Interfaces Free Bound. 27 (2025), 521–573
DOI 10.4171/IFB/545

Minimal and maximal solution maps of elliptic QVIs of obstacle type: Lipschitz stability, differentiability, and optimal control

Amal Alphonse, Michael Hintermüller, Carlos N. Rautenberg, and Gerd Wachsmuth

Abstract. Quasi-variational inequalities (QVIs) of obstacle type in many cases have multiple solutions that can be ordered. We study a multitude of properties of the operator mapping the source term to the minimal or maximal solution of such QVIs. We prove that the solution maps are locally Lipschitz continuous and directionally differentiable and show existence of optimal controls for problems that incorporate these maps as the control-to-state operator. We also consider a Moreau–Yosida-type penalisation for the QVI, wherein we show that it is possible to approximate the minimal and maximal solutions by sequences of minimal and maximal solutions (respectively) of certain PDEs, which have a simpler structure and offer a convenient characterisation in particular for computation. For solution mappings of these penalised problems, we prove a number of properties including Lipschitz and differential stability. Making use of the penalised equations, we derive (in the limit) C-stationarity conditions for the control problem, in addition to the Bouligand stationarity we get from the differentiability result.

1. Introduction

Let $(\Omega, \sigma, \vartheta)$ be a measure space and define $H := L^2(\Omega)$ to be the usual Lebesgue space on this measure space. We utilise the partial ordering $\leq$ defined in the standard almost everywhere sense through ϑ. Take V to be a separable Hilbert space with $V \hookrightarrow H$ (a continuous embedding) and the property that $v \in V$ implies $v^+ \in V$ and that there exists a $C > 0$ with $\|v^+\|_V \leq C\|v\|_V$ for all $v \in V$. Here, $(\cdot)^+ = \max(0, \cdot)$ denotes the positive part of a function. Let $A : V \to V^*$ be a bounded, linear, coercive, and T-monotone operator and suppose that $\Phi : H \to V$ is a given obstacle map that is increasing. Given a source term $f \in V^*$, consider the quasi-variational inequality (QVI)

$$\text{find } u \in V,\ u \leq \Phi(u) \quad \text{such that} \quad \langle Au - f, u - v\rangle \leq 0, \quad \forall v \in V \text{ with } v \leq \Phi(u). \tag{1}$$

Under certain circumstances, this inequality has solutions that can be ordered and we denote by $\mathsf{M}(f)$ the maximal solution of (1) and by $\mathsf{m}(f)$ the minimal solution.

Mathematics Subject Classification 2020: 49J40 (primary); 47J20, 49J52, 49J21, 49K21 (secondary).
Keywords: quasi-variational inequality, optimal control, stationary system, ordered solution, directional differentiability.

In this paper, we study the sensitivity and directional differentiability of these extremal solution maps M and m, in addition to deriving stationarity conditions for optimisation problems with QVI constraints of the form

$$\min_{f \in U_{\mathrm{ad}}} J(\mathsf{M}(f), \mathsf{m}(f), f). \tag{2}$$

Regarding particular instances of J, we have in mind optimisation problems such as

$$\min_{f \in U_{\mathrm{ad}}} \frac{1}{2}\|\mathsf{M}(f) - \mathsf{m}(f)\|_H^2 + \frac{\nu}{2}\|f\|_H^2 \quad \text{and} \quad \min_{f \in U_{\mathrm{ad}}} \frac{1}{2}\|\mathsf{M}(f) - y_d\|_H^2 + \frac{\nu}{2}\|f\|_H^2. \tag{3}$$

The first is a formulation aiming to minimise the variation in solutions, first modelled and motivated in [3], and the second is the typical tracking-type problem.

Inspired in part by our interest in deriving stationarity conditions for the control problems and in part by some results of Lions and Bensoussan in [8, Chapter 4], a substantial portion of this paper is devoted to the study of the following penalised problem associated to QVI (1):

$$Au + \frac{1}{\rho}\sigma_\rho(u - \Phi(u)) = f, \tag{4}$$

where $\rho > 0$ is a parameter and σ_ρ is the below smoothed approximation of $(\cdot)^+$,

$$\sigma_\rho(r) := \begin{cases} 0 & \text{if } r \leq 0, \\ \frac{r^2}{2\rho} & \text{if } 0 < r < \rho, \\ r - \frac{\rho}{2} & \text{if } r \geq \rho. \end{cases} \tag{5}$$

It turns out that (4) also has multiple solutions that can be ordered and we can again find a maximal solution $\mathsf{M}_\rho(f)$ and a minimal one $\mathsf{m}_\rho(f)$. We provide a substantive analysis of the properties of these maps $\mathsf{M}_\rho, \mathsf{m}_\rho$ and also their limiting behaviour as $\rho \searrow 0$.

For convenience, we summarise our most important findings.

- We show that $\mathsf{M}_\rho(f)$ and $\mathsf{m}_\rho(f)$ converge to $\mathsf{M}(f)$ and $\mathsf{m}(f)$ respectively under some assumptions. Along the way, we prove that $\mathsf{M}_\rho(f)$ and $\mathsf{m}_\rho(f)$ can themselves be approximated by iterative sequences of solutions of PDEs, opening up the possibility for computation and numerical simulation (see Remark 4.11 for details).
- We prove that all four of these extremal solution maps (M_ρ, m_ρ, M, and m) are locally Lipschitz from V^* into V (by a bootstrapping and contraction argument; we also utilise some sharp estimates from [29] to ensure that our assumptions are kept as unobtrusive as possible).
- We demonstrate that the four maps are directionally differentiable for more general directions than in previous works, and also Hadamard differentiable in a certain sense (the proof is along the lines of the iterative approach of [2] with some modifications from [29]).

- Using the differentiability results on M and m, we derive first-order conditions of Bouligand type for the control problem. We also derive C-stationarity conditions, which is possible thanks to the various results on M_ρ and m_ρ that we obtain (we approximate (2) with a penalised control problem and then pass to the limit).

For precise details of all the main results, see Section 2 where we present them in full. Now, let us highlight the novelty and positioning of our work among the literature.

- Continuity of the minimal and maximal solution maps with perturbations in an L^∞-type space was first proved in [3, Theorem 4] under the structural assumption that $\lambda\Phi(u) \leq \Phi(\lambda u)$ for all $\lambda \in (0,1)$ and for $u \in H_+$. In [10, Theorem 3.2], Lipschitz continuity of these maps was shown, again under this setup and for source terms belonging to a subset of L^∞.
  In contrast, our result shows Lipschitz stability with respect to the V norm and for sources in V^* (thus we do not need to restrict to the L^∞ setting) and we do not require the homogeneity-type assumption on Φ (we do however ask for a local small Lipschitz assumption; see (8)).
  In particular, if Φ has a small Lipschitz constant around $\mathsf{M}(f)$, we already know that locally there is a stable (with respect to the norm in V) solution of the QVI [2, 5, 28], but it is not clear whether these are the maximal solutions. On the other hand, there are results [3, 10] showing that the maximal solution is stable (with respect to L^∞). Now, our new results show that the maximal solution is indeed V-stable.
- The first work on directional differentiability for solutions of QVIs in infinite dimensions is, to the best of our knowledge, [2] where it was shown for localised solutions and for non-negative directions. Subsequent work in [28] and [5] relaxed the assumptions of [2] greatly. All three papers use a type of smallness assumption on Φ (locally) similar to the one in this paper. However, neither paper tackled the case of extremal solutions. Regarding in particular differentiability for the minimal and maximal maps, this was proved in [4] under some sign conditions on the direction and a QVI characterisation of the derivative was given. In [10], again in an L^∞-type setting and with Φ assumed to be concave, a differentiability result for the maximal solution appears and under assumptions that entail the unique global solvability of the QVI, a characterisation of the derivative is given.
  In this work, we provide a unique QVI characterisation of the directional derivative of the minimal and maximal solution maps under a general and natural function space setting and with relatively agreeable assumptions. In contrast to the two previous works [4, 10] on extremal solution maps, we require neither sign restrictions on the perturbation directions nor concavity or homogeneity-type assumptions on Φ, nor an embedding into L^∞.
- The study of the specifics of the maps M_ρ and m_ρ in this general setting seems entirely new, although we should once again remind the reader that [8] contains some results on the convergence behaviour of these maps in a specific setting (and not in generality

like ours). The results on the sensitivity and differentiability of the maps are new, as are the convergence results in this generality.

- The stationarity conditions for the control problem involving minimal and maximal solution maps are also entirely new. The works [5, 28] have addressed stationarity for control problems in a QVI setting but not for the extremal solution maps. Furthermore, our C-stationarity system in some sense improves the one in [5] because we are able to show that the multipliers for the adjoints vanish on the inactive set (formally speaking; see Proposition 7.8) without requiring any additional strong assumptions.
- On this note, we are for the first time able to treat problems like the first one in (3) in a substantial way.
- Our results remain valid when the obstacle mapping $\Phi \equiv \psi$ is constant, that is, in the case where (1) is a variational inequality. We note in particular that Proposition 7.8 improves the $\mathcal{E}$-almost conditions derived in [14, Theorem 3.4] for control of the obstacle problem; see Proposition 7.9.

Although we have specified the functional framework of this paper with the base space chosen as $L^2(\Omega)$, let us stress that in fact, many of our results will apply in far greater generality, with a much more general function space setting (than $H = L^2(\Omega)$ as taken above) and also with far more general maps σ_ρ (provided certain crucial properties are satisfied) than the one above – for example, $\sigma_\rho(u) := u^+$. For simplicity and clarity of exposition, we have decided to present our work with the choice of H as above and with σ_ρ as in (5) in the paper. We will not present the details here but invite interested readers to work out the details.

Regarding the organisation of the paper, we begin in Section 1.1 with some basic definitions, notations, and fundamental results. In Section 2 we state all of our main results for the convenience of the reader, also including some useful or interesting remarks and providing some examples in Section 2.4.1. In Section 3, we study (4) and an iterative sequence of associated problems and show that (4) does indeed possess extremal solutions. Section 4 is devoted to the study of the limit $\rho \to 0$ in (4), both with and without a locally small Lipschitz assumption on Φ. Using these obtained results, we prove our claims on the Lipschitzness of all the maps in Section 5 and directional differentiability in Section 6. In Section 7, we study the optimisation problem in (2) and prove B-stationarity and various forms of C-stationarity. In Section 8, we finish the main part of the paper with some final remarks.

1.1. Notation and preliminaries

Define the set $H_+ := \{h \in H : h \geq 0\}$ of non-negative elements of $H = L^2(\Omega)$, and define V_+ similarly. We write $h^+ = P_{H_+}h$ to denote the orthogonal projection of $h \in H$ onto H_+ and define $h^- := h^+ - h$. The infimum and supremum of two elements $h_1, h_2 \in H$ are defined as usual: $\inf(h_1, h_2) := h_1 - (h_1 - h_2)^+$ and $\sup(h_1, h_2) := h_1$

$+ (h_2 - h_1)^+$. We define an order on the dual space V^* via

$$f \leq g \iff \langle f - g, v \rangle \leq 0, \quad \forall v \in V_+.$$

Here $\langle \cdot, \cdot \rangle$ is the duality pairing between V^* and V. Regarding the elliptic operator in (1), as mentioned, we take $A : V \to V^*$ to be a linear operator that satisfies the following properties for all $u, v \in V$:

$$\langle Au, v \rangle \leq C_b \|u\|_V \|v\|_V, \tag{boundedness}$$

$$\langle Au, u \rangle \geq C_a \|u\|_V^2, \tag{coercivity}$$

$$\langle Au^+, u^- \rangle \leq 0, \tag{T-monotonicity}$$

where $C_a, C_b > 0$ are constants.

With $\mathbf{K}(u) := \{v \in V : v \leq \Phi(u)\}$, QVI (1) can be written as

$$u \in \mathbf{K}(u) : \langle Au - f, u - v \rangle \leq 0, \quad \forall v \in \mathbf{K}(u).$$

We introduce $S : V^* \times H \to V$ as the solution map of the associated variational inequality, that is, $u = S(f, \psi)$ if and only if

$$u \in \mathbf{K}(\psi) : \langle Au - f, u - v \rangle \leq 0, \quad \forall v \in \mathbf{K}(\psi).$$

Thus the solutions of (1) are precisely the fixed points of $S(f, \cdot)$.

Assumption 1.1. Given $f \in V^*$, assume that there exist $\underline{u}, \overline{u} \in V$ such that

$$\underline{u} \leq S(f, \underline{u}), \quad \overline{u} \geq S(f, \overline{u}), \quad \text{and} \quad \underline{u} \leq \overline{u}.$$

The element $\underline{u}$ is called a *subsolution* for $S(f, \cdot)$ and $\overline{u}$ is called a *supersolution* for $S(f, \cdot)$.

We come now to an existence result for (1). For more existence results under different assumptions, see [5, §2].

Proposition 1.2. *Under Assumption 1.1, there exists a minimal solution* $\mathsf{m}(f)$ *and maximal solution* $\mathsf{M}(f)$ *to* (1) *on the interval* $[\underline{u}, \overline{u}] := \{v \in V : \underline{u} \leq v \leq \overline{u}$ *a.e. in* $\Omega\}$.

Proof. We apply the Birkhoff–Tartar theorem [6, §15.2.2, Proposition 2], which gives existence of fixed points for increasing maps that possess subsolutions and supersolutions to the map $S(f, \cdot)$ (which is increasing; see [20, §4:5, Theorem 5.1]). See also [24], [7, §11.2], and [19, Chapter 2]. ■

We will use the notation $B_r(x)$ to denote the (closed) ball of radius r centred at x. It should be clear from the context the function space in which the ball is taken, but typically when we use δ (or a variant such as $\overline{\delta}$) for the radius, it refers to the V^* ball, whereas the radius being ε (or a variant) refers to the V ball.

2. Main results

Let us discuss our main results. As a matter of notation, to handle both cases (of the minimal and maximal solution maps) simultaneously, we denote by Z one of the maps M or m. Note that Z is defined at all points f satisfying Assumption 1.1.

2.1. On directional differentiability

Our first result concerns local Lipschitz continuity of Z. For this, we need Z to be defined not just at a solitary point but in a neighbourhood. Thus, we need to expand Assumption 1.1 to take this into account.

Assumption 2.1. Let $f \in V^*$ and take a set $W \subseteq V^*$ containing f and assume that there exist $\underline{u}, \overline{u} \in V$ and $\overline{\delta}$ such that

$$\underline{u} \leq \overline{u}, \tag{6a}$$

$$\underline{u} \leq S(g, \underline{u}), \quad \forall g \in B_{\overline{\delta}}(f) \cap W, \tag{6b}$$

$$\overline{u} \geq S(g, \overline{u}), \quad \forall g \in B_{\overline{\delta}}(f) \cap W. \tag{6c}$$

The intersection with the set W that appears in the assumption above is inspired by applications where the source terms may lie in some given ordered interval and it ensures that natural candidates for the sub- and supersolutions (namely those arising from the boundary of the ordered interval) indeed qualify as sub- and supersolutions; see the next remark.

Remark 2.2. Consider the example in Section 2.4.1. If we had asked for (6) to hold for all $g \in B_{\overline{\delta}}(f)$ (i.e., without the intersection with a set W), then $\overline{u} = 0$ does not satisfy (6b) for the element $f = 0 \in V^*$ since $B_{\overline{\delta}}(0)$ contains negative functions, so that $0 \leq S(g, 0)$ may not hold for all $g \in B_{\overline{\delta}}(0)$. Even worse, due to $S(g, \underline{u}) \leq A^{-1}g$ for all $g \in V^*$ and $\underline{u} \in V$, we would need $\underline{u} \leq A^{-1}g$ for all $g \in B_{\overline{\delta}}(f)$, but this is not possible since $A^{-1}g$ could have negative singularities at arbitrary points. Hence, the intersection with W is necessary for the existence of sub- and supersolutions.

The next theorem will be proved in Section 5.

Theorem 2.3 (Local Lipschitz continuity of Z). *Let $f \in V^*$ and $W \subseteq V^*$ satisfy Assumption 2.1. Assume*

$$\Phi : V \to V \text{ is completely continuous,} \tag{7}$$

there exists $\varepsilon^* > 0$ such that $\Phi : B_{\varepsilon^*}(\mathsf{Z}(f)) \to V$ has a Lipschitz constant C_L satisfying

$$C_L < \frac{C_a}{C_b} \text{ or } A \text{ is self-adjoint and } C_L < 2\frac{\sqrt{C_b/C_a}}{1 + C_b/C_a}. \tag{8}$$

Then then there exists $\delta \in (0, \bar{\delta})$ *such that for all* $g \in B_\delta(f) \cap W$,

$$\|\mathsf{Z}(f) - \mathsf{Z}(g)\|_V \le C\|f - g\|_{V^*}$$

where $C > 0$ *is a constant (which depends only on* C_L, C_a, C_b *and the self-adjointness of* A*).*

In the assumption in (8) above, "self-adjoint" essentially means that the associated bilinear form is symmetric. Note that (8) is indeed a rather strong assumption as it asks for a smallness condition on the Lipschitz constant of Φ, albeit only locally. In particular, the assumption implies local uniqueness on the ball $B_{\varepsilon^*}(\mathsf{Z}(f))$: any solution that exists in the ball is isolated. The nature of QVIs where non-uniqueness appears seems to necessitate such assumptions. We will demonstrate a real-world application in which such an assumption is satisfied in Section 2.4.2.

With the addition of just one more assumption (namely the differentiability of Φ at a point) we can secure directional differentiability. Before we state the result, let us recall that the *radial cone* of a set $C \subset X$ of a Banach space X at a point $x \in C$ is defined as

$$\mathcal{R}_C(x) := \{y \in X \mid \exists s_0 > 0 : x + sy \in C, \quad \forall s \in [0, s_0]\}.$$

The *tangent cone* is defined as

$$\mathcal{T}_C(x) := \{y \in X \mid \exists s_k \searrow 0, \exists y_k \to y \text{ in } X : x + s_k y_k \in C, \quad \forall k\}.$$

In the case that C is additionally convex, the tangent cone is the closure of the radial cone in X, written $\mathcal{T}_C(x) = \overline{\mathcal{R}_C(x)}$.

Theorem 2.4 (Hadamard differentiability of Z). *Let* $f \in V^*$ *and* $W \subseteq V^*$ *satisfy Assumption 2.1. Assume (7), (8), and*

$$\Phi \textit{ is directionally differentiable at } \mathsf{Z}(f). \tag{9}$$

Then

(i) *the map* Z *is Hadamard differentiable in the sense that if* $d \in \mathcal{T}_W(f)$, *then for any sequence* $d_k \to d$ *in* V^* *with* $f + s_k d_k \in W$ *where* $s_k \searrow 0$,

$$\frac{\mathsf{Z}(f + s_k d_k) - \mathsf{Z}(f)}{s_k} \to \mathsf{Z}'(f)(d);$$

(ii) *the derivative* $\mathsf{Z}'(f)(d)$ *is the unique solution of the QVI*

$$\alpha \in \mathcal{K}^u(\alpha) : \langle A\alpha - d, \alpha - v\rangle \le 0, \quad \forall v \in \mathcal{K}^u(\alpha) \tag{10}$$

where, writing $u = \mathsf{Z}(f)$,

$$\mathcal{K}^u(\alpha) := \Phi'(u)(\alpha) + \mathcal{T}_{\mathbf{K}(u)}(u) \cap [f - Au]^\perp;$$

(iii) *the map* $\mathsf{Z}'(f) : \mathcal{T}_W(f) \to V$ *can be extended to a bounded and continuous mapping from* V^* *to* V *by defining it via* (10) *for all* $d \in V^*$.

For the proof, see Section 6.2.

Remark 2.5 (Directional differentiability of Z). A simple corollary of Theorem 2.4(i) is that the map $\mathsf{Z} : B_{\overline{\delta}}(f) \cap W \to V$ is directionally differentiable at f in every direction $d \in \mathcal{R}_W(f) \subset V^*$:

$$\lim_{s \searrow 0} \frac{\mathsf{Z}(f + sd) - \mathsf{Z}(f)}{s} = \mathsf{Z}'(f)(d).$$

Taking the direction from $\mathcal{R}_W(f)$ ensures that the perturbed solution $\mathsf{Z}(f + sd)$ is well defined via Assumption 2.1.

Example 2.6 (The radial cone $\mathcal{R}_W(f)$). Similarly to Section 2.4.1, let us consider

$$W := \left\{g \in V^* : 0 \leq g \leq F\right\}$$

with $F \geq k_0$ for some constant $k_0 > 0$, and $\underline{u} := 0$ and $\overline{u} := A^{-1}F$. Let us try to describe the radial cone at different points in W.

- Take $d \in L^\infty_+(\Omega)$. Then $sd \geq 0$ for all $s > 0$ and if $s \leq k_0/\|d\|_{L^\infty(\Omega)}$, we have, for all $\varphi \in V_+$,

$$\langle sd - F, \varphi\rangle = \langle sd - k_0, \varphi\rangle + \langle k_0 - F, \varphi\rangle \leq 0,$$

  so that $sd \in W$ for sufficiently small s. This shows that $L^\infty_+(\Omega) \subset \mathcal{R}_W(0)$.
- In a similar way, take $d \in L^\infty_-(\Omega)$. For all $s \geq 0$, we have $F + sd \leq F$ and if $s \leq k_0/\|d\|_{L^\infty(\Omega)}$, we have

$$\langle F + sd, \varphi\rangle = \langle F - k_0, \varphi\rangle + \langle k_0 + sd, \varphi\rangle$$

  and $k_0 + sd \geq k_0 + k_0 d/\|d\|_{L^\infty(\Omega)} = k_0(1 + d/\|d\|_{L^\infty(\Omega)}) \geq 0$, thus $F + sd \geq 0$ and we have shown that $L^\infty_-(\Omega) \subset \mathcal{R}_W(F)$;
- Now consider a point f such that $0 < c_0 \leq f \leq c_1 < F$ where c_0 and c_1 are constants. Using similar arguments to the above, we can show $L^\infty(\Omega) \subset \mathcal{R}_W(f)$.

2.2. On the penalised problem

In this section, we address results for the penalised problem in (4), that is,

$$Au + \frac{1}{\rho}\sigma_\rho(u - \Phi(u)) = f.$$

We denote by $T_\rho : V^* \times H \to V$ the solution map $(f, w) \mapsto u$ of the corresponding equation

$$Au + \frac{1}{\rho}\sigma_\rho(u - \Phi(w)) = f.$$

Here, we associate with the real-valued function σ_ρ defined in (5) the operator $\sigma_\rho : V \to V^*$ defined as

$$\langle \sigma_\rho(u), v\rangle = \int_\Omega \sigma_\rho(u)v.$$

If, given $f \in V^*$ and a fixed $\rho > 0$, we have the availability of $\underline{u}, \overline{u}$ such that

$$\underline{u} \leq T_\rho(f, \underline{u}), \quad \overline{u} \geq T_\rho(f, \overline{u}), \quad \text{and} \quad \underline{u} \leq \overline{u}, \tag{11}$$

then there exist a minimal solution $\mathsf{m}_\rho(f)$ and maximal solution $\mathsf{M}_\rho(f)$ to (4) on $[\underline{u}, \overline{u}]$. We will show this in Proposition 3.10. In a similar way as before, we use Z_ρ to denote either M_ρ or m_ρ.

Since we want to consider the limit $\rho \searrow 0$, we need Z_ρ to be defined for sufficiently small ρ, and hence (11) (which holds for a fixed ρ) needs to be modified. We do this in the next assumption, which kills two birds with one stone: it also ensures that both Z_ρ and Z are defined on a neighbourhood (just like we argued for Assumption 2.1) and not just at one point.

Assumption 2.7. Let $f \in V^*$ and take a set $W \subseteq V^*$ containing f and assume that there exist $\underline{u}, \overline{u} \in V$ and $\overline{\delta}, \rho_0 > 0$ such that

$$\underline{u} \leq \overline{u}, \tag{12a}$$

$$\underline{u} \leq S(g, \underline{u}), \qquad \forall g \in B_{\overline{\delta}}(f) \cap W, \tag{12b}$$

$$\overline{u} \geq T_{\rho_0}(g, \overline{u}), \quad \forall g \in B_{\overline{\delta}}(f) \cap W. \tag{12c}$$

The fundamental question is whether $\mathsf{M}_\rho(f)$ and $\mathsf{m}_\rho(f)$ converge (in some sense) to $\mathsf{M}(f)$ and $\mathsf{m}(f)$. In fact, we can even prove something stronger with the following joint (in ρ and the source term) continuity result, the proof of which appears in Section 4.3:

Theorem 2.8 (Convergence of $\mathsf{Z}_\rho(g)$ to $\mathsf{Z}(f)$). *Let $f \in V^*$ and $W \subseteq V^*$ satisfy Assumption 2.7. Assume (7)[1] and (8). Then*

$$\lim_{\substack{\rho \searrow 0 \\ g \to f}} \mathsf{Z}_\rho(g) = \mathsf{Z}(f)$$

where the convergence $g \to f$ is understood in V^* and for $g \in W$.

As we said in Remark 2.2, having $W \neq V^*$ in Assumption 2.7 above makes it a weaker assumption than if it held with W equal to the entire space V^*, and leads to a convergence result with respect to g that is perhaps weaker than one might first expect, but this is obviously natural since the extremal maps only exist for such source terms.

By choosing $W = \{f\}$ in the statement of the theorem, we get the corollary below. Note that the assumption below essentially asks for the inequalities in (12) to hold only at (g replaced with) the particular point f.

[1] Instead of (7) we could assume that $\Phi : V \to V$ is continuous, (29), (30) and (32).

Corollary 2.9 (Convergence of $\mathsf{Z}_\rho(f)$ to $\mathsf{Z}(f)$). *Let $f \in V^*$ and $W := \{f\}$ satisfy Assumption 2.7 and assume (7) and (8). Then $\mathsf{Z}_\rho(f) \to \mathsf{Z}(f)$ in V.*

Remark 2.10. This result uses the small Lipschitz condition in (8), but it is not necessary to obtain the convergence of $\mathsf{M}_\rho(f)$ to $\mathsf{M}(f)$; see Theorem 4.8. It is an open problem whether $\mathsf{m}_\rho(f)$ converges to $\mathsf{m}(f)$ under the general assumptions of Theorem 4.8.

In a similar fashion to Theorems 2.3 and 2.4, we have the following local Lipschitz and differentiability results for Z_ρ, proven in Sections 5 and 6.1, respectively:

Theorem 2.11 (Local Lipschitz continuity of Z_ρ). *Let $f \in V^*$ and $W \subseteq V^*$ satisfy Assumption 2.7. Assume (7) and (8). Then there exist ρ_0 and $\delta > 0$ such that for all $\rho \leq \rho_0$ and $g \in B_\delta(f) \cap W$,*

$$\|\mathsf{Z}_\rho(f) - \mathsf{Z}_\rho(g)\|_V \leq C\|f - g\|_{V^*}$$

where $C > 0$ is a constant (which depends only on C_L, C_a, C_b and the self-adjointness of A).

Theorem 2.12 (Hadamard differentiability of Z_ρ). *Let $f \in V^*$ and $W \subseteq V^*$ satisfy Assumption 2.7. Assume (7), (8), and*

$$\Phi \text{ is directionally differentiable at } \mathsf{Z}_\rho(f). \tag{13}$$

Then for ρ sufficiently small,

(i) *the map Z_ρ is Hadamard differentiable in the sense that if $d \in \mathcal{T}_W(f)$, then for any sequence $d_k \to d$ in V^* with $f + s_k d_k \in W$ where $s_k \searrow 0$,*

$$\frac{\mathsf{Z}_\rho(f + s_k d_k) - \mathsf{Z}_\rho(f)}{s_k} \to \mathsf{Z}_\rho'(f)(d);$$

(ii) *the derivative $\mathsf{Z}_\rho'(f)(d)$ is the unique solution of the equation*

$$A\alpha + \frac{1}{\rho}\sigma_\rho'(u - \Phi(u))(\alpha - \Phi'(u)(\alpha)) = d \tag{14}$$

where $u = \mathsf{Z}_\rho(f)$;

(iii) *the map $\mathsf{Z}_\rho'(f) : \mathcal{T}_W(f) \to V$ can be extended to a bounded and continuous mapping from V^* to V by defining it via (14) for all $d \in V^*$.*

Exactly as in Remark 2.5, we obtain from Theorem 2.12(i) the directional differentability of $\mathsf{Z}_\rho : B_{\bar{\delta}}(f) \cap W \to V$ at f in every direction $d \in \mathcal{R}_W(f) \subset V^*$:

$$\lim_{s \searrow 0} \frac{\mathsf{Z}_\rho(f + sd) - \mathsf{Z}_\rho(f)}{s} = \mathsf{Z}_\rho'(f)(d).$$

2.3. On optimal control

Regarding existing literature on the derivation of stationarity systems for optimal control with QVI constraints, we mention [28] and [5] in particular. The first work contains a strong stationarity system characterisation in the absence of control constraints, while the latter work includes the derivation of various forms of stationarity systems (including strong) with potential box constraints on the control. In this work, we extend these results to the setting of minimal and maximal solution mappings and derive a C-stationarity system.

Suppose that

$$V \overset{c}{\hookrightarrow} H \hookrightarrow V^* \text{ is a Gelfand triple}$$

($\overset{c}{\hookrightarrow}$ means a compact embedding; by definition of the Gelfand triple, $V \overset{d}{\hookrightarrow} H$ is a dense embedding) and let $U_{\text{ad}} \subset H$ be a non-empty, closed, and convex set[2]. Recall the control problem in (2), reproduced here:

$$\min_{f \in U_{\text{ad}}} J(\mathsf{M}(f), \mathsf{m}(f), f).$$

We make the next standing assumption, which guarantees the well-definedness of (2).

Assumption 2.13. There exist $\underline{u}, \overline{u} \in V$ such that

$$\begin{aligned} &\underline{u} \leq \overline{u}, \\ &\underline{u} \leq S(g, \underline{u}), \quad \forall g \in U_{\text{ad}}, \\ &\overline{u} \geq S(g, \overline{u}), \quad \forall g \in U_{\text{ad}}. \end{aligned}$$

Regarding the objective functional J, we need the following assumptions in place. Observe that the last two assumptions below are conditions that involve Φ.

Assumption 2.14. Regarding $J(y, z, f)$, assume that

(i) $J : V \times V \times H \to \mathbb{R}$ is continuously Fréchet differentiable and bounded from below.

(ii) If $(y_n, z_n) \to (y, z)$ in $V \times V$ and $f_n \rightharpoonup f$ in H, then

$$J(y, z, f) \leq \liminf_{n \to \infty} J(y_n, z_n, f_n).$$

(iii) If $\{J(y_n, z_n, f_n)\}$ is bounded for a sequence $\{(y_n, z_n, f_n)\} \subset V \times V \times U_{\text{ad}}$, then $\{f_n\}$ is bounded in H.

(iv) If $J_y \not\equiv 0$, for every $f \in U_{\text{ad}}$, there exists $\varepsilon^* > 0$ such that $\Phi : B_{\varepsilon^*}(\mathsf{M}(f)) \to V$ has a Lipschitz constant satisfying $C_L < C_a/C_b$ or A is self-adjoint and $C_L < 2\sqrt{C_b/C_a}(1 + C_b/C_a)^{-1}$.

[2] It would suffice to replace "closed and convex" here with "weakly sequentially closed" (which is a weaker requirement) for the existence results below.

(v) If $J_z \not\equiv 0$, for every $f \in U_{\text{ad}}$, there exists $\varepsilon^* > 0$ such that $\Phi : B_{\varepsilon^*}(\mathsf{m}(f)) \to V$ has a Lipschitz constant satisfying $C_L < C_a/C_b$ or A is self-adjoint and $C_L < 2\sqrt{C_b/C_a}(1 + C_b/C_a)^{-1}$.

An example of J satisfying items (i)–(iii) above is

$$J(y, z, f) = \frac{1}{2}\|ay + bz - y_d\|_H^2 + \frac{\nu}{2}\|f\|_H^2$$

given constants $a, b \in \mathbb{R}$, $\nu > 0$, and for some given $y_d \in H$. When we choose $a = 1$, $b = -1$, and $y_d \equiv 0$, we recover the first objective functional in (3) and when $a = 1$ and $b = 0$ or vice versa, we recover the second one in (3).

We remark that the assumptions in (iv) and (v) are unfortunately rather unsatisfactory, because they impose local uniqueness around the extremal solutions for *every* source term in U_{ad}.

Theorem 2.15 (Existence of optimal controls). *Assume* (7)*, and Assumptions* 2.13 *and* 2.14*. Then there exists an optimal control $f^* \in U_{\text{ad}}$ to the problem in* (2)*.*

The proof (see Section 7) is more or less standard and uses the direct method in the calculus of variations. From now on, let

$$(y^*, z^*, f^*) \text{ be an arbitrary local minimiser of (2)}$$

with $y^* = \mathsf{M}(f^*)$ and $z^* = \mathsf{m}(f^*)$. We begin with the following primal characterisation of the minimiser:

Proposition 2.16 (Bouligand stationarity). *Assume* (7)*, Assumptions* 2.13*, and* 2.14*, and*

$$\text{if } J_y \not\equiv 0,\ \Phi \text{ is directionally differentiable at } \mathsf{M}(f^*),$$
$$\text{if } J_z \not\equiv 0,\ \Phi \text{ is directionally differentiable at } \mathsf{m}(f^*).$$

Then

$$\langle J_y(y^*, z^*, f^*), \mathsf{M}'(f^*)(h)\rangle + \langle J_z(y^*, z^*, f^*), \mathsf{m}'(f^*)(h)\rangle + \langle J_f(y^*, z^*, f^*), h\rangle \\ \geq 0, \quad \forall h \in \mathcal{T}_{U_{\text{ad}}}(f^*).$$

The proof of the proposition appears in Section 7.1.

For numerics, it is convenient to derive other forms of stationarity systems, like C-stationarity. For this purpose, we consider the penalised control problem

$$\min_{f \in U_{\text{ad}}} J(\mathsf{M}_\rho(f), \mathsf{m}_\rho(f), f). \tag{15}$$

The following standing assumption is stronger than Assumption 2.13 and it implies the assumptions of Theorem 2.11, which is needed for the existence of controls for the above control problem.

Assumption 2.17. There exist $\underline{u}, \overline{u} \in V$ and $\rho_0 > 0$ such that

$$\begin{aligned} &\underline{u} \leq \overline{u}, \\ &\underline{u} \leq S(g, \underline{u}), \qquad \forall g \in U_{\text{ad}}, \\ &\overline{u} \geq T_{\rho_0}(g, \overline{u}), \quad \forall g \in U_{\text{ad}}. \end{aligned}$$

We need some further regularity on Φ in the form of the next assumption. When Φ is continuously Fréchet differentiable, these assumptions follow from Assumptions 2.14(iv) and (v). See the discussion around (40) and the proof of [5, Lemma 5.9].

Assumption 2.18. Assume the following:

(i) If $J_y \not\equiv 0$, assume that there exists $\varepsilon > 0$ such that

$$\text{for all } w \in B_\varepsilon(y^*),\ \Phi \text{ is directionally differentiable at } w \text{ and } \Phi'(w) \text{ is linear.}$$

If $J_z \not\equiv 0$, the above holds with y^* replaced by z^*.

(ii) If $J_y \not\equiv 0$, assume that for sequences $v_n \to v$, $w_n \to w$, and $q_n \rightharpoonup q$ in V with $v_n, v \in B_\varepsilon(y^*)$, we have

$$(\text{Id} - \Phi'(v_n))^{-1} q_n \rightharpoonup (\text{Id} - \Phi'(v))^{-1} q \quad \text{in } V, \tag{16}$$

$$(\text{Id} - \Phi'(v_n))^{-1} w_n \to (\text{Id} - \Phi'(v))^{-1} w \quad \text{in } V, \tag{17}$$

If $J_z \not\equiv 0$, the above holds with y^* replaced by z^*.

We also need additional structure on the function spaces in the form of a Dirichlet space.

Assumption 2.19. Let V be a regular Dirichlet space and suppose that $(\cdot)^+ : V \to V$ is continuous.

We will not enter into an exposition about Dirichlet spaces here (see [5, Example 3.5] for a convenient definition and comments on this as well as further references), but let us give some examples that satisfy the above assumption. Suppose that $D \subset \mathbb{R}^n$ is a bounded Lipschitz domain. We can take $H = L^2(D)$ and $V = H_0^1(D)$ (thus $\Omega \equiv D$), or $H = L^2(\overline{D})$ and $V = H^1(D)$ (thus $\Omega \equiv \overline{D}$). Fractional spaces are also a possibility. Indeed, $V = H^s(D)$ for $s \in (0, 1)$ and $H = L^2(\overline{D})$ is valid (i.e., $\Omega := \overline{D}$), where the fractional Sobolev space $H^s(D)$ is defined as usual as the space of measurable functions $u : D \to \mathbb{R}$ such that the norm

$$\|u\|_{H^s(D)} := \left(\int_D u^2 + \int_D \int_D \frac{|u(x) - u(y)|^2}{|x - y|^{n+2s}} \right)^{\frac{1}{2}} \tag{18}$$

is finite. On the plane, we could also pick $V = H^s(\mathbb{R}^d)$ and $H = L^2(\mathbb{R}^d)$ (V is defined similarly to the above via (18) but with D replaced with $\mathbb{R}^d$); thus, here $\Omega = \mathbb{R}^d$. In these two cases the natural operator to choose for A would be the fractional Laplacian $(-\Delta)^s$. We refer to [2, §1.2.1] for a QVI example involving the fractional Laplacian in an application involving fluid flow.

Assuming a Dirichlet space structure enables us to define notions of capacity and quasi-continuity (capacity is, loosely speaking, a way to measure sets finer than through the Lebesgue measure), see [11, §2.1] or [26, Section 2] for the $H^1_0(\Omega)$ setting. In addition, it allows us to explicitly characterise the critical cone appearing in Theorem 2.4 (see Remark 6.7) using capacity, and (more pertinently for us in this section) the tangent cone as well, which is something we will use to prove a statement in the stationarity system below. On that topic, note for any $y \in V$ we can define[3]

$$\{y = \Phi(y)\} \equiv \{x \in \Omega : y(x) = \Phi(y)(x)\},$$

which, when y is a solution of the QVI, is called the *active* or *coincidence set*. This set is defined up to sets of capacity zero.

We will prove a version of C-stationarity (but note that this terminology is used somewhat inconsistently in the literature). Before we proceed, let us record that owing to the complementarity characterisation of solutions of QVIs (see, e.g., [5, Proposition 2.1]), the statements $y^* = \mathsf{M}(f^*)$ and $z^* = \mathsf{m}(f^*)$ imply (but are not necessarily equivalent to) that

$$\begin{aligned} Ay^* - f^* + \xi_1^* &= 0,\\ Az^* - f^* + \xi_2^* &= 0,\\ \xi_1^* \geq 0 \text{ in } V^*, \quad y^* \leq \Phi(y^*), \quad \langle \xi_1^*, y^* - \Phi(y^*)\rangle &= 0,\\ \xi_2^* \geq 0 \text{ in } V^*, \quad z^* \leq \Phi(z^*), \quad \langle \xi_2^*, z^* - \Phi(z^*)\rangle &= 0. \end{aligned} \tag{19}$$

The main result in this section is the following, which will be proved through a succession of results in Section 7.3:

Theorem 2.20 (C-stationarity). *Assume (7), and Assumptions 2.14, 2.17, 2.18, and 2.19. Take any local minimiser (y^*, z^*, f^*) of (2) and define ξ_1^*, ξ_2^* as in (19). Then there exist multipliers $(p^*, q^*, \lambda^*, \zeta^*) \in V \times V \times V^* \times V^*$ satisfying the C-stationarity system*

$$y^* = \mathsf{M}(f^*), \tag{20a}$$

$$z^* = \mathsf{m}(f^*), \tag{20b}$$

$$A^*p^* + (\mathrm{Id} - \Phi'(y^*))^*\lambda^* = -J_y(y^*, z^*, f^*), \tag{20c}$$

$$A^*q^* + (\mathrm{Id} - \Phi'(z^*))^*\zeta^* = -J_z(y^*, z^*, f^*), \tag{20d}$$

$$f^* \in U_{\mathrm{ad}} : \langle J_f(y^*, z^*, f^*) - p^* - q^*, f^* - v\rangle \leq 0 \quad \forall v \in U_{\mathrm{ad}}, \tag{20e}$$

$$\langle \lambda^*, p^*\rangle \geq 0, \tag{20f}$$

$$\langle \zeta^*, q^*\rangle \geq 0, \tag{20g}$$

[3] Strictly speaking, every $y \in V$ has a quasi-continuous representative and we identify it with its representative. Then the set $\{y = \Phi(y)\}$ is quasi-closed.

$$\langle \lambda^*, v\rangle = 0, \quad \forall v \in V : v = 0 \textit{ q.e. on } \{y^* = \Phi(y^*)\}, \tag{20h}$$

$$\langle \zeta^*, v\rangle = 0, \quad \forall v \in V : v = 0 \textit{ q.e. on } \{z^* = \Phi(z^*)\}, \tag{20i}$$

$$\langle \xi_1^*, (p^*)^+\rangle = \langle \xi_1^*, (p^*)^-\rangle = \langle \xi_2^*, (q^*)^+\rangle = \langle \xi_2^*, (q^*)^-\rangle = 0. \tag{20j}$$

The “q.e.” appearing in (20h) and (20i) means *quasi-everywhere* and a statement holds “q.e.” if it holds everywhere except on a set of capacity zero. Let us observe that (20h) and (20i) imply

$$\langle \lambda^*, y^* - \Phi(y^*)\rangle = 0, \quad \langle \zeta^*, z^* - \Phi(z^*)\rangle = 0.$$

It is worth noting that if Assumption 2.19 is not available, it is still possible to show that a subset of the conditions above (called *weak C-stationarity*) are satisfied; see Proposition 7.4. Therein, (20h)–(20j) are missing. By assuming just the continuity of $(\cdot)^+ : V \to V$, we can further show some substitutes for the missing relations, see Proposition 7.5 and Lemma 7.7.

2.4. Examples

2.4.1. Obstacle map as solution map of PDE. It is illustrative to give an example that occurs commonly in applications and that satisfies all of the assumptions (for the subsolution and supersolution) that appear in this paper. Let Φ be increasing and satisfy $\Phi(0) \geq 0$ and let $F \in V_+^*$ be a given function. Define

$$\underline{u} := 0, \quad \overline{u} := A^{-1}F.$$

We define the set of source terms

$$W := \{g \in V^* : 0 \leq g \leq F\}.$$

With these choices, we in fact satisfy *every* assumption on the existence of sub- and supersolutions that is mentioned in the paper. We will prove this later in Lemma 3.8.

Regarding specific choices of the function spaces, we can take $\Omega \subset \mathbb{R}^n$ to be a bounded Lipschitz domain and set $V = H_0^1(\Omega)$. A concrete example of $\Phi : H \to V$ could be $\Phi(w) = \phi$ defined via

$$\begin{aligned} -\Delta\phi &= w && \text{in } \Omega, \\ \phi &= 0 && \text{on } \partial\Omega \end{aligned} \tag{21}$$

where $-\Delta : V \to V^*$ denotes the weak Laplacian. Clearly, $\Phi(0) = 0$. Regarding the operator A, we could take, for example, $A(u) = -\nabla \cdot (a\nabla u)$ where the coefficient $a : \Omega \to \mathbb{R}$ is a function satisfying $a \in L^\infty(\Omega)$ and $a \geq a_0 > 0$ almost everywhere for a constant a_0.

Further applications and examples of QVIs can be found in, for example, [5, 10].

2.4.2. An application in thermoforming. We consider now an application of our theory in thermoforming. Thermoforming is a manufacturing process in which mould shapes that are to be reproduced are forced into contact with heated membranes (which are typically plastic sheets): the membrane deforms and takes on the shape of the mould. In some circumstances, the ensuing heat exchange between the materials leads also to a deformation of the mould, giving rise to the QVI nature of the problem. For further details, we refer to [2]. We look at a concrete one-dimensional realisation from [1, §4.3], which is co-authored by two of the present authors. Let $D \equiv \Omega = (0,1)$, $A = -\Delta$, $V := H_0^1(\Omega)$, and consider the QVI

$$u \in V, u \leq \Phi(u), \quad \langle -\Delta u - f, u - v\rangle_{V^*,V} \leq 0, \quad \forall v \in V, v \leq \Phi(u), \tag{22}$$

for $f \in L^2(\Omega)$ defined by $f(x) = \pi^2 \sin(\pi x)$ and $\Phi(u)$ given by $\Phi(u) := \varphi T$ where $T \in H^1(\Omega)$ is the unique weak solution of

$$\begin{aligned} kT - \Delta T &= g(\psi T - u) && \text{in } \Omega, \\ \partial_\nu T &= 0 && \text{on } \partial\Omega, \end{aligned}$$

and where

$$\varphi(x) = \frac{10\pi^2 \sin(\pi x)}{5 - \cos(2\pi x)}, \quad k = \pi^2, \quad g(s) = 4\min(0,s)^2, \quad \psi(x) = \frac{5\pi^2 \sin(\pi x)}{5 - \cos(2\pi x)}.$$

Here, u refers to the displacement of the membrane, $\Phi(u)$ is the displacement of the mould and T is the temperature of the membrane. The above model is valid for one time step in the time discretisation of the thermoforming process; see again [2] for full details.

Regarding existence for (22), first note that the fact that g is decreasing implies that $u \mapsto T$ is increasing (the argument is the same as in [2, Lemma 6.3]) and hence as is the map Φ. Regarding $S(f, \cdot)$, since $f \in L^2(\Omega)$ with $f \geq 0$, it is easy to check that 0 is a subsolution and $A^{-1}f$ is a supersolution with $0 \leq A^{-1}f$, by non-negativity of f. By Proposition 1.2, QVI (22) possesses minimal and maximal solutions on $[0, A^{-1}f]$. It is not difficult to see that 0 is a solution of the QVI, and hence $\mathsf{m}(f) = 0$ for all non-negative $f \in L^2(\Omega)$. Note that (22) has the second explicit solution $\sin(\pi x)$; see [1, Lemma 4.3].

The assumption in (7) on the complete continuity of Φ follows from the compact embedding of V into $L^2(\Omega)$ and the inequality

$$\|\Phi(u_1) - \Phi(u_2)\|_V \leq \operatorname{Lip}(g)\big(\|\varphi\|_{L^\infty(\Omega)} k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)} k^{-1}\big)\|u_1 - u_2\|_{L^2(\Omega)},$$

which was shown in the proof of [1, Theorem 3.9].

The smallness condition in (8) on $B_{\varepsilon^*}(0)$ is satisfied thanks to the next result.

Lemma 2.21. *There exists an $\varepsilon^* > 0$ such that $\Phi : B_{\varepsilon^*}(0) \to V$ has a Lipschitz constant C_L satisfying $C_L < 1$.*

Proof. Let $B \subset V$ be a closed ball such that $B \subset \{v \in V \mid \|v\|_V < R^*\}$, where

$$R^* := \frac{3}{10(1+\pi)}\left(\sqrt{\frac{13\pi^2 + 8\pi}{80}} - \frac{\pi}{4}\right).$$

If we take $R < R^*$ and set

$$M_R = \frac{1}{2}R + \frac{10(1+\pi)}{3\pi}R^2, \tag{23}$$

then we have, using the notation $\mathrm{Lip}(g, I)$ to mean the Lipschitz constant of $g : I \to \mathbb{R}$ on the interval I,

$$\begin{aligned}\mathrm{Lip}(g, [-M_R, M_R])\big(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2} + \|\varphi'\|_{L^\infty(\Omega)}k^{-1}\big)\frac{1}{\pi}\\ = \frac{50}{3}\Big(R + \frac{20(1+\pi)}{3\pi}R^2\Big) < 1\end{aligned} \tag{24}$$

(see [1, Lemma 4.3] for the equality; the inequality holds because we took $R < R^*$ and is not difficult to see by using the quadratic formula). From the estimate in [1, Theorem 3.12],

$$\begin{aligned}&\|\Phi(u_1) - \Phi(u_2)\|_V\\ &\quad\le \frac{1}{\pi}\mathrm{Lip}(g, [-M_R, M_R])\big(\|\varphi\|_{L^\infty(\Omega)}k^{-1/2}\\ &\qquad + \|\varphi'\|_{L^\infty(\Omega)}k^{-1}\big)\|u_1 - u_2\|_V, \quad \forall u_1, u_2 \in B_R(0),\end{aligned}$$

and (24), it follows that there exists $\gamma_B \in [0, 1)$ such that

$$\|\Phi(u_1) - \Phi(u_2)\|_V \le \gamma_B\|u_1 - u_2\|_V, \quad \forall u_1, u_2 \in B. \tag{25}$$

This implies the claim. ∎

Let us now address (9). First we remark that Φ is Newton differentiable from V into $V \cap H^2(\Omega)$; see [1, Theorem 3.9(iii)]. Indeed, defining ξ by

$$\begin{aligned}k\xi - \Delta\xi - g'(\psi T - u)\psi\xi &= -g'(\psi T - u)h &&\text{in } \Omega,\\ \partial_\nu T &= 0 &&\text{on } \partial\Omega,\end{aligned}$$

we have that the Newton derivative of Φ is $G\Phi(u)(h) = \varphi\xi$. In fact, we can prove the following stronger result:

Lemma 2.22. *The map $\Phi : V \to V$ is continuously Fréchet differentiable.*

Proof. This relies on applying the implicit function theorem to the map $\mathcal{F} : V \times H^1(\Omega) \to H^1(\Omega)^*$ defined by $\mathcal{F}(u, T) := kT - \Delta T - g(\psi T - u)$, and is essentially the same as the proof of [2, Theorem 8], except with two differences. We modify the first step of the cited proof and show the Fréchet differentiability of g as follows. Using the mean value theorem, for $x, y \in \mathbb{R}$,

$$g(x + y) - g(x) - g'(x)y = \int_0^1 g'(x + (1-\lambda)y)y - g'(x)y \,\mathrm{d}\lambda,$$

and hence if we take now $v, d \in H^1(\Omega)$, using the Lipschitzness of g',

$$
\begin{aligned}
&\|g(v+d) - g(v) - g'(v)d\|_{H^1(\Omega)^*} \\
&= \sup_{w \in H^1(\Omega) \|w\|_{H^1(\Omega)}=1} \int_\Omega \Big(\int_0^1 g'(v + (1-\lambda)d)d - g'(v)d \,\mathrm{d}\lambda\Big) w \\
&\leq 8 \sup_{w \in H^1(\Omega) \|w\|_{H^1(\Omega)}=1} \int_\Omega \Big(\int_0^1 (1-\lambda)d^2 \,\mathrm{d}\lambda\Big) w \leq 8 \sup_{\substack{w \in H^1(\Omega) \\ \|w\|_{H^1(\Omega)}=1}} \int_\Omega d^2 w \\
&\leq 8\|d\|^2_{L^4(\Omega)} \sup_{w \in H^1(\Omega) \|w\|_{H^1(\Omega)}=1} \|w\|_{L^2(\Omega)} \\
&\leq C_1 \|d\|^2_{H^1(\Omega)} \qquad (\text{using } H^1(\Omega) \hookrightarrow L^4(\Omega))
\end{aligned}
$$

for some constant C_1; this shows that $g : H^1(\Omega) \to H^1(\Omega)^*$ is Fréchet differentiable. In the second step of the proof of [2, Theorem 8], we can use the Lipschitzness of g' (instead of the mean value theorem as utilised there) to show the continuity of $g' : H^1(\Omega) \to \mathcal{L}(H^1(\Omega), H^1(\Omega)^*)$. The rest of the proof follows as in [2, Theorem 8]. ∎

Now, if we take $W := [0, F]$ where $F \in V^*$ is such that $F \geq f$, and set $\underline{u} := 0$ and $\overline{u} := S(F, \infty) = T_\rho(F, \infty)$ as the solution of the unconstrained problem, we see that Assumption 2.7 is satisfied. Therefore, all of the results in Section 2 up to and including Section 2.2 are applicable.

3. Properties of the penalised problem

This section culminates in a result that shows the existence (in a constructive way) of extremal solutions to (4). To arrive at such a result, we first have to study some intermediary problems, which will also be of considerable use in later sections.

Recalling σ_ρ from (5), let us point out that $\sigma_\rho : V \to V^*$ is bounded (in the sense of non-linear operators), increasing, T-monotone, and hemicontinuous[4]. T-monotonicity and the fact that σ_ρ is increasing will be needed for the comparison results that are required for this paper. Note that the T-monotonicity condition implies monotonicity [19, Chapter 2, Lemma 2.1]. Another important property is the following, which shows that σ_ρ is indeed a penalty operator:

Lemma 3.1. *We have that*

$$z_\rho \rightharpoonup z \text{ in } V \text{ and } \sigma_\rho(z_\rho) \to 0 \text{ in } V^* \implies z \leq 0.$$

[4] T-monotonicity in the non-linear setting means $\langle \sigma_\rho(u) - \sigma_\rho(v), (u-v)^+ \rangle \geq 0$ and hemicontinuity means $s \mapsto \langle \sigma_\rho(u + sv), w \rangle$ is continuous for all $u, v, w \in V$.

Proof. First observe that for any $h \in H$, we have $\sigma_\rho(h) \to h^+$ in H. This is an immediate consequence of the estimate

$$0 \leq r^+ - \sigma_\rho(r) \leq \frac{\rho}{2}$$

(see [14, Lemma 2.1(iv)]). Suppose that $z_\rho \rightharpoonup z$ in V and $\sigma_\rho(z_\rho) \to 0$ in V^*. By monotonicity, we have for any $\lambda > 0$,

$$\langle \sigma_\rho(z_\rho) - \sigma_\rho(z + \lambda v), z_\rho - z - \lambda v \rangle \geq 0, \quad \forall v \in V.$$

Passing to the limit $\rho \searrow 0$ here using the strong convergence of $\sigma_\rho(z_\rho)$ and the fact that $\sigma_\rho(z + \lambda v) \to (z + \lambda v)^+$ in H, we obtain

$$\langle (z + \lambda v)^+, \lambda v \rangle \geq 0, \quad \forall v \in V.$$

Dividing through by λ and using (hemi)continuity of $(\cdot)^+ : H \to H$, we derive

$$\langle z^+, v \rangle \geq 0, \quad \forall v \in V.$$

The arbitrariness of v then implies that $z^+ = 0$. ∎

3.1. Results on a semilinear elliptic PDE

For $f \in V^*$ and $\varphi \in H$, consider the equation

$$Au + \frac{1}{\rho}\sigma_\rho(u - \Phi(\varphi)) = f, \tag{26}$$

the solution map of which we write as

$$u = T_\rho(f, \varphi),$$

so that $T_\rho : V^* \times H \to V$. Equation (26) has a unique solution (for fixed f and φ): the non-linearity is monotone, radially continuous, and bounded, giving pseudomonotonicity of the full elliptic operator by [21, Lemmas 2.9 and 2.11], whereas coercivity follows from

$$\langle Au, u - \Phi(\varphi) \rangle + \frac{1}{\rho}\langle \sigma_\rho(u - \Phi(\varphi)), u - \Phi(\varphi)) \geq C_a \|u\|_V^2 - C_b \|u\|_V \|\Phi(\varphi)\|_V,$$

leading to existence via [21, Theorem 2.6].

In the next two lemmas, we utilise the results of [29] to obtain Lipschitz estimates for T_ρ.

Lemma 3.2. *Assume that Φ is Lipschitz on $U \subset V$ with Lipschitz constant $C_L \geq 0$ satisfying*

$$C_L < \frac{C_a}{C_b} \quad \text{or} \quad A \text{ is self-adjoint and } C_L < 2\frac{\sqrt{C_b/C_a}}{1 + C_b/C_a}. \tag{27}$$

Then, there exist constants $C \geq 0$, $\tilde{c} \in [0, 1)$ (depending only on C_L, C_a, C_b, and the self-adjointness of A) such that for all $u, v \in V$ and $\varphi, \psi \in U$, we have

$$\langle A(u - v), u - \Phi(\varphi) - v + \Phi(\psi) \rangle \geq C\left(\|u - v\|_V^2 - \tilde{c}^2 \|\varphi - \psi\|_V^2\right).$$

Proof. This is precisely [29, Lemma 20]. Note that the linear and continuous operator A is a derivative of a convex function if and only if A is self-adjoint. ■

If the constant C_L is larger than or equal to the allowed threshold from Lemma 3.2, the result no longer holds (cf. [28, Theorems 3.6 and 3.7]). Note that the latter constant in (27) is larger than the former one. If $C_L < C_a/C_b$, we may choose

$$C = \frac{C_a}{2}, \quad \tilde{c} = \frac{C_b C_L}{C_a},$$

whereas in the other case we could choose

$$C = \frac{C_a C_b}{C_a + C_b}, \quad \tilde{c} = \frac{(C_a + C_b)C_L}{2\sqrt{C_a C_b}}.$$

The next result will be crucial, since it shows that the map $u \mapsto T_\rho(f, u)$ is a contraction under appropriate assumptions.

Proposition 3.3. *For all $f, g \in V^*$ and $\varphi, \psi \in H$, we have*

$$\|T_\rho(f,\varphi) - T_\rho(g,\psi)\|_V \le \sqrt{2}C_a^{-1}\|f-g\|_{V^*} + C_a^{-1}(\sqrt{2}C_b)\|\Phi(\psi) - \Phi(\varphi)\|_V.$$

In the case that $\Phi : V \to V$ is locally Lipschitz in $U \subset V$ with small Lipschitz constant C_L satisfying (27) and if $\varphi, \psi \in U$, then

$$\|T_\rho(f,\varphi) - T_\rho(g,\psi)\|_V \le \hat{C}\|f-g\|_{V^*} + \hat{c}\|\psi - \varphi\|_V$$

for some constants $\hat{C} \ge 0$, $\hat{c} \in [0, 1)$, depending only on C_L, C_a, C_b and the self-adjointness of A.

Proof. Let $u = T_\rho(f, \varphi)$ and $v = T_\rho(g, \psi)$. We have that

$$Au + \frac{1}{\rho}\sigma_\rho(u - \Phi(\varphi)) = f \quad \text{and} \quad Av + \frac{1}{\rho}\sigma_\rho(v - \Phi(\psi)) = g.$$

Testing the difference with $u - \Phi(\varphi) - v + \Phi(\psi)$ and using monotonicity leads to

$$\langle A(u-v), u - \Phi(\varphi) - v + \Phi(\psi)\rangle \le \langle f - g, u - \Phi(\varphi) - v + \Phi(\psi)\rangle \tag{28}$$

and, consequently,

$$\begin{aligned} C_a\|u-v\|_V^2 - C_b\|u-v\|_V\|\Phi(\varphi) - \Phi(\psi)\|_V \\ \le \|f-g\|_{V^*}(\|u-v\|_V + \|\Phi(\varphi) - \Phi(\psi)\|_V). \end{aligned}$$

Together with the estimates

$$C_b\|u-v\|_V\|\Phi(\varphi) - \Phi(\psi)\|_V \le \frac{C_a}{4}\|u-v\|_V^2 + \frac{C_b^2}{C_a}\|\Phi(\varphi) - \Phi(\psi)\|_V^2,$$

$$\|f-g\|_{V^*}\|u-v\|_V \le \frac{C_a}{4}\|u-v\|_V^2 + \frac{1}{C_a}\|f-g\|_{V^*}^2,$$

we get

$$\begin{aligned}\frac{1}{2}\|u-v\|_V^2 &\le \frac{C_b^2}{C_a^2}\|\Phi(\varphi)-\Phi(\psi)\|_V^2 + \frac{1}{C_a}\|f-g\|_{V^*}\|\Phi(\varphi)-\Phi(\psi)\|_V \\ &\quad + \frac{1}{C_a^2}\|f-g\|_{V^*}^2 \\ &\le \Big(\frac{C_b}{C_a}\|\Phi(\varphi)-\Phi(\psi)\|_V + \frac{1}{C_a}\|f-g\|_{V^*}\Big)^2.\end{aligned}$$

This shows the first estimate.

In order to arrive at the second estimate, we use Lemma 3.2 in (28) to obtain

$$C\big(\|u-v\|_V^2 - \tilde{c}^2\|\varphi-\psi\|_V^2\big) \le \|f-g\|_{V^*}(\|u-v\|_V + C_L\|\varphi-\psi\|_V).$$

Together with

$$\|f-g\|_{V^*}\|u-v\|_V \le C\frac{1-\tilde{c}^2}{2}\|u-v\|_V^2 + \frac{1}{2C(1-\tilde{c}^2)}\|f-g\|_{V^*}^2$$

we get

$$\begin{aligned}C\frac{1+\tilde{c}^2}{2}\|u-v\|_V^2 &\le \frac{1}{2C(1-\tilde{c}^2)}\|f-g\|_{V^*}^2 + \|f-g\|_{V^*}C_L\|\varphi-\psi\|_V \\ &\quad + C\tilde{c}^2\|\varphi-\psi\|_V^2 \\ &\le \Big(\Big(\frac{1}{\sqrt{2C(1-\tilde{c}^2)}} + \frac{C_L}{2\sqrt{C}\tilde{c}}\Big)\|f-g\|_{V^*} + \sqrt{C}\tilde{c}\|\varphi-\psi\|_V\Big)^2.\end{aligned}$$

This yields the claim. ■

The next lemma shows that the solution of the PDE converges to the solution of the associated VI.

Lemma 3.4. *For $f \in V^*$ and $\varphi \in H$, we have $T_\rho(f,\varphi) \to S(f,\varphi)$ in V as $\rho \searrow 0$.*

Proof. This is an extension of the classical penalty theory (see [12, Theorem 3.1] or [16, §5.3, Chapter 3]) to the varying σ_ρ setting given in [5]. More precisely, since σ_ρ is hemicontinuous (and hence radially continuous) and bounded, this follows by [5, Theorem 2.18]. ■

3.2. Order properties

In this section, we discuss various properties related to the partial order. The next lemma is fundamental: it will be used to show that (4) has minimal and maximal solutions.

Lemma 3.5. *The map $T_\rho(\cdot,\cdot) : V^* \times H \to V$ is increasing.*

Proof. Let $f \geq g$, $\varphi \geq \psi$ and consider $u = T_\rho(f, \varphi)$ and $v = T_\rho(g, \psi)$. Testing the equation for $v - u$ with $(v-u)^+$, we have

$$\langle A(v-u), (v-u)^+\rangle + \frac{1}{\rho}\langle \sigma_\rho(v - \Phi(\psi)) - \sigma_\rho(u - \Phi(\varphi)), (v-u)^+\rangle = \langle g - f, (v-u)^+\rangle.$$

Since $\Phi(\varphi) \geq \Phi(\psi)$, we have $v - \Phi(\psi) \geq v - \Phi(\varphi)$ and hence by the increasing property, $\sigma_\rho(v - \Phi(\psi)) \geq \sigma_\rho(v - \Phi(\varphi))$. This implies from above that

$$\langle A(v-u), (v-u)^+\rangle + \frac{1}{\rho}\langle \sigma_\rho(v - \Phi(\varphi)) - \sigma_\rho(u - \Phi(\varphi)), (v-u)^+\rangle \leq 0$$

and hence, using T-monotonicity, we get $(v-u)^+ = 0$ so that $v \leq u$. ■

Lemma 3.6. *We have*

$$\rho \leq \kappa \implies T_\rho(f, \varphi) \leq T_\kappa(f, \varphi).$$

Proof. Let $u_\rho = T_\rho(f, \varphi)$ and $u_\kappa = T_\kappa(f, \varphi)$. We have

$$A(u_\rho - u_\kappa) + \frac{1}{\rho}\sigma_\rho(u_\rho - \Phi(\varphi)) - \frac{1}{\kappa}\sigma_\kappa(u_\kappa - \Phi(\varphi)) = 0,$$

and we manipulate

$$\begin{aligned}
&\frac{1}{\rho}\sigma_\rho(u_\rho - \Phi(\varphi)) - \frac{1}{\kappa}\sigma_\kappa(u_\kappa - \Phi(\varphi)) \\
&\quad= \Big(\frac{1}{\rho} - \frac{1}{\kappa}\Big)\sigma_\rho(u_\rho - \Phi(\varphi)) + \frac{1}{\kappa}\big(\sigma_\rho(u_\rho - \Phi(\varphi)) - \sigma_\kappa(u_\kappa - \Phi(\varphi))\big) \\
&\quad= \Big(\frac{1}{\rho} - \frac{1}{\kappa}\Big)\sigma_\rho(u_\rho - \Phi(\varphi)) + \frac{1}{\kappa}\big(\sigma_\rho(u_\rho - \Phi(\varphi)) - \sigma_\rho(u_\kappa - \Phi(\varphi))\big) \\
&\qquad+ \frac{1}{\kappa}\big(\sigma_\rho(u_\kappa - \Phi(\varphi)) - \sigma_\kappa(u_\kappa - \Phi(\varphi))\big)
\end{aligned}$$

which, when tested with $(u_\rho - u_\kappa)^+$, is non-negative (the first term by $\rho \leq \kappa$, the second by T-monotonicity, and the third because σ_ρ satisfies $\rho \leq \kappa \implies \sigma_\rho \geq \sigma_\kappa$). ■

We should expect that the solution of the VI is dominated by the solution of the penalised equation.

Lemma 3.7. *We have $S(f, \varphi) \leq T_\rho(f, \varphi)$.*

Proof. Let $u_\rho = T_\rho(f, \varphi)$ and $v = S(f, \varphi)$. Take as test function in the VI for v the function $v - (v - u_\rho)^+$ and combine to get

$$\langle A(v - u_\rho), (v - u_\rho)^+\rangle - \frac{1}{\rho}\langle \sigma_\rho(u_\rho - \Phi(\varphi)), (v - u_\rho)^+\rangle \leq 0.$$

Since $v \le \Phi(\varphi)$, we have $u_\rho - \Phi(\varphi) \le u_\rho - v$ and the increasing property of σ_ρ as well as the fact that $\sigma_\rho \equiv 0$ on $(-\infty, 0]$ implies that

$$\langle \sigma_\rho(u_\rho - \Phi(\varphi)), (v - u_\rho)^+ \rangle \le \langle \sigma_\rho(u_\rho - v), (v - u_\rho)^+ \rangle \le 0.$$

Using this fact above, we deduce that $\langle A(v - u_\rho), (v - u_\rho)^+ \rangle \le 0$, which gives the claim. ■

Before we move on, let us prove, with the aid of a result from this section, a claim we made earlier in Section 2.4.1.

Lemma 3.8. *The example in Section 2.4.1 satisfies every assumption on sub- and supersolutions in the paper. More precisely, with W, $\underline{u}$, and $\overline{u}$ as defined in Section 2.4.1, any $f \in W$ and the set W satisfy Assumptions 1.1, 2.1, 2.7, 2.13, 2.17 (with $W = U_{\text{ad}}$), 3.9, and 4.6.*

Proof. It suffices to show that $\underline{u}$ and $\overline{u}$ are sub- and supersolutions for $S(f, \cdot)$ and $T_\rho(f, \cdot)$ for all $f \in W$. It is not difficult to see this:

- Since Φ is increasing, for all $\rho \ge 0$, we have $\overline{u} = T_\rho(F, \infty) \ge T_\rho(F, \overline{u}) \ge T_\rho(f, \overline{u})$ for any $f \le F$ because of Lemma 3.5 (for $\rho > 0$) and [20, §4:5, Theorem 5.1] (for $\rho = 0$). Thus, $\overline{u}$ is a supersolution of $S(f, \cdot)$ and $T_\rho(f, \cdot)$ for all $f \in W$.
- If $f \in W$, for all $\rho \ge 0$, we have $T_\rho(f, 0) \ge T_\rho(0, 0) = 0$ again by the above-cited results and since $f \ge 0$. Hence, $\underline{u}$ is a subsolution for $S(f, \cdot)$ and $T_\rho(f, \cdot)$ for all $f \in W$.

The proof is complete. ■

3.3. Minimal and maximal solutions of PDEs

Recall (4):

$$Au + \frac{1}{\rho}\sigma_\rho(u - \Phi(u)) = f.$$

Let us assume the existence of a sub- and supersolution for $T_\rho(f, \cdot)$ and prove our earlier claim that (4) has extremal solutions.

Assumption 3.9 (Well-definedness of $\mathsf{Z}_\rho(f)$). Given $f \in V^*$, assume that there exist $\underline{u}, \overline{u} \in V$ such that

$$\underline{u} \le T_\rho(f, \underline{u}), \quad \overline{u} \ge T_\rho(f, \overline{u}), \quad \text{and} \quad \underline{u} \le \overline{u}.$$

This assumption is exactly (11). Arguing like in Proposition 1.2, we have the following:

Proposition 3.10. *Under Assumption 3.9, there exist a minimal solution $\mathsf{m}_\rho(f)$ and maximal solution $\mathsf{M}_\rho(f)$ to equation (4) on $[\underline{u}, \overline{u}]$.*

Proof. Due to Lemma 3.5, it follows by the Birkhoff–Tartar theorem [6, §15.2, Proposition 2] that the set of fixed points of $u \mapsto T_\rho(f, u)$ is non-empty and possesses a minimal and maximal solution on the interval $[\underline{u}, \overline{u}]$. ■

Now, we focus on ways to approximate these extremal solutions by sequences.

Definition 3.11. Define the iterative sequence $\{\overline{u}_\rho^n\}$ by

$$\overline{u}_\rho^n = T_\rho(f, \overline{u}_\rho^{n-1}), \quad \overline{u}_\rho^0 = \overline{u},$$

and $\{\underline{u}_\rho^n\}$ by

$$\underline{u}_\rho^n = T_\rho(f, \underline{u}_\rho^{n-1}), \quad \underline{u}_\rho^0 = \underline{u}.$$

Note that $\{\overline{u}_\rho^n\}$ is a decreasing sequence and $\{\underline{u}_\rho^n\}$ is an increasing sequence (see the proof of the next result). As a matter of fact, $\overline{u}_\rho^n$ approaches $\mathsf{M}_\rho(f)$ from above and $\underline{u}_\rho^n$ approaches $\mathsf{m}_\rho(f)$ from below.

Proposition 3.12 (Strong convergence). *Under Assumption 3.9, assume (7) or that*

$$\Phi : V \to V \text{ is weakly sequentially continuous,} \tag{29}$$

$$V \overset{c}{\hookrightarrow} H. \tag{30}$$

Then

$$\overline{u}_\rho^n \searrow \mathsf{M}_\rho(f) \quad \text{and} \quad \underline{u}_\rho^n \nearrow \mathsf{m}_\rho(f) \qquad \text{strongly in } V \text{ as } n \to \infty.$$

Proof. For readability, let us write u^n instead of $\overline{u}_\rho^n$. Each u^n satisfies

$$Au^n + \frac{1}{\rho}\sigma_\rho(u^n - \Phi(u^{n-1})) = f.$$

By the definition of supersolution, $u^0 = \overline{u} \geq T_\rho(f, \overline{u}) = u^1$, and since we have shown above that $T_\rho(f, \cdot)$ is increasing, we obtain in this fashion that $u^n \geq u^{n+1}$, so that $\{u^n\}$ is a decreasing sequence.

Note also that $u^1 \geq T_\rho(f, \underline{u}) \geq \underline{u}$, and hence $u^n \geq \underline{u}$ for all n. Define $v_0 = \Phi(\underline{u})$. Then we have $v_0 \leq \Phi(u^n)$ for all n since Φ is increasing, therefore,

$$\langle \sigma_\rho(u^n - \Phi(u^{n-1})), u^n - v_0 \rangle = \langle \sigma_\rho(u^n - \Phi(u^{n-1})) - \sigma_\rho(v_0 - \Phi(u^{n-1})), u^n - v_0 \rangle \geq 0$$

by monotonicity. Testing the u^n equation with $u^n - v_0$,

$$C_a \|u^n\|_V^2 \leq \|f\|_{V^*}\|u^n\|_V + \|f\|_{V^*}\|v_0\|_V + C_b\|u^n\|_V\|v_0\|_V$$

and this leads to a uniform bound in V. Thus, $u^n \rightharpoonup u$ in V for some u, for the entire sequence by monotonicity (see, e.g., [5, Lemma 2.3]). Take any solution $u^* = T_\rho(f, u^*)$ with $u^* \leq u^0$. It follows that $u^* \leq u^1$ by applying $T_\rho(f, \cdot)$ to both sides. Likewise, $u^* \leq u^n$ and hence $u^* \leq u$, so if u is a solution of the limiting problem, it must be the largest solution. Let us show now that u does solve the limiting equation, that is, $u = T_\rho(f, u)$.

Satisfaction of the equation. Case 1. Under complete continuity (see (7)), making the transformation $w^n = u^n - \Phi(u^{n-1})$, we can write the equation for u^n as

$$Aw^n + \frac{1}{\rho}\sigma_\rho(w^n) = f - A\Phi(u^{n-1}).$$

Call the operator on the left-hand side $\widehat{A}$. By monotonicity, we have for all $v \in V$,

$$0 \le \langle \widehat{A}(w^n) - \widehat{A}(v), w^n - v\rangle = \langle f - A\Phi(u^{n-1}) - \widehat{A}(v), w^n - v\rangle,$$

and hence, noting that $w^n \rightharpoonup u - \Phi(u) =: w$ and $\Phi(u^{n-1}) \to \Phi(u)$ by (7) (observe that it suffices to have this complete continuity only for monotonic sequences),

$$0 \le \langle f - A\Phi(u) - \widehat{A}(v), w - v\rangle, \quad \forall v \in V.$$

Since $\widehat{A}$ is radially continuous, by Minty's trick [21, Lemma 2.13], we obtain $\widehat{A}(w) = f - A\Phi(u)$, that is,

$$Aw + \frac{1}{\rho}\sigma_\rho(w) = f - A\Phi(u).$$

Since $w = u - \Phi(u)$, we see that $u = T_\rho(f, u)$.

Case 2. Otherwise, by (29), the Lipschitz continuity of $\sigma_\rho : H \to H$, and the fact that $V \overset{c}{\hookrightarrow} H$, we obtain $\sigma_\rho(u^n - \Phi(u^{n-1})) \rightharpoonup \sigma_\rho(u - \Phi(u))$ in V^*. This lets us pass to the limit in the equation for u^n.

Strong convergence. It remains for us to show that $u_n \to u$ in V strongly.

Case 1. By using Lemma 3.3, we obtain the continuous dependence estimate

$$\|u^n - u\|_V \le C\|\Phi(u^{n-1}) - \Phi(u)\|_V,$$

and we can pass to the limit on the right-hand side using (7), yielding $u^n \to u$.

Case 2. In the second case, we test the equation for $u^n - u$ with $u^n - u$ and manipulate

$$C_a\rho\|u^n - u\|_V^2 \le \langle \sigma_\rho(u - \Phi(u)) - \sigma_\rho(u^n - \Phi(u^{n-1})), u^n - u\rangle_{H^*,H} \to 0$$

with the convergence because we have $\sigma_\rho(u^n - \Phi(u^{n-1})) \to \sigma_\rho(u - \Phi(u))$ in H^* by the compact embedding (see (30)), and $u^n - u \to 0$ in H for the same reason. ■

Remark 3.13. If we assume that $\Phi : H \to V$ is continuous, (30) implies (7). Since the aforementioned continuity of Φ and (30) typically do hold in examples, the above result is rather a powerful property that we attain without cost.

In some sense, the conclusion of Proposition 3.12 improves the similar convergence result of [5, Theorem 2.18] where it was shown that, in greater generality and in the absence of the assumption that Φ is increasing, solutions of (4) converge along a subsequence to some solution of QVI (1). Here, we are able to select precisely the minimal or maximal solution as the limiting objects thanks to the strengthened structure.

4. Convergence to the QVIs

We now consider the limiting behaviour of M_ρ and m_ρ as $\rho \searrow 0$ and show that they converge to the expected limits under some circumstance. First, we need some more properties.

4.1. Properties with respect to varying ρ

In the next lemma, we show that $\rho \mapsto \mathsf{Z}_\rho$ is increasing. In other words, Z_ρ shrinks as ρ gets smaller (this is natural, since we expect Z_ρ to converge to the solution of the constrained problem). Recall Definition 3.11.

Lemma 4.1. *Let $\rho, \kappa > 0$ and assume that $\underline{u}$ is a subsolution and $\overline{u}$ is a supersolution of both $T_\rho(f, \cdot)$ and $T_\kappa(f, \cdot)$ with $\underline{u} \leq \overline{u}$. If $\rho \leq \kappa$, then*

$$\overline{u}_\rho^n \leq \overline{u}_\kappa^n \quad \text{and} \quad \underline{u}_\rho^n \leq \underline{u}_\kappa^n.$$

Thus, if the assumptions of Proposition 3.12 hold, then

$$\mathsf{M}_\rho(f) \leq \mathsf{M}_\kappa(f) \quad \text{and} \quad \mathsf{m}_\rho(f) \leq \mathsf{m}_\kappa(f).$$

Proof. Set $\overline{u}_\rho := \mathsf{M}_\rho(f)$, $\overline{u}_\kappa := \mathsf{M}_\kappa(f)$. We have $\overline{u}_\rho^1 = T_\rho(f, \overline{u}) \leq T_\kappa(f, \overline{u}) = \overline{u}_\kappa^1$ by Lemma 3.6. Hence, $\overline{u}_\rho^2 = T_\rho(f, \overline{u}_\rho^1) \leq T_\rho(f, \overline{u}_\kappa^1) \leq T_\kappa(f, \overline{u}_\kappa^1) = \overline{u}_\kappa^2$ by the increasing property of Lemma 3.5 and again Lemma 3.6. The same holds when one replaces the supersolution by the subsolution. This implies the first claim and then taking $n \to \infty$, using Proposition 3.12 implies the second. ■

Remark 4.2. In the above lemma, we could consider two different pairs of sub/supersolution for T_ρ and T_κ. We can prove the same result (but Z_ρ and Z_κ would be defined on different intervals of course) if we assume the subsolution (supersolution) for the ρ problem is less than or equal to than the subsolution (supersolution) for the κ problem. We leave the details to the reader.

In a similar fashion to Definition 3.11, we introduce the following:

Definition 4.3. Define the sequences

$$\hat{u}^n = S(f, \hat{u}^{n-1}), \quad \hat{u}^0 = \overline{u},$$

and

$$\tilde{u}^n = S(f, \tilde{u}^{n-1}), \quad \tilde{u}^0 = \underline{u}.$$

The map S can be thought of as T_0 (i.e., T_ρ with $\rho = 0$).

Proposition 4.4 ([4, Lemmas 3.2 and 3.3]). *Under Assumption 1.1, assume (7). Then $\hat{u}^n \searrow \mathsf{M}(f)$ and $\tilde{u}^n \nearrow \mathsf{m}(f)$ in V.*

This proposition corresponds to the convergence results of Proposition 3.12 for the $\rho = 0$ case.

The next lemma shows that the unconstrained iterates (which solve PDEs) are greater than the constrained iterates which solve VIs).

Lemma 4.5. *If $\{\overline{u}^n_\rho\}, \{\hat{u}^n\}$ and $\{\underline{u}^n_\rho\}, \{\tilde{u}^n\}$ are defined as above with the same initial elements $\overline{u}$ and $\underline{u}$ respectively, then $\overline{u}^n_\rho \geq \hat{u}^n$ and $\underline{u}^n_\rho \geq \tilde{u}^n$.*

Proof. This is essentially Lemma 4.1 with $\rho = 0$.

We have, using Lemma 3.7 , $\overline{u}^1_\rho = T_\rho(f, \overline{u}) \geq S(f, \overline{u}) = \hat{u}_1$, and hence $\overline{u}^2_\rho = T_\rho(f, \overline{u}^1_\rho) \geq S(f, \overline{u}^1_\rho) \geq S(f, \hat{u}_1) = \hat{u}_2$, and so on. Here, we have used the increasing property of $S(f, \cdot)$. The same applies with the supersolution replacing the subsolution. ■

4.2. The $\rho \searrow 0$ limit

We want to prove that the penalised extremal solutions converge to the (non-penalised) extremal solutions in the limit $\rho \searrow 0$. First, we have to guarantee that all these objects exist.

Assumption 4.6 (Well-definedness of $\mathsf{Z}(f)$ and $\mathsf{Z}_\rho(f)$ for all ρ sufficiently small). Given $f \in V^*$, assume that there exist $\underline{u}, \overline{u} \in V$ and $\rho_0 > 0$ such that

$$\underline{u} \leq \overline{u}, \quad \underline{u} \leq S(f, \underline{u}), \quad \overline{u} \geq T_{\rho_0}(f, \overline{u}).$$

Remark 4.7. The statements

$$\underline{u} \leq S(f, \underline{u})$$

and

$$\underline{u} \leq T_\rho(f, \underline{u}), \quad \forall \rho \leq \rho_0$$

are equivalent. One direction follows from the convergence result (as $\rho \to 0$) of Lemma 3.4 and the other from Lemma 3.7.

Under this assumption, $\underline{u} \leq S(f, \underline{u}) \leq T_\rho(f, \underline{u})$ for all ρ (see the above remark), and

$$\overline{u} \geq T_{\rho_0}(f, \overline{u}) \geq T_\rho(f, \overline{u}) \geq S(f, \overline{u}), \quad \forall \rho \leq \rho_0$$

so that $(\underline{u}, \overline{u})$ are a sub- and supersolution pair for both $T_\rho(f, \cdot)$ (for all $\rho \leq \rho_0$) and $S(f, \cdot)$. This means that both Assumption 1.1 and (11) (i.e., Assumption 3.9) are satisfied and both $\mathsf{Z}_\rho(f)$ and $\mathsf{Z}(f)$ are well-defined objects in $[\underline{u}, \overline{u}]$, for all $\rho \leq \rho_0$.

Theorem 4.8. *Let Assumption 4.6 and the weak sequential continuity (see (29)) hold. Then $\mathsf{M}_\rho(f) \searrow \mathsf{M}(f)$ weakly in V and $\mathsf{m}_\rho(f) \searrow u$ weakly in V, where $u \in V$ is a solution of (1). If (7) holds, then the convergences are strong.*

Proof. As mentioned above, $\mathsf{M}(f)$ and $\mathsf{M}_\rho(f)$ exist for all $0 < \rho \leq \rho_0$ by Assumption 4.6. Define $u_\rho := \mathsf{M}_\rho(f)$. Since $\underline{u} \leq S(f, \underline{u}) \leq \Phi(\underline{u}) \leq \Phi(u_\rho)$ for all ρ, if we set $v_0 := \underline{u}$, we can test the u_ρ equation with $u_\rho - v_0$ and we get that $\{u_\rho\}$ is bounded by using

$$\langle \sigma_\rho(u_\rho - \Phi(u_\rho)), u_\rho - v_0 \rangle = \langle \sigma_\rho(u_\rho - \Phi(u_\rho)) - \sigma_\rho(v_0 - \Phi(u_\rho)), u_\rho - v_0 \rangle \geq 0.$$

Hence, $u_\rho \rightharpoonup u$ in V for a subsequence that we have relabelled. This implies that $\rho(f - Au_\rho) = \sigma_\rho(u_\rho - \Phi(u_\rho)) \to 0$ in V^*. Then, testing the equation for u_ρ with $u_\rho - v$ for $v \in V$ and using the monotonicity formula

$$\langle \sigma_\rho(u_\rho - \Phi(u_\rho)), u_\rho - v \rangle \geq \langle \sigma_\rho(v - \Phi(u_\rho)), u_\rho - v \rangle, \quad \forall v \in V \tag{31}$$

(this follows by the monotonicity of σ_ρ; see the proof of [5, Theorem 2.18]), we have

$$\langle Au_\rho, u_\rho \rangle + \frac{1}{\rho} \langle \sigma_\rho(v - \Phi(u_\rho)), u_\rho - v \rangle \leq \langle f, u_\rho - v \rangle + \langle Au_\rho, v \rangle.$$

Let $v \in V$ be such that $v \leq \Phi(u)$. Since $u \leq u_\rho$ (as Lemma 4.1 shows that $\{u_\rho\}$ is a decreasing sequence), $v \leq \Phi(u_\rho)$, and hence the second term on the left disappears. We can pass to the limit and use weak lower semicontinuity to obtain that u solves the expected inequality. By (29), $u_\rho - \Phi(u_\rho) \rightharpoonup u - \Phi(u)$ in V, which, in conjunction with the fact that $\sigma_\rho(u_\rho - \Phi(u_\rho)) \to 0$, implies by Lemma 3.1 that $u \leq \Phi(u)$, so that u solves (1).

We also have, using Lemma 4.5 for the first inequality and with the limits below being weak,

$$u_\rho = \mathsf{M}_\rho(f) = \lim_n \overline{u}^n_\rho \geq \lim_n \hat{u}^n \geq \mathsf{M}(f)$$

(recall $\hat{u}^n = S(f, \hat{u}^{n-1})$ is defined above) where we used Lemma 4.4 for the final inequality. Passing to weak limit in ρ proves the result.

For strong convergence, we begin by defining $v_\rho := u + \Phi(u_\rho) - \Phi(u)$, which satisfies

$$v_\rho \to u \text{ in } V, \quad v_\rho \leq \Phi(u_\rho), \quad u_\rho - v_\rho = (u_\rho - u) + (\Phi(u) - \Phi(u_\rho)) \rightharpoonup 0 \text{ in } V,$$

with the first part holding thanks to (7). Testing the equation for u_ρ with $u_\rho - v_\rho$, we have

$$\langle A(u_\rho - v_\rho), u_\rho - v_\rho \rangle = \langle f, u_\rho - v_\rho \rangle - \frac{1}{\rho} \langle \sigma_\rho(u_\rho - \Phi(u_\rho)), u_\rho - v_\rho \rangle - \langle Av_\rho, u_\rho - v_\rho \rangle,$$

and to this we apply the monotonicity formula (see (31)) and coercivity of A to find

$$\begin{aligned} C_a \|u_\rho - v_\rho\|_V^2 &\leq \langle f, u_\rho - v_\rho \rangle - \frac{1}{\rho} \langle \sigma_\rho(v_\rho - \Phi(u_\rho)), u_\rho - v_\rho \rangle - \langle Av_\rho, u_\rho - v_\rho \rangle \\ &= \langle f, u_\rho - v_\rho \rangle - \langle Av_\rho, u_\rho - v_\rho \rangle \quad (\text{since } v_\rho \leq \Phi(u_\rho)). \end{aligned}$$

The right-hand side converges to zero, and hence $u_\rho - v_\rho \to 0$ strongly in V, implying $u_\rho \to u$.

The claim for the minimal solution follows similar lines to the above to deduce that the weak limit u is a solution to (1). ∎

Note that we are not (yet) able to identify u above as the minimal solution and prove that $\mathsf{m}_\rho(f) \rightharpoonup \mathsf{m}(f)$ in the above theorem under such general circumstances. It appears difficult to do this because $\underline{u}_\rho^n \geq \tilde{u}^n$ (by virtue of $\underline{u}_\rho^n$ being a solution of the unconstrained problem) and thus in the limit we do not obtain anything useful. This has been an open problem as identified in [8, Chapter 4, Remark 1.4]. But we can identify the desired limit with a contractive argument under different assumptions – see Theorem 2.9, which we will prove below.

4.3. The $\rho \searrow 0$ limit under a contraction assumption

By assuming the small Lipschitz constant assumption in (8), we can prove the convergence result that we wanted. It is convenient to introduce the following notation: considering the cases $\mathsf{Z} = \mathsf{m}$ or $\mathsf{Z} = \mathsf{M}$, we define the mappings $Z^n : \mathbb{R}^+ \cup \{0\} \times V^* \to V$ via

$$Z^n(\rho, g) := T_\rho(g, Z^{n-1}(\rho, g)), \quad Z^0(\rho, g) := \begin{cases} \underline{u} & \text{if } \mathsf{Z} = \mathsf{m}, \\ \overline{u} & \text{if } \mathsf{Z} = \mathsf{M}, \end{cases}$$

where $\underline{u}, \overline{u}$ are given and independent of ρ and g (to be fixed later). For convenience, we define $T_0(g, \varphi) := S(g, \varphi)$. With this, note that $Z^n(0, g) = S(g, Z^{n-1}(0, g))$.

Lemma 4.9. *If $\Phi : V \to V$ is continuous, then for each n, the map $Z^n : \mathbb{R}^+ \times V^* \to V$ is continuous at all points from the set $\{0\} \times V^*$.*

Proof. We show this by induction. It is clear for $n = 0$, as Z^0 is constant in its arguments. Assume that $Z^n(\rho, g) \to Z^n(0, f)$ as $(\rho, g) \to (0, f)$. The inductive step is:

$$\begin{aligned} Z^{n+1}(\rho, g) - Z^{n+1}(0, f) &= T_\rho(g, Z^n(\rho, g)) - T_0(f, Z^n(0, f)) \\ &= [T_\rho(g, Z^n(\rho, g)) - T_\rho(f, Z^n(0, f))] \\ &\quad + [T_\rho(f, Z^n(0, f)) - T_0(f, Z^n(0, f))]. \end{aligned}$$

The first bracket is continuous, due to Lemma 3.3 and the induction hypothesis, and the second bracket is continuous due to Lemma 3.4. ■

Lemma 4.10. *Assume that $\Phi : V \to V$ is continuous and $f \in V^*$ is such that (8) and*

$$Z^n(0, f) \to \mathsf{Z}(f) \tag{32}$$

hold. Then for every $\varepsilon > 0$, there exist $N \in \mathbb{N}$ and $\rho_0, \delta > 0$ such that

$$Z^n(\rho, g) \in B_\varepsilon(\mathsf{Z}(f)), \quad \forall n \geq N, \rho \in [0, \rho_0], g \in B_\delta(f). \tag{33}$$

Regarding the above assumptions, note that we are implicitly assuming that $\mathsf{Z}(f)$ is defined (it would be true if f, $\underline{u}$, and $\overline{u}$ satisfy Assumption 1.1; see Lemma 4.4).

Proof. Without loss of generality, we assume $\varepsilon \le \varepsilon^*$. Let us first record some useful estimates. Due to (32), we get $N \in \mathbb{N}$ such that $n \ge N$ gives $Z^n(0, f) \in B_{\varepsilon/2}(\mathsf{Z}(f))$. For $g \in V^*$ and $\rho \ge 0$, we have

$$\begin{aligned}\|Z^N(\rho, g) - \mathsf{Z}(f)\|_V &\le \|Z^N(\rho, g) - Z^N(0, f)\|_V + \|Z^N(0, f) - \mathsf{Z}(f)\|_V \\ &\le \|Z^N(\rho, g) - Z^N(0, f)\|_V + \frac{\varepsilon}{2}.\end{aligned}$$

The function Z^N is continuous at $(0, f)$ due to the previous lemma. This, together with Lemma 3.4, means that we can choose $\rho_0 > 0$ and $\delta > 0$ such that

$$\begin{aligned} Z^N(\rho, g) &\in B_\varepsilon(\mathsf{Z}(f)), && \forall \rho \in [0, \rho_0], g \in B_\delta(f), \\ \|T_\rho(f, \mathsf{Z}(f)) - T_0(f, \mathsf{Z}(f))\|_V &\le \frac{1 - \hat{c}}{2}\varepsilon, && \forall \rho \in [0, \rho_0], \\ \delta &\le \frac{1 - \hat{c}}{1 + \hat{C}} \frac{\varepsilon}{2}. \end{aligned}$$

Here, the constants $\hat{C} \ge 0$ and $\hat{c} \in [0, 1)$ are chosen as in Lemma 3.3.

By induction over n, we show that (33) holds. To this end, suppose that $n \ge N$, $\rho \in [0, \rho_0]$, and $g \in B_\delta(f)$ are given such that $Z^n(\rho, g) \in B_\varepsilon(\mathsf{Z}(f))$. Using the definition of Z^{n+1}, we find

$$\begin{aligned}\|Z^{n+1}(\rho, g) - \mathsf{Z}(f)\|_V &= \|T_\rho(g, Z^n(\rho, g)) - T_0(f, \mathsf{Z}(f))\|_V \\ &\le \|T_\rho(g, Z^n(\rho, g)) - T_\rho(f, \mathsf{Z}(f))\|_V + \|T_\rho(f, \mathsf{Z}(f)) - T_0(f, \mathsf{Z}(f))\|_V \\ &\le \|T_\rho(g, Z^n(\rho, g)) - T_\rho(f, \mathsf{Z}(f))\|_V + \frac{1 - \hat{c}}{2}\varepsilon.\end{aligned}$$

Since $Z^n(\rho, g) \in B_\varepsilon(\mathsf{Z}(f)) \subset B_{\varepsilon^*}(\mathsf{Z}(f))$, we can apply Lemma 3.3 and get

$$\begin{aligned}\|Z^{n+1}(\rho, g) - \mathsf{Z}(f)\|_V &\le \hat{C}\|f - g\|_{V^*} + \hat{c}\|Z^n(\rho, g) - \mathsf{Z}(f)\|_V + \frac{1 - \hat{c}}{2}\varepsilon \\ &\le \hat{C}\delta + \hat{c}\varepsilon + \frac{1 - \hat{c}}{2}\varepsilon \le \varepsilon.\end{aligned}$$

This shows $Z^{n+1}(\rho, g) \in B_\varepsilon(\mathsf{Z}(f))$. By induction, (33) follows. ■

The important point in the previous result is that (33) holds for $\rho \le \rho_0$ and $g \in B_\delta(f)$ uniformly in n.

Let us now prove the theorem on the convergence of $\mathsf{Z}_\rho(g) \to \mathsf{Z}(f)$.

Proof of Theorem 2.8. We take $\underline{u}$ and $\overline{u}$ in the definition of Z^0 to satisfy Assumption 2.7. By Assumption 2.7, Assumption 4.6 is satisfied for every source term in $B_{\overline{\delta}}(f) \cap W$ and we get that Z and Z_ρ are well defined on the set $B_{\overline{\delta}}(f) \cap W$ for small ρ.

Step 1. By Lemma 4.10 (note that Φ is continuous by (7)), there exist $N \in \mathbb{N}$ and $\rho_0, \delta > 0$ such that $Z^n(\rho, g) \in B_{\varepsilon^*/2}(\mathsf{Z}(f))$ as long as $n \ge N$, $\rho \le \rho_0$, and $g \in B_\delta(f)$.

Without loss of generality, we can assume that $\delta \le \widehat{\delta}$. For $g \in B_{\delta}(f)$, let us define the sequence

$$y^n := T_\rho(g, y^{n-1}), \quad y^0 := Z^N(\rho, g).$$

that is, $y^n = Z^{N+n}(\rho, g)$. As noted, we have $y^0 \in B_{\varepsilon^*/2}(\mathsf{Z}(f))$ under the stated conditions on ρ and g. We claim that $T_\rho(g, \cdot) : B_{\varepsilon^*/2}(\mathsf{Z}(f)) \to B_{\varepsilon^*/2}(\mathsf{Z}(f))$ is a contraction for sufficiently small ρ and g sufficiently close to f. Indeed, take $\varphi \in B_{\varepsilon^*/2}(\mathsf{Z}(f))$, $g \in B_{\delta^*}(f)$ where

$$\delta^* = \frac{(1-\widehat{c})\varepsilon^*}{4(\widehat{C}+1)},$$

and take ρ small enough (let us say $\rho \le \rho_1$) so that $T_\rho(f, \mathsf{Z}(f)) \in B_{\widehat{C}\delta^*}(\mathsf{Z}(f))$ (this is possible by Lemma 3.4). Then using Lemma 3.3 and (8),

$$\begin{aligned}\|T_\rho(g,\varphi) - \mathsf{Z}(f)\|_V &\le \|T_\rho(g,\varphi) - T_\rho(f, \mathsf{Z}(f))\|_V + \|T_\rho(f, \mathsf{Z}(f)) - \mathsf{Z}(f)\|_V \\ &\le \widehat{C}\|g - f\|_{V^*} + \widehat{c}\|\varphi - \mathsf{Z}(f)\|_V + \frac{(1-\widehat{c})\varepsilon^*}{4} \\ &\le \frac{(1-\widehat{c})\varepsilon^*}{4} + \frac{\widehat{c}\varepsilon^*}{2} + \frac{(1-\widehat{c})\varepsilon^*}{4} = \frac{\varepsilon^*}{2}\end{aligned}$$

so that $T_\rho(g, \cdot)$ is invariant on the ball in question. For the contraction property, due to Lemma 3.3 and (8),

$$\|T_\rho(g,\varphi) - T_\rho(g,\psi)\|_V \le \widehat{c}\|\varphi - \psi\|_V, \quad \forall \varphi, \psi \in B_{\varepsilon^*/2}(\mathsf{Z}(f))$$

where $\widehat{c} \in [0, 1)$.

Hence, by Banach's fixed point theorem, we obtain $y^n \to y$ in V where $y = T_\rho(g, y)$. That is, $Z^n(\rho, g) \to Z^\infty(\rho, g)$ for some $Z^\infty(\rho, g)$. Furthermore, we have

$$\|y^n - y\|_V \le \widehat{c}\|y^{n-1} - y\|_V,$$

which implies

$$\|y^n - y\|_V \le \widehat{c}^n\|y_0 - y\|_V,$$

that is,

$$\|Z^{N+n}(\rho, g) - Z^\infty(\rho, g)\|_V \le \widehat{c}^n\|Z^N(\rho, g) - Z^\infty(\rho, g)\|_V \le \varepsilon\widehat{c}^n$$

since $Z^N(\rho, g), Z^\infty(\rho, g) \in B_{\varepsilon/2}(\mathsf{Z}(f))$. Recall that the above holds as long as $n \ge N$, $\rho \le \min(\rho_0, \rho_1)$ and $g \in B_{\min(\delta,\delta^*)}(f)$.

We can rewrite this as

$$\|Z^n(\rho, g) - Z^\infty(\rho, g)\|_V \le \varepsilon\widehat{c}^{n-N}$$

for $n \ge 2N$, $\rho \le \min(\rho_0, \rho_1)$, and $g \in B_{\min(\delta,\delta^*)}(f)$, where we note that the right-hand side of the inequality is independent of ρ and g.

Step 2. Assumption 2.7 implies that Assumption 3.9 is satisfied for all $g \in B_{\bar{\delta}}(f) \cap W$ and small enough ρ, and thus for g taken in $B_{\bar{\delta}}(f) \cap W$, we can apply Proposition 3.12, which allows us to identify $Z^\infty(\rho, g) = \mathsf{Z}_\rho(g)$.

Conclusion. To summarise, we have shown

$$Z^n(\rho, g) \to \mathsf{Z}_\rho(g) \quad \text{uniformly in } \rho,\, g \in B_\delta(f) \cap W \text{ as } n \to \infty,$$

while from Lemma 4.9, we have

$$Z^n(\rho, g) \to Z^n(0, f) \quad \text{as } \rho \searrow 0,\, g \to f.$$

Thus, we can interchange the iterated limits and get (the limit $g \to f$ below should be understood for $g \in W$)

$$\lim_{\substack{\rho\searrow 0\\ g\to f}} \mathsf{Z}_\rho(g) = \lim_{\substack{\rho\searrow 0\\ g\to f}} \lim_{n\to\infty} Z^n(\rho, g) = \lim_{n\to\infty} \lim_{\substack{\rho\searrow 0\\ g\to f}} Z^n(\rho, g) = \lim_{n\to\infty} Z^n(0, f) = \mathsf{Z}(f),$$

which concludes the proof. ■

By taking $\underline{u}, \overline{u}$ in the definition of Z^0 to satisfy Assumption 4.6 (rather than Assumption 2.7) and arguing similarly to above, we obtain Theorem 2.9.

Remark 4.11. Examining Sections 3 and 4, we see that there is a constructive way to approach the minimal and maximal solutions: we start at a subsolution or a supersolution, solve iteratively to get $\underline{u}^n_\rho$ or $\overline{u}^n_\rho$ (see Definition 3.11) for a large n, take ρ small and we will be close to the minimal solution or the maximal solution, thanks to the results of Proposition 3.12 and either Theorem 4.8 or, in the case of the maximal solution, Theorem 2.9. This can be useful for numerical realisations.

5. Local Lipschitz continuity of Z and Z_ρ

Local Lipschitz continuity for these maps does not immediately follow from the continuous dependence estimate of Lemma 3.3 if we impose only the local Lipschitz condition of Φ (as in the statement of the result below), since we do not know a priori that (even if f and g are close enough) $\mathsf{Z}_\rho(f)$ and $\mathsf{Z}_\rho(g)$ are in the neighbourhood where Φ is Lipschitz with a small Lipschitz constant. Instead, we have to argue using the results of the above section.

Proof of Theorem 2.11. Since by Theorem 2.8, $\mathsf{Z}_\rho(g) \in B_{\varepsilon^*}(\mathsf{Z}(f))$ for all $g \in W$ sufficiently close to f and ρ sufficiently small, we obtain via Lemma 3.3 and (8), for $\hat{c} < 1$, the estimate

$$\|\mathsf{Z}_\rho(f) - \mathsf{Z}_\rho(g)\|_V \le \hat{C}\|f - g\|_{V^*} + \hat{c}\|\mathsf{Z}_\rho(f) - \mathsf{Z}_\rho(g)\|_V. \qquad ■$$

Regarding Lipschitz continuity for Z, a first thought might be that we could pass to the limit in ρ in the inequality of Theorem 2.11, but the assumptions with respect to T_ρ would still be needed with that approach. We argue differently.

Proof of Theorem 2.3. The idea is that if we had $\mathsf{Z}(g) \in B_{\varepsilon^*}(\mathsf{Z}(f))$ for g sufficiently close to f, we can, like in the above proof, once again apply Lemma 3.3 (with $\rho = 0$) and the smallness assumption in (8) to obtain the result.

Thus, we need the result of Theorem 2.8 for $\rho = 0$ (without any assumptions on σ_ρ or other ρ-dependent quantities. This can be achieved by simply noting that the arguments of Section 4.3 still hold with $\rho = 0$ and with Assumption 2.7 replaced by Assumption 2.1. The proofs of the results can be modified in the obvious way, but let us point out that in the proof of Theorem 2.8, we need to use Lemma 4.4 in place of Proposition 3.12 (observe that (6) implies Assumption 1.1). ■

It is worth noting that the Lipschitz constants in Theorems 2.3 and 2.11 are both exactly

$$\frac{\hat{C}}{1-\hat{c}},$$

with $\hat{c}$ and $\hat{C}$ given in Lemma 3.3.

6. Directional differentiability

In this section, we shall prove that Z_ρ and Z are directionally differentiable maps (and also Hadamard differentiable in a certain sense). Our line of attack is based on the iteration approach from [2] (where we approximate the QVI solutions by a sequence of solutions of VIs, derive an expansion formula for the elements of the sequence and then pass to the limit) combined with some refinements from [29]. We start with the analysis for Z_ρ.

6.1. Differentiability for Z_ρ

An essential task is to obtain differentiability for T_ρ in its arguments. In the equation defining T_ρ, observe that the non-linearity $\sigma_\rho : V \to V^*$ is Hadamard differentiable and the derivative is bounded in the direction: in fact, when seen as a real-valued function, σ_ρ is C^1, and by using the Lipschitzness and boundedness of σ_ρ', we have that $\sigma_\rho : V \to V^*$ is Gâteaux differentiable [13, Theorem 8] (thus, it is also Hadamard differentiable, since σ_ρ is Lipschitz). We use this fact below.

Lemma 6.1. *Let $f \in V^*$ and assume that Φ is directionally differentiable at $\varphi \in V$. Then $T_\rho : V^* \times V \to V$ is directionally differentiable at (f, φ), that is,*

$$\lim_{s \searrow 0} \frac{T_\rho(f + sd, \varphi + sh) - T_\rho(f, \varphi)}{s} = T_\rho'(f, \varphi)(d, h) \quad \text{for } d \in V^* \text{ and } h \in V,$$

where $T'_\rho(f,\varphi)(d,h) = \delta$ *is the unique solution of the equation*

$$A\delta + \frac{1}{\rho}\sigma'_\rho(u - \Phi(\varphi))(\delta - \Phi'(\varphi)(h)) = d. \tag{34}$$

Proof. First, it is easy to see that (34) has a unique solution: if we make a transformation $\widehat{\delta} = \delta - \Phi'(\varphi)(h)$, we have

$$A\widehat{\delta} + \frac{1}{\rho}\sigma'_\rho(u - \Phi(\varphi))(\widehat{\delta}) = d - A\Phi'(\varphi)(h)$$

and this is uniquely solvable by the Lax–Milgram lemma because the linear operator $A + \frac{1}{\rho}\sigma'_\rho(u - \Phi(\varphi))$ is coercive and bounded[5].

Let $y := T_\rho(f + sd, \varphi + sh)$, $u := T_\rho(f,\varphi)$ and define δ as the solution of (34). Let us make the transformation $\widehat{y} = y - \Phi(\varphi + sh)$, $\widehat{u} = u - \Phi(\varphi)$ and (as above) $\widehat{\delta} = \delta - \Phi'(\varphi)(h)$ so that

$$A\widehat{y} + \frac{1}{\rho}\sigma_\rho(\widehat{y}) = f + sd - A\Phi(\varphi + sh), \quad A\widehat{u} + \frac{1}{\rho}\sigma_\rho(\widehat{u}) = f - A\Phi(\varphi),$$
$$A\widehat{\delta} + \frac{1}{\rho}\sigma'_\rho(\widehat{u})(\widehat{\delta}) = d - A\Phi'(\varphi)(h).$$

Multiplying the last equation by s, subtracting the latter two equations from the first, and adding and subtracting $\rho^{-1}\sigma_\rho(\widehat{u} + s\widehat{\delta})$, we obtain

$$\begin{aligned}A(\widehat{y} - \widehat{u} - s\widehat{\delta}) &+ \frac{1}{\rho}(\sigma_\rho(\widehat{y}) - \sigma_\rho(\widehat{u} + s\widehat{\delta})) + \frac{1}{\rho}(\sigma_\rho(\widehat{u} + s\widehat{\delta}) - \sigma_\rho(\widehat{u}) - s\sigma'_\rho(\widehat{u})(\widehat{\delta}))\\ &= -Al_s(\varphi,h),\end{aligned}$$

where l_s is the remainder term associated to Φ. The above is, using the fact that σ_ρ is directionally differentiable,

$$A(\widehat{y} - \widehat{u} - s\widehat{\delta}) + \frac{1}{\rho}(\sigma_\rho(\widehat{y}) - \sigma_\rho(\widehat{u} + s\widehat{\delta})) + \frac{1}{\rho}o_s^m(\widehat{u},\widehat{\delta}) = -Al_s(\varphi,h),$$

where o_s^m denotes the remainder term of σ_ρ. Testing with $\widehat{y} - \widehat{u} - s\widehat{\delta}$ and using monotonicity,

$$C_a\|\widehat{y} - \widehat{u} - s\widehat{\delta}\|_V \le \frac{1}{\rho}\|o_s^m(\widehat{u},\widehat{\delta})\|_{V^*} + C_b\|l_s(\varphi,h)\|_V.$$

[5] In general, if σ_ρ is directionally differentiable at $w \in V$, then $\sigma'_\rho(w) : V \to V^*$ is a monotone operator; this follows from

$$\begin{aligned}\langle\sigma'_\rho(w)(a) - \sigma'_\rho(b), a - b\rangle &= \frac{1}{s}\langle\sigma_\rho(w + sa) - o^m(s,a) - \sigma_\rho(w + sb) + o^m(s,b), a - b\rangle\\ &\ge \frac{1}{s}\langle o^m(s,b) - o^m(s,a), a - b\rangle.\end{aligned}$$

Now note that $\widehat{y} - \widehat{u} - s\widehat{\delta} = y - u - s\delta - l_s(\varphi, h)$ so that

$$C_a\|y - u - s\delta\|_V \leq \frac{1}{\rho}\|o_s^m(u - \Phi(\varphi), \delta - \Phi'(\varphi)(h))\|_{V^*} + (C_a + C_b)\|l_s(\varphi, h)\|_V.$$

Dividing by s and sending $s \to 0$ proves the result. ■

Suppose that we are

$$\text{given } f \in V^* \text{ and a set } W \subseteq V^* \text{ satisfying Assumption 2.7}$$

so that $\mathsf{Z}_\rho(g)$ and $\mathsf{Z}(g)$ are well defined for all $g \in B_{\bar{\delta}}(f) \cap W$ and sufficiently small ρ. Let $d \in \mathcal{T}_W(f)$, so there exist $\{d_k\}$ with $d_k \to d$ in V^* and $\{s_k\}$ with $s_k \searrow 0$ such that $f + s_k d_k \in W$. Define

$$u_n^k := T_\rho(f + s_k d_k, u_{n-1}^k), \quad u_0^k := \mathsf{Z}_\rho(f).$$

For convenience, let us also define

$$u := \mathsf{Z}_\rho(f).$$

We have omitted writing the dependence on ρ in these definitions for ease of reading. In the following, we need, in particular, that Φ is locally Lipschitz on $B_{\varepsilon^*}(\mathsf{Z}(f))$ and take C_L as in (27) from Lemma 3.2, that is, we assume (8). An alternative approach could be to instead assume it is locally Lipschitz on $B_{\varepsilon^*}(\mathsf{Z}_\rho(f))$; this would entail a different set of assumptions from the below.

Lemma 6.2. *Assume* (7), (8), *and Assumption* 2.7. *If ρ is sufficiently small and k is sufficiently large, we have*

$$u_n^k \to u^k := \mathsf{Z}_\rho(f + s_k d_k) \quad \textit{in } V \textit{ as } n \to \infty.$$

Proof. We take ρ small enough and $K > 1$ (to be specified later) such that $u_0^k = u = \mathsf{Z}_\rho(f) \in B_{\varepsilon^*/K}(\mathsf{Z}(f)) \subset B_{\varepsilon^*}(\mathsf{Z}(f))$, which is possible thanks to Theorem 2.9[6].

If k is sufficiently large, we have $f + s_k d_k \in B_{\bar{\delta}}(f)$, and we have by assumption that $f + s_k d_k \in W$. Hence, by Assumption 2.7, for ρ sufficiently small and k sufficiently large, $\mathsf{Z}_\rho(f + s_k d_k)$ is well defined and $\mathsf{Z}_\rho(f + s_k d_k) \in B_{\varepsilon^*}(\mathsf{Z}(f))$, due to the local Lipschitz property for Z_ρ; see Theorem 2.11.

We next show that the operator $T_\rho(f + s_k d_k, \cdot)$ maps the ball $B_{\varepsilon^*}(\mathsf{Z}(f))$ onto itself, if k and K are large enough. We take an arbitrary $\varphi \in B_{\varepsilon^*}(\mathsf{Z}(f))$. By using $\mathsf{Z}_\rho(f) = T_\rho(f, \mathsf{Z}_\rho(f))$ and by utilising Lemma 3.3, we get

$$\begin{aligned}\|T_\rho(f + s_k d_k, \varphi) - \mathsf{Z}(f)\|_V &\leq \|T_\rho(f + s_k d_k, \varphi) - \mathsf{Z}_\rho(f)\|_V + \|\mathsf{Z}_\rho(f) - \mathsf{Z}(f)\|_V \\ &\leq s_k \widehat{C}\|d_k\|_{V^*} + \hat{c}\|\varphi - \mathsf{Z}_\rho(f)\|_V + \frac{\varepsilon^*}{K}.\end{aligned}$$

[6] We could instead apply Theorem 4.8 if $\mathsf{Z}_\rho = \mathsf{M}_\rho$ is under consideration; this would lead to different assumptions being required for this result.

Here, $\widehat{C} \geq 0$ and $\widehat{c} \in [0,1)$ are given by Lemma 3.3. For the second term on the right-hand side, we employ the triangle inequality to get

$$\|\varphi - \mathsf{Z}_\rho(f)\|_V \leq \|\varphi - \mathsf{Z}(f)\|_V \|\mathsf{Z}(f) - \mathsf{Z}_\rho(f)\|_V \leq \varepsilon^* + \frac{\varepsilon^*}{K}.$$

Altogether, we arrive at

$$\|T_\rho(f + s_k d_k, \varphi) - \mathsf{Z}(f)\|_V \leq s_k \widehat{C} C_1 + \widehat{c}\Big(\varepsilon^* + \frac{\varepsilon^*}{K}\Big) + \frac{\varepsilon^*}{K}$$

where C_1 is the uniform bound on the d_k. The right-hand side is less than ε^* if k is sufficiently large and K is chosen large enough (we need $K > (1+\widehat{c})(1-\widehat{c})^{-1}$).

This proves the mapping property $T_\rho(f + s_k d_k, \cdot) : B_{\varepsilon^*}(\mathsf{Z}(f)) \to B_{\varepsilon^*}(\mathsf{Z}(f))$. Using Lemma 3.3 again, we find $T_\rho(f + s_k d_k, \cdot)$ is a contraction on $B_{\varepsilon^*}(\mathsf{Z}(f))$. Hence, the assertions follow from the celebrated Banach fixed point theorem, since $\mathsf{Z}_\rho(f + s_k d_k)$ is a fixed point of $T_\rho(f + s_k d_k; \cdot)$ on $B_{\varepsilon^*}(\mathsf{Z}(f))$. ■

The next proposition shows that if Φ is differentiable at $u = \mathsf{Z}_\rho(f)$, we can obtain a Taylor expansion for u_n^k.

Proposition 6.3. *Let (8) and (13) hold. For ρ sufficiently small, we have for each n,*

$$\lim_{k\to\infty} \frac{u_n^k - u}{s_k} = \alpha_n$$

where $\alpha_n := T_\rho'(f,u)(d,\alpha_{n-1})$, that is,

$$A\alpha_n + \frac{1}{\rho}\sigma_\rho'(u - \Phi(u))(\alpha_n - \Phi'(u)(\alpha_{n-1})) = d.$$

Proof. First of all, due to Lemma 3.3 (which gives local Lipschitzness for T_ρ around $V^* \times B_{\varepsilon^*}(\mathsf{Z}(f))$) and Lemma 6.1 (which gives directional differentiability of T_ρ at (f,u)) we find that T_ρ is Hadamard differentiable at (f,u) because we have taken ρ such that $u \in B_{\varepsilon^*}(\mathsf{Z}(f))$.

We use a proof by induction. The base case is obviously true (with $\alpha_n = 0$). Assume $(1/s_k)(u_n^k - u) \to \alpha_n$. Then we have

$$\begin{aligned}\frac{u_{n+1}^k - u}{s_k} &= \frac{T_\rho(f + s_k d_k, u_n^k) - T_\rho(f,u)}{s_k}\\ &= \frac{T_\rho(f + s_k d_k, u + s_k(\frac{u_n^k - u}{s_k})) - T_\rho(f,u)}{s_k} \to T_\rho'(f,u)(d,\alpha_n),\end{aligned}$$

where we used that T_ρ is Hadamard differentiable and $d_k \to d$. ■

Lemma 6.4. *Let (8) and (13) hold. We have that $\alpha_n \to \alpha$ in V, where α is the unique solution of*

$$A\alpha + \frac{1}{\rho}\sigma_\rho'(u - \Phi(u))(\alpha - \Phi'(u)(\alpha)) = d.$$

Furthermore, the map $d \mapsto \alpha$ is bounded and continuous from V^* to V.

Proof. Consider the map $\beta \mapsto \alpha$ defined as the solution mapping of

$$A\alpha + \frac{1}{\rho}\sigma'_\rho(u - \Phi(u))(\alpha - \Phi'(u)(\beta)) = d,$$

that is, the map $T'_\rho(f, u)(d, \cdot)$. We show that it is a contraction. By using $\beta, \widehat{\beta} \in V$ and the associated solutions $\alpha, \widehat{\alpha} \in V$, we get

$$A(\alpha - \widehat{\alpha}) + \frac{1}{\rho}\big(\sigma'_\rho(u - \Phi(u))(\alpha - \Phi'(u)(\beta)) - \sigma'_\rho(u - \Phi(u))(\widehat{\alpha} - \Phi'(u)(\widehat{\beta}))\big) = 0.$$

Testing with $\alpha - \Phi'(u)(\beta) - \widehat{\alpha} + \Phi'(u)(\widehat{\beta})$ and using monotonicity, we obtain

$$\langle A(\alpha - \widehat{\alpha}), \alpha - \Phi'(u)(\beta) - \widehat{\alpha} + \Phi'(u)(\widehat{\beta})\rangle \leq 0.$$

Now, since $\Phi'(u) : V \to V$ is Lipschitz with the same Lipschitz constant C_L as Φ, we obtain via similar arguments to [29] (see also the proof of Lemma 3.3 following (28)) that $T'_\rho(f, u)(d, \cdot)$ is a contraction, since C_L satisfies (27). The Banach fixed point theorem gives the result.

The map $d \mapsto \alpha$ defined through (14) is sensible for all $d \in V^*$ by the above procedure, and it is bounded, as can be seen by testing with $\alpha - \Phi'(u)(\alpha)$ and using the smallness condition on C_L of Lemma 3.2. For continuity, if $d_n \to d$ in V^* and α_n and α are the associated derivatives, we have

$$\begin{aligned}&\langle A(\alpha_n - \alpha), \alpha_n - \Phi'(u)(\alpha_n) - \alpha + \Phi'(u)(\alpha)\rangle\\ &\quad \leq \langle d_n - d, \alpha_n - \Phi'(u)(\alpha_n) - \alpha + \Phi'(u)(\alpha)\rangle\end{aligned}$$

and making use again of the Lipschitz property of $\Phi'(u)$, we conclude the claim from

$$\|\alpha_n - \alpha\|_V \leq C\|d_n - d\|_{V^*}. \qquad \blacksquare$$

Lemma 6.5. *Let the assumptions of Theorem 2.11 hold. If k is sufficiently large and ρ is sufficiently small, we have*

$$\lim_{n\to\infty} \frac{u_n^k - u}{s_k} = \frac{u^k - u}{s_k} \qquad \textit{uniformly in } k \textit{ and } \rho.$$

In [2, §5.3], three of the present authors showed that (under a different setup to what we have here) the limit $\lim_{s\searrow 0} \frac{u_n^s - s}{s}$ is uniform in n. Here though, like in [29, Theorem 32], we will show uniformity in k (and ρ) in the limit $n \to \infty$.

Proof. We argue similarly to the proof of [29, Theorem 32]. In the proof of Lemma 6.2, we have used the Banach fixed point theorem to obtain the convergence of the sequence u_n^k. This directly yields the a priori estimate

$$\|u_n^k - u^k\|_V \leq \widehat{c}^k \|u_0^k - u^k\|_V \leq \widehat{c}^n C s_k \|d_k\|_{V^*},$$

where we used $u_0^k - u^k = \mathsf{Z}_\rho(f) - \mathsf{Z}_\rho(f + s_k d_k)$ and the estimate from Theorem 2.11. This shows that

$$\frac{\|u_n^k - u^k\|_V}{s_k} \to 0 \quad \text{as } n \to \infty \text{ uniformly in } k \text{ and } \rho.$$

Using $u_n^k - u^k = u_n^k - u - (u^k - u)$, we deduce the result. ∎

We now prove our differentiability result for Z_ρ.

Proof of Theorem 2.12. The above results allow us to switch limits:

$$\lim_{k\to\infty} \frac{u^k - u}{s_k} = \lim_{k\to\infty} \lim_{n\to\infty} \frac{u_n^k - u}{s_k} = \lim_{n\to\infty} \lim_{k\to\infty} \frac{u_n^k - u}{s_k} = \lim_{n\to\infty} \alpha_n = \alpha.$$

This is exactly the Hadamard differentiability claim (see Theorem 2.12(i)). The remaining assertions on the derivative have been shown in Lemma 6.4. ∎

6.2. Differentiability for Z

We cannot pass to the limit in ρ to deduce that Z is differentiable because we do not have uniformity in ρ or s of the appropriate expression, but we may repeat the arguments in Section 6.1 with ρ taken to be zero and with Assumption 2.1 rather than Assumption 2.7. Let us point out the changes. For the $\rho = 0$ version of Lemma 6.1, we have from similar arguments to [2, Proposition 1] the following, making use of the differentiability of the VI solution map result in [27] given under a general vector lattice setting, which generalises Mignot's result in [17]:

Lemma 6.6. *Let Φ be directionally differentiable at $\varphi \in V$ and take $f \in V^*$. Then $S : V^* \times V \to V$ is directionally differentiable at (f, φ) and we have*

$$\frac{S(f + sd, \varphi + sh) - S(f, \varphi)}{s} \to S'(f, \varphi)(d, h) \quad \textit{for } d \in V^* \textit{ and } h \in V,$$

where the derivative $S'(f, \varphi)(d, h) = \delta$ is the solution of the inequality

$$\delta \in \mathcal{K}^u(\varphi, h) : \langle A\delta - d, \delta - v\rangle \le 0, \quad \forall v \in \mathcal{K}^u(\varphi, h),$$

where $u = S(f, \varphi)$ and

$$\mathcal{K}^u(\varphi, h) := \Phi'(\varphi)(h) + \mathcal{T}_{\mathbf{K}(\varphi)}(u) \cap [f - Au]^\perp.$$

In the above, recall that $\mathcal{T}_{\mathbf{K}(\varphi)}(u)$ is the tangent cone, which can be defined as the closure $\overline{\mathcal{R}_{\mathbf{K}(\varphi)}(u)}$.

Remark 6.7. When we are in a Dirichlet space setting (see the discussion around Assumption 2.19), we obtain an explicit expression for the tangent cone and we in fact have that

$$\mathcal{K}^u(\varphi, h) = \big\{w \in V : w \le \Phi'(\varphi)(h) \text{ q.e. on } \{u = \Phi(\varphi)\} \\ \text{and } \langle Au - f, w - \Phi'(\varphi)(h)\rangle = 0\big\}.$$

Let us assume (7), (8), (9), and Assumption 2.1. Similarly to before, we let $d \in \mathcal{T}_W(f)$, so that there exist $\{d_k\}$ with $d_k \to d$ in V^* and $\{s_k\}$ with $s_k \searrow 0$ such that $f + s_k d_k \in W$. Define

$$u_n^k := S(f + s_k d_k, u_{n-1}^k), \quad u_0^k := \mathsf{Z}(f),$$

and $u = \mathsf{Z}(f)$. Then

- Lemma 6.2 still holds with $u_n^k \to u^k := \mathsf{Z}(f + s_k d_k)$ if we use Theorem 2.3 instead of Theorem 2.11.
- In Proposition 6.3, we may use Lemma 6.6 instead of Lemma 6.1 and we have instead that $\alpha_n := S'(f,u)(d,\alpha_{n-1})$, which satisfies $d - A\alpha_n \in \mathcal{N}_{\mathcal{K}^u(u,\alpha_{n-1})}(\alpha_n)$, that is,

$$\alpha_n \in \mathcal{K}^u(u,\alpha_{n-1}) : \langle A\alpha_n - d, \alpha_n - v\rangle \le 0, \quad \forall v \in \mathcal{K}^u(u,\alpha_{n-1}).$$

- The result of Lemma 6.4 still holds. The map for the derivative is well defined for all $d \in V^*$: we can consider the VI

$$\alpha \in \mathcal{K}^u(\beta) : \langle A\alpha - d, \alpha - v\rangle \le 0, \quad \forall v \in \mathcal{K}^u(\beta)$$

  and use a fixed point approach, just like in the proof of Lemma 6.4 (or see [5, Proposition 3.9]). For the continuity of the derivative, a similar argument to that above works (or see [5, Proposition 3.12]). Lemma 6.5 also holds if we again use Theorem 2.3.
- Finally, arguing similarly to the proof of Theorem 2.12, we can prove Theorem 2.4.

7. Optimal control and stationarity

The proof for the existence of optimal points is straightforward.

Proof of Theorem 2.15. Let $\{f_n\} \subset U_{\text{ad}}$ be an infimising sequence with $y_n = \mathsf{M}(f_n)$ and $z_n = \mathsf{m}(f_n)$, that is,

$$J(y_n, z_n, f_n) \to \inf_{\substack{f \in U_{\text{ad}},\\ y=\mathsf{M}(f),\\ z=\mathsf{m}(f)}} J(y,z,f).$$

Then by Assumption 2.14(iii), $\{f_n\}$ is bounded in H and therefore there exists $f^* \in H$ such that, for a subsequence,

$$f_{n_j} \rightharpoonup f^* \text{ in } H.$$

The weak sequential closedness of U_{ad} yields that $f^* \in U_{\text{ad}}$. By Assumption 2.14(iv) and (v), (8) holds in a ball around the points $\mathsf{M}(f^*)$ and $\mathsf{m}(f^*)$. Using $H \overset{c}{\hookrightarrow} V^*$, we have $f_{n_j} \to f^*$ in V^* so $f_{n_j} \in B_\delta(f^*)$ sufficiently far along the sequence.

Since $B_{\bar{\delta}}(f) \cap U_{\text{ad}} \subset U_{\text{ad}}$ for any $f \in U_{\text{ad}}$, by Assumption 2.13, we have that Assumption 2.1 holds (with W selected as U_{ad}). Thus we can use Theorem 2.3, and pass to the limit to discover $(y_{n_j}, z_{n_j}) = (\mathsf{M}(f_{n_j}), \mathsf{m}(f_{n_j})) \to (\mathsf{M}(f^*), \mathsf{m}(f^*)) = (y^*, z^*)$ in V.

To see that this point is optimal, we observe that (dispensing with the subsequence notation now), using Assumption 2.14(ii),

$$J(y^*, z^*, f^*) \leq \liminf_{n\to\infty} J(y_n, z_n, f_n) \leq \lim_{n\to\infty} J(y_n, z_n, f_n) = \min_{\substack{f\in U_{\text{ad}}\\ y=\mathsf{M}(f),\\ z=\mathsf{m}(f)}} J(y, z, f). \quad \blacksquare$$

7.1. Bouligand stationarity

Working directly with the non-smooth optimisation problem, we can obtain a Bouligand stationarity characterisation of local minimisers (as in the case for variational inequalities, see [17, §5] and [18, Lemma 3.1]).

Proof of Lemma 2.16. Take h in the radial cone of U_{ad} at f^* so that it is an admissible direction. Writing $y_s = \mathsf{M}(f^* + sh)$ and $z_s = \mathsf{m}(f^* + sh)$, we obtain by Theorem 2.4 that

$$y_s = y^* + s\alpha + o(s) \quad \text{and} \quad z_s = z^* + s\beta + o(s),$$

where o is a remainder term and $\alpha = \mathsf{M}'(f^*)(h)$ and $\beta = \mathsf{m}'(f^*)(h)$. It follows that $(f^* + sh, y_s, z_s)$ can be made arbitrarily close to (f^*, y^*, z^*) if s is sufficiently small.

By the definition of local minimiser, we have $J(y_s, z_s, f^* + sh) - J(y^*, z^*, f^*) \geq 0$ for s sufficiently small. Dividing by s and taking the limit, using the fact that J is (at least) Hadamard differentiable, this yields

$$J_y(y^*, z^*, f^*)(\alpha) + J_z(y^*, z^*, f^*)(\beta) + J_f(y^*, z^*, f^*)(h) \geq 0, \quad \forall h \in \mathcal{R}_{U_{\text{ad}}}(f^*),$$

and by density and continuity of the derivatives appearing above with respect to the direction, also for $h \in \mathcal{T}_{U_{\text{ad}}}(f^*)$. $\blacksquare$

7.2. The penalised problem

We will not work directly with the penalised problem in (15), but instead a modified problem in order to prove that *every* minimiser is a stationarity point. This is a classical localisation approach.

Proposition 7.1. *Assume* (7). *For any local minimiser* (y^*, z^*, f^*) *of* (2), *there exists a sequence of locally optimal points* $(y^*_\rho, z^*_\rho, f^*_\rho)$ *of*

$$\min_{f\in U_{\text{ad}}} J(\mathsf{M}_\rho(f), \mathsf{m}_\rho(f), f) + \frac{1}{2}\|f - f^*\|^2_H \tag{35}$$

such that $(y^*_\rho, z^*_\rho, f^*_\rho) \to (y^*, z^*, f^*)$ *in* $V \times V \times H$.

Proof. Denote by γ the radius such that f^* is a minimiser on $U_{\text{ad}} \cap B^H_\gamma(f^*)$ (the latter object is the closed ball in H of radius γ with centre f^*).

Define the augmented functional $\bar{J}(y,z,f) := J(y,z,f) + \frac{1}{2}\|f - f^*\|_H^2$ that appears in (35) and consider the problem

$$\min_{f \in U_{\mathrm{ad}} \cap B_\gamma^H(f^*)} \bar{J}(\mathsf{M}_\rho(f), \mathsf{m}_\rho(f), f). \tag{36}$$

By the same proof as for Theorem 2.15 with the obvious modifications, we find that there exists an optimal point to this problem, which we denote by $(\bar{y}_\rho, \bar{z}_\rho, \bar{f}_\rho)$. From

$$\bar{J}(\bar{y}_\rho, \bar{z}_\rho, \bar{f}_\rho) \le \bar{J}(\mathsf{M}_\rho(f^*), \mathsf{m}_\rho(f^*), f^*), \tag{37}$$

and using $\mathsf{Z}_\rho(f^*) \to \mathsf{Z}(f^*)$ (due to Theorem 2.8) and the continuity of $\bar{J}$, we have

$$\limsup_{\rho \to 0} \bar{J}(\bar{y}_\rho, \bar{z}_\rho, \bar{f}_\rho) \le J(y^*, z^*, f^*).$$

On the other hand, it follows from (37) and Theorem 2.8 that $\bar{J}(\bar{y}_\rho, \bar{z}_\rho, \bar{f}_\rho)$ is uniformly bounded, and hence, due to Assumption 2.14(iii), we obtain the existence of $\hat{f}$ such that (for a subsequence that we have relabelled) $\bar{f}_\rho \rightharpoonup \hat{f}$ in H with the convergence strong in V^*.

We have

$$\mathsf{M}_\rho(\bar{f}_\rho) - \mathsf{M}(\hat{f}) = (\mathsf{M}_\rho(\bar{f}_\rho) - \mathsf{M}_\rho(\hat{f})) + (\mathsf{M}_\rho(\hat{f}) - \mathsf{M}(\hat{f}))$$

and availing ourselves of the Lipschitz estimate of Theorem 2.11 (with the Lipschitz constant independent of ρ), we have that the first term above converges to zero and the second term does also due to Theorem 2.9. Hence $\bar{y}_\rho \to \hat{y} := \mathsf{M}(\hat{f})$ and, arguing similarly, $\bar{z}_\rho \to \hat{z} := \mathsf{m}(\hat{f})$. By the inequality $\limsup(a_n) + \liminf(b_n) \le \limsup(a_n + b_n)$ and weak lower semicontinuity, this gives

$$\begin{aligned}&\limsup_{\rho \to 0} \bar{J}(\bar{y}_\rho, \bar{z}_\rho, \bar{f}_\rho) \\ &\quad \ge J(\hat{y}, \hat{z}, \hat{f}) + \limsup_{\rho \to 0} \|\bar{f}_\rho - f^*\|_H^2 \ge J(y^*, z^*, f^*) + \limsup_{\rho \to 0} \|\bar{f}_\rho - f^*\|_H^2,\end{aligned}$$

with the last inequality because (y^*, z^*, f^*) is a local minimiser and $\hat{f} \in B_\gamma^H(f^*)$. Combining these two inequalities shows that $\hat{f} = f^*$ and $\bar{f}_\rho \to f^*$ in H. The latter fact implies that for ρ sufficiently small, $\bar{f}_\rho \in B_\gamma^H(f^*)$ automatically and, hence, the feasible set in (36) can be taken to be just U_{ad}. Finally, since the limit point $\hat{f} = f^*$ is independent of the subsequence that was taken, it follows by the subsequence principle that the entire sequence $\{\bar{f}_\rho\}$ converges. From this, we also gain convergence for $\{\bar{y}_\rho\}$ and $\{\bar{z}_\rho\}$ (by repeating the above arguments). ■

7.3. C-stationarity

Via Proposition 7.1, we obtain the existence of minimisers $(y_\rho^*, z_\rho^*, f_\rho^*)$ of (35) such that

$$(y_\rho^*, z_\rho^*, f_\rho^*) \to (y^*, z^*, f^*) \text{ in } V \times V \times H.$$

Thus, for any $\varepsilon > 0$, we can find a ρ_0 such that $\rho \le \rho_0$ implies

$$(y_\rho^*, z_\rho^*) \in B_\varepsilon(y^*) \times B_\varepsilon(z^*).$$

We make the standing assumption (Assumption 2.18 (i)) on the local differentiability of Φ and linearity of the derivative on the above balls. Observe that (13) (which in this context is the assumption that Φ is differentiable at $\mathsf{Z}_\rho(f_\rho^*)$) follows from these assumptions: since $\mathsf{Z}_\rho(f_\rho^*) \to \mathsf{Z}(f^*)$ (thanks to Theorem 2.8), for sufficiently small ρ, $\mathsf{Z}_\rho(f_\rho^*) \in B_\varepsilon(\mathsf{Z}(f^*))$ and Φ is differentiable at these points too.

In the next result, we meet the conditions to apply the directional differentiability result of Theorem 2.12.

Proposition 7.2. *Let* (7) *hold. For any optimal point* $(y_\rho^*, z_\rho^*, f_\rho^*)$ *of* (35)*, there exists* $(p_\rho^*, q_\rho^*) \in V \times V$ *such that*

$$\begin{aligned}
&A^* p_\rho^* + \frac{1}{\rho}(\mathrm{Id} - \Phi'(y_\rho^*))^* \sigma_\rho'(y_\rho^* - \Phi(y_\rho^*)) p_\rho^* = -J_y(y_\rho^*, z_\rho^*, f_\rho^*),\\
&A^* q_\rho^* + \frac{1}{\rho}(\mathrm{Id} - \Phi'(z_\rho^*))^* \sigma_\rho'(z_\rho^* - \Phi(z_\rho^*)) q_\rho^* = -J_z(y_\rho^*, z_\rho^*, f_\rho^*),\\
&\langle J_f(y_\rho^*, z_\rho^*, f_\rho^*) - p_\rho^* - q_\rho^*, f_\rho^* - v\rangle + (f_\rho^* - f^*, f_\rho^* - v)_H \le 0, \quad \forall v \in U_{\mathrm{ad}}.
\end{aligned} \tag{38}$$

Proof. Defining $\hat{J}(f) := \bar{J}(\mathsf{M}_\rho(f), \mathsf{m}_\rho(f), f)$ we consider the reduced problem

$$\min_{f \in U_{\mathrm{ad}}} \hat{J}(f).$$

Note that we may use the chain rule (see, e.g., [9, Proposition 2.47]) to differentiate $\hat{J}$ since it is the composition of a C^1 map with a directionally differentiable map. Now, at the optimal point f_ρ^*, we have $\hat{J}(f_\rho^* + sh) - \hat{J}(f_\rho^*) \ge 0$ for all $h \in \mathcal{R}_{U_{\mathrm{ad}}}(f_\rho^*)$, and hence

$$\langle \hat{J}'(f_\rho^*), h\rangle \ge 0, \quad \forall h \in \mathcal{R}_{U_{\mathrm{ad}}}(f_\rho^*).$$

We calculate, with $y_\rho^* = \mathsf{M}_\rho(f_\rho^*)$ and $z_\rho^* = \mathsf{m}_\rho(f_\rho^*)$,

$$\begin{aligned}
\langle \hat{J}'(f_\rho^*), h\rangle &= \langle J_y(y_\rho^*, z_\rho^*, f_\rho^*), \mathsf{M}_\rho'(f_\rho^*)(h)\rangle + \langle J_z(y_\rho^*, z_\rho^*, f_\rho^*), \mathsf{m}_\rho'(f_\rho^*)(h)\rangle\\
&\quad + \langle J_f(y_\rho^*, z_\rho^*, f_\rho^*), h\rangle\\
&= \langle \mathsf{M}_\rho'(f_\rho^*)^* J_y(y_\rho^*, z_\rho^*, f_\rho^*) + \mathsf{m}_\rho'(f_\rho^*)^* J_z(y_\rho^*, z_\rho^*, f_\rho^*), h\rangle\\
&\quad + \langle J_f(y_\rho^*, z_\rho^*, f_\rho^*), h\rangle + (f_\rho^* - f^*, h)_H
\end{aligned}$$

with the adjoint well defined since $\mathsf{Z}_\rho'(f_\rho^*)$ is a bounded linear map thanks to Assumption 2.18(i) (which implies that the derivative satisfies a linear PDE; see (14)). It is easy to see that the previous inequality in fact holds for all $h \in \mathcal{T}_{U_{\mathrm{ad}}}(f_\rho^*)$ by a simple density argument.

Defining $\theta_\rho^* := -(\mathsf{M}'_\rho(f_\rho^*)^*(J_y(y_\rho^*, z_\rho^*, f_\rho^*)) + \mathsf{m}'_\rho(f_\rho^*)^*(J_z(y_\rho^*, z_\rho^*, f_\rho^*)))$, we write the above as

$$\langle J_f(y_\rho^*, z_\rho^*, f_\rho^*) - \theta_\rho^*, h\rangle + (f_\rho^* - f^*, h)_H \geq 0, \quad \forall h \in \mathcal{T}_{U_{\text{ad}}}(f).$$

Take $v \in U_{\text{ad}}$, then $h := v - f_\rho^*$ is in the tangent cone. With this choice of h, we recover

$$\langle J_f(y_\rho^*, z_\rho^*, f_\rho^*) - \theta_\rho^*, v - f_\rho^*\rangle + (f_\rho^* - f^*, v - f_\rho^*)_H \geq 0, \quad \forall v \in U_{\text{ad}}.$$

Let us characterise each term in θ_ρ. First, observe that

$$\begin{aligned} &p := \mathsf{M}'_\rho(g)^*(d) \\ &\iff A^*p + \frac{1}{\rho}(\text{Id} - \Phi'(v_\rho))^*\sigma'_\rho(v_\rho - \Phi(v_\rho))p = d \quad \text{where } v_\rho = \mathsf{M}_\rho(g) \end{aligned}$$

and a similar formula holds for $\mathsf{m}'_\rho(f)^*(w)$. Note that these adjoint maps (which are solution maps of linear PDEs) are linear in w. Hence, if we define

$$p_\rho^* := \mathsf{M}'_\rho(f_\rho^*)^*(-J_y(y_\rho^*, z_\rho^*, f_\rho^*)), \quad q_\rho^* := \mathsf{m}'_\rho(f_\rho^*)^*(-J_z(y_\rho^*, z_\rho^*, f_\rho^*)),$$

they satisfy $\theta_\rho^* = p_\rho^* + q_\rho^*$ and the equations stated in the proposition. ∎

Before proceeding, let us record some facts. Due to the Lipschitz condition (see Assumption 2.14(iv)–(v)), we have

$$\begin{aligned} (\text{Id} - \Phi'(w)) : V \to V \text{ is invertible for } w \in B_\varepsilon(y^*) \text{ if } J_y \not\equiv 0,& \\ \text{and for } w \in B_\varepsilon(z^*) \text{ if } J_z \not\equiv 0,& \end{aligned} \tag{39}$$

which follows from the Neumann series, and the inverse satisfies $\|(\text{Id} - \Phi'(w))^{-1}v\|_V \leq (1 - C_L)^{-1}\|v\|_V$ for all $v \in V$. For an arbitrary $v \in V$, we set $u = (\text{Id} - \Phi'(w))^{-1}v$. Then we have

$$\langle A(\text{Id} - \Phi'(w))^{-1}v, v\rangle = \langle Au, (\text{Id} - \Phi'(w))u\rangle \geq C'_a\|u\|_V^2 \geq \frac{C'_a}{(1 + C_L)^2}\|v\|_V^2$$

for some C'_a depending only on C_L, C_a, C_b, and the self-adjointedness of A, by using Lemma 3.2 (see also [29]) adapted to the operator $\Phi'(w)$. Thus, we have shown that

$$A(\text{Id} - \Phi'(w))^{-1} : V \to V^* \text{ is uniformly bounded and uniformly coercive} \tag{40}$$

for w belonging to the same sets as in (39); see also [28, Lemmas 3.3 and 3.5].

Lemma 7.3. *Under Assumption 2.18(ii), if $J_y \not\equiv 0$, for sequences $v_n \to v$ and $q_n \rightharpoonup q$ in V with $v_n, v \in B_\varepsilon(y^*)$, we have*

$$\liminf_{n\to\infty}\langle A(\text{Id} - \Phi'(v_n))^{-1}q_n, q_n\rangle \geq \langle A(\text{Id} - \Phi'(v))^{-1}q, q\rangle. \tag{41}$$

A similar result holds if $J_z \not\equiv 0$ with the obvious modifications.

Proof. Let $T_n := (\mathrm{Id} - \Phi'(v_n))^{-1}$. We have, due to the coercivity above,

$$0 \leq \langle AT_n(q_n - q), q_n - q\rangle = \langle AT_n q_n, q_n\rangle - \langle AT_n q_n, q\rangle - \langle AT_n q, q_n\rangle + \langle AT_n q, q\rangle$$

Rearranging, we have

$$\langle AT_n q_n, q_n\rangle \geq \langle AT_n q_n, q\rangle + \langle AT_n q, q_n\rangle - \langle AT_n q, q\rangle.$$

Taking the limit inferior, using on the right-hand side (16) for the first and last terms and (17) for the second term, we obtain the desired statement. ∎

For convenience and because of structural reasons, the proof of Theorem 2.20 will be realised via the next three propositions. First, we show that a system of so-called *weak C-stationarity* is satisfied; see [5, §5] for the terminology.

Proposition 7.4 (Weak C-stationarity). *There exists* $(p^*, q^*, \lambda^*, \zeta^*) \in V \times V \times V^* \times V^*$ *satisfying*

$$\begin{aligned}
y^* &= \mathsf{M}(f^*), && \text{(42a)}\\
z^* &= \mathsf{m}(f^*), && \text{(42b)}\\
A^* p^* + (\mathrm{Id} - \Phi'(y^*))^* \lambda^* &= -J_y(y^*, z^*, f^*), && \text{(42c)}\\
A^* q^* + (\mathrm{Id} - \Phi'(z^*))^* \zeta^* &= -J_z(y^*, z^*, f^*), && \text{(42d)}\\
f^* \in U_{\mathrm{ad}} : \langle J_f(y^*, z^*, f^*) - p^* - q^*, f^* - v\rangle &\leq 0, \quad \forall v \in U_{\mathrm{ad}}, && \text{(42e)}\\
\langle \lambda^*, p^*\rangle &\geq 0, && \text{(42f)}\\
\langle \zeta^*, q^*\rangle &\geq 0. && \text{(42g)}
\end{aligned}$$

In this and the following proofs, for ease of reading, we will omit the stars in ρ-dependent notation as p_ρ^* and simply write this as p_ρ.

Proof. By construction, we already know that $(y_\rho, z_\rho, f_\rho) \to (y^*, z^*, f^*)$ in $V \times V \times H$ due to Proposition 7.1. We now need to pass to the limit in system (38) for the adjoint states and the optimal control. We write the arguments just for p_ρ; obvious modifications will work for the q_ρ equation too.

1. Satisfaction of the equation. The weak form of the equation for p_ρ is

$$\langle A^* p_\rho, \varphi\rangle + \frac{1}{\rho}\langle \sigma_\rho'(y_\rho - \Phi(y_\rho)) p_\rho, (\mathrm{Id} - \Phi'(y_\rho))\varphi\rangle = -\langle J_y(y_\rho, z_\rho, f_\rho), \varphi\rangle, \quad \forall \varphi \in V.$$

By defining $v := (\mathrm{Id} - \Phi'(y_\rho))\varphi$, thanks to the invertibility property in (39), this can be transformed to

$$\begin{aligned}
&\langle A^* p_\rho, (\mathrm{Id} - \Phi'(y_\rho))^{-1} v\rangle + \frac{1}{\rho}\langle \sigma_\rho'(y_\rho - \Phi(y_\rho)) p_\rho, v\rangle\\
&\quad = -\langle J_y(y_\rho, z_\rho, f_\rho), (\mathrm{Id} - \Phi'(y_\rho))^{-1} v\rangle
\end{aligned}$$

for all $v \in V$. Now, selecting $v = p_\rho$, using the coercivity in (40), the monotonicity of σ_ρ (which implies that $\langle \sigma'_\rho(v)(h), h\rangle \geq 0$ for all $v, h \in V$), Young's inequality with $\gamma > 0$, and the uniform boundedness of J_y (see Assumption 2.14(i)) and of $(\mathrm{Id} - \Phi'(y_\rho))^{-1}$ (see the discussion above (40)), we obtain

$$C'_a \|p_\rho\|_V^2 \leq C_\gamma + \gamma \|p_\rho\|_V^2.$$

Selecting γ sufficiently small so that the right-most term is absorbed onto the left, we obtain a bound on $\{p_\rho\}$ independent of ρ. This gives rise to the convergence (for a subsequence that has been relabelled)

$$p_\rho \rightharpoonup p.$$

In a similar way, we also obtain $q_\rho \rightharpoonup q$. Define

$$\begin{aligned}
\lambda_\rho &:= \frac{1}{\rho}\sigma'_\rho(y_\rho - \Phi(y_\rho))^* p_\rho,\\
\mu_\rho &:= \frac{1}{\rho}(\mathrm{Id} - \Phi'(y_\rho))^* \sigma'_\rho(y_\rho - \Phi(y_\rho))^* p_\rho = -J_y(y_\rho, z_\rho, f_\rho) - A^* p_\rho,
\end{aligned}$$

the latter of which, since the right-hand side converges, satisfies

$$\mu_\rho \rightharpoonup \mu := -J_y(y, z, f) - A^* p. \tag{43}$$

Setting $\lambda := (\mathrm{Id} - \Phi'(y)^*)^{-1}\mu$ in (43), we get (42c).

2. Inequality relating multiplier to adjoint. Again using the monotonicity of σ_ρ,

$$\begin{aligned}
\langle J_y(y_\rho, z_\rho, f_\rho) + A^* p_\rho, (\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho\rangle &= -\langle \mu_\rho, (\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho\rangle\\
&= -\frac{1}{\rho}\langle \sigma'_\rho(y_\rho - \Phi(y_\rho))^* p_\rho, p_\rho\rangle \leq 0,
\end{aligned}$$

and taking the limit superior of this, we obtain (noting that $(\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho \rightharpoonup (\mathrm{Id} - \Phi'(y))^{-1} p$ by (16))

$$\begin{aligned}
0 &\geq \limsup_{\rho\to 0} \langle J_y(y_\rho, z_\rho, f_\rho) + A^* p_\rho, (\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho\rangle\\
&\geq \limsup_{\rho\to 0} \langle J_y(y_\rho, z_\rho, f_\rho), (\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho\rangle + \liminf_{\rho\to 0}\langle A(\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho, p_\rho\rangle\\
&\qquad \text{(using } \limsup(a_n + b_n) \geq \limsup(a_n) + \liminf(b_n))\\
&\geq \langle J_y(y, z, f), (\mathrm{Id} - \Phi'(y))^{-1} p\rangle + \langle A(\mathrm{Id} - \Phi'(y))^{-1} p, p\rangle\\
&= \langle -\mu^*, (\mathrm{Id} - \Phi'(y))^{-1} p\rangle
\end{aligned}$$

using the continuity of the Fréchet derivative from Assumption 2.14(i) and (41) for the final inequality. This shows (42f).

3. VI relating control to adjoint. Finally, writing the VI relating u_ρ and $\theta_\rho := p_\rho + q_\rho$ in (38) as

$$0 \le \langle J_f(y_\rho, z_\rho, f_\rho) - \theta_\rho, v - f_\rho\rangle + (f_\rho - f^*, v - f_\rho)_H = \langle J_f(y_\rho, z_\rho, f_\rho), v - f_\rho\rangle \\ - \langle \theta_\rho, v - f_\rho\rangle + (f_\rho - f^*, v - f_\rho)_H, \quad \forall v \in U_{\text{ad}}$$

and taking the limit inferior here, using the continuity of J_f from Assumption 2.14(i) and the inequality $\liminf_n (a_n + b_n) \le \limsup_n a_n + \liminf b_n$, we get the desired inequality. ■

The next results (until the end of this section) use the fact that $(\cdot)^+ : V \to V$ is continuous. Furthermore, the next proposition also uses weak sequential continuity of the map.

Proposition 7.5 (Orthogonality conditions). *We have*

$$\langle \xi_1^*, (p^*)^+\rangle = \langle \xi_1^*, (p^*)^-\rangle = \langle \xi_2^*, (q^*)^+\rangle = \langle \xi_2^*, (q^*)^-\rangle = 0.$$

In the proof, we use specific properties of the fact that H is a Lebesgue space. The proof is almost identical to that of [5, Theorem 5.11], but we give it here for completeness.

Proof. Let us introduce the sets

$$M_1(\rho) := \big\{0 \le y_\rho - \Phi(y_\rho) < \varepsilon\big\} \quad \text{and} \quad M_2(\rho) := \big\{y_\rho - \Phi(y_\rho) \ge \varepsilon\big\}.$$

Since $\langle \xi_\rho, y_\rho - \Phi(y_\rho)\rangle \to \langle \xi^*, y - \Phi(y)\rangle = 0$, we find

$$\frac{1}{\rho}\int_{M_1(\rho)} \frac{(y_\rho - \Phi(y_\rho))^3}{2\varepsilon} + \frac{1}{\rho}\int_{M_2(\rho)} \Big(y_\rho - \Phi(y_\rho) - \frac{\varepsilon}{2}\Big)(y_\rho - \Phi(y_\rho)) \to 0,$$

and as both integrands are non-negative,

$$\left\| \frac{\chi_{M_1(\rho)}(y_\rho - \Phi(y_\rho))^{\frac{3}{2}}}{\sqrt{\rho\varepsilon}} \right\| \to 0 \quad \text{and} \quad \left\| \frac{\chi_{M_2(\rho)}(y_\rho - \Phi(y_\rho) - \frac{\varepsilon}{2})}{\sqrt{\rho}} \right\| \to 0, \tag{44}$$

where for the second convergence we used the fact that $y_\rho - \Phi(y_\rho) \ge y_\rho - \Phi(y_\rho) - \varepsilon/2 \ge 0$. We calculate

$$\begin{aligned}\langle \xi_\rho, p_\rho\rangle &= \frac{1}{\rho}\int_{M_1(\rho)} \frac{(y_\rho - \Phi(y_\rho))^2}{2\varepsilon} p_\rho + \frac{1}{\rho}\int_{M_2(\rho)} \Big(y_\rho - \Phi(y_\rho) - \frac{\varepsilon}{2}\Big) p_\rho \\ &\le \frac{1}{2}\left\| \chi_{M_1(\rho)} \frac{(y_\rho - \Phi(y_\rho))^{3/2}}{\sqrt{\rho\varepsilon}} \right\| \left\| \frac{(y_\rho - \Phi(y_\rho))^{1/2}}{\sqrt{\rho\varepsilon}} \chi_{M_1(\rho)} p_\rho \right\| \\ &\quad + \left\| \frac{\chi_{M_2(\rho)}(y_\rho - \Phi(y_\rho) - \frac{\varepsilon}{2})}{\sqrt{\rho}} \right\| \left\| \frac{\chi_{M_2(\rho)} p_\rho}{\sqrt{\rho}} \right\|.\end{aligned} \tag{45}$$

Now, using (44), the first factor in each term above converges to zero and hence the above right-hand side will converge to zero if we are able to show that the second factor in each term remains bounded. Since μ_ρ and $(\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho$ are bounded (for the latter, see (39) and the discussion), so is their duality product, and therefore

$$\begin{aligned} C \geq |\langle \mu_\rho, (\mathrm{Id} - \Phi'(y_\rho))^{-1} p_\rho \rangle| &= \frac{1}{\rho}\left| \int_\Omega \sigma_\rho'(y_\rho - \Phi(y_\rho))(p_\rho)^2 \right| \\ &= \frac{1}{\rho} \int_\Omega \chi_{M_1(\rho)} \frac{y_\rho - \Phi(y_\rho)}{\varepsilon}(p_\rho)^2 + \frac{1}{\rho}\int_\Omega \chi_{M_2(\rho)}(p_\rho)^2. \end{aligned}$$

Both of the terms on the right-hand side are individually bounded uniformly in ρ as the integrands are non-negative. This fact then implies from (45) that

$$\langle \xi^*, p^* \rangle = 0.$$

Replacing p_ρ by $(p_\rho)^+$ in (45) and in the above calculation, we also obtain in the same way (utilising the fact[7] that $v_n \rightharpoonup v$ in V implies that $v_n^+ \rightharpoonup v^+$ in V)

$$\langle \xi^*, (p^*)^+ \rangle = 0. \qquad \blacksquare$$

We are left to show conditions (20h) and (20i) on the multipliers. To do so, we will follow an approach motivated by [26, Lemma 2.6].

Lemma 7.6. *If $g_n \rightharpoonup g$ in V^* and $s_n \to s$ in V with $s_n \geq 0$ and*

$$\langle g_n, v \rangle = 0, \quad \forall v \in V, \ 0 \leq v \leq s_n,$$

then

$$\langle g, v \rangle = 0, \quad \forall v \in V, \ 0 \leq v \leq s.$$

Proof. Let $v \in V$ with $0 \leq v \leq s$ be given. Set $v_n := \inf(v, s_n)$, which satisfies $0 \leq v_n \leq s_n$ and $v_n \to v$. Thus,

$$0 = \langle g_n, \inf(v, s_n) \rangle \to \langle g, v \rangle. \qquad \blacksquare$$

In the next lemma, we use the fact that

$$\sigma_\rho(z) = \sigma_\rho(z - v), \quad \forall v \in V, \ 0 \leq v \leq z^-.$$

This essentially means that $\sigma_\rho(z)$ ignores changes of z in the regions where z is already negative.

Lemma 7.7. *We have*

$$\langle \lambda^*, v \rangle = 0, \quad \forall v \in V, \ 0 \leq v \leq \Phi(y^*) - y^*.$$

[7] This is due to the compact embedding $V \overset{c}{\hookrightarrow} H$ and the fact that $(\cdot)^+ : H \to H$ is continuous, as well as the boundedness of $(\cdot)^+ : V \to V$ that we assumed in the introduction.

The condition on λ^* means, roughly speaking, that λ^* vanishes on the inactive set on which $y^* - \Phi(y^*) < 0$.

Proof. The property on σ_ρ stated above immediately implies

$$\sigma_\rho'(z)v = 0, \quad \forall v \in V, \;\; 0 \le v \le z^-.$$

Using the definition of λ_ρ, we find

$$\langle \lambda_\rho, v\rangle = \frac{1}{\rho}\langle \sigma_\rho'(y_\rho - \Phi(y_\rho))v, p_\rho\rangle = 0, \quad \forall v \in V, \;\; 0 \le v \le (y_\rho - \Phi(y_\rho))^-$$

by the above property of σ_ρ'. As $(y_\rho - \Phi(y_\rho))^- \to (y^* - \Phi(y^*))^- = \Phi(y^*) - y^*$, Lemma 7.6 yields the claim. ■

Obviously, a similar condition also holds for ζ^*.

Proposition 7.8. *Let Assumption 2.19 hold. We have*

$$\langle \lambda^*, v\rangle = 0 \; \forall v \in V : v = 0 \text{ q.e. on } \{y^* = \Phi(y^*)\},$$
$$\langle \zeta^*, v\rangle = 0 \; \forall v \in V : v = 0 \text{ q.e. on } \{z^* = \Phi(z^*)\}.$$

Hence, system (20) is satisfied.

Proof. Set $\hat{y} := \Phi(y^*) - y^*$. Since $\hat{y} \in V$, it has a quasi-continuous representative and we will identify $\hat{y}$ with its representative. Define the active set

$$A := \{\hat{y} = 0\}.$$

Let $v \in V$ with $v \ge 0$ and $v = 0$ quasi-everywhere on A be given. Since we have the following expression for the tangent cone of V_+ (see [9, Theorem 6.57] in the $V = H_0^1(\Omega)$ setting or [17, Lemma 3.2] in the general Dirichlet space setting):

$$\mathcal{T}_{V_+}(\hat{y}) = \{\varphi \in V : \varphi \ge 0 \text{ q.e. on } A\},$$

it follows that $v \in \mathcal{T}_{V+}(\hat{y})$ and hence, there exists a sequence $\{v_n\}$ with $v_n \to v$ in V and $v_n \le t_n \hat{y}$ for some $t_n > 0$. Thus,

$$0 \le \max(0, v_n/t_n) \le \hat{y}$$

and we can apply the conclusion of Lemma 7.7 and get

$$\left\langle \lambda^*, \max\left(0, \frac{v_n}{t_n}\right)\right\rangle = 0.$$

Multiplying by t_n and passing to the limit $n \to \infty$ gives

$$\langle \lambda^*, v\rangle = 0 \text{ for all } v \ge 0, \; v = 0 \text{ q.e. on } A.$$

Then, using the decomposition $v = v^+ - v^-$, we obtain the result. ■

Remark 7.9 ($\mathcal{E}$-almost C-stationarity). Define the inactive sets

$$\mathcal{I}_1 = \{y^* < \Phi(y^*)\} \quad \text{and} \quad \mathcal{I}_2 = \{z^* < \Phi(z^*)\}.$$

The following argument shows that the conditions

$$\begin{aligned} &\forall \tau > 0, \exists E^\tau \subset \mathcal{I}_1 \text{ with } |\mathcal{I}_1 \setminus E^\tau| \le \tau : \langle \lambda^*, v\rangle = 0 \\ &\qquad\qquad\qquad\qquad \forall v \in V : v = 0 \text{ a.e. on } \Omega \setminus E^\tau, \end{aligned} \tag{46}$$

$$\begin{aligned} &\forall \tau > 0, \exists E^\tau \subset \mathcal{I}_2 \text{ with } |\mathcal{I}_2 \setminus E^\tau| \le \tau : \langle \zeta^*, v\rangle = 0 \\ &\qquad\qquad\qquad\qquad \forall v \in V : v = 0 \text{ a.e. on } \Omega \setminus E^\tau, \end{aligned} \tag{47}$$

are an easy consequence of Proposition 7.8 and of the regularity of the Lebesgue measure: for every $\tau > 0$, there exists an open set O^τ with $\{y^* = \Phi(y^*)\} \subset O^\tau$ and $|O^\tau \setminus \{y^* = \Phi(y^*)\}| \le \tau$. Using $\mathcal{I}_1 = \Omega \setminus \{y^* = \Phi(y^*)\}$ and $E^\tau := \Omega \setminus O^\tau$, we get $E^\tau \subset \mathcal{I}_1$ and $|\mathcal{I}_1 \setminus E^\tau| \le \tau$ by taking complements. Next, we take an arbitrary function $v \in V$ with $v = 0$ almost everywhere on $\Omega \setminus E^\tau = O^\tau$. Since O^τ is open, this gives $v = 0$ quasi-everywhere on O^τ and, in particular, $v = 0$ quasi-everywhere on $\{y^* = \Phi(y^*)\}$. Thus, Proposition 7.8 yields $\langle \lambda^*, v\rangle = 0$ and we get (46).

Remark 7.10 (Regularity of optimal control). Suppose we have $J_f(y, z, f) = \nu f$ and we take U_{ad} to be of the box constraint type

$$U_{\text{ad}} = \{u \in H : u_a \le u \le u_b \text{ a.e. in } \Omega\}$$

for given functions $u_a, u_b \in H$. The VI relating f^* and p^* is, in this case,

$$f^* \in U_{\text{ad}} : \langle \nu f^* - p^* - q^*, f^* - v\rangle \le 0, \quad \forall v \in U_{\text{ad}},$$

Using the characterisation in [15, §II.3],

$$\frac{1}{\nu}(p^* + q^*) + \left(u_a - \frac{p^* + q^*}{\nu}\right)^+ - \left(\frac{p^* + q^*}{\nu} - u_b\right)^+ = f^*$$

and it follows that $f^* \in V$ if u_a and u_b belong to V.

7.4. Alternative stationarity conditions

In some papers, for example, [22], in direct analogy with the finite-dimensional setting, rather than the inequality condition in (20f), the stronger condition

$$\langle \lambda^*, \psi p^*\rangle \ge 0 \quad \text{for all sufficiently smooth and non-negative } \psi$$

is required in order to satisfy the terminology *C-stationarity*. We can show this holds under an additional assumption.

Proposition 7.11 (Satisfaction of alternative criterion in C-stationarity). *Under the conditions of Theorem 2.20, assume also that for $q_\rho \rightharpoonup q$ in V,*

$$\liminf_{\rho\to 0}\langle A^* q_\rho, (\mathrm{Id}-\Phi'(y^*_\rho))^{-1}(\psi q_\rho)\rangle \geq \langle A^* q, (\mathrm{Id}-\Phi'(y^*))^{-1}(\psi q)\rangle, \quad \forall \psi \in W^{1,\infty}(\Omega) \text{ with } \psi \geq 0. \tag{48}$$

Then inequality condition (20f) can be strengthened to

$$\langle \lambda^*, \psi p^*\rangle \geq 0, \quad \forall \psi \in W^{1,\infty}(\Omega) \text{ with } \psi \geq 0.$$

Under the obvious modifications to the above assumption, (20g) can also be strengthened similarly.

Proof. Testing the equation for p_ρ with $(\mathrm{Id}-\Phi'(y_\rho))^{-1}(\psi p_\rho)$, noticing that $\psi p_\rho \rightharpoonup \psi p$ in V and arguing in a similar way to the proof of Proposition 7.4,

$$\begin{aligned}
\limsup_{\rho\to 0} -\langle \mu_\rho, (\mathrm{Id}-\Phi'(y_\rho))^{-1}(\psi p_\rho)\rangle &= \limsup_{\rho\to 0}\langle J_y(y_\rho, z_\rho, f_\rho), (\mathrm{Id}-\Phi'(y_\rho))^{-1}(\psi p_\rho)\rangle \\
&\quad + \liminf_{\rho\to 0}\langle A^* p_\rho, (\mathrm{Id}-\Phi'(y_\rho))^{-1}(\psi p_\rho)\rangle \\
&\geq \langle J_y(y,z,f), (\mathrm{Id}-\Phi'(y))^{-1}(\psi p)\rangle \\
&\quad + \langle A^* p, (\mathrm{Id}-\Phi'(y))^{-1}(\psi p)\rangle \\
&\qquad \text{(using Assumption 2.14(i) and (48))} \\
&= -\langle \mu, (\mathrm{Id}-\Phi'(y))^{-1}(\psi p)\rangle = -\langle \lambda, \psi p\rangle.
\end{aligned}$$

On the other hand, we have

$$\langle \mu_\rho, (\mathrm{Id}-\Phi'(y_\rho))^{-1}(\psi p_\rho)\rangle = \langle \lambda_\rho, \psi p_\rho\rangle = \frac{1}{\rho}\int_\Omega \sigma'_\rho(y_\rho - \Phi(y_\rho))(p_\rho)^2\psi \geq 0,$$

which implies the result. ∎

Remark 7.12. Some works (such as [14]) call system (20) C-stationarity only if the “q.e.” in conditions (20h) and (20i) are replaced by “a.e”. Note that this is a stronger condition.

8. Conclusion

In conclusion, we have provided a thorough theory of Lipschitz and differential stability for M and m and the penalised versions. We studied in depth the penalised problem in (4) and its properties and used it to derive stationarity conditions for a general class of optimisation problems with the extremal maps as constraints. We conclude with the following remarks:

- Applying this theory to other real-world phenomena (such as applications in biomedicine [23]) in this context and studying numerical schemes in line with Remark 4.11 are natural next steps.
- It would be interesting to derive strong stationarity conditions for (2) using the approaches of [5, 25].
- Resolving whether in Theorem 4.8 $\mathtt{m}_\rho(f)$ indeed converges (weakly) to $\mathtt{m}(f)$ or providing a counterexample is open.
- We aim to investigate whether the convergence result for $\mathtt{M}_\rho(f)$ in Theorem 4.8 can be used to obtain differentiability results for $\mathtt{M}$ without the small Lipschitz assumption in (2.14).

Acknowledgements. We thank the referees for their helpful comments.

Funding. CNR was partially supported by NSF grant DMS-2012391.

References

[1] A. Alphonse, C. Christof, M. Hintermüller, and I. P. A. Papadopoulos. A globalized inexact semismooth Newton method for nonsmooth fixed-point equations involving variational inequalities. 2024, arXiv:2409.19637

[2] A. Alphonse, M. Hintermüller, and C. N. Rautenberg, Directional differentiability for elliptic quasi-variational inequalities of obstacle type. *Calc. Var. Partial Differential Equations* **58** (2019), no. 1, article no. 39 Zbl 1439.49014 MR 3903798

[3] A. Alphonse, M. Hintermüller, and C. N. Rautenberg, Stability of the solution set of quasi-variational inequalities and optimal control. *SIAM J. Control Optim.* **58** (2020), no. 6, 3508–3532 Zbl 1516.49004 MR 4178375

[4] A. Alphonse, M. Hintermüller, and C. N. Rautenberg, On the differentiability of the minimal and maximal solution maps of elliptic quasi-variational inequalities. *J. Math. Anal. Appl.* **507** (2022), no. 1, article no. 125732 Zbl 1477.49011 MR 4324299

[5] A. Alphonse, M. Hintermüller, and C. N. Rautenberg, Optimal control and directional differentiability for elliptic quasi-variational inequalities. *Set-Valued Var. Anal.* **30** (2022), no. 3, 873–922 Zbl 07563231 MR 4455150

[6] J.-P. Aubin, *Mathematical methods of game and economic theory*. Stud. Math. Appl. 7, North-Holland Publishing Co., Amsterdam-New York, 1979 Zbl 0452.90093 MR 556865

[7] C. Baiocchi and A. Capelo, *Variational and quasivariational inequalities*. Wiley-Intersci. Publ., John Wiley & Sons, New York, 1984 Zbl 0551.49007 MR 745619

[8] A. Bensoussan and J.-L. Lions, *Impulse control and quasivariational inequalities*. μ, Gauthier-Villars, Montrouge; Heyden & Son, Philadelphia, PA, 1984 MR 756234

[9] J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*. Springer Ser. Oper. Res., Springer, New York, 2000 Zbl 0966.49001 MR 1756264

[10] C. Christof and G. Wachsmuth, Lipschitz stability and Hadamard directional differentiability for elliptic and parabolic obstacle-type quasi-variational inequalities. *SIAM J. Control Optim.* **60** (2022), no. 6, 3430–3456 Zbl 1512.35323 MR 4522870

[11] M. Fukushima, Y. Oshima, and M. Takeda, *Dirichlet forms and symmetric Markov processes*. extended edn., De Gruyter Stud. Math. 19, Walter de Gruyter & Co., Berlin, 2011 Zbl 1227.31001 MR 2778606

[12] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical analysis of variational inequalities*. Stud. Math. Appl. 8, North-Holland Publishing Co., Amsterdam-New York, 1981 Zbl 0463.65046 MR 635927

[13] H. Goldberg, W. Kampowsky, and F. Tröltzsch, On Nemytskij operators in L_p-spaces of abstract functions. *Math. Nachr.* **155** (1992), 127–140 Zbl 0760.47031 MR 1231260

[14] M. Hintermüller and I. Kopacka, A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs. *Comput. Optim. Appl.* **50** (2011), no. 1, 111–145 Zbl 1229.49032 MR 2822818

[15] D. Kinderlehrer and G. Stampacchia, *An introduction to variational inequalities and their applications*. Pure and Applied Mathematics 88, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1980 Zbl 0457.35001 MR 567696

[16] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod, Paris; Gauthier-Villars, Paris, 1969 Zbl 0189.40603 MR 259693

[17] F. Mignot, Contrôle dans les inéquations variationelles elliptiques. *J. Functional Analysis* **22** (1976), no. 2, 130–185 Zbl 0364.49003 MR 423155

[18] F. Mignot and J.-P. Puel, Optimal control in some variational inequalities. *SIAM J. Control Optim.* **22** (1984), no. 3, 466–476 Zbl 0561.49007 MR 739836

[19] U. Mosco, Implicit variational problems and quasi variational inequalities. In *Nonlinear operators and the calculus of variations (Summer School, Univ. Libre Bruxelles, Brussels, 1975)*, pp. 83–156, Lecture Notes in Math. 543, Springer, Berlin-New York, 1976 Zbl 0346.49003 MR 513202

[20] J.-F. Rodrigues, *Obstacle problems in mathematical physics*. North-Holland Math. Stud. 134, North-Holland Publishing Co., Amsterdam, 1987 Zbl 0606.73017 MR 880369

[21] T. Roubíček, *Nonlinear partial differential equations with applications*. Internat. Ser. Numer. Math. 153, Birkhäuser, Basel, 2005 Zbl 1087.35002 MR 2176645

[22] A. Schiela and D. Wachsmuth, Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints. *ESAIM Math. Model. Numer. Anal.* **47** (2013), no. 3, 771–787 Zbl 1266.65112 MR 3056408

[23] S.-M. Stengl. Combined regularization and discretization of equilibrium problems and primal-dual gap estimators. 2021, arXiv:2110.02817

[24] L. Tartar, Inéquations quasi variationnelles abstraites. *C. R. Acad. Sci. Paris Sér. A* **278** (1974), 1193–1196 Zbl 0334.49003 MR 344964

[25] G. Wachsmuth, Strong stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Optim.* **24** (2014), no. 4, 1914–1932 Zbl 1328.49007 MR 3274378

[26] G. Wachsmuth, Towards M-stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Control Optim.* **54** (2016), no. 2, 964–986 Zbl 1337.49042 MR 3484394

[27] G. Wachsmuth, A guided tour of polyhedric sets: basic properties, new results on intersections, and applications. *J. Convex Anal.* **26** (2019), no. 1, 153–188 Zbl 1412.52006 MR 3847219

[28] G. Wachsmuth, Elliptic quasi-variational inequalities under a smallness assumption: uniqueness, differential stability and optimal control. *Calc. Var. Partial Differential Equations* **59** (2020), no. 2, article no. 82 Zbl 1444.49006 MR 4083198

[29] G. Wachsmuth, From resolvents to generalized equations and quasi-variational inequalities: existence and differentiability. *J. Nonsmooth Anal. Optim.* **3** (2022), article no. 8537
Zbl 1547.49019 MR 4409538

Received 14 February 2024; revised 26 February 2025.

Amal Alphonse
Weierstrass Institute, Mohrenstraße 39, 10117 Berlin, Germany; alphonse@wias-berlin.de

Michael Hintermüller
Weierstrass Institute, Mohrenstraße 39, 10117 Berlin; Humboldt-Universität zu Berlin, Unter den Linden 6, 10117 Berlin, Germany; hintermueller@wias-berlin.de

Carlos N. Rautenberg
Department of Mathematical Sciences and the Center for Mathematics and Artificial Intelligence (CMAI), George Mason University, 4400 University Drive, Fairfax, VA 22030, USA; crautenb@gmu.edu

Gerd Wachsmuth
Institute of Mathematics, Brandenburgische Technische Universität Cottbus-Senftenberg, 03046 Cottbus, Germany; gerd.wachsmuth@b-tu.de