TOPOLOGY OPTIMIZATION VIA DENSITY BASED APPROACHES

J. HAUBNER*, F. NEUMANN†, AND M. ULBRICH‡

Key words. Topology optimization, phase field, Stokes flow

AMS subject classifications. 35R30, 49K20, 49Q10, 65K10

Abstract. A new method for performing density based topology optimization for Stokes flow is presented, which differs from previous approaches in the way the underlying mixed integer problem is relaxed. It is theoretically justified by a thorough theoretical investigation regarding existence of solutions and differentiability. Based on these results a numerical realization is presented which applies an H^s-regularization for the control.

1. Introduction.

Shape and topology optimization denotes a family of optimization problems aiming to find the optimal shape with respect to a given objective function on a set of admissible shapes $\mathcal{O}_{ad}$. Shape optimization problems are given by

$$\min_{\Omega \in \mathcal{O}_{ad}} \tilde{j}(\Omega),$$

where $\tilde{j}: \mathcal{O}_{ad} \to \mathbb{R}$ denotes a shape functional [26, Def. 4.3.1] and $\mathcal{O}_{ad}$ denotes a set of admissible shapes. There are many different applications in fluid mechanics and structural optimization such as weight reduction or airplane optimization, see [44, 13]. In [1], shape optimization is utilized in a biomedical engineering setting to analyze blood flow.

In order to have well-definedness in a classical sense, and to develop optimization theory and methods, a metric structure has to be imposed. This can be realized in various ways leading to different concepts. One possibility is given via transformations, which leads to the concept of shape derivatives, e.g. [49], and the method of mappings, e.g. [45]. Another way is the imposition of a metric via characteristic functions, or, equivalently, via functions that attain values in $\{-1, 1\}$, on a domain $D \subset \mathbb{R}^d$. The latter, which is the focus of this work, is abstractly given by

$$\min_{\chi} j(\chi), \quad \text{s.t.} \quad g(\chi) \leq 0, \ \chi \in \{-1, 1\} \text{ a.e. },$$

where g represents constraints, e.g. the volume constraints, and $j(\chi_\Omega) := \tilde{j}(\Omega)$ for every characteristic function χ_Ω, $\Omega \in \mathcal{O}_{ad}$, defined by $\chi_\Omega = \begin{cases} 1 & \text{for } x \in \Omega, \\ -1 & \text{for } x \in D \setminus \Omega. \end{cases}$
It naturally allows for shapes with different topologies and is, therefore, referred to as topology optimization. However, due to its infinite dimensional mixed integer nature, this optimization problem is hard to solve. On that account, different relaxation techniques were introduced to handle this problem. In this paper, we examine a relaxation that is different from existing approaches.

While topology optimization was initially introduced and studied for structure mechanical problems (e.g. [12, 50, 4]), [15] was the pioneering work in applying

*Department of Numerical Analysis and Scientific Computing, Simula Research Laboratory, Lysaker, Norway (haubnerj@simula.no)

†Institute of Automotive Technology, Technical University of Munich, Boltzmannstr. 15, 85748 Garching b. München, Germany (neumann@ftm.mw.tum.de)

‡Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching b. München, Germany (mulbrich@ma.tum.de)

this method to the fluid mechanical setting based on the Stokes equations. The theoretical analysis was complemented [31] and extended to the steady state Navier-Stokes equations [32, 30]. For a survey on applications of topology optimization in fluid mechanical problems, see, e.g., [8, 44]. Here, we consider topology optimization for the Stokes problem, using the setting proposed in [15].

Problems in shape and topology optimization are highly complex and have to be treated carefully. Numerical methods typically rely on relaxation techniques [15, 33], however, one still has to deal with the nonlinear nature of the problems which typically leads to many local minima. The approach in [15] is restricted to a specific objective function. [33] can deal with more general objective functions and is based on the reformulation of the $\{-1, 1\}$-constraint

$$\min_{\rho} j(\rho), \quad \text{s.t.} \;\; g(\rho) \leq 0, \; \int_D (\rho^2 - 1) d\xi = 0, \; -1 \leq \rho \leq 1 \text{ a.e.}$$

and the relaxation

$$\min_{\rho \in Y} j(\rho) + \gamma \int_D (\rho^2 - 1) d\xi + \frac{\eta}{2} \|\rho\|_Y^2, \quad \text{s.t.} \;\; g(\rho) \leq 0, \; -1 \leq \rho \leq 1 \text{ a.e.}$$

for $Y = H^1(D)$. Here, the sphere constraint $\int_D (\rho^2 - 1) d\xi = 0$ is penalized in the objective function value and intermediate values between -1 and 1 are allowed. Since an interfacial layer is included, it is called phase field approach. For numerical investigations, in [34] the problem is further relaxed by penalizing the constraint $-1 \leq \rho \leq 1$ using

$$\begin{aligned} \Upsilon(\rho) :&= \frac{1}{2} \| \max(0, \rho - 1) \|_{L^2(D)}^2 + \frac{1}{2} \| \min(0, \rho + 1) \|_{L^2(D)}^2 \\ &= \frac{1}{2} \| \max(0, \rho - 1)^2 \|_{L^1(D)} + \frac{1}{2} \| \max(0, -\rho - 1)^2 \|_{L^1(D)}. \end{aligned}$$

Moreover, this approach was also applied for problems that are governed by the steady Navier-Stokes flow, see, e.g., [36, 35].

A similar formulation that—to the best of the authors' knowledge—has not been investigated so far and is worth examination is given by

$$\text{(1.1)} \qquad \min_{\rho \in Y} \bar{j}(\rho) := j(\rho) + \gamma \Upsilon_p(\rho) + \frac{\eta}{2} \|\rho\|_Y^2, \quad \text{s.t.} \;\; g(\rho) \leq 0, \; \int_D (|\rho|^q - 1) d\xi = 0$$

for $p, q > 1$. In contrast to previous approaches, the sphere constraint and the penalization

$$\text{(1.2)} \qquad \Upsilon_p(\rho) := \frac{1}{p} \| \max(0, \rho - 1)^p \|_{L^1(D)} + \frac{1}{p} \| \max(0, -\rho - 1)^p \|_{L^1(D)}.$$

are generalized (basically in order to be able to work with BV-spaces). In addition, the sphere constraint is kept as an equality constraint. In this paper, we consider (1.1) for minimizing the total potential power in the Stokes flow on a Lipschitz domain $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$ with outer unit normal n. More precisely, we investigate the PDE constrained optimization problem

$$\text{(1.3)} \qquad j(\rho) := J(\rho, S(\rho)),$$

with $S:\rho\mapsto u$ being the solution operator of the generalized Stokes equations, compare [15],

$$\begin{aligned}\alpha(\rho)u-\mu\Delta u+\nabla p &= f \quad \text{on } D,\\ \operatorname{div}(u) &= 0 \quad \text{on } D,\\ u &= u_D \quad \text{on } \partial D,\end{aligned} \tag{1.4}$$

where $f\in H^{-1}(D)^d$ denotes a source term and $u_D\in H^{\frac{1}{2}}(\partial D)^d$ denotes Dirichlet boundary conditions. Moreover, $\tilde{a}:\mathbb{R}\to\mathbb{R}$ is chosen such that $\tilde{a}\geq 0$, $\tilde{a}(x)=0$ for $x\geq 1$ and $\tilde{a}(x)\gg 1$ for $x\leq -1$, and the Nemytskii operator α is defined by

$$\alpha:\rho\mapsto\alpha(\rho),\ \alpha(\rho)(\xi):=\tilde{a}(\rho(\xi)) \tag{1.5}$$

for a.e. $\xi\in D$. Hence, where $\rho(\xi)\geq 1$, the standard Stokes equations are solved, whereas for $\rho(\xi)\leq -1$ the $\alpha(\rho)u$ term dominates and forces u to be small. The cost functional that we consider is the total potential power function defined by

$$J(\rho,u):=\frac{1}{2}(\alpha(\rho)u,u)_D+\frac{\mu}{2}(\nabla u,\nabla u)_D-(f,u)_D, \tag{1.6}$$

compare [15]. Here, $(u,v)_D:=\int_D u\cdot v d\xi$ denotes the L^2-inner product on D. We will pose the volume constraint

$$g(\rho):=\frac{1}{2}\int_D(\rho(\xi)+1)d\xi-V\leq 0, \tag{1.7}$$

which upper bounds the volume of the fluid domain by a constant $V>0$.

In Section 2 we consider the solution operator for the generalized Stokes equations. We extend the results in [34] to less restrictive choices of α and prove a differentiability result. Section 3 presents a continuity and differentiability result for superposition operators that will be used to show differentiability of Υ_p, and later also for showing a differentiability result for α. These results are used to show existence of solutions and, in Section 4, differentiability of the reduced objective under assumptions on the Banach space Y. Section 5 considers the limit behavior for increasing penalization parameter γ. Section 6 motivates different settings, that fulfill all requirements that are needed for the theoretical analysis. In Section 7 we discuss the numerical realization. Section 8 presents the results.

2. On the solution operator for the generalized Stokes equations. Let $d\in\{2,3\}$, X be a Banach space and

$$\begin{aligned}U:=&\{u\in H^1(D)^d\ :\ u=u_D \text{ on } \partial D,\ \operatorname{div}(u)=0\},\\ V:=&\{v\in H_0^1(D)^d\ :\ \operatorname{div}(v)=0\}.\end{aligned}$$

The weak formulation of (1.4) is given by: find $u\in U$ such that

$$E(\rho,u)(v):=(\alpha(\rho)u,v)_D+\mu(\nabla u,\nabla v)_D=\langle f,v\rangle_{H^{-1}(D)^d,H^1(D)^d} \tag{2.1}$$

for all $v\in V$, see e.g. [39, Remark 5.1]. In this section, we show well-definedness (Lemma 2.1), continuity (Lemma 2.2) and Fréchet differentiability (Lemma 2.3) of the solution operator $S:X\to U\subset H^1(D)^d,\rho\mapsto u$ of (2.1) under general assumptions on the superposition operator α.

LEMMA 2.1 (Well-definedness of the solution operator). *Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u_D \in H^{\frac{1}{2}}(\partial D)^d$ with $\int_{\partial D} u_D \cdot n ds = 0$ and $f \in H^{-1}(D)^d$. Moreover, assume that for every $\rho \in X$, $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is bounded on an open neighborhood around ρ, i.e. there exists an open subset $\tilde{X} \subset X$ with $\rho \in \tilde{X}$, and a constant $C > 0$ depending on ρ such that $\|\alpha(\tilde{\rho})\|_{L^s(D)} \leq C$ for all $\tilde{\rho} \in \tilde{X}$. Then, for every $\tilde{\rho} \in \tilde{X}$, there exists a unique $u = u(\tilde{\rho}) \in U$ such that (1.4) is fulfilled and a constant $c > 0$ (that depends on ρ) such that*

$$\|u\|_{H^1(D)^d} \leq c(\|f\|_{H^{-1}(D)^d} + \|u_D\|_{H^{\frac{1}{2}}(\partial D)^d}).$$

Proof. The proof is based on [39, Lemma 5.1] and Lax–Milgram's theorem. First, we reduce the variational equation (2.1) to a homogenous problem. By [39, Lemma 4.1], there exists a continuous extension operator ext : $\{\tilde{g} \in H^{\frac{1}{2}}(\partial D)^d : \int_{\partial D} \tilde{g} \cdot n ds = 0\} \to \{u \in H^1(D)^d : \operatorname{div}(u) = 0\}$, $\tilde{g} \mapsto \operatorname{ext}(\tilde{g})$ such that $\operatorname{ext}(\tilde{g})|_{\partial D} = \tilde{g}$. Let $w := \operatorname{ext}(u_D)$, i.e., there exists a constant C such that $\|w\|_{H^1(D)^d} \leq C\|u_D\|_{H^{\frac{1}{2}}(\partial D)^d}$ and $w|_{\partial D} = u_D$. Hence, $u \in U$ solves (2.1) if and only if $u_0 := u - w \in V$ solves

$$\begin{aligned} a(u_0, v) :&= (\alpha(\rho)u_0, v)_D + \mu(\nabla u_0, \nabla v)_D \\ &= \langle f, v\rangle_{H^{-1}(D)^d, H^1(D)^d} - (\alpha(\rho)w, v)_D - \mu(\nabla w, \nabla v)_D \\ &=: \langle \bar{f}, v\rangle_{H^{-1}(D)^d, H^1(D)^d}. \end{aligned} \tag{2.2}$$

Let $\rho \in X$ and $\tilde{\rho} \in \tilde{X} = \tilde{X}(\rho)$. Since $\alpha(\tilde{\rho}) \geq 0$, with Poincaré's inequality, we obtain coercivity of the bilinear form $a : V \times V \to \mathbb{R}$. By $H^1(D) \hookrightarrow L^{\frac{2s}{s-1}}(D)$, the assumptions on α, and Hölder's inequality there exists a constant $C > 0$ such that

$$(\alpha(\tilde{\rho})u, v)_D \leq C\|\alpha(\rho)\|_{L^s(D)}\|u\|_{H^1(D)^d}\|v\|_{H^1(D)^d}. \tag{2.3}$$

The properties of α yield a constant C depending on ρ such that

$$(\alpha(\tilde{\rho})u, v)_D \leq C\|u\|_{H^1(D)^d}.\|v\|_{H^1(D)^d} \tag{2.4}$$

This implies continuity of $a : V \times V \to \mathbb{R}$, and in combination with continuity of ext

$$\begin{aligned} \|\bar{f}\|_{H^{-1}(D)^d} &\leq C(\|f\|_{H^{-1}(D)^d} + \|\alpha(\tilde{\rho})\|_{L^s(D)}\|w\|_{H^1(D)^d} + \|w\|_{H^1(D)^d}) \\ &\leq C(\|f\|_{H^{-1}(D)^d} + \|u_D\|_{H^{\frac{1}{2}}(D)^d}) \end{aligned} \tag{2.5}$$

with a generic constant $C > 0$ depending on ρ. Applying Lax–Milgram's theorem yields a unique solution $u_0 \in V$ that fulfills (2.2) and there exists a constant C such that

$$\|u_0\|_{H^1(D)^d} \leq C(\|f\|_{H^{-1}(D)^d} + \|u_D\|_{H^{\frac{1}{2}}(D)^d}). \tag{2.6}$$

Hence, $u = u_0 + w$ is a solution of (2.1) and, with (2.6), continuity of ext and the triangle inequality, there exists a constant $C > 0$ depending on ρ such that

$$\begin{aligned} \|u\|_{H^1(D)^d} = \|u_0 + w\|_{H^1(D)^d} &\leq \|u_0\|_{H^1(D)^d} + \|w\|_{H^1(D)^d} \\ &\leq C(\|f\|_{H^{-1}(D)^d} + \|u_D\|_{H^{\frac{1}{2}}(D)^d}). \end{aligned}$$ ∎

Lemma 2.1 gives bijectivity of $E(\rho, u)$ as a mapping $X \times H^1(D)^d \to H^{-1}(D)^d$ and thus the well-definedness of the solution operator $S:\ X \to H^1(D)^d,\ \rho \mapsto u$, where (u, p) is the solution to the partial differential equation (1.4).

Lemma 2.2 (Continuity of the solution operator). *Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u_D \in H^{\frac{1}{2}}(\partial D)^d$ with $\int_{\partial D} u_D \cdot n ds = 0$ and $f \in H^{-1}(D)^d$. Moreover, assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous. Then, $S:\ X \to H^1(D)^d,\ \rho \mapsto u$ is continuous.*

Proof. Let $\rho_1, \rho_2 \in X$. By Lemma 2.1 we know that there exist unique $u_1, u_2 \in U$ such that

$$(\alpha(\rho_1)u_1, v)_D + \mu(\nabla u_1, \nabla v)_D = (f, v)_D,$$
$$(\alpha(\rho_2)u_2, v)_D + \mu(\nabla u_2, \nabla v)_D = (f, v)_D,$$

for all $v \in V$. Substracting the two equations gives

$$(\alpha(\rho_1)(u_2 - u_1), v)_D + \mu(\nabla(u_2 - u_1), \nabla v)_D = -((\alpha(\rho_2) - \alpha(\rho_1))u_2, v)_D. \tag{2.7}$$

Testing with $v = u_2 - u_1$, using $\alpha(\rho_1) \geq 0$, the Poincaré inequality, and (2.3) yields

$$\|u_2 - u_1\|_{H^1(D)^d} \leq C\|\alpha(\rho_2) - \alpha(\rho_1)\|_{L^s(D)}\|u_2\|_{H^1(D)^d} \tag{2.8}$$

for a constant $C > 0$. Continuity of α implies boundedness on an open neighborhood around ρ_1. Hence, by Lemma 2.1, there exists a constant $C_{\rho_1} > 0$ and $\delta > 0$ such that $\|u_2\|_{H^1(D)^d} \leq C_{\rho_1}$ for all $\rho_2 \in B_\delta(\rho_1)$. Thus (2.8) and the continuity of α yield continuity of S. ◻

Lemma 2.3 (Fréchet differentiability of the solution operator). *Let $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u_D \in H^{\frac{1}{2}}(\partial D)^d$ with $\int_{\partial D} u_D \cdot n ds = 0$ and $f \in H^{-1}(D)^d$. Moreover, assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuously differentiable. Let $\rho_0 \in X$. Then, $S:\ X \to H^1(D)^d$ is Fréchet differentiable in an open neighborhood of ρ_0.*

Proof. By Lemma 2.1, for $u_0 = S(\rho_0)$ it holds that $E(\rho_0, u_0) = 0$. Using Hölder's inequality it can be verified that $(w, u) \mapsto w \cdot u$ is continuously differentiable as a mapping $L^s(D) \times H^1(D)^d \to L^r(D)^d$, with $r = \frac{2s}{s+1}$, and since $H^1(D)^d \hookrightarrow L^{\frac{1}{1-\frac{1}{r}}}(D)^d$ it is also Fréchet differentiable as a mapping $L^s(D) \times H^1(D)^d \to H^{-1}(D)^d$. Linearity of $u \mapsto \nabla u$ as a mapping $H^1(D)^d \to L^2(D)^{d\times d}$, continuous differentiability of α and the chain rule, therefore, yield continuous differentiability of $(\rho, u) \to E(\rho, u)$ as a mapping $X \times H^1(D)^d \to H^{-1}(D)^d$. By linearity of $u \mapsto E(\rho, u)$ and Lemma 2.1, $E_u(\rho_0, u_0) \in \mathcal{L}(U, H^{-1}(D)^d)$ is bijective, i.e. continuously invertible. The implicit function theorem thus yields Fréchet differentiability of S in an open neighborhood of ρ_0. ◻

3. Existence of solutions of the relaxed problem. This section gives an existence result for the relaxed problem if Y is reflexive (Theorem 3.9) or $Y = BV(D)$ (Theorem 3.10). For deriving these results, continuity results for j (Lemma 3.6, Lemma 3.7, Lemma 3.8) and Υ_p (Corollary 3.2) are needed. We also show differentiability (see Section 4) in the following lemma, from which continuity and differentiability of Υ_p follow (Corollary 3.2).

Lemma 3.1. *Let $D \subset \mathbb{R}^d$ be a measurable, bounded subset of $\mathbb{R}^d$, $p \geq r \geq 1$ and $z \in \{-1, 1\}$. Then the mapping $h : \mathbb{R} \to \mathbb{R}, h(x) := \frac{1}{r}\max(0, zx - 1)^r$ is convex,* *non-negative and continuous. Furthermore, the associated superposition operator*

$$T_h : L^p(D) \to L^1(D),\ T_h(\rho)(\xi) = \frac{1}{r}\max(0, z\rho(\xi) - 1)^r \tag{3.1}$$

is convex and continuous. Assume that $p \geq r > 1$, then h is continuously differentiable and T_h defined in (3.1) is Fréchet differentiable with derivative

$$T_h'(\rho) \in \mathcal{L}(L^p(D), L^1(D)),\ [T_h'(\rho)w](\xi) = z\max(0, z\rho(\xi) - 1)^{r-1}w(\xi).$$

Moreover, T_h is continuously differentiable.

Proof. Let $h_1(\cdot) := z \cdot -1$ and $h_2(\cdot) := \max(0, \cdot)^r$. Then the mapping

$$x \mapsto h(x) = h_2(h_1(x))$$

for all $x \in \mathbb{R}$ is convex, since h_1 is affine linear and h_2 is convex for $r \geq 1$. Moreover, it is continuous and, for $r > 1$, continuously differentiable with $h'(x) = z\max(0, zx - 1)^{r-1}$. Convexity of T_h is inherited from the convexity of h. Since $|T_h(\rho)(\xi)| \leq \frac{|z|}{r}|\rho(\xi)|^r$, [51, Section 4.3.3] implies continuity of $T_h : L^p(D) \to L^1(D)$ for $p = r$, and, since D is bounded, for $p \geq r$. The superposition operator $T_{h'}$ associated to h' and given by

$$T_{h'}(\rho)(\xi) := z\max(0, z\rho(\xi) - 1)^{r-1}$$

fulfills the growth condition

$$|T_{h'}(\rho)(\xi)| \leq C(1 + |\rho(\xi)|^{r-1}) \tag{3.2}$$

for a constant $C > 0$. Hence, $T_{h'} : L^p(D) \to L^s(D)$ with $1 \leq s \leq \frac{p}{r-1}$ and, therefore, in particular, for $s = \frac{p}{p-1}$ since $p \geq r > 1$. By [51, Section 4.3.3], this implies Fréchet differentiability of $T_h : L^p(D) \to L^1(D)$ with derivative

$$T_h'(\rho) \in \mathcal{L}(L^p(D), L^1(D)),\ [T_h'(\rho)w](\xi) = T_{h'}(\rho)(\xi)w(\xi).$$

Since h' is continuous and, therefore, fulfills the Carathéodory condition, and the growth condition (3.2) holds, $T_{h'} : L^p(D) \to L^{\frac{p}{p-1}}(D)$ is well-defined and thus continuous by [51, Section 4.3.3], [10, Theorem 3.1]. Hence, T_h is continuously (Fréchet) differentiable. ∎

With this lemma continuity and differentiability of Υ_p follow directly.

COROLLARY 3.2. *Let $D \subset \mathbb{R}^d$ be a measurable, bounded subset of $\mathbb{R}^d$, and $p > 1$. Then the mapping Υ_p is convex, continuous as a mapping $L^p(D) \to \mathbb{R}$ and continuously differentiable with derivative*

$$\begin{aligned}
&\Upsilon_p'(\rho) \in \mathcal{L}(L^p(D), \mathbb{R}),\\
&\Upsilon_p'(\rho)w = \int_D (\max(0, \rho(\xi) - 1)^{p-1} - \max(0, -\rho(\xi) - 1)^{p-1})w(\xi)d\xi.
\end{aligned}$$

Moreover, $\Upsilon_p : L^p(D) \to \mathbb{R}$ is weakly lower semicontinuous.

Proof. Follows from Lemma 3.1 and the fact that continuity and convexity imply weak lower semicontinuity [21, Corollary 3.9]. ∎

LEMMA 3.3. *Let $d \in \{2, 3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let Y be a reflexive Banach space such that Y embeds compactly in a Banach space X, $Y \hookrightarrow L^p(D)$ for $p > 1$. Let $(y_k)_{k\in\mathbb{N}}$ be a bounded sequence in Y. Then there exists a subsequence $(y_k)_{k\in K}$, $K \subset \mathbb{N}$, and $\bar{y} \in Y$ such that $y_m \rightharpoonup \bar{y}$ in Y, $y_m \to \bar{y}$ in X, and $y_m \rightharpoonup \bar{y}$ in $L^p(D)$ for $m \to \infty$.*

Proof. Since $(y_k)_{k\in\mathbb{N}}$ is bounded in Y, there exists a Y-weakly convergent subsequence $(y_k)_{k\in K}$, $K \subset \mathbb{N}$, that converges to a $\bar{y} \in Y$ [21, Theorem 3.17]. Since $Y \hookrightarrow L^p(D)$, this subsequence also converges weakly in $L^p(D)$ to the same limit. [5, Lemma 10.2(1)] concludes the proof. ∎

LEMMA 3.4. *Let $d \in \{2,3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $Y = BV(D)$, $p \in (1, \frac{d}{d-1}]$, $q \in [1, \frac{d}{d-1})$, $X = L^q(D)$ and $(y_k)_{k\in\mathbb{N}}$ be a bounded sequence in Y. Then there exists a subsequence $(y_k)_{k\in\mathbb{K}}$, $K \subset \mathbb{N}$ and $\bar{y} \in Y$ such that $y_k \rightharpoonup^* \bar{y}$ in Y, $y_k \to \bar{y}$ in X, and $y_k \rightharpoonup \bar{y}$ in $L^p(D)$ for $K \ni k \to \infty$.*

Proof. By [20, Lemma 6.108], $BV(D)$ embeds continuously into $L^{\frac{d}{d-1}}(D)$ and compactly into $L^q(D)$ for $q \in [1, \frac{d}{d-1})$. Since $BV(D)$ is compactly embedded into $L^q(D)$, there exists $\bar{y} \in L^q(D)$ and a subsequence $(y_k)_{k\in K_1}$, $K_1 \subset \mathbb{N}$ that converges $L^q(D)$-strongly to $\bar{y}$. Since D is bounded, $L^q(D)$ is continously embedded into $L^1(D)$ and, therefore, $(y_k)_{k\in K_1}$ converges L^1-strongly to $\bar{y}$.
Since $(y_k)_{k\in K_1}$ is bounded in $BV(D)$ and converges L^1-strongly to $\bar{y}$, $(y_k)_{k\in K_1}$ converges BV-weakly* to $\bar{y}$ [7, Proposition 3.13].
By the continuous embedding of $BV(D)$ into $L^p(D)$, $(y_k)_{k\in K_1}$ is bounded in $L^p(D)$. Thus, there exists a limit point $\bar{x}$ and a weakly convergent subsequence $(y_k)_{k\in K_2}$, $K_2 \subset K_1$, that converges L^p-weakly to $\bar{x}$. Since weak convergence in L^p implies weak convergence in L^1 for bounded D, $(y_k)_{k\in K_2}$ converges weakly to $\bar{x}$. Since $(y_k)_{k\in K_1}$ converges strongly to $\bar{y}$ in $L^1(D)$, it also converges L^1-weakly to $\bar{y}$, which implies $\bar{x} = \bar{y}$.

LEMMA 3.5 (Continuity and boundedness from below of J). *Let $d \in \{2,3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let X be a Banach space. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous. Then $J : X \times H^1(D)^d \to \mathbb{R}$ is continuous. Moreover, J is bounded from below.*

Proof. Recall that J is defined in (1.6). We exemplarily show continuity of the first summand of $J(\rho,u)$ given by $(\alpha(\rho)u,u)_D$. Consider the multilinear form $m(w_1,w_2,w_3) := (w_1w_2,w_3)_D$. By Hölder's inequality and the continuous embedding $H^1(D)^d \hookrightarrow L^{\frac{2s}{s-1}}(D)^d$, there exists a constant $C > 0$ such that

$$\begin{aligned} |m(w_1,w_2,w_3)| &\leq \|w_1\|_{L^s(D)} \|w_2\|_{L^{\frac{2s}{s-1}}(D)^d} \|w_3\|_{L^{\frac{2s}{s-1}}(D)^d} \\ &\leq C\|w_1\|_{L^s(D)}\|w_2\|_{H^1(D)^d}\|w_3\|_{H^1(D)^d}. \end{aligned} \tag{3.3}$$

Therefore, $m : L^s(D) \times H^1(D)^d \times H^1(D)^d \to \mathbb{R}$ is well-defined. Due to continuity of α, the mapping $(\rho,u) \mapsto (\alpha(\rho),u,u)$ is continuous as a mapping $X \times H^1(D)^d \to L^s(D) \times H^1(D)^d \times H^1(D)^d$. Thus, we obtain continuity of $(\rho,u) \to m(\alpha(\rho),u,u) = (\alpha(\rho)u,u)_D$. The other terms can be handled analogously.

It holds that J is bounded from below since by Poincaré's inequality and Young's inequality there exists a constant $C > 0$ such that

$$\begin{aligned} \frac{1}{2}(\alpha(\rho)u,u)_D + \frac{\mu}{2}(\nabla u, \nabla u)_D - (f,u)_D &\geq C\|u\|_{H^1(D)^d} - \|f\|_{H^{-1}(D)^d}\|u\|_{H^1(D)^d} \\ &\geq \frac{C}{2}\|u\|_{H^1(D)^d} - \frac{1}{2C}\|f\|_{H^{-1}(D)^d}. \end{aligned}$$ ∎

LEMMA 3.6 (Continuity and boundedness from below of j). *Let $d \in \{2,3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let X be a Banach space. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous. Then $j : X \to \mathbb{R}$ is continuous. Moreover, j is bounded from below.*

Proof. Follows from Lemma 3.5 and Lemma 2.2. ∎

LEMMA 3.7 (Weak lower semicontinuity of j). *Let $d \in \{2,3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let X be a Banach space, Y be a reflexive Banach space such that Y embeds compactly in X, $Y \hookrightarrow L^p(D)$ for $p > 1$. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous. Then $j : Y \to \mathbb{R}$ is weakly lower semicontinuous.*

Proof. We proof this statement via contradiction. Assume that j is not weakly lower semicontinuous. Then there exists $\delta > 0$ and a sequence $(\rho_k)_{k\in\mathbb{N}}$ with Y-weak limit $\bar{\rho} \in Y$ such that $j(\rho_k) \leq j(\bar{\rho}) - \delta$ for all $k \in \mathbb{N}$. Since every weakly convergent sequence is bounded, $(\rho_k)_{k\in\mathbb{N}}$ is bounded. By Lemma 3.3 we obtain a subsequence $(\rho_k)_{k\in K}$, $K \subset \mathbb{N}$, that converges X-strongly to $\bar{\rho}$. Lemma 3.6 implies

$$|j(\rho_k) - j(\bar{\rho})| \to 0$$

for $K \ni k \to \infty$. This yields a contradiction. ∎

LEMMA 3.8 (Weak* lower semicontinuity of j). *Let $d \in \{2,3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $Y = BV(D)$, $p \in (1, \frac{d}{d-1}]$, $q \in (1, \frac{d}{d-1})$, $X = L^q(D)$. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous. Then $j : Y \to \mathbb{R}$ is weakly* lower semicontinuous.*

Proof. The proof of Lemma 3.7 can be adapted using Lemma 3.4 instead of Lemma 3.3. ∎

THEOREM 3.9. *Let $d \in \{2,3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u_D \in H^{\frac{1}{2}}(\partial D)^d$ with $\int_{\partial D} u_D \cdot n ds = 0$ and $f \in H^{-1}(D)^d$. Moreover, let $p, q > 1$, X, Z be Banach spaces and Y be a reflexive Banach space such that Y is compactly embedded into X and $X \hookrightarrow L^q(D)$ and $Y \hookrightarrow L^p(D)$. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous, $g : X \to Z$ is continuous and $\{\rho \in Y \ : \ g(\rho) \leq 0, \ \int_D (|\rho|^q - 1) d\xi = 0\}$ is non-empty. Then, for fixed $\gamma, \eta > 0$, the optimization problem defined by (1.1) - (1.6) attains a solution.*

Proof. Due to Lemmas 2.1 and 2.2 we can directly look at the reduced problem (1.1). By Lemma 3.6, $j(\rho)$ is bounded from below. Hence, a minimizing sequence $(\rho_k)_{k\in\mathbb{N}} \subset Y$ can be chosen such that $g(\rho_k) \leq 0$, $\int_D (|\rho_k|^q - 1) d\xi = 0$ for all $k \in \mathbb{N}$, the objective function values of the minimizing sequence are monotonically decreasing, and

$$\lim_{k\to\infty} j(\rho_k) + \gamma \Upsilon_p(\rho_k) + \frac{\eta}{2}\|\rho_k\|_Y^2 = \min_{\rho\in Y, \ g(\rho)\leq 0, \ \int_D (|\rho|^q - 1) d\xi = 0} j(\rho) + \gamma \Upsilon_p(\rho) + \frac{\eta}{2}\|\rho\|_Y^2. \tag{3.4}$$

Due to the regularization term in the objective function, $(\|\rho_k\|_Y)_{k\in\mathbb{N}}$ is bounded and, therefore, by Lemma 3.3 there exists $\bar{\rho} \in Y$ and a subsequence $(\rho_k)_{k\in K}$, $K \subset \mathbb{N}$, such that $\rho_k \rightharpoonup \bar{\rho}$ in Y and $\rho_k \to \bar{\rho}$ in X for $K \ni k \to \infty$. Since $Y \hookrightarrow L^p(D)$, $\rho_k \rightharpoonup \bar{\rho}$ in $L^p(D)$. The mapping $\rho \to \Upsilon_p(\rho)$ is weakly lower semicontinuous as mapping $L^p(D) \to L^1(D)$ by Corollary 3.2. With $Y \hookrightarrow L^p(D)$, the weak lower semicontinuity of j (Lemma 3.7), and the weak lower semicontinuity of the norm we hence obtain

$$j(\bar{\rho}) + \gamma \Upsilon_p(\bar{\rho}) + \frac{\eta}{2}\|\bar{\rho}\|_Y^2 \leq j(\rho_k) + \gamma \Upsilon_p(\rho_k) + \frac{\eta}{2}\|\rho_k\|_Y^2 \tag{3.5}$$

for all $k \in K$ and due to the monotonicity of the minimizing sequence also for all

$k \in \mathbb{N}$. What remains to show is the admissibility of $\bar{\rho}$. Since X embeds into $L^q(D)$,

$$\int_D (|\bar{\rho}|^q - 1) d\xi = \lim_{m\to\infty} \int_D (|\rho_m|^q - 1) d\xi = 0.$$

Together with continuity of $g : X \to Z$ we know that $\bar{\rho}$ is admissible, which proves that $\bar{\rho}$ is a minimizer. ∎

THEOREM 3.10. *Let $d \in \{2, 3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $u_D \in H^{\frac{1}{2}}(\partial D)^d$ with $\int_{\partial D} u_D \cdot n ds = 0$ and $f \in H^{-1}(D)^d$. Moreover, let $Y = BV(D)$, $p \in (1, \frac{d}{d-1}]$, $q \in (1, \frac{d}{d-1})$, $X = L^q(D)$, and Z be a Banach space such that $g : X \to Z$ is continuous. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuous, and $\{\rho \in Y \ : \ g(\rho) \leq 0, \ \int_D (|\rho|^q - 1) d\xi = 0\}$ is non-empty. Then, for fixed $\gamma, \eta > 0$, the optimization problem defined by (1.1) - (1.6) attains a solution.*

Proof. The adaption of the proof of Theorem 3.9 is straightforward, using that the BV-norm is weak* lower semicontinuous [54, Theorem 5.2.1], Lemma 3.4 and Lemma 3.8 instead of Lemma 3.3 and Lemma 3.7. ∎

4. On the differentiability of the reduced objective.

LEMMA 4.1 (Continuous differentiability of J). *Let $d \in \{2, 3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let X be a Banach space. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuously differentiable. Then $J : X \times H^1(D)^d \to \mathbb{R}$ is continuously differentiable.*

Proof. The multilinear form $m : (w_1, w_2, w_3) \mapsto (w_1 w_2, w_3)_D$ is well-defined as a mapping $L^s(D) \times H^1(D)^d \times H^1(D)^d \to \mathbb{R}$ by (3.3). Due to continuous differentiability of α, the mapping $(\rho, u) \mapsto (\alpha(\rho), u, u)$ is continuously differentiable as a mapping $X \times H^1(D)^d \to L^s(D) \times H^1(D)^d \times H^1(D)^d$. Hence, by the chain rule, $(\rho, u) \to m(\alpha(\rho), u, u)$, which corresponds to the first summand of J, is continuously differentiable as a mapping $X \times H^1(D)^d \to \mathbb{R}$. Continuous differentiability of the other terms can be proven analogously. ∎

LEMMA 4.2 (Fréchet differentiability of j). *Let $d \in \{2, 3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Let X be a Banach space. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuously differentiable. Then j is Fréchet differentiable.*

Proof. Follows from Lemma 4.1 and Lemma 2.3 using the chain rule. ∎

LEMMA 4.3 (Fréchet differentiability of $\bar{j}$). *Let $d \in \{2, 3\}$, $D \subset \mathbb{R}^d$ be a bounded Lipschitz domain, X and Y be Banach spaces such that the requirements of Lemma 3.3 are fulfilled. Assume that $\alpha : X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$, is continuously differentiable and $\cdot \mapsto \|\cdot\|_Y^2$ is continuously differentiable as a mapping $Y \to \mathbb{R}$. Then $\bar{j}$ is Fréchet differentiable.*

Proof. Follows from Lemma 4.2, Corollary 3.2 and continuous differentiability of the norm. ∎

Remark 4.4. Due to the non-differentiability of the BV-norm the adaption of Lemma 4.3 requires either smoothing techniques, see e.g. [2], or, in convex cases, working with nonsmooth optimization approaches [23, 18].

5. Limit considerations for increasing penalty parameter. We now show that, in the limit, ρ attains almost everywhere the values 1 or -1 if the penalty parameter γ is sent to infinity.

THEOREM 5.1. *Let Y be a reflexive Banach space. Let $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ be a strictly monotonically increasing sequence with $\lim_{k\to\infty} \gamma_k = \infty$. Let the prerequisites of Theorem 3.9 be fulfilled and $(\rho_k)_{k\in\mathbb{N}} \subset Y$ be a sequence of global optimal solutions of (1.1) for $\gamma = \gamma_k$ (which exists due to Theorem 3.9). Assume that*

$$\Phi_{ad} = \{\rho \in Y \ : \ g(\rho) \le 0,\ \int_D (|\rho|^q - 1)\mathrm{d}\xi = 0,\ -1 \le \rho \le 1 \ a.e.\} \tag{5.1}$$

is non-empty. Then there exists a subsequence $(\rho_k)_{k\in K}$, $K \subset \mathbb{N}$, that converges X-strongly and Y-weakly to $\bar\rho \in Y$, which is an optimal solution of

$$\min_{\rho\in\Phi_{ad}} j(\rho) + \frac{\eta}{2}\|\rho\|_Y^2. \tag{5.2}$$

Proof. The proof is inspired by the proof of [52, Theorem 18.2] and consists of several steps. Define $P_{\gamma_k}(\rho) := \hat{j}(\rho) + \gamma_k \cdot \Upsilon_p(\rho)$ and $\hat{j}(\rho) := j(\rho) + \frac{\eta}{2}\|\rho\|_Y^2$.
Step 1: The sequence $(P_{\gamma_k}(\rho_k))_{k\in\mathbb{N}}$ is monotonically increasing.
Since $\gamma_k < \gamma_{k+1}$, $\Upsilon_p(\rho_{k+1}) \ge 0$ and ρ_k is a global optimal solution of (1.1) for $\gamma = \gamma_k$ it holds

$$\begin{aligned} P_{\gamma_k}(\rho_k) \le P_{\gamma_k}(\rho_{k+1}) &= \hat{j}(\rho_{k+1}) + \gamma_k \Upsilon_p(\rho_{k+1}) \\ &\le \hat{j}(\rho_{k+1}) + \gamma_{k+1}\Upsilon_p(\rho_{k+1}) = P_{\gamma_{k+1}}(\rho_{k+1}). \end{aligned}$$

Step 2: The sequence $(\Upsilon_p(\rho_k))_{k\in\mathbb{N}}$ is monotonically decreasing.
We know that $P_{\gamma_k}(\rho_k) \le P_{\gamma_k}(\rho_{k+1})$ and $P_{\gamma_{k+1}}(\rho_{k+1}) \le P_{\gamma_{k+1}}(\rho_k)$. Adding both inequalities leads to the inequality

$$\gamma_k \Upsilon_p(\rho_k) + \gamma_{k+1}\Upsilon_p(\rho_{k+1}) \le \gamma_k \Upsilon_p(\rho_{k+1}) + \gamma_{k+1}\Upsilon_p(\rho_k).$$

This is equivalent to the inequality

$$\gamma_k(\Upsilon_p(\rho_k) - \Upsilon_p(\rho_{k+1})) \le \gamma_{k+1}(\Upsilon_p(\rho_k) - \Upsilon_p(\rho_{k+1})).$$

Since $\gamma_k < \gamma_{k+1}$, we have

$$\Upsilon_p(\rho_k) - \Upsilon_p(\rho_{k+1}) \ge 0.$$

Step 3: The sequence $(\hat{j}(\rho_k))_{k\in\mathbb{N}}$ is monotonically increasing.
It holds $P_{\gamma_k}(\rho_k) \le P_{\gamma_k}(\rho_{k+1})$, and by step 2, $\Upsilon_p(\rho_k) \ge \Upsilon_p(\rho_{k+1})$. In combination with $\gamma_k > 0$ for all $k \in \mathbb{N}$, this leads to the inequality

$$\begin{aligned} 0 \le P_{\gamma_k}(\rho_{k+1}) - P_{\gamma_k}(\rho_k) &= \hat{j}(\rho_{k+1}) - \hat{j}(\rho_k) + \gamma_k(\Upsilon_p(\rho_{k+1}) - \Upsilon_p(\rho_k)) \\ &\le \hat{j}(\rho_{k+1}) - \hat{j}(\rho_k). \end{aligned}$$

Step 4: It holds $\lim_{k\to\infty} \Upsilon_p(\rho_k) = 0$.
The set Φ_{ad} is non-empty, and thus, there exists $\hat\rho \in \Phi_{ad}$ and a corresponding $\hat{u} = S(\hat\rho)$ such that $\Upsilon_p(\hat\rho) = 0$. Using optimality of ρ_k and step 3, we have

$$\hat{j}(\hat\rho) = P_{\gamma_k}(\hat\rho) \ge P_{\gamma_k}(\rho_k) = \hat{j}(\rho_k) + \gamma_k \Upsilon_p(\rho_k) \ge \hat{j}(\rho_0) + \gamma_k \Upsilon_p(\rho_k)$$

for all $k \in \mathbb{N}$. Therefore, $\hat{j}(\rho_0) + \gamma_k \Upsilon_p(\rho_k)$ is bounded and for $\gamma_k \xrightarrow{k\to\infty} \infty$ we have $\Upsilon_p(\rho_k) \xrightarrow{k\to\infty} 0$.

Step 5: There exists a subsequence $(\rho_k)_{k\in K}$, $K \subset \mathbb{N}$ that converges X-strongly and Y-weakly to $\bar\rho \in Y$, which is an optimal solution of (5.2).
Due to continuity of $\hat j : Y \to \mathbb{R}$ (Lemma 3.6 and continuity of the norm), optimality of ρ_k, $\Upsilon_p(\rho_k) \geq 0$, and non-emptyness of Φ_{ad}, $(\hat j(\rho_k))_{k\in\mathbb{N}}$ is bounded since there exists $\hat\rho \in \Phi_{ad}$ such that

$$\infty > \hat j(\hat\rho) = P_{\gamma_k}(\hat\rho) \geq P_{\gamma_k}(\rho_k) = \hat j(\rho_k) + \gamma_k \Upsilon_p(\rho_k) \geq \hat j(\rho_k) \geq \hat j(\rho_0).$$

Since j is bounded from below (Lemma 3.6), $(\|\rho_k\|_Y)_{k\in\mathbb{N}}$ is bounded. Due to the compact embedding of Y in X, by Lemma 3.3 there exists a subsequence $(\rho_k)_{k\in K}$, $K \subset \mathbb{N}$, $\bar\rho \in Y$ such that $\rho_k \to \bar\rho$ in X and $\rho_k \rightharpoonup \bar\rho$ in Y for $K \ni k \to \infty$. Step 4 and weak lower semicontinuity of $\Upsilon_p : L^p(D) \to \mathbb{R}$ (Corollary 3.2) imply that

$$0 \leq \Upsilon_p(\bar\rho) \leq \liminf_{K\ni k\to\infty} \Upsilon_p(\rho_k) = 0,$$

and, therefore, $-1 \leq \bar\rho \leq 1$ a.e. Continuity of g implies $g(\bar\rho) \leq 0$. Since $X \hookrightarrow L^q(D)$, we also know that $\int_D (|\bar\rho|^q - 1)d\xi = 0$. Hence, $\bar\rho \in \Phi_{ad}$.
For the following inequalities, we use that $\gamma_k \Upsilon_p(\rho_k) \geq 0$ for all $k \in \mathbb{N}$ and that ρ_k is optimal for P_{γ_k}. Since for all $\rho \in \Phi_{ad}$ it holds that $\Upsilon_p(\rho) = 0$, we have

$$\hat j(\rho_k) \leq P_{\gamma_k}(\rho_k) \leq P_{\gamma_k}(\rho) = \hat j(\rho).$$

In combination with the weak lower semicontinuity of $\hat j$ as a mapping $Y \to \mathbb{R}$ (Lemma 3.7), $\bar\rho$ is an optimal solution of (5.2) since

$$\hat j(\bar\rho) \leq \liminf_{K\ni k\to\infty} \hat j(\rho_k) \leq \hat j(\rho) \quad \text{for all } \rho \in \Phi_{ad}. \qquad \square$$

Remark 5.2. Theorem 5.1 can also be proven if we replace the reflexivity of Y with the requirements of Lemma 3.4.

Remark 5.3. Theorem 5.1 requires the global optima of the relaxed problems. In practice, due to the nonlinear nature of the optimization problems, one typically obtains local optima. The quality of these optima typically depends on the intial point for the optimization and on appropriate regularization techniques, such as a term that corresponds to a penalization of the perimeter of the resulting optimal shapes. Another approach is using deflation techniques [46].

6. Choice of Y, X, Z, g, p, q and α. Summarizing the requirements of the previous sections, we obtain the following assumptions.

Assumption 1. Let Y, X, Z, g, p, q and a superposition operator α defined by (1.5), satisfy
- Y is either reflexive or $Y = BV(D)$,
- Y embeds compactly in X, with $X \hookrightarrow L^q(D)$ with $q > 1$,
- $Y \hookrightarrow L^p(D)$ with $p > 1$,
- α is continuously differentiable as a mapping $X \to L^s(D)$, defined by (1.5), with $s > 1$ for $d = 2$ and $s \geq \frac{3}{2}$ for $d = 3$,
- Φ_{ad}, defined in (5.1), contains an element $\hat\rho$ such that $j(\hat\rho) < \infty$. In particular, Y should allow for jumps of ρ along hypersurfaces,
- g is continuous as a mapping $X \to Z$.

The following lemma will be helpful to prove Fréchet differentiability of α.

LEMMA 6.1. *Let $p > q > 1$ and $t > 1$ be such that $qt \leq p$. Let α be a superposition operator defined by (1.5) and* $\tilde{\alpha}(x) := \begin{cases} \bar{\alpha}|x|^t & \text{if } x < 0, \\ 0 & \text{else}, \end{cases}$ *with $\bar{\alpha} > 0$. Then, $\alpha : L^p(D) \to L^q(D)$ is continuously differentiable.*

Proof. Can be shown analogously to Lemma 3.1. It holds $\tilde{\alpha}(x) := \max(0, (-x)^t)$, which is locally Lipschitz continuous with $\tilde{\alpha}'(x) = t \max(0, (-x)^{t-1})$. We consider the superposition operator $\alpha(\rho)(\xi) := \tilde{\alpha}(\rho(\xi))$, which fulfills the growth condition $|\alpha(\rho)(\xi)| \leq \bar{\alpha}|\rho(\xi)|^t$. This implies continuity of α for $t = \frac{p}{q}$ [51, Section 4.3.3]. In addition, $\alpha'(\rho)(\xi) := \tilde{\alpha}'(\rho(\xi))$ maps $\rho \in L^p(D)$ to $L^r(D)$ with $r = \frac{p}{t-1}$. Hence α is continuously differentiable [51, Section 4.3.3]. ∎

One choice of Y that allows for fulfilling the above requirements is the space $BV(D)$. The corresponding total variation (TV) term in the regularization promotes piecewise constant behavior of optimal solutions, see e.g. [47, 19].

LEMMA 6.2. *The choice $d \in 2$, $Y = BV(D)$, $X = L^q(D)$, $p = 2$, $q \in (1, \frac{3}{2})$, g given by (1.7), $Z = \mathbb{R}$,* $\tilde{\alpha}(x) = \begin{cases} \bar{\alpha}|x|^{\frac{3}{2q}} & \text{if } x < 0 \\ 0 & \text{else} \end{cases}$*, and $\alpha(\rho)(\xi) := \tilde{\alpha}(\rho(\xi))$ for all $\xi \in D$, with $\bar{\alpha} \gg 1$ satisfies Assumption 1.*

Proof. The assumptions on Y are fulfilled due to [7, Proposition 3.13, Definition 3.11], and [54, Theorem 5.2.1], see also proof of Lemma 3.4. By [20, Lemma 6.108], $BV(D)$ embeds continuously into $L^{\frac{d}{d-1}}(D)$ and compactly into $L^r(D)$ for $r \in (1, \frac{d}{d-1})$. The Fréchet differentiability of α follows from Lemma 6.1. ∎

The total variation is, in general, not accessible for computation. For an indicator function of a subset $\Omega \subset D$ it corresponds to the perimeter of Ω, see [7, Section 3.3]. If we consider smoother functions $u \in W^{1,1}(D)$, then $TV(u)$ can be computed via

$$\int_D |\nabla u|_2 \, d\xi,$$

see [2, Section 2]. However, $TV(u)$ is not differentiable, which is disadvantageous for optimization, in particular Lemma 4.3 is not applicable, see also Remark 4.4.

Another choice for Y is the space $H^\sigma(D)$, $\sigma < \frac{d}{2}$.

LEMMA 6.3. *Let D be a bounded Lipschitz domain. The choice $d = 2$, $Y = H^\sigma(D)$, $\sigma = \frac{7}{8}$, $X = L^8(D)$, $p = 8$, $q = 2$, $s = 2$, g given by (1.7), $Z = \mathbb{R}$,* $\tilde{\alpha}(x) = \begin{cases} \bar{\alpha}x^4 & \text{if } x < 0 \\ 0 & \text{else} \end{cases}$*, and $\alpha(\rho)(\xi) := \tilde{\alpha}(\rho(\xi))$ for all $\xi \in D$, with $\bar{\alpha} \gg 1$ satisfies Assumption 1.*

Proof. It holds that with $\sigma < 1$, $H^\sigma(D) \hookrightarrow L^{\tilde{p}}(D)$ for $\tilde{p} = \frac{2d}{d-2\sigma}$ and $H^\sigma(D)$ embeds compactly into $L^q(D)$ for any $q \in [1, \tilde{p})$ [6, Theorem 4.4], [3, Theorem 7.34], [27, Theorem 6.7]. Hence, we can choose $p = 8$ and $q = 2$, which also gives continuity and differentiability of α according to Lemma 6.1. ∎

The $H^\sigma(D)$-norm is given as

$$\| \cdot \|_{H^\sigma(D)} = (\| \cdot \|^2_{L^2(D)} + | \cdot |^2_\sigma)^{\frac{1}{2}},$$

where the H^σ-seminorm is, e.g., given by the Sobolev-Slobodeckij seminorm

$$| \cdot |_\sigma = \left(\int_D \int_D \frac{|u(x) - u(y)|^2}{\|x - y\|^{d+2\sigma}} dy \, dx\right)^{\frac{1}{2}}. \tag{6.1}$$

[28, 40] propose to work with a slightly different norm, which - under assumptions on the weighting $\kappa(x,y)$ - is equivalent to the original H^σ-norm according to [40, Lemma 2.1]:

$$|||\cdot|||_{H^\sigma(D)} := (\|\cdot\|^2_{L^2(D)} + |\cdot|^2_{\kappa,\sigma})^{\frac{1}{2}}, \tag{6.2}$$

where

$$|\cdot|_{\kappa,\sigma} := (\int_D\int_D \frac{|u(x)-u(y)|^2}{\|x-y\|^{d+2\sigma}}\kappa(x,y)dy\,dx)^{\frac{1}{2}}, \tag{6.3}$$

and $\kappa(x,y)$ fulfills [40, Assumption 2.1], e.g.

$$\begin{cases} 1 & \text{if } \|x-y\| \le \delta, \\ 0 & \text{else,} \end{cases} \tag{6.4}$$

see [40, Remark 2.2]. For convenience, we work with the continuously differentiable approximation

$$\kappa(x,y) = \begin{cases} 1 & \text{if } \|x-y\| \le \delta, \\ f(\frac{\|x-y\|^2-\delta^2}{\frac{9}{16}\delta^2}) & \text{if } \|x-y\| \in (\delta, \frac{5}{4}\delta), \\ 0 & \text{else,} \end{cases} \tag{6.5}$$

for $f(r) := 2r^3 - 3r^2 + 1$, which also fulfills [40, Assumption 2.1]. Working with this definition of the norm reduces the computational effort to assemble the $H^\sigma(D)$-matrix. However, for fixed δ, the bandwidth of the matrix increases for decreasing mesh size. It might be convenient to have a matrix with fixed bandwidth. This requires to choose $\delta = \mathcal{O}(h)$. In the following, we motivate that this is justified in our application as long as $\sigma = \sigma(\delta)$ is adapted correspondingly.

One is often interested in shapes with bounded total variation, compare Remark 5.3. When working with $H^\sigma(D)$, $\sigma < \frac{d}{2}$, it is a priorily not clear if the optimal shape has bounded variation. For this reason, we take a closer look into the theory. In [16] it is shown that a function $u \in L^1(D)$ is an element of $BV(D)$ if and only if

$$\liminf_{\tilde\sigma\to 1}(1-\tilde\sigma)\int_D\int_D \frac{|u(x)-u(y)|}{|x-y|^{d+\tilde\sigma}}dydx < \infty.$$

More precisely, if D is a Lipschitz domain, there exists a constant c that depends on d such that

$$\lim_{\tilde\sigma\to 1^-}(1-\tilde\sigma)\int_D\int_D \frac{|u(x)-u(y)|}{|x-y|^{d+\tilde\sigma}}dydx \to c\,TV(u) \tag{6.6}$$

for all $u \in BV(D)$, where $TV(u)$ denotes the total variation of u [25, 42]. A similar result can also obtained for the seminorm (6.3). Let $\sigma = \frac{1}{2}\tilde\sigma$, $u \in BV(D, \{-1,1\})$ and let $\delta > 0$ be chosen arbitrarily. Then, since $|u| = 1$ a.e., and

$$\begin{aligned} &\int_D\int_D \frac{|u(x)-u(y)|}{|x-y|^{d+2\sigma}}dydx = \frac{1}{2}\int_D\int_D \frac{|u(x)-u(y)|^2}{|x-y|^{d+2\sigma}}dydx \\ &= \frac{1}{2}\int_D\int_D \frac{|u(x)-u(y)|^2}{|x-y|^{d+2\sigma}}\kappa(x,y)dydx + \frac{1}{2}\int_D\int_D \frac{|u(x)-u(y)|^2}{|x-y|^{d+2\sigma}}(1-\kappa(x,y))dydx. \end{aligned}$$

Since

$$
\begin{aligned}
&|\frac{1}{2}(1-2\sigma)\int_D\int_D \frac{|u(x)-u(y)|^2}{|x-y|^{d+2\sigma}}(1-\kappa(x,y))dydx\ | \\
&\leq |\frac{1}{2}(1-2\sigma)\int_D\int_{D\setminus B_\delta(x)} \frac{|u(x)-u(y)|^2}{|x-y|^{d+2\sigma}}dydx| \leq 2|D|^2(1-2\sigma)\delta^{-d-2\sigma},
\end{aligned}
$$

we have

$$
\lim_{\sigma\to\frac{1}{2}^-}\frac{1}{2}(1-2\sigma)\int_D\int_D \frac{|u(x)-u(y)|^2}{|x-y|^{d+2\sigma}}\kappa(x,y)dydx = cTV(D)
$$

if $\sigma=\sigma(\delta)$ such that $\sigma\to\frac{1}{2}^-$ and $(1-2\sigma)\delta^{-d-2\sigma}\to 0$ for $\delta\to 0$. This motivates to consider the following setting.

LEMMA 6.4. *Let D be a bounded Lipschitz domain. The choice $d\in\{2,3\}$, $Y = H^\sigma(D)$, $\frac{3}{8}\leq\sigma\leq\frac{1}{2}$, $X = L^{\frac{15}{6}}(D)$, $p=\frac{8}{3}$, $q=2$, $s=\frac{3}{2}$, g given by (1.7), $Z=\mathbb{R}$,*

$$
\tilde{a}(x) = \begin{cases} \bar{\alpha}|x| & \text{if } \ x<-1, \\ \bar{\alpha}(-\frac{1}{16}x^4+\frac{3}{8}x^2-\frac{1}{2}x+\frac{3}{16}) & \text{if } \ -1\leq x<1, \\ 0 & \text{else,} \end{cases}
$$

and $\alpha(\rho)(\xi) := \tilde{a}(\rho(\xi))$ for all $\xi\in D$, with $\bar{\alpha}\gg 1$ satisfies Assumption 1.

Proof. Follows as in Lemma 6.3. Since continuous differentiability of α is not directly covered by Lemma 6.1, we prove it here. It holds that $p>q>1$. Furthermore, α is Lipschitz continuous with

$$
\tilde{a}'(x) = \begin{cases} -\bar{\alpha} & \text{for } \xi<-1, \\ \bar{\alpha}(-\frac{1}{4}x^3+\frac{3}{4}x-\frac{1}{2}) & \text{for } -1\leq\xi<1, \\ 0 & \text{else.} \end{cases}
$$

It fulfills the growth condition $|\alpha(\rho)(\xi)|\leq\bar{\alpha}(|\rho(\xi)|+1)$. Hence, continuity of $\alpha: L^p(D)\to L^q(D)$ follows with [51, Section 4.3.3]. Since $\alpha'(\rho)(\xi) := \tilde{a}'(\rho(\xi))$ maps $\rho\in L^p(D)$ to $L^\infty(D)$ and $p>q$, continuous differentiability follows with [51, Section 4.3.3]. ◻

Remark 6.5. Our experiments indicate that the $\frac{1}{2}(\alpha(\rho)u,u)_D$ term in the objective function is important for the numerical performance. Moreover, choosing $\tilde{a}(x)>0$ for $x\in(-1,1)$ shows faster convergence than having a plateau by choosing $\tilde{a}(x)=0$ for $x\in(0,1)$. This relates to the observations in connection with [34, Figure 7].

7. Numerical realization. In the scope of the work we realize the setting of Lemma 6.4 for the particular choice $d=2$, $\sigma=\frac{7}{16}$. To discretize the states (u,p) we use mixed Taylor-Hood finite elements, i.e. piecewise quadratic continuous Lagrange finite elements (CG2 FEM) for the velocity u and piecewise linear continuous Lagrange finite elements (CG1 FEM) for the pressure p. The design variable ρ is discretized with piecewise constant discontinuous Lagrange finite elements (DG0 FEM) in order to for the discretized space to be a subset of Y.

Remark 7.1. The volume constraint prevents ρ from being constantly 1 or -1. Using CG1 FEM for ρ enforces an interfacial region with width of at least $O(h)$, in which $\rho\in(-1,1)$. Hence, the sphere constraint enforces values of ρ^h which are bigger than 1 or smaller than -1. This, however, leads for a fixed mesh size h to an optimal objective function $\bar{j}$ value that diverges to ∞ for $\gamma\to\infty$.

By (ψ_l) we denote the nodal basis functions of the CG1 FEM space $S_1^h \subset H^1(D)$, by (ϕ_k) the nodal basis functions of the CG2 FEM space $S_2^h \subset H^1(D)$, and by (Φ_k) the nodal basis functions of DG0 FEM space $S_0^h \subset H^\sigma(D)$. Therefore, the discrete representations of the velocity $u \in H^1(D)^d$, the pressure $p \in L_0^2(D)$ and of the design variable $\rho \in H^\sigma(D)$ are the following:

$$u_i^h(\xi) = \sum_k (\mathbf{u}_i)_k \phi_k(\xi),\ p^h(\xi) = \sum_\ell \mathbf{p}_\ell \psi_\ell(\xi),\ \rho^h(\xi) = \sum_j \boldsymbol{\rho}_j \Phi_j(\xi),$$

for $i \in \{1, \ldots, d\}$ with coefficient vectors $(\mathbf{u}_i), \mathbf{p}, \boldsymbol{\rho}$. Since it is discretized with CG1 FEM, for p^h the entries of $\mathbf{p}$ correspond to the nodal values. For ρ^h, $\boldsymbol{\rho}$ contains the values on the cells.

7.1. State Equation. The pressure solving (1.4) is only unique up to an additive constant [14]. Therefore, we choose p to be in the Banach space $L_0^2(D) = \{p \in L^2(D) : \int_D p \mathrm{d}\xi = 0\}$. Thus, the variational problem of the state equation, having a unique solution, is:
Find $u \in U = \{u \in H^1(D)^d : u = u_D \text{ on } \partial D\}$ and $p \in \Pi = L_0^2(D)$ s.t.

$$\begin{aligned} \mu \int_D \nabla u : \nabla v \mathrm{d}\xi + \int_D \alpha(\rho) u \cdot v \mathrm{d}\xi - \int_D p \operatorname{div}(v) \mathrm{d}\xi &= \int_D f \cdot v \mathrm{d}\xi, \\ \int_D q \operatorname{div}(u) \mathrm{d}\xi &= 0, \end{aligned}$$

for all $v \in H_0^1(D)^d$, $q \in L_0^2(D)$. We define the bilinear forms

$$\begin{aligned} &a : H^1(D) \times H^1(D) \to \mathbb{R},\ a(u,v) := \langle \nabla u, \nabla v \rangle_{L^2(D)}, \\ &b_i : H^1(D) \times L^2(D) \to \mathbb{R},\ b_i(v,q) := \langle \partial_i v, q \rangle_{L^2(D)} \end{aligned}$$

and the linear form

$$F_i : H^1(D) \to \mathbb{R},\ F_i(v) = \langle f_i, v \rangle_{L^2(D)^d}.$$

Additionally, we have the nonlinear form

$$r : H^\sigma(D) \times H^1(D) \times H^1(D) \to \mathbb{R},\ r(\rho; u, v) := \langle \alpha(\rho) u, v \rangle_{L^2(D)}.$$

Therefore, the variational formulation of the state equation for $D \subset \mathbb{R}^d$ can be written as:
Find $u \in U$ and $p \in \Pi$ s.t.

$$\begin{aligned} &\langle E(\rho, u, p), (v, q) \rangle_{H^{-1}(D)^d \times \Pi^*, H_0^1(D)^d \times \Pi} \\ &= \sum_{i=1}^d \mu a(u_i, v_i) + r(\rho; u_i, v_i) - b_i(v_i, p) + b_i(u_i, q) - \sum_{i=1}^d F_i(v_i) = 0 \end{aligned}$$

for all $v \in H_0^1(D)^d$ and $q \in L_0^2(D)$. This variational problem with Dirichlet boundary condition can be reduced to a homogeneous problem by choosing a function $u_D \in U$ and setting $u = w + u_D$ with $(w, p) \in H_0^1(D)^d \times L_0^2(D)$ solving

$$\begin{aligned} &\sum_{i=1}^d \mu a(w_i, v_i) + r(\rho; w_i, v_i) - b_i(v_i, p) + b_i(w_i, q) \\ &= \sum_{i=1}^d F_i(v_i) - \mu a(u_{Di}, v_i) - r(\rho; u_{Di}, v_i) - b_i(u_{Di}, q) \end{aligned}$$

for all $(v,q) \in H_0^1(D)^d \times L_0^2(D)$.

The discrete version of the nonlinear terms $r(\rho; u_i, v_i)$, $i \in \{1,\dots,d\}$, is

$$r(\rho^h; u_i^h, v_i^h) = \sum_{j,k,\ell} \alpha(\boldsymbol{\rho}_\ell)(\mathbf{u}_i)_j(\mathbf{v}_i)_k \int_D \Phi_\ell(\xi)\phi_j(\xi)\phi_k(\xi)\mathrm{d}\xi = \mathbf{u}_i^\top R(\boldsymbol{\rho})\mathbf{v}_i,$$

with $R_{jk}(\boldsymbol{\rho}) = r(\rho^h; \phi_j, \phi_k)$. To get the discrete equations of the variational problem we assemble

$$A_{ij} = a(\phi_j, \phi_i), \quad B_{ij}^\ell = b_\ell(\phi_i, \psi_j), \quad \text{and } (\mathbf{f}_\ell)_i = F_\ell^h(\phi_i),$$

where $F_\ell^h(\phi_i) := \langle f_i^h, v\rangle_{L^2(D)^d}$ and f_i^h is a piecewise linear or quadratic, continuous interpolation of the function f_i. Since w_i^h and v_i^h fulfill the homogeneous Dirichlet boundary conditions, it holds $w_i^h = \sum_k (\mathbf{w_i})_k\phi_k(\xi) = \sum_{k\in I}(\mathbf{w_i})_k\phi_k(\xi)$ and $v_i^h = \sum_{k\in I}(\mathbf{v_i})_k\phi_k(\xi)$, where I denotes the set of non-Dirichlet boundary nodes. Then for $d=2$, the FEM discretization of the state equation in matrix-vector form reads as

$$\begin{aligned} \mu A_{II}(\mathbf{w}_1)_I + R(\boldsymbol{\rho})_{II}(\mathbf{w}_1)_I - B^1_{I\bullet}\mathbf{p} &= (\mathbf{f}_1)_I - \mu A_{I\bullet}\mathbf{u}_{D1} - R(\boldsymbol{\rho})_{I\bullet}\mathbf{u}_{D1} \\ \mu A_{II}(\mathbf{w}_2)_I + R(\boldsymbol{\rho})_{II}(\mathbf{w}_2)_I - B^2_{I\bullet}\mathbf{p} &= (\mathbf{f}_2)_I - \mu A_{I\bullet}\mathbf{u}_{D2} - R(\boldsymbol{\rho})_{I\bullet}\mathbf{u}_{D2} \\ \left(B^1_{I\bullet}\right)^\top(\mathbf{w_1})_I + \left(B^2_{I\bullet}\right)^\top(\mathbf{w_2})_I &= -\left(B^1\right)^\top\mathbf{u}_{D1} - \left(B^2\right)^\top\mathbf{u}_{D2} \end{aligned} \tag{7.1}$$

For a given ρ^h, these equations define a unique solution for $u^h = w^h + u_D^h$ and p^h if we fix one degree of freedom of the pressure p^h.

7.2. $H^\sigma(D)$-norm. In this section we discuss how we realize the H^σ-norm on uniform meshes based on the Sobolev-Slobodeckij norm, see the discussion in Section 6 and [40, 28]. There are also other possibilities to realize (norms that are equivalent to) fractional order Sobolev norms, e.g. working with inverse estimates on a hierarchy of nested subspaces [22] or fractional powers of the stiffness matrix (for DG finite elements obtained by a discontinuous Galerkin discretization of the Laplacian) [11, 9, 37, 29, 43].

The problem in the numerical realization is the non-locality of the $H^\sigma(D)$-norm, which makes it hard to compute. To assemble the matrix corresponding to the H^σ-seminorm $|\cdot|_{\kappa,\sigma}$, consider the symmetric bilinear form

$$a_\sigma(\rho_1,\rho_2) := \int_D \rho_1(\xi)\rho_2(\xi)d\xi + \langle \rho_1,\rho_2\rangle_{\kappa,\sigma},$$

with

$$\langle \rho_1,\rho_2\rangle_{\kappa,\sigma} := \int_D\int_D \frac{(\rho_1(x)-\rho_1(y))(\rho_2(x)-\rho_2(y))}{\|x-y\|^{d+2\sigma}}\kappa(x,y)dy\,dx. \tag{7.2}$$

When we consider the discretized functions $\rho_\ell^h(\xi) = \sum_i(\boldsymbol{\rho}_\ell)_i\Phi_i(\xi)$, $\ell \in \{1,2\}$, we obtain $\langle \rho_1^h,\rho_2^h\rangle_{\kappa,\sigma} = \sum_{i,j}(\boldsymbol{\rho_1})_iM_{i,j}(\boldsymbol{\rho_2})_j$ with

$$M_{i,j} = \langle\!\langle \Phi_i\Phi_j, 1\rangle\!\rangle + \langle\!\langle 1, \Phi_i\Phi_j\rangle\!\rangle - \langle\!\langle \Phi_j,\Phi_i\rangle\!\rangle - \langle\!\langle \Phi_i,\Phi_j\rangle\!\rangle$$

for $i \neq j$, where

$$\langle\!\langle \rho_1,\rho_2\rangle\!\rangle := \int_D\int_D \frac{\rho_1(x)\rho_2(y)}{\|x-y\|^{d+2\sigma}}\kappa(x,y)dy\,dx \tag{7.3}$$

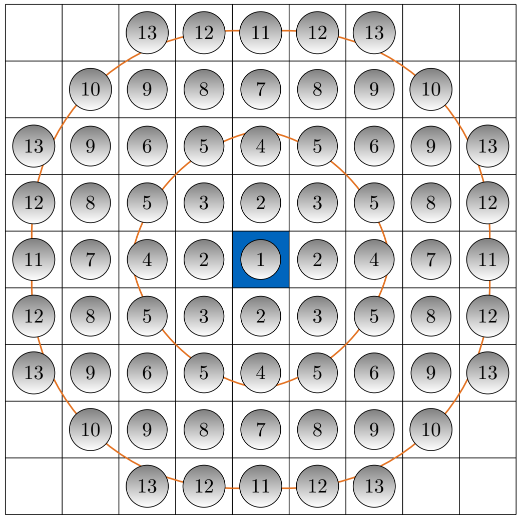


Figure 7.1: Local stencil for $|\cdot|_{\kappa,\sigma}$, between the orange lines κ attains values in $(0,1)$

for all $\rho_1, \rho_2 \in H^\sigma(D)$ and κ defined by (6.5). Using symmetry of κ we obtain

$$M_{i,j} = 2\langle\!\langle \Phi_i \Phi_j, 1 \rangle\!\rangle - 2\langle\!\langle \Phi_i, \Phi_j \rangle\!\rangle. \tag{7.4}$$

If Φ_i and Φ_j have disjoint interior supports, this further simiplifies to

$$M_{i,j} = -2\langle\!\langle \Phi_i, \Phi_j \rangle\!\rangle.$$

In order to minimize the computational effort, we consider $\delta = O(h)$ in (6.5), which yields local, h-dependent equivalent norms of the non-local H^σ-norm. Keeping the motivation in Section 6 in mind, this is justified if $(1-2\sigma) = o(h^{d+2\sigma})$.

For simplicity, we consider uniform rectangular meshes, which is, e.g., obtained for uniform triangular meshes if we choose - for piecewise constant finite elements - the degrees of freedom of two neighboring elements forming a rectangle equally. Moreover, we coose $\delta = 2\sqrt{2}h$ in the definition of κ such that in all neighboring elements the weighting is constantly 1. Figure 7.1 illustrates the local stencil. Due to symmetry arguments and the κ-term 13 integrals have to be determined. However, when using quadrature rules for determining the integrals, one has to take care that singularities appear for 2 and 3. The κ-term is different from being constantly 1 or 0 on the cells 4 - 13. Let f be defined as in (6.5) and

$$\tilde{\kappa}(x,y) = \begin{cases} 1 & \text{if } \|x-y\| \le 2\sqrt{2}, \\ f\left(\frac{\|x-y\|^2 - 8}{\frac{9}{2}}\right) & \text{if } \|x-y\| \in (2\sqrt{2}, \frac{5}{2}\sqrt{2}), \\ 0 & \text{else.} \end{cases} \tag{7.5}$$

Let $d = 2$, and

$$I_{i,j} := -2h^{2-2\sigma} \int_0^1 \int_0^1 \int_i^{i+1} \int_j^{j+1} \|x-y\|^{-2-2\sigma} \tilde{\kappa}(x,y) dy_2 dy_1 dx_2 dx_1,$$

k	g	d	b	a	a	a	a
k	g	d	b	a	a	a	a
k	g	d	b	a	a	a	a
k	g	d	b	a	a	a	a
ℓ	h	e	c	b	b	b	b
m	i	f	e	d	d	d	d
n	j	i	h	g	g	g	g
o	n	m	ℓ	k	k	k	k

$$
\begin{aligned}
1_a &= 1\\
1_b &= 1_a + 11 + 2\,12 + 2\,13\\
1_c &= 1_a + 2\,11 + 4\,12 + 4\,13\\
1_d &= 1_b + 7 + 2\,8 + 2\,9 + 2\,10\\
1_e &= 1_d + 11 + 2\,12 + 2\,13\\
1_f &= 1_c + 2\,7 + 4\,8 + 4\,9 + 3\,10\\
1_g &= 1_d + 4 + 2\,5 + 2\,6 + 2\,9 + 2\,13\\
1_h &= 1_g + 11 + 2\,12 + 13\\
1_i &= 1_h + 7 + 2\,8 + 9 + 10\\
1_j &= 1_i + 4 + 2\,5 + 6 + 9 + 13\\
1_k &= 1_g + 2 + 2\,3 + 2\,5 + 2\,8 + 2\,12\\
1_\ell &= 1_k + 11 + 12 + 13\\
1_m &= 1_\ell + 7 + 8 + 9 + 10\\
1_n &= 1_m + 4 + 5 + 6 + 9 + 13\\
1_o &= 1_n + 2 + 3 + 5 + 8 + 12
\end{aligned}
$$

Figure 7.2: Classification of elements near the boundary of the rectangular domain D.

where $x = (x_1, x_2)^\top$ and $y = (y_1, y_2)^\top$. We obtain the integrals over functions with singularities and $\tilde{\kappa} \equiv 1$

$$2 = I_{-1,0}, \quad 3 = I_{-1,-1},$$

and, with $\tilde{\kappa}$-term not constantly equal to 1,

$$
\begin{aligned}
&4 = I_{2,0}, \quad 5 = I_{2,1}, \quad 6 = I_{2,2}, \quad 7 = I_{3,0}, \quad 8 = I_{3,1},\\
&9 = I_{3,2}, \quad 10 = I_{3,3}, \quad 11 = I_{4,0}, \quad 12 = I_{4,1}, \quad 13 = I_{4,2}.
\end{aligned}
$$

Since (7.2) is zero for $\rho_1 = \Phi_i$ and $\rho_2 \equiv 1$, we obtain for elements that are sufficiently far away from the boundary such that all neighboring elements of the local stencil exist

$$
\begin{aligned}
1 = 0 &- 4\,2 - 4\,3 - 4\,4 - 8\,5 - 4\,6 - 4\,7\\
&- 8\,8 - 8\,9 - 4\,10 - 4\,11 - 8\,12 - 8\,13.
\end{aligned}
$$

Hence, $M_{i,i} = 1 + \int_D \Phi_i(\xi)\Phi_i(\xi)d\xi$, for all elements i sufficiently far away from the boundary.

7.2.1. Modification of local stencil near boundary. For elements i close to the boundary we obtain $M_{i,i} = 1_* + \int_D \Phi_i(\xi)\Phi_i(\xi)d\xi$, where $*$ denotes the classification of the element i and the modified formulas near the boundary are given according to Figure 7.2.

7.2.2. Computation of the entries of the local stencil. From the local stencil, the global matrix M can be assembled such that

$$a_\sigma(\rho_1^h, \rho_2^h) = \boldsymbol{\rho}_1^\top M \boldsymbol{\rho}_2.$$

To compute the integrals ② and ③ we use the procedure described in [24]. Therefore, we first transform the integrals appropriately such that we integrate over the $2d$-hypercube and the singularity is isolated in the first coordinate direction. In order to compute ② we have to evaluate

$$
\begin{aligned}
&\int_0^1\int_0^1\int_{-1}^0\int_0^1 \|x-y\|^{-2-2\sigma} dy_2 dy_1 dx_2 dx_1 \\
&= \int_0^1\int_0^1\int_0^1\int_0^1 ((x_1+y_1)^2+(x_2-y_2)^2)^{-1-\sigma} dy_2 dy_1 dx_2 dx_1 \\
&= \int_0^1\int_0^1\int_0^1\int_{x_2-1}^{x_2} ((x_1+y_1)^2+z^2)^{-1-\sigma} dz dx_2 dy_1 dx_1 \\
&= \int_0^1\int_0^1\int_{-1}^1\int_{\max(0,z)}^{\min(1+z,1)} ((x_1+y_1)^2+z^2)^{-1-\sigma} dx_2 dz dy_1 dx_1 \\
&= \int_0^1\int_0^1\int_0^1\int_z^1 ((x_1+y_1)^2+z^2)^{-1-\sigma} dx_2 dz dy_1 dx_1 \\
&\quad + \int_0^1\int_0^1\int_{-1}^0\int_0^{1+z} ((x_1+y_1)^2+z^2)^{-1-\sigma} dx_2 dz dy_1 dx_1 \\
&= 2\int_0^1\int_0^1\int_0^1\int_0^1 (1-z)((x_1+y_1)^2+z^2)^{-1-\sigma} dx_2 dz dy_1 dx_1 \\
&= 2\int_0^1\int_0^1\int_0^1 h_2(x_1,y_1,z) dz dy_1 dx_1
\end{aligned}
$$

with $h_2(x_1,y_1,z) := (1-z)((x_1+y_1)^2+z^2)^{-1-\sigma}$, where we did formal computations assuming that Fubini's theorem is applicable. This integral is singular if $(x_1,y_1,z)=0$. This singularity of radial type located in the corner of the integration domain $[0,1]^3$ is isolated in a single variable by partitioning $[0,1]^3$ into pyramids and applying a high-dimensional Duffy transformation in each pyramid, which parametrizes each pyramid by the hypercube, see [24, Figure 2, Section 3.5]:

$$
\begin{aligned}
&2\int_0^1\int_0^1\int_0^1 (1-z)((x_1+y_1)^2+z^2)^{-1-\sigma} dz dy_1 dx_1 \\
&= 2\int_0^1\int_0^1\int_0^1 (h_2(s,s\xi_1,s\xi_2)+h_2(s\xi_1,s,s\xi_2)+h_2(s\xi_1,s\xi_2,s)) s^2 d\xi_1 d\xi_2 ds \\
&= 2\int_0^1 s^{-2\sigma} ds (\int_0^1\int_0^1 (\tilde{h}_2(1,\xi_1,\xi_2)+\tilde{h}_2(\xi_1,1,\xi_2)+\tilde{h}_2(\xi_1,\xi_2,1)) d\xi_1 d\xi_2) \\
&\quad + 2\int_0^1 s^{1-2\sigma} ds (\int_0^1\int_0^1 (\hat{h}_2(1,\xi_1,\xi_2)+\hat{h}_2(\xi_1,1,\xi_2)+\hat{h}_2(\xi_1,\xi_2,1)) d\xi_1 d\xi_2) \\
&= \frac{2}{1-2\sigma}\int_0^1\int_0^1 (\tilde{h}_2(1,\xi_1,\xi_2)+\tilde{h}_2(\xi_1,1,\xi_2)+\tilde{h}_2(\xi_1,\xi_2,1)) d\xi_1 d\xi_2 \\
&\quad + \frac{2}{2-2\sigma}\int_0^1\int_0^1 (\hat{h}_2(1,\xi_1,\xi_2)+\hat{h}_2(\xi_1,1,\xi_2)+\hat{h}_2(\xi_1,\xi_2,1)) d\xi_1 d\xi_2
\end{aligned}
$$

with $\tilde{h}_2(\xi_1,\xi_2,\xi_3) = ((\xi_1+\xi_2)^2+\xi_3^2)^{-1-\sigma}$ and $\hat{h}_2(\xi_1,\xi_2,\xi_3) = -\xi_3((\xi_1+\xi_2)^2+\xi_3^2)^{-1-\sigma}$

For $\sigma = \frac{7}{16}$, we obtain with MATLAB

$$\int_0^1 \int_0^1 (\tilde{h}_2(1,\xi_1,\xi_2) + \tilde{h}_2(\xi_1,1,\xi_2) + \tilde{h}_2(\xi_1,\xi_2,1))d\xi_1 d\xi_2$$
$$\approx 2 \cdot 3.0959 \cdot 10^{-1} + 4.2072 \cdot 10^{-1} = 1.0399 \cdot 10^0,$$
$$\int_0^1 \int_0^1 (\hat{h}_2(1,\xi_1,\xi_2) + \hat{h}_2(\xi_1,1,\xi_2) + \hat{h}_2(\xi_1,\xi_2,1))d\xi_1 d\xi_2$$
$$\approx 2 \cdot (-1.3763 \cdot 10^{-1}) - 4.2072 \cdot 10^{-1} = -6.9598 \cdot 10^{-1}.$$

For ③ we have

$$\int_0^1 \int_0^1 \int_{-1}^0 \int_{-1}^0 \|x-y\|^{-2-2\sigma} dy_2 dy_1 dx_2 dx_1$$
$$= \int_0^1 \int_0^1 \int_0^1 \int_0^1 h_3(x_1,x_2,y_1,y_2) dy_2 dy_1 dx_2 dx_1$$

with $h_3(x_1,x_2,y_1,y_2) := ((x_1+y_1)^2 + (x_2+y_2)^2)^{-1-\sigma}$ and thus, using again [24, Section 3.5],

$$\int_0^1 \int_0^1 \int_0^1 \int_0^1 h_3(x_1,x_2,y_1,y_2) dy_2 dy_1 dx_2 dx_1$$
$$= \int_0^1 \int_0^1 \int_0^1 \int_0^1 s^{1-2\sigma}(h_3(1,\xi_1,\xi_2,\xi_3) + h_3(\xi_1,1,\xi_2,\xi_3)$$
$$+ h_3(\xi_1,\xi_2,1,\xi_3) + h_3(\xi_1,\xi_2,\xi_3,1))d\xi_1 d\xi_2 d\xi_3 ds$$
$$= \frac{4}{2-2\sigma} \int_0^1 \int_0^1 \int_0^1 h_3(1,\xi_1,\xi_2,\xi_3) d\xi_1 d\xi_2 d\xi_3.$$

For $\sigma = \frac{7}{16}$, we obtain with MATLAB

$$\int_0^1 \int_0^1 \int_0^1 h_3(1,\xi_1,\xi_2,\xi_3) d\xi_1 d\xi_2 d\xi_3 \approx 2.1065 \cdot 10^{-1}.$$

Hence,

$$② \approx -2h^{2-2\sigma}(\frac{2}{1-2\sigma} 1.0399 \cdot 10^0 - \frac{2}{2-2\sigma} 6.9598 \cdot 10^{-1}),$$
$$③ \approx -2h^{2-2\sigma}(\frac{4}{2-2\sigma} 2.1065 \cdot 10^{-1}).$$

For $\sigma = \frac{7}{16}$, we obtain the approximations for the integrals ④ – ⑬ with MATLAB:

$$④ \approx -2h^{2-2\sigma} 1.6422 \cdot 10^{-1}, \qquad ⑨ \approx -2h^{2-2\sigma} 6.9627 \cdot 10^{-3},$$
$$⑤ \approx -2h^{2-2\sigma} 1.1512 \cdot 10^{-1}, \qquad ⑩ \approx -2h^{2-2\sigma} 2.1142 \cdot 10^{-4},$$
$$⑥ \approx -2h^{2-2\sigma} 4.8272 \cdot 10^{-2}, \qquad ⑪ \approx -2h^{2-2\sigma} 9.2385 \cdot 10^{-4},$$
$$⑦ \approx -2h^{2-2\sigma} 3.5498 \cdot 10^{-2}, \qquad ⑫ \approx -2h^{2-2\sigma} 3.7609 \cdot 10^{-4},$$
$$⑧ \approx -2h^{2-2\sigma} 2.5427 \cdot 10^{-2}, \qquad ⑬ \approx -2h^{2-2\sigma} 1.1380 \cdot 10^{-5}.$$

7.3. Objective Function. We consider the objective function

$$\bar{j}(\rho) := J(\rho, S(\rho)) + \gamma\Upsilon_p(\rho) + \frac{\eta}{2}|||\rho|||^2_{H^\sigma(D)},$$

with

$$J : H^\sigma(D) \times H^1(D)^d \to \mathbb{R},$$
$$J(\rho, u) = \frac{1}{2}\int_D \alpha(\rho) u \cdot u \mathrm{d}\xi + \frac{\mu}{2}\int_D \nabla u : \nabla u \mathrm{d}\xi - \int_D f \cdot u \mathrm{d}\xi,$$

compare (1.2)–(1.6). Using the bilinear forms, the linear and the nonlinear form defined in section 7.1, J can be written as

$$J(\rho, u) = \sum_{i=1}^d \left(\frac{1}{2} r(\rho; u_i, u_i) + \frac{\mu}{2} a(u_i, u_i) - F_i(u_i)\right).$$

The discrete version of the objective function is the following:

$$\bar{J}^h(\rho^h, u^h) := J^h(\rho^h, u^h) + \gamma\Upsilon_p(\rho^h) + \frac{\eta}{2} a_\sigma(\rho^h, \rho^h),$$

$$\begin{aligned} J^h(\rho^h, u^h) &= \sum_{i=1}^d \left(\frac{1}{2} r(\rho^h; u_i^h, u_i^h) + \frac{\mu}{2} a(u_i^h, u_i^h) - F_i^h(u_i)\right) \\ &= \sum_{i=1}^d \left(\frac{1}{2} \mathbf{u}_i^\top R(\boldsymbol{\rho}) \mathbf{u}_i + \frac{\mu}{2} \mathbf{u}_i^\top A \mathbf{u}_i - \mathbf{f}_i^\top \mathbf{u}_i\right), \end{aligned}$$

and, for $p = 2$,

$$\Upsilon_p(\rho^h) = \frac{1}{2}\sum_\ell \max(0, \boldsymbol{\rho}_\ell - 1)^2 \int_D \Phi_\ell(\xi)\mathrm{d}\xi + \frac{1}{2}\sum_\ell \min(0, \boldsymbol{\rho}_\ell + 1)^2 \int_D \Phi_\ell(\xi)\mathrm{d}\xi.$$

7.4. Lagrangian and Adjoint Equation. Let $\lambda_i^h = \sum_k (\boldsymbol{\lambda_i})_k \phi_k \in H^1(D)$ for $i \in \{1, \ldots, d\}$ and $\nu^h = \sum_\ell \boldsymbol{\nu}_\ell \psi_\ell \in L_0^2(D)$. The discretized Lagrangian is given by

$$\begin{aligned} &L^h(\rho^h, w^h, p^h, \lambda^h, \nu^h) \\ &= \bar{J}^h(\rho^h, w^h + u_D^h) + \sum_{i=1}^d \big(\mu a(w_i^h + u_{Di}^h, \lambda_i^h) + r(\rho^h; w_i^h + u_{Di}^h, \lambda_i^h) \\ &\quad - b_i(\lambda_i^h, p^h) + b_i(w_i^h + u_{Di}^h, \nu^h) - F_i^h(\lambda_i^h)\big). \end{aligned}$$

To compute the gradient of the reduced objective function we need the solution of the adjoint equation. The discrete adjoint state is defined by the following equations:

$$\begin{aligned} \langle \frac{d}{dw_i^h} L^h, v_i^h \rangle &= r(\rho^h; w_i^h + u_{Di}^h, v_i^h) + \mu a(w_i^h + u_{Di}^h, v_i^h) - F_i^h(v_i) \\ &\quad + \mu a(v_i^h, \lambda_i^h) + r(\rho^h; v_i^h, \lambda_i^h) + b_i(v_i^h, \nu^h) = 0, \\ \langle \frac{d}{dp^h} L^h, q^h \rangle &= -\sum_{i=1}^d b_i(\lambda_i^h, q^h) = 0, \end{aligned}$$

for all $v_i^h \in S_2^h$, $i \in \{1,\dots,d\}$ and $q^h \in S_1^h$. Written in a matrix-vector form, for $d=2$, the adjoint equation is

$$\begin{aligned}
&R(\boldsymbol{\rho})_{II}(\boldsymbol{\lambda}_1)_I + \mu A_{II}(\boldsymbol{\lambda}_1)_I - B^1_{I\bullet}\boldsymbol{\nu} = (\mathbf{f}_1)_I - (\mu A_{I\bullet} + R(\boldsymbol{\rho})_{I\bullet})(\mathbf{w}_1 + \mathbf{u}_{D1})\\
&R(\boldsymbol{\rho})_{II}(\boldsymbol{\lambda}_2)_I + \mu A_{II}(\boldsymbol{\lambda}_2)_I - B^2_{I\bullet}\boldsymbol{\nu} = (\mathbf{f}_2)_I - (\mu A_{I\bullet} + R(\boldsymbol{\rho})_{I\bullet})(\mathbf{w}_2 + \mathbf{u}_{D2})\\
&\left(B^1_{I\bullet}\right)^\top (\boldsymbol{\lambda}_1)_I + \left(B^2_{I\bullet}\right)^\top (\boldsymbol{\lambda}_2)_I = 0.
\end{aligned} \tag{7.6}$$

For fixed ρ^h, u^h and p^h, the adjoint state (λ^h, ν^h) is the unique solution of these equations if we fix one degree of freedom for the pressure.

7.5. Derivative of the Reduced Objective Function. Since it holds

$$\bar{j}(\rho) = \bar{J}(\rho, u(\rho)) = L(\rho, u(\rho), p(\rho), \lambda, \nu) \quad \forall (\lambda, \nu) \in H^1(D)^d \times L_0^2(D)$$

we choose (λ, ν) as the solution of the adjoint equation such that we get for the derivative of the reduced objective function

$$\bar{j}'(\rho) = \frac{d}{d\rho} L(\rho, u(\rho), p(\rho), \lambda, \nu).$$

Thus, the discrete derivative of the reduced objective function is

$$\begin{aligned}
\langle (\bar{j}^h)'(\rho^h), d^h\rangle &= \langle \frac{d}{d\rho^h} L^h(\rho^h, w^h, p^h, \lambda^h, \nu^h), d^h\rangle = \\
&= \sum_{i=1}^d \frac{1}{2} \langle \frac{d}{d\rho^h} r(\rho^h; w_i^h + u_{Di}^h, w_i^h + u_{Di}^h), d^h\rangle + \frac{\eta}{2}\langle \frac{d}{d\rho^h} a_\sigma(\rho^h, \rho^h), d^h\rangle \\
&\quad + \gamma \langle \frac{d}{d\rho^h} \Upsilon_p(\rho^h), d^h\rangle + \sum_{i=1}^d \langle \frac{d}{d\rho^h} r(\rho^h; w_i^h + u_{Di}^h, \lambda_i^h), d^h\rangle \\
&= \sum_{i=1}^d (\mathbf{w}_i + \mathbf{u}_{Di})^\top \left(R(\boldsymbol{\rho})(\frac{1}{2}(\mathbf{w}_i + \mathbf{u}_{Di}) + \boldsymbol{\lambda_i})\right)_\rho \mathbf{d} + \eta \boldsymbol{\rho}^T M \boldsymbol{d} \\
&\quad + \gamma \sum_\ell (\max(0, \boldsymbol{\rho}_\ell - 1) + \min(0, \boldsymbol{\rho}_\ell + 1)) \mathbf{d}_\ell \int_D \Phi_\ell(\xi) \mathrm{d}\xi.
\end{aligned}$$

The derivative of the nonlinear term r w.r.t. ρ^h can be derived as follows: First, the derivative of of $R_{ij}(\boldsymbol{\rho})$ w.r.t. $\boldsymbol{\rho}_\ell$ is

$$\frac{\partial}{\partial \boldsymbol{\rho}_\ell} R_{ij}(\boldsymbol{\rho}) = \int_D \alpha'(\boldsymbol{\rho}_\ell) \Phi_\ell(\xi) \phi_i(\xi) \phi_j(\xi) \mathrm{d}\xi.$$

Thus, it holds

$$\left(R'(\boldsymbol{\rho})\mathbf{d}\right)_{ij} = \sum_\ell \int_D \alpha'(\boldsymbol{\rho}_\ell) \mathbf{d}_\ell \Phi_\ell(\xi) \phi_i(\xi) \phi_j(\xi) \mathrm{d}\xi$$

and

$$\left((R(\boldsymbol{\rho})\mathbf{w})_\rho \mathbf{d}\right)_i = \sum_\ell \sum_j \int_D \alpha'(\boldsymbol{\rho}_\ell) \mathbf{d}_\ell \Phi_\ell(\xi) \phi_i(\xi) \mathbf{w}_j \phi_j(\xi) \mathrm{d}\xi.$$

We define

$$S(\mathbf{w})_{i\ell} = \int_D \Phi_\ell(\xi)\phi_i(\xi)\Big(\sum_j \mathbf{w}_j\phi_j(\xi)\Big)\mathrm{d}\xi.$$

Then, with $\boldsymbol{\alpha'(\rho)}$ denoting the vector with the components $\alpha'(\boldsymbol{\rho}_\ell)$, we can write

$$\big(R(\boldsymbol{\rho})\mathbf{w}\big)_\rho\mathbf{d} = S(\mathbf{w})\,\mathrm{Diag}(\boldsymbol{\alpha'(\rho)})\mathbf{d}$$

where Diag generates a diagonal matrix from a vector. Hence,

(7.7)
$$\begin{aligned}\langle(\bar{j}^h)'(\rho^h), d^h\rangle &= \sum_{i=1}^{d}(\mathbf{w}_i+\mathbf{u}_{Di})^\top S(\frac{1}{2}(\mathbf{w}_i+\mathbf{u}_{Di})+\boldsymbol{\lambda_i})\,\mathrm{Diag}(\boldsymbol{\alpha'(\rho)})\mathbf{d}+\eta\boldsymbol{\rho}^\top M\mathbf{d}\\ &\quad+\gamma\sum_\ell(\max(0,\boldsymbol{\rho}_\ell-1)+\min(0,\boldsymbol{\rho}_\ell+1))\mathbf{d}_\ell\int_D\Phi_\ell(\xi)\mathrm{d}\xi.\end{aligned}$$

To compute the derivative $(j^h)'(\rho^h)$ one has to determine the solution of the forward problem (7.1) for the $\boldsymbol{\rho}$ corresponding to the given ρ^h to get $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{p}$. Having these solutions at hand, the adjoint equations (7.6) have to be solved to get $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\nu}$. Finally, the derivative $(j^h)'(\rho^h)$ can be determined by inserting the computed values into (7.7).

7.6. Choice of initial value. As already discussed in Remark 5.3, a good initial point for the optimization has an impact on the quality of the solution since many local minima exist and gradient based optimization algorithms typically only yield local solutions. To compute a starting point, we further relax the problem, ignore the simple bound constraint, and reformulate the sphere constraint as inequality constraint such that we have a convex feasible set. Under suitable assumptions, the existence of an optimal solution $\bar{\rho}\in Y$ of the optimization problem

$$\min_{\rho\in Y}\bar{j}(\rho) := j(\rho)+\frac{\eta}{2}\|\rho\|_Y^2,\quad \text{s.t.}\;\; g(\rho)\le 0,\;\int_D(|\rho|^2-1)d\xi\le 0 \tag{7.8}$$

can be shown similarly to Section 3. For linear g, $\bar{\rho}$ is identified with a feasible point of (1.1) by using the following procedure. First determine $\bar{\rho}_0$, the L^2-projection of 0 onto the hyperplane $\bar{\mathcal{H}} := \{\rho \;:\; g(\rho) = g(\bar{\rho})\}$. Then define the initial point $\rho_0 := \bar{\rho}_0 + t(\bar{\rho}-\bar{\rho}_0)$, where $t\ge 1$ is chosen such that $\int_D(|\rho_0|^2-1)d\xi = 0$. Since $\bar{\rho}_0$ is the projection of 0 onto $\bar{\mathcal{H}}$, $\int_D\bar{\rho}_0(\bar{\rho}-\bar{\rho}_0)d\xi = 0$. Hence,

$$\begin{aligned}0 = \int_D\rho_0^2-1d\xi &= \int_D(\bar{\rho}_0+t(\bar{\rho}-\bar{\rho}_0))^2-1d\xi\\ &= \int_D\bar{\rho}_0^2d\xi+t^2(\int_D(\bar{\rho}-\bar{\rho}_0)^2d\xi)-\int_D 1d\xi\end{aligned}$$

with

$$t_{1,2} = \pm\sqrt{\frac{\int_D 1d\xi-\int_D\bar{\rho}_0^2d\xi}{\int_D(\bar{\rho}-\bar{\rho}_0)^2d\xi}}.$$

7.7. Solving the discretized optimization problem using IPOPT. As many other existing implementations of optimization methods, IPOPT [53] assumes that the problem is posed in the Euclidean space. Therefore, directly solving the discretized optimization problem with IPOPT leads to a loss of information since it is no longer taken into account that the control is the discretization of a function with a certain regularity (here H^σ-regularity). The correct discrete inner product for functions $\rho_1(\xi) = \sum_i (\boldsymbol{\rho}_1)_i \Phi_i(\xi)$ and $\rho_2 = \sum_i (\boldsymbol{\rho}_2)_i \Phi_i(x)$ is given by

$$(\rho_1, \rho_2)_{H^\sigma(D)} = (\rho_1, \rho_2)_{L^2(D)} + \langle \rho_1, \rho_2 \rangle_{\kappa,\sigma} = a_\sigma(\rho_1, \rho_2) = \boldsymbol{\rho}_1^\top M \boldsymbol{\rho}_2$$

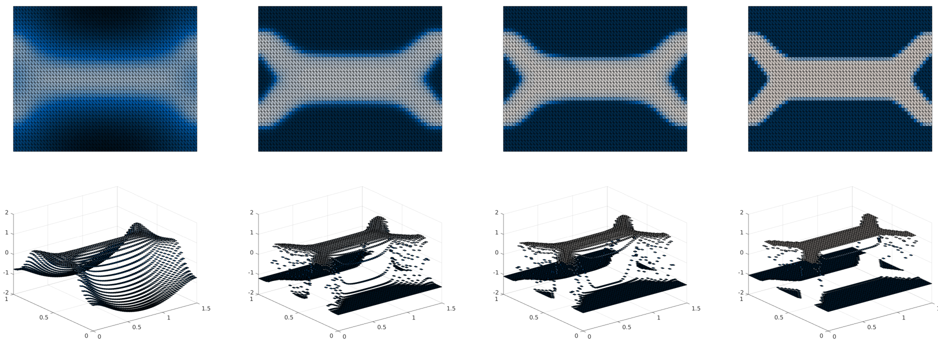
instead of $\boldsymbol{\rho}_1^\top \boldsymbol{\rho}_2$. In order to include this information during the optimization, we work on the space of transformed coordinates

$$\check{\boldsymbol{\rho}} = \check{M} \boldsymbol{\rho},$$

where $\check{M}$ is chosen such that $\check{M}^\top \check{M} = M$. This is, e.g., obtained for $\check{M} = M^{\frac{1}{2}}$ (which is impracticable if the size of M is large) or by a (sparse) Cholesky decomposition, see e.g. [38, Section 5.3.3]. There are other works that use this approach, e.g. [17]. Alternatively, one can also use optimization methods that directly work with the correct inner product, e.g., in the context of the BFGS method, [41, 48].

8. Numerical results. To test our approach numerically, we consider the double pipe example presented in [15, Section 4.5]. The task is to minimize the dissipated power in the fluid, which is modeled by the Stokes equations, for a given inflow and outflow profile. Additionally, we have the constraint that only $\frac{1}{3}$ of the given volume should be filled with fluid. The domain $D = (0, 1.5) \times (0, 1.0)$ is a rectangle in $\mathbb{R}^2$ with length 1.5 and heigth 1.0. Two inlets with center points $(0, \frac{1}{4})^\top$, $(0, \frac{3}{4})^\top$ and width $\ell = \frac{1}{6}$ are located on the left boundary of the domain, and two outlets with center points $(1.5, \frac{1}{4})^\top$, $(1.5, \frac{3}{4})^\top$ and width $\ell = \frac{1}{6}$ are located on the opposite boundary. On each of the four the parabolic flow profile $g(t) = \bar{g}(1 - \frac{2}{\ell}(y - c_y)^2)$ is imposed as Dirichlet boundary condition on the fluid velocity, where $\bar{g} = 1$ and c_y denotes the y-coordinate of the center of the corresponding in- or outlet. On the rest of the boundary no-slip conditions are imposed. As in [15] we choose $\mu = 1$ and $\bar{\alpha} = 25000$. We discretize the domain uniformly with 60×40 (150×100) rectangular cells, i.e. 61×41 (151×101) vertices for the uniform triangular mesh. Hence, $h = 0.025$ ($h = 0.01$).

We implemented the setting described in Lemma 6.4 in MATLAB for $\sigma = \frac{7}{16}$ and with a suitable regularization parameter $\eta = 10$. We have seen in our numerical experiments that a too large or too small choice of the regularization parameter can result in convergence to a different local optimum. Table 8.1 (Table 8.2) gives the number of iterations, the optimal objective function value j, the number of objective function evaluations and the number of gradient evaluations until IPOPT converges with an overall NLP error smaller than 10^{-4}. The initial optimization problem relaxes the sphere constraint to a ball constraint. The solution of this problem is moved onto the sphere as described in Subsection 7.6 in order to obtain an initial guess. Since directly solving with a very large γ yields an ill-conditioned problem, we solve the optimization problems for an increasing sequence of penalty parameters. The solution of the previous optimization problem serves as starting point for the next optimization problem. First, we choose $\gamma = 1000$ and then we increase it twice by a factor 5 (that the last value for γ is 25000 and corresponds to the choice of $\bar{\alpha}$ is coincidence). Figure 8.1 (Figure 8.2) shows the solution of the optimization problems. In the top row one can see the top view of the plots that are presented in the bottom



(a) $\gamma = 0$ on ball (b) $\gamma = 1000$ (c) $\gamma = 5000$ (d) $\gamma = 25000$

Figure 8.1: Optimal solution for a discretization with 60×40 cells

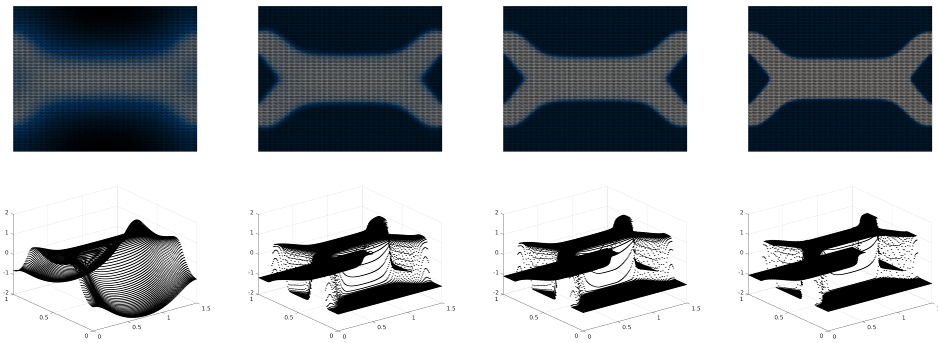
row. The obtained results are virtually identical the results presented in [15]. Also with respect to the iteration numbers our algorithm seems to compare well to the results reported in [15] (which needs 236 iterations). As expected and forced by the penalization term, the smallest and largest values converge to -1 and $+1$. In addition, due to the sphere constraint, the number of cells with values in $(-1, +1)$ decreases. Moreover, as expected for penalty methods, the optimal objective function value j increases for increasing γ.

The choice of the inner product in this example is crucial for obtaining convergence. While using CG1 FEM with H^1-regularization shows good convergence behaviour for computing the initial value (where the sphere constraint is relaxed to a ball constraint and $\gamma = 0$), it shows poor convergence properties with the sphere constraint and $\gamma > 0$. Remark 7.1 discusses a possible reason for this and motivates to consider DG0 FEM. Using L^2-regularization shows poor convergence behavior and oscillatory iterates. H^1-regularization is not available for DG0 FEM since jumps along hypersurfaces are prohibited for H^1-functions. These observations motivate the use of H^σ-regularization.

The approximation of the H^σ-norm is mesh-dependent. We have to keep the considerations in Section 6 in mind if we refine the mesh. Nevertheless, besides the computation of the initial value the iteration numbers of IPOPT seem to be comparable for the presented refinement. This initial guess can also be computed using CG1 FEM on triangles with H^1-regularization and a performing a post-processing step applying a projection onto DG0 FEM on rectangles. The corresponding results are shown in Table 8.3, Table 8.4, Figure 8.3, and Figure 8.4.

Moreover, the approach of finding a good initial point and also the strategy for increaseing the penalization parameter γ presented in this work are heuristics. Even though it works well for the presented example, more sophisticated methods are desirable. Since one is only interested in a good starting point for performing the optimization on the sphere, the optimization on the ball can, e.g., be terminated with a higher tolerance.

9. Conclusion and Outlook. Based on ideas of classical topology optimization and phase field approaches, we presented a novel relaxation of a topology optimization



(a) $\gamma = 0$ on ball (b) $\gamma = 1000$ (c) $\gamma = 5000$ (d) $\gamma = 25000$

Figure 8.2: Optimal solution for a discretization with 150×100 cells

	# iterations	opt. obj. val. j	# obj. eval.	# grad. eval.
$\gamma = 0$ on ball	47	86.79	75	48
$\gamma = 1000$	43	33.00	75	44
$\gamma = 5000$	50	38.00	111	51
$\gamma = 25000$	157	44.04	518	158

Table 8.1: Optimization with IPOPT using a discretzation with 60×40 cells

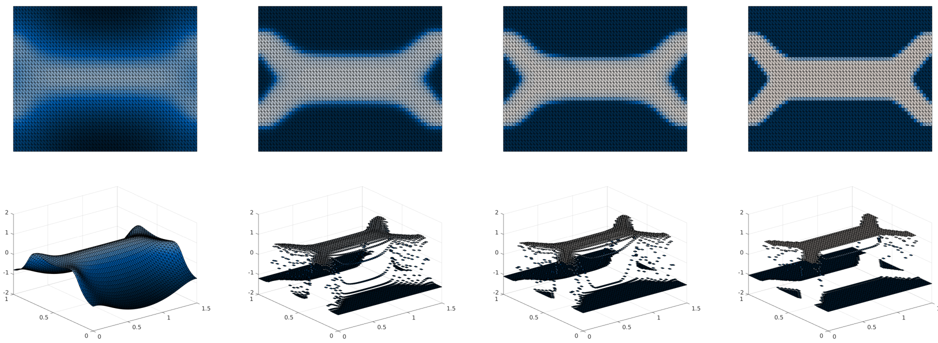
	# iterations	opt. obj. val. j	# obj. eval.	# grad. eval.
$\gamma = 0$ on ball	153	50.16	395	154
$\gamma = 1000$	52	31.86	88	53
$\gamma = 5000$	60	36.19	161	61
$\gamma = 25000$	122	43.20	333	123

Table 8.2: Optimization with IPOPT using a discretization with 150×100 cells

	# iterations	opt. obj. val. j	# obj. eval.	# grad. eval.
$\gamma = 0$ on ball	39	114.75	63	40
$\gamma = 1000$	40	33.00	69	41
$\gamma = 5000$	51	38.00	125	52
$\gamma = 25000$	154	44.04	484	155

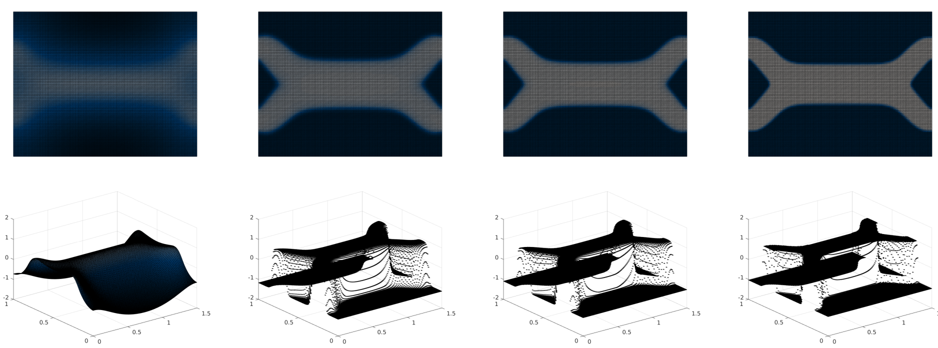
Table 8.3: Optimization with IPOPT using a discretization with 60×40 cells using CG1 FEM and H^1-regularization for the initial problem on the ball

problem for fluid flows. We showed existence of solutions and differentiability results, which allow for the application of gradient based optimization methods. We motivated that it is reasonable to discretize the control with DG0 finite elements. Connections



(a) $\gamma = 0$ on ball (b) $\gamma = 1000$ (c) $\gamma = 5000$ (d) $\gamma = 25000$

Figure 8.3: Optimal solution for a discretization with 60×40 cells using CG1 FEM and H^1-regularization for the initial problem on the ball



(a) $\gamma = 0$ on ball (b) $\gamma = 1000$ (c) $\gamma = 5000$ (d) $\gamma = 25000$

Figure 8.4: Optimal solution for a discretization with 150×100 cells using CG1 FEM and H^1-regularization for the initial problem on the ball

	# iterations	opt. obj. val. j	# obj. eval.	# grad. eval.
$\gamma = 0$ on ball	88	115.30	224	89
$\gamma = 1000$	54	31.86	103	55
$\gamma = 5000$	58	36.19	189	59
$\gamma = 25000$	123	43.20	367	124

Table 8.4: Optimization with IPOPT using a discretization with 150×100 cells using CG1 FEM and H^1-regularization for the initial problem on the ball

between the H^σ- and BV-norm justify the use of a localized H^σ-regularization if σ is adapted to the mesh size. Numerical results show the viability of the proposed method. Even though we focus in the discussion and numerical realization on a steady state Stokes flow and a specific choice of the objective, the conceptual algorithm can be applied also to other state equations and cost functions. Our results provide encouragement to expect that also in other settings it can perform well and be underpinned by an analysis in the spirit developed here. Moreover, examining (adaptive) refinement techniques numerically, and improving the heuristics for the initial guess and the adaption of the penalization parameter are left for future research. It might also be worth investigation to use different optimization algorithms such as optimization on manifolds or augmented Lagrange methods.

Acknowledgments. Johannes Haubner, Franziska Neumann and Michael Ulbrich received support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the International Research Training Group IGDK 1754 "Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures" – Project Number 188264188/GRK1754.

REFERENCES

[1] F. Abraham, M. Behr, and M. Heinkenschloss, *Shape optimization in steady blood flow: A numerical study of non-Newtonian effects*, Computer Methods in Biomechanics and Biomedical Engineering, 8 (2005), pp. 127–137, https://doi.org/10.1080/10255840500180799, https://doi.org/10.1080/10255840500180799.

[2] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229, https://doi.org/10.1088/0266-5611/10/6/003, https://doi.org/10.1088%2F0266-5611%2F10%2F6%2F003.

[3] R. A. Adams and J. J. F. Fournier, *Sobolev spaces*, New York, NY: Academic Press, 2nd ed., 2003.

[4] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve, *Shape optimization by the homogenization method*, Numerische Mathematik, 76 (1997), pp. 27–68.

[5] H. W. Alt, *Linear Functional Analysis - An Application-Oriented Introduction*, Springer, Berlin, Heidelberg, 2016.

[6] H. Amann, *Compact embeddings of vector-valued Sobolev and Besov spaces*, Glas. Mat., III. Ser., 35 (2000), pp. 161–177.

[7] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford: Clarendon Press, 2000.

[8] C. S. Andreasen, A. R. Gersborg, and O. Sigmund, *Topology optimization of microfluidic mixers*, International Journal for Numerical Methods in Fluids, 61 (2009), pp. 498–513, https://doi.org/10.1002/fld.1964, https://doi.org/10.1002/fld.1964.

[9] H. Antil, J. Pfefferer, and S. Rogovs, *Fractional operators with inhomogeneous boundary conditions: analysis, control, and discretization*, Communications in Mathematical Sciences, 16 (2018).

[10] J. Appell and P. P. Zabrejko, *Nonlinear Superposition Operators*, Cambridge University Press, July 1990, https://doi.org/10.1017/cbo9780511897450, https://doi.org/10.1017/cbo9780511897450.

[11] M. Arioli and D. Loghin, *Discrete interpolation norms with applications*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 2924–2951, https://doi.org/10.1137/080729360, https://doi.org/10.1137/080729360.

[12] M. P. Bendsøe and N. Kikuchi, *Generating optimal topologies in structural design using a homogenization method*, Computer methods in applied mechanics and engineering, 71 (1988), pp. 197–224.

[13] M. P. Bendsøe and O. Sigmund, *Topology Optimization*, Springer Berlin Heidelberg, 2004, https://doi.org/10.1007/978-3-662-05086-6, https://doi.org/10.1007/978-3-662-05086-6.

[14] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Berlin Heidelberg, 2013, https://doi.org/10.1007/978-3-642-36519-5, https://doi.org/10.1007/978-3-642-36519-5.

[15] T. Borrvall and J. Petersson, *Topology optimization of fluids in Stokes flow*, International Journal for Numerical Methods in Fluids, 41 (2002), pp. 77–107, https://doi.org/10.1002/fld.426, https://doi.org/10.1002/fld.426.

[16] J. Bourgain, H. Brezis, and P. Mironescu, *Another look at sobolev spaces*, in in Optimal Control and Partial Differential Equations, 2001, pp. 439–455.

[17] C. Brandenburg, F. Lindemann, M. Ulbrich, and S. Ulbrich, *Advanced numerical methods for pde constrained optimization with application to optimal design in navier stokes flow*, in Constrained optimization and optimal control for partial differential equations, Springer, 2012, pp. 257–275.

[18] K. Bredies and M. Holler, *A total variation–based JPEG decompression model*, SIAM Journal on Imaging Sciences, 5 (2012), pp. 366–393, https://doi.org/10.1137/110833531, https://doi.org/10.1137/110833531.

[19] K. Bredies and M. Holler, *A pointwise characterization of the subdifferential of the total variation functional*, 2016, https://arxiv.org/abs/arXiv:1609.08918.

[20] K. Bredies and D. Lorenz, *Mathematical Image Processing*, Springer International Publishing, 2018, https://doi.org/10.1007/978-3-030-01458-2, https://doi.org/10.1007/978-3-030-01458-2.

[21] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer New York, 2010, https://doi.org/10.1007/978-0-387-70914-7, https://doi.org/10.1007/978-0-387-70914-7.

[22] C. Burstedde, *On the numerical evaluation of fractional sobolev norms*, Communications on Pure & Applied Analysis, 6 (2007), p. 587.

[23] E. Casas, K. Kunisch, and C. Pola, *Regularization by functions of bounded variation and applications to image enhancement*, Applied Mathematics and Optimization, 40 (1999), pp. 229–257, https://doi.org/10.1007/s002459900124, https://doi.org/10.1007/s002459900124.

[24] A. Chernov and A. Reinarz, *Numerical quadrature for high-dimensional singular integrals over parallelotopes*, Computers & Mathematics with Applications, 66 (2013), pp. 1213–1231, https://doi.org/10.1016/j.camwa.2013.07.017, https://doi.org/10.1016/j.camwa.2013.07.017.

[25] J. Dávila, *On an open question about functions of bounded variation*, Calculus of Variations and Partial Differential Equations, 15 (2002), pp. 519–527, https://doi.org/10.1007/s005260100135, https://doi.org/10.1007/s005260100135.

[26] M. C. Delfour and J.-P. Zolesio, *Shapes and Geometries - Metrics, Analysis, Differential Calculus, and Optimization, Second Edition*, SIAM, Philadelphia, 2011.

[27] E. Di Nezza, G. Palatucci, and E. Valdinoci, *Hitchhiker's guide to the fractional Sobolev spaces*, Bull. Sci. Math., 136 (2012), pp. 521–573.

[28] Q. Du, M. Gunzburger, R. B. Lehoucq, and K. Zhou, *Analysis and approximation of nonlocal diffusion problems with volume constraints*, SIAM Review, 54 (2012), pp. 667–696, https://doi.org/10.1137/110833294, https://doi.org/10.1137/110833294.

[29] B. Duan, R. D. Lazarov, and J. E. Pasciak, *Numerical approximation of fractional powers of elliptic operators*, IMA Journal of Numerical Analysis, 40 (2019), pp. 1746–1771, https://doi.org/10.1093/imanum/drz013, https://doi.org/10.1093/imanum/drz013.

[30] A. Evgrafov, *On the theoretical background for the topology optimization of navier-stokes equations*, in 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, American Institute of Aeronautics and Astronautics, Aug. 2004, https://doi.org/10.2514/6.2004-4522, https://doi.org/10.2514/6.2004-4522.

[31] A. Evgrafov, *The limits of porous materials in the topology optimization of stokes flows*, Applied Mathematics and Optimization, 52 (2005), pp. 263–277, https://doi.org/10.1007/s00245-005-0828-z, https://doi.org/10.1007/s00245-005-0828-z.

[32] A. Evgrafov, *Topology optimization of slightly compressible fluids*, ZAMM, 86 (2006), pp. 46–62, https://doi.org/10.1002/zamm.200410223, https://doi.org/10.1002/zamm.200410223.

[33] H. Garcke and C. Hecht, *Shape and topology optimization in Stokes flow with a phase field approach*, Applied Mathematics & Optimization, 73 (2016), pp. 23–70, https://doi.org/10.1007/s00245-015-9291-7, https://doi.org/10.1007/s00245-015-9291-7.

[34] H. Garcke, C. Hecht, M. Hinze, and C. Kahle, *Numerical approximation of phase field based shape and topology optimization for fluids*, SIAM Journal on Scientific Computing, 37 (2015), pp. A1846–A1871, https://doi.org/10.1137/140969269, https://doi.org/10.1137/140969269.

[35] H. Garcke, C. Hecht, M. Hinze, C. Kahle, and K. F. Lam, *Shape optimization for surface functionals in navier-stokes flow using a phase field approach*, Interfaces and Free Boundaries, 18 (2016), pp. 219–261, https://doi.org/10.4171/ifb/363, https://doi.org/10.4171/ifb/363.

[36] H. Garcke, M. Hinze, C. Kahle, and K. F. Lam, *A phase field approach to shape optimization in navier–stokes flow with integral state constraints*, Advances in Computa-

tional Mathematics, 44 (2018), pp. 1345–1383, https://doi.org/10.1007/s10444-018-9586-8, https://doi.org/10.1007/s10444-018-9586-8.

[37] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov, and J. Pasciak, *Analysis of numerical methods for spectral fractional elliptic equations based on the best uniform rational approximation*, Journal of Computational Physics, 408 (2020), p. 109285, https://doi.org/10.1016/j.jcp.2020.109285, https://doi.org/10.1016/j.jcp.2020.109285.

[38] J. Haubner, *Shape Optimization for Fluid-Structure Interaction*, dissertation, Technische Universität München, München, 2020.

[39] C. Hecht, *Shape and topology optimization in fluids using a phase field approach and an application in structural optimization*, Mai 2014, https://epub.uni-regensburg.de/29869/.

[40] G. Holler and K. Kunisch, *Learning nonlocal regularization operators*, 2020, https://arxiv.org/abs/arXiv:2001.09092.

[41] C. T. Kelley and E. W. Sachs, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM Journal on Control and Optimization, 25 (1987), pp. 1503–1516, https://doi.org/10.1137/0325083, https://doi.org/10.1137/0325083.

[42] A. Kreuml and O. Mordhorst, *Fractional sobolev norms and BV functions on manifolds*, Nonlinear Analysis, 187 (2019), pp. 450–466, https://doi.org/10.1016/j.na.2019.06.014, https://doi.org/10.1016/j.na.2019.06.014.

[43] M. Kuchta, M. Nordaas, J. C. G. Verschaeve, M. Mortensen, and K.-A. Mardal, *Preconditioners for saddle point systems with trace constraints coupling 2d and 1d domains*, SIAM Journal on Scientific Computing, 38 (2016), pp. B962–B987, https://doi.org/10.1137/15m1052822, https://doi.org/10.1137/15m1052822.

[44] B. Mohammadi and O. Pironneau, *Shape Optimization in Fluid Mechanics*, Annual Review of Fluid Mechanics, 36 (2004), pp. 255–279, https://doi.org/10.1146/annurev.fluid.36.050802.121926, https://doi.org/10.1146/annurev.fluid.36.050802.121926.

[45] F. Murat and J. Simon, *Etude de problèmes d'optimal design*, in Optimization Techniques Modeling and Optimization in the Service of Man Part 2: Proceedings, 7th IFIP Conference Nice, September 8–12, 1975, J. Cea, ed., Springer-Verlag, Berlin, Heidelberg, 1976, pp. 54–62.

[46] I. P. A. Papadopoulos, P. E. Farrell, and T. M. Surowiec, *Computing multiple solutions of topology optimization problems*, 2020, https://arxiv.org/abs/arXiv:2004.11797.

[47] W. Ring, *Structural properties of solutions to total variation regularization problems*, M2AN, Math. Model. Numer. Anal., 34 (2000), pp. 799–810.

[48] T. Schwedes, D. A. Ham, S. W. Funke, and M. D. Piggott, *Mesh dependence in PDE-constrained optimisation*, in Mesh Dependence in PDE-Constrained Optimisation, Springer, 2017, pp. 53–78.

[49] J. Sokolowski and J.-P. Zolesio, *Introduction to Shape Optimization*, Springer Berlin Heidelberg, 1992, https://doi.org/10.1007/978-3-642-58106-9, https://doi.org/10.1007/978-3-642-58106-9.

[50] K. Suzuki and N. Kikuchi, *A homogenization method for shape and topology optimization*, Computer methods in applied mechanics and engineering, 93 (1991), pp. 291–318.

[51] F. Tröltzsch, *Optimal Control of Partial Differential Equations*, American Mathematical Society, Apr. 2010, https://doi.org/10.1090/gsm/112, https://doi.org/10.1090/gsm/112.

[52] M. Ulbrich and S. Ulbrich, *Nichtlineare Optimierung*, Springer Basel, 2012, https://doi.org/10.1007/978-3-0346-0654-7, https://doi.org/10.1007/978-3-0346-0654-7.

[53] A. Wächter and L. Biegler, *On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57.

[54] W. P. Ziemer, *Weakly Differentiable Functions*, Springer New York, 1989, https://doi.org/10.1007/978-1-4612-1015-3, https://doi.org/10.1007/978-1-4612-1015-3.