Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

A globalized inexact semismooth Newton method for nonsmooth fixed-point equations involving variational inequalities

Amal Alphonse[1], Constantin Christof[2], Michael Hintermüller[1,3],

Ioannis P. A. Papadopoulos[1]

submitted: October 2, 2024

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: amal.alphonse@wias-berlin.de
michael.hintermueller@wias-berlin.de
ioannis.papadopoulos@wias-berlin.de

[2] CIT
Department of Mathematics
Technical University of Munich
Boltzmannstraße 3
85748 Garching b. München
Germany
E-Mail: christof@cit.tum.de

[3] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
E-Mail: hint@math.hu-berlin.de

No. 3132
Berlin 2024

2020 *Mathematics Subject Classification.* 35J86, 47J20, 49J40, 49J52, 49M15.

Key words and phrases. Semismooth Newton method, quasi-variational inequality, thermoforming, nonsmooth analysis, obstacle problem, Newton differentiability, semismoothness, superlinear convergence.

IP was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

A globalized inexact semismooth Newton method for nonsmooth fixed-point equations involving variational inequalities

Amal Alphonse, Constantin Christof, Michael Hintermüller,
Ioannis P. A. Papadopoulos

Abstract

We develop a semismooth Newton framework for the numerical solution of fixed-point equations that are posed in Banach spaces. The framework is motivated by applications in the field of obstacle-type quasi-variational inequalities and implicit obstacle problems. It is discussed in a general functional analytic setting and allows for inexact function evaluations and Newton steps. Moreover, if a certain contraction assumption holds, we show that it is possible to globalize the algorithm by means of the Banach fixed-point theorem and to ensure q-superlinear convergence to the problem solution for arbitrary starting values. By means of a localization technique, our Newton method can also be used to determine solutions of fixed-point equations that are only locally contractive and not uniquely solvable. We apply our algorithm to a quasi-variational inequality which arises in thermoforming and which not only involves the obstacle problem as a source of nonsmoothness but also a semilinear PDE containing a nondifferentiable Nemytskii operator. Our analysis is accompanied by numerical experiments that illustrate the mesh-independence and q-superlinear convergence of the developed solution algorithm.

1 Introduction

This paper is concerned with the design, analysis, and numerical realization of semismooth Newton methods for fixed-point equations of the type

$$\text{Find } \bar{x} \in X \text{ such that } \bar{x} = H(\bar{x}) \tag{F}$$

that are posed in a real Banach space X and involve a Newton differentiable operator $H\colon X \to X$. A main focus of our work is on the case where the map H can be written as the composition

$$H = S \circ \Phi \tag{1}$$

of two Newton differentiable functions S and Φ, with a special emphasis on the situation where S is the solution map of an elliptic variational inequality (VI) with pointwise constraints. Our main motivation for considering this kind of structure is that it arises naturally when studying elliptic quasi-variational inequalities (QVIs) of obstacle type, i.e., variational problems of the form

$$\begin{aligned} &\text{Find } u \in K(\Phi(u)) \text{ such that } \langle -\Delta u - f, v - u\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \ \forall v \in K(\Phi(u)), \\ &\text{with } K(\Phi(u)) := \{v \in H_0^1(\Omega) \mid v \leq \Phi(u) + \Phi_0 \text{ a.e. in } \Omega\}, \end{aligned} \tag{Q}$$

where $\Omega \subset \mathbb{R}^d$ for $d \in \mathbb{N}$ denotes a nonempty open bounded set, the Sobolev space $H_0^1(\Omega)$ is defined as in [11, §5.1], $\Phi\colon H_0^1(\Omega) \to H_0^1(\Omega)$ is a given operator, f and Φ_0 are given functions, and the symbols Δ and $\langle \cdot, \cdot \rangle_{H^{-1}(\Omega),H_0^1(\Omega)}$ denote the distributional Laplacian and the dual pairing in

DOI 10.20347/WIAS.PREPRINT.3132 Berlin 2024

$H_0^1(\Omega)$, respectively (see Section 3 for the precise setting). Indeed, if we define S to be the solution operator $S\colon H_0^1(\Omega) \to H_0^1(\Omega)$, $\phi \mapsto u$, of the classical obstacle problem

$$\text{Find } u \in K(\phi) \text{ such that } \langle -\Delta u - f, v - u\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \ \forall v \in K(\phi), \tag{2}$$

then (Q) can clearly be recast as $u = H(u)$ with $H := S \circ \Phi\colon H_0^1(\Omega) \to H_0^1(\Omega)$, and thus (Q) can immediately be seen as a fixed-point equation of the form (F).

Note that the salient feature of QVIs of the type (Q) is that the solution u enters the problem not only via the Laplacian but also via the obstacle $\Phi(u) + \Phi_0$ defining the pointwise constraint in $K(\Phi(u))$. This creates a variational structure in which the set of admissible test functions depends implicitly on the problem solution and which distinguishes (Q) quite drastically from standard partial differential equations (PDEs) and VIs. In applications, the dependence of the admissible set $K(\Phi(u))$ of (Q) on u allows to incorporate feedback effects into the problem formulation that reflect, for instance, how the obstacle $\Phi(u) + \Phi_0$ interacts with u in zones of contact. Because of the ability to capture such an interplay between the problem solution and the constraints, QVIs have proved to be important instruments for modeling processes in various areas of physics and economics, e.g., thermoforming [7], sandpile growth [15, 17, 54], impulse control [19, 20, 21, 43], and superconductivity [53, 16, 57]. Compare also with the classical works [21, 12, 46, 14] in this context.

While the feedback effects in a QVI like (Q) are very desirable from the application and modeling point of view, mathematically, they often pose serious challenges. Establishing the Hadamard well-posedness of problems of the type (Q), for example, is typically a hard task. In fact, in many situations, it is possible that a QVI possesses multiple solutions—indeed a whole continuum of solutions—or no solutions at all. Compare, e.g., with the results on the Lipschitz and differential stability of QVI-solutions in [7, 8, 27, 10, 6, 61, 9] in this context, and in particular with the examples in [27, §6.1] and Section 4.3. The implicit relationship between the problem solution and the set of test functions also causes the numerical solution of QVIs to be a very delicate topic. As far as problems posed in infinite-dimensional spaces are concerned, the algorithmic approaches that are currently used for this purpose in the literature are primarily based on fixed-point arguments or regularization/penalization techniques—and, as a consequence, are either slow or inexact. See, for example, [18, Chapter IV], [8, §2.1], [7, §6.4], [39, §5], [62], and the references therein for particular instances of such algorithms.

A main goal of the present paper is to show that recent advances in the field of generalized differentiability properties of solution maps of obstacle-type VIs make it possible to set up and analyze semismooth Newton methods for the numerical solution of QVIs of the type (Q) and, thus, to solve obstacle-type quasi-variational inequalities in function space with superlinear convergence speed. More generally, we develop a semismooth Newton framework for fixed-point equations of the type (F) that is tailored to applications in the field of obstacle-type QVIs. A main feature of our semismooth Newton method for (F) is that it is provably locally q-superlinearly convergent, robust with respect to inexactness, and mesh-independent. Moreover, if a certain contractivity assumption holds, the algorithm can be made globally convergent; see Theorems 2.4 and 3.7. Note that the inclusion of inexactness is of special importance in the context of obstacle-type QVIs as evaluations of the outer map S in (1) arising in the context of (Q) are typically subject to numerical errors due to a discretization of (2) or the introduction of an easy-to-evaluate surrogate model S_ϵ; see Remark 2.14.

To the best of our knowledge, this paper is the first to develop a semismooth Newton method that is suitable for the solution of unregularized obstacle-type QVIs in the infinite-dimensional setting and provably q-superlinearly convergent. For related approaches in finite dimensions, we refer the reader to [31, 44, 47, 38] and the references therein. We remark that the example that we construct in Section 4.3 to obtain an instance of a (generalized) thermoforming QVI with multiple solutions is also of independent interest as it provides an important benchmark problem for numerical solution algorithms.

As we conduct our convergence analysis in a general Banach space setting and for the abstract fixed-point equation (F), the results that we prove in this paper are, of course, not only applicable to QVIs but also to other problems with comparable continuity and contractivity properties. We mention exemplarily VIs and PDEs with semilinearities, implicit VIs, and operator equations that arise as optimality conditions of optimal control problems with $H_0^1(\Omega)$-controls; see [28, §5].

1.1 Main results

The main results of this paper are concerned with the convergence of semismooth Newton methods for the solution of fixed-point equations of the type (F) and the applicability of the developed abstract theory to QVIs of the form (Q). For the highlights, we refer the reader to:

- Theorem 2.4, which establishes the global q-superlinear convergence/finite convergence to a given tolerance of an inexact globalized semismooth Newton method (Algorithm 2) for the solution of (F). This result relies on the assumption that $H\colon X \to X$ is Newton differentiable and globally contractive, i.e., γ-Lipschitz for some $\gamma \in [0, 1)$; see Assumption 2.2.
- Theorem 2.8, which localizes Algorithm 2 by means of a projection onto a nonempty closed convex set $B \subset X$. This result makes it possible to apply our convergence analysis to equations (F) that satisfy a contraction assumption only locally and have multiple solutions, a structure that is prevalent in many QVI-applications; cf. [7, 61, 9, 10, 6, 8]. For the precise setting for this result, see Assumption 2.5.
- Theorems 3.7, 3.9 and 3.12, which demonstrate that obstacle-type QVIs of the form (Q) are indeed covered by our general abstract semismooth Newton framework provided the involved quantities are sufficiently well behaved (see Assumptions 3.1 and 3.8).

We remark that the numerical realization of semismooth Newton methods for obstacle-type QVIs is also an interesting field on its own, in particular as the residue function $R(x) := x - H(x)$ arising in the context of problems like (Q) involves the solution map S of the variational inequality (2) and since the Newton derivatives that appear depend on the active, inactive, and strictly active set of the current Newton iterate. For a detailed discussion of how we tackle these challenging problems in our numerical implementation, we refer the reader to Section 4.

1.2 Notation and basic concepts

Throughout this paper, we denote by $\|\cdot\|_X$ the norm of a (real) normed vector space X. For the closed ball of radius $r > 0$ centered at a point $c \in X$, we write $B_r^X(c)$. A sequence $\{x_n\} \subset X$ is said to converge q-superlinearly to $x \in X$ if $x_n \to x$ and $\|x_{n+1} - x\|_X \leq o(1)\|x_n - x\|_X$ hold for $n \to \infty$, where the Landau notation $o(1)$ represents a term that vanishes in the limit. Given two normed spaces X and Y, we use the symbol $\mathcal{L}(X, Y)$ to denote the space of linear and continuous functions from X to Y. We write $X^* := \mathcal{L}(X, \mathbb{R})$ for the topological dual space of X. The evaluation of an element $x^* \in X^*$ at $x \in X$ is denoted by the dual pairing $\langle x^*, x\rangle_{X^*,X}$. For the identity map, we use the symbol Id. If X is Hilbert, then $(\cdot, \cdot)_X$ denotes the inner product of X, $V^\perp$ stands for the orthogonal complement of a closed subspace V of X, and $P_B(x) := \operatorname{argmin}_{z\in B}\|x - z\|_X$ is the metric projection onto a closed convex nonempty set $B \subset X$.

Given mappings $F\colon X \to Y$ and $G\colon Y \to Z$ between normed spaces X, Y, and Z, the composition of G and F is denoted by $G \circ F\colon X \to Z$. In the case of linear operators, the symbol $\circ$ is often dropped. The image of a set $D \subset X$ under F is denoted by $F(D)$. Recall that a function $F\colon D \to Y$ defined on a nonempty subset D of a normed space X with values in a normed space Y is called *Newton differentiable* with *(Newton) derivative* $G_F\colon D \to \mathcal{L}(X,Y)$ if

$$\lim_{\substack{0<\|h\|_X\to 0,\\ x+h\in D}} \frac{\|F(x+h)-F(x)-G_F(x+h)h\|_Y}{\|h\|_X} = 0 \qquad \forall x \in D. \tag{3}$$

We remark that Newton derivatives are often defined as set-valued mappings in the literature; see, e.g., [28, Definition 2.11]—we make a slight abuse of notation and assume G_F is the realization of one of the elements in the set.

Given a nonempty open set $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we denote by $C(\bar{\Omega})$, $C_c^\infty(\Omega)$, and $C^m(\Omega)$ for $m \in \mathbb{N}$ the usual subspaces of the space $C(\Omega)$ of real-valued continuous functions on Ω. For the boundary, Lebesgue measure, and diameter of Ω, we use the symbols $\partial\Omega$, $|\Omega|$, and $\operatorname{diam}(\Omega)$, respectively. The real Lebesgue and Sobolev spaces on Ω are denoted as usual by $L^p(\Omega)$, $W^{m,p}(\Omega)$, and $H^m(\Omega)$ for $m \in \mathbb{N}$, $p \in [1,\infty]$. If Ω is bounded, then we define $H_0^1(\Omega)$ to be the closure of $C_c^\infty(\Omega)$ in $(H^1(\Omega), \|\cdot\|_{H^1(\Omega)})$ and endow it with the norm $\|v\|_{H_0^1(\Omega)} := \||\nabla v|\|_{L^2(\Omega)}$. Here, ∇ denotes the weak gradient and $|\cdot|$ the Euclidean norm. We write $H^{-1}(\Omega)$ for the dual space of $H_0^1(\Omega)$. The (distributional) Laplacian and the (weak) normal derivative are denoted by Δ and ∂_ν, respectively. For $d = 1$, derivatives are denoted by a prime. We use $C_P(\Omega)$ to denote the constant in the Poincaré–Friedrichs inequality $\|v\|_{L^2(\Omega)} \le C_P(\Omega)\|v\|_{H_0^1(\Omega)}$ for $v \in H_0^1(\Omega)$.

Given a locally Lipschitz continuous function $g\colon \mathbb{R} \to \mathbb{R}$, we define $\operatorname{Lip}(g,[a,b]) := \min\{c \ge 0\colon |g(s_1)-g(s_2)| \le c|s_1-s_2|\ \forall s_1,s_2 \in [a,b]\}$ to be the Lipschitz constant of g on $[a,b]$, $a<b$. In this case, we further write $\partial_c g(x) \subset \mathbb{R}$ for Clarke's generalized differential of g at x in the sense of [29, §2.1]. If g is globally Lipschitz, then $\operatorname{Lip}(g)$ denotes the Lipschitz constant of g on $\mathbb{R}$.

2 Semismooth Newton methods for fixed-point problems

In this section, we develop an inexact semismooth Newton framework for the obstacle-type QVI (Q) by addressing the more general fixed-point equation (F).

2.1 Vanilla inexact semismooth Newton method

We begin with Algorithm 1 which constitutes a standard inexact semismooth Newton method for solving (F) and its convergence properties are stated in Theorem 2.1. Throughout this paper, we use the symbol R to denote the residue function

$$R\colon X \to X, \qquad R(x) := x - H(x),$$

of the equation (F).

Algorithm 1 Vanilla inexact semismooth Newton method for the solution of (F)

Input: Initial guess $x_0 \in X$, tolerance $\texttt{tol} \geq 0$, and sequence $\{\rho_i\} \subset [0, \infty)$.
Output: $x^* \in X$ satisfying $\|R(x^*)\|_X \leq \texttt{tol}$, where $R(x) := x - H(x)$.
for $i = 0, 1, 2, 3, \ldots$ **do**
  if $\|R(x_i)\|_X \leq \texttt{tol}$ **then**
    Set $x^* := x_i$ and stop the iteration (convergence reached).
  else
    Compute $z_i \in X$ that satisfies $R(x_i) + G_R(x_i)z_i \approx 0$ in the following sense:

$$\|R(x_i) + G_R(x_i)z_i\|_X \leq \rho_i \|R(x_i)\|_X.$$

    Set $x_{i+1} := x_i + z_i$.
  end if
end for

Theorem 2.1 (Local convergence of Algorithm 1)**.** *Consider the fixed-point equation* (F) *involving a Banach space X and a map $H\colon X \to X$. Let $R\colon X \to X$, $R(x) := x - H(x)$, denote the residue function of* (F)*. Suppose that the following holds:*

i) $B \subset X$ *is an open set containing a solution $\bar{x}$ of* (F)*, i.e., $\bar{x} = H(\bar{x})$;*

ii) $R\colon X \to X$ *is Newton differentiable on B with Newton derivatives $G_R(x)$ for $x \in B$;*

iii) $R\colon X \to X$ *is L-Lipschitz on B for some $L \in [0, \infty)$, i.e.,*

$$\|R(x_1) - R(x_2)\|_X \leq L\|x_1 - x_2\|_X \quad \forall x_1, x_2 \in B; \tag{4}$$

iv) $G_R(x)$ *is invertible for all $x \in B$ and there exists a number $M \in [0, \infty)$ such that*

$$\left\|G_R(x)^{-1}\right\|_{\mathcal{L}(X,X)} \leq M \quad \forall x \in B;$$

v) $\{\rho_i\}$ *satisfies $\{\rho_i\} \subset [0, \rho^*]$ for some $\rho^* \in [0, \infty)$ with $ML\rho^* < 1$.*

Then there exists $r > 0$ such that the standard semismooth Newton method, i.e., Algorithm 1, satisfies the following for all $x_0 \in B_r^X(\bar{x})$:

i) If $\texttt{tol} > 0$ holds, then Algorithm 1 terminates after finitely many steps.

ii) If $\texttt{tol} = 0$ holds, then Algorithm 1 produces iterates that converge finitely or q-linearly to $\bar{x}$.

iii) If $\texttt{tol} = 0$ holds and $\rho_i \to 0$ for $i \to \infty$, then Algorithm 1 produces a sequence of iterates $\{x_i\}$ that converges finitely or q-superlinearly to $\bar{x}$.

Since the proof is relatively standard, we give it in Appendix B. A few remarks:

- **Selection principle.** We assume that a selection principle has been applied for the Newton derivative and hence $G_R(x_i)$ is not set-valued.
- **Implementation of step 7 of Algorithm 1.** The realization of step 7 (evaluating $R(x_i)$ and computing the action of $G_R(x_i)$ or $G_R(x_i)^{-1}$, respectively) is dependent on the precise form of R, G_R, and X. In Section 4.1, we detail how to implement step 7 for a piecewise (bi)linear finite element discretization for a particular instance of the obstacle-type QVI (Q).
- **Measure of inexactness.** Expressing the accuracy requirements on the update steps in terms of the norm of the residue is common practice in the analysis of inexact Newton methods; cf. [30] and [59, §3.2.4]. We remark that, in the context of obstacle-type QVIs, ensuring that the iterates are sufficiently accurate can be a delicate matter. For details on this topic, we refer to Section 4.
- **Choice of the forcing sequence $\{\rho_i\}$.** Some choices for the forcing sequence $\{\rho_i\}$ are

  $$\rho_i = \|x_i - x_{i-1}\|_X \qquad \text{or} \qquad \rho_i = \min(\|x_i - x_{i-1}\|_X, \alpha_i),$$

  where $\{\alpha_i\} \subset (0, \infty)$ is a sequence satisfying $\alpha_i \to 0$ used for safeguarding.

  Note that the forcing sequence $\{\rho_i\}$ must be chosen carefully to avoid oversolving. If it decays too quickly, then the convergence rate does not justify the computational cost whereas decaying too slowly will impair the q-superlinear convergence.
- **Globalization.** Algorithm 1 only guarantees convergence if the initial guess is sufficiently close to the solution. Globalizing semismooth Newton methods is a delicate topic and there is often a tradeoff between the assumptions a particular globalization technique requires and its computational cost. A globalization will likely require many more evaluations of H which might be prohibitively expensive. For example, in the context of obstacle-type QVIs (Q), each evaluation of H requires the solve of an obstacle problem. In the next subsection, we explore a cheap globalization technique that requires a contraction assumption. We also consider a globalization based on a merit function and an Armijo linesearch [48, 32] in Section 4.

2.2 Global q-superlinear convergence for contractive equations

In this subsection, we show that if (F) satisfies a contraction condition, then a small modification of Algorithm 1 will guarantee global convergence. Consider the following assumptions:

Assumption 2.2 (Global contraction assumptions)**.**

i) X is a Banach space;

ii) $H\colon X \to X$ is a Newton differentiable function with Newton derivative $G_H\colon X \to \mathcal{L}(X,X)$ and the residue function $R\colon X \to X$ is endowed with the Newton derivative $G_R := \mathrm{Id} - G_H$;

iii) there exists $\gamma \in [0,1)$ such that $H\colon X \to X$ is globally γ-Lipschitz, i.e.,

$$\|H(x_1) - H(x_2)\|_X \leq \gamma \|x_1 - x_2\|_X \quad \forall x_1, x_2 \in X, \tag{5}$$

and

$$\sup_{x\in X} \|G_H(x)\|_{\mathcal{L}(X,X)} \leq \gamma. \tag{6}$$

Note that the contraction conditions in point iii) of Assumption 2.2 are restrictive in general applications. However, in the field of QVIs, they are assumed anyway in many situations to guarantee the Hadamard well-posedness of the problem; see [7, 61, 9, 10, 6, 8]. Regarding (6), it should be noted that, in practice, the uniform γ-bound on $G_H(x)$ is often a direct consequence of the Lipschitz estimate (5). In Section 2.3, we discuss techniques for localizing the contractivity assumption (5).

We begin our analysis by noting that, in the situation of Assumption 2.2, the existence and uniqueness of solutions of (F) are immediate consequences of the Banach fixed-point theorem.

Lemma 2.3 (Unique solvability of (F))**.** *Suppose that Assumption 2.2 holds. Then the problem* (F) *has a unique solution $\bar{x} \in X$.*

The globalized semismooth inexact Newton method that we propose for the solution of (F), under the assumptions (5) and (6), is stated in Algorithm 2.

Algorithm 2 Globally convergent inexact semismooth Newton method for the solution of (F)

Input: Initial guess $x_0 \in X$, arbitrary constant $\tau^* \in [0,\infty]$, tolerance $\texttt{tol} \geq 0$, and sequences $\{\tau_i\} \subset [0,\infty) \cap [0,\tau^*]$, $\{\rho_i\} \subset [0,\infty)$ satisfying $\tau_i \to 0$, $\rho_i \to 0$.
Output: $x^* \in X$ satisfying $\|R(x^*)\|_X \leq \texttt{tol}$.
for $i = 0, 1, 2, 3, \ldots$ **do**
  if $\|R(x_i)\|_X \leq \texttt{tol}$ **then**
    Set $x^* := x_i$ and stop the iteration (convergence reached).
  else
    Compute $x_B \in X$ that satisfies $x_B \approx H(x_i)$ in the following sense:

$$\|x_B - H(x_i)\|_X \leq \tau_i \|R(x_i)\|_X. \tag{7}$$

    Compute $x_N \in X$ that satisfies $R(x_i) + G_R(x_i)(x_N - x_i) \approx 0$ in the following sense:

$$\|R(x_i) + G_R(x_i)(x_N - x_i)\|_X \leq \rho_i \|R(x_i)\|_X. \tag{8}$$

    if $\|R(x_N)\|_X \leq \|R(x_B)\|_X$ **then**
      Set $x_{i+1} := x_N$.
    else
      Set $x_{i+1} := x_B$.
    end if
  end if
end for

Before we state the convergence properties of Algorithm 2, a few remarks:

- **Vanilla semismooth Newton.** By choosing a very large constant τ^*, a sequence $\{\tau_i\}$ that goes to zero very slowly, and trial iterates x_B with large residues $\|R(x_B)\|_X$, one can essentially switch off the safeguarding by means of the fixed-point iterates in Algorithm 2. If run in such a configuration, Algorithm 2 effectively behaves like a vanilla semismooth Newton method in numerical experiments, i.e., like Algorithm 1 in Section 2.1.
- **Choice of the sequence $\{\tau_i\}$.** The proof of Theorem 2.4 on the convergence of Algorithm 2 below hinges on the fact that, either for all or for all sufficiently large i, one has $\tau_i+\gamma+\gamma\tau_i < 1$, where γ is the contraction factor. If, when solving a particular problem, the value of γ is known, then one can simply choose the constant sequence $\tau_i = (\lambda-\gamma)/(1+\gamma)$ for an arbitrary $\lambda \in (\gamma, 1)$ to ensure this inequality and then the property $\tau_i \to 0$ is not needed.
- **Mesh-independence.** Note that the bound on the iteration index in (9) below is independent of discretization quantities. This is a strong indicator for mesh-independence.

Theorem 2.4 (Finite and global q-superlinear convergence of Algorithm 2)**.** *Suppose that Assumption 2.2 holds. Let $x_0 \in X$ and $\tau^* \in [0,\infty]$ be arbitrary. Let $\{\tau_i\} \subset [0,\infty) \cap [0,\tau^*]$ and $\{\rho_i\} \subset [0,\infty)$ satisfy $\tau_i \to 0$ and $\rho_i \to 0$.*

i) If $\mathtt{tol} > 0$ holds and τ^* satisfies $\tau^* \le (\lambda-\gamma)/(1+\gamma)$ for some $\lambda \in (\gamma,1)$, then the termination criterion in line 4 of Algorithm 2 is triggered for an iteration index $i \in \mathbb{N}_0$ satisfying

$$0 \le i \le \begin{cases} 0 & \text{if } \|R(x_0)\|_X \le \mathtt{tol}, \\ \left\lceil \frac{\ln(\mathtt{tol})-\ln(\|R(x_0)\|_X)}{\ln(\lambda)} \right\rceil & \text{if } \|R(x_0)\|_X > \mathtt{tol}, \end{cases} \tag{9}$$

and the last produced iterate $x^* = x_i$ satisfies

$$\|R(x^*)\|_X \le \mathtt{tol} \qquad \text{and} \qquad \|x^* - \bar{x}\|_X \le \frac{\mathtt{tol}}{1-\gamma}.$$

Here, $\lceil\cdot\rceil : (0,\infty) \to \mathbb{N}$ denotes the operation of rounding up to the nearest larger integer.

ii) If $\mathtt{tol} = 0$ holds, then Algorithm 2 either terminates after finitely many steps with the solution $x^* = \bar{x}$ of (F) *or produces a sequence of iterates $\{x_i\}$ that satisfies*

$$x_i \to \bar{x} \text{ } q\text{-superlinearly in } X \qquad \text{and} \qquad R(x_i) \to 0 \text{ } q\text{-superlinearly in } X.$$

The proof of Theorem 2.4 is a careful intertwining of the convergence proofs of a fixed-point iteration and a semismooth Newton method. We defer it to Appendix B.

2.3 Localization of the contraction assumptions and multiple solutions

Next, we discuss possibilities to localize the contraction conditions in Assumption 2.2iii) whilst still retaining global convergence. The assumptions of this subsection are motivated by results on QVIs with multiple, locally stable solutions (see [7, 61, 9, 10, 6, 8]) and they make it possible to apply our Newton framework also to fixed-point equations (F) that are not uniquely solvable. We consider the following setting.

Assumption 2.5 (Local contraction assumptions)**.**

i) X is a Hilbert space;

ii) $H\colon X \to X$ *is a Newton differentiable function with Newton derivative* $G_H\colon X \to \mathcal{L}(X,X)$*;*

iii) there exist a nonempty closed convex set $B \subset X$ *and a number* $\gamma \in [0,1)$ *such that*

$$\|H(x_1) - H(x_2)\|_X \leq \gamma \|x_1 - x_2\|_X \quad \forall x_1, x_2 \in B \tag{10}$$

and

$$\sup_{x\in B} \|G_H(x)\|_{\mathcal{L}(X,X)} \leq \gamma; \tag{11}$$

iv) the metric projection $P_B\colon X \to B$ *in* $(X, \|\cdot\|_X)$ *onto* B *is Newton differentiable with Newton derivative* $G_{P_B}\colon X \to \mathcal{L}(X,X)$ *and it holds* $\|G_{P_B}(x)\|_{\mathcal{L}(X,X)} \leq 1$ *for all* $x \in X$*.*

We will show in Lemma 2.9 below that Assumption 2.5iv) holds in particular if B is a closed ball in X.

Remark 2.6. *It is possible to drop the requirement that* X *is Hilbert in Assumption 2.5. If this is done, however, one needs additional assumptions on* X *and* B *to ensure that the projection* P_B *is well defined, single-valued, and Lipschitz continuous; see [5] and the references therein. (Note that projections are typically not one-Lipschitz in the Banach space setting; cf. [33, Example 6.1].) We focus on the Hilbert space case in this subsection because it simplifies the presentation, covers almost all practical applications, and yields easier-to-track estimates due to the non-expansiveness of* P_B*.*

The main idea of the following analysis is to resort to the global situation studied in Section 2.2 by composing the function H in (F) with the projection P_B. That is, instead of (F), we consider the fixed-point equation

$$\text{Find } \hat{x} \in X \text{ such that } \hat{x} = H_B(\hat{x}), \tag{F$_{loc}$}$$

where $H_B\colon X \to X$ is defined by $H_B := H \circ P_B$. Note that this approach only works because our algorithm is able to handle nonsmooth functions. As we will see below, by applying Algorithm 2 to the modified equation (F$_{loc}$), we obtain a numerical method that is able to determine precisely the intersection of the solution set $\{x \in X \mid x = H(x)\}$ of the fixed-point equation (F) with the set B. In practical applications, the set B in (F$_{loc}$) could, for example, be a closed ball in which the estimates in (10) and (11) can be proven to hold; see Section 4.3. If one is interested in an abstract local convergence result similar to the classical one discussed in Section 2.1, then one can also assume that B is a closed ball $B_\varepsilon^X(\bar{x})$ that is centered at an isolated solution $\bar{x}$ of (F). In the latter case, the conditions (10) and (11) take a form that is also often encountered in the sensitivity analysis of obstacle-type QVIs; cf. [7, 61, 10]. That Assumption 2.2 holds for the composition $H_B = H \circ P_B$ is proven in the following lemma.

Lemma 2.7 (Properties of H_B)**.** *Suppose that Assumption 2.5 holds. Then:*

i) H_B *is Newton differentiable with Newton derivative* $G_{H_B}(x) := G_H(P_B(x))G_{P_B}(x)$*.*

ii) It holds

$$\|H_B(x_1) - H_B(x_2)\|_X \leq \gamma \|x_1 - x_2\|_X \qquad \forall x_1, x_2 \in X \tag{12}$$

and

$$\sup_{x\in X} \|G_{H_B}(x)\|_{\mathcal{L}(X,X)} \leq \gamma. \tag{13}$$

Proof. Assertion i) follows from the conditions in Assumption 2.5, the chain rule for Newton derivatives (see Lemma A.1), and the global one-Lipschitz continuity of P_B. To prove ii), we first note that (10), (again) the one-Lipschitz continuity of P_B, and the definition of H_B imply

$$\|H_B(x_1) - H_B(x_2)\|_X \leq \gamma \|P_B(x_1) - P_B(x_2)\|_X \leq \gamma \|x_1 - x_2\|_X \quad \forall x_1, x_2 \in X.$$

This establishes (12). Similarly, we obtain from assertion i) of this lemma, (11), and the bound $\|G_{P_B}(x)\|_{\mathcal{L}(X,X)} \leq 1$ that

$$\|G_{H_B}(x)\|_{\mathcal{L}(X,X)} = \|G_H(P_B(x)) G_{P_B}(x)\|_{\mathcal{L}(X,X)} \leq \|G_H(P_B(x))\|_{\mathcal{L}(X,X)} \, \|G_{P_B}(x)\|_{\mathcal{L}(X,X)} \leq \gamma$$

holds for all $x \in X$. This implies (13) and completes the proof. □

Lemma 2.7 shows that (F_{loc}) is covered by the semismooth Newton framework of Section 2.2. This allows us to deduce the following from Theorem 2.4.

Theorem 2.8 (Convergence in the localized setting)**.** *Suppose that Assumption 2.5 holds. Let $x_0 \in X$ and $\tau^* \in [0, \infty]$ be arbitrary. Let $\{\tau_i\} \subset [0, \infty) \cap [0, \tau^*]$ and $\{\rho_i\} \subset [0, \infty)$ satisfy $\tau_i \to 0$ and $\rho_i \to 0$. Then Algorithm 2, applied to the fixed-point problem* (F_{loc}) *with parameters x_0, $\mathtt{tol} = 0$, τ^*, $\{\tau_i\}$, and $\{\rho_i\}$, converges finitely or q-superlinearly to a point $\hat{x} \in X$ and the following is true:*

i) If $\hat{x} \in B$ holds, then (F) *possesses precisely one solution $\bar{x}$ in B and it holds $\bar{x} = \hat{x}$.*

ii) If $\hat{x} \notin B$ holds, then (F) *does not possess a solution in B.*

If, in addition, H satisfies $H(B) \subset B$, then only case i) occurs.

Proof. As the results of Section 2.2 are applicable to (F_{loc}) by Assumption 2.5 and Lemma 2.7, we obtain from Lemma 2.3 and Theorem 2.4 that there is a unique $\hat{x} \in X$ satisfying $\hat{x} = H_B(\hat{x})$ and that Algorithm 2 with $\mathtt{tol} = 0$ converges finitely or q-superlinearly to $\hat{x}$ when applied to (F_{loc}). Suppose now that $\hat{x} \in B$ holds. Then we have $\hat{x} = P_B(\hat{x})$ and $\hat{x}$ is also a solution of (F). Moreover, (F) cannot possess any further solutions $\bar{x} \neq \hat{x}$ in B as those would also solve (F_{loc}) which is uniquely solvable by Lemma 2.3. This proves i). The assertion in ii) is obtained along the same lines. That only case i) occurs if $H(B) \subset B$ holds follows from the structure of (F_{loc}). □

Note that Theorem 2.8 expresses that, if a sufficiently nice set $B \subset X$ (e.g., a ball) satisfying (10) and (11) is given, then Algorithm 2—applied to (F_{loc})—is able to determine precisely whether (F) possesses a solution in B and, in the case of its existence, identify the unique solution of (F) in B with superlinear convergence speed. In applications in which it is important to decide whether solutions are present in certain sets or to determine distinguished solutions of (F) (e.g., maximal and minimal solutions), this type of localization of the framework in Section 2.2 offers an attractive alternative to classical localization approaches; see [25, Proof of Theorem 3.4] and [59, Proof of Theorem 3.13], and compare also with Theorem 3.12 and the experiments in Section 4.3.

We conclude this subsection by establishing that balls $B_r^X(c)$, $r > 0$, $c \in X$, in X indeed satisfy the conditions in Assumption 2.5iv). The proof of the following result relies heavily on the chain and product rule for Newton derivatives. We recall these calculus rules in Appendix A of this paper for the convenience of the reader.

Lemma 2.9 (Projections onto closed balls)**.** *Let $c \in X$ and $r > 0$ be given. Define $B := B_r^X(c)$. Then the projection $P_B \colon X \to X$ is Newton differentiable with Newton derivative*

$$G_{P_B}(x)h := \begin{cases} h & \text{if } \|x-c\|_X \le r, \\ \frac{r}{\|x-c\|_X}\left[h - \left(\frac{x-c}{\|x-c\|_X}, h\right)_X \frac{x-c}{\|x-c\|_X}\right] & \text{if } \|x-c\|_X > r, \end{cases}$$

and it holds

$$\|G_{P_B}(x)\|_{\mathcal{L}(X,X)} \le \min\left(1, \frac{r}{\|x-c\|_X}\right) \qquad \forall x \in X. \tag{14}$$

Proof. Due to the chain rule in Lemma A.1, it suffices to prove the claim for $c = 0$, i.e., $B = B_r^X(0)$. For such a set B, we have

$$P_B(x) = \begin{cases} x & \text{if } \|x\|_X \le r, \\ r\frac{x}{\|x\|_X} & \text{if } \|x\|_X > r, \end{cases} \tag{15}$$

and, thus, $P_B(x) = g(\|x\|_X^2)x$ with $g \colon [0,\infty) \to \mathbb{R}$, $g(s) := r\max(r, s^{1/2})^{-1}$. Note that g is piecewise C^1 with a single kink at $s = r^2$. This implies that $g \colon [0,\infty) \to \mathbb{R}$ is Newton differentiable with Newton derivative

$$G_g(s) := \begin{cases} 0 & \text{if } 0 \le s^{1/2} \le r, \\ -\frac{r}{2s^{3/2}} & \text{if } s^{1/2} > r; \end{cases}$$

see [49, Proposition 2.1]. (This Newton differentiability can also be checked directly by a simple calculation.) As the function $X \ni x \mapsto \|x\|_X^2 \in [0,\infty)$ is smooth and locally Lipschitz continuous, it follows from the chain rule of Lemma A.1 that $f \colon X \to \mathbb{R}$, $x \mapsto g(\|x\|_X^2)$, is Newton differentiable with derivative

$$G_f(x)h := \begin{cases} 0 & \text{if } 0 \le \|x\|_X \le r, \\ -\frac{r}{\|x\|_X^3}(x,h)_X & \text{if } \|x\|_X > r. \end{cases}$$

As f is continuous, the product rule of Lemma A.2 (with $U = X$, $V = \mathbb{R}$, $W = X$, $Z = X$, $P = f$, $Q = \mathrm{Id}$, and a as the multiplication with a scalar in X) now yields that the map $P_B(x) = f(x)x$ is Newton differentiable with Newton derivative

$$G_{P_B}(x)h := \begin{cases} h & \text{if } 0 \le \|x\|_X \le r, \\ -\frac{r(x,h)_X}{\|x\|_X^3}x + \frac{rh}{\|x\|_X} = \frac{r}{\|x\|_X}\left[h - \left(\frac{x}{\|x\|_X}, h\right)_X \frac{x}{\|x\|_X}\right] & \text{if } \|x\|_X > r. \end{cases} \tag{16}$$

This proves the assertion on the Newton differentiability of P_B and the formula for G_{P_B}. To obtain (14), it suffices to note that the expression in the square brackets on the right-hand side of (16) is the projection of h onto the orthogonal complement of the line $\mathbb{R}x \subset X$ and, thus, bounded in the X-norm by $\|h\|_X$. This completes the proof. □

Remark 2.10. *If X is merely Banach (and not necessarily Hilbert), then one can still proceed along the lines of the proof of Lemma 2.9 to establish that the radial projection given by the formula* (15) *is Newton differentiable, provided the norm of X is C^1 away from the origin. If this is done, however, then one cannot bound the $\mathcal{L}(X,X)$-norm of the Newton derivatives $G_{P_B}(x)$ by the right-hand side of* (14) *but only by a worse constant; cf. the Lipschitz estimate proven in [33, §6.1]. In combination with the lost non-expansiveness of P_B, this makes it necessary to impose more restrictive assumptions on the number γ in* (12) *and* (13) *if a Banach space setting is considered.*

2.4 Composite fixed-point equations

With the general convergence theory of Sections 2.1 to 2.3 at hand, we can turn our attention to the special case that the function H in (F) is of the type $S \circ \Phi$, i.e., that the considered fixed-point equation has the form

$$\text{Find } \bar{x} \in X \text{ such that } \bar{x} = S(\Phi(\bar{x})). \tag{F$_c$}$$

As discussed before, this problem formulation is motivated by the structure of the QVI (Q) which can be recast as an equation of the type (F$_c$) by means of the solution operator S of the variational inequality (2) and the inner obstacle map Φ; see the concrete examples in Sections 3 and 4 and the discussion in Section 1. The next three corollaries make precise under which assumptions on the functions S and Φ the problem (F$_c$) is covered by the convergence results in Theorems 2.1, 2.4 and 2.8. Their proofs boil down to applications of the chain rule for Newton derivatives (Lemma A.1) and elementary estimates and are thus omitted.

Corollary 2.11 (Local convergence of Algorithm 1 for (F$_c$))**.** *Suppose that the following assumptions are satisfied:*

i) X *and* Y *are real Banach spaces, and* $D \subset Y$ *is a nonempty set;*

ii) $\Phi\colon X \to D$ *is a Newton differentiable function with Newton derivative* $G_\Phi\colon X \to \mathcal{L}(X,Y)$ *and, for every* $x \in X$*, there exist constants* $C, \varepsilon > 0$ *satisfying*

$$\|\Phi(x+h) - \Phi(x)\|_Y \leq C\|h\|_X \quad \forall h \in B_\varepsilon^X(0);$$

iii) $S\colon D \to X$ *is a Newton differentiable function with Newton derivative* $G_S\colon D \to \mathcal{L}(Y,X)$ *and, for every* $y \in D$*, there exist constants* $C, \varepsilon > 0$ *satisfying*

$$\sup_{w \in D \cap B_\varepsilon^Y(y)} \|G_S(w)\|_{\mathcal{L}(Y,X)} \leq C;$$

iv) there exist a nonempty open set $B \subset X$ *and* $\bar{x} \in B$ *such that* $\bar{x} = S(\Phi(\bar{x}))$*;*

v) there exists a number $L \in [0,\infty)$ *such that* $R = \mathrm{Id} - S \circ \Phi\colon X \to X$ *is* L*-Lipschitz on* B*, i.e.,*

$$\|R(x_1) - R(x_2)\|_X \leq L\|x_1 - x_2\|_X \quad \forall x_1, x_2 \in B;$$

vi) $G_R(x) := \mathrm{Id} - G_S(\Phi(x))G_\Phi(x) \in \mathcal{L}(X,X)$ *is invertible for all* $x \in B$ *and there exists a number* $M \in [0,\infty)$ *with*

$$\left\|G_R(x)^{-1}\right\|_{\mathcal{L}(X,X)} \leq M \quad \forall x \in B;$$

vii) $\{\rho_i\}$ *satisfies* $\{\rho_i\} \subset [0,\rho^*]$ *for some* $\rho^* \in [0,\infty)$ *with* $ML\rho^* < 1$*.*

Then the convergence result in Theorem 2.1 applies to (F$_c$)*. In particular, the sequence of iterates* $\{x_i\}$ *produced by Algorithm 1 converges finitely/*q*-superlinearly to* $\bar{x}$ *if* x_0 *is sufficiently close to* $\bar{x}$*,* `tol` *is chosen as zero, and the forcing sequence* $\{\rho_i\}$ *satisfies* $\rho_i \to 0$*.*

Corollary 2.12 (Global convergence of Algorithm 2 for (F$_c$) with global contraction)**.** *Suppose that the following assumptions are satisfied:*

i) $X, Y, D, S,$ *and* Φ *satisfy the conditions in points i) to iii) of Corollary 2.11;*

ii) there exists $\gamma \in [0,1)$ *such that* $S \circ \Phi \colon X \to X$ *is globally* γ*-Lipschitz, i.e.,*

$$\|S(\Phi(x_1)) - S(\Phi(x_2))\|_X \leq \gamma \|x_1 - x_2\|_X \quad \forall x_1, x_2 \in X, \tag{17}$$

and it holds

$$\sup_{x \in X} \|G_S(\Phi(x)) G_\Phi(x)\|_{\mathcal{L}(X,X)} \leq \gamma. \tag{18}$$

Then all of the results in Section 2.2 apply to (F_c) *with* $H := S \circ \Phi$*. In particular, Algorithm 2 applied to* (F_c) *satisfies the finite and global* q*-superlinear convergence result of Theorem 2.4.*

Corollary 2.13 (Global convergence of Algorithm 2 for (F_c) with local contraction)**.** *Suppose that the following assumptions are satisfied:*

i) $X, Y, D, S,$ *and* Φ *satisfy the conditions in points i) to iii) of Corollary 2.11;*

ii) X *is additionally a Hilbert space;*

iii) there exist a nonempty closed convex set $B \subset X$ *and a number* $\gamma \in [0,1)$ *satisfying*

$$\|S(\Phi(x_1)) - S(\Phi(x_2))\|_X \leq \gamma \|x_1 - x_2\|_X \quad \forall x_1, x_2 \in B \tag{19}$$

and

$$\sup_{x \in B} \|G_S(\Phi(x)) G_\Phi(x)\|_{\mathcal{L}(X,X)} \leq \gamma; \tag{20}$$

iv) the metric projection $P_B \colon X \to B$ *in* $(X, \|\cdot\|_X)$ *onto* B *is Newton differentiable with Newton derivative* $G_{P_B} \colon X \to \mathcal{L}(X,X)$ *and it holds* $\|G_{P_B}(x)\|_{\mathcal{L}(X,X)} \leq 1$ *for all* $x \in X$*.*

Then the results of Section 2.3 apply to (F_c)*. In particular, Algorithm 2, applied to the fixed-point problem* (F_{loc}) *with* $H = S \circ \Phi$*, satisfies the convergence result in Theorem 2.8.*

We conclude this section with a remark on inexact function evaluations in the context of (F_c).

Remark 2.14. *In practice, one typically cannot evaluate the composition* $H(x_i) = S(\Phi(x_i))$ *(and, as a consequence, the residue* $R(x_i) = x_i - S(\Phi(x_i))$*) exactly, as required, e.g., in steps 4 and 7 of Algorithm 1. Instead, one only has access to approximations* $H_\epsilon = S_\epsilon \circ \Phi_\epsilon \colon X \to X$*,* $\epsilon > 0$*, of the function* $H = S \circ \Phi \colon X \to X$*. If, for example,* S *is the solution map of the obstacle problem* (2) *(as in the case of the concrete application* (Q)*), then* S_ϵ *might be the solution map of a PDE-approximation of* (2) *obtained via penalization; see [37, 58, 42]. In the presence of such an inexact oracle* H_ϵ*, one can still easily ensure the accuracy requirements in Algorithms 1 and 2, provided the error* $H - H_\epsilon$ *is controllable by means of an* a-priori *estimate. If we assume, for example, that* $\|H(x_i) - H_\epsilon(x_i)\|_X \leq C\epsilon$ *holds for* x_i *with a known constant* $C > 0$*, then the triangle inequality implies that the following holds for the condition in* (7)*:*

$$\begin{aligned} &\|x_B - H_\epsilon(x_i)\|_X \leq \tau_i \|x_i - H_\epsilon(x_i)\|_X - (1+\tau_i) C\epsilon \\ &\quad \Rightarrow \|x_B - H(x_i)\|_X - C\epsilon \leq \tau_i \|x_i - H(x_i)\|_X + \tau_i C\epsilon - (1+\tau_i) C\epsilon \\ &\quad \Rightarrow \|x_B - H(x_i)\|_X \leq \tau_i \|R(x_i)\|_X. \end{aligned}$$

This shows that, by determining x_B *and* ϵ *with* $\|x_B - H_\epsilon(x_i)\|_X \leq \tau_i \|x_i - H_\epsilon(x_i)\|_X - (1+\tau_i) C\epsilon$*, one can calculate a trial iterate* x_B *satisfying* (7) *without having precise access to* H *and* R*.*

3 Elliptic obstacle-type QVIs

We are now in a position to prove the convergence of Algorithms 1 and 2 applied to the obstacle-type QVIs (Q). Recall that elliptic QVIs of obstacle type correspond to fixed-point equations of the form (F_c) that involve as the map S the solution operator $S\colon H_0^1(\Omega) \to H_0^1(\Omega)$, $\phi \mapsto u$, of the unilateral obstacle problem

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } u \in K(\phi),\ \langle -\Delta u - f, v - u\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0\ \forall v \in K(\phi), \quad (21)$$

where $K(\phi) := \{v \in H_0^1(\Omega) \mid v \leq \phi + \Phi_0 \text{ a.e. in } \Omega\}$. In Section 3.1 below, we show that the solution operator S of (21) satisfies the Newton differentiability and Lipschitz continuity requirements of Corollaries 2.11 to 2.13. Here, we also specify the properties that the obstacle function Φ has to possess in the context of (Q) so that Algorithms 1 and 2 can be applied; see Theorem 3.7. In Section 3.2, we then establish that obstacle mappings Φ arising from semilinear elliptic PDEs satisfy the requirements on the inner function Φ in Corollaries 2.11 to 2.13. Finally, in Section 3.3, we demonstrate that our results can also be employed in the context of nonlinear and implicit obstacle-type VIs.

3.1 The solution operator of the obstacle problem

Throughout this subsection, we consider the following situation.

Assumption 3.1 (QVI-assumptions)**.**

i) $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, *is a nonempty open bounded set;*

ii) p *is a given exponent satisfying* $\max(1, 2d/(d+2)) < p \leq \infty$*;*

iii) $f \in L^p(\Omega)$ *is a given function;*

iv) $\Phi_0\colon \Omega \to (-\infty, \infty]$ *is a quasi-lower semicontinuous, Borel-measurable function such that there exists* $w \in H_0^1(\Omega)$ *satisfying* $w \leq \Phi_0$ *a.e. in* Ω *and such that, for every* $v \in H_0^1(\Omega)$*, it holds* $v \leq \Phi_0$ *a.e. in* Ω *if and only if* $v \leq \Phi_0$ *quasi-everywhere (q.e.) in* Ω*;*

v) Y_p *is the space defined by* $Y_p := \{v \in H_0^1(\Omega) \mid \Delta v \in L^p(\Omega)\}$ *and equipped with the norm*

$$\|v\|_{Y_p} := \|v\|_{H_0^1(\Omega)} + \|\Delta v\|_{L^p(\Omega)}. \quad (22)$$

Note that $(Y_p, \|\cdot\|_{Y_p})$ is a Banach space in the above situation (as one may easily check). For the precise definitions of the terms "quasi-lower semicontinuous" and "quasi-everywhere" and details on the related notion of $H_0^1(\Omega)$-capacity, we refer to [23, §6.4.3]. We remark that Assumption 3.1iv) is automatically satisfied if Ω is a bounded Lipschitz domain and Φ_0 an element of $C(\bar{\Omega}) \cap H^1(\Omega)$ with a nonnegative trace.

Lemma 3.2 (Well-definedness of S)**.** *Suppose that Assumption 3.1 holds. Then* (21) *has a unique solution* $S(\phi) := u$ *for all* $\phi \in H_0^1(\Omega)$*. The associated solution map* $S\colon H_0^1(\Omega) \to H_0^1(\Omega)$, $\phi \mapsto u$*, satisfies*

$$\|S(\phi_1) - S(\phi_2)\|_{H_0^1(\Omega)} \leq \|\phi_1 - \phi_2\|_{H_0^1(\Omega)} \qquad \forall \phi_1, \phi_2 \in H_0^1(\Omega). \quad (23)$$

Proof. We know that $\Phi_0 + \phi \geq w + \phi \in H_0^1(\Omega)$ holds for all $\phi \in H_0^1(\Omega)$, where $w \in H_0^1(\Omega)$ is the function from Assumption 3.1iv). Thus, $K(\phi) \neq \emptyset$ for all $\phi \in H_0^1(\Omega)$. The unique solvability of (21) for all $\phi \in H_0^1(\Omega)$ now follows immediately from [56, Theorem 4:3.1]. Let us now assume that $\phi_1, \phi_2 \in H_0^1(\Omega)$ are given. Define $u_j := S(\phi_j)$, $j = 1, 2$. Then it holds $u_1 - \phi_1 + \phi_2 \leq \phi_2 + \Phi_0$ and $u_2 - \phi_2 + \phi_1 \leq \phi_1 + \Phi_0$ a.e. in Ω and we obtain from the VIs satisfied by u_1 and u_2 that

$$\begin{aligned} 0 &\leq \langle -\Delta u_1 - f, (u_2 - \phi_2 + \phi_1) - u_1 \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} + \langle -\Delta u_2 - f, (u_1 - \phi_1 + \phi_2) - u_2 \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \\ &\leq -\|u_1 - u_2\|^2_{H_0^1(\Omega)} + \|u_1 - u_2\|_{H_0^1(\Omega)} \|\phi_1 - \phi_2\|_{H_0^1(\Omega)}. \end{aligned}$$

This yields (23) and completes the proof. □

Note that, via a simple variable transformation, we can shift the function ϕ in $K(\phi)$ into the source term of (21) and, thus, rewrite S in terms of the solution map $S_0 \colon H^{-1}(\Omega) \to H_0^1(\Omega)$, $w \mapsto u$, of the variational inequality

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } u \in K(0), \ \ \langle -\Delta u - w, v - u \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \quad \forall v \in K(0). \tag{24}$$

Indeed, we have:

Lemma 3.3. *Suppose that Assumption 3.1 holds. Then*

$$S(\phi) = \phi + S_0(f + \Delta\phi) \qquad \forall \phi \in H_0^1(\Omega). \tag{25}$$

Proof. Let $\phi \in H_0^1(\Omega)$ be given. Define $u := S(\phi)$. Then $\tilde{u} := u - \phi$ satisfies $\tilde{u} \in K(0)$ and, by (21),

$$\langle -\Delta\tilde{u} - f - \Delta\phi, v - \tilde{u} \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = \langle -\Delta u - f, v + \phi - u \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \quad \forall v \in K(0).$$

As (24) is uniquely solvable by [56, Theorem 4:3.1], this yields $S_0(f + \Delta\phi) = \tilde{u} = S(\phi) - \phi$. □

Due to (25), the function S inherits various important properties from S_0. To formulate these properties, we require some notation. First, we define the *active set* associated with (21) via

$$A(\phi) := \{x \in \Omega \mid S(\phi)(x) = \phi(x) + \Phi_0(x)\}. \tag{26}$$

(This set is also known as the *coincidence set*.) Following the lines of [41, §1] and [55, §2.2], we interpret the identity (26) in the sense of capacity theory, i.e., we define $A(\phi)$ up to polar sets and w.r.t. quasi-(lower semi)continuous representatives. Note that, due to the quasi-lower semicontinuity of Φ_0 and the inequality $S(\phi) \leq \phi + \Phi_0$, this implies that $A(\phi)$ is a quasi-closed set. Its complement, the quasi-open set $\Omega \setminus A(\phi)$, is called the *inactive set* and denoted by $I(\phi)$ in the following. The quasi-openness of $I(\phi)$ makes it possible to sensibly define the space $H_0^1(I(\phi)) \subset H_0^1(\Omega)$ of all elements of $H_0^1(\Omega)$ that vanish in $A(\phi)$; see again [55, §2.1] or [41, §2]. If $I(\phi)$ is open in the classical sense, then $H_0^1(I(\phi))$ coincides with the closure of $C_c^\infty(I(\phi))$ in $H_0^1(\Omega)$; see [4, Theorem 9.1.3]. By means of $I(\phi)$, we can introduce:

Definition 3.4. *We denote by $Z_S \colon H_0^1(\Omega) \to \mathcal{L}(H^{-1}(\Omega), H_0^1(\Omega))$, $\phi \mapsto Z_S(\phi)$, the map defined by*

$$Z_S(\phi)g = z \qquad \textit{if and only if} \qquad z \in H_0^1(I(\phi)), \ \ \langle -\Delta z - g, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = 0 \ \forall v \in H_0^1(I(\phi))$$

for given $g \in H^{-1}(\Omega)$ and $\phi \in H_0^1(\Omega)$. We further define $\tilde{G}_S \colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))$ via

$$\tilde{G}_S(\phi)h := h + Z_S(\phi)\Delta h \qquad \forall h, \phi \in H_0^1(\Omega). \tag{27}$$

Informally speaking and modulo an extension by zero to the whole of Ω, the operator $Z_S(\phi)$ defined above can be interpreted as the solution map of the Poisson problem with homogeneous Dirichlet boundary conditions on the inactive set $I(\phi)$, i.e., $z = Z_S(\phi)g$ can be characterized as the solution of the boundary value problem

$$-\Delta z = g \text{ in } I(\phi), \qquad z = 0 \text{ on } \partial(I(\phi)).$$

As the notation suggests, the function $\tilde{G}_S$ provides a Newton derivative G_S for S. Indeed, we have:

Lemma 3.5 (Newton differentiability of S)**.** *Suppose that Assumption 3.1 holds. Then the operator S is Newton differentiable as a function from Y_p to $H_0^1(\Omega)$ when endowed with the derivative $G_S := \tilde{G}_S$, $G_S \colon Y_p \to \mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))$.*

Proof. Let $\phi \in Y_p$ be given. Then it holds

$$\begin{aligned}
0 &\leq \limsup_{0<\|h\|_{Y_p}\to 0} \frac{\|S(\phi+h) - S(\phi) - G_S(\phi+h)h\|_{H_0^1(\Omega)}}{\|h\|_{Y_p}} \\
&= \limsup_{0<\|h\|_{Y_p}\to 0} \frac{\|S_0(f+\Delta\phi+\Delta h) - S_0(f+\Delta\phi) - Z_S(\phi+h)\Delta h\|_{H_0^1(\Omega)}}{\|h\|_{Y_p}} \\
&\leq \limsup_{0<\|\Delta h\|_{L^p(\Omega)}\to 0, h\in Y_p} \frac{\|S_0(f+\Delta\phi+\Delta h) - S_0(f+\Delta\phi) - Z_S(\phi+h)\Delta h\|_{H_0^1(\Omega)}}{\|\Delta h\|_{L^p(\Omega)}} \\
&\leq \limsup_{0<\|g\|_{L^p(\Omega)}\to 0} \sup_{H\in\partial_B^{ss} S_0(f+\Delta\phi+g)} \frac{\|S_0(f+\Delta\phi+g) - S_0(f+\Delta\phi) - Hg\|_{H_0^1(\Omega)}}{\|g\|_{L^p(\Omega)}}.
\end{aligned} \tag{28}$$

Here, we have used Lemma 3.3, the definition of $\|\cdot\|_{Y_p}$, and the fact that $Z_S(\phi+h)$ is an element of the so-called *strong-strong Bouligand differential* $\partial_B^{ss} S_0(f+\Delta\phi+\Delta h)$ of S_0 at $f+\Delta\phi+\Delta h \in L^p(\Omega)$ by [55, Theorem 4.3] (and a trivial direct argument in the case $d = 1$). See [55, Definition 2.10] for the precise definition of $\partial_B^{ss} S_0$. From the Newton differentiability properties of S_0 proven in [26, Theorem 4.4] and the inclusions in [55, Proposition 2.11], we obtain that the right-hand side of (28) is equal to zero. This completes the proof. □

For convenience, we drop the tilde in $\tilde{G}_S$ everywhere in the following. From (27), we obtain:

Lemma 3.6 (Boundedness of G_S)**.** *Suppose that Assumption 3.1 holds. Then, for every $\phi \in H_0^1(\Omega)$, it holds $\|G_S(\phi)\|_{\mathcal{L}(H_0^1(\Omega),H_0^1(\Omega))} \leq 1$.*

Proof. Let $\phi \in H_0^1(\Omega)$ be given. Due to its definition, $-Z_S(\phi)\Delta \colon H_0^1(\Omega) \to H_0^1(\Omega)$ is precisely the $H_0^1(\Omega)$-orthogonal projection onto $H_0^1(I(\phi))$. This implies that $G_S(\phi)h = h + Z_S(\phi)\Delta h$ is the $H_0^1(\Omega)$-orthogonal projection onto $H_0^1(I(\phi))^\perp$. The assertion now follows immediately from the fact that projections in Hilbert spaces are non-expansive. □

If we combine Lemmas 3.2, 3.5 and 3.6 and compare with the requirements of the convergence results in Corollaries 2.11 to 2.13, then we arrive at the following main theorem.

Theorem 3.7 (Convergence of Algorithms 1 and 2 for (Q))**.** *Suppose that Assumption 3.1 holds and $\Phi \colon H_0^1(\Omega) \to Y_p$, $\gamma \in [0,1)$, and $B \subset H_0^1(\Omega)$ are given such that the following is true:*

i) Φ is Newton differentiable from $H_0^1(\Omega)$ to Y_p with derivative $G_\Phi \colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), Y_p)$;

ii) Φ *is locally Lipschitz continuous from* $H_0^1(\Omega)$ *to* Y_p*;*

iii) B *is a closed ball of radius* $r > 0$ *in* $H_0^1(\Omega)$ *(not necessarily centered at zero) or* $B = H_0^1(\Omega)$*;*

iv) Φ *satisfies* $\|\Phi(v_1) - \Phi(v_2)\|_{H_0^1(\Omega)} \leq \gamma \|v_1 - v_2\|_{H_0^1(\Omega)}$ *for all* $v_1, v_2 \in B$*;*

v) G_Φ *satisfies* $\|G_\Phi(v)\|_{\mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))} \leq \gamma$ *for all* $v \in B$*.*

Then the QVI

$$\begin{aligned} &\text{Find } u \in H_0^1(\Omega) \text{ such that } u \in K(\Phi(u)),\ \langle -\Delta u - f, v - u \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \ \forall v \in K(\Phi(u)), \\ &\qquad \text{with } K(\Phi(u)) := \{ v \in H_0^1(\Omega) \mid v \leq \Phi(u) + \Phi_0 \text{ a.e. in } \Omega \}, \end{aligned} \tag{Q}$$

is equivalent to the fixed-point equation $u = S(\Phi(u))$ *and the following is true:*

I) If $B = H_0^1(\Omega)$ *holds, then the assumptions of Corollaries 2.11 and 2.12 are satisfied with* $X = H_0^1(\Omega), Y = D = Y_p$*, and* G_S *as in* (27)*. In particular, the results of Section 2.2 apply to* (Q)*,* (Q) *possesses a unique solution* $\bar{u} \in H_0^1(\Omega)$*, and* $\bar{u}$ *can be identified by means of Algorithms 1 and 2, with the convergence guarantees in Theorems 2.1 and 2.4, respectively.*

II) If B *has finite radius, then the assumptions of Corollary 2.13 are satisfied (with the same* X*,* Y*, etc. as in I)), and the results of Section 2.3 apply to* (Q)*. In particular, Algorithm 2, applied to the fixed-point problem* $u = S(\Phi(P_B(u)))$ *satisfies the convergence result in Theorem 2.8.*

Proof. The assertions of the theorem follow immediately from Lemmas 2.9, 3.2, 3.5 and 3.6; the assumptions on Φ; and the fact that $\|v\|_{H_0^1(\Omega)} \leq \|v\|_{Y_p}$ holds for all $v \in Y_p$. □

In the context of the QVI (Q), the Newton derivative G_R appearing in steps 7 and 8 of Algorithms 1 and 2, respectively, is given by

$$G_R(u)h = (\mathrm{Id} - G_\Phi(u))h - Z_S(\Phi(u))\Delta G_\Phi(u)h \qquad \forall u, h \in H_0^1(\Omega).$$

Details on how this and related objects can be realized numerically are given in Section 4.1.

Note that Theorem 3.7 leaves considerable freedom regarding the precise form of Φ and also covers cases in which the pointwise-a.e. constraint in (Q) is only imposed in certain parts of Ω as Assumption 3.1iv) allows to consider functions Φ_0 that take the value $+\infty$. In what follows, we focus primarily on the case that Φ is the solution map of a (potentially nonsmooth) PDE.

3.2 Solution operators of semilinear PDEs as obstacle maps

Next, we show that solution operators of certain semilinear PDEs give rise to obstacle maps Φ that satisfy the assumptions of Theorem 3.7 and, thus, yield obstacle-type QVIs that are covered by the general semismooth Newton framework of Section 2. The operators Φ that we focus on in this subsection are given by

$$\Phi(u) := \varphi T \text{ with } T \text{ as the solution of } \quad kT - \Delta T = g(\Psi_0 + \psi T - u) \text{ in } \Omega, \quad \partial_\nu T = 0 \text{ on } \partial\Omega. \tag{29}$$

Assumption 3.8.

i) $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, *is a bounded Lipschitz domain;*

ii) $\varphi \in C^2(\bar{\Omega})$ *is a given function satisfying* $\varphi = 0$ *on* $\partial\Omega$*;*

iii) $k > 0$ *is a given constant;*

iv) Ψ_0 *is a given function satisfying* $\Psi_0 \in L^{2+\varepsilon}(\Omega)$ *for some* $\varepsilon > 0$*;*

v) $\psi \in L^\infty(\Omega)$ *is a given function satisfying* $\psi \geq 0$ *a.e. in* Ω*;*

vi) $g\colon \mathbb{R} \to \mathbb{R}$ *is locally Lipschitz continuous, nonincreasing, and Newton differentiable with a derivative* $G_g\colon \mathbb{R} \to \mathbb{R}$ *that is Borel-measurable and satisfies* $G_g(s) \in \partial_c g(s)$ *for all* $s \in \mathbb{R}$*;*

vii) Y_2 *is defined by* $Y_2 := \{v \in H_0^1(\Omega) \mid \Delta v \in L^2(\Omega)\}$ *and equipped with the norm* $\|\cdot\|_{Y_2}$ *from* (22)*.*

Note that, as the choice $g(s) = -s$ satisfies Assumption 3.8vi), (29) also covers linear elliptic PDEs with homogeneous Neumann boundary conditions. We remark that the arguments that we use in the following can be extended easily to, e.g., more general differential operators and Dirichlet boundary conditions; cf. the general assumptions in Theorem 3.7. We focus on (29) because this setting covers QVIs such as thermoforming problems [7, §6]. We first consider (29) for arbitrary $d \in \mathbb{N}$ in the next result. (We later consider $d = 1$ as a special case.)

Theorem 3.9 (Properties of (29))**.** *Assume, in addition to Assumption 3.8, that the function* $g\colon \mathbb{R} \to \mathbb{R}$ *is globally Lipschitz continuous. Then the PDE in* (29)*, i.e.,*

$$kT - \Delta T = g(\Psi_0 + \psi T - u) \text{ in } \Omega, \qquad \partial_\nu T = 0 \text{ on } \partial\Omega, \tag{30}$$

possesses a unique (weak) solution $T \in H^1(\Omega)$ *for all* $u \in H_0^1(\Omega)$*. This solution satisfies* $\varphi T \in Y_2$*. Furthermore, the operator* $\Phi\colon H_0^1(\Omega) \to Y_2$, $u \mapsto \varphi T$*, possesses the following properties:*

i) It holds

$$\begin{aligned}&\|\Phi(u_1) - \Phi(u_2)\|_{H_0^1(\Omega)}\\ &\quad \leq C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)} k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)} k^{-1}\right)\|u_1 - u_2\|_{H_0^1(\Omega)} \quad \forall u_1, u_2 \in H_0^1(\Omega).\end{aligned} \tag{31}$$

ii) There exists a constant $C > 0$ *satisfying*

$$\|\Phi(u_1) - \Phi(u_2)\|_{Y_2} \leq C\|u_1 - u_2\|_{H_0^1(\Omega)} \qquad \forall u_1, u_2 \in H_0^1(\Omega). \tag{32}$$

iii) Let $G_\Phi\colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), Y_2)$ *be defined by* $G_\Phi(u)h = \varphi\xi_h$ *with* ξ_h *as the weak solution of*

$$k\xi_h - \Delta\xi_h - G_g(\Psi_0 + \psi T - u)\psi\xi_h = -G_g(\Psi_0 + \psi T - u)h \text{ in } \Omega, \qquad \partial_\nu \xi_h = 0 \text{ on } \partial\Omega, \tag{33}$$

and T *as the solution of* (30)*. Then* Φ *is Newton differentiable as a function from* $H_0^1(\Omega)$ *to* Y_2 *with derivative* G_Φ *and, for all* $u \in H_0^1(\Omega)$*, it holds*

$$\|G_\Phi(u)\|_{\mathcal{L}(H_0^1(\Omega), H_0^1(\Omega))} \leq C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)} k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)} k^{-1}\right). \tag{34}$$

Proof. From Young's inequality and our assumptions on g, Ψ_0, ψ, and Ω, it follows that

$$\int_\Omega |g(\Psi_0+\psi v-u)|^2 \mathrm{d}x \le 2\int_\Omega |g(0)|^2+\mathrm{Lip}(g)^2|\Psi_0+\psi v-u|^2 \mathrm{d}x < \infty \quad \forall v \in H^1(\Omega) \quad \forall u \in H_0^1(\Omega). \tag{35}$$

Thus, $g(\Psi_0+\psi v-u) \in L^2(\Omega)$ for all $v \in H^1(\Omega)$ and $u \in H_0^1(\Omega)$, and we obtain that the operator

$$A_u \colon H^1(\Omega) \to H^1(\Omega)^*, \qquad \langle A_u(v), w\rangle_{H^1(\Omega)^*,H^1(\Omega)} := \int_\Omega kvw+\nabla v\cdot\nabla w - g(\Psi_0+\psi v-u)w \mathrm{d}x,$$

is well defined for all $u \in H_0^1(\Omega)$. Due to the global Lipschitz continuity and monotonicity of g and the nonnegativity and essential boundedness of ψ, we further have

$$\begin{aligned}
&\|A_u(v_1) - A_u(v_2)\|_{H^1(\Omega)^*} \\
&\quad \le \max(k,1)\|v_1-v_2\|_{H^1(\Omega)} + \|g(\Psi_0+\psi v_1-u) - g(\Psi_0+\psi v_2-u)\|_{L^2(\Omega)} \\
&\quad \le \left(\max(k,1) + \mathrm{Lip}(g)\|\psi\|_{L^\infty(\Omega)}\right)\|v_1-v_2\|_{H^1(\Omega)} \quad \forall v_1,v_2 \in H^1(\Omega) \quad \forall u \in H_0^1(\Omega)
\end{aligned}$$

and

$$\begin{aligned}
&\langle A_u(v_1) - A_u(v_2), v_1-v_2\rangle_{H^1(\Omega)^*,H^1(\Omega)} \\
&\quad \ge \int_\Omega k(v_1-v_2)^2 + |\nabla(v_1-v_2)|^2 - \left(g(\Psi_0+\psi v_1-u) - g(\Psi_0+\psi v_2-u)\right)(v_1-v_2)\mathrm{d}x \\
&\quad \ge k\|v_1-v_2\|_{L^2(\Omega)}^2 + \||\nabla(v_1-v_2)|\|_{L^2(\Omega)}^2 \quad \forall v_1,v_2 \in H^1(\Omega) \quad \forall u \in H_0^1(\Omega).
\end{aligned} \tag{36}$$

This shows that $A_u \colon H^1(\Omega) \to H^1(\Omega)^*$ is globally Lipschitz continuous and coercive and, by [56, Theorem 4:3.1], that the equation $A_u(T) = 0$ is uniquely solvable for all $u \in H_0^1(\Omega)$. Due to the definition of A_u, the latter shows that (30) possesses a unique solution $T \in H^1(\Omega)$ for all $u \in H_0^1(\Omega)$. That this solution satisfies $\varphi T \in H_0^1(\Omega)$ and $\Delta(\varphi T) \in L^2(\Omega)$ for all $u \in H_0^1(\Omega)$ follows immediately from the properties of φ, the definition of Δ, (30), and (35). Thus, $\varphi T \in Y_2$ as claimed.

It remains to prove points i), ii), and iii). To this end, let us assume that $u_1, u_2 \in H_0^1(\Omega)$ are given and that $T_1, T_2 \in H^1(\Omega)$ are the associated solutions of (30). Then (36) yields

$$\begin{aligned}
k\|T_1-T_2\|_{L^2(\Omega)}^2 + \||\nabla T_1 - \nabla T_2|\|_{L^2(\Omega)}^2 &\le \langle A_{u_1}(T_1) - A_{u_1}(T_2), T_1-T_2\rangle_{H^1(\Omega)^*,H^1(\Omega)} \\
&= (g(\Psi_0+\psi T_2-u_1) - g(\Psi_0+\psi T_2-u_2), T_1-T_2)_{L^2(\Omega)} \\
&\le \mathrm{Lip}(g)\|u_1-u_2\|_{L^2(\Omega)}\|T_1-T_2\|_{L^2(\Omega)}.
\end{aligned}$$

This implies

$$\|T_1-T_2\|_{L^2(\Omega)} \le \mathrm{Lip}(g)k^{-1}\|u_1-u_2\|_{L^2(\Omega)} \tag{37}$$

and

$$\||\nabla T_1 - \nabla T_2|\|_{L^2(\Omega)} \le \mathrm{Lip}(g)k^{-1/2}\|u_1-u_2\|_{L^2(\Omega)}. \tag{38}$$

In combination with the definition of Φ, we now obtain

$$\begin{aligned}
\|\Phi(u_1)-\Phi(u_2)\|_{H_0^1(\Omega)} &\le \|\varphi\|_{L^\infty(\Omega)}\||\nabla T_1 - \nabla T_2|\|_{L^2(\Omega)} + \||\nabla\varphi|\|_{L^\infty(\Omega)}\|T_1-T_2\|_{L^2(\Omega)} \\
&\le \mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)}k^{-1}\right)\|u_1-u_2\|_{L^2(\Omega)} \\
&\le C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)}k^{-1}\right)\|u_1-u_2\|_{H_0^1(\Omega)}.
\end{aligned}$$

This proves i). To prove ii), we note that (37), the PDEs satisfied by T_1 and T_2, and the global Lipschitz continuity of g imply that

$$\begin{aligned}
\|\Delta T_1 - \Delta T_2\|_{L^2(\Omega)} &= \|kT_1 - kT_2 + g(\Psi_0+\psi T_2-u_2) - g(\Psi_0+\psi T_1-u_1)\|_{L^2(\Omega)} \\
&\le k\|T_1-T_2\|_{L^2(\Omega)} + \mathrm{Lip}(g)\left(\|\psi\|_{L^\infty(\Omega)}\|T_1-T_2\|_{L^2(\Omega)} + \|u_1-u_2\|_{L^2(\Omega)}\right) \\
&\le C\|u_1-u_2\|_{H_0^1(\Omega)}
\end{aligned}$$

holds with some constant $C > 0$. In combination with the assumptions on φ, (37), and (38), this yields (32) (with a potentially larger constant $C > 0$). It remains to prove iii). To this end, let us denote the solution map of the PDE (30) by $P\colon H_0^1(\Omega) \to H^1(\Omega)$, $u \mapsto T$. We claim that

$$\lim_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\|P(u+h) - P(u) - G_P(u+h)h\|_{H^1(\Omega)}}{\|h\|_{H_0^1(\Omega)}} = 0 \qquad \forall u \in H_0^1(\Omega) \tag{39}$$

and

$$\lim_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\|\Delta P(u+h) - \Delta P(u) - \Delta G_P(u+h)h\|_{L^2(\Omega)}}{\|h\|_{H_0^1(\Omega)}} = 0 \qquad \forall u \in H_0^1(\Omega) \tag{40}$$

hold, where $G_P\colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), H^1(\Omega))$ is defined by $G_P(u)h := \xi_h$ for all $u, h \in H_0^1(\Omega)$ with ξ_h being the solution of (33). To establish (39) and (40), let us assume that $u \in H_0^1(\Omega)$ is fixed. From the PDEs solved by $P(u+h)$, $P(u)$, and $G_P(u+h)h$, we obtain that the function $\delta_h := P(u+h) - P(u) - G_P(u+h)h \in H^1(\Omega)$ satisfies

$$\begin{aligned} &k\delta_h - \Delta\delta_h - G_g(\Psi_0 + \psi P(u+h) - u - h)\psi\delta_h \\ &\quad = g(\Psi_0 + \psi P(u+h) - u - h) - g(\Psi_0 + \psi P(u) - u) \\ &\qquad - G_g(\Psi_0 + \psi P(u+h) - u - h)\,(\psi P(u+h) - \psi P(u) - h) =: \rho_h \end{aligned} \tag{41}$$

in Ω and $\partial_\nu \delta_h = 0$ on $\partial\Omega$ for all $h \in H_0^1(\Omega)$. Due to (37) and (38) and our assumptions on ψ, Ψ_0, and Ω, we further know that there exist constants $C, \varepsilon > 0$ such that $\Psi_0 \in L^{2+\varepsilon}(\Omega)$ holds, such that $H^1(\Omega)$ embeds continuously into $L^{2+\varepsilon}(\Omega)$, and such that, for all $h \in H_0^1(\Omega)$, we have

$$\|\psi P(u+h) - \psi P(u) - h\|_{L^{2+\varepsilon}(\Omega)} \le \|\psi\|_{L^\infty(\Omega)}\|P(u+h) - P(u)\|_{L^{2+\varepsilon}(\Omega)} + \|h\|_{L^{2+\varepsilon}(\Omega)} \le C\|h\|_{H_0^1(\Omega)}.$$

In combination with the properties of g and [59, Theorem 3.49], the above entails that

$$\limsup_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\|\rho_h\|_{L^2(\Omega)}}{\|h\|_{H_0^1(\Omega)}} \le C \limsup_{0<\|\tilde{h}\|_{L^{2+\varepsilon}(\Omega)}\to 0} \frac{\left\|g(\tilde{u}+\tilde{h}) - g(\tilde{u}) - G_g(\tilde{u}+\tilde{h})\tilde{h}\right\|_{L^2(\Omega)}}{\|\tilde{h}\|_{L^{2+\varepsilon}(\Omega)}} = 0$$

holds, where $\tilde{u}$ is defined by $\tilde{u} := \Psi_0 + \psi P(u) - u$ and $\tilde{h}$ has been used to replace the perturbation $\psi P(u+h) - \psi P(u) - h$ appearing in the definition of ρ_h. By choosing δ_h as the test function in the weak form of (41)—keeping in mind that $G_g(s) \in \partial_c g(s) \subset [-\mathrm{Lip}(g), 0]$ holds for all $s \in \mathbb{R}$ by Assumption 3.8vi) and that ψ is nonnegative—we now obtain that

$$\frac{k\|\delta_h\|_{L^2(\Omega)}^2 + \||\nabla\delta_h|\|_{L^2(\Omega)}^2}{\|h\|_{H_0^1(\Omega)}} \le \frac{\|\rho_h\|_{L^2(\Omega)}\|\delta_h\|_{L^2(\Omega)}}{\|h\|_{H_0^1(\Omega)}} \qquad \forall h \in H_0^1(\Omega) \setminus \{0\}$$

and, after applying Young's inequality, that there exists a constant $C > 0$ satisfying

$$\limsup_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\|\delta_h\|_{H^1(\Omega)}}{\|h\|_{H_0^1(\Omega)}} \le C \limsup_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\|\rho_h\|_{L^2(\Omega)}}{\|h\|_{H_0^1(\Omega)}} = 0. \tag{42}$$

By revisiting (41) and by exploiting that $|G_g|$ is bounded by $\mathrm{Lip}(g)$, it now also follows that

$$\limsup_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\|\Delta\delta_h\|_{L^2(\Omega)}}{\|h\|_{H_0^1(\Omega)}} \le \limsup_{0<\|h\|_{H_0^1(\Omega)}\to 0} \frac{\big(k + \mathrm{Lip}(g)\|\psi\|_{L^\infty(\Omega)}\big)\,\|\delta_h\|_{L^2(\Omega)} + \|\rho_h\|_{L^2(\Omega)}}{\|h\|_{H_0^1(\Omega)}} = 0. \tag{43}$$

Due to the definition of δ_h, the estimates (42) and (43) establish (39) and (40). Since $\Phi(u) = \varphi P(u)$ holds with φ satisfying $\varphi \in C^2(\bar{\Omega})$ and $\varphi = 0$ on $\partial\Omega$, this yields that Φ is Newton differentiable as a function from $H_0^1(\Omega)$ to Y_2 with derivative $G_\Phi(u)h := \varphi G_P(u)h$ as claimed. It remains to prove the bound in (34). To this end, we note that, for all $u, h \in H_0^1(\Omega)$, we can choose $\xi_h = G_P(u)h$ as the test function in the weak form of (33) to obtain $k\|\xi_h\|_{L^2(\Omega)}^2 + \||\nabla\xi_h|\|_{L^2(\Omega)}^2 \le \mathrm{Lip}(g)\|\xi_h\|_{L^2(\Omega)}\|h\|_{L^2(\Omega)}$. Here, we have again exploited that G_g maps into the interval $[-\mathrm{Lip}(g), 0]$ and that ψ is nonnegative. From the exact same arguments as in (37) and (38), we now obtain $\|\xi_h\|_{L^2(\Omega)} \le \mathrm{Lip}(g)k^{-1}\|h\|_{L^2(\Omega)}$ and $\||\nabla\xi_h|\|_{L^2(\Omega)} \le \mathrm{Lip}(g)k^{-1/2}\|h\|_{L^2(\Omega)}$ for all $u, h \in H_0^1(\Omega)$ and, as a consequence,

$$\begin{aligned}\|G_\Phi(u)h\|_{H_0^1(\Omega)} &\le \|\varphi\|_{L^\infty(\Omega)}\||\nabla\xi_h|\|_{L^2(\Omega)} + \||\nabla\varphi|\|_{L^\infty(\Omega)}\|\xi_h\|_{L^2(\Omega)}\\ &\le \mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)}k^{-1}\right)\|h\|_{L^2(\Omega)}\\ &\le C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)}k^{-1}\right)\|h\|_{H_0^1(\Omega)}.\end{aligned}$$

This establishes (34) and completes the proof. □

Remark 3.10. *If Ω is contained in an open cube with sides of length l, then the Poincaré constant satisfies $C_P(\Omega) \le l$; see [24, Theorem 1.5]. For $d = 1$, one easily checks that $C_P(\Omega) = \mathrm{diam}(\Omega)/\pi$.*

By comparing Theorem 3.9 with Theorem 3.7, we arrive at:

Corollary 3.11. *Suppose that g is globally Lipschitz continuous and that*

$$\gamma := C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)}k^{-1}\right) \in [0, 1).$$

Then the map $\Phi\colon H_0^1(\Omega) \to Y_2$ in Theorem 3.9 satisfies the conditions i) to v) in Theorem 3.7 with the derivative $G_\Phi\colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), Y_2)$ in Theorem 3.9iii), $p = 2$, and $B := H_0^1(\Omega)$.

In the special case $d = 1$, we can exploit that $H^1(\Omega)$ embeds into $L^\infty(\Omega)$ to obtain the following variant of Theorem 3.9 that only requires g to be locally Lipschitz continuous.

Theorem 3.12 (Properties of (29) when $d = 1$)**.** *Assume, in addition to Assumption 3.8, that $d = 1$ and $\Psi_0 \equiv 0$. Then* (30) *possesses a unique (weak) solution $T \in H^1(\Omega)$ for all $u \in H_0^1(\Omega)$. This solution satisfies $\varphi T \in Y_2$. If, further, we define $N_R := |\Omega|^{1/2}R/2 = \mathrm{diam}(\Omega)^{1/2}R/2$ and*

$$M_R := N_R + \|\psi\|_{L^\infty(\Omega)}\left(|\Omega|^{-1/2}k^{-1} + |\Omega|^{1/2}k^{-1/2}\right)\left(\mathrm{Lip}(g, [-N_R, N_R])\frac{2N_R}{\pi} + |g(0)|\right)|\Omega|^{1/2} \tag{44}$$

for all $R > 0$, then the following statements are true for the operator $\Phi\colon H_0^1(\Omega) \to Y_2$, $u \mapsto \varphi T$:

i) For all $R > 0$ and all $u_1, u_2 \in B_R^{H_0^1(\Omega)}(0)$, it holds

$$\begin{aligned}&\|\Phi(u_1) - \Phi(u_2)\|_{H_0^1(\Omega)}\\ &\quad\le \mathrm{Lip}(g, [-M_R, M_R])\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)}k^{-1}\right)\frac{|\Omega|}{\pi}\|u_1 - u_2\|_{H_0^1(\Omega)}.\end{aligned} \tag{45}$$

ii) For all $R > 0$, there exists a constant $C_R > 0$ satisfying

$$\|\Phi(u_1) - \Phi(u_2)\|_{Y_2} \le C_R\|u_1 - u_2\|_{H_0^1(\Omega)} \qquad \forall u_1, u_2 \in B_R^{H_0^1(\Omega)}(0).$$

iii) Let G_Φ be defined as in Theorem 3.9iii). Then Φ is Newton differentiable as a function from $H_0^1(\Omega)$ to Y_2 with derivative G_Φ and, for every $R > 0$, it holds

$$\begin{aligned}&\|G_\Phi(u)\|_{\mathcal{L}(H_0^1(\Omega),H_0^1(\Omega))}\\ &\quad\leq \lim_{t\searrow M_R} \operatorname{Lip}(g,[-t,t])\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)}k^{-1}\right)\frac{|\Omega|}{\pi} \quad \forall u \in B_R^{H_0^1(\Omega)}(0).\end{aligned} \tag{46}$$

Proof. Let $R > 0$ be arbitrary and suppose that $u \in B_R^{H_0^1(\Omega)}(0)$ is given. Define

$$\hat{g}(s) := \begin{cases} g(-M_R) & \text{if } s \leq -M_R,\\ g(s) & \text{if } s \in (-M_R, M_R),\\ g(M_R) & \text{if } s \geq M_R,\end{cases} \qquad \forall s \in \mathbb{R},$$

and consider the differential equation

$$k\hat{T} - \hat{T}'' = \hat{g}(\psi\hat{T} - u) \text{ in } \Omega, \qquad \partial_\nu \hat{T} = 0 \text{ on } \partial\Omega. \tag{47}$$

Then $\hat{g}$ is globally Lipschitz continuous with Lipschitz constant $\operatorname{Lip}(\hat{g}) = \operatorname{Lip}(g,[-M_R,M_R])$ and we obtain from Theorem 3.9 that (47) has a unique (weak) solution $\hat{T} \in H^1(\Omega)$. Note that, by proceeding along the lines of (36), we obtain

$$k\|\hat{T}\|_{L^2(\Omega)}^2 + \|\hat{T}'\|_{L^2(\Omega)}^2 \leq \langle k\hat{T} - \hat{T}'' - \hat{g}(\psi\hat{T} - u) + \hat{g}(-u), \hat{T}\rangle_{H^1(\Omega)^*,H^1(\Omega)} = (\hat{g}(-u),\hat{T})_{L^2(\Omega)}.$$

Analogously to (37) and (38), this yields

$$\|\hat{T}\|_{L^2(\Omega)} \leq k^{-1}\|\hat{g}(-u)\|_{L^2(\Omega)} \qquad \text{and} \qquad \|\hat{T}'\|_{L^2(\Omega)} \leq k^{-1/2}\|\hat{g}(-u)\|_{L^2(\Omega)}.$$

From the mean value theorem and the fact that elements of $H^1(\Omega)$ possess a $C(\bar{\Omega})$-representative for $d = 1$, we further obtain that there exist $\hat{x}, \bar{x} \in \bar{\Omega}$ satisfying

$$\|\hat{T}\|_{L^\infty(\Omega)} = |\hat{T}(\hat{x})| \qquad \text{and} \qquad \hat{T}(\bar{x}) = \frac{1}{|\Omega|}\int_\Omega \hat{T}\mathrm{d}x.$$

Due to the inequality of Cauchy–Schwarz and the fundamental theorem of calculus, this yields

$$\|\hat{T}\|_{L^\infty(\Omega)} = \left|\hat{T}(\bar{x}) + \int_{\bar{x}}^{\hat{x}} \hat{T}'\mathrm{d}x\right| = \left|\frac{1}{|\Omega|}\int_\Omega \hat{T}\mathrm{d}x + \int_{\bar{x}}^{\hat{x}} \hat{T}'\mathrm{d}x\right| \leq |\Omega|^{-1/2}\|\hat{T}\|_{L^2(\Omega)} + |\Omega|^{1/2}\|\hat{T}'\|_{L^2(\Omega)}.$$

In combination with the (sharp) estimate $\|v\|_{L^\infty(\Omega)} \leq |\Omega|^{1/2}\|v\|_{H_0^1(\Omega)}/2$ for all $v \in H_0^1(\Omega)$ (that is easily established by variational calculus), $\|u\|_{H_0^1(\Omega)} \leq R$, and Remark 3.10, it now follows that

$$\begin{aligned}\|\hat{T}\|_{L^\infty(\Omega)} &\leq |\Omega|^{-1/2}\|\hat{T}\|_{L^2(\Omega)} + |\Omega|^{1/2}\|\hat{T}'\|_{L^2(\Omega)}\\ &\leq \left(|\Omega|^{-1/2}k^{-1} + |\Omega|^{1/2}k^{-1/2}\right)\|\hat{g}(-u)\|_{L^2(\Omega)}\\ &\leq \left(|\Omega|^{-1/2}k^{-1} + |\Omega|^{1/2}k^{-1/2}\right)\left(\operatorname{Lip}(\hat{g},[-\|u\|_{L^\infty(\Omega)},\|u\|_{L^\infty(\Omega)}])\|u\|_{L^2(\Omega)} + \|\hat{g}(0)\|_{L^2(\Omega)}\right)\\ &\leq \left(|\Omega|^{-1/2}k^{-1} + |\Omega|^{1/2}k^{-1/2}\right)\left(\operatorname{Lip}(\hat{g},[-N_R,N_R])\frac{2N_R}{\pi} + |g(0)|\right)|\Omega|^{1/2},\end{aligned}$$

and, due to the identity $\operatorname{Lip}(\hat{g},[-N_R,N_R]) = \operatorname{Lip}(g,[-N_R,N_R])$, that

$$\|\psi\hat{T} - u\|_{L^\infty(\Omega)} \leq \|\psi\|_{L^\infty(\Omega)}\|\hat{T}\|_{L^\infty(\Omega)} + \|u\|_{L^\infty(\Omega)} \leq \|\psi\|_{L^\infty(\Omega)}\|\hat{T}\|_{L^\infty(\Omega)} + N_R \leq M_R.$$

As $\hat{g}$ coincides with g on $[-M_R, M_R]$, the last estimate implies that $\hat{T}$ is also a solution of (30). Since (30) can have at most one solution (as one may easily check by means of a contradiction argument), this shows that (30) is uniquely solvable for all $u \in H_0^1(\Omega)$. Note that, if we denote by $P\colon H_0^1(\Omega) \to H^1(\Omega)$, $u \mapsto T$, the solution operator of (30), by $P_R\colon H_0^1(\Omega) \to H^1(\Omega)$, $u \mapsto \hat{T}$, the solution operator of (47), and by $\Phi, \Phi_R\colon H_0^1(\Omega) \to H_0^1(\Omega)$ the functions $\Phi(u) := \varphi P(u)$ and $\Phi_R(u) := \varphi P_R(u)$, respectively, then it follows from the above considerations that $P(u) = P_R(u)$ and $\Phi(u) = \Phi_R(u)$ hold for all $u \in H_0^1(\Omega)$ with $\|u\|_{H_0^1(\Omega)} \leq R$. As Theorem 3.9 applies to Φ_R for all $R > 0$ and since $\mathrm{Lip}(\hat{g}) = \mathrm{Lip}(g, [-M_R, M_R])$, it now follows immediately that $\varphi P(u) \in Y_2$ holds for all $u \in H_0^1(\Omega)$ and that Φ satisfies the assertions in i), ii), and iii). (Note that, in (45) and (46), we have used the identity $C_P(\Omega) = \mathrm{diam}(\Omega)/\pi = |\Omega|/\pi$ obtained from Remark 3.10, and that the limit in (46) is necessary since $\partial_c \hat{g}(s) = \partial_c g(s)$ is only true for $s \in (-M_R, M_R)$.) □

Remark 3.13. *Theorem 3.12 can be extended straightforwardly to the case $\Psi_0 \in L^\infty(\Omega)$. We have assumed that $\Psi_0 \equiv 0$ holds for the sake of simplicity and to avoid additional technicalities.*

Corollary 3.14. *Suppose that $d = 1$ holds, that $\Psi_0 \equiv 0$, and that $R > 0$ is a number such that*

$$\gamma_R := \lim_{t \searrow M_R} \mathrm{Lip}(g, [-t, t]) \left(\|\varphi\|_{L^\infty(\Omega)} k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)} k^{-1} \right) \frac{|\Omega|}{\pi} \in [0, 1),$$

where M_R is defined as in (44)*. Then the map $\Phi\colon H_0^1(\Omega) \to Y_2$, $u \mapsto \varphi T$, in Theorem 3.12 satisfies the conditions i) to v) in Theorem 3.7 with the derivative $G_\Phi\colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), Y_2)$ in Theorem 3.12iii), $p = 2$, and $B := B_R^{H_0^1(\Omega)}(0)$.*

As we will see in Section 4.3, Theorem 3.12 and Corollary 3.14 cover examples of obstacle-type QVIs (Q) that possess several solutions and, as a consequence, can only be studied within the localized framework of Section 2.3.

3.3 An alternative application: semilinear VIs

We conclude this section by demonstrating that the semismooth Newton framework developed in Section 2.4 can also be applied to variational problems that are not of QVI-type. To this end, we consider the semilinear variational inequality

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } u \in K,\ \langle -\Delta u - b_1(u) b_2(u) - f, v - u \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \ \forall v \in K, \tag{48}$$

with $K := \{v \in H_0^1(\Omega) \mid v \leq \Phi_0 \text{ a.e. in } \Omega\}$ and with $b_1(u)b_2(u)$ as the pointwise-a.e. product of two Nemytskii operators induced by (potentially nonsmooth) functions $b_1, b_2\colon \mathbb{R} \to \mathbb{R}$.

Variational inequalities of the type (48) and their PDE-counterparts $-\Delta u = b_1(u)b_2(u) + f$ in Ω, $u = 0$ on $\partial\Omega$, are prototypical examples of nonlinear variational problems that are widely studied in the literature. The existence of solutions to such problems is typically established by proving that a linearization of the solution mapping is contractive on a suitably chosen fixed ball. The localization assumptions in Corollary 2.13 ask for precisely this kind of contraction property. Hence, working with Corollary 2.13 (and, in particular, with the smallness condition iii) in this corollary) is very natural when studying these kinds of nonlinear variational problems.

Throughout this subsection, we work with the following setting:

Assumption 3.15.

i) $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, $d \leq 5$, *is a nonempty open bounded set;*

ii) Φ_0 *is as in Assumption 3.1iv);*

iii) p *is an exponent satisfying* $\max(1, 2d/(d+2)) < p < \infty$ *and, in the case* $d \geq 3$, $p < d/(d-2)$*;*

iv) $f \in L^p(\Omega)$ *is a given function;*

v) $b_1, b_2 \colon \mathbb{R} \to \mathbb{R}$ *are globally Lipschitz continuous and Newton differentiable with derivatives* $G_{b_i} \colon \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, *that are Borel-measurable and satisfy* $G_{b_i}(s) \in \partial_c b_i(s)$ *for all* $s \in \mathbb{R}$.

Note that the condition $d \leq 5$ arises naturally from the requirements on p in Assumption 3.15iii) and that the assumptions on b_1 and b_2 allow to model nonsmooth semilinearities in (48) with linear/quadratic growth (e.g., $\max(0, u)(u + \cos(u))$ with $b_1(s) := \max(0, s)$, $b_2(s) := s + \cos(s)$). Recall further that standard Sobolev embeddings imply that $H_0^1(\Omega)$ is continuously embedded into $L^q(\Omega)$ for all

$$1 \leq q \begin{cases} \leq \infty & \text{if } d = 1, \\ < \infty & \text{if } d = 2, \\ \leq \frac{2d}{d-2} & \text{if } d \geq 3. \end{cases}$$

In combination with Assumption 3.15iii), this yields that $H_0^1(\Omega)$ embeds continuously into $L^{2p}(\Omega)$, that $L^p(\Omega)$ embeds continuously into $H^{-1}(\Omega)$, and, since the Hölder conjugate $p' := p/(p-1)$ of p satisfies $p' < 2d/(d-2)$ for $d \geq 3$, that $H_0^1(\Omega)$ embeds continuously into $L^{p'}(\Omega)$.

To see that (48) is equivalent to a fixed-point problem of the form (F_c) and, thus, covered by the analysis of Section 2.4, we have to study the mapping $H_0^1(\Omega) \ni u \mapsto b_1(u)b_2(u) + f \in L^p(\Omega)$ which we shall denote by Φ, i.e.,

$$\Phi \colon H_0^1(\Omega) \to L^p(\Omega), \qquad \Phi(u) := b_1(u)b_2(u) + f. \tag{49}$$

Lemma 3.16. *Suppose that Assumption 3.15 holds. Then the function* $\Phi \colon H_0^1(\Omega) \to L^p(\Omega)$ *in* (49) *is well defined and Newton differentiable with the derivative* $G_\Phi \colon H_0^1(\Omega) \to \mathcal{L}(H_0^1(\Omega), L^p(\Omega))$ *given by*

$$G_\Phi(u)h := b_1(u)G_{b_2}(u)h + b_2(u)G_{b_1}(u)h \qquad \forall u, h \in H_0^1(\Omega). \tag{50}$$

Further, it holds

$$\|G_\Phi(u)\|_{\mathcal{L}(H_0^1(\Omega), L^p(\Omega))} \leq \left(\mathrm{Lip}(b_1) \|b_2(u)\|_{L^{2p}(\Omega)} + \mathrm{Lip}(b_2) \|b_1(u)\|_{L^{2p}(\Omega)} \right) \|\iota_{2p}\|_{\mathcal{L}(H_0^1(\Omega), L^{2p}(\Omega))} \tag{51}$$

for all $u \in H_0^1(\Omega)$ *and*

$$\begin{aligned} &\|\Phi(u_1) - \Phi(u_2)\|_{L^p(\Omega)} \\ &\leq \left(\mathrm{Lip}(b_1) \|b_2(u_2)\|_{L^{2p}(\Omega)} + \mathrm{Lip}(b_2) \|b_1(u_1)\|_{L^{2p}(\Omega)} \right) \|\iota_{2p}\|_{\mathcal{L}(H_0^1(\Omega), L^{2p}(\Omega))} \|u_1 - u_2\|_{H_0^1(\Omega)} \end{aligned} \tag{52}$$

for all $u_1, u_2 \in H_0^1(\Omega)$, *where* $\iota_{2p} \in \mathcal{L}(H_0^1(\Omega), L^{2p}(\Omega))$ *denotes the embedding of* $H_0^1(\Omega)$ *into* $L^{2p}(\Omega)$.

Proof. From Hölder's inequality, the triangle inequality, our assumptions on p, the Lipschitz continuity of b_1 and b_2, and the Sobolev embeddings, we obtain that

$$\begin{aligned}
&\|b_1(u)b_2(u) + f\|_{L^p(\Omega)} \\
&\quad \le \|f\|_{L^p(\Omega)} + \|b_1(u)\|_{L^{2p}(\Omega)} \|b_2(u)\|_{L^{2p}(\Omega)} \\
&\quad \le \|f\|_{L^p(\Omega)} + \Big(\operatorname{Lip}(b_1) \|u\|_{L^{2p}(\Omega)} + \|b_1(0)\|_{L^{2p}(\Omega)}\Big) \Big(\operatorname{Lip}(b_2) \|u\|_{L^{2p}(\Omega)} + \|b_2(0)\|_{L^{2p}(\Omega)}\Big) \\
&\quad \le \|f\|_{L^p(\Omega)} + \prod_{i=1}^{2} \Big(\operatorname{Lip}(b_i)\|\iota_{2p}\|_{\mathcal{L}(H_0^1(\Omega),L^{2p}(\Omega))} \|u\|_{H_0^1(\Omega)} + \|b_i(0)\|_{L^{2p}(\Omega)}\Big)
\end{aligned}$$

holds for all $u \in H_0^1(\Omega)$. This shows that Φ is well defined as a function from $H_0^1(\Omega)$ to $L^p(\Omega)$. Consider now an exponent $q > 2p$ that satisfies $q < \infty$ in the case $d \le 2$ and $q < 2d/(d-2)$ in the case $d \ge 3$. (Such a q exists by Assumption 3.15iii).) Then it follows from [59, Theorem 3.49] that the maps $F_i \colon L^q(\Omega) \to L^{2p}(\Omega)$, $u \mapsto b_i(u)$, $i = 1, 2$, are Newton differentiable with Newton derivatives $G_{F_i}(u)h = G_{b_i}(u)h$. In combination with Lemma A.2 (applied to $U = L^q(\Omega)$, $V = W = L^{2p}(\Omega)$, $Z = L^p(\Omega)$, and $a \colon L^{2p}(\Omega) \times L^{2p}(\Omega) \to L^p(\Omega)$ as the pointwise-a.e. multiplication), this yields that the function $F \colon L^q(\Omega) \to L^p(\Omega)$, $u \mapsto b_1(u)b_2(u)$, is Newton differentiable with Newton derivative $G_F(u)h = b_1(u)G_{b_2}(u)h + b_2(u)G_{b_1}(u)h$. That Φ is Newton differentiable as a function from $H_0^1(\Omega)$ to $L^p(\Omega)$ with the derivative in (50) now follows immediately from the Sobolev embeddings and the sum rule for Newton derivatives. Note that (51) is an immediate consequence of the estimate

$$\begin{aligned}
&\|G_\Phi(u)\|_{\mathcal{L}(H_0^1(\Omega),L^p(\Omega))} \\
&\quad = \sup_{h \in H_0^1(\Omega), \|h\|_{H_0^1(\Omega)} \le 1} \|b_1(u)G_{b_2}(u)h + b_2(u)G_{b_1}(u)h\|_{L^p(\Omega)} \\
&\quad \le \sup_{h \in H_0^1(\Omega), \|h\|_{H_0^1(\Omega)} \le 1} \Big(\|b_1(u)\|_{L^{2p}(\Omega)} \operatorname{Lip}(b_2) + \|b_2(u)\|_{L^{2p}(\Omega)} \operatorname{Lip}(b_1)\Big) \|h\|_{L^{2p}(\Omega)} \\
&\quad \le \Big(\operatorname{Lip}(b_1)\|b_2(u)\|_{L^{2p}(\Omega)} + \operatorname{Lip}(b_2)\|b_1(u)\|_{L^{2p}(\Omega)}\Big) \|\iota_{2p}\|_{\mathcal{L}(H_0^1(\Omega),L^{2p}(\Omega))} \quad \forall u \in H_0^1(\Omega).
\end{aligned}$$

Similarly, we also obtain

$$\begin{aligned}
&\|\Phi(u_1) - \Phi(u_2)\|_{L^p(\Omega)} \\
&\quad = \|b_1(u_1)b_2(u_1) - b_1(u_1)b_2(u_2) + b_1(u_1)b_2(u_2) - b_1(u_2)b_2(u_2)\|_{L^p(\Omega)} \\
&\quad \le \|b_1(u_1)\|_{L^{2p}(\Omega)} \|b_2(u_1) - b_2(u_2)\|_{L^{2p}(\Omega)} + \|b_2(u_2)\|_{L^{2p}(\Omega)} \|b_1(u_1) - b_1(u_2)\|_{L^{2p}(\Omega)} \\
&\quad \le \Big(\operatorname{Lip}(b_2)\|b_1(u_1)\|_{L^{2p}(\Omega)} + \operatorname{Lip}(b_1)\|b_2(u_2)\|_{L^{2p}(\Omega)}\Big) \|u_1 - u_2\|_{L^{2p}(\Omega)} \\
&\quad \le \Big(\operatorname{Lip}(b_2)\|b_1(u_1)\|_{L^{2p}(\Omega)} + \operatorname{Lip}(b_1)\|b_2(u_2)\|_{L^{2p}(\Omega)}\Big) \|\iota_{2p}\|_{\mathcal{L}(H_0^1(\Omega),L^{2p}(\Omega))}\|u_1 - u_2\|_{H_0^1(\Omega)}
\end{aligned}$$

for all $u_1, u_2 \in H_0^1(\Omega)$. This establishes (52) and completes the proof. □

Recall that our assumptions on p imply that $L^p(\Omega)$ embeds continuously into $H^{-1}(\Omega)$. In combination with Lemma 3.16, this allows us to recast the semilinear VI (48) in the form

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } u = S_0(\Phi(u)), \tag{53}$$

where $\Phi \colon H_0^1(\Omega) \to L^p(\Omega)$ is defined as in (49) and $S_0 \colon H^{-1}(\Omega) \to H_0^1(\Omega)$ again denotes the solution map of (24), i.e., the function that maps a source term $w \in H^{-1}(\Omega)$ to the solution u of

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } u \in K, \ \ \langle -\Delta u - w, v - u\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \ge 0 \quad \forall v \in K. \tag{54}$$

From [26, 28], we (again) obtain that S_0 is Newton differentiable as a function from $L^p(\Omega)$ to $H_0^1(\Omega)$. For the convenience of the reader, we restate this Newton differentiability property in a way that fits to the application context of (48).

Lemma 3.17. *Suppose that Assumption 3.15 holds. Then the solution map $S_0\colon H^{-1}(\Omega) \to H_0^1(\Omega)$, $w \mapsto u$, of the obstacle problem* (54) *is well defined. Further, S_0 is Newton differentiable as a function from $L^p(\Omega)$ to $H_0^1(\Omega)$ with the Newton derivative $G_{S_0}\colon L^p(\Omega) \to \mathcal{L}(L^p(\Omega), H_0^1(\Omega))$ defined by*

$$G_{S_0}(w)h = z \quad \text{if and only if} \quad z \in H_0^1(I(w)),\ \ \langle -\Delta z - h, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0\ \forall v \in H_0^1(I(w))$$

for all $w, h \in L^p(\Omega)$. Here, $I(w) := \Omega \setminus \{x \in \Omega \mid S_0(w)(x) = \Phi_0(x)\}$ denotes the inactive set associated with w, defined in the same sense as in (26)*. Moreover, it holds*

$$\|G_{S_0}(w)\|_{\mathcal{L}(L^p(\Omega),H_0^1(\Omega))} \le \|\iota_{p'}\|_{\mathcal{L}(H_0^1(\Omega),L^{p'}(\Omega))} \qquad \forall w \in L^p(\Omega) \tag{55}$$

and

$$\|S_0(w_1) - S_0(w_2)\|_{H_0^1(\Omega)} \le \|\iota_{p'}\|_{\mathcal{L}(H_0^1(\Omega),L^{p'}(\Omega))}\|w_1 - w_2\|_{L^p(\Omega)} \qquad \forall w_1, w_2 \in L^p(\Omega), \tag{56}$$

where $p' := p/(p-1)$ denotes the Hölder conjugate of p and $\iota_{p'}$ the embedding of $H_0^1(\Omega)$ into $L^{p'}(\Omega)$.

Proof. That S_0 is well defined again follows from our assumptions on Φ_0 and [56, Theorem 4:3.1]. That S_0 is Newton differentiable as a function $S_0\colon L^p(\Omega) \to H_0^1(\Omega)$ with derivative G_{S_0} is a consequence of [55, Theorem 4.3], the inclusions in [55, Proposition 2.11], and [26, Theorem 4.4]; cf. the proof of Lemma 3.5. The estimate (55) follows trivially from

$$\begin{aligned}\|G_{S_0}(w)h\|_{H_0^1(\Omega)}^2 = \langle h, G_{S_0}(w)h\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} &\le \|h\|_{L^p(\Omega)}\|G_{S_0}(w)h\|_{L^{p'}(\Omega)} \\ &\le \|\iota_{p'}\|_{\mathcal{L}(H_0^1(\Omega),L^{p'}(\Omega))}\|h\|_{L^p(\Omega)}\|G_{S_0}(w)h\|_{H_0^1(\Omega)}\end{aligned}$$

for all $w, h \in L^p(\Omega)$. The Lipschitz estimate (56) is obtained along the exact same lines by choosing $S_0(w_1)$ as the test function in the VI for $S_0(w_2)$, by choosing $S_0(w_2)$ as the test function in the VI for $S_0(w_1)$, and by adding the resulting inequalities. □

From Lemmas 3.16 and 3.17, it follows immediately that the VI (48)—or, more precisely, its reformulation (53)—is covered by the analysis of Section 2.4.

Corollary 3.18. *Suppose that Assumption 3.15 holds and define*

$$\begin{aligned}\gamma_R := \|\iota_{2p}\|_{\mathcal{L}(H_0^1(\Omega),L^{2p}(\Omega))}&\|\iota_{p'}\|_{\mathcal{L}(H_0^1(\Omega),L^{p'}(\Omega))} \\ &\cdot \sup_{v_i \in H_0^1(\Omega), \|v_i\|_{H_0^1(\Omega)} \le R, i=1,2} \Big(\mathrm{Lip}(b_1)\|b_2(v_1)\|_{L^{2p}(\Omega)} + \mathrm{Lip}(b_2)\|b_1(v_2)\|_{L^{2p}(\Omega)}\Big) \quad \forall R \in (0,\infty],\end{aligned}$$

where ι_{2p} and $\iota_{p'}$ denote the embeddings introduced in Lemmas 3.16 and 3.17. Let $R \in (0,\infty)$ be given such that $\gamma_R \in [0,1)$ holds. Then the maps $\Phi\colon H_0^1(\Omega) \to L^p(\Omega)$ and $S_0\colon L^p(\Omega) \to H_0^1(\Omega)$ from Lemmas 3.16 and 3.17 satisfy the assumptions of Corollary 2.13 with

$$X = H_0^1(\Omega), \qquad Y = D = L^p(\Omega), \qquad S = S_0, \qquad B = B_R^{H_0^1(\Omega)}(0), \quad \text{and} \quad \gamma = \gamma_R. \tag{57}$$

Furthermore, if $\gamma_\infty \in [0,1)$ holds, then $\Phi\colon H_0^1(\Omega) \to L^p(\Omega)$ and $S_0\colon L^p(\Omega) \to H_0^1(\Omega)$ satisfy the assumptions of Corollary 2.12 with $\gamma := \gamma_\infty$ and X, Y, D, and S as in (57)*.*

Proof. The assertions of the corollary follow straightforwardly from Lemmas 2.9, 3.16 and 3.17 by comparing with the assumptions of Corollaries 2.12 and 2.13. □

From Corollary 3.18, we obtain that (53) is amenable to Algorithm 2 provided the Lipschitz constants and the growth behavior of the functions b_1 and b_2 are suitable; analogously to the obstacle-type QVIs in Corollaries 3.11 and 3.14.

4 Numerical experiments for obstacle-type QVIs

If we combine the results of Sections 3.1 and 3.2, then we obtain that the analysis of Section 2.4 applies to obstacle-type QVIs of the form

Find $u \in H_0^1(\Omega)$ satisfying

$$u \leq \Phi_0 + \Phi(u), \quad \langle -\Delta u - f, v - u\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \; \forall v \in H_0^1(\Omega), v \leq \Phi_0 + \Phi(u),$$

with $\Phi(u)$ given by $\Phi(u) := \varphi T$ and T as the solution of

$$kT - \Delta T = g(\Psi_0 + \psi T - u) \text{ in } \Omega, \qquad \partial_\nu T = 0 \text{ on } \partial\Omega, \tag{58}$$

provided the quantities Ω, Φ_0, f, φ, k, g, Ψ_0, and ψ in (58) satisfy the conditions in Assumptions 3.1 and 3.8 for $p = 2$ and the data is such that Lemmas 3.2, 3.5 and 3.6 and Theorems 3.9 and 3.12 allow to establish the conditions (17) and (18) (or (19) and (20), respectively) for $X = H_0^1(\Omega)$. Note that (58) covers the so-called thermoforming problem [7, §6] as a special case.

Remark 4.1 (Thermoforming QVI)**.** *For $d = 2$,* (58)*, with $\Psi_0 \equiv \Phi_0$ and $\psi \equiv \varphi$, provides a simple model for the problem of determining the displacement u of an elastic membrane, clamped at the boundary $\partial\Omega$, that has been heated and is pushed by means of an external force f into a metallic mould with original shape Φ_0 and deformation $\Phi(u)$. The deformation is due to the mould's temperature field T which varies according to the membrane's temperature. The function g then models how the temperature T is affected by the distance between the membrane and the mould. For more details on the thermoforming problem, we refer to [7, §6] and the references therein.*

In the present section, we provide several examples of QVIs of the type (58) for which all necessary assumptions are satisfied and present the results that are obtained when Algorithms 1 and 2 are applied. Our goal is in particular to demonstrate the q-superlinear convergence and mesh-independence of our semismooth Newton method. We begin with a detailed description of how we realized and implemented Algorithms 1 and 2 in the situation of (58).

4.1 Implementation details

Data availability and packages used: We implement our experiments in the open-source language Julia [22]. Our implementation makes use of the `Gridap` package for the finite element discretization of the variational problems [13, 60] as well as the `NLsolve` [1] and `LineSearches` [2] packages, for the solution of the (smooth) nonlinear equations that arise when evaluating, e.g., the solution map of the obstacle problem (21) by means of a regularization approach. For the sake of reproducibility, the scripts used to generate the tables and plots depicted in the following subsections can be found in the `SemismoothQVIs` package [51]. The version of `SemismoothQVIs` run in our experiments is archived on Zenodo [52]. A Python implementation of Algorithms 1 and 2 is also available with Firedrake [35] as the finite element backend; cf. [50].

Evaluation of $S \circ \Phi$: The semismooth Newton methods in Algorithms 1 and 2 require the computation of the quantity $S(\Phi(u_i))$ in each iteration. (Here and in what follows, we write u_i etc. instead of x_i etc. to conform with the notation in (58).) The computation of $S(\Phi(u_i))$ can be split into two steps:

(E1) Given u_i, compute $\Phi(u_i)$. (E2) Given ϕ, compute $S(\phi)$.

For the QVI (58), (E1) is equivalent to the solution of a (potentially nonsmooth) semilinear PDE and hence (E1) can be realized by means of a (semismooth) Newton method.

The evaluation of S in (E2) is equivalent to solving the obstacle problem (21). As this is a classical problem, a myriad of algorithms exist for its numerical solution. In this paper, we opt for a path-following smoothed Moreau–Yosida regularization (PFMY) [3] followed by a feasibility restoration by means of iterations of the primal-dual active set (PDAS) method [36]. By combining these algorithms, we find that the number of iterations needed for the evaluation (E2) does not grow uncontrollably as the mesh width goes to zero and the feasibility of $S(\Phi(u_i))$ is guaranteed.

Recall that the PFMY-algorithm for the solution of the obstacle problem relies on the Huber-type function $\sigma_\rho \colon \mathbb{R} \to \mathbb{R}$ defined by

$$\sigma_\rho(u) := \begin{cases} 0 & \text{if } u \le 0, \\ u^2/(2\rho) & \text{if } 0 < u < \rho, \\ u - \rho/2 & \text{if } u \ge \rho, \end{cases} \qquad \text{for } \rho > 0. \tag{59}$$

A PFMY-algorithm for approximating the solution $u = S(\phi)$ of (21) for given f, ϕ, and Φ_0 chooses a sequence of Moreau–Yosida parameters $\rho_0 > \rho_1 > \ldots > \rho_J > 0$ and solves for each j the subproblem

$$u_{\rho_j} \in H^1_0(\Omega), \quad (u_{\rho_j}, v)_{H^1_0(\Omega)} + \left(\rho_j^{-1}\sigma_{\rho_j}(u_{\rho_j} - \phi - \Phi_0) - f, v\right)_{L^2(\Omega)} = 0 \quad \forall v \in H^1_0(\Omega). \tag{60}$$

When u_{ρ_j} is computed, it serves as the initial guess for the iterative solution of the subsequent subproblem with parameter ρ_{j+1}. In our implementation, all the subproblems that appear were solved by means of a classical Newton algorithm. The last PFMY-iterate u_{ρ_J} is used as the initial guess for the PDAS-method in our solver which restores the feasibility of the solution.

Action of $G_R(u_i)^{-1}$: Next, we detail how to compute the Newton iterate u_N in steps 7 and 8 of Algorithms 1 and 2, respectively. In view of Lemma B.1ii), given the previous iterate u_i and $u_B := S(\Phi(u_i))$, to determine u_N, one has to (approximately) solve the linear system

$$(\mathrm{Id} - G_S(\Phi(u_i))G_\Phi(u_i))\,\delta u_N = u_B - u_i \tag{61}$$

for δu_N and set $u_N := u_i + \delta u_N$. In order to realize the composition $G_S(\Phi(u_i))G_\Phi(u_i)$ in (61), we introduce the auxiliary variables $\eta := G_\Phi(u_i)\delta u_N$ and $\mu := G_S(\Phi(u_i))\eta - \eta = Z_S(\Phi(u_i))\Delta\eta$, where Z_S is defined as in Definition 3.4. As discussed in Theorem 3.9iii), η then satisfies $\eta = \varphi\xi$ with $\xi \in H^1(\Omega)$ as the solution of (33) with $h := \delta u_N$. From Definition 3.4, we further obtain that μ is the weak solution of $-\Delta\mu - \Delta\eta = 0$ in I_i, $\mu = 0$ in $\bar\Omega \setminus I_i$, where I_i and A_i denote the inactive and active set associated with the iterate u_i, respectively, i.e., $A_i = A(\Phi(u_i)) = \{x \in \Omega \mid u_B(x) = \Phi_0(x) + \Phi(u_i)(x)\}$ and $I_i = \Omega \setminus A_i$. By rewriting all of these PDEs in variational form and substituting, it follows that (61) can be recast as:

Find $(\delta u_N, \xi, \mu) \in H^1_0(\Omega) \times H^1(\Omega) \times H^1_0(I_i)$ such that

$$(\delta u_N - \mu - \varphi\xi, v)_{L^2(\Omega)} - (u_B - u_i, v)_{L^2(\Omega)} = 0 \qquad \forall v \in H^1_0(\Omega), \tag{62}$$

$$(\nabla\xi, \nabla\zeta)_{L^2(\Omega)} + (k\xi + G_g(\Psi_0 + \psi T_i - u_i)(\delta u_N - \psi\xi), \zeta)_{L^2(\Omega)} = 0 \qquad \forall \zeta \in H^1(\Omega), \tag{63}$$

$$(\nabla\mu + \nabla(\varphi\xi), \nabla q)_{L^2(\Omega)} = 0 \qquad \forall q \in H^1_0(I_i), \tag{64}$$

where $T_i \in H^1(\Omega)$ denotes the solution of (30) with $u = u_i$. Note that the system (62)–(64) is linear in δu_N, ξ, and μ and, therefore, reduces to a linear system solve after discretization. The act of encoding the inactive set in (64) in the discretized linear system is discretization dependent. How this is achieved for a piecewise (bi)linear finite element discretization is discussed at the end of this subsection.

Comparison methods: In our numerical experiments, we compare Algorithms 1 and 2 with three alternative approaches for the numerical solution of (58), namely:

(C1) a pure fixed-point method;

(C2) a Newton method applied to a smoothed Moreau–Yosida regularization of the QVI (58) with fixed ρ;

(C3) Algorithm 1 but with a backtracking Armijo linesearch applied to each Newton update. We use the backtracking Armijo linesearch as described in [48, §3.5] and implemented in the `LineSearches` package [2] with the merit function $\|R(u_i)\|_{H^1(\Omega)}$.

The iterates of the fixed-point method are defined by $u_{i+1} = S(\Phi(u_i))$ for $i = 1, 2, \ldots$, where $S(\Phi(u_i))$ is computed as described above. If the map $S \circ \Phi$ is contractive, then this type of algorithm can be expected to converge linearly to the QVI-solution.

For a given $\rho > 0$, a smoothed Moreau–Yosida regularization of the QVI (58) corresponds to replacing the solution operator S of the obstacle problem (21) by the solution map S_ρ of the mollified problem (60) wherever it appears. For the inner solver, this means that $u_B = S(\Phi(u_i))$ is approximately calculated by replacing (21) with its Moreau–Yosida regularization (60) and by subsequently applying Newton's method. In (61), we then approximate the update formula $u_N := u_i + \delta u_N$ via $u_{N,\rho} = u_i + \delta u_{N,\rho}$, where $\delta u_{N,\rho}$ is obtained by solving the system

$$
\begin{aligned}
&\text{Find } (\delta u_{N,\rho}, \xi, w) \in H^1_0(\Omega) \times H^1(\Omega) \times H^1_0(\Omega) \text{ such that} \\
&(\delta u_{N,\rho} - w, v)_{L^2(\Omega)} - (u_B - u_i, v)_{L^2(\Omega)} = 0 && \forall v \in H^1_0(\Omega), \quad (65)\\
&(\nabla \xi, \nabla \zeta)_{L^2(\Omega)} + (k\xi + G_g(\Psi_0 + \psi T_i - u_i)(\delta u_{N,\rho} - \psi\xi), \zeta)_{L^2(\Omega)} = 0 && \forall \zeta \in H^1(\Omega), \quad (66)\\
&(\nabla w, \nabla q)_{L^2(\Omega)} + \rho^{-1} \left(G_{\sigma_\rho}(u_i - \Phi_0 - \varphi T_i)(w - \varphi\xi), q\right)_{L^2(\Omega)} = 0 && \forall q \in H^1_0(\Omega). \quad (67)
\end{aligned}
$$

Here, $T_i \in H^1(\Omega)$ again denotes the solution of (30) with $u = u_i$ and $G_{\sigma_\rho} \colon \mathbb{R} \to \mathbb{R}$ the derivative of the function σ_ρ in (59). The key difference between (65)–(67) and (62)–(64) is that (65)–(67) does not contain an active set. In particular, (C2) does not require the semismoothness results that we derived in this paper for the map $S \circ \Phi$. However, we shall see that this regularization is unfavourable. In particular, the approximated QVI-solution is dependent on the Moreau–Yosida parameter ρ and is typically infeasible. Moreover, the convergence of the algorithm is slower than when computing δu_N via (62)–(64) directly, and the convergence rate degrades in the limit $\rho \to 0$ due to ill-conditioning; see Figure 3(d) in Section 4.4.

Finite element discretization: To discretize the variational problems, we consider uniform subdivisions of the domain Ω into intervals and quadrilateral cells, respectively, depending on whether the dimension is one or two. In one dimension, we discretize with a continuous piecewise linear finite element $\mathcal{P}_1$ and in two dimensions we choose the tensor-product of the one-dimensional basis $\mathcal{P}_1 \times \mathcal{P}_1$ and define the mesh size h as the length of the edge of a cell. This discretization applies to u, T, and the auxiliary functions ξ, η, μ, and w.

Discretization of the active set: The choice of a $\mathcal{P}_1$-, respectively, $\mathcal{P}_1 \times \mathcal{P}_1$-discretization means that the active sets in the PDAS-algorithm for the obstacle problem as well as the sets A_i and I_i in (64) may be found by examining the coefficient vectors of the involved finite element functions with respect to the nodal basis. In particular, enforcing (64) for all $q \in H_0^1(I_i)$ simplifies to deleting the rows and columns in the corresponding finite element matrices and the rows in the right-hand side vector that are associated with nodal basis functions that belong to active nodes. More explicitly, let $\mathbf{u}_B$, $\boldsymbol{\xi}$, $\boldsymbol{\mu}$, $\boldsymbol{\delta}\mathbf{u}_N$, $\mathbf{r}$, and $\bar{\boldsymbol{\Theta}} \in \mathbb{R}^M$ denote the coefficient vectors of the finite element functions $u_{B,h}$, ξ_h, μ_h, $\delta u_{N,h}$, $u_h - u_{B,h}$, and $\Phi_{0,h} + \varphi_h T_h$, respectively, that correspond to the quantities in the system (62)–(64). Let $\mathfrak{N}_h = \{1, 2, \dots, M\}$ be the index set of the set of the degrees of freedom and denote the discrete active set by $\mathfrak{A}_h = \{i \in \mathfrak{N}_h \mid \mathbf{u}_{B,i} = \bar{\boldsymbol{\Theta}}_i\}$. Denote further the discrete inactive set by $\mathfrak{I}_h = \mathfrak{N}_h \setminus \mathfrak{A}_h$ and suppose that the finite element linear system (before removing the active set) induced by (62)–(64) is given by

$$\begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & \mathbf{0} \\ \mathbf{0} & K_{32} & K_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}\mathbf{u}_N \\ \boldsymbol{\xi} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} -\mathbf{r} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \tag{68}$$

where $K_{ij} \in \mathbb{R}^{M \times M}$ for $i, j \in \{1, 2, 3\}$. Then modifying (68) such that it corresponds to a discretization of (62)–(64) with $\mu \in H_0^1(I_i)$ and test functions $q \in H_0^1(I_i)$ is equivalent to deleting the rows and columns in K_{33}, the columns in K_{13}, and the rows in K_{32} and $\boldsymbol{\mu}$ whenever the row or column index is an element of $\mathfrak{A}_h$. In other words, (68) gets reduced to

$$\begin{pmatrix} K_{11} & K_{12} & [K_{13}]_{\mathfrak{N}_h, \mathfrak{I}_h} \\ K_{21} & K_{22} & \mathbf{0}_{\mathfrak{N}_h, \mathfrak{I}_h} \\ \mathbf{0}_{\mathfrak{I}_h, \mathfrak{N}_h} & [K_{32}]_{\mathfrak{I}_h, \mathfrak{N}_h} & [K_{33}]_{\mathfrak{I}_h, \mathfrak{I}_h} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}\mathbf{u}_N \\ \boldsymbol{\xi} \\ \boldsymbol{\mu}_{\mathfrak{I}_h} \end{pmatrix} = \begin{pmatrix} -\mathbf{r} \\ \mathbf{0} \\ \mathbf{0}_{\mathfrak{I}_h} \end{pmatrix}, \qquad \boldsymbol{\mu}_{\mathfrak{A}_h} = \mathbf{0}.$$

Note that the realization of (64) for higher-order discretizations is a far more delicate topic. We leave the study of such higher-order finite elements for future research.

4.2 Test 1: a one-dimensional QVI with a known solution

We are now in position to present our numerical experiments. We begin with a simple one-dimensional instance of the QVI (58) which is covered by the global framework of Section 2.4 and possesses a unique solution that is known in closed form. We choose the quantities in the QVI (58) as follows:

$$\begin{gathered} \Omega = (0,1), \qquad \Phi_0(x) = \max(0, |x - 0.5| - 0.25), \\ f(x) = \pi^2 \sin(\pi x) + 100 \max(0, -|x - 0.625| + 0.125), \qquad \varphi(x) = \frac{1}{\alpha_1} \sin(\pi x), \\ k = 1, \qquad g(s) = \alpha_1 + \arctan\left(\frac{1}{\alpha_2} \min\left(\frac{1-s}{2}, 1 - s\right)\right), \qquad \Psi_0 \equiv 1, \qquad \psi \equiv \varphi. \end{gathered} \tag{69}$$

Here, $\alpha_1, \alpha_2 > 0$ are parameters that will be fixed later. Note that the conditions in Assumptions 3.1 and 3.8 are trivially satisfied in the situation of (69) (with $p = 2$). Using direct calculations and Remark 3.10, it is furthermore easy to establish the following lemma.

Lemma 4.2. *In the situation of* (69)*, the QVI* (58) *possesses the solution* $\bar{u}(x) = \sin(\pi x)$*,* $\bar{T} \equiv \alpha_1$*. For the solution* $(\bar{u}, \bar{T})$*, the inactive set, the strictly active set, and the biactive set are given by*

$$\begin{aligned} \{x \in \Omega \mid \bar{u}(x) < \Phi_0(x) + \Phi(\bar{u})(x)\} &= (0, 0.25) \cup (0.75, 1), \\ \{x \in \Omega \mid \bar{u}(x) = \Phi_0(x) + \Phi(\bar{u})(x), -\bar{u}''(x) - f(x) \neq 0\} &= (0.5, 0.75), \end{aligned}$$

and

$$\{x \in \Omega \mid \bar{u}(x) = \Phi_0(x) + \Phi(\bar{u})(x), -\bar{u}''(x) - f(x) = 0\} = [0.25, 0.5] \cup \{0.75\},$$

respectively. Here, the function evaluations are defined w.r.t. the continuous representatives of $\bar{u}$ Φ_0, $\Phi(\bar{u})$, *and* $-\bar{u}'' - f$. *The function* $\Psi_0 + \psi\bar{T} - \bar{u}$ *takes values only in the set of points of nondifferentiability of* g. *Further, the constant appearing in* (31) *and* (34) *satisfies*

$$C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)}k^{-1}\right) = \frac{1+\pi}{\pi\alpha_1\alpha_2}. \tag{70}$$

Recall that the solution operator of the obstacle problem is Gâteaux differentiable if and only if strict complementarity holds; see [55, Lemma 2.6]. In combination with the fact that the function $\Psi_0 + \psi\bar{T} - \bar{u}$ takes values only in the set of nondifferentiable points of g, this means that the solution $(\bar{u}, \bar{T})$ corresponds to a worst-case example. From (70), Corollary 3.11, and the global Lipschitz continuity of g, it follows further that the mapping $\Phi\colon H_0^1(\Omega) \to H_0^1(\Omega)$ appearing in (58) in the situation of (69) satisfies the conditions i) to v) in Theorem 3.7 whenever $\alpha_1, \alpha_2 > 0$ are chosen such that $1 + \pi^{-1} < \alpha_1\alpha_2$ holds (with $p = 2$, the Newton derivative G_Φ defined in Theorem 3.9iii), and $\gamma := (1+\pi)(\pi\alpha_1\alpha_2)^{-1}$). In particular, the QVI is uniquely solvable for $1 + \pi^{-1} < \alpha_1\alpha_2$, $(\bar{u}, \bar{T})$ is its unique solution, and the assertions of Theorem 2.4 hold when we apply Algorithm 2.

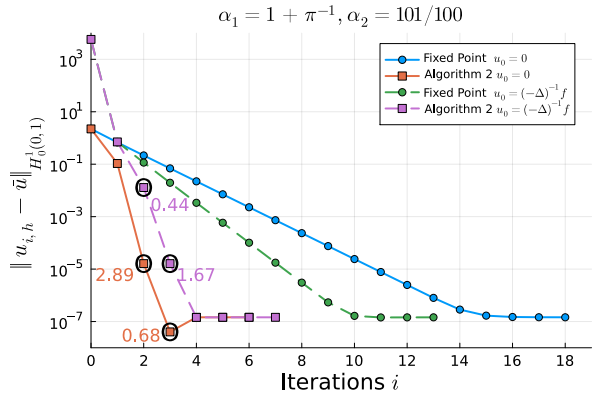
The results that are obtained when our globalized semismooth Newton method is applied to the QVI (58) in the situation of (69) (or, more precisely, to the operator equation $u = S(\Phi(u))$ that the QVI may be recast as) can be seen in Figure 1.

Here, we considered two different parameter choices: (I) $(\alpha_1, \alpha_2) = (1 + \pi^{-1}, 101/100)$ and (II) $(\alpha_1, \alpha_2) = (10^{-2}, 10^{-2})$. Note that, for choice (I), we trivially have $1 + \pi^{-1} < \alpha_1\alpha_2$, so that Theorem 2.4 is applicable and global q-superlinear convergence to the problem solution $\bar{u}$ is guaranteed. For (II), we have $1 + \pi^{-1} \gg \alpha_1\alpha_2 = 10^{-4}$. We use this second configuration to investigate how our algorithm behaves in situations that are beyond the scope of our analysis. In addition, we also considered two different choices for the initial guess u_0, namely, $u_0 = 0$ and $u_0 = (-\Delta)^{-1}f$. Figure 1(a) shows the behavior of Algorithm 2 and the pure fixed-point method (C1) for the two different starting values in configuration (I). It can be seen that both algorithms are able to identify the solution $\bar{u}$ of the QVI up to the discretization error for both choices of u_0. The semismooth Newton method requires far fewer iterations, however, and exhibits q-superlinear convergence speed, as predicted by Theorem 2.4. This is also confirmed by the experimental orders of convergence (EOCs) given by the formula

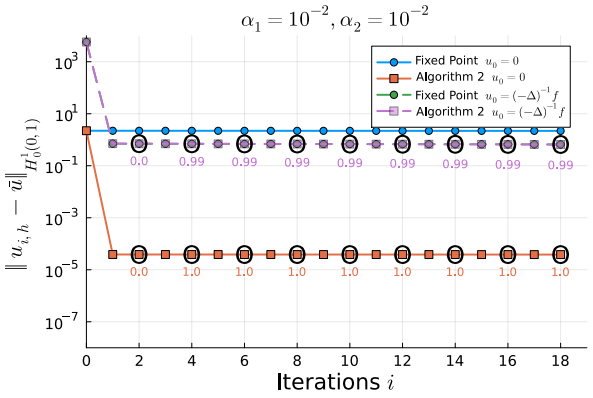
$$\mathrm{EOC}_i := \log\left(\frac{\|u_i - \bar{u}\|_{H_0^1(\Omega)}}{\|u_{i-1} - \bar{u}\|_{H_0^1(\Omega)}}\right)\log\left(\frac{\|u_{i-1} - \bar{u}\|_{H_0^1(\Omega)}}{\|u_{i-2} - \bar{u}\|_{H_0^1(\Omega)}}\right)^{-1}, \qquad i \geq 2. \tag{71}$$

We remark that, in this test case, Algorithm 2 always chooses the iterate u_N in step 9.

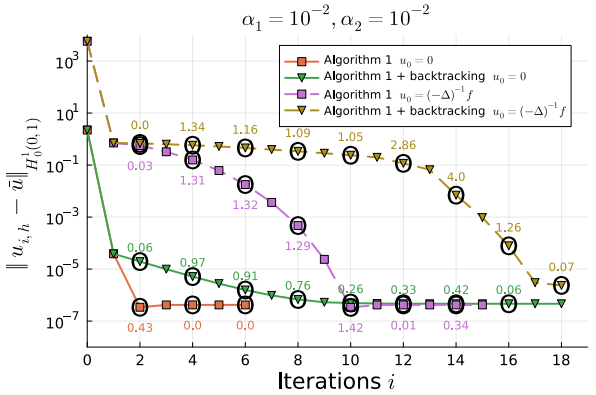
Figure 1(b) shows the convergence behavior of the two algorithms for the parameter choice (II). It can be seen here that both Algorithm 2 and the fixed-point method stagnate/converge very slowly. For this configuration, Algorithm 2 always chooses x_B in step 9 for $i \geq 2$, and the lack of contractivity of the composition $S \circ \Phi$ results in a very poor convergence performance. Figure 1(c) shows what happens in (II) with Algorithm 1 with and without a backtracking linesearch (C3). As can be seen, both methods are able to identify $\bar{u}$ with superlinear convergence speed even in the absence of contractivity. Notably, Algorithm 1 without a linesearch achieves convergence with the fewest number of iterations and with a significantly smaller computational cost since the backtracking linesearch requires many evaluations of S. This highlights that it is worth to consider our semismooth Newton approach for the solution of QVIs even in those situations where a rigorous convergence analysis is not possible.



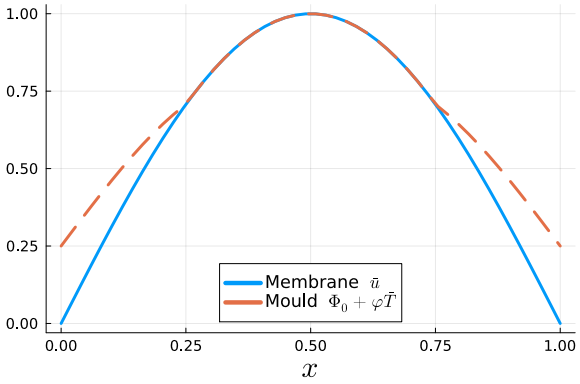
(a) Algorithm 2 and (C1) in test configuration (I)



(b) Algorithm 2 and (C1) in test configuration (II)



(c) Vanilla & backtracking SSN iterations in test case (II)



(d) The solution $\bar{u}$ and its mould $\Phi_0 + \varphi\bar{T}$.

Figure 1: (Test 1) Convergence behavior and results of the fixed-point method (C1) and Algorithm 2 for the parameter choices (I) and (II) as well as Algorithm 1 and (C3) (Algorithm 1 with a backtracking linesearch), for the parameter choice (II) with the initial guesses $u_0 = 0$ and $u_0 = (-\Delta)^{-1} f$ in the situation of (69). The mesh size was chosen as $h = 5 \times 10^{-4}$. The numbers next to the graphs are the EOCs in (71).

4.3 Test 2: a one-dimensional QVI with two known solutions

Next, we consider the QVI (58) in a situation in which there are several solutions and in which the localized setting of Section 2.3 has to be employed. We choose the data in (58) as follows:

$$\begin{aligned} &\Omega = (0,1), \qquad \Phi_0 \equiv 0, \qquad f(x) = \alpha_1 \pi^2 \sin(\pi x), \qquad \varphi(x) = \alpha_2 \frac{10\pi^2 \sin(\pi x)}{5 - \cos(2\pi x)}, \\ &k = \pi^2, \qquad g(s) = \frac{4}{\alpha_1} \min(0, s)^2, \qquad \Psi_0 \equiv 0, \qquad \psi(x) = \frac{5\pi^2 \sin(\pi x)}{5 - \cos(2\pi x)}. \end{aligned} \tag{72}$$

Here, $\alpha_1 > 0$ and $\alpha_2 \geq 1$ are again parameters that can be chosen arbitrarily. Note that the conditions in Assumptions 3.1 and 3.8 are all satisfied in the situation of (72) (with $p = 2$) and that, in contrast to the example in Section 4.2, the function g in (72) is only locally Lipschitz continuous (but C^1). Similarly to Lemma 4.2, we obtain:

Lemma 4.3. *In the situation of* (72)*, the QVI* (58) *possesses (at least) the two solutions*

$$\bar{u}_1 \equiv 0, \qquad \bar{T}_1 \equiv 0$$

and

$$\bar{u}_2(x) := \alpha_1 \sin(\pi x), \qquad \bar{T}_2(x) := \alpha_1 \frac{5 - \cos(2\pi x)}{10\pi^2}.$$

If M_R is defined as in Theorem 3.12, then it holds

$$M_R = \frac{1}{2}R + \frac{10(1+\pi)}{3\pi\alpha_1}R^2 \tag{73}$$

and

$$\begin{aligned} &\operatorname{Lip}(g, [-M_R, M_R]) \left(\|\varphi\|_{L^\infty(\Omega)} k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)} k^{-1} \right) \frac{|\Omega|}{\pi} \\ &= \lim_{t \searrow M_R} \operatorname{Lip}(g, [-t, t]) \left(\|\varphi\|_{L^\infty(\Omega)} k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)} k^{-1} \right) \frac{|\Omega|}{\pi} \\ &= \frac{50\alpha_2}{3\alpha_1} \left(R + \frac{20(1+\pi)}{3\pi\alpha_1} R^2 \right). \end{aligned} \tag{74}$$

Furthermore, the right-hand side of (74) *is an element of the interval $[0, 1)$ if and only if*

$$R < R_{\alpha_1,\alpha_2} := \frac{3\alpha_1}{10(1+\pi)} \left(\sqrt{\frac{(5\alpha_2 + 8)\pi^2 + 8\pi}{80\alpha_2}} - \frac{\pi}{4} \right). \tag{75}$$

Proof. To see that $(\bar{u}_1, \bar{T}_1)$ solves (58) in the situation of (72), it suffices to plug everything in (72) and the formulas $\bar{u}_1 \equiv 0$, $\bar{T}_1 \equiv 0$ into (58). That $(\bar{u}_2, \bar{T}_2)$ is another solution is proved analogously. Note that, to see that the differential equation in (58) is satisfied for $(\bar{u}_2, \bar{T}_2)$, one has to employ the double-angle-formula for the cosine, i.e.,

$$\begin{aligned} k\bar{T}_2 - \bar{T}_2'' &= \pi^2 \left(\alpha_1 \frac{5 - \cos(2\pi x)}{10\pi^2} \right) - \frac{4\pi^2 \alpha_1 \cos(2\pi x)}{10\pi^2} \\ &= \frac{\alpha_1}{2}(1 - \cos(2\pi x)) = \alpha_1 \sin(\pi x)^2 \\ &= g\left(-\frac{\alpha_1}{2} \sin(\pi x) \right) = g(\Psi_0 + \psi \bar{T}_2 - \bar{u}_2). \end{aligned}$$

To check that M_R is given by (73) in the situation of (72), it suffices to note that $|\Omega|^{1/2} = 1$ holds for $\Omega = (0, 1)$, to calculate that $\|\psi\|_{L^\infty(\Omega)} = 5\pi^2/6$, to note that the function g in (72) satisfies $\operatorname{Lip}(g, [-t, t]) = 8t/\alpha_1$ for all $t > 0$, and to plug into (44). To obtain (74), one proceeds analogously, using that $\|\varphi\|_{L^\infty(\Omega)} = 5\alpha_2\pi^2/3$ and $\|\varphi'\|_{L^\infty(\Omega)} = 5\alpha_2\pi^3/2$. The estimate (75) finally follows from an application of the pq-formula. □

That the QVI (58) possesses two solutions in the situation of (72) shows that this test case is beyond the scope of the analysis of Section 2.2 and that problems of the type (58) may indeed possess several solutions if the conditions in (17) and (18) are violated. To see that we may apply the results of Section 2.3, we first note that the spaces $X = H_0^1(\Omega)$, $Y_2 = \{v \in H_0^1(\Omega) \mid v'' \in L^2(\Omega)\} = H_0^1(\Omega) \cap H^2(\Omega)$, the set $D = Y_2$, the solution operator $S\colon H_0^1(\Omega) \to H_0^1(\Omega)$ of (21), and the map $\Phi\colon H_0^1(\Omega) \to Y_2$ in (58) satisfy the conditions in points i) and ii) of Corollary 2.13 in the situation of (58) by Lemmas 3.5 and 3.6 and Theorem 3.12. It remains to find a set B with the properties in points iii) and iv) of Corollary 2.13. To this end, let us suppose that B is a closed ball in $H_0^1(\Omega)$ (not necessarily centered at zero) that is contained in $\{v \in H_0^1(\Omega) \mid \|v\|_{H_0^1(\Omega)} < R_{\alpha_1,\alpha_2}\}$ with R_{α_1,α_2} defined by the formula on the right-hand side of (75). From Lemma 2.9, it follows that the projection $P_B\colon H_0^1(\Omega) \to B$ in $H_0^1(\Omega)$ onto B satisfies all of the conditions in point iv) of Corollary 2.13. From

Theorem 3.12, the properties of R_{α_1,α_2}, and Lemmas 3.2 and 3.6, we further obtain that there exists $\gamma_B \in [0,1)$ such that

$$\|S(\Phi(u_1)) - S(\Phi(u_2))\|_{H_0^1(\Omega)} \leq \|\Phi(u_1) - \Phi(u_2)\|_{H_0^1(\Omega)} \leq \gamma_B \|u_1 - u_2\|_{H_0^1(\Omega)} \quad \forall u_1, u_2 \in B \tag{76}$$

and

$$\sup_{u\in B} \|G_S(\Phi(u))G_\Phi(u)\|_{\mathcal{L}(H_0^1(\Omega),H_0^1(\Omega))} \leq \sup_{u\in B} \|G_\Phi(u)\|_{\mathcal{L}(H_0^1(\Omega),H_0^1(\Omega))} \leq \gamma_B.$$

Putting everything together, it follows that the conditions in Corollary 2.13 are all satisfied for B in the situation of (72). In particular, Theorem 2.8 is applicable and we may employ Algorithm 2 to determine the intersection of the solution set $\{u \in H_0^1(\Omega) \mid u = S(\Phi(u))\}$ of (58) with B by solving the localized fixed-point equation

$$\text{Find } \hat{x} \in X \text{ such that } \hat{x} = S(\Phi_B(\hat{x})), \tag{77}$$

where $\Phi_B := \Phi \circ P_B$. (This is just (F_{loc}) for $H = S \circ \Phi$.) Observe that since we already know that $\bar{u}_1 \equiv 0$ solves the QVI (58), we obtain from Theorem 2.8 that our algorithm should converge q-superlinearly to $\bar{u}_1$ if $0 \in B$ and to a point $\hat{u} \notin B$ if $0 \notin B$ when applied to (77).

To put these predictions to the test, we have implemented Algorithm 2 for the solution of (77) in the situation of (72) for the case that B is a closed ball $B_R(0)$ of radius $R > 0$ in $H_0^1(\Omega)$ centered at the origin. As it turns out, for this configuration, there are two distinct regimes: (I) $\alpha_2 = 1$ and (II) $\alpha_2 > 1$. In regime (I), the biactive set of the second solution $(\bar{u}_2, \bar{T}_2)$ is the whole domain, i.e., it holds $\bar{u}_2 = (-\Delta)^{-1} f = \Phi_0 + \varphi \bar{T}_2$ in Ω. For this case, the solution $(\bar{u}_2, \bar{T}_2)$ is highly unstable with respect to perturbations of the data. Furthermore, the QVIs obtained from the discretization of (58) apparently do not possess any discrete solutions that approximate $(\bar{u}_2, \bar{T}_2)$. The outcome is that, after a discretization, the second closed-form solution $(\bar{u}_2, \bar{T}_2)$ cannot be discovered by any method considered in this work, even if the interpolant $I_h \bar{u}_2$ of $\bar{u}_2$ is used as the initial guess and the mesh width h is very small; see Figure 2.

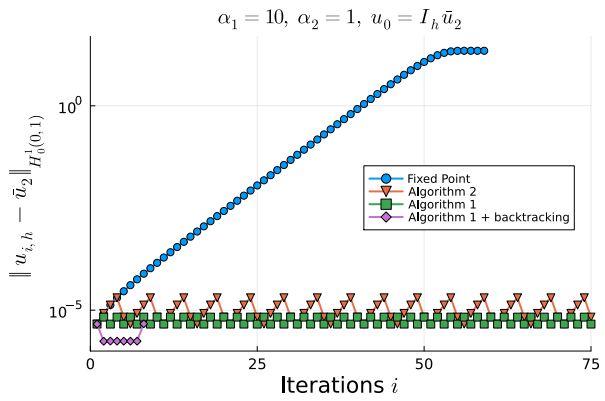


Figure 2: (Test 2) Divergence behavior of the fixed-point method (C1), Algorithm 2, and Algorithm 1 with and without a backtracking linesearch (C3) in the situation of the QVI (58) with setup (72) and $(\alpha_1, \alpha_2) = (10, 1)$. No projection is used and the initial guess is (the Lagrange interpolant of) $\bar{u}_2$. The mesh size is $h = 5 \times 10^{-4}$. It can be checked that the pure fixed-point method actually converges to $\bar{u}_1$ although it is initialized as close to $\bar{u}_2$ as possible. Algorithm 2 and Algorithm 1 both with and without a backtracking linesearch stagnate/enter a cycle and are unable to reduce the residue below 10^{-6} in this test case—a bound much larger than observed in the other experiments; cf. Figure 3. This indicates that, on the discrete level, there is no solution corresponding to $\bar{u}_2$.

Note that this observation is also of independent interest as it shows that, for $\alpha_2 = 1$ and the data in (72), the QVI (58) is ill posed; cf. [7, 61, 9, 10, 6, 8]. In the second regime (II), the solution $(\bar{u}_2, \bar{T}_2)$ satisfies $(-\Delta)^{-1} f = \bar{u}_2 < \Phi_0 + \varphi \bar{T}_2$ in $(0,1)$ and the active set of $\bar{u}_2$ is empty. Here, the numerical identification of $\bar{u}_2$ works without problems; see below.

For our numerical experiments in regime (II), we considered the parameters $(\alpha_1, \alpha_2) = (10, 3/2)$ and the initial guess $u_0(x) = 100(1-x)x$. With this choice of α_1 and α_2, we have $R_{\alpha_1,\alpha_2} \approx 0.3136$ and $\|\bar{u}_2\|_{H_0^1(0,1)} = \sqrt{50}\pi \approx 22.21$. In particular, the q-superlinear convergence of Algorithm 2 to $\bar{u}_1$ is guaranteed by our analysis for all $R < 0.3136$. Note that, due to (76), we also immediately obtain that the fixed-point method (C1) converges to $\bar{u}_1$ for all $R < 0.3136$ when applied to (77).

When running Algorithm 2 and the fixed-point method (C1) in this configuration, one observes that both algorithms converge in one iteration to $(\bar{u}_2, \bar{T}_2)$ for all $R \geq \sqrt{50}\pi$. This effect is caused by the inactivity of $\bar{u}_2$ in regime (II). The behavior for $1/100 \leq R < \sqrt{50}\pi$ is tabulated in Table 1.

R	0.01	2.477	4.944	7.411	9.879	12.346	14.813	17.28
(C1)	4 (1.91)	7 (2.01)	9 (2.0)	12 (2.01)	-	-	-	-
Algorithm 2	2 (1.56)	6 (2.01)	8 (2.0)	11 (2.01)	-	-	-	-
Algorithm 1	2 (1.56)	3 (6.91)	3 (9.16)	3 (11.29)	3 (13.59)	3 (16.42)	3 (20.26)	-
(C3)	2 (1.56)	3 (6.91)	3 (9.16)	3 (11.29)	3 (13.59)	3 (16.42)	3 (20.26)	-

Table 1: (Test 2) Number of iterations needed by the pure fixed-point method (C1), Algorithms 1 and 2, and (C3) (Algorithm 1 with a backtracking linesearch) to converge to $\bar{u}_1$ for the QVI-problem (58) with setup (72) and $(\alpha_1, \alpha_2) = (10, 3/2)$ in dependence of the projection radius R. The bracketed numbers are the largest EOC-values measured for the respective regime. We chose a mesh size of $h = 5 \times 10^{-4}$ and terminated the algorithm when $\|u_i\|_{H_0^1(0,1)} \leq 10^{-13}$. A dash implies that the algorithm stagnated without convergence. All of the considered algorithms failed to converge for $17 \lesssim R < \sqrt{50}\pi$.

As can be seen, the number of iterations that Algorithm 2 and the fixed-point method (C1) need to approximate $\bar{u}_1$ in $H_0^1(\Omega)$ up to the tolerance 10^{-13} grows as the projection radius increases. Interestingly, both methods still converge to the zero solution for $R \lesssim 9$ despite R being greater than the critical radius R_{α_1,α_2}. For larger R, both methods stagnate without convergence. This behavior is caused by the loss of contractivity on $B_R(0)$ for large projection radii R; cf. the behavior in Section 4.2.

As can be seen in Table 1, Algorithm 1 with and without a backtracking linesearch converges to $\bar{u}_1$ for all $R \lesssim 15$. Similarly to Section 4.2, they both handle the lack of contractivity for large R far better than Algorithm 2 and the fixed-point method (C1) and both converge in a maximum of three iterations when $R \lesssim 15$. To allow for a comparison of convergence speeds, Table 1 also shows the largest EOC-values that were measured for the four algorithms during each run (defined as in (71)). As can be seen, Algorithm 2 and Algorithm 1 with and without a backtracking linesearch exhibit superlinear convergence speed. However, in contrast to the example in Section 4.2, this time we also observe q-superlinear convergence for the fixed-point method (C1). The reason for this effect is that the Lipschitz constant $\mathrm{Lip}(g, [-t,t]) = 8t/\alpha_1$ of the nonlinearity in (72) goes to zero for $0 < t \to 0$. (Note that,

for Algorithm 1 and (C3), the number of iterations is so small that the EOC-values in Table 1 should be taken with a grain of salt.)

4.4 Test 3: the thermoforming QVI in two dimensions

As a third test case, we consider the QVI (58) in the following situation:

$$\begin{aligned} \Omega = (0,1)^2, \qquad \Phi_0(x_1,x_2) = 1 - 2\max(|x_1 - 0.5|, |x_2 - 0.5|), \qquad f \equiv 25, \\ \varphi(x_1,x_2) = \sin(\pi x_1)\sin(\pi x_2), \qquad k = 1, \qquad \Psi_0 \equiv \Phi_0, \qquad \psi \equiv \varphi, \\ g(s) = \begin{cases} 1/5 & \text{if } s \le 0, \\ (1-s)/5 & \text{if } 0 < s < 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{78}$$

Note that the above setting corresponds to an example of the thermoforming problem described in Remark 4.1. From the application point of view, it models the situation that a hot thin square plastic sheet is pushed by a constant pressure into a pyramidal metal mould. For the data in (78), one can check that the conditions in Assumptions 3.1 and 3.8 are satisfied with $p = 2$. From a straightforward calculation and Remark 3.10, we further obtain that

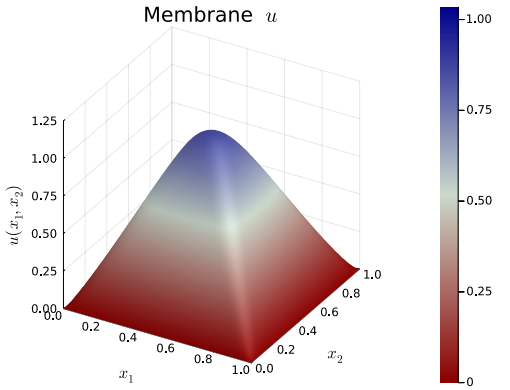
$$C_P(\Omega)\,\mathrm{Lip}(g)\left(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \||\nabla\varphi|\|_{L^\infty(\Omega)}k^{-1}\right) \le \frac{1+\pi}{5} \approx 0.8283 < 1.$$

In combination with Corollary 3.11 and Theorems 3.7 and 3.9, this shows that the setting (78) is covered by the analysis of Section 2.4 and that our global convergence result for Algorithm 2 applies. In particular, (58) possesses a unique solution $(\bar{u}, \bar{T})$ in the case of (78) by Lemma 2.3.

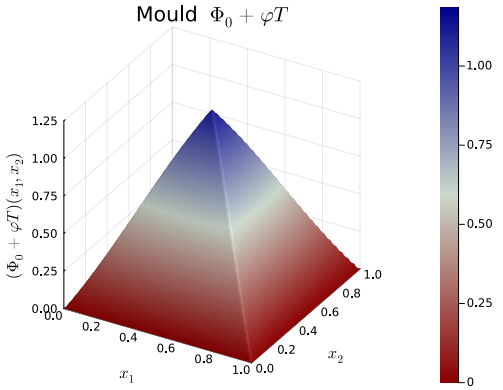
The results that we have obtained in the situation (78) for the QVI (58) can be seen in Figure 3 and Table 2. Figure 3 shows surface plots of the membrane $\bar{u}$ and the mould $\Phi_0 + \varphi\bar{T}$ as well as a slice of the membrane, mould, temperature $\bar{T}$, and mould deformation $\varphi\bar{T}$ at $x_2 = 1/2$. It also depicts the convergence behavior of Algorithm 2, the fixed-point method (C1), and the Moreau–Yosida-based regularization method (C2) for various choices of ρ. We see that Algorithm 2 converges the fastest and that the fixed-point method (C1) is the slowest, exhibiting linear convergence speed (as expected). For the regularization method (C2), the convergence degrades for $\rho \to 0$. Recall that ρ must be driven to zero in (C2) in order to get an approximate solution close to the true QVI-solution. This, however, causes ill-conditioning effects that reduce the convergence speed.

Table 2 shows the number of outer semismooth Newton steps as well as the overall number of inner semismooth Newton iterations, PFMY-iterations, and PDAS-feasibility restoration steps that Algorithm 2 requires in the situation of (78) for various mesh widths h to drive the residue $\|R(u_i)\|_{H^1(\Omega)}$ below the tolerance 10^{-12}. As can be seen, the number of semismooth Newton steps that Algorithm 2 needs is four or fewer for all considered mesh widths h. This shows that—on the level of the solver for the fixed-point equation (F_c)—we observe *mesh-independence* for our semismooth Newton method. This very desirable property is characteristic for solution algorithms whose convergence can be established not only for the discrete problems obtained from a discretization but also on the function space level; see Section 2.4. Note that, for the inner semismooth-Newton solver used for the evaluation of Φ (i.e., the solution of the nonsmooth semilinear PDE in (58)), we also observe a mesh-independent convergence behavior. For the hybrid PFMY-PDAS-algorithm used for the evaluation of the solution map S of the obstacle problem, a mild form of mesh-dependence is present that, however, is manageable. Note that the mesh-dependence of solution algorithms for the obstacle problem is a known problem. Nevertheless other experimentally mesh-independent solution methods, for the evaluation of

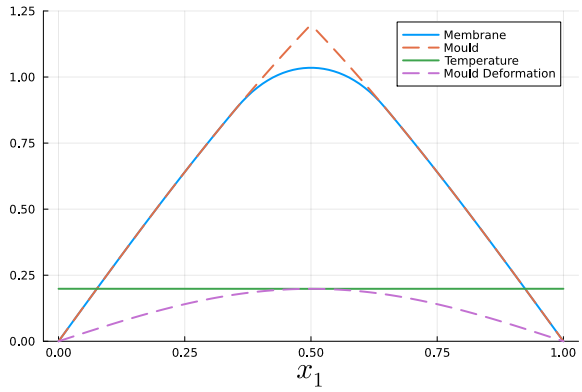
S, exist than just the one considered in this paper [40, 34]. The fact that Algorithms 1 and 2 reduce the number of outer iterations and, thus, required evaluations of S to a minimum is a main advantage of our semismooth Newton approach in comparison with fixed-point-based methods that have to evaluate S far more often due to their slow convergence speed; cf. Figure 3(d).



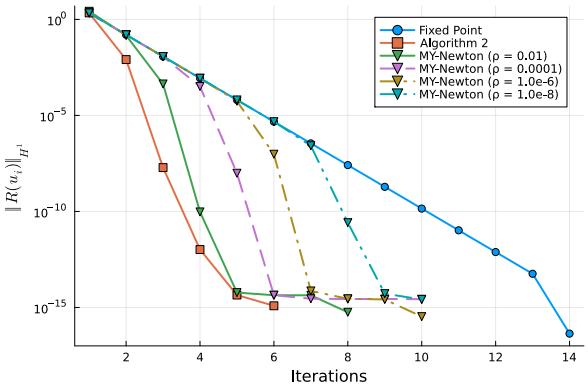
(a) The membrane $\bar{u}$



(b) The mould $\Phi_0 + \varphi\bar{T}$



(c) Slice at $x_2 = 1/2$



(d) Convergence behavior

Figure 3: (Test 3) Surface plots of the membrane (a) and corresponding mould (b) together with a slice plot at $x_2 = 1/2$ (c) for the thermoforming setting (78). Figure (d) depicts the outer loop convergence of Algorithm 2, the fixed-point method (C1), and the regularization method (C2) for $h = 0.02$ and various ρ. The residue depicted for (C2) is that of the smoothed system. It can be seen that Algorithm 2 converges the fastest and that the convergence speed of the Moreau–Yosida-based method (C2) degrades for $\rho \to 0$ due to ill-conditioning effects.

4.5 Test 4: a nonlinear VI

Finally, we briefly consider the nonlinear VI-example (53) of Section 3.3. We omit conducting detailed numerical experiments along the lines of Sections 4.1 to 4.4 here to avoid overloading the paper and simply present a short feasibility study for a particular instance of (53); see Table 3 and Figure 4. It can be observed that Algorithm 2 behaves similar for the semilinear VI (53) as for the QVI-examples in Sections 4.2 to 4.4.

We remark that, analogously to the approach presented in Section 3.3, one can also establish that the analysis of Section 2.4 covers nonsmooth semilinear and quasilinear partial differential equations (for

	Outer loop	Inner solver to evaluate Φ	Inner VI-solver to evaluate S	
h	Semismooth Newton	Semismooth Newton	PFMY	PDAS
0.04	4 (4)	13 (9)	293 (159)	17 (10)
0.02	4 (4)	13 (9)	340 (185)	31 (17)
0.01	3 (3)	12 (8)	277 (150)	20 (11)
0.00667	3 (3)	12 (8)	283 (158)	17 (11)
0.005	3 (3)	12 (8)	285 (158)	29 (17)
0.004	3 (4)	11 (8)	289 (199)	29 (21)
0.00333	3 (4)	10 (7)	262 (184)	29 (21)

Table 2: (Test 3) Number of iterations of the outer loop and cumulative number of iterations for the inner loops when Algorithm 2 and Algorithm 1 (in brackets) are applied to (58) in the situation of (78) for various mesh widths h. The algorithm is terminated once $\|R(u_i)\|_{H^1(\Omega)} \leq 10^{-12}$. We observe that the number of outer loop iterations and the number of inner semismooth Newton steps needed for the evaluation of Φ are mesh-independent and that the number of PFMY- and PDAS-iterations does not grow in an uncontrollable manner.

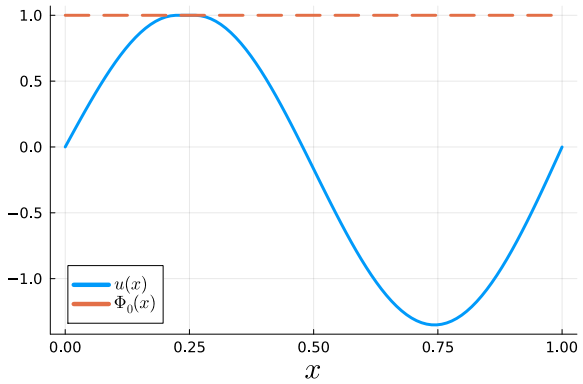


Figure 4: (Test 4) Solution of the semilinear VI (48) in the situation of Table 3. The mesh size is $h = 1.56 \times 10^{-4}$.

which, as mentioned, contraction assumptions of the type used in Section 2.4 are standard tools in existence proofs). A further interesting application area for the framework of Section 2.4 is the study of fixed-point problems (F_c) in which both S and Φ arise from a variational inequality. We leave both these topics for future research.

Appendix A Calculus rules for Newton derivatives

The following calculus rules are well known and can be found in various variants in the literature; see, e.g., [59, §3.3.7] for a version of the chain rule in Lebesgue spaces.

Lemma A.1 (Chain rule for Newton derivatives)**.** *Let $(U, \|\cdot\|_U)$, $(V, \|\cdot\|_V)$, and $(W, \|\cdot\|_W)$ be real normed spaces and let $E \subset V$ be nonempty. Suppose that $\Psi\colon U \to E$ and $T\colon E \to W$ are Newton differentiable functions with Newton derivatives $G_\Psi\colon U \to \mathcal{L}(U, V)$ and $G_T\colon E \to \mathcal{L}(V, W)$, respectively. Suppose that, for every $u \in U$, there exist constants $C, \varepsilon > 0$ such that*

$$\|\Psi(u+h) - \Psi(u)\|_V \leq C\|h\|_U \quad \forall h \in B_\varepsilon^U(0), \tag{79}$$

	Outer loop	Inner VI-solver to evaluate S	
h	Semismooth Newton	PFMY	PDAS
0.02	4 (5)	205 (136)	10 (7)
0.01	4 (4)	248 (142)	17 (9)
0.005	3 (4)	210 (152)	7 (5)
0.0025	4 (4)	264 (149)	9 (5)
0.00125	4 (4)	296 (166)	17 (9)
0.00062	4 (4)	314 (176)	9 (5)
0.00031	3 (4)	262 (188)	7 (5)
0.00016	3 (4)	269 (194)	7 (5)

Table 3: (Test 4) Number of iterations of the outer loop and cumulative numbers of iterations for the inner loops when Algorithm 2 and Algorithm 1 (in brackets) are applied to (48) for various mesh widths h. The data was chosen as $\Omega = (0,1)$, $f(x) = 50\sin(2\pi x)$, $b_1(s) = \max(0,s)$, $b_2(s) = s + \cos s$, and $\Phi_0 \equiv 1$. The initial guess was $u_0(x) = 0$. The algorithm was terminated once $\|R(u_i)\|_{H^1(\Omega)} \leq 10^{-10}$. It can be seen that the algorithms converge and that the number of outer loop semismooth Newton steps and inner PFMY-/PDAS-iterations does not grow in an uncontrollable manner; similarly to the results in Table 2.

and that, for every $v \in E$, there exist constants $C, \varepsilon > 0$ such that

$$\sup_{w \in E \cap B_\varepsilon^V(v)} \|G_T(w)\|_{\mathcal{L}(V,W)} \leq C. \tag{80}$$

Define $K\colon U \to W$ by $K := T \circ \Psi$. Then the function K is Newton differentiable with Newton derivative $G_K\colon U \to \mathcal{L}(U,W)$, $G_K(u) := G_T(\Psi(u))G_\Psi(u)$.

Proof. If $u \in U$ is fixed and $\{h_n\} \subset U$ is an arbitrary sequence satisfying $\Psi(u+h_n) - \Psi(u) \neq 0$ for all $n \in \mathbb{N}$ and $0 < \|h_n\|_U \to 0$ for $n \to \infty$, then, for all large enough n, we have

$$\begin{aligned}
0 &\leq \frac{\|K(u+h_n) - K(u) - G_K(u+h_n)h_n\|_W}{\|h_n\|_U} \\
&= \frac{\|T(\Psi(u+h_n)) - T(\Psi(u)) - G_T(\Psi(u+h_n))G_\Psi(u+h_n)h_n\|_W}{\|h_n\|_U} \\
&\leq \frac{\|T(\Psi(u+h_n)) - T(\Psi(u)) - G_T(\Psi(u+h_n))(\Psi(u+h_n) - \Psi(u))\|_W}{\|\Psi(u+h_n) - \Psi(u)\|_V} \\
&\quad \times \frac{\|\Psi(u+h_n) - \Psi(u)\|_V}{\|h_n\|_U} \\
&\quad + \|G_T(\Psi(u+h_n))\|_{\mathcal{L}(V,W)} \frac{\|\Psi(u+h_n) - \Psi(u) - G_\Psi(u+h_n)h_n\|_V}{\|h_n\|_U} \\
&\leq C \frac{\|T(\Psi(u+h_n)) - T(\Psi(u)) - G_T(\Psi(u+h_n))(\Psi(u+h_n) - \Psi(u))\|_W}{\|\Psi(u+h_n) - \Psi(u)\|_V} \\
&\quad + C \frac{\|\Psi(u+h_n) - \Psi(u) - G_\Psi(u+h_n)h_n\|_V}{\|h_n\|_U}.
\end{aligned} \tag{81}$$

Here, we have used (79) and (80) in the last step. Due to the Newton differentiability of T and Ψ, the right-hand side of (81) goes to zero for $n \to \infty$. By adjusting the estimates in (81) slightly, we also obtain this convergence for sequences satisfying $\Psi(u+h_n) - \Psi(u) = 0$ for some/all n. In combination with the arbitrariness of $\{h_n\}$ and $u \in U$, this proves the lemma. □

Lemma A.2 (Product rule for Newton derivatives)**.** *Let $(U, \|\cdot\|_U)$, $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$, and $(Z, \|\cdot\|_Z)$ be normed spaces and let $a\colon V \times W \to Z$ be a bilinear and continuous mapping, i.e., it holds $\|a(v,w)\|_Z \leq C_a \|v\|_V \|w\|_W$ for all $(v,w) \in V \times W$ with a constant $C_a > 0$. Assume that $P\colon U \to V$ and $Q\colon U \to W$ are Newton differentiable with Newton derivatives $G_P\colon U \to \mathcal{L}(U,V)$ and $G_Q\colon U \to \mathcal{L}(U,W)$, respectively. Suppose that P and Q are continuous with one of them being locally Lipschitz. Then the function $K\colon U \to Z$, $K(u) := a(P(u), Q(u))$, is Newton differentiable with derivative*

$$G_K\colon U \to \mathcal{L}(U,Z), \qquad G_K(u)h := a(P(u), G_Q(u)h) + a(G_P(u)h, Q(u)) \quad \forall u, h \in U.$$

Proof. For all $u \in U$ and $h \in U \setminus \{0\}$, we have

$$\begin{aligned}
&K(u+h) - K(u) - G_K(u+h)h \\
&\quad = a(P(u+h), Q(u+h)) - a(P(u), Q(u)) \\
&\qquad - a(P(u+h), G_Q(u+h)h) - a(G_P(u+h)h, Q(u+h)) \\
&\quad = a(P(u+h) - P(u) - G_P(u+h)h, Q(u+h)) \\
&\qquad + a(P(u+h), Q(u+h) - Q(u) - G_Q(u+h)h) \\
&\qquad + a(P(u) - P(u+h), Q(u+h) - Q(u)),
\end{aligned}$$

and, thus,

$$\begin{aligned}
&\frac{\|K(u+h) - K(u) - G_K(u+h)h\|_Z}{\|h\|_U} \\
&\qquad \leq C_a \|Q(u+h)\|_W \frac{\|P(u+h) - P(u) - G_P(u+h)h\|_V}{\|h\|_U} \\
&\qquad + C_a \|P(u+h)\|_V \frac{\|Q(u+h) - Q(u) - G_Q(u+h)h\|_W}{\|h\|_U} \\
&\qquad + C_a \frac{\|P(u+h) - P(u)\|_V \, \|Q(u+h) - Q(u)\|_W}{\|h\|_U}.
\end{aligned} \tag{82}$$

Due to the Newton differentiability and continuity properties of P and Q, the right-hand side of (82) tends to zero for $0 < \|h\|_U \to 0$. This proves the assertion; see (3). □

Appendix B Convergence of semismooth Newton methods

Proof of Theorem 2.1. The proof is standard; see, e.g., [45, Theorem 3] but we give it here for the sake of completeness. The openness of B and the Newton differentiability of R on B imply that, for every $\epsilon > 0$, there exists $r > 0$ such that $B_r^X(\bar{x}) \subset B$ and

$$x \in B_r^X(\bar{x}) \implies \|R(x) - R(\bar{x}) - G_R(x)(x - \bar{x})\|_X \leq \epsilon \|x - \bar{x}\|_X . \tag{83}$$

Choose $\epsilon > 0$ such that $\alpha := M\epsilon + ML\rho^* < 1$ holds and let $r > 0$ be such that (83) is satisfied for this ϵ. Let $x_i \in B_r^X(\bar{x})$ be arbitrary and let x_{i+1} and z_i be as in steps 7 and 8 of Algorithm 1. Then

$$\begin{aligned}
&\|x_{i+1} - \bar{x}\|_X \\
&\quad = \left\|x_i - \bar{x} - G_R(x_i)^{-1}R(x_i) + G_R(x_i)^{-1}(G_R(x_i)z_i + R(x_i))\right\|_X \\
&\quad = \left\|G_R(x_i)^{-1}G_R(x_i)(x_i - \bar{x}) - G_R(x_i)^{-1}R(x_i) + G_R(x_i)^{-1}(G_R(x_i)z_i + R(x_i))\right\|_X \\
&\quad \leq \left\|G_R(x_i)^{-1}\right\|_{\mathcal{L}(X,X)} \left(\|R(x_i) - R(\bar{x}) - G_R(x_i)(x_i - \bar{x})\|_X + \|G_R(x_i)z_i + R(x_i)\|_X\right) \\
&\quad \leq M\left(\epsilon\|x_i - \bar{x}\|_X + \rho_i\|R(x_i)\|_X\right) \\
&\quad \leq (M\epsilon + ML\rho_i)\|x_i - \bar{x}\|_X \\
&\quad \leq \alpha\|x_i - \bar{x}\|_X .
\end{aligned} \tag{84}$$

This shows that $x_{i+1} \in B_r^X(\bar{x})$ holds and, after a trivial induction, that the iterates produced by Algorithm 1 satisfy $\|x_i - \bar{x}\|_X \leq \alpha^i\|x_0 - \bar{x}\|_X$ for all $x_0 \in B_r^X(\bar{x})$ and all i. The assertions in i) and ii) follow immediately from this estimate and (4). The q-superlinear convergence in iii) is obtained by revisiting the estimates in (84) with the knowledge that $x_i \to \bar{x}$. □

Next, we wish to prove Theorem 2.4 concerning the convergence of Algorithm 2. We first require the following lemma on the properties of the residue function R of (F).

Lemma B.1 (Properties of R)**.** *Suppose that Assumption 2.2 holds. Then the function $R\colon X \to X$ satisfies the following:*

i) R is bijective and its inverse $R^{-1}\colon X \to X$ satisfies

$$\|R^{-1}(x_1) - R^{-1}(x_2)\|_X \leq \frac{1}{1-\gamma}\|x_1 - x_2\|_X \quad \forall x_1, x_2 \in X \tag{85}$$

and

$$\|x_1 - x_2\|_X \leq (1+\gamma)\|R^{-1}(x_1) - R^{-1}(x_2)\|_X \quad \forall x_1, x_2 \in X. \tag{86}$$

ii) R is Newton differentiable on X with Newton derivative $G_R(x) := \mathrm{Id} - G_H(x)$.

iii) For every $x \in X$, the inverse $G_R(x)^{-1}$ exists and it holds $\|G_R(x)^{-1}\|_{\mathcal{L}(X,X)} \leq (1-\gamma)^{-1}$.

Proof. To prove i), suppose that $y \in X$ is given. Then $y = R(x)$ is equivalent to the fixed-point equation $x = y + H(x)$, which possesses a unique solution $x := R^{-1}(y)$ by the Banach fixed-point theorem. Thus, R is bijective and $R^{-1}\colon X \to X$ exists. Consider now some $x_1, x_2 \in X$. Then it holds $x_j = R(R^{-1}(x_j)) = R^{-1}(x_j) - H(R^{-1}(x_j))$, $j = 1, 2$, by the definition of R. Thus, by the triangle inequality and (5),

$$\begin{aligned}
\|R^{-1}(x_1) - R^{-1}(x_2)\|_X &= \|x_1 - x_2 + H(R^{-1}(x_1)) - H(R^{-1}(x_2))\|_X \\
&\leq \|x_1 - x_2\|_X + \gamma\|R^{-1}(x_1) - R^{-1}(x_2)\|_X.
\end{aligned}$$

This establishes (85). The second estimate (86) follows from

$$\|x_1 - x_2\|_X = \|R(R^{-1}(x_1)) - R(R^{-1}(x_2))\|_X \leq (1+\gamma)\|R^{-1}(x_1) - R^{-1}(x_2)\|_X.$$

This proves i). The assertion of ii) follows from the sum rule for Newton differentiable functions and the Newton differentiability of H. To finally establish iii), it suffices to note that the $\mathcal{L}(X,X)$-norm of the operator $G_H(x)$ appearing in the definition of $G_R(x)$ is bounded by $\gamma \in [0,1)$ due to Assumption 2.2iii) and to use Neumann's series. This completes the proof. □

Using the properties in Lemma B.1, we can now prove Theorem 2.4.

Proof of Theorem 2.4. Let $x_i \in X$ be given and let $x_B, x_N \in X$ be chosen such that (7) and (8) hold. Then the γ-Lipschitz continuity of H in (5), (7), the triangle inequality, and the definition of R imply that

$$\begin{aligned}\min\left(\|R(x_B)\|_X, \|R(x_N)\|_X\right) &\leq \|R(x_B)\|_X \\ &= \|x_B - H(x_i) + H(x_i) - H(x_B)\|_X \\ &\leq \tau_i \|R(x_i)\|_X + \gamma \|x_i - H(x_i) + H(x_i) - x_B\|_X \\ &\leq \left(\tau_i + \gamma + \gamma\tau_i\right) \|R(x_i)\|_X \\ &= \beta_i \|R(x_i)\|_X\end{aligned}$$

holds, where $\beta_i := \tau_i + \gamma + \gamma\tau_i$. Due to the acceptance criterion in Algorithm 2, the above yields that, if Algorithm 2 does not terminate in the iterations $i = 0, 1, \ldots, n-1$, $n \in \mathbb{N}$, then x_n satisfies

$$\|R(x_n)\|_X \leq \theta_{n-1} \|R(x_0)\|_X \tag{87}$$

with $\theta_{n-1} := \beta_0 \beta_1 \cdots \beta_{n-1}$.

Let us now first consider the situation in i), i.e., the case $\texttt{tol} > 0$ and $\tau^* \leq (\lambda - \gamma)/(1+\gamma)$ for some $\lambda \in (\gamma, 1)$. Then we have

$$\beta_i = \tau_i + \gamma + \gamma\tau_i \leq \tau^* + \gamma + \gamma\tau^* \leq \frac{\lambda - \gamma}{1+\gamma} + \gamma + \gamma\frac{\lambda - \gamma}{1+\gamma} = \lambda \in (\gamma, 1) \quad \forall i \in \mathbb{N}_0,$$

and we obtain from (87) that $\|R(x_n)\|_X \leq \lambda^n \|R(x_0)\|_X$ holds when the termination criterion $\|R(x_i)\|_X \leq \texttt{tol}$ in step 4 is not triggered for $i = 0, 1, \ldots, n-1$, $n \in \mathbb{N}$. That Algorithm 2 has to terminate after the number of iterations in (9) follows immediately from this estimate. From (85), we further obtain that the iterate x^* that Algorithm 2 returns in this situation satisfies

$$\begin{aligned}\|x^* - \bar{x}\|_X &= \|R^{-1}(R(x^*)) - R^{-1}(R(\bar{x}))\|_X \\ &\leq \frac{1}{1-\gamma} \|R(x^*) - R(\bar{x})\|_X \\ &= \frac{1}{1-\gamma} \|R(x^*)\|_X \\ &\leq \frac{\texttt{tol}}{1-\gamma}.\end{aligned} \tag{88}$$

This completes the proof of i).

Let us now assume that $\texttt{tol} = 0$ holds and that Algorithm 2 does not terminate after finitely many iterations. Then it follows from the inequality $\|R(x_i)\|_X \leq \theta_{i-1} \|R(x_0)\|_X$ for all $i \in \mathbb{N}$, the definitions $\theta_i := \beta_0 \beta_1 \cdots \beta_i$ and $\beta_i := \tau_i + \gamma + \gamma\tau_i$, the convergence $\tau_i \to 0$, and $\gamma \in (0, 1)$, that β_i converges to γ, that θ_i goes to zero, and that $\|R(x_i)\|_X \to 0$ holds for $i \to \infty$. In combination with the estimate $\|x_i - \bar{x}\|_X \leq (1-\gamma)^{-1} \|R(x_i)\|_X$, that is obtained along the exact same lines as (88), this yields that $0 < \|x_i - \bar{x}\|_X \to 0$ for $i \to \infty$. From (8), the properties in Lemma B.1, the acceptance criterion

in Algorithm 2, and the convergence $\rho_i \to 0$, we further obtain that

$$\begin{aligned}
&\|R(x_{i+1})\|_X \\
&\le \|R(x_N)\|_X \\
&= \|R(x_N) - R(\bar{x})\|_X \\
&\le (1+\gamma)\|x_N - \bar{x}\|_X \\
&\le \frac{1+\gamma}{1-\gamma}\|G_R(x_i)(x_N - \bar{x})\|_X \\
&\le \frac{1+\gamma}{1-\gamma}\|R(\bar{x}) - R(x_i) + G_R(x_i)(x_i - \bar{x})\|_X + \frac{1+\gamma}{1-\gamma}\|R(x_i) + G_R(x_i)(x_N - x_i)\|_X \\
&\le \frac{1+\gamma}{1-\gamma}\frac{\|R(\bar{x}) - R(x_i) + G_R(x_i)(x_i - \bar{x})\|_X}{\|x_i - \bar{x}\|_X}\|x_i - \bar{x}\|_X + \frac{1+\gamma}{1-\gamma}\rho_i\|R(x_i)\|_X \\
&\le \frac{1+\gamma}{(1-\gamma)^2}\frac{\|R(\bar{x}) - R(x_i) + G_R(x_i)(x_i - \bar{x})\|_X}{\|x_i - \bar{x}\|_X}\|R(x_i) - R(\bar{x})\|_X + \frac{1+\gamma}{1-\gamma}\rho_i\|R(x_i)\|_X \\
&= o(1)\|R(x_i)\|_X \qquad \forall i \in \mathbb{N}_0,
\end{aligned}$$

where the Landau notation $o(1)$ refers to the limit $i \to \infty$. This proves that $\|R(x_i)\|_X$ converges q-superlinearly to zero. To obtain that the convergence $\|x_i - \bar{x}\|_X \to 0$ is q-superlinear too, it suffices to note that the same arguments as in (88) and the estimates (85) and (86) imply

$$\|x_{i+1} - \bar{x}\|_X \le \frac{1}{1-\gamma}\|R(x_{i+1})\|_X \le o(1)\|R(x_i)\|_X = o(1)\|R(x_i) - R(\bar{x})\|_X \le o(1)\|x_i - \bar{x}\|_X$$

for $i \to \infty$. This establishes ii) and completes the proof of Theorem 2.4. □

References

[1] 2020. URL: https://github.com/JuliaNLSolvers/NLsolve.jl.

[2] 2022. URL: https://github.com/JuliaNLSolvers/LineSearches.jl.

[3] L. Adam, M. Hintermüller, and T. M. Surowiec. "A semismooth Newton method with analytical path-following for the H^1-projection onto the Gibbs simplex". In: *IMA J. Numer. Anal.* 39.3 (2019), pp. 1276–1295. DOI: 10.1093/imanum/dry034.

[4] D. R. Adams and L. I. Hedberg. *Function Spaces and Potential Theory*. Vol. 314. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, 1996. ISBN: 3-540-57060-8.

[5] Y. I. Alber. "A Bound for the Modulus of Continuity for Metric Projections in a Uniformly Convex and Uniformly Smooth Banach Space". In: *J. Approx. Theory* 85.3 (1996), pp. 237–249. ISSN: 0021-9045. DOI: https://doi.org/10.1006/jath.1996.0040.

[6] A. Alphonse, M. Hintermüller, and C. N. Rautenberg. "On the differentiability of the minimal and maximal solution maps of elliptic quasi-variational inequalities". In: *J. Math. Anal. Appl.* 507.1 (2022), p. 125732. ISSN: 0022-247X. DOI: https://doi.org/10.1016/j.jmaa.2021.125732.

[7] A. Alphonse, M. Hintermüller, and C. N. Rautenberg. "Directional differentiability for elliptic quasi-variational inequalities of obstacle type". In: *Calc. Var. Partial Differential Equations* 58.1 (2019), Paper No. 39, 47. ISSN: 0944-2669. DOI: 10.1007/s00526-018-1473-0.

[8] A. Alphonse, M. Hintermüller, and C. N. Rautenberg. “Optimal control and directional differentiability for elliptic quasi-variational inequalities”. In: *Set-Valued Var. Anal.* 30.3 (2022), pp. 873–922. ISSN: 1877-0533,1877-0541. DOI: 10.1007/s11228-021-00624-x.

[9] A. Alphonse, M. Hintermüller, and C. N. Rautenberg. “Stability of the Solution Set of Quasi-variational Inequalities and Optimal Control”. In: *SIAM J. Control Optim.* 58.6 (2020), pp. 3508–3532. DOI: 10.1137/19M1250327.

[10] A. Alphonse et al. “Minimal and maximal solution maps of elliptic QVIs: penalisation, Lipschitz stability, differentiability and optimal control”. In: *arXiv e-prints*, arXiv:2312.13879 (Dec. 2023), arXiv:2312.13879. DOI: 10.48550/arXiv.2312.13879. arXiv: 2312.13879.

[11] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces*. Philadelphia: SIAM, 2006.

[12] J.-P. Aubin. *Mathematical Methods of Game and Economic Theory*. Vol. 7. Studies in Mathematics and its Applications. North-Holland Publishing Co., Amsterdam-New York, 1979, pp. xxxii+619. ISBN: 0-444-85184-4.

[13] S. Badia and F. Verdugo. “Gridap: An extensible Finite Element toolbox in Julia”. In: *Journal of Open Source Software* 5.52 (2020), p. 2520. DOI: 10.21105/joss.02520. URL: https://doi.org/10.21105/joss.02520.

[14] C. Baiocchi and A. Capelo. *Variational and Quasivariational Inequalities*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1984, pp. ix+452. ISBN: 0-471-90201-2.

[15] J. W. Barrett and L. Prigozhin. “A quasi-variational inequality problem arising in the modeling of growing sandpiles”. In: *ESAIM: M2AN* 47.4 (2013), pp. 1133–1165. DOI: 10.1051/m2an/2012062.

[16] J. W. Barrett and L. Prigozhin. “A quasi-variational inequality problem in superconductivity”. In: *Math. Models Methods Appl. Sci.* 20.05 (2010), pp. 679–706. DOI: 10.1142/S0218202510004404.

[17] J. W. Barrett and L. Prigozhin. “Sandpiles and superconductors: nonconforming linear finite element approximations for mixed formulations of quasi-variational inequalities”. In: *IMA J. Numer. Anal.* 35.1 (Dec. 2013), pp. 1–38. ISSN: 0272-4979. DOI: 10.1093/imanum/drt062.

[18] A. Bensoussan. *Stochastic Control by Functional Analysis Methods*. North-Holland, 1982.

[19] A. Bensoussan and J. L. Lions. “Nouvelles methodes en contrôle impulsionnel”. In: *Appl. Math. Optim.* 1.4 (1975), pp. 289–312. DOI: 10.1007/bf01447955.

[20] A. Bensoussan and J. L. Lions. “Optimal impulse and continuous control: method of nonlinear quasi-variational inequalities”. In: *Trudy Mat. Inst. Steklov.* 134 (1975), pp. 5–22.

[21] A. Bensoussan and J.-L. Lions. *Impulse Control and Quasivariational Inequalities*. Gauthier-Villars, Montrouge; Heyden & Son, Inc., Philadelphia, PA, 1984, pp. xiv+684. ISBN: 2-04-015577-5.

[22] J. Bezanson et al. “Julia: A fresh approach to numerical computing”. In: *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671.

[23] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer-Verlag, New York, 2000, pp. xviii+601. ISBN: 0-387-98705-3. DOI: 10.1007/978-1-4612-1394-9.

[24] D. Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 2001.

[25] X. Chen, Z. Nashed, and L. Qi. “Smoothing Methods and Semismooth Methods for Nondifferentiable Operator Equations”. In: *SIAM J. Numer. Anal.* 38.4 (2000), pp. 1200–1216. DOI: 10.1137/s0036142999356719.

[26] C. Christof and G. Wachsmuth. “Energy Space Newton Differentiability for Solution Maps of Unilateral and Bilateral Obstacle Problems”. In: *arXiv e-prints*, arXiv:2308.15289 (Aug. 2023), arXiv:2308.15289. DOI: 10.48550/arXiv.2308.15289. arXiv: 2308.15289.

[27] C. Christof and G. Wachsmuth. “Lipschitz Stability and Hadamard Directional Differentiability for Elliptic and Parabolic Obstacle-Type Quasi-variational Inequalities”. In: *SIAM J. Control Optim.* 60.6 (2022), pp. 3430–3456. DOI: 10.1137/21M1419635.

[28] C. Christof and G. Wachsmuth. “Semismoothness for Solution Operators of Obstacle-Type Variational Inequalities with Applications in Optimal Control”. In: *SIAM J. Control Optim.* 61.3 (2023), pp. 1162–1186. DOI: 10.1137/21M1467365.

[29] F. H. Clarke. *Optimization and Nonsmooth Analysis*. 2nd ed. SIAM, 1990. DOI: 10.1137/1.9781611971309.

[30] R. S. Dembo, S. C. Eisenstat, and T. Steihaug. “Inexact Newton Methods”. In: *SIAM J. Numer. Anal.* 19.2 (1982), pp. 400–408. DOI: 10.1137/0719025.

[31] F. Facchinei et al. “The semismooth Newton method for the solution of quasi-variational inequalities”. In: *Comput. Optim. Appl.* 62.1 (2015), pp. 85–109. ISSN: 0926-6003,1573-2894. DOI: 10.1007/s10589-014-9686-4.

[32] M. Gerdts, S. Horn, and S.-J. Kimmerle. “Line search globalization of a semismooth Newton method for operator equations in Hilbert spaces with application in optimal control”. In: *J. Ind. Manag. Optim.* 13.1 (2017). DOI: 10.3934/jimo.2016003.

[33] K. Goebel and S. Prus. *Elements of Geometry of Balls in Banach Spaces*. Oxford University Press, 2018.

[34] C. Gräser and R. Kornhuber. “Multigrid methods for obstacle problems”. In: *J. Comput. Math.* (2009), pp. 1–44.

[35] D. A. Ham et al. *Firedrake User Manual*. First. Imperial College London et al. May 2023. DOI: 10.25561/104839.

[36] M. Hintermüller, K. Ito, and K. Kunisch. “The primal-dual active set strategy as a semismooth Newton method”. In: *SIAM J. Optim.* 13.3 (2003), pp. 865–888. DOI: 10.1137/S1052623401383558.

[37] M. Hintermüller and I. Kopacka. “A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs”. In: *Comput. Optim. Appl.* 50.1 (2011), pp. 111–145. ISSN: 0926-6003. DOI: 10.1007/s10589-009-9307-9. URL: http://dx.doi.org/10.1007/s10589-009-9307-9.

[38] A. F. Izmailov, A. L. Pogosyan, and M. V. Solodov. “Semismooth Newton method for the lifted reformulation of mathematical programs with complementarity constraints”. In: *Comput. Optim. Appl.* 51.1 (2012), pp. 199–221. ISSN: 0926-6003,1573-2894. DOI: 10.1007/s10589-010-9341-7.

[39] C. Kanzow and D. Steck. “Quasi-Variational Inequalities in Banach Spaces: Theory and Augmented Lagrangian Methods”. In: *SIAM J. Optim.* 29.4 (2019), pp. 3174–3200. DOI: 10.1137/18M1230475.

[40] B. Keith and T. M. Surowiec. “Proximal Galerkin: A structure-preserving finite element method for pointwise bound constraints”. In: *arXiv e-prints*, arXiv:2307.12444 (2023). arXiv: 2307.12444.

[41] T. Kilpeläinen and J. Malý. “Supersolutions to degenerate elliptic equations on quasi open sets”. In: *Comm. Partial Differential Equations* 17.3-4 (1992), pp. 371–405. ISSN: 0360-5302. DOI: 10.1080/03605309208820847.

[42] K. Kunisch and D. Wachsmuth. “Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities”. In: *ESAIM Control Optim. Calc. Var.* 18.2 (2012), pp. 520–547. ISSN: 1292-8119. DOI: 10.1051/cocv/2011105. URL: https://doi.org/10.1051/cocv/2011105.

[43] P. L. Lions and B. Perthame. “Quasi-variational inequalities and ergodic impulse control”. In: *SIAM J. Control Optim.* 24.4 (1986), pp. 604–615. DOI: 10.1137/0324036.

[44] M. Mandlmayr. “Semismooth* Newton methods for quasi-variational inequalities and contact problems with friction”. PhD thesis. Universität Linz, 2022.

[45] J. Martínez and L. Qi. “Inexact Newton methods for solving nonsmooth equations”. In: *J. Comput. Appl. Math* 60.1 (1995). Proceedings of the International Meeting on Linear/Nonlinear Iterative Methods and Verification of Solution, pp. 127–145. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(94)00088-I. URL: https://www.sciencedirect.com/science/article/pii/037704279400088I.

[46] U. Mosco. “Implicit variational problems and quasi variational inequalities”. In: *Nonlinear operators and the calculus of variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975)*. Lecture Notes in Math., Vol. 543. Springer, Berlin, 1976, pp. 83–156.

[47] T. Ni and J. Zhai. “A regularized smoothing Newton-type algorithm for quasi-variational inequalities”. In: *Comput. Math. Appl.* 68.10 (2014), pp. 1312–1324. ISSN: 0898-1221. DOI: https://doi.org/10.1016/j.camwa.2014.08.026.

[48] J. Nocedal and S. J. Wright. *Numerical Optimization*. 2nd ed. Springer, 2000. ISBN: 978-0387-30303-1. DOI: 10.1007/b98874.

[49] J. S. Pang and L. Qi. “A globally convergent Newton method for convex SC1 minimization problems”. In: *J. Optim. Theory Appl.* 85.3 (1995), pp. 633–648.

[50] I. P. A. Papadopoulos. *semismoothQVIs: a Firedrake implementation, v0.0.1*. 2024. URL: https://github.com/ioannisPApapadopoulos/semismoothQVIs.

[51] I. P. A. Papadopoulos. *SemismoothQVIs.jl, v0.0.1*. 2024. URL: https://github.com/ioannisPApapadopoulos/SemismoothQVIs.jl.

[52] I. P. A. Papadopoulos. *SemismoothQVIs.jl, v0.0.1 (Zenodo)*. Version v0.0.1. 2024. DOI: 10.5281/zenodo.13757145.

[53] L. Prigozhin. “On the Bean critical-state model in superconductivity”. In: *Eur. J. Appl. Math.* 7.3 (1996), pp. 237–247. DOI: 10.1017/s0956792500002333.

[54] L. Prigozhin. “Variational model of sandpile growth”. In: *Eur. J. Appl. Math.* 7.3 (1996), pp. 225–235. DOI: 10.1017/s0956792500002321.

[55] A.-T. Rauls and G. Wachsmuth. “Generalized derivatives for the solution operator of the obstacle problem”. In: *Set-Valued Var. Anal.* 28.2 (2020), pp. 259–285. ISSN: 1877-0533. DOI: 10.1007/s11228-019-0506-y.

[56] J.-F. Rodrigues. *Obstacle Problems in Mathematical Physics*. Vol. 134. North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam, 1987, pp. xvi+352. ISBN: 0-444-70187-7.

[57] J. F. Rodrigues and L. Santos. “A parabolic quasi-variational inequality arising in a superconductivity model”. en. In: *Ann. Sc. Norm. Super. Pisa Cl. Sci.* Ser. 4, 29.1 (2000), pp. 153–169.

[58] A. Schiela and D. Wachsmuth. “Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints”. In: *ESAIM Math. Model. Numer. Anal.* 47.3 (2013), pp. 771–787. ISSN: 0764-583X. DOI: 10.1051/m2an/2012049. URL: https://doi.org/10.1051/m2an/2012049.

[59] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Vol. 11. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011, pp. xiv+308. ISBN: 978-1-611970-68-5. DOI: 10.1137/1.9781611970692.

[60] F. Verdugo and S. Badia. “The software design of Gridap: A Finite Element package based on the Julia JIT compiler”. In: *Comput. Phys. Commun.* 276 (July 2022), p. 108341. DOI: 10.1016/j.cpc.2022.108341. URL: https://doi.org/10.1016/j.cpc.2022.108341.

[61] G. Wachsmuth. “Elliptic quasi-variational inequalities under a smallness assumption: uniqueness, differential stability and optimal control”. In: *Calc. Var. Partial Differential Equations* 59.2 (2020), Paper No. 82, 15. ISSN: 0944-2669. DOI: 10.1007/s00526-020-01743-3.

[62] L. C. Zeng. “On a general projection algorithm for variational inequalities”. In: *J. Optim. Theory Appl.* 97.1 (1998), pp. 229–235. ISSN: 0022-3239,1573-2878. DOI: 10.1023/A:1022687403403.