A semismooth Newton method for a class of semilinear optimal control problems with box and volume constraints

Samuel Amstutz, Antoine Laurain

A SEMISMOOTH NEWTON METHOD FOR A CLASS OF SEMILINEAR OPTIMAL CONTROL PROBLEMS WITH BOX AND VOLUME CONSTRAINTS

SAMUEL AMSTUTZ AND ANTOINE LAURAIN

Abstract. In this paper we consider optimal control problems subject to a semilinear elliptic state equation together with the control constraints $0 \leq u \leq 1$ and $\int u = m$. Optimality conditions for this problem are derived and reformulated as a nonlinear, nonsmooth equation which is solved using a semismooth Newton method. A regularization of the nonsmooth equation is necessary to obtain the superlinear convergence of the semismooth Newton method. We prove that the solutions of the regularized problems converge to a solution of the original problem and a path-following technique is used to ensure a constant decrease rate of the residual. We show that, in certain situations, the optimal controls take $0-1$ values, which amounts to solving a topology optimization problem with volume constraint.

1. Introduction

This paper is dedicated to the numerical solution of minimization problems of the form

$$\min_{(u,y)\in U_{ad}\times Y} J(y) \qquad \text{subject to } E(u,y)=0, \tag{1.1}$$

where $J : Y \to \mathbb{R}$ and $E : U \times Y \to Z$ are appropriate functionals, Y and Z are Banach spaces, the sets U and U_{ad} are defined by

$$U := \{u \in L^2(D),\ 0 \leq u \leq 1 \text{ a.e. in } D\},$$
$$U_{ad} := \left\{ u \in U, \int_D u = m \right\}, \qquad 0 < m < |D|,$$

and D is a bounded domain of $\mathbb{R}^N$, $N \in \{2, 3\}$, with N-dimensional Lebesgue measure $|D|$. In [2] a semismooth Newton method was introduced for a control problem subject to a linear elliptic state equation and an L^1 control cost, with the feature that the control u, a priori searched for within U, eventually takes $0-1$ values. Such a problem is actually a topology optimization problem [1, 4] since u may be written as the characteristic function of a measurable domain $\Omega \subset D$. We speak of topology optimization rather than shape optimization since the topology of Ω is not imposed and may be complex. The control cost $\int_D u$ is interpreted as a volume penalization, which is standard in topology optimization. In the present paper we extend the approach of [2] mainly in two directions. Firstly, the volume term is now treated as an equality constraint instead of a simple penalization. Secondly, we consider a class of semilinear state equations, for which the optimal controls are not necessarily in $0-1$.

Nonsmooth control costs or constraints such as the L^1-norm usually lead to optimal controls whose structure is fundamentally different than when using smooth control costs such as L^p norms with $p > 1$. Nonsmooth control costs have received a great deal of attention recently and have been used for different purposes. The bounded variation norm has been employed primarily in

2010 *Mathematics Subject Classification.* 35J61,49K20,49M05,49M15,49Q10.

Key words and phrases. optimal control, topology optimization, semilinear equation, semismooth Newton method, volume constraint.

image processing and inverse problems [11, 17, 24] in order to preserve sharp edges and recover nonsmooth data. Recently, it has been shown that the L^1-norm [21, 25, 28] or the measure norm of the control [13] provide sparse optimal controls. Sparsity is a property that may be desirable in certain applications where simple structure or easy storage are required for instance. The L^1-norm is also a more natural measure of the cost of the control in some applications. In shape and topology optimization, L^1 or total variation control costs are the natural regularizations as they correspond to volume and perimeter constraints on the geometry, respectively.

Unlike smooth, for instance L^2, regularizations, the treatment of the nonsmooth control cost is technical but nevertheless well-understood nowadays from the theoretical and numerical point of view for linear PDE-constraints. Using convex duality, one considers the predual problem which corresponds to the minimization of a smooth functional with box constraints, for which standards optimization techniques are available [13]. For the numerical solution, a Moreau-Yosida approximation of the predual problem may be employed and can be solved using a semismooth Newton method. A continuation technique is then necessary to obtain the solution of the non-regularized dual problem. Alternatively, the problem can be regularized by adding the L^2-norm of the control to the functional to be minimized, without loosing the sparse properties of the L^1-norm; see [9, 25, 29] for details.

The main contribution of our paper is to develop a fast and efficient algorithm to solve (1.1) when E is nonlinear. In particular we study the case where $E(u, y) = 0$ is a certain class of semilinear equations. Our algorithm is based on a reformulation of the optimality conditions for Problem (1.1) in the form $\Phi(u, y, p, \lambda) = 0$, where (p, λ) are Lagrange multipliers appearing in the optimality conditions and Φ is a nonsmooth, nonlinear vector function. Although the L^1-norm is in our case differentiable due to the box constraint $0 \leq u \leq 1$, this constraint itself leads to a non-smoothness and the generalized Jacobian of Φ exhibit singularities which call for a regularization. The nonlinearity of the state equation does not allow to have a convenient reduced problem formulation where the control is the only variable as in [2], and the problem becomes considerably more involved. To cope with the nonsmoothness of Φ some tools of nonsmooth analysis are needed. In particular we rely here on the use of a semismooth Newton method [15, 23] which exploits generalized differentiability properties of Φ, the so-called *Newton differentiability*, related to the notion of *semismoothness*. In some particular cases which are relevant for applications, we show that we obtain binary solutions, in other words the problem is equivalent to a topology optimization problem. In this case the constraint $\int_D u = m$ allows to exactly control the sparsity of u, whose support decreases with m. In the general case, one cannot expect binary solutions to (1.1). However, we observe in numerical experiments that the optimal control often presents a piecewise constant behavior.

The paper is organized as follows. First of all we write in Section 2 the optimality conditions for the general optimization problem (1.1) under reasonable assumptions on E and J. These conditions are rewritten as a nonlinear, nonsmooth equation. In Section 3 we describe the semismooth Newton method employed to solve the nonlinear equation. We specialize then the problem in Section 4 by considering a semilinear elliptic problem. We prove the superlinear convergence of the semismooth Newton method applied to an approprialety regularized problem, and, at the end of the section, we also prove the convergence of the regularized solutions to the solution of the original problem (1.1). In Section 5 the numerical algorithm is described, and a path-following strategy to steer the regularization parameter so as to ensure a constant decrease rate of some merit function is explained. Finally, numerical results which illustrate both the convergence of the method and the binary or piecewise constant nature of the optimal controls are given in Section 6.

2. Problem statement and optimality conditions

In order to derive optimality conditions in a general setting, we make the following assumptions on the functionals and spaces appearing in Problem (1.1). These assumptions cover a large spectrum of applications.

Assumption 2.1. (a) $J : Y \to \mathbb{R}$ and $E : U \times Y \to Z$ are continuously Fréchet-differentiable and Y, Z are Banach spaces with Y reflexive.
(b) The equation $E(u, y) = 0$ has a single-valued solution operator $u \in V \mapsto y(u) \in Y_{ad}$, where V is a neighborhood of U_{ad} in U and Y_{ad} is a bounded subset of Y.
(c) $(u, y) \in U_{ad} \times Y_{ad} \mapsto E(u, y) \in Z$ is continuous under weak convergence.
(d) The partial Fréchet-derivative of E with respect to y at the point $(u, y(u))$, denoted by $E_y(u, y(u)) \in \mathcal{L}(Y, Z)$, has a bounded inverse for all $u \in V$.
(e) J is sequentially weakly lower semicontinuous.
(f) The partial Fréchet-derivative of E with respect to u, denoted by E_u, can be extended in a continuous linear map from $L^1(D)$ into Z.

Subsequently we denote, given any normed vector space $\mathcal{X}$, by $\mathcal{X}'$ the continuous dual space of $\mathcal{X}$, by $\langle ., . \rangle$ the duality pairing between $\mathcal{X}'$ and $\mathcal{X}$, and by f^* the adjoint of a linear map f. The following result is easily proved by standards arguments of the calculus of variations, see e.g. [18] (Theorem 1.45 and Corollary 1.3).

Theorem 2.2. *Let Assumption 2.1 hold. Then Problem* (1.1) *has an optimal solution* $(\bar{u}, \bar{y})$. *Moreover, there exists* $\bar{p} \in Z'$ *such that*

$$E_y(\bar{u}, \bar{y})^* \bar{p} = -J_y(\bar{y}), \tag{2.1}$$

$$\langle E_u(\bar{u}, \bar{y})^* \bar{p}, u - \bar{u} \rangle \geq 0 \qquad \forall u \in U_{ad}. \tag{2.2}$$

We shall reformulate the conditions (2.1) and (2.2) in a more convenient way. To this aim we introduce the Lagrangian $L : U \times Y \times Z' \to \mathbb{R}$ defined by

$$L(u, y, p) = J(y) + \langle p, E(u, y) \rangle,$$

and whose partial derivatives are

$$L_u(u, y, p) = E_u(u, y)^* p, \tag{2.3}$$

$$L_y(u, y, p) = J_y(y) + E_y(u, y)^* p, \tag{2.4}$$

$$L_p(u, y, p) = E(u, y). \tag{2.5}$$

For every $u \in U_{ad}$ we define the cone $K(u) \subset L^2(D)$ by

$$\forall v \in L^2(D), \qquad v \in K(u) \Longleftrightarrow \begin{cases} v = 0 \text{ a.e. in } [0 < u < 1], \\ v \geq 0 \text{ a.e. in } [u = 0], \\ v \leq 0 \text{ a.e. in } [u = 1]. \end{cases}$$

Theorem 2.3. *Let Assumption 2.1 hold and* $(\bar{u}, \bar{y})$ *be an optimal solution of* (1.1). *Then there exists* $(\bar{\lambda}, \bar{p}) \in \mathbb{R} \times Z'$ *such that*

$$L_u(\bar{u}, \bar{y}, \bar{p}) + \bar{\lambda} \in K(\bar{u}), \tag{2.6}$$

$$L_y(\bar{u}, \bar{y}, \bar{p}) = 0, \tag{2.7}$$

$$L_p(\bar{u}, \bar{y}, \bar{p}) = 0, \tag{2.8}$$

$$\int_D \bar{u} = m. \tag{2.9}$$

Proof. In view of (2.4) and (2.5), the equations (2.7)-(2.9) are straightforward consequences of (2.1) together with the constraints. Therefore we focus on (2.6). For simplicity we set $\bar{g} := L_u(\bar{u}, \bar{y}, \bar{p}) \in L^2(D)$. From (2.2) and (2.3) we infer

$$\langle \bar{g}, u - \bar{u} \rangle \geq 0 \qquad \forall u \in U_{ad}.$$

In other words,

$$\bar{u} \in \operatorname*{argmin}_{u \in U_{ad}} \langle \bar{g}, u - \bar{u} \rangle.$$

By standard Lagrangian duality theory (see e.g. [7, Theorem 3.6]), the constraint $\int_D \bar{u} = m$ can be eliminated by means of a Lagrange multiplier, namely, there exists $\bar{\lambda} \in \mathbb{R}$ such that

$$\bar{g} + \bar{\lambda} \in -\mathcal{N}_U(\bar{u}) \text{ where } \mathcal{N}_U(\bar{u}) := \{v \in L^2(D) : \langle v, u - \bar{u} \rangle \leq 0, \forall u \in U\}$$

is the normal cone of U at $\bar{u}$. According to [7, Lemma 6.34] we actually have $-\mathcal{N}_U(\bar{u}) = K(\bar{u})$ which leads to $\bar{g} + \bar{\lambda} \in K(\bar{u})$. □

In order to reformulate the conditions (2.6)-(2.9) in a tractable way we consider a functional

$$T : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \tag{2.10}$$

$$(s, t) \mapsto T(s, t) \tag{2.11}$$

which satisfies the following assumption.

Assumption 2.4. (a) For all $(s, t) \in [0, 1] \times \mathbb{R}$

$$T(s, t) = 0 \Longleftrightarrow s \in \Theta(t), \tag{2.12}$$

with the set-valued mapping

$$\Theta : t \in \mathbb{R} \mapsto \begin{cases} \{1\} & \text{if } t < 0, \\ [0, 1] & \text{if } t = 0, \\ \{0\} & \text{if } t > 0. \end{cases}$$

(b) The superposition operator made from T maps $L^2(D) \times L^\infty(D)$ onto $L^2(D)$.

We give two examples of functions T satisfying Assumption 2.4. The first one, introduced in [19], is

$$T_{(1)}(s, t) := t - \max(0, t - cs) - \min(0, t - c(s - 1)),$$

for some arbitrary constant $c > 0$. The second one, proposed in [2], is

$$T_{(2)}(s, t) = s \max(0, t) + (1 - s) \min(0, t).$$

Also, for all $(u, y, p, \lambda) \in L^2(D) \times Y \times Z' \times \mathbb{R}$ we set

$$\Phi(u, y, p, \lambda) := \begin{pmatrix} T(u, L_u(u, y, p) + \lambda) \\ L_y(u, y, p) \\ L_p(u, y, p) \\ \int_D u - m \end{pmatrix}.$$

Note that, by Assumption 2.1 we have $L_u(u, y, p) \in L^\infty(D)$ for all $(u, y, p) \in L^2(D) \times Y \times Z'$. Therefore, with Assmption 2.4, Φ maps $L^2(D) \times Y \times Z' \times \mathbb{R}$ into $L^2(D) \times L^2(D) \times Z \times \mathbb{R}$.

Proposition 2.5. *Let $(\bar{u}, \bar{y}, \bar{p}, \bar{\lambda}) \in U \times Y \times Z' \times \mathbb{R}$. The conditions* (2.6)-(2.9) *are equivalent to*

$$\Phi(\bar{u}, \bar{y}, \bar{p}, \bar{\lambda}) = 0.$$

Proof. We only have to prove that $g \in K(u) \Leftrightarrow T(u, g) = 0$, which, by virtue of (2.12), amounts to proving that $g \in K(u) \Leftrightarrow u \in \Theta(g)$. This is a straightforward consequence of the definition of Θ. □

3. Solution strategy

3.1. **Standard results on semismooth Newton methods.** We briefly recall a few useful results concerning semismooth Newton methods [10, 15, 18, 20]. Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces and $\mathcal{U}$ be an open subset of $\mathcal{X}$.

Definition 3.1. A function $F : \mathcal{U} \to \mathcal{Y}$ is called Newton differentiable if there exists a map $G : \mathcal{U} \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$, referred to as Newton derivative, such that

$$\lim_{h\to 0} \frac{1}{\|h\|_{\mathcal{X}}} \|F(u+h) - F(u) - G(u+h)h\|_{\mathcal{Y}} = 0$$

for all $u \in \mathcal{U}$.

Note that the Newton derivative is not necessarily unique. Of course, functions which are $\mathcal{C}^1$ in the sense of Fréchet are Newton differentiable. The following theorem [15, Proposition 4.1] provides another particularly useful example for our purposes.

Theorem 3.2. *The maps* $\max(0,\cdot)$ *and* $\min(0,\cdot) : L^q(D) \to L^p(D)$ *with* $1 \le p < q \le +\infty$ *are Newton differentiable on* $L^q(D)$*, and*

$$\mathscr{G}^+_\varpi : u \mapsto \mathbb{1}_{[u>0]} + \varpi \mathbb{1}_{[u=0]},$$
$$\mathscr{G}^-_\varpi : u \mapsto \mathbb{1}_{[u<0]} + \varpi \mathbb{1}_{[u=0]},$$

are their respective Newton derivatives for any $\varpi \in \mathbb{R}$*.*

The following theorem [15, Theorem 1.1] asserts the local convergence of the semismooth Newton method applied to a Newton differentiable function.

Theorem 3.3. *Suppose that* u^* *solves* $F(u^*) = 0$ *and that* $F : \mathcal{X} \to \mathcal{Y}$ *is Newton differentiable in an open set* $\mathcal{U}$ *containing* u^**, with Newton derivative* G*. If* $G(u)$ *is nonsingular for all* $u \in \mathcal{U}$ *and* $\{\|G(u)^{-1}\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})}, u \in \mathcal{U}\}$ *is bounded, then the Newton iteration*

$$u_{n+1} = u_n - G(u_n)^{-1} F(u_n)$$

converges superlinearly to u^**, provided that* $\|u_0 - u^*\|_{\mathcal{X}}$ *is sufficiently small.*

3.2. **Differentiability properties of the optimality system and regularization.** In specific cases, provided that attention is paid to the choice of the norms, the function Φ will be indeed Newton differentiable. However, we shall see in Section 4 that the Newton derivative of Φ may fail to be invertible. This is typical of the absence of quadratic control cost $\int_D u^2$ in the objective functional or in the constraint [15, 19]. For this reason we regularize Φ by introducing

$$\Phi^\varepsilon(u,y,p,\lambda) := \begin{pmatrix} T^\varepsilon(u, L_u(u,y,p) + \lambda) \\ L_y(u,y,p) \\ L_p(u,y,p) \\ \langle 1, u\rangle - m \end{pmatrix},$$

where the locally Lipschitz function $T^\varepsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is an appropriate regularization of T. Specifically, it is assumed the following.

Assumption 3.4. (a) The superposition operator $T^\varepsilon : L^2(D) \times L^\infty(D) \to L^2(D)$ is well-defined, locally Lipschitz and Newton-differentiable.

(b) There exists a non-increasing function $\theta^\varepsilon \in W^{1,\infty}(\mathbb{R}, [0,1])$ with $\lim_{t\to-\infty} \theta^\varepsilon(t) = 1$ and $\lim_{t\to+\infty} \theta^\varepsilon(t) = 0$ such that, for all $(s,t) \in [0,1] \times \mathbb{R}$,

$$T^\varepsilon(s,t) = 0 \iff s = \theta^\varepsilon(t).$$

In addition, there holds for every $t \in \mathbb{R}$

$$\lim_{\varepsilon \to 0} \theta^\varepsilon(t) = \theta(t) := \begin{cases} 1 & \text{if } t < 0, \\ \theta_0 \in [0,1] & \text{if } t = 0, \\ 0 & \text{if } t > 0, \end{cases}$$

and the convergence is monotone in the sense that, when ε decreases, $\theta^\varepsilon(t)$ is nondecreasing if $t < 0$ and nonincreasing if $t > 0$.

(c) There exists $\alpha_\varepsilon > 0$ such that, for a.e. $(s,t) \in \mathbb{R} \times \mathbb{R}$, the partial derivative of T^ε w.r.t. s satisfies

$$T^\varepsilon_s(s,t) \geq \alpha_\varepsilon.$$

(d) There exists $\beta_\varepsilon > 0$ such that, for a.e. $(s,t) \in \mathbb{R} \times \mathbb{R}$, the partial derivative of T^ε w.r.t. t satisfies

$$T^\varepsilon_t(s,t) \geq -\beta_\varepsilon \operatorname{dist}(s,[0,1]).$$

(e) There exists $a_\varepsilon, b_\varepsilon > 0$ such that, for a.e. $(s,t) \in \mathbb{R} \times \mathbb{R}$,

$$|T^\varepsilon_t(s,t)| \leq a_\varepsilon |s| + b_\varepsilon.$$

(f) For all $\delta > 0$ there exists $\eta > 0$ such that, for a.e. $s \in \mathbb{R}$,

$$\delta \leq \theta^\varepsilon(s) \leq 1 - \delta \Rightarrow (\theta^\varepsilon)'(s) \leq -\eta.$$

(g) For all $\tau > 0$ there exists $k_s, k_t > 0$ such that, for a.e. $(s_1, s_2, t) \in \mathbb{R} \times \mathbb{R} \times [-\tau, \tau]$,

$$\begin{aligned} |T^\varepsilon_s(s_1,t) - T^\varepsilon_s(s_2,t)| &\leq k_s |s_1 - s_2|, \\ |T^\varepsilon_t(s_1,t) - T^\varepsilon_t(s_2,t)| &\leq k_t |s_1 - s_2|. \end{aligned}$$

Let us give some examples.

(1) The standard Tikhonov regularization of (1.1) consists in replacing $J(y)$ by $J_\varepsilon(u,y) := J(y) + \frac{\varepsilon}{2}\|u - \frac{1}{2}\|^2_{L^2}$. The corresponding optimality system is the same as in Proposition 2.5, with $T(s,t)$ replaced by $T\left(s, t + \varepsilon(s - \frac{1}{2})\right)$. For $T = T_{(1)}$ we have

$$T_{(1)}\left(s, t + \varepsilon(s - \frac{1}{2})\right) = t + \varepsilon\left(s - \frac{1}{2}\right) - \max\left(0, t + (\varepsilon - c)s - \frac{\varepsilon}{2}\right) - \min\left(0, t + (\varepsilon - c)s + c - \frac{\varepsilon}{2}\right).$$

We immediately observe that, when choosing $T^\varepsilon_{(1)}(s,t) = T_{(1)}\left(s, t + \varepsilon(s - \frac{1}{2})\right)$, items (a) and (g) of Assumption 3.4 will be satisfied only if $c = \varepsilon$. We then arrive at

$$T^\varepsilon_{(1)}(s,t) = t + \varepsilon\left(s - \frac{1}{2}\right) - \max\left(0, t - \frac{\varepsilon}{2}\right) - \min\left(0, t + \frac{\varepsilon}{2}\right).$$

In this case all the other items of Assumption 3.4 are also fulfilled, with (see Figure 1)

$$\theta^\varepsilon_{(1)}(t) = \frac{1}{2} - \frac{t}{\varepsilon} + \max\left(0, \frac{t}{\varepsilon} - \frac{1}{2}\right) + \min\left(0, \frac{t}{\varepsilon} + \frac{1}{2}\right).$$

Convergence results of the semismooth Newton method applied to the solution of the corresponding system $\Phi^\varepsilon_{(1)}(u,y,p,\lambda) = 0$ for ε fixed and without the constraint $\int_D u = m$ (i.e. with $\lambda = 0$ fixed) are established in [19]. The convergence of the solutions when $\varepsilon \to 0$ is studied in [29] for linear problems including an L^1 control cost.

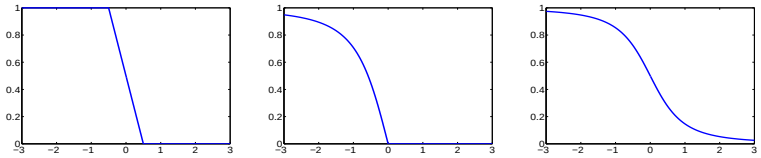

FIGURE 1. From left to right: functions $\theta^\varepsilon_{(1)}$, $\theta^\varepsilon_{(2a)}$ and $\theta^\varepsilon_{(2b)}$ for $\varepsilon = 1$.

(2) Noticing that $T_{(2)}(s,t) = s|t| + \min(0,t)$, it has been proposed in [2] the function

$$T^\varepsilon_{(2a)}(s,t) = s\sqrt{\varepsilon^2 + t^2} + \min(0,t).$$

Of course, many other regularizations of the absolute value would be possible. In order to preserve the symmetric role played by the bounds 0 and 1, noting that $T_{(2)}(s,t) = \frac{t}{2} + (s - \frac{1}{2})|t|$, one may prefer

$$T^\varepsilon_{(2b)}(s,t) = \frac{t}{2} + \left(s - \frac{1}{2}\right)\sqrt{\varepsilon^2 + t^2}.$$

Both functions $T^\varepsilon_{(2a)}$ and $T^\varepsilon_{(2b)}$ satisfy Assumption 3.4, with (see Figure 1)

$$\theta^\varepsilon_{(2a)}(t) = -\frac{\min(0,t)}{\sqrt{\varepsilon^2 + t^2}}, \qquad \theta^\varepsilon_{(2b)}(t) = \frac{1}{2} - \frac{t}{2\sqrt{\varepsilon^2 + t^2}}.$$

Our strategy is to apply the semismooth Newton method to the solution of $\Phi^\varepsilon(u,y,p,\lambda) = 0$, then let ε go to zero by a continuation technique.

4. Study of a semilinear elliptic problem

4.1. **Problem formulation.** Using the framework developed in the previous sections, we specialize to the following spaces

$$Y = (H^2 \cap H^1_0)(D), \qquad Z = (H^2 \cap H^1_0)(D)',$$

and functionals

$$J(y) = \frac{1}{2}\int_D (y - y^\dagger)^2,$$
$$E(u,y) = Ay + \psi(y) - u,$$

where $y^\dagger \in L^2(D)$, A denotes the negative Dirichlet Laplacian on D and the function $\psi : \mathbb{R} \to \mathbb{R}$ is continuous, non-decreasing and has bounded derivatives up to the order 3. More precisely, we have for all $(u,y,z) \in L^2(D) \times H^1_0(D) \times H^1_0(D)$

$$\langle E(u,y), z\rangle = \int_D \nabla y \cdot \nabla z + \psi(y)z - uz.$$

We set

$$M^k_\psi := \|\psi^{(k)}\|_{L^\infty}, \qquad k = 1,2,3. \tag{4.1}$$

We assume that D is of class $\mathcal{C}^2$ or convex. Then, according to classical results on semilinear partial differential equations, see [5, 8, 14] for instance, we get, for all $u \in L^2(D)$, the existence of a unique solution $y(u) \in H^2(D) \cap H^1_0(D)$ to the equation $E(u, y(u)) = 0$. Moreover, the map $u \in L^2(D) \mapsto y(u) \in H^2(D)$ is of class $\mathcal{C}^2$, and we have the following estimate.

Lemma 4.1. *There exist constants $a, b > 0$ such that, for all $f \in L^2(D)$, the solution of*

$$Ay + \psi(y) = f$$

satisfies

$$\|y\|_{H^2} \leq a + b\|f\|_{L^2}.$$

Proof. By [18, Theorem 1.25], we have

$$\|y\|_{H^1} \leq c\|f - \psi(0)\|_{L^2}. \tag{4.2}$$

Next, from

$$Ay = f - \psi(y),$$

we obtain by elliptic regularity, using that D is of class $\mathcal{C}^2$ or convex,

$$\|y\|_{H^2} \leq c\|f - \psi(y)\|_{L^2} \leq c(\|f\|_{L^2} + \|\psi(y)\|_{L^2}), \tag{4.3}$$

where, here and throughout the paper, c denotes a generic positive constant. Using that

$$|\psi(t)| \leq |\psi(0)| + M_\psi^1|t|$$

we infer

$$\|\psi(y)\|_{L^2} \leq c + c\|y\|_{L^2}. \tag{4.4}$$

Combining (4.2)-(4.4) completes the proof. □

We obtain for the Lagrangian and its derivatives:

$$L(u, y, p) = \frac{1}{2}\int_D (y - y^\dagger)^2 + \langle p, Ay + \psi(y) - u\rangle, \tag{4.5}$$

$$L_u(u, y, p) = -p, \tag{4.6}$$

$$L_y(u, y, p) = B(y)p + y - y^\dagger, \tag{4.7}$$

$$L_p(u, y, p) = Ay + \psi(y) - u, \tag{4.8}$$

where

$$B(y) := A + \psi'(y).$$

We have the following useful lemma concerning the continuity of the operator $B(y)^{-1}$.

Lemma 4.2. *For all $(y, f) \in L^1(D) \times L^2(D)$, there exists a unique $z \in (H_0^1 \cap H^2)(D)$ such that $B(y)z = f$. The solution operator mapping f to z, denoted by $B(y)^{-1}$, satisfies*

$$\|B(y)^{-1}(f)\|_{H^2} \leq c\|f\|_{L^2}, \tag{4.9}$$

where c is a constant independent of y.

Proof. The function z must satisfy

$$\int_D \nabla z \cdot \nabla\varphi + \psi'(y)z\varphi = \langle f, \varphi\rangle \qquad \forall\varphi \in H_0^1(D).$$

The bilinear form on the left-hand side of the above equation is clearly continuous on $H_0^1(D) \times H_0^1(D)$, and coercive by virtue of the nonnegativity of ψ' and the Poincaré inequality. The existence and uniqueness of z results from the Lax-Milgram theorem. Due to the assumption $\psi'(y) \geq 0$ we have for some constant $c > 0$

$$c\|z\|_{H^1}^2 \leq \int_D |\nabla z|^2 + \psi'(y)z^2 = \langle f, z\rangle \leq \|f\|_{L^2}\|z\|_{L^2},$$

from which we deduce

$$\|z\|_{H^1} \leq c\|f\|_{L^2}. \tag{4.10}$$

To obtain the H^2 estimate we write

$$Az = f - \psi'(y)z,$$

which implies, using that D is of class $\mathcal{C}^2$ or convex,

$$\|z\|_{H^2} \leq c\|f - \psi'(y)z\|_{L^2} \leq c(\|f\|_{L^2} + M_\psi^1\|z\|_{L^2}).$$

Using (4.10) completes the proof. □

With the help of the above results we straightforwardly check that Assumption 2.1 is fulfilled. Therefore, by Theorem 2.2, we get the existence of optimal solutions. For later purposes, we need another technical assumption.

Assumption 4.3. There exists $\gamma > 0$ such that, for all $(u, y, p) \in U \times (H^2 \cap H_0^1)(D) \times (H^2 \cap H_0^1)(D)$ satisfying

$$Ay + \psi(y) = u, \tag{4.11}$$

$$B(y)p = -(y - y^\dagger), \tag{4.12}$$

there holds

$$1 + \psi''(y)p \geq \gamma.$$

By lemma 4.1 the norm $\|y\|_{H^2}$ is uniformly bounded when $u \in U$. Using Lemma 4.2 and the Sobolev embedding of H^2 into L^∞, we get that the norm $\|p\|_{L^\infty}$ is also uniformly bounded. Therefore, to fulfill Assumption 4.3, one may just require that M_ψ^2 is small enough for instance.

In what follows we denote by $(y(u), p(u))$ the solution (y, p) of (4.11)-(4.12) for a given $u \in L^2(D)$.

4.2. **Binary controls.** In this section we show that, in some important particular cases, the optimal controls necessarily take their values in $\{0, 1\}$. Therefore these problems fall into the framework of topology optimization [1, 4], with the constraint $\int_D u = m$ acting as a volume constraint.

We shall use the following "almost everywhere" definition of the interior:

$$x \in \mathrm{Int}[0 < u < 1] \Longleftrightarrow \exists r > 0 \mid 0 < u(x') < 1 \text{ a.e. } x' \in B(x, r) \cap D.$$

Theorem 4.4. *Suppose that $\psi \equiv 0$ and $-\Delta y^\dagger = 0$ in D. Then every solution (u, y) of (1.1) satisfies*

$$\mathrm{Int}[0 < u < 1] = \emptyset.$$

Proof. Let (u, y) be a solution of (1.1), and assume that $x \in \mathrm{Int}[0 < u < 1]$. By definition there exists $r > 0$ such that

$$0 < u(x') < 1 \text{ a.e. } x' \in B(x, r) \cap D. \tag{4.13}$$

We denote by p and λ the adjoint state and the Lagrange multiplier associated to (u, y), according to Theorem 2.3. Thus we have $-p + \lambda = 0$ a.e. in $B(x, r) \cap D$ in view of (2.6). Yet there holds $-\Delta p + y - y^\dagger = 0$, which implies $y = y^\dagger$ a.e. in $B(x, r) \cap D$. Since $y^\dagger$ is harmonic we also have $-\Delta y = 0$ a.e. in $B(x, r) \cap D$. Then the state equation implies $u = 0$ a.e. in $B(x, r) \cap D$, which contradicts (4.13). □

4.3. **Existence of regularized solutions.** In this section, we prove (Theorem 4.7) the existence of solutions to the equation $\Phi^\varepsilon(u, y, p, \lambda) = 0$. In Lemma 4.5 we provide two useful estimates for the adjoint state $p(u)$. In Lemma 4.6 we show the existence of solutions to the system deprived of the volume constraint, for a fixed Lagrange multiplier λ, and in Theorem 4.7 we show that this volume constraint is achieved for a certain λ.

Lemma 4.5. *Let $u, \bar{u} \in U$ and Assumption 4.3 hold. For all $t \in [0,1]$ set $u_t = \bar{u} + t(u - \bar{u})$, $y_t = y(u_t)$, $p_t = p(u_t)$. Then we have*

$$-\langle p(u) - p(\bar{u}), u - \bar{u}\rangle \geq \gamma \int_0^1 \left\| B(y_t)^{-1}(u - \bar{u})\right\|_{L^2}^2 dt, \tag{4.14}$$

$$\|p(u) - p(\bar{u})\|_{H^1(D)} \leq \beta \int_0^1 \left\| B(y_t)^{-1}(u - \bar{u})\right\|_{L^2} dt, \tag{4.15}$$

where the above constant $\beta > 0$ is independent of u and $\bar{u}$.

Proof. We have already seen that the map $u \in L^2(D) \mapsto y(u) \in H_0^1(D)$ is Fréchet-differentiable. By composition, and using the implicit function theorem, the map $u \in L^2(D) \mapsto p(u) \in H_0^1(D)$ is also Fréchet-differentiable. Differentiating (4.11) in the direction $\delta u \in L^2(D)$ yields

$$B(y)\frac{dy}{du}\delta u = \delta u,$$

and differentiating (4.12) in the direction $\delta y \in H_0^1(D)$ yields

$$B(y)\frac{dp}{dy}\delta y + \psi''(y)p\delta y = -\delta y.$$

Then the chain rule entails

$$\frac{dp}{du}\delta u = \frac{dp}{dy}\left(\frac{dy}{du}\delta u\right) = -B(y)^{-1}\left[1 + \psi''(y)p\right]B(y)^{-1}\delta u. \tag{4.16}$$

We now write

$$p(u) - p(\bar{u}) = \int_0^1 \frac{dp}{du}(\bar{u} + t(u - \bar{u}))(u - \bar{u})dt. \tag{4.17}$$

We have by the Fubini theorem

$$\langle p(u) - p(\bar{u}), u - \bar{u}\rangle = \int_0^1 \left\langle \frac{dp}{du}(\bar{u} + t(u - \bar{u}))(u - \bar{u}), u - \bar{u}\right\rangle dt.$$

Then using (4.16) we get

$$-\langle p(u) - p(\bar{u}), u - \bar{u}\rangle = \int_0^1 \left\langle B(y_t)^{-1}\left[1 + \psi''(y_t)p_t\right]B(y_t)^{-1}(u - \bar{u}), u - \bar{u}\right\rangle dt.$$

Since $B(y_t)$ is self-adjoint we arrive at

$$-\langle p(u) - p(\bar{u}), u - \bar{u}\rangle = \int_0^1 \left\langle \left[1 + \psi''(y_t)p_t\right]B(y_t)^{-1}(u - \bar{u}), B(y_t)^{-1}(u - \bar{u})\right\rangle dt.$$

Using Assumption 4.3 we obtain (4.14). Going back to (4.17), we have

$$\|p(u) - p(\bar{u})\|_{H^1(D)} \leq \int_0^1 \left\| B(y_t)^{-1}\left[1 + \psi''(y_t)p_t\right]B(y_t)^{-1}(u - \bar{u})\right\|_{H^1(D)} dt.$$

By Lemma 4.2 and the uniform boundedness of $\|p_t\|_{L^\infty}$ we obtain (4.15). □

Lemma 4.6. *Let Assumption 4.3 hold. For all $\lambda \in \mathbb{R}$ there exists a unique $u(\lambda) \in L^2(D)$ such that $u(\lambda) = \theta^\varepsilon(-p(u(\lambda)) + \lambda)$. In addition, the map $\lambda \mapsto \int_D u(\lambda)$ is continuous.*

Proof. **Existence.** We fix $\lambda \in \mathbb{R}$. The superposition operator

$$\begin{aligned} \widetilde{\theta^\varepsilon} : L^2(D,[0,1]) &\to L^2(D,[0,1]) \\ u &\mapsto \theta^\varepsilon(-p(u)+\lambda) \end{aligned}$$

is clearly Lipschitz-continuous, as $L^2 \ni u \mapsto p(u) \in L^2$ is itself Lipschitz-continuous and θ^ε is also Lipschitz. In addition, if $u \in L^2(D)$, then $p(u) \in H^1(D)$ and since $[\theta^\varepsilon]'$ is in L^∞ we have

$$\nabla[\widetilde{\theta^\varepsilon}(u)] = -[\theta^\varepsilon]'(-p(u)+\lambda)\nabla p(u) \in L^2(D).$$

Therefore $\widetilde{\theta^\varepsilon}(u) \in H^1(D)$. Furthermore, there exists $c > 0$ such that $\|\widetilde{\theta^\varepsilon}(u)\|_{H^1} \le c$ for all $u \in L^2(D,[0,1])$. It follows by the Rellich theorem that $\widetilde{\theta^\varepsilon}(L^2(D,[0,1]))$ is a relatively compact subset of $L^2(D,[0,1])$. By the Schauder fixed point theorem, there exists $u \in L^2(D,[0,1])$ such that $\widetilde{\theta^\varepsilon}(u) = u$.

Uniqueness. Assume that $\lambda, \bar\lambda \in \mathbb{R}$ and $u, \bar u \in L^2(D)$ satisfy $\theta^\varepsilon(-p(u)+\lambda) = u$ and $\theta^\varepsilon(-p(\bar u)+\bar\lambda) = \bar u$. We have

$$\begin{aligned} -\langle p(u)-p(\bar u), u-\bar u\rangle &= -\langle p(u)-p(\bar u), \theta^\varepsilon(-p(u)+\lambda)-\theta^\varepsilon(-p(\bar u)+\bar\lambda)\rangle \\ &= \langle (-p(u)+\lambda)-(-p(\bar u)+\bar\lambda), \theta^\varepsilon(-p(u)+\lambda)-\theta^\varepsilon(-p(\bar u)+\bar\lambda)\rangle \\ &\quad -\langle \lambda-\bar\lambda, \theta^\varepsilon(-p(u)+\lambda)-\theta^\varepsilon(-p(\bar u)+\bar\lambda)\rangle. \end{aligned}$$

As θ^ε is nonincreasing, the first term is nonpositive. Using also that θ^ε is L_θ-Lipschitz continuous we obtain

$$-\langle p(u)-p(\bar u), u-\bar u\rangle \le L_\theta \|(-p(u)+\lambda)-(-p(\bar u)+\bar\lambda)\|_{L^1}|\lambda-\bar\lambda|.$$

Using the triangle inequality and the Cauchy-Schwarz inequality yields

$$-\langle p(u)-p(\bar u), u-\bar u\rangle \le L_\theta\left(\sqrt{|D|}\|p(u)-p(\bar u)\|_{L^2} + |D||\lambda-\bar\lambda|\right)|\lambda-\bar\lambda|.$$

Using (4.14) and (4.15) from Lemma 4.5 we get

$$\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2}^2 dt \le c_1|\lambda-\bar\lambda|\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2} dt + c_2|\lambda-\bar\lambda|^2$$

for some constants $c_1, c_2 > 0$, possibly depending on ε. By the Cauchy-Schwarz inequality we obtain

$$\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2}^2 dt \le c_1|\lambda-\bar\lambda|\left[\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2}^2\right]^{1/2} dt + c_2|\lambda-\bar\lambda|^2.$$

The Young inequality yields for any $\kappa > 0$

$$\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2}^2 dt \le \frac{c_1\kappa}{2}\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2}^2 dt + \left(\frac{c_1}{2\kappa}+c_2\right)|\lambda-\bar\lambda|^2.$$

Choosing κ small enough we infer the existence of a positive constant c such that

$$\int_0^1 \left\|B(y_t)^{-1}(u-\bar u)\right\|_{L^2}^2 dt \le c|\lambda-\bar\lambda|^2. \tag{4.18}$$

When $\lambda = \bar\lambda$, we derive $B(y_t)^{-1}(u-\bar u) = 0$ for almost every $t \in [0,1]$, and consequently $u = \bar u$.

Continuity. Assume that $\lambda_n \to \bar\lambda$. We have $u_n := u(\lambda_n) \in L^2(D,[0,1])$ for every n, thus the sequence (u_n) is weakly compact in $L^2(D)$. Let $\tilde u \in L^2(D,[0,1])$ be a cluster point of (u_n). There exists a subsequence, not relabeled, such that $u_n \rightharpoonup \tilde u$ weakly in $L^2(D)$. By (4.18), denoting $\bar u := u(\bar\lambda)$, we obtain

$$\int_0^1 \left\|B(y(\bar u + t(u_n-\bar u)))^{-1}(u_n-\bar u)\right\|_{L^2}^2 dt \le c|\lambda_n-\bar\lambda|^2 \to 0.$$

Hence there exists a subsequence, still not relabeled, such that

$$\left\|B(y(\bar{u}+t(u_n-\bar{u})))^{-1}(u_n-\bar{u})\right\|_{L^2}\to 0$$

for almost every $t\in[0,1]$. Thus there exists $t_0\in[0,1]$ such that

$$\left\|B(y_n)^{-1}(u_n-\bar{u})\right\|_{L^2}\to 0, \tag{4.19}$$

with $y_n=y(\bar{u}+t_0(u_n-\bar{u}))$. Since $\|y_n\|_{H^2}$ is bounded, there exists a subsequence and $\tilde{y}\in H^s(D)$, $s<2$, such that $y_n\to\tilde{y}$ in H^s. Therefore, choosing the appropriate s, we may apply Lemma B.1 to obtain, for all $\eta\in L^2(D)$,

$$\begin{aligned}\langle B(y_n)^{-1}(u_n)-B(\tilde{y})^{-1}(\tilde{u}),\eta\rangle &= \langle [B(y_n)^{-1}-B(\tilde{y})^{-1}](u_n)+B(\tilde{y})^{-1}(u_n-\tilde{u}),\eta\rangle\\ &= \langle u_n,[B(y_n)^{-1}-B(\tilde{y})^{-1}]\eta\rangle+\langle u_n-\tilde{u},B(\tilde{y})^{-1}\eta\rangle\\ &\to 0.\end{aligned}$$

Hence $B(y_n)^{-1}(u_n)\rightharpoonup B(\tilde{y})^{-1}(\tilde{u})$ weakly in $L^2(D)$. By compactness of $\{B(y_n)^{-1}(u_n)\}$ in $L^2(D)$, the convergence holds actually strongly. The convergence $B(y_n)^{-1}\bar{u}\to B(\tilde{y})^{-1}\bar{u}$ in $L^\infty(D)$ also follows from Lemma B.1. Using (4.19) we obtain $B(\tilde{y})^{-1}(\tilde{u})=B(\tilde{y})^{-1}(\bar{u})$ and subsequently $\tilde{u}=\bar{u}$. The uniqueness of the cluster point implies that the whole sequence $\{u_n\}$ converges to $\bar{u}$ weakly in $L^2(D)$. We derive straightforwardly that $\int_D u_n\to\int_D\bar{u}$. □

Theorem 4.7. *Let Assumption 4.3 hold. For each $\varepsilon>0$ there exists $(u,y,p,\lambda)\in L^2(D)\times H_0^1(D)\times H_0^1(D)\times\mathbb{R}$ such that $\Phi^\varepsilon(u,y,p,\lambda)=0$. In addition, every such solution belongs to $L^2(D,[0,1])\times(H^2\cap H_0^1)(D)\times(H^2\cap H_0^1)(D)\times\mathbb{R}$.*

Proof. With the notation introduced before, we have

$$\Phi^\varepsilon(u,y,p,\lambda)=0\Longleftrightarrow\left\{\begin{array}{l}\displaystyle\int_D u(\lambda)=m,\\ u=u(\lambda)=\theta^\varepsilon(-p(u(\lambda))+\lambda),\ y=y(u),\ p=p(u).\end{array}\right.$$

In view of Lemma 4.6, for all $\lambda\in\mathbb{R}$, there exists $u(\lambda)\in L^2(D)$ such that $u(\lambda)=\theta^\varepsilon(-p(u(\lambda))+\lambda)$. In addition $\|p(u(\lambda))\|_{L^\infty(D)}$ is uniformly bounded with respect to λ. Thus we have

$$\lim_{\lambda\to-\infty}u(\lambda)=1\text{ and }\lim_{\lambda\to+\infty}u(\lambda)=0\text{ a.e. in }D.$$

By the dominated convergence theorem it follows that

$$\lim_{\lambda\to-\infty}\int_D u(\lambda)=|D|\text{ and }\lim_{\lambda\to+\infty}\int_D u(\lambda)=0.$$

As $0<m<|D|$ and the map $\lambda\mapsto\int_D u(\lambda)$ is continuous, the intermediate value theorem implies the existence of $\lambda\in\mathbb{R}$ such that $\int_D u(\lambda)=m$. □

4.4. **Convergence of the Newton algorithm.** For simplicity we subsequently denote by $\zeta=(u,y,p,\lambda)$ the primal-dual variable. We define the spaces

$$\mathscr{E}:=L^2(D)\times(H^2\cap H_0^1)(D)\times(H^2\cap H_0^1)(D)\times\mathbb{R},$$

$$\mathscr{F}:=L^2(D)\times L^2(D)\times L^2(D)\times\mathbb{R},$$

so that Φ^ε maps $\mathscr{E}$ into $\mathscr{F}$. We endow $\mathscr{E}$ and $\mathscr{F}$ with arbitrary product norms, simply denoted by $\|.\|$ when no confusion is possible. Assumption 3.4 and the chain rule for Newton differentiability

(see e.g. [16]) provide the Newton derivative of Φ^ε

$$D\Phi^\varepsilon(\zeta) = \begin{pmatrix} T_s^\varepsilon + T_t^\varepsilon L_{uu} & T_t^\varepsilon L_{yu} & T_t^\varepsilon L_{pu} & T_t^\varepsilon \\ L_{uy} & L_{yy} & L_{py} & 0 \\ L_{up} & L_{yp} & L_{pp} & 0 \\ \Lambda & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} T_s^\varepsilon & 0 & -T_t^\varepsilon & T_t^\varepsilon \\ 0 & \psi''(y)p+1 & A+\psi'(y) & 0 \\ -1 & A+\psi'(y) & 0 & 0 \\ \Lambda & 0 & 0 & 0 \end{pmatrix}. \tag{4.20}$$

Above Λ denotes the integral operator

$$\Lambda : L^1(D) \ni f \mapsto \int_D f \in \mathbb{R}.$$

One of the main results of our paper is stated in the following theorem, where the local convergence of the Newton algorithm is established.

Theorem 4.8. *Assume that Assumption 4.3 holds. Let $\varepsilon > 0$ be fixed and ζ^ε be a solution of $\Phi^\varepsilon(\zeta^\varepsilon) = 0$. Then the Newton iteration*

$$\zeta_{n+1} = \zeta_n - D\Phi^\varepsilon(\zeta_n)^{-1}\Phi^\varepsilon(\zeta_n) \tag{4.21}$$

is well-defined and converges superlinearly to ζ^ε as long as $\|\zeta_0 - \zeta^\varepsilon\|$ is sufficiently small.

Proof. In order to apply Theorem 3.3 we need to prove the invertibility of the generalized Jacobian $D\Phi^\varepsilon(\zeta) : \mathscr{E} \to \mathscr{F}$ and to obtain a uniform bound on the norm $\|D\Phi^\varepsilon(\zeta)^{-1}\|_{\mathcal{L}(\mathscr{F},\mathscr{E})}$ in a neighborhood of $\zeta^\varepsilon = (u^\varepsilon, y^\varepsilon, p^\varepsilon, \lambda^\varepsilon)$. Let $\zeta = (u, y, p, \lambda) \in \mathscr{E}$ be for the moment arbitrary, and set

$$h := T_s^\varepsilon(u, L_u(u, y, p) + \lambda) = T_s^\varepsilon(u, -p + \lambda) \geq \alpha_\varepsilon, \tag{4.22}$$

$$g := -p + \lambda, \qquad w := \frac{T_t^\varepsilon(u, g)}{h} \in L^2(D).$$

Given an arbitrary right-hand side $(\tilde{u}, \tilde{y}, \tilde{p}, \tilde{\lambda}) \in \mathscr{F}$, we study the solvability of the system

$$D\Phi^\varepsilon(\zeta) \begin{pmatrix} \delta u \\ \delta y \\ \delta p \\ \delta\lambda \end{pmatrix} = \begin{pmatrix} \tilde{u} \\ \tilde{y} \\ \tilde{p} \\ \tilde{\lambda} \end{pmatrix},$$

with unknown $(\delta u, \delta y, \delta p, \delta\lambda) \in \mathscr{E}$. This leads to the following equations

$$\begin{aligned} \delta u - w\delta p + w\delta\lambda &= h^{-1}\tilde{u}, \\ (\psi''(y)p + 1)\delta y + (A + \psi'(y))\delta p &= \tilde{y}, \\ -\delta u + (A + \psi'(y))\delta y &= \tilde{p}, \\ \Lambda(\delta u) &= \tilde{\lambda}. \end{aligned}$$

For simplicity we define the diagonal operator

$$C(y, p) := 1 + \psi''(y)p.$$

Recall that $B(y) = A + \psi'(y)$ is invertible by virtue of Lemma 4.2. Substitution leads to

$$\delta y = B^{-1}(\delta u + \tilde{p}), \tag{4.23}$$

$$\delta p = -B^{-1}CB^{-1}(\delta u + \tilde{p}) + B^{-1}\tilde{y}, \tag{4.24}$$

$$(I + wB^{-1}CB^{-1})\delta u + w\delta\lambda = h^{-1}\tilde{u} - wB^{-1}CB^{-1}\tilde{p} + wB^{-1}\tilde{y}, \tag{4.25}$$

$$\Lambda(\delta u) = \tilde{\lambda}. \tag{4.26}$$

We shall focus on solving Equations (4.25)-(4.26), which are decoupled from (4.23)-(4.24). We begin by studying the operator $I + wB^{-1}CB^{-1}$.

Step 1 (invertibility): We have for all $\varphi \in L^2(D)$

$$\begin{aligned}\langle (I + wB^{-1}CB^{-1})\varphi, B^{-1}CB^{-1}\varphi\rangle &= \langle B^{-1}CB^{-1}\varphi,\varphi\rangle + \langle wB^{-1}CB^{-1}\varphi, B^{-1}CB^{-1}\varphi\rangle \\ &\geq \langle CB^{-1}\varphi, B^{-1}\varphi\rangle - \|\min(0,w)\|_{L^2}\|B^{-1}CB^{-1}\varphi\|^2_{L^4}.\end{aligned}$$

The operator C is diagonal and thus self-adjoint. It is positive definite as well if $\|\zeta - \zeta^\varepsilon\|$ is small enough. To see this, we introduce the sets

$$Y_\varepsilon := \{y \in H^1_0(D) \cap H^2(D), \|y - y_\varepsilon\|_{H^2} \leq M_Y\},$$

$$P_\varepsilon := \{p \in H^1_0(D) \cap H^2(D), \|p - p_\varepsilon\|_{H^2} \leq M_P\},$$

with $M_Y, M_P > 0$ to be fixed later. Thanks to the Sobolev embedding $H^2(D) \subset L^\infty(D)$ which is valid for $N \in \{2,3\}$, $(y,p) \in Y_\varepsilon \times P_\varepsilon$ implies $\|y - y_\varepsilon\|_{L^\infty} \leq cM_Y$ and $\|p - p_\varepsilon\|_{L^\infty} \leq cM_P$, with $c > 0$ depending only on D. Using Assumption 4.3, we have then the estimates

$$\begin{aligned}1 + \psi''(y)p &= 1 + \psi''(y_\varepsilon)p_\varepsilon + [\psi''(y) - \psi''(y_\varepsilon)]p + \psi''(y_\varepsilon)[p - p_\varepsilon] \\ &\geq \gamma - \|\psi''(y) - \psi''(y_\varepsilon)\|_{L^\infty}\|p\|_{L^\infty} - \|\psi''(y_\varepsilon)\|_{L^\infty}\|p - p_\varepsilon\|_{L^\infty} \\ &\geq \gamma - M^3_\psi\|y - y_\varepsilon\|_{L^\infty}\|p\|_{L^\infty} - M^2_\psi\|p - p_\varepsilon\|_{L^\infty} \\ &\geq \gamma - M^3_\psi\|y - y_\varepsilon\|_{L^\infty}(\|p_\varepsilon\|_{L^\infty} + \|p - p_\varepsilon\|_{L^\infty}) - M^2_\psi\|p - p_\varepsilon\|_{L^\infty} \\ &\geq \gamma - M^3_\psi cM_Y(\|p_\varepsilon\|_{L^\infty} + cM_P) - M^2_\psi cM_P.\end{aligned}$$

Therefore, when M_Y and M_P are chosen sufficiently small, we have

$$1 + \psi''(y)p \geq c > 0 \qquad \forall (y,p) \in M_Y \times M_P,$$

and thus, assuming henceforth that $(y,p) \in M_Y \times M_P$, C is positive definite. We can then define the squareroot $C^{1/2}$ of C which is also self-adjoint and write

$$\langle (I + wB^{-1}CB^{-1})\varphi, B^{-1}CB^{-1}\varphi\rangle \geq \|C^{1/2}B^{-1}\varphi\|^2_{L^2} - \|\min(0,w)\|_{L^2}\|B^{-1}CB^{-1}\varphi\|^2_{L^4}.$$

Next we utilize the estimate

$$\|B^{-1}CB^{-1}\varphi\|_{L^4} \leq \|B^{-1}C^{1/2}\|_{\mathcal{L}(L^2,L^4)}\|C^{1/2}B^{-1}\varphi\|_{L^2}.$$

Going back to the main inequality we obtain

$$\langle (I + wB^{-1}CB^{-1})\varphi, B^{-1}CB^{-1}\varphi\rangle \geq (1 - \|\min(0,w)\|_{L^2}\|B^{-1}C^{1/2}\|^2_{\mathcal{L}(L^2,L^4)})\|C^{1/2}B^{-1}\varphi\|^2_{L^2}.$$

Using Lemma 4.2 and the above considerations on the uniform boundedness of C, we have $\|B^{-1}C^{1/2}\|_{\mathcal{L}(L^2,L^4)} \leq M_C$ for some constant M_C. Therefore, whenever $(w,p) \in W^- \times P_\varepsilon$ with

$$W^- := \{w \in L^2(D), \|\min(0,w)\|_{L^2} \leq M_{W^-}\},$$

and $0 < M_{W^-} < (M_C)^{-2}$, the operator $I + wB^{-1}CB^{-1} : L^2(D) \to L^2(D)$ is injective, and subsequently invertible by virtue of the Fredholm alternative.

Step 2 (collective compactness): We first examine under which condition on ζ we have $w \in W^-$. Let $u \in L^2(D)$. In view of Assumption 3.4(d) we have

$$T^\varepsilon_t(u,g) \geq -\beta_\varepsilon \operatorname{dist}(u,[0,1]).$$

Using Theorem 4.7 which asserts that $0 \leq u^\varepsilon \leq 1$, we obtain that $\operatorname{dist}(u,[0,1]) \leq |u - u^\varepsilon|$ almost everywhere. This yields that

$$w \geq -\frac{\beta_\varepsilon}{\alpha_\varepsilon}|u - u^\varepsilon| \text{ a.e. in } D,$$

and subsequently

$$\|\min(0,w)\|_{L^2} \leq \frac{\beta_\varepsilon}{\alpha_\varepsilon}\|u - u^\varepsilon\|_{L^2}. \tag{4.27}$$

We define the set

$$U_\varepsilon := \{u \in L^2(D), \|u - u^\varepsilon\|_{L^2} \le \frac{\alpha_\varepsilon}{\beta_\varepsilon} M_{W^-}\}, \tag{4.28}$$

so that

$$u \in U_\varepsilon \Longrightarrow w \in W^-.$$

Furthermore, there holds for all $u \in U_\varepsilon$, using assumption 3.4(e)

$$\begin{aligned} \|w\|_{L^2} &\le \frac{1}{\alpha_\varepsilon}\|T_t^\varepsilon(u,g)\|_{L^2} \le \frac{1}{\alpha_\varepsilon}(a_\varepsilon\|u\|_{L^2} + b_\varepsilon\|1\|_{L^2}) \\ &\le \frac{1}{\alpha_\varepsilon}(a_\varepsilon\|u^\varepsilon\|_{L^2} + a_\varepsilon\|u - u^\varepsilon\|_{L^2} + b_\varepsilon\|1\|_{L^2}) \\ &\le \frac{a_\varepsilon + b_\varepsilon}{\alpha_\varepsilon}\sqrt{|D|} + \frac{a_\varepsilon}{\beta_\varepsilon} M_{W^-} =: M_{W^+}. \end{aligned}$$

We then define

$$W_\varepsilon^+ := \{w \in L^2(D), \|w\|_{L^2} \le M_{W^+}\}, \qquad W_\varepsilon := W_\varepsilon^+ \cap W^-,$$

so that

$$u \in U_\varepsilon \Longrightarrow w \in W_\varepsilon.$$

We now introduce the operator

$$K(y,p,w) : \varphi \in L^2 \mapsto wB(y)^{-1}C(y,p)B(y)^{-1}\varphi \in L^2,$$

whose adjoint is

$$K(y,p,w)^* : \varphi \in L^2 \mapsto B(y)^{-1}C(y,p)B(y)^{-1}(w\varphi) \in L^2.$$

Here, $B(y)^{-1}$ denotes in fact the adjoint of $B(y)^{-1} : L^2 \to H^2 \cap H_0^1$, which in particular defines a compact operator from L^1 into L^2. The same notation has been kept since it is an extension of $B(y)^{-1}$. We define the set of operators

$$\mathcal{K} := \{K(y,p,w)^*, w \in W_\varepsilon, y \in Y_\varepsilon, p \in P_\varepsilon\}.$$

We obtain for all $(w,y,p) \in W_\varepsilon \times Y_\varepsilon \times P_\varepsilon$

$$\|K(y,p,w)^*\varphi\|_{H^1} \le c\|[C(y,p)B(y)^{-1}](w\varphi)\|_{L^2} \le c\|B(y)^{-1}(w\varphi)\|_{L^2} \le c\|w\varphi\|_{L^1} \le c\|\varphi\|_{L^2}. \tag{4.29}$$

This implies by the Rellich theorem that $\mathcal{K}$ is collectively compact; see Appendix A.

Step 3 (uniform bound on the inverse operator): We now check the remaining hypothesis of Theorem A.3, i.e., the pointwise sequential compactness of $\mathcal{K}$. Let (w_n, y_n, p_n) be a sequence of $W_\varepsilon \times Y_\varepsilon \times P_\varepsilon$. Since W_ε is bounded, convex and closed in $L^2(D)$, there exists a subsequence, not relabeled, such that $w_n \rightharpoonup w \in W_\varepsilon$ weakly in $L^2(D)$. By compact Sobolev embedding, for any $s < 2$, we have for subsequences

$$(y_n, p_n) \to (y,p) \text{ strongly in } H^s \times H^s.$$

Choosing s appropriately we get $y_n \to y$ in L^∞. Applying Lemma B.1 leads to

$$\forall \eta \in L^2(D),\ B(y_n)^{-1}\eta = [A + \psi'(y_n)]^{-1}\eta \to [A + \psi'(y)]^{-1}\eta := \bar{B}^{-1}\eta \qquad \text{in } L^\infty(D). \tag{4.30}$$

For all $(\varphi, \eta) \in L^2 \times L^2$ we have

$$\begin{aligned} \langle B(y_n)^{-1}(w_n\varphi) - \bar{B}^{-1}(w\varphi), \eta\rangle &= \langle [B(y_n)^{-1} - \bar{B}^{-1}](w_n\varphi) + \bar{B}^{-1}((w_n - w)\varphi), \eta\rangle \\ &= \langle w_n\varphi, [B(y_n)^{-1} - \bar{B}^{-1}]\eta\rangle + \langle w_n - w, \varphi\bar{B}^{-1}\eta\rangle \\ &\to 0. \end{aligned}$$

Hence $B(y_n)^{-1}(w_n\varphi) \rightharpoonup \bar{B}^{-1}(w\varphi)$ weakly in L^2. By compactness of $\{B(y_n)^{-1}(w_n\varphi)\}$ in L^2, the convergence holds actually strongly.

We have trivially the convergence in operator norm

$$C(y_n,p_n) = I + \psi''(y_n)p_n \to I + \psi''(y)p =: \bar{C} \qquad \text{in } \mathcal{L}(L^2,L^2).$$

Let us fix an arbitrary $\varphi \in L^2$. We have

$$z_n := C(y_n,p_n)B(y_n)^{-1}(w_n\varphi) \to \bar{C}\bar{B}^{-1}(w\varphi) =: z \text{ in } L^2.$$

From $K(y_n,p_n,w_n)^*\varphi = B(y_n)^{-1}z_n$ we write

$$\begin{aligned}\|K(y_n,p_n,w_n)^*\varphi - \bar{B}^{-1}z\|_{L^\infty} &\leq \|B(y_n)^{-1}(z_n-z)\|_{L^\infty} + \|(B(y_n)^{-1}-\bar{B}^{-1})z\|_{L^\infty}\\ &\leq \|B(y_n)^{-1}\|_{\mathcal{L}(L^2,L^\infty)}\|z_n-z\|_{L^2} + \|(B(y_n)^{-1}-\bar{B}^{-1})z\|_{L^\infty}.\end{aligned}$$

By Lemma B.1 and the Banach-Steinhaus theorem, $\|B(y_n)^{-1}\|_{\mathcal{L}(L^2,L^\infty)}$ is uniformly bounded, hence, using also (4.30),

$$K(y_n,p_n,w_n)^*\varphi \to \bar{B}^{-1}z = \bar{B}^{-1}\bar{C}\bar{B}^{-1}(w\varphi) \qquad \text{in } L^\infty. \tag{4.31}$$

We have seen that $I + wB^{-1}CB^{-1}$ is invertible for every $w \in W_\varepsilon$, therefore $I + K(y,p,w)^*$ is also invertible and Theorem A.3 provides

$$\sup_{(w,y,p)\in W_\varepsilon\times Y_\varepsilon\times P_\varepsilon} \|(I+K(y,p,w)^*)^{-1}\|_{\mathcal{L}(L^2,L^2)} < +\infty.$$

Passing to the adjoint yields

$$\sup_{(w,y,p)\in W_\varepsilon\times Y_\varepsilon\times P_\varepsilon} \|(I+K(y,p,w))^{-1}\|_{\mathcal{L}(L^2,L^2)} < +\infty.$$

In other words, there exists $\tau > 0$ such that

$$\|(I + wB(y)^{-1}C(y,p)B(y)^{-1})^{-1}\|_{\mathcal{L}(L^2,L^2)} \leq \tau \qquad \forall (w,y,p) \in W_\varepsilon \times Y_\varepsilon \times P_\varepsilon. \tag{4.32}$$

Step 4 (uniform bound on the Jacobian): From (4.25) and the invertibility of $I+K$ we obtain

$$\delta u = -\delta\lambda(I+K)^{-1}w + (I+K)^{-1}\left(h^{-1}\tilde{u} - K\tilde{p} + wB^{-1}\tilde{y}\right),$$

and using (4.26)

$$\delta\lambda\int_D (I+K)^{-1}w = -\tilde{\lambda} + \int_D (I+K)^{-1}\left(h^{-1}\tilde{u} - K\tilde{p} + wB^{-1}\tilde{y}\right). \tag{4.33}$$

In order to obtain $\delta\lambda$, we need to show that

$$\mathcal{I}(w) := \int_D (I+K)^{-1}w$$

is nonzero. More precisely we look for a uniform lower bound for $\mathcal{I}(w)$ when ζ is close enough to ζ^ε. We write

$$\begin{aligned}\mathcal{I}(w) &= \langle (I + wB^{-1}CB^{-1})^{-1}w, 1\rangle\\ &= \langle w, (I + B^{-1}CB^{-1}(w\cdot))^{-1}1\rangle,\end{aligned}$$

and we set

$$\xi := (I + B^{-1}CB^{-1}(w\cdot))^{-1}1, \qquad \text{i.e.} \qquad \xi + B^{-1}CB^{-1}(w\xi) = 1, \qquad \xi \in L^2(D).$$

Therefore we have

$$\begin{aligned}\mathcal{I}(w) &= \langle w,\xi\rangle = \langle w\xi, 1\rangle\\ &= \langle w\xi,\xi\rangle + \langle B^{-1}CB^{-1}(w\xi), w\xi\rangle\\ &= \int_D w\xi^2 + \|C^{1/2}B^{-1}(w\xi)\|_{L^2}^2.\end{aligned}$$

We now use

$$\|1-\xi\|_{L^2} = \|B^{-1}CB^{-1}(w\xi)\|_{L^2} \le \|B^{-1}C^{1/2}\|_{\mathcal{L}(L^2,L^2)}\|C^{1/2}B^{-1}(w\xi)\|_{L^2},$$

which entails

$$\mathcal{I}(w) \ge \int_D w\xi^2 + \|B^{-1}C^{1/2}\|^{-2}_{\mathcal{L}(L^2,L^2)}\|1-\xi\|^2_{L^2}.$$

By Lemma 4.2 there exists a constant $\mu > 0$ such that $\|B^{-1}C^{1/2}\|^{-2}_{\mathcal{L}(L^2,L^2)} \ge \mu$ therefore

$$\mathcal{I}(w) \ge \int_D w\xi^2 + \mu(1-\xi)^2\,dx.$$

For all $(s,t) \in \mathbb{R}\times\mathbb{R}$ we set

$$\mathscr{W}^\varepsilon(s,t) = \frac{T^\varepsilon_t(s,t)}{T^\varepsilon_s(s,t)}.$$

When ζ is in a neighborhood of ζ^ε, $\|g\|_{L^\infty}$ remains bounded, say $\|g\|_{L^\infty} \le \tau$. Using Assumption 3.4 we obtain that, whenever $|t| \le \tau$,

$$|\mathscr{W}^\varepsilon(s_1,t) - \mathscr{W}^\varepsilon(s_2,t)| \le \frac{k_s}{\alpha^2_\varepsilon}(a_\varepsilon|s_2| + b_\varepsilon)|s_1-s_2| + \frac{k_t}{\alpha_\varepsilon}|s_1-s_2|.$$

This implies by substitution

$$\|\mathscr{W}^\varepsilon(u,g) - \mathscr{W}(u^\varepsilon,g)\|_{L^2} \le \frac{k_s}{\alpha^2_\varepsilon}\|a_\varepsilon|u^\varepsilon| + b_\varepsilon\|_{L^\infty}\|u-u^\varepsilon\|_{L^2} + \frac{k_t}{\alpha_\varepsilon}\|u-u^\varepsilon\|_{L^2}.$$

As $0 \le u^\varepsilon \le 1$ we have

$$\|\mathscr{W}^\varepsilon(u,g) - \mathscr{W}(u^\varepsilon,g)\|_{L^2} \le k_{\mathscr{W}}\|u-u^\varepsilon\|_{L^2},$$

for some $k_{\mathscr{W}} > 0$. Now, we use that

$$\mathscr{W}^\varepsilon(u^\varepsilon,g) = \mathscr{W}^\varepsilon(\theta^\varepsilon(g^\varepsilon),g).$$

Arguing as previously we get

$$\|\mathscr{W}^\varepsilon(\theta^\varepsilon(g^\varepsilon),g) - \mathscr{W}^\varepsilon(\theta^\varepsilon(g),g)\|_{L^2} \le k_{\mathscr{W}}\|\theta^\varepsilon(g^\varepsilon) - \theta^\varepsilon(g)\|_{L^2},$$

and since θ^ε is Lipschitz of constant $k_\theta > 0$,

$$\|\mathscr{W}^\varepsilon(\theta^\varepsilon(g^\varepsilon),g) - \mathscr{W}^\varepsilon(\theta^\varepsilon(g),g)\|_{L^2} \le k_{\mathscr{W}}k_\theta\|g^\varepsilon - g\|_{L^2}.$$

We arrive at

$$w = \mathscr{W}^\varepsilon(u,g) = \bar{w} + R$$

with $\bar{w} = \mathscr{W}^\varepsilon(\theta^\varepsilon(g),g)$ and

$$\|R\|_{L^2} \le k_{\mathscr{W}}\|u-u^\varepsilon\|_{L^2} + k_{\mathscr{W}}k_\theta\|g^\varepsilon - g\|_{L^2} \le k_\zeta\|\zeta-\zeta^\varepsilon\|. \tag{4.34}$$

Next we write the decomposition

$$\mathcal{I}(w) \ge \int_D \bar{w}\xi^2 + \mu(1-\xi)^2 + \int_D R\xi^2. \tag{4.35}$$

As $0 < \int_D u^\varepsilon = m < |D|$ and $u^\varepsilon = \theta^\varepsilon(g^\varepsilon)$ is continuous, there exists $\bar{x} \in D$ such that $0 < \theta^\varepsilon(g^\varepsilon(\bar{x})) < 1$. By continuity, there exists $\delta > 0$ and a neighborhood ω of $\bar{x}$ such that

$$\delta \le \theta^\varepsilon(g^\varepsilon(x)) \le 1-\delta \qquad \forall x \in \omega.$$

As θ^ε is Lipschitz, we also have for $\|g-g^\varepsilon\|_{L^\infty}$ sufficiently small

$$\delta/2 \le \theta^\varepsilon(g(x)) \le 1-\delta/2 \qquad \forall x \in \omega.$$

By Assumption 3.4(f), there exists $\eta > 0$ such that

$$(\theta^\varepsilon)'(g(x)) \le -\eta \qquad \forall x \in \omega.$$

On differentiating with respect to t the equality $T^\varepsilon(\theta^\varepsilon(t),t)=0$ we derive

$$\mathscr{W}^\varepsilon(\theta^\varepsilon(t),t)=-(\theta^\varepsilon)'(t)\qquad\forall t\in\mathbb{R}.$$

This entails $\bar{w}=-(\theta^\varepsilon)'(g)\geq 0$ and also $\bar{w}(x)\geq\eta$ for all $x\in\omega$. Therefore we have

$$\int_D \bar{w}\xi^2+\mu(1-\xi)^2\geq\int_\omega \eta\xi^2+\mu(1-\xi)^2.$$

We easily show that for all $\xi\in\mathbb{R}$,

$$\eta\xi^2+\mu(1-\xi)^2\geq\frac{\eta\mu}{\eta+\mu},$$

whereby

$$\int_D \bar{w}\xi^2+\mu(1-\xi)^2\geq|\omega|\frac{\eta\mu}{\eta+\mu}.$$

As to the second integral in (4.35) we have by the Cauchy-Schwarz inequality

$$\left|\int_D R\xi^2\right|\leq\|R\|_{L^2}\|\xi\|_{L^4}^2.$$

In view of (4.29) and (4.31), the set of operators $\mathcal{K}$ is also collectively compact and pointwise sequentially compact in $\mathcal{L}(L^4,L^4)$, thus, arguing as in Step 3, we have that

$$\sup_{(w,y,p)\in W_\varepsilon\times Y_\varepsilon\times P_\varepsilon}\|(I+K(y,p,w)^*)^{-1}\|_{\mathcal{L}(L^4,L^4)}<+\infty,$$

which yields $\|\xi\|_{L^4}\leq\sigma$ for some constant $\sigma>0$. Using (4.34) we infer

$$\left|\int_D R\xi^2\right|\leq\sigma^2 k_\zeta\|\zeta-\zeta^\varepsilon\|.$$

Altogether we arrive at

$$\mathcal{I}(w)\geq|\omega|\frac{\eta\mu}{\eta+\mu}-\sigma^2k_\zeta\|\zeta-\zeta^\varepsilon\|.$$

Therefore, there exist $\beta,\nu>0$ such that

$$\|\zeta-\zeta^\varepsilon\|\leq\beta\Rightarrow\mathcal{I}(w)\geq\nu.$$

Suppose now that $\|\zeta-\zeta^\varepsilon\|\leq\beta$. From (4.33) we get

$$\delta\lambda=\mathcal{I}(w)^{-1}\left(-\tilde{\lambda}+\int_D(I+K)^{-1}\left(h^{-1}\tilde{u}-K\tilde{p}+wB^{-1}\tilde{y}\right)\right).$$

Then from (4.25), (4.23) and (4.24), respectively, we derive explicit expressions for δu, δy and δp. This means that $D\Phi^\varepsilon(\zeta)$ is invertible. In addition, we obtain by the Cauchy-Schwarz inequality

$$|\delta\lambda|\leq\nu^{-1}\left(|\tilde{\lambda}|+\|(I+K)^{-1}(h^{-1}\tilde{u}-K\tilde{p}+wB^{-1}\tilde{y})\|_{L^2}\right).$$

Then using (4.32) we get

$$|\delta\lambda|\leq\nu^{-1}\left(|\tilde{\lambda}|+\tau\|(h^{-1}\tilde{u}-K\tilde{p}+wB^{-1}\tilde{y})\|_{L^2}\right)\leq c(|\tilde{\lambda}|+\|\tilde{u}\|_{L^2}+\|\tilde{p}\|_{L^2}+\|\tilde{y}\|_{L^2}).$$

We deduce straightforwardly using (4.25), (4.23) and (4.24) that

$$\|(\delta u,\delta y,\delta p,\delta\lambda)\|\leq c\|(\tilde{u},\tilde{y},\tilde{p},\tilde{\lambda})\|,$$

which in turn implies

$$\|D\Phi^\varepsilon(\zeta)^{-1}\|_{\mathcal{L}(\mathscr{F},\mathscr{E})}\leq c,$$

where c is a positive constant which may depend on ε. □

4.5. **Convergence of the regularized solutions.** In this section we study the convergence of the regularized solution $\zeta^\varepsilon = (u^\varepsilon, y^\varepsilon, p^\varepsilon, \lambda^\varepsilon)$ as $\varepsilon \to 0$.

Theorem 4.9. *Let $\{\varepsilon_k\}_{k\in\mathbb{N}}$ be a sequence of positive numbers such that $\varepsilon_k \to 0$ as $k \to \infty$. Denote by $\zeta^{\varepsilon_k} = (u^{\varepsilon_k}, y^{\varepsilon_k}, p^{\varepsilon_k}, \lambda^{\varepsilon_k})$ a solution of $\Phi^{\varepsilon_k}(\zeta^{\varepsilon_k}) = 0$. Then*

(1) For any $s < 2$ there exists a subsequence $\{\varepsilon_{k_l}\}_{l\in\mathbb{N}}$ and $(u^*, \lambda^*) \in L^2(D, [0,1]) \times \mathbb{R}$ such that

$$u^{\varepsilon_{k_l}} \rightharpoonup u^* \quad \text{weakly in } L^2(D), \qquad y^{\varepsilon_{k_l}} \to y^* \quad \text{strongly in } H^s(D),$$
$$p^{\varepsilon_{k_l}} \to p^* \quad \text{strongly in } H^s(D), \qquad \lambda^{\varepsilon_{k_l}} \to \lambda^* \quad \text{in } \mathbb{R},$$

where y^*, p^* are given by

$$Ay^* + \psi(y^*) = u^*, \tag{4.36}$$
$$B(y^*)p^* = -(y^* - y^\dagger). \tag{4.37}$$

(2) Every cluster point $\zeta^* := (u^*, y^*, p^*, \lambda^*)$ of the sequence $\{\zeta^{\varepsilon_k}\}_{k\in\mathbb{N}}$ for the above product topology (for $s < 2$ large enough) satisfies $\Phi(\zeta^*) = 0$, and u^* is a strong cluster point of $\{u^{\varepsilon_k}\}$ in $L^2([-p^* + \lambda^* \neq 0])$.

Proof. Theorem 4.7 asserts that $u^{\varepsilon_k} \in L^2(D, [0,1])$ for each k, which implies the weak convergence of a subsequence $\{u^{\varepsilon_{k_l}}\}_{l\in\mathbb{N}}$. We denote by u^* the weak limit. Since the solution $y^{\varepsilon_{k_l}}$ of $Ay^{\varepsilon_{k_l}} + \psi(y^{\varepsilon_{k_l}}) = u^{\varepsilon_{k_l}}$ is uniformly bounded in $H^2(D)$, for $s < 2$, there exists a $y^* \in H^s(D) \cap H_0^1(D)$ such that $y^{\varepsilon_{k_l}} \to y^*$ in H^s (for a further subsequence). Passing to the limit in the equation $Ay^{\varepsilon_{k_l}} + \psi(y^{\varepsilon_{k_l}}) = u^{\varepsilon_{k_l}}$ integrated against a test function $z \in H_0^1(D)$ we obtain $Ay^* + \psi(y^*) = u^*$. Since $p^{\varepsilon_{k_l}}$ is also uniformly bounded in $H^2(D)$, there exists $p^* \in H^s(D) \cap H_0^1(D)$ for $s < 2$ such that $p^{\varepsilon_{k_l}} \to p^*$ in H^s. We actually have the equation

$$p^{\varepsilon_{k_l}} = -B(y^{\varepsilon_{k_l}})^{-1}(y^{\varepsilon_{k_l}} - y^\dagger), \tag{4.38}$$

and $y^{\varepsilon_{k_l}} \to y^*$ in L^∞ due to Sobolev embedding. Applying Lemma B.1 we obtain, for all $\eta \in L^2$,

$$\begin{aligned}\langle B(y^{\varepsilon_{k_l}})^{-1}(y^{\varepsilon_{k_l}}) - B(y^*)^{-1}(y^*), \eta\rangle &= \langle [B(y^{\varepsilon_{k_l}})^{-1} - B(y^*)^{-1}](y^{\varepsilon_{k_l}}) + B(y^*)^{-1}(y^{\varepsilon_{k_l}} - y^*), \eta\rangle \\ &= \langle y^{\varepsilon_{k_l}}, [B(y^{\varepsilon_{k_l}})^{-1} - B(y^*)^{-1}]\eta\rangle + \langle y^{\varepsilon_{k_l}} - y^*, B(y^*)^{-1}\eta\rangle \\ &\to 0.\end{aligned}$$

Hence $B(y^{\varepsilon_{k_l}})^{-1}(y^{\varepsilon_{k_l}}) \rightharpoonup B(y^*)^{-1}(y^*)$ weakly in L^2. By compactness of $\{B(y^{\varepsilon_{k_l}})^{-1}(y^{\varepsilon_{k_l}})\}$ in L^2, the convergence holds actually strongly (for a subsequence). The convergence of $B(y^{\varepsilon_{k_l}})^{-1}y^\dagger$ in (4.38) also follows from Lemma B.1. Thus, letting $l \to \infty$ in (4.38) we obtain $p^* = -B(y^*)^{-1}(y^* - y^\dagger)$.

Now we show that λ^{ε_k} is uniformly bounded. Assume for instance that there exists a subsequence $\lambda^{\varepsilon_{k_l}} \to +\infty$. In view of Lemma 4.6, we have

$$u^{\varepsilon_{k_l}} = \theta^{\varepsilon_{k_l}}(-p^{\varepsilon_{k_l}} + \lambda^{\varepsilon_{k_l}}). \tag{4.39}$$

Since $\|p^{\varepsilon_{k_l}}\|_{L^\infty(D)}$ is uniformly bounded, (4.39) and Assumption (3.4) provide $u^{\varepsilon_{k_l}} \to 0$ almost everywhere, but this implies $\int_D u^{\varepsilon_{k_l}} \to 0$ since $0 \le u^{\varepsilon_{k_l}} \le 1$ which contradicts the volume constraint $\int_D u^{\varepsilon_{k_l}} = m$. Therefore λ^{ε_k} is bounded from above and with a similar argument, also from below. Thus we have found that λ^{ε_k} is uniformly bounded and it follows that there exists a subsequence $\lambda^{\varepsilon_{k_l}}$ and λ^* such that $\lambda^{\varepsilon_{k_l}} \to \lambda^*$ in $\mathbb{R}$.

We now turn to the second assertion of the theorem. Due to (4.36) and (4.37), we already have $L_y(u^*, y^*, p^*) = 0$ and $L_p(u^*, y^*, p^*) = 0$. Passing to the weak limit in $\langle 1, u^{\varepsilon_k}\rangle = m$ yields $\langle 1, u^*\rangle = m$. Set $g^{\varepsilon_k} := -p^{\varepsilon_k} + \lambda^{\varepsilon_k}$ and $g^* := -p^* + \lambda^*$. For a subsequence we have $g^{\varepsilon_k} \to g^*$ almost everywhere in D. In addition, $T^{\varepsilon_k}(u^{\varepsilon_k}, g^{\varepsilon_k}) = 0$ entails $u^{\varepsilon_k}(x) = \theta^{\varepsilon_k}(g^{\varepsilon_k}(x))$ for almost every $x \in D$. Assume for instance that $g^*(x) > 0$ for some $x \in D$. Then, for k large enough,

$g^{\varepsilon_k}(x)$ stays in a compact subset of $(0,+\infty)$. Yet, by Dini's theorem, $\theta^{\varepsilon_k} \to 0$ uniformly on compact subsets of $(0,+\infty)$. This entails $u^{\varepsilon_k}(x) = \theta^{\varepsilon_k}(g^{\varepsilon_k}(x)) \to 0$. Similarly, if $g^*(x) < 0$, then $u^{\varepsilon_k}(x) = \theta^{\varepsilon_k}(g^{\varepsilon_k}(x)) \to 1$. Thus $u^{\varepsilon_k}(x) \to \theta(g^*(x))$ for a.e. $x \in [g^* \neq 0]$. By dominated convergence, since $0 \leq u^{\varepsilon_k} \leq 1$, this limit holds also strongly in $L^2([g^* \neq 0])$. By uniqueness, we infer that $u^* = \theta(g^*)$ on $[g^* \neq 0]$. This implies that $u^* \in \Theta(g^*)$, and subsequently, by Assumption 2.4, that $T(u^*, g^*) = 0$. □

5. Algorithmic issues

5.1. **Discretization.** In this section we consider the discrete counterpart of the minimization problem (1.1), i.e. where the function spaces U_{ad} and Y as well as the functionals J and E are discretized. For simplicity we keep the notation of the infinite-dimensional setting. We place ourselves in the context of Section 4. We use for A the standard finite difference approximation of the Dirichlet Laplacian. The function spaces become

$$U = Y = Z = \mathbb{R}^n,$$

and the integrals are replaced by discrete sums. The finite-dimensional counterpart of $\mathscr{E}$ and $\mathscr{F}$, defined in Section 4.4, is given by

$$\mathscr{E} = \mathscr{F} = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R},$$

so that Φ^ε maps $\mathscr{E}$ onto $\mathscr{F}$. Then the discrete counterparts of Theorems 4.7, 4.8 and 4.9 can be readily deduced, and checked by inspection of the proofs.

5.2. **Initialization.** We initialize the control and the Lagrange multiplier by the values $u \equiv 1/2$ and $\lambda = 0$. Then we fix y and p by solving the direct and adjoint problems (4.11)-(4.12), and we set $\zeta^{\varepsilon_0} = (u, y, p, \lambda)$. For the first value of the regularization parameter we choose $\varepsilon_1 = \|g\|_{L^\infty}$, with $g = -p + \lambda$.

5.3. **Implementation of the semismooth Newton method.** The Newton algorithm corresponding to the regularization parameter ε_k, $k \geq 1$, is initialized by the variable $\zeta^{\varepsilon_{k-1}}$ and provides at convergence the new iterate ζ^{ε_k}. The Newton iteration is described in Theorem 4.8, where $D\Phi^\varepsilon(\zeta)$ is now a generalized Jacobian matrix whose block structure is given by (4.20). It may depend on the arbitrary parameter ϖ appearing in $\mathscr{G}^\pm_\varpi$ if we use a functional T^ε involving $\max(0,\cdot)$ or $\min(0,\cdot)$, in which case we choose $\varpi = 0$. In order to speed up the numerical solving of the linear system (4.21), by exploiting the special structure of the Jacobian, we use the so-called *dual approach* described in [6, Section 14.4]. The stopping criterion is related to the increment between two Newton steps, namely

$$\frac{\|\zeta_k - \zeta_{k-1}\|_{\mathscr{E}}}{\|\zeta_{k-1}\|_{\mathscr{E}}} < \kappa_N.$$

5.4. **Update of ε.** We would like to ensure a constant rate of convergence of some merit function depending on ζ^ε and which should be driven to zero. To this end we define

$$R(\varepsilon) := \frac{1}{2}\|\Phi(\zeta^\varepsilon)\|^2_{\mathscr{F}}$$

and look for a sequence $\{\varepsilon_k\}$ such that

$$\frac{R(\varepsilon_{k+1})}{R(\varepsilon_k)} \approx \tau, \tag{5.1}$$

where $0 < \tau < 1$ is a user-given coefficient. A preliminary information on the behavior of the merit function is given by the following lemma. It applies for instance to the function $T_{(2)}$ and its regularizations $T^\varepsilon_{(2a)}$ and $T^\varepsilon_{(2b)}$, which will be those considered in the sequel.

Lemma 5.1. *Suppose that, for all $\varepsilon > 0$ and all $(s,t) \in \mathbb{R} \times \mathbb{R}$, there holds*

$$T(s,t) = \varepsilon T(s, \frac{t}{\varepsilon}), \qquad T^\varepsilon(s,t) = \varepsilon T^1(s, \frac{t}{\varepsilon}). \tag{5.2}$$

Then we have

$$\sup_{t\in\mathbb{R}} |T(\theta^\varepsilon(t), t)| = \varepsilon \sup_{t\in\mathbb{R}} |T(\theta^1(t), t)|. \tag{5.3}$$

If in addition $\sup_{t\in\mathbb{R}} |T(\theta^1(t), t)| < \infty$, then there exists a constant $c > 0$ such that

$$R(\varepsilon) \le c\varepsilon^2. \tag{5.4}$$

Proof. We derive from (5.2) that $s = \theta^\varepsilon(t) \Leftrightarrow s = \theta^1(t/\varepsilon)$, whereby $\theta^\varepsilon(t) = \theta^1(t/\varepsilon)$. Then, we have for all $t \in \mathbb{R}$

$$T(\theta^\varepsilon(t), t) = \varepsilon T\left(\theta^\varepsilon(t), \frac{t}{\varepsilon}\right) = \varepsilon T\left(\theta^1(\frac{t}{\varepsilon}), \frac{t}{\varepsilon}\right),$$

which implies (5.3). From $R(\varepsilon) = \frac{1}{2}\|T(u^\varepsilon, g^\varepsilon)\|^2 = \frac{1}{2}\|T(\theta^\varepsilon(g^\varepsilon), g^\varepsilon)\|^2$, with $g^\varepsilon = -p^\varepsilon + \lambda^\varepsilon$, we get (5.4). □

To get a sequence verifying (5.1), we need first to prove that ζ^ε satisfies an appropriate differentiability property with respect to ε. To this end, we use an implicit function theorem for semismooth mappings. Some preliminaries on the notion of semismoothness are necessary. Let $F : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ be a locally Lipschitz mapping. According to Rademacher's theorem, F is almost everywhere differentiable. Let D_F denote the set of all differentiable points of F. Then we call

$$\partial_B F(x) := \left\{H \in \mathbb{R}^{n_1\times n_2} \middle| \exists\{x^k\} \subseteq D_F : x^k \to x, F'(x^k) \to H\right\}$$

the *B-subdifferential* of F at x. Its convex hull

$$\partial F(x) := \operatorname{conv} \partial_B F(x)$$

is Clarke's *generalized Jacobian* of F at x; see [12]. Note that $\partial_B F \subseteq \partial F$. We will say that F is semismooth if it is directionally differentiable and satisfies

$$\|F(x+d) - F(x) - Hd\| = o(\|d\|)$$

for all $d \to 0$ and all $H \in \partial F(x+d)$. Note that this is not the classical definition of semismoothness, but an equivalent one; see [22, 23].

Now consider a mapping $\Psi : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_2}$. Then the projection $\pi_y \partial\Psi(x,y)$ denotes the set of all $n_2 \times n_2$ matrices M such that, for some $n_2 \times n_1$ matrix N, the $n_2 \times (n_1 + n_2)$ matrix $[N, M]$ belongs to $\partial\Psi(x,y)$. The following implicit function theorem is taken from [27, Theorem 2.3]; see also [26, Theorem 2.1].

Theorem 5.2. *Suppose that $\Psi : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_2}$ is locally Lipschitz and semismooth in a neighborhood of a point $(\bar{x}, \bar{y})$ satisfying $\Psi(\bar{x}, \bar{y}) = 0$, and assume that all matrices in $\pi_y \partial\Psi(\bar{x}, \bar{y})$ are nonsingular. Then there exists an open neighborhood X of $\bar{x}$ and a function $y : X \to \mathbb{R}^{n_2}$ which is Lipschitz and semismooth on X such that $y(x) = \bar{y}$ and $\Psi(x, y(x)) = 0$ for all $x \in X$.*

We apply Theorem 5.2 to the function Ψ defined by $\Psi : \mathbb{R} \times \mathscr{E} \ni (\varepsilon, \zeta) \mapsto \Phi^\varepsilon(\zeta) \in \mathscr{F}$. We obtain the following corollary.

Corollary 5.3. *Let $T^\varepsilon \in \{T^\varepsilon_{(1)}, T^\varepsilon_{(2a)}, T^\varepsilon_{(2b)}\}$ and $(\bar{\varepsilon}, \bar{\zeta})$ be such that $\Phi^{\bar{\varepsilon}}(\bar{\zeta}) = 0$. Then there exists an open neighborhood Υ of $\bar{\varepsilon}$ and a function $\zeta : \Upsilon \to \mathscr{E}$ which is Lipschitz and semismooth on Υ such that $\zeta(\bar{\varepsilon}) = \bar{\zeta}$ and $\Phi^\varepsilon(\zeta(\varepsilon)) = 0$ for all $\varepsilon \in \Upsilon$.*

Proof. We present the proof for the operator $T^\varepsilon_{(2a)}$ only, as the operator $T^\varepsilon_{(1)}$ can be treated likewise and the standard implicit function theorem applies to $T^\varepsilon_{(2b)}$. We essentially need to verify that the assumptions of Theorem 5.2 are satisfied. By construction we have $\Psi(\bar\varepsilon,\bar\zeta)=0$. We denote by $\Psi_i(\varepsilon,\zeta), 1\le i\le 4$, the components of $\Psi(\varepsilon,\zeta)$, i.e.

$$\Psi(\varepsilon,\zeta) =: \begin{pmatrix}\Psi_1(\varepsilon,\zeta)\\ \Psi_2(\varepsilon,\zeta)\\ \Psi_3(\varepsilon,\zeta)\\ \Psi_4(\varepsilon,\zeta)\end{pmatrix} = \begin{pmatrix} T^\varepsilon(u,-p+\lambda)\\ (A+\psi'(y))p+y-y^\dagger\\ Ay+\psi(y)-u\\ \langle 1,u\rangle - m\end{pmatrix}.$$

Since T^ε is locally Lipschitz, Ψ_1 is also locally Lipschitz by composition. Using that $\psi\in W^{3,\infty}(\mathbb{R})$ we get that Ψ_i is also locally Lipschitz for $2\le i\le 4$.

Let us now check that Ψ is semismooth. Similarly to Φ^ε, Ψ admits the Newton derivative

$$D\Psi(\varepsilon,\zeta) = \begin{pmatrix} [T^\varepsilon_\varepsilon(u,g)]_{n,1} & [T^\varepsilon_s(u,g)]_{n,n} & \mathbb{O}_{n,n} & [-T^\varepsilon_t(u,g)]_{n,n} & [T^\varepsilon_t(u,g)]_{n,1}\\ \mathbb{O}_{n,1} & \mathbb{O}_{n,n} & [\psi''(y)p+1]_{n,n} & A+[\psi'(y)]_{n,n} & \mathbb{O}_{n,1}\\ \mathbb{O}_{n,1} & -\mathbb{I}_n & A+[\psi'(y)]_{n,n} & \mathbb{O}_{n,n} & \mathbb{O}_{n,1}\\ \mathbb{O}_{1,1} & \mathbb{1}_{1,n} & \mathbb{O}_{1,n} & \mathbb{O}_{1,n} & \mathbb{O}_{1,1}\end{pmatrix}, \tag{5.5}$$

where $g=-p+\lambda$, $\mathbb{O}_{n_1,n_2}$ and $\mathbb{1}_{n_1,n_2}$ denote the $n_1\times n_2$ matrices filled with zeros and ones, respectively, and $\mathbb{I}_n$ is the $n\times n$ identity matrix. Also, the notation $[v]_{n_1,n_2}$ stands for a block of size $n_1\times n_2$, which is obtained by arranging the components of the vector v columnwise when $n_2=1$ and diagonalwise when $n_1=n_2=n$. For the chosen function T^ε, the vectors $T^\varepsilon_s(u,g)$, $T^\varepsilon_t(u,g)$ and $T^\varepsilon_\varepsilon(u,g)$ are given by

$$\begin{aligned} T^\varepsilon_s(u,g) &= \sqrt{\varepsilon^2+g^2},\\ T^\varepsilon_t(u,g) &= ug(\varepsilon^2+g^2)^{-1/2}+\mathscr{G}^-_\varpi(g),\\ T^\varepsilon_\varepsilon(u,g) &= u\varepsilon(\varepsilon^2+g^2)^{-1/2}.\end{aligned}$$

Of course, in the above expressions, the products, powers and square roots apply componentwise. The vector-valued function $\mathscr{G}^-_\varpi:\mathbb{R}^n\to\mathbb{R}^n$ is defined, for an arbitrarily fixed $\varpi\in\mathbb{R}^n$, by

$$\mathscr{G}^-_\varpi(g) = \{\mathbb{1}_{[g_i<0]}+\varpi_i\mathbb{1}_{[g_i=0]}\}_{i=1}^n.$$

The Newton derivative $\mathscr{G}^-_\varpi$ is the finite-dimensional counterpart [15, Lemma 3.1] of the Newton derivative in Theorem 3.2. Note that $D\Psi(\varepsilon,\zeta)$ is a $(3n+1)\times(3n+2)$ matrix which depends on $\varpi\in\mathbb{R}^n$. A quick calculation indicates that

$$\partial\Psi(\varepsilon,\zeta) = \operatorname{conv}\,\partial_B\Psi(\varepsilon,\zeta) = \{D\Psi(\varepsilon,\zeta),\ 0\le\varpi_i\le 1\ \forall i\}.$$

The fact that $D\Psi(\varepsilon,\zeta)$ acts as Newton derivative of Ψ for every $0\le\varpi_i\le 1$ implies that Ψ is semismooth.

It also follows that

$$\pi_\zeta\partial\Psi(\varepsilon,\zeta) = \partial\Phi^\varepsilon(\zeta) = \operatorname{conv}\,\partial_B\Phi^\varepsilon(\zeta) = \{D\Phi^\varepsilon(\zeta),\ 0\le\varpi_i\le 1\ \forall i\},$$

where $\partial_B\Phi^\varepsilon(\zeta)$ is the set of directional derivatives $D\Phi^\varepsilon(\zeta)$ and $D\Phi^\varepsilon(\zeta)$ denotes here the discrete counterpart of the Jacobian computed in (4.20), i.e.

$$D\Phi^\varepsilon(\zeta) = \begin{pmatrix} [T^\varepsilon_s(u,g)]_{n,n} & \mathbb{O}_{n,n} & [-T^\varepsilon_t(u,g)]_{n,n} & [T^\varepsilon_t(u,g)]_{n,1}\\ \mathbb{O}_{n,n} & [\psi''(y)p+1]_{n,n} & A+[\psi'(y)]_{n,n} & \mathbb{O}_{n,1}\\ -\mathbb{I}_n & A+[\psi'(y)]_{n,n} & \mathbb{O}_{n,n} & \mathbb{O}_{n,1}\\ \mathbb{1}_{1,n} & \mathbb{O}_{1,n} & \mathbb{O}_{1,n} & \mathbb{O}_{1,1}\end{pmatrix}.$$

In Theorem 4.8, we have proved that $D\Phi^\varepsilon(\zeta^\varepsilon)$ is invertible for any $\varepsilon>0$ and $0\le\varpi\le 1$ in the infinite dimensional setting. From inspection of the proof we can check that this result remains

true in the finite dimensional setting for any $0 \leq \varpi_i \leq 1$, and therefore for all matrices in $\partial\Phi^{\bar\varepsilon}(\bar\zeta) = \pi_\zeta \partial\Psi(\bar\varepsilon, \bar\zeta)$ if $\bar\varepsilon > 0$. Thus, we can apply Theorem 5.2 and the corollary follows immediately. □

Now we turn to the update of the regularization parameter ε_k. We would like to achieve the decrease (5.1) with a reasonably small τ. On one hand, by Corollary 5.3, there exists a selection of solutions $\zeta^\varepsilon = \zeta(\varepsilon)$ for which the function $\varepsilon \mapsto \zeta(\varepsilon)$ is semismooth, thus in particular directionally differentiable. As the function $\zeta \mapsto \|\Phi(\zeta)\|^2$ is $\mathcal{C}^1$, we deduce that $\varepsilon \mapsto R(\varepsilon)$ is locally Lipschitz as well as directionally differentiable, and the chain rule applies; see [7, Proposition 2.47]. On the other hand, in the cases of application of Lemma 5.1, we have $R(\varepsilon) \leq c\varepsilon^2$ for some constant c. Thus, in order to make a proper linearization, it makes sense to use a logarithmic scale for both $R(\varepsilon)$ and ε. Therefore we set

$$\rho(\ln \varepsilon) := \ln R(\varepsilon)$$

and, for a given ε_k, we are now looking for ε_{k+1} satisfying

$$\rho(\ln \varepsilon_{k+1}) - \rho(\ln \varepsilon_k) \approx \ln \tau.$$

We now linearize ρ about $\ln \varepsilon_k$ in the direction of decreasing arguments, which leads to

$$\rho'(\ln \varepsilon_k)(\ln \varepsilon_{k+1} - \ln \varepsilon_k) \approx \ln \tau.$$

Thus we take the following update for ε_{k+1}:

$$\varepsilon_{k+1} = \varepsilon_k \tau^{\rho'(\ln \varepsilon_k)^{-1}}.$$

We now compute $\rho'(\ln \varepsilon_k)$. For simplicity, we place ourselves in the (generic) case where the function $\varepsilon \mapsto \zeta(\varepsilon)$ is differentiable at the considered point. We have by the chain rule

$$\rho'(\ln \varepsilon_k) = \varepsilon_k \frac{R'(\varepsilon_k)}{R(\varepsilon_k)} \tag{5.6}$$

and

$$R'(\varepsilon_k) = \langle D\Phi(\zeta^{\varepsilon_k})\zeta'(\varepsilon_k), \Phi(\zeta^{\varepsilon_k})\rangle_{\mathscr{F}}$$

where $\langle\cdot,\cdot\rangle_{\mathscr{F}}$ denotes the scalar product in $\mathscr{F}$ and $D\Phi(\zeta^{\varepsilon_k})$ is an arbitrary Newton derivative of Φ at ζ^{ε_k}. In addition, by (Newton) differentiating $\Psi(\varepsilon, \zeta(\varepsilon)) = 0$ we arrive at

$$\begin{aligned}\zeta'(\varepsilon_k) &= -D_\zeta \Psi(\varepsilon_k, \zeta(\varepsilon_k))^{-1} D_\varepsilon \Psi(\varepsilon_k, \zeta(\varepsilon_k)) \\ &= -D\Phi^{\varepsilon_k}(\zeta^{\varepsilon_k})^{-1} D_\varepsilon \Psi(\varepsilon_k, \zeta^{\varepsilon_k}).\end{aligned}$$

Finally we obtain the update

$$\varepsilon_{k+1} = \varepsilon_k \tau^{\beta_k} \text{ with } \beta_k = \frac{-R(\varepsilon_k)}{\varepsilon_k \langle D\Phi(\zeta^{\varepsilon_k}) D\Phi^{\varepsilon_k}(\zeta^{\varepsilon_k})^{-1} D_\varepsilon \Psi(\varepsilon_k, \zeta^{\varepsilon_k}), \Phi(\zeta^{\varepsilon_k})\rangle_{\mathscr{F}}}.$$

Note that $D_\varepsilon \Psi$ is given by the first column of (5.5).

5.5. **Stopping criterion.** The stopping criterion we choose is related to the logarithmic derivative of the function $\varepsilon \mapsto \zeta(\varepsilon)$, namely

$$\left\| \frac{d\zeta}{d \ln \varepsilon}(\varepsilon) \right\| < \kappa_E \|\zeta(\varepsilon)\|,$$

or equivalently

$$\varepsilon \, \|\zeta'(\varepsilon)\| < \kappa_E \|\zeta(\varepsilon)\|.$$

Note that the user-given constant κ_E is dimensionless.

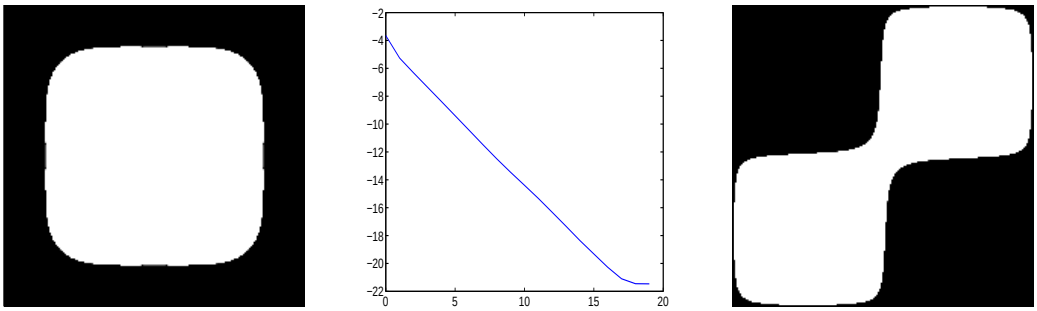

FIGURE 2. Linear case: optimal control for $y^\dagger = y_1^\dagger$ (left), convergence history of $\log_{10} R(\varepsilon)$ for $y^\dagger = y_1^\dagger$ (middle), and optimal control for $y^\dagger = y_2^\dagger$ (right).

6. Numerical experiments

We consider a two dimensional problem on the unit square $D =]0,1[^2$. The target volume is $m = 0.5$. The discretization is done with $n = 39601$ nodes. We choose the function $T^\varepsilon_{(2b)}$ given in Section 3.2 and the parameters $\tau = 0.1$, $\kappa_N = 10^{-8}$, $\kappa_E = 10^{-3}$. We have also tested the functions $T^\varepsilon_{(1)}$ and $T^\varepsilon_{(2a)}$, and our experiments tend to indicate that the algorithm is in average slightly slower with the function $T^\varepsilon_{(2a)}$, while it may be less stable for small values of ε with the function $T^\varepsilon_{(1)}$. For each of the following computations performed in Matlab, the CPU time is of the order of 2 minutes on a standard desktop computer.

6.1. **The linear problem.** To begin with we choose $\psi \equiv 0$ and two functions $y^\dagger$:

$$y_1^\dagger(x_1, x_2) = 0.01,$$
$$y_2^\dagger(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2).$$

The obtained optimal controls u are depicted in Figure 2, where white corresponds to $u = 0$ and black to $u = 1$. Note that the absence of intermediate regions is, for $y^\dagger = y_1^\dagger$, a consequence of Theorem 4.4. The convergence history of the residual $R(\varepsilon)$ is also shown, in semi-logarithmic scale, with the number of updates of ε along the x-axis.

6.2. **Examples of nonlinear problems.** We fix $y^\dagger = y_1^\dagger$, and consider two functions ψ:

$$\psi_1(t) = e^{at} - 1, \qquad a = 10^3,$$
$$\psi_2(t) = \arctan(at), \qquad a = 10^2.$$

Note that ψ_2 has bounded first, second and third derivatives, but it is not the case for ψ_1. The corresponding results are shown on Figure 3. The effect of the nonlinearity is clearly emphasized by the appearance of intermediate regions. We point out that in each case the volume constraint is realized with $|\int_D u - m|/m < 10^{-13}$.

Appendix A. Collectively compact sets of operators

Let $\mathcal{X}$ be a Banach space and $\mathcal{K}$ be a subset of $\mathcal{L}(\mathcal{X})$, where $\mathcal{L}(\mathcal{X})$ is the set of bounded linear operators from $\mathcal{X}$ into itself.

Definition A.1. We say that $\mathcal{K}$ is collectively compact if the set $\{Kx, x \in \mathcal{X}, \|x\| \leq 1, K \in \mathcal{K}\}$ is relatively compact.

Obviously, if $\mathcal{K}$ is collectively compact, every $K \in \mathcal{K}$ is compact. The following result may be found in [3, Theorem 1.6].

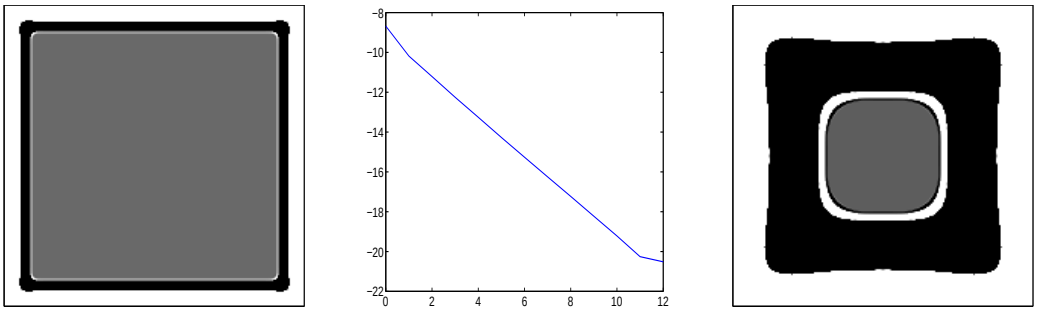

FIGURE 3. Nonlinear cases: optimal control for $\psi = \psi_1$ (left), convergence history of $\log_{10} R(\varepsilon)$ for $\psi = \psi_1$ (middle), and optimal control for $\psi = \psi_2$ (right).

Theorem A.2. *Let K, $(K_n)_{n\in\mathbb{N}} \in \mathcal{L}(\mathcal{X})$. Assume $K_n \to K$ pointwise, $\{K_n\}$ is collectively compact and K is compact. Then $(I-K)^{-1}$ exists if and only if for some n_0 and all $n \geq n_0$ the operators $(I-K_n)^{-1}$ exist and are uniformly bounded, in which case $(I-K_n)^{-1} \to (I-K)^{-1}$ pointwise.*

The following result can be easily deduced from Theorem A.2; see [2].

Theorem A.3. *Let $\mathcal{K}$ be a collectively compact set of bounded linear operators of $\mathcal{X}$. Assume further that $\mathcal{K}$ is pointwise sequentially compact, i.e., for every sequence (K_n) of $\mathcal{K}$ there exists a subsequence (K_{n_p}) and $K \in \mathcal{K}$ such that $K_{n_p}x \to Kx$ for all $x \in \mathcal{X}$. If $I-K$ is invertible for all $K \in \mathcal{K}$, then*

$$\sup_{K\in\mathcal{K}} \|(I-K)^{-1}\| < \infty. \tag{A.1}$$

APPENDIX B. OPERATOR CONVERGENCE

Lemma B.1. *If $y_n \to y$ in $L^\infty(D)$ then, for all $\eta \in L^2(D)$,*

$$B(y_n)^{-1}\eta = [A+\psi'(y_n)]^{-1}\eta \to [A+\psi'(y)]^{-1}\eta = B(y)^{-1}\eta \qquad \text{in } L^\infty(D). \tag{B.1}$$

Proof. With $y_n \to y$ in $L^\infty(D)$ and using $\|\psi''\|_{L^\infty} \leq M_\psi^2$ we obtain

$$\psi'(y_n) \to \psi'(y) \text{ in } L^\infty(D). \tag{B.2}$$

We write

$$B(y_n)^{-1} = A^{-1}[I+\psi'(y_n)A^{-1}]^{-1}.$$

The family of operators $\{\psi'(y_n)A^{-1} : L^2 \to L^2\}$ is collectively compact due to the compactness of A^{-1} and the uniform boundedness of $\|\psi'(y_n)\|_{L^\infty}$. We have for all $\varphi \in L^2(D)$

$$\langle (I+\psi'(y)A^{-1})\varphi, A^{-1}\varphi\rangle = \langle A^{-1}\varphi,\varphi\rangle + \langle \psi'(y)A^{-1}\varphi, A^{-1}\varphi\rangle \geq \langle A^{-1}\varphi,\varphi\rangle,$$

hence $I+\psi'(y)A^{-1}$ is injective and subsequently invertible by the Fredholm alternative. In view of (B.2), we also have the pointwise convergence $\psi'(y_n)A^{-1} \to \psi'(y)A^{-1}$, we may thus apply Theorem A.2 to obtain

$$[I+\psi'(y_n)A^{-1}]^{-1} \to [I+\psi'(y)A^{-1}]^{-1} \text{ pointwise in } L^2(D),$$

which in turn implies (B.1) by composition with A^{-1}. □

References

[1] G. Allaire. *Conception optimale de structures*, volume 58 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2007. With the collaboration of Marc Schoenauer (INRIA) in the writing of Chapter 8.

[2] S. Amstutz. A semismooth Newton method for topology optimization. *Nonlinear Anal.*, 73(6):1585–1595, 2010.

[3] P. M. Anselone. *Collectively compact operator approximation theory and applications to integral equations*. Prentice-Hall Inc., Englewood Cliffs, N. J., 1971. With an appendix by Joel Davis, Prentice-Hall Series in Automatic Computation.

[4] M. P. Bendsøe and O. Sigmund. *Topology optimization*. Springer-Verlag, Berlin, 2003. Theory, methods and applications.

[5] J. F. Bonnans. Second-order analysis for control constrained optimal control problems of semilinear elliptic systems. *Appl. Math. Optim.*, 38(3):303–325, 1998.

[6] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical optimization*. Universitext. Springer-Verlag, Berlin, second edition, 2006. Theoretical and practical aspects.

[7] J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Series in Operations Research. Springer-Verlag, New York, 2000.

[8] E. Casas, J. C. de los Reyes, and F. Tröltzsch. Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints. *SIAM J. Optim.*, 19(2):616–643, 2008.

[9] E. Casas, R. Herzog, and G. Wachsmuth. Optimality conditions and error analysis in semilinear elliptic control problems with L^1 cost functional. *SIAM J. Optim.*, 22(3):795–820, 2012.

[10] X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216 (electronic), 2000.

[11] E. T. Chung, T. F. Chan, and X.-C. Tai. Electrical impedance tomography using level set representation and total variational regularization. *J. Comput. Phys.*, 205(1):357–372, 2005.

[12] F. H. Clarke. *Optimization and nonsmooth analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.

[13] C. Clason and K. Kunisch. A duality-based approach to elliptic control problems in non-reflexive Banach spaces. *ESAIM Control Optim. Calc. Var.*, 17(1):243–266, 2011.

[14] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Springer-Verlag, Berlin, 1977. Grundlehren der Mathematischen Wissenschaften, Vol. 224.

[15] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (electronic) (2003), 2002.

[16] M. Hintermüller and K. Kunisch. Pde-constrained optimization subject to pointwise constraints on the control, the state, and its derivative. *SIAM J. on Optimization*, 20:1133–1156, August 2009.

[17] M. Hintermüller and A. Laurain. Electrical impedance tomography: from topology to shape. *Control Cybernet.*, 37(4):913–933, 2008.

[18] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.

[19] K. Ito and K. Kunisch. The primal-dual active set method for nonlinear optimal control problems with bilateral constraints. *SIAM J. Control Optim.*, 43(1):357–376 (electronic), 2004.

[20] K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[21] K. Ito and K. Kunisch. Novel concepts for nonsmooth optimization and their impact on science and technology. In R. Bhatia, editor, *Proceedings of the International Congress of Mathematicians 2010, Hyderabad, India*, 2010.

[22] J. Outrata, M. Kočvara, and J. Zowe. *Nonsmooth approach to optimization problems with equilibrium constraints*, volume 28 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1998. Theory, applications and numerical results.

[23] L. Q. Qi and J. Sun. A nonsmooth version of Newton's method. *Math. Programming*, 58(3, Ser. A):353–367, 1993.

[24] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60:259–268, November 1992.

[25] G. Stadler. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44(2):159–181, 2009.

[26] D. Sun. A further result on an implicit function theorem for locally Lipschitz functions. *Oper. Res. Lett.*, 28(4):193–198, 2001.

[27] A. von Heusinger and C. Kanzow. Sc1 optimization reformulations of the generalized nash equilibrium problem. *Optimization Methods Software*, 23:953–973, December 2008.
[28] G. Vossen and H. Maurer. On L^1-minimization in optimal control and applications to robotics. *Optimal Control Appl. Methods*, 27(6):301–321, 2006.
[29] G. Wachsmuth and D. Wachsmuth. Convergence and regularization results for optimal control problems with sparsity functional. *ESAIM:COCV*, 17(3):858–886, 2011.

Laboratoire de Mathématiques d'Avignon, Faculté des sciences, 33 rue Louis Pasteur, 84000 Avignon, France.
E-mail address: samuel.amstutz@univ-avignon.fr

Department of Mathematics, Technical University of Berlin, Berlin, Germany
E-mail address: laurain@math.tu-berlin.de