TECHNISCHE UNIVERSITÄT
CHEMNITZ

Optimization problems with complementarity constraints in infinite-dimensional spaces

Cumulative habilitation thesis

Gerd Wachsmuth

November 29, 2016

TU Chemnitz, Faculty of Mathematics,
Research Group Numerical Mathematics (Partial Differential Equations)

http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-227446

Contents

Preface

This thesis is concerned with mathematical programs with complementarity constraints (MPCCs) in infinite-dimensional spaces. MPCCs are optimization problems, in which at least one set of constraints is of complementarity type, e.g.,

$$x \geq 0,\ y \geq 0,\ x\,y = 0$$

in the case of scalars $x, y \in \mathbb{R}$. These constraints violate all constraint qualifications (CQs) of reasonable strength. In particular, the Mangasarian-Fromovitz-CQ (MFCQ) does not hold at any feasible point. This renders the theoretical and numerical treatment of MPCCs challenging.

The content of this thesis is organized as follows. In Part I, we are going to study abstract MPCCs in Banach spaces. The main idea is to transfer the local decomposition approach from finite to infinite dimensions. In the case that the complementarity constraint is defined by a *polyhedric* cone, this is established in Chapter 1. By an additional linearization step, it is also possible to tackle the non-polyhedric situation and this is addressed in Chapter 2. Part I is based on the following publications.

1. G. Wachsmuth. Mathematical programs with complementarity constraints in Banach spaces. *Journal of Optimization Theory and Applications*, 166(2):480–507, 2015. doi: 10.1007/s10957-014-0695-3.
2. G. Wachsmuth. Strong stationarity for optimization problems with complementarity constraints in absence of polyhedricity. *Set-Valued and Variational Analysis*, 2016. doi: 10.1007/s11228-016-0370-y.

Part II deals with the optimal control of the obstacle problem. This is an MPCC in the space $H_0^1(\Omega)$. Under certain assumptions on the data, we are able to prove strong stationarity of local minimizers even in the presence of control constraints, see Chapter 3. If these conditions are not satisfied, we still obtain M-stationarity under a mild condition on the sequence of multipliers associated with a regularized problem, see Chapter 4. Part II is based on the following publications.

3. G. Wachsmuth. Strong stationarity for optimal control of the obstacle problem with control constraints. *SIAM Journal on Optimization*, 24(4):1914–1932, 2014. doi: 10.1137/130925827.
4. G. Wachsmuth. Towards M-stationarity for optimal control of the obstacle problem with control constraints. *SIAM Journal on Control and Optimization*, 54(2):964–986, 2016. doi: 10.1137/140980582.

Finally, we consider the notion of polyhedricity in Part III. Polyhedricity plays an important role in the theory for infinite-dimensional optimization and complementarity problems. Chapter 5 offers an introduction to this topic, new results and counterexamples. Finally, Chapter 6 provides new polyhedricity results for sets in vector-valued Sobolev spaces. Part III is based on the following publications.

5. G. Wachsmuth. A guided tour of polyhedric sets. Preprint, TU Chemnitz, 2016. Submitted.
6. G. Wachsmuth. Pointwise constraints in vector-valued Sobolev spaces. With applications in optimal control. *Applied Mathematics & Optimization*, 2016. doi: 10.1007/s00245-016-9381-1.

All these publications have been typeset by using the original LaTeX sources. Chapters 1 to 4 and 6 are identical to the published manuscripts except for minor changes. Moreover, cross-references between the chapters have been resolved.

All publications on which this thesis is based were written after the completion of the author's Ph.D. degree in December 2011. In the same period of time, the following publications were completed.

7. G. Wachsmuth. On LICQ and the uniqueness of Lagrange multipliers. *Operations Research Letters*, 41(1):78–80, 2013. doi: 10.1016/j.orl.2012.11.009.
8. G. Wachsmuth, M. Lätzer, and E. Leidich. Analytical computation of multiple interference fits under elasto-plastic deformations. *Zeitschrift für Angewandte Mathematik und Mechanik*, 2013. doi: 10.1002/zamm.201300041.
9. D. Wachsmuth and G. Wachsmuth. Necessary conditions for convergence rates of regularizations of optimal control problems. In Dietmar Hömberg and Fredi Tröltzsch, editors, *System Modeling and Optimization*, volume 391 of *IFIP Advances in Information and Communication Technology*, pages 145–154. Springer Berlin Heidelberg, 2013.
10. G. Wachsmuth. The numerical solution of Newton's problem of least resistance. *Mathematical Programming*, 147(1–2):331–350, 2014. doi: 10.1007/s10107-014-0756-2.
11. G. Wachsmuth. Differentiability of implicit functions. *Journal of Mathematical Analysis and Applications*, 414(1):259–272, 2014. doi: 10.1016/j.jmaa.2014.01.007.
12. D. Wachsmuth and G. Wachsmuth. Optimal control of an oblique derivative problem. *Annals of the Academy of Romanian Scientists*, 6(1):50–73, 2014.
13. R. Herzog, C. Meyer, and G. Wachsmuth. Optimal control of elastoplastic processes: Analysis, algorithms, numerical analysis and applications. In *Trends in PDE Constrained Optimization*, International Series of Numerical Mathematics, pages 27–41. Springer, 2014.
14. R. Herzog, J. Obermeier, and G. Wachsmuth. Annular and sectorial sparsity in optimal control of elliptic equations. *Computational Optimization and Applications*, 62(1):157–180, 2015. doi: 10.1007/s10589-014-9721-5.

15. E. Casas, R. Herzog, and G. Wachsmuth. Analysis of spatio-temporally sparse optimal control problems of semilinear parabolic equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 2015. doi: 10.1051/cocv/2015048.

16. R. Schneider and G. Wachsmuth. Achieving optimal convergence order for FEM in control constrained optimal control problems. *PAMM*, 15(1):731–734, 2015. doi: 10.1002/pamm.201510350.

17. S. Frankeser, S. Hiller, G. Wachsmuth, and J. Lutz. Using the on-state-Vbe sat-voltage for temperature estimation of SiC-BJTs during normal operation. In *PCIM Europe*, pages 132–139, 2015.

18. R. Schneider and G. Wachsmuth. A-posteriori error estimation for control-constrained, linear-quadratic optimal control problems. *SIAM Journal on Numerical Analysis*, 54(2):1169–1192, 2016. doi: 10.1137/15M1020460.

19. P. Mehlitz and G. Wachsmuth. Weak and strong stationarity in generalized bilevel programming and bilevel optimal control. *Optimization*, 65(5):907–935, 2016. doi: 10.1080/02331934.2015.1122007.

20. U. Stefanelli, D. Wachsmuth, and G. Wachsmuth. Optimal control of a rate-independent evolution equation via viscous regularization. *Discrete and Continuous Dynamical Systems - Series S*, 10(6):1467–1485, 2017. doi: 10.3934/dcdss.2017076.

21. P. Mehlitz and G. Wachsmuth. The limiting normal cone to pointwise defined sets in Lebesgue spaces. *Set-Valued and Variational Analysis*, pages 1–19, 2016. doi: 10.1007/s11228-016-0393-4.

22. B. González Merino, T. Jahn, and G. Wachsmuth. Hunting for reduced polytopes. Preprint, 2016. arXiv: 1607.08125.

23. A. Rösch and G. Wachsmuth. Mass lumping for the optimal control of elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 55(3):1412–1436, 2017. doi: 10.1137/16M1074473.

24. C. Schneider and G. Wachsmuth. Regularization and discretization error estimates for optimal control of ODEs with group sparsity. *ESAIM: Control, Optimisation and Calculus of Variations*, 2017. doi: 10.1051/cocv/2017049. To appear.

25. D. Wachsmuth and G. Wachsmuth. How not to discretize the control. *PAMM*, 16(1):793–794, 2016. doi: 10.1002/pamm.201610385.

26. G. Wachsmuth. Conforming approximation of convex functions with the finite element method. *Numerische Mathematik*, 2017. doi: 10.1007/s00211-017-0884-8. To appear.

Acknowledgement

The work on which this habilitation is based upon was carried out during my appointment at TU Chemnitz. I express my sincere gratitude to my mentor Prof. Roland Herzog, for

giving me the freedom to work on these interesting research topics, for his support and for many fruitful discussions and hints.

Special thanks go to all (current and former) members of the working group "Numerical mathematics (partial differential equations)": Prof. Roland Herzog, Dr. Martin Bernauer, Jan Blechschmidt, Tommy Etling, Dr. Andreas Günnel, Susann Mach, Felix Ospald, Ilka Riedel, Frank Schmidt and Ailyn Stötzner. It is great to work with you and to have coffee breaks with you. Moreover, I would like to thank all colleagues for the nice working atmosphere.

The last sentences belong to Claudia, Josephine and Alexander. I could not imagine that there is a life without you — at least, I would not call it "life". Thank you for everything.

Summary

In this thesis we consider optimization problems with complementarity constraints in infinite-dimensional spaces.

On the one hand, we deal with the general situation, in which the complementarity constraint is governed by a closed convex cone. We use the local decomposition approach, which is known from finite dimensions, to derive first-order necessary optimality conditions of strongly stationary type. In the non-polyhedric case, stronger conditions are obtained by an additional linearization argument.

On the other hand, we consider the optimal control of the obstacle problem. This is a classical example for a problem with complementarity constraints in infinite dimensions. We are concerned with the control-constrained case. Due to the lack of surjectivity, a system of strong stationarity is not necessarily satisfied for all local minimizers. We identify assumptions on the data of the optimal control problem under which strong stationarity of local minimizers can be verified. Moreover, without any additional assumptions on the data, we show that a system of M-stationarity is satisfied provided that some sequence of multipliers converges in capacity.

Finally, we also discuss the notion of polyhedric sets. These sets have many applications in infinite-dimensional optimization theory. Since the results concerning polyhedricity are scattered in the literature, we provide a review of the known results. Furthermore, we give some new results concerning polyhedricity of intersections and provide counterexamples which demonstrate that intersections of polyhedric sets may fail to be polyhedric. We also prove a new polyhedricity result for sets in vector-valued Sobolev spaces.

Part I.

Infinite-dimensional optimization problems with complementarity constraints

Contents of Part I

Introduction

In finite-dimensional optimization, an MPCC is an optimization problem of the form

$$\begin{aligned} \text{Minimize} \quad & f(x) \\ \text{w.r.t.} \quad & x \in \mathbb{R}^n \\ \text{such that} \quad & g(x) \leq 0, \\ & h(x) = 0, \\ & G(x) \geq 0, \\ & H(x) \geq 0, \\ & G(x)^\top H(x) = 0. \end{aligned}$$

Here, $f\colon \mathbb{R}^n \to \mathbb{R}$, $g\colon \mathbb{R}^n \to \mathbb{R}^k$, $h\colon \mathbb{R}^n \to \mathbb{R}^l$, $G, H\colon \mathbb{R}^n \to \mathbb{R}^m$ are differentiable functions. As already said, it can be checked that MFCQ is violated at all feasible points. Consequently, the standard Karush-Kuhn-Tucker (KKT) conditions may fail to be necessary optimality conditions and many stationarity conditions tailored to MPCCs have been designed in the literature.

The most strict stationarity system is given by the so-called system of strong stationarity. A feasible point $\bar{x}$ of the above MPCC is called strongly stationary, if there exist multipliers $\kappa \in \mathbb{R}^k$, $\lambda \in \mathbb{R}^l$, and $\mu, \nu \in \mathbb{R}^m$ such that

$$\begin{aligned} 0 &= f'(\bar{x}) + g'(\bar{x})^\top \kappa + h'(\bar{x})^\top \lambda + G'(\bar{x})^\top \mu + H'(\bar{x})^\top \nu, \\ \kappa &\geq 0, \quad \kappa^\top g(\bar{x}) = 0, \\ \mu_i &= 0, \quad \forall i \in I^{+0}, \qquad \mu_i \leq 0, \quad \forall i \in I^{00}, \\ \nu_i &= 0, \quad \forall i \in I^{0+}, \qquad \nu_i \leq 0, \quad \forall i \in I^{00}. \end{aligned}$$

Note that there is no multiplier corresponding to the constraint $G(x)^\top H(x) = 0$. In the above system, the index sets I^{+0}, I^{0+} and I^{00} are given by

$$\begin{aligned} I^{+0} &:= \{i \in \{1, \ldots, m\} : G_i(\bar{x}) > 0\}, \\ I^{0+} &:= \{i \in \{1, \ldots, m\} : H_i(\bar{x}) > 0\}, \\ I^{00} &:= \{i \in \{1, \ldots, m\} : G_i(\bar{x}) = H_i(\bar{x}) = 0\}. \end{aligned}$$

The weaker stationarity systems are obtained by choosing different conditions for μ_i, ν_i on the set of biactive indices $i \in I^{00}$.

In this chapter, we generalize the theory to MPCCs in which the complementarity is given by a closed, convex cone $K \subset Z$ and where the Banach space Z is assumed to be reflexive. That is, the complementarity constraint is given by

$$G(x) \in K, \quad H(x) \in K^\circ, \quad \langle G(x), H(x) \rangle = 0.$$

Note that the finite-dimensional MPCC can be obtained by setting $K = [0, \infty)^m$. Since we are not able to work with index sets in the general case governed by the cone K, it is not immediate how to transfer the system of strong stationarity.

In Chapter 1, we use the so-called local decomposition approach to derive optimality conditions in the general case. In the case that the cone K is polyhedric, these conditions possess a reasonable strength. The non-polyhedric situation is considered in Chapter 2. Using an additional linearization argument, we provide stronger optimality conditions. In the situations that K is the second-order cone or the semidefinite cone, we improve the results which are known from the literature.

1. Mathematical programs with complementarity constraints in Banach spaces

Abstract: We consider optimization problems in Banach spaces involving a complementarity constraint, defined by a convex cone K. By transferring the local decomposition approach, we define strong stationarity conditions and provide a constraint qualification, under which these conditions are necessary for optimality. To apply this technique, we provide a new uniqueness result for Lagrange multipliers in Banach spaces. In the case that the cone K is polyhedric, we show that our strong stationarity conditions possess a reasonable strength. Finally, we generalize to the case where K is not a cone and apply the theory to two examples.

Keywords: strong stationarity, mathematical program with complementarity constraints, polyhedricity, optimality conditions

MSC: 49K27, 46N10, 90C33

1.1. Introduction

Mathematical programs with complementarity constraints (MPCCs) are well understood, both theoretically and numerically; we refer to Luo, Pang, Ralph, 1996; Scheel, Scholtes, 2000; Hoheisel, Kanzow, Schwartz, 2013 and the references therein. We are interested in generalizations of the standard complementarity constraints for both finite and infinite-dimensional problems. In particular, we will consider a conic complementarity constraint defined by a *closed, convex cone* in a *reflexive Banach space*, see below for the definitions. The standard case is obtained by using the non-negative orthant. Optimality conditions of optimization problems with conic constraints, as well as the solution of variational inequalities, can be modeled by conic complementarity constraints. To our knowledge, there are no references concerning MPCCs of this general type.

In the case of a standard MPCC, the tightest optimality condition is the system of strong stationarity; see Theorem 2 in Scheel, Scholtes, 2000 and (1.3.1). Our main interest is to obtain similar results for conic MPCCs.

In the *finite-dimensional* case, there exist some very recent contributions for special cases of the cone defining the complementarity (listed in the order of increasing generality):

- the second-order (Lorentz) cone: Outrata, D. Sun, 2008; Liang, Zhu, Lin, 2014 and the references therein.

- the cone of semidefinite matrices: Ding, D. Sun, Ye, 2014; Wu, L. Zhang, Y. Zhang, 2014.
- a symmetric cone in a Euclidean Jordan algebra: Yan, Fukushima, 2011.

We mention that all these contributions except Yan, Fukushima, 2011 contain optimality systems which can be interpreted as strong stationarity.

In the *infinite-dimensional* case, only special cases have been discussed in the literature. To keep the presentation concise, we focus on results concerning strong stationarity. The first such result was obtained in the seminal work Mignot, 1976. The results and proofs were given for the special case of certain cones in Dirichlet spaces, but the generalization to polyhedric cones in general Hilbert spaces is straightforward. However, this approach is limited to a specific structure of the optimization problem, namely that the conic complementarity constraint represents a variational inequality of *first kind*. We refer to (1.6.5) for the presentation of this result in the abstract Hilbert space setting. The same result was reproduced in Hintermüller, Surowiec, 2011 by techniques from variational analysis, and a special case was proven in Outrata, Jarušek, Stará, 2011.

Besides these results, which cover a broad class of problems, there are only two other contributions providing systems of strong stationarity. In Herzog, C. Meyer, G. Wachsmuth, 2013, the authors considered an optimal control problem arising in elasto-plasticity, and in G. Wachsmuth, 2014 (i.e., Chapter 3) the author studies a control constrained optimal control problem governed by the obstacle problem. We also mention that, except Herzog, C. Meyer, G. Wachsmuth, 2013, all these results in infinite dimensions involve polyhedric cones. The definition of polyhedric cones is recalled in Section 1.5.2. Finally, we mention that the case of a certain variational inequality of *second kind* is studied in De los Reyes, C. Meyer, 2016.

One of the main contributions of this work is the definition of strong stationarity conditions for problems involving general conic complementarity constraints; see Definition 1.5.1. In particular, we do not rely on a specific structure of the complementarity condition and we can treat the case of reflexive Banach spaces instead of Hilbert spaces. We briefly mention that the strong stationarity conditions (1.3.1) in the standard case involve various index sets and hence, it is not immediately clear how these conditions can be transfered to the more general conic case.

Moreover, we provide constraint qualifications under which our strong stationarity condition is necessary for optimality; see Section 1.5.3. In difference to the work by Mignot, which involves the implicit-programming approach, we use the local decomposition approach; see, e.g., Luo, Pang, Ralph, 1996; Pang, Fukushima, 1999; Scheel, Scholtes, 2000; Flegel, Kanzow, 2005a; b. As a prerequisite, we provide a new constraint qualification, which is equivalent to the uniqueness of Lagrange multipliers in Banach spaces; see Theorem 1.4.2. This result is also of independent interest.

Under the assumption that the cone defining the complementarity constraint is polyhedric, we show that our optimality condition is equivalent to the so-called linearized B-stationarity; see Theorem 1.5.4 and Lemma 1.5.5. That is, our work extends the theory concerning strong stationarity of standard MPCCs to conic complementarity constraints

defined by a polyhedric cone. In the non-polyhedric case, our optimality condition is still necessary for optimality, provided that a constraint qualification holds. Stronger optimality conditions can be obtained by using an additional linearization argument; see Section 1.6.2.

We also generalize the results to the case that the complementarity constraint is defined by a closed, convex set (and not necessarily by a cone); see Section 1.5.4.

This work is organized as follows. In Section 1.2, we fix some notation. The verification of strong stationarity for standard MPCCs is recalled in Section 1.3. In Section 1.4, we provide a constraint qualification, which is equivalent to the uniqueness of Lagrange multipliers in infinite dimensions, similar to the linear independence constraint qualification in the finite-dimensional case. Section 1.5 is devoted to the main results of this paper; in particular, we define the system of strong stationarity and give constraint qualifications, which render this system necessary for optimality. Finally, we apply the theory to two examples in Section 1.6.

1.2. Notation

Let X be a (real) Banach space. The (norm) closure of a subset $A \subset X$ is denoted by $\mathrm{cl}(A)$. The linear subspace spanned by A is denoted by $\mathrm{lin}(A)$. The duality pairing between X and its topological dual $X^\star$ is denoted by $\langle \cdot, \cdot \rangle : X^\star \times X \to \mathbb{R}$. For subsets $A \subset X$, $B \subset X^\star$, we define their polar cones and annihilators via

$$A^\circ := \{x^\star \in X^\star : \langle x^\star, x\rangle \leq 0, \forall x \in A\}, \quad B^\circ := \{x \in X : \langle x^\star, x\rangle \leq 0, \forall x^\star \in B\},$$
$$A^\perp := \{x^\star \in X^\star : \langle x^\star, x\rangle = 0, \forall x \in A\}, \quad B^\perp := \{x \in X : \langle x^\star, x\rangle = 0, \forall x^\star \in B\}.$$

For convex subsets $C \subset X$ and $x \in C$, we define the cone of feasible directions (sometimes called the radial cone) and the tangent cone by

$$\mathcal{R}_C(x) := \bigcup_{\lambda > 0} \lambda\,(C - x), \qquad \text{and} \qquad \mathcal{T}_C(x) := \mathrm{cl}(\mathcal{R}_C(x)),$$

respectively. In the special case that C is additionally a cone, we find

$$\mathcal{R}_C(x) = C + \mathrm{lin}(x), \qquad \text{and} \qquad \mathcal{T}_C(x) = \mathrm{cl}(C + \mathrm{lin}(x)), \tag{1.2.1}$$

where $\mathrm{lin}(x)$ is the linear subspace spanned by the element $x \in X$; see Example 2.62 in Bonnans, Shapiro, 2000. Moreover, one has

$$\mathcal{T}_C(x)^\circ = C^\circ \cap x^\perp \tag{1.2.2}$$

in this case. Here, $x^\perp$ is short for $\{x\}^\perp$. For closed, convex $C \subset X$, we define the critical cone w.r.t. $x \in C$ and $v \in \mathcal{T}_C(x)^\circ$ by

$$\mathcal{K}_C(x, v) := \mathcal{T}_C(x) \cap v^\perp. \tag{1.2.3}$$

1.3. Strong stationarity for standard MPCCs

In order to motivate the steps, which will be taken in Section 1.5, we briefly recall some results for standard MPCCs. We consider the program

$$\begin{aligned} \text{Minimize} \quad & f(x), \quad \text{w.r.t. } x \in \mathbb{R}^n, \\ \text{s.t.} \quad & G(x) \geq 0,\ H(x) \geq 0,\ G(x)^\top H(x) = 0. \end{aligned} \tag{sMPCC}$$

Here, $f : \mathbb{R}^n \to \mathbb{R}$, $G, H : \mathbb{R}^n \to \mathbb{R}^m$ are assumed to be continuously differentiable and $n, m \geq 1$. For simplicity, we did not include any additional equality or inequality constraints. They can, however, be added in a straightforward way. The prefix "s" in (sMPCC) is short for "standard".

An important technique to derive optimality conditions is the local decomposition approach; see, e.g., Luo, Pang, Ralph, 1996; Pang, Fukushima, 1999; Scheel, Scholtes, 2000; Flegel, Kanzow, 2005a; b. This technique involves several auxiliary problems. Given a feasible point $\bar{x} \in \mathbb{R}^n$, we define the index sets (suppressing the dependence on $\bar{x}$)

$$\begin{aligned} I^{+0} &:= \{i \in \{1, \ldots, m\} : G_i(\bar{x}) > 0\}, \quad I^{0+} := \{i \in \{1, \ldots, m\} : H_i(\bar{x}) > 0\}, \\ I^{00} &:= \{i \in \{1, \ldots, m\} : G_i(\bar{x}) = H_i(\bar{x}) = 0\}. \end{aligned}$$

We are going to introduce four auxiliary problems, depending on these index sets, and thus implicitly on $\bar{x}$. These auxiliary problems are standard nonlinear programs (NLPs). The *relaxed NLP* is given by

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & G_i(x) \geq 0 \text{ for } i \in I^{+0} \cup I^{00}, \quad G_i(x) = 0 \text{ for } i \in I^{0+}, \\ & H_i(x) \geq 0 \text{ for } i \in I^{0+} \cup I^{00}, \quad H_i(x) = 0 \text{ for } i \in I^{+0}. \end{aligned} \tag{sRNLP}$$

The *tightened NLP* is given by

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & G_i(x) \geq 0 \text{ for } i \in I^{+0}, \quad G_i(x) = 0 \text{ for } i \in I^{0+} \cup I^{00}, \\ & H_i(x) \geq 0 \text{ for } i \in I^{0+}, \quad H_i(x) = 0 \text{ for } i \in I^{+0} \cup I^{00}. \end{aligned} \tag{sTNLP}$$

Finally, we introduce

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & G_i(x) \geq 0 \text{ for } i \in I^{+0}, \quad G_i(x) = 0 \text{ for } i \in I^{0+} \cup I^{00}, \\ & H_i(x) \geq 0 \text{ for } i \in I^{0+} \cup I^{00}, \quad H_i(x) = 0 \text{ for } i \in I^{+0}, \end{aligned} \tag{sNLP$_G$}$$

and

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & G_i(x) \geq 0 \text{ for } i \in I^{+0} \cup I^{00}, \quad G_i(x) = 0 \text{ for } i \in I^{0+}, \\ & H_i(x) \geq 0 \text{ for } i \in I^{0+}, \quad H_i(x) = 0 \text{ for } i \in I^{+0} \cup I^{00}. \end{aligned} \tag{sNLP$_H$}$$

Here, the nonlinear programs (sNLP$_G$) and (sNLP$_H$) are the extreme cases of the *restricted NLPs*, which are often denoted by NLP$_*(\beta_1, \beta_2)$; see, e.g., Pang, Fukushima, 1999; Flegel, Kanzow, 2005a. For our analysis it will be sufficient to consider only these extreme cases.

The feasible set of (sMPCC) is locally contained in the feasible set of (sRNLP), whereas the feasible sets of the last three problems are contained in the feasible set of (sMPCC). Hence, if $\bar{x}$ is a local minimizer of (sMPCC), then it is also a local minimizer of these auxiliary problems. Moreover, all these nonlinear programs possess the same Lagrangian, the so-called MPCC-Lagrangian, and this MPCC-Lagrangian does not include a multiplier for the complementarity constraint $G(x)^\top H(x) = 0$.

A feasible point $\bar{x}$ of (sMPCC) is said to be strongly stationary, iff $\bar{x}$ is a Karush-Kuhn-Tucker (KKT) point of (sRNLP). That is, we require the existence of multipliers $\mu, \nu \in \mathbb{R}^m$ such that

$$0 = f'(\bar{x}) + G'(\bar{x})^\top \mu + H'(\bar{x})^\top \nu, \tag{1.3.1a}$$

$$\mu_i = 0, \quad \forall i \in I^{+0}, \qquad \mu_i \leq 0, \quad \forall i \in I^{00}, \tag{1.3.1b}$$

$$\nu_i = 0, \quad \forall i \in I^{0+}, \qquad \nu_i \leq 0, \quad \forall i \in I^{00}. \tag{1.3.1c}$$

It is easy to verify that these conditions are equivalent to $\bar{x}$ being a KKT point of (sMPCC) itself; see also Lemma 1.5.2. Note that a Lagrange multiplier of (sMPCC) contains an additional scalar multiplier for the constraint $G(x)^\top H(x) = 0$. The set of Lagrange multipliers associated to (sMPCC) is always unbounded since the Mangasarian-Fromovitz constraint qualification is violated, whereas the Lagrange multipliers for the relaxed problem (sRNLP) may be bounded.

The following result is well known; see, e.g., Scheel, Scholtes, 2000; Flegel, Kanzow, 2005a; b and the references therein. Nevertheless, we briefly re-state its proof, since we are going to transfer it to the infinite-dimensional case in Section 1.5; see in particular Theorem 1.5.6.

Theorem 1.3.1. Let $\bar{x}$ be a local minimizer of (sMPCC), such that (sTNLP) satisfies the linear independence constraint qualification. Then, $\bar{x}$ is strongly stationary.

Proof. Since (sTNLP), (sNLP$_G$) and (sNLP$_H$) have the same active constraints at $\bar{x}$, the linear independence constraint qualification also holds for (sNLP$_G$) and (sNLP$_H$). Moreover, since $\bar{x}$ is a local minimizer of (sMPCC), it is also a local minimizer of (sNLP$_G$) and (sNLP$_H$). Hence, there exist Lagrange multipliers (μ^G, ν^G) for (sNLP$_G$) and (μ^H, ν^H) for (sNLP$_H$).

Now, it is easy to verify that both pairs of multipliers are also multipliers for (sTNLP). Since the satisfaction of the linear independence constraint qualification implies the uniqueness of these multipliers, we have $(\mu^G, \nu^G) = (\mu^H, \nu^H)$. Collecting the sign conditions of the KKT systems of (sNLP$_G$) and (sNLP$_H$), we find that this pair is also a Lagrange multiplier for (sRNLP). Hence, $\bar{x}$ is strongly stationary.

It is also possible to assume the existence of a Lagrange multiplier for (sTNLP) and require that this multiplier satisfies the strict Mangasarian-Fromovitz condition for the relaxed problem (sTNLP) in order to infer the strong stationarity of $\bar{x}$; see, e.g., Theorem 2 in Scheel, Scholtes, 2000. The strict Mangasarian-Fromovitz condition, however, depends implicitly on the objective f. Therefore, it is not a constraint qualification, but a regularity condition.

1.4. Uniqueness of Lagrange multipliers in Banach spaces

One of the main ingredients in the proof of Theorem 1.3.1 is the well-known result that the linear independence constraint qualification implies the uniqueness of multipliers. In this section, we provide an analogous result for the infinite-dimensional, nonlinear program

$$\text{Minimize} \quad f(x), \quad \text{s.t.} \quad \mathcal{G}(x) \in \mathcal{C}. \tag{1.4.1}$$

Here, $\mathcal{X}$ and $\mathcal{Y}$ are (real) Banach spaces, $f : \mathcal{X} \to \mathbb{R}$ and $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ are Fréchet differentiable and $\mathcal{C} \subset \mathcal{Y}$ is a closed, convex set. The constraint $(\mathcal{G}, \mathcal{C})$ and a feasible point $\bar{x}$ are fixed throughout this section, but we will use objectives f belonging to the set

$$\mathcal{F} := \{f : \mathcal{X} \to \mathbb{R} : f \text{ is Fréchet differentiable}\}.$$

Note that the feasible point $\bar{x}$ may not be a local minimizer of (1.4.1) for all choices of $f \in \mathcal{F}$.

The aim of this section is to state a constraint qualification (i.e., a condition depending only on $\mathcal{G}$, $\mathcal{C}$ and $\bar{x} \in \mathcal{X}$) which implies that Lagrange multipliers for (1.4.1) at the feasible point $\bar{x}$ are unique. As usual, $\lambda \in \mathcal{Y}^\star$ is called a Lagrange multiplier for the objective f at the $\bar{x} \in \mathcal{X}$, if

$$f'(\bar{x}) + \mathcal{G}'(\bar{x})^\star \lambda = 0, \qquad \text{and} \qquad \lambda \in \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ.$$

Here, $\mathcal{G}'(\bar{x})^\star \in \mathcal{L}(\mathcal{Y}^\star, \mathcal{X}^\star)$ denotes the adjoint of $\mathcal{G}'(\bar{x})$.

We start by giving an auxiliary result.

Lemma 1.4.1. Let $\bar{x} \in \mathcal{X}$ be a feasible point of (1.4.1). Then, the following conditions are equivalent.

(a) For all $f \in \mathcal{F}$, there exists at most one Lagrange multiplier of (1.4.1) at $\bar{x}$.

(b) For all $f \in \mathcal{F}$, such that $\bar{x}$ is a local minimum of (1.4.1), there exists at most one Lagrange multiplier at $\bar{x}$.

(c) We have

$$\ker \mathcal{G}'(\bar{x})^\star \cap \operatorname{lin}(\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ) = \{0\}. \tag{1.4.2}$$

Proof. "(a) ⇒ (b)": This implication is obvious.

"(b) ⇒ (c)": We note that $\operatorname{lin}(\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ) = \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ - \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ$, since $\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ$ is a convex cone. Now, let

$$\lambda_1 - \lambda_2 \in \ker \mathcal{G}'(\bar{x})^\star \cap \operatorname{lin}(\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ) \tag{1.4.3}$$

with $\lambda_1, \lambda_2 \in \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ$ be arbitrary. We set $f(x) := -\langle \lambda_1, \mathcal{G}(x) \rangle$ for all $x \in \mathcal{X}$. This implies $f(x) - f(\bar{x}) = -\langle \lambda_1, \mathcal{G}(x) - \mathcal{G}(\bar{x}) \rangle \geq 0$ for all $x \in \mathcal{X}$ which are feasible for (1.4.1). Hence, $\bar{x}$ is a local minimum of (1.4.1) for this choice of f.

By definition of f, we have $f'(\bar{x}) + \mathcal{G}'(\bar{x})^\star \lambda_1 = 0$, which shows that λ_1 is a Lagrange multiplier. Now, (1.4.3) implies that λ_2 is also a Lagrange multiplier. Assertion (b) yields $\lambda_1 = \lambda_2$. This implies assertion (c).

"(c) ⇒ (a)": Let $f : \mathcal{X} \to \mathbb{R}$ and two multipliers $\lambda_1, \lambda_2 \in \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ$ be given. We have $f'(\bar{x}) + \mathcal{G}'(\bar{x})^\star \lambda_i = 0$ for $i \in \{1, 2\}$. Thus, $\mathcal{G}'(\bar{x})^\star(\lambda_1 - \lambda_2) = 0$. Now, assertion (c) implies $\lambda_1 = \lambda_2$. Hence, there exists at most one Lagrange multiplier.

The condition (c) in Lemma 1.4.1 is stated in the dual space $\mathcal{Y}^\star$. In order to obtain an equivalent statement in the primal space $\mathcal{Y}$, we need an additional condition. This is made precise in the following theorem.

Theorem 1.4.2. Let

$$\operatorname{cl}(\mathcal{G}'(\bar{x})\,\mathcal{X} - \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^{\circ\perp}) = \mathcal{Y} \tag{1.4.4}$$

be satisfied. Then, (1.4.2) holds. Conversely, suppose (1.4.2) and

$$\operatorname{lin}(\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ) \text{ is closed in the weak-}\star \text{ topology of } \mathcal{Y}^\star. \tag{1.4.5}$$

Then, (1.4.4) is satisfied.

Proof. We define $A := \ker \mathcal{G}'(\bar{x})^\star$ and $B := \operatorname{lin}(\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ)$, which are subspaces in $\mathcal{Y}^\star$. We find $A^\perp = \operatorname{cl}(\mathcal{G}'(\bar{x})\,\mathcal{X})$ and $B^\perp = \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^{\circ\perp}$. Now, assumption (1.4.4) is equivalent to $\operatorname{cl}(A^\perp - B^\perp) = \mathcal{Y}$, whereas (1.4.2) reads $A \cap B = \{0\}$. For arbitrary subspaces $A, B \subset \mathcal{Y}^\star$, we have

$$\operatorname{cl}(A^\perp - B^\perp) = (\operatorname{cl}_\star A \cap \operatorname{cl}_\star B)^\perp \subset (A \cap B)^\perp, \tag{1.4.6}$$

see, e.g., (2.32) in Bonnans, Shapiro, 2000. Here, $\operatorname{cl}_\star(A)$ denotes the closure of A in the weak-$\star$ topology of $\mathcal{Y}^\star$. Now, the assertion of the theorem follows since A is weak-$\star$ closed.

Note that for a closed, convex cone K, the linear space $K^{\circ\perp}$, which appears in (1.4.4), is just the lineality space $K \cap -K$, which is the largest linear subspace contained in K.

We emphasize that (1.4.4) is always sufficient for the uniqueness of multipliers. The additional assumption (1.4.5) is only needed for the necessity. Moreover, (1.4.5) is always satisfied for a finite-dimensional problem. In infinite dimensions, this assumption is satisfied in some situations. In particular, if $\mathcal{C}$ is the cone of non-negative elements in $L^2(\Omega)$ or $H^{-1}(\Omega)$, then the set $\operatorname{lin}(\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ)$ is closed. However, this is not true for the cone K of non-negative functions in $H_0^1(\Omega)$, since $\operatorname{lin}(K^\circ)$ is not closed, but rather dense in $H^{-1}(\Omega)$.

Remark 1.4.3.

(a) In finite dimensions, the condition (1.4.4) reduces to the so-called non-degeneracy; see, e.g., (4.172) in Bonnans, Shapiro, 2000. In particular, if $\mathcal{C} = \mathbb{R}^n_+$, then (1.4.4) is equivalent to the linear independence constraint qualification. Moreover, the linear hull of a closed, convex cone is always closed in finite dimensions. Hence, Theorem 1.4.2 and Lemma 1.4.1 reduce to the well-known fact that the linear independence constraint qualification is equivalent to the uniqueness of multipliers for arbitrary objectives; see also Theorem 2 in G. Wachsmuth, 2013.

(b) Note that, unlike in finite dimensions, the constraint qualification (1.4.4) does not imply the *existence* of multipliers and neither do the conditions of Lemma 1.4.1. The constraint qualification of Robinson-Zowe-Kurcyusz (assuming $\mathcal{G}$ continuously differentiable) for (1.4.1) reads

$$\mathcal{G}'(\bar{x})\,\mathcal{X} - \mathcal{R}_{\mathcal{C}}(\mathcal{G}(\bar{x})) = \mathcal{Y}, \tag{1.4.7}$$

see (1.4) in Zowe, Kurcyusz, 1979, Theorem 3.1 in Bonnans, Shapiro, 1998. This condition is similar to (1.4.4). However, the cones $\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^{\circ\perp}$ and $\mathcal{R}_{\mathcal{C}}(\mathcal{G}(\bar{x}))$ do, in general, not contain each other.

For a standard, finite-dimensional nonlinear program, the cone $\mathcal{C}$ is

$$\mathcal{C} = \{x \in \mathbb{R}^{n+m} : x_i \le 0, i = 1, \dots, n\}.$$

In this case we have $\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^{\circ\perp} \subset \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x})) = \mathcal{R}_{\mathcal{C}}(\mathcal{G}(\bar{x}))$. Hence, (1.4.4) implies (1.4.7) and, in turn, the existence of multipliers. In the general case of $\mathcal{C}$ being an arbitrary, closed, convex cone in finite dimensions, condition (1.4.4) also implies the existence of minimizers; see the discussion following (4.177) in Bonnans, Shapiro, 2000.

(c) Let us compare our result with Shapiro, 1997a. In this paper, the author studies the question whether *a given* Lagrange multiplier λ is unique. The resulting conditions depend on the Lagrange multiplier λ and, hence, implicitly on the objective f.

Thus, the relation between our constraint qualification (1.4.4) to the conditions of Shapiro, 1997a is similar to the relation between the linear independence constraint qualification and the strict Mangasarian-Fromovitz condition in finite dimensions; see also the discussion in Section 5 in G. Wachsmuth, 2013.

It remains to give an example with unique multipliers, where (1.4.4) is violated. Let $\mathcal{X} := \mathcal{Y} := \ell^2$. We set

$$\mathcal{C} := \{y \in \ell^2 : \sum_{i=1}^{n} y_i \leq 0,\ \forall n \in \mathbb{N}\}.$$

A straightforward calculation shows

$$\mathcal{C}^\circ = \{y \in \ell^2 : y_n \geq y_{n+1},\ \forall n \in \mathbb{N}\}.$$

It is easy to see that $\operatorname{lin}(\mathcal{C}^\circ)$ is dense in ℓ^2. However, for

$$\tilde{y}_i := 1/k, \text{ if } i = k^2 \text{ for some } k \in \mathbb{N}, \qquad \tilde{y}_i := 0, \text{ else},$$

we have $\tilde{y} \in \ell^2 \setminus \operatorname{lin}(\mathcal{C}^\circ)$. Hence, $\operatorname{lin}(\mathcal{C}^\circ)$ cannot be closed.

Now, let $\{\tilde{y}/\|\tilde{y}\|_{\ell^2}, y^{(1)}, y^{(2)}, \ldots\}$ be an orthonormal basis of ℓ^2. We define the bounded, linear map $\mathcal{G} : \ell^2 \to \ell^2$, $x \mapsto \sum_{i=1}^{n} x_i\, y^{(i)}$. Its (Hilbert space) adjoint is given by $(\mathcal{G}^\star x)_i = (y^{(i)}, x)_{\ell^2}$. Since $\mathcal{G}$ is linear, we have $\mathcal{G}'(\bar{x}) = \mathcal{G}$. We set $\bar{x} := 0$, which implies $\mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^\circ = \mathcal{C}^\circ$. Consequently,

$$\operatorname{cl}\left(\mathcal{G}'(\bar{x})\,\ell^2 - \mathcal{T}_{\mathcal{C}}(\mathcal{G}(\bar{x}))^{\circ\perp}\right) = \operatorname{cl}\left(\mathcal{G}\,\ell^2 - \mathcal{C}^{\circ\perp}\right) = \operatorname{cl}\left(\mathcal{G}\,\ell^2\right) = \{\tilde{y}\}^\perp \neq \ell^2.$$

Hence, (1.4.4) is violated at the feasible point $\bar{x} = 0$.

Nevertheless, we can show that Lagrange multipliers for (1.4.1) with this choice of $\mathcal{C}$ and $\mathcal{G}$ are unique. By construction we have $\ker \mathcal{G}^\star = \operatorname{lin}(\tilde{y})$, and hence $\operatorname{lin}(\mathcal{C}^\circ) \cap \ker \mathcal{G}^\star = \{0\}$, which shows (1.4.2). Lemma 1.4.1 yields the uniqueness of multipliers.

1.5. MPCCs in Banach spaces

This section is devoted to the optimization problem with complementarity constraints

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & g(x) \in C,\ G(x) \in K,\ H(x) \in K^\circ,\ \langle G(x), H(x)\rangle = 0. \end{aligned} \tag{MPCC}$$

Here, $f : X \to \mathbb{R}$ is Fréchet differentiable, $g : X \to Y$, $G : X \to Z$ and $H : X \to Z^\star$ are continuously Fréchet differentiable, X, Y, Z are (real) Banach spaces and Z is assumed to be reflexive. Moreover, $C \subset Y$ is a closed, convex set and $K \subset Z$ is a closed, convex cone.

Due to the reflexivity of Z, the problem (MPCC) is symmetric w.r.t. G and H.

A straightforward computation shows that the Robinson-Zowe-Kurcyusz constraint qualification cannot be satisfied at any feasible point. This is similar to the violation of Mangasarian-Fromovitz constraint qualification for standard MPCCs. Hence, the KKT conditions may fail to be necessary for optimality. Therefore, the aim of this section is to provide a stationarity concept for (MPCC) and a constraint qualification, which renders this condition a necessary optimality condition for local minimizers of (MPCC).

The notion of strong stationarity is introduced in Section 1.5.1 by using auxiliary problems similar to those in Section 1.3. By proving some results on polyhedric cones in Section 1.5.2, we show that strong stationarity appears to be a "good" condition if the cone K is polyhedric. We apply the constraint qualification from Section 1.4 to give conditions which render strong stationarity a necessary condition for optimality; see Section 1.5.3. Finally, we generalize to the case that K is not a cone; see Section 1.5.4.

1.5.1. Auxiliary problems and optimality conditions

In this section, we will transfer the ideas of Section 1.3 to the infinite-dimensional case. We start by introducing the relaxations of (MPCC) at the feasible point $\bar{x} \in X$.

We observe that the inequality constraints on $G(x)$ in (sRNLP) can be written as

$$G(x) \geq 0 \qquad G(x)^\top H(\bar{x}) = 0,$$

and similarly for $H(x)$. This formulation, which does not involve the index sets I^{+0}, I^{00}, and I^{0+} is essential, since such index sets are not available for our general program (MPCC). The reformulation motivates the use of

$$G(x) \in K \cap H(\bar{x})^\perp \quad \text{and} \quad H(x) \in K^\circ \cap G(\bar{x})^\perp$$

in the definition of the relaxed NLP. Since K and K° are closed, convex cones, and since Z is reflexive, we have

$$K \cap H(\bar{x})^\perp = \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ \quad \text{and} \quad K^\circ \cap G(\bar{x})^\perp = \mathcal{T}_K(G(\bar{x}))^\circ,$$

see (1.2.2). Hence, we define the relaxed NLP by

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & g(x) \in C, \quad G(x) \in \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ, \quad H(x) \in \mathcal{T}_K(G(\bar{x}))^\circ. \end{aligned} \tag{RNLP}$$

The feasible sets of the remaining auxiliary programs must be contained in the feasible set of (MPCC), cf. the proof of Theorem 1.3.1. To ensure the complementarity, we require $G(x)$ or $H(x)$ to be perpendicular to the entire feasible set of $H(x)$ or $G(x)$ of (RNLP), respectively. In particular, we define

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & g(x) \in C, \quad G(x) \in \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ \cap \mathcal{T}_K(G(\bar{x}))^{\circ\perp}, \\ \text{and} \quad & H(x) \in \mathcal{T}_K(G(\bar{x}))^\circ, \end{aligned} \tag{NLP$_G$}$$

and

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & g(x) \in C, \quad G(x) \in \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ, \\ \text{and} \quad & H(x) \in \mathcal{T}_K(G(\bar{x}))^\circ \cap \mathcal{T}_{K^\circ}(H(\bar{x}))^{\circ\perp}. \end{aligned} \tag{NLP$_H$}$$

Finally, the feasible set of the tightened NLP is the intersection of the feasible sets of (NLP_G) and (NLP_H), i.e.,

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & g(x) \in C, \quad G(x) \in \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ \cap \mathcal{T}_K(G(\bar{x}))^{\circ\perp}, \\ \text{and} \quad & H(x) \in \mathcal{T}_K(G(\bar{x}))^\circ \cap \mathcal{T}_{K^\circ}(H(\bar{x}))^{\circ\perp}. \end{aligned} \tag{TNLP}$$

We emphasize that these NLP relaxations coincide with those of Section 1.3 in the case of a standard MPCC. Moreover, the point $\bar{x}$ is feasible for all auxiliary problems.

As in Section 1.3, we define strong stationarity via the KKT conditions of the relaxed NLP.

Definition 1.5.1 (Strong stationarity)**.** A feasible point $\bar{x}$ of (MPCC) is called strongly stationary iff it is a KKT point of (RNLP), i.e., iff there exist Lagrange multipliers $\lambda \in Y^\star$, $\mu \in Z^\star$ and $\nu \in Z$, such that

$$0 = f'(\bar{x}) + g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu, \tag{1.5.1a}$$
$$\lambda \in \mathcal{T}_C(g(\bar{x}))^\circ, \tag{1.5.1b}$$
$$\mu \in \mathcal{T}_{K^\circ}(H(\bar{x})) \cap G(\bar{x})^\perp = \mathcal{K}_{K^\circ}(H(\bar{x}), G(\bar{x})), \tag{1.5.1c}$$
$$\nu \in \mathcal{T}_K(G(\bar{x})) \cap H(\bar{x})^\perp \ = \mathcal{K}_K(G(\bar{x}), H(\bar{x})). \tag{1.5.1d}$$

Here, we used the critical cones $\mathcal{K}_{K^{(\circ)}}(\cdot,\cdot)$ defined in (1.2.3).

In contrast to the finite-dimensional case, strong stationarity of $\bar{x}$ is, in general, not equivalent to $\bar{x}$ being a classical KKT point of (MPCC).

Lemma 1.5.2. A feasible point $\bar{x}$ of problem (MPCC) is a classical KKT point of (MPCC) if and only if there exist Lagrange multipliers $\lambda \in Y^\star$, $\mu \in Z^\star$ and $\nu \in Z$ satisfying (1.5.1a), (1.5.1b), and

$$\mu \in \mathcal{R}_{K^\circ}(H(\bar{x})) \cap G(\bar{x})^\perp, \qquad \nu \in \mathcal{R}_K(G(\bar{x})) \cap H(\bar{x})^\perp. \tag{1.5.2}$$

Proof. Let $\bar{x}$ be a KKT point of (MPCC). That is, there exist multipliers $\lambda \in Y^\star$, $\tilde{\mu} \in Z^\star$, $\tilde{\nu} \in Z$ and $\tilde{\xi} \in \mathbb{R}$, such that

$$f'(x) + g'(x)^\star \lambda + G'(x)^\star \tilde{\mu} + H'(x)^\star \tilde{\nu} + \tilde{\xi}\,[G'(x)^\star H(x) + H'(x)^\star G(x)] = 0, \tag{1.5.3a}$$
$$\lambda \in \mathcal{T}_C(g(\bar{x}))^\circ, \qquad \tilde{\mu} \in \mathcal{T}_K(G(\bar{x}))^\circ, \qquad \tilde{\nu} \in \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ. \tag{1.5.3b}$$

By setting $\mu := \tilde{\mu} + \tilde{\xi}\, H(\bar{x})$, and $\nu := \tilde{\nu} + \tilde{\xi}\, G(\bar{x})$, (1.5.1a) is satisfied and we find

$$\mu \in K^\circ \cap G(\bar{x})^\perp + \operatorname{lin}(H(\bar{x})) = \mathcal{R}_{K^\circ}(H(\bar{x})) \cap G(\bar{x})^\perp,$$
$$\nu \in K \cap H(\bar{x})^\perp + \operatorname{lin}(G(\bar{x})) = \mathcal{R}_K(G(\bar{x})) \cap H(\bar{x})^\perp.$$

Conversely, let λ, μ, ν satisfy (1.5.1a), (1.5.1b) and (1.5.2). By (1.5.2) and the definition (1.2.1) of the radial cone, we can split the multipliers μ and ν and obtain

$$\begin{aligned} \mu &= \hat{\mu} + \hat{\xi}_1 H(\bar{x}), & \hat{\mu} &\in K^\circ \cap G(\bar{x})^\perp = \mathcal{T}_K(G(\bar{x}))^\circ, \\ \nu &= \hat{\nu} + \hat{\xi}_2 G(\bar{x}), & \hat{\nu} &\in K \cap H(\bar{x})^\perp = \mathcal{T}_{K^\circ}(H(\bar{x}))^\circ. \end{aligned}$$

Now, we set $\tilde{\xi} := \min(\hat{\xi}_1, \hat{\xi}_2)$, $\tilde{\mu} := \mu - \tilde{\xi} H(\bar{x})$, $\tilde{\nu} := \nu - \tilde{\xi} G(\bar{x})$. By

$$\tilde{\mu} = \mu - \tilde{\xi} H(\bar{x}) = \hat{\mu} + (\hat{\xi}_1 - \tilde{\xi}) H(\bar{x}) = \hat{\mu} + \max(0, \hat{\xi}_1 - \hat{\xi}_2) H(\bar{x}) \in K^\circ \cap G(\bar{x})^\perp,$$

and, similarly,

$$\tilde{\nu} = \nu - \tilde{\xi} G(\bar{x}) = \hat{\nu} + (\hat{\xi}_2 - \tilde{\xi}) G(\bar{x}) = \hat{\nu} + \max(0, \hat{\xi}_2 - \hat{\xi}_1) G(\bar{x}) \in K \cap H(\bar{x})^\perp,$$

we find that (1.5.3b) is satisfied. An easy calculation yields (1.5.3a), hence $\lambda, \tilde{\mu}, \tilde{\nu}, \tilde{\xi}$ are KKT multipliers for $\bar{x}$.

We compare the conditions (1.5.1) for strong stationarity with the conditions (1.5.2) for a KKT point. It is immediate that being a KKT point is, in general, a stronger condition than being strongly stationary. Moreover, we refer to p. 54 in Bergounioux, Mignot, 2000 for an example where strong stationarity is satisfied, but the KKT conditions are violated. However, in the case of a standard MPCC, the sets $\mathcal{R}_{K^\circ}(H(\bar{x}))$ and $\mathcal{R}_K(G(\bar{x}))$ are always closed and coincide with the tangent cones. Hence, we recover the result that strong stationarity is equivalent to being a KKT point in this case.

Analogously to the standard, finite-dimensional case, one could introduce weak stationarity via the KKT conditions of (TNLP). However, it is not clear how to define other notions such as A-, C-, and M-stationarity.

1.5.2. Polyhedric cones

In this section, we will consider the case that the cone K is polyhedric. This property enables us to show that strong stationarity implies first-order stationarity; see Theorem 1.5.4. Hence, strong stationarity seems to be a reasonable optimality condition in this case.

We recall that the cone K is called polyhedric w.r.t. $(\bar{G}, \bar{H})$, where $\bar{G} \in K$, $\bar{H} \in K^\circ$, $\langle \bar{G}, \bar{H} \rangle = 0$, iff

$$\operatorname{cl}(\mathcal{R}_K(\bar{G}) \cap \bar{H}^\perp) = \mathcal{T}_K(\bar{G}) \cap \bar{H}^\perp.$$

Note that the right-hand side is just the critical cone $\mathcal{K}_K(\bar{G}, \bar{H})$; see (1.2.3). Similarly, we define the polyhedricity of K° w.r.t. $(\bar{H}, \bar{G})$. This condition was first used by Mignot, 1976; Haraux, 1977 in order to show that the projection onto a polyhedric set is directionally differentiable.

The following lemma gives an important characterization of polyhedricity.

Lemma 1.5.3. Let $\bar{G} \in K$, $\bar{H} \in K^\circ$ with $\langle \bar{G}, \bar{H} \rangle = 0$ be given. The following conditions are equivalent.

(a) The cone K is polyhedric w.r.t. $(\bar{G}, \bar{H})$.

(b) The cone K° is polyhedric w.r.t. $(\bar{H}, \bar{G})$.

(c) $\mathcal{K}_K(\bar{G}, \bar{H})^\circ = \mathcal{K}_{K^\circ}(\bar{H}, \bar{G})$.

(d) $\mathcal{K}_{K^\circ}(\bar{H}, \bar{G})^\circ = \mathcal{K}_K(\bar{G}, \bar{H})$.

Proof. A straightforward calculation shows

$$\begin{aligned}(\mathcal{T}_K(\bar{G}) \cap \bar{H}^\perp)^\circ &= \mathrm{cl}(\mathcal{T}_K(\bar{G})^\circ + \mathrm{lin}(\bar{H})) = \mathrm{cl}(K^\circ \cap \bar{G}^\perp + \mathrm{lin}(\bar{H})) \\ &= \mathrm{cl}\big[(K^\circ + \mathrm{lin}(\bar{H})) \cap \bar{G}^\perp\big] = \mathrm{cl}(\mathcal{R}_{K^\circ}(\bar{H}) \cap \bar{G}^\perp).\end{aligned}$$

This establishes the equivalence of (b) and (c). A similar calculation yields

$$(\mathcal{T}_{K^\circ}(\bar{H}) \cap \bar{G}^\perp)^\circ = \mathrm{cl}(\mathcal{R}_K(\bar{G}) \cap \bar{H}^\perp). \tag{1.5.4}$$

Hence, (a) and (d) are equivalent. Finally, the equivalence of (c) and (d) follows from the bipolar theorem; see, e.g., Proposition 2.40 in Bonnans, Shapiro, 2000.

Note that the last two statements of Lemma 1.5.3 just mean that the critical cones $\mathcal{K}_K(\bar{G}, \bar{H})$ and $\mathcal{K}_{K^\circ}(\bar{H}, \bar{G})$ are polar to each other.

In order to state next theorem, we define the feasible set F of (MPCC)

$$F := \{x \in X : g(x) \in C,\ G(x) \in K,\ H(x) \in K^\circ,\ \langle G(x), H(x) \rangle = 0\}$$

and its linearized cone

$$\mathcal{T}_{\mathrm{lin}}(\bar{x}) := \left\{ \begin{array}{l} d \in X : g'(\bar{x})\, d \in \mathcal{T}_C(g(\bar{x})), \\ \qquad G'(\bar{x})\, d \in \mathcal{T}_K(G(\bar{x})),\ H'(\bar{x})\, d \in \mathcal{T}_{K^\circ}(H(\bar{x})), \\ \qquad \langle G'(\bar{x})\, d,\ H(\bar{x}) \rangle + \langle G(\bar{x}),\ H'(\bar{x})\, d \rangle = 0 \end{array} \right\}.$$

Since we have $H(\bar{x}) \in K^\circ$, $G'(\bar{x})\, d \in \mathcal{T}_K(G(\bar{x})) = \mathrm{cl}(K + \mathrm{lin}(G(\bar{x})))$ and $\langle G(\bar{x}),\ H(\bar{x}) \rangle = 0$, we find $\langle G'(\bar{x})\, d,\ H(\bar{x}) \rangle \leq 0$ and $\langle G(\bar{x}),\ H'(\bar{x})\, d \rangle \leq 0$ is obtained similarly. The sum of these non-positive terms is required to be zero, hence, both addends have to be zero. This implies the characterization

$$\mathcal{T}_{\mathrm{lin}}(\bar{x}) = \left\{ \begin{array}{l} d \in X : g'(\bar{x})\, d \in \mathcal{T}_C(g(\bar{x})),\ \ G'(\bar{x})\, d \in \mathcal{K}_K(G(\bar{x}), H(\bar{x})), \\ \qquad\qquad\qquad\qquad\qquad\qquad H'(\bar{x})\, d \in \mathcal{K}_{K^\circ}(H(\bar{x}), G(\bar{x})) \end{array} \right\}.$$

Theorem 1.5.4. Let $\bar{x}$ be strongly stationary and assume that K is polyhedric w.r.t. $(G(\bar{x}), H(\bar{x}))$. Then, $\bar{x}$ is linearized B-stationary, that is

$$f'(\bar{x})\, d \geq 0, \quad \forall d \in \mathcal{T}_{\mathrm{lin}}(\bar{x}). \tag{1.5.5}$$

Proof. For brevity, we write $\bar{g} = g(\bar{x})$, $\bar{H} = H(\bar{x})$, and $\bar{G} = G(\bar{x})$.

Definition 1.5.1 of strong stationarity directly yields

$$-f'(\bar{x}) \in g'(\bar{x})^\star \mathcal{T}_C(\bar{g})^\circ + G'(\bar{x})^\star \mathcal{K}_{K^\circ}(\bar{H}, \bar{G}) + H'(\bar{x})^\star \mathcal{K}_K(\bar{G}, \bar{H}).$$

We readily obtain

$$\begin{aligned}
&\Big(g'(\bar{x})^\star \mathcal{T}_C(\bar{g})^\circ + G'(\bar{x})^\star \mathcal{K}_{K^\circ}(\bar{H}, \bar{G}) + H'(\bar{x})^\star \mathcal{K}_K(\bar{G}, \bar{H})\Big)^\circ \\
&\quad = \big(g'(\bar{x})^\star \mathcal{T}_C(\bar{g})^\circ\big)^\circ \cap \big(G'(\bar{x})^\star \mathcal{K}_{K^\circ}(\bar{H}, \bar{G})\big)^\circ \cap \big(H'(\bar{x})^\star \mathcal{K}_K(\bar{G}, \bar{H})\big)^\circ \\
&\quad = g'(\bar{x})^{-1} \mathcal{T}_C(\bar{g}) \quad \cap G'(\bar{x})^{-1} \mathcal{K}_{K^\circ}(\bar{H}, \bar{G})^\circ \cap H'(\bar{x})^{-1} \mathcal{K}_K(\bar{G}, \bar{H})^\circ \\
&\quad = g'(\bar{x})^{-1} \mathcal{T}_C(\bar{g}) \quad \cap G'(\bar{x})^{-1} \mathcal{K}_K(\bar{G}, \bar{H}) \quad \cap H'(\bar{x})^{-1} \mathcal{K}_{K^\circ}(\bar{H}, \bar{G}) \\
&\quad = \mathcal{T}_{\mathrm{lin}}(\bar{x}).
\end{aligned}$$

Here, we used $\big(S^\star M^\circ\big)^\circ = S^{-1} M$ for bounded, linear operators S between Banach spaces and closed, convex cones M; see also (1) in Kurcyusz, 1976, and Lemma 1.5.3. This yields the assertion.

In contrast to the case of a standard MPCC, see, e.g., p. 613 in Flegel, Kanzow, 2005a, the converse statement of Theorem 1.5.4 requires a constraint qualification.

Lemma 1.5.5. Suppose that the feasible point $\bar{x}$ satisfies (1.5.5). Assume further that the constraint qualification

$$(g'(\bar{x}), G'(\bar{x}), H'(\bar{x}))\, X + \mathcal{T}_C(\bar{g}) \times \mathcal{K}_K(\bar{G}, \bar{H}) \times \mathcal{K}_{K^\circ}(\bar{H}, \bar{G}) = Y \times Z \times Z^\star$$

is satisfied. Then, $\bar{x}$ is strongly stationary.

Proof. For brevity, we write $\bar{g} = g(\bar{x})$, $\bar{H} = H(\bar{x})$, and $\bar{G} = G(\bar{x})$.

By assertion, we have $-f'(\bar{x}) \in \mathcal{T}_{\mathrm{lin}}(\bar{x})^\circ$. Due to the assumed constraint qualification, we can apply Theorem 2.1 in Kurcyusz, 1976 with the setting (denoting there the cone by M instead of K in order to avoid duplicate use of K)

$$S := (g'(\bar{x}), G'(\bar{x}), H'(\bar{x})) \qquad M := \mathcal{T}_C(\bar{g}) \times \mathcal{K}_K(\bar{G}, \bar{H}) \times \mathcal{K}_{K^\circ}(\bar{H}, \bar{G}),$$

we obtain $(S^{-1} M)^\circ = S^\star M^\circ$, that is

$$\mathcal{T}_{\mathrm{lin}}(\bar{x})^\circ = g'(\bar{x})^\star \mathcal{T}_C(\bar{g})^\circ + G'(\bar{x})^\star \mathcal{K}_K(\bar{G}, \bar{H})^\circ + H'(\bar{x})^\star \mathcal{K}_{K^\circ}(\bar{H}, \bar{G})^\circ.$$

By (1.5.4), we find $\mathcal{K}_{K^\circ}(\bar{H}, \bar{G})^\circ = \mathrm{cl}(\mathcal{R}_K(\bar{G}) \cap \bar{H}^\perp) \subset \mathcal{K}_K(\bar{G}, \bar{H})$, and similarly, $\mathcal{K}_K(\bar{G}, \bar{H})^\circ \subset \mathcal{K}_{K^\circ}(\bar{H}, \bar{G})$. This yields the assertion.

For a standard MPCC, the analog of Theorem 2.1 in Kurcyusz, 1976 follows either by

Farkas' Lemma or by applying the bipolar Theorem to $(S^\star M^\circ)^\circ = S^{-1} M$, since $S^\star M^\circ$ is closed in this setting. We refer to p. 613 in Flegel, Kanzow, 2005a for the proof including the application of Farkas' Lemma. Due to $\mathcal{T}_F(x) \subset \mathcal{T}_{\text{lin}}(x)$, strong stationarity also implies

$$f'(\bar{x})\, d \geq 0, \quad \forall d \in \mathcal{T}_F(\bar{x})$$

in the polyhedric case. Here, $\mathcal{T}_F(\bar{x})$ is the tangent cone of the (possibly non-convex) set F; see (2.84) in Bonnans, Shapiro, 2000. In particular, there are no first-order descent directions if strong stationarity is satisfied.

Theorem 1.5.4 and Lemma 1.5.5 show that our definition of strong stationarity possesses a reasonable strength in the presence of polyhedricity. However, we will see in Section 1.6.2 that our condition is too weak, if K is not polyhedric by means of an example. Nevertheless, every cone is polyhedric w.r.t. $(0, 0)$ and, hence, an additional linearization argument will yield stronger optimality conditions. This is also demonstrated in Section 1.6.2. It seems to be an open question to define strong stationarity for the general problem (MPCC) in the absence of polyhedricity.

1.5.3. Constraint qualifications which imply strong stationarity

With the preparations of Section 1.4, we are able to provide a constraint qualification which implies strong stationarity.

Theorem 1.5.6. Let $\bar{x} \in X$ be a local solution of (MPCC). We further assume that the constraint qualification (1.4.4) is satisfied for (TNLP) at $\bar{x}$, and that (NLP_G) and (NLP_H) satisfy the constraint qualification of Robinson-Zowe-Kurcyusz at $\bar{x}$; see (1.4.7) on page 24. Then, $\bar{x}$ is strongly stationary.

Proof. By using the uniqueness results of Section 1.4, we can directly transfer the proof of Theorem 1.3.1.

We must admit that the verification of the above constraint qualifications can be very complicated. Therefore, we state two stronger conditions which are easier to verify.

Proposition 1.5.7. We assume that (TNLP) satisfies the constraint qualification of Robinson-Zowe-Kurcyusz at the feasible point $\bar{x}$. Then, this constraint qualification is also satisfied for (NLP_G) and (NLP_H).

Proof. This follows easily from the observation that the feasible sets of (NLP_G) and (NLP_H) are larger than the feasible set of (TNLP). If the feasible set grows, the constraint qualification remains satisfied; see (1.4.7).

Proposition 1.5.8. Let us assume that $\bar{x}$ is a feasible point of (TNLP). Moreover, we assume

$$\mathcal{G}'(\bar{x})\, X = Y \times Z \times Z^\star, \tag{1.5.6}$$

where $\mathcal{G}(x) = (g(x), G(x), H(x))$. Then, (TNLP) satisfies (1.4.2) and (1.4.7) at $\bar{x}$. In case $\bar{x}$ is a local minimizer of (MPCC), $\bar{x}$ is strongly stationary.

Proof. We define

$$\mathcal{C} := C \times \left[\mathcal{T}_{K^\circ}(H(\bar{x}))^\circ \cap \mathcal{T}_K(G(\bar{x}))^{\circ\perp}\right] \times \left[\mathcal{T}_K(G(\bar{x}))^\circ \cap \mathcal{T}_{K^\circ}(H(\bar{x}))^{\circ\perp}\right].$$

Then, the constraints in (TNLP) simply read $\mathcal{G}(x) \in \mathcal{C}$. Since $\mathcal{G}'(\bar{x})$ is assumed to be surjective, we have $\ker \mathcal{G}'(\bar{x})^\star = \{0\}$, hence, (1.4.2) is satisfied. The constraint qualification of Robinson-Zowe-Kurcyusz similarly follows from this surjectivity. Now, Theorem 1.5.6 and Proposition 1.5.7 imply the assertion.

For an important class of examples, the left-hand side in (1.5.6) is merely dense in the right-hand side. However, the surjectivity can be obtained after using a density argument; see Section 1.6.1. We emphasize that we did not assume polyhedricity of K in the above results.

1.5.4. Generalization to non-conic MPCCs

In this section, we want to treat the case that the set K is not assumed to be a cone. In absence of this assumption, the complementarity between $G(x)$ and $H(x)$ can no longer be stated via

$$G(x) \in K, \quad H(x) \in K^\circ, \quad \langle H(x), G(x)\rangle = 0.$$

However, if K is a cone, then this is equivalent to

$$G(x) \in K, \quad H(x) \in \mathcal{T}_K(G(x))^\circ, \tag{1.5.7}$$

compare (1.2.2). Note that optimality conditions for constrained optimization problems such as (1.4.1) contain complementarity conditions like (1.5.7). Moreover, we briefly recall that variational inequalities can be written as (1.5.7) in Section 1.6.1. The complementarity relation (1.5.7) is the proper starting point for the generalization in this section. That is, we consider the optimization problem

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{s.t.} \quad & g(x) \in C, \quad G(x) \in K, \quad H(x) \in \mathcal{T}_K(G(x))^\circ. \end{aligned} \tag{1.5.8}$$

We make the same assumptions as in the beginning of Section 1.5, but $K \subset Z$ is just assumed to be a closed, convex set (and not necessarily a cone).

In order to apply the obtained results, we are going to transform the above problem into a problem with a conic complementarity constraint. To this end, we introduce the closed, convex cone

$$\tilde{K} := \text{cl-cone}(\{1\} \times K) \subset \mathbb{R} \times Z, \tag{1.5.9}$$

where cl-cone(A) is the closed, convex, conic hull of a set A. The following lemmas show that there is a close relation between the set K and the cone $\tilde{K}$.

Lemma 1.5.9. For $\bar{G} \in Z$, the condition $\bar{G} \in K$ is equivalent to $(1, \bar{G}) \in \tilde{K}$.

Proof. By definition, $\bar{G} \in K$ gives $(1, \bar{G}) \in \tilde{K}$. We observe

$$\tilde{K}^\circ = (\{1\} \times K)^\circ = \{(s, z^\star) \in \mathbb{R} \times Z^\star : s + \langle z^\star, z\rangle \leq 0,\ \forall z \in K\}. \tag{1.5.10}$$

Now, let us assume $\bar{G} \notin K$. A separation theorem, see, e.g., Theorem 2.14 in Bonnans, Shapiro, 2000, yields the existence of $(s, z^\star) \in \mathbb{R} \times Z^\star$, such that $\langle z^\star, z\rangle \leq -s < \langle z^\star, \bar{G}\rangle$ holds for all $z \in K$. The first inequality yields $(s, z^\star) \in \tilde{K}^\circ$, and then the second one yields $(1, \bar{G}) \notin \tilde{K}^{\circ\circ} = \tilde{K}$.

Lemma 1.5.10. For $\bar{G} \in Z$, $\bar{H} \in Z^\star$, the conditions $\bar{G} \in K$, $\bar{H} \in \mathcal{T}_K(\bar{G})^\circ$ are equivalent to the existence of $s \in \mathbb{R}$ such that

$$(1, \bar{G}) \in \tilde{K}, \quad (s, \bar{H}) \in \tilde{K}^\circ, \quad \big\langle (s, \bar{H}), (1, \bar{G})\big\rangle = 0.$$

Proof. Lemma 1.5.9 shows the equivalency of $\bar{G} \in K$ and $(1, \bar{G}) \in \tilde{K}$. By (1.5.10) we find

$$\begin{aligned}
\mathcal{T}_{\tilde{K}}(1, \bar{G})^\circ &= \tilde{K}^\circ \cap (1, \bar{G})^\perp \\
&= \{(s, z^\star) \in \mathbb{R} \times Z^\star : s = -\langle z^\star, \bar{G}\rangle \text{ and } \langle z^\star, z - \bar{G}\rangle \leq 0, \quad \forall z \in K\} \\
&= \{(s, z^\star) \in \mathbb{R} \times Z^\star : s = -\langle z^\star, \bar{G}\rangle \text{ and } z^\star \in \mathcal{T}_K(\bar{G})^\circ\}.
\end{aligned} \tag{1.5.11}$$

This shows that $\bar{H} \in \mathcal{T}_K(\bar{G})^\circ$ is equivalent to the existence of $s \in \mathbb{R}$ such that $(s, \bar{H}) \in \tilde{K}^\circ$ and $\big\langle (s, \bar{H}), (1, \bar{G})\big\rangle = 0$.

By using (1.5.11) and the bipolar theorem, we also obtain a characterization of the tangent cone of $\tilde{K}$, namely

$$\begin{aligned}
\mathcal{T}_{\tilde{K}}(1, \bar{G}) &= \mathcal{T}_{\tilde{K}}(1, \bar{G})^{\circ\circ} \\
&= \{(t, z) \in \mathbb{R} \times Z : \big\langle (s, z^\star), (t, z)\big\rangle \leq 0, \quad \forall (s, z^\star) \in \mathcal{T}_{\tilde{K}}(1, \bar{G})^\circ\} \\
&= \{(t, z) \in \mathbb{R} \times Z : \langle z^\star, z - t\,\bar{G}\rangle \leq 0, \quad \forall z^\star \in \mathcal{T}_K(\bar{G})^\circ\} \\
&= \{(t, z) \in \mathbb{R} \times Z : z - t\,\bar{G} \in \mathcal{T}_K(\bar{G})\}.
\end{aligned} \tag{1.5.12}$$

Due to Lemma 1.5.10, we can state (1.5.7) equivalently as a complementarity relation involving the cone $\tilde{K}$. Thus, problem (1.5.8) is equivalent to

$$\begin{aligned} \text{Minimize} \quad & f(x), \qquad \text{w.r.t.} \quad (s,x) \in \mathbb{R} \times X, \\ \text{s.t.} \quad & g(x) \in C, \\ & (1, G(x)) \in \tilde{K},\ (s, H(x)) \in \tilde{K}^\circ,\ \langle (s, H(x)),\, (1, G(x)) \rangle = 0. \end{aligned} \tag{1.5.13}$$

Since this is an instance of (MPCC), we can apply the obtained results for conic MPCCs. In order to translate them into results for (1.5.8), we provide the following results.

- The polyhedricity of K implies the polyhedricity of $\tilde{K}$; see Lemma 1.5.11. This enables us to apply Theorem 1.5.4, and it shows the strength of the optimality condition.
- We translate the strong stationarity conditions for (1.5.13) into optimality conditions for (1.5.8); see Lemma 1.5.13.
- We provide a result analogous to Proposition 1.5.8; see Lemma 1.5.14.

Lemma 1.5.11. Assume that the closed, convex set K is polyhedric w.r.t. $(\bar{G}, \bar{H})$ with $\bar{G} \in K$ and $\bar{H} \in \mathcal{T}_K(\bar{G})^\circ$, that is,

$$\operatorname{cl}(\mathcal{R}_K(\bar{G}) \cap \bar{H}^\perp) = \mathcal{T}_K(\bar{G}) \cap \bar{H}^\perp.$$

Then, the cone $\tilde{K}$ is polyhedric w.r.t. $((1, \bar{G}), (-\langle \bar{H},\, \bar{G} \rangle, \bar{H}))$.

Proof. We set $s := -\langle \bar{H},\, \bar{G} \rangle$. We have to show

$$\operatorname{cl}(\mathcal{R}_{\tilde{K}}(1, \bar{G}) \cap (s, \bar{H})^\perp) \supset \mathcal{T}_{\tilde{K}}(1, \bar{G}) \cap (s, \bar{H})^\perp.$$

Let $(t, z) \in \mathcal{T}_{\tilde{K}}(1, \bar{G}) \cap (s, \bar{H})^\perp$ be given. By (1.5.12) we get $z - t\,\bar{G} \in \mathcal{T}_K(\bar{G})$. Since

$$\langle \bar{H},\, z - t\,\bar{G} \rangle = \langle \bar{H},\, z \rangle - t\,\langle \bar{H},\, \bar{G} \rangle = \langle \bar{H},\, z \rangle + t\,s = \langle (s, \bar{H}),\, (t, z) \rangle = 0,$$

we have $z - t\,\bar{G} \in \mathcal{T}_K(\bar{G}) \cap \bar{H}^\perp$. Owing to the polyhedricity of K, we obtain a sequence $\{z_n\} \subset \mathcal{R}_K(\bar{G}) \cap \bar{H}^\perp$ such that $z_n \to z - t\,\bar{G}$ in Z.

Immediately, we obtain $(t, z_n + t\,\bar{G}) \to (t, z)$ in $\mathbb{R} \times Z$ and

$$\langle (s, \bar{H}),\, (t, z_n + t\,\bar{G}) \rangle = -t\,\langle \bar{H},\, \bar{G} \rangle + \langle \bar{H},\, z_n + t\,\bar{G} \rangle = \langle \bar{H},\, z_n \rangle = 0.$$

It remains to show $(t, z_n + t\,\bar{G}) \in \mathcal{R}_{\tilde{K}}(1, \bar{G})$. We have

$$(1, \bar{G}) + \varepsilon\,(t, z_n + t\,\bar{G}) = (1 + \varepsilon\,t)\left(1, \bar{G} + \varepsilon\,(1 + \varepsilon\,t)^{-1}\,z_n\right).$$

Since $z_n \in \mathcal{R}_K(\bar{G})$, $\bar{G} + \varepsilon\,(1 + \varepsilon\,t)^{-1}\,z_n \in K$ for $\varepsilon > 0$ small. By definition (1.5.9) of $\tilde{K}$, this yields $(1, \bar{G}) + \varepsilon\,(t, z_n + t\,\bar{G}) \in \tilde{K}$ for small ε and, hence, $(t, z_n + t\,\bar{G}) \in \mathcal{R}_{\tilde{K}}(1, \bar{G})$.

In order to obtain optimality conditions for (1.5.8) via the strong stationarity conditions of (1.5.13), we state the following lemma.

Lemma 1.5.12. Let $(s, \bar{x})$ be a feasible point of (1.5.13). Then,

$$(0, \nu) \in \mathcal{T}_{\tilde{K}}(1, \bar{G}) \cap (s, \bar{H})^{\perp} \quad \Longleftrightarrow \quad \nu \in \mathcal{T}_K(\bar{G}) \cap \bar{H}^{\perp}, \tag{1.5.14a}$$

$$(t, \mu) \in \mathcal{T}_{\tilde{K}^\circ}(s, \bar{H}) \cap (1, \bar{G})^{\perp} \quad \Longleftrightarrow \quad \mu \in \left(\mathcal{R}_K(\bar{G}) \cap \bar{H}^{\perp}\right)^{\circ}, \quad t = -\langle \mu, \bar{G} \rangle, \tag{1.5.14b}$$

where we set $\bar{G} := G(\bar{x})$, $\bar{H} := H(\bar{x})$.

Proof. The equivalence (1.5.14a) follows immediately from (1.5.12).

To show the first implication of (1.5.14b), let $(t, \mu) \in \mathcal{T}_{\tilde{K}^\circ}(s, \bar{H}) \cap (1, \bar{G})^{\perp}$ be given. The relation $t = -\langle \mu, \bar{G} \rangle$ follows directly from $(t, \mu) \in (1, \bar{G})^{\perp}$. Now, let $v \in \mathcal{R}_K(\bar{G}) \cap \bar{H}^{\perp}$ be given. By definition, there exists $\lambda > 0$ and $k \in K$ with $v = \lambda\,(k - \bar{G})$. Then, $(1, k) \in \tilde{K}$ and

$$\langle (s, \bar{H}),\, (1, k) \rangle = s + \langle \bar{H},\, k \rangle = \langle \bar{H},\, k - \bar{G} \rangle = \lambda^{-1} \langle \bar{H},\, v \rangle = 0.$$

Hence, $(1, k) \in \tilde{K} \cap (s, \bar{H})^{\perp} = \mathcal{T}_{\tilde{K}^\circ}(s, \bar{H})^{\circ}$. Therefore,

$$\lambda^{-1} \langle \mu,\, v \rangle = \langle \mu,\, k - \bar{G} \rangle = \langle \mu,\, k \rangle + t = \langle (t, \mu),\, (1, k) \rangle \leq 0,$$

since $(1, k) \in \mathcal{T}_{\tilde{K}^\circ}(s, \bar{H})^{\circ}$ and $(t, \mu) \in \mathcal{T}_{\tilde{K}^\circ}(s, \bar{H})$. This shows the right-hand side of (1.5.14b).

In order to prove the converse, let $\mu \in \left(\mathcal{R}_K(\bar{G}) \cap \bar{H}^{\perp}\right)^{\circ}$ be given and set $t = -\langle \mu, \bar{G} \rangle$. Then, $(t, \mu) \in (1, \bar{G})^{\perp}$ is immediate. By the bipolar theorem, it remains to show

$$\langle (t, \mu),\, (p, k) \rangle \leq 0, \quad \forall (p, k) \in \tilde{K} \cap (s, \bar{H})^{\perp} = \mathcal{T}_{\tilde{K}^\circ}(s, \bar{H})^{\circ}. \tag{1.5.15}$$

To this end, let $(p, k) \in \tilde{K} \cap (s, \bar{H})^{\perp}$ be given. By definition of $\tilde{K}$, we find $p \geq 0$. Since $\tilde{K}$ is a convex cone and $(1, \bar{G}) \in \tilde{K}$ we have

$$\left(1, (p+1)^{-1}\,(k + \bar{G})\right) = 2\,(p+1)^{-1}\,2^{-1}\,(p + 1, k + \bar{G}) \in \tilde{K}.$$

Hence, Lemma 1.5.9 yields $(k + \bar{G})/(p+1) \in K$. Now, we find

$$k - p\,\bar{G} = (p+1)\left((p+1)^{-1}\,(k + \bar{G}) - \bar{G}\right) \in \mathcal{R}_K(\bar{G})$$

and $\langle \bar{H},\, k - p\,\bar{G} \rangle = \langle \bar{H},\, k \rangle + p\,s = 0$. Hence, $k - p\,\bar{G} \in \mathcal{R}_K(\bar{G}) \cap \bar{H}^{\perp}$. This yields

$$\langle (t, \mu),\, (p, k) \rangle = \langle \mu,\, k \rangle + p\,t = \langle \mu,\, k - p\,\bar{G} \rangle \leq 0.$$

We have shown (1.5.15) and the left-hand side of (1.5.14b) follows by the bipolar theorem.

Now, we are able to obtain optimality conditions for our original problem (1.5.8) via the strong stationarity conditions of the auxiliary problem (1.5.13).

Lemma 1.5.13. Let $(s, \bar{x})$ be a strongly stationary point of (1.5.13). Then, there exist $\lambda \in Y^\star$, $\mu \in Z^\star$, $\nu \in Z$ such that

$$f'(\bar{x}) + g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu = 0, \tag{1.5.16a}$$

$$\lambda \in \mathcal{T}_C(g(\bar{x}))^\circ,\ \mu \in \big(\mathcal{R}_K(G(\bar{x})) \cap H(\bar{x})^\perp\big)^\circ,\ \nu \in \mathcal{T}_K(G(\bar{x})) \cap H(\bar{x})^\perp. \tag{1.5.16b}$$

Proof. Since $(s, \bar{x})$ is a strongly stationary point of (1.5.13), there exist multipliers $\lambda \in Y^\star$, $(t, \mu) \in \mathbb{R} \times Z^\star$, $(r, \nu) \in \mathbb{R} \times Z$, such that $r = 0$ and

$$f'(\bar{x}) + g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu = 0, \qquad \lambda \in \mathcal{T}_C(g(\bar{x}))^\circ,$$
$$(t, \mu) \in \mathcal{T}_{\tilde{K}^\circ}(s, H(\bar{x})) \cap (1, G(\bar{x}))^\perp, \qquad (r, \nu) \in \mathcal{T}_{\tilde{K}}(1, G(\bar{x})) \cap (s, H(\bar{x}))^\perp,$$

cf. Definition 1.5.1. Now, the assertion follows directly from Lemma 1.5.12.

In the case of K being a cone, we obtain exactly the conditions of Definition 1.5.1, cf. (1.5.4). That is, we do not lose any information by the transformation to the problem (1.5.13) in this case.

Finally, we transfer Proposition 1.5.8 to the non-conic case.

Lemma 1.5.14. Let $\bar{x}$ be a local minimizer of (1.5.8) and assume

$$\mathcal{G}'(\bar{x})\, X = Y \times Z \times Z^\star,$$

where $\mathcal{G}(x) = (g(x), G(x), H(x))$. Then, there exist multipliers $\lambda \in Y^\star$, $\mu \in Z^\star$, $\nu \in Z$, such that (1.5.16) is satisfied.

Proof. By Lemma 1.5.10, $(\bar{s}, \bar{x})$ with $\bar{s} = -\langle H(\bar{x}), G(\bar{x})\rangle$ is a local minimizer of (1.5.13). In what follows, we use Theorem 1.5.6 to infer the strong stationarity of the minimizer $(\bar{s}, \bar{x})$ of (1.5.13). We set $\tilde{\mathcal{G}}(s, x) := (g(x), 1, G(x), s, H(x))$ and

$$\mathcal{C} := C \times \Big[\mathcal{T}_{\tilde{K}^\circ}(\bar{s}, H(\bar{x}))^\circ \cap \mathcal{T}_{\tilde{K}}(1, G(\bar{x}))^{\circ\perp}\Big] \times \Big[\mathcal{T}_{\tilde{K}}(1, G(\bar{x}))^\circ \cap \mathcal{T}_{\tilde{K}^\circ}(\bar{s}, H(\bar{x}))^{\circ\perp}\Big].$$

That is, the constraints in (TNLP) associated with (1.5.13) are simply $\tilde{\mathcal{G}}(s, x) \in \mathcal{C}$.

By assumption, we obtain

$$\tilde{\mathcal{G}}'(\bar{s}, \bar{x})\, (\mathbb{R} \times X) = Y \times \{0\} \times Z \times \mathbb{R} \times Z^\star. \tag{1.5.17}$$

Let us abbreviate $V = \Big[\mathcal{T}_{\tilde{K}^\circ}(\bar{s}, H(\bar{x}))^\circ \cap \mathcal{T}_{\tilde{K}}(1, G(\bar{x}))^{\circ\perp}\Big]$. By definition of $\mathcal{C}$, we find $\mathcal{C} \supset \{g(\bar{x})\} \times V \times \{0\}$. Hence, $\mathcal{R}_{\mathcal{C}}(\tilde{\mathcal{G}}(\bar{s}, \bar{x})) \supset \{0\} \times \mathcal{R}_V(1, G(\bar{x})) \times \{0\}$. Since V is a cone, we obtain $\mathcal{R}_{\mathcal{C}}(\tilde{\mathcal{G}}(\bar{s}, \bar{x})) \supset \{0\} \times \operatorname{lin}\big((1, G(\bar{x}))\big) \times \{(0, 0)\}$ by using (1.2.1). This shows

$\mathcal{T}_{\mathcal{C}}(\tilde{\mathcal{G}}(\bar{s},\bar{x}))^{\circ\perp} = \mathcal{R}_{\mathcal{C}}(\tilde{\mathcal{G}}(\bar{s},\bar{x}))^{\circ\perp} \supset \{0\} \times \operatorname{lin}((1, G(\bar{x}))) \times \{(0,0)\}$. Due to this surjectivity in the second component on the right-hand side, (1.5.17) yields

$$\begin{aligned} \tilde{\mathcal{G}}'(\bar{s},\bar{x})\,(\mathbb{R}\times X) - \mathcal{R}_{\mathcal{C}}(\tilde{\mathcal{G}}(\bar{s},\bar{x})) &= Y\times\mathbb{R}\times Z\times\mathbb{R}\times Z^\star,\\ \tilde{\mathcal{G}}'(\bar{s},\bar{x})\,(\mathbb{R}\times X) - \mathcal{T}_{\mathcal{C}}(\tilde{\mathcal{G}}(\bar{s},\bar{x}))^{\circ\perp} &= Y\times\mathbb{R}\times Z\times\mathbb{R}\times Z^\star. \end{aligned}$$

By Proposition 1.5.7, this shows that the assumptions of Theorem 1.5.6 are satisfied. Hence, $(\bar{s},\bar{x})$ is a strongly stationary point of (1.5.13). Lemma 1.5.13 yields the claim.

We cannot directly use Proposition 1.5.8 to infer the above result; see (1.5.17).

1.6. Examples

In this section, we are going to apply the above theory to two problems. The first one is an optimal control problem governed by a variational inequality. The second problem is an MPCC involving the non-polyhedric cone of symmetric, semidefinite matrices.

1.6.1. Optimal control of variational inequalities

This first example shows that our technique is able to reproduce the result by Mignot, 1976 concerning strong stationarity for infinite-dimensional problems. Moreover, we obtain optimality conditions in a more general situation. We assume that

- Y, U, Z are Hilbert spaces,
- the operator $A \in \mathcal{L}(Y, Y^\star)$ is coercive; $B \in \mathcal{L}(U, Y^\star)$; and the operator $C \in \mathcal{L}(Z, Y^\star)$ has a dense range,
- the objective $f : Y \times U \times Z \to \mathbb{R}$ is Fréchet differentiable,
- the set $U_{\mathrm{ad}} \subset U$ is closed and convex, and
- the closed, convex set $K \subset Y$ is polyhedric w.r.t. all (y,ξ) with $y \in K$, $\xi \in \mathcal{T}_K(y)^\circ$.

We consider the optimization problem

$$\begin{aligned} \text{Minimize}\quad & f(y,u,z), \qquad \text{w.r.t.}\quad (y,u,z) \in Y\times U\times Z,\\ \text{s.t.}\quad & u\in U_{\mathrm{ad}},\quad y\in K,\quad B\,u + C\,z - A\,y \in \mathcal{T}_K(y)^\circ. \end{aligned} \tag{1.6.1}$$

The last two constraints represent a complementarity (1.5.7) and they are equivalent to $y = S(B\,u + C\,z)$, where S is the solution operator $Y^\star \ni \omega \mapsto y \in Y$ of the variational inequality

$$\text{find } y\in K \quad \text{s.t.}\quad \langle A\,y - \omega,\, v - y\rangle \geq 0, \quad \forall v \in K.$$

It is well known that this variational inequality is uniquely solvable and that the solution operator $S : Y^\star \to Y$ is Lipschitz continuous. By standard techniques (and further assumptions), one infer the existence of solutions of (1.6.1).

Let us assume that $(\bar{y}, \bar{u}, \bar{z})$ is a local minimizer of (1.6.1). By defining the operator $\mathcal{G}(y, u, z) := (u, y, B\,u + C\,z - A\,y)$, we find that $\mathcal{G}'(\bar{y}, \bar{u}) = \mathcal{G}$ has merely a dense range and is, in general, not surjective. Hence, we cannot apply Lemma 1.5.14. Similarly, one can check that the assumptions of Theorem 1.5.6 are not satisfied for the corresponding problem (1.5.13). To circumvent this, we use a clever linearization argument due to Mignot, 1976.

Indeed, due to the polyhedricity of K, we know from Theorem 2.1 in Mignot, 1976 that S is directionally differentiable and the directional derivative $\delta y = S'(\omega; \delta\omega)$ is given as the solution of the variational inequality

$$\text{find } \delta y \in \mathcal{K} \quad \text{s.t.} \quad \langle A\,\delta y - \delta\omega,\, v - \delta y\rangle \geq 0, \quad \forall v \in \mathcal{K}, \tag{1.6.2}$$

with the critical cone $\mathcal{K} = \mathcal{K}_K(y, \omega - A\,y) = \mathcal{T}_K(y) \cap (\omega - A\,y)^\perp$. Again, the solution operator associated with this variational inequality is Lipschitz continuous from $Y^\star$ to Y. Now, the optimality of $(\bar{y}, \bar{u}, \bar{z})$ implies that

$$f_y(\cdot)\, S'(\bar{\omega}; B\,\delta u + C\,\delta z) + f_u(\cdot)\,\delta u + f_z(\cdot)\,\delta z \geq 0, \quad \forall \delta u \in \mathcal{T}_{U_{\text{ad}}}(\bar{u}), \delta z \in Z, \tag{1.6.3}$$

where f_y, f_u, f_z are the partial derivatives of f, and we abbreviated the argument $(\cdot) = (\bar{y}, \bar{u}, \bar{z})$, and $\bar{\omega} = B\,\bar{u} + C\,\bar{z}$. Following Mignot, 1976 again, we test this variational inequality with $\delta u = 0$ and $\pm\delta z$. We find

$$|f_z(\cdot)\,\delta z| \leq |f_y(\cdot)\, S'(\bar{\omega}; C\,\delta z)| \leq c\,\|C\,\delta z\|_{Y^\star}.$$

Hence, $C\,\delta z \mapsto f_z(\cdot)\,\delta z$ defines a bounded functional on the range of C, which can be extended (by continuity) to a functional $p \in Y^{\star\star} = Y$. In particular, we have $f_z(\cdot) = C^\star p$. Using again the density of the range of C in $Y^\star$ we find that (1.6.3) implies

$$f_y(\cdot)\, S'(\bar{\omega}; B\,\delta u + \delta\zeta) + f_u(\cdot)\,\delta u + \langle p,\, \delta\zeta\rangle \geq 0, \quad \forall \delta u \in \mathcal{T}_{U_{\text{ad}}}(\bar{u}), \delta\zeta \in Y^\star.$$

Together with (1.6.2) it follows that $(\delta y, \delta u, \delta\zeta) = 0$ is a global minimizer of

$$\begin{aligned} \text{Minimize} \quad & f_y(\cdot)\,\delta y + f_u(\cdot)\,\delta u + \langle p,\, \delta\zeta\rangle, \quad \text{w.r.t.} \quad (\delta y, \delta u, \delta\zeta) \in Y \times U \times Y^\star, \\ \text{s.t.} \quad & \delta u \in \mathcal{T}_{U_{\text{ad}}}(\bar{u}), \quad \delta y \in \mathcal{K}, \quad B\,\delta u + \delta\zeta - A\,\delta y \in \mathcal{K}^\circ, \\ \text{and} \quad & \langle B\,\delta u + \delta\zeta - A\,\delta y,\, \delta y\rangle = 0. \end{aligned}$$

Now, we set $\mathcal{G}(\delta y, \delta u, \delta\zeta) := (\delta u, \delta y, B\,\delta u + \delta\zeta - A\,\delta y)$. It is immediate that $\mathcal{G}'(0, 0, 0) = \mathcal{G}$ is surjective. Hence, we can apply Proposition 1.5.8. We evaluate the strong stationarity conditions (1.5.1) for the minimizer $(0, 0, 0)$ and obtain the system

$$\begin{aligned} & f_y(\cdot) + \mu - A^\star\nu = 0, && \lambda \in \mathcal{T}_{U_{\text{ad}}}(\bar{u})^\circ, && \nu \in \mathcal{K}, \\ & f_u(\cdot) + \lambda + B^\star\nu = 0, && p + \nu = 0, && \mu \in \mathcal{K}^\circ. \end{aligned}$$

By eliminating ν and adding the condition $f_z(\cdot) = C^\star p$ for p, we finally obtain the optimality system

$$f_y(\cdot) + \mu + A^\star p = 0, \qquad \lambda \in \mathcal{T}_{U_{\text{ad}}}(\bar{u})^\circ, \tag{1.6.4a}$$

$$f_u(\cdot) + \lambda - B^\star p = 0, \qquad -p \in \mathcal{T}_K(\bar{y}) \cap (B\,\bar{u} + C\,\bar{z} - A\,\bar{y})^\perp, \tag{1.6.4b}$$

$$f_z(\cdot) - C^\star p = 0, \qquad \mu \in (\mathcal{T}_K(\bar{y}) \cap (B\,\bar{u} + C\,\bar{z} - A\,\bar{y})^\perp)^\circ. \tag{1.6.4c}$$

Note that this optimality system is a new result for the problem (1.6.1). We comment on two special cases of the above result.

Dense controls. In the first case, we set $U := U_{\text{ad}} := \{0\}$. The optimization problem (1.6.1) becomes

$$\text{Minimize} \quad f(y,0,z), \quad \text{s.t.} \quad y \in K, \quad C\,z - A\,y \in \mathcal{T}_K(y)^\circ,$$

and we obtained the optimality system

$$\begin{aligned} f_y(\cdot) + \mu + A^\star p = 0, \quad -p \in \mathcal{T}_K(\bar{y}) \cap (C\,\bar{z} - A\,\bar{y})^\perp, \\ f_z(\cdot) - C^\star p = 0, \quad \mu \in (\mathcal{T}_K(\bar{y}) \cap (C\,\bar{z} - A\,\bar{y})^\perp)^\circ. \end{aligned} \tag{1.6.5}$$

This result is well known. In particular, it is straightforward to generalize the arguments leading to Theorem 4.3 in Mignot, 1976 and one obtains the system (1.6.5). The same result was also reproduced in Theorem 4.6 in Hintermüller, Surowiec, 2011 by techniques from variational analysis.

Regularization of control constraints. With $Z = \{0\}$ the problem (1.6.1) reads

$$\text{Minimize} \quad f(y,u), \quad \text{s.t.} \quad u \in U_{\text{ad}}, \quad y \in K, \quad B\,u - A\,y \in \mathcal{T}_K(y)^\circ.$$

This is an optimal control problem of a variational inequality with control constraints. It is known that strong stationarity may not be a necessary optimality condition in this case; see, e.g., the counterexamples in Section 3.6. From this point of view, the problem (1.6.1) is a regularization of the control constrained problem. This regularization is similar to the virtual control regularization introduced in Krumbiegel, Rösch, 2009 for state constrained optimal control problems.

The solution of this regularized problem (1.6.1) satisfies a system of strong stationarity. One could introduce a regularization parameter γ, by setting, e.g.,

$$f(y,u,z) := f(y,u) + \frac{\gamma}{2}\,\|z\|_Z^2$$

and pass to the limit $\gamma \to \infty$ with the optimality system. This is, however, beyond the scope of this paper and subject to further research.

1.6.2. Application to semidefinite complementarity programs

We have seen in Section 1.5.2 that our definition of strong stationarity implies first-order stationarity if K is polyhedric. In this section, we discuss an example including a non-polyhedric cone K. In particular, we consider

$$\begin{aligned} \text{Minimize} \quad f(A,B), \quad \text{w.r.t.} \quad A, B \in \mathbb{S}^n, \\ \text{s.t.} \quad A \in \mathbb{S}^n_+, \quad B \in \mathbb{S}^n_-, \quad (A,\,B)_F = 0. \end{aligned} \tag{1.6.6}$$

Here, $\mathbb{S}^n$ is the set of symmetric $n \times n$ matrices, $\mathbb{S}^n_+$ ($\mathbb{S}^n_-$) are the cones of positive (negative) semidefinite matrices, and $(\cdot, \cdot)_F$ is the Frobenius inner product. The objective $f : \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{R}$ is assumed to be differentiable. In the sequel, we compare four different optimality systems for the problem (1.6.6):

(a) the (classical) KKT conditions,

(b) the strong stationarity conditions which are defined in Definition 5.1 in Ding, D. Sun, Ye, 2014, see also Definition 3.3 in Wu, L. Zhang, Y. Zhang, 2014, which are tailored to problems with semidefinite complementarity constraints,

(c) our strong stationarity conditions applied to (1.6.6),

(d) our strong stationarity conditions applied to a linearization of (1.6.6).

In order to keep the presentation simple, we discuss the case $n = 3$ and assume that the local minimizer $(\bar{A}, \bar{B})$ of (1.6.6) is

$$\bar{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \bar{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

The general case can be discussed analogously, but requires a more complicated notation. The necessary details can be found in the references which are cited in the derivations below.

The KKT conditions. In order to compare the KKT conditions with the other conditions, we will formulate them without a multiplier for the complementarity condition $(A,\, B)_F = 0$ by using Lemma 1.5.2. Since the multipliers are also matrices, we use (U, V) rather than (μ, ν). We obtain the optimality conditions

$$f_A(\bar{A}, \bar{B}) + U = 0, \quad f_B(\bar{A}, \bar{B}) + V = 0, \tag{1.6.7a}$$

$$U \in \mathcal{R}_{\mathbb{S}^n_-}(\bar{B}) \cap \bar{A}^\perp = \left\{ U \in \mathbb{S}^n : \; U = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \leq 0 & * \\ 0 & * & * \end{pmatrix}, \text{ and } U_{23} = 0 \text{ if } U_{22} = 0 \right\}, \tag{1.6.7b}$$

$$V \in \mathcal{R}_{\mathbb{S}^n_+}(\bar{A}) \cap \bar{B}^\perp = \left\{ V \in \mathbb{S}^n : \; V = \begin{pmatrix} * & * & 0 \\ * & \geq 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ and } V_{12} = 0 \text{ if } V_{22} = 0 \right\}. \tag{1.6.7c}$$

Here and in what follows, f_A, f_B are the partial derivatives of f, and we use

$$U = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \leq 0 & * \\ 0 & * & * \end{pmatrix}$$

as a short-hand for $U_{11} = U_{12} = U_{13} = 0$, and $U_{22} \leq 0$.

The tailored conditions from the literature. We give some simple arguments leading to the optimality conditions, which can be found in the above mentioned literature. These arguments also show that the obtained optimality conditions seem to be the "correct" ones.

Proceeding as usual, one finds the necessary first-order condition

$$-f'(\bar{A},\bar{B}) \in \mathcal{T}_F(\bar{A},\bar{B})^\circ, \tag{1.6.8}$$

where F is the feasible set of (1.6.6). This feasible set is just the graph of the (set-valued) normal cone mapping $A \mapsto \mathcal{T}_{\mathbb{S}^n_+}(A)^\circ$, and its normal cone $\mathcal{T}_F(\bar{A},\bar{B})^\circ$ is given by all $(U,V) \in (\mathbb{S}^n)^2$ satisfying

$$U = \begin{pmatrix} 0 & 0 & * \\ 0 & \leq 0 & * \\ * & * & * \end{pmatrix}, \quad V = \begin{pmatrix} * & * & * \\ * & \geq 0 & 0 \\ * & 0 & 0 \end{pmatrix}, \quad U_{13} + V_{13} = 0,$$

see Corollary 3.2 in Wu, L. Zhang, Y. Zhang, 2014. The obtained optimality conditions are (1.6.7a) and $(U,V) \in \mathcal{T}_F(\bar{A},\bar{B})^\circ$. They coincide with the optimality conditions Definition 5.1 in Ding, D. Sun, Ye, 2014, Definition 3.3 in Wu, L. Zhang, Y. Zhang, 2014.

Our strong stationarity applied to (1.6.6). The assumptions of Proposition 1.5.8 are satisfied. By using the expression (9) in Hiriart-Urruty, Malick, 2012 for the tangent cone $\mathcal{T}_{\mathbb{S}^n_-}(\bar{B})$, we obtain the optimality conditions (1.6.7a) and

$$U \in \mathcal{T}_{\mathbb{S}^n_-}(\bar{B}) \cap \bar{A}^\perp = \left\{ U \in \mathbb{S}^n, \;\; U = \begin{pmatrix} 0 & 0 & * \\ 0 & \leq 0 & * \\ * & * & * \end{pmatrix} \right\},$$

$$V \in \mathcal{T}_{\mathbb{S}^n_+}(\bar{A}) \cap \bar{B}^\perp = \left\{ V \in \mathbb{S}^n, \;\; V = \begin{pmatrix} * & * & * \\ * & \geq 0 & 0 \\ * & 0 & 0 \end{pmatrix} \right\}.$$

By comparing this with (1.6.7b), (1.6.7c) we find the well-known fact that $\mathbb{S}^n_+$, $\mathbb{S}^n_-$ are not polyhedric.

Our strong stationarity applied to a linearization. Now, we do not apply our result directly to (1.6.6), but to a certain linearization. We use the first-order condition

$$(f'(\bar{A},\bar{B}),\, h)_F \geq 0, \quad \forall h \in \mathcal{T}_F(\bar{A},\bar{B}).$$

In Corollary 3.2 in Wu, L. Zhang, Y. Zhang, 2014 we find an expression for this tangent cone. Together with the first-order condition, we infer that $(A,B) = (0,0)$ is a global minimizer of

$$\begin{array}{rl} \text{Minimize} & (f'(\bar{A},\bar{B}),\, (A,B))_F, \\ \text{s.t.} & A_{23} = A_{33} = B_{11} = B_{12} = 0, \quad A_{13} - B_{13} = 0, \\ \text{and} & A_{22} \geq 0,\; B_{22} \leq 0,\; A_{22}\, B_{22} = 0. \end{array} \tag{1.6.9}$$

Again, the assumptions of Proposition 1.5.8 are satisfied. We obtain the optimality conditions (1.6.7a) and

$$U = \begin{pmatrix} 0 & 0 & * \\ 0 & \leq 0 & * \\ * & * & * \end{pmatrix}, \quad V = \begin{pmatrix} * & * & * \\ * & \geq 0 & 0 \\ * & 0 & 0 \end{pmatrix}, \quad U_{13} + V_{13} = 0.$$

Note that the cone defining the complementarity constraint in (1.6.9) is polyhedric due to our choice of $(\bar{A}, \bar{B})$. However, similar arguments apply to the more general situation, since all cones are polyhedric w.r.t. the global minimizer $(A, B) = (0, 0)$ of (1.6.9). Hence, in all cases of $(\bar{A}, \bar{B})$, we obtain the same optimality system as with the tailored approach from the literature.

Applicability of the constraint qualifications. Let us briefly recall that the tailored condition (1.6.8) is shown in Proposition 5.1 and Corollary 5.1 of Ding, D. Sun, Ye, 2014 only under the requirement that the classical KKT conditions are valid. However, the classical KKT conditions (1.6.7) are stronger than the tailored conditions (1.6.8).

By means of an example we demonstrate that the classical KKT conditions are *strictly* stronger than the tailored conditions and may not be satisfied by a local minimizer. The unique global minimizer $(\bar{A}, \bar{B})$ of

$$\begin{aligned} \text{Minimize} \quad & (\operatorname{trace}(A) - 1)^2 + (\operatorname{trace}(B) + 1)^2 + A_{12} + A_{22} - B_{11} - B_{12}, \\ \text{s.t.} \quad & A \in \mathbb{S}^2_+, \ B \in \mathbb{S}^2_-, \ (A, B)_F = 0 \end{aligned}$$

is given by $\bar{A}_{11} = 1$, $\bar{A}_{12} = \bar{A}_{22} = 0$, $\bar{B}_{22} = -1$, $\bar{B}_{11} = \bar{B}_{12} = 0$. One can check that the classical KKT conditions (similar to (1.6.7)) cannot be satisfied, whereas the tailored condition (1.6.8) holds.

Hence, the approach from Ding, D. Sun, Ye, 2014 does only apply to this very simple problem, whereas our technique is applicable.

Comparison of the optimality systems. We emphasize that the above arguments suggest that the strong stationarity conditions from Ding, D. Sun, Ye, 2014 are the "correct" ones.

Moreover, the KKT conditions (1.6.7) are too strong for problem (1.6.6). Indeed, the $(1, 3)$-elements of the multipliers U and V are required to be zero, but this is stronger than the first-order condition (1.6.8). Hence, the KKT conditions are, in general, not satisfied. This is also demonstrated by means of an example.

Applying our optimality conditions directly to (1.6.6) yields necessary conditions which are too weak. However, we obtain the correct optimality conditions if the problem is linearized first.

Finally, we want to mention that all differences between the optimality systems pertain to the $(1, 3)$-elements of the multipliers U, V. Hence, these problems does not originate from the biactive component 2, but from the non-polyhedricity of $\mathbb{S}^n_+$. Indeed, in the special case $(\bar{A}, \bar{B}) = 0$, all presented optimality systems coincide.

1.7. Perspectives

We comment on some open problems.

(a) Similar to Definition 1.5.1 we can define weak stationarity conditions by the KKT conditions of (TNLP). Is there any analogue of the other stationarity concepts (such as Clarke- and Mordukhovich-stationarity) in the general situation?

(b) We have seen that the strong stationarity conditions from Definition 1.5.1 may be to weak in the non-polyhedric case. Can we state stronger necessary optimality conditions in this case?

(c) The constraint qualifications given in Section 1.5.3 are rather strong. This is in particular true for the accessible conditions in Proposition 1.5.8. Is it possible to state weaker conditions?

(d) In Section 1.6.1 we have used a density argument to infer the existence of multipliers. Can this density argument be avoided?

1.8. Conclusions

In this paper, we transfer results from standard mathematical programs with complementarity constraints in finite dimensions to general problems with general complementarity constraints

$$G(x) \in K, \qquad H(x) \in K^\circ, \qquad \langle G(x), H(x) \rangle = 0$$

in infinite-dimensional Banach spaces. In particular, we give a definition of strong stationarity; see Definition 1.5.1. We provide conditions, which imply that local optimizers are strongly stationary; see Section 1.5.3. In the case that K is polyhedric, we show that strong stationarity implies the linearized B-stationarity; see Theorem 1.5.4. Some of the results are also valid for the non-polyhedric case, but there remain some open problems in absence of polyhedricity.

Acknowledgment

The author would like to thank Radu Ioan Boţ for the idea leading to the counterexample at the end of Section 1.4.

2. Strong stationarity for optimization problems with complementarity constraints in absence of polyhedricity

Abstract: We consider mathematical programs with complementarity constraints in Banach spaces. In particular, we focus on the situation that the complementarity constraint is defined by a non-polyhedric cone K. We demonstrate how strong stationarity conditions can be obtained in an abstract setting. These conditions and their verification can be made more precise in the case that Z is a Hilbert space and if the projection onto K is directionally differentiable with a derivative as given in Haraux, 1977, Theorem 1. Finally, we apply the theory to optimization problems with semidefinite and second-order-cone complementarity constraints. We obtain that local minimizers are strongly stationary under a variant of the linear-independence constraint qualification, and these are novel results.

Keywords: mathematical programs with complementarity constraints, strong stationarity, conic programming, semidefinite cone, second-order cone, polyhedricity

MSC: 49K10, 49J52, 90C22, 90C33

2.1. Introduction

We consider the optimization problem

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{subject to} \quad & g(x) \in C, \\ & G(x) \in K, \\ & H(x) \in K^\circ, \\ & \langle G(x), H(x) \rangle = 0. \end{aligned} \tag{MPCC}$$

Here, $K \subset Z$ is a convex, closed cone in the reflexive Banach space Z. We refer to Section 2.4 for the precise assumptions on the remaining data of (MPCC). Due to the complementarity constraint

$$G(x) \in K,\ H(x) \in K^\circ,\ \langle G(x), H(x) \rangle = 0,$$

(MPCC) is a mathematical program with complementarity constraints (MPCC) in Banach space. Already in finite dimensions, these complementarity constraints induce several difficulties, see, e.g., Luo, Pang, Ralph, 1996; Scheel, Scholtes, 2000; Hoheisel, Kanzow, Schwartz, 2013 and the references therein. We are going to derive optimality conditions of strongly stationary type for local minimizers $\bar{x}$ of (MPCC).

We give a motivation for the study of (MPCC) with an emphasis on a situation in which the problem (MPCC) has infinite-dimensional components. A very important source of programs with complementarity constraints is given by *bilevel optimization problems*, see Dempe, 2002. Indeed, if we replace a lower-level problem which contains constraints including a cone K by its Karush-Kuhn-Tucker conditions, we arrive at (MPCC). We illustrate this by an example.

Let us consider an *optimal control problem* in which the final state enters an optimization problem as a parameter, for an example, we refer to the *natural gas cash-out problem*, see Kalashnikov, Benita, Mehlitz, 2015; Benita, Dempe, Mehlitz, 2016; Benita, Mehlitz, 2016. A simple prototype for such *bilevel optimal control problems* is

$$\begin{aligned} &\text{Minimize} && f(x,z), \\ &\text{such that} && \dot{x}(t) = F(t,x(t)) \quad \text{for } t \in (0,T), \\ &&& x(0) = x_0, \\ &&& z \text{ solves (2.1.2) with parameter } p = x(T), \end{aligned} \tag{2.1.1}$$

in which the lower-level problem is given by the following finite-dimensional optimization problem

$$\begin{aligned} &\text{Minimize} && j(z,p,q), \\ &\text{with respect to} && z \in \mathbb{R}^n, \\ &\text{such that} && g(z,p,q) \leq 0. \end{aligned} \tag{2.1.2}$$

This lower-level problem depends on the parameter $p = x(T)$ and another parameter q which is fixed for the moment. It is clear that (2.1.1) becomes a problem of form (MPCC) if we replace (2.1.2) by its KKT-conditions. In this particular situation, the cone K is given by $(-\infty, 0]^m$, where $\mathbb{R}^m$ is the range space of the nonlinear mapping g.

Now, let us consider the situation, in which the additional parameter q is *unknown* in the lower-level problem (2.1.2). One possibility to handle this uncertainty is the robust optimization approach from Ben-Tal, Nemirovski, 1998; Ben-Tal, Nemirovski, 2002. That is, the objective in (2.1.2) is replaced by its supremum over $q \in Q$ and the constraint $g(z,p,q) \leq 0$ has to be satisfied for all $q \in Q$, where Q is the uncertainty set. Depending on the type of the uncertainty set Q, this *robustification* of (2.1.2) becomes a problem with second-order cone or semidefinite constraints, see, e.g., Ben-Tal, Nemirovski, 2002, Theorem 1. Consequently, (2.1.1) (together with the Karush-Kuhn-Tucker conditions of the robust counterpart of (2.1.2)) becomes a problem of type (MPCC) with K being the second-order cone or the cone of semidefinite matrices. Finally, we mention that a problem of type (MPCC) with an infinite-dimensional cone K is obtained, if a lower-level problem is attached to (2.1.1) not only for the final time T, but for all $t \in (0,T)$. This

situation leads to the optimal control of a *differential variational inequality*, see Pang, Stewart, 2008 for applications and further references.

Strong stationarity for special instances of (MPCC) are obtained (for infinite-dimensional Z) in Mignot, 1976; Herzog, C. Meyer, G. Wachsmuth, 2013 and G. Wachsmuth, 2014 (i.e., Chapter 3) and in, e.g., Luo, Pang, Ralph, 1996; Ding, D. Sun, Ye, 2014; Ye, Zhou, 2015 (finite-dimensional Z). To our knowledge, G. Wachsmuth, 2015 (i.e., Chapter 1) is the only contribution which considers MPCCs of general type (MPCC). Therein, strong stationarity conditions for (MPCC) are defined and they possess a reasonable strength in the case that K is polyhedric. Moreover, it is shown that local minimizers are strongly stationary if a constraint qualification (CQ) is satisfied.

We are going to extend the results of Chapter 1 to the non-polyhedric situation. This is established by a linearization of (MPCC) and by utilizing the fact that every cone is polyhedric at the origin. We obtain explicit strong stationarity conditions in case that Z is a Hilbert space and if the directional derivative of the projection Proj_K onto K can be characterized as in Haraux, 1977, Theorem 1, see also Definition 2.4.6. Moreover, these conditions hold at local minimizers if a reasonable constraint qualification is satisfied.

We describe the main results of this work. In Section 2.3 we provide two auxiliary results, which are used in Section 2.4 to derive stationarity conditions for (MPCC). We believe that both results are of independent interest. In Section 2.3.1, we consider the set

$$\mathcal{F} := \{x \in \mathcal{C} : \mathcal{G}(x) \in \mathcal{K}\}.$$

Here, $\mathcal{C}$ is a closed convex set and the closed set $\mathcal{K}$ is *not* assumed to be convex. We show that the tangent cone $\mathcal{T}_{\mathcal{F}}(\bar{x})$ of $\mathcal{F}$ at $\bar{x} \in \mathcal{F}$ is obtained by the linearization

$$\{x \in \mathcal{T}_{\mathcal{C}}(\bar{x}) : \mathcal{G}'(\bar{x})\, x \in \mathcal{T}_{\mathcal{K}}(\mathcal{G}(\bar{x}))\}$$

if a certain CQ is satisfied, see Theorem 2.3.6. The second auxiliary result concerns the characterization of the tangent cone of the graph of the normal cone mapping. That is, given a closed convex cone K in a real Hilbert space Z, we are interested in the characterization of the tangent cone to the set

$$\operatorname{gph} \mathcal{T}_K(\cdot)^\circ = \{(z, z^\star) \subset Z \times Z^\star : z \in K, z^\star \in K^\circ, \langle z,\, z^\star \rangle = 0\}.$$

If the metric projection onto the cone K is directionally differentiable as in Haraux, 1977, Theorem 1, we obtain a precise characterization of the tangent cone of $\operatorname{gph} \mathcal{T}_K(\cdot)^\circ$, see Section 2.3.2.

These auxiliary results are utilized in order to prove strong stationarity of minimizers to (MPCC) in Section 2.4. The main result of Section 2.4 is Theorem 2.4.5, in which we prove that local minimizers satisfy the system of strong stationarity (2.4.15) under certain CQs, and this could also be considered the main result of this work.

In order to illustrate the strength of the abstract theory, we study the cases that K is the cone of semidefinite matrices and the second-order cone, see Sections 2.5 and 2.6. In both situations, we obtain novel results. In fact, for the SDPMPCC we obtain that

SDPMPCC-LICQ implies strong stationarity of local minimizers, see Theorem 2.5.9, and this was stated as an open problem in Ding, D. Sun, Ye, 2014, Remark 6.1. For the SOCMPCC we obtain the parallel result that SOCMPCC-LICQ implies strong stationarity, see Theorem 2.6.11. This result was already provided in Ye, Zhou, 2015, Theorem 5.1, but the definition of SOCMPCC-LICQ in Ye, Zhou, 2015 is stronger than our definition. Finally, we consider briefly the case that K is an infinite product of second-order cones in Section 2.7.

2.2. Notation

Let X be a (real) Banach space. The (norm) closure of a subset $A \subset X$ is denoted by $\mathrm{cl}(A)$. The linear subspace spanned by A is denoted by $\mathrm{lin}(A)$. The duality pairing between X and its topological dual $X^\star$ is denoted by $\langle \cdot, \cdot \rangle : X^\star \times X \to \mathbb{R}$. For subsets $A \subset X$, $B \subset X^\star$, we define their polar cones and annihilators via

$$A^\circ := \{x^\star \in X^\star : \langle x^\star, x\rangle \leq 0,\ \forall x \in A\}, \qquad B^\circ := \{x \in X : \langle x^\star, x\rangle \leq 0,\ \forall x^\star \in B\},$$
$$A^\perp := \{x^\star \in X^\star : \langle x^\star, x\rangle = 0,\ \forall x \in A\}, \qquad B^\perp := \{x \in X : \langle x^\star, x\rangle = 0,\ \forall x^\star \in B\}.$$

For a cone $C \subset X$, $C^{\circ\perp} = C \cap -C$ is the lineality space of C, that is, the largest subspace contained in C. For an arbitrary set $C \subset X$ and $x \in C$, we define the cone of feasible directions (also called the radial cone), the Bouligand (or contingent) tangent cone and the adjacent (or inner) tangent cone by

$$\begin{aligned}
\mathcal{R}_C(x) &:= \{h \in X : \exists t > 0 : \forall s \in [0,t] : x + s\,h \in C\},\\
\mathcal{T}_C(x) &:= \{h \in X : \liminf_{t \searrow 0} \mathrm{dist}_C(x + t\,h)/t = 0\},\\
\mathcal{T}_C^\flat(x) &:= \{h \in X : \limsup_{t \searrow 0} \mathrm{dist}_C(x + t\,h)/t = 0\},
\end{aligned}$$

respectively, see Bonnans, Shapiro, 2000, Definition 2.54, Aubin, Frankowska, 2009, Definitions 4.1.1 and 4.1.5. Here, $\mathrm{dist}_C(x)$ is the distance of $x \in X$ to the set C. Recall, that $\mathcal{T}_C(x) = \mathcal{T}_C^\flat(x) = \mathrm{cl}(\mathcal{R}_C(x))$ in case C is convex, see Bonnans, Shapiro, 2000, Proposition 2.55. Moreover, if C is a convex cone, we find

$$\mathcal{T}_C(x)^\circ = C^\circ \cap x^\perp, \tag{2.2.1}$$

see Bonnans, Shapiro, 2000, Example 2.62. Here, $x^\perp$ is short for $\{x\}^\perp$. For a closed, convex set $C \subset X$, we define the critical cone w.r.t. $x \in C$ and $v \in \mathcal{T}_C(x)^\circ$ by

$$\mathcal{K}_C(x, v) := \mathcal{T}_C(x) \cap v^\perp. \tag{2.2.2}$$

The set C is said to be polyhedric w.r.t. $(x, v) \in C \times \mathcal{T}_C(x)^\circ$, if

$$\mathcal{K}_C(x, v) = \mathrm{cl}(\mathcal{R}_C(x) \cap v^\perp) \tag{2.2.3}$$

holds, cf. Haraux, 1977.

A function $g : X \to Y$ is called strictly Fréchet differentiable at $\bar{x} \in X$, if it is Fréchet differentiable at $\bar{x}$ and if for every $\delta > 0$, there exists $\varepsilon > 0$ such that

$$\|g(x_1) - g(x_2) - g'(\bar{x})(x_1 - x_2)\|_Y \leq \delta \|x_1 - x_2\|_X \qquad \forall x_1, x_2 \in B_\varepsilon^X(\bar{x}). \tag{2.2.4}$$

Here, $B_\varepsilon^X(\bar{x})$ is the closed ball around $\bar{x}$ with radius ε. Note that strict Fréchet differentiability is implied by continuous Fréchet differentiability, see Cartan, 1967, Theorem 3.8.1.

2.3. Auxiliary results

In this section, we provide two auxiliary results, which will be utilized in the analysis of (MPCC). However, we believe that the results are of independent interest.

2.3.1. Constraint qualification in absence of convexity

We consider the set

$$\mathcal{F} := \{x \in \mathcal{C} : \mathcal{G}(x) \in \mathcal{K}\} \tag{2.3.1}$$

which is defined by a (possibly) *non-convex* set $\mathcal{K}$. We are going to characterize its tangent cone $\mathcal{T}_\mathcal{F}(\bar{x})$ by means of its linearization cone

$$\mathcal{T}_{\text{lin}}(\bar{x}) := \{h \in \mathcal{T}_\mathcal{C}(\bar{x}) : \mathcal{G}'(\bar{x})\, h \in \mathcal{T}_\mathcal{K}(\mathcal{G}(\bar{x}))\}.$$

Since $\mathcal{K}$ may not be convex, we cannot apply the constraint qualification of Robinson, 1976; Zowe, Kurcyusz, 1979 to obtain a characterization of the tangent cone $\mathcal{T}_\mathcal{F}(\bar{x})$ of $\mathcal{F}$. The main result of this section is Theorem 2.3.6, which provides a constraint qualification that applies to the situation of (2.3.1).

We fix the setting of (2.3.1). Let $\mathcal{X}, \mathcal{Y}$ be (real) Banach spaces. The set $\mathcal{C} \subset \mathcal{X}$ is assumed to be closed and convex and the set $\mathcal{K} \subset \mathcal{Y}$ is closed. We emphasize that the set $\mathcal{K}$ is *not* assumed to be convex. Moreover, the function $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ is strictly Fréchet differentiable at the point $\bar{x} \in \mathcal{F}$.

As in the situation that $\mathcal{K}$ is convex, no condition is required to show that the linearization cone is a superset of the tangent cone and we recall the following lemma.

Lemma 2.3.1. We have $\mathcal{T}_\mathcal{F}(\bar{x}) \subset \mathcal{T}_{\text{lin}}(\bar{x})$.

In order to define a CQ, we need a certain tangent approximation of the non-convex set $\mathcal{K}$.

Definition 2.3.2. A cone $\mathcal{W} \subset \mathcal{Y}$ is called a tangent approximation set of $\mathcal{K}$ at $\bar{y} \in \mathcal{K}$, if for all $\rho \in (0, 1)$, there exists $\delta > 0$, such that

$$\forall y \in \mathcal{K} \cap B_\delta^\mathcal{Y}(\bar{y}), w \in \mathcal{W} \cap B_\delta^\mathcal{Y}(0) : \exists \tilde{w} \in \mathcal{Y} : \|\tilde{w}\|_\mathcal{Y} \leq \rho \|w\|_\mathcal{Y}, y + w + \tilde{w} \in \mathcal{K}. \tag{2.3.2}$$

Note that the set $\mathcal{W}$ is not assumed to be convex or closed. Roughly speaking, all (small) directions in $\mathcal{W}$ are close to tangent directions for all points in a neighborhood of $\bar{y}$. To our knowledge, this definition was not used in the literature so far.

We shall show that a tangent approximation set $\mathcal{W}$ is a subset of the Clarke tangent cone defined by

$$C_{\mathcal{K}}(\bar{y}) := \{d \in \mathcal{Y} : \forall \{y_k\}_{k\in\mathbb{N}} \subset \mathcal{K}, y_k \to \bar{y}, \\ \forall \{t_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^+, t_k \searrow 0 : \operatorname{dist}_{\mathcal{K}}(y_k + t_k\, d)/t_k \to 0\},$$

see, e.g., Aubin, Frankowska, 2009, Definition 4.1.5. Here, $\operatorname{dist}_{\mathcal{K}}$ is the distance function of $\mathcal{K}$. Moreover, in finite dimensions, the Clarke tangent cone is the largest tangent approximation set.

Lemma 2.3.3. We assume that the cone $\mathcal{W} \subset \mathcal{Y}$ is a tangent approximation of $\mathcal{K}$ at $\bar{y} \in \mathcal{K}$. Then, $\mathcal{W} \subset C_{\mathcal{K}}(\bar{y})$. Moreover, $C_{\mathcal{K}}(\bar{y})$ is a tangent approximation set of $\mathcal{K}$ at $\bar{y}$ if $\dim(\mathcal{Y}) < \infty$.

Proof. Let $w \in \mathcal{W}$ and sequences $\{y_k\}$ and $\{t_k\}$ as in the definition of $C_{\mathcal{K}}(\bar{y})$ be given. For a given $\rho \in (0,1)$, there exists $\delta > 0$, such that (2.3.2) holds. Then, $y_k \in \mathcal{K} \cap B^{\mathcal{Y}}_{\delta}(\bar{y})$ and $t_k\, w \in \mathcal{W} \cap B^{\mathcal{Y}}_{\delta}(0)$ for large k. By (2.3.2), there exists $\tilde{w}_k$ with $\|\tilde{w}_k\|_{\mathcal{Y}} \le \rho\, t_k\, \|w\|_{\mathcal{Y}}$ and $y_k + t_k\, w + \tilde{w}_k \in \mathcal{K}$. Hence, $\operatorname{dist}_{\mathcal{K}}(y_k + t_k\, w) \le \|\tilde{w}_k\|_{\mathcal{Y}} \le \rho\, t_k\, \|w\|_{\mathcal{Y}}$. Thus, $\operatorname{dist}_{\mathcal{K}}(y_k + t_k\, w)/t_k \le \rho\, \|w\|_{\mathcal{Y}}$ for large k. Since $\rho \in (0,1)$ was arbitrary, this shows the claim.

To prove that $C_{\mathcal{K}}(\bar{y})$ is a tangent approximation set in case $\dim(\mathcal{Y}) < \infty$, we proceed by contradiction. Hence, there exists $\rho > 0$, a sequence $\{y_k\}_{k\in\mathbb{N}} \subset \mathcal{K}$ with $y_k \to \bar{y}$ and $\{w_k\}_{k\in\mathbb{N}} \in C_{\mathcal{K}}(\bar{y})$ with $w_k \to 0$ such that

$$\operatorname{dist}_{\mathcal{K}}(y_k + w_k) \ge \rho\, \|w_k\|_{\mathcal{Y}} \quad \forall k \in \mathbb{N}.$$

We set $t_k = \|w_k\|_{\mathcal{Y}}$. Since $\mathcal{Y}$ is finite-dimensional, the bounded sequence w_k/t_k possesses a limit w with $\|w\|_{\mathcal{Y}} = 1$ (up to a subsequence, without relabeling). By the closedness of $C_{\mathcal{K}}(\bar{y})$, we have $w \in C_{\mathcal{K}}(\bar{y})$, see Aubin, Frankowska, 2009, p.127. Hence,

$$\frac{\operatorname{dist}_{\mathcal{K}}(y_k + t_k\, w)}{t_k} \ge \frac{\operatorname{dist}_{\mathcal{K}}(y_k + t_k\, (w_k/t_k))}{t_k} - \frac{\|t_k\, w - w_k\|_{\mathcal{Y}}}{t_k} \ge \rho - \|w - w_k/t_k\|_{\mathcal{Y}}.$$

Together with $w_k/t_k \to w$ and $\rho > 0$, this is a contradiction to $w \in C_{\mathcal{K}}(\bar{y})$.

The next lemma suggests that the situation is much more delicate in the infinite-dimensional case. In particular, tangent approximation sets cannot be chosen convex and there might be no largest tangent approximation set.

Lemma 2.3.4. We consider $\mathcal{Y} = L^2(0,1)$ and the closed, convex set $\mathcal{K} = \{y \in L^2(0,1) : y(t) \ge 0 \text{ for a.a. } t \in (0,1)\}$. Then for any $M > 0$, the cone

$$\mathcal{W}_M = \{w \in L^2(0,1) : \|\min(w,0)\|_{L^\infty(0,1)} \le M\, \|w\|_{L^2(0,1)}\}$$

is a tangent approximation set of $\mathcal{K}$ at the constant function $\bar{y} \equiv 1$.

However, $\mathcal{R}_{\mathcal{K}}(\bar{y})$, $\mathcal{T}_{\mathcal{K}}(\bar{y})$, the union of all $\mathcal{W}_M$, $M > 0$ and the convex hull of $\mathcal{W}_M$ for $M \geq \sqrt{2}$ are not tangent approximation sets.

Proof. Let us show that $\mathcal{W}_M$ is a tangent approximation set of $\mathcal{K}$ at $\bar{y}$. Let $\rho \in (0,1)$ be given. We set $\delta = \rho/(2\,M)$ and show that (2.3.2) is satisfied. To this end, let $y \in \mathcal{K}$ and $w \in \mathcal{W}_M$ with $\|y - \bar{y}\|_{L^2(0,1)}, \|w\|_{L^2(0,1)} \leq \delta$ be given. By Chebyshev's inequality, we have

$$\begin{aligned}\mu(\{t \in (0,1) : y(t) \leq 1/2\}) &\leq \mu(\{t \in (0,1) : |y(t) - \bar{y}(t)| \geq 1/2\}) \\ &\leq 2^2 \, \|y - \bar{y}\|^2_{L^2(0,1)} \leq 4\,\delta^2,\end{aligned}$$

where μ denotes the Lebesgue measure on $(0,1)$. Since

$$\|\min(w,0)\|_{L^\infty(0,1)} \leq M \, \|w\|_{L^2(0,1)} \leq 1/2,$$

we have

$$\mu(\{t \in (0,1) : y(t) + w(t) \leq 0\}) \leq \mu(\{t \in (0,1) : y(t) \leq 1/2\}) \leq 4\,\delta^2.$$

We set $\tilde{w} := -\min(0, y + w)$, which yields $y + w + \tilde{w} = \max(0, y + w) \in \mathcal{K}$. It remains to bound $\tilde{w}$ and this is established by

$$\begin{aligned}\|\tilde{w}\|^2_{L^2(0,1)} &\leq \|\min(0, y + w)\|^2_{L^2(0,1)} \\ &= \int_{\{t \in (0,1) : y(t) + w(t) \leq 0\}} (y(t) + w(t))^2 \, \mathrm{d}t \leq \int_{\{t \in (0,1) : y(t) + w(t) \leq 0\}} w(t)^2 \, \mathrm{d}t \\ &\leq \mu(\{t \in (0,1) : y(t) + w(t) \leq 0\}) \, \|\min(w,0)\|^2_{L^\infty(0,1)} \\ &\leq 4\,\delta^2 \, M^2 \, \|w\|^2_{L^2(0,1)} = \rho^2 \, \|w\|^2_{L^2(0,1)}.\end{aligned}$$

This shows that $\mathcal{W}_M$ is a tangent approximation set.

In order to demonstrate the last claim of the lemma, we first show that $L^\infty(0,1)$ is not a tangent approximation set. This is easily established, since for arbitrary $\delta > 0$, we may set

$$y(t) = \begin{cases} 0 & \text{if } t \leq \delta^2, \\ 1 & \text{else,} \end{cases} \qquad \text{and} \qquad w(t) = \begin{cases} -1 & \text{if } t \leq \delta^2, \\ 0 & \text{else.} \end{cases}$$

It is clear that $y \in \mathcal{K}$ and $w \in L^\infty(0,1)$ and $\|y - \bar{y}\|_{L^2(0,1)} = \|w\|_{L^2(0,1)} = \delta$. However, the distance of $y + w$ to $\mathcal{K}$ is δ. Hence, (2.3.2) cannot be satisfied with $\mathcal{W} = L^\infty(0,1)$.

Since $L^\infty(0,1) \subset \mathcal{R}_{\mathcal{K}}(\bar{y}) \subset \mathcal{T}_{\mathcal{K}}(\bar{y})$ and $L^\infty(0,1) \subset \bigcup_{M>0} \mathcal{W}_M$, the sets $\mathcal{R}_{\mathcal{K}}(\bar{y})$, $\mathcal{T}_{\mathcal{K}}(\bar{y})$ and $\bigcup_{M>0} \mathcal{W}_M$ cannot be tangent approximation sets.

Finally, we show that $L^\infty(0,1)$ is contained in the convex hull of $\mathcal{W}_M$, $M \geq \sqrt{2}$. Indeed, for arbitrary $f \in L^\infty(0,1)$, we define $f_1 = 2\,f\,\chi_{[0,1/2]}$, $f_2 = 2\,f\,\chi_{[1/2,1]}$ and it is sufficient

to show that f_1 and f_2 belong to the convex hull $\operatorname{conv}(\mathcal{W}_M)$ of $\mathcal{W}_M$. Therefore, we set $g^{\pm} = f_1 \pm \|f_1\|_{L^\infty(0,1)}\, \chi_{[1/2,1]}$. This gives

$$\|g^{\pm}\|_{L^\infty(0,1)} = \|f_1\|_{L^\infty(0,1)} \qquad \text{and} \qquad \|g^{\pm}\|_{L^2(0,1)} \geq \|f_1\|_{L^\infty(0,1)}/\sqrt{2}.$$

Hence, $g^{\pm} \in \mathcal{W}_M$. This shows $f_1 = (g^+ + g^-)/2 \in \operatorname{conv}(\mathcal{W}_M)$ and similarly we get $f_2 \in \operatorname{conv}(\mathcal{W}_M)$. Hence, $L^\infty(0,1) \subset \operatorname{conv}(\mathcal{W}_M)$ and, thus, $\operatorname{conv}(\mathcal{W}_M)$ is not a tangent approximation set.

We mention that it is also possible to derive non-trivial tangent approximation sets for non-convex sets $\mathcal{K}$. As an example, we mention

$$\mathcal{K} := \{y \in L^2(0,1)^2 : y_1 \geq 0,\ y_2 \geq 0,\ y_1\, y_2 = 0 \text{ a.e. in } (0,1)\}.$$

Given $\bar{y} \in \mathcal{K}$, $M, \varepsilon > 0$, one can use similar arguments as in the proof of Lemma 2.3.4 to show that

$$\mathcal{W} = \left\{ w \in L^2(0,1)^2 : \begin{array}{c} \|w\|_{L^\infty(0,1)^2} \leq M\, \|w\|_{L^2(0,1)^2}, \\ w_1(t) = 0 \text{ if } y_1(t) \leq \varepsilon, \quad \text{for a.a. } t \in (0,1), \\ w_2(t) = 0 \text{ if } y_2(t) \leq \varepsilon, \quad \text{for a.a. } t \in (0,1) \end{array} \right\} \tag{2.3.3}$$

is a tangent approximation set of $\mathcal{K}$ at $\bar{y}$.

In what follows, we define

$$(A)_1 := A \cap B_1^{\mathcal{Z}}(0),$$

where A is a subset of a Banach space $\mathcal{Z}$ and $B_1^{\mathcal{Z}}(0)$ is the (closed) unit ball in $\mathcal{Z}$.

Using the notion of a tangent approximation set, we define a constraint qualification.

Definition 2.3.5. The set $\mathcal{F}$ is called *qualified* at $\bar{x} \in \mathcal{F}$ if there exists a tangent approximation set $\mathcal{W} \subset \mathcal{Y}$ of $\mathcal{K}$ at $\mathcal{G}(\bar{x})$ and $M > 0$ such that

$$B_M^{\mathcal{Y}}(0) \subset \mathcal{G}'(\bar{x})\, (\mathcal{C} - \bar{x})_1 - (\mathcal{W})_1. \tag{2.3.4}$$

If the set $\mathcal{W}$ is actually a closed and convex cone, the condition (2.3.4) is equivalent to

$$\mathcal{Y} = \mathcal{G}'(\bar{x})\, \mathcal{R}_{\mathcal{C}}(\bar{x}) - \mathcal{W}, \tag{2.3.5}$$

see Zowe, Kurcyusz, 1979, Theorem 2.1.

In the case that $\mathcal{Y}$ is finite-dimensional and $\mathcal{K}$ is convex, Definition 2.3.5 reduces to the constraint qualification of Robinson-Zowe-Kurcyusz. Indeed, in the light of Lemma 2.3.3, Definition 2.3.5 is equivalent to

$$B_M^{\mathcal{Y}}(0) \subset \mathcal{G}'(\bar{x})\, (\mathcal{C} - \bar{x})_1 - (\mathcal{T}_{\mathcal{K}}(\mathcal{G}(\bar{x})))_1,$$

which is, in turn, equivalent to the constraint qualification of Robinson-Zowe-Kurcyusz, see Bonnans, Shapiro, 2000, Proposition 2.97.

Theorem 2.3.6. Let us assume that $\mathcal{F}$ is qualified at $\bar{x} \in \mathcal{F}$ in the sense of Definition 2.3.5. Then,

$$\mathcal{T}_{\text{lin}}(\bar{x}) = \mathcal{T}_{\mathcal{F}}(\bar{x}).$$

In the case that $\mathcal{K}$ is convex, this assertion is well known, if we replace $\mathcal{W}$ by $\mathcal{R}_{\mathcal{K}}(\mathcal{G}(\bar{x}))$ in (2.3.5). Note that this assumption might be weaker then the qualification of $\mathcal{F}$ in the sense of Definition 2.3.5, compare Lemma 2.3.4.

Moreover, if $\mathcal{Y}$ is finite-dimensional, the assertion follows from Lemmas 2.3.1 and 2.3.3, and Aubin, Frankowska, 2009, Theorem 4.3.3.

In order to prove Theorem 2.3.6, we provide an auxiliary lemma. Its proof is inspired by the proof of Werner, 1984, Theorem 5.2.5.

Lemma 2.3.7. Let the assumptions of Theorem 2.3.6 be satisfied. Then, there exists $\gamma > 0$, such that for all $u \in \mathcal{X}$, $v \in \mathcal{Y}$ satisfying

$$\|u\|_{\mathcal{X}} \leq \gamma, \qquad \|v\|_{\mathcal{Y}} \leq \gamma, \qquad u \in \frac{1}{2}(\mathcal{C} - \bar{x})_1, \qquad \mathcal{G}(\bar{x}) + \mathcal{G}'(\bar{x})\, u + v \in \mathcal{K},$$

we find $\hat{u} \in \mathcal{X}$ satisfying

$$\hat{u} \in (\mathcal{C} - \bar{x})_1, \qquad \mathcal{G}(\bar{x} + \hat{u}) \in \mathcal{K}$$

and $\|u - \hat{u}\|_{\mathcal{X}} \leq \frac{2}{M} \|\mathcal{G}(\bar{x} + u) - \mathcal{G}(\bar{x}) - \mathcal{G}'(\bar{x})\, u - v\|_{\mathcal{Y}}$.

Proof. We set $\varepsilon := \rho := M/4$. Since $\mathcal{G}$ is strictly differentiable at $\bar{x}$, there exists $\delta \in (0, 1]$, such that

$$\left\|\mathcal{G}(x) - \mathcal{G}(\tilde{x}) - \mathcal{G}'(\bar{x})\,(x - \tilde{x})\right\|_{\mathcal{Y}} \leq \varepsilon \|x - \tilde{x}\|_{\mathcal{X}} \qquad \forall x, \tilde{x} \in B_\delta(\bar{x}) \tag{2.3.6}$$

and such that (2.3.2) is satisfied.

We define the constants

$$L := \max\{\|\mathcal{G}'(\bar{x})\|_{\mathcal{Y}}, \varepsilon\}, \qquad \gamma := \frac{\delta}{4\,(L+1)\,(M^{-1}+1)} \leq \frac{\delta}{4}.$$

Now, let $u \in \mathcal{X}$ and $v \in \mathcal{Y}$ with

$$\|u\|_{\mathcal{X}} \leq \gamma, \qquad \|v\|_{\mathcal{Y}} \leq \gamma, \qquad u \in \frac{1}{2}(\mathcal{C} - \bar{x})_1, \qquad \text{and} \qquad \mathcal{G}(\bar{x}) + \mathcal{G}'(\bar{x})\, u + v \in \mathcal{K}$$

be given.

We define

$$q := \frac{2}{M} \left\|\mathcal{G}(\bar{x} + u) - \mathcal{G}(\bar{x}) - \mathcal{G}'(\bar{x})\, u - v\right\|_{\mathcal{Y}}.$$

For later reference, we provide

$$q \le \frac{2}{M}\left(\varepsilon\,\|u\|_{\mathcal{X}} + \|v\|_{\mathcal{Y}}\right) \le \frac{2}{M}\left(\frac{M}{4}+1\right)\gamma = \left(\frac{1}{2}+\frac{2}{M}\right)\gamma \le \frac{\delta}{2\,(L+1)} \le \frac{\delta}{2} \le \frac{1}{2}. \quad (2.3.7)$$

We set

$$x_0 := 0, \qquad u_0 := u, \qquad v_0 := v$$

and construct sequences $\{x_i\}_{i\in\mathbb{N}}$, $\{u_i\}_{i\in\mathbb{N}}$, $\{v_i\}_{i\in\mathbb{N}}$, such that the following assertions hold

$$u_i \in \frac{q}{2^i}\,(\mathcal{C}-\bar{x})_1 \qquad \forall i \ge 1, \quad (2.3.8a)$$

$$x_i \in \left(1-\frac{1}{2^i}\right)(\mathcal{C}-\bar{x})_1 \qquad \forall i \ge 0, \quad (2.3.8b)$$

$$\|x_i\|_{\mathcal{X}} - \|u\|_{\mathcal{X}} \le \left(1-\frac{1}{2^{i-1}}\right) q \qquad \forall i \ge 1, \quad (2.3.8c)$$

$$\|v_i\|_{\mathcal{Y}} \le \frac{\rho\, q}{2^i} \qquad \forall i \ge 1, \quad (2.3.8d)$$

$$\mathcal{G}(\bar{x}+x_i) + \mathcal{G}'(\bar{x})\,u_i + v_i \in \mathcal{K} \qquad \forall i \ge 0, \quad (2.3.8e)$$

$$\|\mathcal{G}(\bar{x}+x_i+u_i) - \mathcal{G}(\bar{x}+x_i) - \mathcal{G}'(\bar{x})\,u_i - v_i\|_{\mathcal{Y}} \le \frac{M\,q}{2^{i+1}} \qquad \forall i \ge 0. \quad (2.3.8f)$$

Note that (2.3.8b) (2.3.8e) and (2.3.8f), are satisfied for $i = 0$.

For $i = 1, 2, \ldots$ we perform the following. Since (2.3.4) is satisfied, we find

$$\begin{pmatrix} u_i \\ w_i \end{pmatrix} \in \frac{1}{M}\,\|\mathcal{G}(\bar{x}+x_{i-1}+u_{i-1}) - \mathcal{G}(\bar{x}+x_{i-1}) - \mathcal{G}'(\bar{x})\,u_{i-1} - v_{i-1}\|_{\mathcal{Y}} \begin{pmatrix} (\mathcal{C}-\bar{x})_1 \\ (\mathcal{W})_1 \end{pmatrix}, \quad (2.3.9)$$

such that

$$\mathcal{G}'(\bar{x})\,u_i - w_i = -\big(\mathcal{G}(\bar{x}+x_{i-1}+u_{i-1}) - \mathcal{G}(\bar{x}+x_{i-1}) - \mathcal{G}'(\bar{x})\,u_{i-1} - v_{i-1}\big). \quad (2.3.10)$$

Now, (2.3.9) together with (2.3.8f) for $i-1$ shows (2.3.8a) for i. Similarly, we obtain

$$\|w_i\|_{\mathcal{Y}} \le \frac{q}{2^i} \le q \le \delta. \quad (2.3.11)$$

We define

$$x_i := x_{i-1} + u_{i-1}.$$

We have

$$u_0 = u \in \frac{1}{2}\,(\mathcal{C}-\bar{x})_1, \qquad \text{in case } i = 1,$$

$$u_{i-1} \in \frac{q}{2^{i-1}}\,(\mathcal{C}-\bar{x})_1 \subset \frac{1}{2^i}\,(\mathcal{C}-\bar{x})_1, \qquad \text{in case } i > 1 \text{ by (2.3.7), (2.3.8a).}$$

This implies

$$x_i = x_{i-1} + u_{i-1} \in (1 - \frac{1}{2^{i-1}})\,(\mathcal{C} - \bar{x})_1 + (\frac{1}{2^i})\,(\mathcal{C} - \bar{x})_1 \subset (1 - \frac{1}{2^i})\,(\mathcal{C} - \bar{x})_1,$$

which yields (2.3.8b). Similarly, (2.3.8c) follows. Together with (2.3.8a), this implies

$$\|x_i\|_{\mathcal{X}} \le \|u\|_{\mathcal{X}} + q \le \delta \qquad \text{and} \qquad \|x_i + u_i\|_{\mathcal{X}} \le \|u\|_{\mathcal{X}} + q \le \delta. \tag{2.3.12}$$

By (2.3.10) and (2.3.8e) for $i-1$, we have

$$\mathcal{G}(\bar{x} + x_i) + \mathcal{G}'(\bar{x})\,u_i - w_i = \mathcal{G}(\bar{x} + x_{i-1}) + \mathcal{G}'(\bar{x})\,u_{i-1} + v_{i-1} \in \mathcal{K}.$$

Since

$$\begin{aligned}
&\|\mathcal{G}(\bar{x} + x_i) + \mathcal{G}'(\bar{x})\,u_i - w_i - \mathcal{G}(\bar{x})\|_{\mathcal{Y}} \\
&\qquad \le \|\mathcal{G}(\bar{x} + x_i) - \mathcal{G}(\bar{x}) - \mathcal{G}'(\bar{x})\,x_i\|_{\mathcal{Y}} + \|\mathcal{G}'(\bar{x})\,(x_i + u_i)\|_{\mathcal{Y}} + \|w_i\|_{\mathcal{Y}} \\
&\qquad \overset{(2.3.6)}{\le} \varepsilon\,\|x_i\|_{\mathcal{X}} + L\,\|x_i + u_i\|_{\mathcal{X}} + \|w_i\|_{\mathcal{Y}} \\
&\qquad \overset{(2.3.11),\,(2.3.12)}{\le} 2\,(L+1)\,(\|u\|_{\mathcal{X}} + q) \\
&\qquad \overset{(2.3.7)}{\le} 2\,(L+1)\,(1 + \frac{1}{2} + \frac{2}{M})\,\gamma \le 4\,(L+1)\,(1 + \frac{1}{M})\,\gamma = \delta
\end{aligned}$$

and $\|w_i\|_{\mathcal{Y}} \le \delta$ (by (2.3.11)), we can apply (2.3.2) and find v_i with $\|v_i\|_{\mathcal{Y}} \le \rho\,\|w_i\|_{\mathcal{Y}}$ and

$$\mathcal{G}(\bar{x} + x_i) + \mathcal{G}'(\bar{x})\,u_i - w_i + w_i + v_i \in \mathcal{K},$$

which shows (2.3.8e) and (2.3.8d) for i.

It remains to show (2.3.8f). By (2.3.12), we can apply (2.3.6) and find

$$\|\mathcal{G}(\bar{x} + x_i + u_i) - \mathcal{G}(\bar{x} + x_i) - \mathcal{G}'(\bar{x})\,u_i - v_i\|_{\mathcal{Y}} \le \varepsilon\,\|u_i\|_{\mathcal{X}} + \|v_i\|_{\mathcal{Y}} \le (\varepsilon + \rho)\,\frac{q}{2^i} = \frac{M\,q}{2^{i+1}}.$$

Altogether, we have shown (2.3.8).

Since $x_i = \sum_{j=0}^{i-1} u_j$ we get from (2.3.8a) that $\{x_i\}$ is a Cauchy sequence. Hence, there is $\hat{u}$ with $x_i \to \hat{u}$ in $\mathcal{X}$. Moreover, we have $u_i \to 0$ in $\mathcal{X}$ and $v_i \to 0$ in $\mathcal{Y}$ by (2.3.8a) and (2.3.8d).

Hence, passing to the limit $i \to \infty$ in (2.3.8e) and using the closedness of $\mathcal{K}$, we get

$$\mathcal{G}(\bar{x} + \hat{u}) \in \mathcal{K}$$

and from (2.3.8b) we find

$$\hat{u} \in (\mathcal{C} - \bar{x})_1.$$

Finally,

$$\|u - \hat{u}\|_{\mathcal{X}} = \|\sum_{j=1}^{\infty} u_j\|_{\mathcal{X}} \le \sum_{j=1}^{\infty} \frac{q}{2^j} \le q.$$

Proof of Theorem 2.3.6. In view of Lemma 2.3.1, we only need to prove $\mathcal{T}_{\text{lin}}(\bar{x}) \subset \mathcal{T}_{\mathcal{F}}(\bar{x})$. Let $h \in \mathcal{T}_{\text{lin}}(\bar{x})$ be given. Since $\mathcal{G}'(\bar{x})\,h \in \mathcal{T}_{\mathcal{K}}(\mathcal{G}(\bar{x}))$, there exist sequences $\{t_k\}_{k\in\mathbb{N}} \subset (0,\infty)$ and $\{r_k^{\mathcal{K}}\}_{k\in\mathbb{N}} \subset \mathcal{Y}$, such that $t_k \to 0$ and

$$\begin{aligned} \mathcal{G}(\bar{x}) + t_k\,\mathcal{G}'(\bar{x})\,h + r_k^{\mathcal{K}} \in \mathcal{K} \quad &\text{for all } k \in \mathbb{N} \\ \|r_k^{\mathcal{K}}\|_{\mathcal{Y}} = o(t_k) \quad &\text{as } k \to \infty \end{aligned}$$

By the convexity of $\mathcal{C}$ and $h \in \mathcal{T}_{\mathcal{C}}(\bar{x})$, there exists $\{r_k^{\mathcal{C}}\}_{k\in\mathbb{N}} \subset \mathcal{X}$ with

$$\begin{aligned} \bar{x} + 2\,t_k\,h + 2\,r_k^{\mathcal{C}} \in \mathcal{C} \quad &\text{for all } k \in \mathbb{N} \\ \|r_k^{\mathcal{C}}\|_{\mathcal{X}} = o(t_k) \quad &\text{as } k \to \infty \end{aligned}$$

For large k, this implies

$$t_k\,h + r_k^{\mathcal{C}} \in \frac{1}{2}\,(\mathcal{C} - \bar{x})_1 \quad \text{and} \quad \mathcal{G}(\bar{x}) + \mathcal{G}'(\bar{x})\,(t_k\,h + r_k^{\mathcal{C}}) + r_k^{\mathcal{K}} - \mathcal{G}'(\bar{x})\,r_k^{\mathcal{C}} \in \mathcal{K}.$$

For large k, the norms of $u_k := t_k\,h + r_k^{\mathcal{C}}$ and $v_k := r_k^{\mathcal{K}} - \mathcal{G}'(\bar{x})\,r_k^{\mathcal{C}} \in \mathcal{K}$ are small and we can apply Lemma 2.3.7. This yields $\hat{u}_k$ such that

$$\hat{u}_k \in (\mathcal{C} - \bar{x})_1 \qquad \text{and} \qquad \mathcal{G}(\bar{x} + \hat{u}_k) \in \mathcal{K}$$

and

$$\begin{aligned} \|t_k\,h + r_k^{\mathcal{C}} - \hat{u}_k\|_{\mathcal{X}} &= \|u_k - \hat{u}_k\|_{\mathcal{X}} \\ &\le \frac{2}{M}\,\|\mathcal{G}(\bar{x} + t_k\,h + r_k^{\mathcal{C}}) - \mathcal{G}(\bar{x}) - \mathcal{G}'(\bar{x})\,(t_k\,h) - r_k^{\mathcal{K}}\|_{\mathcal{Y}} = o(t_k). \end{aligned}$$

Hence, we have $\bar{x} + \hat{u}_k \in \mathcal{F}$, and

$$\frac{(\bar{x} + \hat{u}_k) - \bar{x}}{t_k} = \frac{\hat{u}_k}{t_k} \to h \qquad \text{as } k \to \infty.$$

This proves the claim $h \in \mathcal{T}_{\mathcal{F}}(\bar{x})$.

An analogous result to Theorem 2.3.6 can be proved for the adjacent/inner tangent cone $\mathcal{T}_{\mathcal{F}}^{\flat}(\bar{x})$. That is, under the conditions of Theorem 2.3.6 we also obtain

$$\mathcal{T}_{\text{lin}}^{\flat}(\bar{x}) := \{h \in \mathcal{T}_{\mathcal{C}}(\bar{x}) : \mathcal{G}'(\bar{x})\,h \in \mathcal{T}_{\mathcal{K}}^{\flat}(\mathcal{G}(\bar{x}))\} = \mathcal{T}_{\mathcal{F}}^{\flat}(\bar{x}).$$

The proof follows the same lines with the obvious modifications.

In the remainder of this section, we show a possibility to obtain a tangent approximation set of the graph of the normal cone mapping in the case that Z is a Hilbert space. That

is, we consider the case $\mathcal{Y} = Z \times Z$ and

$$\begin{aligned}\mathcal{K} := \operatorname{gph} \mathcal{T}_K(\cdot)^\circ &:= \{(z_1, z_2) \in Z^2 : z_2 \in \mathcal{T}_K(z_1)^\circ\} \\ &= \{(z_1, z_2) \in K \times K^\circ : \langle z_1,\, z_2\rangle = 0\},\end{aligned}$$

where $K \subset Z$ is a closed, convex cone. Note that we have identified $Z^\star$ with Z. Due to this identification, we have the well-known and important characterization

$$(z_1, z_2) \in \operatorname{gph} \mathcal{T}_K(\cdot)^\circ \quad \Longleftrightarrow \quad z_1 = \operatorname{Proj}_K(z_1 + z_2) \quad \Longleftrightarrow \quad z_2 = \operatorname{Proj}_{K^\circ}(z_1 + z_2). \tag{2.3.13}$$

In particular, the operators

$$\begin{aligned} P &: Z \to \operatorname{gph} \mathcal{T}_K(\cdot)^\circ, \quad P(z) = (\operatorname{Proj}_K(z), \operatorname{Proj}_{K^\circ}(z)) \\ Q &: \operatorname{gph} \mathcal{T}_K(\cdot)^\circ \to Z, \quad Q(z_1, z_2) = z_1 + z_2 \end{aligned}$$

are inverses of each other.

Lemma 2.3.8. Assume that Z is a Hilbert space and $K \subset Z$ a closed, convex cone. Let $(\bar{z}_1, \bar{z}_2) \in \operatorname{gph} \mathcal{T}_K(\cdot)^\circ$, $T \in \mathcal{L}(Z, Z)$ and a cone $\tilde{W} \subset Z$ be given. Suppose that for all $\varepsilon > 0$ there is $\delta > 0$ with

$$\|\operatorname{Proj}_K(\tilde{z} + w) - \operatorname{Proj}_K(\tilde{z}) - T\,w\|_Z \le \varepsilon \|w\|_Z \qquad \forall w \in B_\delta^Z(0) \cap \tilde{W}, \tilde{z} \in B_\delta^Z(\bar{z}_1 + \bar{z}_2). \tag{2.3.14}$$

Then,

$$W := \{(w_1, w_2) \in Z^2 : w_1 + w_2 \in \tilde{W}, w_1 = T\,(w_1 + w_2)\}$$

is a tangent approximation set of $\operatorname{gph} \mathcal{T}_K(\cdot)^\circ$ at $(\bar{z}_1, \bar{z}_2)$.

Roughly speaking, (2.3.14) asserts that $T\,w$ is the directional derivative of Proj_K for directions $w \in \tilde{W}$ and the remainder term is uniform in a neighborhood of $\bar{z}_1 + \bar{z}_2$.

Proof. Let $\rho \in (0, 1)$ be given. Set $\varepsilon = \rho/2$ and choose $\delta > 0$ such that (2.3.14) is satisfied.

Now, let $(z_1, z_2) \in \operatorname{gph} \mathcal{T}_K(\cdot)^\circ$ with $\|(z_1, z_2) - (\bar{z}_1, \bar{z}_2)\|_{Z^2} \le \delta$ and $(w_1, w_2) \in W$ with $\|(w_1, w_2)\|_{Z^2} \le \delta$ be given.

Then,

$$(\operatorname{Proj}_K(z_1 + w_1 + z_2 + w_2), \operatorname{Proj}_{K^\circ}(z_1 + w_1 + z_2 + w_2)) \in \operatorname{gph} \mathcal{T}_K(\cdot)^\circ$$

To satisfy Definition 2.3.2, it remains to show

$$\begin{aligned}&\|(\operatorname{Proj}_K(z_1 + w_1 + z_2 + w_2) - z_1 - w_1, \operatorname{Proj}_{K^\circ}(z_1 + w_1 + z_2 + w_2) - z_2 - w_2)\|_{Z^2} \\ &\quad \le \rho \|(w_1, w_2)\|_{Z^2}\end{aligned}$$

By (2.3.14) we have

$$\begin{aligned}
&\|\mathrm{Proj}_K(z_1+w_1+z_2+w_2)-z_1-w_1\|_Z\\
&\quad=\|\mathrm{Proj}_K(z_1+z_2+w_1+w_2)-\mathrm{Proj}_K(z_1+z_2)-T\,(w_1+w_2)\|_Z\\
&\quad\le\varepsilon\,\|w_1+w_2\|_Z\le\varepsilon\,(\|w_1\|_Z+\|w_2\|_Z)\le\sqrt{2}\,\varepsilon\,(\|w_1\|_Z^2+\|w_2\|_Z^2)^{1/2}\\
&\quad=\sqrt{2}\,\varepsilon\,\|(w_1,w_2)\|_{Z^2}.
\end{aligned}$$

Since

$$z_1+w_1+z_2+w_2=\mathrm{Proj}_K(z_1+w_1+z_2+w_2)+\mathrm{Proj}_{K^\circ}(z_1+w_1+z_2+w_2),$$

we have

$$\|\mathrm{Proj}_{K^\circ}(z_1+w_1+z_2+w_2)-z_2-w_2\|_Z=\|\mathrm{Proj}_K(z_1+w_1+z_2+w_2)-z_1-w_1\|_Z.$$

Hence,

$$\begin{aligned}
&\big\|\big(\mathrm{Proj}_K(z_1+w_1+z_2+w_2)-z_1-w_1,\mathrm{Proj}_{K^\circ}(z_1+w_1+z_2+w_2)-z_2-w_2\big)\big\|_{Z^2}\\
&\quad=\sqrt{2}\,\|\mathrm{Proj}_K(z_1+w_1+z_2+w_2)-z_1-w_1\|\\
&\quad\le 2\,\varepsilon\,\|(w_1,w_2)\|_{Z^2}=\rho\,\|(w_1,w_2)\|_{Z^2}.
\end{aligned}$$

This shows the claim.

2.3.2. Computation of the tangent cone to the graph of the normal cone map

In this section, we consider a closed convex cone $K\subset Z$, where Z is a (real) Hilbert space. Recall that the graph of the normal cone mapping is given by

$$\mathrm{gph}\,\mathcal{T}_K(\cdot)^\circ:=\{(z_1,z_2)\in Z^2:z_2\in\mathcal{T}_K(z_1)^\circ\}=\{(z_1,z_2)\in K\times K^\circ:\langle z_1,\,z_2\rangle=0\}.$$

We are going to derive a formula for its tangent cone $\mathcal{T}_{\mathrm{gph}\,\mathcal{T}_K(\cdot)^\circ}$. This is also called the *graphical derivative* of the (set-valued) normal-cone mapping.

If the projection onto the cone K is directionally differentiable, we have a well-known characterization of $\mathcal{T}_{\mathrm{gph}\,\mathcal{T}_K(\cdot)^\circ}$.

Lemma 2.3.9. Let us assume that Proj_K is directionally differentiable. Then, for arbitrary $(\bar z_1,\bar z_2)\in\mathrm{gph}\,\mathcal{T}_K(\cdot)^\circ$ we have

$$\mathcal{T}_{\mathrm{gph}\,\mathcal{T}_K(\cdot)^\circ}(\bar z_1,\bar z_2)=\{(z_1,z_2)\in Z^2:\mathrm{Proj}'_K(\bar z_1+\bar z_2;z_1+z_2)=z_1\}.$$

This lemma can be proved similarly as Mordukhovich, Outrata, Ramírez C., 2015b, (3.11) (let g be the identity function therein), or by transferring the proof of Wu, L. Zhang, Y. Zhang, 2014, Theorem 3.1 to the situation at hand.

In order to give an explicit expression of $\mathcal{T}_{\operatorname{gph}\mathcal{T}_K(\cdot)^\circ}$, we use a result by Haraux, 1977.

Lemma 2.3.10. Let us assume that Z is a Hilbert space and let $(\bar{z}_1, \bar{z}_2) \in \operatorname{gph}\mathcal{T}_K(\cdot)^\circ$ be given. We define the critical cone

$$\mathcal{K}_K(\bar{z}_1, \bar{z}_2) = \mathcal{T}_K(\bar{z}_1) \cap \bar{z}_2^\perp$$

and assume that there is a bounded, linear, self-adjoint operator $L : Z \to Z$, such that

$$\begin{aligned} L \circ \operatorname{Proj}_{\mathcal{K}_K(\bar{z}_1,\bar{z}_2)} &= \operatorname{Proj}_{\mathcal{K}_K(\bar{z}_1,\bar{z}_2)} \circ L \\ \operatorname{Proj}_K(\bar{z}_1 + \bar{z}_2 + t\,w) &= \operatorname{Proj}_K(\bar{z}_1 + \bar{z}_2) + t\,L^2\,w + o(t) \quad \forall w \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2). \end{aligned} \tag{2.3.15}$$

Then, for $(z_1, z_2) \in Z^2$ we have

$$(z_1, z_2) \in \mathcal{T}_{\operatorname{gph}\mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2)$$

if and only if there exists $\Pi \in Z$ such that

$$\Pi \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2), \tag{2.3.16a}$$
$$z_1 + z_2 - \Pi \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\circ, \tag{2.3.16b}$$
$$(\Pi,\ z_1 + z_2 - \Pi) = 0, \tag{2.3.16c}$$
$$z_1 - L^2\,\Pi = 0. \tag{2.3.16d}$$

Proof. By Haraux, 1977, Theorem 1, Proj_K is directionally differentiable at $\bar{z}_1 + \bar{z}_2$ and

$$\operatorname{Proj}'_K(\bar{z}_1 + \bar{z}_2; v) = L^2\ \operatorname{Proj}_{\mathcal{K}_K(\bar{z}_1,\bar{z}_2)}(v)$$

for all $v \in Z$.

By Lemma 2.3.9 we obtain

$$\begin{aligned} \mathcal{T}_{\operatorname{gph}\mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2) &= \{(z_1, z_2) \in Z^2 : \operatorname{Proj}'_K(\bar{z}_1, \bar{z}_2; z_1 + z_2) = z_1\} \\ &= \{(z_1, z_2) \in Z^2 : L^2\ \operatorname{Proj}_{\mathcal{K}_K(\bar{z}_1,\bar{z}_2)}(z_1 + z_2) = z_1\}. \end{aligned}$$

Now, we have $\Pi = \operatorname{Proj}_{\mathcal{K}_K(\bar{z}_1,\bar{z}_2)}(z_1 + z_2)$ if and only if

$$\begin{aligned} \Pi &\in \mathcal{K}_K(\bar{z}_1, \bar{z}_2), \\ z_1 + z_2 - \Pi &\in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\circ, \\ (\Pi,\ z_1 + z_2 - \Pi) &= 0. \end{aligned}$$

This implies the assertion.

If K is polyhedric w.r.t. $(\bar{z}_1, \bar{z}_2)$, we can choose $L = I$, since the projection Proj_K satisfies

$$\mathrm{Proj}'_K(\bar{z}_1 + \bar{z}_2, w) = \mathrm{Proj}_{\mathcal{K}_K(\bar{z}_1,\bar{z}_2)}(w) \qquad \forall w \in Z \tag{2.3.17}$$

in this case, see Haraux, 1977, Theorem 2 and Mignot, 1976, Proposition 2.5. The situation in which (2.3.17) is fulfilled was called *projection derivation condition* in Mordukhovich, Outrata, Ramírez C., 2015a, Definition 4.1. This projection derivation condition also holds if K satisfies the extended polyhedricity condition (and another technical condition), see Mordukhovich, Outrata, Ramírez C., 2015a, Proposition 4.2, and in other non-polyhedric situations, see Mordukhovich, Outrata, Ramírez C., 2015a, Example 4.3. Hence, the choice $L = I$ in Lemma 2.3.10 is also possible in these situations.

For later reference, we also provide formulas for the normal cone to $\operatorname{gph} \mathcal{T}_K(\cdot)^\circ$ and for the largest linear subspace which is contained in the tangent cone of $\operatorname{gph} \mathcal{T}_K(\cdot)^\circ$.

Lemma 2.3.11. Under the assumption of Lemma 2.3.10, we have

$$\operatorname{conv}(\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2)) = \begin{pmatrix} L^2 & 0 \\ I - L^2 & I \end{pmatrix} \begin{pmatrix} \mathcal{K}_K(\bar{z}_1, \bar{z}_2) \\ \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\circ \end{pmatrix},$$

where $\operatorname{conv}(\cdot)$ denotes the convex hull. Consequently, the normal cone to $\operatorname{gph} \mathcal{T}_K(\cdot)^\circ$ is given by

$$\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2)^\circ = \begin{pmatrix} L^2 & I - L^2 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\circ \\ \mathcal{K}_K(\bar{z}_1, \bar{z}_2) \end{pmatrix}. \tag{2.3.18}$$

Finally, the largest linear subspace contained in $\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2)$ is given by

$$\begin{pmatrix} L^2 & 0 \\ I - L^2 & I \end{pmatrix} \begin{pmatrix} \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^{\circ\perp} \\ \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\perp \end{pmatrix} \tag{2.3.19}$$

under the assumption that

$$L^2(\Pi^+ + \Pi^-) = 0 \quad \Longrightarrow \quad \Pi^+ + \Pi^- = 0 \qquad \forall \Pi^\pm \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2). \tag{2.3.20}$$

Note that the last result is quite surprising since the tangent cone $\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2)$ is (in general) not convex. Hence, we avoided to call (2.3.19) the *lineality space* of $\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(\bar{z}_1, \bar{z}_2)$, since this is only defined for convex cones.

Proof. We proof the formula for the convex hull, (2.3.18) is then a straightforward consequence, see, e.g., Aubin, Frankowska, 2009, Theorem 2.4.3.

The direction "$\subset$" is clear. To show "$\supset$", we consider $z_1, z_2 \in Z$ such that

$$z_1 = L^2 \Pi_1, \qquad z_2 = (I - L^2) \Pi_1 + \Pi_2$$

for some $\Pi_1 \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)$, $\Pi_2 \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\circ$. Then, it is easy to see that

$$(L^2\,\Pi_1, (I - L^2)\,\Pi_1),\ (0, \Pi_2) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_{K(\cdot)^\circ}}(\bar{z}_1, \bar{z}_2).$$

Since (z_1, z_2) is the sum of these two elements and since $\mathcal{T}_{\operatorname{gph} \mathcal{T}_{K(\cdot)^\circ}}(\bar{z}_1, \bar{z}_2)$ is a cone, the assertion follows.

Finally, we show the last assertion under the assumption (2.3.20). It is clear that (2.3.19) is a linear subspace and that (2.3.19) is contained in $\mathcal{T}_{\operatorname{gph} \mathcal{T}_{K(\cdot)^\circ}}(\bar{z}_1, \bar{z}_2)$. It remains to show the implication

$$\pm\,(z_1, z_2) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_{K(\cdot)^\circ}}(\bar{z}_1, \bar{z}_2) \qquad \Longrightarrow \qquad (z_1, z_2) \in \begin{pmatrix} L^2 & 0 \\ I - L^2 & I \end{pmatrix} \begin{pmatrix} \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^{\circ\perp} \\ \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^{\perp} \end{pmatrix}$$

for all $(z_1, z_2) \in Z^2$. Let $(z_1, z_2) \in Z^2$ with $\pm(z_1, z_2) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_{K(\cdot)^\circ}}(\bar{z}_1, \bar{z}_2)$ be given. By Lemma 2.3.10 we find two elements $\Pi^\pm \in Z$ such that

$$\begin{aligned} \Pi^\pm &\in \mathcal{K}_K(\bar{z}_1, \bar{z}_2), & (\Pi^\pm, \pm(z_1 + z_2) - \Pi^\pm) &= 0, \\ \pm(z_1 + z_2) - \Pi^\pm &\in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^\circ, & \pm z_1 - L^2\,\Pi^\pm &= 0. \end{aligned}$$

cf. (2.3.16). The last equation implies $L^2\,(\Pi^+ + \Pi^-) = 0$, and due to (2.3.20), $\Pi^- = -\Pi^+$ follows. Hence, we have

$$\Pi^+ \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^{\circ\perp}, \qquad z_1 + z_2 - \Pi^+ \in \mathcal{K}_K(\bar{z}_1, \bar{z}_2)^{\perp}$$

and this shows the claim.

2.4. MPCCs in Banach spaces

In this section we apply the results of Section 2.3 to the general MPCC

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{subject to} \quad & g(x) \in C, \\ & G(x) \in K, \\ & H(x) \in K^\circ, \\ & \langle G(x),\, H(x) \rangle = 0. \end{aligned} \tag{MPCC}$$

Here, $f : X \to \mathbb{R}$ is Fréchet differentiable, $g : X \to Y$, $G : X \to Z$ and $H : X \to Z^\star$ are strictly Fréchet differentiable, X, Y, Z are (real) Banach spaces and Z is assumed to be reflexive. Moreover, $C \subset Y$ is a closed, convex set and $K \subset Z$ is a closed, convex cone.

Due to the reflexivity of Z, the problem (MPCC) is symmetric w.r.t. G and H.

The problem (MPCC) was discussed in Chapter 1 and optimality conditions of strongly stationary type were obtained under a CQ. However, these conditions seem to be too

weak in the case that K is not polyhedric, see Section 1.6.2 for an example, but stronger optimality conditions were obtained by an additional linearization argument.

In this section, we apply the same idea to (MPCC). That is, we linearize (MPCC) by means of Theorem 2.3.6, see Section 2.4.1. By means of the results in Section 2.3.2 we recast the linearized problem as a linear MPCC, see Corollary 2.4.3. Finally, we derive optimality conditions via this linearized problem in Section 2.4.2. The main result of this section is Theorem 2.4.5 which asserts that local minimizers of (MPCC) are strongly stationary (in the sense of Definition 2.4.6) if certain CQs are satisfied. Moreover, we demonstrate that this definition of strong stationarity possesses a reasonable strength.

2.4.1. Linearization of the MPCC

In order to apply Theorem 2.3.6, we reformulate the complementarity constraint

$$G(x) \in K,\ H(x) \in K^\circ,\ \langle G(x),\, H(x)\rangle = 0.$$

Indeed, since K is a cone, it is equivalent to

$$H(x) \in \mathcal{T}_K(G(x))^\circ,$$

see (2.2.1), and this can be written as $(G(x), H(x)) \in \operatorname{gph} \mathcal{T}_K(\cdot)^\circ$.

Theorem 2.4.1. Let $\bar{x} \in X$ be a feasible point of (MPCC). Assume that $W \subset Z \times Z^\star$ is a tangent approximation set of the graph of the normal cone mapping of K, i.e., of $\operatorname{gph} \mathcal{T}_K(\cdot)^\circ$, at $(G(\bar{x}), H(\bar{x}))$. If the CQ (2.3.4) with the setting

$$\begin{aligned} &\mathcal{X} := X \times Y, && \mathcal{C} := X \times C, && \mathcal{G}(x,y) := (g(x) - y, G(x), H(x)),\\ &\mathcal{Y} := Y \times (Z \times Z^\star), && \mathcal{K} := \{0\} \times \operatorname{gph} \mathcal{T}_K(\cdot)^\circ, && \mathcal{W} := \{0\} \times W \end{aligned} \tag{2.4.1}$$

is satisfied, then

$$\mathcal{T}_F(\bar{x}) = \mathcal{T}_{\mathrm{lin}}(\bar{x}) := \left\{ h \in X : \begin{array}{r} g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})),\\ (G'(\bar{x})\, h, H'(\bar{x})\, h) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(G(\bar{x}), H(\bar{x})) \end{array} \right\} \tag{2.4.2}$$

holds, where $F \subset X$ is the feasible set of (MPCC). Moreover, if $\bar{x}$ is a local minimizer of (MPCC), then $h = 0$ is a local minimizer of

$$\begin{aligned} \text{Minimize} \quad & f'(\bar{x})\, h,\\ \text{subject to} \quad & g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})),\\ & (G'(\bar{x})\, h, H'(\bar{x})\, h) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(G(\bar{x}), H(\bar{x})). \end{aligned} \tag{2.4.3}$$

An alternative choice to the setting (2.4.1) is

$$\begin{aligned} &\mathcal{X} := X, && \mathcal{C} := X, && \mathcal{G}(x) := (g(x), G(x), H(x)),\\ &\mathcal{Y} := Y \times (Z \times Z^\star), && \mathcal{K} := \mathcal{C} \times \operatorname{gph} \mathcal{T}_K(\cdot)^\circ \end{aligned} \tag{2.4.4}$$

together with a tangent approximation set $\mathcal{W}$ of $\mathcal{K}$. In the case that Y is finite-dimensional, we can choose $\mathcal{W} = \mathcal{T}_C(g(\bar{x})) \times W$, see Lemma 2.3.3, but this may not be possible in the infinite-dimensional case, compare Lemma 2.3.4. Hence, the setting (2.4.4) might require a stronger CQ in this case.

Proof. We set

$$\mathcal{F} := \{(x, y) \in \mathcal{C} : \mathcal{G}(x, y) \in \mathcal{K}\}.$$

An application of Theorem 2.3.6 yields

$$\mathcal{T}_{\mathcal{F}}(\bar{x}, g(\bar{x})) = \{(h_x, h_y) \in \mathcal{T}_{\mathcal{C}}(\bar{x}, g(\bar{x})) : \mathcal{G}'(\bar{x}, g(\bar{x}))\,(h_x, h_y) \in \mathcal{T}_{\mathcal{K}}(\mathcal{G}(\bar{x}, g(\bar{x})))\}. \tag{2.4.5}$$

Now, we need to decode this statement to obtain (2.4.2). To this end, we remark that

$$(x, y) \in \mathcal{F} \quad \Longleftrightarrow \quad y = g(x) \text{ and } x \in F \tag{2.4.6}$$

holds for all $(x, y) \in X \times Y$.

Now, we relate the left-hand sides of (2.4.2) and (2.4.5). Let us show that for arbitrary $h \in X$, the following statements are equivalent.

(i) $h \in \mathcal{T}_F(\bar{x})$.

(ii) $\exists \{t_n\} \subset \mathbb{R}^+, \{h_n\} \subset X : t_n \searrow 0,\ h_n \to h,\ \bar{x} + t_n h_n \in F$.

(iii) $\exists \{t_n\} \subset \mathbb{R}^+, \{h_n\} \subset X : t_n \searrow 0,\ h_n \to h,\ (\bar{x} + t_n h_n, g(\bar{x} + t_n h_n)) \in \mathcal{F}$.

(iv) $(h, g'(\bar{x})\, h) \in \mathcal{T}_{\mathcal{F}}(\bar{x}, g(\bar{x}))$.

Indeed, (i) $\Leftrightarrow$ (ii) is just the definition of $\mathcal{T}_F(\bar{x})$ and (ii) $\Leftrightarrow$ (iii) follows from (2.4.6). The implication (iii) $\Rightarrow$ (iv) is obtained from $(g(\bar{x} + t_n h_n) - g(\bar{x}))/t_n \to g'(\bar{x})\, h$. Finally, (iv) $\Rightarrow$ (i) follows from (2.4.6).

An easy computation shows

$$\mathcal{T}_{\mathcal{C}}(\bar{x}, g(\bar{x})) = X \times \mathcal{T}_C(g(\bar{x})), \qquad \mathcal{T}_{\mathcal{K}}(\mathcal{G}(\bar{x}, g(\bar{x}))) = \{0\} \times \mathcal{T}_{\operatorname{gph} \mathcal{T}_{K(\cdot)^\circ}}(G(\bar{x}), H(\bar{x})).$$

Now, it is straightforward to show that h belongs to the right-hand side of (2.4.2) if and only if $(h, g'(\bar{x})\, h)$ belongs to the right-hand side of (2.4.5).

This shows the equivalency of (2.4.2) and (2.4.5).

To obtain the last assertion, we mention that the local optimality of $\bar{x}$ implies $f'(\bar{x})\, h \geq 0$ for all $h \in \mathcal{T}_F(\bar{x})$. Together with (2.4.2), this yields that $h = 0$ is a local minimizer of (2.4.3).

For convenience, we mention that the CQ (2.3.5) (which is equivalent to (2.3.4) in the case that W is closed and convex) with the setting (2.4.1) is given by

$$\begin{pmatrix} Y \\ Z \\ Z^\star \end{pmatrix} = \begin{pmatrix} g'(\bar{x}) \\ G'(\bar{x}) \\ H'(\bar{x}) \end{pmatrix} X - \begin{pmatrix} \mathcal{R}_C(g(\bar{x})) \\ \\ W \end{pmatrix}. \tag{2.4.7}$$

Note that $W \subset Z \times Z^\star$.

In the case that the linearized constraints of (MPCC) are surjective, we obtain the satisfaction of the CQ (2.4.7) and, consequently, of (2.3.4).

Corollary 2.4.2. Let $\bar{x}$ be a feasible point of (MPCC), and assume that the operator $(g'(\bar{x}), G'(\bar{x}), H'(\bar{x})) \in \mathcal{L}(X, Y \times Z \times Z^\star)$ is surjective. Then, (2.3.4) is satisfied with the setting $W = \{0\} \subset Z \times Z^\star$ and (2.4.1).

The constraints in the linearized problem (2.4.3) contain the tangent cone of the graph of the normal cone mapping. Typically, this tangent cone will again contain complementarity constraints and the linearized problem (2.4.3) will be again an MPCC. Since $h = 0$ is a local minimizer and since every cone is polyhedric in the origin, of Section 1.5.1 will possess a reasonable strength. Hence, we propose to obtain optimality conditions for the local minimizer $\bar{x}$ of (MPCC) by using the optimality conditions of Section 1.5.1 for the minimizer $h = 0$ of (2.4.3). However, it is not possible to characterize this tangent cone $\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}$ in the general case. A typical Hilbert space situation was already discussed in Section 2.3.2, see, in particular, Lemma 2.3.10. In this situation, we have the following result.

Corollary 2.4.3. Let us assume that Z is a Hilbert space and let $\bar{x}$ be a local minimizer of (MPCC). Further, assume that the assertions of Theorem 2.4.1 and of Lemma 2.3.10 (with $(\bar{z}_1, \bar{z}_2) := (G(\bar{x}), H(\bar{x}))$) are satisfied. Then, $(h, \Pi) = 0$ is a local solution of

$$\begin{aligned}
\text{Minimize} \quad & f'(\bar{x})\, h, \\
\text{with respect to} \quad & (h, \Pi) \in X \times Z, \\
\text{subject to} \quad & g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})), \\
& G'(\bar{x})\, h - L^2\, \Pi = 0, \\
& \Pi \in \mathcal{K}_K(G(\bar{x}), H(\bar{x})), \\
& G'(\bar{x})\, h + H'(\bar{x})\, h - \Pi \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^\circ, \\
& (\Pi,\, G'(\bar{x})\, h + H'(\bar{x})\, h - \Pi) = 0.
\end{aligned} \tag{2.4.8}$$

2.4.2. Optimality conditions for the linearized MPCC

In this section we will derive optimality conditions for (MPCC) via optimality conditions for the linearized problem (2.4.3). In the Hilbert space setting of Corollary 2.4.3, (2.4.3) can be written as (2.4.8), which is again an MPCC and all the constraint functions are linear. Moreover, since $(h, \Pi) = 0$ is a solution, and since every cone is polyhedric in the origin, the optimality conditions of Section 1.5.1 possess a reasonable strength.

Hence, we recall the main results from Chapter 1 concerning MPCCs in Banach spaces. We only deal with the linear case and that a local minimizer is given by $h = 0$, since we are going to apply these results to (2.4.8).

We consider the MPCC

$$\begin{aligned} \text{Minimize} \quad & \ell(h), \\ \text{subject to} \quad & \mathcal{A}\,h \in \mathcal{C}, \\ & \mathcal{G}\,h \in \mathcal{K}, \\ & \mathcal{H}\,h \in \mathcal{K}^\circ, \\ & \langle \mathcal{G}\,h,\, \mathcal{H}\,h \rangle = 0. \end{aligned} \tag{2.4.9}$$

Here, $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$, $\mathcal{G} : \mathcal{X} \to \mathcal{Z}$, $\mathcal{H} : \mathcal{X} \to \mathcal{Z}$ are bounded linear maps, $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces, $\mathcal{Z}$ is a Hilbert space, and $\ell \in \mathcal{X}^\star$. The sets $\mathcal{C} \subset \mathcal{Y}$ and $\mathcal{K} \subset \mathcal{Z}$ are assumed to be closed, convex cones. We assume that $h = 0$ is a local minimizer of (2.4.9).

In Section 1.5 the tightened nonlinear program (TNLP) of (2.4.9) (actually, the TNLP is a *linear conic program*, since (2.4.9) was already linear) at $h = 0$ is defined as

$$\begin{aligned} \text{Minimize} \quad & \ell(h), \\ \text{subject to} \quad & \mathcal{A}\,h \in \mathcal{C}, \\ & \mathcal{G}\,h \in \mathcal{K}^{\circ\perp}, \\ & \mathcal{H}\,h \in \mathcal{K}^{\perp}. \end{aligned} \tag{2.4.10}$$

Suppose that the CQs

$$\mathcal{Y} \times \mathcal{Z} \times \mathcal{Z} = (\mathcal{A}, \mathcal{G}, \mathcal{H})\,\mathcal{X} - \mathcal{C} \times \mathcal{K}^{\circ\perp} \times \mathcal{K}^{\perp}, \tag{2.4.11a}$$

$$\mathcal{Y} \times \mathcal{Z} \times \mathcal{Z} = \operatorname{cl}\big[(\mathcal{A}, \mathcal{G}, \mathcal{H})\,\mathcal{X} - \mathcal{C}^{\circ\perp} \times \mathcal{K}^{\circ\perp} \times \mathcal{K}^{\perp}\big] \tag{2.4.11b}$$

are satisfied. Note that the (2.4.11a) is the CQ of Zowe, Kurcyusz, 1979 for (2.4.10) at $h = 0$, whereas (2.4.11b) is a CQ ensuring unique multipliers for (2.4.10), see Section 1.4. This latter CQ (2.4.11b) reduces to the nondegeneracy condition if Y and Z are finite-dimensional, see Remark 1.4.3. Moreover, the nondegeneracy condition (2.4.11b) implies (2.4.11a) in the finite-dimensional case. Then, there exist multipliers $\lambda \in \mathcal{Y}^\star$, $\mu, \nu \in \mathcal{Z}$ associated to the local minimizer $h = 0$, such that the system of strong stationarity

$$\ell + \mathcal{A}^\star\,\lambda + \mathcal{G}^\star\,\mu + \mathcal{H}^\star\,\nu = 0, \tag{2.4.12a}$$

$$\lambda \in \mathcal{C}^\circ, \tag{2.4.12b}$$

$$\mu \in \mathcal{K}^\circ, \tag{2.4.12c}$$

$$\nu \in \mathcal{K} \tag{2.4.12d}$$

is satisfied, see Theorem 1.5.4 and Proposition 1.5.8. Since the cone $\mathcal{K}$ is polyhedric w.r.t. $(0, 0)$, these optimality conditions possess reasonable strength, see Section 1.5.2.

In Theorem 2.4.5, these results will be applied to (2.4.8). We use the setting

$$\begin{aligned} &\mathcal{X} = X \times Z, && \mathcal{A}\,(h, \Pi) = (g'(\bar{x})\,h,\, G'(\bar{x})\,h - L^2\,\Pi), && \mathcal{C} = \mathcal{T}_C(g(\bar{x})) \times \{0\}, \\ &\mathcal{Y} = Y \times Z, && \mathcal{G}\,(h, \Pi) = \Pi, && \mathcal{K} = \mathcal{K}_K(G(\bar{x}), H(\bar{x})), \\ &\mathcal{Z} = Z, && \mathcal{H}\,(h, \Pi) = G'(\bar{x})\,h + H'(\bar{x})\,h - \Pi \end{aligned} \tag{2.4.13}$$

to recast (2.4.8) in the form (2.4.9). The next lemma gives an equivalent characterization of the CQ (2.4.11) applied to (2.4.8) via this setting (2.4.13).

Lemma 2.4.4. By applying the setting (2.4.13), the CQs (2.4.11) are equivalent to

$$\begin{pmatrix} Y \\ Z \\ Z \end{pmatrix} = \begin{pmatrix} g'(\bar{x}) \\ G'(\bar{x}) \\ H'(\bar{x}) \end{pmatrix} X - \begin{pmatrix} I & 0 & 0 \\ 0 & L^2 & 0 \\ 0 & I - L^2 & I \end{pmatrix} \begin{pmatrix} \mathcal{T}_C(g(\bar{x})) \\ \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\circ\perp} \\ \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\perp} \end{pmatrix}, \tag{2.4.14a}$$

$$\begin{pmatrix} Y \\ Z \\ Z \end{pmatrix} = \operatorname{cl}\left\{ \begin{pmatrix} g'(\bar{x}) \\ G'(\bar{x}) \\ H'(\bar{x}) \end{pmatrix} X - \begin{pmatrix} I & 0 & 0 \\ 0 & L^2 & 0 \\ 0 & I - L^2 & I \end{pmatrix} \begin{pmatrix} \mathcal{T}_C(g(\bar{x}))^{\circ\perp} \\ \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\circ\perp} \\ \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\perp} \end{pmatrix} \right\}. \tag{2.4.14b}$$

Here, 0 refer to the zero map between the corresponding spaces.

Proof. "(2.4.14a)⇒(2.4.11a)": We have to show that for all $(y, z_1, z_2, z_3) \in Y \times Z \times Z \times Z$, there are $(\tilde{x}, \tilde{z}_1) \in X \times Z$, $\tilde{y} \in \mathcal{T}_C(g(\bar{x}))$, $\tilde{z}_2 \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\circ\perp}$ and $\tilde{z}_3 \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\perp}$ such that

$$\begin{aligned} y &= g'(\bar{x})\,\tilde{x} - \tilde{y}, & z_2 &= \tilde{z}_1 - \tilde{z}_2, \\ z_1 &= G'(\bar{x})\,\tilde{x} - L^2\,\tilde{z}_1, & z_3 &= (G'(\bar{x}) + H'(\bar{x}))\,\tilde{x} - \tilde{z}_1 - \tilde{z}_3 \end{aligned} \tag{$*$}$$

is satisfied.

Since (2.4.14a) is satisfied, we find $\tilde{x} \in X$, $\tilde{y} \in \mathcal{T}_C(g(\bar{x}))$, $\tilde{z}_2 \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\circ\perp}$ and $\tilde{z}_3 \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^{\perp}$ such that

$$\begin{aligned} y &= g'(\bar{x})\,\tilde{x} - \tilde{y}, & z_1 + L^2\,z_2 &= G'(\bar{x})\,\tilde{x} - L^2\,\tilde{z}_2, \\ & & z_3 - z_1 - (L^2 - I)\,z_2 &= H'(\bar{x})\,\tilde{x} + (I - L^2)\,\tilde{z}_2 - \tilde{z}_3. \end{aligned}$$

Now we set $\tilde{z}_1 := \tilde{z}_2 + z_2$ and it is easy to verify that the above system $(*)$ is satisfied.

The verification of "(2.4.14a)⇒(2.4.11a)" is straightforward and "(2.4.14b)⇔(2.4.11b)" follows from similar arguments.

Note that the lower right block in the right-hand side of the CQs (2.4.14) is just the largest subspace contained in $\mathcal{T}_{\operatorname{gph}\mathcal{T}_{K(\cdot)^\circ}}(\bar{G}, \bar{H})$, see (2.3.19).

The strength of the CQs (2.4.14) is discussed for the applications to SDPMPCCs and SOCMPCCs in the next two section, see, in particular, the remarks after Definitions 2.5.7 and 2.6.9.

Now, we are in position to state the main result of this section.

Theorem 2.4.5. Let the assertions of Theorem 2.4.1 and of Lemma 2.3.10 satisfied and assume that the MPCC (2.4.8) satisfies the CQ (2.4.14). Then, there exist multipliers

$\lambda \in Y^\star$, $\mu, \nu \in Z$, such that the system

$$f'(\bar{x}) + g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu = 0, \tag{2.4.15a}$$
$$\lambda \in \mathcal{T}_C(\bar{x})^\circ, \tag{2.4.15b}$$
$$L^2(\mu - \nu) + \nu \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^\circ, \tag{2.4.15c}$$
$$\nu \in \mathcal{K}_K(G(\bar{x}), H(\bar{x})) \tag{2.4.15d}$$

is satisfied.

Proof. Since the assertions of Theorem 2.4.1 and of Lemma 2.3.10 are satisfied, we can invoke Corollary 2.4.3 and obtain that $(h, \Pi) = 0$ is a local solution of (2.4.8). This linearized MPCC can be transformed to (2.4.10) via the setting (2.4.13) and the corresponding CQ (2.4.11) is satisfied due to Lemma 2.4.4. Hence, the solution $(h, \Pi) = 0$ of (2.4.8) is strongly stationary in the sense of (2.4.12). That is, there exist Lagrange multipliers $\lambda \in Y^\star$, $\rho, \kappa, \nu \in Z$, such that the system

$$f'(\bar{x}) + g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \rho + [G'(\bar{x}) + H'(\bar{x})]^\star \nu = 0, \tag{2.4.16a}$$
$$-L^2 \rho + \kappa - \nu = 0, \tag{2.4.16b}$$
$$\lambda \in \mathcal{T}_C(\bar{x})^\circ, \tag{2.4.16c}$$
$$\kappa \in \mathcal{K}_K(G(\bar{x}), H(\bar{x}))^\circ, \tag{2.4.16d}$$
$$\nu \in \mathcal{K}_K(G(\bar{x}), H(\bar{x})) \tag{2.4.16e}$$

is satisfied. By eliminating $\kappa = \nu + L^2 \rho$ and setting $\mu = \rho + \nu$ we obtain (2.4.15).

We remark that (2.4.16) is obtained from (2.4.15) by defining $\rho := \mu - \nu$ and $\kappa := \nu + L^2 \rho$. That is, the systems (2.4.15) and (2.4.16) are equivalent. Note that conditions (2.4.15c) and (2.4.15d) are equivalent to $(\mu, \nu) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(G(\bar{x}), H(\bar{x}))^\circ$, see Lemma 2.3.11.

Theorem 2.4.5 motivates the following definition of strong stationarity.

Definition 2.4.6. Let $\bar{x}$ be a feasible point of (MPCC) and let $L : Z \to Z$ be a bounded, linear and self-adjoint operator which satisfies (2.3.15). The point $\bar{x}$ is called strongly stationary, if there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in Z$, such that the system (2.4.15) is satisfied.

If K is polyhedric w.r.t. $(G(\bar{x}), H(\bar{x}))$, we can choose $L = I$, since

$$\operatorname{Proj}'_K(G(\bar{x}) + H(\bar{x}), w) = \operatorname{Proj}_{\mathcal{K}_K(G(\bar{x}), H(\bar{x}))}(w) \qquad \forall w \in Z$$

in this case, see Haraux, 1977, Theorem 2. Hence, Definition 2.4.6 is equivalent to Definition 1.5.1 in the polyhedric case.

We remark that it remains an open problem to define strong stationarity in case that no bounded, linear, self-adjoint $L : Z \to Z$ satisfying (2.3.15) exists.

Based on Corollary 2.4.2, it is easy to see that the constraint qualifications of Theorem 2.4.5 are satisfied in the surjective case.

Lemma 2.4.7. Let $\bar{x}$ be a feasible point of (MPCC) and let $L : Z \to Z$ be a bounded, linear and self-adjoint operator which satisfies (2.3.15). Further, assume that the linear operator $(g'(\bar{x}), G'(\bar{x}), H'(\bar{x})) \in \mathcal{L}(X, Y \times Z \times Z)$ is surjective. Then, $\bar{x}$ is strongly stationary in the sense of Definition 2.4.6.

Now, we have seen that the optimality conditions of strongly stationary type (2.4.15) can be obtained under the assumptions of Theorem 2.4.5, and these hold, in particular, if the linearized constraints are surjective, see Lemma 2.4.7. In the remainder of this section, we are going to demonstrate that these strong stationarity conditions possess a reasonable strength and we compare our results to Mordukhovich, Outrata, Ramírez C., 2015b, Section 6.1.

As in the polyhedric case, see Theorem 1.5.4, strong stationarity implies linearized B-stationarity.

Lemma 2.4.8. Let $\bar{x}$ be a feasible point of (MPCC) and let $L : Z \to Z$ be a bounded, linear and self-adjoint operator which satisfies (2.3.15). If $\bar{x}$ is strongly stationary, then

$$f'(x)\, h \geq 0 \qquad \forall h \in \mathcal{T}^2_{\mathrm{lin}}(\bar{x}),$$

where the linearization cone $\mathcal{T}^2_{\mathrm{lin}}(\bar{x})$ is given by

$$\mathcal{T}^2_{\mathrm{lin}}(\bar{x}) := \left\{ \begin{array}{l} h \in X : g'(\bar{x})\, h \in \mathcal{T}_C(\bar{x}), \\ \qquad \exists \Pi \in \mathcal{K}_K(\bar{G}, \bar{H}) : G'(\bar{x})\, h = L^2\, \Pi, \\ \qquad\qquad\qquad\qquad G'(\bar{x})\, h + H'(\bar{x})\, h - \Pi \in \mathcal{K}_K(\bar{G}, \bar{H})^\circ \end{array} \right\}$$

with $(\bar{G}, \bar{H}) = (G(\bar{x}), H(\bar{x}))$. That is, $\bar{x}$ is linearized B-stationary.

Note that the definition of the cone $\mathcal{T}^2_{\mathrm{lin}}(\bar{x})$ does not contain the complementarity condition $(\Pi,\, G'(\bar{x})\, h + H'(\bar{x})\, h - \Pi) = 0$. Hence, it may be larger than $\mathcal{T}_{\mathrm{lin}}(\bar{x})$, see also (2.3.16).

Proof. Since $\bar{x}$ is assumed to be strongly stationary, there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in Z$, such that (2.4.15) is satisfied. Let $h \in \mathcal{T}^2_{\mathrm{lin}}(\bar{x})$ with corresponding $\Pi \in Z$ be given. Then,

$$\begin{aligned} -f'(\bar{x})\, h &= \langle g'(\bar{x})\, h,\, \lambda \rangle + \langle G'(\bar{x})\, h,\, \mu \rangle + \langle H'(\bar{x})\, h,\, \nu \rangle \\ &= \langle g'(\bar{x})\, h,\, \lambda \rangle + \langle G'(\bar{x})\, h,\, \mu - \nu \rangle + \langle G'(\bar{x})\, h + H'(\bar{x})\, h,\, \nu \rangle \\ &= \langle g'(\bar{x})\, h,\, \lambda \rangle + \langle \Pi,\, L^2\, (\mu - \nu) \rangle + \langle G'(\bar{x})\, h + H'(\bar{x})\, h,\, \nu \rangle \\ &= \langle g'(\bar{x})\, h,\, \lambda \rangle + \langle \Pi,\, L^2\, (\mu - \nu) + \nu \rangle + \langle G'(\bar{x})\, h + H'(\bar{x})\, h - \Pi,\, \nu \rangle \\ &\leq 0. \end{aligned}$$

Note that we have

$$\mathcal{T}_F(\bar{x}) \subset \mathcal{T}_{\text{lin}}(\bar{x}) \subset \mathcal{T}^2_{\text{lin}}(\bar{x}),$$

see Lemma 2.3.1 and Lemma 2.3.10. Hence, under the assumptions of Lemma 2.4.8, we have $f'(\bar{x})\,h \geq 0$ for all $h \in \mathcal{T}_F(\bar{x})$ and $h \in \mathcal{T}_{\text{lin}}(\bar{x})$ as well. This shows that the equivalence

$$f'(\bar{x})\,h \geq 0 \quad \forall h \in \mathcal{T}_F(\bar{x}) \qquad \Longleftrightarrow \qquad \bar{x} \text{ is strongly stationary}$$

holds under the assumptions of Theorem 2.4.5 (or of Lemma 2.4.7). Hence, our definition of strong stationarity seems to be of reasonable strength.

Let us compare our results with Mordukhovich, Outrata, Ramírez C., 2015b, Theorem 6.3. Therefore, we consider the problem

$$\begin{array}{ll} \text{Minimize} & f(x,y), \\ \text{such that} & y \in K, \quad H(x,y) \in K^\circ, \quad \text{and} \quad \langle y,\, H(x,y)\rangle = 0. \end{array} \tag{2.4.17}$$

Here, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is Fréchet differentiable and $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is continuously Fréchet differentiable. The closed convex cone K is assumed to satisfy Mordukhovich, Outrata, Ramírez C., 2015b, Assumption (A1) and the assumption of Lemma 2.3.10 such that we can apply the results from Mordukhovich, Outrata, Ramírez C., 2015b as well as our results.

We mention that (2.4.17) is a special case of (MPCC) (in particular, there is no cone constraint $g(x,y) \in C$ and all spaces are finite-dimensional) as well as a special case of Mordukhovich, Outrata, Ramírez C., 2015b, (6.5) (one has to set g to the identity function therein).

In order to obtain optimality conditions for a local minimizer $(\bar{x}, \bar{y})$, the surjectivity of $\nabla_x H(\bar{x}, \bar{y})$ is assumed in Mordukhovich, Outrata, Ramírez C., 2015b, Theorem 6.3. By setting $G(x,y) := y$, we find that $(G'(\bar{x}, \bar{y}), H'(\bar{x}, \bar{y}))$ is surjective as well, and therefore all our CQs are satisfied, see Lemma 2.4.7. By using Lemma 2.3.11 it is easy to check that the resulting optimality conditions of Mordukhovich, Outrata, Ramírez C., 2015b, Theorem 6.3 and Theorem 2.4.5 are equivalent.

However, we might relax the the surjectivity assumption on $\nabla_x H(\bar{x}, \bar{y})$ if we use Theorem 2.4.5 instead of Lemma 2.4.7. In conclusion, both papers deal with slightly different problems and if we apply both theories to a common special case, our CQs might be weaker.

2.5. Optimization with semidefinite complementarity constraints

In this section we apply the theory from Section 2.4 to a problem with semidefinite complementarity constraints. That is, we consider

$$\begin{aligned} \text{Minimize} \quad & f(x), \\ \text{subject to} \quad & g(x) \in C, \\ & G(x) \in \mathbb{S}^n_+, \\ & H(x) \in \mathbb{S}^n_-, \\ & (G(x),\, H(x))_F = 0. \end{aligned} \tag{SDPMPCC}$$

Here, $\mathbb{S}^n_+$ ($\mathbb{S}^n_-$) are the positive (negative) semidefinite symmetric matrices and $(\cdot,\,\cdot)_F$ is the Frobenius inner product on the space $\mathbb{S}^n$ of real symmetric matrices of size $n \times n$, $n \geq 1$. The associated norm on $\mathbb{S}^n$ is denoted by $\|\cdot\|_F$. Moreover, $f : X \to \mathbb{R}$ is Fréchet differentiable, $g : X \to Y$, $G, H : X \to \mathbb{S}^n$ are strictly Fréchet differentiable, X, Y are (real) Banach spaces, and $C \subset Y$ is a closed, convex set.

Note that $\mathbb{S}^n_-$ is the polar cone of $\mathbb{S}^n_+$ w.r.t. the Frobenius inner product, see Ding, D. Sun, Ye, 2014, Section 2.3.

The problem (SDPMPCC) was already considered in Ding, D. Sun, Ye, 2014; Wu, L. Zhang, Y. Zhang, 2014. Both contributions address constraint qualifications and optimality conditions. However, it remained an open problem whether an analogue of MPCC-LICQ implies strong stationarity, see Ding, D. Sun, Ye, 2014, Remark 6.1. In this section, we are going to close this gap.

2.5.1. Notation and preliminaries

We recall some known results concerning the cone of semidefinite matrices, see, e.g., Vandenberghe, Boyd, 1996; Todd, 2001; D. Sun, J. Sun, 2002; Hiriart-Urruty, Malick, 2012; Ding, D. Sun, Ye, 2014 and the references therein.

For a matrix $A \in \mathbb{S}^n$ we call $A = P\,\Lambda\,P^\top$ an ordered eigenvalue decomposition if $P \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, and Λ is a diagonal matrix such that its diagonal entries are ordered decreasingly. Note that this matrix Λ is unique and we denote it by $\Lambda(A)$, whenever appropriate.

For a matrix $A \in \mathbb{S}^n$ and index sets $I, J \subset \{1, \ldots, n\}$, we denote by A_{IJ} or $A_{I,J}$ the submatrix of A containing the rows of A with indices from I and the columns with indices from J. If additionally, P denotes an orthogonal matrix, we define $A^P_{IJ} := A^P_{I,J} := (P^\top A\, P)_{IJ}$.

By $\circ$ we denote the Hadamard product (also called Schur product or entrywise product) of matrices, that is $(A \circ B)_{ij} = A_{ij}\, B_{ij}$.

For two matrices $X \in \mathbb{S}^n_+$, $Y \in \mathbb{S}^n_-$ we have the complementarity $(X, Y)_F = 0$ if and only if there exists a simultaneous ordered eigenvalue decomposition of X and Y and $\Lambda(X)^\top \Lambda(Y) = 0$. That is there exists an orthogonal matrix P with

$$X = P\,\Lambda(X)\,P^\top, \qquad Y = P\,\Lambda(Y)\,P^\top \qquad \text{and} \qquad \Lambda(X)^\top \Lambda(Y) = 0.$$

Let $X \in \mathbb{S}^n_+$, $Y \in \mathbb{S}^n_-$ with $(X, Y)_F = 0$ be given and let $P\,\Lambda\,P^\top = X + Y$ be an ordered eigenvalue decomposition. Note that $P \max(\Lambda, 0)\,P^\top = X$ and $P \min(\Lambda, 0)\,P^\top = Y$ are ordered eigenvalue decompositions of X and Y, see Ding, D. Sun, Ye, 2014, Theorem 2.3. Here, the max and min is understood entrywise. We denote by (α, β, γ) the index sets corresponding to the positive, zero and negative eigenvalues of $X + Y$. Then, we have

$$\mathcal{T}_{\mathbb{S}^n_+}(X) = \{H \in \mathbb{S}^n : H^P_{\beta\cup\gamma,\beta\cup\gamma} \succeq 0\},$$

see, e.g., Shapiro, 1997b, (26) or Hiriart-Urruty, Malick, 2012, (9). Consequently, we have

$$\mathcal{K}_{\mathbb{S}^n_+}(X, Y) = \{H \in \mathbb{S}^n : H^P_{\beta\beta} \succeq 0, H^P_{\beta\gamma} = 0, H^P_{\gamma\gamma} = 0\}, \tag{2.5.1a}$$

$$\mathcal{K}_{\mathbb{S}^n_+}(X, Y)^\circ = \{H \in \mathbb{S}^n : H^P_{\alpha\alpha} = 0, H^P_{\alpha\beta} = 0, H^P_{\alpha\gamma} = 0, H^P_{\beta\beta} \preceq 0\}. \tag{2.5.1b}$$

Moreover, it is well known $\mathrm{Proj}_{\mathbb{S}^n_+}$ is directionally differentiable. For $A \in \mathbb{S}^n$ with ordered eigenvalue decomposition $A = P\,\Lambda\,P^\top$, the directional derivative is given by

$$\mathrm{Proj}'_{\mathbb{S}^n_+}(A; H) = P \begin{bmatrix} H^P_{\alpha\alpha} & H^P_{\alpha\beta} & \Sigma_{\alpha\gamma} \circ H^P_{\alpha\gamma} \\ H^P_{\beta\alpha} & \mathrm{Proj}_{\mathbb{S}^{|\beta|}_+}(H^P_{\beta\beta}) & 0 \\ \Sigma_{\gamma\alpha} \circ H^P_{\gamma\alpha} & 0 & 0 \end{bmatrix} P^\top,$$

see D. Sun, J. Sun, 2002, Theorem 4.7 and Ding, D. Sun, Ye, 2014, (14). Here, $\Sigma \in \mathbb{S}^n$ is defined by

$$\Sigma_{ij} = 1 \qquad (i, j) \in (\alpha \times \alpha) \cup (\alpha \times \beta) \cup (\beta \times \alpha) \cup (\beta \times \beta) \tag{2.5.2a}$$

$$\Sigma_{ij} = 0 \qquad (i, j) \in (\gamma \times \gamma) \cup (\gamma \times \beta) \cup (\beta \times \gamma) \tag{2.5.2b}$$

$$\Sigma_{ij} = \frac{\max(\Lambda_{ii}, 0) - \max(\Lambda_{jj}, 0)}{\Lambda_{ii} - \Lambda_{jj}} \qquad (i, j) \in (\alpha \times \gamma) \cup (\gamma \times \alpha). \tag{2.5.2c}$$

Note that (2.3.15) is satisfied by the bounded, linear, self-adjoint operator $L : \mathbb{S}^n \to \mathbb{S}^n$ defined by

$$L\,H = P\,(\sqrt{\Sigma} \circ (P^\top H\,P))\,P^\top.$$

Here, matrix $\sqrt{\Sigma}$ is defined entrywise by $(\sqrt{\Sigma})_{ij} = \sqrt{\Sigma_{ij}}$. We remark that (2.3.20) is satisfied by this choice of L. For later reference, we remark

$$L^2\,H = P\,(\Sigma \circ (P^\top H\,P))\,P^\top. \tag{2.5.3}$$

2.5.2. Constraint qualifications and strong stationarity

As a first result, we provide a tangent approximation set for $\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$ by means of Lemma 2.3.8.

Lemma 2.5.1. Let $A \in \mathbb{S}^n$ be given and let $A = P\,\Lambda\,P^\top$ be an ordered eigenvalue decomposition. Let (α, β, γ) be the index sets corresponding to the positive, zero and negative eigenvalues of A. We define

$$\tilde{W} = \{H \in \mathbb{S}^n : H^P_{\beta\beta} = 0\}$$

and $\Sigma \in \mathbb{S}^n$ by (2.5.2). Then, for all $\varepsilon > 0$ there is $\delta > 0$ such that

$$\left\|\operatorname{Proj}_{\mathbb{S}^n_+}(\tilde{A} + H) - \operatorname{Proj}_{\mathbb{S}^n_+}(\tilde{A}) - P\,(\Sigma \circ (P^\top H\,P))\,P^\top\right\|_F \leq \varepsilon\,\|H\|_F \tag{2.5.4}$$

for all $\tilde{A} \in \mathbb{S}^n$ and $H \in \tilde{W}$ with

$$\|A - \tilde{A}\|_F \leq \delta \quad \text{and} \quad \|H\|_F \leq \delta.$$

Proof. We first consider the case that A is already diagonal and ordered decreasingly, that is $A = \Lambda$ and $P = I$.

In this proof, M denotes a generic constant which may change from line to line.

Let $\gamma > 0$ be chosen such that

$$\gamma < \frac{\Lambda_{ii}}{2} \qquad \forall i \in \alpha.$$

Following Ding, D. Sun, Ye, 2014, Proof of Proposition 2.6, we define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(t) := \begin{cases} t & \text{if } t > \gamma, \\ 2\,t - \gamma & \text{if } \frac{\gamma}{2} \leq t \leq \gamma, \\ 0 & \text{if } t < \frac{\gamma}{2}. \end{cases}$$

We denote by $F : \mathbb{S}^n \to \mathbb{S}^n$ the corresponding Löwner's operator, that is

$$F(Z) = \sum_{i=1}^{n} f(\lambda_i)\,z_i\,z_i^\top,$$

where λ_i, z_i are the eigenvalues and eigenvectors of the matrix $Z \in \mathbb{S}^n$.

Since f is twice differentiable in a neighborhood of the eigenvalues of A, F is Fréchet differentiable in a neighborhood of A with derivative

$$F'(A)\,H = \Sigma \circ H \qquad \forall H \in \tilde{W},$$

and F' is locally Lipschitz at A, see Bhatia, 1997, Theorem V.3.3, Exercise V.3.9 (ii). In particular, this implies

$$\|F(\tilde{A}+H) - F(\tilde{A}) - F'(\tilde{A})\,H\|_F \leq M\,\|H\|_F^2, \tag{2.5.5a}$$
$$\|F'(\tilde{A})\,H - F'(A)\,H\|_F \leq M\,\|A-\tilde{A}\|_F\,\|H\|_F \tag{2.5.5b}$$

for all $\tilde{A}$ in a neighborhood of A and sufficiently small H.

We are going to prove the desired estimate (2.5.4) by

$$\begin{aligned}&\|\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) - \Sigma\circ H\|_F\\ &\quad\leq \|\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) + F(\tilde{A}) - F(\tilde{A}+H)\|_F\\ &\qquad + \|F(\tilde{A}+H) - F(\tilde{A}) - F'(\tilde{A})\,H\|_F + \|F'(\tilde{A})\,H - \Sigma\circ H\|_F.\end{aligned} \tag{2.5.6}$$

By (2.5.5) we can already bound the second and third term. It remains to study the first term. To this end, we define the operators $\Pi(A)$, $\Pi(\tilde{A})$, $\Pi(\tilde{A}+H)$ to be the sum of the eigenprojections corresponding to the eigenvectors $\Lambda(A)_{\beta\beta}$, $\Lambda(\tilde{A})_{\beta\beta}$, $\Lambda(\tilde{A}+H)_{\beta\beta}$, respectively, see Kato, 1995, (2.1.16). It follows from Kato, 1995, Section 2.4, that

$$\begin{aligned}\|\Pi(\tilde{A}) - \Pi(\tilde{A}+H)\|_F &\leq M\,\|H\|_F,\\ \|\Pi(A) - \Pi(\tilde{A}+H)\|_F &\leq M\,\|A-\tilde{A}-H\|_F\end{aligned}$$

for all $\tilde{A}$ in a neighborhood of A and sufficiently small H. Moreover, since the eigenvalues depends Lipschitz-continuously on the matrix, and $\Lambda(A)_{\beta\beta}=0$, we find

$$\|\Pi(\tilde{A})\,\tilde{A}\|_F = \|\Lambda(\tilde{A})_{\beta\beta}\|_F \leq M\,\|A-\tilde{A}\|_F.$$

By looking at the single eigenspaces, we find

$$\begin{aligned}F(\tilde{A}) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) &= -\mathrm{Proj}_{\mathbb{S}^n_+}(\Pi(\tilde{A})\,\tilde{A}\,\Pi(\tilde{A})),\\ F(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) &= -\mathrm{Proj}_{\mathbb{S}^n_+}(\Pi(\tilde{A}+H)\,(\tilde{A}+H)\,\Pi(\tilde{A}+H)).\end{aligned}$$

This can be used to bound the first term on the right-hand side of (2.5.6)

$$\begin{aligned}&\|\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) + F(\tilde{A}) - F(\tilde{A}+H)\|_F\\ &\quad\leq \|\mathrm{Proj}_{\mathbb{S}^n_+}(\Pi(\tilde{A})\,\tilde{A}\,\Pi(\tilde{A})) - \mathrm{Proj}_{\mathbb{S}^n_+}(\Pi(\tilde{A}+H)\,(\tilde{A}+H)\,\Pi(\tilde{A}+H))\|_F\\ &\quad\leq \|\Pi(\tilde{A})\,\tilde{A}\,\Pi(\tilde{A}) - \Pi(\tilde{A}+H)\,(\tilde{A}+H)\,\Pi(\tilde{A}+H)\|_F.\end{aligned}$$

Here, we used that $\mathrm{Proj}_{\mathbb{S}^n_+}$ is Lipschitz with constant 1.

For $H\in\tilde{W}$ we have $\Pi(A)\,H = H\,\Pi(A) = 0$, and, hence,

$$\begin{aligned}&\Pi(\tilde{A})\,\tilde{A}\,\Pi(\tilde{A}) - \Pi(\tilde{A}+H)\,(\tilde{A}+H)\,\Pi(\tilde{A}+H)\\ &\quad= (\Pi(\tilde{A}) - \Pi(\tilde{A}+H))\,\tilde{A}\,\Pi(\tilde{A}) + \Pi(\tilde{A})\,\tilde{A}\,(\Pi(\tilde{A}) - \Pi(\tilde{A}+H))\\ &\qquad - (\Pi(\tilde{A}) - \Pi(\tilde{A}+H))\,\tilde{A}\,(\Pi(\tilde{A}) - \Pi(\tilde{A}+H))\\ &\qquad + (\Pi(A) - \Pi(\tilde{A}+H))\,H\,(\Pi(A) + \Pi(\tilde{A}+H)).\end{aligned}$$

By the triangle inequality and by the submultiplicativity of the Frobenius norm, we obtain

$$\begin{aligned}
&\left\|\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) + F(\tilde{A}) - F(\tilde{A}+H)\right\|_F \\
&\quad\le \left\|\Pi(\tilde{A})\,\tilde{A}\,\Pi(\tilde{A}) - \Pi(\tilde{A}+H)\,(\tilde{A}+H)\,\Pi(\tilde{A}+H)\right\|_F \\
&\quad\le 2\,\|\Pi(\tilde{A})\,\tilde{A}\|_F\,\|\Pi(\tilde{A}) - \Pi(\tilde{A}+H)\|_F + \|\tilde{A}\|_F\,\|\Pi(\tilde{A}) - \Pi(\tilde{A}+H)\|_F^2 \\
&\qquad + \|\Pi(A) - \Pi(\tilde{A}+H)\|_F\,\|H\|_F\,[\|\Pi(\tilde{A}+H)\|_F + \|\Pi(A)\|_F] \\
&\quad\le M\,(\|A-\tilde{A}\|_F + \|H\|_F)\,\|H\|_F.
\end{aligned}$$

Plugging this into (2.5.6), we obtain

$$\left\|\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) - \Sigma\circ H\right\|_F \le M\,(\|A-\tilde{A}\|_F + \|H\|_F)\,\|H\|_F$$

for sufficiently small $H\in\tilde{W}$ and sufficiently small $A-\tilde{A}$. This shows the claim in the case $A=\Lambda$.

The general case $A = P^\top\Lambda\,P$ will be reduced to the diagonal case. We use that the Frobenius norm is rotationally invariant and find

$$\begin{aligned}
&\left\|\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H) - \mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}) - P\,(\Sigma\circ(P^\top H\,P))\,P^\top\right\|_F \\
&\quad= \left\|P^\top\,\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A}+H)\,P - P^\top\,\mathrm{Proj}_{\mathbb{S}^n_+}(\tilde{A})\,P - \Sigma\circ(P^\top H\,P)\right\|_F \\
&\quad= \left\|\mathrm{Proj}_{\mathbb{S}^n_+}(P^\top\tilde{A}\,P + P^\top H\,P) - \mathrm{Proj}_{\mathbb{S}^n_+}(P^\top\tilde{A}\,P) - \Sigma\circ(P^\top H\,P)\right\|_F.
\end{aligned}$$

Since $(P^\top H\,P)_{\beta\beta} = 0$ we obtain

$$\begin{aligned}
\left\|\mathrm{Proj}_{\mathbb{S}^n_+}(P^\top\tilde{A}\,P + P^\top H\,P) - \mathrm{Proj}_{\mathbb{S}^n_+}(P^\top\tilde{A}\,P) - \Sigma\circ(P^\top H\,P)\right\|_F &\le \varepsilon\,\|P^\top H\,P\|_F \\
&= \varepsilon\,\|H\|_F
\end{aligned}$$

for

$$\|\Lambda - P^\top\tilde{A}\,P\|_F = \|A-\tilde{A}\|_F \le \delta, \qquad \|P^\top H\,P\|_F = \|H\|_F \le \delta.$$

This shows the claim in the general case.

In what follows, we will consider a fixed element $(\bar{G},\bar{H})\in\mathrm{gph}\,\mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$. By

$$\bar{G}+\bar{H} = P\,\Lambda\,P^\top$$

we denote an ordered eigenvalue decomposition, and by (α,β,γ) the index sets corresponding to the positive, zero and negative eigenvalues of $\bar{G}+\bar{H}$. The matrix $\Sigma\in\mathbb{S}^n$ is defined by (2.5.2).

Using Lemma 2.3.8, we can construct a tangent approximation set of $\mathrm{gph}\,\mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$.

Lemma 2.5.2. Let $(\bar{G}, \bar{H}) \in \operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$ be given. Then the set

$$W(\bar{G}, \bar{H}) := \left\{ \begin{array}{ll} (U, V) \in \mathbb{S}^n \times \mathbb{S}^n : & U^P_{\beta\beta} = V^P_{\beta\beta} = 0,\ V^P_{\alpha\alpha} = 0,\ V^P_{\alpha\beta} = 0, \\ & U^P_{\gamma\gamma} = 0,\ U^P_{\gamma\beta} = 0, \\ & U^P_{\alpha\gamma} = \Sigma_{\alpha\gamma} \circ (U^P_{\alpha\gamma} + V^P_{\alpha\gamma}) \end{array} \right\}$$

is a tangent approximation set of $\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$ at $(\bar{G}, \bar{H})$.

Proof. Follows directly from Lemma 2.3.8 and Lemma 2.5.1.

Note that $W(\bar{G}, \bar{H})$ is, in general, smaller than the tangent cone of $\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$, which is given in the following lemma. In fact, $W(\bar{G}, \bar{H})$ is the largest subspace contained in the possibly non-convex cone $\mathcal{T}_{\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ}(\bar{G}, \bar{H})$.

Lemma 2.5.3 (Wu, L. Zhang, Y. Zhang, 2014, Corollary 3.1)**.** Let $(\bar{G}, \bar{H}) \in \operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$ be given. Then,

$$\mathcal{T}_{\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ}(\bar{G}, \bar{H}) = \left\{ \begin{array}{ll} (U, V) \in \mathbb{S}^n \times \mathbb{S}^n : U^P_{\beta\beta} \succeq 0,\ V^P_{\beta\beta} \preceq 0,\ (U^P_{\beta\beta}, V^P_{\beta\beta})_F = 0, \\ \qquad V^P_{\alpha\alpha} = 0,\ V^P_{\alpha\beta} = 0,\ U^P_{\gamma\gamma} = 0,\ U^P_{\gamma\beta} = 0, \\ \qquad U^P_{\alpha\gamma} = \Sigma_{\alpha\gamma} \circ (U^P_{\alpha\gamma} + V^P_{\alpha\gamma}) \end{array} \right\}.$$

We mention that this result also follows from Lemma 2.3.10 by using (2.5.3).

For the calculation of the polar cones of $W(\bar{G}, \bar{H})$ and of $\mathcal{T}_{\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ}(\bar{G}, \bar{H})$, the following lemma is useful.

Lemma 2.5.4. Let $\mathcal{H}$ be a Hilbert space and let $T \in \mathcal{L}(\mathcal{H}, \mathcal{H})$ be given. Then, the subspaces

$$A := \{(G, H) \in \mathcal{H}^2 : G = T(G + H)\} \quad \text{and} \quad B := \{(\mu, \nu) \in \mathcal{H}^2 : \nu = T^\star(\nu - \mu)\}$$

are polars of each other.

Proof. First we show $B \subset A^\circ$. Let $(G, H) \in A$ and $(\mu, \nu) \in B$ be given. Then,

$$\begin{aligned} ((G, H), (\mu, \nu)) &= (G, \mu) + (H + G - G, \nu) = (G, \mu - \nu) + (H + G, \nu) \\ &= (T(G + H), \mu - \nu) + (H + G, \nu) \\ &= (G + H, -\nu) + (H + G, \nu) = 0. \end{aligned}$$

Now, let $(\mu, \nu) \in A^\circ$ be given. For arbitrary $F \in \mathcal{H}$, we have $(T\,F, F - T\,F) \in A$. Hence,

$$0 = ((T\,F, F - T\,F),\, (\mu, \nu)) = (F,\, T^\star \mu) + (F,\, \nu) - (F,\, T^\star \nu) = (F,\, T^\star \mu + \nu - T^\star \nu).$$

Hence, $T^\star \mu + \nu - T^\star \nu = 0$.

Now, we define the SDPMPCC-MFCQ via the CQ (2.4.7).

Definition 2.5.5. Let $\bar{x}$ be a feasible point of (SDPMPCC). We say that the SDPMPCC-MFCQ is satisfied at $\bar{x}$ if

$$Y \times \mathbb{S}^n \times \mathbb{S}^n = (g'(\bar{x}), G'(\bar{x}), H'(\bar{x}))\, X - \mathcal{R}_C(\bar{x}) \times W(G(\bar{x}), H(\bar{x}))$$

holds.

It is easy to see that the corresponding CQ in Wu, L. Zhang, Y. Zhang, 2014, Definition 3.5 is obtained in the special case $X = \mathbb{R}^k$, $Y = \mathbb{R}^{p+q}$ and $C = \{0\}^p \times (\mathbb{R}^-)^q$.

Lemma 2.5.6. Let $\bar{x}$ be a feasible point of (SDPMPCC), such that SDPMPCC-MFCQ is satisfied at $\bar{x}$. Then,

$$\mathcal{T}_F(\bar{x}) = \{h \in X : g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})), (G'(\bar{x})\, h, H'(\bar{x})\, h) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ}(G(\bar{x}), H(\bar{x}))\}. \tag{2.5.7}$$

Moreover, $h = 0$ is a solution of

$$\begin{aligned}
\text{Minimize} \quad & f'(\bar{x})\, h, \\
\text{subject to} \quad & g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})), \\
& (\overline{\mathrm{d}H}\, h)^P_{\alpha\alpha} = 0,\ (\overline{\mathrm{d}H}\, h)^P_{\alpha\beta} = 0, \\
& (\overline{\mathrm{d}G}\, h)^P_{\gamma\gamma} = 0,\ (\overline{\mathrm{d}G}\, h)^P_{\gamma\beta} = 0, \\
& (\overline{\mathrm{d}G}\, h)^P_{\alpha\gamma} = \Sigma_{\alpha\gamma} \circ ((\overline{\mathrm{d}G}\, h)^P_{\alpha\gamma} + (\overline{\mathrm{d}H}\, h)^P_{\alpha\gamma}), \\
& (\overline{\mathrm{d}G}\, h)^P_{\beta\beta} \succeq 0,\ (\overline{\mathrm{d}H}\, h)^P_{\beta\beta} \preceq 0,\ ((\overline{\mathrm{d}G}\, h)^P_{\beta\beta}, (\overline{\mathrm{d}H}\, h)^P_{\beta\beta})_F = 0.
\end{aligned} \tag{2.5.8}$$

Here, we used $\overline{\mathrm{d}G} = G'(\bar{x})$, $\overline{\mathrm{d}H} = H'(\bar{x})$ for brevity.

Proof. Follows from Theorem 2.4.1, see also Corollary 2.4.3.

Note that the satisfaction of (2.5.7) is called SDPMPCC-Abadie-CQ in Wu, L. Zhang, Y. Zhang, 2014, Definition 3.1. We emphasize that Lemma 2.5.6 is a new result for the analysis of (SDPMPCC).

Now, we are going to obtain strong stationarity conditions for (2.5.8). These optimality conditions will then yield an optimality condition for (SDPMPCC). The tightened NLP

of (2.5.8) at $h = 0$ is given by

$$\begin{aligned} \text{Minimize} \quad & f'(\bar{x})\, h, \\ \text{subject to} \quad & g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})), \\ & (G'(\bar{x})\, h, H'(\bar{x})\, h) \in W(\bar{G}, \bar{H}), \end{aligned} \tag{2.5.9}$$

compare (2.4.10). Hence, SDPMPCC-MFCQ implies that the Kurcyusz-Robinson-Zowe-CQ is satisfied for (2.5.9) at $h = 0$, cf. (2.4.11a). It remains to provide an LICQ-variant for (2.5.9), cf. (2.4.11b). Here, we utilize that $W(G(\bar{x}), H(\bar{x}))$ is a subspace.

Definition 2.5.7. Let $\bar{x}$ be a feasible point of (SDPMPCC). We say that the SDPMPCC-LICQ is satisfied at $\bar{x}$ if

$$Y \times \mathbb{S}^n \times \mathbb{S}^n = \operatorname{cl}\big[(g'(\bar{x}), G'(\bar{x}), H'(\bar{x}))\, X - \mathcal{T}_C(\bar{x})^{\circ\perp} \times W(G(\bar{x}), H(\bar{x}))\big]$$

holds.

We remark that SDPMPCC-LICQ implies that there is at most one Lagrange multiplier for (2.5.9) at $h = 0$, see Theorem 1.4.2. In the finite-dimensional case, the SDPMPCC-LICQ is just the nondegeneracy condition for the TNLP (2.5.9). Moreover, it is straightforward to check that SDPMPCC-LICQ implies SDPMPCC-MFCQ in this finite-dimensional case, compare Bonnans, Shapiro, 2000, Corollary 2.98.

Further, SDPMPCC-LICQ implies that there do not exist non-zero multipliers $\lambda \in Y^\star$, $\mu, \nu \in \mathbb{S}^n$ which satisfy the system (see also Lemma 2.5.4)

$$g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu = 0, \tag{2.5.10a}$$

$$\lambda \in \operatorname{lin}(\mathcal{T}_C(g(\bar{x}))^\circ), \tag{2.5.10b}$$

$$\mu^P_{\alpha\alpha} = 0, \quad \mu^P_{\alpha\beta} = 0, \tag{2.5.10c}$$

$$\nu^P_{\gamma\gamma} = 0, \quad \nu^P_{\gamma\beta} = 0, \tag{2.5.10d}$$

$$\Sigma_{\alpha\gamma} \circ (\mu^P_{\alpha\gamma} - \nu^P_{\alpha\gamma}) + \nu^P_{\alpha\gamma} = 0, \tag{2.5.10e}$$

cf. Lemma 1.4.1. Hence, our Definition 2.5.7 is weaker then the LICQ-variants in the literature, which are obtained by replacing $\operatorname{lin}(\mathcal{T}_C(g(\bar{x}))^\circ)$ with $Y^\star$ in (2.5.10b), see Ding, D. Sun, Ye, 2014, (40), Wu, L. Zhang, Y. Zhang, 2014, Definition 3.4. Further, in the important case that Y is finite-dimensional this non-existence of singular multipliers is even equivalent to SDPMPCC-LICQ, see again Theorem 1.4.2.

Theorem 2.5.8. Let $\bar{x}$ be a local solution of (SDPMPCC) which satisfies SDPMPCC-MFCQ and SDPMPCC-LICQ. Then, there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in \mathbb{S}^n$ such that

the strong stationarity system

$$f'(\bar{x}) + g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu = 0, \quad \lambda \in \mathcal{T}_C(g(\bar{x}))^\circ, \tag{2.5.11a}$$
$$\mu^P_{\alpha\alpha} = 0, \quad \mu^P_{\alpha\beta} = 0, \quad \nu^P_{\gamma\gamma} = 0, \quad \nu^P_{\gamma\beta} = 0, \tag{2.5.11b}$$
$$\Sigma_{\alpha\gamma} \circ (\mu^P_{\alpha\gamma} - \nu^P_{\alpha\gamma}) + \nu^P_{\alpha\gamma} = 0, \tag{2.5.11c}$$
$$\mu^P_{\beta\beta} \preceq 0, \quad \nu^P_{\beta\beta} \succeq 0 \tag{2.5.11d}$$

is satisfied.

Proof. By Lemma 2.5.6, $h = 0$ is a local solution to (2.5.8). Since the associated tightened NLP (2.5.9) satisfies the CQ of Kurcyusz-Robinson-Zowe and possesses a unique Lagrange multiplier, the assertion follows, see (2.4.12), Theorem 1.5.6 and Proposition 1.5.7.

We remark that the result also follows from Theorem 2.4.5 by using (2.5.1) and (2.5.3).

We emphasize that Theorem 2.5.8 is a novel result for the analysis of (SDPMPCC), see Ding, D. Sun, Ye, 2014, Remark 6.1. The strong stationarity conditions (2.5.11) are equivalent to Ding, D. Sun, Ye, 2014, Definition 5.1 and Wu, L. Zhang, Y. Zhang, 2014, Definition 3.3. We again stress the fact that, if Y is finite-dimensional, the non-existence of singular multipliers which satisfy (2.5.10) is equivalent to SDPMPCC-LICQ and implies SDPMPCC-MFCQ. Therefore, our definition of SDPMPCC-LICQ is weaker than the corresponding definitions in the literature see Ding, D. Sun, Ye, 2014, (40), Wu, L. Zhang, Y. Zhang, 2014, Definition 3.4. This finite-dimensional situation is summarized in the following theorem.

Theorem 2.5.9. Assume that the constraint space Y is finite-dimensional. Let $\bar{x}$ be a local solution of (SDPMPCC). We assume that there are no non-zero singular multipliers $\lambda \in Y^\star$, $\mu, \nu \in \mathbb{S}^n$ such that (2.5.10) holds. Then, there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in \mathbb{S}^n$ such that the strong stationarity system (2.5.11) is satisfied.

2.6. Optimization with second-order-cone complementarity constraints

In this section we apply the theory from Section 2.4 to a problem with second-order-cone complementarity constraints. That is, we consider the problem

$$\begin{aligned}
\text{Minimize} \quad & f(x), \\
\text{subject to} \quad & g(x) \in C, \\
& G^{(i)}(x) \in \mathcal{K}^{(i)}, && \forall i = 1, \dots, N, \\
& H^{(i)}(x) \in (\mathcal{K}^{(i)})^\circ, && \forall i = 1, \dots, N, \\
& (G^{(i)}(x),\, H^{(i)}(x))_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i = 1, \dots, N.
\end{aligned} \tag{SOCMPCC}$$

Here, $N \geq 1$ and

$$\mathcal{K}^{(i)} := \{(z_1, z_2) \in \mathbb{R} \times \mathcal{H}^{(i)} : z_1 \geq \|z_2\|_{\mathcal{H}^{(i)}}\}$$

is the second-order cone (also called Lorentz cone) in the Hilbert space $\mathbb{R} \times \mathcal{H}^{(i)}$, where $\mathcal{H}^{(i)}$ is an arbitrary Hilbert space, $i = 1, \ldots, N$. Moreover, X, Y are (real) Banach spaces, and $C \subset Y$ is a closed, convex set. The objective function $f : X \to \mathbb{R}$ is Fréchet differentiable, whereas the constraint functions $g : X \to Y$, $G^{(i)}, H^{(i)} : X \to \mathbb{R} \times \mathcal{H}^{(i)}$ are strictly Fréchet differentiable.

Since $(\mathcal{K}^{(i)})^\circ = -\mathcal{K}^{(i)}$, the constraint involving $H^{(i)}$ could also be written as $-H^{(i)}(x) \in \mathcal{K}^{(i)}$.

The problem (SOCMPCC) was already considered in Outrata, D. Sun, 2008; Liang, Zhu, Lin, 2014; Ye, Zhou, 2015. Theorem 5.1 in Ye, Zhou, 2015 shows that local optimizers of (SOCMPCC) are strongly stationary under an LICQ variant. In what follows, we provide the same result under a weaker and more natural version of LICQ, see in particular Theorem 2.6.10.

2.6.1. Preliminaries and notations

We recall some known results concerning the second-order cone $\mathcal{K}$ in $\mathbb{R} \times \mathcal{H}$ for some Hilbert space $\mathcal{H}$. The results are well known in the finite-dimensional case $\mathcal{H} = \mathbb{R}^m$, see, e.g., Alizadeh, Goldfarb, 2003, and the extension to the general case is straightforward, see also Yang, Chang, Chen, 2011. Throughout, a vector $z \in \mathbb{R} \times \mathcal{H}$ is partitioned into $z = (z_1, z_2)$ with $z_1 \in \mathbb{R}$ and $z_2 \in \mathcal{H}$.

By $\operatorname{int} \mathcal{K}$ and $\operatorname{bd} \mathcal{K}$ we denote the interior and boundary of $\mathcal{K}$, respectively. Note that $x \in \operatorname{int} \mathcal{K}$ if and only if $x_1 > \|x_2\|_{\mathcal{H}}$ and $x \in \operatorname{bd} \mathcal{K}$ if and only if $x_1 = \|x_2\|_{\mathcal{H}}$. Two vectors $x \in \mathcal{K}$ and $y \in -\mathcal{K}$ are complementary, i.e., $(x,\, y)_{\mathbb{R}\times\mathcal{H}} = 0$, if and only if one of the (mutually exclusive) following cases occur

$$\begin{aligned}
&\text{iz}: \ x \in \operatorname{int} \mathcal{K},\ y = 0, && \text{bb}: \ x \in \operatorname{bd} \mathcal{K} \setminus \{0\},\ y \in -\operatorname{bd} \mathcal{K} \setminus \{0\},\ x_1\, y_2 = -y_1\, x_2,\\
&\text{zi}: \ x = 0,\ y \in -\operatorname{int} \mathcal{K}, && \text{bz}: \ x \in \operatorname{bd} \mathcal{K} \setminus \{0\},\ y = 0,\\
&\text{zz}: \ x = 0,\ y = 0. && \text{zb}: \ x = 0,\ y \in -\operatorname{bd} \mathcal{K} \setminus \{0\},
\end{aligned} \tag{2.6.1}$$

For the equivalence in case bb, we refer to Ye, Zhou, 2015, Proposition 2.2. The abbreviations of the different cases "z", "i", "b" refer to "zero", "interior" and "boundary" (but not zero), respectively. These six cases will appear frequently in the analysis of this section.

As in (2.3.13), we have the equivalence

$$x \in \mathcal{K},\ y \in -\mathcal{K},\ (x,\, y)_{\mathbb{R}\times\mathcal{H}} = 0 \quad \Leftrightarrow \quad x = \operatorname{Proj}_{\mathcal{K}}(x + y) \quad \Leftrightarrow \quad y = \operatorname{Proj}_{-\mathcal{K}}(x + y). \tag{2.6.2}$$

Using this equivalence, the above six cases can also rephrased in terms of $x + y$:

$$\begin{aligned}
&\text{iz}: \ x + y \in \operatorname{int} \mathcal{K}, && \text{bz}: \ x + y \in \operatorname{bd} \mathcal{K} \setminus \{0\}, && \text{bb}: \ x + y \notin \mathcal{K} \cup (-\mathcal{K}),\\
&\text{zi}: \ x + y \in -\operatorname{int} \mathcal{K}, && \text{zb}: \ x + y \in -\operatorname{bd} \mathcal{K} \setminus \{0\}, && \text{zz}: \ x + y = 0.
\end{aligned}$$

Similar to the eigenvalue decomposition of a matrix, there is a spectral decomposition

$$z = \lambda_1(z)\, c_1(z) + \lambda_2(z)\, c_2(z),$$

with

$$\lambda_i(z) := z_1 + (-1)^i \|z_2\|_{\mathcal{H}}, \qquad \text{and} \qquad c_i(z) := \begin{cases} \frac{1}{2}\,(1, (-1)^i\, z_2/\|z_2\|_{\mathcal{H}}) & \text{if } z_2 \neq 0, \\ \frac{1}{2}\,(1, (-1)^i\, w) & \text{if } z_2 = 0, \end{cases}$$

where $w \in \mathcal{H}$ is a fixed unit vector. We emphasize that the functions λ_i and c_i are smooth on $\{z \in \mathbb{R} \times \mathcal{H} : z_2 \neq 0\}$. Note that $\|c_i(z)\|_{\mathbb{R}\times\mathcal{H}} = 1/\sqrt{2}$.

Using (2.6.2), it is easy to see that the projection $\mathrm{Proj}_{\mathcal{K}}(z)$ of z onto $\mathcal{K}$ can be computed by

$$\mathrm{Proj}_{\mathcal{K}}(z) = \max\{0, \lambda_1(z)\}\, c_1(z) + \max\{0, \lambda_2(z)\}\, c_2(z),$$

see also Fukushima, Luo, Tseng, 2002, Proposition 3.3.

The following lemma is classical in case $\mathcal{H} = \mathbb{R}^m$ and can be found, e.g., in Outrata, D. Sun, 2008, Lemma 2.

Lemma 2.6.1. The projection $\mathrm{Proj}_{\mathcal{K}}$ onto $\mathcal{K}$ is directionally differentiable at all points $z \in \mathbb{R} \times \mathcal{H}$. Moreover, for $h \in \mathbb{R} \times \mathcal{H}$ the following holds.

iz: If $z \in \operatorname{int} \mathcal{K}$ then $\mathrm{Proj}_{\mathcal{K}}$ is differentiable at z and $\mathrm{Proj}'_{\mathcal{K}}(z) = I$.

zi: If $z \in -\operatorname{int} \mathcal{K}$ then $\mathrm{Proj}_{\mathcal{K}}$ is differentiable at z and $\mathrm{Proj}'_{\mathcal{K}}(z) = 0$.

bz: If $z \in \operatorname{bd} \mathcal{K} \setminus \{0\}$ then $\mathrm{Proj}'_{\mathcal{K}}(z; h) = h - 2\, \min\{0, (c_1(z),\, h)_{\mathbb{R}\times\mathcal{H}}\}\, c_1(z)$.

zb: If $z \in -\operatorname{bd} \mathcal{K} \setminus \{0\}$ then $\mathrm{Proj}'_{\mathcal{K}}(z; h) = 2\, \max\{0, (c_2(z),\, h)_{\mathbb{R}\times\mathcal{H}}\}\, c_2(z)$.

bb: If $z \notin \mathcal{K} \cup (-\mathcal{K})$ then $\mathrm{Proj}_{\mathcal{K}}$ is differentiable at z and

$$\mathrm{Proj}'_{\mathcal{K}}(z) = 2\, c_2(z) \otimes c_2(z) + \frac{1}{2}\,\Big(\frac{z_1}{\|z_2\|_{\mathcal{H}}} + 1\Big) \begin{bmatrix} 0 & 0 \\ 0 & I - \frac{z_2 \otimes z_2}{\|z_2\|^2_{\mathcal{H}}} \end{bmatrix}.$$

zz: If $z = 0$, then $\mathrm{Proj}'_{\mathcal{K}}(z; h) = \mathrm{Proj}_{\mathcal{K}}(h)$.

Here, $a \otimes b$ denotes the linear operator $h \mapsto (b,\, h)\, a$.

In the next lemma, we show that (2.3.15) is satisfied by a certain choice of L and, hence, we can apply Lemma 2.3.10.

Lemma 2.6.2. Let $z \in \mathbb{R} \times \mathcal{H}$ be given. We have the following formulas for the critical cone $\mathcal{K}_{\mathcal{K}}(x, y) = \mathcal{T}_{\mathcal{K}}(x) \cap y^{\perp}$, where $x = \mathrm{Proj}_{\mathcal{K}}(z)$, $y = \mathrm{Proj}_{\mathcal{K}^\circ}(z)$.

iz: If $z \in \operatorname{int} \mathcal{K}$ then $\mathcal{K}_{\mathcal{K}}(x, y) = \mathbb{R} \times \mathcal{H}$.

zi: If $z \in -\operatorname{int} \mathcal{K}$ then $\mathcal{K}_{\mathcal{K}}(x, y) = \{0\}$.

bz: If $z \in \operatorname{bd} \mathcal{K} \setminus \{0\}$ then $\mathcal{K}_{\mathcal{K}}(x, y) = \{d \in \mathbb{R} \times \mathcal{H} : (c_1(z),\, d)_{\mathbb{R}\times\mathcal{H}} \geq 0\}$.

zb: If $z \in -\operatorname{bd}\mathcal{K} \setminus \{0\}$ then $\mathcal{K}_{\mathcal{K}}(x,y) = \mathbb{R}^+ c_2(z)$.

bb: If $z \notin \mathcal{K} \cup (-\mathcal{K})$ then $\mathcal{K}_{\mathcal{K}}(x,y) = \{d \in \mathbb{R}\times\mathcal{H} : (c_1(z),\, d)_{\mathbb{R}\times\mathcal{H}} = 0\}$.

zz: If $z = 0$, then $\mathcal{K}_{\mathcal{K}}(x,y) = \mathcal{K}$.

Finally, (2.3.15) and (2.3.20) are satisfied by the choice

$$L = 2\, c_2(z) \otimes c_2(z) + \sqrt{\frac{1}{2}\left(\frac{z_1}{\|z_2\|_{\mathcal{H}}} + 1\right)} \begin{bmatrix} 0 & 0 \\ 0 & I - \frac{z_2 \otimes z_2}{\|z_2\|_{\mathcal{H}}^2} \end{bmatrix}$$

in case $z \notin \mathcal{K} \cup (-\mathcal{K})$, i.e. (bb), and $L = I$ else.

Proof. The validation of the formulas for the critical cone are easily obtained by using

$$\mathcal{T}_{\mathcal{K}}(x) = \{d \in \mathbb{R}\times\mathcal{H} : (c_1(z),\, d)_{\mathbb{R}\times\mathcal{H}} \geq 0\}$$

in case $x \in \operatorname{bd}\mathcal{K} \setminus \{0\}$.

The validation of (2.3.15) is straightforward up to the case $z \notin \mathcal{K} \cup (-\mathcal{K})$. In this case, we can use that the critical cone is the orthogonal complement of $c_1(z)$, hence,

$$\operatorname{Proj}_{\mathcal{K}_{\mathcal{K}}(x,y)} = I - 2\, c_1(z) \otimes c_1(z)$$

and $L\, c_1(z) = 0$. Similarly, (2.3.20) follows.

Finally, we define

$$\hat{\mathcal{H}} := \prod_{i=1}^{N} (\mathbb{R}\times\mathcal{H}^{(i)}) \qquad \text{and} \qquad \hat{\mathcal{K}} := \prod_{i=1}^{N} \mathcal{K}^{(i)}.$$

2.6.2. Constraint qualifications and strong stationarity

As a first result, we provide a tangent approximation set for $\operatorname{gph}\mathcal{T}_{\mathcal{K}}(\cdot)^\circ$ by means of Lemma 2.3.8. As in Section 2.6.1, $\mathcal{K}$ is the second-order cone in the Hilbert space $\mathbb{R}\times\mathcal{H}$.

Lemma 2.6.3. Let $z \in \mathbb{R}\times\mathcal{H}$ be given. We define $\tilde{W} \subset \mathbb{R}\times\mathcal{H}$ and $T \in \mathcal{L}(\mathbb{R}\times\mathcal{H}, \mathbb{R}\times\mathcal{H})$ by the following distinction of cases.

iz: If $z \in \operatorname{int}\mathcal{K}$ then $\tilde{W} = \mathbb{R}\times\mathcal{H}$, $T = I$.

zi: If $z \in -\operatorname{int}\mathcal{K}$ then $\tilde{W} = \mathbb{R}\times\mathcal{H}$, $T = 0$.

bz: If $z \in \operatorname{bd}\mathcal{K} \setminus \{0\}$ then $\tilde{W} = c_1(z)^\perp$, $T = I$.

zb: If $z \in -\operatorname{bd}\mathcal{K} \setminus \{0\}$ then $\tilde{W} = c_2(z)^\perp$, $T = 0$.

bb: If $z \notin \mathcal{K} \cup (-\mathcal{K})$ then $\tilde{W} = \mathbb{R}\times\mathcal{H}$ and $T = \operatorname{Proj}'_{\mathcal{K}}(z)$.

zz: If $z = 0$ then $\tilde{W} = \{0\}$, $T = 0$.

Then, for all $\varepsilon > 0$ there is $\delta > 0$ such that

$$\left\|\mathrm{Proj}_{\mathcal{K}}(\tilde{z} + h) - \mathrm{Proj}_{\mathcal{K}}(\tilde{z}) - T\,h\right\|_{\mathbb{R}\times\mathcal{H}} \leq \varepsilon\,\|h\|_{\mathbb{R}\times\mathcal{H}} \tag{2.6.3}$$

for all $\tilde{z} \in \mathbb{R} \times \mathcal{H}$ and $h \in \tilde{W}$ with

$$\|\tilde{z} - z\|_{\mathbb{R}\times\mathcal{H}} \leq \delta \quad \text{and} \quad \|h\|_{\mathbb{R}\times\mathcal{H}} \leq \delta.$$

Proof. In the cases iz, zi, and bb, the function projection $\mathrm{Proj}_{\mathcal{K}}$ is continuously differentiable at z and $T = \mathrm{Proj}'_{\mathcal{K}}(z)$, see Lemma 2.6.1. This implies (2.6.3).

Further, in case zz the estimate (2.6.3) holds trivially, since $h \in \tilde{W} = \{0\}$.

In the remainder, we discuss the case bz, the case zb can be treated analogously.

Let $z \in \mathrm{bd}\,\mathcal{K} \setminus \{0\}$ be given. We have $z_1 = \|z_2\|_{\mathcal{H}} > 0$, hence $z_2 \neq 0$, as well as $\lambda_1(z) = 0$ and $\lambda_2(z) = 2\,z_1$. This implies that the functions λ_i and c_i are smooth in a neighborhood of z. In particular, $\lambda_2(\tilde{z}) > 0$ for all $\tilde{z}$ in a neighborhood of z. Let $\varepsilon > 0$ be given. For $\tilde{z} \in \mathbb{R} \times \mathcal{H}$, $h \in \tilde{W}$ with $\|\tilde{z} - z\|_{\mathbb{R}\times\mathcal{H}}$ and $\|h\|_{\mathbb{R}\times\mathcal{H}}$ sufficiently small we have

$$\begin{aligned}
&\mathrm{Proj}_{\mathcal{K}}(\tilde{z} + h) - \mathrm{Proj}_{\mathcal{K}}(\tilde{z}) - T\,h \\
&\quad = \max\{0, \lambda_1(\tilde{z} + h)\}\,c_1(\tilde{z} + h) + \lambda_2(\tilde{z} + h)\,c_2(\tilde{z} + h) \\
&\qquad - \max\{0, \lambda_1(\tilde{z})\}\,c_1(\tilde{z}) - \lambda_2(\tilde{z})\,c_2(\tilde{z}) - h.
\end{aligned} \tag{2.6.4}$$

Since λ_2 and c_2 are smooth in a neighborhood of z, the function $s_2 := \lambda_2\,c_2$ is continuously differentiable in z. Hence,

$$\|s_2(\tilde{z} + h) - s_2(\tilde{z}) - s'_2(z)\,h\|_{\mathbb{R}\times\mathcal{H}} \leq \frac{\varepsilon}{2}\,\|h\|_{\mathbb{R}\times\mathcal{H}} \tag{2.6.5}$$

for $\|\tilde{z} - z\|_{\mathbb{R}\times\mathcal{H}}$, $\|h\|_{\mathbb{R}\times\mathcal{H}}$ small enough. A straightforward calculation shows

$$s'_2(z)\,h = \frac{1}{2}\Big(h_1 + \frac{(z_2,\,h_2)_{\mathcal{H}}}{\|z_2\|_{\mathcal{H}}},\ \frac{h_1}{\|z_2\|_{\mathcal{H}}}\,z_2 + \frac{z_1}{\|z_2\|_{\mathcal{H}}}\,h_2 - \frac{z_1\,(h_2,\,z_2)_{\mathcal{H}}}{\|z_2\|^3_{\mathcal{H}}}\,z_2 + h_2\Big).$$

Since $h \in \tilde{W}$, we have $(h,\,c_1(z))_{\mathbb{R}\times\mathcal{H}} = 0$ which implies $\|z_2\|_{\mathcal{H}}\,h_1 = (z_2,\,h_2)_{\mathcal{H}}$. Together with $z_1 = \|z_2\|_{\mathcal{H}}$, we find

$$s'_2(z)\,h = \frac{1}{2}\Big(h_1 + h_1,\ \frac{h_1}{\|z_2\|_{\mathcal{H}}}\,z_2 + h_2 - \frac{h_1}{\|z_2\|_{\mathcal{H}}}\,z_2 + h_2\Big) = h.$$

Using this representation in (2.6.5), we obtain

$$\|\lambda_2(\tilde{z} + h)\,c_2(\tilde{z} + h) - \lambda_2(\tilde{z})\,s_2(\tilde{z}) - h\|_{\mathbb{R}\times\mathcal{H}} \leq \frac{\varepsilon}{2}\,\|h\|_{\mathbb{R}\times\mathcal{H}} \tag{2.6.6}$$

for $h \in \tilde{W}$, and $\|\tilde{z} - z\|_{\mathbb{R}\times\mathcal{H}}$, $\|h\|_{\mathbb{R}\times\mathcal{H}}$ small enough.

It remains to study the terms involving λ_1 in (2.6.4). Since $\|c_1(\tilde{z})\|_{\mathbb{R}\times\mathcal{H}} = 1/\sqrt{2} \leq 1$, we have

$$\begin{aligned}
&\big\|\max\{0,\lambda_1(\tilde{z}+h)\}\,c_1(\tilde{z}+h) - \max\{0,\lambda_1(\tilde{z})\}\,c_1(\tilde{z})\big\|_{\mathbb{R}\times\mathcal{H}}\\
&\quad\leq \big\|\max\{0,\lambda_1(\tilde{z}+h)\}\,(c_1(\tilde{z}+h)-c_1(\tilde{z}))\big\|_{\mathbb{R}\times\mathcal{H}}\\
&\qquad+\big\|(\max\{0,\lambda_1(\tilde{z}+h)\}-\max\{0,\lambda_1(\tilde{z})\})\,c_1(\tilde{z})\big\|_{\mathbb{R}\times\mathcal{H}}\\
&\quad\leq |\lambda_1(\tilde{z}+h)|\,\|c_1(\tilde{z}+h)-c_1(\tilde{z})\|_{\mathbb{R}\times\mathcal{H}} + \big|\max\{0,\lambda_1(\tilde{z}+h)\}-\max\{0,\lambda_1(\tilde{z})\}\big|\\
&\quad\leq |\lambda_1(\tilde{z}+h)|\,\|c_1(\tilde{z}+h)-c_1(\tilde{z})\|_{\mathbb{R}\times\mathcal{H}} + |\lambda_1(\tilde{z}+h)-\lambda_1(\tilde{z})|.
\end{aligned}$$

By using the smoothness of λ_1 and c_1 in a neighborhood of z, we find

$$\begin{aligned}
|\lambda_1(\tilde{z}+h)| &\leq |\lambda_1(z)| + C\,\|\tilde{z}+h-z\|_{\mathbb{R}\times\mathcal{H}}\\
&\leq 0 + C\,(\|\tilde{z}-z\|_{\mathbb{R}\times\mathcal{H}} + \|h\|_{\mathbb{R}\times\mathcal{H}}),\\
\|c_1(\tilde{z}+h)-c_1(\tilde{z})\|_{\mathbb{R}\times\mathcal{H}} &\leq C\,\|h\|_{\mathbb{R}\times\mathcal{H}},\\
|\lambda_1(\tilde{z}+h)-\lambda_1(\tilde{z})| &\leq |\lambda_1'(z)\,h| + \frac{\varepsilon}{4}\,\|h\|_{\mathbb{R}\times\mathcal{H}}\\
&= \left|h_1 - \frac{(z_2,\,h_2)_{\mathcal{H}}}{\|z_2\|_{\mathcal{H}}}\right| + \frac{\varepsilon}{4}\,\|h\|_{\mathbb{R}\times\mathcal{H}} = \frac{\varepsilon}{4}\,\|h\|_{\mathbb{R}\times\mathcal{H}}
\end{aligned}$$

for some constant $C>0$, $h\in\tilde{W}$, and $\|\tilde{z}-z\|_{\mathbb{R}\times\mathcal{H}}$, $\|h\|_{\mathbb{R}\times\mathcal{H}}$ small enough. Hence,

$$\begin{aligned}
&\big\|\max\{0,\lambda_1(\tilde{z}+h)\}\,c_1(\tilde{z}+h) - \max\{0,\lambda_1(\tilde{z})\}\,c_1(\tilde{z})\big\|_{\mathbb{R}\times\mathcal{H}}\\
&\qquad\leq C^2\left(\|\tilde{z}-z\|_{\mathbb{R}\times\mathcal{H}}\,\|h\|_{\mathbb{R}\times\mathcal{H}} + \|h\|^2_{\mathbb{R}\times\mathcal{H}}\right) + \frac{\varepsilon}{4}\,\|h\|_{\mathbb{R}\times\mathcal{H}}.
\end{aligned}$$

Together with (2.6.4) and (2.6.6), we find

$$\big\|\mathrm{Proj}_{\mathcal{K}}(\tilde{z}+h) - \mathrm{Proj}_{\mathcal{K}}(\tilde{z}) - T\,h\big\|_{\mathbb{R}\times\mathcal{H}} \leq \varepsilon\,\|h\|_{\mathbb{R}\times\mathcal{H}}$$

for $h\in\tilde{W}$ and $\|\tilde{z}-z\|_{\mathbb{R}\times\mathcal{H}}$, $\|h\|_{\mathbb{R}\times\mathcal{H}}$ small enough.

Now, we consider a feasible point $\bar{x}\in X$ of (SOCMPCC). For convenience, we set

$$\bar{G} = (\bar{G}^{(1)},\dots,\bar{G}^{(N)}), \qquad \bar{H} = (\bar{H}^{(1)},\dots,\bar{H}^{(N)}),$$

where

$$\bar{G}^{(i)} = G^{(i)}(\bar{x}), \quad \bar{H}^{(i)} = H^{(i)}(\bar{x}), \qquad \forall i = 1,\dots,N.$$

We define the (disjoint) index sets

$$\begin{aligned}
I^{\mathrm{iz}}(\bar{x}) &:= \{i\in\{1,\dots,N\} : \bar{G}^{(i)}\in \operatorname{int}\mathcal{K}^{(i)}\},\\
I^{\mathrm{zi}}(\bar{x}) &:= \{i\in\{1,\dots,N\} : \bar{H}^{(i)}\in -\operatorname{int}\mathcal{K}^{(i)}\},\\
I^{\mathrm{bz}}(\bar{x}) &:= \{i\in\{1,\dots,N\} : \bar{G}^{(i)}\in \operatorname{bd}\mathcal{K}^{(i)}\setminus\{0\}, \bar{H}^{(i)}=0\},\\
I^{\mathrm{zb}}(\bar{x}) &:= \{i\in\{1,\dots,N\} : \bar{G}^{(i)}=0, \bar{H}^{(i)}\in -\operatorname{bd}\mathcal{K}^{(i)}\setminus\{0\}\},
\end{aligned}$$

$$I^{\mathrm{bb}}(\bar{x}) := \{i \in \{1,\dots,N\} : \bar{G}^{(i)} \in \operatorname{bd}\mathcal{K}^{(i)} \setminus \{0\}, \bar{H}^{(i)} \in -\operatorname{bd}\mathcal{K}^{(i)} \setminus \{0\}\},$$
$$I^{\mathrm{zz}}(\bar{x}) := \{i \in \{1,\dots,N\} : \bar{G}^{(i)} = 0, \bar{H}^{(i)} = 0\}.$$

For $i \in I^{\mathrm{bb}}(\bar{x})$ we have (see also (2.6.1))

$$\frac{\bar{G}_2^{(i)} + \bar{H}_2^{(i)}}{\bar{G}_1^{(i)} - \bar{H}_1^{(i)}} = \frac{\bar{G}_2^{(i)} + \bar{H}_2^{(i)}}{\|\bar{G}_2^{(i)} + \bar{H}_2^{(i)}\|_{\mathcal{H}^{(i)}}} = \frac{\bar{G}_2^{(i)}}{\|\bar{G}_2^{(i)}\|_{\mathcal{H}^{(i)}}} = \frac{\bar{G}_2^{(i)}}{\bar{G}_1^{(i)}} = \frac{\bar{H}_2^{(i)}}{\|\bar{H}_2^{(i)}\|_{\mathcal{H}^{(i)}}} = -\frac{\bar{H}_2^{(i)}}{\bar{H}_1^{(i)}}$$

and, hence,

$$c_1(\bar{G}^{(i)} + \bar{H}^{(i)}) = \frac{1}{2}\Big(1, -\frac{\bar{G}_2^{(i)}}{\bar{G}_1^{(i)}}\Big) = \frac{1}{2}\Big(1, \frac{\bar{H}_2^{(i)}}{\bar{H}_1^{(i)}}\Big), \quad c_2(\bar{G}^{(i)} + \bar{H}^{(i)}) = \frac{1}{2}\Big(1, \frac{\bar{G}_2^{(i)}}{\bar{G}_1^{(i)}}\Big).$$

This implies the following representation of the derivative of $\mathrm{Proj}_{\mathcal{K}^{(i)}}$ for $i \in I^{\mathrm{bb}}(\bar{x})$

$$E^{(i)} := \mathrm{Proj}'_{\mathcal{K}^{(i)}}(\bar{G}^{(i)} + \bar{H}^{(i)}) = \frac{1}{2}(1, w^{(i)}) \otimes (1, w^{(i)}) + \kappa^{(i)} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes (w^{(i)}) \end{bmatrix}, \quad (2.6.7)$$

where

$$\kappa^{(i)} := \frac{\bar{G}_1^{(i)}}{\bar{G}_1^{(i)} - \bar{H}_1^{(i)}} \quad \text{and} \quad w^{(i)} := \frac{\bar{G}_2^{(i)}}{\bar{G}_1^{(i)}}. \quad (2.6.8)$$

Similarly, we define

$$L^{(i)} = \begin{cases} \frac{1}{2}(1, w^{(i)}) \otimes (1, w^{(i)}) + \sqrt{\kappa^{(i)}} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes (w^{(i)}) \end{bmatrix} & \text{if } i \in I^{\mathrm{bb}}(\bar{x}), \\ I & \text{else.} \end{cases} \quad (2.6.9)$$

Together with Lemma 2.6.2, the assertion of Lemma 2.3.10 is satisfied by defining $L : \hat{\mathcal{H}} \to \hat{\mathcal{H}}$ componentwise via $L^{(i)}$. Note that the critical cone to $\hat{\mathcal{K}}$ can be obtained by the product of the critical cones of $\mathcal{K}^{(i)}$ and these can be computed via Lemma 2.6.2.

Then, we obtain a straightforward consequence of Lemma 2.6.3.

Corollary 2.6.4. Let $\bar{x}$ be a feasible point of (SOCMPCC). We define $\tilde{W} \subset \hat{\mathcal{H}}$ by

$$\tilde{W} := \left\{ \begin{array}{r} (h^{(1)}, \dots, h^{(N)}) \in \hat{\mathcal{H}} : h^{(i)} = 0,\ \forall i \in I^{\mathrm{zz}}(\bar{x}) \\ (c_1(\bar{G}^{(i)}),\ h^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0,\ \forall i \in I^{\mathrm{bz}}(\bar{x}); \\ (c_2(\bar{H}^{(i)}),\ h^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0,\ \forall i \in I^{\mathrm{zb}}(\bar{x}) \end{array} \right\}$$

and $T : \hat{\mathcal{H}} \to \hat{\mathcal{H}}$ by

$$\begin{aligned} (T\,h)^{(i)} &= 0, && \text{for } i \in I^{\mathrm{zi}}(\bar{x}) \cup I^{\mathrm{zb}}(\bar{x}) \cup I^{\mathrm{zz}}(\bar{x}), \\ (T\,h)^{(i)} &= h^{(i)}, && \text{for } i \in I^{\mathrm{iz}}(\bar{x}) \cup I^{\mathrm{bz}}(\bar{x}), \\ (T\,h)^{(i)} &= E^{(i)}\,h^{(i)}, && \text{for } i \in I^{\mathrm{bb}}(\bar{x}). \end{aligned}$$

Then, for all $\varepsilon > 0$, there is $\delta > 0$ such that

$$\|\mathrm{Proj}_{\hat{\mathcal{K}}}(\tilde{z} + h) - \mathrm{Proj}_{\hat{\mathcal{K}}}(\tilde{z}) - T\,h\|_{\hat{\mathcal{H}}} \leq \varepsilon\,\|h\|_{\hat{\mathcal{H}}}$$

for all $\tilde{z} \in \hat{\mathcal{H}}$, $h \in \tilde{W}$ and $\|\bar{G} + \bar{H} - \tilde{z}\|_{\hat{\mathcal{H}}}, \|h\|_{\hat{\mathcal{H}}} \leq \delta$.

Together with Lemma 2.3.8 we obtain a tangent approximation set of $\mathrm{gph}\,\mathcal{T}_{\hat{\mathcal{K}}}(\cdot)^\circ$.

Corollary 2.6.5. Let $\bar{x}$ be a feasible point of (SOCMPCC). We define $W(\bar{G}, \bar{H}) \subset \hat{\mathcal{H}} \times \hat{\mathcal{H}}$ by

$$(u, v) \in W(\bar{G}, \bar{H}) :\Longleftrightarrow \begin{cases} v^{(i)} = 0, & \forall i \in I^{\mathrm{iz}}(\bar{x}), \\ u^{(i)} = 0, & \forall i \in I^{\mathrm{zi}}(\bar{x}), \\ v^{(i)} = 0,\ (c_1(\bar{G}^{(i)}),\, u^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, & \forall i \in I^{\mathrm{bz}}(\bar{x}), \\ u^{(i)} = 0,\ (c_2(\bar{H}^{(i)}),\, v^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, & \forall i \in I^{\mathrm{zb}}(\bar{x}), \\ u^{(i)} = E^{(i)}\,(u^{(i)} + v^{(i)}), & \forall i \in I^{\mathrm{bb}}(\bar{x}), \\ u^{(i)} = v^{(i)} = 0, & \forall i \in I^{\mathrm{zz}}(\bar{x}). \end{cases}$$

Then, $W(\bar{G}, \bar{H})$ is a tangent approximation set of $\mathrm{gph}\,\mathcal{T}_{\hat{\mathcal{K}}}(\cdot)^\circ$ in the sense of Definition 2.3.2.

An application of Lemma 2.3.9 yields a characterization of the tangent cone of $\mathrm{gph}\,\mathcal{T}_{\hat{\mathcal{K}}}(\cdot)^\circ$.

Lemma 2.6.6. Let a feasible point $\bar{x}$ of (SOCMPCC) be given. Then, we have $(u, v) \in \mathcal{T}_{\mathrm{gph}\,\mathcal{T}_{\hat{\mathcal{K}}}(\cdot)^\circ}(\bar{G}, \bar{H})$ if and only if

$$\begin{aligned}
&v^{(i)} = 0, && \forall i \in I^{\mathrm{iz}}(\bar{x}), \\
&u^{(i)} = 0, && \forall i \in I^{\mathrm{zi}}(\bar{x}), \\
&u^{(i)} \in (R^{(i)})^\circ, \quad v^{(i)} \in R^{(i)}, \quad \text{and} \quad (u^{(i)}, v^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i \in I^{\mathrm{bz}}(\bar{x}), \\
&u^{(i)} \in R^{(i)}, \quad v^{(i)} \in (R^{(i)})^\circ, \quad \text{and} \quad (u^{(i)}, v^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i \in I^{\mathrm{zb}}(\bar{x}), \\
&u^{(i)} = E^{(i)}\,(u^{(i)} + v^{(i)}), && \forall i \in I^{\mathrm{bb}}(\bar{x}), \\
&u^{(i)} \in \mathcal{K}^{(i)}, \quad v^{(i)} \in -\mathcal{K}^{(i)}, \quad \text{and} \quad (u^{(i)}, v^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i \in I^{\mathrm{zz}}(\bar{x}).
\end{aligned}$$

Here, $R^{(i)} \subset \mathbb{R} \times \mathcal{H}^{(i)}$ is defined by

$$R^{(i)} := \mathbb{R}^-\,c_1(\bar{G}^{(i)}) \qquad \forall i \in I^{\mathrm{bz}}, \qquad R^{(i)} := \mathbb{R}^+\,c_2(\bar{H}^{(i)}) \qquad \forall i \in I^{\mathrm{zb}}. \tag{2.6.10}$$

Proof. By Lemma 2.3.9, we have $(u,v) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_{\hat{\mathcal{K}}}(\cdot)^\circ}(\bar{G},\bar{H})$ if and only if $\operatorname{Proj}'_{\hat{\mathcal{K}}}(\bar{G}+\bar{H}; u+v) = u$. The latter relation yields $\operatorname{Proj}'_{\mathcal{K}^{(i)}}(\bar{G}^{(i)}+\bar{H}^{(i)}; u^{(i)}+v^{(i)}) = u^{(i)}$ for all $i = 1,\dots,N$.

Cases $i \in I^{\mathrm{zi}}(\bar{x})$, $i \in I^{\mathrm{iz}}(\bar{x})$, $i \in I^{\mathrm{bb}}(\bar{x})$: By Lemma 2.6.1, $\operatorname{Proj}'_{\mathcal{K}^{(i)}}(\bar{G}^{(i)}+\bar{H}^{(i)})$ is differentiable and the assertion follows, see also (2.6.7).

Case $i \in I^{\mathrm{bz}}(\bar{x})$: By Lemma 2.6.1 and using $\bar{H}^{(i)} = 0$, we have

$$\begin{aligned}&\operatorname{Proj}'_{\mathcal{K}^{(i)}}(\bar{G}^{(i)}+\bar{H}^{(i)}, u^{(i)}+v^{(i)})\\ &\qquad = u^{(i)}+v^{(i)} - 2\min\{0, (c_1(\bar{G}^{(i)}),\, u^{(i)}+v^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}}\}\, c_1(\bar{G}^{(i)}).\end{aligned}$$

Hence, $\operatorname{Proj}'_{\mathcal{K}^{(i)}}(\bar{G}^{(i)}+\bar{H}^{(i)}; u^{(i)}+v^{(i)}) = u^{(i)}$ if and only if

$$v^{(i)} = 2\min\{0, (c_1(\bar{G}^{(i)}),\, u^{(i)}+v^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}}\}\, c_1(\bar{G}^{(i)}).$$

Now, the assertion follows easily.

The case $i \in I^{\mathrm{zb}}(\bar{x})$ can be treated analogously to case $i \in I^{\mathrm{bz}}(\bar{x})$.

Case $i \in I^{\mathrm{zz}}(\bar{x})$: Since $\bar{G}^{(i)}+\bar{H}^{(i)} = 0$, we have $\mathcal{T}_{\operatorname{gph} \mathcal{T}_{\mathcal{K}^{(i)}}(\cdot)^\circ}(\bar{G}^{(i)},\bar{H}^{(i)}) = \operatorname{gph} \mathcal{T}_{\mathcal{K}^{(i)}}(\cdot)^\circ$.

We mention that the same result can be achieved by using Lemma 2.3.10 and (2.6.9).

As in Section 2.5, the tangent approximation set $W(\bar{G},\bar{H})$ is, in general, smaller than the tangent cone of $\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ$. In fact, $W(\bar{G},\bar{H})$ is the largest subspace contained in the possibly non-convex cone $\mathcal{T}_{\operatorname{gph} \mathcal{T}_{\mathbb{S}^n_+}(\cdot)^\circ}(\bar{G},\bar{H})$.

Using the tangent approximation W from Corollary 2.6.5 we can define an MFCQ-variant for (SOCMPCC) via (2.4.7).

Definition 2.6.7. Let $\bar{x}$ be a feasible point of (SOCMPCC). We say that SOCMPCC-MFCQ is satisfied at $\bar{x}$ if

$$Y \times \hat{\mathcal{H}} \times \hat{\mathcal{H}} = (g'(\bar{x}), G'(\bar{x}), H'(\bar{x}))\, X - \mathcal{R}_C(\bar{x}) \times W(\bar{G},\bar{H})$$

holds.

We will not discuss equivalent formulations of SOCMPCC-MFCQ here. We refer to (2.6.14) for a reformulation of SOCMPCC-LICQ.

An application of Theorem 2.4.1 yields a linearization result, similarly, we could use Corollary 2.4.3 and (2.6.9).

Theorem 2.6.8. Let $\bar{x}$ be a feasible point of (SOCMPCC), such that SOCMPCC-MFCQ is satisfied at $\bar{x}$. Then,

$$\mathcal{T}_F(\bar{x}) = \{h \in X : g'(\bar{x})\, h \in \mathcal{T}_C(g(\bar{x})),\ (G'(\bar{x})\, h, H'(\bar{x})\, h) \in \mathcal{T}_{\operatorname{gph} \mathcal{T}_{\hat{\mathcal{K}}}(\cdot)^\circ}(\bar{G},\bar{H})\}. \quad (2.6.11)$$

Moreover, $h = 0$ is a solution of

$$\begin{aligned}
\text{Min.}\quad & f'(\bar{x})\,h, \\
\text{s.t.}\quad & g'(\bar{x})\,h \in \mathcal{T}_C(g(\bar{x})), \\
& \overline{\mathrm{d}G}^{(i)}h = 0, && \forall i \in I^{\mathrm{iz}}(\bar{x}), \\
& \overline{\mathrm{d}G}^{(i)}h = 0, && \forall i \in I^{\mathrm{zi}}(\bar{x}), \\
& \overline{\mathrm{d}G}^{(i)}h \in (R^{(i)})^\circ,\ \overline{\mathrm{d}H}^{(i)}h \in R^{(i)},\ (\overline{\mathrm{d}G}^{(i)}h,\, \overline{\mathrm{d}H}^{(i)}h)_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i \in I^{\mathrm{bz}}(\bar{x}), \\
& \overline{\mathrm{d}G}^{(i)}h \in R^{(i)},\ \overline{\mathrm{d}H}^{(i)}h \in (R^{(i)})^\circ,\ (\overline{\mathrm{d}G}^{(i)}h,\, \overline{\mathrm{d}H}^{(i)}h)_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i \in I^{\mathrm{zb}}(\bar{x}), \\
& \overline{\mathrm{d}G}^{(i)}h = E^{(i)}(\overline{\mathrm{d}G}^{(i)}h + \overline{\mathrm{d}H}^{(i)}h), && \forall i \in I^{\mathrm{bb}}(\bar{x}), \\
& \overline{\mathrm{d}G}^{(i)}h \in \mathcal{K}^{(i)},\ \overline{\mathrm{d}H}^{(i)}h \in -\mathcal{K}^{(i)},\ (\overline{\mathrm{d}G}^{(i)}h,\, \overline{\mathrm{d}H}^{(i)}h)_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, && \forall i \in I^{\mathrm{zz}}(\bar{x}).
\end{aligned} \tag{2.6.12}$$

Here, $\overline{\mathrm{d}G}^{(i)}$, $\overline{\mathrm{d}H}^{(i)}$ refer to the derivatives of $G^{(i)}$ and $H^{(i)}$ at $\bar{x}$, respectively.

The satisfaction of (2.6.11) could be termed SOCMPCC-Abadie-CQ. We emphasize that the statement "SOCMPCC-MFCQ implies SOCMPCC-ACQ" is a new result.

Now, we are going to obtain strong stationarity conditions for (2.6.12). These optimality conditions will then yield an optimality condition for (SOCMPCC). The tightened NLP of (2.6.12) at the local minimizer $h = 0$ is given by

$$\begin{aligned}
\text{Minimize}\quad & f'(\bar{x})\,h, \\
\text{subject to}\quad & g'(\bar{x})\,h \in \mathcal{T}_C(g(\bar{x})), \\
& (G'(\bar{x})\,h, H'(\bar{x})\,h) \in W(\bar{G}, \bar{H}),
\end{aligned} \tag{2.6.13}$$

compare (2.4.10). Hence, SOCMPCC-MFCQ implies that the Kurcyusz-Robinson-Zowe-CQ is satisfied for (2.6.13) at $h = 0$, cf. (2.4.11a). It remains to provide an LICQ-variant for (2.6.13), cf. (2.4.11b). Here, we utilize that $W(\bar{G}, \bar{H})$ is a subspace.

Definition 2.6.9. Let $\bar{x}$ be a feasible point of (SOCMPCC). We say that SOCMPCC-LICQ is satisfied at $\bar{x}$ if

$$Y \times \hat{\mathcal{H}} \times \hat{\mathcal{H}} = \mathrm{cl}\big[(g'(\bar{x}), G'(\bar{x}), H'(\bar{x}))\,X - \mathcal{T}_C(\bar{x})^{\circ\perp} \times W(\bar{G}, \bar{H})\big]$$

holds.

We remark that SOCMPCC-LICQ implies that there is at most one Lagrange multiplier for (2.6.13) at $h = 0$, see Theorem 1.4.2. In the finite-dimensional case, the SOCMPCC-LICQ is just the nondegeneracy condition for the TNLP (2.6.13). Moreover, it is straightforward to check that SOCMPCC-LICQ implies SOCMPCC-MFCQ in the case that Y and $\hat{\mathcal{H}}$ are finite-dimensional, compare Bonnans, Shapiro, 2000, Corollary 2.98.

By Theorem 1.4.2, SOCMPCC-LICQ implies that there are no non-zero multipliers $\lambda \in Y^\star$, $\mu, \nu \in \hat{\mathcal{H}}$ which satisfy

$$g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu = 0, \qquad \lambda \in \operatorname{lin}(\mathcal{T}_C(g(\bar{x}))^\circ), \qquad (\mu, \nu) \in W(\bar{G}, \bar{H})^\perp.$$

It is easy to show that these conditions are equivalent to (see also Lemma 2.5.4 for the cases $I^{\mathrm{bb}}(\bar{x})$)

$$\begin{aligned}
g'(\bar{x})^\star \lambda + G'(\bar{x})^\star \mu + H'(\bar{x})^\star \nu &= 0, && && \text{(2.6.14a)}\\
\lambda &\in \operatorname{lin}(\mathcal{T}_C(g(\bar{x}))^\circ), && && \text{(2.6.14b)}\\
\mu^{(i)} &= 0, && \forall i \in I^{\mathrm{iz}}(\bar{x}), && \text{(2.6.14c)}\\
\nu^{(i)} &= 0, && \forall i \in I^{\mathrm{zi}}(\bar{x}), && \text{(2.6.14d)}\\
\mu^{(i)} &\in \operatorname{lin}(c_1(\bar{G}^{(i)})), && \forall i \in I^{\mathrm{bz}}(\bar{x}), && \text{(2.6.14e)}\\
\nu^{(i)} &\in \operatorname{lin}(c_2(\bar{H}^{(i)})), && \forall i \in I^{\mathrm{zb}}(\bar{x}), && \text{(2.6.14f)}\\
\nu^{(i)} &= E^{(i)}\,(\nu^{(i)} - \mu^{(i)}), && \forall i \in I^{\mathrm{bb}}(\bar{x}). && \text{(2.6.14g)}
\end{aligned}$$

Further, in the important case that Y is finite-dimensional, this non-existence of singular multipliers is even equivalent to SOCMPCC-LICQ, see again Theorem 1.4.2. Hence, our Definition 2.6.9 is significantly weaker then the LICQ-variant in Ye, Zhou, 2015, Definition 5.3.

Theorem 2.6.10. Let $\bar{x}$ be a local solution of (SOCMPCC) which satisfies SOCMPCC-MFCQ and SOCMPCC-LICQ. Then, there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in \hat{\mathcal{H}}$ such that the system

$$\begin{aligned}
0 &= f'(\bar{x}) + g'(\bar{x})^\star\,\lambda + G'(\bar{x})^\star\,\mu + H'(\bar{x})^\star\,\nu, && && \text{(2.6.15a)}\\
\lambda &\in \mathcal{T}_C(g(\bar{x}))^\circ, && && \text{(2.6.15b)}\\
\mu^{(i)} &= 0, && \forall i \in I^{\mathrm{iz}}(\bar{x}), && \text{(2.6.15c)}\\
\nu^{(i)} &= 0, && \forall i \in I^{\mathrm{zi}}(\bar{x}), && \text{(2.6.15d)}\\
\mu^{(i)} &\in R^{(i)}, \quad \text{and} \quad \nu^{(i)} \in (R^{(i)})^\circ, && \forall i \in I^{\mathrm{bz}}(\bar{x}), && \text{(2.6.15e)}\\
\mu^{(i)} &\in (R^{(i)})^\circ, \quad \text{and} \quad \nu^{(i)} \in R^{(i)}), && \forall i \in I^{\mathrm{zb}}(\bar{x}), && \text{(2.6.15f)}\\
\nu^{(i)} &= E^{(i)}\,(\nu^{(i)} - \mu^{(i)}), && \forall i \in I^{\mathrm{bb}}(\bar{x}), && \text{(2.6.15g)}\\
\mu^{(i)} &\in -\mathcal{K}^{(i)}, \quad \text{and} \quad \nu^{(i)} \in \mathcal{K}^{(i)}, && \forall i \in I^{\mathrm{zz}}(\bar{x}) && \text{(2.6.15h)}
\end{aligned}$$

is satisfied.

For convenience, we recall

$$R^{(i)} := \mathbb{R}^-\, c_1(\bar{G}^{(i)}), \qquad \forall i \in I^{\mathrm{bz}}(\bar{x}), \qquad R^{(i)} := \mathbb{R}^+\, c_2(\bar{H}^{(i)}), \qquad \forall i \in I^{\mathrm{zb}}(\bar{x}),$$

and refer to (2.6.7) for the definition of $E^{(i)}$, $i \in I^{\mathrm{bb}}(\bar{x})$.

Proof. One possibility to prove this theorem is to recast the linearized MPCC (2.6.12) in the form (2.4.9). Consequently, (2.6.13) corresponds to (2.4.10) and it is easy to check that the CQs (2.4.11) are given by SOCMPCC-LICQ and SOCMPCC-MFCQ.

Alternatively, Theorem 2.6.10 can be proved via Theorem 2.4.5 by using the operator L from (2.6.9).

We emphasize that Theorem 2.6.10 is a novel result for the analysis of (SOCMPCC). In particular, Ye, Zhou, 2015, Theorem 5.1 required a stronger CQ in order to show the same stationarity system, see Lemma 2.6.12 below.

We discuss the novelty of our result in the case that Y as well as $\hat{\mathcal{H}}$ are finite-dimensional. In this case, the non-existence of singular multipliers which satisfy (2.6.14) is equivalent to SOCMPCC-LICQ and implies SOCMPCC-MFCQ. Therefore, our definition of SOCMPCC-LICQ is significantly weaker than the definition in Ye, Zhou, 2015. Nevertheless, we obtain the same stationarity conditions under this weaker CQ. This finite-dimensional situation is summarized in the following theorem.

Theorem 2.6.11. Assume that the spaces Y and $\hat{\mathcal{H}}$ are finite-dimensional. Let $\bar{x}$ be a local solution of (SOCMPCC). We assume that there are no non-zero singular multipliers $\lambda \in Y^\star$, $\mu, \nu \in \hat{\mathcal{H}}$ such that (2.6.14) holds. Then, there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in \hat{\mathcal{H}}$ such that the strong stationarity system (2.6.15) is satisfied.

It remains to show that the optimality system (2.6.15) is equivalent to the optimality system in Ye, Zhou, 2015, Theorem 5.1. Up to (2.6.15g), this equivalence is straightforward to check. The equivalency of (2.6.15g) with the corresponding condition in Ye, Zhou, 2015, Theorem 5.1, is demonstrated in Lemma 2.6.12 below, see also Ye, Zhou, 2015, Proposition 3.1 for a different proof.

Lemma 2.6.12. Let $i \in I^{\mathrm{bb}}(\bar{x})$ be given. For $\mu^{(i)}, \nu^{(i)} \in \mathbb{R} \times \mathcal{H}^{(i)}$, the condition (2.6.15g) is equivalent to

$$\left((1, w^{(i)}),\, \mu^{(i)}\right)_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, \tag{2.6.16a}$$

$$\left((1, -w^{(i)}),\, \nu^{(i)}\right)_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0, \tag{2.6.16b}$$

$$\kappa^{(i)}\, \hat{\mu}^{(i)} + (\kappa^{(i)} - 1)\, \nu^{(i)} \in \operatorname{lin}((1, w^{(i)})), \tag{2.6.16c}$$

where $\hat{\mu}^{(i)} = (\mu_1^{(i)}, -\mu_2^{(i)})$.

Proof. Let (2.6.16) be satisfied. In particular, there is $\alpha \in \mathbb{R}$ such that

$$\nu^{(i)} = \alpha\, (1, w^{(i)}) - \frac{\kappa^{(i)}}{\kappa^{(i)} - 1}\, \hat{\mu}^{(i)}.$$

This yields

$$\nu^{(i)} - \mu^{(i)} = \alpha\,(1, w^{(i)}) - \frac{1}{\kappa^{(i)} - 1}\,(\kappa^{(i)}\,\hat{\mu}^{(i)} + (\kappa^{(i)} - 1)\,\mu^{(i)}).$$

Now,

$$\begin{aligned}
E^{(i)}\,(\nu^{(i)} - \mu^{(i)}) &= \alpha\,(1, w^{(i)}) + 0 - \frac{1}{2}\,(1, w^{(i)}) \otimes (1, w^{(i)}) \frac{\kappa^{(i)}}{\kappa^{(i)} - 1}\,\hat{\mu}^{(i)} + 0 \\
&\quad - \frac{\kappa^{(i)}}{\kappa^{(i)} - 1} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes (w^{(i)}) \end{bmatrix} (\kappa^{(i)}\,\hat{\mu}^{(i)} + (\kappa^{(i)} - 1)\,\mu^{(i)}) \\
&= \alpha\,(1, w^{(i)}) - \frac{1}{2}\,(1, w^{(i)}) \otimes (1, -w^{(i)}) \frac{\kappa^{(i)}}{\kappa^{(i)} - 1}\,\mu^{(i)} \\
&\quad - \frac{\kappa^{(i)}}{\kappa^{(i)} - 1} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes (w^{(i)}) \end{bmatrix} (-\kappa^{(i)}\,\mu^{(i)} + (\kappa^{(i)} - 1)\,\mu^{(i)}) \\
&= \alpha\,(1, w^{(i)}) - \frac{1}{2}\,\frac{\kappa^{(i)}}{\kappa^{(i)} - 1} \begin{bmatrix} 1 & -w^{(i)} \\ w^{(i)} & -w^{(i)} \otimes (w^{(i)}) \end{bmatrix} \mu^{(i)} \\
&\quad + \frac{\kappa^{(i)}}{\kappa^{(i)} - 1} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes w^{(i)} \end{bmatrix} \mu^{(i)} \\
&= \alpha\,(1, w^{(i)}) + \frac{1}{2}\,\frac{\kappa^{(i)}}{\kappa^{(i)} - 1} \begin{bmatrix} -1 & w^{(i)} \\ -w^{(i)} & 2\,I - w^{(i)} \otimes w^{(i)} \end{bmatrix} \mu^{(i)} \\
&= \alpha\,(1, w^{(i)}) + \frac{1}{2}\,\frac{\kappa^{(i)}}{\kappa^{(i)} - 1} \begin{bmatrix} -2 & 0 \\ 0 & 2\,I \end{bmatrix} \mu^{(i)} \\
&= \alpha\,(1, w^{(i)}) - \frac{\kappa^{(i)}}{\kappa^{(i)} - 1}\,\hat{\mu}^{(i)} = \nu^{(i)}.
\end{aligned}$$

Here, we used $((1, w^{(i)}),\, \mu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0$ frequently. This shows (2.6.15g).

Now, let (2.6.15g) be satisfied. By testing (2.6.15g) with $(1, -w^{(i)})$, we obtain

$$\begin{aligned}
&\frac{1}{\kappa^{(i)}}\,((1, -w^{(i)}),\, \nu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} \\
&\qquad = \left((1, -w^{(i)}),\, \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes (w^{(i)}) \end{bmatrix} (\nu^{(i)} - \mu^{(i)})\right)_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0.
\end{aligned}$$

Similar, by testing (2.6.15g) with $(1, w^{(i)})$, we obtain

$$((1, w^{(i)}),\, \nu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = ((1, w^{(i)}),\, \nu^{(i)} - \mu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}}$$

and this shows $((1, w^{(i)}),\, \mu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} = 0$.

Now, (2.6.15g) implies

$$\nu^{(i)} = \frac{1}{2}\,(1, w^{(i)})\,((1, w^{(i)}),\, \nu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} + \kappa^{(i)} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes w^{(i)} \end{bmatrix} (\nu^{(i)} - \mu^{(i)}).$$

Hence,

$$\begin{aligned}
(1-\kappa^{(i)})\,\nu^{(i)} - \kappa^{(i)}\,\hat{\mu}^{(i)} &= \frac{1}{2}\,(1, w^{(i)})\,((1, w^{(i)}),\, \nu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} \\
&\quad + \kappa^{(i)} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes w^{(i)} \end{bmatrix} \nu^{(i)} - \kappa^{(i)} \begin{bmatrix} 1 & 0 \\ 0 & I \end{bmatrix} \nu^{(i)} \\
&\quad - \kappa^{(i)} \begin{bmatrix} 0 & 0 \\ 0 & I - w^{(i)} \otimes w^{(i)} \end{bmatrix} \mu^{(i)} - \kappa^{(i)} \begin{bmatrix} 1 & 0 \\ 0 & -I \end{bmatrix} \mu^{(i)} \\
&= \frac{1}{2}\,(1, w^{(i)})\,((1, w^{(i)}),\, \nu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} \\
&\quad - \kappa^{(i)} \begin{bmatrix} 1 & 0 \\ 0 & w^{(i)} \otimes w^{(i)} \end{bmatrix} \nu^{(i)} - \kappa^{(i)} \begin{bmatrix} 1 & 0 \\ 0 & -w^{(i)} \otimes w^{(i)} \end{bmatrix} \mu^{(i)} \\
&= \frac{1}{2}\,(1, w^{(i)})\,((1, w^{(i)}),\, \nu^{(i)})_{\mathbb{R}\times\mathcal{H}^{(i)}} \\
&\quad + \kappa^{(i)} \begin{bmatrix} 0 & -w^{(i)} \\ 0 & -w^{(i)} \otimes w^{(i)} \end{bmatrix} \nu^{(i)} - \kappa^{(i)} \begin{bmatrix} 0 & -w^{(i)} \\ 0 & -w^{(i)} \otimes w^{(i)} \end{bmatrix} \mu^{(i)} \\
&\in \operatorname{lin}((1, w^{(i)})).
\end{aligned}$$

This shows (2.6.16).

2.7. A problem with an infinite-dimensional cone complementarity constraint

In this section, we illustrate that the theory from Section 2.4 can also be applied to situations with an infinite-dimensional K beyond the SOC case discussed in Section 2.6. In particular, we focus on the situation

$$\begin{aligned}
Z &= L^2(0,1) \times L^2(0,1)^n \\
K &= \{(z_1, z_2) \in Z : z_1(t) \geq |z_2(t)|_{\mathbb{R}^n} \text{ for a.a. } t \in (0,1)\}.
\end{aligned}$$

Here, $L^2(0,1)$ is the usual Lebesgue space over the unit interval $(0,1)$ and $|\cdot|_{\mathbb{R}^n}$ is the Euclidean norm on $\mathbb{R}^n$. In fact, K could be considered as an *infinite product* of second-order cones $\mathcal{K} \subset \mathbb{R}^{n+1}$.

In this setting, we consider the problem

$$\begin{aligned}
\text{Minimize} \quad & f(x), \\
\text{subject to} \quad & g(x) \in C, \\
& G(x) \in K, \\
& H(x) \in K^\circ, \\
& \langle G(x),\, H(x) \rangle = 0.
\end{aligned} \tag{ISOCMPCC}$$

Here, $f : X \to \mathbb{R}$ is Fréchet differentiable, $g : X \to Y$, $G, H : X \to Z$ are strictly Fréchet differentiable, X, Y are (real) Banach spaces. Moreover, $C \subset Y$ is a closed, convex set.

It is straightforward to check that the projection onto K acts pointwise, that is,

$$\mathrm{Proj}_K(v)(t) = \mathrm{Proj}_{\mathcal{K}}(v(t)) \quad \text{for a.a. } t \in (0,1)$$

for $v \in L^2(0,1)^{n+1}$. Consequently, we obtain the directional differentiability of $\mathrm{Proj}_{\mathcal{K}}$.

Lemma 2.7.1. For any $u, v \in L^2(0,1)^{n+1}$, the mapping $\mathrm{Proj}_{\mathcal{K}} : L^2(0,1)^{n+1} \to L^2(0,1)^{n+1}$ is directionally differentiable at u in direction v and

$$\mathrm{Proj}'_K(u; v)(t) = \mathrm{Proj}'_{\mathcal{K}}(u(t); v(t)) \quad \text{for a.a. } t \in (0,1).$$

Proof. For $r > 0$, we have the pointwise convergence

$$\begin{aligned}\frac{\mathrm{Proj}_K(u + r\,v)(t) - \mathrm{Proj}_K(u)(t)}{r} &= \frac{\mathrm{Proj}_{\mathcal{K}}(u(t) + r\,v(t)) - \mathrm{Proj}_{\mathcal{K}}(u(t))}{r} \\ &\to \mathrm{Proj}'_{\mathcal{K}}(u(t); v(t)) \quad \text{as } r \searrow 0 \text{ for a.a. } t \in (0,1).\end{aligned}$$

Moreover, we have the pointwise bound

$$\left\| \frac{\mathrm{Proj}_K(u + r\,v)(t) - \mathrm{Proj}_K(u)(t)}{r} \right\|_{\mathbb{R}^{n+1}} \leq \|v(t)\|_{\mathbb{R}^{n+1}},$$

which is uniform in r and square-integrable. Now, the claim follows from the dominated convergence theorem.

We consider a feasible point $\bar{x} \in X$ of (ISOCMPCC). For convenience, we set $\bar{G} = G(\bar{x})$ and $\bar{H} = H(\bar{x})$. We define the (disjoint) active sets

$$\begin{aligned}
I^{\mathrm{iz}}(\bar{x}) &:= \{t \in (0,1) : \bar{G}(t) \in \operatorname{int} \mathcal{K}\}, \\
I^{\mathrm{zi}}(\bar{x}) &:= \{t \in (0,1) : \bar{H}(t) \in -\operatorname{int} \mathcal{K}\}, \\
I^{\mathrm{bz}}(\bar{x}) &:= \{t \in (0,1) : \bar{G}(t) \in \operatorname{bd} \mathcal{K} \setminus \{0\}, \bar{H} = 0\}, \\
I^{\mathrm{zb}}(\bar{x}) &:= \{t \in (0,1) : \bar{G}(t) = 0, \bar{H} \in -\operatorname{bd} \mathcal{K} \setminus \{0\}\}, \\
I^{\mathrm{bb}}(\bar{x}) &:= \{t \in (0,1) : \bar{G}(t) \in \operatorname{bd} \mathcal{K} \setminus \{0\}, \bar{H} \in -\operatorname{bd} \mathcal{K} \setminus \{0\}\}, \\
I^{\mathrm{zz}}(\bar{x}) &:= \{t \in (0,1) : \bar{G}(t) = 0, \bar{H} = 0\}.
\end{aligned}$$

where $\mathcal{K} \subset \mathbb{R}^{n+1}$ is the finite-dimensional second-order cone.

It is easy to check that the critical cone to K can be computed pointwise via the critical

cone to $\mathcal{K}$. By using Lemma 2.6.2, we have $v \in \mathcal{K}_K(\bar{G}, \bar{H})$ if and only if

$$\begin{aligned}
v(t) &\in \mathbb{R}^{n+1}, && \text{for a.a. } t \in I^{\mathrm{iz}}(\bar{x}),\\
v(t) &= 0, && \text{for a.a. } t \in I^{\mathrm{zi}}(\bar{x}),\\
(c_1(\bar{G}+\bar{H}),\, v(t))_{\mathbb{R}^{n+1}} &\geq 0, && \text{for a.a. } t \in I^{\mathrm{bz}}(\bar{x}),\\
v(t) &\in \mathbb{R}^{+}\, c_2(\bar{G}+\bar{H}), && \text{for a.a. } t \in I^{\mathrm{zb}}(\bar{x}),\\
(c_1(\bar{G}+\bar{H}),\, v(t))_{\mathbb{R}^{n+1}} &= 0, && \text{for a.a. } t \in I^{\mathrm{bb}}(\bar{x}),\\
v(t) &\in \mathcal{K}, && \text{for a.a. } t \in I^{\mathrm{zz}}(\bar{x}).
\end{aligned}$$

Moreover, the assertion of Lemma 2.3.10 is satisfied if we define $L : L^2(0,1)^{n+1} \to L^2(0,1)^{n+1}$ via $(L\,v)(t) = L(t)\,v(t)$ with $L(t) : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defined analogously to (2.6.9).

By owing to Lemma 2.4.7 we obtain the following results. Note that the optimality conditions are evaluated similarly to Theorem 2.6.10.

Theorem 2.7.2. Let us assume that the operator $(g'(\bar{x}), G'(\bar{x}), H'(\bar{x})) \in \mathcal{L}(X, Y \times L^2(0,1)^{n+1} \times L^2(0,1)^{n+1})$ is surjective. Then, there exist multipliers $\lambda \in Y^\star$, $\mu, \nu \in L^2(0,1)^{n+1}$ such that the system

$$\begin{aligned}
0 &= f'(\bar{x}) + g'(\bar{x})^\star\,\lambda + G'(\bar{x})^\star\,\mu + H'(\bar{x})^\star\,\nu, && && \text{(2.7.1a)}\\
\lambda &\in \mathcal{T}_C(g(\bar{x}))^\circ, && && \text{(2.7.1b)}\\
\mu(t) &= 0, && \text{for a.a. } t \in I^{\mathrm{iz}}(\bar{x}), && \text{(2.7.1c)}\\
\nu(t) &= 0, && \text{for a.a. } t \in I^{\mathrm{zi}}(\bar{x}), && \text{(2.7.1d)}\\
\mu(t) &\in R(t), \quad \text{and} \quad \nu(t) \in (R(t))^\circ, && \text{for a.a. } t \in I^{\mathrm{bz}}(\bar{x}), && \text{(2.7.1e)}\\
\mu(t) &\in (R(t))^\circ, \quad \text{and} \quad \nu(t) \in R(t)), && \text{for a.a. } t \in I^{\mathrm{zb}}(\bar{x}), && \text{(2.7.1f)}\\
\nu(t) &= E(t)\,(\nu(t) - \mu(t)), && \text{for a.a. } t \in I^{\mathrm{bb}}(\bar{x}), && \text{(2.7.1g)}\\
\mu(t) &\in -\mathcal{K}(t), \quad \text{and} \quad \nu(t) \in \mathcal{K}(t), && \text{for a.a. } t \in I^{\mathrm{zz}}(\bar{x}) && \text{(2.7.1h)}
\end{aligned}$$

is satisfied. Here,

$$R(t) := \mathbb{R}^{-}\,c_1(\bar{G}(t)), \quad \text{f.a.a. } t \in I^{\mathrm{bz}}(\bar{x}), \quad R(t) := \mathbb{R}^{+}\,c_2(\bar{H}(t)), \quad \text{f.a.a. } t \in I^{\mathrm{zb}}(\bar{x}),$$

and

$$E(t) = \mathrm{Proj}'_{\mathcal{K}}(\bar{G}(t) + \bar{H}(t)) = \frac{1}{2}\,(1, w(t)) \otimes (1, w(t)) + \kappa(t) \begin{bmatrix} 0 & 0 \\ 0 & I - w(t) \otimes (w(t)) \end{bmatrix},$$

for a.a. $t \in I^{\mathrm{bb}}(\bar{x})$, where $w(t) = \bar{G}_2(t)/\bar{G}_1(t)$ and $\kappa(t) = \bar{G}_1(t)/(\bar{G}_1(t) - \bar{H}_1(t))$.

Theorem 2.7.2 shows that the abstract theory from Section 2.4 is applicable also in situations in with the complementarity is defined by a infinite-dimensional cone.

For simplicity, we only stated the theorem by assuming the surjectivity of the operator $(g'(\bar{x}), G'(\bar{x}), H'(\bar{x}))$. By deriving a tangent approximation set W of K (similarly to (2.3.3)) one can formulate weaker CQs based on (2.4.7), and (2.4.14). In this infinite-dimensional setting, however, we expect that a tangent approximation set is strictly smaller than the largest linear subspace within $\mathcal{T}_{\operatorname{gph} \mathcal{T}_K(\cdot)^\circ}(\bar{G}, \bar{H})$, compare (2.3.3). Thus, (2.4.14a) will be implied by (2.4.7), but (2.4.7) and (2.4.14b) does not imply each other.

2.8. Conclusions

In this work, we study optimization problems with complementarity constraints in Banach spaces. Under reasonable constraint qualifications, we obtain a linearization argument, see Theorem 2.4.1. This can be utilized to obtain strong stationarity conditions, cf. Theorem 2.4.5.

We apply the theory to problems with semidefinite and second-order-cone complementarity constraints. In both situations, we obtain that a reasonable variant of LICQ implies strong stationarity of local minimizers, see Theorems 2.5.8 and 2.6.10, and these are novel results. Also in the finite-dimensional case, the results are superior to the corresponding results in the literature. In fact, the implication SDPMPCC-LICQ implies strong stationarity (see Theorem 2.5.9) was not previously established. In the SOCMPCC-case, the definition of SOCMPCC-LICQ from the literature is stronger than our definition and we are able to show strong stationarity under this weaker CQ, see Theorem 2.6.11.

It is also possible to apply the abstract theory to problems with standard complementarity conditions, i.e., $K = \{v \in \mathbb{R}^n : v \geq 0\}$. In this case, we are able to reproduce some well-known results, e.g., MPCC-MFCQ implies MPCC-ACQ by Theorem 2.4.1, see Flegel, Kanzow, 2005a, Theorem 3.1; or MPCC-LICQ implies strong stationarity of local minimizers by Theorem 2.4.5, see Scheel, Scholtes, 2000, Theorem 2.

Part II.

Optimality conditions for the optimal control of the obstacle problem

Contents of Part II

Introduction

The obstacle problem is a classical variational inequality (VI) in the Sobolev space $H_0^1(\Omega)$. It models the deflection $y \in H_0^1(\Omega)$ of an elastic membrane due to an external force $u \in L^2(\Omega)$ and this deflection is constrained by an obstacle $\psi \in H^1(\Omega)$. This leads to the energy minimization problem

$$\begin{aligned} \text{Minimize} \quad & \int_\Omega \frac{1}{2} |\nabla y|^2 - y\, u \,\mathrm{d}x \\ \text{w.r.t.} \quad & y \in K, \end{aligned}$$

where $K = \{v \in H_0^1(\Omega) \mid y \leq \psi\}$. It is straightforward to check that the solution $\bar{y}$ can be equivalently characterized by the complementarity system

$$\begin{aligned} -\Delta \bar{y} &= u - \bar{\xi}, \\ \bar{y} &\leq \psi, \\ \langle \bar{\xi},\, v - \bar{y} \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} &\leq 0 \quad \forall v \in K, \end{aligned}$$

where $\bar{\xi} \in H^{-1}(\Omega)$ is the multiplier corresponding to the constraint $y \in K$. It can be interpreted as the contact force between the membrane and the obstacle.

Now, we consider the optimal control of the obstacle problem

$$\begin{aligned} \text{Minimize} \quad & J(y, u) \\ \text{w.r.t.} \quad & y \in H_0^1(\Omega), u \in L^2(\Omega), \\ \text{such that} \quad & y \text{ solves the obstacle problem with right-hand side } u \\ \text{and} \quad & u_a \leq u \leq u_b. \end{aligned}$$

By replacing the obstacle problem with the complementarity system, we obtain

$$\begin{aligned} \text{Minimize} \quad & J(y, u) \\ \text{w.r.t.} \quad & y \in H_0^1(\Omega), u \in L^2(\Omega), \xi \in H^{-1}(\Omega), \\ \text{such that} \quad & -\Delta y = u - \xi, \\ & y \leq \psi, \\ & \langle \xi,\, v - y \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \leq 0 \quad \forall v \in K \\ \text{and} \quad & u_a \leq u \leq u_b. \end{aligned}$$

This is an optimization problem with complementarity constraints in infinite-dimensional spaces, cf. Section 1.5.4. In the absence of the control constraints, we know that local

minimizers are strongly stationary, cf. Mignot, 1976. However, in the presence of control constraints, the necessity of strong stationarity is an open problem. Under certain assumptions on the data, we are able to verify strong stationarity of local minimizers, see Chapter 3. Moreover, we present counterexamples which demonstrate that the system of strong stationarity may not hold if these assumptions are not satisfied.

We use a different technique for deriving optimality conditions in Chapter 4. Therein, we regularize the control constraints and pass to the limit with the corresponding optimality systems. Under a mild assumption on the sequence of adjoint states, we arrive at a system of M-stationarity.

3. Strong stationarity for optimal control of the obstacle problem with control constraints

Abstract: We consider the distributed optimal control of the obstacle problem with control constraints. Since Mignot proved in 1976 the necessity of a system which is equivalent to strong stationarity, it has been an open problem whether such a system is still necessary in the presence of control constraints. Using moderate regularity of the optimal control and an assumption on the control bounds (which is implied by $u_a < 0 \leq u_b$ quasi-everywhere (q.e.) in Ω in the case of an upper obstacle $y \leq \psi$), we can answer this question in the affirmative. We also present counterexamples showing that strong stationarity may not hold if $u_a < 0$ or $0 \leq u_b$ are violated.

Keywords: obstacle problem, optimal control, strong stationarity, complementarity conditions, control constraints

MSC: 49K21, 35J86

3.1. Introduction

We consider the distributed optimal control of the obstacle problem with control constraints

$$\begin{aligned} \text{Minimize} \quad & j(y) + \frac{\alpha}{2} \|u\|^2_{L^2(\Omega)}, \\ \text{with respect to} \quad & (y, u, \xi) \in H^1_0(\Omega) \times L^2(\Omega) \times H^{-1}(\Omega), \\ \text{such that} \quad & \mathcal{A}\, y = u - \xi + f, \\ & 0 \geq y - \psi \ \perp \ \xi \geq 0, \\ \text{and} \quad & u_a \leq u \leq u_b \quad \text{a.e. in } \Omega. \end{aligned} \tag{\textbf{P}}$$

Here, the set $\Omega \subset \mathbb{R}^n$, $n \geq 1$, is open and bounded. The objective consists of an Fréchet-differentiable observation term $j : H^1_0(\Omega) \to \mathbb{R}$ of the state y and of an $L^2(\Omega)$-regularization term with $\alpha > 0$. The bounded linear operator $\mathcal{A} : H^1_0(\Omega) \to H^{-1}(\Omega)$ is assumed to be coercive. The right-hand side f belongs to $H^{-1}(\Omega)$. The control bounds satisfy $u_a, u_b \in H^1(\Omega)$. The obstacle $\psi \in H^1(\Omega)$ satisfies $\psi \geq 0$ on Γ in the sense that $\min\{\psi, 0\} \in H^1_0(\Omega)$. The complementarity condition

$$0 \geq y - \psi \ \perp \ \xi \geq 0$$

is to be understood in the dual pairing of $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, that is

$$y - \psi \leq 0 \quad \text{a.e. in } \Omega, \tag{3.1.1a}$$

$$\langle \xi,\, v - y \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \leq 0 \quad \text{for all } v \in H_0^1(\Omega) : v \leq \psi \text{ a.e. in } \Omega. \tag{3.1.1b}$$

Note that both statements $\xi \geq 0$ and $y - \psi \perp \xi$ are contained in (3.1.1b), see Proposition 3.2.5. Since this complementarity condition is a constraint in (**P**), all constraint qualifications (CQs) of a certain strength, e.g., the CQ of Robinson-Zowe-Kurcyusz, see Robinson, 1976; Zowe, Kurcyusz, 1979, are violated by (**P**). Hence, proving necessary optimality conditions is difficult and, as in the case of finite-dimensional mathematical programs with complementarity constraints (MPCCs), there exists a wide variety of stationarity concepts, see Scheel, Scholtes, 2000, Sec. 2 for stationarity concepts in finite dimensions and, e.g., Mignot, 1976; Barbu, 1984; Mignot, Puel, 1984; Hintermüller, Kopacka, 2009; Outrata, Jarušek, Stará, 2011; Herzog, C. Meyer, G. Wachsmuth, 2012; 2013; Schiela, D. Wachsmuth, 2013; Hintermüller, Mordukhovich, Surowiec, 2014 for stationarity conditions for MPCCs in function space.

Among these conditions, the so called *strong stationarity conditions* are the strongest. In finite dimensions, strong stationarity conditions are necessary for optimality if the MPCC satisfies the Guignard-CQ. The Guignard-CQ is in turn implied by MPCC-LICQ (and by even weaker CQs) and hence usually satisfied for MPCCs, see Flegel, Kanzow, 2005b.

For optimal control problems involving infinite-dimensional complementarity systems, however, strong stationarity has been proved so far only under rather restrictive conditions, namely that the set of admissible controls U_{ad} (or, strictly speaking, its tangent cone) has to be dense in the set of right-hand sides of the state equation (here: $H^{-1}(\Omega)$). This condition excludes the case where control constraints are present (and active). For the derivation of strong stationarity we refer to Mignot, 1976, Thm. 5.2, see also Mignot, Puel, 1984, for the control of the obstacle problem with control in $L^2(\Omega)$, Outrata, Jarušek, Stará, 2011, Thm. 6 for the case with controls in $H^{-1}(\Omega)$, and to Herzog, C. Meyer, G. Wachsmuth, 2013, Thm. 4.5 for an optimal-control problem arising in elasto-plasticity. Moreover, Mignot, 1976 was able to derive a strongly stationary system in the cases where the control-to-state map is Fréchet differentiable, see Mignot, 1976, p.161, or if the desired state y_d satisfies a certain condition rendering the objective convex, see Mignot, 1976, p.166.

It has been a long-standing issue to prove or disprove the necessity of strong stationarity also in the case of control constraints. Besides this, we remark that strong stationarity is part of second order sufficient conditions (SSC), see Kunisch, D. Wachsmuth, 2012, Thm. 2.2. These SSCs provide a quadratic growth condition, which is in turn essential to prove discretization error estimates, see C. Meyer, Thoma, 2013, Ass. 5.3, Thm. 5.7.

The main goal of this paper is to provide the necessity of strong stationarity under less restrictive assumptions than previously known. We only require some moderate regularity of the optimizer $\bar{u} \in H_0^1(\Omega)$ and a technical assumption (3.5.1) on the control constraints. This assumption is satisfied, e.g., if $u_a < 0 \leq u_b$ holds q.e. on Ω, see Lemma 3.5.3. We

refer to Section 3.2 for the notion of quasi-everywhere (q.e.). In particular, we do *not* require regularity of the domain Ω or of any of the active sets. For the problem under consideration, strong stationarity is defined in (3.1.3) after the introduction of some notation.

Let us mention that the discussion of the state-constrained problem (3.5.8), which appears as an auxiliary problem, is interesting in its own right. In dependence on the active sets, there may be no interior point of the feasible set w.r.t. the topology of $C(\bar{\Omega})$. We prove the existence of multipliers which belong to $H^{-1}(\Omega)$. This space is different from the measure space $\mathcal{M}(\Omega) = C_0(\Omega)'$, which is typically expected for state-constrained problems, see, e.g., Casas, 1986. A similar phenomenon was observed in Schiela, 2009, where the existence of an interior point, however, was assumed in a space more regular than the state space.

Let us give a brief outline of the paper. In the remainder of the introduction, we fix some notation and introduce the system of strong stationarity. Some basic results on capacity theory are recalled in Section 3.2. In Section 3.3 we consider a linearization of (**P**), which is used in Sections 3.4 and 3.5 to prove additional properties of a local minimizer $\bar{u}$ and to show that strong stationarity is a necessary condition, respectively. We present two counterexamples in Section 3.6 demonstrating that strong stationarity may not hold when the assumption $u_a < 0 \leq u_b$ is violated. In Section 3.A, we give an explicit characterization of the strictly active set $\tilde{A}_s$, which differs from the usual definition of A_s in the literature. Our definition of $\tilde{A}_s$, see also Lemma 3.3.1, is more suited for our analysis, since it allows for a quasi-every formulation (see Section 3.2 for the definition of quasi-everywhere) of the cone $\mathcal{K}(\bar{u})$, which occurs in the linearized state equation, see (3.3.1).

Notation

We define the set of admissible controls

$$U_{\text{ad}} := \{u \in L^2(\Omega) : u_a \leq u \leq u_b \text{ a.e. on } \Omega\},$$

and the closed convex set

$$K := \{y \in H_0^1(\Omega) : y \leq \psi \text{ a.e. on } \Omega\}.$$

For a convex set $M \subset Y$ in a normed space Y and $y \in M$ we denote by $\mathcal{T}_M(y)$ the tangent cone of M at y, which is the closed conic hull of $M - y$. We use this notation for the sets $K \subset H_0^1(\Omega)$ and $U_{\text{ad}} \subset L^2(\Omega)$.

For sets $M \subset H_0^1(\Omega)$ and $N \subset H^{-1}(\Omega)$ we define, as usual, the polar cones

$$M^\circ := \{f \in H^{-1}(\Omega) : \langle f, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \leq 0 \text{ for all } v \in M\},$$
$$N^\circ := \{v \in H_0^1(\Omega) : \langle f, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \leq 0 \text{ for all } f \in N\}.$$

Using this notation, the complementarity condition (3.1.1b) is equivalent to

$$\xi \in \mathcal{T}_K(y)^\circ. \tag{3.1.2}$$

In Section 3.5 we work with a closed subspace $V \subset H_0^1(\Omega)$. For subsets $M \subset V$ and $N \subset V^\star$ we define the polars w.r.t. the V-$V^\star$ duality. We make also use of the polar cone of $\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})$ w.r.t. the $L^2(\Omega)$-inner product which is denoted by

$$\mathcal{N}_{U_{\mathrm{ad}}}(\bar{u}) = \Big\{ v \in L^2(\Omega) : \int_\Omega u\, v \,\mathrm{d}x \le 0 \text{ for all } u \in \mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}) \Big\}.$$

Finally, for $\xi \in H^{-1}(\Omega)$, we define the annihilator

$$\xi^\perp := \{ v \in H_0^1(\Omega) : \langle \xi,\, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = 0 \}.$$

Strong stationarity

Using standard arguments, the existence of minimizers $(\bar{y}, \bar{u}, \bar{\xi})$ of **(P)** can be shown under additional assumptions (in particular, j has to be bounded from below and weakly lower semicontinuous; and $u_a \le u_b$), see, e.g., Mignot, Puel, 1984, Thm. 2.1.

Throughout the paper, we denote by $(\bar{y}, \bar{u}, \bar{\xi})$ a local minimum of **(P)**. We define (up to sets of zero capacity, see Definition 3.2.1 for the notion of capacity) the active sets w.r.t. the control constraints

$$A_a := \{ x \in \Omega : \bar{u}(x) = u_a(x) \},$$
$$A_b := \{ x \in \Omega : \bar{u}(x) = u_b(x) \},$$

as well as the active set w.r.t. the constraint $y - \psi \le 0$ in the obstacle problem

$$A := \{ x \in \Omega : 0 = \bar{y}(x) - \psi(x) \}.$$

By modifying A (or, equivalently, $\bar{y}$) on a set of zero capacity if necessary, we may assume that A is a Borel set, see Lemma 3.2.2. Note that the active sets A_a, A_b, A are quasi-closed, see Definition 3.2.1 for the notion of quasi-closeness.

We say that a feasible point $(\bar{y}, \bar{u}, \bar{\xi})$ of **(P)** is strongly stationary, if there exist multipliers $(p, \mu, \nu) \in H_0^1(\Omega) \times H^{-1}(\Omega) \times L^2(\Omega)$, such that the system of strong stationarity

$$\mathcal{A}^\star p + j'(\bar{y}) + \mu = 0 \quad \text{in } H^{-1}(\Omega), \tag{3.1.3a}$$
$$\alpha\, \bar{u} - p + \nu = 0 \quad \text{in } L^2(\Omega), \tag{3.1.3b}$$
$$-p \in \mathcal{T}_K(\bar{y}) \cap \bar{\xi}^\perp, \tag{3.1.3c}$$
$$\mu \in (\mathcal{T}_K(\bar{y}) \cap \bar{\xi}^\perp)^\circ, \tag{3.1.3d}$$
$$\nu \in \mathcal{N}_{U_{\mathrm{ad}}}(\bar{u}) \tag{3.1.3e}$$

is satisfied. Here, $\mathcal{A}^\star : H_0^1(\Omega) \to H^{-1}(\Omega)$ is the adjoint operator of $\mathcal{A}$. Note that (3.1.3) is a generalization of the necessary conditions in the case without control constraints

Mignot, 1976, Thm. 4.3, Mignot, Puel, 1984, Thm. 2.2, which can be obtained by setting $U_{\text{ad}} = L^2(\Omega)$ and hence $\nu = 0$ in (3.1.3).

Using the representation (3.3.4) of the set $\mathcal{T}_K(\bar{y}) \cap \bar{\xi}^{\perp}$, we can rewrite (3.1.3c) and (3.1.3d) equivalently as

$$p \geq 0 \text{ q.e. on } \tilde{B} \text{ and } p = 0 \text{ q.e. on } \tilde{A}_s, \tag{3.1.3c'}$$

$$\langle \mu, v \rangle_{H^{-1}, H_0^1} \geq 0 \text{ for all } v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \tilde{B} \text{ and } v = 0 \text{ q.e. on } \tilde{A}_s. \tag{3.1.3d'}$$

The strongly active set $\tilde{A}_s$ and the biactive set $\tilde{B} = A \setminus \tilde{A}_s$ are defined in Lemma 3.3.1, see also Section 3.A. We remark that (3.1.3c') and (3.1.3d') are falsely stated in many papers as

$$p \geq 0 \text{ a.e. on } A \setminus A_s \text{ and } p = 0 \text{ a.e. on } A_s, \tag{3.1.3c"}$$

$$\langle \mu, v \rangle_{H^{-1}, H_0^1} \geq 0 \text{ for all } v \in H_0^1(\Omega) : v \geq 0 \text{ a.e. on } B \text{ and } v = 0 \text{ a.e. on } A_s, \tag{3.1.3d"}$$

where $A_s := \{x \in \Omega : \bar{\xi}(x) > 0\}$ for $\bar{\xi} \in L^2(\Omega)$ and $B := A \setminus A_s$. The first condition is weaker than (3.1.3c'), but the second one is stronger than (3.1.3d'). Moreover, it is easy to construct an example with $U_{\text{ad}} = L^2(\Omega)$, where A is a set of measure zero, but non-zero capacity. Then, (3.1.3d') is satisfied by a solution, whereas (3.1.3d") may not hold.

We do not denote the strictly (biactive) set by A_s (B) in order to remind the reader that our definition of it differs from the usual definition in the literature.

3.2. Basics about capacity theory

In this section, we will recall some basic results in capacity theory. First, we give the definitions, see, e.g., Attouch, Buttazzo, Michaille, 2006, Sec. 5.8.2, Bonnans, Shapiro, 2000, Def. 6.47, and Delfour, Zolésio, 2001, Sec. 8.6.1.

Definition 3.2.1. The *capacity* of a set $A \subset \Omega$ (w.r.t. $H_0^1(\Omega)$) is defined as

$$\operatorname{cap}(A) := \inf\{\|\nabla v\|^2_{L^2(\Omega)^n} : v \in H_0^1(\Omega), v \geq 1 \text{ a.e. in a neighbourhood of } A\}. \tag{3.2.1}$$

A function $v : \Omega \to \mathbb{R}$ is called *quasi-continuous* if for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\operatorname{cap}(G_\varepsilon) < \varepsilon$ and v is continuous on $\Omega \setminus G_\varepsilon$.

A set $O \subset \Omega$ is called *quasi-open* if for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\operatorname{cap}(G_\varepsilon) < \varepsilon$ and $O \cup G_\varepsilon$ is open.

Finally, $D \subset \Omega$ is called *quasi-closed* if $\Omega \setminus D$ is quasi-open.

A set of zero capacity has measure zero, but the converse does not hold.

It is known, see Delfour, Zolésio, 2001, Thm. 6.1, that every $v \in H^1(\Omega)$ possesses a quasi-continuous representative. This representative is uniquely determined up to

sets of zero capacity. When we speak about a function $v \in H^1(\Omega)$, we always mean the quasi-continuous representative. For every quasi-continuous function v, the set $\{x \in \Omega : v(x) \leq 0\}$ is quasi-closed, whereas $\{x \in \Omega : v(x) > 0\}$ is quasi-open. Every sequence which converges in $H_0^1(\Omega)$ possesses a pointwise quasi-everywhere convergent subsequence, see Bonnans, Shapiro, 2000, Lem. 6.52.

The next lemma shows that quasi-open (and, similarly, quasi-closed) sets are "almost" Borel sets. This result is classical, but it is not contained in any of the above references. Hence, for the convenience of the reader, we state its proof.

Lemma 3.2.2. Let $O \subset \Omega$ be quasi-open. Then there exists a set $M \subset \Omega$, $\operatorname{cap}(M) = 0$, such that $O \cup M$ is a Borel set.

Proof. By definition, for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\operatorname{cap}(G_\varepsilon) \leq \varepsilon$ and $O \cup G_\varepsilon$ is open. Then, we have $O \subset \bigcap_{i\in\mathbb{N}}(O \cup G_{1/i})$ and

$$\operatorname{cap}\Big(\Big[\bigcap_{i\in\mathbb{N}} O \cup G_{1/i}\Big] \setminus O\Big) \leq \operatorname{cap}\Big(\Big[\bigcap_{i\in\mathbb{N}} G_{1/i}\Big]\Big) = 0,$$

by the monotonicity of the capacity, see Bonnans, Shapiro, 2000, Lem. 6.48. Hence, O differs from the Borel set $\bigcap_{i\in\mathbb{N}}(O \cup G_{1/i})$ only by a set $M \subset \Omega$ with $\operatorname{cap}(M) = 0$.

We say that $v \geq 0$ holds quasi-everywhere (q.e.) on $O \subset \Omega$ if

$$\operatorname{cap}(\{v < 0\} \cap O) = 0.$$

The next lemma is essential for converting a.e.-statements into q.e.-statements. It is a slight generalization of Bonnans, Shapiro, 2000, Lem. 6.49.

Lemma 3.2.3. Let $O \subset \Omega$ be a quasi-open subset and $v : \Omega \to \mathbb{R}$ a quasi-continuous function. Then, $v \geq 0$ a.e. on O implies $v \geq 0$ q.e. on O.

Proof. Let $\varepsilon > 0$ be given. Since O is quasi-open and v is quasi-continuous, there exist open sets $G_\varepsilon, H_\varepsilon$ such that v is continuous on $\Omega \setminus G_\varepsilon$, $O \cup H_\varepsilon$ is open, and $\operatorname{cap}(G_\varepsilon) \leq \varepsilon$, $\operatorname{cap}(H_\varepsilon) \leq \varepsilon$.

We set $U_\varepsilon = G_\varepsilon \cup H_\varepsilon$. Using the continuity of v on $\Omega \setminus U_\varepsilon$, the set $\{v < 0\} \cup U_\varepsilon$ is open. Hence, the set $(\{v < 0\} \cup U_\varepsilon) \cap (O \cup U_\varepsilon) = (\{v < 0\} \cap O) \cup U_\varepsilon$ is open.

Let a function $g \in H_0^1(\Omega)$ with $g \geq 1$ a.e. on U_ε be given. Then, $g \geq 1$ a.e. on $(\{v < 0\} \cap O) \cup U_\varepsilon$, since $\{v < 0\} \cap O$ has measure zero. By the definition of the capacity, this implies (note that both involved sets are open and hence neighborhoods of themselves)

$$\operatorname{cap}((\{v < 0\} \cap O) \cup U_\varepsilon) \leq \operatorname{cap}(U_\varepsilon).$$

Using the monotonicity and subadditivity of the capacity, see Bonnans, Shapiro, 2000, Lem. 6.48, we obtain

$$\operatorname{cap}(\{v < 0\} \cap O) \leq \operatorname{cap}\big((\{v < 0\} \cap O) \cup U_\varepsilon\big) \leq \operatorname{cap}(U_\varepsilon) \leq 2\,\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we have

$$\operatorname{cap}(\{v < 0\} \cap O) = 0.$$

By applying this lemma to $O = \Omega$, we find that $v \geq 0$ a.e. (on Ω) is equivalent to $v \geq 0$ q.e. (on Ω).

Finally, we recall some results on the relation between non-negative functionals in $H^{-1}(\Omega)$ and capacity theory, see Bonnans, Shapiro, 2000, pp. 564–565.

Lemma 3.2.4. Let $\xi \in H^{-1}(\Omega)$ be a non-negative functional (i.e. ξ takes non-negative values on non-negative functions). Then, ξ can be identified with a regular Borel measure on Ω which is, in addition, finite on compact sets. Moreover, for every Borel set $D \subset \Omega$, $\operatorname{cap}(D) = 0$ implies $\xi(D) = 0$.

Finally, the quasi-continuous representative of every $v \in H_0^1(\Omega)$ is ξ-integrable and we have

$$\langle v,\, \xi \rangle_{H_0^1(\Omega), H^{-1}(\Omega)} = \int_\Omega v \,\mathrm{d}\xi. \tag{3.2.2}$$

Note that, in particular, $v \geq 0$ q.e. implies $v \geq 0$ ξ-a.e. for all non-negative $\xi \in H^{-1}(\Omega)$.

Now, we are able to give an expression for the normal cone of K at a point $y \in K$, see also Bonnans, Shapiro, 2000, Thm. 6.57 for the same result in the case $\psi = 0$.

Proposition 3.2.5. For $y \in K$ we have

$$\begin{aligned} \mathcal{T}_K(y) &= \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. on } y = \psi\}, \\ \mathcal{T}_K(y)^\circ &= \{\xi \in H^{-1}(\Omega) : \xi \text{ is non-negative and } y - \psi = 0 \ \xi\text{-a.e. on } \Omega.\} \end{aligned}$$

In particular, we have $\xi(\{x \in \Omega : y(x) < \psi(x)\}) = 0$ for $\xi \in \mathcal{T}_K(y)^\circ$.

Proof. The first identity is given in Mignot, 1976, Lem. 3.2.

Let us prove the second identity.

"$\subset$": Let $\xi \in \mathcal{T}_K(y)^\circ$ be given. We start by proving that ξ is non-negative. For $w \in H_0^1(\Omega)$ with $w \geq 0$ a.e. in Ω, we have $v_w := y - w \leq \psi$ a.e. in Ω. This implies $v_w \in K$ and, hence,

$$\langle \xi,\, w \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = \langle \xi,\, y - v_w \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \quad \text{for all } w \in H_0^1(\Omega) : w \geq 0 \text{ a.e. in } \Omega. \tag{3.2.3}$$

By $y \in K$, we have $y - \psi \leq 0$ a.e. in Ω and Lemma 3.2.3 implies $y - \psi \leq 0$ q.e. in Ω. After modification of y on a set of capacity zero, the sets $\{y = \psi\}$, $\{y < \psi\}$ and $\{y > \psi\}$ are Borel sets, see Lemma 3.2.2. Now, Lemma 3.2.4 implies $y - \psi \leq 0$ ξ-a.e. in Ω.

Now, let a smooth cut-off function $\chi \in C_0^\infty(\Omega)$ with $1 \geq \chi \geq 0$ and $\chi = 1$ on some compact $C \subset \Omega$ be given. By defining $w := \chi\,\psi + (1-\chi)\,y \leq \psi$ we obtain $w \in H_0^1(\Omega)$ and, in particular, $w \in K$. This yields

$$0 \geq \langle \xi,\, w - y \rangle = \int_\Omega \chi\,\psi + (1-\chi)\,y - y \,\mathrm{d}\xi = \int_\Omega \chi\,(\psi - y)\,\mathrm{d}\xi.$$

Since $\chi \geq 0$ everywhere and $\psi - y \geq 0$ ξ-a.e., we infer $\chi\,(y - \psi) = 0$ ξ-a.e., and in particular, $y - \psi = 0$ ξ-a.e. on C. Since Ω can be written as a countable union of compact sets and since ξ is countable additive, we have $y - \psi = 0$ ξ-a.e. on Ω. Finally,

$$\xi(\{y < \psi\}) \leq \xi(\{y \neq \psi\}) = 0.$$

"$\supset$": Let ξ be non-negative with $y - \psi = 0$ ξ-a.e. on Ω, hence $\xi(\{y \neq \psi\}) = 0$.

For arbitrary $v \in \mathcal{T}_K(y)$ we obtain

$$\langle v,\, \xi \rangle_{H_0^1(\Omega),H^{-1}(\Omega)} = \int_\Omega v \,\mathrm{d}\xi = \int_{\{y=\psi\}} v \,\mathrm{d}\xi + \int_{\{y\neq\psi\}} v \,\mathrm{d}\xi \leq 0$$

since $\xi(\{y \neq \psi\}) = 0$ and $v \leq 0$ q.e. on $\{y = \psi\}$ implies $v \leq 0$ ξ-a.e. on $\{y = \psi\}$.

3.3. Linearization of the problem

We denote by $S : H^{-1}(\Omega) \to H_0^1(\Omega), u \mapsto y$ the solution operator of the variational inequality (VI)

$$\text{Find } y \in K, \quad \text{such that } \langle \mathcal{A}\,y - u - f,\, v - y \rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \quad \text{for all } v \in K.$$

The unique solvability of this VI is well-known and follows from Kinderlehrer, Stampacchia, 1980, Thm. II.2.1. It is known, that for given (y, u), there exists $\xi \in H^{-1}(\Omega)$ such that (y, u, ξ) is feasible for **(P)** if and only if $y = S(u)$ and $u \in U_{\mathrm{ad}}$.

Since the obstacle $\psi \in H^1(\Omega)$ has a quasi-continuous representative, we can apply Mignot, 1976, Thm. 3.2 to infer the polyhedricity of K. Hence, Mignot, 1976, Thm. 2.1 yields the directional differentiability of S. Using the Lipschitz continuity of S, we find that S is even Hadamard-differentiable by Shapiro, 1990, Prop. 3.5, see also Bonnans, Shapiro, 2000, Thm. 6.58 for a similar argument in the case $\psi = 0$. The derivative $S'(\bar{u}; h)$ in the direction $h \in H^{-1}(\Omega)$ is the solution of the VI

$$\text{Find } y_h \in \mathcal{K}(\bar{u}), \quad \text{such that } \langle \mathcal{A}\,y_h - h,\, v - y_h \rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \text{ for all } v \in \mathcal{K}(\bar{u}), \quad (3.3.1)$$

where

$$\mathcal{K}(\bar{u}) := \mathcal{T}_K(\bar{y}) \cap \bar{\xi}^{\perp} = \{y_h \in H_0^1(\Omega) : y_h \leq 0 \text{ q.e. in } A \text{ and } \langle y_h, \bar{\xi}\rangle_{H_0^1, H^{-1}} = 0\}, \quad (3.3.2)$$

see Mignot, 1976, Lem. 3.2. The VI (3.3.1) is equivalent to the complementarity system

$$\mathcal{A}\, y_h - h + \xi_h = 0, \quad (3.3.3a)$$
$$y_h \in \mathcal{K}(\bar{u}), \quad (3.3.3b)$$
$$\xi_h \in \mathcal{K}(\bar{u})^{\circ}, \quad (3.3.3c)$$
$$\langle \xi_h, y_h\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = 0. \quad (3.3.3d)$$

The following lemma provides a useful characterization of the closed convex cone $\mathcal{K}(\bar{u})$ in terms of q.e.-(in)equalities.

Lemma 3.3.1. Let $\bar{u} \in H^{-1}(\Omega)$ be given and denote $\bar{y} = S(\bar{u})$, $\bar{\xi} = \bar{u} - \mathcal{A}\,\bar{y} + f$. Then, there exists a set $\tilde{A}_s$, such that $\tilde{A}_s \subset A$ and

$$\begin{aligned}\mathcal{K}(\bar{u}) &= \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. in } A \text{ and } v = 0 \text{ q.e. in } \tilde{A}_s\} \\ &= \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. in } \tilde{B} \text{ and } v = 0 \text{ q.e. in } \tilde{A}_s\},\end{aligned} \quad (3.3.4)$$

where $\tilde{B} := A \setminus \tilde{A}_s$ is the biactive set. In particular, we could choose $\tilde{A}_s$ to be quasi-closed.

Proof. Since $\bar{\xi}$ is a non-negative functional on $H_0^1(\Omega)$, we identify it with a regular Borel measure, see Lemma 3.2.4. Now, let $v \in H_0^1(\Omega)$ be given, satisfying $v \leq 0$ q.e. on A. Lemma 3.2.4 implies $v \leq 0$ $\bar{\xi}$-a.e. on A. Using (3.2.2), we have

$$\langle v, \bar{\xi}\rangle_{H_0^1(\Omega), H^{-1}(\Omega)} = \int_{\Omega} v \,\mathrm{d}\bar{\xi} = \int_A v \,\mathrm{d}\bar{\xi},$$

since $\bar{\xi}(\Omega \setminus A) = 0$, see Proposition 3.2.5. By using $v \leq 0$ $\bar{\xi}$-a.e. on A,

$$\langle v, \bar{\xi}\rangle_{H_0^1(\Omega), H^{-1}(\Omega)} = \int_A v \,\mathrm{d}\bar{\xi} = 0$$

is equivalent to $v = 0$ $\bar{\xi}$-a.e. on A. Since $\bar{\xi}(\Omega \setminus A) = 0$, this is in turn equivalent to $v = 0$ $\bar{\xi}$-a.e. on Ω.

The above reasoning shows

$$\mathcal{K}(\bar{u}) = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. in } A \text{ and } v = 0 \ \bar{\xi}\text{-a.e.}\}, \quad (3.3.5)$$

compare (3.3.2). Finally, Stollmann, 1993, Thm. 1 implies the existence of a quasi-closed set $\tilde{A}_s$, such that

$$\{v \in H_0^1(\Omega) : v = 0 \ \bar{\xi}\text{-a.e.}\} = \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. in } \tilde{A}_s\}.$$

It remains to show $\tilde{A}_s \subset A$. Since $\bar{y} - \psi = 0$ $\bar{\xi}$-a.e., we have $\bar{y} - \psi = 0$ q.e. on $\tilde{A}_s$, hence $\mathrm{cap}(\tilde{A}_s \setminus A) = 0$. Replacing $\tilde{A}_s$ by $\tilde{A}_s \cap A$ yields the claim.

Note that we give a more explicit characterization of the strictly active set $\tilde{A}_s$ in Section 3.A, see in particular Lemma 3.A.5. Moreover, we do not denote the strictly active set by A_s in order to remind the reader that our definition of it differs from the usual definition in the literature.

We consider the reduced formulation of (**P**)

$$\begin{aligned} \text{Minimize} \quad & j(S(u)) + \frac{\alpha}{2} \|u\|^2_{L^2(\Omega)} \\ \text{and} \quad & u \in U_{\text{ad}}. \end{aligned} \tag{$\mathbf{P}_{\text{red}}$}$$

Due to the continuity of S, $(\bar{y}, \bar{u}, \bar{\xi})$ is a local solution of (**P**) if and only if $\bar{u}$ is a local solution of ($\mathbf{P}_{\text{red}}$). The local optimality of $\bar{u}$ for ($\mathbf{P}_{\text{red}}$) and the Hadamard-differentiability of S imply that $h = 0$ is a global solution of the "linearized" problem

$$\begin{aligned} \text{Minimize} \quad & j'(S(\bar{u}))\, S'(\bar{u}; h) + \alpha\, (\bar{u},\, h)_{L^2(\Omega)} \\ \text{such that} \quad & h \in \mathcal{T}_{U_{\text{ad}}}(\bar{u}), \end{aligned} \tag{$\mathbf{P}^{\text{lin}}_{\text{red}}$}$$

where $\mathcal{T}_{U_{\text{ad}}}(\bar{u}) \subset L^2(\Omega)$ is the tangent cone of U_{ad}.

In the sequel, we will consider different restrictions of this linearized problem in order to prove properties of the minimizer $\bar{u}$, see Section 3.4, and to prove the strong stationarity of $\bar{u}$ in Section 3.5.

3.4. Properties of local solutions

We are going to prove properties of a local minimizer $\bar{u}$ by evaluating optimality conditions of a certain restriction of ($\mathbf{P}^{\text{lin}}_{\text{red}}$). From now on, we assume the additional regularity $\bar{u} \in H^1_0(\Omega)$. This property can be shown by penalization arguments, see, e.g., Mignot, Puel, 1984; Schiela, D. Wachsmuth, 2013, or by limiting variational calculus, see Hintermüller, Mordukhovich, Surowiec, 2014, Rem. 1. In order to keep the presentation simple, we just *assume* this regularity, keeping in mind that it can be achieved under rather mild assumptions on the data, in particular one uses $u_a, u_b \in H^1_0(\Omega)$.

We interpret $L^2(\Omega)$ as an subspace of $H^{-1}(\Omega)$ via the canonical embedding $E : L^2(\Omega) \to H^{-1}(\Omega)$, $h \mapsto (v \mapsto \int_\Omega h\, v \, \mathrm{d}x)$. Up to now we did not mention this embedding in favor of a clearer presentation. In order to get sharper optimality conditions for ($\mathbf{P}^{\text{lin}}_{\text{red}}$) we are going to enlarge the feasible set. The closeness of the constraint set in $H^{-1}(\Omega)$ in (3.4.1) will be crucial for rewriting (3.4.5) into (3.4.6).

Lemma 3.4.1. The functional $h = 0 \in H^{-1}(\Omega)$ is a global minimizer of

$$\begin{aligned} \text{Minimize} \quad & j'(S(\bar{u}))\, S'(\bar{u}; h) + \alpha\, \langle \bar{u},\, h \rangle_{H^1_0(\Omega, H^{-1}(\Omega))} \\ \text{such that} \quad & h \in \overline{E\mathcal{T}_{U_{\text{ad}}}(\bar{u})}^{H^{-1}(\Omega)}. \end{aligned} \tag{3.4.1}$$

Proof. By using the canonical embedding $E : L^2(\Omega) \to H^{-1}(\Omega)$ in ($\mathbf{P}^{\mathrm{lin}}_{\mathrm{red}}$), we obtain that $h = 0 \in H^{-1}(\Omega)$ is a global solution of

$$\begin{aligned} &\text{Minimize} && j'(S(\bar{u}))\, S'(\bar{u}; h) + \alpha \, \langle \bar{u},\, h \rangle_{H^1_0(\Omega), H^{-1}(\Omega))} \\ &\text{such that} && h \in E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}). \end{aligned} \tag{3.4.2}$$

We proceed by contradiction and assume that 0 is not a global solution of (3.4.1). This yields the existence of $h \in \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}$ with

$$j'(S(\bar{u}))\, S'(\bar{u}; h) + \alpha \, \langle \bar{u},\, h \rangle_{H^1_0(\Omega, H^{-1}(\Omega))} < 0.$$

Since $E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})$ is dense in $\overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}$ and since the objective in (3.4.1) is continuous w.r.t. $h \in H^{-1}(\Omega)$, this gives the existence of $\tilde{h} \in E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})$ with

$$j'(S(\bar{u}))\, S'(\bar{u}; \tilde{h}) + \alpha \, \langle \bar{u},\, \tilde{h} \rangle_{H^1_0(\Omega, H^{-1}(\Omega))} < 0.$$

This, however, is a contradiction to the fact that $h = 0$ is a global minimizer of (3.4.2).

Using the equivalent reformulation (3.3.3) of the linearized VI (3.3.1), the problem (3.4.1) can be written as

$$\begin{aligned} &\text{Minimize} && j'(S(\bar{u}))\, y_h + \alpha \, \langle \bar{u},\, h \rangle_{H^1_0(\Omega), H^{-1}(\Omega)} \\ &\text{with respect to} && (y_h, h, \xi_h) \in H^1_0(\Omega) \times H^1_0(\Omega) \times H^{-1}(\Omega), \\ &\text{such that} && \mathcal{A}\, y_h - h + \xi_h = 0, \\ & && y_h \in \mathcal{K}(\bar{u}), \\ & && \xi_h \in \mathcal{K}(\bar{u})^\circ, \\ & && \langle y_h,\, \xi_h \rangle_{H^1_0(\Omega), H^{-1}(\Omega)} = 0, \\ & && h \in \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}. \end{aligned} \tag{3.4.3}$$

Since $h = 0$ is a global minimizer of (3.4.1), $(y_h, h, \xi_h) = (0, 0, 0)$ is a global minimizer of (3.4.3). Note that (3.4.3) still contains a complementarity constraint (as long as $\mathcal{K}(\bar{u})$ is not a subspace). By restricting y_h to zero, we obtain that $(h, \xi_h) = (0, 0)$ is a global solution of the auxiliary problem

$$\begin{aligned} &\text{Minimize} && \alpha \, \langle \bar{u},\, h \rangle_{H^1_0(\Omega), H^{-1}(\Omega)} \\ &\text{with respect to} && (h, \xi_h) \in H^{-1}(\Omega) \times H^{-1}(\Omega), \\ &\text{such that} && -h + \xi_h = 0, \\ & && \xi_h \in \mathcal{K}(\bar{u})^\circ, \\ & && h \in \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}. \end{aligned} \tag{3.4.4}$$

Making use of the constraint $h = \xi_h$, $h = 0$ is a global solution of

$$\begin{aligned} &\text{Minimize} && \alpha \langle \bar{u},\, h\rangle_{H_0^1(\Omega),H^{-1}(\Omega)} \\ &\text{such that} && h \in \mathcal{K}(\bar{u})^\circ \cap \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}. \end{aligned} \tag{3.4.5}$$

The optimality condition reads (note that this requires no constraint qualification)

$$\alpha\, \bar{u} \in -\Big[\mathcal{K}(\bar{u})^\circ \cap \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}\Big]^\circ,$$

where the polar cones are to be evaluated w.r.t. the $H^{-1}(\Omega)$-$H_0^1(\Omega)$ duality. By using $(K_1 \cap K_2)^\circ = \overline{K_1^\circ + K_2^\circ}$ for closed, convex cones K_1, K_2 in a reflexive Banach space, see, e.g., Bonnans, Shapiro, 2000, (2.32), we obtain

$$\Big[\mathcal{K}(\bar{u})^\circ \cap \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}\Big]^\circ = \overline{\mathcal{K}(\bar{u}) + \Big(\overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}\Big)^\circ}^{H_0^1(\Omega)}.$$

Since $A^\circ = (\,\overline{A}\,)^\circ$ holds for all sets A, we get

$$\alpha\, \bar{u} \in -\Big[\mathcal{K}(\bar{u})^\circ \cap \overline{E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{H^{-1}(\Omega)}\Big]^\circ = -\overline{\mathcal{K}(\bar{u}) + (E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ}^{H_0^1(\Omega)}. \tag{3.4.6}$$

It remains to evaluate the right-hand side.

Lemma 3.4.2. The polar cone of $E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})$ w.r.t. the $H^{-1}(\Omega)$-$H_0^1(\Omega)$ duality is given by

$$\begin{aligned} (E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ = \{v \in H_0^1(\Omega) : {} & v \le 0 \text{ q.e. on } A_a,\ v \ge 0 \text{ q.e. on } A_b, \text{ and} \\ & v = 0 \text{ q.e. on } \Omega \setminus (A_a \cup A_b)\}. \end{aligned}$$

Proof. A simple calculation shows

$$\begin{aligned} (E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ &= \{v \in H_0^1(\Omega) : \langle v,\, h\rangle_{H_0^1(\Omega),H^{-1}(\Omega)} \le 0 \text{ for all } h \in E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})\} \\ &= \{v \in H_0^1(\Omega) : \int_\Omega v\, u \,\mathrm{d}x \le 0 \text{ for all } u \in \mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})\} \\ &= \{v \in H_0^1(\Omega) : v \le 0 \text{ a.e. on } A_a,\ v \ge 0 \text{ a.e. on } A_b, \text{ and} \\ &\qquad\qquad v = 0 \text{ a.e. on } \Omega \setminus (A_a \cup A_b)\}. \end{aligned}$$

Now, the inclusion "$\supset$" of the assertion follows easily, since $v \le 0$ q.e. on A_a implies $v \le 0$ a.e. on A_a, and analogous arguments for the other conditions.

Let $v \in (E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ$ be given. By the above calculation, we have $v \ge 0$ a.e. on $\Omega \setminus A_a$. The set $\Omega \setminus A_a = \{x \in \Omega : \bar{u}(x) < u_a(x)\}$ is quasi-open. Using Lemma 3.2.3, we find that $v \ge 0$ a.e. on $\Omega \setminus A_a$ implies $v \ge 0$ q.e. on $\Omega \setminus A_a$ and, in particular, $v \ge 0$ q.e. on A_b. Similarly, we obtain $v \le 0$ q.e. on $\Omega \setminus A_b \supset A_a$ and $v = 0$ q.e. on $\Omega \setminus (A_a \cup A_b)$. This shows the claim.

Now, we obtain the announced properties of the local minimizer $\bar{u}$. We recall the definition of the biactive set $\tilde{B} = A \setminus \tilde{A}_s$ from Lemma 3.3.1.

Lemma 3.4.3. If $\bar{u}$ belongs to $H_0^1(\Omega)$, we have the sign conditions

$$\begin{aligned} \bar{u} &= 0 \quad \text{q.e. on } \tilde{A}_s \cap [\Omega \setminus (A_a \cup A_b)], \\ \bar{u} &\le 0 \quad \text{q.e. on } \tilde{A}_s \cap A_b, \\ \bar{u} &\ge 0 \quad \text{q.e. on } (\tilde{A}_s \cap A_a) \cup [\tilde{B} \cap (\Omega \setminus A_b)]. \end{aligned}$$

In particular, $u_b \ge 0$ q.e. on A and $u_a \le 0$ q.e. on $\tilde{A}_s$ imply $\bar{u} \ge 0$ q.e. on $\tilde{B}$ and $\bar{u} = 0$ q.e. on $\tilde{A}_s$.

Proof. By using (3.4.6), there are sequences $\{v_1^{(i)}\} \subset -\mathcal{K}(\bar{u})$ and $\{v_2^{(i)}\} \subset -(E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ$, such that

$$\bar{u} = \lim_{i\to\infty} (v_1^{(i)} + v_2^{(i)}) \quad \text{in } H_0^1(\Omega).$$

After passing to a subsequence, we have the pointwise convergence

$$\bar{u} = \lim_{i\to\infty} (v_1^{(i)} + v_2^{(i)}) \quad \text{q.e. in } \Omega, \tag{3.4.7}$$

see Bonnans, Shapiro, 2000, Lem. 6.52. By using Lemma 3.3.1 we know

$$v_1^{(i)} \in -\mathcal{K}(\bar{u}) = \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. in } \tilde{B} \text{ and } v = 0 \text{ q.e. in } \tilde{A}_s\}$$

and by Lemma 3.4.2 we have

$$-v_2^{(i)} = -(E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ = \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } A_a,\ v \le 0 \text{ q.e. on } A_b, \text{ and } v = 0 \text{ q.e. on } \Omega \setminus (A_a \cup A_b)\}.$$

That is, we have the following q.e. sign conditions:

$$\begin{aligned} &\text{on } \tilde{A}_s \cap [\Omega \setminus (A_a \cup A_b)] : && v_1^{(i)} = 0 \quad \text{and} \quad v_2^{(i)} = 0, \\ &\text{on } \tilde{A}_s \cap A_b : && v_1^{(i)} = 0 \quad \text{and} \quad v_2^{(i)} \le 0, \\ &\text{on } \tilde{A}_s \cap A_a : && v_1^{(i)} = 0 \quad \text{and} \quad v_2^{(i)} \ge 0, \\ &\text{on } \tilde{B} \cap (\Omega \setminus A_b) : && v_1^{(i)} \ge 0 \quad \text{and} \quad v_2^{(i)} \ge 0. \end{aligned}$$

Together with (3.4.7), this gives the desired sign conditions of $\bar{u}$.

3.5. Strong stationarity

We use the results of the previous section together with the KKT conditions of a restriction of $(\mathbf{P}_{\mathrm{red}}^{\mathrm{lin}})$ in order to prove necessity of the strong stationarity system (3.1.3). In addition

to $\bar u \in H_0^1(\Omega)$, we assume

$$u_b \geq 0 \quad \text{q.e. in } \tilde B, \tag{3.5.1a}$$

$$\operatorname{cap}(A_a \cap \tilde B) = 0, \tag{3.5.1b}$$

$$\bar u = 0 \quad \text{q.e. on } \tilde A_s. \tag{3.5.1c}$$

We refer to Lemma 3.5.3 for a simple condition which implies that this assumption is satisfied.

We start by restating $(\mathbf{P}^{\mathrm{lin}}_{\mathrm{red}})$ in a subspace of $H_0^1(\Omega)$. Therefore, we recall the characterization

$$\mathcal{K}(\bar u) = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. in } \tilde B \text{ and } v = 0 \text{ q.e. in } \tilde A_s\}$$

from Lemma 3.3.1. We define

$$V = \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. in } \tilde A_s\}. \tag{3.5.2}$$

Note that the subspace V is closed, since sequences converging in $H_0^1(\Omega)$ contain a pointwise quasi-everywhere convergent subsequence, see Bonnans, Shapiro, 2000, Lem. 6.52. Since $\mathcal{K}(\bar u)$ is a subset of the closed subspace V, we can restate the VI (3.3.1) characterizing the derivative $S'(\bar u; h)$ in the space V. To this end, we introduce the canonical injection

$$I : V \to H_0^1(\Omega),\ v \mapsto v. \tag{3.5.3}$$

The action of its adjoint $I^\star : H^{-1}(\Omega) \to V^\star$ is the restriction of the domain of a functional from $H_0^1(\Omega)$ to V. Further, we introduce the bounded, linear operator

$$\mathcal{A}_V = I^\star \mathcal{A}\, I : V \to V^\star,$$

which inherits the ellipticity from $\mathcal{A}$, and the closed convex cone

$$\mathcal{K}_V(\bar u) = \{v \in V : v \leq 0 \text{ q.e. in } \tilde B\}.$$

Note that $\mathcal{K}(\bar u) = I\,\mathcal{K}_V(\bar u)$. Let us recall the VI (3.3.1) characterizing the derivative $S'(\bar u; h)$ of S in direction $h \in H^{-1}$

$$\text{Find } y_h \in \mathcal{K}(\bar u), \quad \text{such that } \langle \mathcal{A}\, y_h - h,\, v - y_h\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 0 \text{ for all } v \in \mathcal{K}(\bar u).$$

We define $S'_V(\bar u; h)$ for $h \in V^\star$ as the unique solution of

$$\text{Find } y_h \in \mathcal{K}_V(\bar u), \quad \text{such that } \langle \mathcal{A}_V\, y_h - h,\, v - y_h\rangle_{V^\star, V} \geq 0 \text{ for all } v \in \mathcal{K}_V(\bar u), \tag{3.5.4}$$

see Kinderlehrer, Stampacchia, 1980, Thm. II.2.1 for the unique solvability. An immediate consequence is

$$S'(\bar u; h) = I\, S'_V(\bar u; I^\star h)$$

for all $h \in H^{-1}(\Omega)$. Using this equivalence and $\bar u \in V$ by (3.5.1c), we obtain from $(\mathbf{P}^{\mathrm{lin}}_{\mathrm{red}})$ that $h = 0 \in V^\star$ is a global solution of

$$\begin{aligned} &\text{Minimize} && j'(S(\bar u))\, I\, S'_V(\bar u; h) + \alpha\, \langle \bar u,\, h\rangle_{V, V^\star} \\ &\text{such that} && h \in I^\star E \mathcal{T}_{U_{\mathrm{ad}}}(\bar u). \end{aligned} \tag{3.5.5}$$

Arguing similarly as in Lemma 3.4.1, we obtain that $h = 0 \in V^\star$ is a global solution of

$$\begin{aligned} &\text{Minimize} && j'(S(\bar u))\, I\, S_V'(\bar u; h) + \alpha\, \langle \bar u,\, h\rangle_{V,V^\star} \\ &\text{such that} && h \in \overline{I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u)}^{V^\star}. \end{aligned} \tag{3.5.6}$$

Similar to (3.3.3), we can rewrite the VI (3.5.4) as a complementarity system and obtain that $(y_h, h, \xi_h) = (0, 0, 0)$ is a global solution of

$$\begin{aligned} &\text{Minimize} && j'(S(\bar u))\, I\, y_h + \alpha\, \langle \bar u,\, h\rangle_{V,V^\star} \\ &\text{such that} && \mathcal{A}_V\, y_h - h + \xi_h = 0, \\ &&& y_h \in \mathcal{K}_V(\bar u), \\ &&& \xi_h \in \mathcal{K}_V(\bar u)^\circ, \\ &&& \langle y_h,\, \xi_h\rangle_{V,V^\star} = 0, \\ &&& h \in \overline{I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u)}^{V^\star}. \end{aligned} \tag{3.5.7}$$

We restrict the slack variable ξ_h to 0. This enables us to drop the complementarity condition. We obtain that $(h, y_h) = (0, 0)$ is a global solution of the auxiliary problem

$$\begin{aligned} &\text{Minimize} && j'(S(\bar u))\, I\, y_h + \alpha\, \langle \bar u,\, h\rangle_{V,V^\star} \\ &\text{such that} && \mathcal{A}_V\, y_h - h = 0, \\ &&& y_h \in \mathcal{K}_V(\bar u), \\ &&& h \in \overline{I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u)}^{V^\star}. \end{aligned} \tag{3.5.8}$$

Due to this restriction of ξ_h, the optimality system (3.5.10) of the problem (3.5.8) will not contain any information of p on $\tilde B$. However, this information can be recovered by the gradient equation (3.5.10b) and the signs of $\bar u$ from Lemma 3.4.3. Note that this relies heavily on the fact that the control lives on the same domain as the constraint $y \le \psi$.

Lemma 3.5.1. The polar cone of $I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u) \subset V^\star$ is given by

$$\begin{aligned} (I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u))^\circ = \{v \in V : v \le 0 \text{ q.e. on } A_a,\ v \ge 0 \text{ q.e. on } A_b, \text{ and} \\ v = 0 \text{ q.e. on } \Omega \setminus (A_a \cup A_b)\}. \end{aligned} \tag{3.5.9}$$

Proof. A simple calculation, see also Aubin, Frankowska, 2009, Lem. 2.4.3, shows

$$(I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u))^\circ = I^{-1}(E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u))^\circ,$$

where the right-hand side denotes the preimage of $(E\mathcal{T}_{U_{\mathrm{ad}}}(\bar u))^\circ$ w.r.t. the injection $I : V \to H_0^1(\Omega)$. Now, Lemma 3.4.2 yields the claim.

We show that the CQ of Robinson-Zowe-Kurcyusz is satisfied at the solution $(y_h, h) = (0,0)$ of (3.5.8). Let an arbitrary $\mu \in V^\star$ be given. We have to show the existence of $y_h \in \mathcal{K}_V(\bar{u})$, $h \in \overline{I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{V^\star}$, such that $\mathcal{A}_V\, y_h - h = \mu$. We set $y_h = S_V'(\bar{u}, \mu) \in \mathcal{K}_V(\bar{u})$. Then, there exists $h \in -\mathcal{K}_V(\bar{u})^\circ$ such that

$$\mathcal{A}\, y_h - h = \mu$$

and $\langle y_h,\, h\rangle_{V,V^\star} = 0$ (we do not use this condition) are satisfied. Note that we have

$$\begin{aligned}(I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ &\subset \{v \in V : v \le 0 \text{ q.e. on } A_a,\ v \ge 0 \text{ q.e. on } \Omega \setminus A_a\}\\ &\subset \{v \in V : v \ge 0 \text{ q.e. on } \tilde{B}\} = -\mathcal{K}_V(\bar{u})\end{aligned}$$

by (3.5.1b) and the definition of $\mathcal{K}_V(\bar{u})$. Hence, $h \in -\mathcal{K}_V(\bar{u})^\circ \subset (I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^{\circ\circ} = \overline{I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{V^\star}$. This shows that the CQ of Robinson-Zowe-Kurcyusz is satisfied by the problem (3.5.8).

Hence, there exists multipliers $(p, \tilde{\mu}, \nu) \in V \times V^\star \times V$ satisfying the optimality system

$$\mathcal{A}_V^\star\, p + I^\star\, j'(S(\bar{u})) + \tilde{\mu} = 0 \text{ in } V^\star, \tag{3.5.10a}$$

$$\alpha\, \bar{u} - p + \nu = 0 \text{ in } V, \tag{3.5.10b}$$

$$\tilde{\mu} \in \{y \in V : y \le 0 \text{ q.e. in } \tilde{B}\}^\circ, \tag{3.5.10c}$$

$$\nu \in (\overline{I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})}^{V^\star})^\circ = (I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ. \tag{3.5.10d}$$

Now we show that the system (3.1.3) is satisfied, where $\mu \in H^{-1}(\Omega)$ is defined by

$$\mu = -\mathcal{A}^\star p - j'(S(\bar{u})).$$

Due to this definition of μ, (3.1.3a) holds. The gradient equation (3.1.3b) follows from (3.5.10b), since $\bar{u}$, p, and ν are zero on $\tilde{A}_s$.

By definition of p, we have $p = 0$ q.e. on $\tilde{A}_s$. By the gradient equation (3.5.10b), we obtain

$$p = \alpha\, \bar{u} + \nu \ge \alpha\, \bar{u} \ge 0 \quad \text{q.e. on } \tilde{B}.$$

The first inequality follows from (3.5.1b) and (3.5.9), whereas the second one follows from (3.5.1a) and Lemma 3.4.3. Hence, (3.1.3c') is satisfied.

In order to show the sign condition (3.1.3d') on μ, let $v \in H_0^1(\Omega)$, $v \le 0$ q.e. on $\tilde{B}$ and $v = 0$ q.e. on $\tilde{A}_s$ be given. Using $v \in V$, we obtain from the definition of μ and (3.5.10c)

$$\begin{aligned}\langle \mu,\, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} &= \langle -\mathcal{A}^\star p - j'(S(\bar{u})),\, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}\\ &= \langle -\mathcal{A}_V^\star p - I^\star\, j'(S(\bar{u})),\, v\rangle_{V^\star,V} = \langle \tilde{\mu},\, v\rangle_{V^\star,V} \le 0.\end{aligned}$$

This is the desired sign condition on μ.

Since $V \subset L^2(\Omega)$, we have (the following inequalities are to be understood in the a.e.-sense)

$$\begin{aligned}
&\nu \in (I^\star E\mathcal{T}_{U_{\mathrm{ad}}}(\bar{u}))^\circ \\
&\quad = \{v \in V \quad : v \le 0 \text{ on } A_a,\ v \ge 0 \text{ on } A_b, \text{ and } v = 0 \text{ on } \Omega \setminus (A_a \cup A_b)\} \\
&\quad \subset \{v \in L^2(\Omega) : v \le 0 \text{ on } A_a,\ v \ge 0 \text{ on } A_b, \text{ and } v = 0 \text{ on } \Omega \setminus (A_a \cup A_b)\} \\
&\quad = \mathcal{N}_{U_{\mathrm{ad}}}(\bar{u}),
\end{aligned}$$

which is the sign condition (3.1.3e) on ν.

Altogether, we have proven the following theorem.

Theorem 3.5.2. Let $(\bar{y}, \bar{u}, \bar{\xi}) \in H_0^1(\Omega) \times H_0^1(\Omega) \times H^{-1}(\Omega)$ be a local solution of (**P**), such that (3.5.1) holds. Then, there exist multipliers $(p, \mu, \nu) \in H_0^1(\Omega) \times H^{-1}(\Omega) \times H_0^1(\Omega)$, such that the strong stationarity conditions (3.1.3) are satisfied.

Note that the uniqueness of multipliers does not simply follow from (3.1.3) (as in the case without control constraints).

The arguments leading to Theorem 3.5.2 remain valid in the cases $u_a = -\infty$ or $u_b = +\infty$, with the obvious modifications.

As announced, we remark that assumption (3.5.1) is implied by a simple assumption on the control bounds, which can be checked a-priori.

Lemma 3.5.3. If the bounds $u_a, u_b \in H^1(\Omega)$ fulfill

$$u_a < 0 \le u_b \text{ q.e. in } \Omega, \tag{3.5.11}$$

then (3.5.1) is satisfied.

Note that we do not need to assume that u_a is uniformly negative in (3.5.11).

Proof. It is clear that (3.5.1a) holds. Lemma 3.4.3 implies $\bar{u} = 0$ q.e. on $\tilde{A}_s$, i.e. (3.5.1c), and $\bar{u} \ge 0$ q.e. on A. Hence, we have $\bar{u} = 0 > u_a$ q.e. on A. This shows (3.5.1b).

Finally, we give a remark on the condition (3.5.1c). By inspecting the calculation leading to Theorem 3.5.2, we find that this assumption could be replaced by the following weaker one: assume that

$$\tilde{u} = \begin{cases} 0 & \text{on } \tilde{A}_s \\ \bar{u} & \text{on } \Omega \setminus \tilde{A}_s \end{cases} \text{ belongs to } H_0^1(\Omega). \tag{3.5.12}$$

Moreover, we could drop the assumption $\bar{u} = 0$ on $\tilde{A}_s$ if we could discuss an auxiliary problem similar to (3.5.8) directly in $H^{-1}(\Omega) \times H_0^1(\Omega)$. However, we were not able to provide a CQ for such an auxiliary problem.

3.6. Counterexamples

In this section we present two counterexamples, which show that strong stationarity may not hold if $u_a < 0$ or $u_b \geq 0$ are violated. Note that we do not have a counterexample if $\bar{u} = 0$ on $\tilde{A}_s$ is violated. In both examples, the domain is $\Omega = (0,1)$ and $\mathcal{A} = -\Delta$, i.e., $\mathcal{A}\, y = -y''$.

3.6.1. The lower bound is zero and active

This counterexample, which was constructed by the author, can already be found in Schiela, D. Wachsmuth, 2013. We consider

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|y+1\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \\ \text{such that} \quad & \mathcal{A}\, y = u - \xi, \\ & 0 \geq y \ \perp \ \xi \geq 0, \\ \text{and} \quad & u \geq 0. \end{aligned}$$

For all feasible u, the solution of the complementarity system is $(y, \xi) = (0, u)$. Hence, the unique global solution of this problem is $(\bar{y}, \bar{u}, \bar{\xi}) = (0, 0, 0)$. Using $A = A_a = \Omega$ and $\tilde{A}_s = \emptyset$, the system of strong stationarity (3.1.3) reads

$$\begin{aligned} \mathcal{A}\, p + 1 + \mu &= 0 && \text{in } H^{-1}(\Omega), \\ -p + \nu &= 0 && \text{a.e. in } \Omega, \\ p &\geq 0 && \text{q.e. in } \Omega, \\ \langle \mu,\, v \rangle_{H^{-1}(\Omega), H^1_0(\Omega)} &\geq 0 && \text{for all } v \in H^1_0(\Omega), v \geq 0 \text{ q.e. in } \Omega, \\ \nu &\leq 0 && \text{a.e. in } \Omega. \end{aligned}$$

This directly implies $p = \nu = 0$ and $\mu = -1$, which is a contradiction.

3.6.2. The upper bound is negative

We consider

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|y+1\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \\ \text{such that} \quad & \mathcal{A}\, y = u - \xi + 1, \\ & 0 \geq y \ \perp \ \xi \geq 0, \\ \text{and} \quad & u \leq -1. \end{aligned}$$

Since $u - \xi + 1 \leq 0$ for all admissible controls u and all multipliers $\xi \geq 0$, $0 \geq y$ is satisfied trivially by the maximum principle. Since ξ is unique, $\xi = 0$ follows for all admissible u.

Hence, the problem is equivalent to the control constrained problem

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|y+1\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \\ \text{such that} \quad & \mathcal{A}y = u+1, \\ \text{and} \quad & u \le -1. \end{aligned}$$

In the case $\alpha \ge 1/8$, $(\bar{y},\bar{u}) = (0,-1)$ is the unique global solution (this can be proven by checking the first order necessary and sufficient conditions) and hence the solution of the original problem. Then, we have $A = A_b = \Omega$ and $\tilde{A}_s = \emptyset$. However, there are no multipliers p, μ, ν, such that the strong stationarity system (3.1.3)

$$\begin{aligned} \mathcal{A}p + 1 + \mu &= 0 \quad \text{in } H^{-1}(\Omega), \\ \alpha\,\bar{u} - p + \nu &= 0 \quad \text{a.e. in } \Omega, \\ p &\ge 0 \quad \text{q.e. in } \Omega, \\ \langle \mu,\, v\rangle_{H^{-1}(\Omega),H^1_0(\Omega)} &\ge 0 \quad \text{for all } v \in H^1_0(\Omega), v \ge 0 \text{ q.e. in } \Omega, \\ \nu &\ge 0 \quad \text{a.e. in } \Omega \end{aligned}$$

is satisfied, since if p satisfies the first equation with some $\mu \ge 0$, we have $p < 0$ by the maximum principle.

3.A. Discussion of the strictly active set

The aim of this section is to show that the strictly active set $\tilde{A}_s$ defined in Lemma 3.3.1 can chosen to be the fine support (to be defined, see Lemma 3.A.4) of $\bar{\xi}$, in contrast to the implicit definition in the proof of Lemma 3.3.1.

In order to use some results from the literature, we have to define a capacity for arbitrary sets $A \subset \mathbb{R}^n$ by

$$\operatorname{cap}_{\mathbb{R}^n}(A) = \inf\{\|(v,\nabla v)\|^2_{L^2(\mathbb{R}^n)^{n+1}} : v \in H^1(\mathbb{R}^n), v \ge 1 \text{ a.e. in a neighbourhood of } A\},$$

compare Heinonen, Kilpeläinen, Martio, 1993, Sec. 2.35. Note that there are two differences to Definition 3.2.1: $H^1_0(\Omega)$ is replaced by $H^1(\mathbb{R}^n)$ and we use a different norm. Following the proof of Attouch, Buttazzo, Michaille, 2006, Prop.5.8.3 (a), we find that this definition is equivalent to Adams, Hedberg, 1996, Def. 2.2.1–2.2.4.

For sets $A \subset \Omega$, $\operatorname{cap}_{\mathbb{R}^n}(A)$ can be estimated from above by $\operatorname{cap}(A)$:

Lemma 3.A.1. There exists a constant $C > 0$, such that

$$\operatorname{cap}_{\mathbb{R}^n}(A) \le C \operatorname{cap}(A) \tag{3.A.1}$$

holds for all $A \subset \Omega$.

Proof. Let a function $v \in H_0^1(\Omega)$ satisfying $v \geq 1$ in a neighbourhood of A be given. Then, $v \in H^1(\mathbb{R}^n)$ and

$$\|(v, \nabla v)\|^2_{L^2(\mathbb{R}^n)^{n+1}} \leq C \|\nabla v\|^2_{L^2(\Omega)^n}$$

for some $C > 0$ by Poincaré's inequality. Taking the infimum over all such v, we obtain

$$\inf\{\|(v, \nabla v)\|^2_{L^2(\mathbb{R}^n)^{n+1}} : v \in H_0^1(\Omega), v \geq 1 \text{ a.e. in a neighbourhood of } A\} \leq C \operatorname{cap}(A).$$

This implies the claim.

Note that the reverse estimate to (3.A.1) does not hold in the general case, in particular we have $\operatorname{cap}(\Omega) = \infty$, but $\operatorname{cap}_{\mathbb{R}^n}(\Omega) < \infty$. However, we have the following important lemma.

Lemma 3.A.2 (Heinonen, Kilpeläinen, Martio, 1993, Lem. 2.9, Cor. 2.39)**.** For a set $A \subset \Omega$, we have

$$\operatorname{cap}(A) = 0 \quad \Longleftrightarrow \quad \operatorname{cap}_{\mathbb{R}^n}(A) = 0. \tag{3.A.2}$$

Finally, we need the concept of the so-called fine topology in $\mathbb{R}^n$, which is closely related to the notion of capacities. The fine topology is defined as the coarsest topology such that all sub-harmonic functions are continuous. We refer to Adams, Hedberg, 1996, Def. 6.4.1 or Heinonen, Kilpeläinen, Martio, 1993, Chap. 12 for more details. For our purposes it is enough to know that the fine topology possesses the following properties.

- The fine topology is finer than the usual topology on $\mathbb{R}^n$.
- Every $\operatorname{cap}_{\mathbb{R}^n}$-quasi-open set O (defined similarly to Definition 3.2.1) is equivalent to a finely open set $\tilde{O}$, in the sense that $\operatorname{cap}_{\mathbb{R}^n}((O \setminus \tilde{O}) \cup (\tilde{O} \setminus O)) = 0$, and every finely open set is $\operatorname{cap}_{\mathbb{R}^n}$-quasi-open, see Adams, Hedberg, 1996, Prop. 6.4.12, 6.4.13.
- The fine topology has the quasi-Lindelöf property, i.e., for every family $\{A_\alpha\}$ of finely open sets, there exists a countable subfamily $\{A_{\alpha_i}\}_{i \in \mathbb{N}}$, such that

  $$\operatorname{cap}_{\mathbb{R}^n}\Big(\bigcup_\alpha A_\alpha \setminus \bigcup_{i \in \mathbb{N}} A_{\alpha_i}\Big) = 0,$$

  see Adams, Hedberg, 1996, Rem. 6.5.11.

The induced topology on Ω is also called the fine topology. Since Ω is open, it is finely open. Therefore, a set $A \subset \Omega$ is finely open in $\mathbb{R}^n$ if and only if it is finely open in Ω.

Now, we are going to define the support w.r.t. the fine topology of a non-negative $\xi \in H^{-1}(\Omega)$, which is identified with a Borel measure by Lemma 3.2.4. To this end, we have to extend the Borel measure ξ to finely open sets. This requires the definition of a σ-algebra which contains the finely open sets.

Let us remark that every finely open set $O \subset \Omega$ is a Borel set up to a set of zero capacity, compare also Lemma 3.2.2: since O is $\mathrm{cap}_{\mathbb{R}^n}$-quasi-open, there exists, for any $\varepsilon > 0$, an open set G_ε such that $\mathrm{cap}_{\mathbb{R}^n}(G_\varepsilon) \leq \varepsilon$ and $O \cup G_\varepsilon$ is open. Since Ω is open, we can assume that $G_\varepsilon \subset \Omega$. Then, we have $O \subset \bigcap_{i\in\mathbb{N}}(O \cup G_{1/i})$ and

$$\mathrm{cap}_{\mathbb{R}^n}\Big(\Big[\bigcap_{i\in\mathbb{N}} O \cup G_{1/i}\Big] \setminus O\Big) \leq \mathrm{cap}_{\mathbb{R}^n}\Big(\Big[\bigcap_{i\in\mathbb{N}} G_{1/i}\Big]\Big) = 0.$$

Hence, O differs from the Borel set $\bigcap_{i\in\mathbb{N}}(O \cup G_{1/i})$ by a set $M \subset \Omega$ with $\mathrm{cap}_{\mathbb{R}^n}(M) = 0$. By (3.A.2) we have $\mathrm{cap}(M) = 0$.

This motivates the following definition.

Definition 3.A.3. We define the set

$$\mathcal{C} = \{G \cup H \subset \Omega : G \text{ is a Borel set and } \mathrm{cap}(H) = 0\}.$$

Then, $\mathcal{C}$ is a σ-algebra and it contains the finely open sets and all Borel sets.

Proof. We have to prove that $\mathcal{C}$ is a σ-algebra. It is easy to see that $\mathcal{C}$ is closed under countable unions, since the countable union of sets of zero capacity still has zero capacity. In order to show that $\mathcal{C}$ is closed under countable intersections, we remark that

$$\bigcap_{i\in\mathbb{N}} G_i \subset \bigcap_{i\in\mathbb{N}} (G_i \cup H_i) \subset \Big(\bigcap_{i\in\mathbb{N}} G_i\Big) \cup \Big(\bigcup_{i\in\mathbb{N}} H_i\Big).$$

Hence, for $\{G_i \cup H_i\} \subset \mathcal{C}$, the intersection differs from the Borel set $\bigcap_{i\in\mathbb{N}} G_i$ only by a set of zero capacity.

In the following, we simply say "$G \cup H \in \mathcal{C}$", instead of "$G \subset \Omega$ is a Borel set and $H \subset \Omega$ has zero capacity".

Now, let $\xi \in H^{-1}(\Omega)$ be a non-negative functional, which is identified with a Borel measure, see Lemma 3.2.4. Since $\xi(A) = 0$ for Borel sets A with $\mathrm{cap}(A) = 0$, we can extend ξ to $\mathcal{C}$ in a well-defined way by letting

$$\xi(G \cup H) = \xi(G) \quad \text{for all } G \cup H \in \mathcal{C}. \tag{3.A.3}$$

It is easy to show that ξ is additive on $\mathcal{C}$. Moreover, for all $\{G_i \cup H_i\} \subset \mathcal{C}$ we have

$$\xi\Big(\bigcup_{i\in\mathbb{N}} (G_i \cup H_i)\Big) = \xi\Big(\bigcup_{i\in\mathbb{N}} G_i \cup \bigcup_{i\in\mathbb{N}} H_i\Big) = \xi\Big(\bigcup_{i\in\mathbb{N}} G_i\Big) \leq \sum_{i\in\mathbb{N}} \xi(G_i) = \sum_{i\in\mathbb{N}} \xi(G_i \cup H_i).$$

Hence, ξ is countably subadditive on $\mathcal{C}$.

Now, we are in the position to define the fine support of ξ.

Lemma 3.A.4. Let $\xi \in H^{-1}(\Omega)$ be a non-negative functional. There exists a largest finely open set $M \subset \Omega$ with $\xi(M) = 0$. Its complement $\Omega \setminus M$ is called the fine support of ξ and is denoted by f-supp(ξ).

Proof. Let $\{A_\alpha\}$ be the family of finely open sets in Ω, whose ξ-measure is zero. Let $\{A_{\alpha_i}\}_{i\in\mathbb{N}}$ be a subfamily given by the quasi-Lindelöf property. We define

$$M = \bigcup_{\alpha} A_\alpha, \quad \tilde{M} = \bigcup_{i\in\mathbb{N}} A_{\alpha_i}, \quad O = M \setminus \tilde{M}.$$

By the definition of $\{A_{\alpha_i}\}$, we have $\mathrm{cap}_{\mathbb{R}^n}(O) = 0$ and $O \subset \Omega$. Hence, $\mathrm{cap}(O) = 0$ by Lemma 3.A.2. By definition (3.A.3) of ξ, this gives $\xi(O) = 0$. Using that ξ is countably additive, we have

$$\xi(M) = \xi(\tilde{M} \cup O) = \xi(\tilde{M}) \leq \sum_{i\in\mathbb{N}} \xi(A_{\alpha_i}) = 0.$$

This shows that M is the desired finely open set.

With these tools at hand, we can prove a refinement of Lemma 3.3.1.

Lemma 3.A.5. Let $\xi \in H^{-1}(\Omega)$ be a non-negative functional. Then, we have

$$\{v \in H_0^1(\Omega) : v = 0 \ \xi\text{-a.e.}\} = \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on f-supp}(\xi)\}.$$

In particular, we have

$$\mathcal{K}(\bar{u}) = \mathcal{T}_K(\bar{y}) \cap \bar{\xi}^\perp = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. in } A \text{ and } v = 0 \text{ q.e. in } \tilde{A}_s\},$$

where $\tilde{A}_s = \text{f-supp}(\bar{\xi})$.

Proof. We only have to prove the first identity. The second one follows together with (3.3.5).

"$\subset$": Let $v \in H_0^1(\Omega)$, $v = 0$ ξ-a.e. be given. The set $O = \{x \in \Omega : v \neq 0\}$ is quasi-open, hence, $\mathrm{cap}_{\mathbb{R}^n}$-quasi-open by (3.A.1). Therefore, there exists a finely open set F which differs by O only with capacity zero. Thus, $\xi(O) = \xi(F)$ and $\xi(O) = 0$ by assumption. Hence, $F \subset \Omega \setminus \text{f-supp}(\xi)$ and consequently, $\mathrm{cap}(O \cap \text{f-supp}(\xi)) = 0$.

"$\supset$": Let $v \in H_0^1(\Omega)$, $v = 0$ q.e. on f-supp(ξ) be given. We define the set $O = \{x \in \Omega : v \neq 0\}$. By assumption we have $\mathrm{cap}(O \cap \text{f-supp}(\xi)) = 0$. This implies $\xi(O \cap \text{f-supp}(\xi)) = 0$ and hence $\xi(O) = 0$ by $\xi(\Omega \setminus \text{f-supp}(\xi)) = 0$.

Note that the support of ξ (defined similarly by using the usual topology of $\mathbb{R}^n$ on Ω), is larger than the fine support of ξ (since every open set is finely open). Hence, we may not replace the fine support by the support of ξ in Lemma 3.A.5.

4. Towards M-stationarity for optimal control of the obstacle problem with control constraints

Abstract: We consider an optimal control problem, whose state is given as the solution of the obstacle problem. The controls are not assumed to be dense in $H^{-1}(\Omega)$. Hence, local minimizers may not be strongly stationary. By a non-smooth regularization technique similar to the virtual control regularization, we prove a system of C-stationarity using only minimal regularity requirements. We show that even a system of M-stationarity is satisfied under the assumption that the regularized adjoint states converge in capacity. We also give a counterexample, showing that this additional assumption might be crucial.

Keywords: M-stationarity conditions, obstacle problem, control constraints

MSC: 49K21, 35J86

4.1. Introduction

We consider the optimal control of the obstacle problem with control constraints

$$\begin{aligned} \text{Minimize} \quad & J(y,u), \\ \text{with respect to} \quad & (y,u,\xi) \in H_0^1(\Omega) \times U \times H^{-1}(\Omega), \\ \text{such that} \quad & \mathcal{A}\, y = \mathcal{B}\, u - \xi + f, \\ & y - \psi \leq 0 \quad \text{a.e. in } \Omega, \\ & \langle \xi,\, v - y \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \leq 0 \quad \text{for all } v \in H_0^1(\Omega) : v \leq \psi \text{ a.e. in } \Omega, \\ \text{and} \quad & u \in U_{\mathrm{ad}}. \end{aligned} \tag{\textbf{P}}$$

We refer to Assumption 4.1.2 for the precise requirements on the data of the problem. Due to the constraints on y and ξ, problem (**P**) is a mathematical problem with complementarity constraint (MPCC) in function space, see also the discussion at the end of Section 4.1.3, and it is difficult to verify sharp first-order necessary optimality conditions. In particular, the KKT conditions may not be necessary for optimality and one has to use stationarity concepts tailored to MPCCs, see Scheel, Scholtes, 2000. The tightest concept is the so-called *strong stationarity*, see (4.1.7), (4.1.8). It is, however, known that local minimizers of (**P**) may not satisfy the system of strong stationarity in the case that U_{ad} is a proper subset of U, e.g. in the presence of control constraints, or if the

range of $\mathcal{B}$ is not dense, e.g., if the control acts only on a (possibly lower-dimensional) subset of Ω, see Section 3.6 for counterexamples. Hence, one is interested in stationarity concepts which impose weaker conditions on the biactive set. In the finite-dimensional case, we refer to Outrata, 1999; Scheel, Scholtes, 2000 for stationarity concepts and for examples showing the different strength of these conditions. The stationarity systems for the infinite-dimensional problem (**P**) are defined in Section 4.1.4.

We give some references concerning optimality conditions of problems similar to (**P**). If $U_{\mathrm{ad}} = U$ and if the range of $\mathcal{B}$ is dense in $H^{-1}(\Omega)$, it is shown by Mignot, 1976 that minimizers of (**P**) satisfy the system of strong stationarity. The same result was reproduced by Hintermüller, Surowiec, 2011 by techniques from variational analysis and it was slightly generalized in Section 1.6.1. The special case $U = U_{\mathrm{ad}} = H^{-1}(\Omega)$ can be found in Outrata, Jarušek, Stará, 2011. Moreover, a strong stationarity result in presence of control constraints was given in G. Wachsmuth, 2014 (i.e., Chapter 3) under some assumptions on the data and the objective. However, all these results are rather restrictive and cover only special cases of (**P**). In the case $\Omega \subset \mathbb{R}^1$ and $U = H^{-1}(\Omega)$, Jarušek, Outrata, 2007 have shown that the minimizer is an M-stationary point, compare (4.1.11). However, this analysis heavily exploits the compact embedding of $H^1(\Omega)$ in $C(\bar{\Omega})$, which only holds in the one-dimensional case.

Many authors studied the approximation of problems similar to (**P**) with *smooth* problems. Similar to smooth relaxation methods in finite dimensions, see Hoheisel, Kanzow, Schwartz, 2013, at most C-stationarity can be expected in the limit. The most strict system which was obtained by this approach is the C-stationarity system from Schiela, D. Wachsmuth, 2013, see also Lemma 4.4.6. We also refer to Barbu, 1984; Ito, Kunisch, 2000; Hintermüller, Mordukhovich, Surowiec, 2014 for weaker stationarity systems.

In order to obtain stationarity conditions which are sharper than C-stationarity, one has to use different techniques. We mention that there are various approaches for deriving M-stationarity in finite dimensions, see e.g., Outrata, 1999; Flegel, Kanzow, 2006; Hoheisel, Kanzow, Schwartz, 2013, but these methods can not be applied to the infinite-dimensional problem (**P**). Hence, it is necessary to develop a new technique for deriving optimality conditions. In this work, we approximate (**P**) by a sequence of *non-smooth*, surrogate problems ($\mathbf{P}_n^{\mathrm{reg}}$), similar to the virtual control approach from Krumbiegel, Rösch, 2009. These regularized problems satisfy a system of strong stationarity, see Section 1.6.1. By passing to the limit with the regularization parameter, we obtain optimality conditions for the original problem (**P**). A main feature of our technique of proof is that we only use minimal regularity of the data.

Without any further assumptions, we arrive at the system of weak stationarity (4.1.7). In contrast to the literature, our system of weak stationarity contains conditions on the multipliers holding quasi-everywhere (q.e.) on certain sets, and not only almost-everywhere (a.e.). This is established by using results from potential theory. By assuming that the operator $\mathcal{A}$ is an elliptic second-order differential operator in divergence form, we obtain a system of C-stationarity, see (4.1.9), which is equivalent to the system in Schiela, D. Wachsmuth, 2013, see Lemma 4.4.6. We emphasize that, in contrast to Schiela, D. Wachsmuth, 2013, our regularity requirements are much weaker, in particular,

we do not need the Lipschitz continuity of the solution mapping of the obstacle problem from U to $C(\bar{\Omega})$. Finally, if the regularized adjoint states do not only converge weakly in $H_0^1(\Omega)$, but also converge in capacity, we even arrive at a system of M-stationarity, see (4.1.11). We remark that convergence of capacity is a rather weak notion of convergence, see Section 4.2. Moreover, this additional assumption of convergence in capacity is automatically satisfied in the one-dimensional case $d = 1$, see Lemma 4.2.3. Hence, we reproduce the result of Jarušek, Outrata, 2007 in a slightly different setting.

Using results from potential theory, in particular Theorem 4.2.5, is a novel technique for deriving optimality conditions of (**P**). Theorem 4.2.5 is utilized for deriving sign conditions for the multipliers, see, e.g., Lemmas 4.4.2 and 4.4.3. These results may also be applied for different regularization approaches, and are of independent interest. In particular, Lemma 4.4.3 answers an open question which was raised after the proof of Outrata, Jarušek, Stará, 2011, Thm. 16. Our technique allows to work with the basic regularity $y \in H_0^1(\Omega)$, $\xi \in H^{-1}(\Omega)$ of the obstacle problem and we do not need any additional regularity of the obstacle problem. We reproduce the C-stationarity result of Schiela, D. Wachsmuth, 2013 in this low-regularity setting, and we obtain even stronger conditions under the mild Assumption 4.5.1.

These main results of this paper are summarized in the following theorem, which is proven in Lemma 4.4.4, Lemma 4.4.5 and Theorem 4.5.4.

Theorem 4.1.1. Let us denote by $(\bar{y}, \bar{u}, \bar{\xi})$ a local minimizer of (**P**). Then, there exist multipliers, such that the system of weak stationarity (4.1.7) is satisfied.

If the operator $\mathcal{A}$ is an elliptic second-order differential operator, see Lemma 4.4.5, then there exist multipliers such that the system of C-stationarity (4.1.7), (4.1.9) is satisfied.

Now assume that there is a regularization scheme, see Definition 4.3.1, such that the regularized adjoint states p_n converge in capacity, see Assumption 4.5.1. Then there exist multipliers satisfying the system of M-stationarity (4.1.7), (4.1.11).

In the remainder of this section, we fix the assumptions on the data (Section 4.1.1), recall some basic results in capacity theory (Section 4.1.2), and set up the notation (Section 4.1.3). The various optimality concepts are defined in Section 4.1.4. In Section 4.2 we consider the concept of convergence in capacity. The regularization schemes are introduced in Section 4.3. By passing to the limit with the regularization, we obtain optimality systems for (**P**) in Section 4.4 and Section 4.5. Finally, we present a counterexample showing that Assumption 4.5.1 is crucial for deriving M-stationarity with our technique of proof, see Section 4.6.

4.1.1. Assumptions on the data and preliminaries

The set $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is open and bounded. We emphasize, that we do not assume any regularity of Ω in the entire paper.

The requirements on the data of (**P**) are collected in the following assumption.

Assumption 4.1.2. The bounded linear operator $\mathcal{A} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is assumed to be coercive. The right-hand side f belongs to $H^{-1}(\Omega)$. The obstacle $\psi \in H^1(\Omega)$ satisfies $\psi \geq 0$ on Γ in the sense that $\min\{\psi, 0\} \in H_0^1(\Omega)$.

The control space U is a Hilbert space, the control map $\mathcal{B} : U \to H^{-1}(\Omega)$ is a bounded, linear operator and the admissible set $U_{\text{ad}} \subset U$ is closed and convex. Either $\mathcal{B}$ or U_{ad} is assumed to be compact. Throughout the paper, we will identify U with its dual space.

The objective $J : H_0^1(\Omega) \times U \to \mathbb{R}$ is assumed to be continuously Fréchet-differentiable and bounded from below. We require that J is sequentially lower semi-continuous w.r.t. to the strong topology in $H_0^1(\Omega)$ and the weak topology in U, that is $J(y, u) \leq \liminf_{n\to\infty} J(y_n, u_n)$ for all sequences $\{(y_n, u_n)\} \subset H_0^1(\Omega) \times U$ satisfying $y_n \to y$ in $H_0^1(\Omega)$ and $u_n \rightharpoonup u$ in U. Finally, we assume that J is coercive w.r.t. the second variable on the feasible set U_{ad}, that is the boundedness of $\{u_n\}$ in U follows from the boundedness of $\{J(y_n, u_n)\}$ for all sequences $\{(y_n, u_n)\} \subset H_0^1(\Omega) \times U_{\text{ad}}$.

We will not assume more regularity of f and ψ and, up to Lemma 4.4.5, we do not impose any structural assumptions on $\mathcal{A}$.

Two possible choices of the control space are $U = L^2(\Omega)$ or $U = L^2(\Gamma, \gamma)$ where $\Gamma \subset \Omega$ is a smooth manifold of dimension $d - 1$ and γ is the surface measure on Γ.

As a simple example for the objective J, we mention the tracking-type functional

$$J(y, u) = \frac{1}{2} \|y - y_d\|^2_{L^2(\Omega_o)} + \frac{\alpha}{2} \|u\|^2_U,$$

where the observation domain $\Omega_o \subset \Omega$ is measurable, $y_d \in L^2(\Omega_o)$ is the desired state and $\alpha > 0$ is a regularization parameter.

The constraints

$$\mathcal{A}\, y = \mathcal{B}\, u - \xi + f, \tag{4.1.1a}$$

$$y - \psi \leq 0 \quad \text{a.e. in } \Omega, \tag{4.1.1b}$$

$$\langle \xi,\, v - y \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \leq 0 \quad \text{for all } v \in H_0^1(\Omega) : v \leq \psi \text{ a.e. in } \Omega, \tag{4.1.1c}$$

in **(P)** are equivalent to the obstacle problem:

$$\text{find } y \in H_0^1(\Omega), y \leq \psi \quad \text{such that } \langle \mathcal{A}\, y - \mathcal{B}\, u - f,\, v - y \rangle \geq 0, \quad \text{for all } v \in H_0^1(\Omega), v \leq \psi.$$

In presence of Assumption 4.1.2, this problem has a unique solution for every $\mathcal{B}\, u + f \in H^{-1}(\Omega)$ and the solution mapping $\mathcal{B}\, u + f \mapsto y$ is Lipschitz continuous from $H^{-1}(\Omega)$ to $H_0^1(\Omega)$, see Kinderlehrer, Stampacchia, 1980, Theorem II.2.1.

Using the continuity of this solution mapping and Assumption 4.1.2, the existence of solutions of **(P)** follows from standard arguments, see also Lemma 4.3.2.

4.1.2. Capacity theory

In what follows, we recall some basic results in capacity theory. These are crucial to characterize tangent cones in $H^1_0(\Omega)$, see (4.1.3) below and to give a convenient expression of the critical cone, see (4.1.5). The capacity of a set $O \subset \Omega$ is defined as

$$\operatorname{cap}(O) := \inf\{\|\nabla v\|^2_{L^2(\Omega;\mathbb{R}^d)} : v \in H^1_0(\Omega) \text{ and } v \geq 1 \text{ a.e. in a neighbourhood of } O\},$$

see, e.g., Attouch, Buttazzo, Michaille, 2006, Sec. 5.8.2, Bonnans, Shapiro, 2000, Def. 6.47 and Delfour, Zolésio, 2001, Sec. 8.6.1.

A function $v : \Omega \to \mathbb{R}$ is called *quasi-continuous* if for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\operatorname{cap}(G_\varepsilon) < \varepsilon$ and v is continuous on $\Omega \setminus G_\varepsilon$. A set $O \subset \Omega$ is called *quasi-open* if for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\operatorname{cap}(G_\varepsilon) < \varepsilon$ and $O \cup G_\varepsilon$ is open. For every quasi-continuous function v, the set $\{x \in \Omega : v(x) > 0\}$ is quasi-open.

We say that a property P (depending on $x \in \Omega$) holds quasi-everywhere (q.e.), if it is only violated on a set of capacity zero, e.g., $\operatorname{cap}(\{x \in \Omega : P(x) \text{ does not hold}\}) = 0$. We say that P holds q.e. on a subset $K \subset \Omega$, if and only if $\operatorname{cap}(\{x \in K : P(x) \text{ does not hold}\}) = 0$.

It is known, see Delfour, Zolésio, 2001, Thm. 8.6.1, that every $v \in H^1(\Omega)$ possesses a quasi-continuous representative and this representative is uniquely determined up to sets of zero capacity. When we speak about a function $v \in H^1(\Omega)$, we always mean the quasi-continuous representative. Every sequence which converges in $H^1_0(\Omega)$ possesses a pointwise quasi-everywhere convergent subsequence, see Bonnans, Shapiro, 2000, Lem. 6.52.

The so-called fine topology in $\mathbb{R}^d$ is closely related to the notion of capacities. It is defined as the coarsest topology such that all sub-harmonic functions are continuous, we refer to Adams, Hedberg, 1996, Def. 6.4.1 or Heinonen, Kilpeläinen, Martio, 1993, Chap. 12 for more details.

We recall, that a non-negative $\xi \in H^{-1}(\Omega)$ can be represented as a regular Borel measure, see, e.g., Bonnans, Shapiro, 2000, p.564. Moreover, since ξ does not charge sets of capacity zero, it can be extended to finely-open sets and the *fine support*, denoted by f-supp(ξ), is the complement of the largest finely-open set O with $\xi(O) = 0$. We refer to Section 3.A for details.

By following the proof of Heinonen, Kilpeläinen, Martio, 1993, Lemma 4.7, we find

$$\operatorname{cap}(O) = \inf\{\|\nabla v\|^2_{L^2(\Omega;\mathbb{R}^d)} : v \in H^1_0(\Omega) \text{ and } v \geq 1 \text{ q.e. on } O\}. \tag{4.1.2}$$

We use $\{v \geq 0\}$ as a short hand for the set $\{x \in \Omega : v(x) \geq 0\}$, and similarly for other expressions depending on functions. We emphasize that such sets are defined up to sets of zero capacity if v is quasi-continuous, in particular if $v \in H^1(\Omega)$.

4.1.3. Notation

As the norm in $H_0^1(\Omega)$ we choose

$$\|y\|^2_{H_0^1(\Omega)} = \|\nabla y\|^2_{L^2(\Omega;\mathbb{R}^d)} = \int_\Omega |\nabla y|^2 \,\mathrm{d}x,$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^d$.

We define the closed convex set

$$K := \{v \in H_0^1(\Omega) : v \leq \psi \text{ a.e. on } \Omega\}.$$

We denote by $\mathcal{T}_K(y)$ the tangent cone of K at y, which is the closed conic hull of $K - y$. We recall that this tangent cone can be characterized by

$$\mathcal{T}_K(y) = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. on } \{y = \psi\}\} \tag{4.1.3}$$

for $y \in K$, see Mignot, 1976, Lemma 3.2. We would like to emphasize that the notion of "quasi-everywhere" is crucial for this characterization, and it is not possible to rephrase (4.1.3) in terms of "almost-everywhere". This is in particular true if the set $\{y = \psi\}$ has measure zero but positive capacity.

For a set $M \subset H_0^1(\Omega)$ we define, as usual, the polar cone

$$M^\circ := \{f \in H^{-1}(\Omega) : \langle f, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \leq 0 \text{ for all } v \in M\}.$$

Note that (4.1.1c) is equivalent to $\xi \in \mathcal{T}_K(y)^\circ$ and this cone can be characterized by

$$\mathcal{T}_K(y)^\circ = \{\xi \in H^{-1}(\Omega) : \xi \text{ is non-negative and } y - \psi = 0 \ \xi\text{-a.e. on } \Omega\}, \tag{4.1.4}$$

see Mignot, 1976, Lemma 3.1, Proposition 3.2.5 and also Bonnans, Shapiro, 2000, Theorem 6.57 in the case $\psi = 0$.

For $\xi \in H^{-1}(\Omega)$, we define the annihilator

$$\xi^\perp := \{v \in H_0^1(\Omega) : \langle \xi, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0\}.$$

We further define the critical cone $\mathcal{K}(y,\xi) := \mathcal{T}_K(y) \cap \xi^\perp$ for $y \in K$ and $\xi \in \mathcal{T}_K(y)^\circ$. In the case of high regularity $\xi \in L^2(\Omega)$, we have

$$\mathcal{K}(y,\xi) = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. on } \{y = \psi\} \text{ and } v = 0 \text{ a.e. on } \{\xi > 0\}\},$$

but it is very cumbersome to work with this mix of an a.e.-equality and a q.e.-inequality. By employing the notion of the fine support of ξ, we do not need any additional regularity of ξ and find

$$\mathcal{K}(y,\xi) = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. on } \{y = \psi\} \text{ and } v = 0 \text{ q.e. on f-supp}(\xi)\}, \tag{4.1.5}$$

see Lemma 3.A.5. Note that the cone $\mathcal{K}(y, \xi)$ only depends on the active set $A = \{y = \psi\}$ and the strictly active set $A_s = \text{f-supp}(\xi)$. Moreover, we define the inactive set $I = \{y < \psi\}$.

We make also use of the normal cone of U_{ad} at $\bar{u} \in U_{\text{ad}}$ w.r.t. the U-inner product which is denoted by

$$\mathcal{N}_{U_{\text{ad}}}(\bar{u}) = \{v \in U : (v,\, u - \bar{u}) \leq 0 \text{ for all } u \in U_{\text{ad}}\}.$$

Finally, we would like to justify that the conditions (4.1.1b), (4.1.1c) are called "complementarity conditions", although the set K is, in general, not a cone. As already mentioned, these constraints are equivalent to

$$y \in K, \qquad \xi \in \mathcal{T}_K(y)^\circ. \tag{4.1.6}$$

In the case that K is a cone, which happens if and only if $\psi = 0$, this is, in turn, equivalent to the (conic) complementarity condition

$$y \in K, \qquad \xi \in K^\circ, \qquad \langle y,\, \xi \rangle = 0.$$

Hence, (4.1.6) can be seen as a proper generalization of a complementarity condition to the case that K is not a cone. We also refer to Section 1.5.4 for further discussions.

4.1.4. Optimality conditions

Following the nomenclature from finite dimensions, see Scheel, Scholtes, 2000, we say that $(\bar{y}, \bar{u}, \bar{\xi})$ together with multipliers $p \in H_0^1(\Omega)$, $\mu \in H^{-1}(\Omega)$ and $\lambda \in U$ is *weakly stationary*, if the system

$$\begin{aligned}
\mathcal{A}\,\bar{y} &= \mathcal{B}\,\bar{u} - \bar{\xi} + f, && \text{(4.1.7a)}\\
\bar{y} &\in K, && \text{(4.1.7b)}\\
\bar{\xi} &\in \mathcal{T}_K(\bar{y})^\circ, && \text{(4.1.7c)}\\
\bar{u} &\in U_{\text{ad}}, && \text{(4.1.7d)}\\
J_y(\bar{y}, \bar{u}) + \mu + \mathcal{A}^\star p &= 0, && \text{(4.1.7e)}\\
J_u(\bar{y}, \bar{u}) + \lambda - \mathcal{B}^\star p &= 0, && \text{(4.1.7f)}\\
-p &\in \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } A_s\}, && \text{(4.1.7g)}\\
\mu &\in \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } A\}^\circ, && \text{(4.1.7h)}\\
\lambda &\in \mathcal{N}_{U_{\text{ad}}}(u) && \text{(4.1.7i)}
\end{aligned}$$

is satisfied. Here, J_y and J_u denote the partial derivatives of J and $A = \{\bar{y} = \psi\}$, $A_s = \text{f-supp}(\bar{\xi})$. The multiplier associated to $\xi \in \mathcal{T}_K(y)^\circ$ has already been eliminated since it equals $-p$. Note that in difference to weak stationarity systems appearing in the literature, condition (4.1.7g) contains a quasi-everywhere condition on p, and similarly in (4.1.7h) for the test function v. If μ would be a function, (4.1.7h) would read $\mu = 0$ on $\Omega \setminus A = I = \{\bar{y} < \psi\}$.

For a *strongly stationary* point, we additionally require

$$-p \in \mathcal{K}(\bar{y},\bar{\xi}) \;= \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } A \text{ and } v = 0 \text{ q.e. on } A_s\}, \tag{4.1.8a}$$

$$\mu \in \mathcal{K}(\bar{y},\bar{\xi})^\circ = \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } A \text{ and } v = 0 \text{ q.e. on } A_s\}^\circ. \tag{4.1.8b}$$

If the biactive set $B = A \setminus A_s$ has capacity zero, this condition already follows from (4.1.7). Otherwise, strong stationarity is strictly stronger than weak stationarity and both conditions differ only on the biactive set. We mention that the strong and weak stationarity systems are the KKT conditions of the relaxed and tightened nonlinear program associated with (**P**), see Section 1.5.1.

Again, we would like to emphasize that the conditions in (4.1.8) cannot be formulated with the notion of "almost everywhere", see the comment after (4.1.3).

In the unconstrained case $U_{\mathrm{ad}} = U$ and if the range of $\mathcal{B}$ is dense in $H^{-1}(\Omega)$, it is well known that all local minimizers of (**P**) are strongly stationary, see Mignot, 1976. A partial result concerning strong stationarity in the constrained case can be found in Section 3.5. However, there are counterexamples showing that strong stationarity is, in general, not valid in the presence of control constraints, see Section 3.6.

In finite dimensions, there are several systems between weak and strong stationarity, e.g., C- and M-stationarity. It is, however, not directly clear how the finite-dimensional formulations should be transfered to the infinite-dimensional case. For problems of type (**P**), several systems of C-stationarity are defined in the literature, and the tightest system is the one given in Schiela, D. Wachsmuth, 2013. This system is described before Lemma 4.4.6.

Our definition of C-stationarity is slightly different due to the low regularity requirements. We say that the feasible point $(\bar{y},\bar{u},\bar{\xi})$ of (**P**) together with multipliers $p \in H_0^1(\Omega)$, $\mu \in H^{-1}(\Omega)$ and $\lambda \in U$ is C-stationary, if (4.1.7) and

$$\langle \mu,\, \varphi\, p\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \ge 0 \quad \text{for all } \varphi \in W^{1,\infty}(\Omega), \varphi \ge 0 \tag{4.1.9}$$

are satisfied. In difference to the system in Schiela, D. Wachsmuth, 2013, our system contains conditions holding quasi-everywhere in (4.1.7). However, we show that both systems are equivalent under the higher regularity requirements of Schiela, D. Wachsmuth, 2013, see Lemma 4.4.6.

To our knowledge, the only available definition of M-stationarity for problems similar to (**P**) is given in Jarušek, Outrata, 2007, which is, however, limited to the 1-dimensional case.

In order to motivate our notion of M-stationarity, we recall the finite-dimensional situation. For the sake of brevity, we only discuss the multipliers associated with the complementarity constraint and their sign conditions. We refer to Kanzow, Schwartz, 2013, Section 2.2 for a more complete discussion. If the complementarity constraint is $0 \le G(x) \perp H(x) \ge 0$, one introduces the index sets

$$\begin{aligned} I_{+0} &= \{i \in \mathbb{R}^n : G_i(\bar{x}) > 0, H_i(\bar{x}) = 0\},\\ I_{00} &= \{i \in \mathbb{R}^n : G_i(\bar{x}) = 0, H_i(\bar{x}) = 0\},\\ I_{0+} &= \{i \in \mathbb{R}^n : G_i(\bar{x}) = 0, H_i(\bar{x}) > 0\} \end{aligned}$$

given a local minimizer $\bar{x}$, compare Kanzow, Schwartz, 2013, Section 2.2. For the multipliers γ, ν associated with G and H one requires

$$\gamma_i = 0 \text{ for } i \in I_{+0} \quad \text{and} \quad \nu_i = 0 \text{ for } i \in I_{0+}$$

for weak stationarity, see Kanzow, Schwartz, 2013, Definition 2.3. If additionally

$$\gamma_i, \nu_i \geq 0 \text{ or } \gamma_i \, \nu_i = 0 \text{ for } i \in I_{00}$$

is satisfied, the point is called M-stationary. This formulation for M-stationarity is, however, not suited for the infinite-dimensional setting. Therefore, we give an alternative description: there is a disjoint decomposition of the biactive set $I_{00} = \hat{I}_{+0} \cup \hat{I}_{00} \cup \hat{I}_{0+}$, such that

$$\gamma_i = 0 \text{ for } i \in I_{+0} \cup \hat{I}_{+0}, \tag{4.1.10a}$$

$$\nu_i = 0 \text{ for } i \in I_{0+} \cup \hat{I}_{0+}, \tag{4.1.10b}$$

$$\gamma_i, \nu_i \geq 0 \text{ for } i \in \hat{I}_{00} \tag{4.1.10c}$$

are satisfied. It is easy to see that this is equivalent to the usual definition of M-stationarity in finite dimensions. Moreover, it is possible to transfer this definition to the infinite-dimensional setting.

A feasible point $(\bar{y}, \bar{u}, \bar{\xi})$ with multipliers (p, μ, λ) is said to be *M-stationary*, if it satisfies (4.1.7) and there is a disjoint decomposition of the biactive set $B = \hat{I} \cup \hat{B} \cup \hat{A}_s$ such that the conditions

$$p \in -\hat{\mathcal{K}}, \tag{4.1.11a}$$

$$\mu \in \hat{\mathcal{K}}^\circ \tag{4.1.11b}$$

are satisfied, where

$$\hat{\mathcal{K}} = \{v \in H_0^1(\Omega) : v \leq 0 \text{ q.e. on } \hat{B} \text{ and } v = 0 \text{ q.e. on } A_s \cup \hat{A}_s\}. \tag{4.1.12}$$

In the case that μ is even a function, the condition (4.1.11b) asserts $\mu \geq 0$ on $\hat{B}$ and $\mu = 0$ on $I \cup \hat{I}$. Hence, (4.1.11) is the infinite-dimensional analogue of (4.1.10). As in the finite-dimensional case, we can easily show

$$\text{strong stationarity} \;\Rightarrow\; \text{M-stationarity} \;\Rightarrow\; \text{C-stationarity} \;\Rightarrow\; \text{weak stationarity}.$$

4.2. Convergence in capacity

In order to obtain optimality conditions, we employ the notion of convergence in capacity, see, e.g., Casado-Diaz, Dal Maso, 2000. We give the definition and some basic properties. The main result of this section is Lemma 4.2.6 which enables us to pass to the limit with certain duality relations.

Definition 4.2.1. Let $\{v_n\}$ be a sequence of quasi-continuous functions mapping Ω to $\mathbb{R}$ and let $v : \Omega \to \mathbb{R}$ be quasi-continuous. If

$$\operatorname{cap}(\{|v_n - v| \geq \varepsilon\}) \to 0 \quad \text{as } n \to \infty$$

holds for all $\varepsilon > 0$, we say that $\{v_n\}$ converges towards v in capacity.

This notion of convergence is similar to the convergence in measure. An argument similar to Chebyshev's inequality shows that strong convergence in $H_0^1(\Omega)$ implies convergence in capacity.

Lemma 4.2.2. Let $v_n \to v$ in $H_0^1(\Omega)$. Then, v_n converges to v in capacity.

Proof. Let $\varepsilon > 0$ be fixed. For convenience, we set $O_n = \{|v_n - v| \geq \varepsilon\}$. Since $\varepsilon^{-1} |v_n - v| \geq 1$ q.e. on O_n, (4.1.2) implies

$$\operatorname{cap}(O_n) \leq \left\| \varepsilon^{-1} |v_n - v| \right\|^2_{H_0^1(\Omega)} = \varepsilon^{-2} \|v_n - v\|^2_{H_0^1(\Omega)}.$$

This yields the claim.

In the one-dimensional case, one can utilize the compact embedding of $H_0^1(\Omega)$ in $C(\bar{\Omega})$ to weaken the assumptions of Lemma 4.2.2.

Lemma 4.2.3. Let $\Omega \subset \mathbb{R}^1$ be a bounded, open set and assume $v_n \rightharpoonup v$ in $H_0^1(\Omega)$. Then, v_n converges to v in capacity.

Proof. Let $\varepsilon > 0$ be given. Since $H_0^1(\Omega)$ is compactly embedded in $C(\bar{\Omega})$, we can choose $N \in \mathbb{N}$ such that $\|v_n - v\|_{C(\bar{\Omega})} < \varepsilon$ for all $n \geq N$. Hence, the set $\{|v_n - v| \geq \varepsilon\}$ is empty for all $n \geq N$.

In higher dimensions, we need an additional assumption on the sequence v_n. Note that we do not assume $p > d$ and we do not use the compact embedding of $W_0^{1,p}(\Omega)$ in $C(\bar{\Omega})$ in two dimensions for $p > 2$.

Lemma 4.2.4. Let $v_n \rightharpoonup v$ in $H_0^1(\Omega)$ and assume that v_n is bounded in $W_0^{1,p}(\Omega)$ for some $p > 2$. Then, v_n converges to v in capacity.

Proof. This proof is essentially due to Evans, 1990, Thm. 1.3.3. Let $\varepsilon > 0$ be fixed. We define $O_n = \{|v_n - v| \geq \varepsilon\}$, $\tilde{O}_n = \{|v_n - v| \geq \varepsilon/2\}$, and set

$$w_n = \frac{2}{\varepsilon} \max\Big(|v_n - v| - \frac{\varepsilon}{2}, 0\Big).$$

We have $w_n \geq 1$ q.e. on O_n, which implies by (4.1.2)

$$\operatorname{cap}(O_n) \leq \|w_n\|^2_{H^1_0(\Omega)} = \int_{\tilde{O}_n} |\nabla w_n|^2 \,\mathrm{d}x = \frac{4}{\varepsilon^2} \int_{\tilde{O}_n} |\nabla v_n - \nabla v|^2 \,\mathrm{d}x.$$

We choose $q \in (1, \infty)$ such that $1/1 = 2/p + 1/q$ and apply Hölder's inequality

$$\operatorname{cap}(O_n) \leq \frac{4}{\varepsilon^2} \int_{\tilde{O}_n} |\nabla v_n - \nabla v|^2 \,\mathrm{d}x \leq \frac{4}{\varepsilon^2}\, m(\tilde{O}_n)^{1/q} \Big(\int_{\tilde{O}_n} |\nabla v_n - \nabla v|^p \,\mathrm{d}x \Big)^{2/p},$$

where m denotes the Lebesgue measure in $\mathbb{R}^d$. Using the compact embedding of $H^1_0(\Omega)$ in $L^2(\Omega)$ and Chebyshev's inequality, we find $m(\tilde{O}_n) \to 0$. By the boundedness of v_n in $W^{1,p}_0(\Omega)$, the last term is bounded. Hence, the assertion follows.

Note that neither (weak) convergence in $H^1_0(\Omega)$ nor the boundedness in $W^{1,p}_0(\Omega)$, $p > 2$, are necessary for the convergence in capacity. In order to illustrate that convergence in capacity is a rather weak measure of convergence, we give a simple example. Let $\Omega = U_1(0) \subset \mathbb{R}^2$ be the open unit disc. We set

$$\tilde{v}(x,y) = \begin{cases} \log(\frac{1}{\sqrt{x^2+y^2}})^{1/4} - \log(2)^{1/4} & \text{for } x^2 + y^2 \leq 1/4, \\ 0 & \text{else.} \end{cases}$$

Then, it is easy to show that $\tilde{v} \in H^1_0(\Omega) \setminus L^\infty(\Omega)$. We set

$$v_n(x,y) = \tilde{v}(n\,x, n\,y) \quad \text{for } (x,y) \in \Omega.$$

By definition, $\|v_n\|_{H^1_0(\Omega)} = \|\tilde{v}\|_{H^1_0(\Omega)}$ is bounded. Moreover, by using the compact embedding $H^1_0(\Omega) \to L^2(\Omega)$ and the pointwise convergence towards 0, we can show that v_n converges weakly towards 0 in $H^1_0(\Omega)$. Moreover, the set $\{|v_n| \geq \varepsilon\}$ is a ball whose radius goes to zero as $n \to \infty$. Hence, its capacity also converges to zero. This yields the convergence of v_n to 0 in capacity, although v_n has constant distance to 0 in $H^1_0(\Omega)$ and v_n does not even belong to $W^{1,p}_0(\Omega)$ for any $p > 2$.

Finally, we give an application of the convergence in capacity, which enables us to pass to the limit in certain duality relations, see Lemma 4.2.6. To this end, we need the following result which is a replacement for the partition of unity for arbitrary quasi-open sets.

Theorem 4.2.5 (Kilpeläinen, Malý, 1992, Lem. 2.4, Thm. 2.10)**.** Let $U \subset \Omega$ be a quasi-open set. Let $g \in H^1_0(\Omega)$ with $0 \leq g \leq K$ a.e. in Ω for some constant $K \geq 0$ and $g = 0$ q.e. on $\Omega \setminus U$ be given.

Moreover, let $\{U_n\}_{n\in\mathbb{N}}$ be an increasing sequence of quasi-open subsets of Ω such that $\operatorname{cap}(U \setminus \bigcup_{n\in\mathbb{N}} U_n) = 0$.

Then, there exists a sequence $\{g_n\}_{n\in\mathbb{N}} \in H^1_0(\Omega)$ with $0 \leq g_n \leq K$ a.e. in Ω, $g_n = 0$ q.e. on $\Omega \setminus U_n$ and $g_n \to g$ in $H^1_0(\Omega)$.

Lemma 4.2.6. Let there be given sequences $\{s_n\}, \{t_n\} \subset H^1(\Omega)$ converging in capacity to $s, t \in H^1(\Omega)$, respectively. Moreover, let the sequence $\{\kappa_n\} \subset H^{-1}(\Omega)$ converge weakly to $\kappa \in H^{-1}(\Omega)$. Then,

$$\kappa_n \in \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{s_n = 0\} \cap \{t_n = 0\}\}^\circ$$

for all $n \in \mathbb{N}$ implies

$$\kappa \in \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{s = 0\} \cap \{t = 0\}\}^\circ.$$

Proof. Let $v \in H_0^1(\Omega)$ with $v \geq 0$ q.e. on Ω and $v = 0$ q.e. on $\{s = 0\} \cap \{t = 0\}$ be given. In order to show $\langle \kappa, v \rangle \leq 0$, we approximate v by using Theorem 4.2.5.

For arbitrary $K > 0$ we set $v_K = \min(v, K)$ and obtain $\|v - v_K\|_{H_0^1(\Omega)} \to 0$ as $K \to \infty$. Since $v_K = 0$ q.e. on $\{s = 0\} \cap \{t = 0\} = \Omega \setminus (\{s \neq 0\} \cup \{t \neq 0\})$ and by the decomposition

$$\{s \neq 0\} \cup \{t \neq 0\} = \bigcup_{n=1}^{\infty} \{|s| > 1/n\} \cup \{|t| > 1/n\},$$

we can apply Theorem 4.2.5 and obtain a sequence $v_{K,n} \in H_0^1(\Omega)$ with $0 \leq v_{K,n} \leq K$ a.e. in Ω, $\|v_K - v_{K,n}\|_{H_0^1(\Omega)} \to 0$ as $n \to \infty$ and

$$v_{K,n} = 0 \text{ q.e. in } \Omega \setminus (\{|s| > 1/n\} \cup \{|t| > 1/n\}).$$

Now, we set

$$O_{K,n,m} = \{v_{K,n} > 0\} \cap \{s_m = 0\} \cap \{t_m = 0\}.$$

Since $\{v_{K,n} > 0\} \subset \{|s| > 1/n\} \cup \{|t| > 1/n\}$ (up to a set of capacity zero), we find

$$O_{K,n,m} \subset \{|s - s_m| \geq 1/n\} \cup \{|t - t_m| \geq 1/n\}.$$

Since $\{s_m\}, \{t_m\}$ converges in capacity, this gives

$$\operatorname{cap}(O_{K,n,m}) \to 0 \quad \text{as } m \to \infty.$$

By the definition of the capacity, there exists a non-negative $w_{n,m}$ with $w_{n,m} \geq 1$ in a neighbourhood of $O_{K,n,m}$ with $\|w_{n,m}\|^2_{H_0^1(\Omega)} \leq \operatorname{cap}(O_{K,n,m}) + 1/m$. We set

$$v_{K,n,m} = \max(v_{K,n} - K\, w_{n,m}, 0).$$

Since $v \mapsto \max(v, 0)$ is continuous in $H_0^1(\Omega)$ and $\|w_{n,m}\|_{H_0^1(\Omega)} \to 0$, we get $\|v_{K,n} - v_{K,n,m}\|_{H_0^1(\Omega)} \to 0$ as $m \to \infty$. Moreover, we have $v_{K,n,m} \geq 0$ and $v_{K,n,m} = 0$ q.e. on $\{v_{K,n} = 0\} \cup O_{K,n,m} \subset \{s_m = 0\} \cap \{t_m = 0\}$. This yields

$$\langle \kappa_m, v_{K,n,m} \rangle \leq 0.$$

By passing to the limits $m \to \infty$, $n \to \infty$ and $K \to \infty$, we get

$$\langle \kappa, v \rangle \leq 0.$$

4.3. Regularization schemes

In order to derive optimality conditions for the problem (**P**), we will consider a certain class of regularizations. We use an idea similar to the virtual control technique developed in Krumbiegel, Rösch, 2009 for state-constrained problems.

Definition 4.3.1. A regularization scheme is a quadruple $(V, \mathcal{C}, \{\alpha_n\}, \beta)$, where V is a Hilbert space, $\mathcal{C} : V \to H^{-1}(\Omega)$ is a compact linear operator with dense range, the sequence $\{\alpha_n\} \subset (0, \infty)$ converges towards infinity and $\beta > 0$.

We fix a local minimizer $(\bar{y}, \bar{u}, \bar{\xi})$ of (**P**) and denote by $\varepsilon > 0$ its radius of optimality. With each regularization scheme $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ and $n \in \mathbb{N}$ we associate the regularized problem

$$
\begin{aligned}
\text{Minimize} \quad & J(y,u) + \frac{\beta}{2}\|u-\bar{u}\|_U^2 + \frac{\alpha_n}{2}\|v\|_V^2, \\
\text{with respect to} \quad & (y,u,v,\xi) \in H_0^1(\Omega) \times U \times V \times H^{-1}(\Omega), \\
\text{such that} \quad & \mathcal{A}\,y = \mathcal{B}\,u + \mathcal{C}\,v - \xi + f, \\
& y - \psi \le 0 \quad \text{a.e. in } \Omega, \\
& \langle \xi,\, v - y\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \le 0 \quad \text{for all } v \in H_0^1(\Omega) : v \le \psi \text{ a.e. in } \Omega, \\
\text{and} \quad & u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}).
\end{aligned}
\tag{$\mathbf{P}_n^{\mathrm{reg}}$}
$$

Here, $B_\varepsilon(\bar{u})$ is the closed ball centered in $\bar{u}$ with radius ε in the U-norm.

A simple example for a regularization scheme is given by $(L^2(\Omega), \mathcal{I}, \{n\}_{n\in\mathbb{N}}, 1)$, where $\mathcal{I} : L^2(\Omega) \to H^{-1}(\Omega)$ is the canonical embedding.

By using standard arguments, we obtain the existence of solutions of ($\mathbf{P}_n^{\mathrm{reg}}$).

Lemma 4.3.2. Let Assumption 4.1.2 be satisfied and let $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ be a regularization scheme. Then, ($\mathbf{P}_n^{\mathrm{reg}}$) has a global solution for each $n \in \mathbb{N}$.

Proof. Let $n \in \mathbb{N}$ be fixed. First we note that $(\bar{y}, \bar{u}, 0, \bar{\xi})$ is a feasible point of ($\mathbf{P}_n^{\mathrm{reg}}$). Additionally, the objective is bounded from below, hence, the infimum j of ($\mathbf{P}_n^{\mathrm{reg}}$) is finite and there exist a minimizing sequence (y^k, u^k, v^k, ξ^k).

Since J is bounded from below, $\|u^k\|_U$ and $\|v^k\|_V$ are bounded. By the compactness of $\mathcal{C}$, there exists a subsequence (denoted by the same symbol) and $u_n \in U$, $v_n \in V$ with $u^k \rightharpoonup u_n$ in U and $\mathcal{C}v^k \to \mathcal{C}v_n$ in $H^{-1}(\Omega)$ as $k \to \infty$. Since $\mathcal{B}$ or U_{ad} is compact, we can assume $\mathcal{B}\,u^k \to \mathcal{B}\,u_n$ in $H^{-1}(\Omega)$. By the properties of the solution mapping of the obstacle problem, there are $y_n \in H_0^1(\Omega)$, $\xi_n \in H^{-1}(\Omega)$, such that (y_n, u_n, v_n, ξ_n) is feasible for ($\mathbf{P}_n^{\mathrm{reg}}$) and $y^k \to y_n$ in $H_0^1(\Omega)$. By the strongly-weakly lower-semicontinuity of J, (y_n, u_n, v_n, ξ_n) is a global solution.

From now on, we denote by (y_n, u_n, v_n, ξ_n) a fixed global solution of $(\mathbf{P}_n^{\mathrm{reg}})$.

Lemma 4.3.3. Let Assumption 4.1.2 be satisfied and let a regularization scheme $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ be given. Then, the solutions (y_n, u_n, v_n, ξ_n) of $(\mathbf{P}_n^{\mathrm{reg}})$ satisfy

$$\begin{aligned} y_n &\to \bar{y} \quad \text{in } H_0^1(\Omega), \\ u_n &\to \bar{u} \quad \text{in } U, \\ v_n &\to 0 \quad \text{in } V, \\ \xi_n &\to \bar{\xi} \quad \text{in } H^{-1}(\Omega). \end{aligned}$$

The proof is standard, but included for the reader's convenience.

Proof. Let us take an arbitrary subsequence of (y_n, ξ_n, v_n, u_n), which is not relabeled.

Owing to the requirement that J is bounded from below, $(u_n, \sqrt{\alpha_n}\, v_n)$ is bounded in $U \times V$ and we can extract a weakly convergent subsequence (which is not relabeled) with weak limit $(\tilde{u}, \tilde{v})$. In case U_{ad} is compact, u_n converges even strongly in U. This yields the convergence of $\mathcal{B}\, u_n + \mathcal{C}\, v_n$ towards $\mathcal{B}\, \tilde{u}$ in $H^{-1}(\Omega)$. Since the solution operator of the obstacle problem is continuous, this yields the convergence of $(y_n, \xi_n) \to (\tilde{y}, \tilde{\xi})$ in $H_0^1(\Omega) \times H^{-1}(\Omega)$.

Now, let $u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u})$ be arbitrary and denote by y the associated state. By optimality of (y_n, u_n, v_n, ξ_n) we find

$$J(y_n, u_n) + \frac{\beta}{2} \|u_n - \bar{u}\|_U^2 + \frac{\alpha_n}{2} \|v_n\|_V^2 \le J(y, u) + \frac{\beta}{2} \|u - \bar{u}\|_U^2.$$

Passing to the limit $n \to \infty$ and taking into account the lower semi-continuity properties of J, we find

$$J(\tilde{y}, \tilde{u}) + \frac{\beta}{2} \|\tilde{u} - \bar{u}\|_U^2 \le J(y, u) + \frac{\beta}{2} \|u - \bar{u}\|_U^2 \qquad \text{for all } u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}).$$

By plugging in $u = \bar{u}$ and using $J(\tilde{y}, \tilde{u}) \ge J(\bar{y}, \bar{u})$ we find

$$J(\bar{y}, \bar{u}) + \frac{\beta}{2} \|\tilde{u} - \bar{u}\|_U^2 \le J(\bar{y}, \bar{u}) + \frac{\beta}{2} \|\bar{u} - \bar{u}\|_U^2,$$

and, hence, $(\tilde{y}, \tilde{u}, \tilde{\xi}) = (\bar{y}, \bar{u}, \bar{\xi})$.

It remains to show the strong convergence of u_n. Again, by optimality of (y_n, u_n, v_n, ξ_n), we have

$$J(y_n, u_n) + \frac{\beta}{2} \|u_n - \bar{u}\|_U^2 + \frac{\alpha_n}{2} \|v_n\|_V^2 \le J(\bar{y}, \bar{u}).$$

This implies

$$\frac{\beta}{2} \|u_n - \bar{u}\|_U^2 \le J(\bar{y}, \bar{u}) - J(y_n, u_n).$$

Hence, by taking the limes superior on both sides

$$\limsup_{n\to\infty} \frac{\beta}{2} \|u_n - \bar{u}\|_U^2 \leq J(\bar{y}, \bar{u}) - \liminf_{n\to\infty} J(y_n, u_n) \leq 0$$

we get the convergence of u_n towards $\bar{u}$ in U.

Since every subsequence of the original sequence (y_n, u_n, v_n, ξ_n) possesses a subsequence with limit $(\bar{y}, \bar{u}, 0, \bar{\xi})$, we obtain the convergence of the whole sequence.

Due to the dense range of $\mathcal{C}$, the solutions to the regularized problems are strongly stationary, compare the system (4.3.1) with the (unregularized) system of strong stationarity (4.1.7) and (4.1.8).

Lemma 4.3.4. Let Assumption 4.1.2 be satisfied and let a regularization scheme $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ be given. We denote by (y_n, u_n, v_n, ξ_n) a local solution of $(\mathbf{P}_n^{\mathrm{reg}})$. Then, there exists $p_n \in H_0^1(\Omega)$, $\mu_n \in H^{-1}(\Omega)$ and $\lambda_n \in U$, such that

$$J_y(y_n, u_n) + \mu_n + \mathcal{A}^\star p_n = 0, \tag{4.3.1a}$$
$$J_u(y_n, u_n) + \beta\,(u_n - \bar{u}) + \lambda_n - \mathcal{B}^\star p_n = 0, \tag{4.3.1b}$$
$$\alpha_n\, v_n - \mathcal{C}^\star p_n = 0, \tag{4.3.1c}$$
$$\lambda_n \in \mathcal{N}_{U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u})}(u_n), \tag{4.3.1d}$$
$$-p_n \in \mathcal{K}(y_n, \xi_n), \tag{4.3.1e}$$
$$\mu_n \in \mathcal{K}(y_n, \xi_n)^\circ \tag{4.3.1f}$$

is satisfied. Moreover, these multipliers are unique.

Proof. The existence of the multipliers is shown in Section 1.6.1, see in particular Equation (1.6.4).

Since $\mathcal{C}$ is assumed to have a dense range, $\mathcal{C}^\star$ is injective. Hence, the uniqueness of p_n follows from (4.3.1c). The uniqueness of μ_n and λ_n follows from (4.3.1a) and (4.3.1b), respectively.

By proving the boundedness of the multipliers, we obtain their weak convergence.

Lemma 4.3.5. Let a regularization scheme $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ be given. We denote by (y_n, u_n, v_n, ξ_n) a global solution of $(\mathbf{P}_n^{\mathrm{reg}})$ and by p_n, μ_n, λ_n the multipliers satisfying (4.3.1). Then,

$$\|p_n\|_{H_0^1(\Omega)} + \|\mu_n\|_{H^{-1}(\Omega)} + \|\lambda_n\|_U \leq C$$

with $C > 0$ independent of n. In particular, there exist subsequences (denoted by the same symbol) such that $(p_n, \mu_n, \lambda_n) \rightharpoonup (p, \mu, \lambda)$ in $H_0^1(\Omega) \times H^{-1}(\Omega) \times U$.

Proof. We have $\langle \mu_n,\, p_n \rangle \geq 0$ by (4.3.1e) and (4.3.1f). By testing (4.3.1a) with p_n, and using the coercivity of $\mathcal{A}$, we find

$$\|p_n\|^2_{H^1_0(\Omega)} \leq C \, \|J_y(y_n, u_n)\|_{H^{-1}(\Omega)} \, \|p_n\|_{H^1_0(\Omega)}.$$

This shows the boundedness of p_n. The boundedness of λ_n, μ_n follows from (4.3.1a) and (4.3.1b). Since $H^1_0(\Omega)$, $H^{-1}(\Omega)$ and U are Hilbert spaces, we can choose a weakly convergent subsequence of (p_n, μ_n, λ_n).

4.4. Weak and C-stationarity of the limit point

In this section we pass to the limit with the optimality system (4.3.1). As a preparation, we need the following lemma.

Lemma 4.4.1. The mappings $v \mapsto v^+ = \max(v, 0)$ and $v \mapsto v^- = \max(-v, 0)$ are weakly sequentially continuous from $H^1_0(\Omega)$ to $H^1_0(\Omega)$.

Proof. Let a sequence $\{v_n\} \subset H^1_0(\Omega)$ with $v_n \rightharpoonup v$ in $H^1_0(\Omega)$ be given. By the compact embedding of $H^1_0(\Omega)$ in $L^2(\Omega)$, we infer $v_n \to v$ in $L^2(\Omega)$ and, hence, $v_n^+ \to v^+$ in $L^2(\Omega)$.

Since $\|v_n^+\|_{H^1_0(\Omega)} \leq \|v_n\|_{H^1_0(\Omega)}$, see Kinderlehrer, Stampacchia, 1980, Thm. II.A.1, each subsequence of $\{v_n^+\}$ has a weakly convergent subsequence with limit point $w \in H^1_0(\Omega)$. By $v_n^+ \to v^+$ in $L^2(\Omega)$, we infer $w = v^+$. This shows the weak convergence of $\{v_n^+\}$ towards v^+.

Similarly, one can show the weak convergence of $\{v_n^-\}$ towards v^-.

Now, we provide results which enable us to show that the weak limits of the multipliers p_n and μ_n satisfy the conditions (4.1.7g) and (4.1.7h) of the system of weak stationarity. Here, it is essential that we can work with the fine support f-supp(ξ) of ξ.

Lemma 4.4.2. We assume that the sequences $\{\xi_n\} \subset H^{-1}(\Omega)$, with $\xi_n \geq 0$, and $\{p_n\} \subset H^1_0(\Omega)$ satisfy $\xi_n \to \xi$ in $H^{-1}(\Omega)$ and $p_n \rightharpoonup p$ in $H^1_0(\Omega)$. Then, $p_n = 0$ q.e. on f-supp(ξ_n) for all $n \in \mathbb{N}$ implies $p = 0$ q.e. on f-supp(ξ).

Proof. In the following, if we state a condition involving $p^{\pm}_{(n)}$, we mean that this condition holds for both $p^+_{(n)} = \max(p_{(n)}, 0)$ and $p^-_{(n)} = \max(-p_{(n)}, 0)$. Since $p^{\pm}_n = 0$ q.e. on f-supp(ξ_n), we have

$$\langle \xi_n,\, p^{\pm}_n \rangle = \int_\Omega p^{\pm}_n \, \mathrm{d}\xi_n = 0.$$

By Lemma 4.4.1 we infer $p_n^{\pm} \rightharpoonup p^{\pm}$ in $H_0^1(\Omega)$. Passing to the limit in the above identity we get

$$\langle \xi, p^{\pm} \rangle = \int_{\Omega} p^{\pm} \, \mathrm{d}\xi = 0. \tag{4.4.1}$$

Since $p^{\pm} \geq 0$ q.e. in Ω, we infer $p^{\pm} \geq 0$ ξ-a.e. in Ω since ξ does not charges sets of capacity zero. Now, (4.4.1) implies $p^{\pm} = 0$ ξ-a.e. in Ω. By Lemma 3.A.5 we obtain $p^{\pm} = 0$ q.e. on f-supp(ξ) and $p = p^+ - p^- = 0$ q.e. on f-supp(ξ).

By applying Lemma 4.2.6 which uses results from potential theory, we find the condition on μ.

Lemma 4.4.3. We assume that the sequences $\{y_n\} \subset H_0^1(\Omega)$ and $\{\mu_n\} \subset H^{-1}(\Omega)$ satisfy $y_n \to y$ in capacity and $\mu_n \rightharpoonup \mu$ in $H^{-1}(\Omega)$. Then,

$$\mu_n \in \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } \{y_n = \psi\}\}^{\circ}$$

for all $n \in \mathbb{N}$ implies

$$\mu \in \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } \{y = \psi\}\}^{\circ}.$$

Proof. By assumption, we have

$$\pm\mu_n \in \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{y_n = \psi\}\}^{\circ}.$$

Now, we can apply Lemma 4.2.6 with $s_n = y_n - \psi$, $s = y - \psi$, $t_n = t = 0$ and $\kappa_n = \pm\mu_n$. This yields

$$\pm\,\mu \in \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{y = \psi\}\}^{\circ}.$$

The assertion follows.

Note that this lemma can be used to answer an open question raised after the proof of Outrata, Jarušek, Stará, 2011, Thm. 16.

Using these two lemmas and passing to the limit in (4.3.1) we obtain the system of weak stationarity (4.1.7).

Lemma 4.4.4. Let $(\bar{y}, \bar{u}, \bar{\xi})$ be a local minimizer of **(P)** and let the assumptions of Lemma 4.3.5 be satisfied. We denote by (p, μ, λ) a weak limit point of the multipliers as constructed in Lemma 4.3.5. Then, the weak stationarity system (4.1.7) is satisfied by the multipliers (p, μ, λ).

Proof. Since J is continuously differentiable, (4.1.7e) and (4.1.7f) are obtained by passing to the limit in (4.3.1a) and (4.3.1b) and using the strong convergence of (y_n, u_n) and the weak convergence of (p_n, μ_n, λ_n).

The condition (4.1.7g) follows from Lemma 4.4.2. From the regularized optimality system (4.3.1b), we obtain

$$\begin{aligned}\mu_n &\in \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } \{y_n = \psi\} \text{ and } v = 0 \text{ q.e. on f-supp}(\xi_n)\}^\circ \\ &\subset \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } \{y_n = \psi\}\}^\circ.\end{aligned}$$

Hence, we can apply Lemma 4.4.3 to infer (4.1.7h).

Since $\lambda_n \in \mathcal{N}_{U_{\mathrm{ad}} \cap B_\varepsilon(\bar u)}(u_n)$, we have

$$(\lambda_n,\, u - u_n) \le 0 \quad \text{for all } u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar u).$$

Passing to the limit $n \to \infty$ yields

$$(\lambda,\, u - \bar u) \le 0 \quad \text{for all } u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar u),$$

and hence, $\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(\bar u)$. This shows that the system (4.1.7) is satisfied.

In order to obtain the C-stationarity condition (4.1.9), we need some additional structure of the operator $\mathcal{A} : H_0^1(\Omega) \to H^{-1}(\Omega)$. Note that such a structural assumption was also used in Schiela, D. Wachsmuth, 2013, Lemma 3.6. We emphasize that this additional assumption on $\mathcal{A}$ is solely used in Lemma 4.4.5.

Lemma 4.4.5. Let $(\bar y, \bar u, \bar \xi)$ be a local minimizer of **(P)** and let the assumptions of Lemma 4.3.5 be satisfied. We denote by (p, μ, λ) a weak limit point of the multipliers as constructed in Lemma 4.3.5. In addition to the assumptions on the operator $\mathcal{A}$ made in Assumption 4.1.2, we suppose

$$\begin{aligned}\langle \mathcal{A}\, y,\, v\rangle = \int_\Omega \sum_{i,j=1}^n a_{ij}(x)\, \partial_i y(x)\, \partial_j v(x) + \sum_{i=1}^n a_i(x)\, \partial_i y(x)\, v(x) + \sum_{i=1}^n b_i(x)\, y(x)\, \partial_i v(x)\, \mathrm{d}x \\ + \int_\Omega a(x)\, y(x)\, v(x)\, \mathrm{d}x,\end{aligned}$$

where $a_{ij}, a_i, b_i, a \in L^\infty(\Omega)$ and

$$\sum_{i,j=1}^d a_{ij}(x) v_i\, v_j \ge \underline{a}\, \|v\|^2 \quad \text{for all } v \in \mathbb{R}^d \text{ and almost all } x \in \Omega$$

for some $\underline{a} > 0$. Then, the multipliers $p \in H_0^1(\Omega)$, $\mu \in H^{-1}(\Omega)$ satisfy

$$\langle \mu,\, \varphi\, p\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \ge 0 \quad \text{for all } \varphi \in W^{1,\infty}(\Omega), \varphi \ge 0. \tag{4.1.9}$$

Proof. Let $\varphi \in W^{1,\infty}(\Omega)$ with $\varphi \geq 0$ be given. From $-p_n \in \mathcal{K}(y_n, \xi_n)$ we infer also $-\varphi\, p_n \in \mathcal{K}(y_n, \xi_n)$, see (4.1.5). This yields

$$\langle \mu_n,\, \varphi\, p_n \rangle \geq 0.$$

By using the adjoint equation (4.3.1a), we find

$$J_y(y_n, u_n)\,(\varphi\, p_n) + \langle \mathcal{A}^\star p_n,\, \varphi\, p_n \rangle = -\langle \mu_n,\, \varphi\, p_n \rangle \leq 0.$$

Since J is assumed to be continuously Fréchet differentiable, the first term converges towards $J_y(\bar{y}, \bar{u})\,(\varphi\, p)$. By the assumption on $\mathcal{A}$ and the product rule, we find

$$\begin{aligned}
\langle \mathcal{A}^\star p_n,\, \varphi\, p_n \rangle &= \langle \mathcal{A}(\varphi\, p_n),\, p_n \rangle \\
&= \int_\Omega p_n(x) \sum_{i,j=1}^n a_{ij}(x)\, \partial_i \varphi(x)\, \partial_j p_n(x) + p_n(x) \sum_{i=1}^n a_i(x)\, \partial_i \varphi(x)\, p_n(x)\, \mathrm{d}x \\
&\quad + \int_\Omega \varphi(x) \sum_{i,j=1}^n a_{ij}(x)\, \partial_i p_n(x)\, \partial_j p_n(x) + \varphi(x) \sum_{i=1}^n a_i(x)\, \partial_i p_n(x)\, p_n(x)\, \mathrm{d}x \\
&\quad + \int_\Omega \sum_{i=1}^n b_i(x)\, \varphi(x)\, p_n(x)\, \partial_i p_n(x) + a(x)\, \varphi(x)\, p_n(x)\, p_n(x)\, \mathrm{d}x.
\end{aligned}$$

By using $p_n \rightharpoonup p$ in $H_0^1(\Omega)$, $p_n \to p$ in $L^2(\Omega)$ and the assumption on a_{ij}, we find

$$\langle \mathcal{A}^\star p,\, \varphi\, p \rangle \leq \liminf_{n\to\infty} \langle \mathcal{A}^\star p_n,\, \varphi\, p_n \rangle. \tag{4.4.2}$$

Together with the adjoint equation (4.1.7e), we obtain

$$\begin{aligned}
-\langle \mu,\, \varphi\, p \rangle &= J_y(\bar{y}, \bar{u})\,(\varphi\, p) + \langle \mathcal{A}^\star p,\, \varphi\, p \rangle \\
&\leq \lim_{n\to\infty} J_y(y_n, u_n)\,(\varphi\, p_n) + \liminf_{n\to\infty} \langle \mathcal{A}^\star p_n,\, \varphi\, p_n \rangle \leq 0.
\end{aligned}$$

The assumptions on $\mathcal{A}$ are only used to obtain the property (4.4.2) for all $\varphi \in W^{1,\infty}(\Omega)$, $\varphi \geq 0$. Hence, the result of Lemma 4.4.5 is still valid if we only require that $\mathcal{A}$ satisfies (4.4.2). Note that we still get

$$\langle \mu,\, p \rangle \geq 0$$

without any further assumptions on $\mathcal{A}$. This condition, however, is weaker than condition (4.1.9).

We compare our system of C-stationarity (4.1.7), (4.1.9) with the corresponding system in Schiela, D. Wachsmuth, 2013 for the case $U = L^2(\Omega)$. Since higher regularity is required to state that system, we suppose

$$\bar{y}, \psi \in C(\bar{\Omega}), \text{ and } \bar{\xi} \in L^2(\Omega). \tag{4.4.3}$$

They obtain the existence of multipliers $p \in H_0^1(\Omega) \cap L^\infty(\Omega)$, $\mu \in H^{-1}(\Omega) \cap C(\bar\Omega)^\star$, $\nu \in L^2(\Omega)$ satisfying (4.1.7e), (4.1.7f), (4.1.7i) and

$$p = 0 \quad \text{a.e. on } \{\bar\xi > 0\} \tag{4.4.4a}$$
$$\langle \mu,\, \phi\rangle = 0 \quad \text{for all } \phi \in C(\bar\Omega) \text{ with } \phi = 0 \text{ on } \{\bar y = \psi\} \tag{4.4.4b}$$
$$\langle \mu,\, \varphi\, p\rangle \geq 0 \quad \text{for all } \varphi \in W^{1,\infty}(\Omega) \text{ with } \varphi \geq 0 \tag{4.4.4c}$$

see Schiela, D. Wachsmuth, 2013, Prop. 3.5–3.8. Condition (4.4.4c) coincides with (4.1.9). The following lemma demonstrates that both systems of C-stationarity are equivalent.

Lemma 4.4.6. Let $\bar\xi \in L^2(\Omega)$, with $\bar\xi \geq 0$, and $p \in H_0^1(\Omega)$ be given. Then, (4.4.4a) is equivalent to (4.1.7g).

Let $\bar y, \psi \in C(\bar\Omega)$ and $\mu \in H^{-1}(\Omega) \cap C(\bar\Omega)^\star$ be given. Then, (4.1.7h) is equivalent to the existence of $\tilde\mu \in H^{-1}(\Omega) \cap C(\bar\Omega)^\star$ satisfying (4.4.4b) and $\langle \mu,\, v\rangle = \langle \tilde\mu,\, v\rangle$ for all $v \in H_0^1(\Omega)$.

Proof. We first consider the equivalence of (4.4.4a) and (4.1.7g). Since $\bar\xi \in L^2(\Omega)$ is non-negative, we find

$$\int_{\{p\neq 0\}\cap\{\bar\xi>0\}} \mathrm{d}x = 0 \quad \Leftrightarrow \quad \int_{\{p\neq 0\}} \bar\xi \,\mathrm{d}x = 0.$$

Hence, (4.4.4a) is equivalent to

$$p = 0 \quad \bar\xi\text{-a.e. in } \Omega.$$

By Lemma 3.A.5 we obtain the equivalence of (4.4.4a) and (4.1.7g).

Now, we consider the second equivalence. For convenience, we recall

$$\langle \mu,\, \phi\rangle = 0 \quad \text{for all } \phi \in H_0^1(\Omega) \text{ with } \phi = 0 \text{ q.e. on } \{\bar y = \psi\}, \tag{4.1.7h}$$
$$\langle \tilde\mu,\, \phi\rangle = 0 \quad \text{for all } \phi \in C(\bar\Omega) \text{ with } \phi = 0 \text{ on } \{\bar y = \psi\}. \tag{4.4.4b}$$

The active set $A = \{x \in \Omega : \bar y(x) = \psi(x)\}$ is relatively closed in Ω. We find

$$\phi \in H_0^1(\Omega \setminus A) \quad \Leftrightarrow \quad \phi \in H_0^1(\Omega) \text{ and } \phi = 0 \text{ q.e. on } A,$$
$$\phi \in C_0(\Omega \setminus A) \quad \Leftrightarrow \quad \phi \in C_0(\Omega) \text{ and } \phi = 0 \text{ on } A,$$

for the first equivalence we refer to Heinonen, Kilpeläinen, Martio, 1993, Thm. 4.5. Since $C_0(\Omega\setminus A)\cap H_0^1(\Omega\setminus A)$ is dense in $H_0^1(\Omega\setminus A)$ and $C_0(\Omega\setminus A)$ (in their respective norms), see, e.g., Fukushima, Ōshima, Takeda, 1994, p.100, we immediately obtain the equivalence of (4.1.7h) and

$$\langle \mu,\, \phi\rangle = 0 \quad \text{for all } \phi \in C_0(\Omega) \text{ with } \phi = 0 \text{ on } \{\bar y = \psi\}. \tag{4.1.7h'}$$

In order to prove "$\Leftarrow$", let $\tilde\mu \in H^{-1}(\Omega) \cap C(\bar\Omega)^\star$ satisfying (4.4.4b) and $\langle \mu,\, v\rangle = \langle \tilde\mu,\, v\rangle$ for $v \in H_0^1(\Omega)$ be given. Since $\tilde\mu$ satisfies (4.4.4b), it also satisfies (4.1.7h'). By using

the above density argument, $\tilde{\mu}$ satisfies (4.1.7). Since $\langle \mu, v\rangle = \langle \tilde{\mu}, v\rangle$ for $v \in H_0^1(\Omega)$, the functional μ also satisfies (4.1.7h).

It remains to prove the converse. Since μ is assumed to satisfy (4.1.7h), we have (4.1.7h'). By the Riesz representation theorem, see, e.g., Rudin, 1987, Thm. 6.19, we have $C_0(\Omega)^\star = \mathcal{M}(\Omega)$ and $C(\bar{\Omega}) = C_0(\bar{\Omega}) = \mathcal{M}(\bar{\Omega})$, where $\mathcal{M}$ denotes the space of regular, signed Borel measures with bounded variation.

Now, we define $\tilde{\mu} \in \mathcal{M}(\bar{\Omega})$ by $\tilde{\mu}(B) = \mu(B \cap \Omega)$ for all Borel sets $B \subset \bar{\Omega}$.

Let us show that $\tilde{\mu}$ satisfies (4.4.4b). We pick an increasing sequence $\{\Omega_n\}$ of open subsets of Ω, such that $\bar{\Omega}_n \subset \Omega$, and $\Omega = \cup_{n=1}^\infty \Omega_n$. Since μ is countably additive, this yields $\mu(\Omega \setminus \Omega_n) \to 0$. Moreover, for each $n \in \mathbb{N}$ there is $\chi_n \in C_0(\Omega)$ with $0 \le \chi_n \le 1$ and $\chi_n = 1$ on Ω_n.

Now, let $\phi \in C(\bar{\Omega})$ with $\phi = 0$ on A be given. Since $\phi\,\chi_n \in C_0(\Omega)$, we have $\langle \mu, \phi\,\chi_n\rangle = 0$ by (4.1.7h'). We find

$$\begin{aligned}\left|\int_{\bar{\Omega}} \phi - \phi\,\chi_n \,\mathrm{d}\tilde{\mu}\right| &= \left|\int_{\bar{\Omega}\setminus\Omega_n} \phi\,(1-\chi_n)\,\mathrm{d}\tilde{\mu}\right| \\ &\le \int_{\bar{\Omega}\setminus\Omega_n} |\phi|\,\mathrm{d}\tilde{\mu} \le \|\phi\|_{C(\bar{\Omega})}\,\tilde{\mu}(\bar{\Omega}\setminus\Omega_n) = \|\phi\|_{C(\bar{\Omega})}\,\mu(\Omega\setminus\Omega_n) \to 0.\end{aligned}$$

Hence,

$$\langle \tilde{\mu}, \phi\rangle = \int_{\bar{\Omega}} \phi\,\mathrm{d}\tilde{\mu} = \lim_{n\to\infty} \int_{\bar{\Omega}} \phi\,\chi_n\,\mathrm{d}\tilde{\mu} = \lim_{n\to\infty} \langle \mu, \phi\,\chi_n\rangle = 0.$$

This yields (4.4.4b).

Since $\langle \tilde{\mu}, \phi\rangle = \langle \mu, \phi\rangle$ for all $\phi \in C_0^\infty(\Omega)$, we can extend $\tilde{\mu}$ continuously to $H_0^1(\Omega)$ and we obtain $\langle \tilde{\mu}, v\rangle = \langle \mu, v\rangle$ for all $v \in H_0^1(\Omega)$.

4.5. M-stationarity of the limit point

In order to obtain M-stationarity conditions in the limit, we need some additional information. Therefore, we suppose that the regularization scheme $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ satisfies the following assumption.

Assumption 4.5.1. Let Assumption 4.1.2 be satisfied and let $(\bar{y}, \bar{u}, \bar{\xi})$ be a local minimizer of (**P**). We denote by $(V, \mathcal{C}, \{\alpha_n\}, \beta)$ a regularization scheme. We assume that the multipliers (p_n, μ_n, λ_n) associated with a local minimizer (y_n, u_n, v_n, ξ_n) of ($\mathbf{P}_n^{\mathrm{reg}}$) by Lemma 4.3.4 converge weakly towards (p, μ, λ) in $H_0^1(\Omega) \times H^{-1}(\Omega) \times U$ and that p_n converges towards p in capacity, that is, for every $\varepsilon > 0$

$$\mathrm{cap}(\{|p_n - p| \ge \varepsilon\}) \to 0$$

holds.

We recall that each regularization scheme has a subsequence, such that p_n, μ_n and U converge weakly, see Lemma 4.3.5. The crucial assumption is that p_n converges in capacity.

For convenience, we recall the following relations from Lemma 4.3.3 and Lemma 4.3.4

$$y_n \in K, \qquad y_n \to \bar{y} \text{ in } H_0^1(\Omega), \tag{4.5.1a}$$
$$\xi_n \in \mathcal{T}_K(y_n)^\circ, \qquad \xi_n \to \bar{\xi} \text{ in } H^{-1}(\Omega), \tag{4.5.1b}$$
$$p_n \in -\mathcal{K}(y_n, \xi_n), \qquad p_n \rightharpoonup p \text{ in } H_0^1(\Omega), \tag{4.5.1c}$$
$$\mu_n \in \mathcal{K}(y_n, \xi_n)^\circ, \qquad \mu_n \rightharpoonup \mu \text{ in } H^{-1}(\Omega). \tag{4.5.1d}$$

Now, we define the sets

$$\hat{I} = \{p < 0\} \cap \{\bar{y} = \psi\},$$
$$\hat{B} = \{p > 0\} \cap \{\bar{y} = \psi\},$$
$$\hat{A}_s = \{p = 0\} \cap \{\bar{y} = \psi\} \cap B.$$

We show that these sets form a disjoint partition of B (up to a set of zero capacity) as required by our definition of M-stationarity (4.1.11). Since $p = 0$ q.e. on A_s by Lemma 4.4.2 and $\{\bar{y} < \psi\} = I$, we find $\hat{I}, \hat{B} \subset B$. Hence, all three sets $\hat{I}, \hat{B}, \hat{A}_s$ are subsets of B, they are obviously disjoint and

$$\hat{I} \cup \hat{B} \cup \hat{A}_s = \{p \in \mathbb{R}\} \cap \{\bar{y} = \psi\} \cap B = B.$$

As in (4.1.12), we set

$$\hat{\mathcal{K}} = \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } \hat{B} \text{ and } v = 0 \text{ q.e. on } A_s \cup \hat{A}_s\}.$$

By definition of the sets $\hat{I}, \hat{B}, \hat{A}_s$, we have $p \in -\hat{\mathcal{K}}$. It remains to show $\mu \in \hat{\mathcal{K}}^\circ$. By decomposing $v \in \hat{\mathcal{K}}$ into its positive and negative part, we find

$$\begin{aligned}\hat{\mathcal{K}} = {} & \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } I \cup \hat{I} \text{ and } v = 0 \text{ q.e. on } \hat{B} \cup A_s \cup \hat{A}_s\} \\ & + \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } I \cup \hat{I} \cup \hat{B} \text{ and } v = 0 \text{ q.e. on } A_s \cup \hat{A}_s\}.\end{aligned}$$

Hence, its polar is given by

$$\begin{aligned}\hat{\mathcal{K}}^\circ = {} & \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } I \cup \hat{I} \text{ and } v = 0 \text{ q.e. on } \hat{B} \cup A_s \cup \hat{A}_s\}^\circ \\ & \cap \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } I \cup \hat{I} \cup \hat{B} \text{ and } v = 0 \text{ q.e. on } A_s \cup \hat{A}_s\}^\circ.\end{aligned}$$

Now, we can verify that μ belongs to both sets on the right-hand side.

Lemma 4.5.2. Let Assumption 4.5.1 be satisfied. We have

$$\mu \in \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } I \cup \hat{I} \text{ and } v = 0 \text{ q.e. on } \hat{B} \cup A_s \cup \hat{A}_s\}^\circ.$$

Proof. From the regularized optimality system (4.3.1b), see also (4.5.1d), we obtain

$$\begin{aligned}\mu_n &\in \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } \{y_n = \psi\} \text{ and } v = 0 \text{ q.e. on f-supp}(\xi_n)\}^\circ \\ &\subset \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{y_n = \psi\}\}^\circ.\end{aligned}$$

Since $p_n \ge 0$ q.e. on $\{y_n = \psi\}$, this yields

$$\begin{aligned}\mu_n &\in \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{y_n = \psi\} \cap \{p_n \ge 0\}\}^\circ \\ &= \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{y_n = \psi\} \cap \{\min(p_n, 0) = 0\}\}^\circ.\end{aligned}$$

By Assumption 4.5.1 $\min(p_n, 0)$ converges towards $\min(p, 0)$ in capacity. Hence, we can apply Lemma 4.2.6 with the setting $\kappa_n = \mu_n$, $s_n = y_n - \psi$, $t_n = \min(p_n, 0)$. This yields

$$\mu \in \{v \in H_0^1(\Omega) : v \ge 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{\bar{y} = \psi\} \cap \{\min(p, 0) = 0\}\}^\circ$$

which is the assertion.

Lemma 4.5.3. Let Assumption 4.5.1 be satisfied. We have

$$\mu \in \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } I \cup \hat{I} \cup \hat{B} \text{ and } v = 0 \text{ q.e. on } A_s \cup \hat{A}_s\}^\circ.$$

Proof. From the regularized optimality system (4.3.1b), see also (4.5.1d), we obtain

$$\begin{aligned}\mu_n &\in \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } \{y_n = \psi\} \text{ and } v = 0 \text{ q.e. on f-supp}(\xi_n)\}^\circ \\ &\subset \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{y_n = \psi\} \cap \{p_n = 0\}\}^\circ.\end{aligned}$$

By applying Lemma 4.2.6, we find

$$\mu \in \{v \in H_0^1(\Omega) : v \le 0 \text{ q.e. on } \Omega \text{ and } v = 0 \text{ q.e. on } \{\bar{y} = \psi\} \cap \{p = 0\}\}^\circ,$$

which is the assertion.

Altogether, we have proved the following theorem.

Theorem 4.5.4. Let us denote by $(\bar{y}, \bar{u}, \bar{\xi})$ a local minimizer of **(P)**. Moreover, we assume that there exists a regularization scheme satisfying Assumption 4.5.1. Then there exists multipliers $(p, \mu, \lambda) \in H_0^1(\Omega) \times H^{-1}(\Omega) \times U$, such that the M-stationarity system (4.1.7), (4.1.11) is satisfied.

Proof. The assertion follows from Lemma 4.4.4, Lemma 4.5.2 and Lemma 4.5.3.

We emphasize that in the one-dimensional case $d = 1$, Assumption 4.5.1 is automatically satisfied by Lemma 4.2.3. Hence, we reproduced the result Outrata, Jarušek, Stará, 2011,

Thm. 14. Note that our setting is slightly different from Outrata, Jarušek, Stará, 2011. In particular, we required compactness of the control operator $\mathcal{B} : U \to H^{-1}(\Omega)$ (or of U_{ad}), whereas Outrata, Jarušek, Stará, 2011 use $U = H^{-1}(\Omega)$ and $\mathcal{B} = I$, which is not compact.

It is also interesting to have a look on the contrapositive of Theorem 4.5.4. Let us suppose that we have a local minimizer $(\bar{y}, \bar{u}, \bar{\xi})$ of **(P)**, which is *not* M-stationary, i.e., there do not exist multipliers (p, μ, λ) satisfying (4.1.7), (4.1.11). Then, for *any* regularization scheme, no subsequence of the sequence $\{p_n\}$ of multipliers (associated with the regularized solutions) can converge in capacity. In view of Lemma 4.3.5 and Section 4.2, there are subsequences, which converge weakly in $H_0^1(\Omega)$, but no subsequence can converge strongly in $H_0^1(\Omega)$ and no subsequence can be bounded in $W_0^{1,p}(\Omega)$ for any $p > 2$. Finally, we mention that it is not trivial to construct a sequence $\{p_n\} \subset H_0^1(\Omega)$, which converges weakly in $H_0^1(\Omega)$, but not in capacity, see the construction in the next section.

4.6. A counterexample

In the previous section, we have shown the necessity of M-stationarity conditions in the case Assumption 4.5.1 is satisfied. We emphasize that we only used the relations (4.5.1) in addition to Assumption 4.5.1, in order to obtain the sign conditions (4.1.11) on p and μ.

In this section, we construct sequences $\{y_n\}$, $\{\xi_n\}$, $\{p_n\}$, $\{\mu_n\}$ satisfying (4.5.1), but the limits p, μ do not satisfy (4.1.11) for any choice of $\hat{I}, \hat{B}, \hat{A}_s$. This shows that the Assumption 4.5.1 is crucial for our technique of proof.

In order to construct our counterexample, we use results from Cioranescu, Murat, 1997 and limit ourselves to the case $\mathcal{A} = -\Delta$, $\psi \in H_0^1(\Omega)$. We choose $d > 1$, since in case $d = 1$, Assumption 4.5.1 is always satisfied by Lemma 4.2.3.

We construct a perforated domain, as described in Cioranescu, Murat, 1997, Ex. 2.1. That is, we choose a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}} \subset (0, \infty)$ with $\varepsilon_n \to 0$ and set

$$r_n = \begin{cases} \exp^{-\varepsilon_n^{-2}} & \text{if } d = 2, \\ \varepsilon_n^{d/(d-2)} & \text{if } d > 2. \end{cases}$$

For each $\boldsymbol{i} \in \mathbb{Z}^d$, let $T_{\boldsymbol{i}}^n = B_{r_n}(\varepsilon_n \boldsymbol{i})$ be the closed ball with radius r_n centered at $\varepsilon_n \boldsymbol{i}$. Now, the perforated domain is given by

$$\Omega_n = \Omega \setminus \bigcup_{\boldsymbol{i}\in\mathbb{Z}^d} T_{\boldsymbol{i}}^n.$$

As $n \to \infty$, both the distance ε_n and the radius r_n of the holes go to 0.

Now we define $p_n \in H_0^1(\Omega_n)$ as the weak solution of

$$\begin{aligned} -\Delta p_n &= -1 \quad \text{in } \Omega_n, \\ p_n &= 0 \quad \text{on } \partial\Omega_n, \end{aligned}$$

and extend p_n by 0 to a function in $H_0^1(\Omega)$. By Cioranescu, Murat, 1997, Thms. 1.2, 2.2, p_n converges weakly in $H_0^1(\Omega)$ towards the weak solution $p \in H_0^1(\Omega)$ of

$$-\Delta p + \kappa\, p = -1 \quad \text{in } \Omega, \tag{4.6.1a}$$
$$p = 0 \qquad \text{on } \partial\Omega, \tag{4.6.1b}$$

for some $\kappa > 0$. For the precise value of κ, we refer to Cioranescu, Murat, 1997, eq. (2.3).

In order to verify (4.5.1), we choose sequences $\{y_n\} \subset K$, $\{\xi_n\} \subset H^{-1}(\Omega)$, satisfying

$$\xi_n \in \mathcal{T}_K(y_n)^\circ, \qquad y_n \to \bar{y} = \psi \quad \text{in } H_0^1(\Omega),$$
$$\{y_n = \psi\} = \text{f-supp}(\xi_n) = \Omega \setminus \Omega_n, \qquad \xi_n \to \bar{\xi} = 0 \quad \text{in } H^{-1}(\Omega).$$

One possible choice would be

$$y_n = \psi + \frac{1}{n}\, p_n, \quad \xi_n = \frac{1}{n}\, \chi_{\Omega\setminus\Omega_n}.$$

The construction of y_n, ξ_n yields

$$\begin{aligned}\mathcal{K}(y_n, \xi_n) &= \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \{y_n = \psi\} \text{ and } v = 0 \text{ q.e. on f-supp}(\xi_n)\} \\ &= \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } \Omega \setminus \Omega_n\} = H_0^1(\Omega_n).\end{aligned}$$

By definition of p_n, we have

$$p_n \in -\mathcal{K}(y_n, \xi_n) = H_0^1(\Omega_n).$$

Now, we define $\mu_n \in H^{-1}(\Omega)$ by

$$\langle \mu_n,\, v\rangle = -\int_\Omega \nabla p_n\, \nabla v + v\, \mathrm{d}x,$$

i.e., $\mu_n = -\mathcal{A}\, p_n - 1$. By definition of p_n, we obtain

$$\langle \mu_n,\, v\rangle = -\int_\Omega \nabla p_n\, \nabla v + v\, \mathrm{d}x = 0 \qquad \text{for all } v \in H_0^1(\Omega_n) = \mathcal{K}(y_n, \xi_n).$$

Hence, $\mu_n \in \mathcal{K}(y_n, \xi_n)^\circ$.

By passing to the limit with $\mu_n = -\mathcal{A}\, p_n - 1$, we obtain that $\mu_n \rightharpoonup \mu = -\mathcal{A}\, p - 1$. This yields $\mu = \kappa\, p$. Now, all conditions in (4.5.1) are verified.

It remains to show that the limit p, μ does not satisfy (4.1.11). Since p is the solution of (4.6.1), we obtain the interior regularity $p \in C^2(\Omega)$, see Evans, 1998, Thm. 6.3.3. The maximum principle yields $p \leq 0$ in Ω. We will even show $p < 0$ in Ω. To the contrary, assume that for some $x \in \Omega$ we have $p(x) = 0$. Then, x is a local maximizer of p and thus the Hessian of p at x is negative semi-definite. This is a contradiction to $-\Delta p(x) = -1$. This shows that $p(x) < 0$ for all $x \in \Omega$.

Hence, in order to satisfy (4.1.11a), we have to choose

$$\hat{I} = B = \Omega \quad \text{and} \quad \hat{B} = \hat{A}_s = \emptyset.$$

This gives $\hat{\mathcal{K}} = H_0^1(\Omega)$ by (4.1.12). But then, (4.1.11b) requires that $\mu = 0$, which does not hold. This shows that the conclusions of Lemma 4.5.2 and Lemma 4.5.3 are violated, and, in particular, Assumption 4.5.1 cannot hold. Note that even the adjoint equations (4.1.7e) and (4.3.1a) are satisfied, if we choose J such that $J_y \equiv 1$ holds.

However, it is unclear whether such sequences y_n, ξ_n, p_n, μ_n may actually arise as solutions and multipliers of the regularized problem ($\mathbf{P}_n^{\mathrm{reg}}$). It remains an open question if all minimizers of (**P**) satisfy the system of M-stationarity. Since Assumption 4.5.1 is not too strong and has to be satisfied only for one particular regularization scheme, see Theorem 4.5.4, it is very reasonable that the answer to the above question is affirmative. We are also not aware of any counterexamples of a problem like (**P**) with an optimal solution which is not M-stationary.

4.7. Conclusions and perspectives

We have derived optimality conditions for the problem (**P**). By using results from potential theory, we were able to work with the basic regularity of the obstacle problem. In particular, Theorem 4.2.5 is a crucial ingredient of Lemma 4.2.6 and this lemma was used to obtain the sign conditions on the multiplier μ in Lemmas 4.4.4, 4.5.2 and 4.5.3. Also the technique used in Lemma 4.4.3 to derive the condition on the adjoint state p seems to be new.

Under the assumption that the adjoint states p_n converge in capacity towards p, we were able to derive a system of M-stationarity. It is, however, unclear how to prove this assumption. We are also not aware of any counterexamples which violate the system of M-stationarity. Hence, it remains an open question whether all minimizers of (**P**) are M-stationary.

Let us give some comments on generalizations of our results. The key lemmas which are used to derive the optimality systems (e.g. Lemmas 4.4.2, 4.4.3, 4.5.2 and 4.5.3) do not depend on the operators $\mathcal{A}$ and $\mathcal{B}$. Hence, it is possible to transfer the results to, e.g., nonlinear operators $\mathcal{A}$.

The generalization to other boundary conditions (e.g., by replacing $H_0^1(\Omega)$ by $H^1(\Omega)$) seems to be more technical. An important ingredient is Theorem 4.2.5. Hence, one has to find a proper replacement for Theorem 4.2.5 which holds in $H^1(\Omega)$.

Part III.

Polyhedricity of convex sets

Contents of Part III

Introduction

In this part we consider polyhedric sets. A closed convex set K is called polyhedric w.r.t. $x \in K$ and $\mu \in \mathcal{T}_K(x)^\circ$ if

$$\mathcal{T}_K(x) \cap \mu^\perp = \mathrm{cl}(\mathcal{R}_K(x) \cap \mu^\perp)$$

holds. Here, $\mathcal{T}_K(x)$ and $\mathcal{R}_K(x)$ are the tangent cone and radial cone of K at x, respectively. We refer to Section 5.2 for the notation.

It can be easily seen that polyhedral sets are polyhedric. Moreover, in infinite dimensions there are many examples for polyhedric sets which are not polyhedral.

Polyhedricity of the set K has many important applications in infinite-dimensional optimization theory. These applications are described in Section 5.5 and they generalize corresponding results for polyhedral sets.

Thus, polyhedricity can be seen as a generalization of polyhedrality to infinite-dimensional spaces. Due to the importance of the applications, it is worthwhile to study the geometry of polyhedric sets.

In Chapter 5 we review known results concerning polyhedricity. Moreover, we provide new insights concerning the intersection of polyhedric sets. Finally, we address the polyhedricity of sets in vector-valued Sobolev spaces with pointwise constraints in Chapter 6.

5. A guided tour of polyhedric sets – Basic properties, new results on intersections and applications

Abstract: The aim of this contribution is twofold. On the one hand, we give some new results concerning polyhedric sets. In particular, we show that sets with pointwise lower and upper bound are polyhedric in many important function spaces. Moreover, we show that the intersection of such a set with finitely many hyperplanes and half-spaces is polyhedric. We also provide counterexamples demonstrating that the intersection of polyhedric sets may fail to be polyhedric. On the other hand, we gather all important results from the literature concerning polyhedric sets in order to give a complete picture of the current knowledge. In particular, we illustrate the applications of polyhedricity.

Keywords: polyhedricity, polyhedric set, directional differentiability, projection, vector lattice, strong stationarity, second-order conditions

MSC: 49K21, 46A55, 46N10

5.1. Introduction

The notion of polyhedricity of closed convex sets was first used in the seminal works Mignot, 1976; Haraux, 1977. Haraux, 1977 also coined the term "polyhedricity". In these two works, the authors have shown that polyhedricity of a closed convex set in a Hilbert space implies that the metric projection onto this set (or equivalently, the solution map of a variational inequality) is directionally differentiable. Moreover, Mignot, 1976 verified a stationarity system for the optimal control of a variational inequality.

Since then, the notion of polyhedricity found some very important applications in infinite-dimensional optimization. In particular, polyhedricity helps to provide no-gap second-order optimality conditions in the infinite-dimensional case.

However, all the available results concerning polyhedricity are scattered in the literature and it is difficult to get an overview on the current knowledge. Hence, one important goal of the present paper is to bridge this gap. In particular, we collect sets which are known to be polyhedric, see Section 5.3.2, and we provide some known properties in Section 5.3.3.

On the other hand, we study the question whether the intersection of two polyhedric sets is again polyhedric and this is addressed in Section 5.4. We mention that we obtain

completely new results, which generalize the results of Mignot, 1976; Haraux, 1977 concerning the polyhedricity of sets with bounds in vector lattices. We do not need a Dirichlet space setting as in Mignot, 1976 and, thus, our assumptions are easier to verify. In Section 5.4.3 we give some counterexamples which demonstrate that intersections of polyhedric sets may fail to be polyhedric and we present a polyhedric sets in finite dimensions which is not polyhedral.

Finally, we give the most important applications of polyhedricity, see Section 5.5. In particular, we prove the directional differentiability of the projection onto polyhedric sets, strong stationarity for optimal control of the projection, and provide no-gap second-order conditions in the polyhedric case.

5.2. Notation

We denote by $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ the natural numbers.

Let $A \subseteq X$ be a subset of the real Banach space X. We denote by $\operatorname{cl} A$, $\operatorname{conv} A$, $\operatorname{cone} A$, $\overline{\operatorname{conv}} A$, and $\operatorname{lin} A$ the closure, the convex hull, the (convex) conic hull, the closed convex hull and the linear hull of A, respectively. By $X^\star$ we denote the (topological) dual space of X with corresponding dual pairing $\langle \cdot, \cdot \rangle \colon X^\star \times X \to \mathbb{R}$. We define the polar cone and the polar set of A by

$$A^\circ := \{x^\star \in X^\star \mid \forall x \in A : \langle x^\star, x \rangle \leq 0\}, \qquad A^\square := \{x^\star \in X^\star \mid \forall x \in A : \langle x^\star, x \rangle \leq 1\}.$$

Similarly, we use

$$B^\circ := \{x \in X \mid \forall x^\star \in B : \langle x^\star, x \rangle \leq 0\}, \qquad B^\square := \{x \in X \mid \forall x^\star \in B : \langle x^\star, x \rangle \leq 1\}.$$

for any non-empty set $B \subseteq X^\star$. The closure of B w.r.t. the weak-$\star$ topology of $X^\star$ is denoted by $\operatorname{cl}_\star$.

For a functional $\mu \in X^\star$ we denote the annihilator by

$$\mu^\perp := \{x \in X \mid \langle \mu, x \rangle = 0\}.$$

The radial cone and the tangent cone of a closed, convex set $K \subset X$ at $x \in K$ are given by

$$\mathcal{R}_K(x) := \operatorname{cone}(K - x), \qquad \mathcal{T}_K(x) := \operatorname{cl} \mathcal{R}_K(x).$$

Finally, the critical cone of K w.r.t. $(x, \mu) \in K \times \mathcal{T}_K(x)^\circ$ is given by

$$\mathcal{K}_K(x, \mu) := \mathcal{T}_K(x) \cap \mu^\perp.$$

5.3. Definition, examples and basic properties

5.3.1. Definition

We start with the definition of polyhedricity and related concepts.

Definition 5.3.1. Let K be a closed, convex subset of the Banach space X.

(a) We say that K is *polyhedric* at (x, μ) with $x \in K$, $\mu \in \mathcal{T}_K(x)^\circ$, if

$$\mathcal{T}_K(x) \cap \mu^\perp = \mathrm{cl}(\mathcal{R}_K(x) \cap \mu^\perp). \tag{5.3.1}$$

We say that K is polyhedric at $x \in K$, if (5.3.1) holds for all $\mu \in \mathcal{T}_K(x)^\circ$. Finally, K is polyhedric, if it is polyhedric at all $x \in K$.

(b) The set K is called *polyhedral*, if it is the intersection of finitely many half-spaces, i.e., if there exist $n \in \mathbb{N}_0$, $\nu_1, \dots, \nu_n \in X^\star$, and $c_1, \dots, c_n \in \mathbb{R}$ such that

$$K = \{x \in X \mid \langle \nu_i, x_i \rangle \leq c_i \ \forall i = 1, \dots, n\}. \tag{5.3.2}$$

(c) We call K *co-polyhedral*, if it is the convex hull of finitely many points and rays, i.e., if there exist $n, m \in \mathbb{N}_0$, $x_1, \dots, x_n \in X$, $r_1, \dots, r_m \in X$ such that

$$K = \mathrm{conv}\{x_1, \dots, x_n\} + \mathrm{cone}\{r_1, \dots, r_m\}.$$

(d) Finally, K is said to be *generalized polyhedral*, if it is the intersection of a polyhedral set with a closed, affine subspace.

It is well-known that the notions of polyhedrality, co-polyhedrality and generalized polyhedrality coincide in finite-dimensional spaces, see, e.g., Klee, 1959, Theorem 2.12 or Ziegler, 1995, Theorem 1.2.

In infinite dimensions, the concepts of polyhedrality and co-polyhedrality are dual. Indeed, if K satisfies (5.3.2), then we have

$$K^\square = \mathrm{conv}\Big(\{0\} \cup \Big\{\frac{\nu_i}{c_i}\Big\}_{\{i|c_i>0\}}\Big) + \mathrm{cone}\{\nu_i\}_{\{i|c_i\leq 0\}}.$$

Hence, $K^\square$ is co-polyhedral. Similarly, the polar set of a co-polyhedral set is polyhedral.

Now we discuss the notion of polyhedricity, which was introduced in Mignot, 1976; Haraux, 1977. From $\mathcal{R}_K(x) \subset \mathcal{T}_K(x)$ it is clear that the left-hand side of (5.3.1) always contains the right-hand side. Moreover, (5.3.1) follows trivially if $\mathcal{R}_K(x) = \mathcal{T}_K(x)$ and it is straightforward to check that this is the case for polyhedral, co-polyhedral and generalized polyhedral sets.

We will see in Section 5.3.2 that there are many polyhedric sets that are not generalized polyhedral if the space X is infinite-dimensional. In finite dimensions, one is tempted to

conjecture that polyhedric sets are polyhedral. However, one has to restrict this question to bounded polyhedric sets or broaden the notion of polyhedrality, since polyhedricity is a local property whereas polyhedrality is global. Indeed, the set

$$\operatorname{conv}\{(s, s^2)\}_{s\in\mathbb{Z}}$$

is polyhedric, but has countably infinitely many vertices. Thus, it is not polyhedral. In Example 5.4.24 we will present a compact, convex and polyhedric set in $\mathbb{R}^3$ which fails to be a polyhedron. To our knowledge, such an example was not known previously.

5.3.2. Examples

Now, we give the most important examples for polyhedric sets in infinite-dimensional spaces, which can be found in the literature.

A broad class of examples can be found in Banach spaces which possess additionally a lattice structure with strongly-weakly continuous lattice operations, see Section 5.4.2 for references and further details. In fact, this yields the polyhedricity of the sets

$$\begin{aligned}
&\{u \in L^p(\Omega) \mid u_a \le u \le u_b \text{ a.e. in } \Omega\} && \text{in } L^p(\Omega),\ p \in [1,\infty],\ u_a, u_b \in L^p(\Omega),\\
&\{u \in W_0^{1,p}(\Omega) \mid u_a \le u \le u_b \text{ a.e. in } \Omega\} && \text{in } W_0^{1,p}(\Omega),\ p \in [1,\infty),\ u_a, u_b \in W_0^{1,p}(\Omega),\\
&\{u \in W^{1,p}(\Omega) \mid u_a \le u \le u_b \text{ a.e. in } \Omega\} && \text{in } W^{1,p}(\Omega),\ p \in [1,\infty),\ u_a, u_b \in W^{1,p}(\Omega),\\
&\{u \in W^{1,p}(\Omega) \mid u_a \le u \le u_b \text{ a.e. on } \partial\Omega\} && \text{in } W^{1,p}(\Omega),\ p \in [1,\infty),\ u_a, u_b \in W^{1,p}(\Omega).
\end{aligned}$$

In all cases, $\Omega \subset \mathbb{R}^d$ is an open, bounded set. In the last case, we have to assume additionally that Ω is a Lipschitz domain in order to get a well-defined trace and the notion "a.e." refers to the surface measure on the boundary $\partial\Omega$. Such sets where already studied in Mignot, 1976; Haraux, 1977, at least in the Hilbert space case $p = 2$. The regularity requirements on u_a, u_b can be significantly reduced. We refer to Section 5.4.2 for details and to Example 5.4.21 for more polyhedric sets defined by (pointwise) bounds.

The pointwise ordering on $H_0^2(\Omega)$ does not induce a lattice structure. Nevertheless, it is shown in Rao, Sokołowski, 1993 that subsets $K \subset H_0^2(\Omega)$ which are defined by pointwise constraints are polyhedric at some $u \in K$, but not at all $u \in K$. Similarly, $H^{-1/2}(\Omega)$ does not posses a lattice structure, but Sokołowski, 1988, Lemma 1 shows the polyhedricity of the set

$$\{\mu \in H^{-1/2}(\partial\Omega) \mid \mu \in L^\infty(\partial\Omega) \text{ and } -1 \le \mu \le 1 \text{ a.e. on } \partial\Omega\}$$

in $H^{-1/2}(\partial\Omega)$ at (μ, u) for smooth Ω and under some regularity assumptions on (μ, u).

Finally, we mention that there are certain results in the vector-valued case. Using the proof of Bonnans, 1998, Proposition 4.3, one can show that

$$\{u \in L^p(\Omega)^n \mid u(x) \in K(x) \text{ for almost all } x \in \Omega\}$$

is polyhedric in $L^p(\Omega)^n$, $1 \leq p < \infty$, if the set-valued mapping $x \mapsto K(x)$ is measurable and if $K(x)$ is a polyhedral set for almost all $x \in \Omega$, see also Lemma 5.4.20.

A similar result for vector-valued Sobolev spaces was recently obtained in Chapter 6: Let $C \subset \mathbb{R}^n$ be a polyhedral set with $0 \in \operatorname{int}(C)$ which satisfies the linear-independence constraint qualification (in the sense of nonlinear optimization). Then, the set

$$\{u \in H_0^1(\Omega)^n \mid u(x) \in C \text{ for almost all } x \in \Omega\}$$

is polyhedric in $H_0^1(\Omega)^n$. The technique of proof is much more involved as in the $L^p(\Omega)^n$-case due to the spatial coupling of the $H_0^1(\Omega)^n$-norm. It is not clear whether the result can be generalized to a constraint set C which depends on $x \in \Omega$, and whether the assumption $0 \in \operatorname{int}(C)$ or the linear-independence constraint qualification can be dropped.

Finally, by using a surjectivity argument, see also Lemma 5.3.3 below, one can show the polyhedricity of

$$\{u \in H_\Gamma^1(\Omega;\mathbb{R}^d) \mid u^\top \nu \leq b \text{ a.e. on } \Gamma_C\}$$

in $H_\Gamma^1(\Omega;\mathbb{R}^d) := \{u \in H^1(\Omega;\mathbb{R}^d) \mid u = 0 \text{ a.e. on } \Gamma\}$, where Ω is a Lipschitz domain, $\Gamma, \Gamma_C \subset \partial\Omega$ are closed and measurable (w.r.t. the surface measure) and have positive distance, and $\nu\colon \bar{\Omega} \to \mathbb{R}^d$ is Lipschitz with $|\nu|_{\mathbb{R}^d} \geq 1$ on $\bar{\Omega}$, cf. Sokołowski, Zolésio, 1992, Section 4.6 Betz, 2015, Section 3.2.2, and Müller, Schiela, 2016, Corollary 4.16 for similar results.

5.3.3. Basic properties

In this section, we review two basic properties of polyhedric sets. First, we recall a characterization of polyhedricity for cones. Second, we show a stability result of polyhedric sets under linear mappings.

Lemma 5.3.2. Let $K \subset X$ be a closed, convex cone in the reflexive Banach space X. For $x \in K$, $\mu \in \mathcal{T}_K(x)^\circ = K^\circ \cap x^\perp$ the following conditions are equivalent.

(a) K is polyhedric at (x, μ).

(b) K° is polyhedric at (μ, x).

(c) $\mathcal{K}_K(x,\mu)^\circ = \mathcal{K}_{K^\circ}(\mu, x)$.

(d) $\mathcal{K}_{K^\circ}(\mu,x)^\circ = \mathcal{K}_K(x, \mu)$.

We refer to Lemma 1.5.3 for a proof. To our knowledge, there is no similar result which characterizes the polyhedricity of a closed, convex set K with $0 \in K$ via its polar set $K^\square$. In principle, this should be possible since $K = K^{\square\square}$ is determined by $K^\square$. However, it is easily checked that the set

$$K := \overline{\operatorname{conv}}\{(1, 1 - 1/n)\}_{n\in\mathbb{N}} \subset \mathbb{R}^2$$

is polyhedric, but its polar set is not polyhedric at $((0,1),(1,0))$.

The next lemma shows that the preimage of a polyhedric set under a linear mapping is again polyhedric if a certain condition is satisfied.

Lemma 5.3.3. Let X, Y be Banach spaces, $K_Y \subset Y$ be a closed, convex set and $S\colon X \to Y$ be a bounded and linear operator. Set $K_X := S^{-1}(K_Y)$ and let $x \in K_X$, $\mu \in \mathcal{T}_{K_X}(x)^\circ$ be given.

Assume that

$$S\,X - \mathcal{R}_{K_Y}(S\,x) = Y \tag{5.3.3}$$

is satisfied. Then, there is $\lambda \in \mathcal{T}_{K_Y}(S\,x)^\circ$ with $\mu = S^\star\lambda$.

Additionally, we suppose that

$$S\,X - (\mathcal{R}_{K_Y}(S\,x) \cap \lambda^\perp) = Y \tag{5.3.4}$$

holds and that K_Y is polyhedric at $(S\,x, \lambda)$. Then, K_X is polyhedric at (x, μ).

On the other hand, suppose that

$$S\,X - (\mathcal{R}_{K_Y}(S\,x) \cap [-\mathcal{R}_{K_Y}(S\,x)]) = Y \tag{5.3.5}$$

is satisfied and that K_X is polyhedric at (x, μ). Then, K_Y is polyhedric at $(S\,x, \lambda)$.

Proof. In the case that (5.3.3) is satisfied, we have

$$\mathcal{T}_{K_X}(x) = \mathcal{T}_{S^{-1}(K_Y)}(x) = S^{-1}(\mathcal{T}_{K_Y}(S\,x)), \tag{5.3.6}$$

see Bonnans, Shapiro, 2000, Corollary 2.91. Consequently,

$$\mathcal{T}_{K_X}(x)^\circ = S^{-1}(\mathcal{T}_{K_Y}(S\,x))^\circ = S^\star\,\mathcal{T}_{K_Y}(S\,x)^\circ,$$

see Kurcyusz, 1976, Theorem 2.1. This shows the first assertion.

Now, assert (5.3.4) and that K_Y is polyhedric at $(S\,x, \lambda)$. Let $h \in \mathcal{T}_{K_X}(x) \cap \mu^\perp$ be given. This implies $S\,h \in \mathcal{T}_{K_Y}(S\,x)$, see (5.3.6), and

$$\langle \lambda,\, S\,h\rangle_{Y^\star,Y} = \langle S^\star\lambda,\, h\rangle_{X^\star,X} = \langle \mu,\, h\rangle_{X^\star,X} = 0.$$

By the polyhedricity of K_Y, we get a sequence $\{\hat{h}_k\}_{k\in\mathbb{N}} \subset \mathcal{R}_{K_Y}(S\,x) \cap \lambda^\perp$ with $\hat{h}_k \to S\,h$. Due to the assumption (5.3.4) we can apply the generalization of the open mapping theorem Zowe, Kurcyusz, 1979, Theorem 2.1 we find a constant $M > 0$ and sequences $\{h_k\}_{k\in\mathbb{N}} \subset X$, $\{\tilde{h}_k\}_{k\in\mathbb{N}} \subset \mathcal{R}_{K_Y}(S\,x) \cap \lambda^\perp$ with $\|h_k\|_X + \|\tilde{h}_k\|_Y \le M\,\|\hat{h}_k - S\,h\|_Y \to 0$ and

$$\hat{h}_k - S\,h = S\,h_k - \tilde{h}_k.$$

Hence, $h + h_k \to h$ in X and $S\,(h + h_k) = \hat{h}_k + \tilde{h}_k \in \mathcal{R}_{K_Y}(S\,x) \cap \lambda^\perp$. Hence, $h + h_k \in \mathcal{R}_{K_X}(x) \cap \mu^\perp$ and this demonstrates the polyhedricity of K_X at (x, μ).

Now, suppose that (5.3.5) holds and that K_X is polyhedric at (x, μ). Let $v \in \mathcal{T}_{K_Y}(Sx) \cap \lambda^\perp$ be given. Due to (5.3.5), there is $h \in X$, $z \in \mathcal{R}_{K_Y}(Sx) \cap (-\mathcal{R}_{K_Y}(Sx))$ such that $v = Sh - z$. Moreover, $\pm z \in \mathcal{R}_{K_Y}(Sx)$ and $\lambda \in \mathcal{T}_{K_Y}(Sx)^\circ$ implies $z \in \lambda^\perp$. Hence, $Sh = v + z \in \mathcal{T}_{K_Y}(Sx) \cap \lambda^\perp$. By (5.3.6) we find $h \in \mathcal{T}_{K_X}(x) \cap \mu^\perp$. From the polyhedricity of K_X we get a sequence $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{R}_{K_X}(x) \cap \mu^\perp$ with $h_n \to h$ in X. Now, it is easy to see that $Sh_n - z \in \mathcal{R}_{K_Y}(x) \cap \lambda^\perp$ and that this sequence converges to v. This shows the polyhedricity of K_Y at (Sx, λ).

The condition (5.3.3) is the constraint qualification of Robinson, Zowe and Kurcyusz. Conditions (5.3.4) and (5.3.5) are stricter variants, which also imply the uniqueness of the multiplier λ, see Shapiro, 1997a, Theorem 2.2. Finally, we mention that the result of Lemma 5.3.3 was previously only known in the case that S is surjective, see Sokołowski, Zolésio, 1992, pp. 208, 209, Bonnans, Shapiro, 2000, Proposition 3.54 and Müller, Schiela, 2016, Lemma 4.11. Note that the surjectivity of S implies all of the conditions (5.3.3)–(5.3.5).

5.4. Polyhedricity of intersections

In this section, we are interested in the polyhedricity of intersections of (polyhedric) sets. In particular, we are going to study conditions which ensure that the intersection is again polyhedric.

First, we give some general remarks concerning the intersection of a polyhedric set with a polyhedral set in Section 5.4.1. Then, we restrict our attention to sets with bounds in Banach spaces featuring a lattice structure in Section 5.4.2. In particular, we extend the classical results by Mignot, 1976; Haraux, 1977 by showing that intersections of sets with bounds with polyhedral sets are polyhedric. Finally, we give certain counterexamples in Section 5.4.3.

5.4.1. Intersections of a polyhedric set with a polyhedral set

In this section, we study the intersection of a polyhedric set with a polyhedral set. In general, this intersection may fail to be polyhedric, see Example 5.4.24, and we can only give some partial results.

We start with a simple observation.

Lemma 5.4.1. Let $K \subset X$ be polyhedric at x. Then,

$$\mathcal{T}_K(x) \cap \mu^\perp = \operatorname{cl}(\mathcal{R}_K(x) \cap \mu^\perp) \qquad \forall \mu \in X^\star.$$

Proof. If $\mu \in \mathcal{T}_K(x)^\circ$ or $\mu \in -\mathcal{T}_K(x)^\circ$, the claim follows from the definition of polyhedricity.

On the other hand, if $\mu \notin \mathcal{T}_K(x)^\circ$ and $\mu \notin -\mathcal{T}_K(x)^\circ$, then, there are $v^+, v^- \in \mathcal{R}_K(x)$ such that

$$\langle \mu,\, v^+ \rangle > 0 \quad \text{and} \quad \langle \mu,\, v^- \rangle < 0.$$

Now, let $v \in \mathcal{T}_K(x) \cap \mu^\perp$ be given. By definition of $\mathcal{T}_K(x)$, there is a sequence $\{v_k\}_{k\in\mathbb{N}} \subset \mathcal{R}_K(x)$, such that $\|v - v_k\|_X \to 0$. Now, it is easy to see that there are non-negative null-sequences $\{\lambda_k^+\}_{k\in\mathbb{N}}$, $\{\lambda_k^-\}_{k\in\mathbb{N}}$, such that

$$\langle \mu,\, v_k + \lambda_k^+ v^+ + \lambda_k^- v^- \rangle = 0.$$

This shows

$$v_k + \lambda_k^+ v^+ + \lambda_k^- v^- \in \mathcal{R}_K(x) \cap \mu^\perp \qquad \text{and} \qquad v_k + \lambda_k^+ v^+ + \lambda_k^- v^- \to v.$$

Hence, $v \in \mathrm{cl}\big(\mathcal{R}_K(x) \cap \mu^\perp\big)$.

By using this lemma, we get a formula for the tangent cone of the intersection of a polyhedric set and a hyperplane or half-space.

Lemma 5.4.2. Let K be polyhedric at x and let $\mu \in X^\star$ be given. We set

$$K_\mu := \{y \in K \mid \langle \mu,\, y - x \rangle = 0\} \quad \text{and} \quad \hat{K}_\mu := \{y \in K \mid \langle \mu,\, y - x \rangle \leq 0\}$$

Then,

$$\begin{aligned} &\mathcal{R}_{K_\mu}(x) = \mathcal{R}_K(x) \cap \mu^\perp, \quad \mathcal{T}_{K_\mu}(x) = \mathcal{T}_K(x) \cap \mu^\perp, \quad \mathcal{T}_{K_\mu}(x)^\circ = \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ + \mathrm{lin}\{\mu\}), \\ &\mathcal{R}_{\hat{K}_\mu}(x) = \mathcal{R}_K(x) \cap \mu^\circ, \quad \mathcal{T}_{\hat{K}_\mu}(x) = \mathcal{T}_K(x) \cap \mu^\circ, \quad \mathcal{T}_{\hat{K}_\mu}(x)^\circ = \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ + \mathrm{cone}\{\mu\}). \end{aligned}$$

Proof. We have

$$\mathcal{R}_K(x) \cap \mu^\perp = \bigcup_{\lambda>0} \lambda\,(K - x) \cap \mu^\perp = \bigcup_{\lambda>0} \lambda\,(K_\mu - x) = \mathcal{R}_{K_\mu}(x).$$

By using the previous lemma, we find

$$\mathcal{T}_K(x) \cap \mu^\perp = \mathrm{cl}\big(\mathcal{R}_K(x) \cap \mu^\perp\big) = \mathrm{cl}\Big(\mathcal{R}_{K_\mu}(x)\Big) = \mathcal{T}_{K_\mu}(x).$$

The formula for $\mathcal{T}_{K_\mu}(x)^\circ$ follows from taking polars.

The formulas for $\hat{K}_\mu$ follow from similar considerations.

In order to study intersections with more than one hyperplane, we introduce the concept of higher-order polyhedricity.

Definition 5.4.3. Let $K \subset X$ be a closed, convex subset of the Banach space X and let $n \in \mathbb{N}_0$ be given. We call K n-polyhedric at $x \in K$, if

$$\mathcal{T}_K(x) \cap \bigcap_{i=1}^{n} \mu_i^{\perp} = \operatorname{cl}\Big(\mathcal{R}_K(x) \cap \bigcap_{i=1}^{n} \mu_i^{\perp}\Big) \qquad \forall \mu_1, \dots, \mu_n \in X^\star \tag{5.4.1}$$

holds.

Note that polyhedric sets are 1-polyhedric (by Lemma 5.4.1) and all closed, convex sets are 0-polyhedric.

Similarly to Lemma 5.4.2 we can prove the following result.

Lemma 5.4.4. We fix a closed, convex set $K \subset X$ and a polyhedral set

$$P := \{x \in X \mid \langle \nu_i,\, x\rangle = c_i, 1 \le i \le n;\ \langle \mu_j,\, x\rangle \le d_j, 1 \le j \le m\}$$

such that $K \cap P$ is not empty and we fix $x \in K \cap P$. Here, $n, m \in \mathbb{N}_0$, $\nu_i, \mu_j \in X^\star$, $c_i, d_j \in \mathbb{R}$ for $1 \le i \le n$ and $1 \le j \le m$. Moreover, we set $N = m + n$.

For simplicity of the presentation, we assume that all inequality constraints are active at x, i.e., $\langle \mu_j,\, x\rangle = d_j$ holds for all $1 \le j \le m$.

Then,

$$\mathcal{R}_{K\cap P}(x) = \mathcal{R}_K(x) \cap \bigcap_{i=1}^{n} \nu_i^{\perp} \cap \bigcap_{j=1}^{m} \mu_j^{\circ}.$$

In the case that K is N-polyhedric at x, we have

$$\begin{aligned}
\mathcal{T}_{K\cap P}(x) &= \mathcal{T}_K(x) \cap \bigcap_{i=1}^{n} \nu_i^{\perp} \cap \bigcap_{j=1}^{m} \mu_j^{\circ},\\
\mathcal{T}_{K\cap P}(x)^\circ &= \operatorname{cl}_\star(\mathcal{T}_K(x)^\circ + \operatorname{lin}\{\nu_1, \dots, \nu_n\} + \operatorname{cone}\{\mu_1, \dots, \mu_m\}).
\end{aligned}$$

If, additionally, K is $(N+q)$-polyhedric at x for some $q \in \mathbb{N}_0$, then $K \cap P$ is q-polyhedric at x.

Proof. The verification of the formula for $\mathcal{R}_{K\cap P}(x)$ is straightforward.

The inclusion "$\subset$" in the formula for the tangent cone follows from the formula for the radial cone.

Now, let $v \in \mathcal{T}_K(x) \cap \bigcap_{i=1}^{n} \nu_i^{\perp} \cap \bigcap_{j=1}^{m} \mu_j^{\circ}$ be given. W.l.o.g., we assume that there is $\hat{m} \in \mathbb{N}_0$ such that $\langle \mu_j,\, v\rangle = 0$ for $1 \le j \le \hat{m}$ and $\langle \mu_j,\, v\rangle < 0$ for $\hat{m} < j \le m$. Since K is $(n + \hat{m})$-polyhedric, there is a sequence $\{v_k\}_{k\in\mathbb{N}} \subset \mathcal{R}_K(x) \cap \bigcap_{i=1}^{n} \nu_i^{\perp} \cap \bigcap_{j=1}^{\hat{m}} \mu_j^{\perp}$ with $v_k \to v$ in X. Hence, $v_k \in \mathcal{R}_K(x) \cap \bigcap_{i=1}^{n} \nu_i^{\perp} \cap \bigcap_{j=1}^{m} \mu_j^{\circ}$ for m large enough. Hence, $v \in \mathcal{T}_{K\cap P}(x)$.

The formula for the normal cone follows by taking polars.

The assertion concerning polyhedricity of $K \cap P$ is straightforward to check.

The following lemma generalizes Lemma 5.4.1 to higher-order polyhedricity.

Lemma 5.4.5. Let $n \geq 1$ be given and assume that $K \subset X$ is $(n-1)$-polyhedric, but not n-polyhedric. Then, there exist $\mu_1, \ldots, \mu_n \in X^\star$ such that

$$\mu_i \in \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ + \mathrm{lin}\{\mu_1, \ldots, \mu_{i-1}\}) \qquad \forall i \in \{1, \ldots, n\}$$

and

$$\mathcal{T}_K(x) \cap \bigcap_{i=1}^n \mu_i^\perp \neq \mathrm{cl}\Big(\mathcal{R}_K(x) \cap \bigcap_{i=1}^n \mu_i^\perp\Big).$$

Proof. We prove the result by induction over n.

The case $n = 1$ follows from Lemma 5.4.1.

Let $n > 1$ be given and assume that the assertion holds for all sets $\tilde{K} \subset X$ which are $(n-2)$-polyhedric, but not $(n-1)$-polyhedric.

Since K is not n-polyhedric, there exist $\lambda_1, \ldots, \lambda_n \subset X^\star$ such that

$$\mathcal{T}_K(x) \cap \bigcap_{i=1}^n \lambda_i^\perp \neq \mathrm{cl}\Big(\mathcal{R}_K(x) \cap \bigcap_{i=1}^n \lambda_i^\perp\Big).$$

Hence, there exist $v \in \mathcal{T}_K(x) \cap \bigcap_{i=1}^n \lambda_i^\perp$ and $\varepsilon > 0$, such that

$$B_\varepsilon(v) \cap \mathcal{R}_K(x) \cap \bigcap_{i=1}^n \lambda_i^\perp = \emptyset. \tag{$*$}$$

We define the bounded, linear map $P\colon X \to \mathbb{R}^n$, $x \mapsto (\langle \lambda_1, x\rangle, \ldots, \langle \lambda_n, x\rangle)$. Next, we define the convex sets

$$S := B_\varepsilon(v) \cap \mathcal{R}_K(x) \subset X \qquad \text{and} \qquad M := PS \subset \mathbb{R}^n.$$

By $(*)$ we have

$$0 \notin M$$

and it is clear that the set M is convex. Hence, we can separate M and 0, i.e., there exists $t \in \mathbb{R}^n \setminus \{0\}$, such that $M \subset t^\circ$, see Bonnans, Shapiro, 2000, Thm. 2.17. Since $t \neq 0$, we can find a regular matrix $R \in \mathbb{R}^{n\times n}$, such that $R_{1i} = t_i$ for all $i = 1, \ldots, n$. We define

$$\nu_i := \sum_{j=1}^n R_{ij}\lambda_j \qquad \forall i \in \{1, \ldots, n\}$$

and by regularity of R, we get $\operatorname{lin}\{\lambda_1, \dots, \lambda_n\} = \operatorname{lin}\{\nu_1, \dots, \nu_n\}$ and thus $\bigcap_{i=1}^n \lambda_i^\perp = \bigcap_{i=1}^n \nu_i^\perp$.

As a next step, we show that $\nu_1 \in \mathcal{T}_K(x)^\circ$. To the contrary, assume that $\nu_1 \notin \mathcal{T}_K(x)^\circ$. Then, there exists $w \in \mathcal{R}_K(x)$ with $\langle \nu_1, w\rangle > 0$. We set $h = (\langle \nu_i, w\rangle)_{i=1}^n \in \mathbb{R}^n$. Since $h \neq 0$, it is possible to choose vectors $g_1, \dots, g_{n-1} \in \mathbb{R}^n$, such that

$$\operatorname{lin}\{h\} = \bigcap_{i=1}^{n-1} g_i^\perp.$$

Since

$$v \in \mathcal{T}_K(x) \cap \bigcap_{i=1}^n \lambda_i^\perp = \mathcal{T}_K(x) \cap \bigcap_{i=1}^n \nu_i^\perp \subset \mathcal{T}_K(x) \cap \bigcap_{i=1}^{n-1} \Big(\sum_{j=1}^n (g_i)_j \nu_j\Big)^\perp,$$

and since K is $(n-1)$-polyhedric, there exists a sequence

$$\{v_k\} \subset \mathcal{R}_K(x) \cap \bigcap_{i=1}^{n-1} \Big(\sum_{j=1}^n (g_i)_j \nu_j\Big)^\perp,$$

with $v_k \to v$.

Then, we have

$$\big(g_i, (\langle \nu_j, v_k\rangle)_{j=1}^n\big)_{\mathbb{R}^n} = \sum_{j=1}^n (g_i)_j \langle \nu_j, v_k\rangle = \Big\langle \sum_{j=1}^n (g_i)_j \nu_j, v_k \Big\rangle = 0 \qquad \forall i \in \{1, \dots, n-1\}.$$

Hence, $(\langle \nu_j, v_k\rangle)_{j=1}^n$ belongs to the linear hull of h. Thus, there exists a sequence $\{\alpha_k\} \subset \mathbb{R}$ such that

$$\langle \nu_i, v_k\rangle = \alpha_k \langle \nu_i, w\rangle \qquad \forall i = 1, \dots, n. \tag{$**$}$$

For $i = 1$, we find $\langle \nu_1, v_k\rangle = \alpha_k \langle \nu_1, w\rangle$. Since $v_k \in S$ (it holds for k large enough and w.l.o.g. we can drop the terms in the sequence for which $v_k \notin S$), we have

$$\langle \nu_1, v_k\rangle = \sum_{j=1}^n R_{1j} \langle \lambda_j, v_k\rangle = \sum_{j=1}^n t_j \langle \lambda_j, v_k\rangle = (t, P v_k)_{\mathbb{R}^m} \leq 0,$$

since $P v_k \in PS = M \subset t^\circ$. Hence, $\alpha_k = \langle \nu_1, v_k\rangle / \langle \nu_1, w\rangle \leq 0$ and $\alpha_k \to 0$, since $\langle \nu_1, v\rangle = 0$ and $v_k \to v$. Then, the sequence $\{v_k - \alpha_k w\}_{k\in\mathbb{N}}$ belongs to $\mathcal{R}_K(x)$ and converges towards v. By ($**$) we have $\langle \nu_i, v_k - \alpha_k w\rangle = 0$ for all $i = 1, \dots, n$ and $k \in \mathbb{N}$. Hence,

$$v_k - \alpha_k w \in \mathcal{R}_K(x) \cap \bigcap_{i=1}^n \nu_i^\perp = \mathcal{R}_K(x) \cap \bigcap_{i=1}^n \lambda_i^\perp$$

and this is a contradiction to ($*$). Hence, we have $\nu_1 \in \mathcal{T}_K(x)^\circ$.

Now, we set $\mu_1 := \nu_1$. From Lemma 5.4.4, we find that the set $K_{\mu_1} := \{y \in K \mid \langle x - y, \mu_1 \rangle = 0\}$ is $(n-2)$-polyhedric and

$$v \in \mathcal{T}_K(x) \cap \bigcap_{i=1}^{n} \nu_i = \mathcal{T}_{K_{\mu_1}}(x) \cap \bigcap_{i=2}^{n} \nu_i$$
$$v \notin \mathrm{cl}(\mathcal{R}_K(x) \cap \bigcap_{i=1}^{n} \nu_i) = \mathrm{cl}(\mathcal{R}_{K_{\mu_1}}(x) \cap \bigcap_{i=2}^{n} \nu_i),$$

see Lemma 5.4.2. Hence, K_{μ_1} is not $(n-1)$-polyhedric. By the induction hypothesis there exist $\mu_2, \dots, \mu_n \in X$, such that

$$\mu_i \in \mathrm{cl}_\star(\mathcal{T}_{K_{\mu_1}}(x)^\circ + \mathrm{lin}\{\mu_2, \dots, \mu_{i-1}\}) = \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ + \mathrm{lin}\{\mu_1, \dots, \mu_{i-1}\})$$

holds for all $n = 2, \dots, n$ and

$$\mathcal{T}_K(x) \cap \bigcap_{i=1}^{n} \mu_i^\perp = \mathcal{T}_{K_{\mu_1}}(x) \cap \bigcap_{i=1}^{n} \mu_i^\perp \neq \mathrm{cl}(\mathcal{R}_{K_{\mu_1}}(x) \cap \bigcap_{i=1}^{n} \mu_i^\perp) = \mathrm{cl}(\mathcal{R}_K(x) \cap \bigcap_{i=1}^{n} \mu_i^\perp).$$

This finishes the induction step.

Hence, in order to check 2-polyhedricity of a polyhedric set, it is sufficient to show

$$\mathcal{T}_K(x) \cap \mu_1^\perp \cap \mu_2^\perp = \mathrm{cl}(\mathcal{R}_K(x) \cap \mu_1^\perp \cap \mu_2^\perp)$$

for $\mu_1 \in \mathcal{T}_K(\bar{x})^\circ$ and $\mu_2 \in \mathrm{cl}_\star(\mathcal{T}_K(\bar{x})^\circ + \mathrm{lin}\{\mu_1\})$. However, it is not possible to verify this condition in the case of a general polyhedric set. Indeed, we will show the existence of polyhedric sets which are not 2-polyhedric in Example 5.4.24.

5.4.2. Intersections of a set with bounds with a polyhedral set

In this section, we study sets which are compatible to the lattice structure of a Banach space. The precise assumption on this lattice structure is given in the following definition and Assumption 5.4.7.

Definition 5.4.6. Let X be a (real) Banach space and $C \subset X$ be a closed convex cone with $C \cap -C = \{0\}$. For $x, y \in X$ we say $x \geq y$ if and only if $x - y \in C$.

We say that $z \in X$ is the supremum of $x, y \in X$ if $z \geq x$, $z \geq y$ and

$$(w \geq x \text{ and } w \geq y) \quad \Rightarrow \quad w \geq z \qquad \forall w \in X.$$

It is easy to see that the supremum (if it exists) is unique and it will be denoted by $\max(x, y)$.

If the supremum of all $x, y \in X$ exists, we say that X is a vector lattice.

It is easy to check that the relation $\geq$ on X is antisymmetric ($x \geq y$ and $y \geq x$ implies $x = y$), transitive ($x \geq y$ and $y \geq z$ implies $x \geq z$) and reflexive ($x \geq x$ for all $x \in X$). Further, it is compatible with the linear structure on X, i.e.,

$$\begin{aligned} x \geq y \quad &\Longrightarrow \quad x + z \geq y + z \qquad \forall x, y, z \in X, \\ x \geq y \quad &\Longrightarrow \quad \alpha\, x \geq \alpha\, y \qquad \forall x, y \in X, \alpha \geq 0, \\ x \geq y \quad &\Longrightarrow \quad \alpha\, y \geq \alpha\, x \qquad \forall x, y \in X, \alpha \leq 0. \end{aligned}$$

For convenience, we will use the natural notations

$$\min(x, y) := -\max(-x, -y) \quad \text{and} \quad |x| := \max(x, 0) - \min(x, 0),$$

for $x, y \in X$. Note that

$$x = x + \max(0, -x) - \max(-x, 0) = \max(x, 0) + \min(x, 0).$$

This readily implies

$$x + y = x - y + 2\, y = \max(x - y, 0) + \min(x - y, 0) + 2\, y = \max(x, y) + \min(x, y). \quad (5.4.2)$$

For more information on vector lattices, we refer to Schaefer, 1974, Chapter II.

In what follows, we need a slight assumption on the continuity of $\max(\cdot, \cdot)$ and this will be a standing assumption in this section.

Assumption 5.4.7. We assume that X is a vector lattice (induced by the closed, convex cone C). Moreover, we assume that for all sequences $\{x_n\}_{n \in \mathbb{N}} \subset X$ with $x_n \to x$ in X we have

$$\max(0, x_n) \rightharpoonup \max(0, x) \quad \text{in } X.$$

That is, $\max(0, \cdot)$ is assumed to be strongly-weakly sequentially continuous.

Note that this assumption even implies (strong-weak) continuity in both arguments, since

$$\max(x_n, y_n) = \max(x_n - y_n, 0) + y_n \rightharpoonup \max(x - y, 0) + y = \max(x, y)$$

for $x_n \to x$, $y_n \to y$.

We emphasize that we do not require the stronger property that X is a Banach lattice, which would amount to

$$|x| \leq |y| \quad \Rightarrow \quad |||x|||_X \leq |||y|||_X \qquad \forall x, y \in X,$$

see Schaefer, 1974, Section II.5. Indeed, the important space $X = H_0^1(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is a bounded, open set, equipped with its natural, pointwise ordering satisfies Assumption 5.4.7, but it is *not* a Banach lattice.

A convenient result to show the satisfaction of Assumption 5.4.7 is the following results, which is essentially Haraux, 1977, Theorem 3. We give the proof for the convenience of the reader.

Lemma 5.4.8. Let X be a reflexive Banach space which is additionally a vector lattice. Suppose that $x \mapsto \max(x, 0)$ is bounded, i.e., there exists $M > 0$ with $\|\max(x, 0)\|_X \leq M \|x\|_X$ for all $x \in X$. Then, Assumption 5.4.7 is satisfied.

Proof. Let $\{x_n\}_{n\in\mathbb{N}} \subset X$ be a sequence with $x_n \to x$ in X. The boundedness of $x \mapsto \max(x, 0)$ implies the boundedness of $\{\max(x_n, 0)\}_{n\in\mathbb{N}}$. Due to the reflexivity of X, there is a subsequence (without relabeling) such that $\max(x_n, 0) \rightharpoonup y$ in X for some $y \in X$. It remains to show $y = \max(x, 0)$. By the definition of the maximum, we have

$$\max(x_n, 0) + \max(x - x_n, 0) \geq x_n + (x - x_n) = x.$$

Together with

$$\max(x_n, 0) + \max(x - x_n, 0) \geq 0 + 0 = 0$$

we infer

$$\max(x_n, 0) + \max(x - x_n, 0) \geq \max(x, 0).$$

Since C is closed and convex, it is weakly closed and we can pass to the weak limit to obtain $y + 0 \geq \max(x, 0)$, since $\|\max(x - x_n, 0)\|_X \leq M \|x - x_n\|_X \to 0$. Similarly, we show

$$\max(x, 0) + \max(x_n - x, 0) \geq \max(x_n, 0)$$

and $n \to \infty$ gives $\max(x, 0) + 0 \geq y$. Hence, $y = \max(x, 0)$. The uniqueness of the weak limit ensures the weak convergence of the entire sequence $\{\max(x_n, 0)\}_{n\in\mathbb{N}}$.

In the case that X is additionally a Hilbert space, the boundedness of $x \mapsto \max(x, 0)$ follows from $\left(\max(x, 0), \min(x, 0)\right)_X \geq 0$ for all $x \in X$, see Haraux, 1977, Corollary 1.

In what follows, we will consider sets with bounds in the sense of the following definition.

Definition 5.4.9. We assume that the Banach space X satisfies Assumption 5.4.7.

Let $\underline{K} \subset X$ be a closed convex set. We say that $\underline{K}$ is a *set with lower bound*, if $x, y \in \underline{K}$ implies $\min(x, y) \in \underline{K}$ and if we have $\underline{K} + C \subset \underline{K}$.

Similarly, a closed convex set $\overline{K} \subset X$ is called a *set with upper bound*, if $-\overline{K}$ is a set with lower bound.

Finally, $K \subset X$ is said to be a *set with bounds*, if it can be written as the intersection of a set with upper bound and a set with lower bound.

We note that Mignot uses the same definition in the special case of *Dirichlet spaces*, cf. Mignot, 1976, Définition 3.2.

We give some examples illustrating Definition 5.4.9. For all $x \in X$, the set

$$x + C = \{y \in X \mid y \geq x\}$$

is a set with lower bound. Similarly, X itself is a set with lower bound.

It is also possible to consider the case that the lower bound does not belong to X. To illustrate this situation, we use the Lebesgue space $X := L^2(0,1)$ with the natural, pointwise ordering, and pick a measurable function $x\colon (0,1) \to \mathbb{R} \cup \{-\infty\}$. Then,

$$\{y \in X \mid y \geq x \text{ a.e. in } (0,1)\}$$

is a set with lower bound, even in case that $x \notin L^2(0,1)$.

It is easy to see that if $\underline{K}$ is a set with lower bound, $x, y \in \underline{K}$ implies $\max(x,y) \in \underline{K}$, since $\max(x,y) = x + \max(0, y-x) \in x + C \subset \underline{K}$. Hence, $\underline{K}$ is closed under taking maximums and minimums.

The next assumption is a standing assumption in this section.

Assumption 5.4.10. We suppose that Assumption 5.4.7 is satisfied. The sets $\underline{K}$ and $\overline{K}$ denote sets with lower and upper bound, respectively. We set $K := \underline{K} \cap \overline{K}$ and fix some $x \in K$.

Now, we give some straightforward results concerning the structure of the sets $\underline{K}$, $\overline{K}$ and K.

Lemma 5.4.11. For $v \in \mathcal{R}_{\underline{K}}(x)$, we have $\max(v,0), \min(v,0) \in \mathcal{R}_{\underline{K}}(x)$.

For $v \in \mathcal{R}_{\overline{K}}(x)$, we have $\max(v,0), \min(v,0) \in \mathcal{R}_{\overline{K}}(x)$.

For $v \in \mathcal{R}_K(x)$, we have $\max(v,0), \min(v,0) \in \mathcal{R}_K(x)$.

Proof. Due to $\mathcal{R}_{\overline{K}}(x) = -\mathcal{R}_{-\overline{K}}(x)$ and $\mathcal{R}_K(x) = \mathcal{R}_{\underline{K}}(x) \cap \mathcal{R}_{\overline{K}}(x)$, it is sufficient to prove the claim for $v \in \mathcal{R}_{\underline{K}}(x)$.

For a given $v \in \mathcal{R}_{\underline{K}}(x)$, there is $t > 0$, such that $x + t\,v \in \underline{K}$. Together with $x \in \underline{K}$, we find

$$x + t\,\max(v,0) = x + t\,\max(0,v) = x + \max(0, t\,v) = \max(x, x + t\,v) \in \underline{K}$$

and

$$x + t\,\min(v,0) = x + t\,\min(0,v) = x + \min(0, t\,v) = \min(x, x + t\,v) \in \underline{K}.$$

This shows $\max(v,0), \min(v,0) \in \mathcal{R}_{\underline{K}}(x)$.

In the following, we will often invoke Mazur's lemma in order to generate a strongly convergent sequence as convex combinations of a weakly convergent sequence. In particular, if the sequence $\{v_n\} \subset X$ belongs to some convex set $G \subset X$, and $v_n \rightharpoonup v$, Mazur's lemma ensures the existence of a sequence $\{\tilde{v}_n\} \subset G$ with $\tilde{v}_n \to v$.

Lemma 5.4.12. For $v \in \mathcal{T}_{\underline{K}}(x)$, we have $\max(v,0), \min(v,0) \in \mathcal{T}_{\underline{K}}(x)$.

For $v \in \mathcal{T}_{\overline{K}}(x)$, we have $\max(v,0), \min(v,0) \in \mathcal{T}_{\overline{K}}(x)$.

Proof. Similar to the proof of Lemma 5.4.11, it is sufficient to consider the case $v \in \mathcal{T}_{\underline{K}}(x)$. By definition, there is a sequence $\{v_n\} \subset \mathcal{R}_{\underline{K}}(x)$, such that $v_n \to v$ as $n \to \infty$. Using Assumption 5.4.7, we get

$$\max(v_n,0) \rightharpoonup \max(v,0) \qquad \text{and} \qquad \min(v_n,0) \rightharpoonup \min(v,0).$$

By Lemma 5.4.11, we have $\max(v_n,0), \min(v_n,0) \in \mathcal{R}_{\underline{K}}(x)$ and, together with Mazur's lemma, this shows $\max(v,0), \min(v,0) \in \mathcal{T}_{\underline{K}}(x)$, since $\mathcal{R}_{\underline{K}}(x)$ is convex.

The following result generalizes Mignot, 1976, Lemme 3.4(ii). Note that, in general, the tangent cone to an intersection is (strictly) smaller than the intersection of the tangent cones.

Lemma 5.4.13. We have $\mathcal{R}_K(x) = \mathcal{R}_{\underline{K}}(x) \cap \mathcal{R}_{\overline{K}}(x)$ and $\mathcal{T}_K(x) = \mathcal{T}_{\underline{K}}(x) \cap \mathcal{T}_{\overline{K}}(x)$. In particular, $v \in \mathcal{T}_K(x)$ implies $\max(v,0), \min(v,0) \in \mathcal{T}_K(x)$.

Proof. Since $K = \underline{K} \cap \overline{K}$, we have

$$\mathcal{R}_K(x) = \mathcal{R}_{\underline{K}}(x) \cap \mathcal{R}_{\overline{K}}(x) \qquad \text{and} \qquad \mathcal{T}_K(x) \subset \mathcal{T}_{\underline{K}}(x) \cap \mathcal{T}_{\overline{K}}(x).$$

To show the reverse inclusion, let $v \in \mathcal{T}_{\underline{K}}(x) \cap \mathcal{T}_{\overline{K}}(x)$ be given. Then, there exist sequences $\{v_n\} \subset \mathcal{R}_{\underline{K}}(x)$ and $\{w_n\} \subset \mathcal{R}_{\overline{K}}(x)$, such that $v_n \to v$ and $w_n \to v$. In particular, $\min(v_n,0) + \max(w_n,0) \rightharpoonup v$. By Lemma 5.4.11, we have $\min(v_n,0) \in \mathcal{R}_{\underline{K}}(x)$ and together with $\min(v_n,0) \in -C$, we have $\min(v_n,0) \in \mathcal{R}_{\underline{K}}(x) \cap \mathcal{R}_{\overline{K}}(x) = \mathcal{R}_K(x)$. Similarly, $\max(w_n,0) \in \mathcal{R}_K(x)$ follows. Hence, $\min(v_n,0) + \max(w_n,0) \in \mathcal{R}_K(x)$. This shows $v \in \mathcal{T}_K(x)$.

Together with Lemma 5.4.12, we get $\max(v,0), \min(v,0) \in \mathcal{T}_K(x)$ for $v \in \mathcal{T}_K(x)$.

The next result shows that the radial cone of K is compatible with the order structure of X.

Lemma 5.4.14. Let $u, v, w \in X$ with $u \geq v \geq w$ and $u, w \in \mathcal{R}_K(x)$ be given. Then, $v \in \mathcal{R}_K(x)$.

Proof. For $u, w \in \mathcal{R}_K(x)$, there is $t > 0$ with $x + t\,u, x + t\,w \in K$. From $x + t\,u \geq x + t\,v \geq x + t\,w$ we find $v \in \mathcal{R}_K(x)$.

The next lemma demonstrates a possibility to define a projection onto the set $\{v \in X \mid u \geq v \geq w\}$ for $v \geq w$.

Lemma 5.4.15. Let $u, v, w \in X$ be given such that $u \geq w$. Then, $u \geq \min(u, v) + \max(w, v) - v \geq w$.

Proof. We set $\tilde{v} := \min(u, v) + \max(w, v) - v$. From $u \geq w$, we get $\max(u, v) \geq \max(w, v)$. Using (5.4.2) gives $u \geq \min(u, v) + \max(w, v) - v = \tilde{v}$. Similarly, $u \geq w$, implies $\min(u, v) \geq \min(w, v)$ and, consequently, $\tilde{v} = \min(u, v) + \max(v, w) - v \geq w$. Thus, we have shown $u \geq \tilde{v} \geq w$.

Note that the mapping $v \mapsto \min(u, v) + \max(w, v) - v$ is strongly-weakly continuous due to Assumption 5.4.7.

The following lemma is the analogue to Lemma 5.4.14 for the tangent cone. It provides some further insight into the structure of the set K.

Lemma 5.4.16. Let $u, v, w \in X$ with $u \geq v \geq w$ and $u, w \in \mathcal{T}_K(x)$ be given. Then, $v \in \mathcal{T}_K(x)$.

Proof. We start by sequences $\{\tilde{u}_n\}_{n\in\mathbb{N}}, \{\tilde{w}_n\}_{n\in\mathbb{N}} \subset \mathcal{R}_K(x)$ with $\tilde{u}_n \to u$ and $\tilde{w}_n \to w$. By Assumption 5.4.7, we have $\min(\tilde{w}_n, \tilde{u}_n) \rightharpoonup w$. By applying Mazur's lemma to the sequence $\{(\tilde{u}_n, \min(\tilde{w}_n, \tilde{u}_n))\}_{n\in\mathbb{N}}$, we get sequences $\{u_n\}_{n\in\mathbb{N}}, \{w_n\}_{n\in\mathbb{N}} \subset \mathcal{R}_K(x)$ with $u_n \to u$, $w_n \to w$ and $u_n \geq w_n$.

Now, we set $v_n := \min(u_n, v) + \max(w_n, v) - v$. By Lemma 5.4.15 we infer $u_n \geq v_n \geq w_n$ and from Lemma 5.4.14 we get $v_n \in \mathcal{R}_K(x)$.

Finally, Assumption 5.4.7 implies $v_n \rightharpoonup \min(u, v) + \max(w, v) - v = v$. Thus, the claim follows from Mazur's lemma.

Now, we provide an auxiliary lemma which will be the crucial ingredient in the proof of Theorem 5.4.18 below. It contains all the structure of K from Assumption 5.4.10 in the sense that Assumption 5.4.10 is not explicitly used in Theorem 5.4.18 below. It is inspired by the proof of Mignot, 1976, Théorème 3.2.

Lemma 5.4.17. Let $v \in \mathcal{T}_K(x)$ be given. Then, there is a sequence $\{v_k\}_{k\in\mathbb{N}} \subset \mathcal{R}_K(x) \cap (v - \mathcal{T}_K(x))$ and $v_k \to v$. That is,

$$v \in \mathrm{cl}(\mathcal{R}_K(x) \cap (v - \mathcal{T}_K(x)))$$

holds for all $v \in \mathcal{T}_K(x)$.

Proof. We take a sequence $\{w_k\} \subset \mathcal{R}_K(x)$, such that $w_k \to v$. Then,

We define $v_k^+ := \min(\max(v, 0), w_k) + \max(0, w_k) - w_k$. By Assumption 5.4.7, we get $v_k^+ \rightharpoonup \min(\max(v, 0), \max(v, 0)) = \max(v, 0)$ and $\{v_k^+\} \subset \mathcal{R}_K(x)$ follows from $\max(w_k, 0) \geq v_k^+ \geq 0$ and $\max(w_k, 0) \in \mathcal{R}_K(x)$, see Lemmas 5.4.11, 5.4.14 and 5.4.15.

Next, we show $\max(v,0) - v_k^+ \in \mathcal{T}_K(x)$. By Lemma 5.4.15, we find $\max(v,0) \geq v_k^+ \geq 0$ and this implies $\max(v,0) \geq \max(v,0) - v_k^+ \geq 0$. Hence, $\max(v,0) - v_k^+ \in \mathcal{T}_K(x)$ by Lemmas 5.4.13 and 5.4.16.

To summarize, the sequence $\{v_k^+\}$ satisfies $v_k^+ \rightharpoonup \max(v,0)$, $v_k^+ \in \mathcal{R}_K(x)$ and $\max(v,0) - v_k^+ \in \mathcal{T}_K(x)$ for all $k \in \mathbb{N}$.

Similarly, we can show that $v_k^- := \min(0, w_k) + \max(\min(v,0), w_k) - w_k$ satisfies $v_k^- \rightharpoonup \min(v,0)$, $v_k^- \in \mathcal{R}_K(x)$ and $\min(v,0) - v_k^- \in \mathcal{T}_K(x)$. Now, we define $v_k := v_k^+ + v_k^-$ and it is easy to check that $v_k \rightharpoonup v$, $v_k \in \mathcal{R}_K(x)$ and $v - v_k \in \mathcal{T}_K(x)$, i.e., $v_k \in v - \mathcal{T}_K(x)$, hold for all $k \in \mathbb{N}$.

Finally, we can invoke Mazur's lemma to find a convex combination of the sequence $\{v_k\}_{k\in\mathbb{N}} \subset \mathcal{R}_K(x) \cap (v - \mathcal{T}_K(x))$ which converges strongly towards v.

Now we are in position to prove the main theorem of this section.

Theorem 5.4.18. Let Assumption 5.4.10 be satisfied by K. Then, the set K is n-polyhedric for all $n \in \mathbb{N}_0$.

Proof. By induction over $n \in \mathbb{N}_0$, we prove that K is n-polyhedric at all $x \in K$.

Assume that K is n-polyhedric for some $n \in \mathbb{N}_0$. We proceed by contradiction and assume that K is not $(n+1)$-polyhedric. Hence, we can invoke Lemma 5.4.5 and get $\mu_1, \dots, \mu_{n+1} \in X^\star$ such that $\mu_i \in \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ + \mathrm{lin}\{\mu_1, \dots, \mu_{i-1}\})$ for all $i \in \{1, \dots, n+1\}$ and

$$\mathcal{T}_K(x) \cap \bigcap_{i=1}^{n+1} \mu_i^\perp \neq \mathrm{cl}\Big(\mathcal{R}_K(x) \cap \bigcap_{i=1}^{n+1} \mu_i^\perp\Big). \tag{5.4.3}$$

Let $v \in \mathcal{T}_K(x) \cap \bigcap_{i=1}^{n+1} \mu_i^\perp$ be arbitrary. Using Lemma 5.4.17, we get a sequence $\{v_k\} \subset \mathcal{T}_K(x)$ such that $v_k \in \mathcal{R}_K(x)$, $v - v_k \in \mathcal{T}_K(x)$ and $v_k \to v$.

Now, we use induction over j to show that

$$v_k \in \bigcap_{i=1}^{j} \mu_i^\perp \qquad \forall k \in \mathbb{N} \tag{5.4.4}$$

holds for all $j \in \{0, \dots, n+1\}$. The case $j = 0$ is trivially satisfied.

Now, suppose that (5.4.4) holds for some $j \in \{0, \dots, n\}$. Since

$$\mu_{j+1} \in \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ + \mathrm{lin}\{\mu_1, \dots, \mu_j\}) = \Big(\mathcal{T}_K(x) \cap \bigcap_{i=1}^{j} \mu_j^\perp\Big)^\circ, \tag{5.4.5}$$

we have

$$0 \geq \langle \mu_{j+1},\, v - v_k \rangle = \langle \mu_{j+1},\, -v_k \rangle \geq 0,$$

where we have used (5.4.4), (5.4.5), $v \in \bigcap_{i=1}^{j+1} \mu_i^\perp$ and $v_k, v - v_k \in \mathcal{T}_K(x)$. Hence, $v_k \in \mu_{j+1}^\perp$. By induction, (5.4.4) holds for all $j \in \{0, \dots, n+1\}$.

Hence, the sequence $\{v_k\}$ belongs to $\mathcal{R}_K(x) \cap \bigcap_{i=1}^{n+1} \mu_i^\perp$ and converges towards v. Thus, $v \in \mathrm{cl}(\mathcal{R}_K(x) \cap \bigcap_{i=1}^{n+1} \mu_i^\perp)$ and this is a contradiction to (5.4.3). Therefore, K is $(n+1)$-polyhedric and this finishes the induction over n.

Before we give some remarks, we mention an easy corollary which follows together with Lemma 5.4.4.

Corollary 5.4.19. Let $P \subset X$ be a polyhedral set with $x \in P$. Then, $K \cap P$ is polyhedric and

$$\mathcal{T}_{K\cap P}(x) = \mathcal{T}_K(x) \cap \mathcal{T}_P(x), \qquad \text{and} \qquad \mathcal{T}_{K\cap P}(x)^\circ = \mathrm{cl}_\star(\mathcal{T}_K(x)^\circ \cap \mathcal{T}_P(x)^\circ).$$

Now, we give some remarks concerning Theorem 5.4.18.

First, we mention that Theorem 5.4.18 extends previously known results significantly:

(a) Mignot, 1976, Théorème 3.2 shows the polyhedricity of sets with bounds in Dirichlet spaces. In the proof, he uses a pointwise characterization of the tangent cone $\mathcal{T}_K(x)$, which is not available in our general setting.

(b) Haraux, 1977, Corollary 2 shows the polyhedricity of the cone of non-negative vectors $C \subset X$ under Assumption 5.4.7 in the case that X is a Hilbert space. This situation is much easier to handle, since there is only a lower bound and this set C is even a cone.

Moreover, both results show only the polyhedricity of the set under consideration, whereas our result also enables us to show n-polyhedricity, which, in turn, provides additionally the polyhedricity of the intersection with a polyhedral set, see Corollary 5.4.19. To our knowledge, even in special cases, such a result was not available.

Second, by inspecting the proof of Theorem 5.4.18, we see that we only used Lemma 5.4.5, which holds for arbitrary closed, convex sets K, and Lemma 5.4.17. In particular, the lattice structure of the underlying space X was not explicitly used. Hence, the proof of Theorem 5.4.18 provides the n-polyhedricity at $x \in K$ of all closed, convex sets K satisfying

$$v \in \mathrm{cl}(\mathcal{R}_K(x) \cap (v - \mathcal{T}_K(x))) \qquad \forall v \in \mathcal{T}_K(x). \tag{5.4.6}$$

We emphasize that no further assumptions on K are necessary. This enables us to show polyhedricity also in situations in which a lattice structure is not available. We demonstrate this possibility in the following lemma, which significantly relaxes the requirements of Bonnans, 1998, Proposition 4.3, and also shows the n-polyhedricity of K.

Lemma 5.4.20. Let $(\Omega, \Sigma, \mathfrak{m})$ be a complete, σ-finite measure space. Further, let $Q\colon \Omega \rightrightarrows \mathbb{R}^d$ be a measurable (in the sense of Aubin, Frankowska, 2009, Definition 8.1.1), set-valued map such that $Q(\omega) \subset \mathbb{R}^d$ is a polyhedron for a.a. $\omega \in \Omega$. Then, for $1 \leq p < \infty$, the set

$$K = \{u \in L^p(\mathfrak{m}; \mathbb{R}^d) \mid u(\omega) \in Q(\omega) \text{ for a.a. } \omega \in \Omega\}$$

is n-polyhedric at all $u \in K$ for all $n \in \mathbb{N}_0$.

Proof. We follow the proof of Bonnans, 1998, Proposition 4.3. It is easily checked that K is closed in $L^p(\mathfrak{m}; \mathbb{R}^d)$ and convex. Let $v \in \mathcal{T}_K(u)$ be given. For $k \in \mathbb{N}$, we define

$$v_k(\omega) := \begin{cases} v(\omega) & \text{if } u(\omega) + \frac{1}{k} v(\omega) \in Q(\omega), \\ 0 & \text{else.} \end{cases}$$

From the measurability of Q, we get the measurability of v_k. Moreover, $v_k \in \mathcal{R}_K(u)$ by construction. Owing to Aubin, Frankowska, 2009, Corollary 8.5.2, we easily find $v - v_k \in \mathcal{T}_K(u)$. Since $Q(\omega)$ is a polyhedron and $v(\omega) \in \mathcal{T}_{Q(\omega)}(u(\omega)) = \mathcal{R}_{Q(\omega)}(u(\omega))$, we obtain the pointwise convergence $v_k(\omega) \to v(\omega)$ for a.a. $\omega \in \Omega$. Now, $v_k \to v$ in $L^p(\mathfrak{m}; \mathbb{R}^d)$ follows from the dominated convergence theorem, since $|v_k(\omega)|_{\mathbb{R}^d} \leq |v(\omega)|_{\mathbb{R}^d}$. Hence, condition (5.4.6) is satisfied by the set K and the n-polyhedricity follows as in the proof of Theorem 5.4.18.

We mention that the polyhedrality assertion on $Q(\omega)$ can be relaxed to $\mathcal{R}_{Q(\omega)}(u) = \mathcal{T}_{Q(\omega)}(u)$ for all $u \in Q(\omega)$ and a.e. $\omega \in \Omega$. Note that, in difference to the proofs of Lemma 5.4.17 and Theorem 5.4.18, we did use the actual characterization of $\mathcal{T}_K(u)$ in the above proof. We do not know if the result of Lemma 5.4.20 holds for $p = \infty$. Moreover, we mention that the above proof does not generalize to the case of $W^{1,p}(\Omega; \mathbb{R}^d)$, $p \in [1, \infty)$.

We finish this section by giving some examples to which Theorem 5.4.18 can be applied.

Example 5.4.21.

(a) Let $(\Omega, \Sigma, \mathfrak{m})$ be a complete, σ-finite measure space, $a, b : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ measurable and $p \in [1, \infty]$. Then,

$$K = \{u \in L^p(\mathfrak{m}) \mid a \leq u \leq b \text{ a.e.}\}$$

is n-polyhedric for all $n \in \mathbb{N}_0$. Here, the lattice property is evident and Assumption 5.4.7 follows from the fact that $\max(\cdot, 0)$ is globally Lipschitz on $\mathbb{R}$.

(b) Let $\Omega \subset \mathbb{R}^d$ be an open, bounded set (equipped with the Lebesgue measure), $p \in [1, \infty)$ and $a, b\colon \Omega \to \mathbb{R} \cup \{\pm\infty\}$ measurable. Then,

$$K = \{u \in W_0^{1,p}(\Omega) \mid a \leq u \leq b \text{ a.e.}\}$$

is n-polyhedric for all $n \in \mathbb{N}_0$. The lattice property follows from Stampacchia's lemma and Assumption 5.4.7 from Lebesgue's dominated convergence theorem, cf. Bonnans, Shapiro, 2000, Proposition 6.45 and Attouch, Buttazzo, Michaille, 2006, Theorem 5.8.2.

(c) Let $\Omega \subset \mathbb{R}^d$ be an open, bounded set (equipped with the Lebesgue measure), $p \in [1, \infty)$ and $a, b \colon \Omega \to \mathbb{R} \cup \{\pm\infty\}$ measurable. Then,

$$K = \{u \in W^{1,p}(\Omega) \mid a \leq u \leq b \text{ a.e.}\}$$

is n-polyhedric for all $n \in \mathbb{N}_0$. As in the case of $W_0^{1,p}(\Omega)$, the lattice property follows from Stampacchia's lemma and Assumption 5.4.7 from Lebesgue's dominated convergence theorem.

(d) We consider the Besov space $B^s_{p,q}(\mathbb{R}^d)$ with $p, q \in (1, \infty)$ and $s \in (0, 1 + 1/p)$. Then, we know from Bourdaud, Y. Meyer, 1991, Théorème 2 and Oswald, 1992, Theorem 1 that $u \mapsto \max(u, 0)$ is a bounded operator from $B^s_{p,q}(\mathbb{R}^d)$ into itself. Now, since $B^s_{p,q}(\mathbb{R}^d)$ is reflexive, we can use Lemma 5.4.8 to ensure the satisfaction of Assumption 5.4.7. Hence, we obtain the n-polyhedricity of

$$K = \{u \in B^s_{p,q}(\mathbb{R}^d) \mid a \leq u \leq b \text{ a.e.}\},$$

where $a, b \colon \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ are measurable.

If $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, we can use the extension operator from Rychkov, 1999 to get the boundedness of $u \mapsto \max(u, 0)$ from $B^s_{p,q}(\Omega)$ into itself. Hence, we also get n-polyhedricity of sets with pointwise bounds in $B^s_{p,q}(\Omega)$.

The same reasoning can be used in the Triebel-Lizorkin spaces $F^s_{p,q}(\mathbb{R}^d)$ and $F^s_{p,q}(\Omega)$, see Runst, Sickel, 1996, Theorem 5.4.1.

(e) Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. We choose $a, b \colon \partial\Omega \to \mathbb{R} \cup \{\pm\infty\}$, measurable w.r.t. the surface measure on $\partial\Omega$ and denote by $T \colon W^{1,p}(\Omega) \to L^p(\partial\Omega)$ the trace operator, cf. Ziemer, 1989, 190f. Then, it is easy to check that the set

$$K = \{u \in W^{1,p}(\Omega) \mid a \leq T(u) \leq b \text{ a.e. on } \partial\Omega\}$$

is a set with bounds under the natural ordering of $W^{1,p}(\Omega)$ and, thus, n-polyhedric at every $u \in W^{1,p}(\Omega)$ for all $n \in \mathbb{N}_0$.

(f) We use the same setting as in the previous example and, additionally, a vector field $\nu \colon \bar{\Omega} \to \mathbb{R}^d$ such that ν and $\nu/|\nu|^2_{\mathbb{R}^d}$ are Lipschitz continuous on $\bar{\Omega}$. This follows, e.g., if $\nu \in W^{1,\infty}(\Omega)$ is uniformly positive. Then, the operator $R \colon W^{1,p}(\Omega)^d \to W^{1,p}(\Omega)$ defined via $R(v) = v^\top \nu$ is surjective, since $R(w\,\nu/|\nu|^2_{\mathbb{R}^d}) = w$ for all $w \in W^{1,p}(\Omega)$.

Hence, Lemma 5.3.3 together with the previous example shows that the set

$$K = \{u \in W^{1,p}(\Omega; \mathbb{R}^d) \mid a \leq T(u)^\top \nu \leq b \text{ a.e. on } \partial\Omega\}$$

is polyhedric.

(g) Let X be a locally compact Hausdorff space. By the Riesz-Markov theorem Rudin, 1987, Theorem 6.19, we know that the dual space of the Banach lattice $C_0(X)$ (equipped with the natural, pointwise ordering) can be identified with the space $\mathcal{M}(X)$ of regular, countably additive and signed Borel measures equipped with the total variation norm. Owing to Schaefer, 1974, Proposition 5.5, $\mathcal{M}(X)$ is also a Banach lattice and, thus, satisfies Assumption 5.4.7. Hence, for all measures $\mu_a, \mu_b \in \mathcal{M}(X)$, the set

$$K = \{\mu \in \mathcal{M}(X) \mid \mu_a \leq \mu \leq \mu_b\}$$

is n-polyhedric for all $n \in \mathbb{N}_0$. Here, $\mu_a \leq \mu$ is to be understood with the ordering in $\mathcal{M}(X)$ induced by the ordering of $C_0(X)$, i.e., $\int_X f \,\mathrm{d}\mu_a \leq \int_X f \,\mathrm{d}\mu$ for all $f \in C_0(X)$ satisfying $f(x) \geq 0$ for all $x \in X$.

(h) We denote by $BV_0[0,1]$ the space of all functions $f\colon [0,1] \to \mathbb{R}$ of bounded variation with $f(0)$. This set becomes a Banach lattice if it is equipped with the bounded variation norm and the non-negative cone

$$C = \{f \in BV_0[0,1] \mid f \text{ is monotonically non-decreasing}\},$$

see Aliprantis, Border, 2006, Section 9.8, Theorem 9.51. Thus, for every $a, b \in BV_0[0,1]$, the set

$$K = \{f \in BV_0[0,1] \mid f - a, b - f \text{ are monotonically non-decreasing}\}$$

is n-polyhedric for all $n \in \mathbb{N}_0$.

5.4.3. Counterexamples

In this section, we give some counterexamples. To our knowledge, these are the first available counterexamples which show that the intersection of polyhedric sets may fail to be polyhedric.

The first counterexample limits possible generalizations of Corollary 5.4.19. In particular, the intersection of a set with bounds with a co-polyhedral set can, in general, fail to be polyhedric.

Example 5.4.22. We consider the unit interval $(0,1)$ equipped with the Lebesgue measure. We define $b \in L^2(0,1)$ via $b(x) := \left(x^2 + (1-x)^2\right)^{1/2}$ and we set

$$K := \{u \in L^2(0,1) \mid 0 \leq u \leq b \quad \text{a.e. on } (0,1)\}.$$

Moreover, we denote by

$$P := \operatorname{lin}\{e_1, e_2\}$$

the two-dimensional subspace spanned by $e_1(x) := 1 - x$ and $e_2(x) := x$.

Then, it is easy to check that

$$K \cap P = \{\alpha_1 e_1 + \alpha_2 e_2 \mid \alpha_1, \alpha_2 \geq 0 \text{ and } \alpha_1^2 + \alpha_2^2 \leq 1\}.$$

Let $u = \alpha_1 e_1 + \alpha_2 e_2 \in K \cap P$ be given. Then, the set $\{x \in (0,1) \mid u(x) = b(x)\}$ has measure zero. Hence, $\mathcal{T}_K(u) = L^2(0,1)$. This shows that, in general, $\mathcal{T}_{K\cap P}(u) \neq \mathcal{T}_K(u) \cap \mathcal{T}_P(u)$.

Moreover, it is easy to check that $K \cap P$ is not polyhedric, in particular, it is not polyhedric at $(e_1, e_1 - 2\,e_2)$. Here, we identified $L^2(0,1)^\star$ with $L^2(0,1)$.

The same negative results are obtained by intersecting K with the compact co-polyhedral set $\operatorname{conv}\{0, 2\,e_1, 2\,e_2\}$.

The next example shows that sets which are defined by pointwise bounds may fail to be polyhedric if the underlying Banach space does not posses a lattice structure.

Example 5.4.23. Let $\Omega \subset \mathbb{R}^d$ be an open, bounded set. We set $C := \{v \in H_0^1(\Omega) \mid v \geq 0\}$. Then, the cone $-C^\circ \subset H^{-1}(\Omega) := (H_0^1(\Omega))^\star$ induces an order on $H^{-1}(\Omega)$. It is known that $H^{-1}(\Omega)$ equipped with this order fails to be a vector lattice. However, from Lemma 5.3.2 we infer that $-C^\circ$ is polyhedric.

Now, we define

$$\Lambda := \{\lambda \in H^{-1}(\Omega) \mid 1 \geq \lambda \geq -1\},$$

where $1 \in H^{-1}(\Omega)$ is the functional $v \mapsto \int_\Omega v \,\mathrm{d}x$ and "$\geq$" is the ordering induced by $-C^\circ$. Note that Λ is the intersection of the polyhedric sets

$$-C^\circ - 1 = \{\lambda \in H^{-1}(\Omega) \mid \lambda \geq -1\} \quad \text{and} \quad C^\circ + 1 = \{\lambda \in H^{-1}(\Omega) \mid 1 \geq \lambda\}.$$

However, we will show that Λ fails to be polyhedric at some points. Without loss of generality, we assume that the closed unit cube $[-1,1]^d$ belongs to Ω. Then, there exists $u \in C$ with

$$u(x) = |x_1| \qquad \forall x \in [-1,1]^d$$

and such that $\{x \in \Omega \mid u(x) = 0\}$ has Lebesgue measure zero. We set $\lambda := 1$. It is clear that $\lambda \in \Lambda$. Moreover, it is easy to check that $\Lambda \subset L^\infty(\Omega)$ and

$$\mathcal{R}_\Lambda(\lambda) = \{\mu \in L^\infty(\Omega) \mid \mu(x) \leq 0 \text{ for a.a. } x \in \Omega\}.$$

Now, $u \in \mathcal{R}_\Lambda(\lambda)^\circ = \mathcal{T}_\Lambda(\lambda)^\circ$ follows. Moreover, $\mathcal{R}_\Lambda(\lambda) \cap u^\perp = \{0\}$.

We define the surface measure $\mu \in H^{-1}(\Omega)$ via

$$\langle \mu, v \rangle := -\int_S v \,\mathrm{d}s,$$

where $S = \{0\} \times (-1,1)^{d-1}$ is equipped with the $(d-1)$-dimensional Lebesgue measure. Now, we define the functions $\mu_n := -\alpha_n \chi_{A_n} \in \mathcal{R}_\Lambda(\lambda)$ with $A_n := (-1/n, 1/n) \times$

$(-1,1)^{d-1}$ and $\alpha_n := n/2$. We show that $\mu_n \rightharpoonup \mu$ in $H^{-1}(\Omega)$. Indeed, for all $\varphi \in C_0^\infty(\Omega)$ we have $\langle \mu_n, \varphi\rangle \to \langle \mu, \varphi\rangle$ and, thus, it is sufficient to show that the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ is bounded in $H^{-1}(\Omega)$. For $v \in C_0^\infty(\Omega)$ we have

$$v(x) = \int_{-\infty}^{x_1} \frac{\partial v}{\partial x_1}(t, x_2, x_3, \dots)\,\mathrm{d}t \le c\,\|v(\cdot, x_2, x_3, \dots)\|_{L^2(\mathbb{R})} \qquad \forall x \in \Omega.$$

Hence,

$$\int_{-1/n}^{1/n} \frac{n}{2}\, v(x)\,\mathrm{d}x_1 \le c\,\|v(\cdot, x_2, x_3, \dots)\|_{L^2(\mathbb{R})} \qquad \forall x_2, \dots, x_d \in (-1,1).$$

Using Fubini's theorem, we find

$$\langle \mu_n, v\rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-1/n}^{1/n} \frac{n}{2}\, v(x)\,\mathrm{d}x_1\,\mathrm{d}x_2 \,\dots\, \mathrm{d}x_d \le \tilde{c}\,\|v\|_{H_0^1(\Omega)}.$$

Since $\mu_n \in H^{-1}(\Omega)$ and $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, we find $\|\mu_n\|_{H^{-1}(\Omega)} \le \tilde{c}$ for all $n \in \mathbb{N}$. Hence, we have shown $\mu_n \rightharpoonup \mu$ in $H^{-1}(\Omega)$. Using Mazur's lemma, we find $\mu \in \mathcal{T}_\Lambda(\lambda)$. Further, it is clear that $\mu \in u^\perp$. Hence,

$$\mu \in \mathcal{T}_\Lambda(\lambda) \cap u^\perp \neq \mathrm{cl}(\mathcal{R}_\Lambda(\lambda) \cap u^\perp) = \{0\}$$

and this shows that Λ is not polyhedric at (λ, u).

The above example is crucial for studying variational inequalities of the second kind in $H_0^1(\Omega)$, cf. the technique for a similar problem in $H^{1/2}(\partial\Omega)$ in Sokołowski, Zolésio, 1992, Section 4.5. In Christof, C. Meyer, 2015, the authors provided a directional differentiability result for a variational inequality of second kind in $H_0^1(\Omega)$ and additional terms appear in the variational inequality for the derivative. This cannot happen if $\Lambda \subset H^{-1}(\Omega)$ would be polyhedric. However, the results of De los Reyes, C. Meyer, 2016 suggest that Λ may be polyhedric at some (λ, u) possessing certain regularity properties.

Finally, we present a counterexample in $\mathbb{R}^3$ which possesses some very bizarre properties. It is a compact polyhedric set, but fails to be a polyhedron. Moreover, it is not 2-polyhedric and the intersection with a hyperplane (which is polyhedral and co-polyhedral) is not polyhedric.

Example 5.4.24. In $\mathbb{R}^3$ we consider the points

$$\begin{aligned} O &:= (0,0,0), \\ P_n &:= (n^{-2}, n^{-1}, -n^{-3}), && n \in \mathbb{N}, \\ Q_n &:= (-(n/\gamma)^{-2}, (n/\gamma)^{-1}, 0), && n \in \mathbb{N}, \end{aligned}$$

where $\gamma = (2+\sqrt{7})/3$. We set

$$K = \mathrm{conv}(\{O\} \cup \{P_n, Q_n\}_{n\in\mathbb{N}}),$$

see also Figure 5.4.1. Since the sequences $\{P_n\}$ and $\{Q_n\}$ converge towards O, the set K is closed. In what follows, we show that K is polyhedric, but not polyhedral.

First, we show that K is polyhedric at O. We have

$$\mathcal{R}_K(O) = \operatorname{cone} K = \operatorname{cone}\{P_n, Q_n\}_{n\in\mathbb{N}} = \operatorname{cone}\{(n^{-1}, 1, -n^{-2}), (-n^{-1}, 1, 0)\}_{n\in\mathbb{N}}.$$

Now, we can show

$$\begin{aligned}\mathcal{R}_K(O) &= \operatorname{cone}(\{(n^{-1}, 1, -n^{-2})\}_{n\in\mathbb{N}} \cup \{(-1, 1, 0), (0, 1, 0)\}) \setminus \operatorname{cone}\{(0, 1, 0)\},\\ \mathcal{T}_K(O) &= \operatorname{cone}(\{(n^{-1}, 1, -n^{-2})\}_{n\in\mathbb{N}} \cup \{(-1, 1, 0), (0, 1, 0)\}).\end{aligned}$$

The ray $\operatorname{cone}\{(0, 1, 0)\}$ is not an exposed ray of $\mathcal{T}_K(O)$, cf. Figure 5.4.1. In particular, if an intersection of $\mathcal{T}_K(O)$ with a hyperplane $\mu^\perp$, $\mu \in \mathbb{R}^3$, contains $(0, 1, 0)$, we have

$$\operatorname{cone}\{(-1, 1, 0), (0, 1, 0)\} \setminus \operatorname{cone}\{(0, 1, 0)\} \subset \mathcal{R}_K(O) \cap \mu^\perp.$$

This shows $\mathcal{T}_K(O) \cap \mu^\perp = \operatorname{cl}(\mathcal{R}_K(O) \cap \mu^\perp)$ and K is polyhedric at O.

Next, it is a little bit tedious to check that K is the intersection of the half-spaces which are defined by the following inequalities and that the points on the right-hand side are exactly those points of O, P_n, Q_n which lie on the boundary of the half-spaces:

$$\begin{aligned}
x^\top(0, 0, 1) &\le 0, && O, Q_n \quad \forall n \in \mathbb{N},\\
x^\top(1, \gamma, 1+\gamma) &\ge 0, && O, Q_1, P_1,\\
x^\top(2\,k+1, -1, k\,(k+1)) &\le 0, && O, P_k, P_{k+1},\\
x^\top(a_k^{(1)}, b_k^{(1)}, c_k^{(1)}) &\le \gamma^2, && Q_k, Q_{k+1}, P_k,\\
x^\top(a_k^{(2)}, b_k^{(2)}, c_k^{(2)}) &\le 1, && P_k, P_{k+1}, Q_{k+1},
\end{aligned}$$

where $k \in \mathbb{N}$. In the last two lines, we have used the coefficients

$$\begin{aligned}
a_k^{(1)} &:= k\,(k+1), & b_k^{(2)} &:= \frac{\gamma^2\,(3\,k^2+3\,k+1) + (k+1)^2}{(2\,k+1)\,\gamma^2 + (k+1)\,\gamma},\\
b_k^{(1)} &:= \gamma\,(2\,k+1), & a_k^{(2)} &:= 3\,k^2+3\,k+1-(2\,k+1)\,b_k^{(2)},\\
c_k^{(1)} &:= k\,a_k^{(1)} + k^2\,b_k^{(1)} - \gamma^2\,k^3, & c_k^{(2)} &:= k\,a_k^{(2)} + k^2\,b_k^{(2)} - k^3.
\end{aligned}$$

From this representation of K, we learn two things. First, all P_k, Q_k are extreme points of K and, thus, K is not polyhedral. Second, the intersection $K \cap \{x \in \mathbb{R}^3 \mid x^\top(1, 1, 1) \ge \varepsilon\}$ is a polyhedron for all $\varepsilon > 0$, since it can be written as a finite intersection of half-spaces. Thus, $\mathcal{R}_K(x)$ is closed for all $x \in K \setminus \{O\}$.

Hence, we have shown that K is polyhedric, but not polyhedral. Moreover, it is easy to check that the intersection of K with the hyperplane $x^\top(0, 0, 1) = 0$ is not polyhedric. Hence, K is not 2-polyhedric, cf. Lemma 5.4.4.

Finally, we mention that the example can be lifted to $\mathbb{R}^n$, $n > 3$, by considering $K \times \mathbb{R}^{n-3}$.

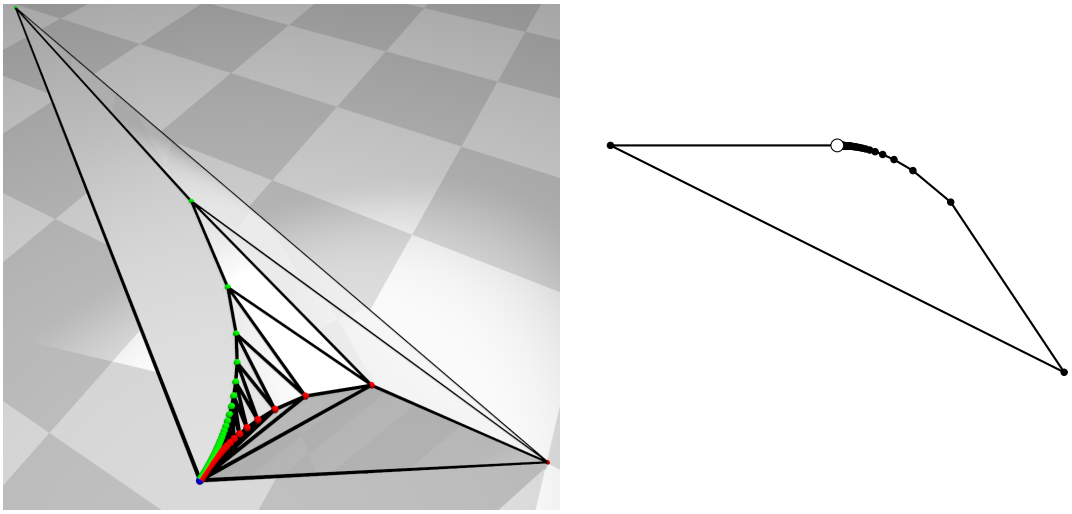

Figure 5.4.1.: The polyhedric, non-polyhedral set K from Example 5.4.24 is shown left. The points Q_n (on the left face) and P_n (on the right) are colored green and red, respectively. The right figure shows the intersection of $\mathcal{R}_K(O)$ with the two-dimensional hyperplane $x^\top(0,1,0) = 1$. Note that the hollow dot (corresponding to the non-exposed ray in direction $(0,1,0)$ of $\mathcal{T}_K(O)$) does not belong to this intersection.

However, it can be shown that all bounded, $(n-1)$-polyhedric sets in $\mathbb{R}^n$ are polyhedral. Indeed, if K is bounded and $(n-1)$-polyhedric, intersections of K with two-dimensional subspaces are bounded and polyhedric by Lemma 5.4.4. In $\mathbb{R}^2$, bounded and polyhedric sets are precisely the bounded polyhedral sets and this is easy to prove. Now, we can invoke Klee, 1959, Theorem 4.7 to obtain the polyhedrality of K. Moreover, the same reasoning shows that $(n-1)$-polyhedric sets in $\mathbb{R}^n$ are boundedly polyhedral in the sense of Klee, 1959, that is, they have polyhedral intersections with all bounded polyhedral sets.

5.4.4. Summary

We briefly summarize the findings of Section 5.4. We provided results concerning the polyhedricity of the intersection of sets. These results are collected in Table 5.4.1. The reader should bear in mind that "with bounds" refer to a set with bounds in the sense of Definition 5.4.9 in a Banach space with a lattice structure satisfying Assumption 5.4.7, see also Example 5.4.23. Moreover, for the polyhedricity of the intersection of two sets with bounds we require that both sets satisfy Assumption 5.4.10 w.r.t. the same ordering in X. Then, the intersection is again a set with bounds and, thus, polyhedric. The other positive results are straightforward (lower right corner) or follow from Corollary 5.4.19. Moreover, Examples 5.4.22 to 5.4.24 show the failure of polyhedricity for all negative cases.

	polyhedric	with bounds	polyhedral	co-polyhedral
polyhedric	✗	✗	✗	✗
with bounds	✗	✓	✓	✗
polyhedral	✗	✓	✓	✓
co-polyhedral	✗	✗	✓	✓

Table 5.4.1.: A summary concerning the polyhedricity of intersections. The notions of the sets are given in Definitions 5.3.1 and 5.4.9. A check mark (✓) indicates intersections which are always polyhedric, whereas a cross (✗) signifies that this type of intersection may fail to be polyhedric.

5.5. Applications of polyhedricity

In order to complete the picture of polyhedricity, we illustrate its important applications. For the convenience of the reader, we included the rather short proofs.

5.5.1. Directional differentiability of projections

The very first application of polyhedricity was a result on the differentiability of (metric) projections onto closed, convex sets. In fact, these differentiability results motivated the introduction of the notion of polyhedricity in Mignot, 1976; Haraux, 1977.

We state the results in the generality of Mignot, 1976, Section 2, but use the proofs of Haraux, 1977, Section 1. Another difference to both works is that we do not use the identification $H \cong H^\star$, but we will explicitly work with the dual space $H^\star$.

In this section, we use the following standing assumption.

Assumption 5.5.1. The space H is a real Hilbert space, $A\colon H \to H^\star$ is a bounded, linear operator which is coercive, i.e., there exists $\alpha > 0$ such that

$$\langle A\,x,\, x\rangle \geq \underline{\alpha}\,\|x\|^2 \quad \forall x \in H,$$

Finally, $K \subset H$ is a closed, convex set.

Note that it is also possible to work with the bounded, coercive and bilinear form a defined via $a(x,y) = \langle A\,x,\, y\rangle$.

We further point out that the coercivity of A already implies that H is isomorphic to a Hilbert space since $a_s(x,y) := (a(x,y) + a(y,x))$ is a scalar product on H and the induced norm is equivalent to the original norm of H.

For a given functional $\ell \in H^\star$, we consider the variational inequality

$$\text{Find} \quad y \in K \qquad \text{such that} \quad \langle A\,y - \ell,\, v - y\rangle \geq 0 \quad \forall v \in K. \tag{5.5.1}$$

In the case that A is symmetric, i.e., $\langle A\,x,\,y\rangle = \langle A\,y,\,x\rangle$ for all $x, y \in H$, this problem is equivalent to a projection problem. In particular, y is the projection of the Riesz representative of ℓ (w.r.t. the a-scalar product) onto K w.r.t. the norm $\|x\|_A^2 := \langle A\,x,\,x\rangle = a(x,x)$.

It is well known that for all $\ell \in H^\star$ there is a unique solution $y := S(\ell)$ of (5.5.1), see Stampacchia, 1964. Moreover, this solution operator $S\colon H^\star \to H$ is globally Lipschitz continuous. From (5.5.1) we immediately find $\ell - A\,y \in \mathcal{T}_K(y)^\circ$. We recall the definition of the critical cone $\mathcal{K}_K(y,\lambda) := \mathcal{T}_K(y) \cap \lambda^\perp$.

Theorem 5.5.2. Let $\ell \in H^\star$ be given and let $y := S(\ell)$, $\lambda := \ell - A\,y$. Moreover, assume that K is polyhedric at (y,λ). Then, S is directionally differentiable at ℓ and the directional derivative $\delta y := S'(\ell;h)$ in direction $h \in H^\star$ solves the variational inequality

$$\delta y \in \mathcal{K}_K(y,\lambda) \qquad \text{and} \qquad \langle A\,\delta y - h,\, v - \delta y\rangle \geq 0 \quad \forall v \in \mathcal{K}_K(y,\lambda). \tag{5.5.2}$$

Proof. Let a direction $h \in H^\star$ be given. For $t > 0$, we define $y_t := S(\ell + t\,h)$ and $\lambda_t := \ell + t\,h - A\,y_t$. We find $y_t \in K$ and $\lambda_t \in \mathcal{T}_K(y_t)^\circ$, see (5.5.1). Using the Lipschitz continuity of S we find

$$\Big\|\frac{y_t - y}{t}\Big\|_H \leq C \qquad \text{and} \qquad \Big\|\frac{\lambda_t - \lambda}{t}\Big\|_{H^\star} \leq C.$$

Since H and $H^\star$ are reflexive, we find a subsequence (without relabeling) and $\delta y \in H$, $\delta\lambda \in H^\star$ such that

$$\frac{y_t - y}{t} \rightharpoonup \delta y \qquad \text{and} \qquad \frac{\lambda_t - \lambda}{t} \rightharpoonup \delta\lambda.$$

From $y_t \in K$ we immediately obtain $\delta y \in \mathcal{T}_K(y)$.

Now, we consider an arbitrary $x \in K$ with $\langle \lambda,\, x - y\rangle = 0$. We obtain

$$\langle \delta\lambda,\, x - y\rangle \leftarrow \Big\langle \frac{\lambda_t - \lambda}{t},\, x - y_t\Big\rangle \leq \Big\langle \frac{-\lambda}{t},\, x - y_t\Big\rangle = \Big\langle \frac{-\lambda}{t},\, y - y_t\Big\rangle \to \langle \lambda,\, \delta y\rangle \leq 0.$$

This shows

$$\delta\lambda \in \{z \in K - y \mid \langle \lambda,\, z\rangle = 0\}^\circ = \big(\mathcal{R}_K(y) \cap \lambda^\perp\big)^\circ$$

and by using $x = y$ we find $\delta y \in \lambda^\perp$.

The definition $\lambda_t = \ell + t\,h - A\,y_t$ implies immediately $\delta\lambda = h - A\,\delta y$. Together with the polyhedricity of K, we find that $\delta y \in \mathcal{T}_K(y) \cap \lambda^\perp = \mathcal{K}_K(y,\lambda)$ solves the VI (5.5.2). Thus, the weak limit δy is independent of the chosen subsequence and a standard argument shows the convergence of the entire sequence.

It remains to show the strong convergence $(y_t - y)/t \to \delta y$. By testing the variational inequality (5.5.1) for y with $v = y_t$ and vice versa, we obtain

$$\Big\langle A\,\frac{y_t - y}{t} - h,\, \frac{y_t - y}{t}\Big\rangle \leq 0.$$

Using the weak convergence of the difference quotient and the coercivity of A, we find

$$0 \le -\langle \delta\lambda,\, \delta y\rangle = \langle A\,\delta y - h,\, \delta y\rangle \le \liminf_{t\searrow 0}\Big\langle A\,\frac{y_t - y}{t} - h,\, \frac{y_t - y}{t}\Big\rangle$$
$$\le \limsup_{t\searrow 0}\Big\langle A\,\frac{y_t - y}{t} - h,\, \frac{y_t - y}{t}\Big\rangle \le 0.$$

This shows the convergence

$$\langle A\,\delta y,\, \delta y\rangle = \lim_{t\searrow 0}\Big\langle A\,\frac{y_t - y}{t},\, \frac{y_t - y}{t}\Big\rangle$$

and, by coercivity of A, the strong convergence $(y_t - y)/t \to \delta y$ follows.

We mention some extensions of Theorem 5.5.2. First of all, the result Haraux, 1977, Theorem 1 also allows for non-polyhedric sets K by including an operator L, which takes into account the curvature of the boundary of K. Levy, 1999, Theorem 1.1 treats nonlinear variational inequalities.

Theorem 5.5.2 can be applied to show the shape differentiability of the solution mapping to variational inequalities, see Sokołowski, Zolésio, 1992, Chapter 4, and to compute the graphical derivative of the normal cone mapping to K, see Lemma 2.3.10, and Mordukhovich, Outrata, Ramírez C., 2015a, Theorem 5.2.

5.5.2. Optimal control of VIs: strong stationarity

A second application of polyhedricity is the derivation of optimality conditions for the optimal control of variational inequalities, and was first considered in Mignot, 1976, Proposition 4.1. This is strongly linked to the differentiability result Theorem 5.5.2. The same technique can be used to derive optimality conditions for control problems involving other non-smooth maps, if the derivative can be written as the solution of a variational inequality.

In addition to Assumption 5.5.1, we use the following setting: $f \in H^\star$, U is a real Banach space, $B \in \mathcal{L}(U, H^\star)$ has a dense range, and $J\colon H \times U \to \mathbb{R}$ is Fréchet differentiable.

We consider the optimal control problem

$$\begin{aligned} &\text{Minimize} && J(y,u)\\ &\text{such that} && y \in K \text{ and } \langle A\,y - B\,u - f,\, v - y\rangle_{H^\star,H} \ge 0 \quad \forall v \in K. \end{aligned} \tag{5.5.3}$$

Theorem 5.5.3. Suppose that $(\bar y, \bar u) \in H \times U$ is locally optimal for (5.5.3). We set $\bar\lambda := B\,\bar u + f - A\,\bar y$ and assume that K is polyhedric at $(\bar y, \bar\lambda)$. Then, there exist $p \in H$,

$\mu \in H^\star$ satisfying the optimality system

$$J_y(\cdot) + A^\star\, p + \mu = 0, \tag{5.5.4a}$$
$$J_u(\cdot) - B^\star\, p = 0, \tag{5.5.4b}$$
$$p \in -\mathcal{K}_K(\bar{y}, \bar{\lambda}), \tag{5.5.4c}$$
$$\mu \in \mathcal{K}_K(\bar{y}, \bar{\lambda})^\circ. \tag{5.5.4d}$$

Here, $J_y(\cdot) \in H^\star$ and $J_u(\cdot) \in U^\star$ are the partial Fréchet derivatives of J at $(\bar{u}, \bar{y})$.

Proof. We follow the proof of Mignot, 1976, Proposition 4.1.

As in Section 5.5.1, we denote by S the solution operator of the variational inequality and use the reduced problem

$$\text{Minimize} \quad J(S(B\,u + f), u).$$

Since $\bar{u}$ is a local minimizer and using Theorem 5.5.2, we have

$$\langle J_y(\cdot),\, S'(B\,\bar{u} + f; B\,h)\rangle_{H^\star,H} + \langle J_u(\cdot),\, h\rangle_{U^\star,U} \geq 0 \qquad \forall h \in U.$$

Since A is coercive, we get

$$\|S'(B\,\bar{u} + f; B\,h)\|_H \leq C\,\|B\,h\|_{H^\star} \qquad \forall h \in U.$$

and this yields

$$|\langle J_u(\cdot),\, h\rangle_{U^\star,U}| \leq C\,\|B\,h\|_{H^\star}. \qquad \forall h \in U.$$

Hence, there is $p \in H^{\star\star} \cong H$ (by defining it as in the next line on the dense subspace $\text{image}(B) \subset H^\star$ and extending it by continuity on the whole space $H^\star$) such that

$$\langle J_u(\cdot),\, h\rangle_{U^\star,U} = \langle p,\, B\,h\rangle_{H,H^\star} \qquad \forall h \in U.$$

This yields (5.5.4b) and

$$\langle J_y(\cdot),\, S'(B\,\bar{u} + f; B\,h)\rangle_{H^\star,H} + \langle p,\, B\,h\rangle_{H,H^\star} \geq 0 \qquad \forall h \in U.$$

Using the density of $\text{image}(B)$ in $H^\star$ we get

$$\langle J_y(\cdot),\, S'(B\,\bar{u} + f; h)\rangle_{H^\star,H} + \langle p,\, h\rangle_{H,H^\star} \geq 0 \qquad \forall h \in H^\star. \tag{$*$}$$

In what follows, we set $\mathcal{K} := \mathcal{K}_K(\bar{y}, \bar{\lambda})$ for convenience. We choose $h \in \mathcal{K}^\circ$ in ($*$). Then, (5.5.2) shows $S'(\bar{u} + f; h) = 0$ and, thus,

$$\langle p,\, h\rangle_{H,H^\star} \geq 0 \qquad \forall h \in \mathcal{K}^\circ,$$

i.e., $p \in -\mathcal{K}$, which shows (5.5.4c).

Now, we choose $h \in A\,\mathcal{K}$ in $(*)$, hence $S'(B\,\bar{u} + f; h) = A^{-1}\,h$ by (5.5.2). We get

$$\langle J_y(\cdot),\, A^{-1}\,h\rangle_{H^\star,H} + \langle p,\, h\rangle_{H,H^\star} \geq 0 \qquad \forall h \in A\,\mathcal{K}.$$

Rearranging terms yields

$$\langle J_y(\cdot) + A^\star\,p,\, v\rangle_{H^\star,H} \geq 0 \qquad \forall v \in \mathcal{K}.$$

Now, we define $\mu := -J_y(\cdot) - A^\star\,p$ and get $\mu \in \mathcal{K}^\circ$. This shows (5.5.4a) and (5.5.4d).

The crucial ingredients in the proof of Theorem 5.5.3 are the density of the image of B and the absence of control constraints on u. In the presence of control constraints, system (5.5.4) is no longer necessary without further assumptions. We refer to Section 3.6 for counterexamples.

Using terminology from finite-dimensional optimization, the optimality system (5.5.4) is of strongly stationary type. By using methods from variational analysis (in particular, Levy, 1999, Theorem 3.1), the same result is shown in Hintermüller, Surowiec, 2011, Theorem 4.6.

For more general optimization problems with complementarity constraints in Banach spaces a notion of strong stationarity was introduced in Chapter 1. In the polyhedric case, this notion is of reasonable strength and it is a necessary optimality condition under some constraint qualification, see Section 1.5.3.

5.5.3. No-gap second-order conditions

As a last application of polyhedricity, we consider second-order optimality conditions. In particular, polyhedricity is one possibility obtain no gap between the necessary and sufficient optimality conditions of second order.

In this section, we consider the optimization problem

$$\begin{aligned} &\text{minimize} && f(x), \\ &\text{subject to} && g(x) \in K. \end{aligned} \tag{5.5.5}$$

Here, X, Y are Banach spaces and the mappings $f\colon X \to \mathbb{R}$, $g\colon X \to Y$ are assumed to be twice Fréchet differentiable at $\bar{x}$, where $\bar{x} \in X$ satisfies $g(\bar{x}) \in K$. The set $K \subset Y$ is closed and convex.

The Lagrangian associated with problem (5.5.5) is

$$L(x, \lambda) := f(x) + \langle \lambda,\, g(x)\rangle$$

and $\bar{\lambda} \in Y^\star$ is said to be a Lagrange multiplier at $\bar{x}$, if $\bar{\lambda} \in \mathcal{T}_K(g(\bar{x}))^\circ$ and $0 = L'(\bar{x}, \bar{\lambda}) = f'(\bar{x}) + g'(\bar{x})^\star\,\bar{\lambda}$. Here, L' is the partial Fréchet-derivative of L w.r.t. x. The second

partial Fréchet-derivative of L w.r.t. x is denoted by L'' and we write $L''(\bar{x}, \bar{\lambda})\, h^2 := L''(\bar{x}, \bar{\lambda})[h, h]$.

The critical cone $C(\bar{x})$ at $\bar{x}$ is defined as

$$C(\bar{x}) := \{h \in X \mid g'(\bar{x})\, h \in \mathcal{T}_K(g(\bar{x})) \text{ and } f'(\bar{x})\, h \leq 0\}.$$

If $\bar{\lambda}$ is a Lagrange multiplier, it is easy to check that

$$C(\bar{x}) = \{h \in X \mid g'(\bar{x})\, h \in \mathcal{T}_K(g(\bar{x})) \cap \bar{\lambda}^{\perp}\} = g'(\bar{x})^{-1}\, \mathcal{K}_K(g(\bar{x}), \bar{\lambda}).$$

We start with a result on second-order necessary conditions. The first part of the following theorem was given in Maurer, Zowe, 1979, Theorem 3.3.

Theorem 5.5.4. We assume that $\bar{x}$ is a local minimizer of (5.5.5) and let $\bar{\lambda} \in \mathcal{T}_K(g(\bar{x}))^{\circ}$ denote a corresponding Lagrange multiplier, i.e.,

$$f'(\bar{x}) + g'(\bar{x})^{\star}\, \bar{\lambda} = 0.$$

We further assume that the condition

$$Y = g'(\bar{x})\, X - (\mathcal{R}_K(g(\bar{x})) \cap \bar{\lambda}^{\perp}) \tag{5.5.6}$$

is satisfied. Then, we have

$$L''(\bar{x}, \bar{\lambda})\, h^2 \geq 0 \qquad \forall h \in X : g'(\bar{x})\, h \in \mathcal{R}_K(\bar{x}), f'(\bar{x})\, h = 0. \tag{5.5.7}$$

If K is polyhedric at $(g(\bar{x}), \bar{\lambda})$, we get

$$L''(\bar{x}, \bar{\lambda})\, h^2 \geq 0 \qquad \forall h \in C(\bar{x}). \tag{5.5.8}$$

We mention that the condition (5.5.6) is, strictly speaking, not a constraint qualification, since it depends on the objective f via the multiplier $\bar{\lambda}$. In a finite-dimensional setting, it reduces to the so-called strict Mangasarian-Fromovitz condition.

The proof of (5.5.7) is straightforward by using

$$\mathcal{T}_F(\bar{x}) = g'(\bar{x})^{-1}\, (\mathcal{T}_K(g(\bar{x})) \cap \bar{\lambda}^{\perp})$$

due to (5.5.6), see Bonnans, Shapiro, 2000, Corollary 2.91. Here, $F := \{x \in X \mid g(x) \in K \text{ and } g(x) - g(\bar{x}) \in \bar{\lambda}^{\perp}\}$. Moreover, (5.5.8) follows from (5.5.7) by using the following lemma which is just a translation of Lemma 5.3.3 to the critical cone $C(\bar{x})$.

Lemma 5.5.5. Let us assume that $\bar{\lambda} \in \mathcal{T}_K(g(\bar{x}))^{\circ}$ satisfies $f'(\bar{x}) + g'(\bar{x})^{\star}\, \bar{\lambda} = 0$ and (5.5.6). Moreover, we suppose that K is polyhedric at $(g(\bar{x}), \bar{\lambda})$. Then, problem (5.5.5)

satisfies the strong extended polyhedricity condition at $\bar{x}$ in the sense of Bonnans, Shapiro, 2000, Definition 3.52. That is,

$$C_R(\bar{x}) := \{h \in X \mid g'(\bar{x})\,\mathcal{R}_K(\bar{x}) \text{ and } f'(\bar{x})\,h = 0\}$$

is dense in the critical cone $C(\bar{x})$.

Proof. We apply Lemma 5.3.3 with the setting

$$S = g'(\bar{x}), \qquad K_y = K - g(\bar{x}), \qquad \mu = g'(\bar{x})\,\bar{\lambda}.$$

Recall that (5.5.6) implies the uniqueness of the multiplier $\bar{\lambda}$, see Shapiro, 1997a, hence, we get $\lambda = \bar{\lambda}$ in Lemma 5.3.3. An easy calculation using (5.3.6) shows

$$C(\bar{x}) = \{h \in X \mid g'(\bar{x})\,h \in \mathcal{T}_{K-g(\bar{x})}(0) \cap \bar{\lambda}^{\perp}\} = \mathcal{T}_{g'(\bar{x})^{-1}(K-g(\bar{x}))}(0) \cap \mu^{\perp}.$$

Hence, Lemma 5.3.3 implies that this set is the closure of

$$\mathcal{R}_{g'(\bar{x})^{-1}(K-g(\bar{x}))}(0) \cap \mu^{\perp} = g'(\bar{x})^{-1}\,\mathcal{R}_K(g(\bar{x})) \cap \mu^{\perp} = C_R(\bar{x}).$$

Another route to second-order necessary conditions is the utilization of so-called second-order tangent sets, see Bonnans, Shapiro, 2000, Section 3.2. Then, the necessary condition is similar to (5.5.8), but contains a term corresponding to the curvature of the set K, see Bonnans, Shapiro, 2000, Theorem 3.45. Under the "extended polyhedricity condition", one arrives at (5.5.8). Indeed, using Lemma 5.5.5, Theorem 5.5.4 is also an easy corollary of Bonnans, Shapiro, 2000, Proposition 3.53.

Next, we provide a second-order sufficient condition.

Theorem 5.5.6. We assume that $\bar{\lambda}$ is a Lagrange multiplier at $\bar{x}$. Further, we assume that X is finite-dimensional or that the constraint qualification

$$Y = g'(\bar{x})\,X - \mathcal{T}_K(g(\bar{x})) \tag{5.5.9}$$

is satisfied. Suppose that

$$L''(\bar{x}, \bar{\lambda})\,h^2 \geq \alpha\,\|h\|_X^2 \quad \forall h \in X : g'(\bar{x})\,h \in \mathcal{T}_K(g(\bar{x})) \text{ and } f'(\bar{x})\,h \leq \eta\,\|h\|_X \tag{5.5.10}$$

holds for some $\alpha > 0$, $\eta > 0$. Then, for all $\tilde{\alpha} \in (0, \alpha/2)$ there is $\varepsilon > 0$, such that

$$f(x) \geq f(\bar{x}) + \tilde{\alpha}\,\|x - \bar{x}\|_X^2 \quad \forall x \in F \cap B_\varepsilon(\bar{x}) \tag{5.5.11}$$

holds.

Proof. We verify (5.5.11) by contradiction. Assume that (5.5.11) does not hold. Then, there is a sequence $\{x_n\}$ with $g(x_n) \in K$ and $x_n \neq \bar{x}$ such that $x_n \to x$ and $f(x_n) < f(\bar{x}) + \tilde{\alpha} \|x_n - \bar{x}\|_X^2$. Next, we construct a sequence $\{h_n\}_{n\in\mathbb{N}} \subset X$, such that $\|h_n\|_X = o(\|x_n - \bar{x}\|_X)$ and $g'(\bar{x})\,(x_n - \bar{x} + h_n) \in \mathcal{T}_K(g(\bar{x}))$ for all $n \in \mathbb{N}$.

(a) In the case that X is finite-dimensional, we have w.l.o.g. $(x_n - \bar{x})/\|x_n - \bar{x}\|_X \to h$, thus $g'(\bar{x})\,h \in \mathcal{T}_K(\bar{x})$. We set $h_n := \|x_n - \bar{x}\|_X\, h - (x_n - \bar{x})$ and have $h_n = o(\|x_n - \bar{x}\|_X)$ and $x_n - \bar{x} + h_n = \|x_n - \bar{x}\|_X\, h \in g'(\bar{x})^{-1}\,\mathcal{T}_K(g(\bar{x}))$.

(b) Now, suppose that the condition (5.5.9) is satisfied. Since g is Fréchet-differentiable, we have $r_n := g(\bar{x}) + g'(\bar{x})\,(x_n - \bar{x}) - g(x_n) = o(\|x_n - \bar{x}\|_X)$. Due to the constraint qualification (5.5.9), we can apply the generalized open mapping theorem Zowe, Kurcyusz, 1979, Theorem 2.1 and obtain the existence of $h_n \in X$, $v_n \in \mathcal{T}_K(g(\bar{x}))$ satisfying

$$-r_n = g'(\bar{x})\,h_n - v_n$$

and $h_n = \mathcal{O}(\|r_n\|_X) = o(\|x_n - \bar{x}\|_X)$. In particular, we have

$$g'(\bar{x})\,(x_n - \bar{x} + h_n) = g(x_n) - g(\bar{x}) + v_n \in \mathcal{T}_K(g(\bar{x})).$$

Thus, $x_n - \bar{x} + h_n \in g'(\bar{x})^{-1}\,\mathcal{T}_K(g(\bar{x}))$.

Since

$$\begin{aligned} f'(\bar{x})\,(x_n - \bar{x}) &= f(x_n) - f(\bar{x}) + o(\|x_n - \bar{x}\|_X) \\ &\leq \tilde{\alpha}\,\|x_n - \bar{x}\|_X^2 + o(\|x_n - \bar{x}\|_X) = o(\|x_n - \bar{x}\|_X) \end{aligned}$$

we conclude $f'(\bar{x})\,(x_n - \bar{x} + h_n) = o(\|x_n - \bar{x}\|_X) = o(\|x_n - \bar{x} + h_n\|_X)$. Hence, $x_n - \bar{x} + h_n$ belongs to the set on the right-hand side of (5.5.10) for n large enough.

Finally, we have for n large enough

$$\begin{aligned} \tilde{\alpha}\,\|x_n - \bar{x}\|^2 &\geq f(x_n) - f(\bar{x}) \geq L(x_n, \bar{\lambda}) - L(\bar{x}, \bar{\lambda}) \\ &= L'(\bar{x}, \bar{\lambda})\,(x_n - \bar{x}) + \frac{1}{2}\,L''(\bar{x}, \bar{\lambda})\,(x_n - \bar{x})^2 + o(\|x_n - \bar{x}\|_X^2) \\ &= \frac{1}{2}\,L''(\bar{x}, \bar{\lambda})\,(x_n - \bar{x} + h_n)^2 - L''(\bar{x}, \bar{\lambda})[x_n - \bar{x}, h_n] \\ &\quad - \frac{1}{2}\,L''(\bar{x}, \bar{\lambda})\,h_n^2 + o(\|x_n - \bar{x}\|_X^2) \\ &= \frac{1}{2}\,L''(\bar{x}, \bar{\lambda})\,(x_n - \bar{x} + h_n)^2 + o(\|x_n - \bar{x}\|_X^2) \\ &\geq \frac{\alpha}{2}\,\|x_n - \bar{x}\|^2 + o(\|x_n - \bar{x}\|_X^2) \end{aligned}$$

and this is a contradiction to $\tilde{\alpha} < \alpha/2$.

Without the constraint qualification (5.5.9) we have to work with a larger critical cone,

cf. Bonnans, Shapiro, 2000, Lemma 3.65 and Remark 3.68. In case that the Lagrange multiplier $\bar{\lambda}$ is not unique, it is easy to see that the left-hand side in (5.5.10) can be replaced by $\sup_\lambda L''(\bar{x}, \lambda)\, h^2$, where the sup ranges over a bounded set of Lagrange multipliers λ, by using essentially the same proof, see also Bonnans, Shapiro, 2000, Section 3.3.1. Recall that the set of Lagrange multipliers is already bounded if the constraint qualification (5.5.9) is satisfied, cf. Zowe, Kurcyusz, 1979, Theorem 4.1.

In the case that $L''(\bar{x}, \bar{\lambda})$ is a Legendre form, see Bonnans, Shapiro, 2000, Definition 3.73, we can apply Bonnans, Shapiro, 2000, Lemma 3.75 and get the equivalence of (5.5.10) with

$$L''(\bar{x}, \bar{\lambda})\, h^2 > 0 \qquad \forall h \in C(\bar{x}) \setminus \{0\},$$

in the case that X is reflexive.

Putting everything together, we get the following second-order conditions with no gap.

Theorem 5.5.7. We assume that $\bar{\lambda}$ is a Lagrange multiplier at $\bar{x}$ and

- K is polyhedric at $(g(\bar{x}), \bar{\lambda})$,
- the condition (5.5.6) holds,
- $L''(\bar{x}, \bar{\mu})$ is a Legendre form and X is reflexive.

Then, the following hold.

(i) If $\bar{x}$ is locally optimal, we have

$$L''(\bar{x}, \bar{\lambda})\, h^2 \geq 0 \qquad \forall h \in C(\bar{x}).$$

(ii) The second-order condition

$$L''(\bar{x}, \bar{\lambda})\, h^2 > 0 \qquad \forall h \in C(\bar{x}) \setminus \{0\}$$

is equivalent to the quadratic growth condition (5.5.11). In particular, $\bar{x}$ is a strict local minimizer.

This results provides an indication that polyhedricity plays a crucial role for deriving second-order conditions with no (or a very small) gap. We refer exemplarily to the contributions Bonnans, 1998; Bonnans, Zidani, 1999; Casas, Tröltzsch, 2012 for similar results by using polyhedricity and further extensions.

5.6. Conclusions

We have given an overview of polyhedric sets. Moreover, we have expanded the classical polyhedricity results from Mignot, 1976; Haraux, 1977 and thus we are able to show the polyhedricity of intersections of sets with bounds with polyhedral sets. On the other hand, we have given some counterexamples which show that intersections of polyhedric

sets might be non-polyhedric in various situations. From this point of view, it would be interesting to have further results concerning the intersections of polyhedric sets.

Acknowledgement

Frédéric Bonnans suggested to look at intersections of the non-negative functions in $L^2(\Omega)$ with polyhedral sets and this led to the discovery of Theorem 5.4.18. The author is indebted to Constantin Christof for providing the main ideas of the non-polyhedricity of Λ in Example 5.4.23. Example 5.4.24 grew from some fruitful discussions with Alexander Shapiro, which are gratefully acknowledged.

6. Pointwise constraints in vector-valued Sobolev spaces with applications in optimal control

Abstract: We consider a set $\mathcal{C}$ with pointwise constraints in a vector-valued Sobolev space. We characterize its tangent and normal cone. Under the additional assumption that the pointwise constraints are affine and satisfy the linear independence constraint qualification, we show that the set $\mathcal{C}$ is polyhedric. The results are applied to the optimal control of a string in a polyhedral tube.

Keywords: tangent cone, normal cone, polyhedricity, vector-valued function, vector-valued measure

MSC: 46N10, 49K21

6.1. Introduction

Let $\Omega \subset \mathbb{R}^d$ be a bounded, open set and let $C \subset \mathbb{R}^m$ be a closed, convex set with $0 \in \operatorname{int}(C)$. We consider the set

$$\mathcal{C} := \{v \in H_0^1(\Omega)^m : v(\omega) \in C \text{ for almost all } \omega \in \Omega\}. \tag{6.1.1}$$

We mention that such sets $\mathcal{C}$ appear, e.g., in the study of vector-valued evolution variational inequalities, see Krejčí, 1999, Section 7, and in vector-valued obstacle problems, see, e.g., Hildebrandt, Widman, 1979; Dal Maso, Musina, 1989; Mancini, Musina, 1989. Note that the latter case also includes phase-field models in which the concentrations of the phases belong to the simplex $C = \{x \in \mathbb{R}^m : x \geq 0,\ \sum_{i=1}^m x_i \leq 1\}$.

In this work, we are going to characterize the tangent and normal cone to the set $\mathcal{C}$. Finally, we prove the polyhedricity of $\mathcal{C}$ in case that C is described by affine constraints which satisfy the linear independence constraint qualification (LICQ). We recall that $\mathcal{C}$ is called polyhedric if

$$\mathcal{T}_{\mathcal{C}}(u) \cap \xi^{\perp} = \operatorname{cl}(\mathcal{R}_{\mathcal{C}}(u) \cap \xi^{\perp}) \tag{6.1.2}$$

holds for all $u \in \mathcal{C}$ and $\xi \in \mathcal{T}_{\mathcal{C}}(u)^{\circ}$, see Haraux, 1977. We refer to Section 6.2.1 for the corresponding notation. These results are applied to an optimal control problem in which the state equation is given by a variational inequality over the set $\mathcal{C}$.

To our knowledge, the convex analysis of the set $\mathcal{C}$ was not studied previously in the vector-valued case $m > 1$. In the scalar case $m = 1$, the set $\mathcal{C}$ was studied in Mignot, 1976.

Therein, the set C was allowed to depend also on the spatial variable $\omega \in \Omega$, and Mignot characterized the tangent and normal cone. Moreover, he proved the polyhedricity of $\mathcal{C}$. However, the arguments in Mignot, 1976 heavily exploit the fact that the non-negative functions induce a lattice structure in $H_0^1(\Omega)$. Similarly, Haraux, 1977, Corollary 2 and Bonnans, Shapiro, 2000, Theorem 3.58 prove the polyhedricity of a cone which induces a lattice structure. These results are extended to more general function spaces in Frémiot et al., 2009, Section 4, see also Rao, Sokołowski, 1993 for the case $H_0^2(\Omega)$.

This approach for proving polyhedricity even fails for some polyhedral cones in finite dimensions. To give an example, the polyhedral cone

$$K = \{x \in \mathbb{R}^3 : (-1,1,0)\, x \leq 0,\ (-1,-1,0)\, x \leq 0,\ (-1,0,1)\, x \leq 0,\ (-1,0,-1)\, x \leq 0\}$$

does not induces a lattice structure on $\mathbb{R}^3$. Hence, the polyhedricity of K cannot be deduced from the above mentioned theorems. On the other hand, since K is a polyhedral set, its polyhedricity is evident.

Finally, we mention the results Bonnans, 1998, Lemma 4.2, Proposition 4.3, Bonnans, Shapiro, 2000, Lemma 6.34, Proposition 6.35. These results characterize the tangent and normal cone and show the polyhedricity of a set with pointwise affine constraints in the Lebesgue space $L^s(\Omega)^m$, $s \in [1,\infty)$. The arguments utilize the Lipschitz continuity of a pointwise projection in $L^s(\Omega)^m$. Since a pointwise projection is not Lipschitz continuous in $H_0^1(\Omega)$, the arguments cannot be generalized to the situation at hand.

We mention that the polyhedricity of a set has some important applications. First of all, one can show that the projection onto a polyhedric set in a Hilbert space is directionally differentiable, see Mignot, 1976; Haraux, 1977. This can also be extended to a (directional) shape sensitivity for the solutions of variational inequalities, see Sokołowski, Zolésio, 1987 and Sokołowski, Zolésio, 1992, Section 4. Second, in the presence of polyhedricity, one can provide no-gap second-order optimality conditions, see Bonnans, Shapiro, 2000, Section 3.3.3. Finally, polyhedricity is utilized to obtain optimality conditions of strongly stationary type for infinite-dimensional optimization problems with complementarity constraints, see, e.g., Mignot, 1976, Proposition 4.1 and Section 6.1.

We briefly describe the content of this work. Some notation from convex analysis and capacity theory is introduced in Section 6.2. We characterize the tangent cone $\mathcal{T}_{\mathcal{C}}(u)$ and the normal cone $\mathcal{T}_{\mathcal{C}}(u)^\circ$ of $\mathcal{C}$ in Section 6.3 and Section 6.4, respectively. Under an additional assumption, we deduce the polyhedricity of $\mathcal{C}$ in Section 6.5 and this result is applied to an optimal control problem in Section 6.6. In the appendices, we give various auxiliary results, which are of independent interest. A chain rule for the Nemytskii operator associated with a vector-valued truncation is given in Section 6.A. In Section 6.B, we show that the positive part of a measure in $H^{-1}(\Omega)$ may not belong to $H^{-1}(\Omega)$. Section 6.C contains auxiliary results on polyhedral sets in $\mathbb{R}^m$, which are utilized to infer the polyhedricity of $\mathcal{C}$ in Section 6.5.

6.2. Notation and preliminaries

The Euclidean norm in $\mathbb{R}^m$ is denoted by $|\cdot|_{\mathbb{R}^m}$, and we use $(\cdot\,,\,\cdot)_{\mathbb{R}^m}$ for the associated scalar product. For $x \in \mathbb{R}^m$ and $\gamma > 0$ we define the closed ball

$$B_\gamma(x) := \{y \in \mathbb{R}^m : |x - y|_{\mathbb{R}^m} \leq \gamma\}$$

and we set

$$B_\gamma := B_\gamma(0).$$

6.2.1. Notation from convex analysis

For a Banach space X and an arbitrary subset $A \subset X$, we denote by $\mathrm{cl}(A)$, $\mathrm{conv}(A)$ and $\mathrm{cone}(A)$ the closure, convex hull (smallest convex superset of A) and the conical hull (smallest convex cone containing A) of A, respectively. For a closed, convex set $A \subset X$, we define the radial cone and the tangent cone of A at $x \in A$ by

$$\mathcal{R}_A(x) := \mathrm{cone}(A - x) = \bigcup_{\lambda > 0} \lambda\,(A - x) \qquad \text{and} \qquad \mathcal{T}_A(x) := \mathrm{cl}(\mathcal{R}_A(x)),$$

respectively. The polar cone of A is given by

$$A^\circ := \{x^\star \in X^\star : \forall x^\star \in X^\star : \langle x,\, x^\star\rangle \leq 0\},$$

where $X^\star$ is the topological dual space of X and $\langle\cdot\,,\,\cdot\rangle : X \times X^\star \to \mathbb{R}$ denotes the dual pairing. Finally, for $x^\star \in X^\star$, we define the annihilator

$$(x^\star)^\perp := \{x \in X : \langle x,\, x^\star\rangle = 0\}.$$

6.2.2. Notation and preliminaries from capacity theory

We recall some basic results in capacity theory. These are crucial to characterize the tangent cone and the normal cone of $\mathcal{C}$.

Throughout the paper, $\Omega \subset \mathbb{R}^d$ denotes a bounded, open set. By $H_0^1(\Omega)$ we denote the usual Sobolev space. As a norm in $H_0^1(\Omega)$ we use

$$\|u\|^2_{H_0^1(\Omega)} := \|\nabla u\|^2_{L^2(\Omega;\mathbb{R}^d)} := \int_\Omega |\nabla u(\omega)|^2_{\mathbb{R}^d}\,\mathrm{d}\omega.$$

The capacity of a set $O \subset \Omega$ is defined as

$$\mathrm{cap}(O) := \inf\{\|\nabla v\|^2_{L^2(\Omega;\mathbb{R}^d)} : v \in H_0^1(\Omega) \text{ and } v \geq 1 \text{ a.e. in a neighbourhood of } O\},$$

see, e.g., Attouch, Buttazzo, Michaille, 2006, Section 5.8.2, Bonnans, Shapiro, 2000, Definition 6.47, and Delfour, Zolésio, 2001, Section 8.6.1.

A function $v : \Omega \to \mathbb{R}^m$ is called *quasi-continuous* if for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\mathrm{cap}(G_\varepsilon) < \varepsilon$ and v is continuous on $\Omega \setminus G_\varepsilon$. A set $O \subset \Omega$ is called *quasi-open* if for all $\varepsilon > 0$, there exists an open set $G_\varepsilon \subset \Omega$, such that $\mathrm{cap}(G_\varepsilon) < \varepsilon$ and $O \cup G_\varepsilon$ is open. For every quasi-continuous function $v : \Omega \to \mathbb{R}^m$ and every open set $U \subset \mathbb{R}^m$, the set $\{\omega \in \Omega : v(\omega) \in U\}$ is quasi-open.

We say that a property P (depending on $\omega \in \Omega$) holds quasi-everywhere (q.e.), if it is only violated on a set of capacity zero, e.g., $\mathrm{cap}(\{\omega \in \Omega : P(\omega) \text{ does not hold}\}) = 0$. We say that P holds q.e. on a subset $K \subset \Omega$, if and only if $\mathrm{cap}(\{\omega \in K : P(\omega) \text{ does not hold}\}) = 0$. Similarly, we use the statement "$P(\omega)$ is satisfied for quasi-all (q.a.) $\omega \in K$."

It is known, see, e.g., Bonnans, Shapiro, 2000, Lemma 6.50, Delfour, Zolésio, 2001, Theorem 8.6.1, that every $v \in H_0^1(\Omega)^m$ possesses a quasi-continuous representative and this representative is uniquely determined up to sets of zero capacity. When we speak about a function $v \in H_0^1(\Omega)^m$, we always refer to the quasi-continuous representative. Every sequence which converges in $H_0^1(\Omega)$ possesses a pointwise quasi-everywhere convergent subsequence, see Bonnans, Shapiro, 2000, Lemma 6.52.

It is easily seen that a set with zero capacity also has measure zero. Sets of measure zero may have positive capacity. However, for quasi-open sets we have the following well-known lemma. For the convenience of the reader, we provide a simple proof.

Lemma 6.2.1. Assume that $A \subset \Omega$ is quasi-open and has measure zero. Then, A has zero capacity.

Proof. Assume that the set A is quasi-open and has measure zero. Then, for the function $f = 0$ we have $f \geq 1$ a.e. on A, since A has measure zero. The quasi-openness of A implies $f \geq 1$ q.e. on A, see Lemma 3.2.3. Hence, f is an admissible function for the computation of the capacity of A, cf. Heinonen, Kilpeläinen, Martio, 1993, Lemma 4.7, and this implies $\mathrm{cap}(A) = 0$.

By using this lemma, we can give a characterization of the set $\mathcal{C}$ in terms of quasi-everywhere.

Lemma 6.2.2. For $\mathcal{C}$ defined in (6.1.1) we have

$$\mathcal{C} = \{v \in H_0^1(\Omega)^m : v(\omega) \in C \text{ for q.a. } \omega \in \Omega\}.$$

Proof. The inclusion "$\supset$" is clear since sets of capacity zero have measure zero.

To prove the converse inclusion, let $v \in \mathcal{C}$ be given. Then, $v(\omega) \in C$ for almost all $\omega \in \Omega$. We set $N = \{\omega \in \Omega : v(\omega) \notin C\}$. Then, N has measure zero and is quasi-open. Now, the assertion follows from Lemma 6.2.1.

6.3. Characterization of the tangent cone

In this section, we are going to characterize the tangent cone of $\mathcal{C}$. We will see that the tangent cone of $\mathcal{C}$ consists of functions whose point values lie in the tangent cone of C. As in the scalar case $m = 1$, we have to use the notion of quasi-everywhere. In particular, we will show that

$$\mathcal{T}_{\mathcal{C}}(u) = \{v \in H_0^1(\Omega)^m : v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ for q.a. } \omega \in \Omega\}. \tag{6.3.1}$$

Throughout this section, we fix $u \in \mathcal{C}$. One inclusion in (6.3.1) is easy to prove.

Lemma 6.3.1. We have

$$\mathcal{T}_{\mathcal{C}}(u) \subset A_1 := \{v \in H_0^1(\Omega)^m : v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ for q.a. } \omega \in \Omega\}.$$

Proof. Let $v \in \mathcal{T}_{\mathcal{C}}(u)$ be given. By definition of the tangent cone, there exist sequences $\{t_n\} \subset (0, \infty)$, $\{u_n\} \subset \mathcal{C}$ such that $t_n \searrow 0$ and $(u_n - u)/t_n \to v$ in $H_0^1(\Omega)^m$ as $n \to \infty$. Since $u_n(\omega), u(\omega) \in C$ for q.a. $\omega \in \Omega$, we have $[u_n(\omega) - u(\omega)]/t_n \in \mathcal{T}_C(u(\omega))$ for q.a. $\omega \in \Omega$. Since $(u_n - u)/t_n \to v$ in $H_0^1(\Omega)^m$, we have (up to a subsequence) $[u_n(\omega) - u(\omega)]/t_n \to v(\omega)$ as $n \to \infty$ for q.a. $\omega \in \Omega$. Since $\mathcal{T}_C(u(\omega))$ is closed, this shows $v(\omega) \in \mathcal{T}_C(u(\omega))$ for q.a. $\omega \in \Omega$.

In what follows, we show the converse embedding $A_1 \subset \mathcal{T}_{\mathcal{C}}(u)$, which is harder to obtain. To this end, we will define sets $A_2, A_3, A_4 \subset H_0^1(\Omega)^m$ with the properties

$$A_1 \subset \mathrm{cl}(A_2), \quad A_2 \subset \mathrm{cl}(A_3), \quad A_3 \subset \mathrm{cl}(A_4), \quad A_4 \subset \mathcal{R}_{\mathcal{C}}(u).$$

Taking the closure, we obtain

$$\mathcal{T}_{\mathcal{C}}(u) \subset A_1 \subset \mathrm{cl}(A_1) \subset \mathrm{cl}(A_2) \subset \mathrm{cl}(A_3) \subset \mathrm{cl}(A_4) \subset \mathrm{cl}(\mathcal{R}_{\mathcal{C}}(u)) = \mathcal{T}_{\mathcal{C}}(u),$$

and the characterization of the tangent cone follows, see Theorem 6.3.12. Note that the same result is achieved in Section 6.4 by owing to the bipolar theorem. However, we present the direct approach via approximation, since it can be adopted to infer the polyhedricity of $\mathcal{C}$ under some additional assumptions on C, see Section 6.5.

First, we use a simple truncation argument in order to work in the space $H_0^1(\Omega)^m \cap L^\infty(\Omega)^m$.

Lemma 6.3.2. Each function $v \in A_1$ can be approximated by functions from

$$A_2 := \{v \in H_0^1(\Omega)^m : v \in L^\infty(\Omega)^m,\ v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ for q.a. } \omega \in \Omega\}.$$

That is, $A_1 \subset \mathrm{cl}(A_2)$.

Proof. The truncation $T_M v : \Omega \to \mathbb{R}^m$ of $v \in A_1$ which is defined via

$$(T_M v)(\omega) := \min\Big(1, \frac{M}{|v(\omega)|_{\mathbb{R}^m}}\Big)\, v(\omega)$$

belongs to $(L^\infty(\Omega) \cap H_0^1(\Omega))^m$, and converges towards v in $H_0^1(\Omega)^m$, see Theorem 6.A.2. Since $\mathcal{T}_C(u(\omega))$ is a cone for all $\omega \in \Omega$, $T_M v \in A_2$ follows.

To proceed, we need an assumption on the set C. Roughly speaking, we assume that for each point $x \in C$ there is a ball of radius c_2 and the distance between x and the center of the ball is at most c_1.

Assumption 6.3.3. There exist $c_1, c_2 > 0$, such that for every $x \in C$, there exists a point $w(x) \in C$ such that $|x - w(x)|_{\mathbb{R}^m} \leq c_1$ and $B_{c_2}(w(x)) \subset C$. Moreover, the mapping w is Lipschitz continuous with modulus L and $w(0) = 0$.

Note that Assumption 6.3.3 implies $0 \in \operatorname{int}(C)$. Moreover, in the case that C is bounded and $0 \in \operatorname{int}(C)$, we can choose $w \equiv 0$. However, the unbounded, closed, convex set

$$\{(x, y, z) : x^2 - y \leq 1,\ x^2/(y+1) + |z| \leq 1\}$$

has $(0, 0, 0)$ in its interior, but it does not satisfy Assumption 6.3.3. Hence, Assumption 6.3.3 is stronger than $0 \in \operatorname{int}(C)$.

Lemma 6.3.4. Let Assumption 6.3.3 be satisfied. For all $x \in C$ and $n \in \mathcal{T}_C(x)^\circ$, we have

$$(n,\, x - w(x))_{\mathbb{R}^m} \geq c_2\, |n|_{\mathbb{R}^m}.$$

Proof. Since $B_{c_2}(w(x)) \subset C$, we have

$$\big(n,\, x - (w(x) + p)\big)_{\mathbb{R}^m} \geq 0$$

for all $|p|_{\mathbb{R}^m} \leq c_2$. Hence,

$$(n,\, x - w(x))_{\mathbb{R}^m} \geq \sup_{|p|_{\mathbb{R}^m} \leq c_2} (n,\, p)_{\mathbb{R}^m} = c_2\, |n|_{\mathbb{R}^m}.$$

Let us consider a function $v \in A_2$. Then, the angle between $v(\omega)$ and $n \in \mathcal{T}_C(u(\omega))$ can be a right angle. The next lemma shows, that we can approximate v with functions $\tilde{v}$, for which the angle between $\tilde{v}(\omega)$ and $n \in \mathcal{T}_C(u(\omega))$ is obtuse, uniformly in $\omega \in \Omega$.

Lemma 6.3.5. Let Assumption 6.3.3 be satisfied. Each function $v \in A_2$ can be approximated by functions from

$$A_3 := \left\{ v \in H_0^1(\Omega)^m \cap L^\infty(\Omega)^m : \exists \varepsilon > 0 : (v(\omega),\, n)_{\mathbb{R}^m} \le -\varepsilon\, |n|_{\mathbb{R}^m} \text{ for all } n \in \mathcal{T}_C(u(\omega))^\circ \text{ for q.a. } \omega \in \Omega \right\}.$$

That is, $A_2 \subset \mathrm{cl}(A_3)$.

Proof. Let $v \in A_2$ be given. By definition, we have $(v(\omega),\, n)_{\mathbb{R}^m} \le 0$ for all $n \in \mathcal{T}_C(u(\omega))^\circ$ for q.a. $\omega \in \Omega$, since $v(\omega) \in \mathcal{T}_C(u(\omega))$. Let $w : C \to C$ be the Lipschitz continuous mapping from Assumption 6.3.3. Then, $w(u) \in H_0^1(\Omega)^m$, see Lemma 6.A.1, and $w(u) \in L^\infty(\Omega)^m$. For $\varepsilon > 0$, we set

$$v_\varepsilon := v - \varepsilon\,(u - w(u)).$$

Now, let $\omega \in \Omega$ and $n \in \mathcal{T}_C(u(\omega))^\circ$ be given. By owing to Lemma 6.3.4, we have

$$(v_\varepsilon(\omega),\, n)_{\mathbb{R}^m} = (v(\omega) - \varepsilon\,(u - w(u)),\, n)_{\mathbb{R}^m} \le -\varepsilon\, c_2\, |n|_{\mathbb{R}^n}.$$

The assertion follows since $v_\varepsilon \to v$ in $H_0^1(\Omega)^m$ as $\varepsilon \to 0$ and $v_\varepsilon \in L^\infty(\Omega)^m$.

The following lemma is a replacement for a partition of unity in the setting of capacities. Under certain conditions, it allows to approximate a function f with $f \le 0$ on some set M_0 with functions f_δ satisfying $f_\delta \le 0$ on some prescribed, larger sets M_δ.

Lemma 6.3.6. Let $f \in H_0^1(\Omega)$ be given. We assume that the quasi-closed sets M_δ, $\delta \ge 0$, are non-decreasing in δ, in the sense that $M_\delta \subset M_{\delta'}$ for $0 < \delta < \delta'$. Moreover, we suppose that $f \le 0$ q.e. on M_0 and $M_0 = \bigcap_{\delta>0} M_\delta$.

(a) Then, for all $\delta > 0$ there exists $f_\delta \in H_0^1(\Omega)$, such that $f_\delta \le 0$ q.e. on M_δ, $f_\delta = 0$ q.e. on $M_\delta \cap \{\omega \in \Omega : f(\omega) = 0\}$ and $f_\delta \to f$ in $H_0^1(\Omega)$ as $\delta \to 0$.

(b) In case $\|f\|_{L^\infty(\Omega)} < \infty$, the same can be achieved with $\|f_\delta\|_{L^\infty(\Omega)} \le \|f\|_{L^\infty(\Omega)}$.

(c) If, additionally, $f \le -\varepsilon$ on M_0 for some $\varepsilon > 0$ and if $\mathrm{cap}(M_\delta) < \infty$ for some $\delta > 0$, then, we find $f_\delta \in H_0^1(\Omega)$ with $f_\delta \le -\varepsilon$ q.e. on M_δ for δ small enough, such that $f_\delta \to f$ in $H_0^1(\Omega)$. In case $f \in L^\infty(\Omega)$, we can achieve $\|f_\delta\|_{L^\infty(\Omega)} \le \|f\|_{L^\infty(\Omega)} + 2\,\varepsilon$.

Proof.

(a) The sets $\Omega \setminus M_\delta$ are quasi-open and $\{\Omega \setminus M_\delta\}_{\delta>0}$ is a covering of $\Omega \setminus M_0$. Moreover, $\max(f, 0) = 0$ q.e. on M_0 implies $\max(f, 0) \in H_0^1(\Omega \setminus M_0)$, see Kilpeläinen, Malý, 1992, Theorem 2.10. Now, we can invoke Kilpeläinen, Malý, 1992, Lemma 2.4 and Theorem 2.10, and find $g_\delta \in H_0^1(\Omega \setminus M_\delta)$, $g_\delta \ge 0$ and $g_\delta \to \max(f, 0)$ in $H_0^1(\Omega)$.

We define $f_\delta := f - (\max(f, 0) - g_\delta)$ and obtain $f_\delta \to f$ in $H_0^1(\Omega)$, $f_\delta = f + \min(-f, 0) + g_\delta = \min(0, f) + g_\delta \le 0$ on M_δ. Since $g_\delta = 0$ q.e. on M_δ, this shows $f_\delta = 0$ q.e. on $M_\delta \cap \{\omega \in \Omega : f(\omega) = 0\}$.

(b) We use the same technique as in (a), and obtain additionally $\|g_\delta\|_{L^\infty(\Omega)} \leq \|\max(f,0)\|_{L^\infty(\Omega)}$, see Kilpeläinen, Malý, 1992, Lemma 2.4. Hence, $\|f_\delta\|_{L^\infty(\Omega)} = \|\min(0,f) + g_\delta\|_{L^\infty(\Omega)} \leq \|f\|_{L^\infty(\Omega)}$.

(c) Let $\hat{\delta} > 0$ be given such that $\mathrm{cap}(M_{\hat{\delta}}) < \infty$. By definition of the capacity, there exists a $h \in H^1_0(\Omega)$, $0 \leq h \leq 1$ q.e. on Ω and $h = 1$ q.e. on $M_{\hat{\delta}}$.

We have $f + \varepsilon\, h \leq 0$ on M_0. By owing to (a), we find $g_\delta \in H^1_0(\Omega)$ with $g_\delta \leq 0$ q.e. on M_δ and $g_\delta \to f + \varepsilon\, h$ in $H^1_0(\Omega)$. We set $f_\delta := g_\delta - \varepsilon\, h$ and obtain the first part of the assertion, since $h = 1$ on M_δ for $\delta \leq \hat{\delta}$. By owing to (b), we have $\|g_\delta\|_{L^\infty(\Omega)} \leq \|f\|_{L^\infty(\Omega)} + \varepsilon\, \|h\|_{L^\infty(\Omega)}$ and, hence, $\|f_\delta\|_{L^\infty(\Omega)} \leq \|f\|_{L^\infty(\Omega)} + 2\,\varepsilon\, \|h\|_{L^\infty(\Omega)}$.

Let us recall the well-known result that the normal cone mapping of the convex set C is upper semicontinuous, since this will be important in the sequel.

Lemma 6.3.7. Let $\{x_i\}_{i\in\mathbb{N}} \subset C$ and $\{n_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^m$ be sequences, such that $x_i \to x$, $n_i \to n$ and $n_i \in \mathcal{T}_C(x_i)^\circ$ for all $i \in \mathbb{N}$. Then, $n \in \mathcal{T}_C(x)^\circ$.

The next lemma is a preparation for the proof of Lemma 6.3.9. We recall that $B_\gamma = \{x \in \mathbb{R}^m : |x|_{\mathbb{R}^m} \leq \gamma\}$ for $\gamma \geq 0$.

Lemma 6.3.8. Suppose that Assumption 6.3.3 is satisfied. Let $\gamma \geq 0$ and $n \in \mathbb{R}^m$ be given. For $\lambda \geq 0$ we define the set

$$M_\lambda := \{\omega \in \Omega : \exists \tilde{u} \in B_\lambda(u(\omega)) : n \in B_\gamma + \mathcal{T}_C(\tilde{u})^\circ\}.$$

Then, the sets M_λ possess the following properties.

(a) The set M_λ is quasi-closed for all $\lambda \geq 0$.

(b) The sets M_λ are non-decreasing in λ and $M_0 = \bigcap_{\lambda>0} M_\lambda$.

(c) In case $\gamma < |n|_{\mathbb{R}^m}$ and $\lambda \leq c_2/2$ we have $\mathrm{cap}(M_\lambda) \leq \infty$.

(d) The inequality $(n,\, u(\omega) - w(u(\omega)))_{\mathbb{R}^m} \geq c_2\, |n|_{\mathbb{R}^m}/3$ is satisfied for all $\omega \in M_\lambda$, $\lambda \leq c_2/(2\,L)$ and $\gamma \leq c_2\, |n|_{\mathbb{R}^m}/(6\,(c_1 + 2\,c_2))$.

Proof.

(a) We have the chain of equivalences

$$\begin{aligned} \omega \in M_\lambda \quad &\Longleftrightarrow \quad n \in B_\gamma + \mathcal{T}_C(u(\omega) + B_\lambda)^\circ \\ &\Longleftrightarrow \quad u(\omega) \in \left[\mathcal{T}_C(\cdot)^\circ\right]^{-1}(n + B_\gamma) + B_\lambda. \end{aligned}$$

Here, $\mathcal{T}_C(u(\omega) + B_\lambda)^\circ$ denotes the image of the set $u(\omega) + B_\lambda$ under the set-valued mapping $\mathcal{T}_C(\cdot)^\circ$. Since u is quasi-continuous, it remains to show that the set $\left[\mathcal{T}_C(\cdot)^\circ\right]^{-1}(n + B_\gamma) + B_\lambda$ is closed. Since B_λ is compact, we have to show that

$[\mathcal{T}_C(\cdot)^\circ]^{-1}(n+B_\gamma)$ is closed. To this end, let $\{\tilde{u}_i\} \subset C$ be a convergent sequence, such that $n + p_i \in \mathcal{T}_C(\tilde{u}_i)^\circ$ for some $p_i \in B_\gamma$. Since B_γ is compact, we have (up to a subsequence) $n + p_i \to n + p$ for some $p \in B_\gamma$. Now, Lemma 6.3.7 implies $n + p \in \mathcal{T}_C(\lim_{i\to\infty} \tilde{u}_i)^\circ$, hence $\lim_{i\to\infty} \tilde{u}_i \in [\mathcal{T}_C(\cdot)^\circ]^{-1}(n + B_\gamma)$. This shows the desired closedness.

(b) The first assertion is clear. Let $\omega \in \bigcap_{\lambda>0} M_\lambda$ be given. For $\lambda > 0$, there exist $\tilde{u}_\lambda \in B_\lambda(u(\omega))$ and $p_\lambda \in B_\gamma$, such that $n \in p_\lambda + \mathcal{T}_C(u_\lambda)^\circ$. By compactness of B_γ, there exists a sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ with $\lambda_i \to 0$ as $i \to \infty$ and $p_{\lambda_i} \to p$ for some $p \in B_\gamma$. Hence, $u_{\lambda_i} \to u(\omega)$. Using $n - p_{\lambda_i} \in \mathcal{T}_C(u_{\lambda_i})^\circ$ and Lemma 6.3.7, we infer $n - p \in \mathcal{T}_C(u(\omega))^\circ$ and this yields $\omega \in M_0$.

(c) For $\gamma < |n|_{\mathbb{R}^m}$ no element in $n + B_\gamma$ is equal to zero. Moreover, Assumption 6.3.3 implies $B_{c_2}(0) \subset C$. Hence, $n \in B_\gamma + \mathcal{T}_C(\tilde{u})^\circ$ implies $|\tilde{u}|_{\mathbb{R}^m} \geq c_2$. Together with $\tilde{u} \in B_\lambda(u(\omega))$, we find $|u(\omega)|_{\mathbb{R}^m} \geq c_2 - \lambda$ for q.a. $\omega \in M_\lambda$. Hence, $|u(\omega)|_{\mathbb{R}^m} \geq c_2/2$ for $\lambda \leq c_2/2$ for q.a. $\omega \in M_\lambda$. Thus, $\mathrm{cap}(M_\lambda) < +\infty$.

(d) Let $\omega \in M_\lambda$ be given. Hence, there exist $\tilde{u} \in B_\lambda(u(\omega))$ such that $n + p \in \mathcal{T}_C(\tilde{u})^\circ$ for some $|p|_{\mathbb{R}^m} \leq \gamma$. This yields

$$\begin{aligned}
(n,\, u(\omega) - w(u(\omega)))_{\mathbb{R}^m} &\geq (n+p,\, u(\omega) - w(u(\omega)))_{\mathbb{R}^m} - \gamma\, c_1 \\
&\geq (n+p,\, \tilde{u} - w(\tilde{u}))_{\mathbb{R}^m} - (|n|_{\mathbb{R}^m} + \gamma)\,(L+1)\,\lambda - \gamma\, c_1 \\
&\geq c_2\, |n+p|_{\mathbb{R}^m} - (|n|_{\mathbb{R}^m} + \gamma)\,(L+1)\,\lambda - \gamma\, c_1 \\
&\geq (c_2 - (L+1)\,\lambda)\, |n|_{\mathbb{R}^m} - \gamma\,(c_1 + c_2 + (L+1)\,\lambda),
\end{aligned}$$

where we used Lemma 6.3.4. This yields the claim.

In the next lemma, we use the approximation result Lemma 6.3.6, in order to approximate $v \in A_3$ with a function $\tilde{v}$, such that $\tilde{v}(\omega)$ does belong to the tangent cone of C not only at $u(\omega)$, but also at all points in the neighborhood of $u(\omega)$.

Lemma 6.3.9. Let Assumption 6.3.3 be satisfied. Each function $v \in A_3$ can be approximated by functions from

$$A_4 := \left\{ \begin{array}{l} v \in H_0^1(\Omega)^m \cap L^\infty(\Omega)^m : \exists \lambda > 0 : \\ \qquad (v(\omega),\, n)_{\mathbb{R}^m} \leq 0 \text{ for all } n \in \mathcal{T}_C(\tilde{u})^\circ, \tilde{u} \in B_\lambda(u(\omega)) \cap C \text{ for q.a. } \omega \in \Omega \end{array} \right\}.$$

That is, $A_3 \subset \mathrm{cl}(A_4)$.

Proof. Let $v \in A_3$ be given. W.l.o.g. assume $\|v\|_{L^\infty(\Omega)} \leq 1$. There exists $\varepsilon > 0$, such that $(v(\omega),\, n)_{\mathbb{R}^m} \leq -\varepsilon$ for all $n \in \mathcal{T}_C(u(\omega))^\circ$, $|n|_{\mathbb{R}^m} = 1$ for q.a. $\omega \in \Omega$. Let $\gamma > 0$ be chosen such that $\gamma \leq \frac{\varepsilon}{2+6\,c_1\,c_2^{-1}\,(1+\varepsilon)} \leq \varepsilon/2$ and $\gamma \leq \frac{c_2}{6\,c_1+12\,c_2}$.

By compactness, we find a γ-net $\{n_i\}_{i=1}^N$ of the unit sphere in $\mathbb{R}^m$. That is, $|n_i|_{\mathbb{R}^m} = 1$ and for every $n \in \mathbb{R}^m$ with $|n|_{\mathbb{R}^m} = 1$, there is $i \in \{1, \ldots, N\}$ with $|n - n_i|_{\mathbb{R}^m} \leq \gamma$.

We define the sets

$$M_\lambda^i := \{\omega \in \Omega : \exists \tilde{u} \in B_\lambda(u(\omega)) : n_i \in B_\gamma + \mathcal{T}_C(\tilde{u})^\circ\}.$$

Let $\omega \in M_0^i$ be given. Then, by definition of M_0^i, there exists $n \in \mathcal{T}_C(u(\omega))^\circ$, such that $|n - n_i|_{\mathbb{R}^m} \leq \gamma$. Hence,

$$(n_i,\, v(\omega))_{\mathbb{R}^m} \leq (n,\, v(\omega))_{\mathbb{R}^m} + \gamma \leq -\varepsilon + \gamma \leq -\frac{\varepsilon}{2} < 0$$

for q.a. $\omega \in M_0^i$. We set $f_i(\omega) := (n_i,\, v(\omega))_{\mathbb{R}^m}$.

Now, we choose $k \in \mathbb{N}$.

By Lemma 6.3.6, Lemma 6.3.8, there exist $\lambda_i^k > 0$ and $\tilde{f}_i^k \in H_0^1(\Omega)$, such that $\|f_i - \tilde{f}_i^k\|_{H_0^1(\Omega)} \leq 1/k$, $\|\tilde{f}_i^k\|_{L^\infty(\Omega)} \leq 1 + 2\,(\varepsilon - \gamma)$ and $\tilde{f}_i^k \leq -\varepsilon + \gamma$ on $M_{\lambda_i^k}^i$.

We set $\lambda^k := \min\{\lambda_1^k, \ldots, \lambda_N^k, c_2/(2\,L)\}$ and

$$\tilde{v}^k(\omega) := v(\omega) - \frac{3}{c_2}\,(u(\omega) - w(u(\omega)))\, \max_{i=1,\ldots,N}[\max(f_i(\omega) - \tilde{f}_i^k(\omega), 0)].$$

Then, for $\omega \in M_{\lambda^k}^i$ we have

$$(n_i,\, \tilde{v}^k(\omega))_{\mathbb{R}^m} = f_i(\omega) - \frac{3}{c_2}\,(n_i,\, u(\omega) - w(u(\omega)))_{\mathbb{R}^m}\, \max_{i=1,\ldots,N}[\max(f_i(\omega) - \tilde{f}_i^k(\omega), 0)].$$

Together with Lemma 6.3.8, we have

$$\begin{aligned}(n_i,\, \tilde{v}^k(\omega))_{\mathbb{R}^m} &\leq f_i(\omega) - \max_{i=1,\ldots,N}[\max(f_i(\omega) - \tilde{f}_i^k(\omega), 0)] \leq f_i(\omega) - (f_i(\omega) - \tilde{f}_i^k(\omega)) \\ &\leq -\varepsilon + \gamma\end{aligned}$$

for q.a. $\omega \in M_{\lambda^k}^i$.

Let us show that $\tilde{v}^k$ belongs to A_4. To this end, let $\omega \in \Omega$, $\tilde{u} \in B_\lambda(u(\omega)) \cap C$ and $n \in \mathcal{T}_C(\tilde{u})^\circ$ be given. By definition of $\{n_i\}$, there exists $i \in \{1, \ldots, N\}$ with $|n - n_i|_{\mathbb{R}^m} \leq \gamma$. This shows $\omega \in M_{\lambda^k}^i$. Hence,

$$\begin{aligned}(n,\, \tilde{v}^k(\omega))_{\mathbb{R}^m} \leq (n_i,\, \tilde{v}^k(\omega))_{\mathbb{R}^m} + \gamma\,\|\tilde{v}^k\|_{L^\infty(\Omega)} &\leq -\varepsilon + \gamma + \gamma\,(1 + 3\,\frac{c_1}{c_2}\,2\,(1 + \varepsilon - \gamma)) \\ &\leq -\varepsilon + \gamma\,(2 + 3\,\frac{c_1}{c_2}\,2\,(1 + \varepsilon)) \leq 0.\end{aligned}$$

This shows $\tilde{v}_k \in A_4$.

Let us estimate $\|v - \tilde{v}^k\|_{H_0^1(\Omega)}$. We have for some generic constant $c > 0$, which may change from line to line,

$$\begin{aligned}\|v - \tilde{v}^k\|_{H_0^1(\Omega)} &\le c \Big(\|u - w(u)\|_{L^\infty(\Omega)} \sum_{i=1}^N \|f_i - \tilde{f}_i^k\|_{H_0^1(\Omega)} \\ &\qquad + \|u - w(u)\|_{H_0^1(\Omega)} \sum_{i=1}^N \|f_i - \tilde{f}_i^k\|_{L^\infty(\Omega)} \Big) \\ &\le c \left(\frac{c_1}{k} + \|u - w(u)\|_{H_0^1(\Omega)} \, 2 \, (1 + \varepsilon - \gamma) \right) \\ &\le \text{const}\end{aligned}$$

and for some $p \in (1, 2)$ we choose $q \in (2, \infty)$ such that $1/p = 1/2 + 1/q$ and $\theta = 2/q \in (0, 1)$ and obtain

$$\begin{aligned}\|v - \tilde{v}^k\|_{W_0^{1,p}(\Omega)} &\le c \Big(\|u - w(u)\|_{L^\infty(\Omega)} \sum_{i=1}^N \|f_i - \tilde{f}_i^k\|_{H_0^1(\Omega)} \\ &\qquad + \|u - w(u)\|_{H_0^1(\Omega)} \sum_{i=1}^N \|f_i - \tilde{f}_i^k\|_{L^q(\Omega)} \Big) \\ &\le c \Big(\frac{c_1}{k} + \|u - w(u)\|_{H_0^1(\Omega)} \sum_{i=1}^N \|f_i - \tilde{f}_i^k\|_{L^\infty(\Omega)}^{1-\theta} \, \|f_i - \tilde{f}_i^k\|_{L^2(\Omega)}^{\theta} \Big) \\ &\le c \, \frac{1}{k^\theta} \to 0.\end{aligned}$$

This shows that the sequence $\tilde{v}^k$ converges weakly towards v in $H_0^1(\Omega)$. By convexity of A_4 and owing to Mazur's lemma, there is a sequence in A_4 converging strongly in $H_0^1(\Omega)$ towards v.

The next lemma will be used in the proof of Lemma 6.3.11, but it is also interesting for itself. It shows that a direction h which belongs to the tangent cone at all points in a neighborhood of $x \in C$ actually belongs to $\mathcal{R}_C(x)$.

Lemma 6.3.10. Let $x \in C$ and $h \in \mathbb{R}^m$, $|h|_{\mathbb{R}^m} \le 1$ be given, such that $h \in \bigcap_{\tilde{x} \in B_\varepsilon(x) \cap C} \mathcal{T}_C(\tilde{x})$ for some $\varepsilon > 0$. Then, $x + \varepsilon\, h \in C$.

Proof. We set $y = \mathrm{Proj}_C(x + \varepsilon\, h)$. Then, $|y - x|_{\mathbb{R}^m} \le \varepsilon$. Hence, $h \in \mathcal{T}_C(y)$. By the properties of the projection, we have $(x + \varepsilon\, h) - y \in \mathcal{T}_C(y)^\circ$. Hence,

$$0 \ge (h,\, (x + \varepsilon\, h) - y)_{\mathbb{R}^m} = (h,\, x - y)_{\mathbb{R}^m} + \varepsilon \ge -\varepsilon + \varepsilon = 0.$$

Hence, we have $(h,\, x - y)_{\mathbb{R}^m} = -\varepsilon \ge -|h|_{\mathbb{R}^m} \, |x - y|_{\mathbb{R}^m}$. This yields $h = -(x - y)/\varepsilon$. Hence, $x + \varepsilon\, h = y \in C$.

Using this lemma, we can show that directions in A_4 belong to the radial cone $\mathcal{R}_{\mathcal{C}}(u)$.

Lemma 6.3.11. Let $v \in A_4$ be given. Then, there exists $\delta > 0$ such that $u + \delta\, v \in \mathcal{C}$. This shows $A_4 \subset \mathcal{R}_{\mathcal{C}}(u)$.

Proof. Let $v \in A_4$ be given. There is $\lambda > 0$, such that $(v(\omega),\, n)_{\mathbb{R}^m} \leq 0$ for all $n \in \mathcal{T}_C(\tilde{u})$, $\tilde{u} \in B_\lambda(u(\omega)) \cap C$ for q.a. $\omega \in \Omega$. This gives $v(\omega) \in \mathcal{T}_C(\tilde{u})$ for all $\tilde{u} \in B_\lambda(u(\omega)) \cap C$ for q.a. $\omega \in \Omega$. By applying the previous lemma, we obtain $u(\omega) + \lambda\, v(\omega)/\|v\|_{L^\infty(\Omega)} \in C$ for q.a. $\omega \in \Omega$. Hence, $u + \lambda\, v/\|v\|_{L^\infty(\Omega)} \in \mathcal{C}$, see Lemma 6.2.2.

The following theorem collects the previous lemmas and is the main result of this section.

Theorem 6.3.12. Suppose Assumption 6.3.3 is satisfied. Then,

$$\mathcal{T}_{\mathcal{C}}(u) = \{v \in H_0^1(\Omega)^m : v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ q.e. in } \Omega\}.$$

Proof. From the previous lemmas, we have

$$\mathcal{T}_{\mathcal{C}}(u) \subset A_1 \subset \mathrm{cl}(A_2), \quad A_2 \subset \mathrm{cl}(A_3), \quad A_3 \subset \mathrm{cl}(A_4), \quad A_4 \subset \mathcal{R}_{\mathcal{C}}(u).$$

Taking the closures yields

$$\mathcal{T}_{\mathcal{C}}(u) \subset A_1 \subset \mathrm{cl}(A_1) \subset \mathrm{cl}(A_2) \subset \mathrm{cl}(A_3) \subset \mathrm{cl}(A_4) \subset \mathrm{cl}(\mathcal{R}_{\mathcal{C}}(u)) = \mathcal{T}_{\mathcal{C}}(u),$$

and the assertion follows.

6.4. Characterization of the normal cone

In order to characterize the normal cone of $\mathcal{C}$, we have to work with vector-valued measures, which act on $H_0^1(\Omega)^m \cap C_0(\Omega)^m$. For an introduction to vector-valued measures, we refer to Diestel, Uhl, 1977. Since the values of our measures will belong to $\mathbb{R}^m$, we can identify a vector-valued measure with a tuple of m scalar-valued measures.

Let $\mu = (\mu_1, \ldots, \mu_m)$ be a tuple of regular, signed Borel measures on Ω, i.e., $\mu_i : \mathcal{B} \to \mathbb{R}$, $i = 1, \ldots, m$, where $\mathcal{B}$ denotes the Borel σ-algebra of Ω. The variation $|\mu|$ is defined for $E \in \mathcal{B}$ by

$$|\mu|(E) := \sup_{\pi} \sum_{A \in \pi} |\mu(A)|_{\mathbb{R}^m},$$

where the supremum ranges over all finite partitions π of E into pairwise disjoint Borel sets. We have $|\mu|(\Omega) < \infty$, see Rudin, 1987, Section 6.6, and the variation $|\mu|$ is again a regular Borel measure, see Diestel, Uhl, 1977, Proposition I.1.9 and Rudin, 1987, Theorem 2.18.

We typically identify μ with its completion, i.e. with the complete measure on the smallest σ-algebra containing $\mathcal{B}$ and the $|\mu|$-null sets, cf. Rudin, 1987, Theorem 1.36.

Similar to the polar decomposition of a complex measure, cf. Rudin, 1987, Theorem 6.12, we can obtain a decomposition of μ over $|\mu|$. In particular, the Radon-Nikodým derivative

$$\mu' := \frac{\mathrm{d}\mu}{\mathrm{d}|\mu|}$$

satisfies $\mu' \in L^\infty(|\mu|)^m$ (in fact, we have $|\mu'(\omega)|_{\mathbb{R}^m} = 1$ everywhere after changing μ' on a $|\mu|$-null set) and

$$\mu(E) = \int_E \mu' \,\mathrm{d}|\mu|$$

holds for all $E \in \mathcal{B}$.

By the Riesz representation theorem, see Rudin, 1987, Theorem 6.19, the dual of $C_0(\Omega)^m$ can be identified with $\mathcal{M}(\Omega)^m$ which is the space of regular Borel measures $\mu : \mathcal{B} \to \mathbb{R}^m$ of bounded variation, i.e. $|\mu|(\Omega) < +\infty$, and the duality pairing is

$$\langle \mu,\, f\rangle_{\mathcal{M}(\Omega)^m, C_0(\Omega)^m} := \int_\Omega f \,\mathrm{d}\mu := \int_\Omega (f,\, \mu')_{\mathbb{R}^m} \,\mathrm{d}|\mu|.$$

Suppose we have a functional $\mu \in H^{-1}(\Omega)^m = (H^1_0(\Omega)^m)^\star$ such that μ is bounded w.r.t. the supremum norm on $H^1_0(\Omega)^m \cap C_0(\Omega)^m$. Since $H^1_0(\Omega) \cap C_0(\Omega)$ is a dense subspace of $C_0(\Omega)$, see Fukushima, Oshima, Takeda, 1994, p. 100, μ can be extended uniquely to $C_0(\Omega)^m$. By the Riesz representation theorem, this functional can be represented as a regular Borel measure μ with bounded variation. With a slight abuse of notation, we shall write $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$. In particular, we have

$$\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} = \langle \mu,\, f\rangle_{\mathcal{M}(\Omega)^m, C_0(\Omega)^m} = \int_\Omega (f,\, \mu')_{\mathbb{R}^m} \,\mathrm{d}|\mu| \tag{6.4.1}$$

for all $f \in H^1_0(\Omega)^m \cap C_0(\Omega)^m$.

It would be tempting to assume that $\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} = \int_\Omega (f,\, \mu')_{\mathbb{R}^m} \,\mathrm{d}|\mu|$ holds for all $f \in H^1_0(\Omega)^m$. This is, in general, not true, since $f \in L^1(|\mu|)^m$ may fail to hold, see Section 6.B. We will prove, however, that this formula holds if $\int_\Omega (f,\, \mu')_{\mathbb{R}^m} \,\mathrm{d}|\mu|$ is interpreted as a suitable limit, see (6.4.4) below.

The following lemma is well-known for scalar, non-negative measures μ, see, e.g., Bonnans, Shapiro, 2000, Lemma 6.56. We state an extension to signed, vector-valued measures which is due to Grun-Rehomme, 1977, Proposition 1. For convenience of the reader, we give its proof.

Lemma 6.4.1. Let $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$ be given. Then, $|\mu|$ does not charge Borel sets of zero capacity.

Moreover, all sets of zero capacity belong to the completion of Borel σ-algebra $\mathcal{B}$ (over Ω) w.r.t. $|\mu|$ and are not charged by $|\mu|$. Hence, a property which holds q.e. (i.e., up to a set of zero capacity) holds $|\mu|$-a.e.

Proof. We follow the ideas of the proof of Grun-Rehomme, 1977, Proposition 1.

We first consider the (scalar) case $m = 1$. By the Hahn decomposition theorem, see, e.g. Rudin, 1987, Theorem 6.14, we can decompose Ω into disjoint Borel sets Ω^+ and Ω^-, such that μ is non-negative on subsets of Ω^+ and non-positive on subsets of Ω^-.

Let $A \subset \Omega$ be a Borel set with zero capacity. Since $|\mu|(A) = |\mu|(A\cap\Omega^+)+|\mu|(A\cap\Omega^-)$ and $\operatorname{cap}(A \cap \Omega^+) = \operatorname{cap}(A \cap \Omega^-) = 0$, it is sufficient to consider the case $A \subset \Omega^+$ (otherwise, apply the following arguments to $A \cap \Omega^+$ and $A \cap \Omega^-$).

Let $\varepsilon > 0$ be arbitrary. Since $|\mu|$ is outer regular, we find an open set $O \subset \Omega$ with $A \subset O$ and $|\mu|(O) \leq |\mu|(A)+\varepsilon$. Since $\operatorname{cap}(A) = 0$, we find a sequence $\{u_n\} \subset L^\infty(O)\cap H^1_0(O) \subset L^\infty(\Omega) \cap H^1_0(\Omega)$ with $0 \leq u_n \leq 1$ in Ω, $\|u_n\|_{H^1_0(\Omega)} \to 0$, and $u_n = 1$ on a neighborhood of A, see Heinonen, Kilpeläinen, Martio, 1993, Lemma 2.9. Using Lemma 6.4.2 below, this yields

$$\begin{aligned}
\|u_n\|_{H^1_0(\Omega)} \|\mu\|_{H^{-1}(\Omega)} \geq \langle \mu, u_n\rangle_{H^{-1}(\Omega),H^1_0(\Omega)} &= \int_\Omega u_n \,\mathrm{d}\mu = \int_O u_n \,\mathrm{d}\mu \\
&= \int_A u_n \,\mathrm{d}\mu + \int_{O\setminus A} u_n \,\mathrm{d}\mu \\
&\geq \mu(A) - |\mu|(O \setminus A) = \mu(A) - |\mu|(O) + |\mu|(A) \\
&\geq \mu(A) - (|\mu|(A) + \varepsilon) + |\mu|(A) \\
&\geq \mu(A) - \varepsilon.
\end{aligned}$$

Since $\|u_n\|_{H^1_0(\Omega)} \to 0$, this yields $\varepsilon \geq \mu(A)$ for all $\varepsilon > 0$. Together with $A \subset \Omega^+$, we get $\mu(A) = 0$.

Now, let us consider the (vector-valued) case $m > 1$. Let $A \subset \Omega$ be a Borel set with zero capacity. By definition of $|\mu|$, we have

$$|\mu|(A) = \sup_\pi \sum_{B\in\pi} |\mu(B)|_{\mathbb{R}^m},$$

where the supremum ranges over all finite partitions π of A into pairwise disjoint Borel sets. By the definition of the Euclidean norm and the first part of the proof, we thus have

$$|\mu|(A) = \sup_\pi \sum_{B\in\pi} |\mu(B)|_{\mathbb{R}^m} = \sup_\pi \sum_{B\in\pi} \Big(\sum_{i=1}^m \mu_i(B)^2\Big)^{1/2} = 0,$$

since $\mu_i(B) = 0$ for all $B \subset A$ and $i \in \{1, \dots, m\}$.

Finally, let $A \subset \Omega$ have capacity zero, but be otherwise arbitrary. By the outer regularity of the capacity, see Attouch, Buttazzo, Michaille, 2006, Definition 5.8.1(b) and Proposition 5.8.3(a), we find

$$\operatorname{cap}(A) = \inf_{O \text{ open},\ A\subset O} \operatorname{cap}(O).$$

Hence, there exist open sets $O_n \subset \Omega$, $A \subset O_n$ with $\operatorname{cap}(O_n) \leq 1/n$, $n \in \mathbb{N}$. Hence, $A \subset B := \bigcap_{n\in\mathbb{N}} O_n$. This readily yields $B \in \mathcal{B}$ and $\operatorname{cap}(B) = 0$. Hence, $|\mu|(B) = 0$ by the first part of the proof. Thus, A belongs to the completion of $\mathcal{B}$ w.r.t. $|\mu|$ and $|\mu|(A) = 0$.

We recall that (the quasi-continuous representative of) $f \in H^1_0(\Omega)$ is uniquely determined quasi-everywhere, hence, $|\mu|$-almost-everywhere, for every $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$.

Now, we prove an integral formula for $\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m}$ with $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$ and $f \in H^1_0(\Omega)^m$. Therefore, we use the truncation $T_M : \mathbb{R}^m \to \mathbb{R}^m$, defined for $M > 0$ via

$$T_M(x) := \max\Big(1, \frac{M}{|x|_{\mathbb{R}^m}}\Big)\, x = \begin{cases} x & \text{if } |x|_{\mathbb{R}^m} \le M, \\ M\, \frac{x}{|x|_{\mathbb{R}^m}} & \text{if } |x|_{\mathbb{R}^m} > M. \end{cases} \tag{6.4.2}$$

Note that $T_M : \mathbb{R}^m \to \mathbb{R}^m$ is Lipschitz continuous with modulus 1. By Theorem 6.A.2, we have $T_M f \in H^1_0(\Omega)^m$ and $T_M f \to f$ in $H^1_0(\Omega)^m$ for all $f \in H^1_0(\Omega)$.

The next lemma is a direct consequence of Brézis, Browder, 1979, Theorem 1.

Lemma 6.4.2. Let $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$ be given. Then, any $f \in H^1_0(\Omega)^m \cap L^\infty(\Omega)^m$ belongs to $L^\infty(|\mu|)^m$ and

$$\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} = \int_\Omega (f,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu|. \tag{6.4.3}$$

Moreover, we have

$$\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} = \lim_{M\to\infty} \int_\Omega (T_M\, f,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu|. \tag{6.4.4}$$

for all $f \in H^1_0(\Omega)^m$.

Proof. Formula (6.4.3) follows directly from Brézis, Browder, 1979, Theorem 1 using $f_i = \mathrm{d}\mu_i/\mathrm{d}|\mu|$ therein. For convenience of the reader, we give the proof.

Let $f \in H^1_0(\Omega)^m \cap L^\infty(\Omega)^m$ be given. Using standard arguments, there is a sequence $\{f_n\}_{n\in\mathbb{N}} \subset C^\infty_0(\Omega)$ such that $f_n \to f$ in $H^1_0(\Omega)^m$ and $\|f_n\|_{L^\infty(\Omega)^m} \le \|f\|_{L^\infty(\Omega)^m}$. In particular, we have (up to a subsequence) $f_n \to f$ pointwise q.e. and, thus, pointwise $|\mu|$-a.e. on Ω. This shows that f is $|\mu|$-measurable and $f \in L^\infty(|\mu|)^m$. The estimate $|f_n(\omega)|_{\mathbb{R}^m} + |f(\omega)|_{\mathbb{R}^m} \le 2\,\|f\|_{L^\infty(\Omega)^m}$ holds q.e. and, thus, $|\mu|$-a.e. on Ω. Hence, $2\,\|f\|_{L^\infty(\Omega)^m}$ is $|\mu|$-integrable and dominates the function $|f - f_n|_{\mathbb{R}^m}$. Thus, we get

$$\|f - f_n\|_{L^1(|\mu|)^m} = \int_\Omega |f - f_n|_{\mathbb{R}^m}\, \mathrm{d}|\mu| \to 0$$

by the dominated convergence theorem. This shows $f_n \to f$ in $L^1(|\mu|)^m$. Finally, we have

$$\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} \leftarrow \langle \mu,\, f_n\rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} = \int_\Omega (f_n,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu| \to \int_\Omega (f,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu|.$$

The first limit follows from $f_n \to f$ in $H^1_0(\Omega)^m$ and the second one by $f_n \to f$ in $L^1(|\mu|)^m$ and $\mu' \in L^\infty(|\mu|)^m$.

Now, let $f \in H_0^1(\Omega)^m$ be given. The formula (6.4.4) follows from

$$\langle \mu,\, f\rangle_{H^{-1}(\Omega)^m,H_0^1(\Omega)^m} = \lim_{M\to\infty} \langle \mu,\, T_M\, f\rangle_{H^{-1}(\Omega)^m,H_0^1(\Omega)^m} = \lim_{M\to\infty} \int_\Omega (T_M\, f,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu|,$$

since $T_M\, f \to f$ in $H_0^1(\Omega)^m$ and by $T_M\, f \in H_0^1(\Omega)^m \cap L^\infty(\Omega)^m$ we can use (6.4.2) for $T_M\, f$.

We emphasize that, in general, we cannot pass to the limit under the integral in (6.4.4). Although we have $T_M\, f \to f$ pointwise $|\mu|$-a.e., we cannot find an integrable function dominating $T_M\, f$, since f may not belong to $L^1(|\mu|)$, see Section 6.B. We also refer to Brézis, Browder, 1979, Example 2, which shows that (6.4.3) does not hold for all $f \in H_0^1(\Omega)^m$.

With this preparation, we are able to compute the normal cone. A similar characterization can be found in Grun-Rehomme, 1977, Théorème 3, but there the Lebesgue decomposition of the measure $\mu \in \mathcal{T}_{\mathcal{C}}(u)^\circ$ is used.

Theorem 6.4.3. Suppose $0 \in \mathrm{int}(C)$. For $u \in \mathcal{C}$ we have

$$\mathcal{T}_{\mathcal{C}}(u)^\circ = \{\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m : \mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ \text{ for } |\mu|\text{-a.a. } \omega \in \Omega\}.$$

Proof. Since $(\mathcal{C}-u)^\circ = \mathcal{T}_{\mathcal{C}}(u)^\circ$, it is sufficient to show

$$(\mathcal{C}-u)^\circ = \{\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m : \mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ \text{ for } |\mu|\text{-a.a. } \omega \in \Omega\}.$$

"$\supset$": Let $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$ be given such that $\mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ$ for $|\mu|$-a.a. $\omega \in \Omega$. For $v \in \mathcal{C}$ we have by using (6.4.4)

$$\langle \mu,\, v-u\rangle_{H^{-1}(\Omega)^m,H_0^1(\Omega)^m} = \lim_{M\to\infty} \int_\Omega (T_M(v-u),\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu| \le 0.$$

Here, we employed $(T_M(v(\omega)-u(\omega)),\, \mu'(\omega))_{\mathbb{R}^m} \le 0$ for $|\mu|$-a.a. $\omega \in \Omega$ since $v(\omega)-u(\omega) \in \mathcal{T}_C(u(\omega))$, $T_M(v(\omega)-u(\omega)) \in \mathcal{T}_C(u(\omega))$ and $\mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ$ for $|\mu|$-a.a. $\omega \in \Omega$. This shows $\mu \in (\mathcal{C}-u)^\circ$.

"$\subset$": Let $\mu \in (\mathcal{C}-u)^\circ$ be given. Now, let $h \in H_0^1(\Omega)^m$ with $\|h\|_{L^\infty(\Omega)^m} \le 1$ be given. Since $B_r(0) \subset C$ for some $r > 0$, we have $\pm r\, h \in \mathcal{C}$. Now,

$$\langle \mu,\, \pm r\, h - u\rangle_{H^{-1}(\Omega)^m,H_0^1(\Omega)^m} \le 0$$

implies

$$|\langle \mu,\, h\rangle_{H^{-1}(\Omega)^m,H_0^1(\Omega)^m}| \le \frac{1}{r}\, \langle \mu,\, u\rangle_{H^{-1}(\Omega)^m,H_0^1(\Omega)^m}.$$

This shows

$$|\langle \mu,\, h\rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m}| \leq \frac{\|h\|_{L^\infty(\Omega)^m}}{r} \langle \mu,\, u\rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m}$$

for all $h \in H_0^1(\Omega)^m \cap C_0(\Omega)^m$. By the Riesz representation theorem, we obtain that μ is a regular Borel measure, i.e., $\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m$.

Now, let $h \in C_0^\infty(\Omega), 0 \leq h \leq 1$, $M \in \mathbb{N}$ and $v \in C$ be arbitrary. Then, $h\, T_M(v-u)+u \in \mathcal{C}$ and by Lemma 6.4.2 this yields

$$\begin{aligned} 0 &\geq \langle \mu,\, h\, T_M(v-u) + u - u\rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m} \\ &= \langle \mu,\, h\, T_M(v-u)\rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m} \\ &= \int_\Omega h\, (T_M(v-u),\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu| \qquad \forall h \in C_0^\infty(\Omega), 0 \leq h \leq 1. \end{aligned}$$

Using a mollification argument and $(T_M(v-u),\, \mu')_{\mathbb{R}^m} \in L^\infty(|\mu|)$, we get

$$0 \geq \int_\Omega h\, (T_M(v-u),\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu| \qquad \forall h \in C_c(\Omega), 0 \leq h \leq 1.$$

Here, $C_c(\Omega)$ is the space of continuous functions whose support is compact in Ω. By scaling, we can drop the upper bound $h \leq 1$. Owing to Rudin, 1987, Theorem 3.14, $C_c(\Omega)$ is dense in $L^2(|\mu|)$, thus

$$0 \geq \int_\Omega h\, (T_M(v-u),\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu| \qquad \forall h \in L^2(|\mu|), 0 \leq h.$$

This yields $(T_M(v-u(\omega)),\, \mu'(\omega))_{\mathbb{R}^m} \leq 0$ for $|\mu|$-a.a. $\omega \in \Omega$. Since $M \in \mathbb{N}$ was arbitrary, this yields $(v-u(\omega),\, \mu'(\omega))_{\mathbb{R}^m} \leq 0$ for $|\mu|$-a.a. $\omega \in \Omega$. Since $v \in C$ was arbitrary, this yields $\mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ$ for $|\mu|$-a.a. $\omega \in \Omega$.

Note that the above theorem does not use any result from Section 6.3.

By the bipolar theorem, we also obtain a characterization of the tangent cone.

Theorem 6.4.4. Suppose that $0 \in \mathrm{int}(C)$ and let $u \in \mathcal{C}$ be given. Then,

$$\mathcal{T}_{\mathcal{C}}(u) = \{v \in H_0^1(\Omega)^m : v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ q.e. in } \Omega\}.$$

Proof. From Lemma 6.3.1 we obtain

$$\mathcal{T}_{\mathcal{C}}(u) \subset \{v \in H_0^1(\Omega)^m : v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ q.e. in } \Omega\}.$$

From Theorem 6.4.3 and the bipolar theorem, we infer

$$\mathcal{T}_{\mathcal{C}}(u) = \{\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m : \mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ \text{ for } |\mu|\text{-a.a. } \omega \in \Omega\}^\circ.$$

Now, a direct calculation yields

$$\begin{aligned}&\{v \in H_0^1(\Omega)^m : v(\omega) \in \mathcal{T}_C(u(\omega)) \text{ q.e. in } \Omega\}\\ &\quad \subset \{\mu \in H^{-1}(\Omega)^m \cap \mathcal{M}(\Omega)^m : \mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ \text{ for } |\mu|\text{-a.a. } \omega \in \Omega\}^\circ\end{aligned}$$

and this shows the claim.

The following remark compares the techniques and results from Section 6.3 and Section 6.4.

Remark 6.4.5. In Section 6.3 as well as in Section 6.4, we obtained a characterization of the tangent cone of $\mathcal{C}$, and both sections used rather different techniques. Whereas in Section 6.3 we directly approximated $v \in \mathcal{T}_\mathcal{C}(u)$ by feasible directions (which was rather subtle), we used the bipolar theorem in Section 6.4, which required to work with vector valued measures in $H^{-1}(\Omega)^m$. To this end, the novel representation result (6.4.4) was crucial.

In the next section, we use similar arguments as in Section 6.3 to show the polyhedricity of $\mathcal{C}$ (under additional assumptions). Since elements in $(\mathcal{T}_\mathcal{C}(u) \cap \mu^\perp)^\circ$ might not be measures, the technique of Section 6.4 cannot be used.

Finally, we highlight that the technique in Section 6.3 requires Assumption 6.3.3 to be satisfied, whereas in Section 6.4 the weaker assumption $0 \in \operatorname{int}(\mathcal{C})$ is sufficient. It is not clear whether the arguments of Section 6.3 can be adapted if Assumption 6.3.3 fails, but we still have $0 \in \operatorname{int}(\mathcal{C})$.

6.5. Polyhedricity under LICQ

In this section, we consider the polyhedricity of $\mathcal{C}$. Recall, that $\mathcal{C}$ is called polyhedric w.r.t. (u, μ), where $u \in \mathcal{C}$, $\mu \in \mathcal{T}_\mathcal{C}(u)^\circ$, if

$$\mathcal{R}_\mathcal{C}(u) \cap \mu^\perp \text{ is dense in } \mathcal{T}_\mathcal{C}(u) \cap \mu^\perp.$$

To this end, we put an additional assumption on $C \subset \mathbb{R}^m$.

Assumption 6.5.1. There exist $N \in \mathbb{N}$ and $n_i \in \mathbb{R}^m$, $b_i \in (0, \infty)$ for $i = 1, \dots, N$, such that

$$C = \{x \in \mathbb{R}^m : (x, n_i)_{\mathbb{R}^m} \leq b_i \ \forall i = 1, \dots, N\}.$$

Further, we assume that the linear independence constraint qualification (LICQ) is satisfied, that is, the family $\{n_i : (x, n_i)_{\mathbb{R}^m} = b_i\}$ is linearly independent for all $x \in C$.

Note that Assumption 6.5.1 implies $0 \in \operatorname{int}(C)$, since we required $b_i > 0$.

Throughout this section, we fix $u \in \mathcal{C}$ and $\mu \in \mathcal{T}_{\mathcal{C}}(u)^\circ$.
First, we provide a characterization of $\mathcal{T}_{\mathcal{C}}(u) \cap \mu^\perp$ which does not require Assumption 6.5.1 to be satisfied.

Lemma 6.5.2. We have $v \in \mathcal{T}_{\mathcal{C}}(u) \cap \mu^\perp$ if and only if

$$v(\omega) \in \mathcal{T}_C(u(\omega)) \quad \text{for q.a. } \omega \in \Omega \qquad \text{and} \qquad (v(\omega),\, \mu'(\omega))_{\mathbb{R}^m} = 0 \quad \text{for } |\mu|\text{-a.a. } \omega \in \Omega.$$

Here, $|\mu|$ is the total variation of μ and μ' the Radon-Nikodým derivative of μ w.r.t. $|\mu|$, see Section 6.4.

Proof. From Theorem 6.4.4, we obtain

$$v \in \mathcal{T}_{\mathcal{C}}(u) \qquad \Longleftrightarrow \qquad v(\omega) \in \mathcal{T}_C(u(\omega)) \quad \text{for q.a. } \omega \in \Omega.$$

It remains to show the equivalence

$$v \in \mu^\perp \qquad \Longleftrightarrow \qquad (v(\omega),\, \mu'(\omega))_{\mathbb{R}^m} = 0 \quad \text{for } |\mu|\text{-a.a. } \omega \in \Omega.$$

for all $v \in \mathcal{T}_{\mathcal{C}}(u)$.
To this end, let $v \in \mathcal{T}_{\mathcal{C}}(u)$ be given. Due to

$$v(\omega) \in \mathcal{T}_C(u(\omega)) \quad \text{and} \quad \mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ \qquad \text{for } |\mu|\text{-a.a. } \omega \in \Omega,$$

see Lemma 6.4.1 and Theorem 6.4.3, we find

$$(v(\omega),\, \mu'(\omega))_{\mathbb{R}^m} \leq 0 \quad \text{for } |\mu|\text{-a.a. } \omega \in \Omega.$$

For $M > 0$, the truncation $T_M v$, see (6.4.2) for the definition of T_M, is given by

$$T_M v(\omega) = \min\Big(1, \frac{M}{|v(\omega)|_{\mathbb{R}^m}}\Big)\, v(\omega).$$

The factor $\min(1, \frac{M}{|v|_{\mathbb{R}^m}})$ is monotonically increasing in M and converges to 1 pointwise as $M \to \infty$. Hence, the function $g : (0, \infty) \to \mathbb{R}$, defined via

$$g(M) := \langle \mu,\, T_M v \rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m} = \int_\Omega \min\Big(1, \frac{M}{|v|_{\mathbb{R}^m}}\Big)(v,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu|,$$

is monotonically decreasing, takes non-positive values and $g(M)$ converges towards $\langle \mu,\, v \rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m}$ as $M \to \infty$ by Lemma 6.4.2.
Hence, we get the chain of equivalences

$$\begin{aligned} v \in \mu^\perp \qquad &\Longleftrightarrow \qquad \lim_{M \to \infty} g(M) = 0 \qquad \Longleftrightarrow \qquad \forall M > 0 : g(M) = 0 \\ &\Longleftrightarrow \qquad (v(\omega),\, \mu'(\omega))_{\mathbb{R}^m} = 0 \quad \text{for } |\mu|\text{-a.a. } \omega \in \Omega. \end{aligned}$$

Note that the last "⇒" follows from $\min(1, \frac{M}{|v|_{\mathbb{R}^m}}) > 0$ for q.a. $\omega \in \Omega$ and hence for $|\mu|$-a.a. $\omega \in \Omega$.

As in Section 6.3, we start with a truncation argument.

Lemma 6.5.3. Each function $v \in \mathcal{T}_{\mathcal{C}}(u) \cap \mu^{\perp}$ can be approximated in $H_0^1(\Omega)^m$ from functions in

$$\mathcal{T}_{\mathcal{C}}(u) \cap \mu^{\perp} \cap L^{\infty}(\Omega)^m.$$

Proof. Let $v \in \mathcal{T}_{\mathcal{C}}(u) \cap \mu^{\perp}$ be given. For $M > 0$, we show that $T_M v \in \mathcal{T}_{\mathcal{C}}(u) \cap \mu^{\perp} \cap L^{\infty}(\Omega)^m$. Then, the claim follows from Theorem 6.A.2. By definition, we have $T_M v \in L^{\infty}(\Omega)^m$.

Since $v \in \mathcal{T}_{\mathcal{C}}(u)$, we have $v(\omega) \in \mathcal{T}_C(u(\omega))$ for q.a. $\omega \in \Omega$. This yields $(T_M v)(\omega) = T_M(v(\omega)) \in \mathcal{T}_C(u(\omega))$ for q.a. $\omega \in \Omega$ since $\mathcal{T}_C(u(\omega))$ is a cone. By Theorem 6.4.4 we find $T_M v \in \mathcal{T}_{\mathcal{C}}(u)$.

Finally, $\langle \mu, T_M v \rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m} = 0$ follows from

$$(v(\omega), \mu'(\omega))_{\mathbb{R}^m} = 0 \quad \text{for } |\mu|\text{-a.a. } \omega \in \Omega,$$

see Lemma 6.5.2.

Now, by using Assumption 6.5.1, we can show that

$$\mathcal{R}_{\mathcal{C}}(u) \cap \mu^{\perp} \text{ is dense in } \mathcal{T}_{\mathcal{C}}(u) \cap \mu^{\perp} \cap L^{\infty}(\Omega)^m.$$

Lemma 6.5.4. Assume that Assumption 6.5.1 is satisfied. Let $v \in \mathcal{T}_{\mathcal{C}}(u) \cap \mu^{\perp} \cap L^{\infty}(\Omega)^m$ be given. Then, v can be approximated in $H_0^1(\Omega)^m$ by functions in $\mathcal{R}_{\mathcal{C}}(u) \cap \mu^{\perp}$.

Proof. W.l.o.g. we assume $|n_i|_{\mathbb{R}^m} = 1$ for all $i \in \{1, \dots, N\}$. By $h_i : C \to \mathbb{R}^m$ we denote the Lipschitz continuous functions from Lemma 6.C.1.

In order to apply Lemma 6.3.6, we define for $i \in \{1, \dots, N\}$ and $\delta \geq 0$ the sets

$$M_\delta^i := \{\omega \in \Omega : (u(\omega), n_i)_{\mathbb{R}^m} \geq b_i - \delta\}$$

and $f^i := (v, n_i)_{\mathbb{R}^m} \in H_0^1(\Omega) \cap L^{\infty}(\Omega)$. Then, the assumptions of Lemma 6.3.6 are satisfied and we find $f_\delta^i \in H_0^1(\Omega) \cap L^{\infty}(\Omega)$ with $f_\delta^i \to f^i$ in $H_0^1(\Omega)$ as $\delta \searrow 0$,

$$f_\delta^i \leq 0 \quad \text{q.e. on } M_\delta^i \qquad \text{and} \qquad f_\delta^i = 0 \quad \text{q.e. on } M_\delta^i \cap \{\omega \in \Omega : f^i(\omega) = 0\}.$$

Moreover, the $L^{\infty}(\Omega)$-norm of f_δ^i is bounded by the $L^{\infty}(\Omega)^m$-norm of v.

Now, we define

$$v_\delta := v + \sum_{i=1}^{N} (f_\delta^i - f^i)\, h_i(u) = v + \sum_{i=1}^{N} (f_\delta^i - (v, n_i)_{\mathbb{R}^m})\, h_i(u).$$

Since h_i is globally bounded and Lipschitz continuous, we get $v_\delta \in H_0^1(\Omega)^m \cap L^{\infty}(\Omega)^m$, see also Lemma 6.A.1.

We proceed by showing the weak convergence of v_δ towards v in $H^1_0(\Omega)^m$. Since we have $f^i_\delta \to f_\delta$ in $H^1_0(\Omega)$ as $\delta \searrow 0$, we find $v_\delta \to v$ in $W^{1,1}_0(\Omega)^m$. Further, we can show that v_δ is bounded in $H^1_0(\Omega)^m$, and this yields $v_\delta \rightharpoonup v$ in $H^1_0(\Omega)^m$.

As a next step, we show $v_\delta \in \mathcal{R}_{\mathcal{C}}(u) \cap \mu^\perp$.

We denote by M, $\hat{\delta}$ the constants from Lemma 6.C.3 and by L the largest Lipschitz constant of the functions h_i. From now on, we assume $\delta \leq \hat{\delta}$. We set $\lambda := \min\{(2\,N\,L\,M\,\|v\|_{L^\infty(\Omega)^m})^{-1}, \delta\,\|v_\delta\|^{-1}_{L^\infty(\Omega)^m}\}$ and show $u + \lambda\, v_\delta \in \mathcal{C}$. By owing to Lemma 6.C.3, we find that for $i \in \{1, \dots, N\}$ and for q.a. $\omega \in M^i_\delta$, there exists $\tilde{x} \in C$, such that

$$(\tilde{x},\, n_i)_{\mathbb{R}^m} = b_i \quad \text{and} \quad |u(\omega) - \tilde{x}|_{\mathbb{R}^m} \leq M\,\big(b_i - (u(\omega),\, n_i)_{\mathbb{R}^m}\big).$$

In particular, we have $(h_j(\tilde{x}),\, n_i)_{\mathbb{R}^m} = \delta_{ij}$. Hence, we have for $i \in \{1, \dots, N\}$ and for q.a. $\omega \in M^i_\delta$ the estimate

$$\begin{aligned}
\big(u(\omega) + \lambda\, v_\delta(\omega),\, n_i\big)_{\mathbb{R}^m} &= \big(u(\omega) + \lambda\, v(\omega),\, n_i\big)_{\mathbb{R}^m} \\
&\quad + \lambda \sum_{j=1}^{N} \big(f^j_\delta(\omega) - f^j(\omega)\big)\,\big(h_j(u(\omega)),\, n_i\big)_{\mathbb{R}^m} \\
&\leq \big(u(\omega) + \lambda\, v(\omega),\, n_i\big)_{\mathbb{R}^m} \\
&\quad + \lambda \sum_{j=1}^{N} \big(f^j_\delta(\omega) - f^j(\omega)\big)\,\big(h_j(\tilde{x}),\, n_i\big)_{\mathbb{R}^m} \\
&\quad + \lambda\, N\, 2\, \|v\|_{L^\infty(\Omega)^m}\, L\, M\,\big(b_i - (u(\omega),\, n_i)_{\mathbb{R}^m}\big) \\
&\leq \big(u(\omega) + \lambda\, v(\omega),\, n_i\big)_{\mathbb{R}^m} + \lambda\,\big(f^i_\delta(\omega) - f^i(\omega)\big) \\
&\quad + \big(b_i - (u(\omega),\, n_i)_{\mathbb{R}^m}\big) \\
&= \lambda\,\big(v(\omega),\, n_i\big)_{\mathbb{R}^m} + \lambda\,\big(f^i_\delta(\omega) - (v(\omega),\, n_i)_{\mathbb{R}^m}\big) + b_i \leq b_i.
\end{aligned}$$

On the other hand, for $i \in \{1, \dots, N\}$ and for q.a. $\omega \notin M^i_\delta$, we have

$$\big(u(\omega) + \lambda\, v_\delta(\omega),\, n_i\big)_{\mathbb{R}^m} = (u(\omega),\, n_i)_{\mathbb{R}^m} + \lambda\,\|v_\delta\|_{L^\infty(\Omega)^m} < b_i - \delta + \delta = b_i.$$

This shows $u + \lambda\, v_\delta \in \mathcal{C}$.

Now, we show $\langle \mu,\, v_\delta \rangle_{H^{-1}(\Omega)^m, H^1_0(\Omega)^m} = 0$. By Theorem 6.4.3, we have $\mu'(\omega) \in \mathcal{T}_C(u(\omega))^\circ$ for $|\mu|$-a.a. $\omega \in \Omega$. By $v \in \mathcal{T}_{\mathcal{C}}(u)$, Theorem 6.4.4 and Lemma 6.4.2, we have $(v(\omega),\, \mu'(\omega))_{\mathbb{R}^m} = 0$ for $|\mu|$-a.a. $\omega \in \Omega$. Moreover, there exist functions $\alpha_i : \Omega \to [0, \infty)$, such that $\alpha_i(\omega) = 0$ if $(u(\omega),\, n_i)_{\mathbb{R}^m} < b_i$ and $\mu'(\omega) = \sum_{i=1}^N \alpha_i(\omega)\, n_i$ holds for $|\mu|$-a.a. $\omega \in \Omega$. We do not claim that the functions α_i are measurable. Hence, we have

$$0 = (v(\omega),\, \mu'(\omega))_{\mathbb{R}^m} = \sum_{i=1}^{N} \alpha_i(\omega)\,(v(\omega),\, n_i)_{\mathbb{R}^m} = \sum_{i=1}^{N} \alpha_i(\omega)\, f^i(\omega). \tag{6.5.1}$$

Note that $\alpha_i = 0$ on $\Omega \setminus M_0^i$ and $f^i \geq 0$ on M_0^i. Thus, all summands in (6.5.1) are non-negative and we have for $|\mu|$-a.a. $\omega \in \Omega$

$$\alpha_i(\omega) \neq 0 \quad \Rightarrow \quad \omega \in M_0^i \text{ and } f^i(\omega) = 0 \quad \Rightarrow \quad f_\delta^i(\omega) = 0.$$

Moreover, $\omega \in M_0^i$ if and only if $(u(\omega),\, n_i)_{\mathbb{R}^m} = b_i$, which implies $(h_j(u(\omega)),\, n_i)_{\mathbb{R}^m} = \delta_{ij}$ for all $j = 1, \dots, N$. This shows

$$\begin{aligned}(v_\delta(\omega),\, \mu'(\omega))_{\mathbb{R}^m} &= \sum_{i=1}^N \alpha_i(\omega)\, (v_\delta(\omega),\, n_i)_{\mathbb{R}^m} = \sum_{i=1}^N \alpha_i(\omega) \sum_{j=1}^N (f_\delta^j - f^j)\, (h_j(u(\omega)),\, n_i)_{\mathbb{R}^m} \\ &= \sum_{i=1}^N \alpha_i(\omega)\, (f_\delta^i - f^i) = 0\end{aligned}$$

for $|\mu|$-a.a. $\omega \in \Omega$. By using Lemma 6.4.2, we get

$$\langle \mu,\, v_\delta \rangle_{H^{-1}(\Omega)^m, H_0^1(\Omega)^m} = \int_\Omega (v_\delta,\, \mu')_{\mathbb{R}^m}\, \mathrm{d}|\mu| = 0.$$

Hence, we have $v_\delta \in \mathcal{R}_{\mathcal{C}}(u) \cap \mu^\perp$.

It remains to show that v can be approximated strongly in $H_0^1(\Omega)^m$ by functions in $\mathcal{R}_{\mathcal{C}}(u) \cap \mu^\perp$. The set $\mathcal{R}_{\mathcal{C}}(u) \cap \mu^\perp$ is convex and $v_\delta \rightharpoonup v$ in $H_0^1(\Omega)^m$. Owing to Mazur's lemma, there is a sequence in $\mathcal{R}_{\mathcal{C}}(u) \cap \mu^\perp$ which converges strongly in $H_0^1(\Omega)$ towards v.

By combining Lemma 6.5.3 and Lemma 6.5.4, we obtain the main result of this section.

Theorem 6.5.5. Assumption 6.5.1 implies that $\mathcal{C}$ is polyhedric w.r.t. all $u \in \mathcal{C}$ and $\mu \in \mathcal{T}_{\mathcal{C}}(u)^\circ$.

It is expected, that $\mathcal{C}$ is polyhedric, if $C \subset \mathbb{R}^m$ is a polyhedral set (i.e., a finite intersection of closed half-spaces), but we were not able to prove the result in this generality. In particular, the violation of LICQ impedes the existence of the functions h_i and these functions are crucial in the proof of Lemma 6.5.4.

6.6. Optimal control of a string in a polyhedral tube

In this section, we consider the optimal control of a string whose deflection is constrained by a polyhedral tube. We apply the results of the previous sections and show that a local minimizer satisfies a system of strong stationarity. For simplicity, we consider the two-dimensional situation $m = 2$.

For a given length $L > 0$, we set $\Omega = (0, L)$. Due to $\Omega \subset \mathbb{R}^1$, we have $\mathrm{cap}(A) > 0$ for all $A \subset \Omega$, $A \neq \emptyset$. Hence, the notion of "quasi-everywhere" is equivalent to "everywhere".

The deflection of the string is modelled by $y \in H_0^1(\Omega)^2$. If we apply a force $u \in L^2(\Omega)^2$, the unconstrained deflection y would satisfy the differential equation

$$-\Delta y = u \quad \text{in } \Omega, \qquad\qquad y = 0 \quad \text{in } \{0, L\}.$$

Now, we choose a polygon $C \subset \mathbb{R}^2$ satisfying $0 \in \operatorname{int}(C)$ and consider a string which is constrained to the tube $\Omega \times C$. Then, the deflection of the constrained string is given by the solution of the energy minimization problem

$$\begin{aligned} &\text{Minimize} && \int_\Omega \frac{1}{2} |\nabla y|_{\mathbb{R}^2}^2 - (y,\, u)_{\mathbb{R}^2} \,\mathrm{d}\omega \\ &\text{such that} && y(\omega) \in C \quad \text{for a.a. } \omega \in \Omega. \end{aligned}$$

The unique solution $y = S(u)$ of this problem is characterized by the (necessary and sufficient) optimality conditions

$$y \in \mathcal{C}, \qquad\qquad \xi \in \mathcal{T}_{\mathcal{C}}(y)^\circ, \qquad\qquad A\,y - u + \xi = 0,$$

where $A : H_0^1(\Omega)^2 \to H^{-1}(\Omega)^2$ is given by

$$\langle A\,y,\, v\rangle_{H^{-1}(\Omega)^2, H_0^1(\Omega)^2} := \int_\Omega (\nabla y,\, \nabla v)_{\mathbb{R}^2} \,\mathrm{d}\omega.$$

For a given objective $f : H_0^1(\Omega)^2 \times L^2(\Omega)^2 \to \mathbb{R}$, which is assumed to be Fréchet differentiable, we consider the optimal control problem

$$\begin{aligned} &\text{Minimize} && f(y, u) && \text{(6.6.1a)} \\ &\text{with respect to} && y \in H_0^1(\Omega)^2, u \in L^2(\Omega)^2, \xi \in H^{-1}(\Omega)^2 && \text{(6.6.1b)} \\ &\text{such that} && y \in \mathcal{C} && \text{(6.6.1c)} \\ & && \xi \in \mathcal{T}_{\mathcal{C}}(y)^\circ && \text{(6.6.1d)} \\ & && A\,y - u + \xi = 0. && \text{(6.6.1e)} \end{aligned}$$

Using additional assumptions on the objective f, we may deduce the existence of global solutions by standard arguments, but this will not be discussed here.

The main result of this section is the following optimality system.

Theorem 6.6.1. Let $(\bar{y}, \bar{u}, \bar{\xi})$ be a locally optimal solution of (6.6.1). Then, there exist $p \in H_0^1(\Omega)^2$ and $\mu \in H^{-1}(\Omega)^2$, such that the system

$$\begin{aligned} f_y(\bar{y}, \bar{u}) + \mu + A^\star p = 0, \qquad & -p \in \mathcal{T}_{\mathcal{C}}(\bar{y}) \cap \bar{\xi}^\perp, && \text{(6.6.2a)} \\ f_u(\bar{y}, \bar{u}) - p = 0, \qquad & \mu \in \big(\mathcal{T}_{\mathcal{C}}(\bar{y}) \cap \bar{\xi}^\perp\big)^\circ && \text{(6.6.2b)} \end{aligned}$$

is satisfied. Here, f_y, f_u denote the partial derivatives of f, w.r.t. y and u, respectively.

Proof. First, we remark that every polygon in $\mathbb{R}^2$ can be described by affine constraints which satisfy LICQ. Hence, Assumption 6.5.1 is satisfied and we can apply Theorem 6.5.5 to obtain the polyhedricity of $\mathcal{C}$.

Now, we can argue as in the proof of Mignot, 1976, Proposition 4.1 and get the existence of $p \in H_0^1(\Omega)^2$ and $\mu \in H^{-1}(\Omega)^2$, such that the system (6.6.2) is satisfied. We also refer to Hintermüller, Surowiec, 2011, Theorem 4.6 and Section 1.6.1 for different approaches to obtain this optimality system.

Using the nomenclature from finite dimensions, the system (6.6.2) is of strongly stationary type. For the extension of the finite-dimensional nomenclature to infinite dimensions, we refer to Section 1.5.4.

For a characterization of the critical cone $\mathcal{T}_{\mathcal{C}}(\bar{y}) \cap \bar{\xi}^{\perp}$, which appears in (6.6.2), we refer to Lemma 6.5.2. Recall that we may replace "quasi-everywhere" in Lemma 6.5.2 by "everywhere" due to $\Omega \subset \mathbb{R}^1$. However, the polyhedricity of $\mathcal{C}$ which is established via Theorem 6.5.5 is a delicate issue even in dimension one.

Finally, we give a higher-dimensional application by means of a simplified phase-field model. The domain which is occupied by the $m+1$ phases is $\Omega \subset \mathbb{R}^d$. The concentration of phase i is denoted by y_i and, thus, we have $y_i \geq 0$ and $\sum_{i=1}^{m+1} y_i = 1$. Further, we assume that we have the Dirichlet boundary condition $y_i = 1/(m+1)$ on $\partial\Omega$ for all i. Now, we eliminate the phase $m+1$ via $y_{m+1} = 1 - \sum_{i=1}^{m} y_i$, and replace each phase y_i by $y_i - 1/(m+1)$. Then, the phase-field vector $(y_1, \dots, y_m)$ belongs to the simplex

$$C := \Big\{ x \in \mathbb{R}^m : x \geq -\frac{1}{m+1} \quad \text{and} \quad \sum_{i=1}^{m} x_i \leq \frac{1}{m+1} \Big\}$$

and satisfies the homogeneous Dirichlet boundary condition $y = 0$ on $\partial\Omega$. It is easy to check that this set C satisfies Assumption 6.5.1 and thus we can proceed as above to obtain a higher-dimensional application of Theorem 6.5.5.

6.A. Nemytskii operators on Sobolev spaces

First, we provide a result that Nemytskii operators associated with globally Lipschitz continuous functions $f : \mathbb{R}^m \to \mathbb{R}^n$, map $H_0^1(\Omega)^m$ to $H_0^1(\Omega)^n$. A more general result is provided in Marcus, Mizel, 1973, Theorem 2.1 in case Ω satisfies the cone condition. Since we are only interested in the case of functions with zero trace, we can drop the assumption on Ω.

Lemma 6.A.1. Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. We assume that the function $f : \mathbb{R}^m \to \mathbb{R}^n$ is globally Lipschitz continuous and $f(0) = 0$. Then, the associated Nemytskii operator T, which is defined for functions $u : \Omega \to \mathbb{R}^m$ via

$$(T\,u)(\omega) = f(u(\omega)),$$

maps $H_0^1(\Omega)^m$ to $H_0^1(\Omega)^n$ and there is a constant $c > 0$, such that $\|T\,u\|_{H_0^1(\Omega)^n} \leq c\,\|u\|_{H_0^1(\Omega)^m}$ holds for all $u \in H_0^1(\Omega)^m$.

Proof. Since Ω is bounded, there is $R > 0$, such that $U_R(0) := \{\omega \in \mathbb{R}^d : |\omega|_{\mathbb{R}^d} < R\}$ contains $\bar{\Omega}$. We identify u with its extension by zero and get $u \in H_0^1(B_R(0))^m$. Since $B_R(0)$ satisfies the cone condition, we can invoke Marcus, Mizel, 1973, Theorem 2.1, and obtain $T\,u \in H^1(B_R(0))^n$. Following the proof, we also find the bound $\|T\,u\|_{H^1(B_R(0))^n} \leq c\,\|u\|_{H^1(B_R(0))^m}$.

It remains to show $(T\,u)_{|\Omega} \in H_0^1(\Omega)^n$. Since $u \in H_0^1(\Omega)^m$, its extension by zero belongs to $H^1(\mathbb{R}^d)^m$ and $u = 0$ q.e. on $\mathbb{R}^d \setminus \Omega$, see Heinonen, Kilpeläinen, Martio, 1993, Theorem 4.5. By using $f(0) = 0$, this shows $T\,u = 0$ q.e. on $\mathbb{R}^d \setminus \Omega$ and thus, by using Heinonen, Kilpeläinen, Martio, 1993, Theorem 4.5 again, we have $T\,u \in H_0^1(\Omega)^m$.

Finally, we provide a chain rule for the truncation of vector-valued Sobolev functions. Similar to the classical argument in the scalar-valued case, we use a smooth approximation of the truncation and pass to the limit by using the dominated convergence theorem.

Theorem 6.A.2. Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. For $M > 0$, we define the truncation $T_M : \mathbb{R}^m \to \mathbb{R}^m$ by

$$T_M(x) := \min\Big(1, \frac{M}{|x|_{\mathbb{R}^m}}\Big)\, x = \begin{cases} x, & \text{if } |x|_{\mathbb{R}^m} \leq M, \\ \frac{M}{|x|_{\mathbb{R}^m}}\, x, & \text{if } |x|_{\mathbb{R}^m} > M. \end{cases}$$

Then, the associated Nemytskii operator, which is denoted by the same symbol, maps $H_0^1(\Omega)^m$ to itself and for all $u \in H_0^1(\Omega)^m$ we have

$$\frac{\partial}{\partial \omega_k}(T_M\,u)_i(\omega) = \begin{cases} \frac{\partial}{\partial \omega_k} u_i(\omega), & \text{if } |u(\omega)|_{\mathbb{R}^m} \leq M, \\ \frac{M}{|u(\omega)|_{\mathbb{R}^m}}\, \frac{\partial}{\partial \omega_k} u_i(\omega) + \frac{M}{|u(\omega)|^3_{\mathbb{R}^m}}\, u_i(\omega) \sum_{j=1}^m u_j(\omega)\, \frac{\partial}{\partial \omega_k} u_j(\omega), & \text{if } |u(\omega)|_{\mathbb{R}^m} > M \end{cases}$$

for almost all $\omega \in \Omega$. Moreover, $T_M\,u \to u$ in $H_0^1(\Omega)^m$ as $M \to \infty$.

Proof. Let $u \in H_0^1(\Omega)^m$ be given. By Lemma 6.A.1, we get $T_M u \in H_0^1(\Omega)^m$. Next, we prove the expression for $\nabla(T_M u)$. Since T_M is not differentiable on the set $\{x \in \mathbb{R}^m : |x|_{\mathbb{R}^m} = M\}$, the chain rule of Marcus, Mizel, 1972, Theorem 2.1 is not applicable.

Thus, we are going to provide a differentiable approximation T_M^σ of T_M. For $\sigma \in (0,1)$, we define $m^\sigma : \mathbb{R} \to \mathbb{R}$ via

$$m^\sigma(x) = \begin{cases} \frac{\sigma}{2} + x, & \text{if } x < 1 - \sigma, \\ 1 - \frac{1}{2\sigma}\,(1-x)^2, & \text{if } 1 - \sigma \leq x < 1, \\ 1, & \text{if } 1 \leq x, \end{cases}$$

which is a differentiable approximation of $x \mapsto \min(1, x)$. We find

$$(m^\sigma)'(x) = \begin{cases} 1, & \text{if } x < 1 - \sigma, \\ \frac{1}{\sigma}(1 - x), & \text{if } 1 - \sigma \le x < 1, \\ 0, & \text{if } 1 \le x. \end{cases}$$

Now, a differentiable approximation $T_M^\sigma : \mathbb{R}^m \to \mathbb{R}^m$ of T_M is given by

$$T_M^\sigma(x) := m^\sigma\Big(\frac{M}{|x|_{\mathbb{R}^m}}\Big)\, x.$$

The partial derivatives of the components of T_M^σ are

$$\frac{\partial}{\partial x_j}(T_M^\sigma)_i(x) = m^\sigma\Big(\frac{M}{|x|_{\mathbb{R}^m}}\Big)\,\delta_{ij} - (m^\sigma)'\Big(\frac{M}{|x|_{\mathbb{R}^m}}\Big)\,\frac{M}{|x|^3_{\mathbb{R}^m}}\,x_i\,x_j,$$

where δ_{ij} is the Kronecker delta. Here and in what follows, we use the conventions

$$m^\sigma\Big(\frac{M}{|x|_{\mathbb{R}^m}}\Big) = 0 \qquad \text{and} \quad (m^\sigma)'\Big(\frac{M}{|x|_{\mathbb{R}^m}}\Big)\,\frac{M}{|x|^3_{\mathbb{R}^m}}\,x_i\,x_j = 0$$

in case $|x|_{\mathbb{R}^m} = 0$. By Lemma 6.A.1, we find $T_M^\sigma u \in H_0^1(\Omega)^m$. Since T_M^σ is differentiable, we can apply the chain rule of Marcus, Mizel, 1972, Theorem 2.1 and obtain

$$\begin{aligned}\frac{\partial}{\partial \omega_k}(T_M^\sigma u)_i(\omega) &= \sum_{j=1}^m \frac{\partial}{\partial \omega_j}(T_M^\sigma)_i(u(\omega))\,\frac{\partial}{\partial \omega_k}\,u_j(\omega) \\ &= m^\sigma\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\,\frac{\partial}{\partial \omega_k}u_i(\omega) \\ &\quad - (m^\sigma)'\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\,\frac{M}{|u(\omega)|^3_{\mathbb{R}^m}}\,u_i(\omega)\sum_{j=1}^m u_j(\omega)\,\frac{\partial}{\partial \omega_k}u_j(\omega),\end{aligned}$$

As an candidate for the derivative $\frac{\partial}{\partial \omega_k}(T_M u)_i$, we define

$$\begin{aligned}v_{ki}(\omega) &= m^0\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\,\frac{\partial}{\partial \omega_k}u_i(\omega) \\ &\quad - (m^0)'\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\,\frac{M}{|u(\omega)|^3_{\mathbb{R}^m}}\,u_i(\omega)\sum_{j=1}^m u_j(\omega)\,\frac{\partial}{\partial \omega_k}u_j(\omega),\end{aligned}$$

where the scalar functions m^0 and $(m^0)'$ are defined by

$$m^0(x) = \min\{1, x\}, \qquad (m^0)'(x) = \begin{cases} 1, & \text{if } x < 1, \\ 0, & \text{if } x \ge 1. \end{cases}$$

Since $m^\sigma(x) \to m^0(x)$ for all $x \in \mathbb{R}$, we get $T_M^\sigma u - T_M u \to 0$ a.e. in Ω. Moreover, this difference is dominated by

$$\begin{aligned}|T_M^\sigma u(\omega) - T_M u(\omega)|_{\mathbb{R}^m} &= \Big|m^\sigma\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\, u(\omega) - m^0\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\, u(\omega)\Big|_{\mathbb{R}^m} \\ &\le \Big|m^\sigma\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big) - m^0\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\Big|\, |u(\omega)|_{\mathbb{R}^m} \le \frac{\sigma}{2}\, |u(\omega)|_{\mathbb{R}^m},\end{aligned}$$

where we used $|m^\sigma(x) - m^0(x)| \le \sigma/2$ for all $\sigma \in (0,1)$ and $x \in \mathbb{R}$. By the dominated convergence theorem, we get $\|T_M^\sigma u(\omega) - T_M u(\omega)\|_{L^2(\Omega)^m} \to 0$.

Since $(m^\sigma)'(x) \to (m^0)'(x)$ for all $x \in \mathbb{R}$, we similarly get $\frac{\partial}{\partial \omega_k}(T_M^\sigma u)_i(\omega) - v_{ki}(\omega) \to 0$ for a.a. $\omega \in \Omega$. Note that

$$|(m^\sigma)'(x) - (m^0)'(x)|\, |x| \le 1$$

for all $\sigma \in (0,1)$ and $x \in \mathbb{R}$. Hence,

$$\begin{aligned}&\Big|\frac{\partial}{\partial \omega_k}(T_M^\sigma u)_i(\omega) - v_{ki}(\omega)\Big| \\ &\quad\le \Big|m^\sigma\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big) - m^0\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\Big|\, \Big|\frac{\partial}{\partial \omega_k} u_i(\omega)\Big| \\ &\qquad + \Big|(m^\sigma)'\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big) - (m^0)'\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\Big|\, \frac{M}{|u(\omega)|_{\mathbb{R}^m}^3}\, \Big|u_i(\omega) \sum_{j=1}^m u_j(\omega)\, \frac{\partial}{\partial \omega_k} u_j(\omega)\Big| \\ &\quad\le \frac{\sigma}{2}\, \Big|\frac{\partial}{\partial \omega_k} u_i(\omega)\Big| + \frac{1}{|u(\omega)|_{\mathbb{R}^m}^2}\, |u_i(\omega)| \sum_{j=1}^m |u_j(\omega)|\, \Big|\frac{\partial}{\partial \omega_k} u_j(\omega)\Big| \\ &\quad\le \frac{\sigma}{2}\, \Big|\frac{\partial}{\partial \omega_k} u_i(\omega)\Big| + \Big(\sum_{j=1}^m \Big|\frac{\partial}{\partial \omega_k} u_j(\omega)\Big|^2\Big)^{1/2}.\end{aligned}$$

Thus, we can apply the dominated convergence theorem and obtain

$$\Big\|\frac{\partial}{\partial \omega_k}(T_M^\sigma u)_i - v_{ki}\Big\|_{L^2(\Omega)} \to 0.$$

Finally, for $\varphi \in C_0^\infty(\Omega)$ we find

$$\int_\Omega \varphi\, v_{ki}\, \mathrm{d}\omega \leftarrow \int_\Omega \varphi\, \frac{\partial}{\partial \omega_k}(T_M^\sigma u)_i\, \mathrm{d}\omega = -\int_\Omega \frac{\partial}{\partial \omega_k}\varphi\, (T_M^\sigma u)_i\, \mathrm{d}\omega \to -\int_\Omega \frac{\partial}{\partial \omega_k}\varphi\, (T_M u)_i\, \mathrm{d}\omega.$$

Hence,

$$v_{ki} = \frac{\partial}{\partial \omega_k}(T_M u)_i,$$

and this shows the first part of the claim.

It remains to show $T_M u \to u$ in $H^1_0(\Omega)$. The convergence $T_M u \to u$ in $L^2(\Omega)$ is clear. Using

$$|m^0(x) - 1| \le 1, \quad \text{and} \quad |(m^0)'(x)\, x| \le 1,$$

for all $x \ge 0$, the difference of the derivatives can be bounded by

$$\begin{aligned}
&\Big|\frac{\partial}{\partial\omega_k}(T_M u)_i(\omega) - \frac{\partial}{\partial\omega_k}u_i(\omega)\Big| \\
&\quad = \Big|m^0\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big) - 1\Big|\,\Big|\frac{\partial}{\partial\omega_k}u_i(\omega)\Big| \\
&\qquad + \Big|(m^0)'\Big(\frac{M}{|u(\omega)|_{\mathbb{R}^m}}\Big)\Big|\,\frac{M}{|u(\omega)|^3_{\mathbb{R}^m}}\,\Big|u_i(\omega)\sum_{j=1}^m u_j(\omega)\,\frac{\partial}{\partial\omega_k}u_j(\omega)\Big| \\
&\quad \le \Big|\frac{\partial}{\partial\omega_k}u_i(\omega)\Big| + \frac{1}{|u(\omega)|^2_{\mathbb{R}^m}}\,\Big|u_i(\omega)\sum_{j=1}^m u_j(\omega)\,\frac{\partial}{\partial\omega_k}u_j(\omega)\Big| \\
&\quad \le \Big|\frac{\partial}{\partial\omega_k}u_i(\omega)\Big| + \Big(\sum_{j=1}^m\Big|\frac{\partial}{\partial\omega_k}u_j(\omega)\Big|^2\Big)^{1/2}
\end{aligned}$$

Hence, we can apply the dominated convergence theorem and obtain $T_M u \to u$ in $H^1_0(\Omega)^m$.

Remark 6.A.3. In the proof of Theorem 6.A.2, it is also possible to use different smooth approximations of T_M, e.g.,

$$\tilde m^\sigma(x) = \begin{cases} x, & \text{if } x < 1, \\ 1 + \frac{\sigma}{2} - \frac{1}{2\sigma}\,(1 + \sigma - x)^2, & \text{if } 1 \le x < 1 + \sigma, \\ 1 + \frac{\sigma}{2}, & \text{if } 1 + \sigma \le x. \end{cases}$$

Similarly, we obtain

$$|\tilde m^\sigma(x) - \tilde m^0(x)| \le \sigma/2 \qquad \text{and} \qquad |(\tilde m^\sigma)'(x) - (\tilde m^0)'(x)|\,|x| \le 1$$

for all $\sigma \in (0,1)$ and $x \in \mathbb{R}$ and the arguments of the proof carry over. However, there is the crucial difference

$$(m^\sigma)'(1) = 0 \ne 1 = (\tilde m^\sigma)'(1).$$

Thus, by using the approximation $\tilde m^\sigma$, the arguments of the proof of Theorem 6.A.2 lead to

$$\begin{aligned}
&\frac{\partial}{\partial\omega_k}(T_M\, u)_i(\omega) \\
&\quad = \begin{cases} \frac{\partial}{\partial\omega_k}u_i(\omega), & \text{if } |u(\omega)|_{\mathbb{R}^m} < M, \\ \frac{M}{|u(\omega)|_{\mathbb{R}^m}}\,\frac{\partial}{\partial\omega_k}u_i(\omega) + \frac{M}{|u(\omega)|^3_{\mathbb{R}^m}}\,u_i(\omega)\sum_{j=1}^m u_j(\omega)\,\frac{\partial}{\partial\omega_k}u_j(\omega), & \text{if } |u(\omega)|_{\mathbb{R}^m} \ge M \end{cases}
\end{aligned}$$

for a.a. $\omega \in \Omega$ for all $u \in H_0^1(\Omega)^m$. Together with the result of Theorem 6.A.2, this shows

$$\frac{1}{2}\,\nabla(|u|^2_{\mathbb{R}^m}) = \sum_{j=1}^{m} u_j\,\nabla u_j = 0 \qquad \text{a.e. on the set } \{\omega \in \Omega : |u(\omega)|_{\mathbb{R}^m} = M\}.$$

Note that, in case $m = 1$, this reduces to the well-known formula

$$\nabla u = 0 \qquad \text{a.e. on the set } \{\omega \in \Omega : |u(\omega)| = M\}$$

for $u \in H_0^1(\Omega)$.

6.B. Decomposition of measures in $H^{-1}(\Omega)$

In this section, we give a counterexample which shows that the positive part of a measure in $H^{-1}(\Omega)$, i.e. of an element of $H^{-1}(\Omega) \cap \mathcal{M}(\Omega)$, may not belong to $H^{-1}(\Omega)$.

Let $\Omega = U_1(0) \subset \mathbb{R}^2$ be the (open) unit ball. We denote by $\delta_{\hat{r}}$ the uniform line measure which is located at the radius $\hat{r} \in (0,1)$ and with total mass $2\,\pi$ (i.e., line density $1/\hat{r}$). Note that $\delta_{\hat{r}} \in H^{-1}(\Omega) \cap \mathcal{M}(\Omega)$ for $\hat{r} > 0$.

By $(-\Delta_0)^{-1}$ we denote the solution mapping associated with the Laplace equation with homogeneous Dirichlet boundary condition on Ω. It is easy to check that

$$v_{\hat{r}}(x,y) := (-\Delta_0)^{-1}(\delta_{\hat{r}})(x,y) = \begin{cases} \log(1/\hat{r}), & \text{if } r \le \hat{r}, \\ \log(1/r), & \text{if } r > \hat{r}. \end{cases}$$

Here and in the sequel, we use $r = \sqrt{x^2 + y^2}$. We find

$$\frac{\partial}{\partial r} v_{\hat{r}}(x,y) = \begin{cases} 0, & \text{if } r \le \hat{r}, \\ -1/r, & \text{if } r > \hat{r}, \end{cases}$$

and, thus,

$$\|v_{\hat{r}}\|^2_{H_0^1(\Omega)} = \int_\Omega |\nabla v|^2_{\mathbb{R}^2}\,\mathrm{d}(x,y) = \int_\Omega (\frac{\partial}{\partial r} v)^2\,\mathrm{d}(x,y) = 2\,\pi \int_{\hat{r}}^1 1/r\,\mathrm{d}r = 2\,\pi\,\log(1/\hat{r}).$$

Now, let $q \in (0,1)$ and a sequence $\{c_i\}_{i=1}^\infty \subset \mathbb{R}^+$ be given. We set $r_i = q^i$. We define a sequence $\{\mu_k\} \subset H^{-1}(\Omega) \cap \mathcal{M}(\Omega)$ by

$$\mu_k := \sum_{i=1}^{k} c_i\,(\delta_{r_{2\,i}} - \delta_{r_{2\,i-1}}),$$

where $\{c_i\}$ is a sequence of positive numbers. Since all line measures have mass $2\,\pi$, the sequence $\{\mu_k\}$ is a Cauchy sequence in $\mathcal{M}(\Omega)$ if and only if $\{c_i\}$ is summable.

In order to compute the $H^{-1}(\Omega)$-norm of μ_k, we set

$$v_k := (-\Delta_0)^{-1}\mu_k$$

and have

$$\|v_k\|_{H_0^1(\Omega)} = \|\mu_k\|_{H^{-1}(\Omega)}.$$

Since

$$\frac{\partial}{\partial r}v_k(x,y) = \begin{cases} -c_i/r, & \text{if } r_{2\,i} \le r \le r_{2\,i-1} \text{ with } i \in \{1,\dots,k\}, \\ 0, & \text{else,} \end{cases}$$

we find for $n \le k$

$$\begin{aligned} \|\mu_n - \mu_k\|^2_{H^{-1}(\Omega)} = \|v_n - v_k\|^2_{H_0^1(\Omega)} &= \int_\Omega |\nabla(v_n - v_k)|^2_{\mathbb{R}^2}\,\mathrm{d}(x,y) \\ &= \int_\Omega \big(\frac{\partial}{\partial r}(v_n - v_k)\big)^2\,\mathrm{d}(x,y) = 2\,\pi \sum_{i=n+1}^{k} c_i^2 \int_{r_{2\,i}}^{r_{2\,i-1}} \frac{1}{r}\,\mathrm{d}r \\ &= 2\,\pi \sum_{i=n+1}^{k} c_i^2 \,\log\Big(\frac{r_{2\,i-1}}{r_{2\,i}}\Big) = 2\,\pi\,\log(1/q) \sum_{i=n+1}^{k} c_i^2. \end{aligned}$$

Hence, the sequence $\{\mu_k\}$ is a Cauchy sequence in $H^{-1}(\Omega)$ if and only if $\{c_i\}$ is square summable.

In case $\{c_i\}$ is summable, the limits of μ_k in $H^{-1}(\Omega) = H_0^1(\Omega)^\star$ and $\mathcal{M}(\Omega)$ coincide, since $C_0^\infty(\Omega)$ is a dense subspace of $H_0^1(\Omega)$ and of $C_0(\Omega)$.

Now, we choose $c_i = i^p$ for some $-3/2 < p < -1$. Then c_i is summable and, thus, square summable. Hence, $\{\mu_k\}$ is a Cauchy sequence in $H^{-1}(\Omega)$ and $\mathcal{M}(\Omega)$ and we set

$$\mu := \lim_{k\to\infty} \mu_k = \sum_{i=1}^{\infty} c_i\,(\delta_{r_{2\,i}} - \delta_{r_{2\,i-1}}).$$

Since the mapping $\nu \mapsto \nu^+$ is continuous on $\mathcal{M}(\Omega)$, the positive part of μ is given by

$$\mu^+ = \sum_{i=1}^{\infty} c_i\,\delta_{r_{2\,i}}.$$

Now, we show that μ^+ is not bounded w.r.t. the $H_0^1(\Omega)$ norm on $C_0(\Omega) \cap H_0^1(\Omega)$.

Let $\varphi \in C_0^\infty(\Omega)$ with $0 \le \varphi \le 1$ and $\varphi \equiv 1$ on $B_q(0)$ be given. For $0 < s < 1/2$, the function

$$v(x,y) = \log(1/r)^s\,\varphi(x,y)$$

belongs to $H_0^1(\Omega)$.

For given $M > \log(1/q)^s$, we consider the truncation $v_M := \min\{v, M\}$ of v at M and have

$$v_M(x,y) = \begin{cases} M & r \le \exp(-M^{1/s}), \\ v(x,y) & r > \exp(-M^{1/s}). \end{cases}$$

Moreover, $v_M \to v$ in $H_0^1(\Omega)$ as $M \to \infty$.

But

$$\begin{aligned}\langle \mu^+, v_M \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} &= 2\,\pi \sum_{i=1}^{\infty} c_i\, v_M(r_{2\,i}) \\ &\geq 2\,\pi \sum_{i=1}^{n(M)} c_i \log(1/q^{2\,i})^s = 2\,\pi \log(1/q^2)^s \sum_{i=1}^{n(M)} i^{p+s},\end{aligned}$$

where $n(M) = \lfloor M^{1/s}/(2\log(1/q)) \rfloor$. Note that $n(M) \to \infty$ as $M \to \infty$ and, hence,

$$\langle \mu^+, v_M \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \geq 2\,\pi \log(1/q^2)^s \sum_{i=1}^{n(M)} i^{p+s} \to \infty$$

as $M \to \infty$ if $p + s \geq -1$. Note that for all $p \in (-3/2, -1)$, we can choose $s \in (0, 1/2)$ such that $p + s \geq -1$.

This shows that $\mu^+ \in \mathcal{M}(\Omega)$ is not bounded on $H_0^1(\Omega) \cap C_0(\Omega)$ w.r.t. the $H_0^1(\Omega)$-norm.

A similar reasoning shows that $v \notin L^1(\mu^+)$. Indeed, if v would belong to $L^1(\mu^+)$, v would be integrable and dominates v_M and, thus,

$$\infty > \int_\Omega v \,\mathrm{d}\mu^+ = \lim_{M\to\infty} \int_\Omega v_M \,\mathrm{d}\mu^+ = \infty.$$

Similarly, we can show $v \notin L^1(|\mu|)$.

Remark 6.B.1. We have constructed a counterexample in dimension $d = 2$. The construction can be adopted to dimensions $d > 2$.

In dimension $d = 1$ however, we have $H_0^1(\Omega) \hookrightarrow C_0(\Omega)$ and this embedding is continuous and dense. Hence, we obtain $\mathcal{M}(\Omega) = C_0(\Omega)^\star \hookrightarrow H_0^1(\Omega)^\star = H^{-1}(\Omega)$. Thus, the positive part of a measure belongs to $\mathcal{M}(\Omega)$ and, in turn, to $H^{-1}(\Omega)$. Therefore, it is not possible to construct a similar counterexample in dimension $d = 1$.

6.C. Lemmas on polyhedral sets satisfying LICQ

In this section, we provide some results for polyhedral sets. In the first lemma, we construct Lipschitz continuous functions satisfying (6.C.1). The existence of these functions is crucial in Section 6.5 to infer the polyhedricity of $\mathcal{C}$.

Lemma 6.C.1. We define the set

$$C := \{x \in \mathbb{R}^m : (x,\, n_i)_{\mathbb{R}^m} \leq b_i \;\forall i = 1, \dots, N\},$$

where $n_i \in \mathbb{R}^m$, $b_i \in \mathbb{R}$ are given for $i = 1, \dots, N$. Further, we assume that LICQ is satisfied, that is, the family $\{n_i : (x,\, n_i)_{\mathbb{R}^m} = b_i\}$ is linearly independent for all $x \in C$.

Then, there exist functions $h_i : C \to \mathbb{R}^m$, $i = 1, \dots, N$, which are globally bounded and Lipschitz continuous, and

$$(h_i(x),\, n_j)_{\mathbb{R}^m} = \delta_{ij} \qquad \forall i, j \in \{1, \dots, N\}, x \in F_j. \tag{6.C.1}$$

Here, F_j is the facet corresponding to inequality j, i.e., $F_j := \{x \in C : (x,\, n_j)_{\mathbb{R}^m} = b_j\}$, and δ_{ij} is the Kronecker delta.

Proof. Step I: Triangulation of C.

First, let L denote the orthogonal complement of the lineality space of C. Then, $C = L^\perp + (C \cap L)$ and all faces of $C \cap L$ have at least one vertex (i.e., extreme point), cf. Grünbaum, 1967, 2.5.6. Then, following Clarkson, 1987, p. 200, we find a triangulation $\mathcal{T} = \{S_k : k = 1, \dots, K\}$ of $C \cap L$, see also Clarkson, 1985. That is, $C \cap L = \bigcup_{k=1}^K S_k$ and each S_k is a generalized $\dim(L)$-simplex, i.e., $S_k = \operatorname{conv}\{v_l^k\}_{l=1}^{N_1(k)} + \operatorname{cone}\{r_l^k\}_{l=1}^{N_2(k)}$, such that $N_1(k) \geq 1$, $N_2(k) \geq 0$, $N_1(k) + N_2(k) = \dim(L) + 1$, all v_l^k are vertices of $C \cap L$ and all r_l^k are extremal rays of $C \cap L$, see Clarkson, 1985, Section 4 for details. Moreover, if $S_{k_1} \cap S_{k_2}$ is not empty, it is a common face of S_{k_1} and S_{k_2}.

Step II: Definition of h_i.

Let v be a vertex of $C \cap L$. By the linear independence assumption, we can find vectors $h_i(v) \in \mathbb{R}^m$, such that $(h_i(v),\, n_j)_{\mathbb{R}^m} = \delta_{ij}$ for all j satisfying $v \in F_j$. For any generalized simplex S_k, we extend h_i to $\operatorname{conv}\{v_l^k\}_{l=1}^{N_1(k)}$ linearly (note that each S_k has at least one vertex). Then, we extend h_i to S_k by $h_i(v + r) := h_i(v)$ for $v \in \operatorname{conv}\{v_l^k\}_{l=1}^{N_1(k)}$, $r \in \operatorname{cone}\{r_l^k\}_{l=1}^{N_2(k)}$. Let us check that h_i is well defined on $C \cap L$. If $x \in S_{k_1} \cap S_{k_2}$, this intersection is a common face of S_{k_1} and S_{k_2}. Hence, $S_{k_1} \cap S_{k_2}$ is the convex hull of some common vertices and some common extremal rays. Since h_i is well defined on the vertices of $C \cap L$, and extended linearly to S_{k_1} and S_{k_2}, both definitions of $h_i(x)$ coincide.

Since h_i is piecewise affine on $C \cap L$ and continuous, it is Lipschitz continuous. The boundedness of h_i follows since the range of h_i is contained in the convex hull of $\{h_i(v) : v \text{ is a vertex of } C \cap L\}$.

Finally, we set $h_i(x) := h_i(\hat{x})$ for $x \in C$, where $\hat{x} \in C \cap L$ and $x - \hat{x} \in L^\perp$. The Lipschitz continuity and the boundedness of h_i on C follows.

Step III: Verification of (6.C.1). Let $x \in F_j$ be given. By definition we have $h_i(x) = h_i(\hat{x})$, where $\hat{x} \in C \cap L$ and $x - \hat{x} \in L^\perp$. It is easy to check that $n_j \in L$ and, thus, we have $(\hat{x},\, n_j)_{\mathbb{R}^m} = (x,\, n_j)_{\mathbb{R}^m}$. This shows $\hat{x} \in F_j$. The point $\hat{x} \in C \cap L$ belongs to S_k for some $k \in \{1, \dots, K\}$. Hence, $h_i(\hat{x})$ is a convex combination of $\{h_i(v) : v \text{ is vertex of } S_k \cap F_j\}$. This shows $(h_i(x),\, n_j)_{\mathbb{R}^m} = (h_i(\hat{x}),\, n_j)_{\mathbb{R}^m} = \delta_{ij}$, since $(h_i(v),\, n_j)_{\mathbb{R}^m} = \delta_{ij}$ for all vertices v of $S_k \cap F_j$.

In the next lemma, we show that LICQ even holds for constraints which are almost active.

Lemma 6.C.2. Suppose that the assumptions of Lemma 6.C.1 are satisfied. Then, there exists $\hat{\delta} > 0$, such that the family $\{n_i : (x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}\}$ is linear independent for all $x \in C$.

Proof. As in the proof of Lemma 6.C.1, we denote by L the orthogonal complement of the lineality space of C and have

$$C = L^\perp + (C \cap L) = L^\perp + \operatorname{conv}\{v_j\}_{j=1,\dots,N_1} + \operatorname{cone}\{r_j\}_{j=1,\dots,N_2},$$

where v_j and r_j are the vertices (extreme points) and extreme rays of $C \cap L$, respectively, see Schneider, 2014, Corollary 1.4.4.

In a first step, we use a compactness argument to show that the assertion holds for all $x \in V := \operatorname{conv}\{v_j\}_{j=1,\dots,N_1}$. Let $x \in V$ be arbitrary. By assumption, the family $\{n_i : (x, n_i)_{\mathbb{R}^m} = b_i\}$ is linear independent. Since there are only finitely many inactive constraints, there is $\delta_x > 0$, such that for all $i = 1, \dots, N$ we have

$$(x, n_i)_{\mathbb{R}^m} = b_i \quad \Longleftrightarrow \quad (x, n_i)_{\mathbb{R}^m} \geq b_i - 2\,\delta_x,$$

and, hence, the family $\{n_i : (x, n_i)_{\mathbb{R}^m} \geq b_i - 2\,\delta_x\}$ is linear independent. By continuity of the scalar product, we find $\varepsilon_x > 0$, such that for all $i = 1, \dots, N$ and all $\tilde{x} \in U_{\varepsilon_x}(x) = \{y \in \mathbb{R}^m : |y - x|_{\mathbb{R}^m} < \varepsilon_x\}$ we have

$$(x, n_i)_{\mathbb{R}^m} = b_i \quad \Longleftrightarrow \quad (\tilde{x}, n_i)_{\mathbb{R}^m} \geq b_i - \delta_x.$$

Hence, the family $\{n_i : (\tilde{x}, n_i)_{\mathbb{R}^m} \geq b_i - \delta_x\}$ is linear independent for all $\tilde{x} \in U_{\varepsilon_x}(x)$. Now, $\{U_{\varepsilon_x}(x)\}_{x \in V}$ is an open cover of the compact set V and there exists a finite subcover. We denote by $\hat{\delta}$ the minimal δ_x corresponding to this subcover and obtain $\hat{\delta} > 0$. This shows that the family $\{n_i : (x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}\}$ is independent for all $x \in V$.

Now, an arbitrary point $x \in C$ can be written as $x = \ell + v + r$ with $\ell \in L^\perp$, $v \in V$ and $r \in \operatorname{cone}\{r_j\}_{j=1,\dots,N_2}$. Since $L^\perp$ is a subspace and $\operatorname{cone}\{r_j\}_{j=1,\dots,N_2}$ is a cone, we get $(\ell, n_i)_{\mathbb{R}^m} = 0$ and $(r, n_i)_{\mathbb{R}^m} \leq 0$ for all $i = 1, \dots, N$. Hence, $(x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}$ implies

$$b_i - \delta \leq (x, n_i)_{\mathbb{R}^m} = (\ell + v + r, n_i)_{\mathbb{R}^m} \leq (v, n_i)_{\mathbb{R}^m} \leq b_i.$$

This shows $\{n_i : (x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}\} \subset \{n_i : (v, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}\}$ and the latter family is linear independent by the first step of the proof.

Finally, we show that if the constraint i is almost active for a point $x \in C$, there exists a point in the neighborhood of x in which the constraint i is active.

Lemma 6.C.3. Suppose that the assumptions of Lemma 6.C.1 are satisfied and let $\hat{\delta} > 0$ be given from the previous lemma. Then, there is a constant $M > 0$, such that $x \in C$ and $(x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}$ implies the existence of $\tilde{x} \in C$ such that $(x, n_i)_{\mathbb{R}^m} = b_i$ and $|x - \tilde{x}|_{\mathbb{R}^m} \leq M\,(b_i - (x, n_i)_{\mathbb{R}^m})$.

Proof. For each subset $I \subset \{1, \dots, N\}$, for which $\{n_i\}_{i \in I}$ is linear independent, we choose vectors $\{p_j^I\}_{j \in I}$, such that

$$(p_j^I, n_i)_{\mathbb{R}^m} = \delta_{ij} \qquad \text{for all } i, j \in I.$$

We set $M = \max_{I, i \in I} |p_i^I|_{\mathbb{R}^m}$.

For $x \in C$, we set $I(x) = \{i \in \{1, \dots, N\} : (x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}\}$. We prove the claim by backward induction over the number of elements $\#I(x)$ in the set $I(x)$. The case $\#I(x) > m$ cannot appear since $\{n_i : i \in I(x)\}$ are linear independent vectors in $\mathbb{R}^m$.

Now, let $x \in C$ be given and suppose that the claim already holds for all $\tilde{x} \in C$ with $\#I(\tilde{x}) > \#I(x)$. Let $i \in \{1, \dots, N\}$ be given, such that $(x, n_i)_{\mathbb{R}^m} \geq b_i - \hat{\delta}$. Then, we have $i \in I(x)$ by definition. Moreover,

$$(x + t\, p_i^{I(x)}, n_j)_{\mathbb{R}^m} = \begin{cases} (x, n_j)_{\mathbb{R}^m} & \text{if } j \in I(x) \setminus \{i\}, \\ (x, n_i)_{\mathbb{R}^m} + t & \text{if } j = i. \end{cases}$$

If $x + T\, p_i^{I(x)} \in C$ for $T = b_i - (x, n_i)_{\mathbb{R}^m}$, we can use $\tilde{x} = x + T\, p_i^{I(x)}$. Otherwise, there is a smallest $t > 0$, such that $I(x + t\, p_i^{I(x)})$ is strictly larger than $I(x)$. By the induction hypothesis, we find $\tilde{x} \in C$, such that $(\tilde{x}, n_i)_{\mathbb{R}^m} = b_i$ and

$$|x + t\, p_i^{I(x)} - \tilde{x}|_{\mathbb{R}^m} \leq M\, (b_i - (x + t\, p_i^{I(x)}, n_i)_{\mathbb{R}^m}).$$

This implies

$$\begin{aligned} |x - \tilde{x}|_{\mathbb{R}^m} &\leq |x + t\, p_i^{I(x)} - \tilde{x}|_{\mathbb{R}^m} + t\, |p_i^{I(x)}|_{\mathbb{R}^m} \leq M\, (b_i - (x + t\, p_i^{I(x)}, n_i)_{\mathbb{R}^m}) + t\, M \\ &\leq M\, (b_i - (x, n_i)_{\mathbb{R}^m} - t + t) = M\, (b_i - (x, n_i)_{\mathbb{R}^m}). \end{aligned}$$

This shows that the claim holds for x and this finishes the induction step.

Acknowledgement

The author would like to thank Daniel Wachsmuth for the idea of the proof of Theorem 6.A.2. Moreover, the author appreciates the fruitful discussions with Frank Göring and Thomas Jahn, which led to the proof of Lemma 6.C.1. Finally, the associate editor drew the attention to the references Grun-Rehomme, 1977; Brézis, Browder, 1979 and this is gratefully acknowledged.

Theses

(1) It is possible to transfer the local decomposition approach for finite-dimensional MPCCs to the infinite-dimensional case. This leads to optimality conditions (cf. Section 1.5), which possess a reasonable strength in the polyhedric case.

(2) Using an additional linearization argument, this technique can also be used in the non-polyhedric case. This leads to new results for problems with second-order and semidefinite complementarity constraints, cf. Sections 2.5 and 2.6.

(3) In the general case, the system of strong stationarity is not a necessary optimality condition for the optimal control of the obstacle problem with control constraints. However, under some conditions on the data of the problem, we can still prove its necessity, Chapter 3.

(4) Currently, it is not clear whether M-stationarity is a necessary optimality condition for the optimal control of the obstacle problem with control constraints. We use a non-smooth regularization technique and, under a very mild assumption on the sequence of multipliers, we arrive at a system of M-stationarity, see Chapter 4. Otherwise, we show the necessity of a system of C-stationarity.

(5) The concept of polyhedricity is of importance in infinite-dimensional optimization theory. We generalize this concept to n-polyhedricity and prove that sets with lower and upper bounds in Banach spaces with a lattice structure are n-polyhedric for all $n \in \mathbb{N}$, cf. Theorem 5.4.18. Moreover, we provide counterexamples showing that the intersection of polyhedric sets may fail to be polyhedric.

(6) We demonstrate that sets of the form

$$\mathcal{C} := \{v \in H_0^1(\Omega)^m : v(\omega) \in C \text{ for almost all } \omega \in \Omega\}$$

are polyhedric in $H_0^1(\Omega)^m$ provided that $C \subset \mathbb{R}^m$ is a polyhedron with $0 \in \operatorname{int}(C)$ which satisfies LICQ, cf. Theorem 6.5.5.

Bibliography

Adams, D. R., Hedberg, L. I. (1996). *Function spaces and potential theory*. Vol. 314. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag. ISBN: 3-540-57060-8.

Aliprantis, C. D., Border, K. C. (2006). *Infinite dimensional analysis*. Third. A hitchhiker's guide. Springer, Berlin, pp. xxii+703. ISBN: 978-3-540-32696-0; 3-540-32696-0.

Alizadeh, F., Goldfarb, D. (2003). "Second-order cone programming". *Mathematical Programming. A Publication of the Mathematical Programming Society* 95.1, Ser. B. ISMP 2000, Part 3 (Atlanta, GA), pp. 3–51. ISSN: 0025-5610. DOI: 10.1007/s10107-002-0339-5.

Attouch, H., Buttazzo, G., Michaille, G. (2006). *Variational analysis in Sobolev and BV spaces*. Vol. 6. MPS/SIAM Series on Optimization. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xii+634. ISBN: 0-89871-600-4.

Aubin, J.-P., Frankowska, H. (2009). *Set-valued analysis*. Modern Birkhäuser Classics. Reprint of the 1990 edition. Boston, MA: Birkhäuser Boston Inc., pp. xx+461. ISBN: 978-0-8176-4847-3.

Barbu, V. (1984). *Optimal Control of Variational Inequalities*. Vol. 100. Research Notes in Mathematics. Boston: Pitman.

Benita, F., Dempe, S., Mehlitz, P. (2016). "Bilevel Optimal Control Problems with Pure State Constraints and Finite-dimensional Lower Level". *SIAM Journal on Optimization* 26.1, pp. 564–588. DOI: 10.1137/141000889.

Benita, F., Mehlitz, P. (2016). "Bilevel Optimal Control With Final-State-Dependent Finite-Dimensional Lower Level". *SIAM Journal on Optimization* 26.1, pp. 718–752. DOI: 10.1137/15M1015984.

Ben-Tal, A., Nemirovski, A. (1998). "Robust convex optimization". *Mathematics of Operations Research* 23.4, pp. 769–805. DOI: 10.1287/moor.23.4.769.

Ben-Tal, A., Nemirovski, A. (2002). "Robust optimization - methodology and applications". *Mathematical Programming* 92.3, Ser. B, pp. 453–480.

Bergounioux, M., Mignot, F. (2000). "Optimal Control of Obstacle Problems: Existence of Lagrange Multipliers". *ESAIM: Control, Optimisation and Calculus of Variations* 5, pp. 45–70. DOI: 10.1051/cocv:2000101.

Betz, T. (2015). "Optimal control of two variational inequalities arising in solid mechanics". PhD thesis. Technische Universität Dortmund. DOI: http://dx.doi.org/10.17877/DE290R-7694.

Bhatia, R. (1997). *Matrix analysis*. Vol. 169. Graduate Texts in Mathematics. New York: Springer-Verlag, pp. xii+347. ISBN: 0-387-94846-5. DOI: 10.1007/978-1-4612-0653-8.

Bonnans, J. F. (1998). "Second-order analysis for control constrained optimal control problems of semilinear elliptic systems". *Applied Mathematics and Optimization. An International Journal with Applications to Stochastics* 38.3, pp. 303–325. ISSN: 0095-4616. DOI: 10.1007/s002459900093.

Bonnans, J. F., Shapiro, A. (1998). "Optimization problems with perturbations: a guided tour". *SIAM Review* 40.2, pp. 228–264. ISSN: 0036-1445. DOI: 10.1137/S0036144596302644.

Bonnans, J. F., Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems*. Berlin: Springer.

Bonnans, J. F., Zidani, H. (1999). "Optimal control problems with partially polyhedric constraints". *SIAM Journal on Control and Optimization* 37.6, pp. 1726–1741. ISSN: 0363-0129. DOI: 10.1137/S0363012998333724.

Bourdaud, G., Meyer, Y. (1991). "Fonctions qui opèrent sur les espaces de Sobolev". *Journal of Functional Analysis* 97.2, pp. 351–360. ISSN: 0022-1236. DOI: 10.1016/0022-1236(91)90006-Q.

Brézis, H., Browder, F. (1979). "A property of Sobolev spaces". *Communications in Partial Differential Equations* 4.9, pp. 1077–1083. ISSN: 0360-5302. DOI: 10.1080/03605307908820120.

Cartan, H. (1967). *Calcul différentiel*. Paris: Hermann, p. 178.

Casado-Diaz, J., Dal Maso, G. (2000). "A weak notion of convergence in capacity with applications to thin obstacle problems". *Calculus of variations and differential equations (Haifa, 1998)*. Vol. 410. Chapman & Hall/CRC Res. Notes Math. Chapman & Hall/CRC, Boca Raton, FL, pp. 56–64.

Casas, E. (1986). "Control of an Elliptic Problem with Pointwise State Constraints". *SIAM Journal on Control and Optimization* 24.6, pp. 1309–1318. DOI: 10.1137/0324078.

Casas, E., Tröltzsch, F. (2012). "Second order analysis for optimal control problems: improving results expected from abstract theory". *SIAM Journal on Optimization* 22.1, pp. 261–279. ISSN: 1052-6234. DOI: 10.1137/110840406.

Christof, C., Meyer, C. (2015). *Differentiability properties of the solution operator to an elliptic variational inequality of the second kind*. Tech. rep. Ergebnisberichte des Instituts für Angewandte Mathematik, Nummer 527. Fakultät für Mathematik, TU Dortmund.

Cioranescu, D., Murat, F. (1997). "A strange term coming from nowhere". *Topics in the mathematical modelling of composite materials*. Vol. 31. Progr. Nonlinear Differential Equations Appl. Boston, MA: Birkhäuser Boston, pp. 45–93.

Clarkson, K. L. (1985). "A Probabilistic Algorithm for the Post Office Problem". *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*. STOC '85. Providence, Rhode Island, USA: ACM, pp. 175–184. DOI: 10.1145/22145.22165.

Clarkson, K. L. (1987). "New applications of random sampling in computational geometry". *Discrete & Computational Geometry. An International Journal of Mathematics and Computer Science* 2.2, pp. 195–222. ISSN: 0179-5376. DOI: 10.1007/BF02187879.

Dal Maso, G., Musina, R. (1989). "An approach to the thin obstacle problem for variational functionals depending on vector valued functions". *Communications in Partial Differential Equations* 14.12, pp. 1717–1743. ISSN: 0360-5302. DOI: 10.1080/03605308908820673.

De los Reyes, J. C., Meyer, C. (2016). "Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind". *Journal of Optimization Theory and Applications* 168.2, pp. 375–409. ISSN: 0022-3239. DOI: 10.1007/s10957-015-0748-2.

Delfour, M., Zolésio, J.-P. (2001). *Shapes and Geometries. Analysis, Differential Calculus, and Optimization*. Philadelphia: SIAM.

Dempe, S. (2002). *Foundations of bilevel programming*. Dordrecht: Kluwer Academic Publishers.

Diestel, J., Uhl, J. (1977). *Vector Measures*. Mathematical Surveys and Monographs. Providence: American Mathematical Society.

Ding, C., Sun, D., Ye, J. J. (2014). "First order optimality conditions for mathematical programs with semidefinite cone complementarity constraints". *Mathematical Programming. A Publication of the Mathematical Optimization Society* 147.1-2, Ser. A, pp. 539–579. ISSN: 0025-5610. DOI: 10.1007/s10107-013-0735-z.

Evans, L. C. (1990). *Weak convergence methods for nonlinear partial differential equations*. Vol. 74. CBMS Regional Conference Series in Mathematics. Published for the Conference Board of the Mathematical Sciences, Washington, DC, pp. viii+80. ISBN: 0-8218-0724-2.

Evans, L. C. (1998). *Partial Differential Equations*. Vol. 19. Graduate Studies in Mathematics. Providence, Rhode Island: American Mathematical Society.

Flegel, M. L., Kanzow, C. (2005a). "Abadie-type constraint qualification for mathematical programs with equilibrium constraints". *Journal of Optimization Theory and Applications* 124.3, pp. 595–614. ISSN: 0022-3239. DOI: 10.1007/s10957-004-1176-x.

Flegel, M. L., Kanzow, C. (2005b). "On the Guignard Constraint Qualification for Mathematical Programs with Equilibrium Constraints". *Optimization* 54.6, pp. 517–534. DOI: 10.1080/02331930500342591.

Flegel, M. L., Kanzow, C. (2006). "A direct proof for M-stationarity under MPEC-GCQ for mathematical programs with equilibrium constraints". *Optimization with multivalued mappings*. Vol. 2. Springer Optim. Appl. New York: Springer, pp. 111–122. DOI: 10.1007/0-387-34221-4_6.

Frémiot, G., Horn, W., Laurain, A., Rao, M., Sokołowski, J. (2009). "On the analysis of boundary value problems in nonsmooth domains". *Dissertationes Mathematicae (Rozprawy Matematyczne)* 462, p. 149. ISSN: 0012-3862. DOI: 10.4064/dm462-0-1.

Fukushima, M., Luo, Z.-Q., Tseng, P. (2002). "Smoothing functions for second-order-cone complementarity problems". *SIAM Journal on Optimization* 12.2, pp. 436–460. ISSN: 1052-6234. DOI: 10.1137/S1052623400380365.

Fukushima, M., Ōshima, Y., Takeda, M. (1994). *Dirichlet forms and symmetric Markov processes*. Vol. 19. de Gruyter Studies in Mathematics. Berlin: Walter de Gruyter & Co., pp. x+392. ISBN: 3-11-011626-X. DOI: 10.1515/9783110889741.

Grünbaum, B. (1967). *Convex polytopes*. With the cooperation of Victor Klee, M. A. Perles and G. C. Shephard. Pure and Applied Mathematics, Vol. 16. Interscience Publishers John Wiley & Sons, Inc., New York, pp. xiv+456.

Grun-Rehomme, M. (1977). "Caractérisation du sous-différentiel d'intégrandes convexes dans les espaces de Sobolev". *Journal de Mathématiques Pures et Appliquées. Neuvième Série* 56.2, pp. 149–156. ISSN: 0021-7824.

Haraux, A. (1977). "How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities". *Journal of the Mathematical Society of Japan* 29.4, pp. 615–631. ISSN: 0025-5645.

Heinonen, J., Kilpeläinen, T., Martio, O. (1993). *Nonlinear potential theory of degenerate elliptic equations*. Oxford Mathematical Monographs. Oxford Science Publications. New York: The Clarendon Press Oxford University Press. ISBN: 0-19-853669-0.

Herzog, R., Meyer, C., Wachsmuth, G. (2012). "C-Stationarity for Optimal Control of Static Plasticity with Linear Kinematic Hardening". *SIAM Journal on Control and Optimization* 50.5, pp. 3052–3082. DOI: 10.1137/100809325.

Herzog, R., Meyer, C., Wachsmuth, G. (2013). "B- and Strong Stationarity for Optimal Control of Static Plasticity with Hardening". *SIAM Journal on Optimization* 23.1, pp. 321–352. DOI: 10.1137/110821147.

Hildebrandt, S., Widman, K.-O. (1979). "Variational inequalities for vector-valued functions". *Journal für die Reine und Angewandte Mathematik* 309, pp. 191–220. ISSN: 0075-4102.

Hintermüller, M., Kopacka, I. (2009). "Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm". *SIAM Journal on Optimization* 20.2, pp. 868–902. ISSN: 1052-6234. DOI: 10.1137/080720681.

Hintermüller, M., Mordukhovich, B. S., Surowiec, T. (2014). "Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints". *Mathematical Programming* 146.1-2, Ser. A, pp. 555–582. ISSN: 0025-5610. DOI: 10.1007/s10107-013-0704-6.

Hintermüller, M., Surowiec, T. (2011). "First-order optimality conditions for elliptic mathematical programs with equilibrium constraints via variational analysis". *SIAM Journal on Optimization* 21.4, pp. 1561–1593. ISSN: 1052-6234. DOI: 10.1137/100802396.

Hiriart-Urruty, J.-B., Malick, J. (2012). "A fresh variational-analysis look at the positive semidefinite matrices world". *Journal of Optimization Theory and Applications* 153.3, pp. 551–577. ISSN: 0022-3239. DOI: 10.1007/s10957-011-9980-6.

Hoheisel, T., Kanzow, C., Schwartz, A. (2013). "Theoretical and Numerical Comparison of Relaxation Methods for Mathematical Programs with Complementarity Constraints". *Mathematical Programming* 137.1–2, pp. 257–288. DOI: 10.1007/s10107-011-0488-5.

Ito, K., Kunisch, K. (2000). "Optimal Control of Elliptic Variational Inequalities". *Applied Mathematics and Optimization* 41, pp. 343–364. DOI: 10.1007/s002459911017.

Jarušek, J., Outrata, J. V. (2007). "On sharp necessary optimality conditions in control of contact problems with strings". *Nonlinear Analysis* 67.4, pp. 1117–1128. DOI: 10.1016/j.na.2006.05.021.

Kalashnikov, V. V., Benita, F., Mehlitz, P. (2015). "The Natural Gas Cash-Out Problem: A Bilevel Optimal Control Approach". *Mathematical Problems in Engineering* 2015, Art. ID 286083, 17. ISSN: 1024-123X. DOI: 10.1155/2015/286083.

Kanzow, C., Schwartz, A. (2013). "A new regularization method for mathematical programs with complementarity constraints with strong convergence properties". *SIAM Journal on Optimization* 23.2, pp. 770–798. DOI: 10.1137/100802487.

Kato, T. (1995). *Perturbation theory for linear operators*. Classics in Mathematics. Reprint of the 1980 edition. Berlin: Springer-Verlag, pp. xxii+619. ISBN: 3-540-58661-X.

Kilpeläinen, T., Malý, J. (1992). "Supersolutions to degenerate elliptic equation on quasi open sets". *Communications in Partial Differential Equations* 17.3-4, pp. 371–405. ISSN: 0360-5302. DOI: 10.1080/03605309208820847.

Kinderlehrer, D., Stampacchia, G. (1980). *An Introduction to Variational Inequalities and Their Applications*. New York: Academic Press.

Klee, V. (1959). "Some characterizations of convex polyhedra". *Acta Mathematica* 102, pp. 79–107. ISSN: 0001-5962.

Krejčí, P. (1999). "Evolution variational inequalities and multidimensional hysteresis operators". *Nonlinear differential equations (Chvalatice, 1998)*. Vol. 404. Chapman & Hall/CRC Res. Notes Math. Chapman & Hall/CRC, Boca Raton, FL, pp. 47–110.

Krumbiegel, K., Rösch, A. (2009). "A virtual control concept for state constrained optimal control problems". *Computational Optimization and Applications* 43.2, pp. 213–233. DOI: 10.1007/s10589-007-9130-0.

Kunisch, K., Wachsmuth, D. (2012). "Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities". *ESAIM. Control, Optimisation and Calculus of Variations* 18.2, pp. 520–547. ISSN: 1292-8119. DOI: 10.1051/cocv/2011105.

Kurcyusz, S. (1976). "On the existence and non-existence of Lagrange multipliers in Banach spaces". *Journal of Optimization Theory and Applications* 20.1, pp. 81–110. ISSN: 0022-3239.

Levy, A. B. (1999). "Sensitivity of solutions to variational inequalities on Banach spaces". *SIAM Journal on Control and Optimization* 38.1, pp. 50–60. ISSN: 0363-0129. DOI: 10.1137/S036301299833985X.

Liang, Y.-C., Zhu, X.-D., Lin, G.-H. (2014). "Necessary Optimality Conditions for Mathematical Programs with Second-Order Cone Complementarity Constraints". *Set-Valued and Variational Analysis* 22.1, pp. 59–78. ISSN: 1877-0533. DOI: 10.1007/s11228-013-0250-7.

Luo, Z.-Q., Pang, J.-S., Ralph, D. (1996). *Mathematical Programs with Equilibrium Constraints*. Cambridge: Cambridge University Press.

Mancini, G., Musina, R. (1989). "Surfaces of minimal area enclosing a given body in $\mathbf{R}^3$". *Annali della Scuola Normale Superiore di Pisa. Classe di Scienze. Serie IV* 16.3, 331–354 (1990). ISSN: 0391-173X. URL: http://www.numdam.org/item?id=ASNSP_1989_4_16_3_331_0.

Marcus, M., Mizel, V. J. (1972). "Absolute continuity on tracks and mappings of Sobolev spaces". *Archive for Rational Mechanics and Analysis* 45, pp. 294–320. ISSN: 0003-9527. DOI: 10.1007/BF00251378.

Marcus, M., Mizel, V. J. (1973). "Nemitsky operators on Sobolev spaces". *Archive for Rational Mechanics and Analysis* 51, pp. 347–370. ISSN: 0003-9527. DOI: 10.1007/BF00263040.

Maurer, H., Zowe, J. (1979). "First and Second Order Necessary and Sufficient Optimality Conditions for Infinite-Dimensional Programming Problems". *Mathematical Programming* 16.1, pp. 98–110. DOI: 10.1007/BF01582096.

Meyer, C., Thoma, O. (2013). "A priori finite element error analysis for optimal control of the obstacle problem". *SIAM Journal on Numerical Analysis* 51.1, pp. 605–628. ISSN: 0036-1429. DOI: 10.1137/110836092.

Mignot, F. (1976). "Contrôle dans les inéquations variationelles elliptiques". *Journal of Functional Analysis* 22.2, pp. 130–185.

Mignot, F., Puel, J.-P. (1984). "Optimal control in some variational inequalities". *SIAM Journal on Control and Optimization* 22.3, pp. 466–476. DOI: 10.1137/0322028.

Mordukhovich, B. S., Outrata, J. V., Ramírez C., H. (2015a). "Graphical Derivatives and Stability Analysis for Parameterized Equilibria with Conic Constraints". *Set-Valued and Variational Analysis* 23.4, pp. 687–704. ISSN: 1877-0533. DOI: 10.1007/s11228-015-0328-5.

Mordukhovich, B. S., Outrata, J. V., Ramírez C., H. (2015b). "Second-order variational analysis in conic programming with applications to optimality and stability". *SIAM Journal on Optimization* 25.1, pp. 76–101. ISSN: 1052-6234. DOI: 10.1137/120903221.

Müller, G., Schiela, A. (2016). *On the Control of Time Discretized Dynamical Contact Problems*. Preprint. Universität Bayreuth. URL: https://epub.uni-bayreuth.de/2731/.

Oswald, P. (1992). "On the boundedness of the mapping $f \to |f|$ in Besov spaces". *Commentationes Mathematicae Universitatis Carolinae* 33.1, pp. 57–66. ISSN: 0010-2628.

Outrata, J. V. (1999). "Optimality conditions for a class of mathematical programs with equilibrium constraints". *Mathematics of Operations Research* 24.3, pp. 627–644. ISSN: 0364-765X. DOI: 10.1287/moor.24.3.627.

Outrata, J. V., Jarušek, J., Stará, J. (2011). "On optimality conditions in control of elliptic variational inequalities". *Set-Valued and Variational Analysis* 19.1, pp. 23–42. ISSN: 1877-0533. DOI: 10.1007/s11228-010-0158-4.

Outrata, J. V., Sun, D. (2008). "On the coderivative of the projection operator onto the second-order cone". *Set-Valued Analysis. An International Journal Devoted to the Theory of Multifunctions and its Applications* 16.7-8, pp. 999–1014. ISSN: 0927-6947. DOI: 10.1007/s11228-008-0092-x.

Pang, J.-S., Stewart, D. E. (2008). "Differential variational inequalities". *Mathematical Programming. A Publication of the Mathematical Programming Society* 113.2, Ser. A, pp. 345–424. ISSN: 0025-5610. DOI: 10.1007/s10107-006-0052-x.

Pang, J.-S., Fukushima, M. (1999). "Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with equilibrium constraints". *Computational Optimization and Applications. An International Journal* 13.1-3. Computational optimization—a tribute to Olvi Mangasarian, Part II, pp. 111–136. ISSN: 0926-6003. DOI: 10.1023/A:1008656806889.

Rao, M., Sokołowski, J. (1993). "Polyhedricity of convex sets in Sobolev space $H_0^2(\Omega)$". *Nagoya Mathematical Journal* 130, pp. 101–110. ISSN: 0027-7630. URL: http://projecteuclid.org/euclid.nmj/1118779582.

Robinson, S. M. (1976). "Stability theory for systems of inequalities. II. Differentiable nonlinear systems". *SIAM Journal on Numerical Analysis* 13.4, pp. 497–513. ISSN: 0036-1429.

Rudin, W. (1987). *Real and Complex Analysis*. McGraw–Hill.

Runst, T., Sickel, W. (1996). *Sobolev spaces of fractional order, Nemytskij operators, and nonlinear partial differential equations*. Vol. 3. de Gruyter Series in Nonlinear Analysis and Applications. Walter de Gruyter & Co., Berlin, pp. x+547. ISBN: 3-11-015113-8. DOI: 10.1515/9783110812411.

Rychkov, V. S. (1999). "On restrictions and extensions of the Besov and Triebel-Lizorkin spaces with respect to Lipschitz domains". *Journal of the London Mathematical Society. Second Series* 60.1, pp. 237–257. ISSN: 0024-6107. DOI: 10.1112/S0024610799007723.

Schaefer, H. H. (1974). *Banach lattices and positive operators*. Die Grundlehren der mathematischen Wissenschaften, Band 215. Springer-Verlag, New York-Heidelberg, pp. xi+376.

Scheel, H., Scholtes, S. (2000). "Mathematical Programs with Complementarity Constraints: Stationarity, Optimality, and Sensitivity". *Mathematics of Operations Research* 25.1, pp. 1–22. DOI: 10.1287/moor.25.1.1.15213.

Schiela, A. (2009). "State constrained optimal control problems with states of low regularity". *SIAM Journal on Control and Optimization* 48.4, pp. 2407–2432. ISSN: 0363-0129. DOI: 10.1137/080727610.

Schiela, A., Wachsmuth, D. (2013). "Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints". *ESAIM: Mathematical Modelling and Numerical Analysis* 47.3, pp. 771–787. DOI: 10.1051/m2an/2012049.

Schneider, R. (2014). *Convex bodies: the Brunn-Minkowski theory*. expanded. Vol. 151. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, pp. xxii+736. ISBN: 978-1-107-60101-7.

Shapiro, A. (1990). "On Concepts of Directional Differentiability". *Journal of Optimization Theory and Applications* 66.3, pp. 477–487.

Shapiro, A. (1997a). "On uniqueness of Lagrange multipliers in optimization problems subject to cone constraints". *SIAM Journal on Optimization* 7.2, pp. 508–518. ISSN: 1052-6234. DOI: 10.1137/S1052623495279785.

Shapiro, A. (1997b). "First and second order analysis of nonlinear semidefinite programs". *Mathematical Programming* 77.2, Ser. B. Semidefinite programming, pp. 301–320. ISSN: 0025-5610. DOI: 10.1016/S0025-5610(96)00081-0.

Sokołowski, J., Zolésio, J.-P. (1992). *Introduction to Shape Optimization*. New York: Springer.

Sokołowski, J. (1988). "Sensitivity analysis of contact problems with prescribed friction". *Applied Mathematics and Optimization. An International Journal with Applications to Stochastics* 18.2, pp. 99–117. ISSN: 0095-4616. DOI: 10.1007/BF01443617.

Sokołowski, J., Zolésio, J.-P. (1987). "Shape sensitivity analysis of unilateral problems". *SIAM Journal on Mathematical Analysis* 18.5, pp. 1416–1437. ISSN: 0036-1410. DOI: 10.1137/0518103.

Stampacchia, G. (1964). "Formes bilinéaires coercitives sur les ensembles convexes". *C. R. Acad. Sci. Paris* 258, pp. 4413–4416.

Stollmann, P. (1993). "Closed ideals in Dirichlet spaces". *Potential Analysis. An International Journal Devoted to the Interactions between Potential Theory, Probability Theory, Geometry and Functional Analysis* 2.3, pp. 263–268. ISSN: 0926-2601. DOI: 10.1007/BF01048510.

Sun, D., Sun, J. (2002). "Semismooth matrix-valued functions". *Mathematics of Operations Research* 27.1, pp. 150–169. ISSN: 0364-765X. DOI: 10.1287/moor.27.1.150.342.

Todd, M. J. (2001). "Semidefinite optimization". *Acta Numerica* 10, pp. 515–560. ISSN: 0962-4929. DOI: 10.1017/S0962492901000071.

Vandenberghe, L., Boyd, S. (1996). "Semidefinite programming". *SIAM review* 38.1, pp. 49–95.

Wachsmuth, G. (2013). "On LICQ and the uniqueness of Lagrange multipliers". *Operations Research Letters* 41.1, pp. 78–80. DOI: 10.1016/j.orl.2012.11.009.

Wachsmuth, G. (2014). "Strong Stationarity for Optimal Control of the Obstacle Problem with Control Constraints". *SIAM Journal on Optimization* 24.4, pp. 1914–1932. DOI: 10.1137/130925827.

Wachsmuth, G. (2015). "Mathematical Programs with Complementarity Constraints in Banach Spaces". *Journal of Optimization Theory and Applications* 166.2, pp. 480–507. ISSN: 0022-3239. DOI: 10.1007/s10957-014-0695-3.

Wachsmuth, G. (2016a). *A guided tour of polyhedric sets*. Preprint. TU Chemnitz.

Wachsmuth, G. (2016b). "Pointwise Constraints in Vector-Valued Sobolev Spaces. With Applications in Optimal Control". *Applied Mathematics & Optimization*. To appear, pp. 1–35. DOI: 10.1007/s00245-016-9381-1.

Wachsmuth, G. (2016c). "Towards M-stationarity for optimal control of the obstacle problem with control constraints". *SIAM Journal on Control and Optimization* 54.2, pp. 964–986. DOI: 10.1137/140980582.

Wachsmuth, G. (2017). "Strong stationarity for optimization problems with complementarity constraints in absence of polyhedricity". *Set-Valued and Variational Analysis* 25.1, pp. 133–175. DOI: 10.1007/s11228-016-0370-y.

Werner, J. (1984). *Optimization theory and applications*. Advanced Lectures in Mathematics. Braunschweig: Friedr. Vieweg & Sohn, pp. vii+233. ISBN: 3-528-08594-0.

Wu, J., Zhang, L., Zhang, Y. (2014). "Mathematical Programs with Semidefinite Cone Complementarity Constraints: Constraint Qualifications and Optimality Conditions". *Set-Valued and Variational Analysis* 22.1, pp. 155–187. ISSN: 1877-0533. DOI: 10.1007/s11228-013-0242-7.

Yan, T., Fukushima, M. (2011). "Smoothing method for mathematical programs with symmetric cone complementarity constraints". *Optimization. A Journal of Mathematical Programming and Operations Research* 60.1-2, pp. 113–128. ISSN: 0233-1934. DOI: 10.1080/02331934.2010.541458.

Yang, C.-Y., Chang, Y.-L., Chen, J.-S. (2011). "Analysis of nonsmooth vector-valued functions associated with infinite-dimensional second-order cones". *Nonlinear Analysis. Theory, Methods & Applications. An International Multidisciplinary Journal. Series A: Theory and Methods* 74.16, pp. 5766–5783. ISSN: 0362-546X. DOI: 10.1016/j.na.2011.05.068.

Ye, J. J., Zhou, J. (2015). *First order optimality conditions for mathematical programs with second-order cone complementarity constraints*. Tech. rep. URL: http://www.optimization-online.org/DB_HTML/2015/04/4888.html.

Ziegler, G. M. (1995). *Lectures on polytopes*. Vol. 152. Graduate Texts in Mathematics. Springer-Verlag, New York, pp. x+370. ISBN: 0-387-94365-X. DOI: 10.1007/978-1-4613-8431-1.

Ziemer, W. P. (1989). *Weakly differentiable functions*. Vol. 120. Graduate Texts in Mathematics. Sobolev spaces and functions of bounded variation. New York: Springer-Verlag, pp. xvi+308. ISBN: 0-387-97017-7. DOI: 10.1007/978-1-4612-1015-3.

Zowe, J., Kurcyusz, S. (1979). "Regularity and stability for the mathematical programming problem in Banach spaces". *Applied Mathematics and Optimization* 5.1, pp. 49–62.