(h)+\mathcal{O}(\|\boldsymbol{u}-\boldsymbol{u}_h\|^r_{L^2(\Omega)}),&\end{aligned} \tag{3.129}$$

where $r = 1-\epsilon$, *for any* $0 < \epsilon < 1$ *in two dimensions and* $r = 3/4$ *in three dimensions.*

Proof. The proof follows the same reasoning as the proof for Corollary 3.4, but applications of (3.124) are replaced with (3.128). □

Remark 3.4. *The best approximation bounds* (3.124) *and* (3.128) *are standard results in interpolation theory if we were to relax the function space of the material distribution from* C_γ *to* $L^\infty(\Omega)$ *[39, Ch. 4.8].*

The L^2-norm error of the velocity on the right-hand sides of (3.125) and (3.129) prevents a direct realization of the convergence rate of the finite element method. In the standard Stokes system, with a sufficiently regular domain and data, it can be shown that, for a Taylor–Hood $(\mathrm{CG}_2)^d \times \mathrm{CG}_1$ discretization of the velocity-pressure pair, the L^2-norm error of the velocity converges at a rate of $\mathcal{O}(h^2)$. Such a result cannot be extrapolated to the Borrvall–Petersson problem at this time. We also remark that there is a discrepancy in the predicted rate of convergence in two and three dimensions. We numerically explore convergence rates in two dimensions in Section 4.6.2.

An algorithm must be seen to be believed.

— Donald Knuth, 1968

4 The deflated barrier method

In this chapter we introduce *the deflated barrier method*; a novel algorithm for computing multiple solutions of topology optimization problems. This is achieved by solving the first-order optimality conditions of a barrier functional with an enlarged feasible set. The deflated barrier method uses deflation to discover different branches of solutions. Then, by using continuation to decrease the barrier parameter to zero, we drive the branches of solutions to different isolated minimizers of the original topology optimization problem.

The subproblems arising during the continuation of the barrier parameter are solved with a primal-dual active set solver to enforce the box constraints on ρ, following an optimize-then-discretize (OTD) approach. This is in contrast to traditional discretize-then-optimize (DTO) primal-dual interior point methods (such as IPOPT [170]), which do not use the structure of the original infinite-dimensional optimization problem. In the context of PDE-constrained optimization, ignoring the problem structure often results in mesh-dependence of the solver. Mesh-dependence is the phenomenon where, with each refinement of the mesh, the number of iterations required by the optimization algorithm increases in an unbounded way [144].

The chapter is organized as follows: in Section 4.1 we formulate barrier functionals for the topology optimization of fluids and compliance of elastic structures. We then discuss the solver that will be used to find solutions to the first-order optimality conditions of the barrier functional in Section 4.2. The deflation mechanism is introduced in Section 4.3 and we discuss the implementation of the deflated barrier method in Section 4.4. We present several numerical examples in Section 4.6 where

we also discuss the performance of the algorithm and explore the convergence of the finite element solutions.

4.1 Formulating a barrier functional

Due to the 0-1 nature of the material distribution, a typical strategy to aid convergence is the use of a continuation scheme. In the deflated barrier method, we relax the original optimization problem by augmenting the objective functional with barrier terms. The goal now becomes to find a pair $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that minimizes

$$J_\mu^{\epsilon_{\log}}(\boldsymbol{u}, \rho) := J(\boldsymbol{u}, \rho) - \mu \int_\Omega (\log(\epsilon_{\log} + \rho) + \log(1 + \epsilon_{\log} - \rho)) \mathrm{d}x, \tag{4.1}$$

subject to the PDE constraints of (BP) (where $\boldsymbol{U} = H^1_{|\partial\Omega,\boldsymbol{g},\mathrm{div}}(\Omega)^d$) or (C) (where $\boldsymbol{U} = H^1_{\Gamma_D,\boldsymbol{0}}(\Omega)^d$). Here, $0 \leq \epsilon_{\log} \ll 1$ enlarges the feasible region enforced by the barrier terms and $\mu \geq 0$ is the barrier parameter. We note that the box constraints imposed by the barrier terms are never active as $0 \leq \rho \leq 1$ a.e. We recover the original optimization problem when $\mu = 0$. In the deflated barrier method we do *not* use the barrier terms to enforce the box constraints on ρ, as in a traditional interior point approach, but rather to perform continuation in the barrier parameter to follow a central path. This provides robust nonlinear convergence and offers an opportunity to find other solutions of the optimization problem, as explained in Section 4.4.

Proposition 4.1 (Γ-convergence)**.** *Suppose that $J(\boldsymbol{u}, \rho)$ is weak $\times$ weak-* lower semicontinuous in the weak $\times$ weak-* topology of $H^1(\Omega)^d \times L^\infty(\Omega)$. Then $J_\mu^{\epsilon_{log}}$ Γ-converges to J as $\mu \to 0$ in the weak $\times$ weak-* topology of $H^1(\Omega)^d \times L^\infty(\Omega)$.*

Proof. If we show that a liminf property holds and there exists a recovery sequence, then the proposition is proven. Firstly, if $(\boldsymbol{v}_\mu, \eta_\mu)$ is a sequence that converges to $(\boldsymbol{v}, \eta) \in H^1(\Omega) \times C_\gamma$ in the weak $\times$ weak-* topology of $H^1(\Omega)^d \times L^\infty(\Omega)$ as $\mu \to 0$, then we require that

$$J(\boldsymbol{v}, \eta) \leq \liminf_{\mu \to 0} J_\mu^{\epsilon_{\log}}(\boldsymbol{v}_\mu, \eta_\mu). \tag{4.2}$$

We note that

$$\begin{aligned} J_\mu^{\epsilon_{\log}}(\boldsymbol{v}_\mu, \eta_\mu) - J(\boldsymbol{v}, \eta) &= J(\boldsymbol{v}_\mu, \eta_\mu) - J(\boldsymbol{v}, \eta) \\ &\quad - \mu \int_\Omega (\log(\epsilon_{\log} + \eta_\mu) + \log(1 + \epsilon_{\log} - \eta_\mu)) \mathrm{d}x. \end{aligned} \tag{4.3}$$

By assumption, J is weak $\times$ weak-* lower semicontinuous and hence it remains to show that the barrier integral goes to zero as $\mu \to 0$. However, since $\eta_\mu \in C_\gamma$, the integral remains finite and as $\mu \to 0$

$$\begin{aligned} &\left| \mu \int_\Omega (\log(\epsilon_{\log} + \eta_\mu) + \log(1 + \epsilon_{\log} - \eta_\mu)) \mathrm{d}x \right| \\ &\quad = \mu \left| \int_\Omega (\log(\epsilon_{\log} + \eta_\mu) + \log(1 + \epsilon_{\log} - \eta_\mu)) \mathrm{d}x \right| \to 0. \end{aligned} \tag{4.4}$$

For any pair $(\boldsymbol{v}, \eta)$, we also need to show that there exists a recovery sequence, $(\boldsymbol{v}_\mu, \eta_\mu)$, such that $J_\mu^{\epsilon_{\log}}(\boldsymbol{v}_\mu, \eta_\mu) \to J(\boldsymbol{v}, \eta)$ as $\mu \to 0$. Choosing the trivial sequence $(\boldsymbol{v}_\mu, \eta_\mu) := (\boldsymbol{v}, \eta)$ satisfies this criterion. □

Remark 4.1. *The space $H^1(\Omega)^d \times L^\infty(\Omega)$ is weak $\times$ weak-* compact and, therefore, the sequence of minimizers of $J_\mu^{\epsilon_{log}}$ must converge. By the properties of Γ-convergence, minimizers converge to minimizers: if $(\boldsymbol{u}_\mu, \rho_\mu)$ is a minimizer for $J_\mu^{\epsilon_{log}}$, then every cluster point of the sequence $(\boldsymbol{u}_\mu, \rho_\mu)$ is a minimizer of J.*

4.1.1 The Borrvall–Petersson model

Proposition 4.2. *Suppose that $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, with $d \in \{2, 3\}$, α satisfies properties (A1)–(A3) and is continuously differentiable. Then, there exists a pair $(\boldsymbol{u}, \rho) \in \boldsymbol{U} \times C_\gamma$ that minimizes $J_\mu^{\epsilon_{log}}$ (as defined in (4.1)).*

The following proof mimics the proof of a similar result by Evgrafov [65, Sec. 4].

Proof. Borrvall and Petersson [36, Th. 3.1] proved the weak $\times$ weak-* sequential lower semicontinuity of J, and the weak $\times$ weak-* compactness of the space $H^1_{\boldsymbol{g},\mathrm{div}}(\Omega)^d \times C_\gamma$. We note that the $-\log(\cdot)$ terms are convex, which by Tonelli's theorem [76, Th. 5.14] implies that they are also weakly-* lower semicontinuous. Hence, we conclude that the relaxation of (BP) to (4.1) still admits a solution. □

As with the original optimization problem, (BP), the minimizers of (4.1) are not necessarily unique.

The result of Proposition 2.1 can be extended to include (4.1) by utilizing the fact that $-\log(x)$, $x > 0$, is a monotonically decreasing function in x. Hence, since we can tighten the inequality volume constraint to an equality volume constraint, we define the enlarged feasible-set barrier functional as:

$$L_\mu^{\epsilon_{\log}}(\boldsymbol{u}, \rho, p, p_0, \lambda) := J_\mu^{\epsilon_{\log}}(\boldsymbol{u}, \rho) - \int_\Omega p \operatorname{div}(\boldsymbol{u}) \mathrm{d}x - \int_\Omega p_0 p \, \mathrm{d}x - \int_\Omega \lambda(\gamma - \rho) \mathrm{d}x; \tag{4.5}$$

where $p \in L_0^2(\Omega)$ denotes the pressure, λ is the Lagrange multiplier for the volume constraint, and $p_0 \in \mathbb{R}$ is the Lagrange multiplier to fix the integral of the pressure (as required by the space $L_0^2(\Omega)$). By a slight modification to the proofs of Propositions 2.4 and 2.5, one can show that minimizers of (4.1) necessarily satisfy the first-order optimality conditions of (4.5), i.e., for all $(\eta, \boldsymbol{v}, q, \zeta) \in C_{[0,1]} \times H_0^1(\Omega)^d \times L_0^2(\Omega) \times \mathbb{R}$, we have:

$$a_\rho(\boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{v}, p) = l_{\boldsymbol{f}}(\boldsymbol{v}), \tag{L1}$$

$$b(\boldsymbol{u}, q) = 0, \tag{L2}$$

$$c_{\boldsymbol{u},\lambda}^{\mu,\epsilon_{\log}}(\rho, \eta - \rho) \geq 0, \tag{L3}$$

$$d_\rho(\lambda, \zeta) = 0, \tag{L4}$$

where

$$c_{\boldsymbol{u},\lambda}^{\mu,\epsilon_{\log}}(\rho, \eta) := \frac{1}{2} \int_\Omega \left(\alpha'(\rho)|\boldsymbol{u}|^2 + \lambda - \frac{\mu}{\epsilon_{\log} + \rho} + \frac{\mu}{1 + \epsilon_{\log} - \rho} \right) \eta \, \mathrm{d}x. \tag{4.6}$$

We see that, when $\mu = 0$, (L3) reduces to (FOC3b).

Remark 4.2. *If we are using a divergence-free DG finite element approximation for the velocity, we instead consider the base functional J_h (as defined in (3.55)) instead of J and construct the barrier functional $L_\mu^{\epsilon_{log}}$ as above. The first-order optimality condition (L1) becomes (FOC1-DG$_h$).*

4.1.2 Mixed boundary conditions in fluid flow

One could argue that fixing the outlet flows with a Dirichlet boundary condition in (BP) is inherently nonphysical and a more realistic model would prescribe natural boundary conditions on the outlets (while keeping the Dirichlet boundary conditions on the inlets) [56]. The correct choice of Neumann boundary conditions is nontrivial. Heywood et al. [87] provided an investigation into various formulations. In this work we opt for the natural boundary condition,

$$(-p\mathsf{I} + 2\nu\mathsf{D}(\boldsymbol{u}))\,\boldsymbol{n} = \mathbf{0} \text{ on } \Gamma_N, \tag{4.7}$$

where $\mathsf{D}(\boldsymbol{u}) := (\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^\top)/2$ denotes the symmetrized gradient, I denotes the $d \times d$ identity matrix, and $\Gamma_N \subset \partial\Omega$ denotes the outlets. Heywood et al. [87] note that such a formulation does not support Poiseuille flow. However, Limache et al. [107] proved that (4.7) does satisfy the principle of objectivity, which is often violated by other common formulations, including $(-p\mathsf{I} + \nu\nabla\boldsymbol{u})\,\boldsymbol{n} = \mathbf{0}$. The natural boundary condition (4.7) is achieved by altering the objective functional J in (BP) to

$$J_N(\boldsymbol{u}, \rho) = \frac{1}{2}\int_\Omega \alpha(\rho)|\boldsymbol{u}|^2 + 2\nu|\mathsf{D}(\boldsymbol{u})|^2 \,\mathrm{d}x. \tag{4.8}$$

By utilizing the identity $\operatorname{div}((\nabla\boldsymbol{u})^\top) = \nabla(\operatorname{div}(\boldsymbol{u}))$ and the fact that $\operatorname{div}(\boldsymbol{u}) = 0$, it can be shown that the minimizers of (4.8) satisfy the first-order optimality conditions of (BP) combined with the natural boundary conditions of (4.7). The barrier functional $L_\mu^{\epsilon_{\log}}$ is then constructed as in (4.5), using J_N as the base power dissipation functional instead of J. The other alteration in the optimization problem is the removal of the Lagrange multiplier, p_0, since the absolute pressure level is set by the outflow boundary condition.

4.1.3 Navier–Stokes and non-Newtonian flow

In the case where we wish to minimize the power dissipation of a fluid flow governed by the Navier–Stokes or a non-Newtonian momentum equation, we are required to introduce extra Lagrange multipliers to enforce the momentum and incompressibility equations. Consider the incompressible Navier–Stokes equations (non-Newtonian

flow follows similarly). Then, we introduce the Lagrange multipliers, $\boldsymbol{v} \in H_0^1(\Omega)^d$, $q \in L_0^2(\Omega)$, and $q_0 \in \mathbb{R}$, and define the barrier functional as

$$\begin{aligned} &L_\mu^{\epsilon_{\log}}(\boldsymbol{u}, \rho, p, p_0, \boldsymbol{v}, q, q_0, \lambda) \\ &\quad = J_\mu^{\epsilon_{\log}}(\boldsymbol{u}, \rho) - \int_\Omega p \operatorname{div}(\boldsymbol{u}) \mathrm{d}x - \int_\Omega p_0 p \, \mathrm{d}x - \int_\Omega \lambda(\gamma - \rho) \mathrm{d}x - \int_\Omega q_0 q \, \mathrm{d}x \\ &\qquad - \int_\Omega \nu \nabla \boldsymbol{u} : \nabla \boldsymbol{v} + \delta(\boldsymbol{u} \cdot \nabla)\boldsymbol{u} \cdot \boldsymbol{v} + \alpha(\rho)\boldsymbol{u} \cdot \boldsymbol{v} - q \operatorname{div}(\boldsymbol{v}) \, \mathrm{d}x, \end{aligned} \tag{4.9}$$

where $J_\mu^{\epsilon_{\log}}$ is as defined in (4.1) (J is chosen as defined in (BP)) and δ denotes the (constant) fluid density. By computing the first-order optimality conditions induced by (4.9), we see that $(\boldsymbol{u}, \rho)$ will satisfy a generalized Navier–Stokes momentum equation, the incompressibility constraint (FOC2), and will also minimize (BP).

4.1.4 Compliance of elastic structures

Considering the minimization problem ($\mathrm{C_{GL}}$) and using the reduction (2.21), one can show the appropriate barrier functional for the compliance of linearly elastic structures is given by

$$\begin{aligned} L_\mu^{\epsilon_{\log}}(\boldsymbol{u}, \rho, \lambda) := 2 \int_{\Gamma_N} \boldsymbol{f} \cdot \boldsymbol{u} \, \mathrm{d}s - \int_\Omega k(\rho) \left[2\mu_l \mathbf{D}(\boldsymbol{u}) : \mathbf{D}(\boldsymbol{u}) + \lambda_l \operatorname{tr}(\mathbf{D}(\boldsymbol{u})) \cdot \operatorname{tr}(\mathbf{D}(\boldsymbol{u})) \right] \mathrm{d}x \\ + \frac{\beta\epsilon}{2} \int_\Omega |\nabla \rho|^2 \mathrm{d}x + \frac{\beta}{2\epsilon} \int_\Omega \rho(1 - \rho) \mathrm{d}x - \int_\Omega \lambda(\gamma - \rho) \mathrm{d}x \\ - \mu \int_\Omega (\log(\epsilon_{\log} + \rho) + \log(1 + \epsilon_{\log} - \rho)) \mathrm{d}x, \end{aligned}$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier for the equality volume constraint.

4.2 Choosing a solver for the subproblems

Approximately solving the first-order optimality conditions of L_μ^0 as $\mu \to 0$ is the classical primal interior point approach to finding the minima of (BP) and (C). Without additional care, however, a naïve implementation results in the following poor numerical behavior:

(B1) The Hessian of $L_{\mu_k}^0(\boldsymbol{z})$ has condition number $\mathcal{O}(1/\mu_k)$. Hence, as μ decreases, the computed Newton updates may become inaccurate and require more solver time [77, Th. 4.2];

(B2) An initial guess of $\boldsymbol{z}_* = \boldsymbol{z}_k$ for the subproblem $\mu = \mu_{k+1}$ is asymptotically infeasible if an exact full Newton update of the primal interior point method is used. More precisely, if $\delta\rho_{k+1}^0$ is the calculated Newton update for ρ at the first iteration of the Newton solver at $\mu = \mu_{k+1}$, then as $\mu \to 0$, we see that $0 \leq \rho_k + \delta\rho_{k+1}^0 \leq 1$ a.e. does not hold [77, Sec. 4.3.3].

Typically, to avoid the poor numerical behavior of (B1) and (B2), the DTO primal interior point method is reformulated as a primal-dual interior point method, eliminating the rational expressions arising from the logarithmic terms in the objective functional. In a discretize-then-optimize approach, the slack variables associated with box constraints are associated to the primal variable component-wise. This manifests as a block identity matrix within the full Hessian. The Hessian can then be reduced and the primal-dual approach is reformulated into a condensed form.

It is well known that PDE-constrained optimization solvers suffer from mesh-dependence when they do not properly treat the structure of the underlying infinite-dimensional problem [144]. In order to obtain accurate solutions, where it is clear if the material distribution indicates material or void, we may require several refinements of the mesh; in this context, it is clear that mesh-dependence would be particularly disadvantageous. The mesh-independence of our algorithm will be carefully studied in the subsequent numerical examples.

In order to properly treat the structure of the underlying infinite-dimensional problem, we opt for an optimize-then-discretize (OTD) method. The full Hessian arising from an OTD primal-dual interior point method is no longer easily reduced, since the block associated with the slack variables is now a mass matrix, rather than the identity. To avoid solving uncondensed large systems involving three times the number of degrees of freedom of a primal approach, the goal is to develop an OTD barrier method that avoids the poor numerical behavior of (B1) and (B2). In a novel approach, we achieve this by solving the subproblems arising from the first-order optimality conditions of the enlarged feasible-set barrier functional $L_\mu^{\epsilon_{\log}}$, while still enforcing the true box constraints, $0 \leq \rho \leq 1$ a.e., with a primal-dual

active set solver. Whereas in a standard barrier method, the barrier terms act as a replacement for the box constraints on ρ, *here we retain the box constraints to be handled by the primal-dual active set solver.* The barrier terms are instead used for continuation of the problem, to aid nonlinear global convergence and to search for other branches of solutions.

The two inner nonlinear solvers we consider for the first-order optimality conditions of $L_\mu^{\epsilon_{\log}}$, for a fixed μ, are Hintermüller et al.'s *primal-dual active set strategy* (HIK) [88] and Benson and Munson's *active-set reduced space strategy* (BM) [31]. We briefly illustrate the basic approach taken to solve the individual subproblems using the log-barrier approach coupled with a primal-dual active set solver. Let $J : \mathbb{R}^n \to \mathbb{R}$ be a twice-continuously differentiable function and consider the following box-constrained nonlinear program:

$$\min_{\mathbf{z}\in\mathbb{R}^n} J(\mathbf{z}) \;\text{ subject to }\; \mathbf{a} \leq \mathbf{z} \leq \mathbf{b}. \tag{4.10}$$

Here, we assume that $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ such that $\mathbf{a} < \mathbf{b}$ (in each component) and we understand the inequality constraints $\mathbf{a} \leq \mathbf{z} \leq \mathbf{b}$ component-wise. Next, we formulate an 'outer approximation' of (4.10) using enlarged feasible-set log-barrier terms (for any $\mu, \epsilon_{\log} > 0$):

$$\min_{\mathbf{z}\in\mathbb{R}^n} \left\{ J(\mathbf{z}) - \mu \sum_{i=1}^{n} [\log(\mathbf{z}_i - (\mathbf{a}_i - \epsilon_{\log})) + \log((\mathbf{b}_i + \epsilon_{\log}) - \mathbf{z}_i)] : \mathbf{a} \leq \mathbf{z} \leq \mathbf{b} \right\}.$$

We emphasize that there are two pairs of box constraints: the true box constraints $[\mathbf{a}_i, \mathbf{b}_i]$, $i = 1, \ldots, n$, and the enlarged feasible-set box constraints $[\mathbf{a}_i - \epsilon_{\log}, \mathbf{b}_i + \epsilon_{\log}]$, $i = 1, \ldots, n$, $\epsilon_{\log} > 0$, that will never be active. For any fixed $\mu > 0$, the associated KKT-system has the form

$$\mathbf{f}(\mathbf{z}) - \boldsymbol{\lambda}^a + \boldsymbol{\lambda}^b = \mathbf{0}, \tag{4.11}$$

$$\boldsymbol{\lambda}^a,\ \boldsymbol{\lambda}^b \geq \mathbf{0}, \tag{4.12}$$

$$\mathbf{z} - \mathbf{a} \geq \mathbf{0},\ \mathbf{b} - \mathbf{z} \geq \mathbf{0}, \tag{4.13}$$

$$\langle \boldsymbol{\lambda}^a, \mathbf{z} - \mathbf{a} \rangle_{(\mathbb{R}^n)^*,\mathbb{R}^n} = \langle \boldsymbol{\lambda}^b, \mathbf{b} - \mathbf{z} \rangle_{(\mathbb{R}^n)^*,\mathbb{R}^n} = 0, \tag{4.14}$$

where, $\boldsymbol{\lambda}^a, \boldsymbol{\lambda}^b \in (\mathbb{R}^n)^*$ are Lagrange multipliers associated with the true box constraints and

$$\mathbf{f}(\mathbf{z}) := J'(\mathbf{z}) - \frac{\mu}{\mathbf{z} - (\mathbf{a} - \epsilon_{\log})} + \frac{\mu}{\mathbf{b} + \epsilon_{\log} - \mathbf{z}}, \tag{4.15}$$

where the rational expressions are interpreted component-wise. The equivalent mixed complementarity problem is given by

$$\text{either } \mathbf{a}_i < \mathbf{z}_i < \mathbf{b}_i \text{ and } \mathbf{f}(\mathbf{z})_i = 0, \tag{4.16}$$

$$\text{or } \mathbf{a}_i = \mathbf{z}_i \text{ and } \mathbf{f}(\mathbf{z})_i \geq 0, \tag{4.17}$$

$$\text{or } \mathbf{z}_i = \mathbf{b}_i \text{ and } \mathbf{f}(\mathbf{z})_i \leq 0. \tag{4.18}$$

Consider the natural residual function $\varphi(x, y) = x - (x - y)_+$ where $(\cdot)_+ := \max(\cdot, 0)$. This is an example of an NCP function, a class of functions that for $x, y \in \mathbb{R}$ satisfy

$$\varphi(x, y) = 0 \text{ if and only if } x, y \geq 0, \quad xy = 0. \tag{4.19}$$

Using φ, we note that (4.11)–(4.14) can be reformulated as the following:

$$\mathbf{f}(\mathbf{z}) - \boldsymbol{\lambda}^a + \boldsymbol{\lambda}^b = \mathbf{0}, \tag{4.20}$$

$$\varphi(\boldsymbol{\lambda}^a, \mathbf{z} - \mathbf{a}) = \boldsymbol{\lambda}^a - (\boldsymbol{\lambda}^a - (\mathbf{z} - \mathbf{a}))_+ = \mathbf{0}, \tag{4.21}$$

$$\varphi(\boldsymbol{\lambda}^b, \mathbf{b} - \mathbf{z}) = \boldsymbol{\lambda}^b - (\boldsymbol{\lambda}^b - (\mathbf{b} - \mathbf{z}))_+ = \mathbf{0}. \tag{4.22}$$

Assuming we are given a strictly enlarged-set feasible iterate $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{a} - \epsilon_{\log} < \mathbf{z} < \mathbf{b} + \epsilon_{\log}$, we linearize around the point $(\mathbf{z}, \boldsymbol{\lambda}^a, \boldsymbol{\lambda}^b)$ using the associated Newton-derivative and reduce the system based on the estimates of the active and inactive sets predicted by the semismooth Newton step.

In HIK, the linearized system in the direction of $(\delta\mathbf{z}, \delta\boldsymbol{\lambda}^a, \delta\boldsymbol{\lambda}^b)$ is given by

$$\mathbf{f}'(\mathbf{z})\delta\mathbf{z} - \delta\boldsymbol{\lambda}^a + \delta\boldsymbol{\lambda}^b = -\mathbf{f}(\mathbf{z}) + \boldsymbol{\lambda}^a - \boldsymbol{\lambda}^b, \tag{4.23}$$

where $\mathbf{f}'(\mathbf{z}) \in \mathbb{R}^{n \times n}$ denotes the Fréchet derivative of $\mathbf{f}$ and

$$\mathbf{z}_i + \delta\mathbf{z}_i = \mathbf{a}_i \qquad \text{if } i \in \mathfrak{A}^a = \{i : \boldsymbol{\lambda}_i^a - \mathbf{z}_i + \mathbf{a}_i > 0\}, \tag{4.24}$$

$$\mathbf{z}_i + \delta\mathbf{z}_i = \mathbf{b}_i \qquad \text{if } i \in \mathfrak{A}^b = \{i : \boldsymbol{\lambda}_i^b - \mathbf{b}_i + \mathbf{z}_i > 0\}, \tag{4.25}$$

$$\boldsymbol{\lambda}_i^a + \delta\boldsymbol{\lambda}_i^a = 0 \qquad \text{if } i \in \mathfrak{I}^a = \{i : \boldsymbol{\lambda}_i^a - \mathbf{z}_i + \mathbf{a}_i \leq 0\}, \tag{4.26}$$

$$\boldsymbol{\lambda}_i^b + \delta\boldsymbol{\lambda}_i^b = 0 \qquad \text{if } i \in \mathfrak{I}^b = \{i : \boldsymbol{\lambda}_i^b - \mathbf{b}_i + \mathbf{z}_i \leq 0\}. \tag{4.27}$$

We define the active set by $\mathfrak{A} = \mathfrak{A}^a \cup \mathfrak{A}^b$ and the inactive set by $\mathfrak{I} = \mathfrak{I}^a \cap \mathfrak{I}^b$.

Consider an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ and the subsets $S_n \subset \{1,\dots,n\}$ and $S_m \subset \{1,\dots,m\}$. Then, the matrix $\mathbf{A}_{S_n,S_m} \in \mathbb{R}^{|S_n|\times|S_m|}$ is defined by eliminating the rows of $\mathbf{A}$ in $\{1,\dots,n\}\backslash S_n$ and the columns in $\{1,\dots,m\}\backslash S_m$. Similarly for any column vector $\mathbf{x} \in \mathbb{R}^n$, the vector $\mathbf{x}_{S_n} \in \mathbb{R}^{|S_n|}$ is constructed by eliminating the rows in the set $\{1,\dots,n\}\backslash S_n$ of $\mathbf{x}$.

Now, by substituting (4.24)–(4.27) into (4.23) and removing the rows associated with the active set, we observe that

$$\mathbf{f}'(\mathbf{z})_{\mathfrak{I},\mathfrak{I}}\delta\mathbf{z}_{\mathfrak{I}} = -\mathbf{f}'(\mathbf{z})_{\mathfrak{I},\mathfrak{A}}\delta\mathbf{z}_{\mathfrak{A}} - \mathbf{f}(\mathbf{z})_{\mathfrak{I}}. \tag{4.28}$$

We can therefore solve the reduced linear system (4.28) to find the inactive set unknown components of $\delta\mathbf{z}$ (the active set components of $\delta\mathbf{z}$ are fixed by (4.24) and (4.25)).

BM attempts to solve (4.16)–(4.18) as follows. Given a feasible iterate $\mathbf{z}$ with respect to the true box constraints, $\mathbf{a} \le \mathbf{z} \le \mathbf{b}$, the active set is defined by

$$\mathcal{A} = \{i : \mathbf{z}_i = \mathbf{a}_i \text{ and } \mathbf{f}(\mathbf{z})_i > 0\} \cup \{i : \mathbf{z}_i = \mathbf{b}_i \text{ and } \mathbf{f}(\mathbf{z})_i < 0\}, \tag{4.29}$$

and the inactive set is given by $\mathcal{I} = \{i\}_{i=1}^n \backslash \mathcal{A}$. The linearized system in the direction of $\delta\mathbf{z}$ takes the form

$$\mathbf{f}'(\mathbf{z})_{\mathcal{I},\mathcal{I}}\delta\mathbf{z}_{\mathcal{I}} = -\mathbf{f}(\mathbf{z})_{\mathcal{I}} \;\text{ and }\; \delta\mathbf{z}_{\mathcal{A}} = 0. \tag{4.30}$$

The next iterate is then given by $\pi(\mathbf{z}+\delta\mathbf{z})$, where π is the component-wise projection onto the true box constraints, i.e.

$$\pi(\mathbf{z}+\delta\mathbf{z})_i = \begin{cases} \mathbf{a}_i & \text{if } \mathbf{z}_i + \delta\mathbf{z}_i < \mathbf{a}_i, \\ \mathbf{z}_i + \delta\mathbf{z}_i & \text{if } \mathbf{a}_i \le \mathbf{z}_i + \delta\mathbf{z}_i \le \mathbf{b}_i, \\ \mathbf{b}_i & \text{if } \mathbf{z}_i + \delta\mathbf{z}_i > \mathbf{b}_i. \end{cases} \tag{4.31}$$

The HIK solver is a well-established method and under suitable assumptions is equivalent to a semismooth Newton method [130, 131, 166] in both finite and infinite-dimensions [88]. This equivalence ensures local superlinear convergence and under further assumptions guarantees mesh-independence [89]. Until now,

the BM solver had no supporting theoretical results, although it is conveniently included in the PETSc solver library [24]. Experimentally, we observe that the BM solver also enjoys superlinear convergence. At first glance, the two solvers may appear quite different, but we now prove that for a linear elliptic control problem, if the active and inactive sets coincide between the two algorithms, then the updates given by HIK and BM are identical.

In the following we show that in certain problems the updates of HIK and BM are a *half-step* out of sync, where we define the notion of a half-step below. If the active and inactive sets of BM were redefined to be the same as HIK, then BM would inherit the provably-good convergence properties of HIK. To our knowledge, this is the first analytical result concerning BM. Although the result does not cover the nonlinear case, it might help build an intuitive understanding as to why BM effectively solves the semismooth formulations found in this work.

Consider the minimization problem

$$\min_{y \in L^2(\Omega)} J(y) := \frac{1}{2}(y, Ay)_{L^2(\Omega)} - (f, y)_{L^2(\Omega)} \quad \text{subject to} \quad y \geq \phi, \tag{4.32}$$

where $(\cdot,\cdot)_{L^2(\Omega)}$ denotes the inner product in $L^2(\Omega)$, f and $\phi \in L^2(\Omega)$, inequalities between $L^2(\Omega)$ functions are understood in the a.e. sense, and $A \in \mathcal{L}(L^2(\Omega))$ is self-adjoint and coercive. It can be shown there exists a unique solution y^* to (4.32) and there exists a Lagrange multiplier $\lambda^* \in L^2(\Omega)$ such that (y^*, λ^*) is the unique solution to

$$\begin{aligned} &Ay - \lambda = f, \\ &y \geq \phi,\ \lambda \geq 0,\ (\lambda, y - \phi)_{L^2(\Omega)} = 0. \end{aligned} \tag{4.33}$$

In order to avoid confusion, we denote the iterates generated by HIK by y_k and the iterates generated by BM by u_k. The active and inactive sets at iteration k, $\mathfrak{A}_k$ and $\mathfrak{I}_k$ in HIK and the active and inactive sets $\mathcal{A}_k$ and $\mathcal{I}_k$ in BM are defined by

$$\mathfrak{A}_k = \{x : \lambda_k - (y_k - \phi_i) > 0\}, \text{ and } \mathfrak{I}_k = \{x : \lambda_k - (y_k - \phi) \leq 0\},$$
$$\mathcal{A}_k = \{x : u_k = \phi \text{ and } F(u_k) > 0\}, \text{ and } \mathcal{I}_k = \{x : u_k > \phi \text{ or } F(u_k) \leq 0\},$$

where $F(u_k) \in L^2(\Omega)$ is the L^2-dual representation of the Fréchet derivative of $J(u_k)$. As in Hintermüller et al. [88, Sec. 4], we define $E_{\mathfrak{A}_k}$ the extension-by-zero operator for $L^2(\mathfrak{A}_k)$ to $L^2(\Omega)$-functions, and its transpose $E^*_{\mathfrak{A}_k}$, the restriction operator of $L^2(\Omega)$ to $L^2(\mathfrak{A}_k)$-functions. Identifying the transpose of the extension operator as the restriction operator is well documented as discussed in [165, Ch. 23]. We define $E_{\mathfrak{I}_k}$, $E^*_{\mathfrak{I}_k}$, $E_{\mathcal{A}_k}$, $E^*_{\mathcal{A}_k}$, $E_{\mathcal{I}_k}$ and $E^*_{\mathcal{I}_k}$ similarly. We note that all these restriction and extension operators are linear. We now present the infinite-dimensional description of the BM strategy:

(BM1) Choose a feasible guess $u_0 \in L^2(\Omega)$ and set $k = 0$;

(BM2) Find $\delta u_k \in L^2(\Omega)$ such that $E^*_{\mathcal{I}_k} A E_{\mathcal{I}_k} E^*_{\mathcal{I}_k} \delta u_k = -E^*_{\mathcal{I}_k}(Au_k - f)$ and $E^*_{\mathcal{A}_k} \delta u_k = 0$;

(BM3) Set $u_{k+1} = \pi(u_k + \delta u_k)$ where π is the L^2-projection onto the constraint, i.e. for any given $u \in L^2(\Omega)$, $\pi(u) \in K := \{v \in L^2(\Omega) : v \geq \phi\}$ satisfies

$$\|u - \pi(u)\|_{L^2(\Omega)} \leq \|u - v\|_{L^2(\Omega)} \text{ for all } v \in K.$$

(BM4) If convergence is reached, terminate; otherwise set $k \leftarrow k + 1$ and go to step (BM2).

Theorem 4.1 (Equivalence of HIK and BM)**.** *Let y_k denote the primal variable of HIK at iteration k and let δy_k denote the update calculated at iteration k. Let λ_k denote the dual variable at iteration k. We define half steps such that the active set is updated first, i.e. $E_{\mathfrak{A}_k} y_{k+1/2} = E_{\mathfrak{A}_k} y_{k+1}$ and $E_{\mathfrak{I}_k} y_{k+1/2} = E_{\mathfrak{I}_k} y_k$.*

Let u_k denote the primal variable of BM at iteration k and let δu_k denote the update calculated at iteration k.

Suppose that $\mathcal{A}_k = \mathfrak{A}_k$, $\mathcal{I}_k = \mathfrak{I}_k$ and $E^*_{\mathfrak{I}_k} y_k = E^*_{\mathcal{I}_k} u_k$. Then the following three equalities hold;

(E1) $y_{k+1/2} = u_k$;

(E2) $E^*_{\mathfrak{I}_k} \delta y_k = E^*_{\mathcal{I}_k} \delta u_k$;

(E3) $y_{k+3/2} = u_{k+1}$.

Proof. It is shown in [88] that the update for the inactive set of HIK satisfies

$$E^*_{\mathfrak{I}_k}(A\delta y_k) = -E^*_{\mathfrak{I}_k}(Ay_k - f).$$

Expanding the left and right-hand sides, we see that

$$E^*_{\mathfrak{I}_k}AE_{\mathfrak{I}_k}E^*_{\mathfrak{I}_k}\delta y_k + E^*_{\mathfrak{I}_k}AE_{\mathfrak{A}_k}E^*_{\mathfrak{A}_k}\delta y_k = -E^*_{\mathfrak{I}_k}AE_{\mathfrak{I}_k}E^*_{\mathfrak{I}_k}y_k - E^*_{\mathfrak{I}_k}AE_{\mathfrak{A}_k}E^*_{\mathfrak{A}_k}y_k + E^*_{\mathfrak{I}_k}f.$$

Subtracting the second term on the left-hand side, we see that

$$E^*_{\mathfrak{I}_k}AE_{\mathfrak{I}_k}E^*_{\mathfrak{I}_k}\delta y_k = -E^*_{\mathfrak{I}_k}AE_{\mathfrak{I}_k}E^*_{\mathfrak{I}_k}y_k - E^*_{\mathfrak{I}_k}AE_{\mathfrak{A}_k}E^*_{\mathfrak{A}_k}(y_k + \delta y_k) + E^*_{\mathfrak{I}_k}f. \tag{4.34}$$

By definition $E^*_{\mathfrak{A}_k}(y + \delta y_k) = E^*_{\mathfrak{A}_k}y_{k+1/2}$ and by assumption $\mathcal{A}_k = \mathfrak{A}_k$, $\mathcal{I}_k = \mathfrak{I}_k$ and $E^*_{\mathfrak{I}_k}y_k = E^*_{\mathfrak{I}_k}u_k$. Furthermore, since by assumption $\mathcal{A}_k = \mathfrak{A}_k$ and since $E^*_{\mathfrak{A}_k}\delta y_k = E^*_{\mathfrak{A}_k}(\phi - y_k)$ as derived in [88], we observe that

$$E^*_{\mathfrak{A}_k}y_{k+1/2} = E^*_{\mathfrak{A}_k}(y_k + \phi - y_k) = E^*_{\mathfrak{A}_k}u_k. \tag{4.35}$$

Since, by definition, the first half step in HIK is only an update on the active set, we see that $E^*_{\mathfrak{I}_k}y_{k+1/2} = E^*_{\mathfrak{I}_k}y_k = E^*_{\mathfrak{I}_k}u_k$. We therefore have that

$$y_{k+1/2} = u_k, \tag{4.36}$$

and (E1) holds. From (4.35), we can see that (4.34) is equivalent to

$$E^*_{\mathfrak{I}_k}AE_{\mathfrak{I}_k}E^*_{\mathfrak{I}_k}\delta y_k = -E^*_{\mathfrak{I}_k}(Au_k - f). \tag{4.37}$$

We note that (4.37) is the linear system solved to calculate the update for the inactive set of BM and hence

$$E^*_{\mathfrak{I}_k}\delta y_k = E^*_{\mathfrak{I}_k}\delta u_k. \tag{4.38}$$

Hence (E2) holds. We now show that $y_{k+3/2} = u_k$ by considering four possible cases.

(First case). Consider $C = \mathfrak{I}_k \cap \mathfrak{I}_{k+1}$. If C has measure zero, then we are done. Suppose that $|C| > 0$. Then since the dual variable is set to zero on the inactive set, we know that $E^*_C\lambda_{k+1} = 0$. Therefore, by definition of $\mathfrak{I}_{k+1}$, we know that $E^*_C y_{k+1} \geq E^*_C\phi$. Hence $E^*_C u_k + E^*_C\delta u_k \geq E^*_C\phi$ and therefore $E^*_C u_{k+1} =$

$E_C^*\pi(u_k+\delta u_k) = E_C^*u_k + E_C^*\delta u_k = E_C^*y_{k+1}$. The first half step in HIK only changes the active set, hence $E_C^*y_{k+3/2} = E_C^*u_{k+1}$.

(Second case). Consider $C = \mathfrak{I}_k \cap \mathfrak{A}_{k+1}$. If C has measure zero, then we are done. Suppose that $|C| > 0$. Then since the dual variable is set to zero on the inactive set, we know that $E_C^*\lambda_{k+1} = 0$. Therefore, by definition of $\mathfrak{A}_{k+1}$, we know that $E_C^*y_{k+1} < E_C^*\phi$. Hence $E_C^*u_k + E_C^*\delta u_k < E_C^*\phi$ and therefore $E_C^*u_{k+1} = E_C^*\pi(u_k + \delta u_k) = E_C^*\phi$. By the half-step update of the active set, $\mathfrak{A}_{k+1}$, $E_C^*y_{k+3/2} = E_C^*\phi$. Hence $E_C^*y_{k+3/2} = E_C^*u_{k+1}$.

(Third case). Consider $C = \mathfrak{A}_k \cap \mathfrak{A}_{k+1}$. If C has measure zero, then we are done. Suppose that $|C| > 0$. This implies that $E_C^*y_{k+3/2} = E_C^*\phi$. Since $\mathfrak{A}_k = \mathcal{A}_k$, we know that $E_C^*u_{k+1} = E_C^*\phi$. Hence $E_C^*y_{k+3/2} = E_C^*u_{k+1}$.

(Fourth case). Consider $C = \mathfrak{A}_k \cap \mathfrak{I}_{k+1}$. If C has measure zero, then we are done. Suppose that $|C| > 0$. By definition of $\mathfrak{A}_k$, this implies that $E_C^*y_{k+1} = E_C^*\phi$. Furthermore, by definition of $\mathfrak{I}_{k+1}$ and since the first half step of HIK only changes the active set, we see that $E_C^*y_{k+3/2} = E_C^*\phi$. By definition of $\mathcal{A}_k$, we know that $E_C^*u_{k+1} = E_C^*\phi$. Hence $E_C^*y_{k+3/2} = E_C^*u_{k+1}$.

From the four cases, we conclude that

$$y_{k+3/2} = u_{k+1}. \tag{4.39}$$

□

Both HIK and BM perform a pointwise projection on the iterates generated by the subproblems of the barrier functional. In the context of a classical OTD primal-dual interior point method applied to a PDE-constrained optimal control problem, under certain assumptions, Ulbrich and Ulbrich [167, 168] proved that local superlinear convergence holds if the iterates of the control and its associated Lagrange multipliers are pointwise projected to a controlled neighborhood of the central path. Although not all their assumptions hold in our case (in particular these problems are not convex), the combination of a primal-dual active set solver and barrier method mimics the computation of a Newton step of a primal-dual

approach and then performing a pointwise projection. An advantage of our method is that our pointwise projection is unique and cheap to compute.

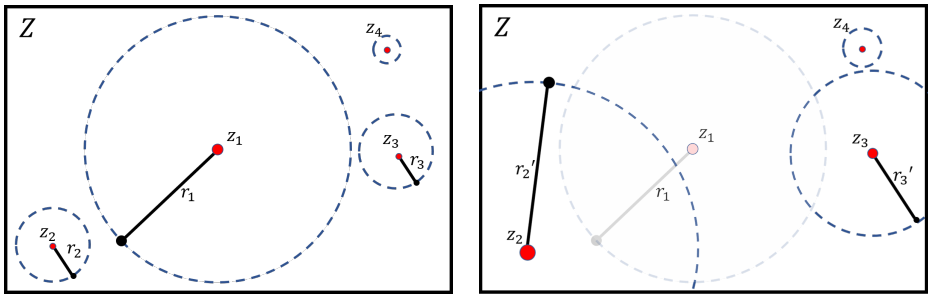
Numerically, applying HIK or BM to solve the first-order optimality conditions induced by the barrier functional, $L_\mu^{c_{\log}}$, only requires solving linear systems that are less than or equal to the size of the linear systems in a standard barrier method. Moreover, in the BM solver, the constrained variables can never reach the bounds of the enlarged feasible-set, ensuring that the Hessian remains bounded. Furthermore, both the BM and the HIK solver removes the rows and columns in the Hessian associated with the active constraints. It is these active constraints which are the source of the unbounded eigenvalues that cause the ill-conditioning of the barrier method as μ approaches zero, which addresses problem (B1). In Fig. 4.5 we give an example demonstrating that the condition number is controlled by the elimination of the active set. Removing rows and columns associated with the active set mimics the principle of Nash et al.'s *stabilized barrier method* [115, 116].

(B2) is observed in numerical examples if we use a Newton solver; however, not when using semismooth HIK or BM. In particular, BM updates can never reach the enlarged box constraints due to the pointwise projection. Hence the logarithmic terms do not influence the step sizes of the active and inactive sets.

4.3 Deflation

Deflation is an algorithm for the calculation of *multiple* solutions of systems of nonlinear equations, starting from the same initial guess. Let V and W be Banach spaces. Suppose a system of PDEs, $F(z) = 0$, $F : V \to W$ has multiple solutions $z = z_1, \ldots, z_n$, that we wish to find. We find the first solution by utilizing a Newton-like algorithm to find z_1 from some initial guess. Now we introduce a modified system $G(z) = 0$ such that:

1. $G(z) = 0$ if and only if $F(z) = 0$, for $z \neq z_1$;
2. A Newton-like solver starting from any initial guess $z_* \neq z_1$ applied to G will not converge to z_1.



(a) Before deflation. **(b)** After the deflation of z_1.

Figure 4.1: The solutions z_1, z_2, z_3 and, z_4 are zeros of the system $F(z)$. The circles around the solutions represent the basins of attraction within which a Newton-like solver converges to that particular solution.

This process is visualized in Fig. 4.1. In principle, one can use the same initial guess to converge to multiple solutions. The modified system is obtained by applying a *deflation operator*, $\mathcal{M}(z; z_1) : W \to W$, to F such that:

(D1) $\mathcal{M}(z; z_1)$ is invertible for all $z \neq z_1$ in a neighborhood of z_1;

(D2) $\liminf_{z \to z_1} \|\mathcal{M}(z; z_1)F(z)\| > 0$.

(D1) ensures that the resulting system has a solution if the original problem has an unknown solution, and (D2) ensures that a Newton-like method applied to the newly deflated system does not converge as $z \to z_1$. In this work we consider the shifted deflation operator $\mathcal{M}(z; z_1) = (\|z - z_1\|_V^{-2} + 1)\mathcal{I}$, where $\mathcal{I} : W \to W$ is the identity operator [66]. In particular, in all the numerical examples discussed in Sections 4.6 and 5.3, deflation is implemented with respect to the material distribution, i.e. $\mathcal{M}(\boldsymbol{z}; \boldsymbol{z}_1) = (\|\rho - \rho_1\|_{L^2(\Omega)}^{-2} + 1)\mathcal{I}$, where $\boldsymbol{z} = (\boldsymbol{u}, \rho, p, p_0, \lambda)$ and $\boldsymbol{z} = (\boldsymbol{u}, \rho, \lambda)$ in fluid and compliance problems, respectively.

Deflation can be implemented very efficiently. In particular, the conditioning of the Jacobian of the deflated system does not cause computational difficulty, since the Newton update of the discrete deflated system is expressed as a scaling of the Newton update of the original discrete undeflated system via the Sherman–Morrison formula [66, Sec. 3]. This is essential for the preconditioning discussed in the next chapter; one can immediately apply preconditioners for the undeflated system to the deflated one.

Let $F_h : V_h \to W_h$ be an approximation to F on the finite-dimensional spaces V_h and W_h. Let δz_h denote the solution of the deflated Newton system evaluated at $z_h \in V_h$, to be computed, and let δy_h denote the solution of the *undeflated* Newton system of F_h, assembled at the same current iterate z_h. Let $\mathbf{z}$, $\delta\mathbf{z}$, and $\delta\mathbf{y}$ be the discrete coefficient vectors of z_h, δz_h, and δy_h, respectively. Moreover, let $m(\mathbf{z}) = \mathcal{M}(z_h, z_{1,h})$ and denote the derivative of m with respect to $\mathbf{z}$ by $m'(\mathbf{z})$. The following proposition is due to Farrell et al. [66, Sec. 3].

Proposition 4.3 (Deflation)**.** *The solution $\delta\mathbf{z}$ of the discrete deflated Newton system can be computed by scaling the solution $\delta\mathbf{y}$ of the discrete undeflated Newton system as follows:*

$$\delta\mathbf{z} = \left(1 + \frac{m^{-1}(m')^\top(\delta\mathbf{y})}{1 - m^{-1}(m')^\top(\delta\mathbf{y})}\right)\delta\mathbf{y}. \tag{4.40}$$

Proof. Let $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{H} \in \mathbb{R}^{n\times n}$ denote the residual and Jacobian of the Newton-like system, respectively, that F_h induces, where n is the number of degrees of freedom of the discretization. By definition $\delta\mathbf{z}$ satisfies

$$(m\mathbf{f})'\delta\mathbf{z} = (m\mathbf{H} + \mathbf{f}(m')^\top)\delta\mathbf{z} = -m\mathbf{f}. \tag{4.41}$$

Now, by an application of the Sherman–Morrison formula, we see that

$$\begin{aligned}
\delta\mathbf{z} &= (m\mathbf{H} + \mathbf{f}(m')^\top)^{-1}(-m\mathbf{f}) \\
&= \left(m^{-1}\mathbf{H}^{-1} - \frac{m^{-1}\mathbf{H}^{-1}(\mathbf{f}(m')^\top)m^{-1}\mathbf{H}^{-1}}{1 + (m')^\top m^{-1}\mathbf{H}^{-1}\mathbf{f}}\right)(-m\mathbf{f}) \\
&= \left((-\mathbf{H}^{-1}\mathbf{f}) + \frac{m^{-1}(-\mathbf{H}^{-1}\mathbf{f})(m')^\top(-\mathbf{H}^{-1}\mathbf{f})}{1 - m^{-1}(m')^\top(-\mathbf{H}^{-1}\mathbf{f})}\right).
\end{aligned} \tag{4.42}$$

The result follows by noting that, by definition, $\delta\mathbf{y} = -\mathbf{H}^{-1}\mathbf{f}$. □

The formula (4.40) applies if multiple solutions have been deflated, i.e. if $m(\mathbf{z}) = \mathcal{M}(z_h, z_{1,h}) \cdots \mathcal{M}(z_h, z_{n,h})$ for $n > 1$. The simple structure of (4.40) arises because the deflated residual is a (nonlinear) scalar multiple of the original residual.

In summary, in order to compute the update $\delta\mathbf{z}$ for the discretized deflated system, only the original, discretized, undeflated system is solved. Its solution $\delta\mathbf{y}$ is then scaled as in (4.40).

Deflation was first introduced in the context of polynomials by Wilkinson [175]. It was then extended to differentiable finite-dimensional maps $F : \mathbb{R}^n \to \mathbb{R}^n$ by Brown and Gearhart [42]. More recently, Farrell et al. extended the original Brown and Gearhart technique to Fréchet-differentiable maps between Banach spaces [66]. Deflation has been used to discover multiple solutions of cholesteric liquid crystals, Bose–Einstein condensates, mechanical metamaterials, aircraft stiffeners, and other applications [50, 61, 112, 138, 177]. It has also been extended to semismooth mappings [68], which is necessary in the current context of topology optimization.

4.4 Implementation

The essential idea is to use deflation to attempt to find other branches during the continuation of the barrier parameter, as visualized in Fig. 4.2. As summarized in Fig. 4.3, the deflated barrier method is divided into three phases: prediction, continuation and deflation.

(Prediction). Given a solution z_{k-1} at $\mu = \mu_{k-1}$, the algorithm calculates an initial guess for the corresponding solution at $\mu = \mu_k < \mu_{k-1}$. This is done via a feasible tangent prediction method (as described in Section 4.5), a classical tangent prediction method [147, Sec. 4.4.1] or a secant prediction method [147, Sec. 4.4.2]. A feasible tangent prediction method is identical to its classical counterpart but with box constraints on the predictor step to ensure the initial guess is feasible.

(Continuation). Given an initial guess for each branch at the new barrier parameter μ_k, the algorithm calculates the new solution along each branch with a primal-dual active set solver whilst deflating away all solutions already known at $\mu = \mu_k$.

(Deflation). At some subset of the continuation steps, the algorithm searches for new branches at $\mu = \mu_k$ using solutions on different branches found at $\mu = \mu_{k-1}$ as initial guesses. The search terminates when all the initial guesses have been exhausted (reached a maximum number of iterations without converging) or when a certain number of branches $\beta_{\max}$ have been found.

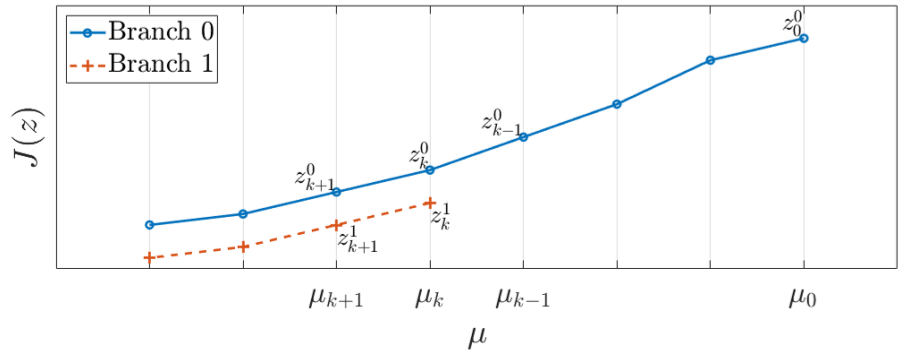


Figure 4.2: A visualization of the deflated barrier method. Branch 0 is discovered at μ_0. A predictor-corrector scheme is used to to follow the branch as μ decreases, denoted by circles. At $\mu = \mu_k$, deflation is used to discover a new solution on a different branch (branch 1), using the solution on branch 0 at $\mu = \mu_{k-1}$ as an initial guess. This newly discovered branch is then also continued as μ decreases, and is denoted by the crosses.

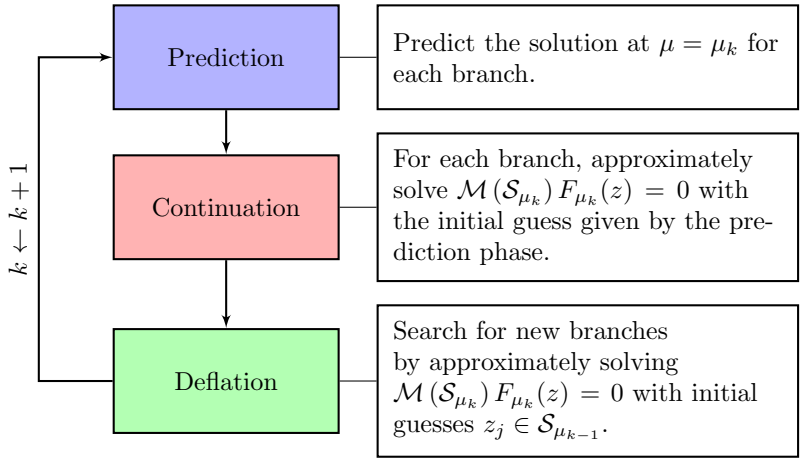


Figure 4.3: A flowchart depicting the three phases involved in the deflated barrier method.

We now explain the notation used in Algorithm 1. Let $\boldsymbol{z} = (\boldsymbol{u}, \rho, p, p_0, \lambda)$ in the Borrvall–Petersson case and $\boldsymbol{z} = (\boldsymbol{u}, \rho, \lambda)$ in the compliance case. The value of the barrier parameter at subproblem iteration k is denoted μ_k. The initial guess for the material distribution is denoted ρ_0 and the initial guess for the volume constraint Lagrange multiplier is denoted λ_0. The generator for the next value of μ is denoted by Θ. The μ-update can be adaptive or chosen a priori, provided it gives a strictly decreasing sequence. Under suitable conditions, the first-order optimality conditions of $L_\mu^{\epsilon\log}(\boldsymbol{z})$ together with the box constraints on ρ can be reformulated into perturbed KKT conditions [168, Rem. 3] which in turn can be reformulated

as a semismooth system of partial differential equations, $F_\mu(\boldsymbol{z})$. Let

$$\boldsymbol{y} = \begin{cases} (\boldsymbol{u}, p, p_0) & \text{in the Borrvall–Petersson case,} \\ \boldsymbol{u} & \text{in the compliance case.} \end{cases} \tag{4.43}$$

Let $'|_{\boldsymbol{z}_i}$ denote the Fréchet derivative with respect to $\boldsymbol{z}_i$ and $\mathcal{S}_{\mu_k}$ denote the set of solutions, $\{\boldsymbol{z}\}_i$, found at μ_k. We denote the deflation operator by $\mathcal{M}(\cdot)$ and the function space of $\boldsymbol{z}$ by Z.

Algorithm 1 Deflated barrier method

Initialize:
- $k \leftarrow 0$ ▷ Initial iteration number
- μ_0 ▷ Initial barrier parameter
- tol ▷ Approximate solve tolerance
- $\beta_{\max}$ ▷ Maximum number of branches sought
- $\rho_0(x) \leftarrow \gamma$ ▷ Constant initial material distribution
- λ_0 ▷ Initial volume constraint multiplier

Approximately solve $(L_{\mu_0}^{\epsilon_{\log}})'|_{\boldsymbol{y}}(\boldsymbol{y}, \rho_0) = 0$ for $\boldsymbol{y}$
$\boldsymbol{z}_* \leftarrow (\boldsymbol{y}, \rho_0, \lambda_0)$ ▷ Initial guess
Approximately solve $F_{\mu_0}(\boldsymbol{z}) = 0$ with initial guess $\boldsymbol{z}_*$.
$S_{\mu_0} \leftarrow S_{\mu_0} \cup \{\boldsymbol{z}\}$ ▷ Include solution in solution set
$\mu_1 \leftarrow \Theta(\mu_0)$, $k \leftarrow 1$ ▷ Update μ and k
while $\mu_k \geq 0$ and $|\mathcal{S}_{\mu_{k-1}}| \neq \varnothing$ **do**
  for $\boldsymbol{z}_i \in \mathcal{S}_{\mu_{k-1}}$ **do**
    ▷ **Prediction**
    Predict solution at μ_k, denoted $\boldsymbol{z}_*$.
    ▷ **Continuation**
    Attempt to solve $\mathcal{M}\left(\mathcal{S}_{\mu_k}\right) F_{\mu_k}(\boldsymbol{z}) = 0$ with initial guess $\boldsymbol{z}_*$.
    if $\|F_{\mu_k}(\boldsymbol{z})\|_{Z^*} \leq \text{tol}$ **then**
      Solve has succeeded; set $S_{\mu_k} \leftarrow S_{\mu_k} \cup \{\boldsymbol{z}\}$.
    end if
  end for
  ▷ **Deflation**
  for $\boldsymbol{z}_j \in \mathcal{S}_{\mu_{k-1}}$ **do**
    if $|\mathcal{S}_{\mu_k}| \geq \beta_{\max}$ **then**
      break
    end if
    Attempt to solve $\mathcal{M}\left(\mathcal{S}_{\mu_k}\right) F_{\mu_k}(\boldsymbol{z}) = 0$ with initial guess $\boldsymbol{z}_j$.
    if $\|F_{\mu_k}(\boldsymbol{z})\|_{Z^*} \leq \text{tol}$ **then**
      Solve has succeeded; set $S_{\mu_k} \leftarrow S_{\mu_k} \cup \{\boldsymbol{z}\}$.
    end if
  end for
  $\mu_{k+1} \leftarrow \Theta(\mu_k)$ ▷ Choose new value of μ
  $k \leftarrow k + 1$
end while

4.5 Feasible tangent prediction

In this section we introduce a novel predictor method that can be used in the prediction phase of the deflated barrier method. Predictor-corrector methods are often used in tracing bifurcation diagrams [147]. The idea is that as the parameter of the problem changes, a cheap predictor generates an initial guess for the solution of the system with the new parameter. A corrector method is then used to converge from this initial guess to the true solution. In our context, the primal-dual active-set solver is the corrector method. Our feasible tangent predictor method draws inspiration from the usual tangent predictor method, which solves a linear equation to find an initial guess, but applies box constraints to ensure the predicted guess is feasible.

The usual tangent predictor is derived as follows. Consider a Fréchet-differentiable equation $F(z^0, \mu^0) = 0$, where $\mu = \mu^0$ is the parameter we wish to vary. Consider a new parameter $\mu = \mu^1$ and let $\delta\mu := \mu^1 - \mu^0$. Furthermore, let $w := (z, \mu)$. The goal is to find δz such that $z^0 + \delta z \approx z^1$ where z^1 is the solution to

$$F(z^1, \mu^1) = 0. \tag{4.44}$$

A first-order approximation of (4.44) is

$$0 = F(z^1, \mu^1) \approx F(z^0, \mu^0) + F'(w)\delta w = F'_z(z^0, \mu^0)\delta z + F'_\mu(z^0, \mu^0)\delta\mu. \tag{4.45}$$

Hence an initial guess, $z_* = z^0 + \delta z$, can be calculated by solving

$$F'_z(z^0, \mu^0)\delta z = -F'_\mu(z^0, \mu^0)\delta\mu, \tag{4.46}$$

for δz. In the context of the deflated barrier method this is equivalent to solving

$$(L^{\epsilon_{\log}}_{\mu^0})''|_{z,z}(\boldsymbol{z}^0)\delta\boldsymbol{z} + (L^{\epsilon_{\log}}_{\mu^0})''|_{z,\mu}(\boldsymbol{z}^0)\delta\mu = 0, \tag{4.47}$$

for $\delta\boldsymbol{z}$. The traditional tangent predictor has no guarantee that $0 \leq \rho^0 + \delta\rho \leq 1$ a.e. To ensure that the initial guess is feasible, we instead transform (4.47) into a complementarity problem. Consider the linear operator, $T(\boldsymbol{w})$ defined by

$$\langle T(\boldsymbol{w}^0), \delta\boldsymbol{w}\rangle = (L^{\epsilon_{\log}}_{\mu^0})''|_{z,z}(\boldsymbol{z}^0)\delta\boldsymbol{z} + (L^{\epsilon_{\log}}_{\mu^0})''|_{z,\mu}(\boldsymbol{z}^0)\delta\mu.$$

Given sufficient regularity of the dual variable $T(\boldsymbol{w})$ and the primal variable $\delta\boldsymbol{w}$, we can consider the following complementarity problem,

$$\delta\rho(x) = -\rho^0(x) \qquad \text{and } T(\boldsymbol{w}^0)(x) \geq 0, \tag{4.48}$$

$$\text{or} \quad -\rho^0(x) < \delta\rho(x) < 1 - \rho^0(x) \quad \text{and } T(\boldsymbol{w}^0)(x) = 0, \tag{4.49}$$

$$\text{or} \quad \delta\rho(x) = 1 - \rho^0(x) \qquad \text{and } T(\boldsymbol{w}^0)(x) \leq 0. \tag{4.50}$$

Solving (4.48)–(4.50) constructs a feasible tangent predictor, $\boldsymbol{z}_*$. We note that this method does not perform a pointwise projection. For example, in the topology optimization of compliance, where we require the material distribution to live in $H^1(\Omega)$, we are instead performing an H^1-projection on the prediction update. In the case where (4.49) holds a.e. in Ω, finding the feasible tangent predictor reduces to solving (4.47).

4.6 Numerical results

All examples of this section were implemented with the finite element software FEniCS [109] and the resulting linear systems were solved by a sparse LU factorization with MUMPS [17] and PETSc [24]. The meshes were either created in FEniCS or Gmsh [81]. We present four different examples of the minimization of the power dissipation of a fluid constrained by the Stokes equations, one constrained by the Navier–Stokes equations, and two examples of the minimization of the compliance constrained by linear elasticity. Throughout the numerical examples, $h_{\min} := \min_{K\in\mathcal{T}_h} h_K$ denotes the minimum diameter of all simplices in the mesh, where the simplex diameter is computed as the maximum edge length. Similarly $h_{\max} := h = \max_{K\in\mathcal{T}_h} h_K$ denotes the maximum diameter of all simplices in the mesh. All solutions depicted are presented as computed by the deflated barrier method, with no truncation or postprocessing of the material distribution.

4.6.1 Borrvall–Petersson double-pipe

The double-pipe problem is the first fluid topology optimization problem found in literature that supports two minimizers [36, Sec. 4.5]. The design domain is a

rectangle $\Omega = (0, 3/2) \times (0, 1)$ with two prescribed flow inputs and two prescribed outputs, and the Dirichlet boundary conditions on $\boldsymbol{u}$ are given by the boundary data

$$\boldsymbol{g}(x,y) = \begin{cases} (1 - 144(y - 3/4)^2, 0)^\top & \text{if } 2/3 \leq y \leq 5/6, x = 0 \text{ or } 3/2, \\ (1 - 144(y - 1/4)^2, 0)^\top & \text{if } 1/6 \leq y \leq 1/3, x = 0 \text{ or } 3/2, \\ (0,0)^\top & \text{elsewhere on } \partial\Omega. \end{cases} \tag{4.51}$$

The volume fraction is $\gamma = 1/3$, the viscosity $\nu = 1$, the forcing term $\boldsymbol{f}(x, y) = (0, 0)^\top$ and we use α as given in (2.22), with $\overline{\alpha} = 2.5 \times 10^4$ and $q = 1/10$. Here q is a penalty parameter which controls the level of intermediate values (between zero or one) in the optimal design, with larger q giving sharper interfaces. The setup of the problem is depicted in Fig. 1.2.

We test our algorithm with two finite element discretizations: a Taylor–Hood $(\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ discretization and a divergence-free Scott–Vogelius $(\mathrm{CG}_2)^2 \times \mathrm{DG}_1$ discretization for the velocity and pressure. In both discretizations we use continuous piecewise linear CG_1 elements for the material distribution. Stability of the Scott–Vogelius discretization is ensured by using a barycentrically-refined mesh [132]. Both of these choices of discretization applied to this problem satisfy the necessary conditions of Theorem 3.1 for the existence of strongly converging sequences to the analytical isolated minimizers.

Deflated barrier method results

The deflated barrier method is applied to the perturbed first-order optimality conditions (L1)–(L4) induced by the barrier functional described in Section 4.1.1. For the BM solver, we begin with $\mu_0 = 100$ and apply deflation immediately to find the second branch of solutions. For HIK, this strategy did not converge to the second branch, although the second branch is discovered with $\mu_0 = 105$. In both cases tangent prediction is used, as well as a damped l^2-minimizing linesearch [43, Alg. 2]. Fig. 4.4 shows the minimizers of the double-pipe problem computed using the deflated barrier method.

In Table 4.1 we explore the mesh-independence of the primal-dual active set solvers. We observe that with each refinement of the mesh, the number of iterations

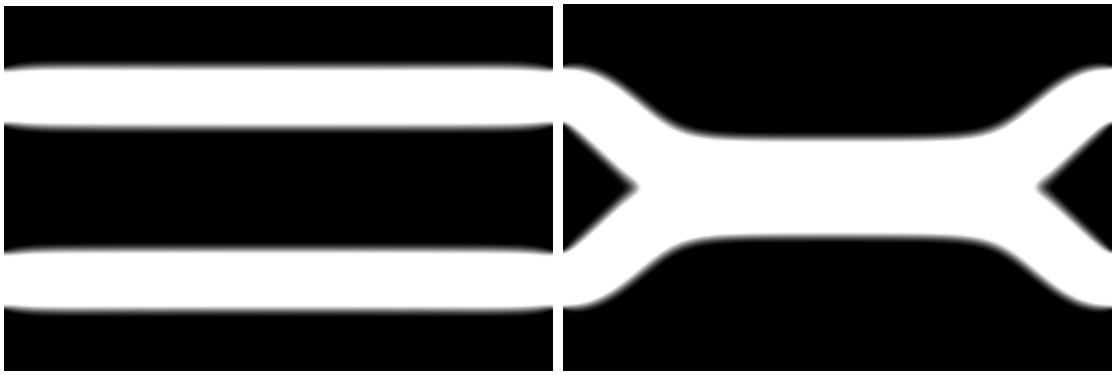


Figure 4.4: The material distribution of the local (left) and global (right) minimizers of the double-pipe optimization problem with mesh size $h = 0.0141$. Black corresponds to a value of $\rho = 0$ and white corresponds to a value of $\rho = 1$. The objective functional values are $J = 32.58$ (left) and $J = 23.87$ (right).

stays roughly constant. In particular, we notice that the behavior is consistent for both HIK and BM in both discretizations. This is a recurring theme and holds in subsequent examples.

In Fig. 4.5 we plot the condition number of the Hessian as in a classical barrier method, and the condition number of the Hessian with the rows and columns associated with the active-set removed. We observe that the condition number of the latter is significantly smaller, accounting for why our proposed methodology does not suffer from ill-conditioning.

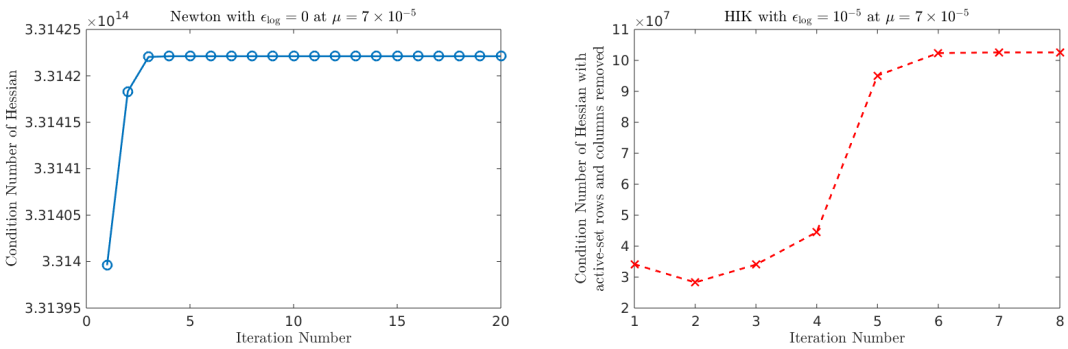


Figure 4.5: The condition number of the Hessian at each iteration of the solver in the subproblem with $\mu = 7 \times 10^{-5}$. The condition number of the Hessian of L_μ^0 arising in the linear systems of a standard Newton solver (left) is six to seven orders of magnitude larger than the condition number of the Hessian of $L_\mu^{\epsilon_{\log}}$ arising in the linear systems of the HIK solver (right).

BM Solver	Taylor–Hood		Branch 0			Branch 1	
h	Dofs	Cont.	Defl.	Pred.	Cont.	Defl.	Pred.
0.0283	38,256	124	0	22	115	30	22
0.0177	97,206	123	0	22	109	30	22
0.0141	151,506	110	0	22	116	29	22
HIK solver	Taylor–Hood		Branch 0			Branch 1	
h	Dofs	Cont.	Defl.	Pred.	Cont.	Defl.	Pred.
0.0283	38,256	174	0	43	261	14	43
0.0177	97,206	189	0	43	223	13	43
0.0141	151,506	173	0	43	197	13	43
BM solver	Scott–Vogelius		Branch 0			Branch 1	
$h_{\min}/h_{\max}$	Dofs	Cont.	Defl.	Pred.	Cont.	Defl.	Pred.
0.0278/0.0501	58,685	155	0	22	139	29	22
0.0139/0.0250	234,005	124	0	22	120	29	22

Table 4.1: The cumulative total numbers of primal-dual active-set solver iterations required in the continuation, deflation and prediction phases of the double-pipe problem. Branch 0 discovers the local minimum shown in Fig. 4.4 and branch 1 discovers the global minimum. As we can see, the number of iterations stays roughly constant for both solvers as we refine the mesh.

Convergence results

Given that the deflated barrier method efficiently computes solutions to (FOC1$_h$)–(FOC3a$_h$), we numerically verify the result of Theorem 3.1, i.e. there exists a sequence of finite element solutions that strongly converges to the straight channel solution and a different sequence of solutions that strongly converges to the double-ended wrench solution.

Analytical solutions for the straight channel and double-ended wrench are not known for choices of the inverse permeability, α, used in practice. Hence, errors are measured with respect to a heavily-refined finite element solution, which is constructed as follows; first the finite element solutions are computed on a mesh with mesh size $h = 0.028$, using the deflated barrier method. Next, the mesh is adaptively refined three times in areas where the material distribution is between 1/10 and 9/10. Each time the mesh is refined, the coarse-mesh solution is interpolated onto the finer mesh as an initial guess and the first-order optimality conditions are

re-solved using the deflated barrier method. The error plots are given in Fig. 4.6.

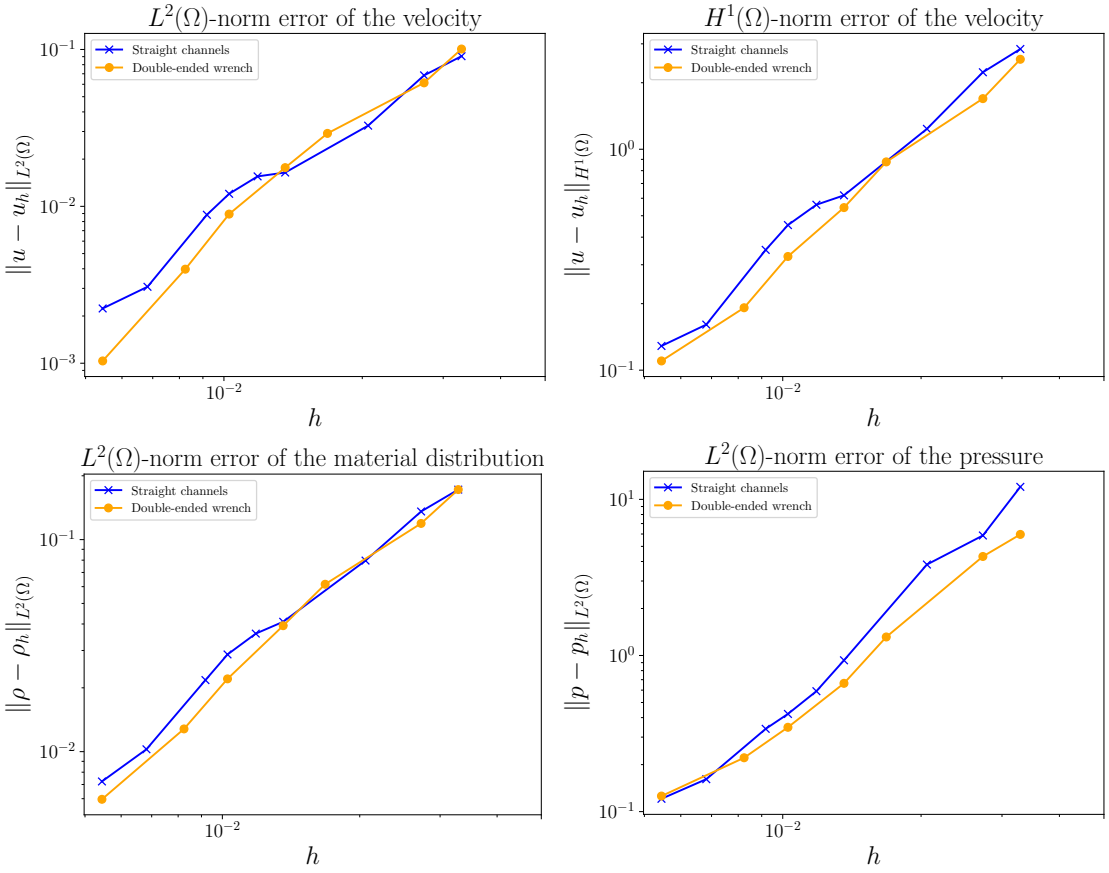


Figure 4.6: The convergence of $\boldsymbol{u}_h$, ρ_h, and p_h for the double-pipe problem for both the straight channel and double-ended wrench solutions on an unstructured mesh with a $\mathrm{CG}_1 \times (\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ discretization for $(\rho_h, \boldsymbol{u}_h, p_h)$.

In principle, there can be infinitely many different subsequences of finite element solutions that strongly converge to the same analytical minimizer at different convergence rates. Separate subsequences cause difficulties in the interpretation of the convergence plots as they present themselves as oscillations in the error. This is observed in practice and appears to be caused by at least the following two observations:

- Multiple finite element solutions can exist on the same mesh that represent the same analytical solution, e.g. Fig. 4.7;
- A fine mesh can align worse than a coarser mesh with the jumps in the analytical material distribution.

The first observation is not surprising in the context of nonlinear PDEs and nonconvex variational problems. In such cases, an additional selection mechanism is required in order to favour one particular solution over others coexisting on the same mesh. Selection mechanisms are problem-dependent. In the case of nonlinear hyperbolic conservation laws, the entropy condition plays this role. In the present context, one might propose choosing the solution, minimizing the modified optimization problem (BP_h), that attains the smallest objective functional value for J, within the basin of attraction of the isolated local minimizer. For sufficiently small h, a minimizer satisfying this selection mechanism must exist. However, it is not necessarily unique and numerically enforcing such a condition can be difficult. In order to promote convergence to the minimizer of (BP_h) with the smallest value J, we interpolate the heavily-refined finite element solutions onto coarser meshes as initial guesses for the deflated barrier method. This strategy was effective in practice. The effects of the second observation are harder to test. However, in Fig. 4.6, we attempt to minimize mesh bias by measuring errors on unstructured meshes.

It may be possible to find a sequence of mesh sizes, (h_i), such that there exist two different sequences of finite element solutions that strongly converge to the same isolated minimizer. In Fig. 4.7, we depict two different straight channel finite element solutions that exist on the same unstructured mesh where $h = 0.04$. Both solutions satisfy the discretized first-order optimality conditions ($\mathrm{FOC1}_h$)–($\mathrm{FOC3a}_h$) and both locally minimize $J(\boldsymbol{v}_h, \eta_h)$. Choosing one over the other would change the convergence pattern of the strongly converging sequence. This may cause difficulty in practice, as optimization strategies are unlikely to discover the discretized global minimum without additional selection mechanisms.

4.6.2 Discontinuous-forcing

The following example is constructed to satisfy the regularity results of Proposition 2.6 and Theorem 2.12 and we solely explore the convergence rates of the finite element solutions. Consider the optimization problem (BP), with a homogeneous

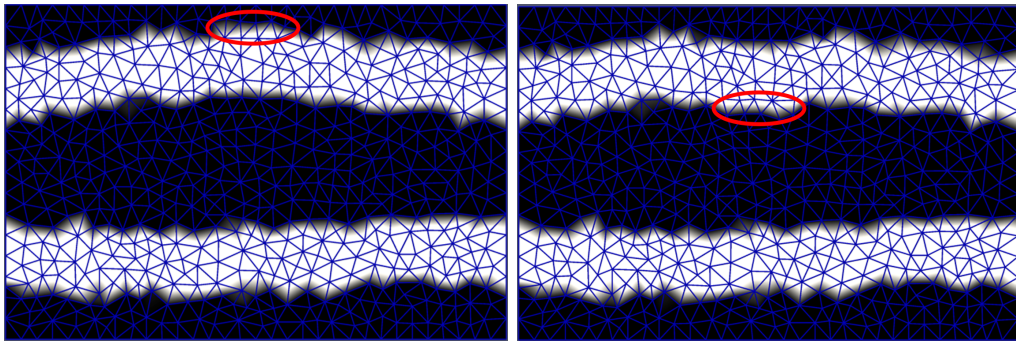


Figure 4.7: Two different straight channel finite element solutions of the double-pipe optimization problem that exist on the same unstructured mesh where $h = 0.04$. The differences can be spotted at the midway point of the top channel.

Dirichlet boundary condition on $\boldsymbol{u}$, $\Omega = (0,1)^2$, volume fraction $\gamma = 1/3$, viscosity $\nu = 1$ and a forcing term given by

$$\boldsymbol{f}(x,y) = \begin{cases} (10,0)^\top & \text{if } 3/10 < x < 7/10 \text{ and } 3/10 < y < 7/10, \\ (0,0)^\top & \text{otherwise.} \end{cases} \tag{4.52}$$

The inverse permeability, α, is as given in (2.22), with $\overline{\alpha} = 2.5 \times 10^4$ and $q = 1/10$, which satisfies (A1)–(A5). Fig. 4.8 depicts the material distribution of three minimizers: one local minimizer is in the shape of a figure eight and the two $\mathbb{Z}_2$ symmetric global minimizers are in the shape of annuli. Since the domain is convex,

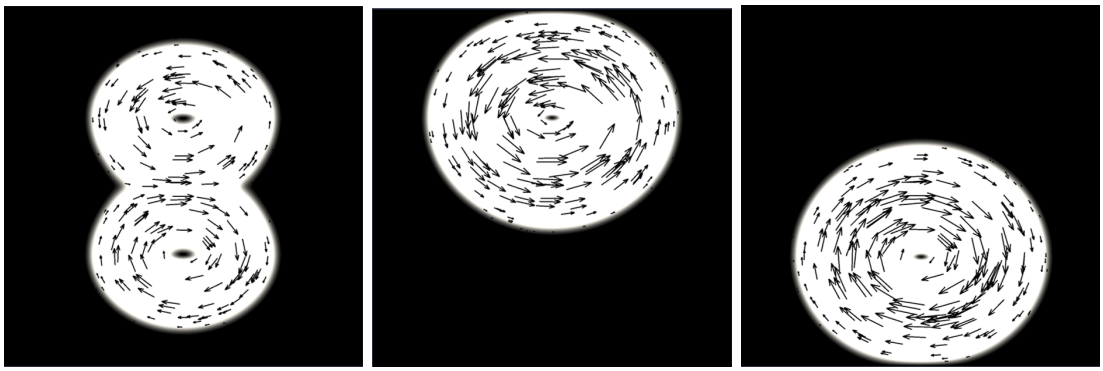


Figure 4.8: The material distribution of a local (left) and the global (middle and right) minimizers of the discontinuous-forcing optimization problem. Black corresponds to a value of $\rho = 0$ and white corresponds to a value of $\rho = 1$, with the grey regions indicating intermediate values. The arrows indicate the velocity profile of the solutions.

$\boldsymbol{g}(x,y) = \boldsymbol{0}$, and $\boldsymbol{f} \in L^2(\Omega)^d$, then, by the regularity results of Proposition 2.6, $\boldsymbol{u} \in H^2(\Omega)^2$ and $p \in H^1(\Omega)$. Moreover, the conditions of Theorem 2.12 hold and, therefore, $\rho \in H^1(U_\theta)$ for every $\theta > 0$. In this particular example, the support

of ρ is compactly contained in the support of the velocity in all three solutions. Therefore, we conclude that $\rho \in H^1(\Omega)$.

Consider a Taylor–Hood $(\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ finite element discretization for the velocity-pressure pair, and a piecewise constant DG_0 finite element discretization for the material distribution. Since all three solutions are isolated local minimizers, by Theorem 3.1, there exists a sequence of finite element solutions to the discretized first-order optimality conditions that strongly converges to the figure eight solution, and different sequences of different finite element solutions that strongly converge to the two annulus solutions. Their existence is confirmed in Fig. 4.9.

Since $\rho \in H^1(\Omega)$ and we are using a DG_0 finite element discretization, a naïve prediction for the convergence rate of the L^2-norm error of the material distribution is $\mathcal{O}(h)$. This rate is observed in the bottom left panel of Fig. 4.9. Moreover, since the minimum regularity of the velocity is $\boldsymbol{u} \in H^2(\Omega)^d$, and we are using a $(\mathrm{CG}_2)^2$ finite element discretization, a prediction for the minimum convergence rates of the velocity are $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ for the H^1-norm and L^2-norm errors of the velocity, respectively.

In the standard Stokes system, the regularity of $\boldsymbol{u}$ is related to the regularity of the forcing term $\boldsymbol{f} \in H^s(\Omega)^d$, such that $\boldsymbol{u} \in H^{s+2}(\Omega)^d$ (assuming the domain and boundary data are also suitably regular). Here, the regularity of the forcing term satisfies $s < 1/2$. If we assume that the velocity has the additional regularity $\boldsymbol{u} \in H^{s+2}(\Omega)^d$, $s \in (0, 1/2)$, in this context, a prediction for the upper limit of the convergence rate is $\mathcal{O}(h^r)$ and $\mathcal{O}(h^{t+1})$, for some $r, t \in [1, s+1]$, for the H^1-norm and L^2-norm errors of the velocity, respectively. The rates observed in the top panels of Fig. 4.9 match this prediction. The H^1-norm error is decreasing at a rate slightly faster than $\mathcal{O}(h^{3/2})$ for all three solutions and the L^2-norm error convergence rate is $\mathcal{O}(h^2)$ for the figure eight solution and $\mathcal{O}(h^{5/2})$ for the annuli solutions. We hypothesize that the upper limit of the convergence rate of the L^2-norm error of the velocity is bounded by the relatively slower rate of the convergence of the material distribution.

Finally, since the minimum regularity of the pressure is $p \in H^1(\Omega)$ and the discretization is CG_1, a prediction for the convergence rate of the L^2-norm error is $\mathcal{O}(h^r)$, for some $r \in [1, s+1]$. Initially, the convergence rate is $\mathcal{O}(h^{3/2})$ which matches our naïve prediction. However, on finer meshes, the convergence rate increases. We hypothesize that this speedup is artificial and is caused by the lack of resolution of the refined finite element solutions that are being used as proxies for the analytical solutions in the error norm estimate. Qualitatively, it can be checked that mesh refinement in areas where the discretized material distribution lies between 1/10 and 9/10 is an ineffective strategy for improving the approximation of the analytical pressure over the whole domain.

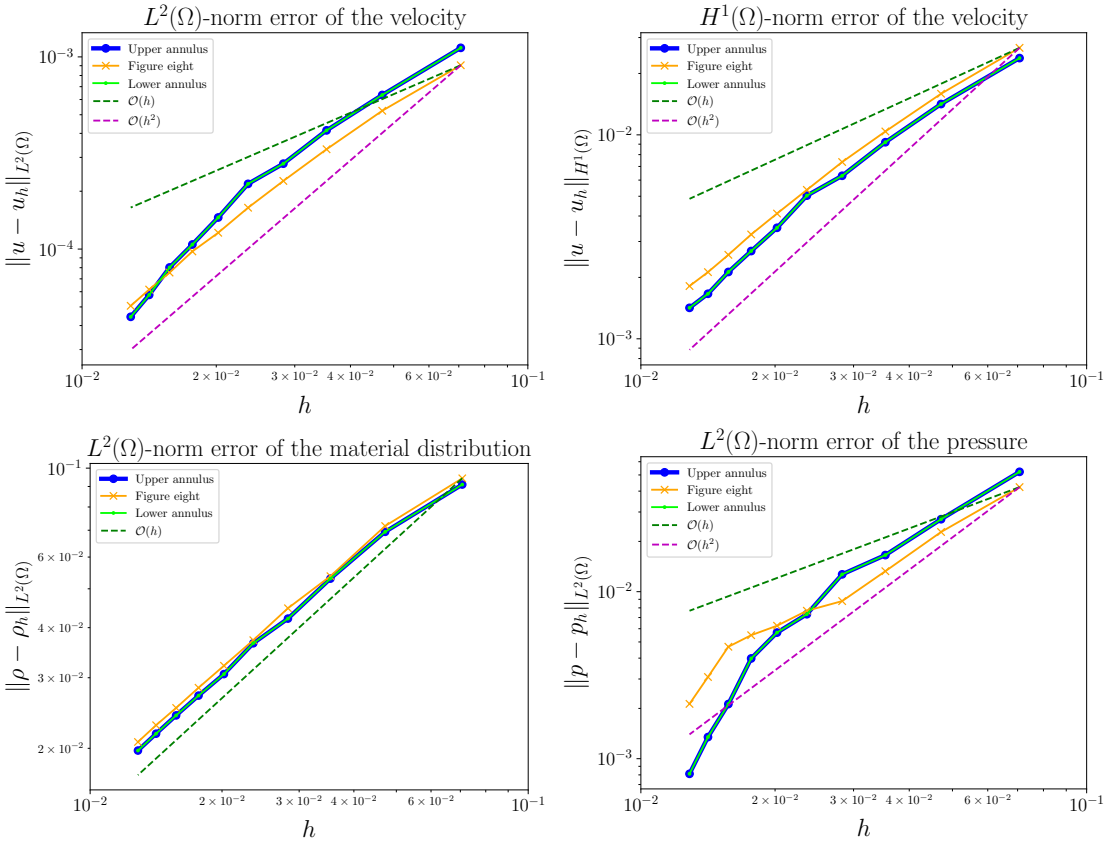


Figure 4.9: The convergence of $\boldsymbol{u}_h$, ρ_h, and p_h in the discontinuous-forcing problem for the figure eight and annulus solutions on structured meshes, with a $\mathrm{DG}_0 \times (\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ discretization for $(\rho_h, \boldsymbol{u}_h, p_h)$.

4.6.3 Neumann-outlet double-pipe

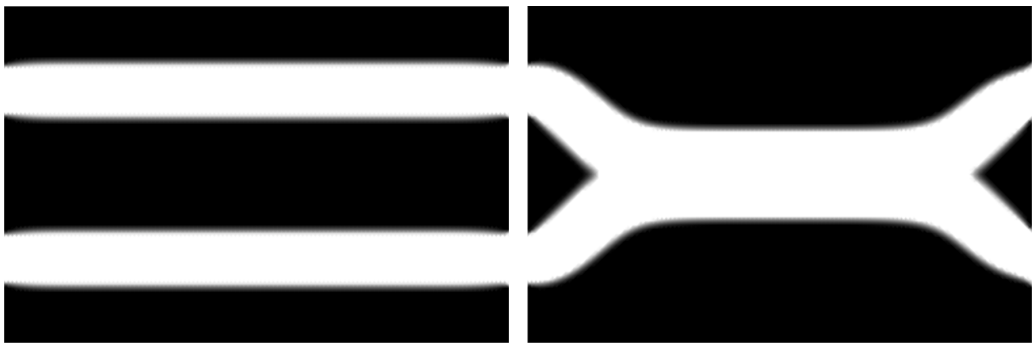
In this example we consider the double-pipe problem with Neumann boundary conditions on the outlets, whilst keeping all other model parameters the same. We employ the Taylor–Hood $(\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ discretization for the velocity-pressure pair and a CG_1 discretization for the material distribution. The barrier functional is described in Section 4.1.2. The barrier parameter is initialized at $\mu_0 = 1000$ and the BM solver is used to solve the perturbed first-order optimality conditions. Deflation finds the second, third, and fourth branches at $\mu = 82.4$. For $h = 0.0333$, deflation discovers branch 2, then branch 1 and 3, whereas for the other mesh sizes, deflation discovers the branches in ascending order.

The removal of an imposed outlet flow has an interesting effect. The global minimizer in the shape of a double-ended wrench is now a local minimizer. Two new $\mathbb{Z}_2$-symmetric global minimizers now exist as shown in Fig. 4.10. This is not entirely surprising. There is a cost associated with the pipe splitting and if the optimization problem does not require the flow to leave both outlets, then it is favorable for the flow to exit via one outlet, not both. This is reflected in the resulting cost.

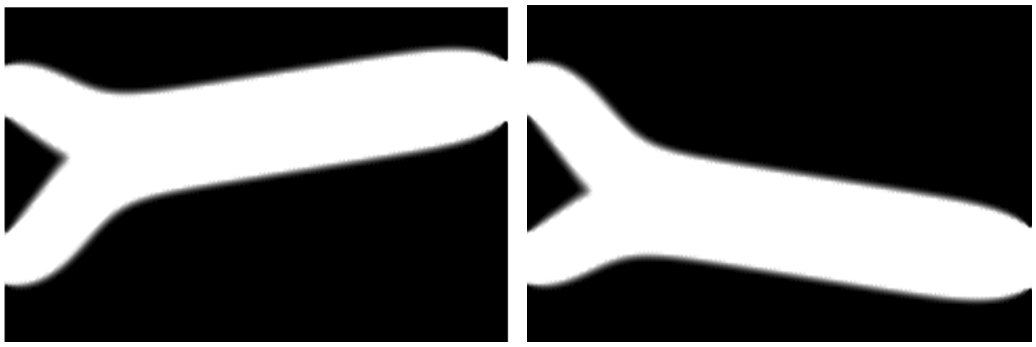
The mesh-independence of the algorithm is investigated in Table 4.2. As before, mesh-independence is observed.

BM Solver		Branch 0			Branch 1		
h	Dofs	Cont.	Defl.	Pred.	Cont.	Defl.	Pred.
0.0333	27,455	118	0	53	108	49	34
0.0250	48,605	136	0	37	107	34	37
0.0125	193,205	113	0	35	106	45	36
		Branch 2			Branch 3		
h	Dofs	Cont.	Defl.	Pred.	Cont.	Defl.	Pred.
0.0333	27,455	166	199	55	166	149	55
0.0250	48,605	145	123	45	145	157	45
0.0125	193,205	128	151	46	128	146	46

Table 4.2: The cumulative total numbers of BM solver iterations required in the continuation, deflation and prediction phases of the double-pipe problem with natural boundary conditions on the outlets.



(a) Branch 0. **(b)** Branch 1.



(c) Branch 2. **(d)** Branch 3.

Figure 4.10: The material distribution of two local and two global minimizers of the double-pipe optimization problem with natural boundary conditions on the outlets, instead of Dirichlet conditions, with $h = 0.0125$. Black corresponds to a value of $\rho = 0$ and white corresponds to a value of $\rho = 1$. Branches 0, 1, 2, and 3 have the objective functional values $J_N = 32.35$, 22.92, 18.46, and 18.46, respectively.

4.6.4 Roller-type pump

In this example problem [56, Sec. 2.1.4.4], the domain is given by

$$\Omega = (0,1)^2 \setminus \left\{(x,y) \in (0,1)^2 : (x-0.5)^2 + (y-0.5)^2 \leq (0.3)^2\right\}.$$

The boundary data $\boldsymbol{g}$ on $\boldsymbol{u}$ is given by:

$$\boldsymbol{g}(x,y) = \begin{cases} (0, 1 - 20(x-0.61)^2)^\top, & \text{if } 0.56 < x < 0.66 \text{ and } y = 0, \\ (1 - 20(y-0.95)^2, 0)^\top, & \text{if } x = 1 \text{ and } 0.9 < y < 1, \\ 10/3(y - 1/2, 1/2 - x)^\top, & \text{if } (x-0.5)^2 + (y-0.5)^2 = (0.3)^2, \\ (0,0)^\top, & \text{elsewhere.} \end{cases}$$

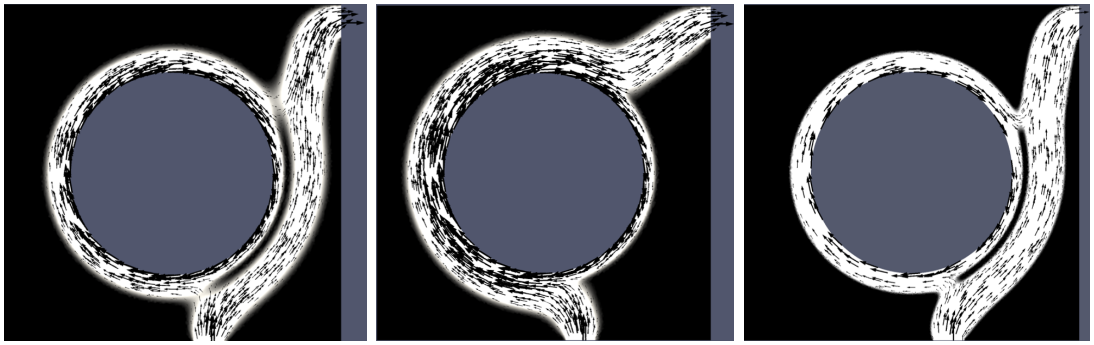
These boundary conditions model an inlet on the bottom of the domain and an outlet on the right of the domain with a pump rotating at a constant velocity in the center of the domain where the fluid experiences no-slip boundary conditions.

We employ the Taylor–Hood discretization and initialize $\mu_0 = 1000$. The barrier functional is as given in Section 4.1.1. Deflation finds the second branch at $\mu = 6.78$.

A global and local minimum of the problem are shown in Fig. 4.11a. The local minimum chooses to avoid the pump in favor of taking the path with the shortest distance from the inlet to the outlet, while the global minimum exploits the rotation given by the pump. The local minimizer for $q = 1/10$ has areas where $\rho \approx 1/2$, which has an ambiguous physical interpretation. In order to verify whether ρ should be equal to zero or one in such areas, a mixture of grid-sequencing and continuation in q was performed, resulting in the solution shown in Fig. 4.11b. The mesh-independence of the algorithm is verified in Table 4.3.

BM solver		Branch 0			Branch 1		
$h_{\min}/h_{\max}$	Dofs	Cont.	Defl.	Pred.	Cont.	Defl.	Pred.
0.0258/0.0509	7388	260	0	55	118	80	35
0.0127/0.0255	29,174	186	0	51	75	117	25
0.0064/0.0127	113,096	177	0	46	83	99	29

Table 4.3: The cumulative total numbers of BM solver iterations required in the continuation, deflation and prediction phases of the roller-type pump problem to find the solutions shown in Fig. 4.11a. The iteration counts are mesh-independent.



(a) The local (left) and global (right) minimizers, ρ. **(b)** Refined local minimizer.

Figure 4.11: (a) The material distribution of the local and global minimizers of the roller-type pump optimization problem, with $h_{\min} = 6.4 \times 10^{-3}$. Black corresponds to a value of $\rho = 0$ and white corresponds to a value of $\rho = 1$. The grey area is the hole removed from the domain. The arrows indicate the direction and magnitude of the velocity, $\boldsymbol{u}$. The values of the objective functional are $J = 26.84$ (left) and $J = 22.67$ (right). (b) A mixture of grid-sequencing of the mesh where $\rho \approx 1/2$ and the continuation of q to larger values was performed on the local minimum of the roller-type pump optimization problem in order to remove areas where $\rho \approx 1/2$. The resulting refined solution has clearly defined areas of $\rho = 0$ and $\rho = 1$. Here $h_{\min} = 3.3 \times 10^{-3}$, $q = 0.65$, and $J = 29.17$.

4.6.5 Five-holes double-pipe with Navier–Stokes

In this example we consider a Navier–Stokes flow through a nonconvex domain with a minimizing power dissipation. The setup is similar to the original Borrvall–Petersson double-pipe problem with Dirichlet outflow conditions, but the domain now includes five small decagonal holes with inscribed radius 0.05 positioned at $(1/2, 1/3)$, $(1/2, 2/3)$, $(1, 1/4)$, $(1, 1/2)$ and $(1, 3/4)$, as shown in Fig. 4.12. The barrier functional is as given in Section 4.1.3. We choose $\nu = 1$ and $\delta = 1$, with other variables equal to those in the original double-pipe problem. We employ the Taylor–Hood discretization for the velocity-pressure pair and a CG_1 discretization for the material distribution. We initialize at $\mu_0 = 200$, use feasible tangent prediction, and apply an l^2-minimizing linesearch in the continuation.

In total we find 42 solutions which are shown in Fig. 4.13. The holes prevent the channels passing directly from the inlets to the outlets and substantially increase the number of local minima. This example reveals that the number of local minima of a topology optimization problem is not always small and that the deflated barrier method is effective in finding many of them. A small number of solutions found exhibited regions of ambiguity $\rho \approx 1/2$, and underwent grid-sequencing and continuation in q in order to remove these areas. We note that there are more solutions that our algorithm did *not* find, since there are missing $\mathbb{Z}_2$ symmetric pairs which must also be solutions. We note that, in this example, the mesh is unstructured and, therefore, not $\mathbb{Z}_2$ symmetric.

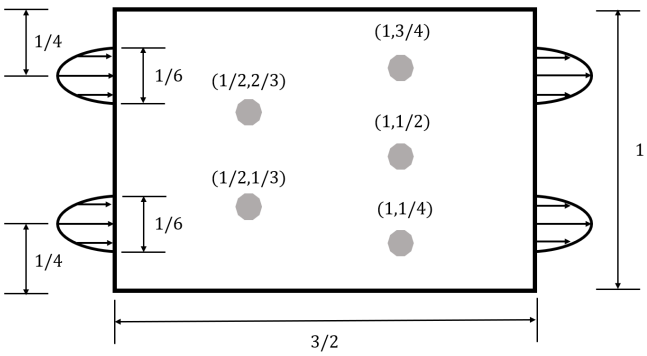


Figure 4.12: Setup of the five-holes double-pipe problem.

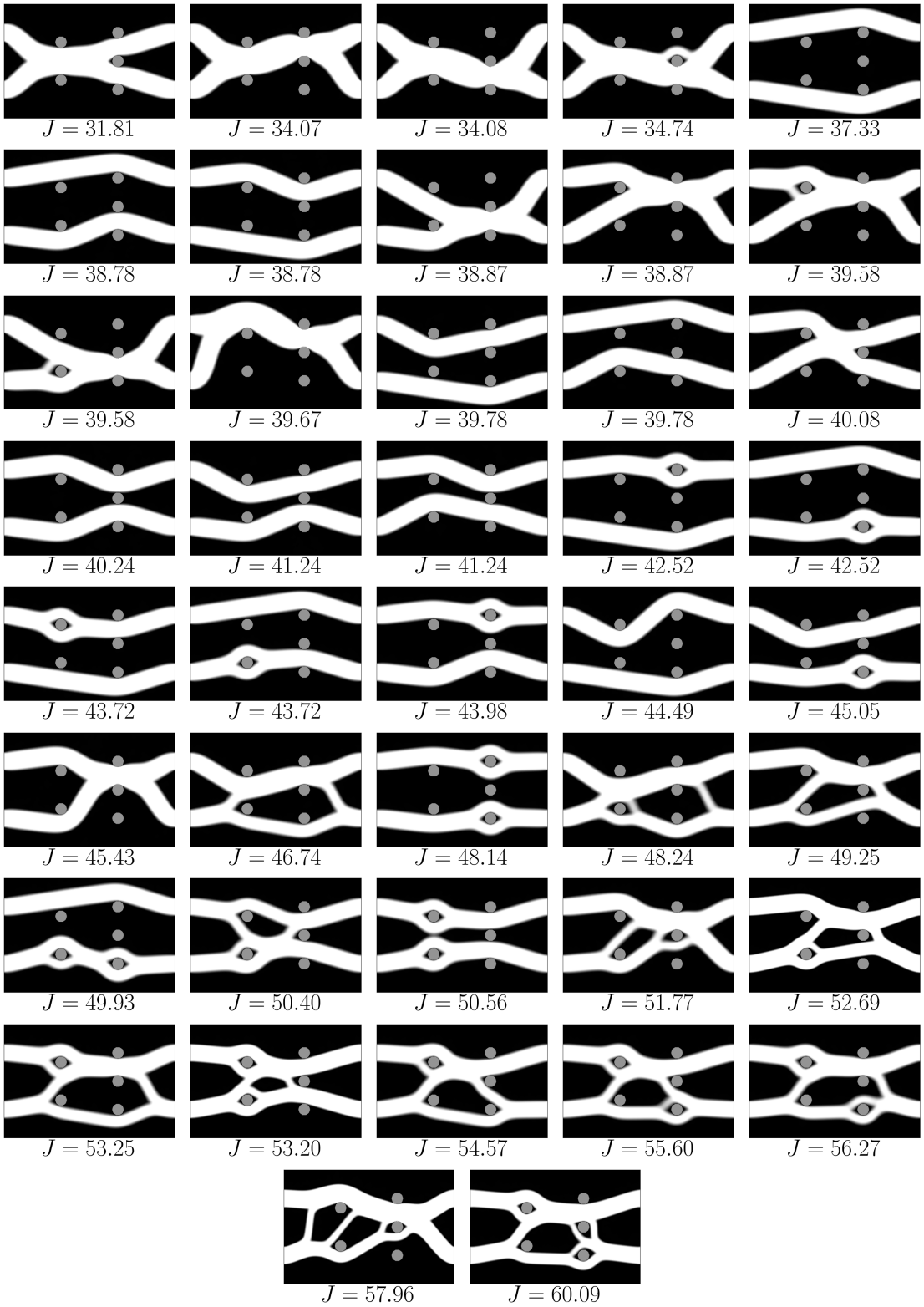


Figure 4.13: The material distribution of 42 solutions of the five-holes double-pipe optimization problem as discovered by the deflated barrier method, and their associated energies J. The fluid flow is governed by the incompressible Navier–Stokes equations. Black corresponds to a value of $\rho = 0$, white corresponds to a value of $\rho = 1$, and the grey regions are the five small holes.

4.6.6 Cantilever beam

In this example we use the deflated barrier method to find multiple stationary points of compliance problems. However, due to the lack of regularity of the Lagrange multipliers associated with the box constraints on ρ, the solver exhibits mesh-dependent behavior. With each refinement of the mesh, the number of iterations required for the solver to converge increases in an unbounded way. This is difficult to resolve, and appropriate techniques to address this are the subject of ongoing research. Practically, we first run the algorithm on a coarse mesh and then use grid-sequencing to obtain refined solutions.

The two-dimensional cantilever beam optimization problem is to find minimizers of (C) that satisfy the boundary conditions

$$\begin{aligned} \mathsf{S}\boldsymbol{n} &= (0,-1)^\top && \text{on } \Gamma_N, \\ \boldsymbol{u} &= (0,0)^\top && \text{on } \Gamma_D, \\ \mathsf{S}\boldsymbol{n} &= (0,0)^\top && \text{on } \partial\Omega\backslash\{\Gamma_N \cup \Gamma_D\}, \end{aligned}$$

with domain $\Omega = (0,1.5) \times (0,1)$, where

$$\Gamma_D = \{(x,y) \in \partial\Omega : x = 0\},$$

$$\Gamma_N = \{(x,y) \in \partial\Omega : 0.1 \leq y \leq 0.2,\ x = 1.5\} \cup \{(x,y) \in \partial\Omega : 0.8 \leq y \leq 0.9,\ x = 1.5\}\,.$$

These boundary conditions describe a cantilever clamped to the y-axis with two traction forces pulling the cantilever vertically downwards in two places at $x = 1.5$. We use CG_1 finite elements for $\boldsymbol{u}$ and ρ. The barrier functional is as given in Section 4.1.4. We initialize the deflated barrier method at $\mu_0 = 10$ and discover the second branch at $\mu = 4.25 \times 10^{-3}$. The two solutions found are shown in Fig. 4.14.

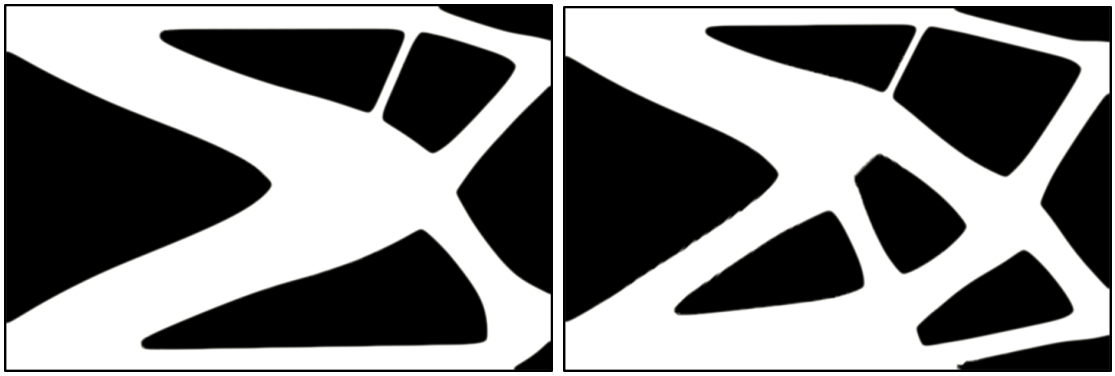


Figure 4.14: The material distribution of two solutions of the cantilever beam. The parameters are $h_{\min} = 3.56 \times 10^{-3}$, $h_{\max} = 5.70 \times 10^{-2}$, $\epsilon = 4.40 \times 10^{-3}$, $\beta = 1.8 \times 10^{-4}$, $\gamma = 0.5$, $\epsilon_{\text{SIMP}} = 10^{-5}$, $p_s = 3$, and the Lamé coefficients are $\mu_l = 75.38$ and $\lambda_l = 64.62$. $J = 6.18 \times 10^{-3}$ (left) and $J = 6.08 \times 10^{-3}$ (right).

4.6.7 Messerschmitt–Bölkow–Blohm (MBB) beam

The two-dimensional MBB beam optimization problem is to find minimizers of (C) that satisfy the boundary conditions

$$\begin{aligned} \boldsymbol{u} \cdot (1,0)^\top &= 0 && \text{on } \Gamma_{D_1}, \\ \boldsymbol{u} \cdot (0,1)^\top &= 0 && \text{on } \Gamma_{D_2}, \\ \mathsf{S}\boldsymbol{n} &= (0,-10)^\top && \text{on } \Gamma_N, \\ \mathsf{S}\boldsymbol{n} &= (0,0)^\top && \text{on } \partial\Omega \backslash \{\Gamma_N \cup \Gamma_{D_1} \cup \Gamma_{D_2}\}, \end{aligned}$$

where $\Omega = (0,3) \times (0,1)$ and

$$\Gamma_{D_1} = \{(x,y) \in \partial\Omega : x = 0\},\ \Gamma_{D_2} = \{(x,y) \in \partial\Omega : y = 0,\ 2.9 \le x \le 3\},$$
$$\Gamma_N = \{(x,y) \in \partial\Omega : y = 1,\ 0 \le x \le 0.1\}.$$

These boundary conditions describe a half-beam that is fixed horizontally on the y-axis and fixed vertically at its bottom right corner on the x-axis. There is a boundary force pushing vertically downwards at the top left corner, which represents the middle of the beam when the half-beam is mirrored. The barrier functional is as given in Section 4.1.4. We use the same finite element discretization and initialize the deflated barrier method at $\mu_0 = 50$. Deflation discovers the second

branch at $\mu = 1.58 \times 10^{-1}$. As in the cantilever problem, the algorithm is mesh-dependent and grid-sequencing is used to find refinements. The two solutions found are shown in Fig. 4.15.

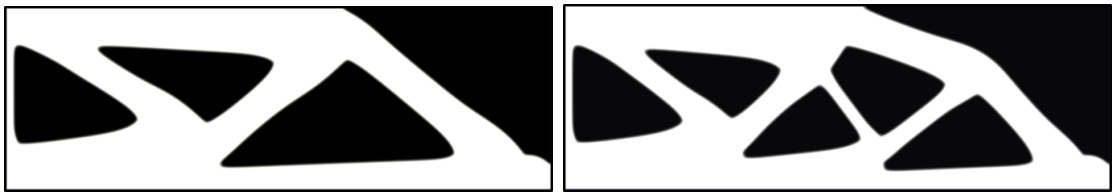


Figure 4.15: The material distribution of two solutions of the MBB beam. The parameters are $h_{\min} = 7.07 \times 10^{-3}$, $h_{\max} = 2.83 \times 10^{-2}$, $\epsilon = 1.90 \times 10^{-2}$, $\beta = 9 \times 10^{-3}$, $\gamma = 0.535$, $\epsilon_{\text{SIMP}} = 10^{-5}$, $p_s = 3$, and the Lamé coefficients are $\mu_l = 75.38$ and $\lambda_l = 64.62$. $J = 0.723$ (left) and $J = 0.681$ (right).

4.7 Code availability

The deflated barrier method algorithm, as used in all the numerical examples in this chapter, has been implemented in a Python library called deflatedbarrier, using FEniCS [109] as the finite element backend. The library can be found at https://github.com/ioannisPApapadopoulos/deflatedbarrier/. For reproducibility, the code used to run these examples has been archived on Zenodo [155, 157].

The library has been designed to quickly facilitate applying the deflated barrier method to new topology optimization problems. In a standard implementation, we first import the FEniCS library [109] and the deflated barrier library. Then, we define a Python `class` that contains information about the mesh, the mixed finite element function space, the barrier functional, and the boundary conditions of the problem as exemplified in Listing 4.1.

```
from dolfin import *
from deflatedbarrier import *

class TopologyOptimizationProblem(PrimalInteriorPoint):

    def mesh(self, comm):
        ...

    def function_space(self, mesh):
        ...

    def lagrangian(self, z, params):
```

```
13          ...
14
15      def boundary_conditions(self, Z, params):
16          ...
```

Listing 4.1: Pseudocode for defining a topology optimization problem in the deflated barrier method library.

Since the underlying library is FEniCS, which utilizes UFL [13] for compact syntax, the implementation is greatly simplified. For instance, the barrier functional for the Borrvall–Petersson problem, discretized with a conforming finite element method, can be implemented as follows:

```
1  def alpha(self, rho, params):
2          (gamma, alphabar, q) = params
3          return alphabar*(1-rho*(q+1)/(rho+q))
4
5  def lagrangian(self, z, params):
6          (rho, u, p, p0, lmbda) = split(z)
7          # rho - material distribution
8          # u - velocity
9          # p - pressure
10         # p0 - constant to fix the integral of the pressure
11         # lmbda - Lagrange multiplier for the volume constraint
12         (gamma, alphabar, q) = params
13         L = (
14                  0.5 * inner(grad(u), grad(u))*dx
15               -  inner(p, div(u))*dx
16               -  inner(p0, p)*dx
17               +  0.5 * self.alpha(rho, params) * inner(u, u)*dx
18               -  inner(lmbda, gamma-rho)*dx
19               )
20         return L
```

Listing 4.2: Implementation of the barrier functional in the Borrvall–Petersson problem.

In lines 1–3, we implement the action of the inverse permeability function $\alpha(\cdot)$ as defined in (2.22). Then, in lines 14–18, we implement the viscous term, the incompressibility constraint, the integral of the pressure, the Brinkman term, and the volume constraint, in that order.

Once the problem has been defined, we call the deflated barrier algorithm via `deflatedbarrier` or, having already computed solutions, we grid-sequence the solutions via `gridsequencing`. Both `deflatedbarrier` and `gridsequencing` are methods implemented in the deflated barrier method library.

```
1  problem = TopologyOptimizationProblem()
2  deflatedbarrier(problem, ...)
```

```
gridsequencing(problem, ...)
```

Listing 4.3: Pseudocode for calling the deflated barrier method or grid-sequencing algorithm.

Preconditioning will always be an art rather than a science.

— Andrew Wathen, 2015

5 Preconditioning

Topology optimization applications tend to be three-dimensional in nature. According to a recent review [9], around a quarter of literature dealing with the topology optimization of pure fluid flow includes the optimization of a three-dimensional example. Strategies that solve fluid three-dimensional topology optimization problems use preconditioning techniques [65], level-set implementations coupled with efficient optimization strategies [49, 54, 189], topological derivatives [141], adaptive methods [26, 95], lattice Boltzmann methods [128, 183], and efficient parallel implementations [2, 3], or enforce symmetry in the problem to reduce the three-dimensional problem to two dimensions [15]. The need for these methods is caused by the increase in the size of the linear problems that are solved during the optimization process. The computational effort is often further impacted by a solvers' mesh-dependence, ultimately rendering three-dimensional topology optimization computationally expensive.

In this chapter we restrict our focus to the three-dimensional Borrvall–Petersson problem and develop iterative methods for the linear systems that arise in the deflated barrier method. We choose a divergence-free DG BDM finite element discretization for the velocity and pressure pair [40, 41]. Preconditioning strategies are required for iterative methods to converge within an acceptable number of iterations. We will show that block preconditioning can reduce the linear systems arising in the deflated barrier method to ones that resemble the systems arising in the discretization of the Stokes–Brinkman equations [65]. Then, we apply modern block preconditioning, pioneered by Wathen and coworkers [60, 158, 174], to further

reduce the linear systems. The block preconditioning is combined with an augmented Lagrangian term to control the innermost Schur complement term. Finally, we develop a geometric multigrid method for the augmented momentum block with a vertex-star patch relaxation that captures the kernel of the augmented Lagrangian term [71, 73, 91, 143]. The multigrid scheme requires a characterization of the active set on all levels of the mesh hierarchy, which we discuss.

Throughout this chapter, we fix our finite element spaces as, for $k \geq 1$:

$$C_{\gamma,h} \subset C_{[0,1],h} = \{\eta_h \in X_{\mathrm{DG}_0} : 0 \leq \eta_h \leq 1 \text{ a.e. in } \Omega\}, \tag{5.1}$$

$$\boldsymbol{X}_h = \boldsymbol{X}_{\mathrm{BDM}_k}, \tag{5.2}$$

$$M_h = X_{\mathrm{DG}_{k-1}}. \tag{5.3}$$

This choice of finite element spaces satisfies Theorem 3.2. Hence, by Corollary 3.3, for each isolated local minimizer of (BP), there exists a sequence of finite element solutions $(\boldsymbol{u}_h, \rho_h, p_h, \lambda_h) \in \boldsymbol{X}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times C_{\gamma,h} \times M_h \times \mathbb{R}$ to the system comprised of (FOC1-DG$_h$), (FOC2-DG$_h$), (FOC3b-DG$_h$), and (FOC4-DG$_h$) such that $(\boldsymbol{u}_h, \rho_h, p_h)$ strongly converges to the minimizer in $H^1_{\boldsymbol{g}}(\mathcal{T}_h)^d \times L^s(\Omega) \times L^2(\Omega)$, $s \in [1, \infty)$, as $h \to 0$.

5.1 Benson–Munson linear system

We restrict our analysis to the BM solver introduced in Section 4.2 since the BM solver generally required fewer iterations than the HIK solver. To recap: the BM solver attempts to find roots of a complementarity problem via linearizations of the residual constrained to the inactive set. First, the discrete Newton system is formed and the indices in the active set are identified. The active set contains the degrees of freedom that satisfy a strict complementarity condition in the primal and residual vectors. Next, the rows and columns of the Jacobian in the Newton system associated with the active set degrees of freedom are set to those of the identity. Finally, the rows on the right-hand side vector associated with the active set degrees of freedom are fixed to zero. Once the update, $\delta \mathbf{z}_k$, of this modified system is computed, the new iterate $\mathbf{z}_{k+1} = \mathbf{z}_k + \delta \mathbf{z}_k$ is component-wise projected onto the box constraints.

For a given barrier parameter $\mu \geq 0$ and interior penalty penalization parameter $\sigma > 0$, the barrier functional we consider in this chapter is given by

$$L_{\mu}^{\epsilon_{\log}}(\boldsymbol{u}, \rho, p, \lambda) := J_{h,\mu}^{\epsilon_{\log}}(\boldsymbol{u}, \rho) - \int_{\Omega} p \operatorname{div}(\boldsymbol{u}) \mathrm{d}x - \int_{\Omega} \lambda(\gamma - \rho) \mathrm{d}x, \tag{5.4}$$

where $J_{h,\mu}^{\epsilon_{\log}}$ is the functional J_h (as defined in (3.55)) augmented with the barrier terms as discussed in Section 4.1.1. Unlike in Section 4.1.1, in this chapter, we fix the integral of the pressure implicitly in the solver and we do not require the addition of an additional Lagrange multiplier $p_0 \in \mathbb{R}$. This is achieved by orthogonalizing against the nullspace of the discretized pressure, which is spanned by the constant vector, within the Krylov method [24]. This strategy was not possible in the previous chapter where we used a direct LU factorization with no outer Krylov method.

By deriving the first-order optimality conditions for $\boldsymbol{u}$, ρ, p, and λ and employing the divergence-free DG finite element discretization, we note that the deflated barrier method subproblem is to find $(\rho_h, \boldsymbol{u}_h, p_h, \lambda_h) \in C_{\gamma,h} \times \boldsymbol{X}_{\boldsymbol{g}_h,\boldsymbol{n},h} \times M_h \times \mathbb{R}$ such that, for all $(\eta_h, \boldsymbol{v}_h, q_h, \zeta_h) \in C_{[0,1],h} \times \boldsymbol{X}_{0,\boldsymbol{n},h} \times M_h \times \mathbb{R}$,

$$a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) = l_{h,\boldsymbol{f},\boldsymbol{g}_h}(\boldsymbol{v}_h), \tag{5.5}$$

$$b(\boldsymbol{u}_h, q_h) = 0, \tag{5.6}$$

$$c_{\boldsymbol{u}_h,\lambda_h}^{\mu,\epsilon_{\log}}(\rho_h, \eta_h - \rho_h) \geq 0, \tag{5.7}$$

$$d_{\rho_h}(\lambda_h, \zeta_h) = 0. \tag{5.8}$$

The forms a_{h,ρ_h}, b, $l_{h,\boldsymbol{f},\boldsymbol{g}_h}$, $c_{\boldsymbol{u}_h,\lambda_h}^{\mu,\epsilon_{\log}}$, and d_{ρ_h} are as defined in (3.58), Proposition 2.4, (3.59), (4.6), and (FOC4), respectively. We note that (5.5), (5.6), and (5.8) are the same equations as (FOC1-DG$_h$), (FOC2-DG$_h$), and (FOC4-DG$_h$), respectively. Moreover, (5.7) reduces to (FOC3b-DG$_h$) when $\mu = 0$.

We now derive the linear systems that arise by using the BM active set strategy to solve the nonlinear system (5.5)–(5.8). Denote the basis functions of the finite element spaces of the material distribution, the velocity, the pressure, and $\mathbb{R}$ by $\eta_i, \boldsymbol{\phi}_i, \psi_i$, and r, respectively and the number of degrees of freedom by $n_\rho, n_{\boldsymbol{u}}$, n_p, and 1, respectively, so that the total number of degrees of freedom of the system is given by $n = n_\rho + n_{\boldsymbol{u}} + n_p + 1$. Consider the finite element BM iterate

$\boldsymbol{z}_{h,k} = (\rho_{h,k}, \boldsymbol{u}_{h,k}, p_{h,k}, \lambda_{h,k})$. Let $\mathbf{f}_k : \mathbb{R}^n \to \mathbb{R}^n$ denote the nonlinear residual induced by complementarity reformulation of (5.5)–(5.8). In the following, we drop the subscript iteration number $_k$ for clarity. Let $\mathbf{z}$ denote the discrete coefficient vector of $\boldsymbol{z}_h$ and define the active set by

$$\mathcal{A} = \{i : \mathbf{z}_i = \mathbf{a}_i \text{ and } \mathbf{f}(\mathbf{z})_i > 0\} \cup \{i : \mathbf{z}_i = \mathbf{b}_i \text{ and } \mathbf{f}(\mathbf{z})_i < 0\}, \tag{5.9}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the lower- and upper-bound box constraints, respectively. In this context, $\mathbf{a}_i = 0$ and $\mathbf{b}_i = 1$ for all degrees of freedom associated with ρ_h and $\mathbf{a}_i = -\infty$, $\mathbf{b}_i = +\infty$, otherwise. Define the inactive set as $\mathcal{I} = \{i\}_{i=1}^n \backslash \mathcal{A}$. Then, the BM updates are computed by solving the linear system:

$$\mathsf{H}_{\rho,\boldsymbol{u},p,\lambda}\delta\mathbf{z} = \begin{pmatrix} \mathsf{C}_\mu & \mathsf{D}^\top & \mathsf{0} & \mathsf{E}^\top \\ \mathsf{D} & \mathsf{A} & \mathsf{B}^\top & \mathsf{0} \\ \mathsf{0} & \mathsf{B} & \mathsf{0} & \mathsf{0} \\ \mathsf{E} & \mathsf{0} & \mathsf{0} & \mathsf{0} \end{pmatrix} \begin{pmatrix} \delta\boldsymbol{\rho} \\ \delta\mathbf{u} \\ \delta\mathbf{p} \\ \delta\boldsymbol{\lambda} \end{pmatrix} = - \begin{pmatrix} \mathbf{f}_\rho \\ \mathbf{f}_{\boldsymbol{u}} \\ \mathbf{f}_p \\ \mathbf{f}_\lambda \end{pmatrix} = -\mathbf{f}, \tag{5.10}$$

where $\delta\boldsymbol{\rho}, \delta\mathbf{u}, \delta\mathbf{p}$ and $\delta\boldsymbol{\lambda}$ denote the discrete coefficient vector BM updates for $\rho, \boldsymbol{u}, p$ and λ, and $\mathbf{f}_\rho$, $\mathbf{f}_{\boldsymbol{u}}$, $\mathbf{f}_p$, and $\mathbf{f}_\lambda$ are the corresponding blocks of the nonlinear residual with the active set rows, $i \in \mathcal{A}$, in $\mathbf{f}_\rho$ zeroed. The entries of A and B are given by

$$[\mathsf{A}]_{ij} = a_{h,\rho_h}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) \text{ and } [\mathsf{B}]_{ij} = b(\boldsymbol{\phi}_j, \psi_i). \tag{5.11}$$

Furthermore, if $j \in \mathcal{I}$, then

$$[\mathsf{D}]_{ij} = \int_\Omega (\alpha'(\rho_h)\boldsymbol{u}_h \cdot \boldsymbol{\phi}_i)\eta_j \, \mathrm{d}x, \quad [\mathsf{E}]_{ij} = -r \int_\Omega \eta_j \, \mathrm{d}x. \tag{5.12}$$

Otherwise, if $j \in \mathcal{A}$, then $[\mathsf{D}]_{ij} = 0$ and $[\mathsf{E}]_{ij} = 0$ for all i. Finally, if $i, j \in \mathcal{I}$, then

$$[\mathsf{C}_\mu]_{ij} = \int_\Omega \left[\frac{1}{2}\alpha''(\rho_h)|\boldsymbol{u}_h|^2 + \frac{\mu}{(\rho_h + \epsilon_{\log})^2} + \frac{\mu}{(1 + \epsilon_{\log} - \rho_h)^2} \right] \eta_i \eta_j \mathrm{d}x. \tag{5.13}$$

Otherwise, if $i \in \mathcal{A}$ or $j \in \mathcal{A}$, then $[\mathsf{C}_\mu]_{ij} = \delta_{ij}$, where δ_{ij} is the Kronecker delta.

Remark 5.1. E *is a row vector of size* $1 \times n_\rho$.

Proposition 5.1. *The matrix* $\mathsf{A} \in \mathbb{R}^{n_{\boldsymbol{u}} \times n_{\boldsymbol{u}}}$ *is symmetric, and provided the penalization parameter* $\sigma > 0$ *is sufficiently large, then it also positive-definite.*

Proof. Symmetry is realized by swapping the indices i and j of the basis functions in their respective definitions and noting that the resulting integrals are equal. Positive-definiteness of $\mathbf{A}$, for sufficiently large $\sigma > 0$, follows from $\alpha(\rho) \geq 0$ and [91, Sec. 3.3]. □

Proposition 5.2. *Suppose that $\mu > 0$. Then, the matrix $\mathbf{C}_\mu$ is symmetric positive-definite.*

Proof. The symmetry of $\mathbf{C}_\mu$ is realized by swapping the indices i and j of the basis functions in its definition and noting that the resulting integrals are equal. Consider the unmodified matrix $\hat{\mathbf{C}}_\mu$, defined by (5.13) for all i and j. Pick an arbitrary function $\eta_h \in X_{\mathrm{DG}_0}$ and its discrete coefficient vector $\boldsymbol{\eta} \in \mathbb{R}^{n_\rho}$. We note that

$$\boldsymbol{\eta}^\top \hat{\mathbf{C}}_\mu \boldsymbol{\eta} = \int_\Omega \left[\frac{1}{2}\alpha''(\rho_h)|\boldsymbol{u}_h|^2 + \frac{\mu}{(\rho_h + \epsilon_{\log})^2} + \frac{\mu}{(1 + \epsilon_{\log} - \rho_h)^2} \right] |\eta_h|^2 \mathrm{d}x. \tag{5.14}$$

Assumption (A5) implies that $\alpha''(\rho_h) \geq 0$ and by definition $|\boldsymbol{u}_h|^2 \geq 0$. Moreover, the rational expressions are greater than zero for $\mu > 0$ as $0 \leq \rho_h \leq 1$. Hence, the right-hand side of (5.14) is equal to zero if and only if $\eta_h = 0$, which is true if and only if $\boldsymbol{\eta} = \mathbf{0}$. Therefore, $\boldsymbol{\eta}^\top \hat{\mathbf{C}}_\mu \boldsymbol{\eta} \geq 0$ with equality if and only if $\boldsymbol{\eta} = \mathbf{0}$. Hence, $\hat{\mathbf{C}}_\mu$ is symmetric positive-definite. Since the discretization for ρ is piecewise constant, $\hat{\mathbf{C}}_\mu$ is a diagonal matrix and, therefore, all diagonal entries must be positive. The procedure of zeroing rows and columns associated with the BM active set and replacing the diagonal entry with a one will result in a diagonal matrix with positive diagonal entries. We conclude that $\mathbf{C}_\mu$ must be symmetric positive-definite. □

5.2 Preconditioning

In this section, we develop a preconditioner for solving (5.10). As discussed in Section 4.3, preconditioning strategies that are robust for the undeflated system can also be used to compute solutions of the deflated systems. On the outermost level of the deflated barrier method, we perform continuation in the barrier parameter μ. Next, at a given μ, we use the BM solver to find a solution of (5.5)–(5.8). At each BM iteration, we use a preconditioned FGMRES method [142].

A direct sparse LU factorization of (5.10) is infeasible on fine meshes of three-dimensional problems. Thus we turn to preconditioning techniques to reduce the cost of each inner linear solve. The preconditioning is made difficult by the saddle point nature of the matrix in (5.10) and the barrier terms in C_μ. In the following subsections we introduce a nested block preconditioning method for solving (5.10), where the Schur complements are controlled with an augmented Lagrangian term.

5.2.1 Block preconditioning

Consider the well-posed linear system

$$\begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{C} & \mathbb{D} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix}, \tag{5.15}$$

where $\mathbb{A} \in \mathbb{R}^{n_1 \times n_1}$ is invertible, $\mathbb{B} \in \mathbb{R}^{n_1 \times n_2}$, $\mathbb{C} \in \mathbb{R}^{n_2 \times n_1}$ and $\mathbb{D} \in \mathbb{R}^{n_2 \times n_2}$. Then, the inverse of the matrix in (5.15) admits a full block factorization of the form [32]

$$\begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{C} & \mathbb{D} \end{pmatrix}^{-1} = \begin{pmatrix} I & -\mathbb{A}^{-1}\mathbb{B} \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbb{A}^{-1} & 0 \\ 0 & \mathbb{S}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\mathbb{C}\mathbb{A}^{-1} & I \end{pmatrix}, \tag{5.16}$$

where $\mathbb{S} = \mathbb{D} - \mathbb{C}\mathbb{A}^{-1}\mathbb{B}$. Good preconditioners for (5.15) can be found by developing cheap approximations to $\mathbb{A}^{-1}$ and $\mathbb{S}^{-1}$ and substituting them into (5.16) [114, 174].

The subspace spanned by the volume constraint Lagrange multiplier λ is one-dimensional and can be handled by at most one iteration of a Krylov subspace solver or via block preconditioning. Experimentally, we found that a full block preconditioning of the real block performed best. Consider the density-momentum-pressure block:

$$\mathsf{H}_{\rho,\boldsymbol{u},p} := \begin{pmatrix} \mathsf{C}_\mu & \mathsf{D}^\top & \mathbf{0} \\ \mathsf{D} & \mathsf{A} & \mathsf{B}^\top \\ \mathbf{0} & \mathsf{B} & \mathbf{0} \end{pmatrix}. \tag{5.17}$$

Then, we choose $\mathbb{A} = \mathsf{H}_{\rho,\boldsymbol{u},p}$, $\mathbb{B} = \mathsf{E}^\top$, $\mathbb{C} = \mathsf{E}$, and $\mathbb{D} = \mathbf{0}$. The Schur complement,

$$\mathbb{S} = \mathsf{S}_0 := -\mathsf{E}\mathsf{H}^{-1}_{\rho,\boldsymbol{u},p}\mathsf{E}^\top, \tag{5.18}$$

is one-dimensional and can be inverted by taking its reciprocal. Hence, the difficulty now lies in solving linear systems involving (5.17). Since we have only decreased the

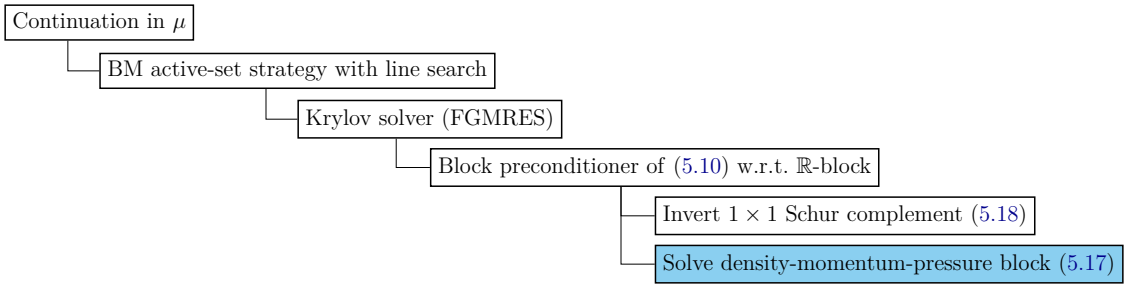


Figure 5.1: Main components of the deflated barrier method solver. The remainder of this section will focus on developing preconditioners for the item in blue.

size of the linear system by one dimension, an LU factorization is still infeasible and we consider block preconditioners for (5.17). We summarize the initial components of the solver in Fig. 5.1.

A block preconditioner approach for (5.17) is to take the Schur complement of (5.17) with respect to the momentum-pressure block. This approach was utilized by Evgrafov for preconditioning the linear systems arising in a similar solver [65, Sec. 5]. In the notation of (5.15), $\mathbb{A} = \mathsf{C}_\mu$, $\mathbb{B} = (\mathsf{D}^\top \;\; \mathbf{0})$, $\mathbb{C} = (\mathsf{D} \;\; \mathbf{0})^\top$, and

$$\mathbb{D} = \begin{pmatrix} \mathsf{A} & \mathsf{B}^\top \\ \mathsf{B} & \mathbf{0} \end{pmatrix}. \tag{5.19}$$

We know that C_μ is invertible by Proposition 5.2. Hence, we can write $\mathbb{A}^{-1} = \mathsf{C}_\mu^{-1}$. We define $\hat{\mathsf{S}}_1$ by

$$\hat{\mathsf{S}}_1 := \begin{pmatrix} \mathsf{A} & \mathsf{B}^\top \\ \mathsf{B} & \mathbf{0} \end{pmatrix}, \tag{5.20}$$

and the true Schur complement is given by

$$\mathbb{S} = \mathsf{S}_1 := \hat{\mathsf{S}}_1 - \begin{pmatrix} \mathsf{D} \\ \mathbf{0} \end{pmatrix} \mathsf{C}_\mu^{-1} \begin{pmatrix} \mathsf{D}^\top & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathsf{A} - \mathsf{D}\mathsf{C}_\mu^{-1}\mathsf{D}^\top & \mathsf{B}^\top \\ \mathsf{B} & \mathbf{0} \end{pmatrix}. \tag{5.21}$$

The reason we use a DG_0 piecewise constant discretization for the material distribution is to ensure that S_1 is sparse. Since the material distribution is discretized with DG_0 finite elements, C_μ is a diagonal matrix, $\mathsf{D}\mathsf{C}_\mu^{-1}\mathsf{D}^\top$ is still sparse and hence S_1 is also sparse.

Remark 5.2. *The Schur complement approximation $\hat{\mathsf{S}}_1$ resembles the linear system that arises in the discretization of the Stokes–Brinkman equations (introduced in Remark 2.1).*

The proposed application of block preconditioning has reduced solving the linear system (5.10) to the following:

1. An outer FGMRES solver;
2. Apply the reciprocal of $\mathbf{S}_0 \in \mathbb{R}$;
3. Invert the diagonal matrix $\mathbf{C}_\mu$;
4. Apply the action of the inverse of the 2×2 block matrix $\mathbf{S}_1$. $\mathbf{S}_1$ is the same size as the matrix that arises in a discretized pure Stokes problem.

We must now develop solvers for $\mathbf{S}_1$ as given in (5.21). One option is to use a direct solver. However, we can further reduce the computational work with another application of block preconditioning. Consider taking the inner Schur complement in $\mathbf{S}_1$ with respect to the pressure block. Using the notation of (5.15), $\mathbb{B} = \mathbf{B}^\top$, $\mathbb{C} = \mathbf{B}$, and $\mathbb{A} = \mathbf{A} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$. The inner Schur complement takes the form

$$\mathbb{S} = \mathbf{S}_2 := -\mathbf{B}(\mathbf{A} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top)^{-1}\mathbf{B}^\top. \tag{5.22}$$

This time, $\mathbf{S}_2$ is dense, and we require a sparse approximation. Let $-\Delta_h$ denote the negative discretized Laplacian matrix. In the context of the incompressible Stokes equations, it is well known [60, Th. 5.22], that $\mathbf{B}(-\Delta_h)^{-1}\mathbf{B}^\top$ is spectrally equivalent to the viscosity-scaled pressure mass matrix $\nu^{-1}\mathbf{M}_p$, i.e. there exist constants c_1 and c_2, independent of the dimension of the problem such that

$$c_1^2 \leq \frac{\mathbf{q}^\top \mathbf{B}(-\Delta_h)^{-1}\mathbf{B}^\top \mathbf{q}}{\nu^{-1}\mathbf{q}^\top \mathbf{M}_p \mathbf{q}} \leq c_2^2 \quad \text{for any} \quad \mathbf{q} \in \mathbb{R}^{n_p}. \tag{5.23}$$

In the case of homogeneous Dirichlet boundary conditions on $\boldsymbol{u}_h$, $c_2 = 1$ and $c_1 = c_b$, where c_b is the inf-sup constant. $\mathbf{M}_p$ is a sparse mass matrix and can be cheaply factorized or solved with a multigrid method. Therefore, a good approximation to the Schur complement of the pure Stokes problem is given by $\nu^{-1}\mathbf{M}_p$. The idea is that the momentum block can then be solved with a direct solver, multigrid methods, or other alternative solvers. Unfortunately, in the context of the Stokes–Brinkman equations, Popov [129] noted that the presence of the Brinkman term $\alpha(\rho_h)\boldsymbol{u}_h$ in the momentum block $\mathbf{A}$ renders the approximation given by $\nu^{-1}\mathbf{M}_p$ ineffective.

Popov proposed a Schur complement preconditioning technique based on incomplete LU factorization, but such factorizations do not generally yield mesh-independent preconditioners. An alternative is a preconditioning scheme utilized by Borrvall and Petersson in their original paper [36, Sec. 2.6] based on the work of Cahouet and Chabard [48]. However, we found that an augmented Lagrangian approach performed better. We compare the effectiveness of the augmented Lagrangian approach with the Cahouet–Chabard preconditioning strategy in Section 5.3.1 below.

In the remainder of this subsection, we propose an augmented Lagrangian strategy to control the second Schur complement $\mathbf{S}_2$. The augmented Lagrangian strategy converges in $\mathcal{O}(1)$ outer FGMRES iterations per BM iteration with no inner Krylov iterations required when a direct solver is used for the augmented momentum block, scaled pressure mass matrix, and density block (which is a diagonal matrix). The augmented Lagrangian approach has also been shown to be robust for a variety of difficult saddle-point systems such as the stationary Navier–Stokes equations at high Reynolds number [73]. Hence, this approach has potential for extension to different fluid topology optimization problems.

There are two possible augmented Lagrangian approaches: continuous and discrete. These approaches are mathematically equivalent for exactly divergence-free elements such as the BDM mixed finite element. We choose to introduce the method in the discrete setting. Post-discretization, the matrix $\mathbf{A}$ in (5.10) is modified by adding an augmented Lagrangian term

$$\mathbf{A}_{\gamma_d} := \mathbf{A} + \gamma_d \mathbf{B}^\top \mathbf{M}_p^{-1} \mathbf{B}, \tag{5.24}$$

where $\gamma_d \gg 0$, and the right-hand side of (5.10) is modified so that the solution of linear system remains unchanged (since $\mathbf{Bu}$ is known). In particular, if the current velocity iterate is divergence-free, then $\gamma_d \mathbf{B}^\top \mathbf{M}_p^{-1} \mathbf{B} \delta\mathbf{u} = \mathbf{0}$. While it does not change the solution, the addition of the augmented Lagrangian term influences the nature of the inner Schur complements. In particular, $\mathbf{S}_1$ becomes

$$\mathbf{S}_{1,\gamma_d} = \begin{pmatrix} \mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{pmatrix}, \tag{5.25}$$

and $\mathbf{S}_2$ becomes

$$\mathbf{S}_{2,\gamma_d} = -\mathbf{B}(\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top)^{-1}\mathbf{B}^\top \approx -\mathbf{B}(\gamma_d\mathbf{B}^\top\mathbf{M}_p^{-1}\mathbf{B})^{-1}\mathbf{B}^\top \approx -\gamma_d^{-1}\mathbf{M}_p. \tag{5.26}$$

The approximation of $\mathbf{S}_{2,\gamma_d}$ by the scaled pressure mass matrix improves as $\gamma_d \to \infty$ [33, Th. 4.2]. If assembled naïvely, the triple matrix product $\mathbf{B}^\top\mathbf{M}_p^{-1}\mathbf{B}$, as it occurs in the augmented Lagrangian term, is expensive to compute. However, it can be checked that the augmented Lagrangian term $\gamma_d\mathbf{B}^\top\mathbf{M}_p^{-1}\mathbf{B}$ corresponds to augmenting the weak form $a_{h,\rho_h}(\boldsymbol{u}_h, \boldsymbol{v}_h)$ in (5.5) by

$$\gamma_d \int_\Omega \Pi_{M_h}(\mathrm{div}(\boldsymbol{u}_h))\Pi_{M_h}(\mathrm{div}(\boldsymbol{v}_h))\mathrm{d}x, \tag{5.27}$$

where Π_{M_h} is the projection onto the discretized pressure space M_h. The projection is the identity for the BDM-DG pair. Therefore, assembling $\mathbf{B}\mathbf{M}_p^{-1}\mathbf{B}^\top$ is equivalent to assembling the matrix associated with the bilinear form $\int_\Omega \mathrm{div}(\boldsymbol{\phi}_i)\mathrm{div}(\boldsymbol{\phi}_j)\mathrm{d}x$, where $\boldsymbol{\phi}_i$, $i = 1, \ldots, n_{\boldsymbol{u}}$, are the basis functions of the velocity finite element space.

With the proposed nested block preconditioning, we have reduced solving linear systems involving (5.10) into the following steps:

1. An outer FGMRES solver;
2. Apply the reciprocal of $\mathbf{S}_0 \in \mathbb{R}$;
3. Invert the diagonal matrix $\mathbf{C}_\mu$;
4. Factorize and solve the block-diagonal pressure mass matrix $\mathbf{M}_p$;
5. Apply the action of the inverse of the augmented momentum block $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$.

Factorizing $\mathbf{C}_\mu$, $\mathbf{M}_p$, and $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ with a direct solver such as MUMPS [17] is faster than factorizing the full matrix in (5.10). We note that $\mathbf{M}_p$ only needs to be factorized once at the start of the algorithm. Hence, most of the computational time during the run of the deflated barrier method is spent on factorizing $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ at each BM iteration.

Remark 5.3. *A property of divergence-free DG discretizations utilized in our proposed nested block preconditioning approach is the identification of the projection* Π_{M_h} *as the identity. This is used in* (5.27) *to cheaply construct the triple matrix product* $\mathbf{B}\mathbf{M}_p^{-1}\mathbf{B}^\top$ *in the augmented momentum block. We note that the projection* Π_{M_h} *can also been characterized in other discretizations, e.g. the Scott–Vogelius* $(\mathrm{CG}_k)^d \times \mathrm{DG}_{k-1}$ *pair (paired with a barycentrically refined mesh if* $k \in [d, 2d]$*) [72], the* $(\mathrm{CG}_d)^d \times \mathrm{DG}_0$ *pair, and the* $(\mathrm{CG}_1 \oplus B_3^F)^3 \times \mathrm{DG}_0$ *pair where* $(\mathrm{CG}_1 \oplus B_3^F)^3$ *represents a piecewise linear velocity space enriched with bubble functions on each facet [73].*

Remark 5.4. *In a continuous augmented Lagrangian approach, the projection* Π_{M_h} *is neglected and the barrier functional* $L_\mu^{\epsilon_{log}}(\boldsymbol{u}, \rho, p, \lambda)$ *in* (5.4) *is augmented with the term*

$$\frac{\gamma_c}{2}\int_\Omega \mathrm{div}(\boldsymbol{u}_h)\,\mathrm{div}(\boldsymbol{u}_h)\mathrm{d}x. \tag{5.28}$$

After a suitable discretization, for sufficiently large γ_c*, the approximation* (5.26) *still holds; although for methods that are not divergence-free, the velocity finite element solution* $\boldsymbol{u}_h$ *and update* $\delta\mathbf{u}$ *are different to that of the unaugmented system. An advantage of the continuous approach is that* (5.28) *can be used for preconditioning the BM systems arising in any conforming finite element discretization, such as the Taylor–Hood mixed finite element. Therefore, our block preconditioning approach would still yield an effective decrease in the computational work to solve linear systems in* (5.10)*, when a direct solver is used for the augmented momentum block. A continuous approach comes at the cost that it can be difficult to develop an effective multigrid cycle for the augmented momentum block when* Π_{M_h} *is not the identity.*

In Fig. 5.2, we summarize the block preconditioning strategy for solving linear systems involving the density-momentum-pressure block (5.17). In the next section we develop a specialized geometric multigrid scheme to efficiently solve linear systems involving $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ (highlighted in pink in Fig. 5.2) in order to reduce the computational time further when the problem is discretized on a fine mesh.

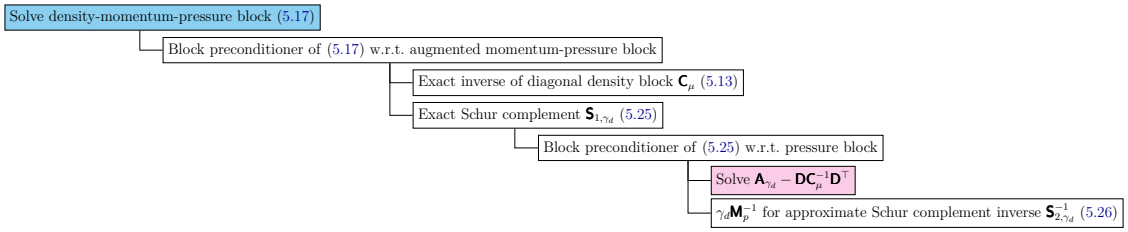


Figure 5.2: The preconditioning strategy to solve the density-momentum-pressure block (5.17). We develop a geometric multigrid scheme for the item in pink in Section 5.2.2.

5.2.2 A specialized multigrid scheme for $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$

The tradeoff for using an augmented Lagrangian term to control the Schur complement $\mathbf{S}_{2,\gamma_d}$, defined in (5.26), is that $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ becomes difficult to solve, due to the semi-definite term with a large coefficient $\gamma_d \gg 0$. Recently, there has been progress on specialized multigrid schemes, based on the work of Schöberl [143], to handle the effects of the augmented Lagrangian term in $\mathbf{A}_{\gamma_d}$. Such a strategy has been shown to be extremely effective in parameter-robust preconditioning of the three-dimensional incompressible Navier–Stokes equations [72, 73], Oseen–Frank models of cholesteric liquid crystals [178], implicitly-constituted non-Newtonian incompressible flow [67, 69], and magnetohydrodynamics [104]. The multigrid scheme has also been analyzed in the context of the $\boldsymbol{H}(\mathrm{div};\Omega)$ Riesz map [21] and, more relevant to our problem, an $\boldsymbol{H}(\mathrm{div};\Omega)$-conforming discretization of the Stokes equations [91].

Recall that a typical multigrid solver requires the following components:

(MG1) A relaxation method;

(MG2) A solver for the coarse-level correction;

(MG3) Transfer operators to inject functions from fine to coarse levels and prolong functions from coarse to fine levels;

(MG4) Construction of the coarse-level operators, i.e. the representation of $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ on coarser levels.

For a discussion on multigrid methods in the context of preconditioning we refer to [174]. In this work, we construct a mesh hierarchy and construct our multigrid solver as a geometric multigrid method. The coarse-level operators are induced

from (5.5)–(5.8) via rediscretization. Moreover, in all our examples, we choose a direct solver for the coarse-level solver. Schöberl's analysis gives requirements on (MG1) and (MG3) to achieve robustness in the context of multigrid cycles applied to symmetric positive-definite problems augmented with a parameter-dependent positive semi-definite term. The first is that the relaxation method must stably capture the kernel of the semi-definite term. The second requirement is that the prolongation operator must have a continuity constant that is independent of γ_d. As noted by Hong et al. [91, Sec. 1], in a nested mesh hierarchy, an exactly divergence-free function on the coarse-grid will be divergence-free on the fine-grid. Therefore, in our context, the natural prolongation operator suffices thanks to our choice of a divergence-free DG discretization, and we only discuss the relaxation method in this work. The kernel of the semi-definite term involving γ_d is

$$\mathcal{N}_h = \{\boldsymbol{w}_h \in \boldsymbol{X}_{\mathrm{BDM}_k} : (\mathrm{div}(\boldsymbol{w}_h), \mathrm{div}(\boldsymbol{v}_h))_{L^2(\Omega)} = 0 \text{ for all } \boldsymbol{v}_h \in \boldsymbol{X}_{\mathrm{BDM}_k}\}, \quad (5.29)$$

i.e. all functions with divergence zero. For large γ_d, A_{γ_d} becomes increasingly singular. Common relaxation methods like Jacobi and Gauss-Seidel do not offer γ_d-robust smoothing and yield ineffective multigrid cycles. To understand the degradation of Jacobi and Gauss–Seidel as $\gamma_d \to \infty$, it is fruitful to view the relaxation method as a subspace correction method [181, 182]. Consider the space decomposition

$$\boldsymbol{X}_{\mathrm{BDM}_k} = \sum_i \boldsymbol{X}_i, \quad (5.30)$$

where the sum is not necessarily direct. In Jacobi and Gauss-Seidel, the decomposition, (5.30), is given by $\{\boldsymbol{X}_i\} = \{\boldsymbol{\phi}_i\}$ where $\boldsymbol{\phi}_i$, $i = 1, \ldots, n_{\boldsymbol{u}}$, are the velocity basis functions. The difference between Jacobi and Gauss-Seidel is whether the updates are applied in parallel (Jacobi) or sequentially (Gauss-Seidel).

A necessary condition for the subspace correction method induced by the decomposition (5.30) to be robust in γ_d for a symmetric positive-definite matrix, is that the decomposition captures the kernel $\mathcal{N}_h$ in the following sense:

$$\mathcal{N}_h = \sum_i \boldsymbol{X}_i \cap \mathcal{N}_h. \quad (5.31)$$

In other words, the decomposition must be sufficiently rich so that any divergence-free velocity can be written as a combination of divergence-free functions from the subspaces $\boldsymbol{X}_i$. Jacobi fails this criterion, as each $\boldsymbol{\phi}_i$ is not divergence-free. A decomposition satisfying (5.31) was developed by Hong et al. [91, Sec. 4.5], where the decomposition is the star patch around every vertex of the mesh. This decomposition is visualized in Fig. 5.3 for a BDM_1 discretization in two dimensions and is easily extended to higher orders and three dimensions. Since $\mathsf{A}_{\gamma_d} - \mathsf{D}\mathsf{C}_\mu^{-1}\mathsf{D}^\top$ is not guaranteed to be positive-definite, the theory does not guarantee robust convergence. Nevertheless, we find that a small number of FGMRES iterations preconditioned with the vertex-star patch iteration is very effective as a smoother, as reported in [73] and subsequent works.

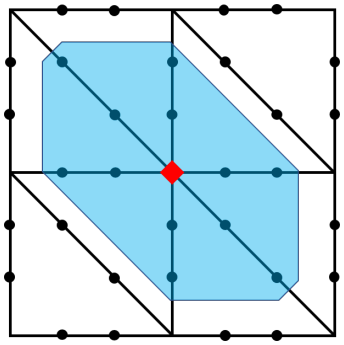


Figure 5.3: The patch of degrees of freedom (black dots inside the blue patch) around a vertex (red diamond) used in the multigrid relaxation for a BDM_1 discretization in two dimensions. Each vertex-star patch contains 12 degrees of freedom in two dimensions.

Injecting the active set

A complication arises in the representation of $\mathsf{A}_{\gamma_d} - \mathsf{D}\mathsf{C}_\mu^{-1}\mathsf{D}^\top$ on the coarser levels. By first ignoring the BM active set, we note that $\mathsf{A}_{\gamma_d} - \mathsf{D}\mathsf{C}_\mu^{-1}\mathsf{D}^\top$ can be assembled by injecting the current finite element iterates, $\boldsymbol{u}_h$ and ρ_h, to the correct level on the mesh hierarchy, assembling the submatrices A, D, $\mathsf{D}^\top$ and C_μ, applying the Dirichlet boundary conditions of the injected velocity to the relevant rows and columns of A, D and $\mathsf{D}^\top$, and subtracting the triple matrix product $\mathsf{D}\mathsf{C}_\mu^{-1}\mathsf{D}^\top$ from A. The triple matrix product is sparse and cheap to compute as C_μ is diagonal on all levels. However, we found that an accurate representation of the active set on the coarser

levels is essential for the convergence of the multigrid scheme. Hence, the difficulty lies in defining the active set on the coarser levels. One choice is to use the definitions of $\mathcal{A}$ and $\mathcal{I}$ in (5.9) defined via the injected material distribution iterate. However, we found that this choice resulted in poor iteration counts as verified in Section 5.3.1.

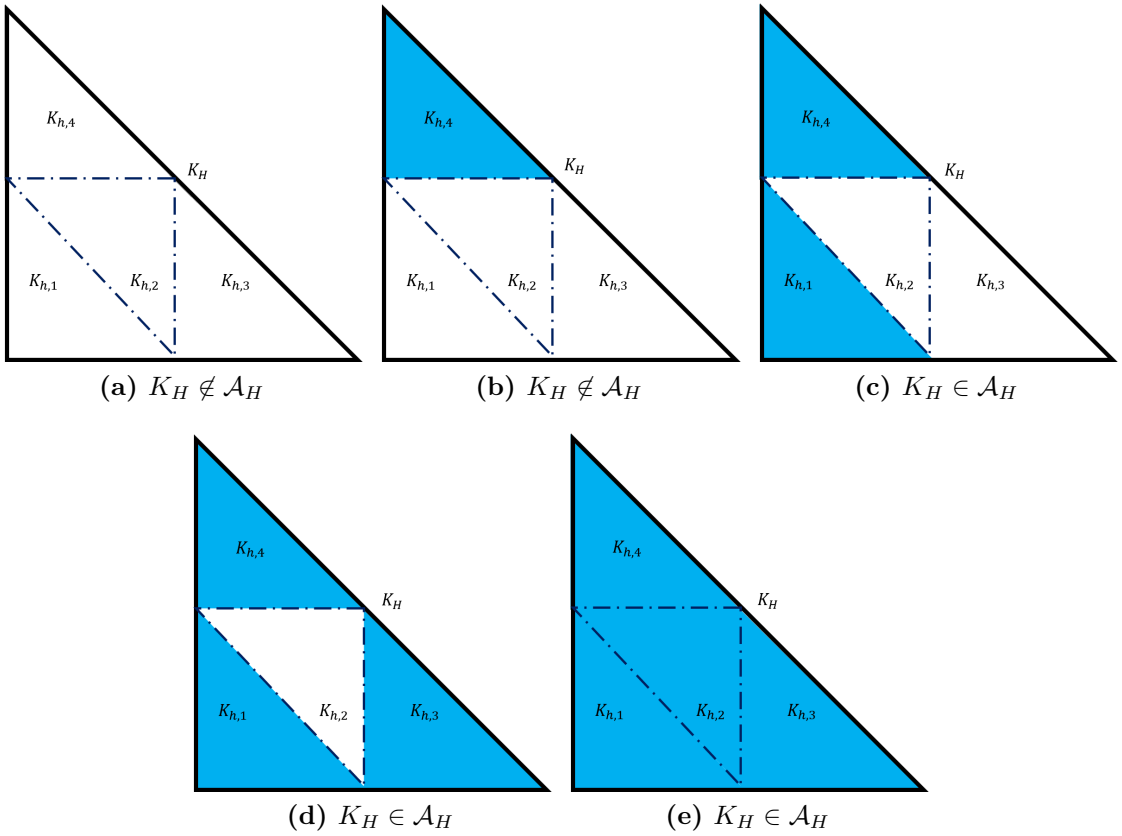
Consider a two-grid method with the fine-level triangulation $\mathcal{T}_h$, $h = H/2$, obtained by a uniform refinement of the simplices in coarse-level triangulation $\mathcal{T}_H$. As the material distribution is discretized with DG_0 elements, each degree of freedom i associated with the fine-level material distribution iterate can be associated with an element $K_h \in \mathcal{T}_h$ in the fine level and analogously with the degrees of freedom of the coarse-level material distribution iterate with elements in the coarse level. We say that a fine-level element $K_h \in \mathcal{T}_h$ is in the active set $\mathcal{A}_h$ (written as $K_h \in \mathcal{A}_h$) if the degree of freedom associated with K_h is in the active set $\mathcal{A}_h$. This definition naturally extends to the coarse-level elements and active set.

We now utilize an idea inspired by the work of Hoppe [93] and Engel and Griebel [62] to define the coarse-level active sets. A coarse-level element, $K_H \in \mathcal{T}_H$ containing the parent fine-level elements $K_{h,1}, \ldots, K_{h,s} \in \mathcal{T}_h$ is defined to be in the coarse-level active set $\mathcal{A}_H$ if

$$|\{K_{h,j} \in \mathcal{A}_h : j = 1, \ldots, s\}| \geq m, \tag{5.32}$$

where $m \in [1, s]$ and $s = 4$ in two dimensions and $s = 8$ in three dimensions. In other words, the coarse-level element is in the coarse-level active set if it contains m or more fine-level parent elements that are in the fine-level active set. By starting at the finest-level active set that is defined by (5.9), we recursively define all the active sets in mesh hierarchy via (5.32). Experiments revealed that a good choice for m is $m = s/2$, i.e. a coarse-level element is active if at least half of its parent fine-level elements are active. This choice is exemplified in Fig. 5.4 and a summary of the multigrid strategy is given in Fig. 5.5.

Remark 5.5. *The choice of* (5.32) *is more generous than utilizing the definition of the fine-level active set directly with the injected material distribution iterate. In particular, some coarse cells that are "borderline" between the active set and*



(a) $K_H \notin \mathcal{A}_H$ (b) $K_H \notin \mathcal{A}_H$ (c) $K_H \in \mathcal{A}_H$

(d) $K_H \in \mathcal{A}_H$ (e) $K_H \in \mathcal{A}_H$

Figure 5.4: Characterization of a two-dimensional coarse element based on whether half or more of its parent fine-level elements are in the fine-level active set. Blue and white fine-level elements are in the fine-level active set and inactive set, respectively.

inactive set are placed in the active set by (5.32) *but in the inactive set when defining the active set via the injected material distribution. Numerically, we see that the iteration counts suffer if the criteria for a coarse cell to be in the coarse-level active set are too strict.*

5.3 Numerical results

All examples in this chapter were implemented with the finite element software Firedrake [135]. Block preconditioning and Krylov subspace methods were implemented using Firedrake [135] and PETSc [24], and sparse LU factorizations were performed with MUMPS [17]. Vertex-star patch relaxation is implemented via the PCPATCH functionality [71] that was recently introduced to PETSc. The meshes were created in Firedrake or Gmsh [81]. The uniqueness of the pressure was enforced

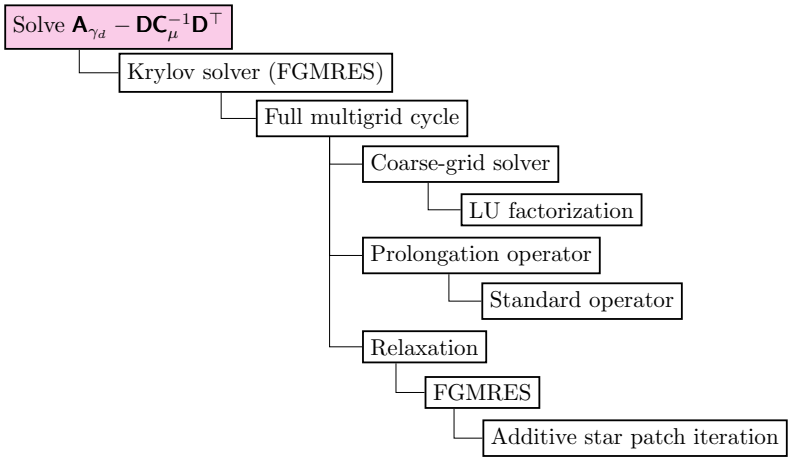


Figure 5.5: The multigrid solver strategy of Section 5.2.2 to solve $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$.

by orthogonalizing against the nullspace of constants in the Krylov method. The coarse-level correction in the multigrid scheme of Section 5.2.2 is computed via an LU factorization. The BM updates are scaled with a (damped) l^2-minimizing linesearch [43, Alg. 2] and we do not use a prediction step in any examples.

5.3.1 Double-pipe

The first example is the two-dimensional double-pipe problem as introduced in Section 4.6.1. As before, the double-pipe problem is posed on a rectangular domain $\Omega = (0, 3/2) \times (0, 1)$ with two inlets and two outlets fixed by the Dirichlet boundary condition given in (4.51). We choose a volume fraction of $\gamma = 1/3$ and the inverse permeability α is given in (2.22), with $\bar{\alpha} = 2.5 \times 10^4$ and $q = 1/10$. The problem supports two minima: a local minimum of two straight channels from each inlet to its opposite outlet, and a global minimum in the shape of a double-ended wrench. These solutions are depicted in Fig. 4.4. The two strategies we utilize for solving the linear systems are the following:

(aL1) The nested block preconditioning approach of Section 5.2.1 on (5.10) with $\gamma_d = 10^4$, and an LU factorization for the pressure mass matrix $\mathbf{M}_p$, diagonal matrix $\mathbf{C}_\mu$, and augmented momentum block $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$;

(aL2) The nested block preconditioning approach of Section 5.2.1 on (5.10) with $\gamma_d = 10^4$, an LU factorization for the pressure mass matrix $\mathbf{M}_p$ and diagonal

matrix $\mathbf{C}_\mu$, and the geometric multigrid method of Section 5.2.2 to approximate the action of the inverse of $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$. We fix the relaxation to 5 FGMRES iterations preconditioned with a vertex-star patch iteration and a full multigrid cycle is used.

We opt for a first-order Brezzi–Douglas–Marini $\mathrm{BDM}_1 \times \mathrm{DG}_0$ mixed finite element discretization for the velocity-pressure pair, with interior penalty parameter $\sigma = 10$, and a DG_0 discretization for the material distribution. For all mesh sizes, we initialize the deflated barrier method at $\mu_0 = 105$ and perform deflation immediately to find the second branch. The first and second branches converge to the straight channels and double-ended wrench solutions, respectively, as $\mu \to 0$. If $h > 10^{-2}$, for $\mu > 0$ and $\mu = 0$, each subproblem is solved to an absolute tolerance of 10^{-6} and 10^{-7}, respectively. If $h \leq 10^{-2}$, each subproblem is solved to an absolute tolerance of 10^{-5}. In Tables 5.1 and 5.2, we list the iteration counts for the strategies (aL1) and (aL2) on meshes with decreasing mesh sizes. In the (aL2) strategy, the augmented block solve is approximated to an absolute tolerance of 10^{-8} or a relative tolerance of 10^{-9}.

		(aL1)	
h	Dofs	BM	OK
0.0361	25,201	562	1055 (1.88)
0.0180	100,401	660	1227 (1.86)
0.0090	400,801	733	1283 (1.75)
0.0045	1,601,601	809	1607 (1.99)

Table 5.1: The total cumulative number of iterations to compute both minimizers of the double-pipe problem over all the subproblems with the (aL1) preconditioner. BM stands for the number of Benson–Munson iterations and OK stands for the number of outer Krylov FGMRES iterations. The numbers in brackets in the OK column are the number of average Krylov iterations per BM iteration.

We see that the Krylov iterations per BM iteration are robust to the mesh size for both preconditioning strategies. The preconditioning strategy with an LU factorization for the augmented momentum block (aL1) averages to under 2 preconditioned FGMRES iterations per BM iteration. Similarly with (aL2), where the augmented block solve is approximated with FGMRES preconditioned with a

		(aL2) 2-grid		
h	Dofs	BM	OK	IK
0.0180	100,401	660	1315 (1.99)	113,881 (12.67)
0.0090	400,801	779	1416 (1.82)	108,740 (10.47)

Table 5.2: The total cumulative number of iterations to compute both minimizers of the double-pipe problem over all the subproblems with the (aL2) preconditioner. BM stands for the number of Benson–Munson iterations, OK stands for the number of outer Krylov FGMRES iterations, and IK is the number of inner Krylov FGMRES iterations preconditioned with the geometric multigrid method of Section 5.2.2 to solve linear systems involving $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$. The numbers in brackets in the OK and IK columns are the number of average Krylov iterations per BM iteration and per $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ solve, respectively.

2-grid multigrid cycle, the outer FGMRES iterations remain under 2 preconditioned FGMRES iterations per BM iteration on average. Moreover, the inner FGMRES iterations remain under 13 iterations per augmented momentum block solve over all mesh sizes. We note that, unlike the conforming discretizations in Section 4.6, the number of BM iterations slowly increases with decreasing mesh size. This may be due to the divergence-free DG discretization or could be related to the fact that the linear systems are not being solved exactly.

We now compare our choice for the definition of the coarse-level active sets with other approaches. We consider the final continuation step in the application of the deflated barrier method to the double-pipe problem on a mesh where $h = 0.0180$. The linear systems are solved with the preconditioner (aL2) with a 2-grid multigrid cycle for the augmented momentum block. Given the two branches of solutions at $\mu = 3.81 \times 10^{-4}$, we use these as initial guesses for the nonlinear solves at $\mu = 0$. In Table 5.3, we report the cumulative number of BM iterations, outer FGMRES iterations, and inner FGMRES iterations preconditioned with the multigrid cycle for varying choices of the definition of the coarse-level active set. We consider two families of strategies. The first is defined by (5.32). In the work of Hoppe [93] and Engel and Griebel [62], a coarse-level cell was only active if all the parent cells where active ($m = 4$), whereas we found that the iteration counts were lower when only half the parent cells are required to be in the active set ($m = 2$). The other strategy is defined by the fine-level definition (5.9) combining the injected material distribution

iterate ρ_h and the restricted residual $\mathbf{f}$. We found that a direct application of this definition did not result in converging multigrid cycles; the first augmented momentum block solve reached 500 inner FGMRES iterations without converging. However, we found this could be remedied if the tolerances were loosened, i.e. we made the following modification to the definition of the coarse-level active set:

$$\mathcal{A}_c = \{i : \mathbf{z}_i \leq \mathbf{a}_i + c \text{ and } \mathbf{f}(\mathbf{z})_i > 0\} \cup \{i : \mathbf{z}_i \geq \mathbf{b}_i - c \text{ and } \mathbf{f}(\mathbf{z})_i < 0\}. \tag{5.33}$$

We report the iteration counts for this choice of definition for the coarse-level active set with varying choices of c. We conclude that the definition described in Section 5.2.2, where $m = 2$, outperforms all the other choices for the definition of the coarse-level active set.

Strategy	BM	OK	IK
(5.32), $m = 2$	12	21 (1.75)	1832 (14.54)
(5.32), $m = 4$	13	23 (1.77)	3185 (23.08)
(5.33), $c = 10^{-10}$	12	21 (1.75)	2976 (23.62)
(5.33), $c = 10^{-5}$	12	23 (1.92)	3841 (27.91)

Table 5.3: The total cumulative number of iterations to compute both solutions of the double-pipe problem at $\mu = 0$ using the solutions at $\mu = 3.81 \times 10^{-4}$ as initial guesses, with the (aL2) preconditioner. BM stands for the number of Benson–Munson iterations, OK stands for the number of outer Krylov FGMRES iterations, and IK is the number of inner Krylov FGMRES iterations preconditioned with the geometric multigrid method to solve linear systems involving $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$. The numbers in brackets in the OK and IK columns are the number of average Krylov iterations per BM iteration and per $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ solve, respectively.

Next, we test the robustness of the preconditioning strategy to higher order BDM discretizations. On the same mesh, we interpolate the two solutions at $h = 0.0180$ discretized with a first-order $\mathrm{BDM}_1 \times \mathrm{DG}_0$ discretization (100,401 degrees of freedom) to higher order $\mathrm{BDM}_k \times \mathrm{DG}_{k-1}$ discretizations for the velocity-pressure pairs, and reinitialize the deflated barrier method at $\mu_0 = 0$. The first-order optimality conditions are solved to an absolute residual tolerance of 10^{-6}. In Table 5.4 we give iteration counts for $k = 2, 3$, and 4 using the preconditioning strategy (aL2) with a 2-grid and 3-grid multigrid scheme. The augmented block solve tolerances remain the same. We see that the number of outer FGMRES iterations per BM iteration

averages to below 2 for any order of the discretization. The 2-grid multigrid scheme is effective in solving the augmented momentum block, regardless of the order of the discretization, averaging between 10 and 12 iterations for any order. The iteration counts increase for the 3-grid cycle due to the quality of the coarse-grid representation, but they are also robust to the polynomial order, averaging to less than 25 iterations per augmented momentum block solve. We note that high-order

		(aL2) 2-grid			(aL2) 3-grid		
k	Dofs	BM	OK	IK	BM	OK	IK
2	230,601	10	16 (1.60)	1112 (11.58)	10	16 (1.60)	2343 (24.41)
3	420,801	11	17 (1.55)	1095 (10.74)	10	18 (1.80)	2620 (24.26)
4	671,001	10	19 (1.90)	1135 (10.91)	10	19 (1.90)	2607 (22.87)

Table 5.4: The total cumulative iteration counts for both solutions when using the (aL2) preconditioner for polynomial order refinement. The initial guesses are the two solutions approximated by the first-order discretization ($k = 1$) on the same mesh. The deflated barrier method is initialized at $\mu_0 = 0$. The columns BM, OK, and IK, as well as the numbers in brackets, have the same meaning as in Table 5.2.

discretizations for the velocity-pressure pair are not practical since our solver relies upon a DG_0 discretization for the material distribution.

Comparison with the Cahouet–Chabard preconditioner

We now compare our augmented Lagrangian preconditioner with the Cahouet–Charbard preconditioning strategy utilized by Borrvall and Petersson in their original work [36]. Borrvall and Petersson used a nested approach via the first-order MMA algorithm [162] to find minimizers of their problems. MMA relies on sensitivity analysis by solving forward/adjoint problems. Hence, given a material distribution iterate, the most computationally expensive step in MMA is the repeated forward solve of (5.5) and (5.6) to compute the velocity and pressure iterates. Borrvall and Petersson used the Cahouet–Chabard preconditioner to precondition the momentum-pressure block. In our framework, the natural application for the Cahouet–Chabard preconditioner is for preconditioning S_1. We emphasize that the Cahouet–Chabard preconditioner was designed to handle the momentum block A and not A augmented with $-\mathsf{D}\mathsf{C}_\mu\mathsf{D}^\top$ as found in the top left block of S_1. This is reflected in our numerical results.

The Cahouet–Chabard preconditioner requires an H^1-conforming discretization of the pressure since the preconditioner involves assembling a matrix induced by a weak form that features $\nabla\psi_i$, where ψ_i, $i = 1, \ldots, n_p$, are the pressure basis functions. Hence, we test the preconditioner with a DG_0 discretization for the material distribution and a Taylor–Hood $(\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ discretization for the velocity-pressure pair. This strategy follows the same steps as our strategy up to and including block preconditioning with respect to the augmented momentum-pressure block step in Fig. 5.2, except without the addition of the augmented Lagrangian term. Hence, the solve is reduced to an outer FGMRES solver, taking the reciprocal of the real number $\mathbf{S}_0$, inverting the diagonal matrix $\mathbf{C}_\mu$, and applying the action of the inverse of the 2×2 block matrix $\mathbf{S}_1$.

Here, we apply an inner FGMRES method to approximate the inverse action of $\mathbf{S}_1$. The FGMRES method is preconditioned by $\hat{\mathbf{S}}_1$ as defined in (5.20). We apply block preconditioning to approximate the action of $\hat{\mathbf{S}}_1^{-1}$, reducing the solve to a preconditioned outer FGMRES solver, taking the reciprocal of the real number $\mathbf{S}_0$, inverting the diagonal matrix $\mathbf{C}_\mu$, an inner preconditioned FGMRES solver for $\mathbf{S}_1$, factorizing the momentum block $\mathbf{A}$, and approximating the inverse action of the innermost Schur complement $\hat{\mathbf{S}}_2 := -\mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$.

As we no longer include an augmented Lagrangian term, solving linear systems involving the momentum block $\mathbf{A}$ is much easier due to the lack of the semi-definite term $\gamma_d\mathbf{B}^\top\mathbf{M}_p^{-1}\mathbf{B}$. However, the difficulty now shifts to approximating the Schur complement $\hat{\mathbf{S}}_2$. Here, $\hat{\mathbf{S}}_2$ is dense and should not be assembled. Hence, the action of $\hat{\mathbf{S}}_2^{-1}$ is approximated by a (preconditioned) conjugate gradient algorithm (CG). The Cahouet–Chabard strategy is used to precondition this CG solver, where the Cahouet–Chabard preconditioner is

$$\mathbf{K}_{p,\rho}^{-1} + \nu\mathbf{M}_p^{-1}, \quad [\mathbf{K}_{p,\rho}]_{ij} := \int_\Omega \alpha(\rho_h)^{-1}\nabla\psi_i \cdot \nabla\psi_j \,\mathrm{d}x, \tag{5.34}$$

where ψ_i, $i = 1, \ldots, n_p$, are the basis functions of the pressure discretization. Since the reciprocal of $\alpha(\cdot)$ appears in $\mathbf{K}_{p,\rho}$, we must have a positive lower bound for

α. Hence, in this example we choose

$$\alpha(\rho) = \bar{\alpha} + (\underline{\alpha} - \bar{\alpha})\frac{\rho(1+q)}{\rho+q}, \tag{5.35}$$

with $\bar{\alpha} = 2.5 \times 10^4$, $q = 1/10$, and $\underline{\alpha} = 10^{-5}$. The action of $\hat{\mathbf{S}}_2$ in the CG iteration requires the action of a momentum block inverse $\mathbf{A}^{-1}$ which is already computed as part of the strategy. In summary, the strategy is the following:

1. An outer FGMRES solver;
2. Apply the reciprocal of $\mathbf{S}_0 \in \mathbb{R}$;
3. Invert the diagonal matrix $\mathbf{C}_\mu$;
4. An inner FGMRES solver for $\mathbf{S}_1$;
5. Factorize and solve the momentum block $\mathbf{A}$;
6. A CG solver for the innermost Schur complement $\hat{\mathbf{S}}_2$;
7. Factorize and solve $\mathbf{K}_{p,\rho}$;
8. Factorize and solve $\mathbf{M}_p$ (only required once and then cached).

In our tests we use an LU factorization for $\mathbf{K}_{p,\rho}$, $\mathbf{M}_p$, and $\mathbf{A}$, and CG preconditioned with (5.34) to approximate the inverse action of $\hat{\mathbf{S}}_2$ to an absolute tolerance of 10^{-8} or relative tolerance of 10^{-5}. The actions of $\mathbf{A}^{-1}$ and $\hat{\mathbf{S}}_2^{-1}$ are then used to precondition the inner FGMRES method that approximates the inverse of the outermost Schur complement $\mathbf{S}_1^{-1}$ to a relative tolerance of 10^{-5}. The outer FGMRES solver was terminated once an absolute tolerance of 10^{-8} was reached.

In Table 5.5, we apply this strategy to the deflated barrier method applied in order to compute the solution on the first branch of the double-pipe problem at the subproblem where $\mu = 105$. We compare it with the augmented Lagrangian approach of (aL1). The runtime measurements are as recorded using all four cores of a computer with an i5-7500T CPU running at 2.7 GHz. We note that, for the Cahouet–Chabard strategy, the number of outermost FGMRES iterations per BM iteration is 1, whereas for the augmented Lagrangian strategy it averages to 2 outermost FGMRES iterations per BM iteration. However, in the Cahouet–Chabard

strategy, we require $\mathcal{O}(10)$ inner FGMRES iteration per outer FGMRES iteration to approximate the inverse action of $\mathbf{S}_1$ and a further 7–9 inner preconditioned CG iterations per approximate $\hat{\mathbf{S}}_2$ solve. These inner iterations are not required in the augmented Lagrangian approach. Hence, each outer FGMRES iteration is significantly more expensive and our augmented Lagrangian preconditioner is faster and more effective for the problems discussed in this chapter. We note the BM iterations differ due to the different discretizations (Taylor–Hood vs. BDM) of the problem in the two strategies.

	Cahouet–Chabard					(aL1)		
h	BM	OK	IK	CG	Time (s)	BM	OK	Time (s)
0.0361	8	8	569	4272	73.75	11	23	22.08
0.0180	7	7	487	3459	307.02	12	23	78.61

Table 5.5: The cumulative total number of iterations and runtime for deflated barrier method at $\mu = 105$ to find the first branch of the double-pipe problem utilizing the Cahouet–Chabard preconditioning strategy vs. the augmented Lagrangian (aL1) strategy. BM stands for the number of Benson–Munson iterations, OK stands for the number of outer preconditioned FGMRES iterations to solve linear systems involving the full matrix (5.10), IK stands for the number of inner preconditioned FGMRES iterations to solve linear systems involving $\mathbf{S}_1$, and CG stands for the number of preconditioned CG iterations to approximate the inverse of $\hat{\mathbf{S}}_2$.

Convergence results

Akin to Section 4.6.1, we investigate the convergence of the finite element solutions to the double-pipe problem. If we assume that the traces of $\nabla \boldsymbol{u}$ are well-defined on the faces of each element in both solutions, then the consistency result in Proposition 3.7 holds. Since we are using a BDM discretization for the velocity-pressure pair and a DG_0 discretization for the material distribution, all the conditions of Theorem 3.2 are satisfied and hence there exists a sequence of solutions to (FOC1-DG_h)–(FOC3a-DG_h) that converges strongly to the straight channels solution and a different sequence of solutions that converges to the double-ended wrench. The existence of these sequences is numerically verified in Fig. 5.6 for a $\mathrm{DG}_0 \times \mathrm{BDM}_1 \times \mathrm{DG}_0$ discretization for $(\rho_h, \boldsymbol{u}_h, p_h)$. The linear systems are preconditioned using the augmented Lagrangian strategy (aL1). As for the conforming finite element

method case, the two solutions are not known analytically. Hence, the errors are measured with respect to the most heavily-refined finite element solution where $h = 1.41 \times 10^{-3}$ resulting in 16,389,121 degrees of freedom for the first-order BDM discretization. To aid the convergence to solutions on the same sequence, we first compute the solutions on the coarsest mesh, uniformly refine the mesh, and successively interpolate the solutions onto the finer mesh as initial guesses for the deflated barrier method initialized at $\mu_0 = 0$.

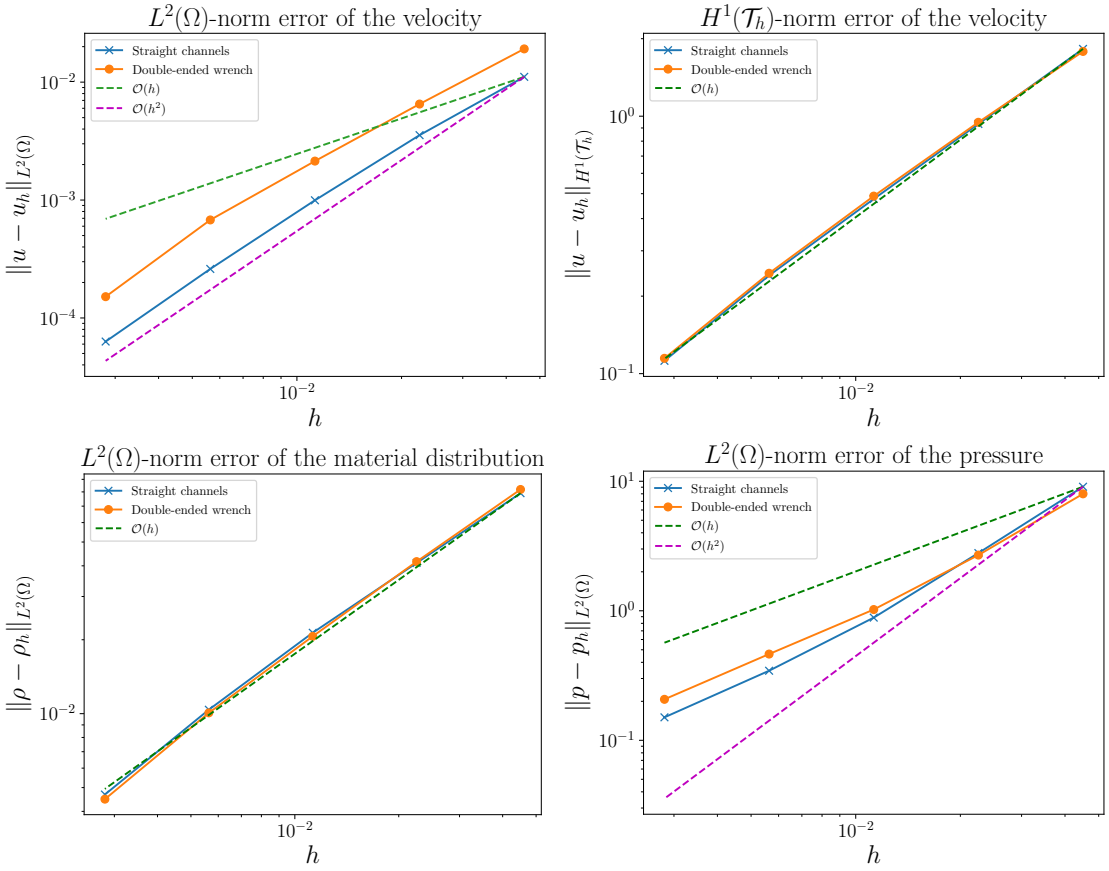


Figure 5.6: The convergence of $\boldsymbol{u}_h$, ρ_h, and p_h for the double-pipe problem for both the straight channels solution and double-ended wrench solution on a sequence of uniformly refined meshes with a $\mathrm{DG}_0 \times \mathrm{BDM}_1 \times \mathrm{DG}_0$ discretization for $(\rho_h, \boldsymbol{u}_h, p_h)$.

A key property of the BDM discretization that was also utilized in our preconditioning strategy is that the incompressibility constraint is satisfied pointwise. To test the difference between the BDM discretization and a finite element method that is not divergence-free, we report the values of $\|\mathrm{div}(\boldsymbol{u}_h)\|_{L^2(\Omega)}$ in Table 5.6 for the BDM discretization alongside the equivalent solutions computed with a Taylor–Hood

$(\mathrm{CG}_2)^2 \times \mathrm{CG}_1$ discretization for the velocity-pressure pair and a DG_0 discretization for the material distribution on the same meshes. Even on coarse meshes, the L^2-norm of the divergence of the velocity in the BDM discretization is small with values in the range of $10^{-6} \sim 10^{-8}$ for both minimizers. By contrast, the pointwise violation of the incompressibility constraint for the Taylor–Hood discretization manifests as relatively large values of $\|\mathrm{div}(\boldsymbol{u}_h)\|_{L^2(\Omega)}$. Even on the finest mesh where $h = 2.82 \times 10^{-3}$, which results in 4,512,004 degrees of freedom, the L^2-norm is still $\mathcal{O}(10^{-2})$, 5 orders of magnitude larger than the equivalent BDM discretization.

	Straight channels		Double-ended wrench	
h	BDM	Taylor–Hood	BDM	Taylor–Hood
4.51×10^{-2}	7.16×10^{-7}	2.10×10^{-1}	3.72×10^{-6}	3.12×10^{-1}
2.25×10^{-2}	8.31×10^{-8}	1.03×10^{-1}	3.06×10^{-8}	1.30×10^{-1}
1.13×10^{-2}	2.56×10^{-8}	6.21×10^{-2}	1.27×10^{-8}	6.61×10^{-2}
5.63×10^{-3}	7.00×10^{-8}	3.28×10^{-2}	2.52×10^{-7}	3.34×10^{-2}
2.82×10^{-3}	8.01×10^{-8}	1.72×10^{-2}	4.74×10^{-7}	1.72×10^{-2}

Table 5.6: Reported values for $\|\mathrm{div}(\boldsymbol{u}_h)\|_{L^2(\Omega)}$ in a BDM and Taylor–Hood discretization for the double-pipe problem, as measured on five meshes, in a uniformly refined mesh hierarchy.

5.3.2 3D cross-channel

The first three-dimensional example we consider is the cross-channel problem as found in Sá et al. [141, Sec. 7.5]. The domain is the unit cube, $\Omega = (0,1)^3$, with two circular inlets and two circular outlets that are arranged in a cross pattern as visualized in Fig. 5.7. The volume fraction is given by $\gamma = 1/10$ and we use (2.22) as our choice of α, with $\bar{\alpha} = 2.5 \times 10^4$ and $q = 1/10$. The Dirichlet boundary is given by

$$\boldsymbol{g}(x,y,z) = \Big(1 - 12\pi((y-a)^2 + (z-b)^2), 0, 0\Big)^\top, \tag{5.36}$$

if $12\pi((y-a)^2 + (z-b)^2) \leq 1$ and $x = 0$ with $a = 1/2$, $b \in \{1/4, 3/4\}$ or $x = 1$ with $a \in \{1/4, 3/4\}$, $b = 1/2$, and $\boldsymbol{g}(x,y,z) = (0,0,0)^\top$ elsewhere on $\partial\Omega$.

We apply the same first-order BDM discretization, with interior penalty penalization parameter $\sigma = 10$, and run the deflated barrier method twice. The first pass is on a $20 \times 20 \times 20$ mesh resulting in 391,201 degrees of freedom. The

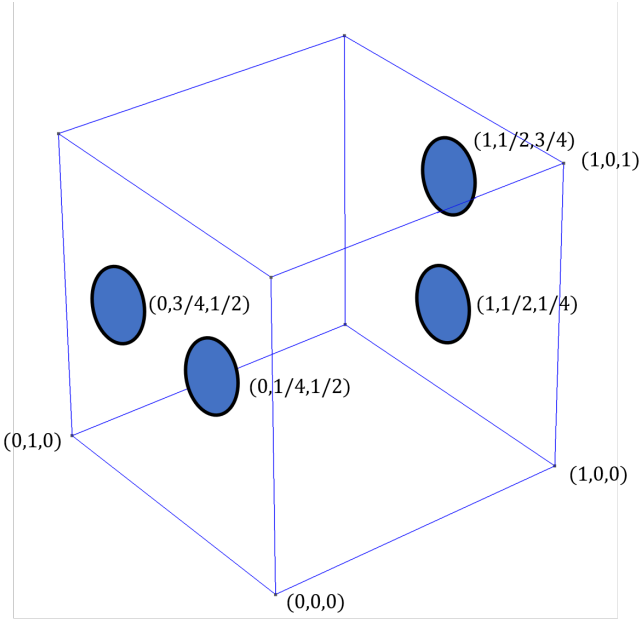


Figure 5.7: Setup of the 3D cross-channel problem. This problem features a unit cube domain with two inlets and two outlets arranged in a cross pattern.

augmented Lagrangian parameter is chosen to be $\gamma_d = 10^6$. Due to the nested block preconditioning, the action of the inverse of the augmented momentum block $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ must be applied six times per outer FGMRES iteration. On this relatively coarse mesh, it is cheaper to factorize $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ with MUMPS at the start of each BM iteration and reuse the factorization, rather than iteratively solve $\mathbf{A}_{\gamma_d} - \mathbf{D}\mathbf{C}_\mu^{-1}\mathbf{D}^\top$ with multigrid each time an inverse action is required, i.e. we use the augmented Lagrangian preconditioner (aL1) for the linear systems. We initialize the barrier parameter at $\mu_0 = 100$. A second branch of solutions is found at $\mu = 38.74$ and a third branch at $\mu = 34.87$.

These coarse-mesh solutions are interpolated onto a finer $40 \times 40 \times 40$ mesh resulting in 3,100,801 degrees of freedom. The deflated barrier method is then reinitialized at $\mu_0 = 10^{-6}$; we found that the BM solver often diverges if initialized at $\mu_0 = 0$. On this finer mesh, we again apply the augmented Lagrangian preconditioner ($\gamma_d = 10^5$). Now a direct solve of the augmented momentum block is prohibitive and we switch to the (aL2) strategy where each approximate inverse of the augmented momentum block is solved to an absolute or relative tolerance of 10^{-8} or 10^{-9}, respectively, with the (2-grid) multigrid scheme of Section 5.2.2. For the relaxation

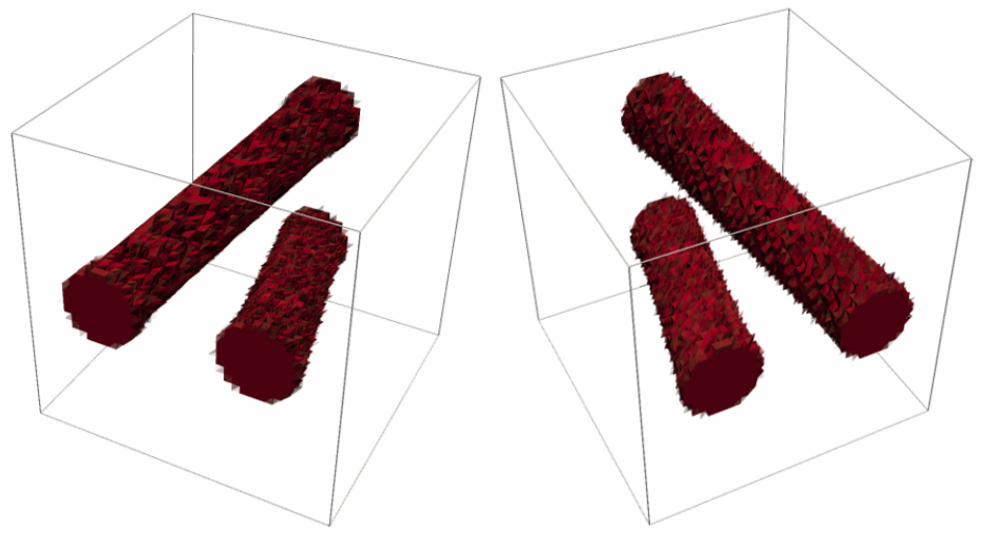
on the fine level, we use 5 FGMRES iterations preconditioned with the vertex-star patch relaxation.

The resulting three solutions are shown in Fig. 5.8. Two of these are symmetric straight channel solutions where the inlets swap which outlet they exit from. Their symmetry results in similar costs. A third global minimizer comes in the form of a merged channel solution; the two channels briefly merge in the middle of the box domain before splitting to exit via the two outlets.

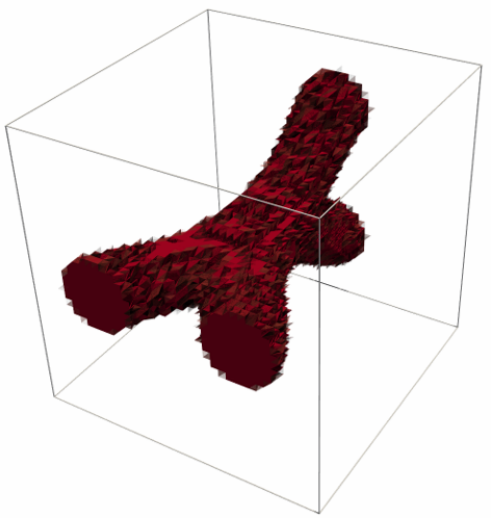
The iteration counts for the initial search on a $20 \times 20 \times 20$ mesh and the refinement on a $40 \times 40 \times 40$ mesh are given in Table 5.7. We see that our preconditioner is effective in both cases. When using a direct solve of the augmented momentum block, we average slightly more than one outer FGMRES iterations per BM iteration. Similarly, the tolerances for the augmented momentum block solve on the fine mesh are strict enough so that the outer FGMRES iterations average between 2.24 and 2.55. Moreover, each augmented momentum block solve requires an average in the range of 15.77–16.77 multigrid preconditioned FGMRES iterations to reach the prescribed tolerances.

	Coarse mesh, $\gamma_d = 10^6$		Fine mesh, $\gamma_d = 10^5$		
Branch	BM	OK	BM	OK	IK
0	828	829 (1.00)	49	110 (2.24)	10,405 (15.77)
1	444	445 (1.00)	52	121 (2.33)	12,176 (16.77)
2	409	422 (1.03)	53	135 (2.55)	13,085 (16.15)

Table 5.7: Cumulative number of BM iterations, outer FGMRES iterations (OK), and for the fine mesh, inner FGMRES iterations preconditioned with the multigrid scheme of Section 5.2.2 (IK) for the 3D cross-channel problem. The bracketed numbers in the OK and IK columns are the average number of outer FGMRES iterations per BM iteration and average number of inner FGMRES iterations per augmented momentum block solve, respectively. The barrier parameter is initialized at $\mu_0 = 100$ on the coarse mesh and $\mu_0 = 10^{-6}$ on the fine mesh.



(a) Branch 0 **(b)** Branch 1



(c) Branch 2

Figure 5.8: The material distribution of the solutions discovered by the deflated barrier method to the 3D cross-channel optimization problem discretized with 3,100,801 degrees of freedom. The power dissipation values are $J_h = 14.51$, 14.62 and 13.08 for branches 0, 1, and 2, respectively.

5.3.3 3D five-holes quadruple-pipe

In Section 4.6.5, we saw that introducing holes in a rectangular domain caused a significant increase in the number of solutions. We now extend this idea to three dimensions and introduce the generalization of the fives-holes double-pipe problem. This problem features a box domain $\Omega = (0, 3/2) \times (0, 1) \times (0, 1)$ with five internal

holes in the shape of cubes, of edge length $1/10$, with centres at $(3/4, 1/4, 1/4)$, $(3/4, 1/4, 3/4)$, $(3/4, 3/4, 1/4)$, $(3/4, 3/4, 3/4)$, and $(3/4, 1/2, 1/2)$. There are four inlets and four outlets. The circular inlets of radius $1/\sqrt{12\pi}$ are positioned on the face where $x = 0$ with the centres $(y, z) = (1/4, 1/4)$, $(1/4, 3/4)$, $(3/4, 1/4)$, and $(3/4, 3/4)$. The circular outlets of the same radius are positioned on the face where $x = 3/2$ with the same centres. The domain setup is depicted in Fig. 5.9. We impose a parabolic Dirichlet boundary condition on the inlets and outlets and a zero Dirichlet boundary condition elsewhere on the boundary (including the boundary of the five internal holes), i.e. the Dirichlet boundary condition is given by:

$$\boldsymbol{g}(x, y, z) = \left(1 - 12\pi((y-a)^2 + (z-b)^2), 0, 0\right)^\top, \tag{5.37}$$

if $12\pi((y-a)^2 + (z-b)^2) \leq 1$, where $a, b \in \{1/4, 3/4\}$, $x = 0$ or $3/2$ and

$$\boldsymbol{g}(x, y, z) = (0, 0, 0)^\top, \tag{5.38}$$

elsewhere on $\partial\Omega$, including the boundaries of the five internal holes. We choose a volume fraction of $\gamma = 1/5$ and the inverse permeability term α is given by (2.22), with $\bar{\alpha} = 2.5 \times 10^4$ and $q = 1/10$.

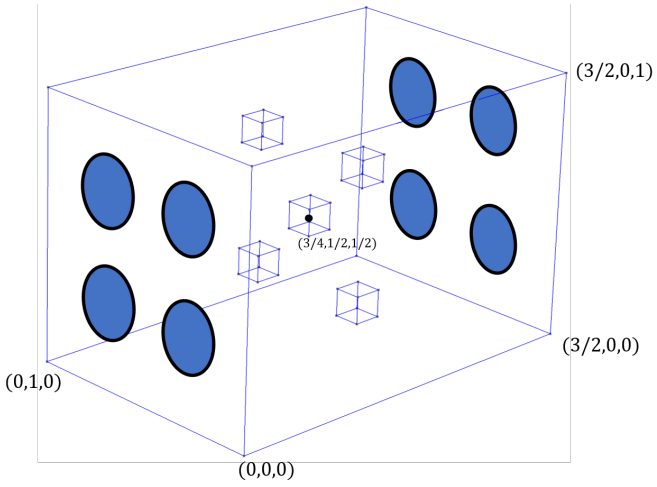


Figure 5.9: Setup of the 3D five-holes quadruple-pipe problem. This problem features 4 inlets and 4 outlets. The domain is a box, $\Omega = (0, 3/2) \times (0, 1) \times (0, 1)$, with five internal holes in the shape of cubes with edge length $1/10$ that are centred at $(3/4, 1/4, 1/4)$, $(3/4, 1/4, 3/4)$, $(3/4, 3/4, 1/4)$, $(3/4, 3/4, 3/4)$, and $(3/4, 1/2, 1/2)$.

We choose a first-order BDM discretization for the velocity-pressure pair, with interior penalty penalization parameter $\sigma = 10^3$, and run the deflated barrier method

twice. The larger choice for σ is required to sufficiently enforce the boundary conditions in the tangential directions. The first run is on a (relatively) coarse mesh with 30,848 elements which results in 256,745 degrees of freedom. The barrier parameter is initialized at $\mu_0 = 200$ and we use the augmented Lagrangian preconditioner (aL1) for the linear systems, with an augmented Lagrangian parameter value of $\gamma_d = 10^5$.

In total we find 11 solutions. Branches 1 and 2 are found at $\mu = 53.81$. Branches 3, 4, 5, and 6 are found at $\mu = 11.39$. Branch 7 is found at $\mu = 10.25$. Branch 8 is found at $\mu = 9.23$. Branch 9 is found at $\mu = 6.73$ and branch 10 is found at $\mu = 6.05$.

We uniformly refine the mesh and interpolate the coarse-level solutions onto the finer mesh which results in 2,014,113 degrees of freedom. We reinitialize the deflated barrier method at $\mu_0 = 10^{-6}$, using the coarse-level solutions as initial guesses, and apply the augmented Lagrangian (2-grid) multigrid preconditioner (aL2) with 5 FGMRES iterations for the relaxation of the fine level and $\gamma_d = 10^5$. The augmented momentum block solve was terminated after 20 iterations. The resulting iteration counts are given in Table 5.8 and the resulting fine mesh solutions are shown in Fig. 5.10 and Fig. 5.11. In Fig. 5.12, we plot the crinkled cross-sections of the discovered solutions at $x = 3/4$. As expected, the five holes obstruct the channels and prevent a large channel passing through the centre. The best solutions found are branches 0, 1, and 10 where the channels form one large channel and either move to the left, upwards or downwards to avoid the middle internal hole. As in the five-holes double-pipe example, there are remaining solutions that we have not yet computed as there are missing symmetrical solutions.

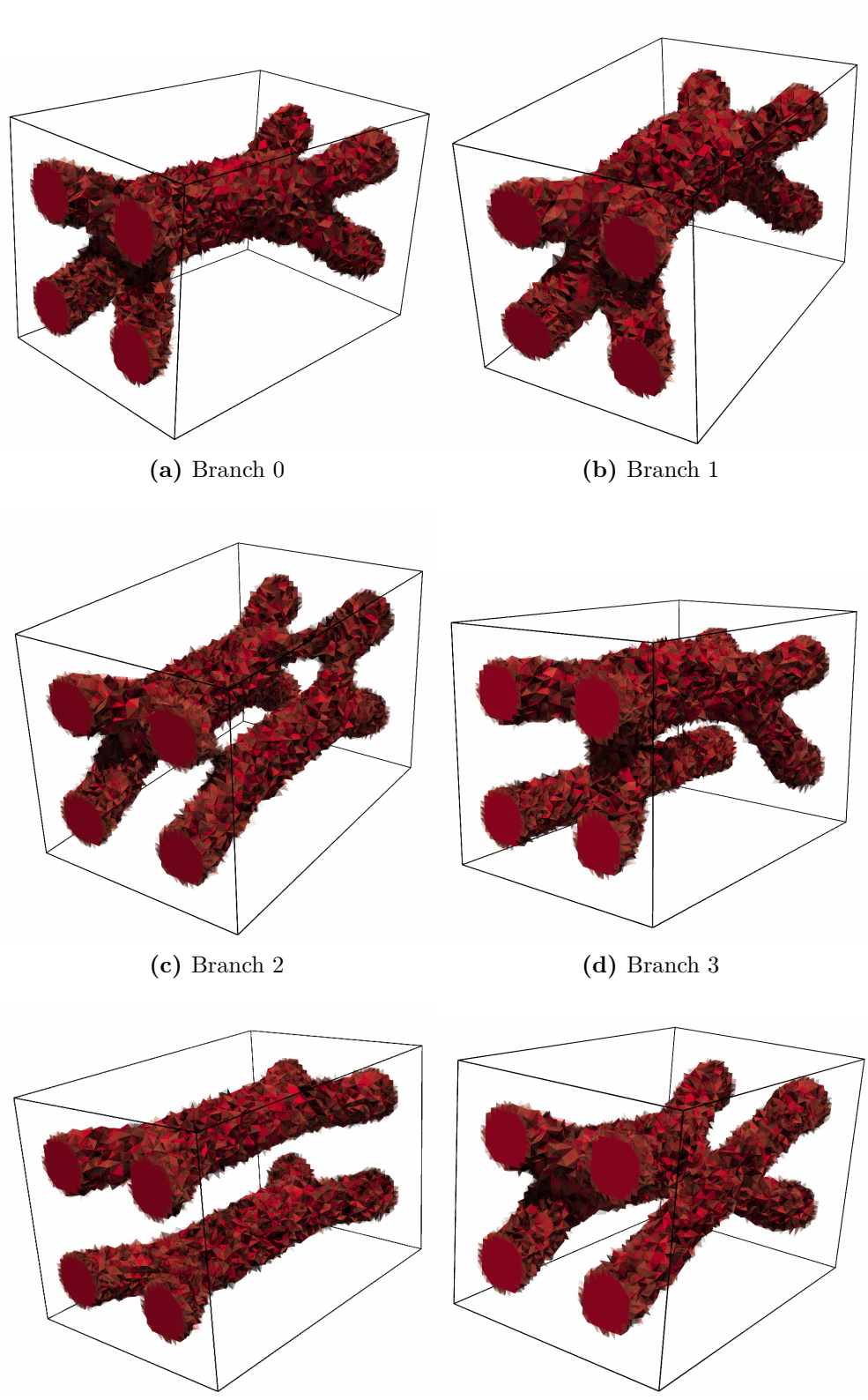
5.4 Code availability

The deflated barrier method algorithm, as used in all the numerical examples in this chapter, has been implemented in a Python library called fir3dab using Firedrake [135] as the finite element backend. The implementation required changes to Firedrake and PETSc [24] in order to facilitate applying the BM active-set strategy [31], together with preconditioners, to the topology optimization problems. The fir3dab library can be found at https://github.com/ioannisPApapadopoulos/

	Coarse mesh		Fine mesh		
Branch	BM	OK	BM	OK	IK
0	438	438 (1)	22	52 (2.36)	6240 (20)
1	338	338 (1)	26	72 (2.77)	8640 (20)
2	396	396 (1)	27	107 (3.96)	12,840 (20)
3	254	254 (1)	44	127 (2.89)	15,240 (20)
4	236	236 (1)	29	81 (2.79)	9720 (20)
5	222	222 (1)	45	113 (2.51)	13,560 (20)
6	223	223 (1)	31	111 (3.58)	13,320 (20)
7	221	221 (1)	51	185 (3.63)	22,200 (20)
8	224	224 (1)	52	164 (3.15)	19,680 (20)
9	227	227 (1)	33	117 (3.55)	14,040 (20)
10	183	183 (1)	48	135 (2.81)	16,200 (20)

Table 5.8: Cumulative number of BM iterations, outer FGMRES iterations (OK), and for the fine mesh, inner FGMRES iterations preconditioned with the multigrid scheme of Section 5.2.2 (IK) for the 3D five-holes quadruple-pipe problem. The bracketed numbers in the OK and IK columns are the average number of outer FGMRES iterations per BM iteration and average number of inner FGMRES iterations per augmented momentum block solve, respectively. The barrier parameter is initialized at $\mu_0 = 200$ on the coarse mesh and $\mu_0 = 10^{-6}$ on the fine mesh.

`fir3dab/`. For reproducibility, the code used to run these examples has been archived on Zenodo [154, 156].

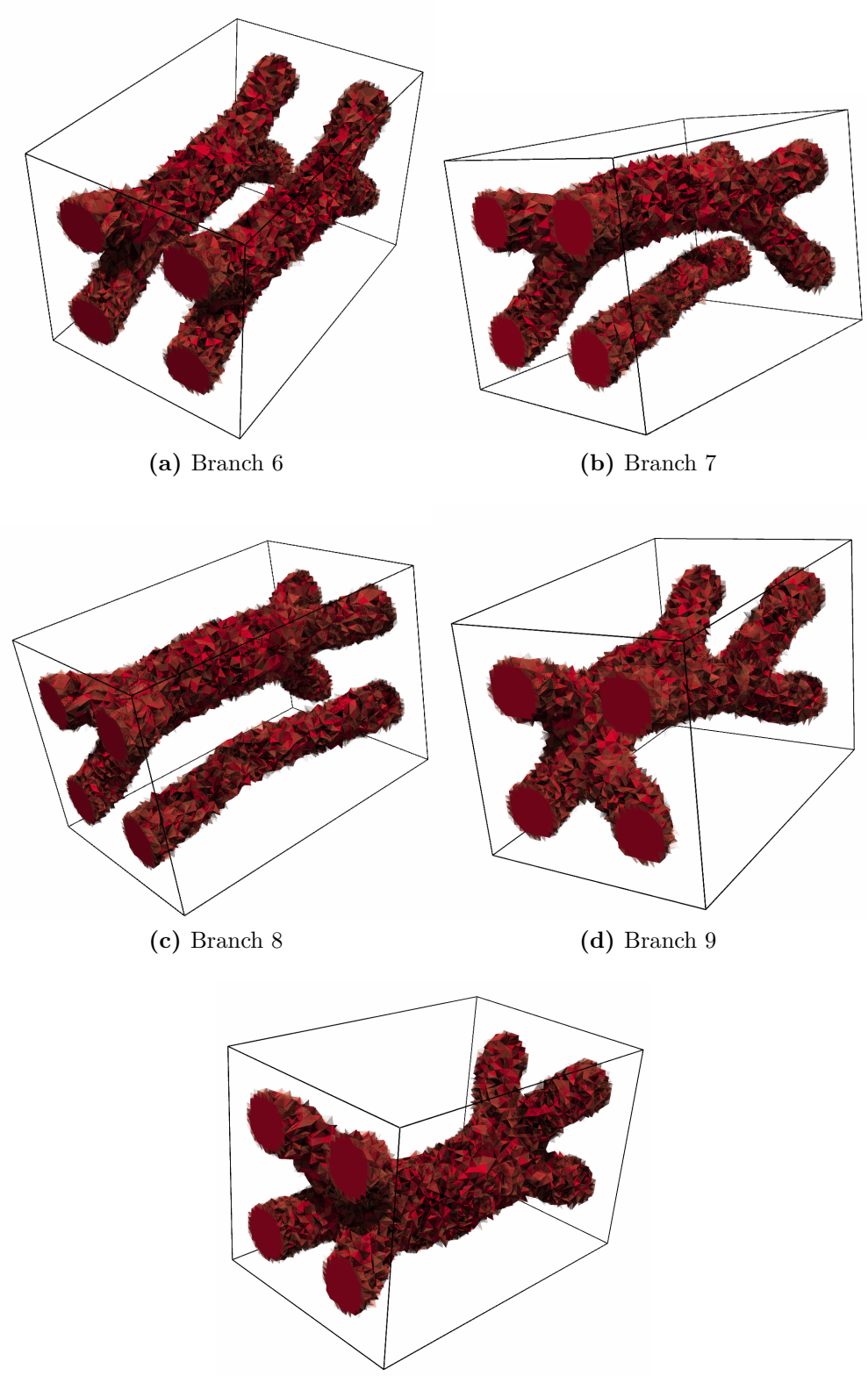


(a) Branch 0 **(b)** Branch 1

(c) Branch 2 **(d)** Branch 3

(e) Branch 4 **(f)** Branch 5

Figure 5.10: The material distribution of the first six solutions discovered by the deflated barrier method to the 3D fives-holes quadruple-pipe optimization problem discretized with 2,014,113 degrees of freedom. The resulting power dissipation values for branches 0–5 are $J_h = 55.03$, 54.73, 63.78, 62.22, 59.56, and 63.31, respectively.

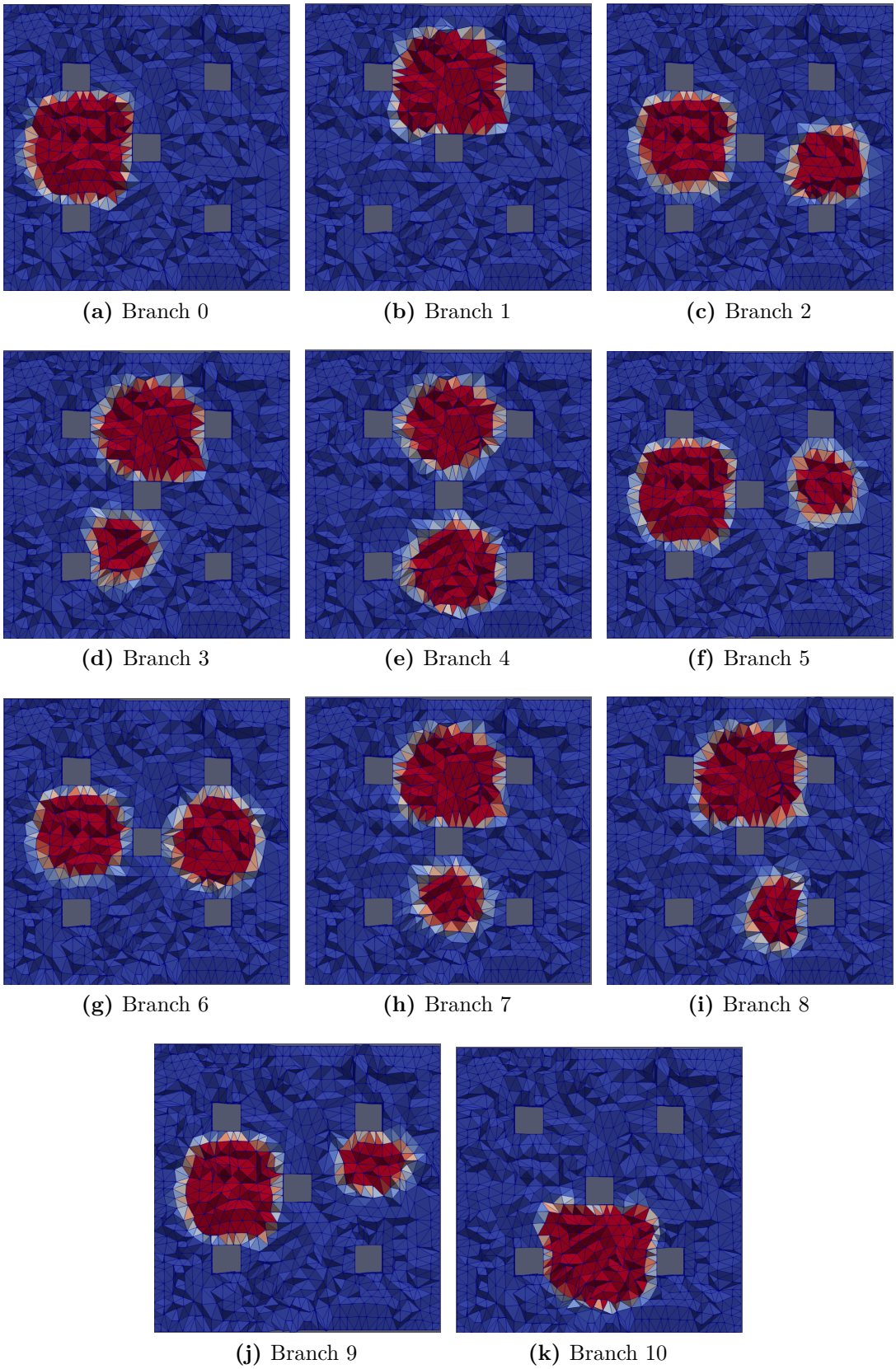


(a) Branch 6 **(b)** Branch 7

(c) Branch 8 **(d)** Branch 9

(e) Branch 10

Figure 5.11: The material distribution of the final five solutions discovered by the deflated barrier method to the 3D fives-holes quadruple-pipe optimization problem discretized with 2,014,113 degrees of freedom. The resulting power dissipation values for branches 6–10 are $J_h = 59.55$, 63.35, 62.52, 62.37, and 55.38, respectively.



(a) Branch 0 **(b)** Branch 1 **(c)** Branch 2

(d) Branch 3 **(e)** Branch 4 **(f)** Branch 5

(g) Branch 6 **(h)** Branch 7 **(i)** Branch 8

(j) Branch 9 **(k)** Branch 10

Figure 5.12: The crinkled cross sections at $x = 3/4$ for the discovered solutions of the 3D five-holes quadruple-pipe. The grey regions are part of the five cuboid holes in the box domain. The material distribution has a value of one in the red regions and zero in the blue regions, with intermediate values for the intermediate coloured regions.

6
Conclusions and outlook

In this thesis we analyzed the nonconvexity of topology optimization problems, with a particular emphasis on the Borrvall–Petersson model for the topology optimization of fluid flow. We developed a framework for the analysis of finite element discretizations of all the isolated analytical minimizers of the problem and constructed a solver that can systemically compute these multiple minimizers.

6.1 Analysis

The first part of this thesis was concerned with deriving analytical and numerical results concerning isolated minimizers of the Borrvall–Petersson model for the topology optimization of the power dissipation of fluid flow. In Chapter 2 we showed that isolated minimizers $(\boldsymbol{u}, \rho, p)$ of the Borrvall–Petersson problem satisfy the following properties:

- The volume constraint on the material distribution is binding;
- The support of the material distribution function is contained within the support of the velocity function;
- The isolated minimizers satisfy first-order optimality conditions consisting of two equations and a variational inequality;
- The volume constraint on the test functions in the variational inequality can be relaxed by introducing another real-valued variable;

- If the domain is convex and the data is suitably regular, then the velocity lives in $H^2(\Omega)^d$ and the pressure lives in $H^1(\Omega)$;
- (Main result) If the inverse permeability term satisfies a strong convexity assumption and the problem features a homogeneous Dirichlet boundary condition on the velocity, then the material distribution is weakly differentiable inside any compact subset of the support of the velocity.

The first five results are expected properties of the solutions. In particular, the first-order optimality conditions can be found in the literature without proof. The additional regularity of the material distribution is surprising but was likely anticipated by Borrvall and Petersson in their initial derivation of the model. Borrvall and Petersson remark that if the inverse permeability term $\alpha(\cdot)$ is linear (and hence does not satisfy the strong convexity assumption), then the range of the material distribution is necessarily $\{0, 1\}$. Given that the material distribution is not the zero function and there is a volume constraint, this means there must be sharp jumps. This is ideal for the interpretation of the solutions, but is difficult to deal with numerically. Sharp interfaces are difficult to resolve and the numerical method will often fail to converge if a linear inverse permeability term is used. On the other hand, if a strongly convex inverse permeability term is used, then the interfaces become regular and the numerical scheme converges with more ease. The standard choice of inverse permeability term (2.22) is controlled via a "greyness" parameter $q > 0$. As $q \to \infty$ the inverse permeability tends to a linear function. This fact is reflected in the proof of Theorem 2.12 as the regularity constants degrade as $q \to \infty$. A potential future objective is to extend the material distribution regularity result to the case of inhomogeneous Dirichlet boundary conditions on the velocity. Numerical evidence suggests that the support of the material distribution is compactly contained in the support of the velocity except at the boundaries where the velocity boundary condition is nonzero. An argument to extend the regularity result will need to resolve how to extend the functions outside the domain.

In Chapter 3 we considered finite element discretizations of the Borrvall–Petersson problem. The two families we considered are a conforming finite element

method, where the finite element spaces for the velocity, material distribution, and pressure are contained within $H^1(\Omega)^d$, C_γ, and $L^2_0(\Omega)$, respectively, as well as a divergence-free DG method where the finite element spaces for the material distribution and pressure remain conforming, but now the velocity is approximated by discontinuous piecewise polynomials. Borrvall and Petersson's [36] original discretization was a conforming finite element method. Their proof of convergence showed that the sequence of finite element solutions to the optimization problem, as the mesh size tended to zero, had a weakly(-*) converging subsequence whose limit was a solution to the original analytical problem. This framework is typical for the convergence results of finite element methods for topology optimization problems [35, 36, 124–127]. However, the analysis left a number of open problems. For example, can all the isolated local minimizers be approximated by a sequence of computable finite element functions and can these sequences be shown to converge strongly? We proved a positive result. By fixing an isolated minimizer and considering its basin of attraction, we showed that a sequence of finite element minimizers, satisfying a modified optimization problem, will strongly converge to the isolated minimizer. Moreover, a subsequence of the finite element minimizers of the modified optimization problem also satisfy the discretized first-order optimality conditions of the original Borrvall–Petersson problem. Hence, by computing multiple solutions of the discretized first-order optimality conditions, we can approximate the multiple solutions of the analytical nonconvex problem. The convergence results also resolve the lack of checkerboarding in the material distribution approximation.

There are several open problems left to consider. These include whether similar finite element convergence results can be proven for finite volume discretizations of the Borrvall–Petersson model, Borrvall–Petersson models where the fluid satisfies Navier–Stokes or non-Newtonian flow, and other topology optimization problems, e.g. density models for the compliance of elastic structures. Moreover, an estimate for the convergence rate of the $L^2(\Omega)$-norm error of the velocity in a Taylor–Hood mixed finite element discretization for the velocity-pressure pair would automatically give

convergence rates for the H^1-norm error of the velocity, L^2-norm error of the material distribution, and the L^2-norm error of the pressure, due to our results in Section 3.3.

A key step in the proof for the finite element convergence is the extraction of an $L^2(\Omega)^d$-strongly converging subsequence of the velocity finite element minimizers. In the conforming finite element method case, this followed by applying the Rellich–Kondrachov theorem to the $H^1(\Omega)^d$-weakly converging sequence. In the divergence-free DG methods case, we required alternative compactness results as found in the work of Buffa and Ortner [46]. We hypothesize that our arguments can be extended to discretizations of the Borrvall–Petersson problem where this extraction is possible. This has potential applications in resolving convergence results for finite volume discretizations. For Borrvall–Petersson models with more complicated fluid flow, careful consideration would be required, and traditional results from standard Navier–Stokes and non-Newtonian flow would need to be utilized [27, 70, 82, 83].

6.2 Solvers for computing multiple solutions of topology optimization problems

Having resolved whether the finite element method can approximate all the different isolated minimizers of Borrvall–Petersson problem, we then developed a solver that can compute these multiple minimizers in a systematic way. To this end, we developed the deflated barrier method in Chapter 4. The deflated barrier method can be applied to a general density-based topology optimization problem and consists of three main components:

- deflation, a mechanism that prevents a Newton-like solver from converging to a previously found solution;
- a primal-dual active set strategy, a Newton-like solver that can enforce the box constraints on the material distribution;
- barrier terms that aid the global nonlinear convergence of the algorithm.

First the barrier functional of the optimization problem is constructed. This consists of the Lagrangian of the problem appended by log-barrier terms. Then the first-order optimality conditions of the barrier functional are (automatically) derived and the system is solved with the primal-dual active set strategy to enforce the box constraints on the material distribution. Here, the discovered solution is deflated and the first-order optimality conditions are solved again. Since the algorithm can no longer converge to the discovered solution, the primal-dual active set strategy (hopefully) converges to a different solution. The deflation process can be repeated any number of times to discover multiple solutions. The barrier parameter is then decreased and the previous solutions are used as initial guesses in the continuation scheme. In this way, we follow branches of solutions as the barrier parameter goes to zero. The algorithm is terminated when the first-order optimality conditions are solved with a barrier parameter value of zero.

The deflated barrier method was applied to a number of problems. Our highlighted examples are the discovery of 42 solutions of a five-holes double-pipe (a Borrvall–Petersson problem constrained by the Navier–Stokes equations), 4 solutions of the double-pipe problem with natural boundary conditions on the outlets, 2 solutions of a fluid problem with a roller pump, 2 solutions of a cantilever compliance problem, and 2 solutions of an MBB compliance problem. In a conforming discretization, the primal-dual active set strategy iteration counts are mesh independent for the Borrvall–Petersson problem, i.e. they remained roughly constant as the mesh is refined, irrespective of the choice of the finite element discretization. Moreover, the method converged superlinearly at each subproblem. We explored how the use of the active set helps control the ill-conditioning caused by the barrier terms as the barrier parameter approaches zero. The iteration counts were not mesh independent for the compliance problems, but a simple grid-sequencing strategy can be used to obtain solutions with sharp interfaces on fine meshes with relative ease.

The mesh dependence in the case of compliance problems is likely caused by the presence of $\nabla\rho$ in the formulation. It is well understood that primal-dual active set

strategies applied to infinite-dimensional obstacle-like problems are not semismooth Newton methods [88, Sec. 4]. The discrepancy of being a valid semismooth Newton method in finite dimensions and not a valid semismooth Newton method in infinite dimensions manifests precisely as mesh dependence. A future objective would be to develop a deflated barrier method variant that is mesh independent for compliance problems. One suggestion is to swap the primal-dual active set strategy and barrier terms for a different continuation scheme that is mesh independent for obstacle problems, e.g. Newton's method coupled with a Moreau–Yosida regularization [7, 68]. However, we note that developing a converging continuation scheme is nontrivial.

A natural question is whether deflation could be combined with a more traditional optimization strategy in topology optimization such as MMA [162, 163]. An issue that arises is that the deflation modification to the update of the optimization process (see Proposition 4.3) is tied to the choice of a Newton-like solver. It is not immediately clear how to translate this to nested first-order methods. Implementations found in the literature tend to quotient the discovered solution on the level of the objective functional [85, 185]. For example, if $\mathbf{x}_* \in \mathbb{R}^n$ is an already discovered solution, then the original problem is modified as follows: for $a, b > 0$,

$$\min_{\mathbf{x}} J(\mathbf{x}) \underbrace{\longrightarrow}_{\text{deflate } \mathbf{x}_*} \min_{\mathbf{x}} \left(\frac{1}{a\|\mathbf{x} - \mathbf{x}_*\|^p} + b \right) J(\mathbf{x}). \tag{6.1}$$

However, unlike the deflation mechanism described in Section 4.3, (6.1) may introduce minimizers that do not exist in the original problem or remove undiscovered minimizers. Moreover, since the energy landscape is modified, the minimizers of the original problem are perturbed. To help remedy this, the constants a and b must be chosen adaptively throughout the optimization process, whereas in our implementation they are constant, in not just the optimization process, but across all examples. Since deflation is ideally a post-processing step after the computation of the update, these real-time modifications are undesirable.

In order to apply the deflated barrier method to three-dimensional Borrvall–Petersson problems, we developed preconditioners for the linear systems that are solved during the optimization process in Chapter 5. This was made possible as

we only solve undeflated systems, even if we are deflating discovered solutions [66]. We fix the discretization to a piecewise constant discretization of the material distribution and a Brezzi–Douglas–Marini discretization for the velocity-pressure pair [40, 41]. By deriving the linear systems arising during the Benson–Munson primal dual active-set strategy, we showed that we are solving a 4×4 block matrix system at each iteration. One block is a row/column vector and can be row eliminated. After applying block preconditioning to the remaining 3×3 block, we reduced the solve to the size of a standard Stokes system [65]. Then, by using an augmented Lagrangian control term, we applied another block preconditioning that reduced the solve to inverting a diagonal matrix, factorizing a block diagonal matrix (that can be cached), and applying the inverse action of an augmented momentum block [60, 158, 174]. In order to solve the augmented momentum block on fine meshes, we developed a multigrid cycle with a specialized relaxation scheme that can handle the parameter-dependent semi-definite augmented Lagrangian term [71, 73, 91, 143] and also required a characterization of the active set of the Benson–Munson strategy on the coarser levels [62, 93].

We applied the preconditioner to three different examples. The first was the double-pipe problem that was introduced in Section 4.6.1. Here we showed that the preconditioner was robust to the mesh size and the polynomial order of the discretization. Moreover, we compared our preconditioner to the Cahouet–Chabard strategy [36, 48] and showed that our preconditioner was more effective for the problems considered in this thesis. The convergence results in Chapter 3 for divergence-free DG discretizations were also numerically verified and we compared the violation of the incompressibility constraint between a Taylor–Hood and BDM discretization for the velocity-pressure pair. The first three-dimensional example we considered was a cross-channel problem [141, Sec. 7.5]. For computational efficiency, we first discretized the problem on a coarse mesh and used an LU factorization for the augmented momentum block. This led to the discovery of three distinct solutions, which were subsequently grid-sequenced onto a finer mesh. On the finer mesh, we used FGMRES preconditioned with our specialized multigrid cycle for the

augmented momentum block solve. Finally, the work in this chapter culminated in the computation of eleven distinct solutions to a three-dimensional five-holes quadruple-pipe problem (an extension of the two-dimensional five-holes double-pipe problem of Section 4.6.5). Again, the solutions were originally found on a coarse mesh and grid-sequenced to a fine mesh.

A key extension for industrial applications would be to develop preconditioners for the deflated barrier method applied to the Borrvall–Petersson problem with a Navier–Stokes constraint. In this case, the linear systems would involve a 7×7 block matrix. It may be possible to utilize similar ideas including using block preconditioning for the volume constraint row and material distribution block.

Thinking more broadly, the deflated barrier method is an attractive approach for a number of problems including time-dependent Borrvall–Petersson problems, the topology optimization of heat transfer, and compliance problems involving materials that are hyperelastic. The deflated barrier method, accompanied with the theoretical finite element convergence theorems, provide a bedrock for exploring the solution landscape of density-based topology optimization problems. Once computed, the solutions can be utilized as initial guesses for industrial shape and topology optimization algorithms, where other practical considerations can be taken into account.

References

[1] N. Aage, E. Andreassen, B. S. Lazarov, and O. Sigmund, "Giga-voxel computational morphogenesis for structural design", *Nature*, vol. 550, no. 7674, pp. 84–86, 2017. DOI: 10.1038/nature23911.

[2] N. Aage and B. S. Lazarov, "Parallel framework for topology optimization using the method of moving asymptotes", *Structural and Multidisciplinary Optimization*, vol. 47, no. 4, pp. 493–505, 2013. DOI: 10.1007/s00158-012-0869-2.

[3] N. Aage, T. H. Poulsen, A. Gersborg-Hansen, and O. Sigmund, "Topology optimization of large scale Stokes flow problems", *Structural and Multidisciplinary Optimization*, vol. 35, no. 2, pp. 175–180, 2007. DOI: 10.1007/s00158-007-0128-0.

[4] W. Achtziger and M. Stolpe, "Global optimization of truss topology with discrete bar areas—Part I: Theory of relaxed problems", *Computational Optimization and Applications*, vol. 40, no. 2, pp. 247–280, 2008. DOI: 10.1007/s10589-007-9138-5.

[5] ——, "Global optimization of truss topology with discrete bar areas—Part II: Implementation and numerical results", *Computational Optimization and Applications*, vol. 44, no. 2, pp. 315–341, 2009. DOI: 10.1007/s10589-007-9152-7.

[6] L. Adam, M. Hintermüller, D. Peschka, and T. M. Surowiec, "Optimization of a multiphysics problem in semiconductor laser design", *SIAM Journal on Applied Mathematics*, vol. 79, no. 1, pp. 257–283, 2019. DOI: 10.1137/18M1179183.

[7] L. Adam, M. Hintermüller, and T. M. Surowiec, "A semismooth Newton method with analytical path-following for the H^1-projection onto the Gibbs simplex", *IMA Journal of Numerical Analysis*, vol. 39, no. 3, pp. 1276–1295, 2019. DOI: 10.1093/imanum/dry034.

[8] R. A. Adams, *Sobolev spaces*. New York: Academic Press, 1975, ISBN: 978-0-120-44150-1.

[9] J. Alexandersen and C. S. Andreasen, "A review of topology optimisation for fluid-based problems", *Fluids*, vol. 5, no. 1, p. 29, 2020. DOI: 10.3390/fluids5010029.

[10] G Allaire and G. Francfort, "A numerical algorithm for topology and shape optimization", in *Topology Design of Structures*, Springer, 1993, pp. 239–248. DOI: 10.1007/978-94-011-1804-0_16.

[11] G. Allaire, F. Gournay, F. Jouve, and A.-M. Toader, "Structural optimization using topological and shape sensitivity via a level set method", *Control and Cybernetics*, vol. 34, pp. 59–80, 2005.

[12] G. Allaire, F. Jouve, and A.-M. Toader, "Structural optimization using sensitivity analysis and a level-set method", *Journal of Computational Physics*, vol. 194, no. 1, pp. 363–393, 2004. DOI: 10.1016/j.jcp.2003.09.032.

[13] M. S. Alnæs, A. Logg, K. B. Ølgaard, M. E. Rognes, and G. N. Wells, "Unified form language: A domain-specific language for weak formulations of partial differential equations", *ACM Transactions on Mathematical Software (TOMS)*, vol. 40, no. 2, pp. 1–37, 2014. DOI: 10.1145/2566630.

[14] D. H. Alonso, L. F. N. de Sá, J. S. R. Saenz, and E. C. N. Silva, "Topology optimization applied to the design of 2D swirl flow devices", *Structural and Multidisciplinary Optimization*, vol. 58, no. 6, pp. 2341–2364, 2018. DOI: 10.1007/s00158-018-2078-0.

[15] ——, "Topology optimization based on a two-dimensional swirl flow model of Tesla-type pump devices", *Computers & Mathematics with Applications*, vol. 77, no. 9, pp. 2499–2533, 2019. DOI: 10.1016/j.camwa.2018.12.035.

[16] D. H. Alonso, J. S. R. Saenz, and E. C. N. Silva, "Non-Newtonian laminar 2D swirl flow design by the topology optimization method", *Structural and Multidisciplinary Optimization*, pp. 1–23, 2020. DOI: 10.1007/s00158-020-02499-2.

[17] P. R. Amestoy, I. S. Duff, J.-Y. L'Excellent, and J. Koster, "A fully asynchronous multifrontal solver using distributed dynamic scheduling", *SIAM Journal on Matrix Analysis and Applications*, 2001. DOI: 10.1137/S0895479899358194.

[18] E. Andreassen, A. Clausen, M. Schevenels, B. S. Lazarov, and O. Sigmund, "Efficient topology optimization in MATLAB using 88 lines of code", *Structural and Multidisciplinary Optimization*, vol. 43, no. 1, pp. 1–16, 2011. DOI: 10.1007/s00158-010-0594-7.

[19] D. N. Arnold, *Finite element exterior calculus*. SIAM, 2018, ISBN: 978-1-61197-553-6. DOI: 10.1137/1.9781611975543.

[20] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, "Unified analysis of discontinuous Galerkin methods for elliptic problems", *SIAM Journal on Numerical Analysis*, vol. 39, no. 5, pp. 1749–1779, 2002. DOI: 10.1137/S0036142901384162.

[21] D. N. Arnold, R. S. Falk, and R. Winther, "Multigrid in $H(\mathrm{div})$ and $H(\mathrm{curl})$", *Numerische Mathematik*, vol. 85, no. 2, pp. 197–217, 2000. DOI: 10.1007/PL00005386.

[22] R Balamurugan, C. Ramakrishnan, and N. Singh, "Performance evaluation of a two stage adaptive genetic algorithm (TSAGA) in structural topology optimization", *Applied Soft Computing*, vol. 8, no. 4, pp. 1607–1624, 2008. DOI: 10.1016/j.asoc.2007.10.022.

[23] R Balamurugan, C. Ramakrishnan, and N Swaminathan, "A two phase approach based on skeleton convergence and geometric variables for topology optimization using genetic algorithm", *Structural and Multidisciplinary Optimization*, vol. 43, no. 3, pp. 381–404, 2011. DOI: 10.1007/s00158-010-0560-4.

[24] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, V. Eijkout, W. Gropp, R. Tran Mills, T. Munson, K. Rupp, P. Sana, B. Smith, S. Zampini, H. Zhang, and H. Zhang, "PETSc users manual", Argonne National Laboratory, Tech. Rep. ANL-95/11 - Revision 3.11, 2019.

[25] S. Banach, *Théorie des Opérations Linéaires*. 1932, ISBN: 9782876471481.

[26] R. Behrou, R. Ranjan, and J. K. Guest, "Adaptive topology optimization for incompressible laminar flow problems with mass flow constraints", *Computer Methods in Applied Mechanics and Engineering*, vol. 346, pp. 612–641, 2019. DOI: 10.1016/j.cma.2018.11.037.

[27] L. Belenki, L. C. Berselli, L. Diening, and M. Růžička, "On the finite element approximation of p-Stokes systems", *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 373–397, 2012. DOI: 10.1137/10080436X.

[28] M. P. Bendsøe, "Optimal shape design as a material distribution problem", *Structural Optimization*, vol. 1, no. 4, pp. 193–202, 1989. DOI: 10.1007/BF01650949.

[29] M. P. Bendsøe and O. Sigmund, *Topology Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ISBN: 978-3-642-07698-5. DOI: 10.1007/978-3-662-05086-6.

[30] M. P. Bendsøe and N. Kikuchi, "Generating optimal topologies in structural design using a homogenization method", *Computer Methods in Applied Mechanics and Engineering*, vol. 71, no. 2, pp. 197–224, 1988. DOI: 10.1016/0045-7825(88)90086-2.

[31] S. J. Benson and T. S. Munson, "Flexible complementarity solvers for large-scale applications", *Optimization Methods and Software*, vol. 21, no. 1, pp. 155–168, 2003. DOI: 10.1080/10556780500065382.

[32] M. Benzi, G. H. Golub, and J. Liesen, "Numerical solution of saddle point problems", *Acta Numerica*, vol. 14, pp. 1–137, May 2005. DOI: 10.1017/S0962492904000212.

[33] M. Benzi and M. A. Olshanskii, "An augmented Lagrangian-based approach to the Oseen problem", *SIAM Journal on Scientific Computing*, vol. 28, no. 6, pp. 2095–2113, 2006. DOI: 10.1137/050646421.

[34] T. Borrvall, "Topology optimization of elastic continua using restriction", *Archives of Computational Methods in Engineering*, vol. 8, no. 4, pp. 351–385, 2001. DOI: 10.1007/BF02743737.

[35] T. Borrvall and J. Petersson, "Topology optimization using regularized intermediate density control", *Computer Methods in Applied Mechanics and Engineering*, vol. 190, no. 37-38, pp. 4911–4928, 2001. DOI: 10.1016/S0045-7825(00)00356-X.

[36] ——, "Topology optimization of fluids in Stokes flow", *International Journal for Numerical Methods in Fluids*, vol. 41, no. 1, pp. 77–107, 2003. DOI: 10.1002/fld.426.

[37] B. Bourdin, "Filters in topology optimization", *International Journal for Numerical Methods in Engineering*, vol. 50, no. 9, pp. 2143–2158, 2001. DOI: 10.1002/nme.116.

[38] B. Bourdin and A. Chambolle, "The phase-field method in optimal design", in *IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials*, Springer, 2006, pp. 207–215. DOI: 10.1007/1-4020-4752-5_21.

[39] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 3rd ed., ser. Texts in Applied Mathematics. New York, NY: Springer New York, 2008, vol. 15, ISBN: 978-0-387-75933-3. DOI: 10.1007/978-0-387-75934-0.

[40] F. Brezzi, J. Douglas, R. Durán, and M. Fortin, "Mixed finite elements for second order elliptic problems in three variables", *Numerische Mathematik*, vol. 51, no. 2, pp. 237–250, 1987. DOI: 10.1007/BF01396752.

[41] F. Brezzi, J. Douglas, and L. D. Marini, "Two families of mixed finite elements for second order elliptic problems", *Numerische Mathematik*, vol. 47, no. 2, pp. 217–235, 1985. DOI: 10.1007/BF01389710.

[42] K. M. Brown and W. B. Gearhart, "Deflation techniques for the calculation of further solutions of a nonlinear system", *Numerische Mathematik*, vol. 16, no. 4, pp. 334–342, 1971. DOI: 10.1007/BF02165004.

[43] P. R. Brune, M. G. Knepley, B. F. Smith, and X. Tu, "Composing scalable nonlinear algebraic solvers", *SIAM Review*, vol. 57, no. 4, pp. 535–565, 2015. DOI: 10.1137/130936725.

[44] T. Bruns, "A reevaluation of the SIMP method with filtering and an alternative formulation for solid–void topology optimization", *Structural and Multidisciplinary Optimization*, vol. 30, no. 6, pp. 428–436, 2005. DOI: 10.1007/s00158-005-0537-x.

[45] T. E. Bruns and D. A. Tortorelli, "Topology optimization of non-linear elastic structures and compliant mechanisms", *Computer Methods in Applied Mechanics and Engineering*, vol. 190, no. 26-27, pp. 3443–3459, 2001. DOI: 10.1016/S0045-7825(00)00278-4.

[46] A. Buffa and C. Ortner, "Compact embeddings of broken Sobolev spaces and applications", *IMA Journal of Numerical Analysis*, vol. 29, no. 4, pp. 827–855, 2009. DOI: 10.1093/imanum/drn038.

[47] M. Burger, B. Hackl, and W. Ring, "Incorporating topological derivatives into level set methods", *Journal of Computational Physics*, vol. 194, no. 1, pp. 344–362, 2004. DOI: 10.1016/j.jcp.2003.09.033.

[48] J. Cahouet and J.-P. Chabard, "Some fast 3D finite element solvers for the generalized Stokes problem", *International Journal for Numerical Methods in Fluids*, vol. 8, no. 8, pp. 869–895, 1988. DOI: 10.1002/fld.1650080802.

[49] V. J. Challis and J. K. Guest, "Level set topology optimization of fluids in Stokes flow", *International Journal for Numerical Methods in Engineering*, vol. 79, no. 10, pp. 1284–1308, 2009. DOI: 10.1002/nme.2616.

[50] E. G. Charalampidis, P. G. Kevrekidis, and P. E. Farrell, "Computing stationary solutions of the two-dimensional Gross-Pitaevskii equation with deflated continuation.", *Communications in Nonlinear Science and Numerical Simulation*, vol. 54, pp. 482–499, 2018. DOI: 10.1016/j.cnsns.2017.05.024.

[51] B. Cockburn, G. Kanschat, and D. Schötzau, "A note on discontinuous Galerkin divergence-free solutions of the Navier–Stokes equations", *Journal of Scientific Computing*, vol. 31, no. 1, pp. 61–73, 2007. DOI: 10.1007/s10915-006-9107-7.

[52] B. Cockburn, G. Kanschat, D. Schötzau, and C. Schwab, "Local discontinuous Galerkin methods for the Stokes system", *SIAM Journal on Numerical Analysis*, vol. 40, no. 1, pp. 319–343, 2002. DOI: 10.1137/S0036142900380121.

[53] J. D. Deaton and R. V. Grandhi, "A survey of structural and multidisciplinary continuum topology optimization: Post 2000", *Structural and Multidisciplinary Optimization*, vol. 49, no. 1, pp. 1–38, 2014. DOI: 10.1007/s00158-013-0956-z.

[54] Y. Deng, Z. Liu, J. Wu, and Y. Wu, "Topology optimization of steady Navier–Stokes flow with body force", *Computer Methods in Applied Mechanics and Engineering*, vol. 255, pp. 306–321, 2013. DOI: 10.1016/j.cma.2012.11.015.

[55] Y. Deng, Z. Liu, P. Zhang, Y. Liu, and Y. Wu, "Topology optimization of unsteady incompressible Navier–Stokes flows", *Journal of Computational Physics*, vol. 230, no. 17, pp. 6688–6708, 2011. DOI: 10.1016/j.jcp.2011.05.004.

[56] Y. Deng, Y. Wu, and Z. Liu, *Topology Optimization Theory for Laminar Flow: Applications in Inverse Design of Microfluidics*. Singapore: Springer Singapore, 2018. DOI: 10.1007/978-981-10-4687-2.

[57] C. B. Dilgen, S. B. Dilgen, D. R. Fuhrman, O. Sigmund, and B. S. Lazarov, "Topology optimization of turbulent flows", *Computer Methods in Applied Mechanics and Engineering*, vol. 331, pp. 363–393, 2018. DOI: 10.1016/j.cma.2017.11.029.

[58] S. B. Dilgen, C. B. Dilgen, D. R. Fuhrman, O. Sigmund, and B. S. Lazarov, "Density based topology optimization of turbulent flow heat transfer systems", *Structural and Multidisciplinary Optimization*, vol. 57, no. 5, pp. 1905–1918, 2018. DOI: 10.1007/s00158-018-1967-6.

[59] Z. Doubrovski, J. C. Verlinden, and J. M. P. Geraedts, "Optimal design for additive manufacturing: Opportunities and challenges", ser. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, vol. 9: 23rd International Conference on Design Theory and Methodology; 16th Design for Manufacturing and the Life Cycle Conference, 2011, pp. 635–646. DOI: 10.1115/DETC2011-48131.

[60] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Oxford University Press, 2014, ISBN: 978-0199678808. DOI: 10.1093/acprof:oso/9780199678792.001.0001.

[61] D. B. Emerson, P. E. Farrell, J. H. Adler, S. P. MacLachlan, and T. J. Atherton, "Computing equilibrium states of cholesteric liquid crystals in elliptical channels with deflation algorithms", *Liquid Crystals*, vol. 45, no. 3, pp. 341–350, 2018. DOI: 10.1080/02678292.2017.1365385.

[62] M. Engel and M. Griebel, "A multigrid method for constrained optimal control problems", *Journal of Computational and Applied Mathematics*, vol. 235, no. 15, pp. 4368–4388, 2011. DOI: 10.1016/j.cam.2011.04.002.

[63] L. C. Evans, *Partial Differential Equations*, 2nd ed. American Mathematical Society, 2010, ISBN: 978-0821849743.

[64] A. Evgrafov, "Topology optimization of slightly compressible fluids", *ZAMM - Journal of Applied Mathematics and Mechanics*, vol. 86, no. 1, pp. 46–62, 2006. DOI: 10.1002/zamm.200410223.

[65] ——, "State space Newton's method for topology optimization", *Computer Methods in Applied Mechanics and Engineering*, vol. 278, pp. 272–290, 2014. DOI: 10.1016/j.cma.2014.06.005.

[66] P. E. Farrell, Á. Birkisson, and S. W. Funke, "Deflation techniques for finding distinct solutions of nonlinear partial differential equations", *SIAM Journal on Scientific Computing*, vol. 37, no. 4, A2026–A2045, 2015. DOI: 10.1137/140984798.

[67] P. E. Farrell, P. A. G. Orozco, and E. Süli, *Finite element approximation and augmented Lagrangian preconditioning for anisothermal implicitly-constituted non-Newtonian flow.* arXiv:2011.03024, 2020.

[68] P. E. Farrell, M. Croci, and T. M. Surowiec, "Deflation for semismooth equations", *Optimization Methods and Software*, pp. 1–24, 2019. DOI: 10.1080/10556788.2019.1613655.

[69] P. E. Farrell and P. A. Gazca-Orozco, "An augmented Lagrangian preconditioner for implicitly constituted non-Newtonian incompressible flow", *SIAM Journal on Scientific Computing*, vol. 42, no. 6, B1329–B1349, 2020. DOI: 10.1137/20M1336618.

[70] P. E. Farrell, P. A. Gazca-Orozco, and E. Süli, "Numerical analysis of unsteady implicitly constituted incompressible fluids: 3-field formulation", *SIAM Journal on Numerical Analysis*, vol. 58, no. 1, pp. 757–787, 2020. DOI: 10.1137/19M125738X.

[71] P. E. Farrell, M. G. Knepley, L. Mitchell, and F. Wechsung, "PCPATCH: Software for the topological construction of multigrid relaxation methods", *ACM Transactions on Mathematical Software (TOMS)*, vol. 47, no. 3, pp. 1–22, 2021. DOI: 10.1145/3445791.

[72] P. E. Farrell, L. Mitchell, L. R. Scott, and F. Wechsung, "A Reynolds-robust preconditioner for the Scott-Vogelius discretization of the stationary incompressible Navier-Stokes equations", *The SMAI Journal of Computational Mathematics*, vol. 7, pp. 75–96, 2021. DOI: 10.5802/smai-jcm.72.

[73] P. E. Farrell, L. Mitchell, and F. Wechsung, "An augmented Lagrangian preconditioner for the 3D stationary incompressible Navier–Stokes equations at high Reynolds number", *SIAM Journal on Scientific Computing*, vol. 41, no. 5, A3073–A3096, 2019. DOI: 10.1137/18M1219370.

[74] F. Ferrari and O. Sigmund, "A new generation 99 line Matlab code for compliance topology optimization and its extension to 3D", *Structural and Multidisciplinary Optimization*, vol. 62, no. 4, pp. 2211–2228, 2020. DOI: 10.1007/s00158-020-02629-w.

[75] P. Fitzpatrick, *Advanced calculus*, 2nd ed. American Mathematical Soc., 2009, vol. 5, ISBN: 978-0-8218-4791-6.

[76] I. Fonseca and G. Leoni, *Modern Methods in the Calculus of Variations: Lp Spaces*, ser. Springer Monographs in Mathematics. New York, NY: Springer New York, 2006, ISBN: 978-0-387-35784-3. DOI: 10.1007/978-0-387-69006-3.

[77] A. Forsgren, P. E. Gill, and M. H. Wright, "Interior methods for nonlinear optimization", *SIAM Review*, vol. 44, no. 4, pp. 525–597, 2002. DOI: 10.1137/S0036144502414942.

[78] N. R. Gauger, A. Linke, and P. W. Schroeder, "On high-order pressure-robust space discretisations, their advantages for incompressible high Reynolds number generalised Beltrami flows and beyond", *The SMAI Journal of Computational Mathematics*, vol. 5, pp. 89–129, 2019. DOI: 10.5802/smai-jcm.44.

[79] A. Gersborg-Hansen, M. P. Bendsøe, and O. Sigmund, "Topology optimization of heat conduction problems using the finite volume method", *Structural and Multidisciplinary Optimization*, vol. 31, no. 4, pp. 251–259, 2006. DOI: 10.1007/s00158-005-0584-3.

[80] A. Gersborg-Hansen, O. Sigmund, and R. B. Haber, "Topology optimization of channel flow problems", *Structural and Multidisciplinary Optimization*, vol. 30, no. 3, pp. 181–192, 2005. DOI: 10.1007/s00158-004-0508-7.

[81] C. Geuzaine and J.-F. Remacle, "Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities", *International Journal for Numerical Methods in Engineering*, vol. 79, no. 11, pp. 1309–1331, 2009. DOI: 10.1002/nme.2579.

[82] V. Girault and P.-A. Raviart, *Finite element methods for Navier–Stokes equations: theory and algorithms*. Springer-Verlag Berlin Heidelberg, 1986, vol. 5. DOI: 10.1007/978-3-642-61623-5.

[83] R. Glowinski and O. Pironneau, "Finite element methods for Navier–Stokes equations", *Annual Review of Fluid Mechanics*, vol. 24, no. 1, pp. 167–204, 1992. DOI: 10.1146/annurev.fl.24.010192.001123.

[84] J. Greifenstein and M. Stingl, "Simultaneous parametric material and topology optimization with constrained material grading", *Structural and Multidisciplinary Optimization*, vol. 54, no. 4, pp. 985–998, 2016. DOI: 10.1007/s00158-016-1457-7.

[85] Y. Gu, C. Wang, and H. Yang, "Structure probing neural network deflation", *Journal of Computational Physics*, vol. 434, p. 110 231, 2021. DOI: 10.1016/j.jcp.2021.110231.

[86] J. Haslinger and R. A. Mäkinen, "On a topology optimization problem governed by two-dimensional Helmholtz equation", *Computational Optimization and Applications*, vol. 62, no. 2, pp. 517–544, 2015. DOI: 10.1007/s10589-015-9746-4.

[87] J. G. Heywood, R. Rannacher, and S. Turek, "Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations", *International Journal for Numerical Methods in Fluids*, vol. 22, no. 5, pp. 325–352, 1996. DOI: 10.1002/(SICI)1097-0363(19960315)22:5<325::AID-FLD307>3.0.CO;2-Y.

[88] M. Hintermüller, K. Ito, and K. Kunisch, "The primal-dual active set strategy as a semismooth Newton method", *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 865–888, 2003. DOI: 10.1137/S1052623401383558.

[89] M. Hintermüller and M. Ulbrich, "A mesh-independence result for semismooth Newton methods", *Mathematical Programming*, vol. 101, no. 1, pp. 151–184, 2004. DOI: 10.1007/s10107-004-0540-9.

[90] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*. Springer Science & Business Media, 2008, vol. 23, ISBN: 978-1-4020-8839-1. DOI: 10.1007/978-1-4020-8839-1.

[91] Q. Hong, J. Kraus, J. Xu, and L. Zikatanov, "A robust multigrid method for discontinuous Galerkin discretizations of Stokes and linear elasticity equations", *Numerische Mathematik*, vol. 132, no. 1, pp. 23–49, 2016. DOI: 10.1007/s00211-015-0712-y.

[92] R. H. Hoppe, S. I. Petrova, and V. Schulz, "Primal-dual Newton-type interior-point method for topology optimization", *Journal of Optimization Theory and Applications*, vol. 114, no. 3, pp. 545–571, 2002. DOI: 10.1023/A:1016070928600.

[93] R. H. Hoppe, "Multigrid algorithms for variational inequalities", *SIAM Journal on Numerical Analysis*, vol. 24, no. 5, pp. 1046–1065, 1987. DOI: 10.1137/0724069.

[94] C. Jain and A. Saxena, "An improved material-mask overlay strategy for topology optimization of structures and compliant mechanisms", *Journal of Mechanical Design*, 2010. DOI: 10.1115/1.4001530.

[95] K. E. Jensen, "Topology optimization of Stokes flow on dynamic meshes using simple optimizers", *Computers & Fluids*, vol. 174, pp. 66–77, 2018. DOI: 10.1016/j.compfluid.2018.07.011.

[96] V. John, A. Linke, C. Merdon, M. Neilan, and L. G. Rebholz, "On the divergence constraint in mixed finite element methods for incompressible flows", *SIAM Review*, vol. 59, no. 3, pp. 492–544, 2017. DOI: 10.1137/15M1047696.

[97] R. B. Kellogg and J. E. Osborn, "A regularity result for the Stokes problem in a convex polygon", *Journal of Functional Analysis*, vol. 21, no. 4, pp. 397–431, 1976. DOI: 10.1016/0022-1236(76)90035-5.

[98] J. Koch, E. Papoutsis-Kiachagias, and K. Giannakoglou, "Transition from adjoint level set topology to shape optimization for 2D fluid mechanics", *Computers & Fluids*, vol. 150, pp. 123–138, 2017. DOI: 10.1016/j.compfluid.2017.04.001.

[99] J. Könnö and R. Stenberg, "$H(\mathrm{div})$-conforming finite elements for the Brinkman problem", *Mathematical Models and Methods in Applied Sciences*, vol. 21, no. 11, pp. 2227–2248, 2011. DOI: 10.1142/S0218202511005726.

[100] ——, "Numerical computations with $H(\mathrm{div})$-finite elements for the Brinkman problem", *Computational Geosciences*, vol. 16, no. 1, pp. 139–158, 2012. DOI: 10.1007/s10596-011-9259-x.

[101] E. Kontoleontos, E. Papoutsis-Kiachagias, A. Zymaris, D. Papadimitriou, and K. Giannakoglou, "Adjoint-based constrained topology optimization for viscous flows, including heat transfer", *Engineering Optimization*, vol. 45, no. 8, pp. 941–961, 2013. DOI: 10.1080/0305215X.2012.717074.

[102] V. A. Kozlov, V. G. Maz'ya, and C. Schwab, "On singularities of solutions to the Dirichlet problem of hydrodynamics near the vertex of a cone", *Journal für die reine und angewandte Mathematik*, vol. 456, pp. 65–97, 1994. DOI: 10.1515/crll.1994.456.65.

[103] S. Kreissl, G. Pingen, and K. Maute, "Topology optimization for unsteady flow", *International Journal for Numerical Methods in Engineering*, vol. 87, no. 13, pp. 1229–1253, 2011. DOI: 10.1002/nme.3151.

[104] F. Laakmann, P. E. Farrell, and L. Mitchell, "An augmented Lagrangian preconditioner for the magnetohydrodynamics equations at high Reynolds and coupling numbers", *arXiv preprint arXiv:2104.14855*, 2021.

[105] B. S. Lazarov and O. Sigmund, "Filters in topology optimization based on Helmholtz-type differential equations", *International Journal for Numerical Methods in Engineering*, vol. 86, no. 6, pp. 765–781, 2011. DOI: 10.1002/nme.3072.

[106] A. V. Levy and S. Gómez, "The tunneling method applied to global optimization", in *Numerical Optimization*, P. T. Boggs, Ed., Society for Industrial and Applied Mathematics, 1984, ISBN: 9780898710540.

[107] A. Limache, S. Idelsohn, R. Rossi, and E. Oñate, "The violation of objectivity in Laplace formulations of the Navier–Stokes equations", *International Journal for Numerical Methods in Fluids*, vol. 54, no. 6-8, pp. 639–664, 2007. DOI: 10.1002/fld.1480.

[108] A. Linke and L. G. Rebholz, "Pressure-induced locking in mixed methods for time-dependent (Navier–) Stokes equations", *Journal of Computational Physics*, vol. 388, pp. 350–356, 2019. DOI: 10.1016/j.jcp.2019.03.010.

[109] A. Logg, K.-A. Mardal, and G. Wells, "Automated solution of differential equations by the finite element method: The FEniCS book", *Springer Science and Business Media*, vol. 84, 2012. DOI: 10.1007/978-3-642-23099-8.

[110] J. A. Madeira, H. Pina, and H. Rodrigues, "GA topology optimization using random keys for tree encoding of structures", *Structural and Multidisciplinary Optimization*, vol. 40, no. 1-6, p. 227, 2010. DOI: 10.1007/s00158-008-0353-1.

[111] G. Marck, M. Nemer, and J.-L. Harion, "Topology optimization of heat and mass transfer problems: Laminar flow", *Numerical Heat Transfer, Part B: Fundamentals*, vol. 63, no. 6, pp. 508–539, 2013. DOI: 10.1080/10407790.2013.772001.

[112] E. Medina, P. E. Farrell, K. Bertoldi, and C. Rycroft, "Navigating the landscape of nonlinear mechanical metamaterials for advanced programmability.", *Physical Review B*, vol. 101, no. 6, 2020. DOI: 10.1103/PhysRevB.101.064101.

[113] L. Modica, "The gradient theory of phase transitions and the minimal interface criterion", *Archive for Rational Mechanics and Analysis*, vol. 98, no. 2, pp. 123–142, 1987. DOI: 10.1007/BF00251230.

[114] M. F. Murphy, G. H. Golub, and A. J. Wathen, "A note on preconditioning for indefinite linear systems", *SIAM Journal on Scientific Computing*, vol. 21, no. 6, pp. 1969–1972, 2000. DOI: 10.1137/S1064827599355153.

[115] S. G. Nash, R. Polyak, and A. Sofer, "A numerical comparison of barrier and modified barrier methods for large-scale bound-constrained optimization", in *Large Scale Optimization*, vol. 1, Boston, MA: Springer US, 1994, pp. 319–338. DOI: 10.1007/978-1-4613-3632-7_16.

[116] S. G. Nash and A. Sofer, "A barrier method for large-scale constrained optimization", *ORSA Journal on Computing*, vol. 5, no. 1, pp. 40–53, 1993. DOI: 10.1287/ijoc.5.1.40.

[117] J.-C. Nédélec, "Mixed finite elements in $\mathbb{R}^3$", *Numerische Mathematik*, vol. 35, no. 3, pp. 315–341, 1980. DOI: 10.1007/BF01396415.

[118] L. H. Olesen, F. Okkels, and H. Bruus, "A high-level programming-language implementation of topology optimization applied to steady-state Navier–Stokes flow", *International Journal for Numerical Methods in Engineering*, vol. 65, no. 7, pp. 975–1001, 2006. DOI: 10.1002/nme.1468.

[119] W. S. Ożański, "The Lagrange multiplier and the stationary Stokes equations", *Journal of Applied Analysis*, vol. 23, no. 2, pp. 137–140, 2017. DOI: 10.1515/jaa-2017-0017.

[120] I. P. A. Papadopoulos, "Numerical analysis of a discontinuous Galerkin method for the Borrvall–Petersson topology optimization problem", *arXiv preprint arXiv:2108.03930*, 2021.

[121] I. P. A. Papadopoulos and P. E. Farrell, "Preconditioners for computing multiple solutions in three-dimensional fluid topology optimization", *in preparation*, 2021.

[122] I. P. A. Papadopoulos, P. E. Farrell, and T. M. Surowiec, "Computing multiple solutions of topology optimization problems", *SIAM Journal on Scientific Computing*, vol. 43, no. 3, A1555–A1582, 2021. DOI: 10.1137/20M1326209.

[123] I. P. A. Papadopoulos and E. Süli, "Numerical analysis of a topology optimization problem for Stokes flow", *arXiv preprint arXiv:2102.10408*, 2021.

[124] J. Petersson, "A finite element analysis of optimal variable thickness sheets", *SIAM Journal on Numerical Analysis*, vol. 36, no. 6, pp. 1759–1778, 1999. DOI: 10.1137/S0036142996313968.

[125] ——, "Some convergence results in perimeter-controlled topology optimization", *Computer Methods in Applied Mechanics and Engineering*, vol. 171, no. 1-2, pp. 123–140, 1999. DOI: 10.1016/S0045-7825(98)00248-5.

[126] J. Petersson and J. Haslinger, "An approximation theory for optimum sheets in unilateral contact", *Quarterly of Applied Mathematics*, pp. 309–325, 1998.

[127] J. Petersson and O. Sigmund, "Slope constrained topology optimization", *International Journal for Numerical Methods in Engineering*, vol. 41, no. 8, pp. 1417–1434, 1998. DOI: 10.1002/(SICI)1097-0207(19980430)41:8<1417::AID-NME344>3.0.CO;2-N.

[128] G. Pingen, A. Evgrafov, and K. Maute, "Topology optimization of flow domains using the lattice Boltzmann method", *Structural and Multidisciplinary Optimization*, vol. 34, no. 6, pp. 507–524, 2007. DOI: 10.1007/s00158-007-0105-7.

[129] P. Popov, "Preconditioning of linear systems arising in finite element discretizations of the Brinkman equation", in *Large-Scale Scientific Computing. LSSC 2011. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2012, pp. 381–389. DOI: 10.1007/978-3-642-29843-1_43.

[130] L. Qi, "Convergence analysis of some algorithms for solving nonsmooth equations", *Mathematics of Operations Research*, vol. 18, no. 1, pp. 227–244, 1993. DOI: 10.1287/moor.18.1.227.

[131] L. Qi and J. Sun, "A nonsmooth version of Newton's method", *Mathematical Programming*, vol. 58, no. 1-3, pp. 353–367, 1993. DOI: 10.1007/BF01581275.

[132] J. Qin, "On the convergence of some low order mixed finite elements for incompressible fluids", PhD thesis, Pennsylvania State University, 1994.

[133] D. Ramalingom, P.-H. Cocquet, and A. Bastide, "A new interpolation technique to deal with fluid-porous media interfaces for topology optimization of heat transfer", *Computers & Fluids*, vol. 168, pp. 144–158, 2018. DOI: 10.1016/j.compfluid.2018.04.005.

[134] M. Rasmussen and M. Stolpe, "Global optimization of discrete truss topology design problems using a parallel cut-and-branch method", *Computers & Structures*, vol. 86, no. 13-14, pp. 1527–1538, 2008. DOI: 10.1016/j.compstruc.2007.05.019.

[135] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. Kelly, "Firedrake: automating the finite element method by composing abstractions", *ACM Transactions on Mathematical Software*, vol. 43, no. 3, pp. 1–27, 2016. DOI: 10.1145/2998441.

[136] P.-A. Raviart and J.-M. Thomas, "A mixed finite element method for 2nd order elliptic problems", in *Mathematical Aspects of Finite Element Methods*, Springer, 1977, pp. 292–315. DOI: 10.1007/BFb0064470.

[137] H. Rezayat, J. R. Bell, A. J. Plotkowski, and S. S. Babu, "Multi-solution nature of topology optimization and its application in design for additive manufacturing", *Rapid Prototyping Journal*, 2019. DOI: 10.1108/RPJ-01-2018-0009.

[138] M. Robinson, C. Luo, P. E. Farrell, R. Erban, and A. Majumdar, "From molecular to continuum modelling of bistable liquid crystal devices", *Liquid Crystals*, vol. 44, no. 14-15, pp. 2267–2284, 2017. DOI: 10.1080/02678292.2017.1290284.

[139] S. Rojas-Labanda and M. Stolpe, "Benchmarking optimization solvers for structural topology optimization", *Structural and Multidisciplinary Optimization*, vol. 52, no. 3, pp. 527–547, 2015. DOI: 10.1007/s00158-015-1250-z.

[140] G. Rozvany and M. Zhou, "The COC algorithm, part I: Cross-section optimization or sizing", *Computer Methods in Applied Mechanics and Engineering*, vol. 89, no. 1-3, pp. 281–308, 1991. DOI: 10.1016/0045-7825(91)90045-8.

[141] L. F. N. Sá, R. C. R. Amigo, A. A. Novotny, and E. C. N. Silva, "Topological derivatives applied to fluid flow channel design optimization problems", *Structural and Multidisciplinary Optimization*, vol. 54, no. 2, pp. 249–264, 2016. DOI: 10.1007/s00158-016-1399-0.

[142] Y. Saad, "A flexible inner-outer preconditioned GMRES algorithm", *SIAM Journal on Scientific Computing*, vol. 14, no. 2, pp. 461–469, 1993. DOI: 10.1137/0914028.

[143] J. Schöberl, "Multigrid methods for a parameter dependent problem in primal variables", *Numerische Mathematik*, vol. 84, no. 1, pp. 97–119, 1999. DOI: 10.1007/s002110050465.

[144] T. Schwedes, D. A. Ham, S. W. Funke, and M. D. Piggott, *Mesh dependence in PDE-constrained optimisation*. Springer, 2017, pp. 53–78, ISBN: 978-3-319-59482-8. DOI: 10.1007/978-3-319-59483-5_2.

[145] L. R. Scott and M. Vogelius, "Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials", *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 19, no. 1, pp. 111–143, 1985. DOI: 10.1051/m2an/1985190101111.

[146] J. A. Sethian and A. Wiegmann, "Structural boundary design via level set and immersed interface methods", *Journal of Computational Physics*, vol. 163, no. 2, pp. 489–528, 2000. DOI: 10.1006/jcph.2000.6581.

[147] R. Seydel, *Practical Bifurcation and Stability Analysis*, 3rd ed., ser. Interdisciplinary Applied Mathematics. New York, NY: Springer New York, 2010, vol. 5, p. 476, ISBN: 978-1-4419-1739-3. DOI: 10.1007/978-1-4419-1740-9.

[148] P. Y. Shim and S. Manoochehri, "Generating optimal configurations in structural design using simulated annealing", *International Journal for Numerical Methods in Engineering*, vol. 40, no. 6, pp. 1053–1069, 1997. DOI: 10.1002/(SICI)1097-0207(19970330)40:6<1053::AID-NME97>3.0.CO;2-I.

[149] O. Sigmund, "Design of material structures using topology optimization", PhD thesis, Technical University of Denmark, 1994.

[150] ——, "A 99 line topology optimization code written in Matlab", *Structural and Multidisciplinary Optimization*, vol. 21, no. 2, pp. 120–127, 2001. DOI: 10.1007/s001580050176.

[151] ——, "On the usefulness of non-gradient approaches in topology optimization", *Structural and Multidisciplinary Optimization*, vol. 43, no. 5, pp. 589–596, 2011. DOI: 10.1007/s00158-011-0638-7.

[152] O. Sigmund and K. Maute, "Topology optimization approaches", *Structural and Multidisciplinary Optimization*, vol. 48, no. 6, pp. 1031–1055, 2013. DOI: 10.1007/s00158-013-0978-6.

[153] O. Sigmund and J. Petersson, "Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima", *Structural Optimization*, vol. 16, no. 1, pp. 68–75, 1998. DOI: 10.1007/BF01214002.

[154] *Software used in Chapter 5 of the PhD Thesis 'Computing multiple solutions of topology optimization problems' of Ioannis Papadopoulos*, 2021. DOI: 10.5281/zenodo.5175642.

[155] *Software used in 'Computing multiple solutions of topology optimization problems'*, 2020. DOI: 10.5281/zenodo.3710963.

[156] *Software used in 'Numerical analysis of a discontinuous Galerkin method for the Borrvall–Petersson topology optimization problem'*, 2021. DOI: 10.5281/zenodo.5146324.

[157] *Software used in 'Numerical analysis of a topology optimization problem for Stokes flow'*, 2021. DOI: 10.5281/zenodo.4514054.

[158] M. Stoll and A. J. Wathen, "Preconditioning for active set and projected gradient methods as semi-smooth Newton methods for PDE-constrained optimization with control constraints", 2009.

[159] M. Stolpe and K. Svanberg, "On the trajectories of penalization methods for topology optimization", *Structural and Multidisciplinary Optimization*, vol. 21, no. 2, pp. 128–139, 2001. DOI: 10.1007/s001580050177.

[160] M. Stolpe and K. Svanberg, "An alternative interpolation scheme for minimum compliance topology optimization", *Structural and Multidisciplinary Optimization*, vol. 22, no. 2, pp. 116–124, 2001. DOI: 10.1007/s001580100129.

[161] A. Stück and T. Rung, "Adjoint complement to viscous finite-volume pressure-correction methods", *Journal of Computational Physics*, vol. 248, pp. 402–419, 2013. DOI: 10.1016/j.jcp.2013.01.002.

[162] K. Svanberg, "The method of moving asymptotes - a new method for structural optimization", *International Journal for Numerical Methods in Engineering*, vol. 24, no. 2, pp. 359–373, 1987. DOI: 10.1002/nme.1620240207.

[163] ——, "A class of globally convergent optimization methods based on conservative convex separable approximations", *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 555–573, 2002. DOI: 10.1137/S1052623499362822.

[164] C. Talischi and G. H. Paulino, "A closer look at consistent operator splitting and its extensions for topology optimization", *Computer Methods in Applied Mechanics and Engineering*, vol. 283, pp. 573–598, 2015. DOI: 10.1016/j.cma.2014.07.005.

[165] F. Treves, *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics, Vol. 25*. Academic Press, Inc, 1967, vol. 25, ISBN: 978-0-12-699450-6.

[166] M. Ulbrich, "Semismooth Newton methods for operator equations in function spaces", *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 805–841, 2003. DOI: 10.1137/s1052623400371569.

[167] M. Ulbrich and S. Ulbrich, "Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds", *SIAM Journal of Control and Optimization*, vol. 38, pp. 1938–1984, 2000. DOI: 10.1137/S0363012997325915.

[168] ——, "Primal-dual interior-point methods for PDE-constrained optimization", *Mathematical Programming*, vol. 117, no. 1-2, pp. 435–485, 2009. DOI: 10.1007/s10107-007-0168-7.

[169] A. Vadakkepatt, S. R. Mathur, and J. Y. Murthy, "Efficient automatic discrete adjoint sensitivity computation for topology optimization–heat conduction applications", *International Journal of Numerical Methods for Heat & Fluid Flow*, 2018. DOI: 10.1108/HFF-01-2017-0011.

[170] A. Wächter and L. T. Biegler, "On the implementation of primal-dual interior point filter line search algorithm for large-scale nonlinear programming", *Mathematical Programming*, vol. 106, pp. 25–57, 2006. DOI: 10.1007/s10107-004-0559-y.

[171] S. Y. Wang and K Tai, "Structural topology design optimization using genetic algorithms with a bit-array representation", *Computer Methods in Applied Mechanics and Engineering*, vol. 194, no. 36-38, pp. 3749–3770, 2005. DOI: 10.1016/j.cma.2004.09.003.

[172] X. Wang, S. Xu, S. Zhou, W. Xu, M. Leary, P. Choong, M. Qian, M. Brandt, and Y. M. Xie, "Topological design and additive manufacturing of porous metals for bone scaffolds and orthopaedic implants: A review", *Biomaterials*, vol. 83, pp. 127–141, 2016. DOI: 10.1016/j.biomaterials.2016.01.012.

[173] X. Wang, Y. Mei, and M. Wang, "Incorporating topological derivatives into level set methods for structural topology optimization", in *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, 2004, p. 4564. DOI: 10.2514/6.2004-4564.

[174] A. J. Wathen, "Preconditioning", *Acta Numerica*, vol. 24, 2015. DOI: 10.1017/S0962492915000021.

[175] J. H. Wilkinson, *Rounding errors in algebraic processes*. Dover Publications, Inc., New York, 1994, pp. viii+161, Reprint of the 1963 original [Prentice-Hall, Englewood Cliffs, NJ], ISBN: 0-486-67999-3.

[176] C.-Y. Wu and K.-Y. Tseng, "Topology optimization of structures using modified binary differential evolution", *Structural and Multidisciplinary Optimization*, vol. 42, no. 6, pp. 939–953, 2010. DOI: 10.1007/s00158-010-0523-9.

[177] J. Xia, P. E. Farrell, and S. G. P. Castro, "Nonlinear bifurcation analysis of stiffener profiles via deflation techniques", *Thin Walled Structures*, vol. 149, p. 106 662, 2020. DOI: 10.1016/j.tws.2020.106662.

[178] J. Xia, P. E. Farrell, and F. Wechsung, "Augmented Lagrangian preconditioners for the Oseen–Frank model of nematic and cholesteric liquid crystals", *BIT Numerical Mathematics*, vol. 61, no. 2, pp. 607–644, 2021. DOI: 10.1007/s10543-020-00838-9.

[179] Y. M. Xie and G. P. Steven, "Basic evolutionary structural optimization", in *Evolutionary Structural Optimization*, Springer, 1997, pp. 12–29. DOI: 10.1007/978-1-4471-0985-3_2.

[180] Y. Xie and G. Steven, "A simple evolutionary procedure for structural optimization", *Computers and Structures*, vol. 49, no. 5, pp. 885–896, 1993. DOI: 10.1016/0045-7949(93)90035-C.

[181] J. Xu, "Iterative methods by space decomposition and subspace correction", *SIAM Review*, vol. 34, no. 4, pp. 581–613, 1992. DOI: 10.1137/1034116.

[182] ——, "The method of subspace corrections", *Journal of Computational and Applied Mathematics*, vol. 128, no. 1-2, pp. 335–362, 2001. DOI: 10.1016/S0377-0427(00)00518-5.

[183] K. Yaji, T. Yamada, M. Yoshino, T. Matsumoto, K. Izui, and S. Nishiwaki, "Topology optimization using the lattice Boltzmann method incorporating level set boundary expressions", *Journal of Computational Physics*, vol. 274, pp. 158–181, 2014. DOI: 10.1016/j.jcp.2014.06.004.

[184] K. Yonekura and Y. Kanno, "Global optimization of robust truss topology via mixed integer semidefinite programming", *Optimization and Engineering*, vol. 11, no. 3, pp. 355–379, 2010. DOI: 10.1007/s11081-010-9107-1.

[185] S. Zhang and J. A. Norato, "Finding better local optima in topology optimization via tunneling", in *Volume 2B: 44th Design Automation Conference*, vol. 17, American Society of Mechanical Engineers, 2018, p. 103, ISBN: 978-0-7918-5176-0. DOI: 10.1115/DETC2018-86116.

[186] S. Zhang, "Divergence-free finite elements on tetrahedral grids for $k \geq 6$", *Mathematics of Computation*, vol. 80, no. 274, pp. 669–695, 2011. DOI: 10.1090/S0025-5718-2010-02412-3.

[187] H. Zhou, "Topology optimization of compliant mechanisms using hybrid discretization model", *Journal of Mechanical Design*, 2010. DOI: 10.1115/1.4002663.

[188] M. Zhou and G. Rozvany, "The COC algorithm, Part II: Topological, geometrical and generalized shape optimization", *Computer Methods in Applied Mechanics and Engineering*, vol. 89, no. 1-3, pp. 309–336, 1991. DOI: 10.1016/0045-7825(91)90046-9.

[189] S. Zhou and Q. Li, "A variational level set method for the topology optimization of steady-state Navier–Stokes flow", *Journal of Computational Physics*, vol. 227, no. 24, pp. 10 178–10 195, 2008. DOI: 10.1016/j.jcp.2008.08.022.