On bilevel optimization problems in infinite-dimensional spaces

Von der Fakultät 1 – MINT – Mathematik, Informatik, Physik, Elektro- und Informationstechnik der Brandenburgischen Technischen Universität Cottbus-Senftenberg genehmigte Dissertation
zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

vorgelegt von

Felix Harder

geboren am 4. Mai 1993 in Kassel

Vorsitzende:	Prof. Dr. Sabine Pickenhain
Gutachter:	Prof. Dr. Gerd Wachsmuth
Gutachter:	Prof. Dr. Christian Meyer
Tag der mündlichen Prüfung:	17.12.2020

DOI: https://doi.org/10.26127/BTUOpen-5375

Abstract

In this thesis we consider bilevel optimization problems in infinite-dimensional spaces. In particular, we are interested in providing first-order necessary optimality conditions. We consider bilevel optimization problems both in abstract Banach spaces and in some special situations. This includes the optimal control of the obstacle problem, which is a typical bilevel optimization problem in Sobolev spaces, as well as a class of inverse optimal control problems. We obtain optimality conditions for these more specific optimization problems by applying our results from the abstract setting.

Our main approach for deriving optimality conditions in the abstract setting utilizes the relaxation of a reformulation of the bilevel optimization problem via the optimal value function. We also introduce the so-called normal-cone-preserving operators and show how this concept can be applied.

We also consider other topics that arise in this context. For instance, we investigate the so-called limiting normal cone to a complementarity set in Sobolev spaces. This complementarity set plays a central role in the context of the optimal control of the obstacle problem. The limiting normal cone is a concept which appears in the area of variational analysis and generalizes the usual normal cone from convex analysis. We also investigate in which spaces Legendre forms and Legendre-$\star$ forms can exist. We show that if a Legendre-$\star$ form exists in a reflexive Banach space or a space with a separable predual space, then this space is already isomorphic to a Hilbert space. We also consider a discretization of a bilevel optimization problem in Lebesgue spaces. We present both theoretical error estimates and numerical experiments.

The new results in this thesis are illustrated by examples and counterexamples. In order to present the topics in a self-contained way, we review some known concepts and their basic properties. A particular focus for this is on the definitions and properties from the area of capacity theory.

Zusammenfassung

In dieser Arbeit werden Zwei-Ebenen-Optimierungsprobleme in unendlichdimensionalen Räumen betrachtet. Insbesondere sind wir an notwendigen Optimalitätsbedingungen erster Ordnung für solche Optimierungsprobleme interessiert. Wir studieren Zwei-Ebenen-Optimierungsprobleme sowohl in abstrakten Banachräumen als auch in konkreteren Situationen. So betrachten wir etwa die optimale Steuerung des Hindernisproblems, welches ein typisches Zwei-Ebenen-Optimierungsproblem in Sobolevräumen ist, sowie auch eine Klasse von inversen Optimalsteuerungsproblemen. Dabei wenden wir unsere Ergebnisse aus den Betrachtungen in abstrakten Banachräumen an, um Optimalitätsbedingungen zu zeigen.

Der Ansatz, um Optimalitätsbedingungen in der abstrakten Situation herzuleiten, benutzt die Relaxierung einer Reformulierung des Zwei-Ebenen-Optimierungsproblems, welche auf der Optimalwertfunktion basiert. Dabei führen wir auch die sogenannten normalkegelerhaltenden Operatoren ein und zeigen, wie dieses Konzept zur Anwendung kommt.

Wir gehen auch auf andere Themen, die in diesem Zusammenhang auftreten, ein. So wird der sogenannte „limiting normal cone“ zu einer Komplementaritätsmenge in Sobolevräumen betrachtet. Diese Komplementaritätsmenge spielt eine zentrale Rolle für die optimale Steuerung des Hindernisproblems. Der „limiting normal cone“ ist ein Konzept, welches in der variationellen Analysis vorkommt und den gewöhnlichen Normalkegel aus der konvexen Analysis verallgemeinert. Ebenfalls untersucht wird die Fragestellung, in welchen Räumen Legendre-Formen und Legendre-$\star$-Formen existieren können. Es wird gezeigt, dass die Existenz von Legendre-$\star$-Formen in einem reflexiven Banachraum oder einem Raum mit einem separablen Prädualraum impliziert, dass diese Räume bereits isomorph zu einem Hilbertraum sind. Wir betrachten auch eine Diskretisierung eines Zwei-Ebenen-Optimierungsproblems in Lebesgueräumen. Dabei werden sowohl Fehlerschätzungen gezeigt als auch numerische Experimente präsentiert.

Die erhaltenen neuen Resultate in dieser Arbeit werden mit Beispielen und Gegenbeispielen bereichert. Damit die Themen in einer in sich abgeschlossenen Form präsentiert werden können, werden auch bereits bekannte Konzepte eingeführt und grundlegende Eigenschaften diskutiert. Ein besonderer Schwerpunkt wird dabei auf die Kapazitätstheorie gelegt.

Acknowledgments

First, I would like to thank my supervisor Prof. Gerd Wachsmuth for giving me the opportunity to be his PhD student, our mathematical collaboration, and his overall guidance during my PhD time. During my studies, I was supported by the DFG (grants WA 3636/4-1 and WA 3636/4-2) within the priority program SPP 1962 (Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization), for which I am grateful.

I would like to express my gratitude towards Prof. Christian Meyer for his willingness to be a co-referee of this thesis. A special thanks goes also to my working groups and colleagues in Chemnitz and Cottbus for the pleasant working atmosphere. Moreover, I would like to thank Patrick, Helen, Paula, Michael, and Rosa for proofreading parts of this thesis.

Last, but not least, I want to thank my parents and my sister for their help and support during the last years.

Contents

1 Introduction

This thesis deals with first-order necessary optimality conditions for bilevel optimization problems in infinite-dimensional spaces. Bilevel optimization problems are optimization problems that consist of two levels. The so-called lower level optimization problem is given by the parameter-dependent optimization problem

$$\begin{aligned} \min_{x} \quad & f(x,p) \\ \text{s.t.} \quad & g(x,p) \in \Phi, \end{aligned} \tag{LL(p)}$$

where $p \in V$ is a parameter, $f : X \times V \to \mathbb{R}$, $g : X \times V \to Y$ are functions, $\Phi \subset Y$ is a closed set, and X, V, Y are Banach spaces. The so-called upper level optimization problem is given by

$$\begin{aligned} \min_{x,p} \quad & F(x,p) \\ \text{s.t.} \quad & x \text{ solves (LL}(p)\text{)}, \\ & p \in \Phi_{UL}, \end{aligned} \tag{UL}$$

where $F : X \times V \to \mathbb{R}$ is the objective function and $\Phi_{UL} \subset V$ is a closed set. Since the upper level optimization problem includes the lower level optimization problem in its constraints, we also call the upper level optimization problem a bilevel optimization problem. In general, bilevel optimization problems are nonsmooth and nonconvex optimization problems.

One possible interpretation for bilevel optimization problems involves two players, namely leader and follower. First, the leader chooses a point $p \in \Phi_{UL}$. Then the follower chooses a point $x \in X$ such that $g(x,p) \in \Phi$ with the goal of minimizing his objective function $f(\cdot, p)$. The leader wants to minimize his objective function F, and since this function can depend on the choice x of the follower, the leader should therefore take the actions of the follower into account. We mention that (UL) is equivalent with respect to global minimizers to the so-called optimistic formulation of a bilevel optimization problem, see also Section 3.3.1. This means that if the lower level optimization problem produces multiple solutions x, the leader can choose freely among those solutions.

Bilevel optimization problems in finite-dimensional spaces have been studied extensively in the literature. We exemplarily refer to the book [Dempe, 2002], where many aspects of bilevel optimization problems in finite dimensions are discussed, including stationarity conditions, numerical algorithms, and applications. An overview over the literature of bilevel optimization problems can be found in [Dempe, 2018].

If we replace the lower level optimization problem with the corresponding system of Karush-Kuhn-Tucker (KKT) conditions (which is not always an equivalent reformulation), we obtain a so-called mathematical program with complementarity constraints, or MPCC for short. This is another class of problems that has been studied for several decades, see, for example, the monographs [Luo, Pang, Ralph, 1996; Outrata, Kočvara, Zowe, 1998].

Bilevel optimization problems and MPCCs have also been studied in infinite-dimensional spaces. An abstract setting in general Banach spaces was used in [G. Wachsmuth, 2015; Mehlitz, G. Wachsmuth, 2016; Mehlitz, 2017]. A well-known instance of an infinite-dimensional bilevel optimization problem is the optimal control of the obstacle problem, which is a bilevel optimization problem in Sobolev spaces. Here, the lower level optimization problem is the so-called obstacle problem, where one wants to minimize the elastic energy of a thin membrane that is subject to an external force and constrained by an obstacle, see (OP(u)) on page 159. In particular, obtaining necessary optimality conditions for local minimizers of the optimal control of the obstacle problem received great interest, see, for example, [Mignot, 1976; Jarušek, Outrata, 2007; Schiela, D. Wachsmuth, 2013; G. Wachsmuth, 2016; Harder, G. Wachsmuth, 2018a]. The optimal control of the obstacle problem can also be interpreted as a special instance of optimal control problems of variational inequalities.

Another class of infinite-dimensional bilevel optimization problems are so-called inverse optimal control problems. Here, the lower level optimization problem is an optimal control problem, and the upper level optimization problem is used to identify parameters of the lower level optimization problem. Usually, these parameters appear in the objective function of the lower level optimization problem. For inverse optimal control problems with ordinary differential equations as constraints in the lower level optimization problem, see [Mombaur, Truong, Laumond, 2010; Albrecht, Leibold, Ulbrich, 2012; Hatz, Schlöder, Bock, 2012; Albrecht, Ulbrich, 2017]. Similarly, in [Holler, Kunisch, Barnard, 2018] a bilevel formulation was used to identify parameters of an inverse problem.

In this thesis, we study infinite-dimensional bilevel optimization problems in abstract Banach spaces and also consider the optimal control of the obstacle problem as well as a class of inverse optimal control problems, which are more specific bilevel optimization problems in Sobolev and Lebesgue spaces. In particular, the focus is on first-order necessary optimality conditions. Our main approach for deriving optimality conditions in abstract Banach spaces will be the relaxation of the optimal value reformulation. The optimal value reformulation is a reformulation of the bilevel optimization problem which utilizes the so-called optimal value function.

We are interested in situations, where some of the analytical difficulties that are associated with bilevel optimization problems in infinite-dimensional spaces such as nonconvexity nonsmoothness are possible, but the problems are still sufficiently well-behaved so that we are able to study analytical properties of the bilevel optimization problems in a rigorous way. Thus, we require some assumptions for our bilevel optimization problems in an abstract setting, such as strong convexity of the lower level objective function. For our examples in Lebesgue and Sobolev spaces we consider bilevel optimization problems that

satisfy these assumptions. A benefit of our approach in abstract Banach spaces is that we also gain some knowledge on what assumptions are relevant for our approach for deriving optimality conditions.

While in some bilevel optimization problems strong stationarity can be shown for local minimizers, the bilevel optimization problems that we consider in this thesis are complicated and difficult enough such that strong stationarity can usually not be obtained. Instead, our results lead to stationarity conditions of C-stationarity type for local minimizers.

1.1 An illustrating example

Let us give a more specific example for an infinite-dimensional bilevel optimization problem, which is also of the class of inverse optimal control problems. Suppose that $\Omega \subset \mathbb{R}^3$ is an open and bounded set which describes a solid body. Let $y(\omega)$ denote the temperature at the point $\omega \in \Omega$. The temperature is being controlled by a distributed heat source $u : \Omega \to \mathbb{R}$, which is restricted by control constraints $u_a, u_b : \Omega \to \mathbb{R}$ (with $u_a, u_b \in L^2(\Omega)$) in the sense that $u_a(\omega) \leq u(\omega) \leq u_b(\omega)$ holds for almost all $\omega \in \Omega$. Further, let $y_a \in \mathbb{R}$ be the ambient temperature and $c : \partial\Omega \to [0, \infty)$ a heat transfer coefficient with $c \in L^\infty(\partial\Omega)$ and $\|c\|_{L^\infty(\partial\Omega)} > 0$. Suppose that a person wants to control u in such a way that y is as close as possible to a temperature $\alpha_1 \in \mathbb{R}$ on Ω_1 and as close as possible to a temperature $\alpha_2 \in \mathbb{R}$ on Ω_2, where $\Omega_1, \Omega_2 \subset \Omega$ are disjoint measurable subsets of Ω. Additionally, the person wants to avoid excessive control costs and thus keep the term $\|u\|^2_{L^2(\Omega)}$ small. If we consider the vector of desired temperatures $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$ as a parameter, then the corresponding (parameter-dependent) optimal control problem can be modeled by

$$\begin{aligned} \min_{y \in H^1(\Omega),\, u \in L^2(\Omega)} \quad & \tfrac{1}{2} \int_{\Omega_1 \cup \Omega_2} \Big(y(\omega) - \big(\alpha_1 \chi_{\Omega_1}(\omega) + \alpha_2 \chi_{\Omega_2}(\omega)\big) \Big)^2 \,\mathrm{d}\omega + \tfrac{\sigma}{2} \|u\|^2_{L^2(\Omega)} \\ \text{s.t.} \quad & -\Delta y = u && \text{on } \Omega, \\ & \frac{\partial y}{\partial n} + c\,(y - y_a) = 0 && \text{on } \partial\Omega, \\ & u_a \leq u \leq u_b && \text{a.e. on } \Omega, \end{aligned} \tag{OC(α)}$$

where $\sigma > 0$ is a constant and $\chi_{\Omega_i}(\omega)$ denotes the characteristic function of Ω_i for $i = 1, 2$, i.e. $\chi_{\Omega_i}(\omega) = 1$ holds if $\omega \in \Omega_i$ and $\chi_{\Omega_i}(\omega) = 0$ holds if $\omega \notin \Omega_i$. Now, we imagine that we observe a (possibly perturbed) measurement $y_m \in L^2(\Omega)$ of the temperature $\hat{y} \in H^1(\Omega)$ that is the result of (OC($\hat{\alpha}$)), where the real parameter vector $\hat{\alpha} \in \mathbb{R}^2$ is not known to us. Our goal is to identify the parameter vector of desired temperatures $\hat{\alpha}$ from the measurement y_m. This can be done using the optimization problem

$$\begin{aligned} \min_{y \in H^1(\Omega),\, u \in L^2(\Omega), \alpha \in \mathbb{R}^2} \quad & \tfrac{1}{2} \int_\Omega (y(\omega) - y_m(\omega))^2 \,\mathrm{d}\omega \\ \text{s.t.} \quad & (y, u) \text{ solves (OC}(\alpha)\text{)}. \end{aligned} \tag{IOC}$$

This problem is now a bilevel optimization problem and since we use it to identify parameters of an optimal control problem we call it an inverse optimal control problem. Using a substitution, we can (without loss of generality) assume that $y_a = 0$. If Ω is also a sufficiently regular domain, this bilevel optimization problem fits into the setting used in Chapter 6 and satisfies Assumption 6.1.1, see also (6.5), (6.6), and Corollary 6.1.3. Moreover, this example can also be extended in various ways.

1.2 Outline of this thesis

We give a short overview of the structure and content of this thesis.

In Chapter 2 we establish notation, provide definitions, give some preliminary technical and elementary results, and review some concepts that are already known in the literature. This concerns the topics of functional analysis, convex analysis, quadratic forms, strong convexity, coercivity, function spaces such as Lebesgue and Sobolev spaces, and infinite-dimensional optimization. In order to provide better context for our stationarity conditions in Chapters 5 and 6, we provide Sections 2.4 and 2.5, which briefly discuss some concepts of variational analysis and stationarity concepts for MPCCs in finite-dimensional spaces.

A particular emphasis has been placed on the study of capacity theory in Section 2.6. Although most of the results are already known in some form in the literature, we provide a thorough and self-contained introduction to this topic which only requires some basic knowledge of Sobolev spaces. Proofs are included for all results. We also introduce the so-called quasi-support, based on the approach with closed lattice ideals that was used in [Harder, G. Wachsmuth, 2018a]. Capacity theory plays an important role for our studies in Chapter 5.

In Chapter 3 we study various aspects of the optimization theory that arises in the context of bilevel optimization problems in an abstract setting. We start with some theory for parameter-dependent optimization problems in Section 3.1. This includes some new counterexamples to illustrate the lack of continuity of the solution operator in some situations. Our study of bilevel optimization problems makes use of the so-called optimal value function, and we discuss its properties in Section 3.2. Some general remarks on bilevel optimization problems and its reformulations can be found in Section 3.3. Next, Section 3.4 is dedicated to obtaining stationarity conditions for bilevel optimization problems in an abstract setting. Our approach uses a relaxation of the so-called optimal value reformulation. We also introduce and use the novel concept of normal-cone-preserving operators.

In the study of differentiability properties of solution operators, the concept of Legendre forms (or Legendre-$\star$ forms) appears. These also play a role in the literature in the context of second-order sufficient optimality conditions, see [Bonnans, Shapiro, 2000]. In order to assess in which spaces this concept is useful, we discuss Legendre forms and

Legendre-$\star$ forms in Chapter 4. This chapter is partially based on [Harder, 2018], but generalizes the findings to nonreflexive Banach spaces.

Chapter 5 deals with the optimal control of the obstacle problem, which is a bilevel optimization problem in Sobolev spaces. The concepts of weak, C-, M-, and strong stationarity are defined for this optimization problem. We show that the abstract theory of Chapter 3 can be applied to this problem. This results in C-stationarity for local minimizers of the optimal control of the obstacle problem. In variational analysis, the concept of the limiting normal cone is sometimes used to obtain stationarity conditions for nonsmooth and nonconvex optimization problems. Thus, we investigate the limiting normal cone to a complementarity set in Sobolev spaces that arises from the study of the optimal control of the obstacle problem in Section 5.3. This section is based on the article [Harder, G. Wachsmuth, 2018c]. We provide lower estimates to the limiting normal cone which show that it is unfortunately rather large. This also indicates that the limiting normal cone is not as useful in infinite-dimensional spaces as in finite-dimensional spaces.

In Chapter 6 we consider another instance of an infinite-dimensional bilevel optimization problem, which we call an inverse optimal control problem. Here, Lebesgue spaces play an important role, and the space V is a finite-dimensional space. We apply the abstract theory from Chapter 3, which allows us to derive stationarity conditions of C-stationarity type for local minimizers. The results generalize and are partially influenced by the results from [Harder, G. Wachsmuth, 2018b; Dempe, Harder, et al., 2019]. We also consider a discretization of the bilevel optimization problem in Lebesgue spaces in Section 6.3. This includes theoretical error estimates and numerical experiments.

Finally, in Chapter 7 we provide some concluding remarks and discuss some open questions. We also provide a list of the notations that are used in this thesis and a bibliography at the end of this thesis.

1.3 Contributions from published papers

Some parts of this thesis originate from published articles where the author of this thesis was a co-author or the sole author. We will list these publications and describe which content was used in this thesis, as well as talk about the contributions from the author and other co-authors from these parts. It is also mentioned in the text whenever parts in this thesis are taken from a published article. We emphasize that the vast majority of this thesis is not taken from any published or unpublished articles.

- In the article [Dempe, Harder, et al., 2019] an inverse optimal control problem is considered which would fit into the setting discussed in Chapter 6. The approach of using the relaxation of the optimal value reformulation in order to derive optimality conditions is generalized for abstract Banach spaces in Section 3.4.

- The article [Harder, G. Wachsmuth, 2018a] compares various stationarity systems for the optimal control of the obstacle. It contains an introduction to capacity theory, which has some influences on Section 2.6. However, our treatment of capacity theory in this thesis has a different structure and is much more self-contained. We mention that Corollary 2.6.20 and its proof are taken from [Harder, G. Wachsmuth, 2018a, Lemma 3.5 (c)], and the idea of using closed lattice ideals to define the quasi-support (or fine support) originates from that article. This idea was due to the author of this thesis. In general, the introductory section on capacity theory in [Harder, G. Wachsmuth, 2018a] is joint work with the co-author Gerd Wachsmuth.
- The article [Harder, G. Wachsmuth, 2018b] contains a counterexample that is almost the same as Example 6.2.12 and which was constructed by the author of this thesis.
- In the article [Harder, G. Wachsmuth, 2018c] the limiting normal cone to a complementarity set in $H_0^1(\Omega) \times H^{-1}(\Omega)$ is investigated. We mention Section 5.3 is taken mostly from [Harder, G. Wachsmuth, 2018c, Section 4]. Additionally, the technical results in Lemmas 2.2.4, 2.2.14, and 2.6.29 and their proofs are taken from the appendix of that article, although the proof of Lemma 2.2.14 can only be found in the preprint version of the article, see [Harder, G. Wachsmuth, 2017, Lemma A.1]. The results in the article are joint work with Gerd Wachsmuth and the author of this thesis made significant contributions, such as having the idea for the adaptive (and nonuniform) size of the holes T_i^n, and constructing a large part of the proofs.
- The results from the article [Harder, 2018] are generalized in Chapter 4, mostly from Legendre forms to Legendre-$\star$ forms. We mention that Lemma 4.1.4 and Sections 4.2 and 4.4 are mostly taken from [Harder, 2018].

2 Preliminaries

2.1 Concepts of functional analysis and convex analysis

In this section we collect definitions, notations, and technical results from the areas of functional analysis and convex analysis that are needed in the rest of this thesis. Also included are sections about lattices, quadratic forms, strong convexity, and coercivity.

2.1.1 Definitions and notations for functional analysis

We start with some very basic definitions, notations, and conventions for normed spaces and analysis in general. Most of the notation is standard. We also refer to the list of notations at the end of this thesis.

Normed spaces. All vector spaces used in this thesis are vector spaces over the real numbers. Let X, Y be normed spaces. For a point $x \in X$ we denote the X-norm of x by $\|x\|_X$, or $\|x\|$ if the space is clear from the context. We denote the space of bounded linear operators from X to Y by $\mathbb{L}(X, Y)$. This space is equipped with the norm

$$\|T\|_{\mathbb{L}(X,Y)} := \sup\{\|Tx\|_Y \mid x \in X, \|x\|_X \leq 1\} \qquad \forall T \in \mathbb{L}(X, Y).$$

We denote the dual space of the normed space X by $X^\star$, which can be defined via $X^\star := \mathbb{L}(X, \mathbb{R})$. In order to provide a more intuitive notation for the function application of continuous linear functionals, we define the bilinear function

$$\langle \cdot, \cdot \rangle_{X^\star \times X} : X^\star \times X \to \mathbb{R}, \quad \langle x^\star, x \rangle_{X^\star \times X} := x^\star(x).$$

In some instances we will write $\langle \cdot, \cdot \rangle$ instead of $\langle \cdot, \cdot \rangle_{X^\star \times X}$ if the spaces are clear.

For a bounded linear operator $T \in \mathbb{L}(X, Y)$ we denote its adjoint by $T^\star \in \mathbb{L}(Y^\star, X^\star)$.

We can equip the Cartesian product $X \times Y$ with the norm given by $\|(x, y)\|_{X \times Y} = \|x\|_X + \|y\|_Y$ for $x \in X, y \in Y$ so that $X \times Y$ becomes a normed space again. If no other norm on a Cartesian product of normed spaces is specified, we will always assume this norm. In some contexts we will use the equivalent norm given by $\|(x, y)\|_{X \times Y} = (\|x\|_X^2 + \|y\|_Y^2)^{1/2}$ instead. This has the nice property that the Cartesian product of Hilbert spaces is again a Hilbert space.

A Banach space is called *Hilbertizable*, if it is isomorphic to a Hilbert space, i.e. there exists an inner product on the space such that the norm induced by this inner product is equivalent to the original norm.

For $n \in \mathbb{N}$ normed spaces $X_1, \dots, X_n$ and subsets $A_i \subset X_i$, $i = 1, \dots, n$ it can occasionally be convenient to use the notation

$$\begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix}$$

in order to represent the set $A_1 \times \cdots \times A_n \subset X_1 \times \cdots \times X_n$.

Let $x \in X$ and $\alpha \geq 0$ be given. We write

$$B_\alpha(x) := \{\hat{x} \in X \mid \|x - \hat{x}\|_X \leq \alpha\}$$

for the closed ball with radius α centered at x.

Let $A \subset X$ be a subset. We define the distance of a point $x \in X$ to A via

$$\operatorname{dist}(x, A) := \inf\{\|x - y\|_X \mid y \in A\}.$$

We also denote the closure, interior, and boundary of A by $\operatorname{cl} A$, $\operatorname{int} A$, and ∂A, respectively.

Examples of Banach spaces. An important class of Banach spaces are the finite-dimensional spaces $\mathbb{R}^n$ for $n \in \mathbb{N}$. Unless stated otherwise, these spaces are equipped with the Euclidean norm, which is denoted by $\|x\|$ or $|x|$ for a vector $x \in \mathbb{R}^n$.

For $p \in [1, \infty]$ we introduce the sequence space ℓ^p of sequences $x = \{x_i\}_{i \in \mathbb{N}} \subset \mathbb{R}$ such that $\|x\|_{\ell^p}$ is finite, where $\|x\|_{\ell^p}$ is defined via

$$\|x\|_{\ell^p} := \begin{cases} \left(\sum_{i \in \mathbb{N}} |x_i|^p\right)^{1/p} & \text{if } p \in [1, \infty), \\ \sup\{|x_i| \mid i \in \mathbb{N}\} & \text{if } p = \infty. \end{cases}$$

We equip the space ℓ^p with the norm $\|\cdot\|_{\ell^p}$ for all $p \in [1, \infty]$.

In the spaces $\mathbb{R}^n$ and ℓ^p we denote by e_i the i-th unit vector, where $i \in \{1, \dots, n\}$ or $i \in \mathbb{N}$.

Differentiation in Normed spaces. For a function $f : X \to Y$ with normed spaces X, Y we say that f is *directionally differentiable* at a point $x \in X$ in direction $h \in X$ if the limit

$$f'(x;h) := \lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t}$$

exists. Then $f'(x;h) \in Y$ is called the *directional derivative* at x in direction h.

If f is directionally differentiable at $x \in X$ in every direction $h \in X$ and if there exists a bounded linear operator $T \in \mathbb{L}(X,Y)$ such that

$$Th = f'(x;h) \qquad \forall\, h \in X$$

holds, then f is said to be *Gâteaux differentiable* at $x \in X$. In this case the operator $f'(x) := T$ denotes the Gâteaux derivative. If f is Gâteaux differentiable at a point $x \in X$ and additionally

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - f'(x)h\|}{\|h\|} = 0$$

holds, then we say that f is *Fréchet differentiable* at x. If f is Fréchet differentiable at x then we call $f'(x)$ the Fréchet derivative. If f is Fréchet differentiable at every point $x \in X$ and $x \mapsto f'(x)$ is continuous from X to $\mathbb{L}(X,Y)$ then we say that f is continuously Fréchet differentiable.

We say that f is directionally differentiable if it is directionally differentiable at every point $x \in X$ and in every direction $h \in X$. Likewise, we say that f is Gâteaux/Fréchet differentiable if it is Gâteaux/Fréchet differentiable at every point $x \in X$.

In some cases we will also need partial Gâteaux and Fréchet derivatives. For a function $f : X_1 \times \cdots \times X_n \to Y$ with normed spaces $X_1, \ldots, X_n, Y$ we denote the partial Gâteaux or Fréchet derivative of the function $(x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n)$ with respect to the variable x_i by $f'_{x_i}(x_1, \ldots, x_n)$. For partial second derivatives, we denote the partial Gâteaux or Fréchet derivative of the function $(x_1, \ldots, x_n) \mapsto f'_{x_i}(x_1, \ldots, x_n)$ with respect to the variable x_j by $f''_{x_i x_j}(x_1, \ldots, x_n)$.

General Notation. If Y_1, Y_2, Y_3 are linear subspaces of a vector space X, we write $Y_3 = Y_1 \dot{+} Y_2$ if Y_3 is the direct sum of the linear subspaces Y_1 and Y_2, i.e. if $Y_3 = Y_1 + Y_2$ and $Y_1 \cap Y_2 = \{0\}$.

For a linear operator $T : X \to Y$ between vector spaces we denote its kernel by $\ker T$, which is defined as the preimage $T^{-1}(\{0\})$.

If $A \subset X$ is a subset of a vector space X, we write $\operatorname{conv} A$ for the convex hull of A and $\operatorname{lin} A$ for the linear hull of A.

If $F : X \to \mathcal{P}(Y)$ is a set-valued mapping from a set X to a set Y then we denote its graph $\{(x,y) \in X \times Y \mid y \in F(x)\}$ by $\operatorname{gph} F$.

If $A \subset X$ is a subset of a set X then we define the corresponding *indicator function* or *characteristic function* χ_A via

$$\chi_A : X \to \{0,1\}, \qquad x \mapsto \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases}$$

The sign function $\operatorname{sgn} : \mathbb{R} \to \mathbb{R}$ is defined via $\operatorname{sgn}(\alpha) = \chi_{(0,\infty)}(\alpha) - \chi_{(-\infty,0)}(\alpha)$.

As for the supremum and infimum of empty sets we agree to set

$$\inf \emptyset := \infty \quad \text{and} \quad \sup \emptyset := -\infty.$$

In many estimates, we will use C as a generic positive constant, which might not always refer to the same constant.

2.1.2 Technical results for functional analysis

We collect some technical results in the area of functional analysis. Many of the results in this section can also be found in the literature.

The first lemma allows us to find a common Lipschitz constant for a locally Lipschitz continuous function in a neighborhood of a compact set.

Lemma 2.1.1. Let X, Y be a normed spaces, $K \subset X$ a compact set, and $f : X \to Y$ a locally Lipschitz continuous function. Then there is an $\varepsilon > 0$ such that f is (globally) Lipschitz continuous on $K + B_\varepsilon(0)$.

Proof. Suppose that for all $\varepsilon > 0$ there is no Lipschitz constant of f on $K + B_\varepsilon(0)$. Thus, for each $i \in \mathbb{N}$, there exist points $x_i, y_i \in K + B_{1/i}(0)$ such that $\|f(x_i) - f(y_i)\| > i\|x_i - y_i\|$ holds. We observe that there exist points $\hat{x}_i, \hat{y}_i \in K$ such that $\|x_i - \hat{x}_i\| \leq 1/i$ and $\|y_i - \hat{y}_i\| \leq 1/i$ hold for all $i \in \mathbb{N}$. Since $K \times K$ is compact one can find a converging subsequence $\{(\hat{x}_{i_j}, \hat{y}_{i_j})\}_{j\in\mathbb{N}}$ of $\{(\hat{x}_i, \hat{y}_i)\}_{i\in\mathbb{N}} \subset K \times K$. We denote the limit of this converging subsequence by $(\hat{x}, \hat{y}) \in K \times K$. Due to $\|x_{i_j} - \hat{x}_{i_j}\| \to 0$ and $\|y_{i_j} - \hat{y}_{i_j}\| \to 0$ the convergences $x_{i_j} \to \hat{x}$ and $y_{i_j} \to \hat{y}$ follow. Since f is continuous we have

$$\|\hat{x} - \hat{y}\| = \lim_{j\to\infty} \|x_{i_j} - y_{i_j}\| \leq \lim_{j\to\infty} i_j^{-1} \|f(x_{i_j}) - f(y_{i_j})\| = 0$$

and therefore $\hat{x} = \hat{y}$ holds. Because f is locally Lipschitz continuous, there exist $\varepsilon, C_L > 0$ such that C_L is a Lipschitz constant of f on $B_\varepsilon(\hat{x})$. Since $x_{i_j}, y_{i_j} \in B_\varepsilon(\hat{x})$ and $i_j > C_L$ hold for sufficiently large $j \in \mathbb{N}$ we obtain

$$i_j \|x_{i_j} - y_{i_j}\| < \|f(x_{i_j}) - f(y_{i_j})\| \leq C_L \|x_{i_j} - y_{i_j}\| \leq i_j \|x_{i_j} - y_{i_j}\|$$

for sufficiently large $j \in \mathbb{N}$, which is impossible. Thus, our initial assumption is wrong and the claim of the lemma was shown.

Let us continue with a small lemma that is taken from [Harder, 2018, Lemma 4.3].

Lemma 2.1.2. Let $X = Y_1 \dotplus Y_2$ be a Banach space with closed linear subspaces Y_1, Y_2 and $\dim Y_2 < \infty$. If Y_1 is Hilbertizable, then X is Hilbertizable.

We continue with several lemmas that are related to the weak-$\star$ topology.

Lemma 2.1.3. Let X be a normed space and $Y_1 \subset X^\star$ be a finite-dimensional linear subspace of its dual space. Then the following holds.

(a) The linear subspace Y_1 is weakly-$\star$ closed.

(b) There exists a weakly-$\star$ closed linear subspace $Y_2 \subset X^\star$ such that $X^\star = Y_1 \dotplus Y_2$.

Proof. Part (a) follows from [Rudin, 1991, Theorem 1.21 (b)] and part (b) follows from [Rudin, 1991, Lemma 4.21 (a)].

Lemma 2.1.4. Let X be a Banach space. Further, let $\{x_i\}_{i\in\mathbb{N}} \subset X^\star$ be a sequence that converges weakly-$\star$ to an element $x \in X^\star$ and $\{x_i^\star\}_{i\in\mathbb{N}} \subset X^{\star\star}$ be a sequence that converges (in norm) to an element $x^\star \in X \subset X^{\star\star}$. Then we have

$$\langle x_i^\star, x_i\rangle_{X^{\star\star}\times X^\star} \to \langle x^\star, x\rangle_{X^{\star\star}\times X^\star} \tag{2.1}$$

as $i \to \infty$.

Proof. Since X is complete, there exists a constant $C > 0$ such that $\|x_i\|_{X^\star} \leq C$ holds for all $i \in \mathbb{N}$. Then, using various common estimates, the claim follows from

$$\begin{aligned}\left|\langle x_i^\star, x_i\rangle_{X^{\star\star}\times X^\star} - \langle x^\star, x\rangle_{X^{\star\star}\times X^\star}\right| &\leq \left|\langle x_i^\star - x^\star, x_i\rangle_{X^{\star\star}\times X^\star}\right| + \left|\langle x^\star, x_i - x\rangle_{X^{\star\star}\times X^\star}\right| \\ &\leq C\|x^\star - x_i^\star\|_{X^{\star\star}} + \left|\langle x_i - x, x^\star\rangle_{X^\star\times X}\right| \to 0.\end{aligned}$$

Here the assumption that $x^\star$ is in the linear subspace $X \subset X^{\star\star}$ is an important assumption. Otherwise there can be a counterexample. If we take $X = c_0$ (where $c_0 \subset \ell^\infty$ denotes the closed linear subspace of ℓ^∞ of sequences that converge to 0), $x_i = e_i \in \ell^1 \cong X^\star$, $x = 0 \in \ell^1 \cong X^\star$, and $x_i^\star = x^\star = (1, 1, \ldots) \in \ell^\infty \cong X^{\star\star}$, then the convergence (2.1) does not hold, even though $x_i \overset{\star}{\rightharpoonup} x$ and $x_i^\star \to x^\star$ hold.

Lemma 2.1.5. Let X be a normed space. Then there exists a Banach space $\hat{X}$ such that there is an isometric isomorphism $\iota \in \mathbb{L}(\hat{X}^\star, X^\star)$ and $\{\iota\hat{x}_i^\star\}_{i\in\mathbb{N}} \subset X^\star$ converges weakly-$\star$ for all weakly-$\star$ convergent sequences $\{\hat{x}_i^\star\}_{i\in\mathbb{N}} \subset \hat{X}^\star$.

Proof. We define $\hat{X}$ as the $X^{\star\star}$-closure of $X \subset X^{\star\star}$. Then $\hat{X}$ is a Banach space. Next, we will show that the operator

$$\iota : \hat{X}^\star \to X^\star, \qquad \hat{x}^\star \mapsto (\hat{X} \supset X \ni x \mapsto \langle \hat{x}^\star, x \rangle_{\hat{X}^\star \times \hat{X}})$$

is an isometric isomorphism. To show that ι is surjective, let $x^\star \in X^\star$ be given. Using the inclusion $X \subset X^{\star\star}$ it can be calculated that the functional

$$\hat{x}^\star := (X^{\star\star} \supset \hat{X} \ni x^{\star\star} \mapsto \langle x^{\star\star}, x^\star \rangle_{X^{\star\star} \times X^\star}) \in \hat{X}^\star$$

satisfies $\iota \hat{x}^\star = x^\star$. Thus, ι is bijective. Next, we calculate $\|\iota \hat{x}^\star\|_{X^\star}$ for a given $\hat{x}^\star \in \hat{X}^\star$. Using the density of X in $\hat{X}$ again we obtain

$$\begin{aligned}
\|\iota \hat{x}^\star\|_{X^\star} &= \sup\{\langle \iota \hat{x}^\star, x \rangle_{X^\star \times X} \mid x \in X,\, \|x\|_X \leq 1\} \\
&= \sup\{\langle \hat{x}^\star, x \rangle_{\hat{X}^\star \times \hat{X}} \mid x \in X \subset \hat{X},\, \|x\|_{\hat{X}} \leq 1\} \\
&= \sup\{\langle \hat{x}^\star, \hat{x} \rangle_{\hat{X}^\star \times \hat{X}} \mid \hat{x} \in \hat{X},\, \|\hat{x}\|_{\hat{X}} \leq 1\} = \|\hat{x}^\star\|_{\hat{X}^\star}.
\end{aligned}$$

Thus, ι is an isometry. It follows that ι is injective and therefore an isometric isomorphism. Finally, let $\{\hat{x}_i^\star\}_{i \in \mathbb{N}} \subset \hat{X}^\star$ be a sequence such that $\hat{x}_i^\star \overset{\star}{\rightharpoonup} \hat{x}^\star$ for some $\hat{x}^\star \in \hat{X}^\star$ and let $x \in X$ be given. Then we have

$$\langle \iota \hat{x}_i^\star, x \rangle_{X^\star \times X} = \langle \hat{x}_i^\star, x \rangle_{\hat{X}^\star \times \hat{X}} \to \langle \hat{x}^\star, x \rangle_{\hat{X}^\star \times \hat{X}} = \langle \iota \hat{x}^\star, x \rangle_{X^\star \times X}$$

and therefore $\iota \hat{x}_i^\star \overset{\star}{\rightharpoonup} \iota \hat{x}^\star$.

Note that the converse implication for weakly-$\star$ convergent sequences does not need to be true. This is because even though $X^\star$ and $\hat{X}^\star$ are isometrically isomorphic, they have different predual spaces and thus their weak-$\star$ topologies can differ.

The following lemma is a simple consequence of the uniform boundedness principle.

Lemma 2.1.6. Let X, Y be a Banach spaces, $T \in \mathbb{L}(X, Y^\star)$ an operator, and $\{T_i\}_{i \in \mathbb{N}} \subset \mathbb{L}(X, Y^\star)$ a sequence of operators such that $T_i x \overset{\star}{\rightharpoonup} Tx$ holds for all $x \in X$. Then

$$T_i x_i \overset{\star}{\rightharpoonup} T\bar{x}$$

holds if $\{x_i\}_{i \in \mathbb{N}} \subset X$ is a sequence such that $x_i \to \bar{x}$ for some $\bar{x} \in X$.

Proof. Since Y is complete, the sequence $\{T_i x\}_{i \in \mathbb{N}}$ is bounded in $Y^\star$ for each $x \in X$. Thus, by the uniform boundedness principle, there exists a constant $C > 0$ such that $\|T_i\| \leq C$ holds for all $i \in \mathbb{N}$. Therefore, the convergence $T_i(x_i - \bar{x}) \to 0$ holds. Then the weak-$\star$ convergence

$$T_i x_i - T\bar{x} = (T_i \bar{x} - T\bar{x}) + T_i(x_i - \bar{x}) \overset{\star}{\rightharpoonup} 0$$

follows from the weak-$\star$ convergence $T_i\bar{x} - T\bar{x} \overset{\star}{\rightharpoonup} 0$.

The next two lemmas are related with surjective operators and operators in the neighborhood of a surjective operator.

Lemma 2.1.7. Let X, Y be Banach spaces and let $T \in \mathbb{L}(X, Y)$ be surjective operator. Then there exists a constant $C > 0$ such that for all $T_1, T_2 \in \mathbb{L}(X, Y)$ in a neighborhood of T the following statements are true.

(a) The estimate
$$\|y^\star\| \leq C\|T_1^\star y^\star\|$$
holds for all $y^\star \in Y^\star$.

(b) The operator T_1 is surjective and $B_1(0) \subset T_1 B_{2C}(0)$ holds.

(c) The estimate
$$\|y_1^\star - y_2^\star\| \leq C(\|T_1^\star y_1^\star - T_2^\star y_2^\star\| + \|T_1 - T_2\|\|y_1^\star\|)$$
holds for all $y_1^\star, y_2^\star \in Y^\star$.

Proof. By [Rudin, 1991, Theorem 4.13] we know that there is a constant C such that
$$\|y_1^\star\| \leq C\|T^\star y_1^\star\| \tag{2.2}$$
holds for all $y_1 \in Y^\star$. Since T_1, T_2 are in a neighborhood of T the estimate (2.2) also holds for T_1, T_2 if we make the constant C a little bit larger. Thus, we have shown part (a). Then part (b) follows by applying [Rudin, 1991, Theorem 4.13] to T_1. Note that the constant $C > 0$ does not depend on the choice of T_1.

Part (c) follows from part (a) by
$$\begin{aligned}\|y_1^\star - y_2^\star\| &\leq C\|T_2^\star y_1^\star - T_2^\star y_2^\star\| \\ &\leq C(\|T_1^\star y_1^\star - T_2^\star y_2^\star\| + \|T_1^\star y_1^\star - T_2^\star y_1^\star\|) \\ &\leq C(\|T_1^\star y_1^\star - T_2^\star y_2^\star\| + \|T_1 - T_2\|\|y_1^\star\|).\end{aligned}$$

Lemma 2.1.8. Let X, Y be Banach spaces, $\{T_i\}_{i\in\mathbb{N}} \subset \mathbb{L}(X, Y)$ be a sequence of operators such that $T_i \to T$ for some surjective operator $T \in \mathbb{L}(X, Y)$, and $\{y_i^\star\}_{i\in\mathbb{N}} \subset Y^\star$ be a sequence of functionals.

(a) If the sequence $\{T_i^\star y_i^\star\}_{i\in\mathbb{N}}$ is bounded, then the sequence $\{y_i^\star\}_{i\in\mathbb{N}}$ is bounded.

(b) If we have the convergence $T_i^\star y_i^\star \to x^\star$ for some $x^\star$, then there is a functional $y^\star \in Y^\star$ such that $T^\star y^\star = x^\star$ and $y_i^\star \to y^\star$.

(c) If we have the convergence $T_i^\star y_i^\star \overset{\star}{\rightharpoonup} x^\star$ for some $x^\star$, then there is a functional $y^\star \in Y^\star$ such that $T^\star y^\star = x^\star$ and $y_i^\star \overset{\star}{\rightharpoonup} y^\star$.

Proof. If T_i is in a sufficiently small neighborhood of T then we can apply Lemma 2.1.7. Since $T_i \to T$, we know that this is the case for large $i \in \mathbb{N}$.

Part (a) follows from Lemma 2.1.7 (a).

For part (b) we know that the sequence $\{y_i^\star\}_{i\in\mathbb{N}}$ is bounded due to part (a). Then we can use Lemma 2.1.7 (c) to show that $\{y_i^\star\}_{i\in\mathbb{N}}$ is a Cauchy sequence. Thus, there exists a functional $y^\star \in Y^\star$ such that $y_i^\star \to y^\star$. Then $T^\star y^\star = x^\star$ follows from $T_i^\star \to T^\star$.

We continue with part (c). Using part (a) we know that $\{y_i^\star\}_{i\in\mathbb{N}}$ is bounded. Thus, the convergence $T^\star y_i^\star \overset{\star}{\rightharpoonup} x^\star$ holds. We define $y^\star \in Y^\star$ via

$$\langle y^\star, y\rangle := \lim_{i\to\infty} \langle y_i^\star, y\rangle \qquad \forall y \in Y. \tag{2.3}$$

Let us argue that $y^\star$ is well-defined. For an arbitrary $y \in Y$ we choose $x \in X$ such that $Tx = y$, and due to

$$\langle y_i^\star, y\rangle = \langle y_i^\star, Tx\rangle = \langle T^\star y_i^\star, x\rangle \to \langle x^\star, x\rangle$$

the limit in the definition (2.3) exists. It can also be shown that $y^\star$ is a linear function, and because $\{y_i^\star\}_{i\in\mathbb{N}}$ is bounded the functional $y^\star$ is also bounded. The convergence $y_i^\star \overset{\star}{\rightharpoonup} y^\star$ follows from the definition of $y^\star$, so it remains to show that $T^\star y^\star = x^\star$. Indeed, for $x \in X$ we have

$$\langle x^\star, x\rangle = \lim_{i\to\infty} \langle T_i^\star y_i^\star, x\rangle = \lim_{i\to\infty} \langle y_i^\star, T_i x\rangle = \langle y^\star, Tx\rangle = \langle T^\star y^\star, x\rangle.$$

For the next several lemmas we need the concept of the convergence of a sequence of bounded linear operators in the strong operator topology.

Definition 2.1.9. Let X, Y, Z be normed spaces and $\{T_i\}_{i\in\mathbb{N}} \subset \mathbb{L}(X, Y)$ be a sequence of operators. We say that $\{T_i\}_{i\in\mathbb{N}}$ converges to an operator $T \in \mathbb{L}(X, Y)$ in the *strong operator topology* if $T_i x \to Tx$ holds for all $x \in X$.

We say that a function $f : Z \to \mathbb{L}(X, Y)$ is continuous in the strong operator topology (of $\mathbb{L}(X, Y)$) if $f(z_i)$ converges to $f(z)$ in the strong operator topology for all converging sequences $\{z_i\}_{i\in\mathbb{N}} \subset Z$ with $z_i \to z \in Z$.

Lemma 2.1.10. Let X, Y be Banach spaces, $\{y_i^\star\}_{i\in\mathbb{N}} \subset Y^\star$ be a sequence such that $y_i^\star \overset{\star}{\rightharpoonup} y^\star$ for some $y^\star \in Y^\star$, $\{T_i\}_{i\in\mathbb{N}} \subset \mathbb{L}(X, Y)$ be a sequence of operators such that T_i

converges to T in the strong operator topology, where $T \in \mathbb{L}(X, Y)$. Then

$$T_i^\star y_i^\star \overset{\star}{\rightharpoonup} T^\star y^\star$$

as $i \to \infty$.

Proof. Let $x \in X$ be given. Then we have

$$\begin{aligned} |\langle T_i^\star y_i^\star - T^\star y^\star, x\rangle_{X^\star \times X}| &\leq |\langle T_i^\star y_i^\star - T^\star y_i^\star, x\rangle_{X^\star \times X}| + |\langle T^\star y_i^\star - T^\star y^\star, x\rangle_{X^\star \times X}| \\ &= |\langle y_i^\star, (T_i - T)x\rangle_{Y^\star \times Y}| + |\langle y_i^\star - y^\star, Tx\rangle_{Y^\star \times Y}| \\ &\leq \|y_i^\star\|_{Y^\star} \|(T_i - T)x\|_Y + |\langle y_i^\star - y^\star, Tx\rangle_{Y^\star \times Y}| \to 0 \end{aligned}$$

as $i \to \infty$, where the first term converges because $\{y_i^\star\}_{i \in \mathbb{N}}$ is bounded and $T_i x$ converges to Tx.

In the next two lemmas we discuss the symmetry of second derivatives in normed spaces. It turns out that the continuity of the second derivatives in the strong operator topology is a useful condition. Recall that in finite-dimensional spaces the continuity of the second partial derivatives is a sufficient condition for the symmetry of the second derivatives.

Lemma 2.1.11. Let X, Y be normed spaces and let $g : X \to Y$ be a function that is twice Gâteaux differentiable and whose second Gâteaux derivative is continuous in the strong operator topology of $\mathbb{L}(X, \mathbb{L}(X, Y))$. Then $g''(x) \in \mathbb{L}(X, \mathbb{L}(X, Y))$ is symmetric for every $x \in X$ in the sense that

$$(g''(x)h_1)h_2 = (g''(x)h_2)h_1 \qquad \forall\, h_1, h_2 \in X.$$

Proof. Let $x, h_1, h_2 \in X$ and $y^\star \in Y^\star$ be given. We define the function $\hat{g} : \mathbb{R}^2 \to \mathbb{R}$ via

$$\alpha = (\alpha_1, \alpha_2) \mapsto \langle y^\star, g(x + \alpha_1 h_1 + \alpha_2 h_2)\rangle_{Y^\star \times Y}.$$

It is easy to see that $\hat{g}$ is twice continuously differentiable. By Schwarz's theorem we have $\hat{g}''_{\alpha_1 \alpha_2}(\alpha) = \hat{g}''_{\alpha_2 \alpha_1}(\alpha)$ for all $\alpha \in \mathbb{R}^2$. Then

$$\langle y^\star, (g''(x)h_1)h_2\rangle_{Y^\star \times Y} = \langle y^\star, (g''(x)h_2)h_1\rangle_{Y^\star \times Y}$$

can be concluded. Since $y^\star \in Y^\star$ was arbitrary the claim follows.

Lemma 2.1.12. Let X, V, Y be normed spaces and let $g : X \times V \to Y$ be a function such that the second partial Gâteaux derivatives g''_{xp}, g''_{px} exist and are continuous in the strong operator topologies of $\mathbb{L}(V, \mathbb{L}(X, Y))$ and $\mathbb{L}(X, \mathbb{L}(V, Y))$. Then these partial

second derivatives are symmetric in the sense that

$$(g''_{xp}(x,p)h_p)h_x = (g''_{px}(x,p)h_x)h_p \qquad \forall\,(x,p),(h_x,h_p) \in X \times V.$$

Proof. Let $(x,p),(h_x,h_p) \in X \times V$ and $y^\star \in Y^\star$ be given. We define the function $\hat{g} : \mathbb{R}^2 \to \mathbb{R}$ via

$$\alpha = (\alpha_1,\alpha_2) \mapsto \langle y^\star, g(x+\alpha_1 h_x, p+\alpha_2 h_p)\rangle_{Y^\star \times Y}.$$

It is easy to see that the second partial derivatives $\hat{g}''_{\alpha_1\alpha_2}$, $\hat{g}''_{\alpha_2\alpha_1}$ exist and are continuous. By [Rudin, 1976, Theorem 9.41] these partial derivatives are equal (even if the partial derivatives $\hat{g}''_{\alpha_1\alpha_1}$, $\hat{g}''_{\alpha_2\alpha_2}$ do not exist). Then

$$\langle y^\star, (g''_{xp}(x,p)h_p)h_x\rangle_{Y^\star\times Y} = \langle y^\star, (g''_{px}(x,p)h_x)h_p\rangle_{Y^\star\times Y}$$

can be concluded. Since $y^\star \in Y^\star$ was arbitrary the claim follows.

Finally, we provide a lemma which is an implicit function theorem and also includes some technical Lipschitzian estimates for the resulting implicit function.

Lemma 2.1.13. Let X be a Banach space, Y, V be normed spaces, and $(\bar{x},\bar{p}) \in X \times V$ be a point. Further, let $g : X \times V \to Y$ be a locally Lipschitz continuous function that is partially Gâteaux differentiable with respect to the x variable and whose partial Gâteaux derivative g'_x is locally Lipschitz continuous.

(a) If $g'_x(\bar{x},\bar{p})$ is surjective then there exists a neighborhood V_0 of $\bar{p}$ and a function $\hat{x} : V_0 \to X$ such that

$$g(\hat{x}(p),p) = g(\bar{x},\bar{p}) \qquad \forall\, p \in V_0$$

holds.

(b) If there are constants $\varepsilon_0 > 0$, $\alpha > 0$, $C_L > 0$ such that $B_1(0) \subset g'_x(x,p)B_\alpha(0)$ holds for all $x \in B_{\varepsilon_0}(\bar{x}), p \in B_{\varepsilon_0}(\bar{p})$ and C_L is a Lipschitz constant of g and g'_x on $B_{\varepsilon_0}(\bar{x}) \times B_{\varepsilon_0}(\bar{p})$, then there is a function $\hat{x} : B_{\varepsilon_1}(\bar{p}) \to X$ such that

$$g(\hat{x}(p),p) = g(\bar{x},\bar{p}) \qquad \text{and} \qquad \|\hat{x}(p)-\bar{x}\| \le 2\alpha C_L\|p-\bar{p}\| \tag{2.4}$$

hold for all $p \in B_{\varepsilon_1}(\bar{p})$, where $\varepsilon_1 := \min(\alpha^{-2}C_L^{-2}, \frac{1}{2}\varepsilon_0\alpha^{-1}C_L^{-1}, \varepsilon_0)$.

Proof. We start with part (b). Without loss of generality we can assume that $g(\bar{x},\bar{p}) = 0$. Let $p \in B_{\varepsilon_1}(\bar{p}) \subset B_{\varepsilon_0}(\bar{p})$ be given. We will construct a sequence $\{x_i\}_{i\in\mathbb{N}} \subset B_{\varepsilon_0}(\bar{x})$ that has the properties

$$\|x_i - \bar{x}\| \le \alpha C_L(2-2^{2-i})\|p-\bar{p}\| \tag{2.5a}$$

$$\|g(x_i,p)\| \le C_L 2^{1-i}\|p-\bar{p}\| \tag{2.5b}$$

for all $i \in \mathbb{N}$. For $i = 1$, we have to choose $x_1 := \bar{x}$ due to (2.5a). The estimate (2.5b) holds for $i = 1$ because of $g(x_1, \bar{p}) = 0$ and the Lipschitz continuity of g.

Suppose x_i is already constructed and (2.5) holds for some $i \in \mathbb{N}$. From (2.5a) we obtain $\|x_i - \bar{x}\| \leq 2\alpha C_L \varepsilon_1 \leq \varepsilon_0$ which gives us $x_i \in B_{\varepsilon_0}(\bar{x})$. Then we choose x_{i+1} such that

$$g'_x(x_i, p)(x_{i+1} - x_i) + g(x_i, p) = 0$$

and

$$\|x_{i+1} - x_i\| \leq \alpha \|g(x_i, p)\| \tag{2.6}$$

are satisfied, which is possible due to $B_1(0) \subset g'_x(x_i, p)B_\alpha(0)$. By (2.5) we have

$$\|x_{i+1} - \bar{x}\| \leq \|x_i - \bar{x}\| + \alpha\|g(x_i, p)\| \leq \alpha C_L(2 - 2^{1-i})\|p - \bar{p}\|$$

which shows that (2.5a) also holds for $i + 1$. Thus, we know that $x_{i+1} \in B_{\varepsilon_0}(\bar{x})$ holds. In order to show (2.5b) for $i + 1$ we continue with

$$\begin{aligned} g(x_{i+1}, p) &= g(x_{i+1}, p) - g(x_i, p) - g'_x(x_i, p)(x_{i+1} - x_i) \\ &= \int_0^1 (g'_x(x_i + s(x_{i+1} - x_i), p) - g'_x(x_i, p))(x_{i+1} - x_i)\, \mathrm{d}s \end{aligned}$$

and by taking norms and using the Lipschitz continuity of g'_x we get

$$\|g(x_{i+1}, p)\| \leq C_L \int_0^1 \|s(x_{i+1} - x_i)\|\|x_{i+1} - x_i\|\, \mathrm{d}s = \tfrac{1}{2} C_L \|x_{i+1} - x_i\|^2.$$

Because of (2.6) and (2.5b) this leads to

$$\begin{aligned} \|g(x_{i+1}, p)\| &\leq \tfrac{1}{2} C_L \alpha^2 \|g(x_i, p)\|^2 \leq C_L^3 \alpha^2 2^{1-2i} \|p - \bar{p}\|^2 \\ &\leq C_L^3 \alpha^2 2^{1-2i} \|p - \bar{p}\| \varepsilon_1 \leq C_L 2^{1-(i+1)} \|p - \bar{p}\| \end{aligned}$$

which completes the proof of (2.5) for $i + 1$.

If (2.6) and (2.5b) are combined for consecutive indices then it can be seen that $\{x_i\}_{i \in \mathbb{N}}$ is a Cauchy sequence. We denote its limit by $\hat{x}(p)$. Clearly, we have $\hat{x}(p) \in B_{\varepsilon_0}(\bar{x})$. The properties of $\hat{x}(p)$ claimed in (2.4) follow by taking the limit $i \to \infty$ in (2.5).

We turn now to part (a). Because g'_x is continuous and $g'_x(\bar{x}, \bar{p})$ is surjective, there exists a constant $\alpha > 0$, $\varepsilon_0 > 0$ such that $B_1(0) \subset g'_x(x, p)B_\alpha(0)$ holds for all $x \in B_{\varepsilon_0}(\bar{x}), p \in B_{\varepsilon_0}(\bar{p})$, see Lemma 2.1.7 (b). Then the claim follows from part (b).

2.1.3 Definitions and lemmas for convex analysis

We want to state some definitions from the area of convex analysis that will be important in the rest of this thesis. We also state some technical lemmas that are needed in later chapters and that make use of the newly introduced definitions.

Definition 2.1.14. Let X be a normed space. We define the *polar cone* of a set $A \subset X$ as

$$A^\circ := \left\{ x^\star \in X^\star \;\middle|\; \langle x^\star, x\rangle_{X^\star \times X} \leq 0 \; \forall x \in A \right\}$$

and the *annihilator* of a set $A \subset X$ as

$$A^\perp := \left\{ x^\star \in X^\star \;\middle|\; \langle x^\star, x\rangle_{X^\star \times X} = 0 \; \forall x \in A \right\}.$$

For a point $x \in X$ we use the notation $x^\perp$ which abbreviates $\{x\}^\perp$.

For a set $\hat{A} \subset X^\star$ we instead define the polar cone via

$$\hat{A}^\circ := \left\{ x \in X \;\middle|\; \langle x^\star, x\rangle_{X^\star \times X} \leq 0 \; \forall x^\star \in \hat{A} \right\}$$

and the annihilator of $\hat{A}$ via

$$\hat{A}^\perp := \left\{ x \in X \;\middle|\; \langle x^\star, x\rangle_{X^\star \times X} = 0 \; \forall x^\star \in \hat{A} \right\}.$$

Likewise, for a point $x^\star \in X^\star$ we use the notation $x^{\star\perp}$ which abbreviates $\{x^\star\}^\perp$.

The polar cone will mostly be used for sets that are already convex cones.

Although the above definition could lead to ambiguity in some cases, it will be clear from the context, which version of the polar cone or annihilator is used. Note that in reflexive spaces X the two definitions for the polar cone or annihilator coincide.

It is easy to see that if $A \subset X$ is a linear subspace of a normed space X then $A^\perp = A^\circ$ holds. Likewise, if $\hat{A} \subset X^\star$ is a linear subspace of the dual space $X^\star$ of a normed space X then $\hat{A}^\perp = \hat{A}^\circ$ holds.

We continue with more definitions from convex analysis.

Definition 2.1.15. Let X be a normed space. For a convex set $A \subset X$ and a point $x \in A$ we define the *radial cone* at x as

$$\mathcal{R}_A(x) := \bigcup_{\alpha > 0} \alpha(A - x),$$

the *tangent cone* at x as

$$\mathcal{T}_A(x) := \mathrm{cl}(\mathcal{R}_A(x)),$$

and the *normal cone* at x as

$$\mathcal{N}_A(x) := \mathcal{R}_A(x)^\circ.$$

For $x \in X \setminus A$ we set $\mathcal{N}_A(x) := \emptyset$. For $x \in A$ and $x^\star \in X^\star$ we define the critical cone at $(x, x^\star)$ as

$$\mathcal{K}_A(x, x^\star) := \mathcal{T}_A(x) \cap x^{\star\perp}.$$

We say that a set A is *polyhedric* at $x \in A$ if

$$\mathcal{K}_A(x, x^\star) = \mathrm{cl}(\mathcal{R}_A(x) \cap x^{\star\perp})$$

holds for all $x^\star \in \mathcal{N}_A(x)$. We say that A is polyhedric if it is polyhedric at all points $x \in A$.

Note that the inclusion "$\supset$" in the definition of polyhedricity is always true. We also mention that for a fixed convex set $A \subset X$ we often interpret the normal cone as a set-valued mapping $\mathcal{N}_A : X \to \mathcal{P}(X^\star)$.

Let us continue with some results that use the concepts defined in Definitions 2.1.14 and 2.1.15. We start with the well-known bipolar theorem.

Theorem 2.1.16. Let X be a normed space. If $A \subset X$ is a nonempty convex cone, then the equality

$$A^{\circ\circ} = \mathrm{cl}(A)$$

holds. In particular, if A is a nonempty closed convex cone then

$$A^{\circ\circ} = A$$

holds. Similarly, if $\hat{A} \subset X^\star$ is a weakly-$\star$ closed convex cone and not empty, then the equality

$$\hat{A}^{\circ\circ} = \hat{A}$$

holds.

A proof of this theorem can be obtained from [Fabian et al., 2001, Theorem 4.32].

Let us discuss some useful relations for the objects that were introduced in Definitions 2.1.14 and 2.1.15. Note that we will not always reference these relations when we use them in the rest of this thesis.

An application of Theorem 2.1.16 yields the relations

$$\mathcal{N}_A(x)^\circ = \mathcal{T}_A(x) \qquad \text{and} \qquad \mathcal{N}_A(x) = \mathcal{T}_A(x)^\circ \tag{2.7}$$

for the normal cone and tangent cone to a convex set $A \subset X$ at a point $x \in A$. For two nonempty convex cones A_1, A_2 in X or $X^\star$ the relation

$$A_1^\circ \cap A_2^\circ = (A_1 + A_2)^\circ \tag{2.8}$$

is known, see [Bonnans, Shapiro, 2000, (2.31)]. Similarly, if A_1, A_2 are closed convex cones in X, the relation

$$(A_1 \cap A_2)^\circ = \mathrm{cl}(A_1^\circ + A_2^\circ)$$

holds, see [Bonnans, Shapiro, 2000, (2.32)]. We remark that for convex subsets $A_1 \subset X_1$, $A_2 \subset X_2$ of normed spaces X_1, X_2 and points $x_1 \in A_1$, $x_2 \in A_2$ the radial cone to $A_1 \times A_2$ at $(x_1, x_2) \in A_1 \times A_2$ can be calculated via

$$\mathcal{R}_{A_1 \times A_2}((x_1, x_2)) = \mathcal{R}_{A_1}(x_1) \times \mathcal{R}_{A_2}(x_2).$$

This can be shown using a short calculation which utilizes the convexity of A_1 and A_2.

In the next lemma we provide an expression for the normal cone of a convex cone, which can be found in [Bonnans, Shapiro, 2000, (2.110)].

Lemma 2.1.17. Let X be a normed space and let $A \subset X$ be a closed convex cone with $x \in A$. Then the normal cone of A at x is given by

$$\mathcal{N}_A(x) = A^\circ \cap x^\perp.$$

Let us also state the following preimage rule for polar cones.

Lemma 2.1.18. Let X, Y be Banach spaces, $A \subset X$ be a closed convex cone, and $T : X \to Y$ a bounded linear operator. Under the assumption

$$T(X) - A = Y$$

the equality

$$(T^{-1}(A))^\circ = T^\star(A^\circ)$$

holds.

Proof. Due to our assumption it follows from [Schirotzek, 2007, Lemma 2.4.3] that $T^\star(A^\circ)$ is weakly-$\star$ closed.

Then the claimed equality follows from [Schirotzek, 2007, Lemma 2.4.1].

Related to this preimage rule, we are also interested in normal cones for convex sets that are formulated as preimages of convex sets under nonlinear but differentiable functions.

Lemma 2.1.19. Let X, Y be Banach spaces, $\Phi \subset Y$ a closed and convex set, and $g : X \to Y$ a Gâteaux differentiable function such that $g^{-1}(\Phi)$ is convex. Then for $x \in X$ with $g(x) \in \Phi$ we have the inclusion

$$g'(x)^\star(\mathcal{N}_\Phi(g(x))) \subset \mathcal{N}_{g^{-1}(\Phi)}(x).$$

Proof. Let $y^\star \in \mathcal{N}_\Phi(g(x))$ and $x_1 \in X$ with $g(x_1) \in \Phi$ be given. We need to show that $\langle g'(x)^\star y^\star, x_1 - x\rangle_{X^\star \times X} \leq 0$ holds. For $t \in [0,1]$ we define $x_t := x + t(x_1 - x)$. Note that by the convexity of $g^{-1}(\Phi)$ we have $g(x_t) \in \Phi$ for all $t \in [0,1]$. Using the relation

$y^\star \in \mathcal{N}_\Phi(g(x))$ and the Gâteaux differentiability of g we obtain

$$0 \geq \langle y^\star, (g(x_t) - g(x))/t \rangle_{Y^\star \times Y} \to \langle y^\star, g'(x)(x_1 - x) \rangle_{Y^\star \times Y} = \langle g'(x)^\star y^\star, x_1 - x \rangle_{X^\star \times X}$$

for $t \downarrow 0$. Since $y^\star \in \mathcal{N}_\Phi(g(x))$ and $x_1 \in g^{-1}(\Phi)$ were chosen arbitrarily, the claim follows.

2.1.4 Definitions and basic properties for lattices

In this section we mainly introduce the concept of a lattice. This will be an important concept in some parts in the rest of this thesis. This concept allows us to have generalized inequalities in Banach spaces.

Let X be a normed space and let $K \subset X$ be a nonempty closed convex cone. We define the relation $\leq_K$ on X via

$$x \leq_K y \quad :\Leftrightarrow \quad y - x \in K \qquad \forall\, x, y \in X.$$

The relation $\geq_K$ is defined in an analogous way. If K is *pointed*, i.e. $K \cap -K = \{0\}$, then it can be seen that $\leq_K$ constitutes a partial order on X.

We say that the function $\max_K : X \times X \to X$ is well-defined if for all $x_1, x_2 \in X$ there exists a unique $y \in X$ such that

$$x_1 \leq_K y \ \wedge \ x_2 \leq_K y$$

and

$$x_1 \leq_K x \ \wedge \ x_2 \leq_K x \quad \Rightarrow \quad y \leq_K x \qquad \forall\, x \in X$$

hold. If this is the case we set $\max_K(x_1, x_2) := y$. If the function $\max_K$ is well-defined then we can also define the functions $\min_K$ and $|\cdot|_K$ via

$$\begin{aligned} \min_K(x, y) &:= -\max_K(-x, -y), \\ |x|_K &:= \max_K(x, -x), \end{aligned}$$

where $x, y \in X$. We can also define the generalized minimum or maximum with respect to Φ for three or more arguments recursively via

$$\begin{aligned} \max_K(x_1, \ldots, x_{n+1}) &:= \max_K(\max_K(x_1, \ldots, x_n), x_{n+1}) \qquad \forall n \geq 2, \\ \min_K(x_1, \ldots, x_{n+1}) &:= \min_K(\min_K(x_1, \ldots, x_n), x_{n+1}) \qquad \forall n \geq 2. \end{aligned}$$

If it is clear from the context which cone K is used, then we will occasionally use $\max, \min, |\cdot|$ instead of $\max_K$, $\min_K$, $|\cdot|_K$. We also use the notation

$$x^+ := \max_K(x, 0), \qquad x^- := -\min_K(x, 0).$$

We mention that $\max_K$, $\min_K$, $|\cdot|_K$ share many properties with the functions max, min, $|\cdot|$ for real numbers.

The following definition is taken from [Bonnans, Shapiro, 2000, Definition 3.56].

Definition 2.1.20. Let X be a normed space and let $K \subset X$ be a closed convex cone. We say that K induces a *lattice structure* on X if K is pointed and the function $\max_K$ is well-defined and continuous.

Clearly, if X has a lattice structure induced by K then the functions $\min_K$ and $|\cdot|_K$ are continuous, too. The next lemma tells us that such convex cones K are polyhedric.

Lemma 2.1.21. Let X be a Banach space and let $K \subset X$ be a closed convex cone that induces a lattice structure on X. Then K is polyhedric.

We refer to [Bonnans, Shapiro, 2000, Theorem 3.58] for a proof of this lemma.

Next, we introduce the notion of a closed lattice ideal.

Definition 2.1.22. Let X be a normed space and let $K \subset X$ be a closed convex cone that induces a lattice structure on X. Then a linear subspace $V \subset X$ is called a *lattice ideal* if for all $v \in V, x \in X$ we have the property

$$|x|_K \leq_K |v|_K \quad \Rightarrow \quad x \in V.$$

If V is a closed linear subspace of X and a lattice ideal, then we call it a *closed lattice ideal*.

Finally, let us provide some technical results related to lattice ideals and closed lattice ideals.

Lemma 2.1.23. Let X be a Banach space and let $K \subset X$ be a closed convex cone that induces a lattice structure on X.

(a) Let $V \subset X$ be a linear subspace such that

$$x \leq_K |v|_K \quad \Rightarrow \quad x \in V$$

holds for all $x \in K$, $v \in V$. Then V is a lattice ideal.

(b) If $V \subset X$ is a lattice ideal then $\mathrm{cl}(V)$ is a closed lattice ideal.

(c) Let $\bar{x} \in K$ be given. Then the set

$$\mathrm{cl}(\mathrm{lin}\{y \in K \mid y \leq_K \bar{x}\})$$

is a closed lattice ideal in X.

Proof. For part (a), let $x \in X$ and $v \in V$ be given with $|x|_K \leq_K |v|_K$. We note that we have $0 \leq_K |x|_K$, i.e. $|x|_K \in K$. Thus, our assumption for V implies $|x|_K \in V$. Since we

have $0 \leq_K |x|_K$ and $x \leq_K |x|_K$ it follows that the generalized inequalities

$$0 \leq_K \max_K(x, 0) \leq_K |x|_K \leq_K |v|_K$$

hold. By applying the assumption we obtain $\max_K(x, 0) \in V$. Then $x \in V$ follows from the equation $x + |x|_K = 2 \max_K(x, 0)$, which shows that V is a lattice ideal.

We continue with part (b). Let $v \in \mathrm{cl}(V)$ and $x \in K$ be given with $x \leq_K |v|_K$. Clearly, there exists a sequence $\{v_i\}_{i\in\mathbb{N}} \subset V$ such that $v_i \to v$ as $i \to \infty$. For each $i \in \mathbb{N}$ we have $0 \leq_K \min_K(x, |v_i|_K) \leq_K x \leq_K |v|_K$ and thus $\min_K(x, |v_i|_K) \in V$. Then we can use the continuity of $\min_K$ and $|\cdot|_K$ to obtain

$$x = \min_K(x, |v|_K) = \lim_{i\to\infty} \min_K(x, |v_i|_K) \in \mathrm{cl}(V).$$

The claim then follows from part (a).

In order to show part (c) it suffices to show that the linear subspace $V := \mathrm{lin}\{y \in K \mid y \leq_K \bar{x}\}$ is a lattice, see part (b).

Let $v \in V$, $x \in K$ be given with $x \leq_K |v|_K$. We want to show that $x \in V$ holds. From $v \in V$ we obtain the existence of finitely many points $\{v_i\}_{i=1}^n \subset K$ and weights $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}$ for some $n \in \mathbb{N}$ such that $v = \sum_{i=1}^n \alpha_i v_i$ and $v_i \leq_K \bar{x}$ for all $i \in \{1, \dots, n\}$. Then we obtain the lattice inequalities

$$x \leq_K |v|_K = \Big|\sum_{i=1}^n \alpha_i v_i\Big|_K \leq_K \sum_{i=1}^n |\alpha_i v_i|_K = \sum_{i=1}^n |\alpha_i| v_i \leq_K \sum_{i=1}^n |\alpha_i| \bar{x}.$$

This implies $x \in \mathrm{lin}\{y \in K \mid y \leq_K \bar{x}\} = V$. Thus, V is a lattice ideal by part (a). The claim then follows from part (b).

2.1.5 Quadratic forms

In this section we will discuss some basics about quadratic forms. Quadratic forms will play an important role in Chapter 4, because Legendre forms and Legendre-$\star$ forms are a special class of quadratic forms. We start with the definition of a quadratic form.

Definition 2.1.24. Let X be a normed space. A function $Q : X \to \mathbb{R}$ is a *quadratic form* if there exists a bilinear function $B : X \times X \to \mathbb{R}$ such that $Q(x) = B(x, x)$ holds for all $x \in X$.

We mention that in some cases a quadratic form is defined so that only nonnegative function values are allowed. However, for our purposes a quadratic form is allowed to take negative values.

Let us start with some simple properties of quadratic forms.

Lemma 2.1.25. Let X be a normed space and let $Q : X \to \mathbb{R}$ be a quadratic form. Then the following statements are true.

(a) The parallelogram law

$$Q(x+y) + Q(x-y) = 2Q(x) + 2Q(y)$$

holds for all $x, y \in X$.

(b) The quadratic form Q is 2-homogeneous, i.e. $Q(tx) = t^2 Q(x)$ holds for all $t \in \mathbb{R}$, $x \in X$.

Proof. The claims follow from elementary calculations using the underlying bilinear function.

As for the parallelogram law, one could wonder whether the other implication is true, i.e. whether a function that satisfies the parallelogram law is already a quadratic form. Indeed, one can show that this is true if the function is continuous.

Let us provide a useful description for quadratic forms using a linear operator $T \in \mathbb{L}(X, X^\star)$.

Lemma 2.1.26. Let X be a normed space and let $Q : X \to \mathbb{R}$ be a continuous quadratic form. Then there exists a unique operator $T \in \mathbb{L}(X, X^\star)$ such that

$$Q(x) = \langle Tx, x\rangle_{X^\star \times X} \qquad \forall\, x \in X \tag{2.9}$$

and T is symmetric, i.e.

$$\langle Tx, y\rangle_{X^\star \times X} = \langle Ty, x\rangle_{X^\star \times X}$$

holds for all $x, y \in X$. Additionally, the operator T satisfies the equations

$$\langle Tx, y\rangle_{X^\star \times X} = \tfrac{1}{2}(B(x,y) + B(y,x)) \tag{2.10}$$
$$\langle Tx, y\rangle_{X^\star \times X} = \tfrac{1}{2}(Q(x+y) - Q(x) - Q(y)) \tag{2.11}$$
$$\langle Tx, y\rangle_{X^\star \times X} = \tfrac{1}{4}(Q(x+y) - Q(x-y)) \tag{2.12}$$

for all $x, y \in X$, where $B : X \times X \to \mathbb{R}$ is a bilinear function that satisfies $Q(x) = B(x,x)$ for all $x \in X$.

Proof. Using elementary calculations one can show that the right-hand sides of (2.10), (2.11), and (2.12) are equal for all $x, y \in X$. Thus, $T : X \to X^\star$ can be defined via these equations. The function T is well-defined and a linear and continuous operator because B is bilinear and Q is continuous. Clearly, T is also symmetric.

On the other hand, if T is a symmetric operator that satisfies (2.9), then an elementary

calculation shows that (2.11) is also satisfied. Therefore, T has to be unique.

Let us provide further definitions that relate to quadratic forms.

Definition 2.1.27. Let X be a normed space and let $Q : X \to \mathbb{R}$ be a quadratic form. Then we say that two subsets $A_1, A_2 \subset X$ are *Q-orthogonal* if

$$Q(x_1 + x_2) = Q(x_1) + Q(x_2) \qquad \forall\, x_1 \in A_1, x_2 \in A_2$$

holds. In this case we write

$$A_1 \perp^Q A_2.$$

For elements $x_1, x_2 \in X$ and a subset $A_1 \subset X$ we also use the simplified notation

$$\begin{aligned} A_1 \perp^Q x_2 \quad &:\Leftrightarrow \quad A_1 \perp^Q \{x_2\}, \\ x_1 \perp^Q x_2 \quad &:\Leftrightarrow \quad \{x_1\} \perp^Q \{x_2\}. \end{aligned}$$

We say that Q is *positive* if $Q(x) > 0$ for all $x \in X \setminus \{0\}$. Similarly, we say that Q is *negative* if $Q(x) < 0$ for all $x \in X \setminus \{0\}$.

From the definition of Q-orthogonality and (2.11) one can also obtain the following corollary.

Corollary 2.1.28. Let X be a normed space, let $Q : X \to \mathbb{R}$ be a continuous quadratic form and $T \in \mathbb{L}(X, X^\star)$ the corresponding operator from Lemma 2.1.26. Then two subsets $A_1, A_2 \subset X$ are Q-orthogonal if and only if

$$\langle Tx_1, x_2 \rangle_{X^\star \times X} = 0 \qquad \forall\, x_1 \in A_1, x_2 \in A_2$$

holds.

2.1.6 Strong convexity and coercivity

An important class of functions in optimization is the class of strongly convex functions. We start with the definition of a strongly convex function.

Definition 2.1.29. Let X be a normed space and $A \subset X$ be a convex subset. We call a function $f : X \to \mathbb{R}$ *strongly convex* on A with parameter $\gamma > 0$ if

$$\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha)x_2) + \alpha(1-\alpha)\tfrac{\gamma}{2}\|x_1 - x_2\|_X^2 \tag{2.13}$$

holds for all $x_1, x_2 \in A$ and $\alpha \in [0,1]$. If no subset A is specified, we use $A = X$ in the above.

Let us discuss some basic properties of strongly convex functions.

Lemma 2.1.30. Let X be a normed space, $A \subset X$ a convex subset and $f : X \to \mathbb{R}$ a function.

(a) Suppose that f is continuous on A or continuous on every line segment in A. Then it is strongly convex with parameter $\gamma > 0$ on A if and only if

$$f(x_1) + f(x_2) \geq 2f((x_1 + x_2)/2) + \tfrac{\gamma}{4}\|x_1 - x_2\|_X^2 \tag{2.14}$$

holds for all $x_1, x_2 \in A$.

(b) Suppose that f is Gâteaux differentiable. Then it is strongly convex with parameter $\gamma > 0$ on A if and only if

$$\langle f'(x_1) - f'(x_2), x_1 - x_2\rangle_{X^\star \times X} \geq \gamma\|x_1 - x_2\|_X^2 \tag{2.15}$$

holds for all $x_1, x_2 \in A$.

(c) If f is strongly convex on A and lower semi-continuous then it is bounded from below on A.

(d) If f is strongly convex on A and lower semi-continuous then

$$\lim_{\|x\| \to \infty, x \in A} f(x) = \infty$$

holds.

(e) If f is strongly convex then it is strictly convex.

Proof. We start with part (a). If f is strongly convex then (2.14) follows by setting $\alpha = 1/2$ in (2.13). Now suppose that (2.14) holds for all $x_1, x_2 \in A$. Let $x_1, x_2 \in A$ be given and let $I \subset [0,1]$ denote the set of $\alpha \in [0,1]$ such that (2.13) holds. Next, we will show that for given $\alpha_1, \alpha_2 \in I$ we also have $\alpha_3 := (\alpha_1 + \alpha_2)/2 \in I$. For this purpose we define $\beta_i := 1 - \alpha_i \in [0,1]$ and $z_i := \alpha_i x_1 + \beta_i x_2$ for $i = 1, 2, 3$. Then we have

$$\begin{aligned}
\alpha_3 f(x_1) + (1 - \alpha_3) f(x_2) &= \tfrac{1}{2}\Big(\alpha_1 f(x_1) + \beta_1 f(x_2) + \alpha_2 f(x_1) + \beta_2 f(x_2)\Big) \\
&\geq \tfrac{1}{2}\Big(f(z_1) + f(z_2) + (\alpha_1\beta_1 + \alpha_2\beta_2)\tfrac{\gamma}{2}\|x_1 - x_2\|^2\Big) \\
&\geq f(z_3) + \tfrac{\gamma}{8}\|z_1 - z_2\|^2 + (\alpha_1\beta_1 + \alpha_2\beta_2)\tfrac{\gamma}{4}\|x_1 - x_2\|^2 \\
&= f(z_3) + \big((\alpha_1 - \alpha_2)^2/4 + (\alpha_1\beta_1 + \alpha_2\beta_2)/2\big)\tfrac{\gamma}{2}\|x_1 - x_2\|^2 \\
&= f(z_3) + \alpha_3(1 - \alpha_3)\tfrac{\gamma}{2}\|x_1 - x_2\|^2
\end{aligned}$$

and therefore $\alpha_3 \in I$ holds. Clearly, we also have $0 \in I$ and $1 \in I$. Thus, by induction one can show that $n2^{-i} \in I$ holds for all $i, n \in \mathbb{N} \cup \{0\}$ with $n \leq 2^i$. Because f is continuous on the line segment $\operatorname{conv}\{x_1, x_2\}$ we also obtain that I must be closed. Since the set $\{n2^{-i} \mid i, n \in \mathbb{N} \cup \{0\},\ n \leq 2^i\} \subset I$ is dense in $[0,1]$ this implies that $I = [0,1]$. Therefore,

(2.13) holds for all $\alpha \in [0,1]$ and f is strongly convex on A with parameter $\gamma > 0$.

For part (b), suppose that f is strongly convex with parameter γ. Let $x_1, x_2 \in A$ and $\alpha \in (0,1)$ be given. If we rearrange some terms in the definition of strong convexity we get the inequality

$$f(x_2) - f(x_2 + \alpha(x_1 - x_2)) \geq \alpha(f(x_2) - f(x_1) + (1-\alpha)\tfrac{\gamma}{2}\|x_1 - x_2\|^2).$$

If we divide by α and take the limit for $\alpha \downarrow 0$ this results in the inequality

$$-\langle f'(x_2), x_1 - x_2\rangle_{X^\star \times X} \geq f(x_2) - f(x_1) + \tfrac{\gamma}{2}\|x_1 - x_2\|^2.$$

Switching the roles of x_1 and x_2 in the above inequality yields the inequality

$$-\langle f'(x_1), x_2 - x_1\rangle_{X^\star \times X} \geq f(x_1) - f(x_2) + \tfrac{\gamma}{2}\|x_1 - x_2\|^2.$$

Then (2.15) follows by adding these two inequalities.

To show the other direction of part (b) we can use part (a). Note that we can apply part (a) because f is continuous on every line due to the Gâteaux differentiability of f. Let $x_1, x_2 \in A$ be given and let us define $y_t := tx_1 + (1-t)x_2$ for $t \in [0,1]$. By the fundamental theorem of calculus we have

$$\begin{aligned} f(x_1) - f(y_{1/2}) &= \int_0^1 \langle f'(y_{1/2} + t(x_1 - y_{1/2})), x_1 - y_{1/2}\rangle \,\mathrm{d}t \\ &= \int_0^1 \tfrac{1}{2t}\langle f'(y_{(1+t)/2}), y_{(1+t)/2} - y_{(1-t)/2}\rangle \,\mathrm{d}t \end{aligned}$$

and

$$\begin{aligned} f(x_2) - f(y_{1/2}) &= \int_0^1 \langle f'(y_{1/2} + t(x_2 - y_{1/2})), x_2 - y_{1/2}\rangle \,\mathrm{d}t \\ &= \int_0^1 \tfrac{1}{2t}\langle f'(y_{(1-t)/2}), y_{(1-t)/2} - y_{(1+t)/2}\rangle \,\mathrm{d}t. \end{aligned}$$

By adding these equalities and applying (2.15) we get

$$\begin{aligned} f(x_1) + f(x_2) - 2f(y_{1/2}) &= \int_0^1 \tfrac{1}{2t}\langle f'(y_{(1+t)/2}) - f'(y_{(1-t)/2}), y_{(1+t)/2} - y_{(1-t)/2}\rangle \,\mathrm{d}t \\ &\geq \int_0^1 \tfrac{1}{2t}\gamma\|y_{(1+t)/2} - y_{(1-t)/2}\|^2 \,\mathrm{d}t \\ &= \int_0^1 \tfrac{1}{2t}t^2\gamma\|x_1 - x_2\|^2 \,\mathrm{d}t \\ &= \gamma\|x_1 - x_2\|^2 \int_0^1 \tfrac{1}{2}t \,\mathrm{d}t = \tfrac{\gamma}{4}\|x_1 - x_2\|^2. \end{aligned}$$

Thus, we have shown that (2.14) holds. Therefore, f is strongly convex with parameter γ by part (a).

For part (c) and part (d) we can without loss of generality assume that $0 \in A$ holds. Because f is lower semi-continuous it is bounded from below in a neighborhood of 0, i.e. there exist $\varepsilon \in (0,1), \beta > -\infty$ such that $f(x) \geq \beta$ holds for all $x \in B_\varepsilon(0)$. Let $x \in A$ be given. We define $\alpha := \varepsilon(\|x\| + 2)^{-1} \in (0, 1/2)$. Then, by the definition of strong convexity, we have

$$\alpha f(x) \geq f(\alpha x) - (1-\alpha)f(0) + \alpha(1-\alpha)\tfrac{\gamma}{2}\|x\|^2$$

for some $\gamma > 0$. If we divide by α and use that $\alpha x \in B_\varepsilon(0)$ we obtain

$$\begin{aligned} f(x) &\geq \alpha^{-1}(f(\alpha x) - (1-\alpha)f(0)) + (1-\alpha)\tfrac{\gamma}{2}\|x\|^2 \\ &\geq \alpha^{-1}(\beta - |f(0)|) + \tfrac{\gamma}{4}\|x\|^2 \\ &= (\|x\| + 2)\varepsilon^{-1}(\beta - |f(0)|) + \tfrac{\gamma}{4}\|x\|^2. \end{aligned}$$

Because the right-hand side of this inequality is a quadratic polynomial in $\|x\|$ with a positive factor before the quadratic term, the claims of part (c) and part (d) follow.

Finally, part (e) follows directly from the definition.

We give an example for a class of strongly convex functions. We note that for $p \in [1, \infty)$ the space $L^p(\Omega)$ refers to the space of (equivalence classes of) p-integrable functions $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$ is a measurable subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Recall that the space $L^p(\Omega)$ is equipped with the norm $\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(\omega)|^p \, \mathrm{d}\omega\right)^{1/p}$.

Example 2.1.31. Let $p \in (1, 2]$ be given. We define the functions $f : \ell^p \to \mathbb{R}$, $x \mapsto \frac{1}{2}\|x\|_{\ell^p}^2$ and $g : L^p(\Omega) \to \mathbb{R}$, $x \mapsto \frac{1}{2}\|x\|_{L^p(\Omega)}^2$, where $\Omega \subset \mathbb{R}^d$ is a measurable subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Then the functions f and g are strongly convex with parameter $p - 1$.

Proof. We provide a proof for the strong convexity of g. The proof for the strong convexity of f works in the same way. Let $x_1, x_2 \in L^p(\Omega)$ be given. In [Ball, Carlen, Lieb, 1994, Proposition 3] the inequality

$$(\|y + z\|_{L^p(\Omega)}^2 + \|y - z\|_{L^p(\Omega)}^2)/2 \geq \|y\|_{L^p(\Omega)}^2 + (p-1)\|z\|_{L^p(\Omega)}^2$$

is shown for all $y, z \in L^p(\Omega)$. If we use this with $y = (x_1 + x_2)/2$ and $z = (x_1 - x_2)/2$ we get

$$g(x_1) + g(x_2) \geq 2g((x_1 + x_2)/2) + (p-1)\tfrac{1}{4}\|x_1 - x_2\|_{L^p(\Omega)}^2.$$

Then g is strongly convex with parameter $p - 1$ due to Lemma 2.1.30 (a).

Let us now return to the topic of quadratic forms and introduce the notion of coercivity for quadratic forms.

Definition 2.1.32. Let X be a normed space. We call a quadratic form $Q : X \to \mathbb{R}$ coercive with coercivity constant $\gamma > 0$ if

$$Q(x) \geq \gamma \|x\|_X^2$$

holds for all $x \in X$. We call an operator $T \in \mathbb{L}(X, X^\star)$ coercive with coercivity constant $\gamma > 0$ if the estimate

$$\langle Tx, x \rangle_{X^\star \times X} \geq \gamma \|x\|_X^2$$

holds for all $x \in X$.

The next lemma explores the relationship between coercivity of quadratic forms and strongly convex functions.

Lemma 2.1.33. Let X be a normed space and $Q : X \to \mathbb{R}$ a quadratic form.

(a) The quadratic form Q is coercive with coercivity constant $\gamma > 0$ if and only if Q is strongly convex with parameter $2\gamma > 0$.

(b) The quadratic form Q is nonnegative if and only if it is convex.

(c) The quadratic form Q is coercive if and only if

$$\lim_{\|x\| \to \infty} Q(x) = \infty$$

holds.

Proof. For part (a), let Q be coercive with coercivity constant $\gamma > 0$. Using the underlying bilinear function $B : X \times X \to \mathbb{R}$ one can show that

$$\alpha Q(x) + (1 - \alpha) Q(y) = Q(\alpha x + (1 - \alpha) y) + \alpha(1 - \alpha) Q(x - y)$$

holds for all $x, y \in X$ and $\alpha \in [0, 1]$. Then the strong convexity of Q with parameter $2\gamma > 0$ follows by using the estimate $Q(x - y) \geq \gamma \|x - y\|^2$ in this equation. If, on the other hand, Q is strongly convex with parameter $2\gamma > 0$, then we have

$$Q(x) = \tfrac{1}{2} Q(x) + \tfrac{1}{2} Q(-x) \geq Q(0) + \tfrac{\gamma}{4} \|2x\|^2 = \gamma \|x\|^2$$

and thus Q is coercive with coercivity constant $\gamma > 0$.

Part (b) can be proven simply by using $\gamma = 0$ in the proof for part (a).

If Q is coercive, then it is clear that $Q(x) \to \infty$ as $\|x\| \to \infty$. Suppose Q is not coercive. Then there exists a sequence $\{x_i\}_{i \in \mathbb{N}}$ with $Q(x_i) < \frac{1}{i^2} \|x_i\|^2$ for all $i \in \mathbb{N}$. Thus, $x_i \neq 0$ for all $i \in \mathbb{N}$. Then the sequence $\{y_i\}_{i \in \mathbb{N}}$ defined via $y_i := i \|x_i\|^{-1} x_i$ satisfies $\|y_i\| \to \infty$ and we have

$$Q(y_i) = i^2 \|x_i\|^{-2} Q(x_i) < 1 \qquad \forall\, i \in \mathbb{N}$$

which is a contradiction to $Q(y_i) \to \infty$. Thus, we have shown part (c).

We mention that Lemma 2.1.33 (b) can also be found in [Bonnans, Shapiro, 2000, Proposition 3.71].

If one wants to use the concept of a coercive quadratic form in applications, it can be useful to know in which spaces it is even possible to find a coercive quadratic form. The following proposition shows that if this is possible for a Banach space, the Banach space is already Hilbertizable. This is a well-known result, which can be found in [Harder, 2018, Proposition 4.7] or in the remarks on [Bonnans, Shapiro, 2000, p. 195].

Proposition 2.1.34. Let X be a Banach space and let $Q : X \to \mathbb{R}$ be a coercive and continuous quadratic form on X. Then X is Hilbertizable, i.e. X is isomorphic to a Hilbert space.

Proof. Since Q is continuous, there exists a symmetric operator $T \in \mathbb{L}(X, X^\star)$ such that (2.9) is satisfied, see Lemma 2.1.26. We will show that

$$B : X \times X \to \mathbb{R}, \qquad (x, y) \mapsto \langle Tx, y\rangle_{X^\star \times X}$$

describes an inner product on X whose induced norm is equivalent to the original norm. The symmetry and bilinearity of B is due to the symmetry and linearity of T. It follows directly from the definition of coercivity that B is also positive definite. It remains to show that the norm induced by the inner product is equivalent to the original norm. Since T is bounded, we obtain the estimate $Q(x) = \langle Tx, x\rangle \leq \|T\|\|x\|^2$ for all $x \in X$. On the other hand, we have the estimate

$$\|x\|^2 \leq \tfrac{1}{\gamma} Q(x) = \tfrac{1}{\gamma} B(x, x),$$

where $\gamma > 0$ is the coercivity constant of Q. Thus, we have shown that the norm $Q(\cdot)^{1/2}$ is equivalent to $\|\cdot\|$.

The next lemma about coercive linear operators can be found in [Hinze et al., 2009, Lemma 1.8] and is known as the Lax-Milgram lemma.

Lemma 2.1.35. Let X be a Hilbert space and $T \in \mathbb{L}(X, X^\star)$ be coercive. Then T is continuously invertible, i.e. T^{-1} exists and is bounded.

We mention that this lemma can also be formulated in Banach spaces due to Proposition 2.1.34.

We discussed previously in what spaces coercive and continuous quadratic forms can even exist. We can also ask a similar question about strongly convex functions. It turns out that continuous and strongly convex functions cannot exists in nonreflexive Banach spaces.

Lemma 2.1.36. Let X be a Banach space such that a strongly convex and continuous function $f : X \to \mathbb{R}$ exists. Then the following holds.

(a) The space X is a reflexive Banach space.

(b) If f is additionally twice Gâteaux differentiable in a point $x_0 \in X$, then X is Hilbertizable.

Proof. According to [Borwein et al., 2008, Theorem 2.4] the space X admits an equivalent uniformly convex norm. This implies that the space X is reflexive, see, e.g. [Pettis, 1939].

In order to prove part (b), we will show that the quadratic form

$$Q : X \to \mathbb{R}, \qquad h \mapsto \langle f''(x_0)h, h\rangle_{X^\star \times X}$$

is a coercive quadratic form. Let $h \in X$ be given. Suppose that f is strongly convex with parameter $\gamma > 0$. According to Lemma 2.1.30 (b) we have

$$\begin{aligned} Q(h) &= \langle f''(x_0)h, h\rangle_{X^\star \times X} \\ &= \lim_{t \downarrow 0} t^{-1} \langle f'(x_0 + th) - f'(x_0), h\rangle_{X^\star \times X} \\ &= \lim_{t \downarrow 0} t^{-2} \langle f'(x_0 + th) - f'(x_0), x_0 + th - x_0\rangle_{X^\star \times X} \\ &\geq \lim_{t \downarrow 0} t^{-2} \gamma \|x_0 + th - x_0\|_X^2 = \gamma \|h\|_X^2. \end{aligned}$$

Thus, Q is a coercive quadratic form. Clearly, Q is also continuous. Then we obtain from Proposition 2.1.34 that X is Hilbertizable.

One might ask whether the requirement that f is twice Gâteaux differentiable in a point is really necessary for the Hilbertizability of X. The next example shows that strongly convex functions that are nowhere twice Gâteaux differentiable can exist in reflexive Banach spaces which are not Hilbertizable.

Example 2.1.37. Let $p \in (1, 2)$ be given. We define the function $f : \ell^p \to \mathbb{R}$, $x \mapsto \frac{1}{2}\|x\|_{\ell^p}^2$. Then the function f is strongly convex even though ℓ^p is not Hilbertizable.

Proof. The strong convexity of the function f was already shown in Example 2.1.31. It remains to show that ℓ^p is not Hilbertizable. According to [Albiac, Kalton, 2016, Remark 6.2.11 (g)], the so-called type (or Rademacher type) of ℓ^p is at most $p < 2$. However, according to [Albiac, Kalton, 2016, Theorem 7.4.1] any Banach space that is Hilbertizable must have type 2. Therefore, the Banach space ℓ^p is not Hilbertizable.

2.2 Lebesgue and Sobolev spaces

2.2.1 Notation and definitions for function spaces

In this section we establish notation that is related to function spaces. In later sections we will discuss some technical results in these function spaces.

By $\Omega \subset \mathbb{R}^d$ we denote a bounded and open set in d-dimensional space with an integer $d \geq 1$. Both d and Ω can be considered fixed throughout this thesis. We call Ω the domain.

For functions $f, g : \Omega \to \mathbb{R}$ we define the abbreviating notation

$$\begin{aligned}
\{f < g\} &:= \{\omega \in \Omega \mid f(\omega) < g(\omega)\},\\
\{f \leq g\} &:= \{\omega \in \Omega \mid f(\omega) \leq g(\omega)\},\\
\{f = g\} &:= \{\omega \in \Omega \mid f(\omega) = g(\omega)\},\\
\{f \neq g\} &:= \{\omega \in \Omega \mid f(\omega) \neq g(\omega)\}.
\end{aligned}$$

Furthermore, we will use the notation

$$f^+ := \max(f, 0) \qquad \text{and} \qquad f^- := -\min(f, 0)$$

for the pointwise positive and negative part of f.

Unless noted otherwise, we equip Ω with the d-dimensional Lebesgue measure. For a Lebesgue measurable set $A \subset \mathbb{R}^d$ we write $\operatorname{meas}(A)$ for its Lebesgue measure. For a set $B \subset \Omega$ we say that a property P that depends on $\omega \in \Omega$ holds *almost everywhere* (a.e.) on B if the set $\{\omega \in B \mid P(\omega) \text{ does not hold}\}$ has Lebesgue measure zero. We will also say that $P(\omega)$ holds for *almost all* (a.a.) $\omega \in B$, which has the same meaning. If no such set B is specified in this context, we mean the set $B = \Omega$.

Next, we will introduce several function spaces and normed spaces that will be important in this thesis. For $p \in (0, \infty)$ and a Lebesgue measurable function $f : \Omega \to \mathbb{R}$ we define

$$\|f\|_{L^p(\Omega)} := \Big(\int_\Omega |f|^p \,\mathrm{d}\omega\Big)^{1/p}$$

and for the case $p = \infty$ we define

$$\|f\|_{L^\infty(\Omega)} := \inf\{c \geq 0 \mid |f| < c \text{ a.e. on } \Omega\},$$

which is also called the *essential supremum* of $|f|$ on Ω.

For $p \in [1, \infty]$ we denote the Lebesgue spaces $L^p(\Omega)$ as the set of (equivalence classes of) Lebesgue measurable functions $f : \Omega \to \mathbb{R}$ such that $\|f\|_{L^p(\Omega)}$ is finite. We equip the space $L^p(\Omega)$ with the norm $\|\cdot\|_{L^p(\Omega)}$. It is well-known that this space is a Banach space.

This definition can also be extended to measurable sets $\Omega \subset \mathbb{R}^d$ that are not open or not bounded. In particular, if $\Omega_0 \subset \Omega$ is a measurable set then we have

$$L^p(\Omega_0) := \{v \in L^p(\Omega) \mid v = 0 \text{ a.e. on } \Omega \setminus \Omega_0\}.$$

We also define the Sobolev spaces $W^{1,p}(\Omega)$ for $p \in [1, \infty]$ via

$$W^{1,p}(\Omega) := \{w \in L^p(\Omega) \mid w_i' \in L^p(\Omega) \; \forall\, i = 1, \dots, d\},$$

where w_i' denotes the weak partial derivative of w with respect to the i-th coordinate. We equip these spaces with the norm

$$\|w\|_{W^{1,p}(\Omega)} := \Big(\|w\|_{L^p(\Omega)}^p + \sum_{i=1}^d \|w_i'\|_{L^p(\Omega)}^p\Big)^{1/p}$$

for $p \in [1, \infty)$ and

$$\|w\|_{W^{1,\infty}(\Omega)} := \max\{\|w\|_{L^\infty(\Omega)}, \|w_1'\|_{L^\infty(\Omega)}, \dots, \|w_d'\|_{L^\infty(\Omega)}\}$$

for the case $p = \infty$. Important linear subspaces of these Sobolev spaces are the Sobolev spaces with zero boundary conditions. For $p \in [1, \infty)$ these are defined via

$$W_0^{1,p}(\Omega) := \operatorname{cl}(C_c^\infty(\Omega)) \subset W^{1,p}(\Omega),$$

where the closure is taken with respect to the $W^{1,p}(\Omega)$-norm and $C_c^\infty(\Omega)$ denotes the space of infinitely often differentiable functions with compact support in Ω. Recall that the support of a function $f : \Omega \to \mathbb{R}$ is defined via $\operatorname{supp}(f) := \operatorname{cl}(\{f \neq 0\})$. We equip the spaces $W_0^{1,p}(\Omega)$ with the norm

$$\|w\|_{W_0^{1,p}(\Omega)} := \Big(\sum_{i=1}^d \|w_i'\|_{L^p(\Omega)}^p\Big)^{1/p}.$$

These norms are equivalent to the $W^{1,p}(\Omega)$ norms on the space $W_0^{1,p}(\Omega)$ due to the Poincaré inequality. For $p \in (1, \infty)$ we also define the Sobolev spaces of negative order via

$$W^{-1,p}(\Omega) := \Big(W_0^{1,q}(\Omega)\Big)^\star,$$

where $q \in (1, \infty)$ is the conjugate exponent of p, i.e. $1/p + 1/q = 1$ holds. For more details on Sobolev spaces and weak derivatives we refer to [Adams, Fournier, 2003].

Let us introduce some more function spaces. First, we define the Hilbert spaces $H^1(\Omega)$, $H_0^1(\Omega)$, and $H^{-1}(\Omega)$ via

$$H^1(\Omega) := W^{1,2}(\Omega), \qquad H_0^1(\Omega) := W_0^{1,2}(\Omega), \qquad H^{-1}(\Omega) := W^{-1,2}(\Omega)$$

with the corresponding norms. The vector space $C^k(\Omega)$ denotes the space of functions $f : \Omega \to \mathbb{R}$ that are k times continuously differentiable. The space $C^k(\operatorname{cl}\Omega)$ then denotes

the space of functions $f \in C^k(\Omega)$ such that f and all its partial derivatives (up to order k) have a continuous extension to $\operatorname{cl}\Omega$. We also write $C(\operatorname{cl}\Omega) := C^0(\operatorname{cl}\Omega)$ for the space of continuous functions on $\operatorname{cl}\Omega$. The space $C(\operatorname{cl}\Omega)$ is usually equipped with the $\|\cdot\|_{L^\infty(\Omega)}$-norm. Moreover, we define the space $C_c(\Omega)$ of continuous functions with compact support in Ω.

We also introduce an important class of convex cones in these function spaces. The sets $C_c(\Omega)_+$ and $C_c^\infty(\Omega)_+$ denote the subset of nonnegative functions in $C_c(\Omega)$ and $C_c^\infty(\Omega)$. Similarly, the sets $L^p(\Omega)_+$, $W^{1,p}(\Omega)_+$, $W_0^{1,p}(\Omega)_+$ denote the subsets of $L^p(\Omega)$, $W^{1,p}(\Omega)$, $W_0^{1,p}(\Omega)$ that consist of (equivalence classes of) functions $v : \Omega \to \mathbb{R}$ such that $v \geq 0$ a.e. on Ω. Note that the sets $L^p(\Omega)_+$, $W^{1,p}(\Omega)_+$, $W_0^{1,p}(\Omega)_+$ can be shown to be closed convex cones in their respective Lebesgue or Sobolev spaces. The cone $W^{-1,p}(\Omega)_+$ of nonnegative functionals in $W^{-1,p}(\Omega)$ is introduced via duality, i.e.

$$\begin{aligned} W^{-1,p}(\Omega)_+ &:= \left(-W_0^{1,q}(\Omega)_+\right)^\circ \\ &= \left\{\mu \in W^{-1,p}(\Omega) \;\middle|\; \langle \mu, v\rangle_{W^{-1,p}(\Omega)\times W_0^{1,q}(\Omega)} \geq 0 \;\forall v \in W_0^{1,q}(\Omega)_+\right\}, \end{aligned}$$

where $q \in (1,\infty)$ is the conjugate exponent of p, i.e. $1/p + 1/q = 1$ holds. Finally, if K_+ denotes one of the above convex cones of nonnegative functions, we define the corresponding convex cone K_- of nonpositive functions via $K_- := -K_+$.

For some pairs of function spaces one can show that one function space $X(\Omega)$ is continuously embedded in another function space $Y(\Omega)$, in which case we write $X(\Omega) \hookrightarrow Y(\Omega)$ and denote the embedding operator by $\mathcal{I}_{X(\Omega)\to Y(\Omega)} \in \mathbb{L}(X(\Omega), Y(\Omega))$. Examples for such embedding operators include $\mathcal{I}_{H_0^1(\Omega)\to L^2(\Omega)} \in \mathbb{L}(H_0^1(\Omega), L^2(\Omega))$ and its adjoint, $\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} \in \mathbb{L}(L^2(\Omega), H^{-1}(\Omega))$. We mention that these particular embedding operators are even compact operators. In some instances we will omit the embedding operator between function spaces if the choice of function spaces is clear from the context.

We remark that all partial differential equations that appear in this thesis are meant in the weak sense. In most cases, these partial differential equations involve the Laplace operator $\Delta \in \mathbb{L}(H_0^1(\Omega), H^{-1}(\Omega))$ that is implicitly given by

$$\langle -\Delta v, w\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \int_\Omega \nabla v^\top \nabla w \,\mathrm{d}\omega$$

for all $v, w \in H_0^1(\Omega)$.

2.2.2 Technical results for Lebesgue spaces

In this section we give some technical results for $L^p(\Omega)$ spaces. First, we show that the cone of nonnegative functions induces a lattice structure.

Lemma 2.2.1. Let $\Omega_0 \subset \Omega$ be a measurable set and $p \in [1, \infty]$. Then $L^p(\Omega_0)_+$ induces a lattice structure on $L^p(\Omega_0)$.

Proof. Clearly, $L^p(\Omega_0)_+$ is a closed and convex cone that is also pointed. From the definition of $L^p(\Omega_0)_+$ it follows that that the generalized inequality $v_1 \leq_{L^p(\Omega_0)_+} v_2$ is equivalent to $v_1 \leq v_2$ a.e. on Ω_0 for all $v_1, v_2 \in L^p(\Omega_0)$. Therefore, the function $\max_{L^p(\Omega_0)_+}$ is well-defined and coincides with the pointwise maximum, i.e. $\max_{L^p(\Omega_0)_+}(v_1, v_2)(\omega) = \max(v_1(\omega), v_2(\omega))$ holds for almost all $\omega \in \Omega_0$. It remains to show that the maximum operator is continuous. Let $v_1, v_2, w \in L^p(\Omega_0)$ be given. Note that the pointwise estimate

$$\left|\max(v_1(\omega), w(\omega)) - \max(v_2(\omega), w(\omega))\right| \leq |v_1(\omega) - v_2(\omega)|$$

holds for all $\omega \in \Omega_0$. Thus, by integrating or taking the essential supremum the Lipschitzian estimate

$$\|\max(v_1, w) - \max(v_2, w)\|_{L^p(\Omega)} \leq \|v_1 - v_2\|_{L^p(\Omega)}$$

follows. If we apply this inequality twice we obtain

$$\begin{aligned}\|\max(v_1, w_1) - \max(v_2, w_2)\|_{L^p(\Omega)} &\leq \|\max(v_1, w_1) - \max(v_2, w_1)\|_{L^p(\Omega)} \\ &\quad + \|\max(v_2, w_1) - \max(v_2, w_2)\|_{L^p(\Omega)} \\ &\leq \|v_1 - v_2\|_{L^p(\Omega)} + \|\max(w_1, v_2) - \max(w_2, v_2)\|_{L^p(\Omega)} \\ &\leq \|v_1 - v_2\|_{L^p(\Omega)} + \|w_1 - w_2\|_{L^p(\Omega)}\end{aligned}$$

for all $v_1, v_2, w_1, w_2 \in L^p(\Omega)$. Thus, the function $\max_{L^p(\Omega_0)_+} : L^p(\Omega_0) \times L^p(\Omega_0) \to L^p(\Omega_0)$ is Lipschitz continuous with Lipschitz constant 1.

We continue with a definition of the so-called Haar system in $L^2(\Omega)$.

Definition 2.2.2. We consider the space $L^2(\Omega)$ with $\Omega = (0, 1)$. For indices $i \in \mathbb{N}$, $j \in \{0, \dots, 2^{i-1} - 1\}$ we define the functions

$$v_{i,j}(\omega) := \begin{cases} 2^{(i-1)/2} & \text{if } \omega \in (2j2^{-i}, (2j+1)2^{-i}), \\ -2^{(i-1)/2} & \text{if } \omega \in ((2j+1)2^{-i}, (2j+2)2^{-i}), \\ 0 & \text{else,} \end{cases}$$

in $L^2(\Omega)$ and additionally we define the function $v_{0,0}(\omega) := 1 \in L^2(\Omega)$. Then the set

$$\{v_{0,0}\} \cup \{v_{i,j} \mid i \in \mathbb{N}, j \in \{0, \dots, 2^{i-1} - 1\}\}$$

is called the *Haar system* in $L^2((0, 1))$.

This system was introduced in [Haar, 1910]. It is often useful for constructing counterexamples. The next lemma, which can be concluded from [Haar, 1910], shows that the

Haar system constitutes an orthonormal basis in $L^2((0,1))$.

Lemma 2.2.3. The Haar system as defined in Definition 2.2.2 is an orthonormal basis, i.e. its members are pairwise orthogonal and have $L^2((0,1))$-norm 1, and their linear hull

$$\operatorname{lin}\Big(\{v_{0,0}\} \cup \{v_{i,j} \mid i \in \mathbb{N}, j \in \{0,\dots,2^{i-1}-1\}\}\Big)$$

is dense in $L^2((0,1))$.

The next result shows that each function $g \in L^1(\Omega)$ can be approximated by simple functions given by local averages over small cubes. The lemma together with its proof is taken from [Harder, G. Wachsmuth, 2018c, Lemma A.2].

Lemma 2.2.4. For each $n \in \mathbb{N}$ let $\{P_i^n\}_{i\in I_n}$ be a collection of closed cubes such that each P_i^n is a translation of $[-\frac{1}{n},\frac{1}{n}]^d$, the collection $\{P_i^n\}_{i\in I_n}$ covers Ω and is pairwise disjoint up to sets of measure zero. Here, the set I_n is a finite index set. We denote by $\operatorname{avg}(P_i^n, v) = \operatorname{meas}(P_i^n)^{-1} \int_{P_i^n} v \,\mathrm{d}\omega$ the average of a function $v \in L^1(\Omega)$, which is extended by zero outside of Ω, over P_i^n.

(a) Let $v \in L^1(\Omega)$ be given. Then

$$\sum_{i\in I_n} \operatorname{avg}(P_i^n, v)\chi_{P_i^n} \to v \quad \text{in } L^1(\Omega).$$

(b) Let $q \geq 1$ and $v \in L^1(\Omega)$ with $v \geq 0$ a.e. on Ω be given. Then

$$\sum_{i\in I_n} \operatorname{avg}(P_i^n, v)^{\frac{1}{q}}\chi_{P_i^n} \to v^{\frac{1}{q}} \quad \text{in } L^q(\Omega).$$

Proof. We start with part (a). Since $C_c^\infty(\Omega)$ is dense in $L^1(\Omega)$, we can find a sequence $\{f_m\}_{m\in\mathbb{N}} \subset C_c^\infty(\Omega)$ such that $f_m \to v$ in $L^1(\Omega)$. Because f_m is uniformly continuous, the convergence

$$\sum_{i\in I_n} \operatorname{avg}(P_i^n, f_m)\chi_{P_i^n} \to f_m \quad (n \to \infty)$$

in $L^\infty(\Omega)$ and therefore in $L^1(\Omega)$ holds for all $m \in \mathbb{N}$. We have

$$\begin{aligned}\Big\|v - \sum_{i\in I_n} \operatorname{avg}(P_i^n, v)\chi_{P_i^n}\Big\|_{L^1(\Omega)} &\leq \|v - f_m\|_{L^1(\Omega)} + \Big\|f_m - \sum_{i\in I_n} \operatorname{avg}(P_i^n, f_m)\chi_{P_i^n}\Big\|_{L^1(\Omega)} \\ &\quad + \Big\|\sum_{i\in I_n} \operatorname{avg}(P_i^n, v - f_m)\chi_{P_i^n}\Big\|_{L^1(\Omega)} \\ &\leq 2\|v - f_m\|_{L^1(\Omega)} + \Big\|f_m - \sum_{i\in I_n} \operatorname{avg}(P_i^n, f_m)\chi_{P_i^n}\Big\|_{L^1(\Omega)}.\end{aligned}$$

Now, we can choose $m \in \mathbb{N}$ such that the first term becomes small and, afterwards, we can choose $n \in \mathbb{N}$ such that the second term is small. The convergence in $L^1(\Omega)$ follows.

Now we turn to the proof of part (b). For real numbers $a, b \geq 0$ we have the inequality

$$|a-b|^q \leq |a^q - b^q|. \tag{2.16}$$

Indeed, without loss of generality suppose that $a > b \geq 0$ holds. Then (2.16) follows from

$$1 = \frac{a-b}{a} + \frac{b}{a} \geq \Big(\frac{a-b}{a}\Big)^q + \Big(\frac{b}{a}\Big)^q$$

by multiplying with a^q and subtracting b^q. By applying (2.16) pointwise, we get

$$\begin{aligned}
\Big\| \sum_{i \in I_n} \operatorname{avg}(P_i^n, v)^{\frac{1}{q}} \chi_{P_i^n} - v^{\frac{1}{q}} \Big\|_{L^q(\Omega)}^q &\leq \sum_{i \in I_n} \int_{P_i^n} |\operatorname{avg}(P_i^n, v)^{\frac{1}{q}} - v^{\frac{1}{q}}|^q \, \mathrm{d}\omega \\
&\leq \sum_{i \in I_n} \int_{P_i^n} |\operatorname{avg}(P_i^n, v) - v| \, \mathrm{d}\omega \\
&= \Big\| \sum_{i \in I_n} \operatorname{avg}(P_i^n, v) \chi_{P_i^n} - v \Big\|_{L^1(\Omega)},
\end{aligned}$$

which yields the claimed convergence by applying part (a).

2.2.3 Technical results for smooth functions

In this brief section we will give two lemmas related to functions in $C_c^\infty(\Omega)$.

Lemma 2.2.5. Let $O \subset \mathbb{R}^d$ be an open set and $K \subset O$ be a compact set. Then there exists a function $f \in C_c^\infty(O)_+$ with $f \leq 1$ in O and $f = 1$ on K.

We refer to [Dobrowolski, 2010, Lemma 5.12] for a proof of this lemma.

Lemma 2.2.6. Let $f \in C_c(\Omega)$ be a function. Then there exists a sequence $\{f_i\}_{i \in \mathbb{N}}$ in $C_c^\infty(\Omega)$ such that $f_i \to f$ uniformly and $\operatorname{supp}(f_i) \subset \operatorname{supp}(f) + B_{1/i}(0)$ for all $i \in \mathbb{N}$. Moreover, if $f \geq 0$, then it is possible to choose $\{f_i\}_{i \in \mathbb{N}}$ such that $f_i \geq 0$ for all $i \in \mathbb{N}$.

Proof. According to [Adams, Fournier, 2003, Theorem 2.29] we can approximate f uniformly on Ω with infinitely often differentiable functions f_i. From the construction of the approximation (using convolutions with a mollifier) it can be seen that we can choose f_i such that $\operatorname{supp}(f_i) \subset \operatorname{supp}(f) + B_{1/i}(0)$ holds for all $i \in \mathbb{N}$. For large $i \in \mathbb{N}$ this implies $f_i \in C_c^\infty(\Omega)$. Thus, we can construct a sequence $\{f_i\}_{i \in \mathbb{N}}$ in $C_c^\infty(\Omega)$ with the desired properties. According to the construction using convolutions with a mollifier, it also holds that $f_i \geq 0$ for all $i \in \mathbb{N}$ if $f \geq 0$.

2.2.4 Technical results for Sobolev spaces

We move on to some technical results in Sobolev spaces. A particular focus will be on the role of the pointwise min and max operators in $H_0^1(\Omega)$ and $H^1(\Omega)$. The basis for many of these results is the following lemma, which is often called Stampacchia's lemma.

Lemma 2.2.7. Let $w \in H^1(\Omega)$. Then $w^+ \in H^1(\Omega)$ is true, where the weak gradient ∇w^+ of w^+ is given by

$$\nabla w^+(\omega) = \chi_{\{w>0\}}(\omega)\nabla w(\omega).$$

If additionally $w \in H_0^1(\Omega)$, then we also have $w^+ \in H_0^1(\Omega)$.

The claim about functions in $H^1(\Omega)$ can be found in [Kinderlehrer, Stampacchia, 1980, Theorem II.A.1], whereas the claim about functions in $H_0^1(\Omega)$ can be found in [Stampacchia, 1963-1964, Lemme 1.1] or [Bonnans, Shapiro, 2000, Proposition 6.45].

As a direct consequence of Lemma 2.2.7 we can collect some helpful statements for further use.

Lemma 2.2.8. (a) If $v, w \in H^1(\Omega)$ then $v^-, |v|, \max(v,w), \min(v,w) \in H^1(\Omega)$.

(b) If $v, w \in H_0^1(\Omega)$ then $v^-, |v|, \max(v,w), \min(v,w) \in H_0^1(\Omega)$.

(c) The equality

$$\||v|\|^2_{H_0^1(\Omega)} = \|v^+\|^2_{H_0^1(\Omega)} + \|v^-\|^2_{H_0^1(\Omega)} = \|v\|^2_{H_0^1(\Omega)}$$

and the inequalities

$$\|v^+\|_{H_0^1(\Omega)} \leq \|v\|_{H_0^1(\Omega)}, \quad \|v^-\|_{H_0^1(\Omega)} \leq \|v\|_{H_0^1(\Omega)}$$

hold for all $v \in H_0^1(\Omega)$.

(d) The equality

$$\|\max(v,w)\|^2_{H_0^1(\Omega)} + \|\min(v,w)\|^2_{H_0^1(\Omega)} = \|v\|^2_{H_0^1(\Omega)} + \|w\|^2_{H_0^1(\Omega)}$$

and the inequalities

$$\begin{aligned}\|\max(v,w)\|^2_{H_0^1(\Omega)} &\leq \|v\|^2_{H_0^1(\Omega)} + \|w\|^2_{H_0^1(\Omega)},\\ \|\min(v,w)\|^2_{H_0^1(\Omega)} &\leq \|v\|^2_{H_0^1(\Omega)} + \|w\|^2_{H_0^1(\Omega)}\end{aligned}$$

hold for all $v, w \in H_0^1(\Omega)$.

(e) The maps $v \mapsto v^+$, $v \mapsto v^-$, $v \mapsto |v|$ are continuous from $H^1(\Omega)$ to $H^1(\Omega)$ and from $H^1_0(\Omega)$ to $H^1_0(\Omega)$. The maps $(v,w) \mapsto \max(v,w)$, $(v,w) \mapsto \min(v,w)$ are continuous from $H^1(\Omega)^2$ to $H^1(\Omega)$ and from $H^1_0(\Omega)^2$ to $H^1_0(\Omega)$.

(f) For $w \in H^1_0(\Omega)$ and $v \in H^1_0(\Omega)$ we have $\min(v, 1-w) \in H^1_0(\Omega)$. Moreover, the map $(v,w) \mapsto \min(v, 1-w)$ is continuous from $H^1_0(\Omega)^2$ to $H^1_0(\Omega)$.

(g) The inequality
$$\|\min(v,1)\|_{H^1_0(\Omega)} \leq \|v\|_{H^1_0(\Omega)}$$
holds for all $v \in H^1_0(\Omega)$.

(h) For $v \in H^1_0(\Omega)$ we have $\lim_{i\to\infty} \|v - \min(v,i)\|_{H^1_0(\Omega)} = 0$.

Proof. The claims in parts (a) and (b) follow from Lemma 2.2.7 by utilizing the relations $v^- = v^+ - v$, $|v| = v^+ + v^-$, $\max(v,w) = w + (v-w)^+$, $\min(v,w) = w - (w-v)^+$, respectively. By using Lemma 2.2.7 and linearity arguments we can also obtain

$$\nabla v^- = -\chi_{\{v<0\}} \nabla v = -\chi_{\{v\leq 0\}} \nabla v \qquad \text{a.e. in } \Omega, \tag{2.17a}$$
$$\nabla v^+ = \chi_{\{v>0\}} \nabla v = \chi_{\{v\geq 0\}} \nabla v \qquad \text{a.e. in } \Omega, \tag{2.17b}$$
$$\chi_{\{v=0\}} \nabla v = 0 \qquad \text{a.e. in } \Omega. \tag{2.17c}$$

For part (c), we can calculate the norms directly. By using (2.17) we obtain

$$\begin{aligned} |||v|||^2_{H^1_0(\Omega)} &= \int_\Omega \left|\nabla |v|\right|^2 \mathrm{d}\omega = \int_{\{v>0\}} |\nabla v|^2 \,\mathrm{d}\omega + \int_{\{v\leq 0\}} |\nabla v|^2 \,\mathrm{d}\omega \\ &= \|v^+\|^2_{H^1_0(\Omega)} + \|v^-\|^2_{H^1_0(\Omega)} \end{aligned}$$

as well as

$$\|v^+\|^2_{H^1_0(\Omega)} + \|v^-\|^2_{H^1_0(\Omega)} = \int_{\{v>0\}} |\nabla v|^2 \,\mathrm{d}\omega + \int_{\{v\leq 0\}} |\nabla v|^2 \,\mathrm{d}\omega = \|v\|^2_{H^1_0(\Omega)}.$$

The claimed inequalities follow directly.

The equality in part (d) follows from the fact that

$$|\nabla \max(v,w)|^2 + |\nabla \min(v,w)|^2 = |\nabla v|^2 + |\nabla w|^2$$

holds a.e. in Ω for all $v, w \in H^1_0(\Omega)$ by integrating. The claimed inequalities follow directly.

For part (e), consider a sequence $\{v_i\}_{i\in\mathbb{N}} \subset H^1(\Omega)$ such that $v_i \to v$ for some $v \in H^1(\Omega)$. Without loss of generality we can assume that $v_i \to v$ and $\nabla v_i \to \nabla v$ pointwise a.e. in

Ω. Due to (2.17b) the pointwise a.e. convergence $v_i \to v$ implies $\chi_{\{v_i>0\}}\nabla v \to \chi_{\{v>0\}}\nabla v$ pointwise a.e. in Ω. Then we have

$$\begin{aligned}\|v_i^+ - v^+\|_{H^1(\Omega)}^2 &= \|v_i^+ - v^+\|_{L^2(\Omega)}^2 + \int_\Omega |\chi_{\{v_i>0\}}\nabla v_i - \chi_{\{v>0\}}\nabla v|^2 \,\mathrm{d}\omega \\ &\le \|v_i - v\|_{L^2(\Omega)}^2 + 2\int_\Omega |\chi_{\{v_i>0\}}(\nabla v_i - \nabla v)|^2 \,\mathrm{d}\omega \\ &\quad + 2\int_\Omega |(\chi_{\{v_i>0\}} - \chi_{\{v>0\}})\nabla v|^2 \,\mathrm{d}\omega \\ &\le 2\|v_i - v\|_{H^1(\Omega)}^2 + 2\int_\Omega |(\chi_{\{v_i>0\}} - \chi_{\{v>0\}})\nabla v|^2 \,\mathrm{d}\omega \to 0,\end{aligned}$$

where the last term converges due to Lebesgue's dominated convergence theorem. The convergences $v_i^- \to v^-$ and $|v_i| \to |v|$ follow. The continuity of max and min can be shown by utilizing the descriptions $\max(v,w) = w + (v-w)^+$ and $\min(v,w) = w - (w-v)^+$. The continuity in $H_0^1(\Omega)$ of all these maps follows simply from the continuity in $H^1(\Omega)$.

For part (f), consider sequences $\{v_i\}_{i\in\mathbb{N}}, \{w_i\}_{i\in\mathbb{N}} \subset C_c^\infty(\Omega)$ such that $v_i \to v$ and $w_i \to w$ in $H_0^1(\Omega)$. By Lemma 2.2.5 there exists a function $f_i \in C_c^\infty(\Omega)_+$ for each $i \in \mathbb{N}$ such that $f_i \le 1$ in Ω and $f_i = 1$ on the compact set $\operatorname{supp}(w_i) \cup \operatorname{supp}(v_i) \subset \Omega$. Then it is easy to see that $\min(v_i, 1-w_i) = \min(v_i, f_i - w_i)$ holds for all $i \in \mathbb{N}$. Thus, by part (b) we have $\min(v_i, 1-w_i) = \min(v_i, f_i - w_i) \in H_0^1(\Omega)$. Since $\min : H^1(\Omega)^2 \to H^1(\Omega)$ is continuous by part (e), taking the limit for $i \to \infty$ yields $\min(v, 1-w) \in H_0^1(\Omega)$. The claimed continuity of the map $(v,w) \mapsto \min(v, 1-w)$ follows from part (e).

Since we now know that $\min(v,1) \in H_0^1(\Omega)$ for $v \in H_0^1(\Omega)$, part (g) can be concluded by integrating the inequality

$$|\nabla \min(v,1)|^2 = |\chi_{\{v<1\}}\nabla v|^2 \le |\nabla v|^2$$

that holds a.e. in Ω.

Finally, for part (h) we can obtain $\min(v,i) \in H_0^1(\Omega)$ from part (f) for all $i \in \mathbb{N}$. Then the claim follows from the expression for $\nabla \min(v,i)$ and Lebesgue's dominated convergence theorem.

From Lemma 2.2.8 (e) and Lemma 2.1.21 we can obtain a little corollary.

Corollary 2.2.9. The convex cone $H_0^1(\Omega)_+$ induces a lattice structure on $H_0^1(\Omega)$. Additionally, the set $H_0^1(\Omega)_+$ is polyhedric.

The lattice structure induced by $H_0^1(\Omega)_+$ will play an important role in this thesis. Note that the pointwise meaning of $\max : H_0^1(\Omega)^2 \to H_0^1(\Omega)$ coincides with the lattice maximum $\max_{H_0^1(\Omega)_+} : H_0^1(\Omega)^2 \to H_0^1(\Omega)$. The same is also true for min and $|\cdot|$.

We can extend the results from Lemma 2.2.8 (e) further in the sense that we can also use weakly converging sequences instead of strongly converging sequences.

Lemma 2.2.10. Let $\{v_i\}_{i\in\mathbb{N}}$, $\{w_i\}_{i\in\mathbb{N}} \subset H_0^1(\Omega)$ be weakly converging sequences with weak limits $v, w \in H^1(\Omega)$. Then we have the weak convergences

$$v_i^+ \rightharpoonup v^+, \qquad v_i^- \rightharpoonup v^-, \qquad |v_i| \rightharpoonup |v|$$

as well as

$$\max(v_i, w_i) \rightharpoonup \max(v, w) \qquad \text{and} \qquad \min(v_i, w_i) \rightharpoonup \min(v, w).$$

Proof. For a proof of $v_i^+ \rightharpoonup v^+$, see [G. Wachsmuth, 2016, Lemma 4.1]. The other claims follow by simple calculations.

In the next lemma we will apply Lemma 2.2.8 (e) to show that a function in $H_0^1(\Omega)_+$ can be approximated by functions in $C_c^\infty(\Omega)_+$.

Lemma 2.2.11. Let $v \in H_0^1(\Omega)_+$ be a nonnegative function. Then there exists a sequence $\{f_i\}_{i\in\mathbb{N}}$ in $C_c^\infty(\Omega)_+$ such that $f_i \to v$ in $H_0^1(\Omega)$.

Proof. Since $v \in H_0^1(\Omega)_+$, there exists a sequence $\{v_i\}_{i\in\mathbb{N}} \subset C_c^\infty(\Omega)$ such that $v_i \to v$ in $H_0^1(\Omega)$. For a fixed $i \in \mathbb{N}$, the relation $v_i^+ \in H_0^1(\Omega)$ holds. Let us approximate v_i^+ with a function in $C_c^\infty(\Omega)_+$.

There exists an infinitely often differentiable function $\pi : \mathbb{R} \to \mathbb{R}$ such that $\pi(t) = t$ for $t \geq 1$, $\pi(t) = 0$ for $t \leq 1/2$, and $\pi(t) \in [0, 1]$ for $t \in [1/2, 1]$ hold. Then there exists a constant $C_\pi > 0$ such that $|\pi'(t)| \leq C_\pi$ for all $t \in \mathbb{R}$. For $j \in \mathbb{N}$ we define $w_{i,j}(\omega) := j^{-1}\pi(jv_i(\omega))$. From the properties of π it can be seen that $w_{i,j} \in C_c^\infty(\Omega)_+$ and $w_{i,j}(\omega) = j^{-1}\pi(jv_i^+(\omega))$ hold. Then we have

$$\begin{aligned}
\|v_i^+ - w_{i,j}\|_{H_0^1(\Omega)}^2 &= \int_\Omega \left|\nabla v_i^+(\omega)(1 - \pi'(jv_i^+(\omega)))\right|^2 \mathrm{d}\omega \\
&= \int_{\{0<jv_i^+<1\}} \left|\nabla v_i^+(\omega)(1 - \pi'(jv_i^+(\omega)))\right|^2 \mathrm{d}\omega \\
&\leq (1 + C_\pi)^2 \|v_i^+\|_{H_0^1(\Omega)}^2 \operatorname{meas}(\{0 < jv_i^+ < 1\}) \to 0.
\end{aligned}$$

Therefore, one can pick a suitable index j_i such that $\|w_{i,j_i} - v_i^+\|_{H_0^1(\Omega)} \leq i^{-1}$. Then the convergence $w_{i,j_i} \to v^+ = v$ in $H_0^1(\Omega)$ follows from Lemma 2.2.8 (e).

It is known that for each open set $O \subset \Omega$ one can find a continuous function $v \in C(\operatorname{cl}\Omega)$ such that $O = \{v > 0\}$. In our next lemma we will show that this function can be chosen such that $v \in H_0^1(\Omega)_+$.

Lemma 2.2.12. Let $O \subset \Omega$ be an open set. Then there exists a function $v \in H_0^1(\Omega)_+ \cap C(\operatorname{cl}\Omega)$ such that $O = \{v > 0\}$.

Proof. Without loss of generality we can assume that O is nonempty. Let $\{K_i\}_{i\in\mathbb{N}}$ be a compact exhaustion of O, i.e. an increasing sequence of compact sets such that $\cup_{i\in\mathbb{N}} K_i = O$. Without loss of generality we can assume that $\operatorname{meas}(K_1) > 0$. By Lemma 2.2.5 there exists a function $v_i \in C_c^\infty(\Omega)_+$ such that $v_i = 1$ on K_i, $v_i \in [0,1]$ on $O \setminus K_i$, and $v_i = 0$ on $\Omega \setminus O$ for each $i \in \mathbb{N}$. Then we define v via

$$v := \sum_{i\in\mathbb{N}} 2^{-i}(1 + \|v_i\|_{H_0^1(\Omega)})^{-1} v_i.$$

Since the above sum converges in $H_0^1(\Omega)$ and uniformly on $\operatorname{cl}\Omega$, the relation $v \in H_0^1(\Omega) \cap C(\operatorname{cl}\Omega)$ holds. The properties $v \in H_0^1(\Omega)_+$ and $\{v > 0\} = O$ can also be concluded.

Let us move on to the topic of partial differential equations. The following theorem is a well-known theorem that is related to the so-called maximum principle.

Theorem 2.2.13. Let $\xi \in H^{-1}(\Omega)_+$ and $y \in H_0^1(\Omega)$ be given such that $-\Delta y = \xi$ holds. Then $y \geq 0$ holds a.e. on Ω. In particular, the above holds if $\xi \in L^2(\Omega) \subset H^{-1}(\Omega)$ with $\xi \geq 0$ a.e. on Ω.

Proof. If we test the variational formulation of the partial differential equation $-\Delta y = \xi$ with $y^- \in H_0^1(\Omega)_+$ and then use (2.17a) we get

$$\begin{aligned} 0 \leq \langle \xi, y^- \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} &= \langle -\Delta y, y^- \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \int_\Omega \nabla y^\top \nabla y^- \,\mathrm{d}\omega \\ &= -\int_{\{y<0\}} \nabla y^\top \nabla y \,\mathrm{d}\omega = -\int_\Omega (\nabla y^-)^\top \nabla y^- \,\mathrm{d}\omega = -\|y^-\|_{H_0^1(\Omega)}^2, \end{aligned}$$

Thus, we have $\|y^-\|_{H_0^1(\Omega)} = 0$ and therefore $y \geq 0$ a.e. on Ω.

We will also give a lemma to provide properties of solutions of certain partial differential equations. The claims can also be found in [Harder, G. Wachsmuth, 2018c, Lemma A.1].

Lemma 2.2.14. Suppose that $d \geq 2$ holds. We define an auxiliary function L^d via

$$L^d(s) := \begin{cases} -\log(s)^{-1} & \text{for } s \in (0,1) \text{ and } d = 2 \\ s^{d-2} & \text{for } s \in (0,\infty) \text{ and } d \geq 3. \end{cases} \tag{2.18}$$

Furthermore, let $O := \operatorname{int} B_b(0) \subset \Omega \subset \mathbb{R}^d$ be an open ball with radius $b \leq 1$ and $T := B_a(0) \subset O$ be a closed ball with radius $a \in (0, b)$.

(a) We consider the problem

$$-\Delta v = \chi_T - a^d b^{-d} \chi_O, \qquad v \in H_0^1(\Omega).$$

Then the solution $v \in H_0^1(\Omega)$ vanishes on $\Omega \backslash O$ and, under the additional requirement $a < 1/e$ in the case $d = 2$, we get the estimate

$$\|v\|^2_{H_0^1(\Omega)} \leq C a^{2d} L^d(a)^{-1}, \tag{2.19}$$

where $C > 0$ is a constant that does not depend on a or b but can depend on d.

(b) We consider the problem

$$v = 1 \text{ on } T, \quad -\Delta v = 0 \text{ on } O \setminus T, \quad v = 0 \text{ on } \Omega \setminus O.$$

Then there is a solution $v \in H_0^1(\Omega)$ with $0 \leq v \leq 1$. Under the additional requirements

$$a < b^2 \quad \text{if } d = 2 \qquad \text{and} \qquad a < \frac{b}{2} \quad \text{if } d \geq 3, \tag{2.20}$$

there is a constant C depending only on the dimension d, such that the inequalities

$$\|v\|^2_{H_0^1(\Omega)} \leq C\, L^d(a), \tag{2.21}$$

$$\|v\|_{L^1(\Omega)} \leq C\, b L^d(a) \tag{2.22}$$

hold. Moreover, the (outer) normal derivative of v at ∂O is given by

$$\left.\frac{\partial v}{\partial n}\right|_{\partial O} = \frac{-\max(1, d-2) b^{1-d}}{L^d(a)^{-1} - L^d(b)^{-1}}. \tag{2.23}$$

Proof. We use the proof of [Harder, G. Wachsmuth, 2017, Lemma A.1]. A frequently occurring constant in this proof is the surface measure of the boundary of the d-dimensional unit ball $B_1(0) \subset \mathbb{R}^d$, which we denote by S_d.

For part (a), we can give an explicit solution of the partial differential equation. It turns out that the solution satisfies $v(\omega) = \tilde{v}(|\omega|)$, where

$$\tilde{v}(r) := \begin{cases} c_1 - \frac{1}{2d}(1 - a^d b^{-d}) r^2 & \text{if } 0 \leq r < a, \\ c_2 + \frac{a^d b^{-d}}{2d} r^2 + c_3 \frac{1}{L^d(r)} & \text{if } a \leq r < b, \\ 0 & \text{if } r \geq b \end{cases}$$

with coefficients $c_1, c_2 \in \mathbb{R}$ and $c_3 = a^d/(d \max(1, d-2))$. The coefficients c_1, c_2 have to be chosen in such a way that $\tilde{v}$ is continuous. Note that our choice of c_3 guarantees that

$\tilde{v}$ is continuously differentiable. For the norm of v we have

$$\begin{aligned}\|v\|_{H_0^1(\Omega)}^2 &= \int_O |\nabla v|^2 \,\mathrm{d}\omega = S_d \int_0^b |\tilde{v}'(r)|^2 r^{d-1} \,\mathrm{d}r \\ &= \frac{S_d}{d^2}(1 - a^d b^{-d})^2 \int_0^a r^{d+1} \,\mathrm{d}r + S_d \int_a^b \Big(\frac{a^d b^{-d}}{d} r - c_3 \frac{\max(1, d-2)}{r^{d-1}}\Big)^2 r^{d-1} \,\mathrm{d}r \\ &\le S_d \int_0^a r^{d+1} \,\mathrm{d}r + 2S_d a^{2d} b^{-2d} \int_0^b r^{d+1} \,\mathrm{d}r + 2S_d a^{2d} \int_a^1 r^{1-d} \,\mathrm{d}r \\ &\le S_d a^{d+2} + 2S_d a^{2d} b^{2-d} + 2S_d a^{2d} L^d(a)^{-1} \le 5S_d a^{2d} L^d(a)^{-1},\end{aligned}$$

where the last inequality uses $a < 1/e$ in the case of $d = 2$. Thus, we have shown (2.19).

For part (b) we can again give an explicit representation of v. Since v is rotationally invariant, we can write $v(\omega) = \tilde{v}(|\omega|)$ and find

$$\tilde{v}(r) := \begin{cases} 1 & \text{if } 0 \le r \le a, \\ \frac{L^d(r)^{-1} - L^d(b)^{-1}}{L^d(a)^{-1} - L^d(b)^{-1}} & \text{if } a < r < b, \\ 0 & \text{if } b \le r, \end{cases} \qquad \tilde{v}'(r) := \begin{cases} 0 & \text{if } 0 \le r \le a, \\ \frac{-\max(1,d-2)\, r^{1-d}}{L^d(a)^{-1} - L^d(b)^{-1}} & \text{if } a < r < b, \\ 0 & \text{if } b \le r. \end{cases}$$

Additionally, (2.23) follows from $\frac{\partial v}{\partial n}\Big|_{\partial O} = \lim_{r \uparrow b} \tilde{v}'(r)$. By using the above expression for $\tilde{v}'(r)$ and $\int_a^b \tilde{v}'(r) \,\mathrm{d}r = -1$, we find

$$\begin{aligned}\|v\|_{H_0^1(\Omega)}^2 &= \int_{O \setminus T} |\nabla v|^2 \,\mathrm{d}\omega = S_d \int_a^b \tilde{v}'(r)^2 r^{d-1} \,\mathrm{d}r \\ &= \frac{-S_d \max(1, d-2)}{L^d(a)^{-1} - L^d(b)^{-1}} \int_a^b \tilde{v}'(r) \,\mathrm{d}r = \frac{S_d \max(1, d-2)}{L^d(a)^{-1} - L^d(b)^{-1}}.\end{aligned}$$

By using (2.18) it can be shown that the requirements (2.20) imply the inequality

$$\frac{1}{L^d(a)^{-1} - L^d(b)^{-1}} \le 2L^d(a). \tag{2.24}$$

The claim (2.21) follows. Next, we calculate $\|v\|_{L^1(\Omega)}$. We have

$$\begin{aligned}\int_\Omega |v| \,\mathrm{d}\omega &= \int_{O \setminus T} v \,\mathrm{d}\omega + \mathrm{meas}(T) = S_d \int_a^b \tilde{v}(r) r^{d-1} \,\mathrm{d}r + \mathrm{meas}(T) \\ &\le \frac{S_d}{L^d(a)^{-1} - L^d(b)^{-1}} \int_a^b L^d(r)^{-1} r^{d-1} \,\mathrm{d}r + 2^d a^d.\end{aligned}$$

By calculating the integral for both the cases $d \geq 3$ and $d = 2$, it can be seen that $\int_a^b L^d(r)^{-1} r^{d-1}\,\mathrm{d}r \leq b$. Therefore, using (2.24) and $a^d \leq aL^d(a)$ results in

$$\int_\Omega |v|\,\mathrm{d}\omega \leq \frac{S_d b}{L^d(a)^{-1} - L^d(b)^{-1}} + 2^d a^d \leq C\, b\, L^d(a)$$

for a suitable constant $C > 0$.

2.2.5 Radon measures and positive linear functionals

In this section we want to talk about how positive linear functionals over $C_c(\Omega)$ can be interpreted as measures. A positive linear functional over $C_c(\Omega)$ is a linear functional $\mu : C_c(\Omega) \to \mathbb{R}$ such that $\mu(f) \geq 0$ holds for all $f \in C_c(\Omega)_+$. Positive linear functionals over other function spaces are defined in a similar way.

In order to specify the properties of the resulting measure, we provide the following definition.

Definition 2.2.15. (a) A measure on Ω that is defined on the Borel σ-algebra of Ω is called a *Borel measure*.

(b) A Borel measure is called *regular* if the inner regularity

$$\mu(A) = \sup\{\mu(K) \mid K \subset A \subset \Omega,\ K \text{ compact}\} \tag{2.25}$$

and outer regularity

$$\mu(A) = \inf\{\mu(O) \mid A \subset O \subset \Omega,\ O \text{ open}\}$$

hold for all Borel measurable sets $A \subset \Omega$.

(c) A regular Borel measure is called a *Radon measure* if it is finite on all compact subsets of Ω.

It should be noted that the above definition can vary in the literature. Our definition for the regularity of a measure can be found in [Folland, 1999, p. 212]. For the definition of a Radon measure in [Folland, 1999, p. 212] the condition (2.25) only has to hold for all open sets $A \subset \Omega$. However, this is equivalent to our definition because Ω is the countable union of compact sets, see [Folland, 1999, Corollary 7.6]. We also mention that a Borel measure which is finite on all compact subsets of Ω is already a regular Borel measure, see [Rudin, 1987, Theorem 2.18].

The following theorem is known as the Riesz representation theorem.

Theorem 2.2.16. Let μ be a positive linear functional on $C_c(\Omega)$. Then there exists a unique (nonnegative) measure $\tilde{\mu}$ which is defined on the Borel σ-algebra of Ω and that

satisfies

$$\mu(f) = \int_\Omega f \, \mathrm{d}\tilde{\mu}$$

for all $f \in C_c(\Omega)$. Moreover, $\tilde{\mu}$ is a Radon measure.

For a proof of the existence and uniqueness of $\tilde{\mu}$, see [Rudin, 1987, Theorem 2.14]. For the properties of the Radon measure, we need to combine [Rudin, 1987, Theorem 2.14] with [Rudin, 1987, Theorem 2.18].

2.3 Infinite-dimensional optimization

2.3.1 Existence of minimizers

Let X be a normed space. We consider the optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in \Phi \end{aligned} \tag{P}$$

for an objective function $f : X \to \mathbb{R}$ and a set $\Phi \subset X$. We are interested in the existence of (global) minimizers of (P).

A frequently used tool for the existence of minimizers of infinite-dimensional optimization problems is the well-known Banach-Alaoglu theorem. We state it here along with its sequential and reflexive versions.

Theorem 2.3.1. (a) Let X be a normed space. Then the closed unit ball in $X^\star$ is weakly-$\star$ compact.

(b) Let X be a separable normed space. Then every bounded sequence in $X^\star$ possesses a weakly-$\star$ convergent subsequence.

(c) Let X be a reflexive Banach space. Then every bounded sequence in X possesses a weakly convergent subsequence.

For a proof of parts (a) and (b) we refer to [Conway, 1985, Theorem V.3.1] and [Rudin, 1991, Theorem 3.17]. A proof of part (c) can be obtained by a simple combination of [Conway, 1985, Theorems V.4.2 and V.13.1],

We state some well-known results that concern the existence of minimizers of this problem in an infinite-dimensional setting.

Lemma 2.3.2. (a) If f is strictly convex and Φ is convex then (P) has at most one minimizer.

(b) If X is a Banach space, f is lower semi-continuous and strongly convex on Φ, and Φ is convex, closed, and nonempty, then there exists a unique minimizer of (P).

(c) If Φ is convex, f is strongly convex on Φ with parameter $\gamma > 0$, and $\bar{x}$ is a minimizer of (P), then the quadratic growth condition

$$f(x) \geq f(\bar{x}) + \tfrac{\gamma}{2}\|x - \bar{x}\|^2 \quad \forall x \in \Phi$$

holds.

(d) If X is a reflexive Banach space, f is sequentially weakly lower semi-continuous, and Φ is nonempty, bounded, and sequentially weakly closed, then there exists a minimizer of (P).

Proof. The proof of part (a) is straightforward: If there are two different minimizers $x_1, x_2 \in \Phi$, then the convex combination $\frac{1}{2}(x_1 + x_2) \in \Phi$ would yield a lower function value, which would be a contradiction to the optimality of x_1, x_2.

We continue with part (b). Since Φ is nonempty, an infimal sequence $\{x_i\}_{i\in\mathbb{N}} \subset \Phi$ with $f(x_i) \to a := \inf\{f(x) \mid x \in \Phi\}$ exists. We can without loss of generality assume that $\{f(x_i)\}_{i\in\mathbb{N}}$ is a nonincreasing sequence. Note that due to Lemma 2.1.30 (c) we have $a > -\infty$. The strong convexity of f on Φ yields

$$\tfrac{1}{2}f(x_i) + \tfrac{1}{2}f(x_j) \geq f((x_i + x_j)/2) + \tfrac{\gamma}{8}\|x_i - x_j\|_X^2$$

for all $i, j \in \mathbb{N}$, where $\gamma > 0$ is a constant independent of $i, j \in \mathbb{N}$. For $i, j \in \mathbb{N}$ with $i \leq j$ this implies

$$\|x_i - x_j\|_X^2 \leq \tfrac{4}{\gamma}\big(f(x_i) + f(x_j) - 2f((x_i + x_j)/2)\big) \leq \tfrac{8}{\gamma}(f(x_i) - a), \tag{2.26}$$

where we used $f((x_i + x_j)/2) \geq a$, which is valid due to the convexity of Φ. Because the right-hand side of (2.26) converges to zero as $i \to \infty$ we obtain that $\{x_i\}_{i\in\mathbb{N}}$ is a Cauchy sequence. Since X is a Banach space and Φ is closed, there exists $\bar{x} \in \Phi$ such that $x_i \to \bar{x}$ as $i \to \infty$. Because of $f(\bar{x}) \leq \liminf_{i\to\infty} f(x_i) = a$ this implies that $\bar{x}$ is a minimizer of (P). The uniqueness of the minimizer follows from part (a).

For part (c), let $x \in \Phi$ and $\alpha \in (0, 1)$ be given. Then, according to the definition of strong convexity we have

$$\begin{aligned} \alpha f(x) + (1-\alpha)f(\bar{x}) &\geq f(\alpha x + (1-\alpha)\bar{x}) + \alpha(1-\alpha)\tfrac{\gamma}{2}\|x - \bar{x}\|^2. \\ &\geq f(\bar{x}) + \alpha(1-\alpha)\tfrac{\gamma}{2}\|x - \bar{x}\|^2, \end{aligned}$$

where we used the optimality of $\bar{x}$ in the last step. Therefore,

$$\alpha f(x) \geq \alpha f(\bar{x}) + \alpha(1-\alpha)\tfrac{\gamma}{2}\|x - \bar{x}\|^2$$

holds. If we divide by α and consider the limit for $\alpha \downarrow 0$ then the claimed growth condition follows.

Finally, for part (d), let $\{x_i\}_{i\in\mathbb{N}} \subset \Phi$ be again an infimal sequence, i.e. $f(x_i) \to a := \inf\{f(x) \mid x \in \Phi\}$ holds. Since $\{x_i\}_{i\in\mathbb{N}}$ is bounded, by Theorem 2.3.1 (c) there exists a weakly convergent subsequence $\{x_{i_j}\}_{j\in\mathbb{N}} \subset \{x_i\}_{i\in\mathbb{N}}$ with weak limit $\bar{x} \in \Phi$. Because f is sequentially weakly lower semi-continuous, we have

$$a \leq f(\bar{x}) \leq \liminf_{j\to\infty} f(x_{i_j}) = a$$

which implies $f(\bar{x}) = a$ and that $\bar{x}$ is a minimizer of (P).

Note that a common proof of Lemma 2.3.2 (b) often requires the space X to be reflexive (and then proceed as in the proof of Lemma 2.3.2 (d)). However, in our proof of Lemma 2.3.2 (b) we did not use the argument that X is reflexive and thus we do not need it as an assumption.

2.3.2 KKT conditions and constraint qualifications

Let X, Y be Banach spaces. We consider the optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in \Phi_1, \\ & g(x) \in \Phi_2, \end{aligned} \tag{P}$$

with an objective function $f : X \to \mathbb{R}$, a function $g : X \to Y$, and closed convex sets $\Phi_1 \subset X$, $\Phi_2 \subset Y$. Similar to KKT conditions in finite dimensions, we can define KKT conditions in the infinite-dimensional setting of this section.

Definition 2.3.3. Let $\bar{x} \in X$ be a point and let f and g be Gâteaux differentiable at $\bar{x}$. We say that the *Karush-Kuhn-Tucker conditions* or *KKT conditions* are satisfied at $\bar{x} \in X$ if there exist $\lambda_1 \in X^\star$, $\lambda_2 \in Y^\star$ that satisfy

$$\begin{aligned} f'(\bar{x}) + g'(\bar{x})^\star \lambda_2 + \lambda_1 &= 0, \\ \lambda_1 &\in \mathcal{N}_{\Phi_1}(\bar{x}), \\ \lambda_2 &\in \mathcal{N}_{\Phi_2}(g(\bar{x})), \\ \bar{x} &\in \Phi_1, \\ g(\bar{x}) &\in \Phi_2. \end{aligned}$$

The functionals λ_1, λ_2 are called *Lagrange multipliers*.

From the theory in finite-dimensional spaces it is known that a constraint qualification is needed to ensure that the KKT conditions are first-order necessary optimality conditions. Let us introduce an important constraint qualification that is commonly used in the infinite-dimensional setting.

Definition 2.3.4. Let $\bar{x} \in X$ be a feasible point of (P) and let g be Gâteaux differentiable at $\bar{x}$. We say that the *Robinson-Zowe-Kurcyusz constraint qualification* or *RZKCQ* is satisfied at $\bar{x}$ if

$$g'(\bar{x})\mathcal{R}_{\Phi_1}(\bar{x}) - \mathcal{R}_{\Phi_2}(g(\bar{x})) = Y$$

holds.

This condition was first introduced in [Robinson, 1976, Definition 2] in an equivalent formulation. Later it was shown in [Zowe, Kurcyusz, 1979] that this condition can be used to show the existence of Lagrange multipliers for local minimizers of (P).

Theorem 2.3.5. Let $\bar{x} \in X$ be a local minimizer of (P) such that f is Fréchet differentiable at $\bar{x}$, g is continuously Fréchet differentiable at $\bar{x}$, and $\bar{x}$ satisfies RZKCQ. Then there exist Lagrange multipliers $\lambda_1 \in \mathcal{N}_{\Phi_1}(\bar{x})$, $\lambda_2 \in \mathcal{N}_{\Phi_2}(g(\bar{x}))$ such that the KKT conditions are satisfied.

Proof. In [Zowe, Kurcyusz, 1979, Theorem 3.1] this result was shown under the condition that Φ_2 is a closed convex cone. We generalize the result to arbitrary closed convex sets for Φ_2 using an equivalent formulation of (P). We consider the modified problem

$$\begin{aligned} \min_{x,y} \quad & \hat{f}(x,y) \\ \text{s.t.} \quad & (x,y) \in \hat{\Phi}_1, \\ & \hat{g}(x,y) \in \hat{\Phi}_2, \end{aligned} \tag{$\widehat{\text{P}}$}$$

where we use the Banach spaces $\hat{X} := X \times Y$ and $\hat{Y} := Y$, the functions $\hat{f} : \hat{X} \to \mathbb{R}$, $(x,y) \mapsto f(x)$ and $\hat{g} : \hat{X} \to \hat{Y}$, $(x,y) \mapsto g(x) - y$, and the sets $\hat{\Phi}_1 = \Phi_1 \times \Phi_2 \subset \hat{X}$ and $\hat{\Phi}_2 = \{0\} \subset \hat{Y}$. It can be seen that ($\widehat{\text{P}}$) is an equivalent reformulation of (P), and if we set $\bar{y} = g(\bar{x})$ then $(\bar{x}, \bar{y})$ is a local minimizer of ($\widehat{\text{P}}$). Due to

$$\begin{aligned} \hat{g}'(\bar{x},\bar{y})\mathcal{R}_{\hat{\Phi}_1}(\bar{x},\bar{y}) - \mathcal{R}_{\hat{\Phi}_2}(\hat{g}(\bar{x},\bar{y})) &= \begin{bmatrix} g'(\bar{x}) & -\mathrm{id}_Y \end{bmatrix} \mathcal{R}_{\Phi_1 \times \Phi_2}(\bar{x}, g(\bar{x})) - \{0\} \\ &= g'(\bar{x})\mathcal{R}_{\Phi_1}(\bar{x}) - \mathcal{R}_{\Phi_2}(g(\bar{x})) \\ &= Y = \hat{Y} \end{aligned}$$

RZKCQ is satisfied for the ($\widehat{\text{P}}$). Since $\hat{\Phi}_2$ is now a closed convex cone, we can apply [Zowe, Kurcyusz, 1979, Theorem 3.1], which yields the existence of Lagrange multipliers $\hat{\lambda}_1 \in \mathcal{N}_{\hat{\Phi}_1}((\bar{x},\bar{y}))$, $\hat{\lambda}_2 \in \mathcal{N}_{\hat{\Phi}_2}((\bar{x},\bar{y})) = \hat{Y}^\star$ such that the KKT conditions for ($\widehat{\text{P}}$) are satisfied. If we choose $\lambda_1 \in X^\star, \lambda_2 \in Y^\star$ such that $\hat{\lambda}_1 = (\lambda_1, \lambda_2)$ holds, then it turns out that the KKT conditions for the original problem (P) hold. Indeed, we have

$$0 = \hat{f}'(\bar{x},\bar{y}) + \hat{g}'(\bar{x},\bar{y})^\star \hat{\lambda}_2 + \hat{\lambda}_1 = \begin{pmatrix} f'(\bar{x}) \\ 0 \end{pmatrix} + \begin{bmatrix} g'(\bar{x})^\star \\ -\mathrm{id}_Y^\star \end{bmatrix} \hat{\lambda}_2 + \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

which implies $\lambda_2 = \hat{\lambda}_2$ and $f'(\bar{x}) + g'(\bar{x})^\star \lambda_2 + \lambda_1 = 0$. We also have

$$(\lambda_1, \lambda_2) = \hat{\lambda}_1 \in \mathcal{N}_{\hat{\Phi}_1}((\bar{x}, \bar{y})) = \mathcal{N}_{\Phi_1 \times \Phi_2}((\bar{x}, g(\bar{x}))) = \mathcal{N}_{\Phi_1}(\bar{x}) \times \mathcal{N}_{\Phi_2}(g(\bar{x}))$$

and thus the KKT conditions for (P) are satisfied.

With this theorem we know that the KKT conditions are first-order necessary optimality conditions if RZKCQ is satisfied.

We remark that RZKCQ is equivalent to the Mangasarian-Fromovitz constraint qualification in a finite-dimensional setting. Furthermore, if $\Phi_2 = Y$ holds then RZKCQ is always satisfied. This is often the case for convex optimization problems. For the case that $\Phi_1 = X$ and $g'(\bar{x})$ is surjective we can even claim that the Lagrange multipliers are unique.

Corollary 2.3.6. Let $\bar{x} \in X$ be a local minimizer of (P) such that f is Fréchet differentiable at $\bar{x}$ and g is continuously Fréchet differentiable at $\bar{x}$. Additionally, suppose that $\Phi_1 = X$ and that $g'(\bar{x})$ is surjective. Then there exist unique Lagrange multipliers $\lambda_1 = 0 \in X^\star, \lambda_2 \in Y^\star$ such that the KKT conditions are satisfied.

Proof. The assumption $\Phi_1 = X$ implies $\mathcal{R}_{\Phi_1}(\bar{x}) = X$. Then RZKCQ follows from the surjectivity of $g'(\bar{x})$. If $\bar{x}$ is a local minimizer then Lagrange multipliers $\lambda_1 \in X^\star$, $\lambda_2 \in Y^\star$ exist due to Theorem 2.3.5. Because of $\lambda_1 \in \mathcal{N}_{\Phi_1}(\bar{x}) = \{0\}$ we have the unique value $\lambda_1 = 0$, and the uniqueness of λ_2 follows from the injectivity of $g'(\bar{x})^\star$.

There is an equivalent description of RZKCQ which was used to prove Theorem 2.3.5. This result is due to Zowe and Kurcyusz, see [Zowe, Kurcyusz, 1979, Theorem 2.1].

Theorem 2.3.7. Let $\bar{x} \in X$ be a feasible point of (P) and let g be Gâteaux differentiable at $\bar{x}$. Then the following are equivalent:

(a) $\bar{x}$ satisfies RZKCQ,

(b) there exists $\alpha > 0$ such that

$$B_\alpha(0) \subset g'(\bar{x})((\Phi_1 - \bar{x}) \cap B_1(0)) - (\Phi_2 - g(\bar{x})) \cap B_1(0).$$

We remark that the direction "(a) implies (b)" is a generalization of the open mapping theorem (which follows if we set $\Phi_1 = X, \Phi_2 = \{0\}$).

In the case that we have a convex optimization problem and the KKT system has a solution, we can use the KKT conditions to find global minimizers of (P) without any constraint qualifications (such as RZKCQ).

Proposition 2.3.8. Let $\bar{x} \in X$ be a point and let f and g be Gâteaux differentiable at $\bar{x}$. Suppose that the function f and the set $g^{-1}(\Phi_2)$ are convex. If $\bar{x}$ satisfies the KKT conditions, then it is a global minimizer of (P).

Proof. Let $x \in \Phi_1 \cap g^{-1}(\Phi_2)$ be given. Since f is convex we have

$$f(x) - f(\bar{x}) \geq \langle f'(\bar{x}), x - \bar{x} \rangle_{X^\star \times X}.$$

Using the KKT conditions this implies

$$f(x) - f(\bar{x}) \geq \langle g'(\bar{x})^\star \lambda_2, \bar{x} - x \rangle_{X^\star \times X} + \langle \lambda_1, \bar{x} - x \rangle_{X^\star \times X}$$

with $\lambda_1 \in \mathcal{N}_{\Phi_1}(\bar{x})$ and $\lambda_2 \in \mathcal{N}_{\Phi_2}(g(\bar{x}))$. Because $g'(\bar{x})^\star \lambda_2 \in \mathcal{N}_{g^{-1}(\Phi_2)}(\bar{x})$ holds due to Lemma 2.1.19, we obtain $f(x) \geq f(\bar{x})$ from $\langle g'(\bar{x})^\star \lambda_2, \bar{x} - x \rangle \geq 0$ and $\langle \lambda_1, \bar{x} - x \rangle \geq 0$. Since x was an arbitrary feasible point, this implies that $\bar{x}$ is a global minimizer.

Suppose that (P) is a convex optimization problem and that RZKCQ is satisfied for all feasible points. Then a simple combination of Theorem 2.3.5 and Proposition 2.3.8 yields that we can characterize global minimizers of (P) by the KKT conditions.

2.4 Generalized normal cones

The normal cone to a convex set can be a very useful concept in optimization. However, for a nonconvex set the situation is more complicated. In order to generalize the concept of the normal cone to nonconvex sets, we provide the following definition.

Definition 2.4.1. Let X be a Banach space, $A \subset X$ be a subset and $\bar{x} \in A$ be a point.

(a) For $\varepsilon \geq 0$ the *set of ε-normals* to A at $\bar{x}$ is defined via

$$\widehat{\mathcal{N}}_A^\varepsilon(\bar{x}) := \left\{ x^\star \in X^\star \;\middle|\; \limsup_{x \to \bar{x},\, x \in A} \frac{\langle x^\star, x - \bar{x} \rangle}{\|x - \bar{x}\|} \leq \varepsilon \right\}.$$

(b) The *Fréchet normal cone* to A at $\bar{x}$ is defined via

$$\mathcal{N}_A^{\text{Fréchet}}(\bar{x}) := \widehat{\mathcal{N}}_A^0(\bar{x}).$$

(c) We define the *limiting normal cone* to A at $\bar{x}$ via

$$\mathcal{N}_A^{\lim}(\bar{x}) := \left\{ x^\star \in X^\star \;\middle|\; \begin{array}{l} \exists \{\varepsilon_i\}_{i \in \mathbb{N}} \subset [0, \infty),\, \{x_i\}_{i \in \mathbb{N}} \subset A,\, \{x_i^\star\}_{i \in \mathbb{N}} \subset X^\star : \\ \varepsilon_i \downarrow 0,\, x_i \to \bar{x},\, x_i^\star \overset{\star}{\rightharpoonup} x^\star,\, x_i^\star \in \widehat{\mathcal{N}}_A^{\varepsilon_i}(x_i)\, \forall i \in \mathbb{N} \end{array} \right\}.$$

(d) If X is a reflexive Banach space and A is closed, then we define the *strong limiting normal cone* to A at $\bar{x}$ via

$$\mathcal{N}_A^{\text{s-lim}}(\bar{x}) := \left\{ x^\star \in X^\star \;\middle|\; \begin{array}{l} \exists \{x_i\}_{i\in\mathbb{N}} \subset A,\ \{x_i^\star\}_{i\in\mathbb{N}} \subset X^\star : \\ \quad x_i \to \bar{x},\ x_i^\star \to x^\star,\ x_i^\star \in \mathcal{N}_A^{\text{Fréchet}}(x_i)\, \forall i \in \mathbb{N} \end{array} \right\}.$$

(e) If X is a reflexive Banach space and A is closed, then the *Clarke normal cone* to A at $\bar{x}$ is defined via

$$\mathcal{N}_A^{\text{Clarke}}(\bar{x}) := \operatorname{cl}\big(\operatorname{conv} \mathcal{N}_A^{\text{lim}}(\bar{x})\big).$$

If $\bar{x} \notin A$ then we set $\widehat{\mathcal{N}}_A^{\varepsilon}(\bar{x}) = \mathcal{N}_A^{\text{Fréchet}}(\bar{x}) = \mathcal{N}_A^{\text{lim}}(\bar{x}) = \mathcal{N}_A^{\text{s-lim}}(\bar{x}) = \mathcal{N}_A^{\text{Clarke}}(\bar{x}) = \emptyset$.

Parts (a) to (c) of this definition come from [Mordukhovich, 2006, Definition 1.1]. The strong limiting normal cone was defined this way in [Mehlitz, G. Wachsmuth, 2018, Section 2.1.3]. The Clarke normal cone is usually defined as the polar cone of the so-called Clarke tangent cone. However, in reflexive Banach spaces this is equivalent to our definition, see [Mordukhovich, 2006, Theorem 3.57]. Clearly, in finite-dimensional spaces the strong limiting normal cone to closed sets is the same as the limiting normal cone. We call $\mathcal{N}_A^{\text{Fréchet}}(\bar{x})$, $\mathcal{N}_A^{\text{lim}}(\bar{x})$, $\mathcal{N}_A^{\text{s-lim}}(\bar{x})$, $\mathcal{N}_A^{\text{Clarke}}(\bar{x})$ generalized normal cones. Note that $\widehat{\mathcal{N}}_A^{\varepsilon}(\bar{x})$ is usually not a cone for $\varepsilon > 0$.

For a fixed set $A \subset X$ we can understand the objects that were introduced in Definition 2.4.1 as set-valued mappings from X to $\mathcal{P}(X^\star)$. Unlike the usual normal cone, these objects are defined in such a way that they only depend on the local behavior of the set A near the point $\bar{x} \in A$.

In the next lemma we show that the generalized normal cones from Definition 2.4.1 are indeed generalizations of the normal cone (which is usually only defined for convex sets).

Lemma 2.4.2. Let X be a Banach space, $A \subset X$ be a convex subset and $\bar{x} \in A$ be a point. Then

$$\mathcal{N}_A(\bar{x}) = \mathcal{N}_A^{\text{Fréchet}}(\bar{x}) = \mathcal{N}_A^{\text{lim}}(\bar{x})$$

holds. If additionally X is reflexive and $A \subset X$ is a closed convex set, then

$$\mathcal{N}_A(\bar{x}) = \mathcal{N}_A^{\text{Clarke}}(\bar{x}) = \mathcal{N}_A^{\text{s-lim}}(\bar{x})$$

holds.

Proof. The equality $\mathcal{N}_A(\bar{x}) = \mathcal{N}_A^{\text{Fréchet}}(\bar{x})$ can be found in [Mordukhovich, 2006, Proposition 1.3], and $\mathcal{N}_A(\bar{x}) = \mathcal{N}_A^{\text{lim}}(\bar{x})$ follows from [Mordukhovich, 2006, Proposition 1.5]. The other equalities follow after relatively simple calculations.

The description of the limiting normal cone in Definition 2.4.1 is quite complicated. If we are in a reflexive Banach space (or more general, in an Asplund space) and the set A is closed, then we can find an alternative description of the limiting normal cone which is a little bit easier to handle.

Lemma 2.4.3. Let X be a Banach space, $A \subset X$ be a closed subset and $\bar{x} \in A$ be a point. If X is reflexive or an Asplund space (which is a space where every separable linear subspace has a separable dual space), then the equality

$$\mathcal{N}_A^{\lim}(\bar{x}) = \left\{ x^\star \in X^\star \;\middle|\; \begin{array}{l} \exists \{x_i\}_{i\in\mathbb{N}} \subset A,\ \{x_i^\star\}_{i\in\mathbb{N}} \subset X^\star : \\ x_i \to \bar{x},\ x_i^\star \overset{\star}{\rightharpoonup} x^\star,\ x_i^\star \in \mathcal{N}_A^{\text{Fréchet}}(x_i)\ \forall i \in \mathbb{N} \end{array} \right\}$$

holds.

This result can be found in [Mordukhovich, 2006, Theorem 2.35]. We mention that all reflexive Banach spaces are Asplund spaces.

2.5 MPCCs in finite dimensions

Let us briefly discuss mathematical programs with complementarity constraints, or MPCCs for short. In finite-dimensional spaces, these problems are well understood. We exemplarily refer to [Luo, Pang, Ralph, 1996; Outrata, Kočvara, Zowe, 1998]. Often, MPCCs are also called mathematical programs with equilibrium constraints, or MPECs for short. While the focus in this thesis is usually on infinite-dimensional optimization problems, some of which have a complementarity-like structure, we are interested in finite-dimensional MPCCs for the purpose of comparisons.

The mathematical program with complementarity constraints is given by

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0, \\ & h(x) = 0, \\ & G(x) \ge 0, \\ & H(x) \ge 0, \\ & G(x)^\top H(x) = 0. \end{aligned} \tag{MPCC}$$

Here, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^l$, $h : \mathbb{R}^n \to \mathbb{R}^m$, $G, H : \mathbb{R}^n \to \mathbb{R}^p$ are assumed to be continuously differentiable functions. The constraints involving the functions G and H are called the complementarity constraints. These complementarity constraints introduce some difficulties that are not present in many ordinary nonlinear optimization problems. For example, constraint qualifications such as the Mangasarian-Fromovitz constraint qualification are usually not satisfied, and the KKT conditions might not be useful for

finding local minimizers. Therefore, one usually considers constraint qualifications and stationarity conditions that are specifically tailored to MPCCs. In order to define some of these stationarity conditions, we first introduce the MPCC-tailored Lagrange function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ via

$$\mathcal{L}(x, \lambda, \eta, \mu, \nu) := f(x) + \lambda^\top g(x) + \eta^\top h(x) + \mu^\top G(x) + \nu^\top H(x).$$

Definition 2.5.1. Let $\bar{x} \in \mathbb{R}^n$ be given. We say that $\bar{x}$ is *weakly stationary* or *W-stationary* for (MPCC) if there exist multipliers $\bar{\lambda} \in \mathbb{R}^l$, $\bar{\eta} \in \mathbb{R}^m$, $\bar{\mu}, \bar{\nu} \in \mathbb{R}^p$ such that the conditions

$$\begin{aligned}
\mathcal{L}'_x(\bar{x}, \bar{\lambda}, \bar{\eta}, \bar{\mu}, \bar{\nu}) &= 0, \\
g(\bar{x}) &\leq 0, \\
h(\bar{x}) &= 0, \\
G(\bar{x}) &\geq 0, \\
H(\bar{x}) &\geq 0, \\
G(\bar{x})^\top H(\bar{x}) &= 0, \\
\bar{\lambda} &\geq 0, \\
\bar{\lambda}^\top g(\bar{x}) &= 0, \\
\bar{\mu}_i G_i(\bar{x}) &= 0 \qquad \forall i \in \{1, \ldots, p\}, \\
\bar{\nu}_i H_i(\bar{x}) &= 0 \qquad \forall i \in \{1, \ldots, p\}
\end{aligned}$$

are satisfied. We call the point $\bar{x}$ *C-stationary* if it is W-stationary and the multipliers $\bar{\mu}, \bar{\nu}$ satisfy the additional condition

$$\bar{\mu}_i \bar{\nu}_i \geq 0 \qquad \forall i \in \{1, \ldots, p\}$$

The point $\bar{x}$ is called *M-stationary* if it is W-stationary and additionally the condition

$$G_i(\bar{x}) = H_i(\bar{x}) = 0 \quad \Rightarrow \quad (\bar{\mu}_i < 0 \wedge \bar{\nu}_i < 0) \vee \bar{\mu}_i \bar{\nu}_i = 0 \qquad \forall i \in \{1, \ldots, p\}. \tag{2.28}$$

holds. Finally, for *strong stationarity* or *S-stationarity* of the point $\bar{x}$ we require that it is W-stationary and that the multipliers $\bar{\mu}, \bar{\nu}$ satisfy the condition

$$G_i(\bar{x}) = H_i(\bar{x}) = 0 \quad \Rightarrow \quad \bar{\mu}_i \leq 0 \wedge \bar{\nu}_i \leq 0 \qquad \forall i \in \{1, \ldots, p\}.$$

As for the relationship between these stationarity concepts one can easily verify the hierarchy

$$\text{S-stationary} \quad \Rightarrow \quad \text{M-stationary} \quad \Rightarrow \quad \text{C-stationary} \quad \Rightarrow \quad \text{W-stationary}.$$

These four notions of stationarity are equivalent if the biactive set $\{i \in \{1, \ldots, p\} \mid G_i(\bar{x}) = H_i(\bar{x}) = 0\}$ is empty.

For use in Section 5.2, we remark that there is an alternative but equivalent way to define M-stationarity. Namely, it can be seen that the condition (2.28) holds for a W-stationary point $\bar{x}$ if and only if there exist index sets $A, B \subset \{1, \dots, p\}$ such that the conditions

$$\{i \mid H_i(\bar{x}) > 0\} \subset A \subset B \subset \{i \mid G_i(\bar{x}) = 0\}, \tag{2.29a}$$

$$\bar{\mu} \in \{x \in \mathbb{R}^p \mid x_i \geq 0 \, \forall i \in B,\, x_i = 0 \, \forall i \in A\}^\circ, \tag{2.29b}$$

$$-\bar{\nu} \in \{x \in \mathbb{R}^p \mid x_i \geq 0 \, \forall i \in B,\, x_i = 0 \, \forall i \in A\} \tag{2.29c}$$

hold. Some of the stationarity conditions from Definition 2.5.1 have a relationship to the cones defined in Section 2.4. For this purpose we define the complementarity set $\mathbb{K} \subset \mathbb{R}^{2p}$ via

$$\mathbb{K} := \{(y, z) \in \mathbb{R}^p \times \mathbb{R}^p \mid 0 \leq y,\, 0 \leq z,\, y^\top z = 0\},$$

i.e. $x \in \mathbb{R}^n$ is a feasible point of (MPCC) if and only if $g(x) \leq 0$, $h(x) = 0$, and $(G(x), H(x)) \in \mathbb{K}$ hold.

Remark 2.5.2. Let $\bar{x}$ and multipliers $\bar{\lambda} \in \mathbb{R}^l$, $\bar{\eta} \in \mathbb{R}^m$, $\bar{\mu}, \bar{\nu} \in \mathbb{R}^p$ be given such that

$$\begin{aligned} \mathcal{L}'_x(\bar{x}, \bar{\lambda}, \bar{\eta}, \bar{\mu}, \bar{\nu}) &= 0, \\ g(\bar{x}) &\leq 0, \\ h(\bar{x}) &= 0, \\ (G(\bar{x}), H(\bar{x})) &\in \mathbb{K}, \\ \bar{\lambda} &\geq 0, \\ \bar{\lambda}^\top g(\bar{x}) &= 0 \end{aligned}$$

holds. Then we have the following relations between stationarity conditions and generalized normal cones.

(a) The point $\bar{x}$ is weakly stationary if and only if

$$(\bar{\mu}, \bar{\nu}) \in \mathcal{N}_{\mathbb{K}}^{\text{Clarke}}((G(\bar{x}), H(\bar{x})))$$

holds.

(b) The point $\bar{x}$ is M-stationary if and only if

$$(\bar{\mu}, \bar{\nu}) \in \mathcal{N}_{\mathbb{K}}^{\text{lim}}((G(\bar{x}), H(\bar{x})))$$

holds. The same is true if we use $\mathcal{N}_{\mathbb{K}}^{\text{s-lim}}$ instead of $\mathcal{N}_{\mathbb{K}}^{\text{lim}}$.

(c) The point $\bar{x}$ is strongly stationary if and only if

$$(\bar{\mu}, \bar{\nu}) \in \mathcal{N}_{\mathbb{K}}^{\text{Fréchet}}((G(\bar{x}), H(\bar{x})))$$

holds.

These results can be obtained by calculating the relevant cones.

For the problem (MPCC) it is known from the literature that a local minimizer is M-stationary if an MPCC-tailored constraint qualification holds. The next theorem is taken from [Flegel, Kanzow, 2006, Theorem 3.1]. For the definition of the MPCC-tailored Guignard constraint qualification, or MPCC-GCQ, we refer to [Flegel, Kanzow, 2006, Definition 2.1], where it is called MPEC-GCQ.

Theorem 2.5.3. Let $\bar{x} \in \mathbb{R}^d$ be a local minimizer of (MPCC) at which MPCC-GCQ holds. Then $\bar{x}$ is M-stationary.

The proof of this theorem utilizes the limiting normal cone. We remark that MPCC-GCQ is a very weak constraint qualification and that MPCC-GCQ holds at every point if the functions g, h, G, H are affine functions.

The M-stationarity condition can also be used for numerical algorithms, see [Harder, Mehlitz, G. Wachsmuth, 2020].

2.6 Capacity theory

In this section we want to explore the topic of capacity theory. Capacity theory has some similarities to measure theory. It will be particularly helpful for Chapter 5. For optimality conditions or stationarity systems involving functions from $H_0^1(\Omega)$ it can be seen that capacity theory is the right concept for pointwise conditions in Ω. As a contrast, the ordinary Lebesgue measure is often the right concept for pointwise conditions in Ω when working with functions in the space $L^p(\Omega)$, where $p \in [1, \infty]$. In [Harder, G. Wachsmuth, 2018a, Section 5.7] it is shown that conditions that hold pointwise almost everywhere are not suitable for stationarity conditions in Sobolev spaces and that conditions that hold pointwise "quasi-everywhere" are preferable.

Our goal for this section is to provide a self-contained presentation of the topic in such a way that it will be useful for our further studies in this thesis. We also give proofs for most results, although many results are not novel. Some lemmas and examples serve to enhance the understanding of capacity theory. The sections on capacity theory in [Harder, 2016; Bonnans, Shapiro, 2000; Harder, G. Wachsmuth, 2018a; c] influenced the content of this section.

2.6.1 Definition and basic properties

We start with the definition of the capacity of a set.

Definition 2.6.1. For a set $A \subset \Omega$ we define its *capacity* via

$$\operatorname{cap}(A) := \inf\{\|v\|^2_{H^1_0(\Omega)} \mid v \in H^1_0(\Omega),\ v \geq 1 \text{ a.e. in an open neighborhood of } A\}.$$

If there is no function $v \in H^1_0(\Omega)$ such that $v \geq 1$ a.e. in an open neighborhood of A, then we set $\operatorname{cap}(A) = \infty$.

This definition can be found in [Bonnans, Shapiro, 2000, Definition 6.47], however, it is only used to define the capacity of Borel measurable sets. Our definition for general subsets of Ω can also be found in slightly different formulations in [Attouch, Buttazzo, Michaille, 2014, Definition 5.8.1] (although technically the case $d = 1$ is not covered there), or [Fukushima, Oshima, Takeda, 2010, p. 66]. In the literature, there is some variety of definitions for the capacity of sets. For example, in [Delfour, Zolésio, 2011, Definition 8.6.2] the capacity is first defined on compact subsets of Ω, then extended to open subsets via supremums over compact subsets and finally extended to arbitrary subsets of Ω via infimums over open sets. This approach is equivalent to our definition of capacity, see Lemma 2.6.4. Our version of the definition of the capacity is simpler than the mentioned definitions from the literature.

It is also possible to extend the definition of the capacity to the spaces $W^{1,p}_0(\Omega)$ for $p \in [1, \infty)$ by replacing $\|v\|^2_{H^1_0(\Omega)}$ by $\|v\|^p_{W^{1,p}_0(\Omega)}$ in the above definition. However, for our purposes it suffices to only consider the case of $H^1_0(\Omega)$, i.e. $p = 2$.

There is an analogy that compares the capacity with the Lebesgue measure on Ω. For a Lebesgue measurable set $A \subset \Omega$ it is easy to see that

$$\operatorname{meas}(A) := \inf\{\|v\|^2_{L^2(\Omega)} \mid v \in L^2(\Omega),\ v \geq 1 \text{ a.e. in an open neighborhood of } A\}$$

holds. Thus, $\operatorname{meas}(\cdot)$ relates to $L^2(\Omega)$ similarly as $\operatorname{cap}(\cdot)$ relates to $H^1_0(\Omega)$.

We give some examples to highlight the difference between measure and capacity. However, the calculations to prove these claims have been omitted.

Example 2.6.2. (a) Let $a < \omega_0 < b$ be real numbers. Then for $d := 1$ and $\Omega := (a, b)$ we have

$$\operatorname{cap}(\{\omega_0\}) = \frac{1}{\omega_0 - a} + \frac{1}{b - \omega_0} > 0.$$

(b) For $d \geq 2$ and $\omega_0 \in \Omega$, the singleton $\{\omega_0\}$ has zero capacity.

(c) The domain Ω has infinite capacity.

As can be seen by Example 2.6.2 (a), there can be sets that have measure zero but a nonzero capacity. We also observe that in Example 2.6.2 (a) the capacity of a set can depend on the surrounding domain Ω.

In the next lemma, we will collect some basic properties of capacities, which will also illustrate the relationship to the Lebesgue measure.

Lemma 2.6.3. (a) The empty set has zero capacity.

(b) For sets A, B with $A \subset B \subset \Omega$ we have $\operatorname{cap}(A) \leq \operatorname{cap}(B)$, i.e. the capacity is monotone.

(c) For sets $A, B \subset \Omega$ we have

$$\operatorname{cap}(A \cup B) + \operatorname{cap}(A \cap B) \leq \operatorname{cap}(A) + \operatorname{cap}(B).$$

In particular, the inequality $\operatorname{cap}(A \cup B) \leq \operatorname{cap}(A) + \operatorname{cap}(B)$ holds for all subsets $A, B \subset \Omega$, i.e. the capacity is subadditive.

(d) For a nondecreasing sequence of sets $\{A_i\}_{i\in\mathbb{N}}$ with $A_i \subset \Omega$ the capacity of the union of this sequence can be expressed via

$$\operatorname{cap}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \lim_{i\to\infty} \operatorname{cap}(A_i),$$

i.e. the capacity is continuous from below.

(e) For a sequence of sets $\{A_i\}_{i\in\mathbb{N}}$ with $A_i \subset \Omega$ the inequality

$$\operatorname{cap}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) \leq \sum_{i\in\mathbb{N}} \operatorname{cap}(A_i)$$

holds, i.e. the capacity is countably subadditive.

(f) There exists a constant $C > 0$ (that depends only on Ω) such that for every Lebesgue measurable set $A \subset \Omega$ the inequality $\operatorname{meas}(A) \leq C \operatorname{cap}(A)$ holds.

(g) If a set has zero capacity, then it is included in a Borel measurable set of measure zero.

Proof. Parts (a) and (b) follow directly from the definition.

For part (c), let $v_A, v_B \in H_0^1(\Omega)$ be such that $v_A \geq 1$ a.e. on an open neighborhood of A and $v_B \geq 1$ a.e. on an open neighborhood of B. Then the function $v_\cup := \max(v_A, v_B) \in H_0^1(\Omega)$ satisfies $v_\cup \geq 1$ a.e. on an open neighborhood of $A \cup B$ and the function $v_\cap := \min(v_A, v_B) \in H_0^1(\Omega)$ satisfies $v_\cap \geq 1$ a.e. on an open neighborhood of $A \cap B$. Due to Lemma 2.2.8 (d) we also have the equality $\|v_\cup\|^2_{H_0^1(\Omega)} + \|v_\cap\|^2_{H_0^1(\Omega)} = \|v_A\|^2_{H_0^1(\Omega)} + \|v_B\|^2_{H_0^1(\Omega)}$. The claim follows by taking the infimum over the functions v_A, v_B that are admissible in the definition of the capacity of A and B.

We continue with part (d). From part (b) we obtain that the inequality

$$\operatorname{cap}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) \geq \lim_{i\to\infty} \operatorname{cap}(A_i)$$

holds. We also observe that $\sup_{i\in\mathbb{N}} \operatorname{cap}(A_i) = \lim_{i\to\infty} \operatorname{cap}(A_i)$ holds. In order to show the remaining inequality, we can without loss of generality assume that $\sup_{i\in\mathbb{N}} \operatorname{cap}(A_i) < \infty$ holds. Let $\varepsilon > 0$ be given. For each $i \in \mathbb{N}$, let $\tilde{O}_i \subset \Omega$ be an open set such that $A_i \subset \tilde{O}_i$ and $\operatorname{cap}(\tilde{O}_i) \leq \operatorname{cap}(A_i) + 2^{-i}\varepsilon$ hold. It is possible to find such an open set $\tilde{O}_i$ because the functions $v \in H_0^1(\Omega)$ in the definition of the capacity for the set A_i have the property $v \geq 1$ a.e. on an open neighborhood of A_i. Next, we recursively define a sequence $\{O_i\}_{i\in\mathbb{N}}$ of open subsets of Ω via $O_1 := \tilde{O}_1$, $O_{i+1} := O_i \cup \tilde{O}_{i+1}$ for all $i \in \mathbb{N}$. Clearly, $A_i \subset O_i$ is true for all $i \in \mathbb{N}$. We claim that

$$\operatorname{cap}(O_i) \leq \operatorname{cap}(A_i) + (1 - 2^{-i})\varepsilon \tag{2.30}$$

holds for all $i \in \mathbb{N}$. Indeed, the claim is clearly true for $i = 1$. If the claim (2.30) is true for $i \in \mathbb{N}$, then we can use part (c) to obtain

$$\begin{aligned}\operatorname{cap}(O_{i+1}) &\leq \operatorname{cap}(O_i) + \operatorname{cap}(\tilde{O}_{i+1}) - \operatorname{cap}(O_i \cap \tilde{O}_{i+1}) \\ &\leq \operatorname{cap}(A_i) + (1 - 2^{-i})\varepsilon + \operatorname{cap}(A_{i+1}) + 2^{-(i+1)}\varepsilon - \operatorname{cap}(A_i) \\ &= \operatorname{cap}(A_{i+1}) + (1 - 2^{-(i+1)})\varepsilon,\end{aligned}$$

which shows that (2.30) also holds for $i+1$. According to the definition of $\operatorname{cap}(O_i)$, there exists a sequence $\{v_i\}_{i\in\mathbb{N}}$ in $H_0^1(\Omega)$ such that $v_i \geq 1$ a.e. on O_i and $\|v_i\|^2_{H_0^1(\Omega)} \leq \operatorname{cap}(A_i)+\varepsilon$ holds for all $i \in \mathbb{N}$. Clearly, $\{v_i\}_{i\in\mathbb{N}}$ is a bounded sequence in $H_0^1(\Omega)$ and therefore has a subsequence that converges weakly in $H_0^1(\Omega)$ and strongly in $L^2(\Omega)$ to some function $v \in H_0^1(\Omega)$. This subsequence has a further subsequence that converges pointwise almost everywhere to v. Without loss of generality we again denote this subsequence by $\{v_i\}_{i\in\mathbb{N}}$. Because $\{O_i\}_{i\in\mathbb{N}}$ is a nondecreasing sequence of open sets, we obtain that $v \geq 1$ holds a.e. on the open set $O := \bigcup_{i\in\mathbb{N}} O_i$. Then the weak convergence in $H_0^1(\Omega)$ and $\bigcup_{i\in\mathbb{N}} A_i \subset O$ imply

$$\operatorname{cap}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) \leq \operatorname{cap}(O) \leq \|v\|^2_{H_0^1(\Omega)} \leq \liminf_{i\to\infty} \|v_i\|^2_{H_0^1(\Omega)} \leq \lim_{i\to\infty} \operatorname{cap}(A_i) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the claim follows.

For part (e) we note that a repeated application of part (c) yields

$$\operatorname{cap}\Big(\bigcup_{j=1}^{i} A_j\Big) \leq \sum_{j=1}^{i} \operatorname{cap}(A_j)$$

for all $i \in \mathbb{N}$. Then the claim follows by applying part (d) to the nondecreasing sequence $\{\bigcup_{j=1}^{i} A_j\}_{i\in\mathbb{N}}$.

For part (f), let $A \subset \Omega$ be Lebesgue measurable. We only need to consider the case where $\operatorname{cap}(A) < \infty$. Let $v \in H_0^1(\Omega)$ be a function such that $v \geq 1$ a.e. in an open neighborhood

of A. Then, by the Poincaré inequality, we have

$$\operatorname{meas}(A) \leq \|v\|^2_{L^2(\Omega)} \leq C\|v\|^2_{H^1_0(\Omega)},$$

where the constant only depends on Ω. Taking the infimum over all $v \in H^1_0(\Omega)$ such that $v \geq 1$ a.e. in an open neighborhood of A yields the desired inequality.

Finally, for a set $A \subset \Omega$ with zero capacity it can be concluded from the definition, that there is a sequence $\{O_i\}_{i\in\mathbb{N}}$ of open supersets of A with arbitrarily small capacity. By part (f), those supersets must have arbitrarily small Lebesgue measure. Hence, A is included in the Borel measurable set $\bigcap_{i\in\mathbb{N}} O_i$ which has measure zero. This concludes the proof of part (g).

Note that Lemma 2.6.3 (g) is not just a special case of Lemma 2.6.3 (f), because the measurability of the set is not an assumption in Lemma 2.6.3 (g).

We further note that the capacity satisfies the definition of an outer measure (which is defined precisely via the properties shown in parts (a), (b), and (e) of Lemma 2.6.3). Thus, the capacity can be considered as a generalization of a measure.

In Definition 2.2.15 we introduced regularity properties for measures. We can show outer regularity for arbitrary sets and inner regularity for open sets in a straightforward way.

Lemma 2.6.4. (a) The capacity is outer regular in the sense that

$$\operatorname{cap}(A) = \inf\{\operatorname{cap}(O) \mid A \subset O \subset \Omega,\ O \text{ open}\}$$

holds for all $A \subset \Omega$.

(b) The capacity is inner regular for open sets in the sense that

$$\operatorname{cap}(O) = \sup\{\operatorname{cap}(K) \mid K \subset O \subset \Omega,\ K \text{ compact}\}$$

holds for all open sets $O \subset \Omega$.

Proof. Part (a) follows directly from the definition of the capacity because the functions $v \in H^1_0(\Omega)$ in the definition of the capacity for the set A have the property $v \geq 1$ a.e. on an open set $O \supset A$.

We recall that every open set $O \subset \Omega$ has a compact exhaustion, i.e. there exists a sequence $\{K_i\}_{i\in\mathbb{N}}$ of compact subsets of O such that $O = \bigcup_{i\in\mathbb{N}} K_i$. Then part (b) follows from Lemma 2.6.3 (d).

In the next definition, we introduce some notions that are similar to the expressions "almost everywhere" and "for almost all", but are based on capacities instead of measures.

Definition 2.6.5. (a) We say that a property P that depends on $\omega \in \Omega$ holds *quasi-everywhere* (q.e.) on a set $A \subset \Omega$ if the set $\{\omega \in A \mid P(\omega) \text{ does not hold}\}$ has zero capacity. We will also say that $P(\omega)$ holds for *quasi-all* (q.a.) $\omega \in A$, which has the same meaning. If no such set A is specified in this context, we mean the set $A = \Omega$.

(b) For sets $A, B \subset \Omega$ we write $A \subset_q B$ if $\operatorname{cap}(A \setminus B) = 0$.

(c) For sets $A, B \subset \Omega$ we write $A =_q B$ if $\operatorname{cap}((B \setminus A) \cup (A \setminus B)) = 0$.

(d) We say that a set $A \subset \Omega$ is *unique up to a set of zero capacity* with respect to a property P if A satisfies P and we have $A =_q B$ for every set $B \subset \Omega$ that also satisfies P.

We give some simple results to show that the relations "$=_q$" and "$\subset_q$" behave as one might expect.

Corollary 2.6.6. (a) For $A, B \subset \Omega$ we have $A =_q B$ if and only if $A \subset_q B$ and $B \subset_q A$.

(b) For $A, B \subset \Omega$ we have $A \subset_q B$ if $A \subset B$ and $A =_q B$ if $A = B$.

(c) The relation "$=_q$" is an equivalence relation.

(d) The relation "$\subset_q$" is reflexive and transitive.

These statements follow directly from a combination of Lemma 2.6.3 (a), Lemma 2.6.3 (c), and the definitions of "$=_q$" and "$\subset_q$".

2.6.2 Quasi-open sets and quasi-continuous functions

We move to the topic of quasi-open and quasi-closed sets as well as quasi-continuous functions. These objects have similarities with ordinary open/closed sets or continuous functions, but in some situations these quasi-concepts are more useful. We start with the definition, which relies on capacity theory discussed in Section 2.6.1.

Definition 2.6.7. (a) A set $O \subset \Omega$ is called *quasi-open* if for every $\varepsilon > 0$ there is an open set $G_\varepsilon \subset \Omega$ with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ such that $O \cup G_\varepsilon$ is open.

(b) A set $A \subset \Omega$ is called *quasi-closed* if $\Omega \setminus A$ is quasi-open.

(c) A function $f : \Omega \to \mathbb{R}$ is called *quasi-continuous* if for every $\varepsilon > 0$ there is an open set $G_\varepsilon \subset \Omega$ with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ such that f is continuous on $\Omega \setminus G_\varepsilon$.

It is also possible to define quasi-open or quasi-continuous without the requirement that the set G_ε is open, see Lemma 2.6.8 (c) and Lemma 2.6.8 (d).

By choosing $G_\varepsilon = \emptyset$ in the definition above, it can be seen that open sets $O \subset \Omega$ are quasi-open and continuous functions $f : \Omega \to \mathbb{R}$ are quasi-continuous.

Next, we show equivalences to alternative definitions for the notions "quasi-open", "quasi-closed", and "quasi-continuous". For example, Lemma 2.6.8 (b) is used as a definition for quasi-continuous functions in [Bonnans, Shapiro, 2000, Definition 6.47], whereas Lemma 2.6.8 (e) is used as a definition for quasi-open and quasi-closed sets in [Fuglede, 1971, Definition 2.1].

Lemma 2.6.8. (a) A set O is quasi-open if and only if there is a nonincreasing sequence $\{G_i\}_{i\in\mathbb{N}}$ of open subsets of Ω such that $O \cup G_i$ is open for every $i \in \mathbb{N}$ and $\operatorname{cap}(G_i) \to 0$ as $i \to \infty$.

(b) A function $f : \Omega \to \mathbb{R}$ is quasi-continuous if and only if there is a nonincreasing sequence $\{G_i\}_{i\in\mathbb{N}}$ of open subsets of Ω such that f is continuous on $\Omega \setminus G_i$ for all $i \in \mathbb{N}$ and $\operatorname{cap}(G_i) \to 0$ as $i \to \infty$.

(c) A set $O \subset \Omega$ is quasi-open if and only if for every $\varepsilon > 0$ there is a set $G_\varepsilon \subset \Omega$ with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ such that $O \cup G_\varepsilon$ is open.

(d) A function $f : \Omega \to \mathbb{R}$ is quasi-continuous if and only if for every $\varepsilon > 0$ there is a set $G_\varepsilon \subset \Omega$ with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ such that f is continuous on $\Omega \setminus G_\varepsilon$.

(e) A set $O \subset \Omega$ is quasi-open if and only if

$$\inf\{\operatorname{cap}((O \setminus \tilde{O}) \cup (\tilde{O} \setminus O)) \mid \tilde{O} \subset \Omega \text{ open}\} = 0.$$

A set $A \subset \Omega$ is quasi-closed if and only if

$$\inf\{\operatorname{cap}((A \setminus \tilde{A}) \cup (\tilde{A} \setminus A)) \mid \tilde{A} \subset \Omega \text{ relatively closed}\} = 0.$$

Proof. For part (a), let O be quasi-open. For every $j \in \mathbb{N}$, we find an open set $\tilde{G}_j \subset \Omega$ such that $O \cup \tilde{G}_j$ is open and $\operatorname{cap}(\tilde{G}_j) < 1/j$. Then, for $i \in \mathbb{N}$, we define $G_i := \bigcap_{j=1}^{i} \tilde{G}_j$. It is clear that G_i and $O \cup G_i$ are open for every $i \in \mathbb{N}$ and that $\{G_i\}_{i\in\mathbb{N}}$ is a nonincreasing sequence. Due to Lemma 2.6.3 (b) we also obtain $\operatorname{cap}(G_i) < 1/i$ for all $i \in \mathbb{N}$ and therefore $\operatorname{cap}(G_i) \to 0$ as $i \to \infty$. Thus, we have shown all the necessary properties of $\{G_i\}_{i\in\mathbb{N}}$. The other direction of part (a) follows directly.

For part (b), let $f : \Omega \to \mathbb{R}$ be a quasi-continuous function. For every $j \in \mathbb{N}$, we find an open set $\tilde{G}_j \subset \Omega$ such that f is continuous on $\Omega \setminus \tilde{G}_j$ and $\operatorname{cap}(\tilde{G}_j) < 1/j$. Then, for $i \in \mathbb{N}$, we define $G_i := \bigcap_{j=1}^{i} \tilde{G}_j$. It is clear that G_i is open and that f is continuous on $\Omega \setminus G_i$ for every $i \in \mathbb{N}$ and that $\{G_i\}_{i\in\mathbb{N}}$ is a nonincreasing sequence. Due to Lemma 2.6.3 (b) we also obtain $\operatorname{cap}(G_i) < 1/i$ for all $i \in \mathbb{N}$ and therefore $\operatorname{cap}(G_i) \to 0$ as $i \to \infty$. Thus, we have shown all the necessary properties of $\{G_i\}_{i\in\mathbb{N}}$. The other direction of part (b) follows directly.

For parts (c) and (d) it suffices to show that for each set $G_\varepsilon \subset \Omega$ with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ there

is an open set $\tilde{G}_\varepsilon \subset \Omega$ with $G_\varepsilon \subset \tilde{G}_\varepsilon$ and $\operatorname{cap}(\tilde{G}_\varepsilon) < \varepsilon$. This, however, is true due to Lemma 2.6.4 (a).

For part (e) we focus on the claim regarding quasi-open sets. The claim for quasi-closed sets follows by taking complements (with respect to Ω) of open or quasi-open sets. Suppose $O \subset \Omega$ is given such that $\inf\{\operatorname{cap}((O \setminus \tilde{O}) \cup (\tilde{O} \setminus O)) \mid \tilde{O} \subset \Omega \text{ open}\} = 0$. Let $\varepsilon > 0$ be given. Then there is an open set $\tilde{O}$ such that $\operatorname{cap}((O \setminus \tilde{O}) \cup (\tilde{O} \setminus O)) < \varepsilon$. Due to Lemma 2.6.4 (a) we can find an open set $G_\varepsilon \subset \Omega$ with $(O \setminus \tilde{O}) \cup (\tilde{O} \setminus O) \subset G_\varepsilon$ and $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Then, using elementary considerations, one can show that $O \cup G_\varepsilon = \tilde{O} \cup G_\varepsilon$ holds. Therefore, the set $O \cup G_\varepsilon$ is open. Since $\varepsilon > 0$ was chosen arbitrarily, it follows that O is quasi-open.

Conversely, suppose O is quasi-open and let $\varepsilon > 0$ be given. Then by part (c) there is a set G_ε such that $O \cup G_\varepsilon$ is open and $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Now we can choose the open set $\tilde{O} := O \cup G_\varepsilon$ and observe that $\operatorname{cap}((O \setminus \tilde{O}) \cup (\tilde{O} \setminus O)) = \operatorname{cap}(G_\varepsilon \setminus O) < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, this proves the claim.

It is well-known that a function $f : \Omega \to \mathbb{R}$ is continuous if and only if all preimages of open sets under f are open. The question arises whether an analogous result holds true for quasi-continuous functions and quasi-open sets. As the following lemma shows, this is indeed true. A similar result can be found in [Kilpeläinen, Malý, 1992, Theorem 1.4].

Lemma 2.6.9. A function $f : \Omega \to \mathbb{R}$ is quasi-continuous if and only if all preimages of open sets under f are quasi-open.

Proof. Let $f : \Omega \to \mathbb{R}$ be a quasi-continuous function and let $O \subset \mathbb{R}$ be an open set. We want to show that $f^{-1}(O)$ is quasi-open. Let $\varepsilon > 0$ be given. Then there exists an open set $G_\varepsilon \subset \Omega$ such that f is continuous on $\Omega \setminus G_\varepsilon$ and $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Therefore, $f^{-1}(O)$ is relatively open with respect to $\Omega \setminus G_\varepsilon$. Thus, the set $f^{-1}(O) \cup G_\varepsilon$ is open. Since $\varepsilon > 0$ was arbitrary, it follows that $f^{-1}(O)$ is quasi-open.

Now we assume that all preimages of open sets under a function $f : \Omega \to \mathbb{R}$ are quasi-open and we want to show that f is quasi-continuous. Let $\varepsilon > 0$ be given. We consider intervals $(a, b) \subset \mathbb{R}$ with $a, b \in \mathbb{Q}$. Since $\mathbb{Q} \times \mathbb{Q}$ is countable, there exists a bijection $n : \mathbb{Q} \times \mathbb{Q} \to \mathbb{N}$. For given points $a, b \in \mathbb{Q}$ we know that $f^{-1}((a, b))$ is quasi-open and therefore there exists an open set $G_{\varepsilon,a,b} \subset \Omega$ such that $f^{-1}((a, b)) \cup G_{\varepsilon,a,b}$ is open and $\operatorname{cap}(G_{\varepsilon,a,b}) < \varepsilon 2^{-n(a,b)}$. We define $G_\varepsilon := \bigcup_{a,b\in\mathbb{Q}} G_{\varepsilon,a,b}$ and note that Lemma 2.6.3 (e) implies that $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Clearly, G_ε is also open. We observe that the preimage under f of every interval in $\mathbb{R}$ with rational endpoints is relatively open in $\Omega \setminus G_\varepsilon$. Since every open set in $\mathbb{R}$ can be written as the union of intervals with rational endpoints, it follows that the preimage under f of every open subset of $\mathbb{R}$ is relatively open in $\Omega \setminus G_\varepsilon$. Thus, f is continuous on $\Omega \setminus G_\varepsilon$. Since $\varepsilon > 0$ was arbitrary, f is quasi-continuous.

It is well-known that arbitrary unions and finite intersections of open sets are open. For quasi-open sets we can show something similar.

Lemma 2.6.10. The countable union of quasi-open sets is quasi-open. The intersection of finitely many quasi-open sets is quasi-open.

Proof. Let $\{O_i\}_{i\in\mathbb{N}}$ be a sequence of quasi-open sets in Ω and let $\varepsilon > 0$ be given. For every $i \in \mathbb{N}$ we can find an open set $G_{\varepsilon,i} \subset \Omega$ such that $O_i \cup G_{\varepsilon,i}$ is open and $\operatorname{cap}(G_{\varepsilon,i}) < 2^{-i}\varepsilon$. Then we define $G_\varepsilon := \bigcup_{i\in\mathbb{N}} G_{\varepsilon,i}$ and note that Lemma 2.6.3 (e) implies that $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Because the sets G_ε and $G_\varepsilon \cup \bigcup_{i\in\mathbb{N}} O_i$ are open, the claim follows.

For the finite intersection of quasi-open sets it suffices to show that the intersection $O_1 \cap O_2$ of two quasi-open sets $O_1, O_2 \subset \Omega$ is quasi-open. Again, let $\varepsilon > 0$ be given. For $i = 1, 2$ we can find open sets $G_{\varepsilon,i} \subset \Omega$ such that $O_i \cup G_{\varepsilon,i}$ is open and $\operatorname{cap}(G_{\varepsilon,i}) < \varepsilon/2$. Then the open set $G_\varepsilon := G_{\varepsilon,1} \cup G_{\varepsilon,2}$ satisfies that $(O_1 \cap O_2) \cup G_\varepsilon$ is open and $\operatorname{cap}(G_\varepsilon) < \varepsilon$ follows from Lemma 2.6.3 (c). Thus, $O_1 \cap O_2$ is quasi-open.

One might wonder if uncountable unions of quasi-open sets are quasi-open. The next example shows that this is not always the case.

Example 2.6.11. (a) If $d = 2$ and $\Omega = \operatorname{int} B_2(0) \subset \mathbb{R}^2$ then the compact set $B_1(0) \subset \Omega$ is not quasi-open.

(b) We consider the case that $d = 2$ and $\Omega = \operatorname{int} B_2(0) \subset \mathbb{R}^2$. Then the set $\{\omega\}$ is quasi-open for all $\omega \in \Omega$. However, the uncountable union $B_1(0) = \bigcup_{\omega\in B_1(0)} \{\omega\}$ is not quasi-open.

We omit a detailed proof of the claims in this example.

Due to Example 2.6.11 (b) we know that the quasi-open sets are not a topology. However, in the literature there exists the concept of the fine topology, which is defined as the coarsest topology on $\mathbb{R}^n$ which makes all superharmonic functions continuous. We refer to [Heinonen, Kilpeläinen, Martio, 1993, Chapter 12] for details. The finely open sets are related to quasi-open sets. In particular it is possible to show that for every quasi-open set there exists a finely open set that differs only by a set of zero capacity.

From Lemma 2.6.3 (g) it is clear that if a property is true q.e. on a set, it also true a.e. on that set. The next lemma shows, that in some situations we also obtain the other direction.

Lemma 2.6.12. (a) Let $O \subset \Omega$ be a quasi-open set. Then O is also a Lebesgue measurable set.

(b) Let $O \subset \Omega$ be a quasi-open set with measure zero. Then O has zero capacity.

(c) Let $O \subset \Omega$ be a quasi-open set and let $f : \Omega \to \mathbb{R}$ be quasi-continuous. Then the equivalences

$$f \geq 0 \text{ a.e. on } O \qquad \Leftrightarrow \qquad f \geq 0 \text{ q.e. on } O$$

and

$$f = 0 \text{ a.e. on } O \qquad \Leftrightarrow \qquad f = 0 \text{ q.e. on } O$$

hold.

Proof. For part (a), let $\{G_i\}_{i\in\mathbb{N}}$ be chosen according to Lemma 2.6.8 (a), i.e. it is a nonincreasing sequence of open subsets of Ω such that $O \cup G_i$ is open for all $i \in \mathbb{N}$ and $\operatorname{cap}(G_i) \to 0$ holds. Then the set $\tilde{O} := O \cup \bigcap_{i\in\mathbb{N}} G_i$ is Borel measurable, and we have $\tilde{O} \setminus O \subset \bigcap_{i\in\mathbb{N}} G_i$. Due to $\operatorname{cap}(G_i) \to 0$ as $i \to \infty$, this implies $\operatorname{cap}(\tilde{O} \setminus O) = 0$. From Lemma 2.6.3 (g) we obtain that $\tilde{O} \setminus O$ is Lebesgue measurable with Lebesgue measure 0. The claim follows from $O = \tilde{O} \setminus (\tilde{O} \setminus O)$.

For part (b), we first remark that O is Lebesgue measurable due to part (a). Let $\varepsilon > 0$ be arbitrary and let $G_\varepsilon \subset \Omega$ be an open set with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ such that $O \cup G_\varepsilon$ is open. Furthermore, let $v \in H_0^1(\Omega)$ be such that $v \geq 1$ a.e. on G_ε and $\|v\|^2_{H_0^1(\Omega)} < \varepsilon$. Since O has Lebesgue measure zero, we also have that $v \geq 1$ a.e. on $O \cup G_\varepsilon$. Thus, $\operatorname{cap}(O) \leq \operatorname{cap}(O \cup G_\varepsilon) < \varepsilon$ holds. Since $\varepsilon > 0$ was arbitrary this implies that $\operatorname{cap}(O) = 0$.

The "$\Leftarrow$" directions of part (c) follow from Lemma 2.6.3 (g). For the "$\Rightarrow$" directions we note that the sets $\{f < 0\} \cap O, \{f \neq 0\} \cap O$ are quasi-open according to Lemmas 2.6.9 and 2.6.10. Then the result follows directly from part (b).

The next proposition will illustrate how the concept of quasi-continuous functions is helpful in the Sobolev space $H_0^1(\Omega)$. Before that, we provide the definition of quasi-continuous representatives of functions in $H_0^1(\Omega)$.

Definition 2.6.13. Let $v \in H_0^1(\Omega)$ be given. A quasi-continuous and Borel measurable function $\tilde{v} : \Omega \to \mathbb{R}$ with the property that $v = \tilde{v}$ a.e. in Ω is called a *quasi-continuous representative* of v.

As the next proposition shows, a quasi-continuous representative of functions in $H_0^1(\Omega)$ always exists. This result can be found in [Bonnans, Shapiro, 2000, Lemma 6.50] or [Heinonen, Kilpeläinen, Martio, 1993, Theorem 4.4]

Proposition 2.6.14. For every function $v \in H_0^1(\Omega)$ there exists a quasi-continuous representative. Two quasi-continuous representatives of a function $v \in H_0^1(\Omega)$ are equal quasi-everywhere.

Proof. Because $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, there exists a sequence of functions $v_i \in C_c^\infty(\Omega)$ such that $v_i \to v$ in $H_0^1(\Omega)$. Without loss of generality we can assume that $v_i \to v$ pointwise a.e. in Ω and $\|v_i - v_{i+1}\|_{H_0^1(\Omega)} \leq 4^{-i}$ for all $i \in \mathbb{N}$. We define the open sets

$$\tilde{G}_i := \{|v_i - v_{i+1}| > 2^{-i}\}, \qquad G_i := \bigcup_{j\geq i} \tilde{G}_j,$$

for each $i \in \mathbb{N}$ and the set $G_\infty := \bigcap_{i\in\mathbb{N}} G_i$. By using the function $2^i|v_i - v_{i+1}|$ in the definition of the capacity it can be seen that $\operatorname{cap}(\tilde{G}_i) \leq 4^{-i}$ and hence $\operatorname{cap}(G_i) \to 0$ hold.

Now we fix $i \in \mathbb{N}$. We note that $|v_j - v_{j+1}| \leq 2^{-j}$ holds everywhere on $\Omega \setminus G_i$ for all $j \geq i$. Thus, the sequence of continuous functions $\{v_j\}_{j\in\mathbb{N}}$ converges uniformly on $\Omega \setminus G_i$ and therefore there exists a continuous function $\tilde{v}_i$ as the (uniform) limit of $\{v_j\}_{j\in\mathbb{N}}$ on $\Omega \setminus G_i$. By doing the same for $i+1$ it is clear that $\tilde{v}_{i+1} = \tilde{v}_i$ on $\Omega \setminus G_i$. By recursive extension we can define the Borel measurable function $\tilde{v}$ such that $\tilde{v} = \tilde{v}_i$ on $\Omega \setminus G_i$ and $\tilde{v} = 0$ on G_∞. Since G_i is a nonincreasing sequence, it follows that $\tilde{v}$ is quasi-continuous by Lemma 2.6.8 (b). Since $\operatorname{meas}(G_\infty) = \operatorname{cap}(G_\infty) = 0$ it follows from the pointwise convergences that $\tilde{v} = v$ holds a.e. on Ω. Thus, we have shown that $\tilde{v}$ is a quasi-continuous representative of v.

It remains to consider that there is another quasi-continuous representative $\tilde{w}$ of v. Then Lemma 2.6.12 (c) applied to $\tilde{w} - \tilde{v}$ implies that $\tilde{w} = \tilde{v}$ q.e. in Ω.

For the rest of this thesis, whenever we talk about a function $v \in H_0^1(\Omega)$ we refer to a quasi-continuous representative of v. One example of the benefit of this convention is that we can use Lemma 2.6.12 (c) on Ω to obtain the helpful identity

$$H_0^1(\Omega)_+ = \{v \in H_0^1(\Omega) \mid v \geq 0 \text{ q.e. in } \Omega\}.$$

For a converging sequence in a Lebesgue space one can find a pointwise a.e.-converging subsequence. It will be helpful to have a similar result for convergent sequences in $H_0^1(\Omega)$. The following lemma shows that we can extract a subsequence that converges pointwise except on a set of zero capacity. This will be useful for showing that certain subsets of $H_0^1(\Omega)$ are closed. The statement can also be found in [Bonnans, Shapiro, 2000, Lemma 6.52]. However, our proof is different.

Lemma 2.6.15. Let $\{v_i\}_{i\in\mathbb{N}}$ be a sequence of functions in $H_0^1(\Omega)$ that converge to a function $v \in H_0^1(\Omega)$ in $H_0^1(\Omega)$-norm. Then there exists a subsequence of $\{v_i\}_{i\in\mathbb{N}}$ that converges pointwise quasi-everywhere to v.

Proof. Without loss of generality we can assume that $v = 0$. Since we are only looking for a subsequence, we can also assume that $\|v_i\|^2_{H_0^1(\Omega)} < 2^{-i}$.

The set of points $\omega \in \Omega$ where $\{v_i(\omega)\}_{i\in\mathbb{N}}$ does not converge to 0 can be written as

$$B := \bigcup_{k\in\mathbb{N}} \bigcap_{i\in\mathbb{N}} \bigcup_{j\geq i} \{|v_j| > 1/k\}.$$

Let $k \in \mathbb{N}$ be fixed for now. For $i \in \mathbb{N}$ we consider the set $O_i := \{|v_i| > 1/k\}$ which is quasi-open according to Lemma 2.6.9. Thus, there is an open set $G_i \subset \Omega$ such that $O_i \cup G_i$ is open and a function $w_i \in H_0^1(\Omega)$ with $w_i \geq 1$ a.e. on G_i and $\|w_i\|^2_{H_0^1(\Omega)} < k^2 2^{-i}$.

Then we have $k|v_i| + w_i^+ \geq 1$ a.e. on $O_i \cup G_i$. The estimate

$$\operatorname{cap}(O_i) \leq \operatorname{cap}(O_i \cup G_i) \leq \|k|v_i| + w_i^+\|^2_{H_0^1(\Omega)} \leq 4k^2 2^{-i}$$

follows. Due to Lemma 2.6.3 (e) we obtain that

$$\operatorname{cap}\Big(\bigcup_{j\geq i}\{|v_j| > 1/k\}\Big) = \operatorname{cap}\Big(\bigcup_{j\geq i} O_j\Big) \leq \sum_{j\geq i} k^2 2^{-j} = k^2 2^{1-i} \to 0$$

as $i \to \infty$. Thus, the set $B_k := \bigcap_{i\in\mathbb{N}} \bigcup_{j\geq i}\{|v_j| > 1/k\}$ has zero capacity. Another application of Lemma 2.6.3 (e) yields that $\operatorname{cap}(B) = \operatorname{cap}(\bigcup_{k\in\mathbb{N}} B_k) = 0$, which completes the proof.

We can see from the proof that we do not need to consider subsequences if the sequence converges already fast enough.

The next lemma is an application of Lemma 2.6.15. It gives us an alternative description of the capacity that has similarities to the original definition but utilizes the "q.e."-notion of capacity theory and relies on our convention to consider only quasi-continuous representatives of functions in $H_0^1(\Omega)$.

Lemma 2.6.16. For a set $A \subset \Omega$ its capacity can be described by

$$\operatorname{cap}(A) = \inf\{\|v\|^2_{H_0^1(\Omega)} \mid v \in H_0^1(\Omega),\ 0 \leq v \leq 1 \text{ q.e. on } \Omega \text{ and } v = 1 \text{ q.e. on } A\}$$

and if A has finite capacity then the infimum is attained at a unique minimizer.

Proof. We roughly follow the proof of [Harder, G. Wachsmuth, 2018a, Lemma 3.4 (d)]. We define the set

$$M := \{v \in H_0^1(\Omega) \mid 0 \leq v \leq 1 \text{ q.e. on } \Omega \text{ and } v = 1 \text{ q.e. on } A\}.$$

Let $v \in M$ and $\varepsilon \in (0,1)$ be given. Then the set $\{v > 1-\varepsilon\}$ is quasi-open. Therefore, there exists an open set G_ε such that $\{v > 1-\varepsilon\} \cup G_\varepsilon$ is open and $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Thus, there exists a function $w_\varepsilon \in H_0^1(\Omega)$ such that $w_\varepsilon \geq 1$ a.e. on G_ε and $\|w_\varepsilon\|^2_{H_0^1(\Omega)} < \varepsilon$. Then the function $v_\varepsilon := (1-\varepsilon)^{-1}v + w_\varepsilon^+$ satisfies $\|v_\varepsilon\|_{H_0^1(\Omega)} \leq (1-\varepsilon)^{-1}\|v\|_{H_0^1(\Omega)} + \varepsilon^{1/2}$ and $v_\varepsilon \geq 1$ a.e. on $\{v > 1-\varepsilon\} \cup G_\varepsilon$, which is an open neighborhood of A. Thus, $\operatorname{cap}(A) \leq \big((1-\varepsilon)^{-1}\|v\|_{H_0^1(\Omega)} + \varepsilon^{1/2}\big)^2$ holds. By taking the infimum over $\varepsilon > 0$ and $v \in M$ it follows that $\operatorname{cap}(A) \leq \inf\{\|v\|^2_{H_0^1(\Omega)} \mid v \in M\}$.

For the other inequality let $w \in H_0^1(\Omega)$ be given such that $w \geq 1$ a.e. in an open neighborhood of A. By Lemma 2.6.12 (c) it follows that $w \geq 1$ holds even q.e. in that open neighborhood. Now we define $v := \min(w^+, 1)$. According to Lemma 2.2.8 (c) and Lemma 2.2.8 (g) we have $\|v\|_{H_0^1(\Omega)} \leq \|w\|_{H_0^1(\Omega)}$. Since $v \in M$ this implies $\operatorname{cap}(A) \geq$

$\inf\{\|v\|^2_{H^1_0(\Omega)} \mid v \in M\}$.

Now let us consider the case that A has finite capacity, i.e. M is nonempty. It is easy to see that M is convex. Let us argue that M is also closed. For a sequence $\{v_i\}_{i\in\mathbb{N}}$ in M that converges to a function $v \in H^1_0(\Omega)$ in the $H^1_0(\Omega)$-norm, we can find a subsequence of $\{v_i\}_{i\in\mathbb{N}}$ that converges pointwise quasi-everywhere to v according to Lemma 2.6.15. Thus, the limit v also has to satisfy the conditions $0 \le v \le 1$ q.e. on Ω and $v = 1$ q.e. on A and hence $v \in M$. Since it is well-known that nonempty closed convex subsets of Hilbert spaces contain a unique element with the smallest norm, the claim follows.

For a compact set $K \subset \Omega$ it can be shown that $v \in H^1_0(\Omega)$ is the unique minimizer from Lemma 2.6.16 if and only if v solves the partial differential equation

$$v = 1 \text{ on } K, \qquad -\Delta v = 0 \text{ in } \Omega \setminus K, \qquad v = 0 \text{ on } \partial\Omega.$$

For an open set $O \subset \Omega$ we can find a function $v \in H^1_0(\Omega)$ such that $O = \{v > 0\}$, see Lemma 2.2.12. Using capacity theory, we can extend this result to quasi-open sets $O \subset \Omega$.

Lemma 2.6.17. For every quasi-open set $O \subset \Omega$ there exists a function $v \in H^1_0(\Omega)_+$ such that $O =_q \{v > 0\}$.

Proof. We follow the proof of [Velichkov, 2013, Proposition 2.3.14]. From Lemma 2.6.8 (a) we obtain a sequence $\{G_i\}_{i\in\mathbb{N}}$ of open sets such that $O \cup G_i$ is open and $\operatorname{cap}(G_i) \to 0$ as $i \to \infty$. For each $i \in \mathbb{N}$ let $v_i \in H^1_0(\Omega)$ be a chosen such that $v_i = 1$ q.e. on G_i, $0 \le v_i \le 1$ q.e. on Ω, and $\|v_i\|_{H^1_0(\Omega)} = \operatorname{cap}(G_i)$. This is possible due to Lemma 2.6.16. Since $O \cup G_i$ is open, according to Lemma 2.2.12 there exists a function $w_i \in H^1_0(\Omega)_+$ such that $O \cup G_i = \{w_i > 0\}$ holds. Next, we define the sequence $\{\tilde{w}_i\}_{i\in\mathbb{N}}$ via $\tilde{w}_i := \min(w_i, 1 - v_i)$. Due to Lemma 2.2.8 (f) we know that $\tilde{w}_i \in H^1_0(\Omega)$ for each $i \in \mathbb{N}$. Then we have the estimate

$$\operatorname{cap}(O \setminus \{\tilde{w}_i > 0\}) = \operatorname{cap}(\{v_i = 1\}) \le \|v_i\|^2_{H^1_0(\Omega)}$$

for each $i \in \mathbb{N}$. Finally, we define

$$v := \sum_{i\in\mathbb{N}} 2^{-i}(\|\tilde{w}_i\|_{H^1_0(\Omega)})^{-1}\tilde{w}_i \ \in H^1_0(\Omega).$$

It can be seen that $v = 0$ q.e. on $\Omega \setminus O$. Since $\{\tilde{w}_i > 0\} \subset \{v > 0\}$, we obtain the estimate

$$\operatorname{cap}(O \setminus \{v > 0\}) \le \operatorname{cap}(O \setminus \{\tilde{w}_i > 0\}) \le \|v_i\|^2_{H^1_0(\Omega)} \to 0$$

for $i \to \infty$, which concludes the proof.

2.6.3 Functionals on Sobolev spaces as measures

The goal in this section is to find a representation of positive or negative linear functionals in $H^{-1}(\Omega)_+$ or $H^{-1}(\Omega)_-$ as measures. The general approach is similar to the one presented in [Bonnans, Shapiro, 2000, Section 6.4.3]. We start with a lemma that allows us to extend positive linear functionals on $C_c(\Omega) \cap H_0^1(\Omega)$ to positive linear functionals on $C_c(\Omega)$. Recall that positive linear functionals over $C_c(\Omega) \cap H_0^1(\Omega)$ or $C_c(\Omega)$ are linear functionals $\mu : C_c(\Omega) \cap H_0^1(\Omega) \to \mathbb{R}$, $\tilde{\mu} : C_c(\Omega) \to \mathbb{R}$ with $\mu(f) \geq 0$ for all $f \in C_c(\Omega)_+ \cap H_0^1(\Omega)_+$ or $\tilde{\mu}(f) \geq 0$ for all $f \in C_c(\Omega)_+$. Later, we restrict positive linear functionals over $H_0^1(\Omega)$ to $C_c(\Omega) \cap H_0^1(\Omega)$ and use the Riesz representation theorem to obtain measures.

Lemma 2.6.18. A positive linear functional over $C_c(\Omega) \cap H_0^1(\Omega)$ can be extended to a positive linear functional over $C_c(\Omega)$.

Proof. Let $\mu : C_c(\Omega) \cap H_0^1(\Omega) \to \mathbb{R}$ be a positive linear functional and let $K \subset \Omega$ be a compact set. Then there exists $i_0 \in \mathbb{N}$ such that $\tilde{K} := K + B_{1/i_0} \subset \Omega$ is also a compact set in Ω. We know that there exists a function $g_{\tilde{K}} \in C_c^\infty(\Omega)_+$ such that $g_{\tilde{K}} = 1$ on $\tilde{K}$, see Lemma 2.2.5. For a function $f \in C_c(\Omega)$ with $\operatorname{supp}(f) \subset K$ we define the extension of μ via

$$\tilde{\mu}(f) := \lim_{i \to \infty} \mu(f_i)$$

where $\{f_i\}_{i \in \mathbb{N}}$ is a sequence in $C_c(\Omega) \cap H_0^1(\Omega)$ that converges uniformly to f and such that $\operatorname{supp}(f_i) \subset \tilde{K}$ for all $i \in \mathbb{N}$. We have to argue that $\tilde{\mu}$ is well-defined. First, the existence of $\{f_i\}_{i \in \mathbb{N}}$ follows from Lemma 2.2.6. For $i, j \in \mathbb{N}$ we can use the positivity of μ to obtain

$$|\mu(f_i) - \mu(f_j)| \leq \mu(|f_i - f_j|) \leq \mu(g_{\tilde{K}}) \|f_i - f_j\|_{L^\infty(\Omega)}.$$

This implies that $\{\mu(f_i)\}_{i \in \mathbb{N}}$ is a Cauchy sequence, which guarantees the existence of the limit. Furthermore, if there is another sequence $\{\tilde{f}_i\}_{i \in \mathbb{N}}$ in $C_c(\Omega) \cap H_0^1(\Omega)$ converging uniformly to f with $\operatorname{supp}(\tilde{f}_i) \subset \tilde{K}$, then

$$|\mu(f_i) - \mu(\tilde{f}_i)| \leq \mu(|f_i - \tilde{f}_i|) \leq \mu(g_{\tilde{K}}) \|f_i - \tilde{f}_i\|_{L^\infty(\Omega)} \to 0 \quad (i \to \infty)$$

implies $\lim_{i \to \infty} \mu(f_i) = \lim_{i \to \infty} \mu(\tilde{f}_i)$. It is also easy to see that the definition of $\tilde{\mu}(f)$ is independent of $\tilde{K}$ or K and that $\tilde{\mu}(f) = \mu(f)$ if $f \in C_c(\Omega) \cap H_0^1(\Omega)$. It remains to show that $\tilde{\mu} : C_c(\Omega) \to \mathbb{R}$ is a positive linear functional. The positivity follows from Lemma 2.2.6. For functions $f, \tilde{f} \in C_c(\Omega)$ we consider a common compact set K such that $\operatorname{supp}(f) \cup \operatorname{supp}(\tilde{f}) \subset K$. Then the linearity follows from the linearity on $C_c(\Omega) \cap H_0^1(\Omega)$ via the limit.

Now we can proceed with the main result of this section, which allows us to represent positive linear functionals over $H_0^1(\Omega)$ as Radon measures (i.e. regular Borel measures that are finite on compact sets).

Proposition 2.6.19. Let $\mu \in H^{-1}(\Omega)_+$ be given. Then there exists a unique Radon measure $\tilde{\mu}$ such that every function in $H_0^1(\Omega)$ is $\tilde{\mu}$-integrable and

$$\langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \int_\Omega v \, \mathrm{d}\tilde{\mu} \tag{2.31}$$

holds for all $v \in H_0^1(\Omega)$. Moreover, $\tilde{\mu}(A) = 0$ holds for all Borel sets $A \subset \Omega$ with zero capacity.

Proof. Clearly, μ is a positive linear functional over $C_c(\Omega) \cap H_0^1(\Omega)$. By Lemma 2.6.18 we can extend μ (without renaming) to a positive linear functional over $C_c(\Omega)$. By Theorem 2.2.16 we find a Radon measure $\tilde{\mu}$ that represents μ in the sense that

$$\mu(f) = \int_\Omega f \, \mathrm{d}\tilde{\mu}$$

holds for all $f \in C_c(\Omega)$. In particular, (2.31) holds for all $v \in C_c(\Omega) \cap H_0^1(\Omega)$.

In order to show that the integral in (2.31) does not depend on the quasi-continuous representative of v, we have to argue that $\tilde{\mu}$ does not charge sets of zero capacity. Let $K \subset \Omega$ be a compact set with $\operatorname{cap}(K) = 0$ and let $\varepsilon > 0$. Using the definition of the capacity, there exists a function $w \in H_0^1(\Omega)$ with $\|w\|^2_{H_0^1(\Omega)} < \varepsilon$ such that $w \geq 1$ a.e. on an open set O with $K \subset O \subset \Omega$. By Lemma 2.2.5 there exists a function $f \in C_c^\infty(\Omega)_+$ such that $f = 1$ on K and $f = 0$ on $\Omega \setminus O$. Then we have

$$\tilde{\mu}(K) \leq \int_\Omega f \, \mathrm{d}\tilde{\mu} = \langle \mu, f \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}.$$

By using $w - f \in H_0^1(\Omega)_+$ and $\mu \in H^{-1}(\Omega)_+$ this yields

$$\tilde{\mu}(K) \leq \langle \mu, w \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \leq \|\mu\|_{H^{-1}(\Omega)} \|w\|_{H_0^1(\Omega)} \leq \varepsilon^{1/2} \|\mu\|_{H^{-1}(\Omega)}.$$

Since $\varepsilon > 0$ was arbitrary, this implies $\tilde{\mu}(K) = 0$. Using the inner regularity of $\tilde{\mu}$ it follows that $\tilde{\mu}(A) = 0$ for all Borel measurable sets $A \subset \Omega$ with $\operatorname{cap}(A) = 0$.

Now, let $v \in H_0^1(\Omega)$ be given. Then there exists a sequence $\{f_i\}_{i\in\mathbb{N}}$ in $C_c^\infty(\Omega)$ that converges to v in $H_0^1(\Omega)$-norm. Without loss of generality we can assume that $\{f_i\}_{i\in\mathbb{N}}$ converges pointwise quasi-everywhere, see Lemma 2.6.15. Note that for $i, j \in \mathbb{N}$ we have $|f_i - f_j| \in C_c(\Omega) \cap H_0^1(\Omega)$ due to Lemma 2.2.8 (b). Thus, we have

$$\int_\Omega |f_i - f_j| \, \mathrm{d}\tilde{\mu} = \langle \mu, |f_i - f_j| \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \leq \|\mu\|_{H^{-1}(\Omega)} \|f_i - f_j\|_{H_0^1(\Omega)}.$$

Hence, the sequence $\{f_i\}_{i\in\mathbb{N}}$ is a Cauchy sequence in $L^1_{\tilde{\mu}}(\Omega)$, where $L^1_{\tilde{\mu}}(\Omega)$ denotes the Lebesgue space with exponent 1 for the measure $\tilde{\mu}$. Because $\{f_i\}_{i\in\mathbb{N}}$ converges to v

quasi-everywhere and therefore $\tilde{\mu}$-almost everywhere, this implies that $f_i \to v$ in $L^1_{\tilde{\mu}}(\Omega)$. Therefore, v is $\tilde{\mu}$-integrable and (2.31) follows from

$$\int_\Omega v \,\mathrm{d}\tilde{\mu} = \lim_{i\to\infty} \int_\Omega f_i \,\mathrm{d}\tilde{\mu} = \lim_{i\to\infty} \langle \mu, f_i\rangle_{H^{-1}(\Omega)\times H^1_0(\Omega)} = \langle \mu, v\rangle_{H^{-1}(\Omega)\times H^1_0(\Omega)}.$$

For the uniqueness of $\tilde{\mu}$ let $\tilde{\mu}_1, \tilde{\mu}_2$ be Radon measures that satisfy (2.31) and let $K \subset \Omega$ be a compact set. Then there exists a sequence $\{f_i\}_{i\in\mathbb{N}}$ in $C_c^\infty(\Omega)$ such that $0 \le f_i \le 1$ in Ω, $f_i = 1$ on K for each $i \in \mathbb{N}$, $\{\operatorname{supp}(f_i)\}_{i\in\mathbb{N}}$ is a nonincreasing sequence of sets, and $K = \bigcap_{i\in\mathbb{N}} \operatorname{supp}(f_i)$, see Lemma 2.2.5. From (2.31) we obtain that

$$\tilde{\mu}_{3-j}(K) \le \int_\Omega f_i \,\mathrm{d}\tilde{\mu}_{3-j} = \int_\Omega f_i \,\mathrm{d}\tilde{\mu}_j \le \tilde{\mu}_j(\operatorname{supp}(f_i)) \to \tilde{\mu}_j(K) \quad (i \to \infty)$$

holds for $j = 1, 2$. This implies $\tilde{\mu}_1(K) = \tilde{\mu}_2(K)$, and the equality of the measures for arbitrary Borel measurable sets follows from the inner regularity of $\tilde{\mu}_1, \tilde{\mu}_2$.

We note that it is important to use a quasi-continuous representative in the integral in (2.31), because otherwise the integral might not be well-defined if we obtain a measure that charges sets of Lebesgue measure zero.

Due to Proposition 2.6.19 we can identify a functional in $H^{-1}(\Omega)_+$ with a Radon measure via (2.31). For the rest of this thesis, whenever we have a functional in $H^{-1}(\Omega)_+$ we will use the same name to refer to the Radon measure that satisfies (2.31). As an example, the function $1 \in H^{-1}(\Omega)_+$ can be interpreted as the Lebesgue measure. For a functional $\mu \in H^{-1}(\Omega)_-$ it is clear that there exists a unique signed (or negative) Radon measure $\tilde{\mu}$ that satisfies (2.31). Again, we will use μ to refer to both the functional and the signed measure $\tilde{\mu}$.

From the representation (2.31) we can also obtain an estimate on the μ-measure of Borel sets. This was already presented in [Harder, G. Wachsmuth, 2018a, Lemma 3.5 (c)].

Corollary 2.6.20. Let $A \subset \Omega$ be a Borel measurable set and $\mu \in H^{-1}(\Omega)_+$. Then the inequality

$$\mu(A) \le \operatorname{cap}(A)^{1/2} \|\mu\|_{H^{-1}(\Omega)}$$

holds.

Proof. We follow the proof of [Harder, G. Wachsmuth, 2018a, Lemma 3.5 (c)]. Without loss of generality we assume that $\operatorname{cap}(A) < \infty$. According to Lemma 2.6.16 there exists a function $v \in H^1_0(\Omega)$ such that $v = 1$ q.e. on A, $0 \le v \le 1$ q.e. on Ω, and $\operatorname{cap}(A) = \|v\|^2_{H^1_0(\Omega)}$. Since the measure μ does not charge sets of zero capacity, this implies $v \ge \chi_A$ μ-a.e. on Ω. Then by Proposition 2.6.19 we have

$$\mu(A) \le \int_\Omega v \,\mathrm{d}\mu = \langle \mu, v\rangle_{H^{-1}(\Omega)\times H^1_0(\Omega)} \le \|v\|_{H^1_0(\Omega)} \|\mu\|_{H^{-1}(\Omega)} = \operatorname{cap}(A)^{1/2} \|\mu\|_{H^{-1}(\Omega)}.$$

2.6.4 The quasi-support

In this section we want to introduce the so-called quasi-support of a Borel measure on Ω. This allows us to describe a (quasi-closed) subset of Ω where the measure is "active". This is similar to the notion of the support of continuous functions.

Although the quasi-support is defined for Borel measures that do not charge sets of zero capacity, we will use it mostly for functionals in $H^{-1}(\Omega)_+$ or $H^{-1}(\Omega)_-$. We recall that such functionals can be interpreted as (positive or negative) measures by Proposition 2.6.19.

The notion of the quasi-support is useful for providing a more intuitive understanding of the pointwise nature of functionals in $H^{-1}(\Omega)_+$ or $H^{-1}(\Omega)_-$. This will be helpful for understanding optimality conditions of optimization problems where the set $H_0^1(\Omega)_+$ plays a role. It also allows us to give a nice expression for the normal and critical cone of the set $H_0^1(\Omega)_+$, see Proposition 2.6.26.

The quasi-support of a functional $\mu \in H^{-1}(\Omega)_+$ is also sometimes known as the *fine support* and denoted as f-supp(μ). It was defined in [G. Wachsmuth, 2014, Lemma A.4] as the complement of the largest finely-open set that is not charged by the measure μ. The approach in [G. Wachsmuth, 2014, Appendix A] for introducing the fine support requires some background theory because it relies on the concept of the fine topology on Ω. For details of the fine topology we refer to [Heinonen, Kilpeläinen, Martio, 1993, Chapter 12].

In this section we will take an alternative approach which is more elementary. It is based on the result [Stollmann, 1993, Theorem 1]. This approach was already discussed in [Harder, G. Wachsmuth, 2018a, Section 3.2]. However, we provide some more details here and define the quasi-support for more Borel measures. Note that what was called "fine support" in [Harder, G. Wachsmuth, 2018a, Definition 3.8] is now called the quasi-support. The notion of a quasi-support of measures is also introduced in [Fukushima, Oshima, Takeda, 2010, p. 190]. However, this book uses so-called symmetric Hunt processes (which are stochastic processes) in their analysis of capacity theory.

Closed lattice ideals in $H_0^1(\Omega)$ will play an important role in this section. Lattice ideals were defined in Definition 2.1.22. When talking about the lattice in $H_0^1(\Omega)$ we refer to the lattice induced by $H_0^1(\Omega)_+$ unless noted otherwise. The following theorem allows us to create a relationship between closed lattice ideals in $H_0^1(\Omega)$ and quasi-closed subsets of Ω. This will be helpful for introducing the quasi-support later. The theorem can be found in a slightly different version (without uniqueness of the set A) in [Stollmann, 1993, Theorem 1]. For the convenience of the reader we will provide a proof.

Theorem 2.6.21. Let $V \subset H_0^1(\Omega)$ be a closed lattice ideal. Then there exists a quasi-closed set $A \subset \Omega$ such that

$$V = \{w \in H_0^1(\Omega) \mid w = 0 \text{ q.e. on } A\} \tag{2.32}$$

holds. Moreover, the set A is unique up to a set of zero capacity.

Proof. Let $\{v_i\}_{i\in\mathbb{N}}$ be a sequence in V such that its linear hull is dense in V. From Definition 2.1.22 it follows that $|v_i| \in V$ for all $i \in \mathbb{N}$. We define the function

$$v := \sum_{i\in\mathbb{N}} (2^i \|v_i\|_{H_0^1(\Omega)})^{-1} |v_i|.$$

Because V is closed it is clear that $v \in V$. We set $A := \{v = 0\}$. From Lemma 2.6.9 it follows that A is quasi-closed.

Now that we have a candidate for A, we need to prove that this set satisfies the condition (2.32). We denote the right-hand side of (2.32) by W. Using Lemma 2.6.15 it can be seen that W is a closed linear subspace of $H_0^1(\Omega)$. Since W also contains v_i for all $i \in \mathbb{N}$ and the linear hull of $\{v_i\}_{i\in\mathbb{N}}$ is dense in V it follows that $V \subset W$.

Our next goal is to prove that $W \subset V$. Let $\varepsilon > 0$ be given. Since v is quasi-continuous, there exists an open set $G_\varepsilon \subset \Omega$ with $\operatorname{cap}(G_\varepsilon) < \varepsilon$ such that v is continuous on $\Omega \setminus G_\varepsilon$. Due to Lemma 2.6.16 there exists a function $w_\varepsilon \in H_0^1(\Omega)$ such that $\|w_\varepsilon\|^2_{H_0^1(\Omega)} < \varepsilon$, $0 \le w_\varepsilon \le 1$ q.e. on Ω, and $w_\varepsilon = 1$ q.e. on G_ε. We observe that the set $\Omega_1 := \{v > 0\} \cup G_\varepsilon$ is open. Let $f \in C_c^\infty(\Omega_1)_+$ be given. Then we have $\min(f, 1 - w_\varepsilon) \in H_0^1(\Omega)$ due to Lemma 2.2.8 (f) and using the sign conditions for w_ε and f we even obtain $\min(f, 1 - w_\varepsilon) \in W$. Now let $\alpha \in \mathbb{R}$ be the minimizer of v on the compact set $\operatorname{supp}(f) \setminus G_\varepsilon \subset \Omega_1$. Due to the continuity of v on that set this is well-defined and we obtain $\alpha > 0$. It follows that $\min(f, 1 - w_\varepsilon) \le \alpha^{-1} v$ q.e. on Ω and therefore $\min(f + w_\varepsilon, 1) - w_\varepsilon = \min(f, 1 - w_\varepsilon) \in V$. Because $C_c^\infty(\Omega_1)_+$ is dense in $H_0^1(\Omega_1)_+$ by Lemma 2.2.11 and $\min(\cdot, 1)$ is continuous in $H_0^1(\Omega)$ by Lemma 2.2.8 (f) this implies that $\min(w^+ + w_\varepsilon, 1) - w_\varepsilon \in V$ for all $w \in W \subset H_0^1(\Omega_1)$. If we consider the facts that $w_\varepsilon \to 0$ in $H_0^1(\Omega)$ and V is closed, we obtain $\min(w^+, 1) = \lim_{\varepsilon \downarrow 0} \min(w^+ + w_\varepsilon, 1) - w_\varepsilon \in V$ for all $w \in W$ (where Lemma 2.2.8 (f) was applied again). By linearity of W it follows that $\min(w^+, i) - \min(w_-, i) \in V$ for all $w \in W, i \in \mathbb{N}$. Finally, by Lemma 2.2.8 (h) it follows that $w = w^+ - w_- \in V$ for all $w \in W$, concluding the proof that $W \subset V$.

For the uniqueness of the set A, assume that there are two quasi-closed sets A_1, A_2 that satisfy (2.32). Then by Lemma 2.6.17 there is a function $v_1 \in H_0^1(\Omega)$ such that $\Omega \setminus A_1 =_q \{v_1 > 0\}$. This is equivalent to $\{v_1 = 0\} =_q A_1$. Thus, we have $v_1 \in V$ and therefore $v_1 = 0$ q.e. on A_2. Hence, we obtain $A_2 \subset_q A_1$ and by switching the roles of A_1 and A_2 in the previous steps we also obtain $A_1 \subset_q A_2$. It follows that $A_1 =_q A_2$ which concludes the proof.

There are similar results for closed lattice ideals in other spaces. For example, for each closed lattice ideal in $L^2(\Omega)$ (with respect to the closed convex cone $L^2(\Omega)_+ \subset L^2(\Omega)$) there is a measurable set $A \subset \Omega$ such that the closed lattice ideal is given as $\{w \in L^2(\Omega) \mid w = 0 \text{ a.e. on } A\}$. It can also be seen that each closed lattice ideal in

$C(\operatorname{cl}\Omega)$ (with respect to the closed convex cone $C(\operatorname{cl}\Omega)_+ \subset C(\operatorname{cl}\Omega)$) can be written as $\{w \in C(\operatorname{cl}\Omega) \mid w = 0 \text{ on } A\}$ for a closed set $A \subset \Omega$.

Now that we have characterized closed lattice ideals in $H_0^1(\Omega)$, we can continue working towards the existence and definition of the quasi-support.

Proposition 2.6.22. Let μ be a Borel measure on Ω such that $\mu(A) = 0$ for all Borel sets $A \subset \Omega$ with $\operatorname{cap}(A) = 0$. Then there exists a quasi-closed set $A \subset \Omega$ such that

$$\mu(\{v \neq 0\}) = 0 \quad \Leftrightarrow \quad v = 0 \text{ q.e. on } A \tag{2.33}$$

holds for all $v \in H_0^1(\Omega)$. Moreover, the set A is unique up to a set of zero capacity.

Proof. We define the set $V := \{v \in H_0^1(\Omega) \mid \mu(\{v \neq 0\}) = 0\}$. Let us show that V is a closed lattice ideal. Because for $v, \tilde{v} \in V, \alpha \in \mathbb{R}$ we have $\{\alpha v + \tilde{v} \neq 0\} \subset \{v \neq 0\} \cup \{\tilde{v} \neq 0\}$, we know that V is a linear subspace. In order to prove that V is a closed set, let $\{v_i\}_{i\in\mathbb{N}}$ be a sequence in V that converges to an element $v \in H_0^1(\Omega)$. Due to Lemma 2.6.15 there is a subsequence of $\{v_i\}_{i\in\mathbb{N}}$ that converges quasi-everywhere to v. This implies $\{v \neq 0\} \subset_q \bigcup_{i\in\mathbb{N}} \{v_i \neq 0\}$ and $v \in V$ follows. For the lattice property, let $v \in V$, $w \in H_0^1(\Omega)$ be given such that $|w| \leq |v|$. Then we have $\{w \neq 0\} \subset \{v \neq 0\}$ and therefore $w \in V$. Thus, V is a closed lattice ideal. Therefore, the existence and uniqueness (up to a set of zero capacity) of a quasi-closed set A follow directly from Theorem 2.6.21.

We can use this proposition to define the quasi-support.

Definition 2.6.23. Let μ be a Borel measure on Ω such that $\mu(A) = 0$ for all Borel sets $A \subset \Omega$ with $\operatorname{cap}(A) = 0$. Then we choose a quasi-closed set $A \subset \Omega$ that satisfies (2.33) and define the *quasi-support* of μ as $\text{q-supp}(\mu) := A$. For a functional $\mu \in H^{-1}(\Omega)_-$ we define $\text{q-supp}(\mu) := \text{q-supp}(-\mu)$.

Note that the definition of the quasi-support can depend on the particular chosen set $A \subset \Omega$ from Proposition 2.6.22. However, this will not be a problem when working with the quasi-support, because its properties are only relevant up to a set of zero capacity. Due to the axiom of choice the above definition is well-defined.

We continue with some equivalent descriptions of the quasi-support of a functional in $H^{-1}(\Omega)_+$.

Lemma 2.6.24. Let $\mu \in H^{-1}(\Omega)_+$ be a functional and let $A \subset \Omega$ be a quasi-closed set. Then the following are equivalent:

(a) $A =_q \text{q-supp}(\mu)$,

(b) $\mu(\{v \neq 0\}) = 0 \Leftrightarrow v = 0$ q.e. on A for all $v \in H_0^1(\Omega)$,

(c) $\langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0 \Leftrightarrow v = 0$ q.e. on A for all $v \in H_0^1(\Omega)_+$,

(d) $\mu(O) = 0 \Leftrightarrow O \cap A =_q \emptyset$ for all quasi-open sets $O \subset \Omega$.

Proof. The equivalence of (a) and (b) follows directly from the definition of the quasi-support (which relies on (2.33)).

From the identification (2.31) it follows that (b) implies (c). To show that (c) implies (d) we can use a function $v \in H_0^1(\Omega)_+$ such that $O =_q \{v > 0\}$ which is possible due to Lemma 2.6.17. The remaining implication (d) $\Rightarrow$ (b) follows because the set $\{v \neq 0\}$ is quasi-open.

The equivalent description Lemma 2.6.24 (d) can be interpreted as A being the complement of the largest quasi-open set that is not charged by μ.

For future use, we mention the following observations about properties of the quasi-support here.

Corollary 2.6.25. Let $v \in H_0^1(\Omega)$ and $\mu \in H^{-1}(\Omega)_-$ be given such that $v \geq 0$ q.e. on q-supp(μ). Then

(a) $\langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \leq 0$

(b) $\langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0 \quad \Leftrightarrow \quad v = 0$ q.e. on q-supp$(\mu) \quad \Leftrightarrow \quad \mu(\{v > 0\}) = 0$

hold.

Proof. If we apply Lemma 2.6.24 (c) to v^-, then we obtain that $\langle \mu, v^- \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$. Thus, $\langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle \mu, v^+ \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \leq 0$ shows the claim of part (a).

Using $\langle \mu, v^+ \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$ we can now apply Lemma 2.6.24 (c) and Lemma 2.6.24 (b) to v^+. Then the equivalencies of part (b) follow.

2.6.5 Applications

In this section we want to apply the theory of the previous sections and provide results that we use in Chapter 5. An important application of capacity theory is the calculation of the normal, tangent, and critical cone of the convex cone $H_0^1(\Omega)_+ \subset H_0^1(\Omega)$. These results can be found in [Bonnans, Shapiro, 2000, Theorem 6.57] or [Harder, G. Wachsmuth, 2018a, Lemma 3.11].

Proposition 2.6.26. Let $v \in H_0^1(\Omega)_+$ and $\mu \in \mathcal{N}_{H_0^1(\Omega)_+}(v) \subset H^{-1}(\Omega)_-$ be given.

(a) The normal cone of $H_0^1(\Omega)_+$ at v can be described by

$$\begin{aligned}\mathcal{N}_{H_0^1(\Omega)_+}(v) &= \{\xi \in H^{-1}(\Omega)_- \mid \xi(\{v > 0\}) = 0\} \\ &= \{\xi \in H^{-1}(\Omega)_- \mid v = 0 \text{ q.e. on q-supp}(\xi)\}.\end{aligned}$$

(b) The tangent cone of $H_0^1(\Omega)_+$ at v can be described by

$$\mathcal{T}_{H_0^1(\Omega)_+}(v) = \{w \in H_0^1(\Omega) \mid w \geq 0 \text{ q.e. on } \{v = 0\}\}. \tag{2.34}$$

(c) The critical cone to $H_0^1(\Omega)_+$ at (v, μ) can be described by

$$\mathcal{K}_{H_0^1(\Omega)_+}(v, \mu) = \{w \in H_0^1(\Omega) \mid w = 0 \text{ q.e. on q-supp}(\mu),\ w \geq 0 \text{ q.e. on } \{v = 0\}\}.$$

Proof. We start with calculating the normal cone. Due to Lemma 2.1.17 we have

$$\mathcal{N}_{H_0^1(\Omega)_+}(v) = \{\xi \in H^{-1}(\Omega)_- \mid \langle \xi, v \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} = 0\}.$$

Now the claimed expressions follow directly from Corollary 2.6.25 (b).

We move on to the tangent cone. Let $w \in H_0^1(\Omega)$ with $w \geq 0$ q.e. on $\{v = 0\}$ and $\xi \in \mathcal{N}_{H_0^1(\Omega)_+}(v)$ be given. Since q-supp$(\xi) \subset_q \{v = 0\}$ we have that $w \geq 0$ q.e. on q-supp(ξ). Then it follows from Corollary 2.6.25 (a) that $\langle \xi, w \rangle \leq 0$. Since ξ was arbitrary, this implies that $w \in \mathcal{N}_{H_0^1(\Omega)_+}(v)^\circ = \mathcal{T}_{H_0^1(\Omega)_+}(v)$. For the other inclusion we first observe that the right-hand side of (2.34) is closed. This can be seen by using Lemma 2.6.15. Thus, it suffices to show that $\mathcal{R}_{H_0^1(\Omega)_+}(v)$ is included in the right-hand side of (2.34). Indeed, if $w = \alpha(z - v) \in \mathcal{R}_{H_0^1(\Omega)_+}(v)$ is given with $z \in H_0^1(\Omega)_+$ and $\alpha > 0$, then $w = \alpha z \geq 0$ holds q.e. on $\{v = 0\}$.

The expression for the critical cone follows directly from the expression for the tangent cone and the first equivalence in Corollary 2.6.25 (b).

We can use these results to calculate the normal cone to a wider class of convex cones in $H_0^1(\Omega)$.

Proposition 2.6.27. For a quasi-closed subset $A \subset \Omega$ we consider the convex cone

$$M := \{w \in H_0^1(\Omega) \mid w \geq 0 \text{ q.e. on } A\}.$$

Then its normal cone at a point $v \in M$ can be described by

$$\mathcal{N}_M(v) = \{\xi \in H^{-1}(\Omega)_- \mid \text{q-supp}(\xi) \subset_q A \cap \{v = 0\}\}.$$

Proof. Our first goal is to calculate M°. According to Lemma 2.6.17 there exists a function $\hat{v} \in H_0^1(\Omega)_+$ such that $A =_q \{\hat{v} = 0\}$. Then, by Proposition 2.6.26 (b) we obtain that $M = \mathcal{T}_{H_0^1(\Omega)_+}(\hat{v})$. Thus, by Proposition 2.6.26 (a) we obtain

$$\begin{aligned} M^\circ &= \mathcal{T}_{H_0^1(\Omega)_+}(\hat{v})^\circ = \mathcal{N}_{H_0^1(\Omega)_+}(\hat{v}) \\ &= \{\xi \in H^{-1}(\Omega)_- \mid \hat{v} = 0 \text{ q.e. on q-supp}(\xi)\} \end{aligned}$$

$$= \{\xi \in H^{-1}(\Omega)_- \mid \text{q-supp}(\xi) \subset_q \{\hat{v} = 0\}\}$$
$$= \{\xi \in H^{-1}(\Omega)_- \mid \text{q-supp}(\xi) \subset_q A\}.$$

Now, let $v \in M$ be given. Due to Lemma 2.1.17 we have

$$\mathcal{N}_M(v) = M^\circ \cap v^\perp.$$

Thus, it remains to show that the equivalence

$$\text{q-supp}(\xi) \subset_q A \quad \wedge \quad \langle \xi, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0 \qquad \Leftrightarrow \qquad \text{q-supp}(\xi) \subset_q A \cap \{v = 0\}$$

holds for all $\xi \in H^{-1}(\Omega)_-$. Since $v \geq 0$ q.e. on q-supp(ξ) holds if the left-hand side or right-hand side of this equivalence holds, this equivalence follows from Corollary 2.6.25 (b).

Let us also state a technical lemma that involves the quasi-support and functions in $C_c^\infty(\Omega)_+$.

Lemma 2.6.28. Let $v \in H_0^1(\Omega)$ and $\mu \in H^{-1}(\Omega)_+$ be given such that

$$\langle \mu, fv \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$$

holds for all $f \in C_c^\infty(\Omega)_+$. Then $v = 0$ q.e. on q-supp(μ).

Proof. Let $O \subset \Omega$ be an open subset. First, we want to show that

$$\int_O v \, \mathrm{d}\mu = 0$$

holds. We recall that every open set $O \subset \Omega$ has a compact exhaustion, i.e. there exists an increasing sequence $\{K_i\}_{i\in\mathbb{N}}$ of compact subsets of O such that $O = \bigcup_{i\in\mathbb{N}} K_i$. Then there exists a sequence of functions $\{f_i\}_{i\in\mathbb{N}} \subset C_c^\infty(O)_+ \subset C_c^\infty(\Omega)$ such that $0 \leq f_i \leq 1$ in Ω, $f_i = 0$ on $\Omega \setminus O$, and $f_i = 1$ on K_i holds for all $i \in \mathbb{N}$, see Lemma 2.2.5. Then we have $f_i(\omega)v(\omega) \to v(\omega)$ for all $\omega \in O$. Since $|v|$ is μ-integrable and $|f_i v| \leq |v|$ holds for all $i \in \mathbb{N}$ we can utilize Lebesgue's dominated convergence theorem which yields

$$\int_O v \, \mathrm{d}\mu = \lim_{i\to\infty} \int_O f_i v \, \mathrm{d}\mu = \lim_{i\to\infty} \langle \mu, f_i v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0. \tag{2.35}$$

Since the set $\{v > 0\}$ is quasi-open, there exists a nonincreasing sequence $\{G_i\}_{i\in\mathbb{N}}$ of open subsets of Ω such that $\{v > 0\} \cup G_i$ is open for every $i \in \mathbb{N}$ and $\mathrm{cap}(G_i) \to 0$ as $i \to \infty$, see Lemma 2.6.8 (a). When we use $O = \{v > 0\} \cup G_i$ in (2.35) we obtain

$$\int_{\{v>0\}\cup G_i} v \, \mathrm{d}\mu = 0 \qquad \forall i \in \mathbb{N}.$$

Then

$$\begin{aligned}\langle \mu, v^+ \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} &= \int_\Omega v^+ \,\mathrm{d}\mu = \int_{\{v>0\}\cup G_i} v^+ \,\mathrm{d}\mu \\ &= \int_{\{v>0\}\cup G_i} v \,\mathrm{d}\mu - \int_{\{v>0\}\cup G_i} v^- \,\mathrm{d}\mu = -\int_{G_i} v^- \,\mathrm{d}\mu\end{aligned}$$

follows for all $i \in \mathbb{N}$. Because v^- is μ-integrable and $\mu(G_i) \to 0$ holds due to Corollary 2.6.20, this implies $\langle \mu, v^+ \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$. If we use (2.35) again with $O = \Omega$ we also obtain $\langle \mu, v \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$. Therefore, $\langle \mu, |v| \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$ holds. Then the claim follows by applying Corollary 2.6.25 (b) to $-\mu$ and $|v|$.

Finally, we provide a density result on quasi-open sets. Note that the result is classical for open sets. The lemma together with its proof are taken from [Harder, G. Wachsmuth, 2018c, Lemma A.2].

Lemma 2.6.29. Let $O \subset \Omega$ be a quasi-open set. Then

$$V := \{w \in H_0^1(\Omega) \mid w = 0 \text{ q.e. on } \Omega \setminus O\}$$

is dense in $L^q(O)$, where $q \in [1, \infty)$ is such that $H_0^1(\Omega) \hookrightarrow L^q(\Omega)$.

Proof. We recall that O is Lebesgue measurable due to Lemma 2.6.12 (a). We note that the linear hull of the set

$$\{f \in L^q(O) \mid 0 \le f \le 1\}$$

is dense in $L^q(O)$. Hence, it is sufficient to show that $f \in L^q(O)$ with $0 \le f \le 1$ can be approximated by functions from V.

Let $\varepsilon > 0$ be given. Then we can find an open set G_ε such that $O \cup G_\varepsilon$ is open and $\operatorname{cap}(G_\varepsilon) < \varepsilon$. Since $H_0^1(O \cup G_\varepsilon)$ is dense in $L^q(O \cup G_\varepsilon)$, we can find a function $w_\varepsilon \in H_0^1(O \cup G_\varepsilon) \subset H_0^1(\Omega)$ such that $0 \le w_\varepsilon \le 1$ q.e. on Ω and $\|w_\varepsilon - f\|_{L^q(\Omega)} < \varepsilon$. Using Lemma 2.6.16 yields the existence of $v_\varepsilon \in H_0^1(\Omega)_+$ such that $\|v_\varepsilon\|_{H_0^1(\Omega)} < \sqrt{\varepsilon}$ and $v_\varepsilon = 1$ q.e. on G_ε. We define the function

$$\tilde{w}_\varepsilon := \max(0, w_\varepsilon - v_\varepsilon) \in H_0^1(\Omega).$$

From $w_\varepsilon = 0$ q.e. on $\Omega \setminus (O \cup G_\varepsilon)$, $w_\varepsilon \le 1$, $v_\varepsilon \ge 0$ and $v_\varepsilon \ge 1$ q.e. on G_ε, we find $\tilde{w}_\varepsilon = 0$ q.e. on $(\Omega \setminus (O \cup G_\varepsilon)) \cup G_\varepsilon$. This implies $\tilde{w}_\varepsilon \in V$. Moreover, we have

$$\begin{aligned}\|\tilde{w}_\varepsilon - w_\varepsilon\|_{L^q(\Omega)}^q = \|\max(-w_\varepsilon, -v_\varepsilon)\|_{L^q(\Omega)}^q &= \int_{\{w_\varepsilon \le v_\varepsilon\}} |w_\varepsilon|^q \,\mathrm{d}\omega + \int_{\{v_\varepsilon < w_\varepsilon\}} |v_\varepsilon|^q \,\mathrm{d}\omega \\ &\le \|v_\varepsilon\|_{L^q(\Omega)}^q \le C \|v_\varepsilon\|_{H_0^1(\Omega)}^q \le C\, \varepsilon^{\frac{q}{2}}.\end{aligned}$$

Using the triangle inequality yields

$$\|f - \tilde{w}_\varepsilon\|_{L^q(\Omega)} \leq \|f - w_\varepsilon\|_{L^q(\Omega)} + \|w_\varepsilon - \tilde{w}_\varepsilon\|_{L^q(\Omega)} \leq \varepsilon + C\sqrt{\varepsilon}.$$

Thus, we can approximate f with functions in $\tilde{w}_\varepsilon \in V$, and this proves the claim.

3 Optimization theory for bilevel optimization problems

3.1 Optimization problems with parameters

3.1.1 Notation and setting

In this section we want to discuss abstract optimization problems that contain a parameter. Therefore, we establish some notation and basic assumptions.

Throughout this section X, Y, V will denote abstract Banach spaces. For each parameter $p \in V$ we consider the optimization problem

$$\begin{aligned} \min_x \quad & f(x,p) \\ \text{s.t.} \quad & g(x,p) \in \Phi, \end{aligned} \tag{P(p)}$$

where $f : X \times V \to \mathbb{R}$ is the objective function and $g : X \times V \to Y$ as well as the closed and nonempty set $\Phi \subset Y$ constitute the constraint. Note that both the objective function and the constraint depend on a parameter, but we minimize only with respect to the variable $x \in X$.

Definition 3.1.1. We denote by $\Psi : V \to \mathcal{P}(X)$ the set-valued mapping that maps a parameter $p \in V$ to the set of (global) solutions of (P(p)).

If for each $p \in V$ there exists a unique solution of (P(p)), we denote the *solution operator* that maps a parameter p to the unique solution of (P(p)) by $\psi : V \to X$.

Another important object is the optimal value function.

Definition 3.1.2. We define the *value function* or *optimal value function* $\varphi : V \to \mathbb{R} \cup \{-\infty, +\infty\}$ via

$$\varphi(p) := \inf\{f(x,p) \mid x \in X,\, g(x,p) \in \Phi\} \qquad \forall\, p \in V.$$

For later use we define the parameter-dependent Lagrange function associated with the problem (P(p)).

Definition 3.1.3. We define the (parameter-dependent) Lagrange function $\mathcal{L} : X \times Y^\star \times V \to \mathbb{R}$ that corresponds to the parametrized optimization problem (P($\cdot$)) via

$$\mathcal{L}(x, \lambda, p) := f(x, p) + \langle \lambda, g(x, p) \rangle_{Y^\star \times Y} \qquad \forall\, x \in X, \lambda \in Y^\star, p \in V.$$

3.1.2 Continuity properties of solution operators

A natural question to ask regarding an optimization problem is whether solutions exist and, if this is the case, whether there exists a unique solution. For a parametrized problem such as (P(p)) it is also of interest how the solutions change when the parameter p is perturbed.

We focus on the case where Ψ is single-valued, i.e. ψ is defined. This has the advantage that we can focus on the conventionally known properties of functions. For set-valued mappings, we would need to discuss concepts that are less fundamental, e.g. the so-called upper hemicontinuity instead of continuity. Moreover, all applications of parametrized optimization problems in this thesis only require the case, in which there is at most one global minimizer.

In this section we consider continuity properties of the solution operator ψ. We will see that in many cases the strong convexity of the objective function of (P(p)) plays an important role, but there are also more assumptions needed to guarantee continuity of ψ. In some cases we will only consider the situation where the feasible set does not depend on p, i.e. $g(x, p) = x$. To illustrate the importance of these assumptions we will give some counterexamples. For strong convexity we remark that strongly convex functions that are also continuous can only exist in reflexive Banach spaces, see Lemma 2.1.36 (a). Thus, when functions exist that are strongly convex on subsets, the requirement that the underlying space is reflexive is not a very strong requirement.

We start with a result that shows continuity of ψ for the case where the feasible set can depend on the parameter p. However, some assumptions on the behavior of g are needed. A result in the literature that also discusses continuity properties of the solution operator can be found in [Bonnans, Shapiro, 2000, Proposition 4.4]. However, since compact sets are rare in infinite-dimensional spaces, this result is not as useful for our purposes as our own result.

Lemma 3.1.4. Let $\bar{p} \in V$ be given and suppose that

(a) there exists a constant $\gamma > 0$ such that for all $p \in V$ the function $f(\cdot, p)$ is strongly convex with parameter γ on the feasible set $g(\cdot, p)^{-1}(\Phi)$,

(b) $g(\cdot, p)^{-1}(\Phi)$ is convex, closed, and nonempty for all $p \in V$,

(c) for every sequence $\{p_i\}_{i \in \mathbb{N}} \subset V$ with $p_i \to \bar{p}$ there exists a sequence $\{x_i\}_{i \in \mathbb{N}} \subset X$ such that $g(x_i, p_i) \in \Phi$ for all $i \in \mathbb{N}$ and $x_i \to \psi(\bar{p})$,

(d) f is continuous,

(e) X is reflexive,

(f) there exists an $\varepsilon > 0$ such that for all sequences $\{p_i\}_{i\in\mathbb{N}} \subset V$, $\{x_i\}_{i\in\mathbb{N}} \subset B_\varepsilon(\psi(\bar{p}))$ with $p_i \to \bar{p}$ and $x_i \rightharpoonup x_0$ for some $x_0 \in X$ we have $f(x_0, \bar{p}) \leq \liminf_{i\to\infty} f(x_i, p_i)$ as well as $g(x_0, \bar{p}) \in \Phi$ if $g(x_i, p_i) \in \Phi$ for all $i \in \mathbb{N}$.

Then ψ exists and is continuous at $\bar{p}$.

Proof. The existence of ψ follows from Lemma 2.3.2 (b).

Let a sequence $\{p_i\}_{i\in\mathbb{N}} \subset V$ with $p_i \to \bar{p}$ be given. We modify the parametrized problem (P(p)) by adding the constraint $x \in B_\varepsilon(\psi(\bar{p}))$. For large $i \in \mathbb{N}$ assumption (c) guarantees that the modified feasible set for the problem with parameter p_i is nonempty. Without loss of generality we can assume that this is the case for all $i \in \mathbb{N}$.

Thus, for each $i \in \mathbb{N}$ there exists a unique solution of the modified problem with parameter p_i (see Lemma 2.3.2 (b) again), which we denote by $\psi_\varepsilon(p_i)$. Note that due to the convexity of the problem, if $\|\psi_\varepsilon(p) - \psi(\bar{p})\| < \varepsilon$ holds for some $p \in V$, then the newly added constraint is irrelevant and therefore $\psi_\varepsilon(p) = \psi(p)$.

Since $\{\psi_\varepsilon(p_i)\}_{i\in\mathbb{N}}$ is a bounded sequence, there exists a weakly convergent subsequence. Without loss of generality we can assume that $\psi_\varepsilon(p_i) \rightharpoonup x_0$ for some $x_0 \in B_\varepsilon(\psi(\bar{p}))$. Then from assumption (f) we obtain

$$f(x_0, \bar{p}) \leq \liminf_{i\to\infty} f(\psi_\varepsilon(p_i), p_i) \qquad \text{and} \qquad g(x_0, \bar{p}) \in \Phi. \tag{3.1}$$

Now, let $\{x_i\}_{i\in\mathbb{N}}$ be a sequence that satisfies $g(x_i, p_i) \in \Phi$, $x_i \in B_\varepsilon(\psi(\bar{p}))$, and $x_i \to \psi(\bar{p})$, see assumption (c). Then by Lemma 2.3.2 (c) the quadratic growth condition

$$f(\psi_\varepsilon(p_i), p_i) + \tfrac{\gamma}{2}\|x_i - \psi_\varepsilon(p_i)\|^2 \leq f(x_i, p_i)$$

holds for all $i \in \mathbb{N}$. Taking the limes inferior and using assumption (d) and (3.1) yields

$$f(x_0, \bar{p}) + \liminf_{i\to\infty} \tfrac{\gamma}{2}\|x_i - \psi_\varepsilon(p_i)\|^2 \leq \liminf_{i\to\infty} f(x_i, p_i) = f(\psi(\bar{p}), \bar{p}) \leq f(x_0, \bar{p}).$$

Thus, $\|x_i - \psi_\varepsilon(p_i)\| \to 0$ holds along a subsequence. Since the limit does not depend on the chosen subsequence the convergence $\psi_\varepsilon(p_i) \to \psi(\bar{p})$ follows. As mentioned above, we have $\psi_\varepsilon(p_i) = \psi(p_i)$ for large $i \in \mathbb{N}$. Thus, we also obtain the convergence $\psi(p_i) \to \psi(\bar{p})$.

To highlight the importance of assumption (f) in Lemma 3.1.4, we provide an example of an unconstrained parametrized optimization problem where the other assumptions are satisfied, and ψ is not continuous.

Example 3.1.5. We consider the case that $X = Y := L^2(\Omega)$ with $\Omega := (0,1)$ and $V := \mathbb{R}$. We set $g(x,p) := x$ and $\Phi := Y$, i.e. we consider the unconstrained case. We define the function f via

$$f(x,p) := \begin{cases} \int_0^1 \frac{1}{2}x(\omega)^2 - \sin(p^{-1}\omega)x(\omega)\,\mathrm{d}\omega & \text{if } p > 0, \\ \int_0^1 \frac{1}{2}x(\omega)^2\,\mathrm{d}\omega & \text{if } p \leq 0. \end{cases}$$

Then f is continuous, $f(\cdot,p)$ is strongly convex with parameter $\gamma = 1$ for all $p \in V$, and the solution operator ψ exists. However, ψ is not continuous.

Proof. The strong convexity of $f(\cdot,p)$ with parameter $\gamma = 1$ follows from Lemma 2.1.30 (b).

It is easy to see that the (unique) minimizer exists and satisfies

$$\psi(p)(\omega) = \begin{cases} \sin(p^{-1}\omega) & \text{if } p > 0, \\ 0 & \text{if } p \leq 0 \end{cases} \tag{3.2}$$

for almost all $\omega \in \Omega$. Then it can be calculated that $\|\psi(p)\|^2_{L^2(\Omega)} \geq 1/4$ holds for $p \in (0,1)$. Since $\psi(0) = 0$ this shows that ψ is not continuous.

It remains to show that f is continuous. Let $\{x_i\}_{i\in\mathbb{N}} \subset L^2(\Omega)$ be a sequence such that $x_i \to x_0$ for some element $x_0 \in L^2(\Omega)$ and let $\{p_i\}_{i\in\mathbb{N}} \subset \mathbb{R}$ be a sequence such that $p_i \to p_0$ for some element $p_0 \in \mathbb{R}$. If $p_0 < 0$ then $f(x_i,p_i) \to f(x_0,p_0)$ is trivial, and if $p_0 > 0$ the convergence $f(x_i,p_i) \to f(x_0,p_0)$ can be obtained with the triangle inequality and the dominated convergence theorem. If $p_0 = 0$, then it suffices to only consider the case where $p_i > 0$ for all $i \in \mathbb{N}$. It can be shown that $\sin(p_i^{-1}\omega) \rightharpoonup 0$ as $i \to \infty$ in the weak topology on $L^2(\Omega)$. Then it follows that $\int_0^1 \sin(p_i^{-1}\omega)x_i(\omega)\,\mathrm{d}\omega \to 0$ as $i \to \infty$, and since the continuity of $x \mapsto \frac{1}{2}\|x\|^2_{L^2(\Omega)}$ is obvious, the claim follows.

We are also interested in results which give us better properties of the solution operator ψ than mere continuity. In the next lemma we show (under some assumptions) that ψ is locally Lipschitz continuous. The assumptions include that we only consider the case of a fixed feasible set. The result has some similarities to the one in [Bonnans, Shapiro, 2000, Proposition 4.36]. We prove local Lipschitz continuity of ψ, which is a stronger property than the Lipschitzian stability at a point $\bar{p} \in V$ shown in [Bonnans, Shapiro, 2000, Proposition 4.36], but we also use stronger assumptions. The case where f is a simple quadratic function is important for applications.

Lemma 3.1.6. Suppose that $X = Y$, $g(x,p) = x$, Φ is convex, and $f(\cdot,p)$ is strongly convex with parameter $\gamma > 0$ (which is independent of p) on the feasible set Φ and Gâteaux differentiable for every $p \in V$. Moreover, we assume that the partial Gâteaux derivative $f'_x : X \times V \to X^\star$ is locally Lipschitz continuous. Then the solution operator ψ exists and is locally Lipschitz continuous.

Moreover, if C_L is a local Lipschitz constant of f'_x then the local Lipschitz constant of ψ can be chosen as $C_L\gamma^{-1}$. If f is quadratic and continuous then ψ is globally Lipschitz continuous with Lipschitz constant $\gamma^{-1}\|f''_{xp}(0,0)\|$.

Proof. The existence of ψ follows from Lemma 2.3.2 (b).

Let $p_0 \in V$ be given. Let $C_L \geq 0$ be a Lipschitz constant of f'_x on the set $B_{\varepsilon_1}(\psi(p_0)) \times B_{\varepsilon_1}(p_0)$ for some $\varepsilon_1 > 0$. Then we introduce the constant

$$\alpha := \sup\{\|p_1 - p_2\|^{-1}\|f'_x(x,p_1) - f'_x(x,p_2)\| \mid x \in B_{\varepsilon_1}(\psi(p_0)), p_1, p_2 \in B_{\varepsilon_1}(p_0), p_1 \neq p_2\}.$$

It can be seen that $\alpha \leq C_L$ and that $\alpha = \|f''_{xp}(0,0)\|$ if f is quadratic.

We define the modified feasible set $\hat{\Phi} := B_{\varepsilon_1}(\psi(p_0)) \cap \Phi$. Let us consider fixed elements $p_1, p_2 \in B_{\varepsilon_1}(p_0)$ and let $i \in \{1,2\}$ be given. Then there exists a unique minimizer $x_i \in \hat{\Phi}$ of $f(\cdot, p_i)$ over $\hat{\Phi}$, see Lemma 2.3.2 (b). From the optimality conditions of this convex optimization problem we obtain that

$$\langle f'_x(x_i, p_i), \hat{x} - x_i\rangle_{X^\star \times X} \geq 0 \quad \forall \hat{x} \in \hat{\Phi} \tag{3.3}$$

holds. Moreover, since $f(\cdot, p_i)$ is strongly convex on Φ, we also have

$$\gamma\|\hat{x} - x_i\|_X^2 \leq \langle f'_x(\hat{x}, p_i) - f'_x(x_i, p_i), \hat{x} - x_i\rangle_{X^\star \times X} \quad \forall \hat{x} \in \hat{\Phi},$$

see Lemma 2.1.30 (b). If we add the inequality (3.3) twice and use x_{3-i} for $\hat{x}$, we obtain the inequality

$$\gamma\|x_{3-i} - x_i\|_X^2 \leq \langle f'_x(x_{3-i}, p_i) + f'_x(x_i, p_i), x_{3-i} - x_i\rangle_{X^\star \times X}.$$

Now, this inequality can be added for the cases $i = 1, 2$. We obtain

$$\begin{aligned} 2\gamma\|x_2 - x_1\|^2 &\leq \langle f'_x(x_2, p_1) - f'_x(x_2, p_2) - f'_x(x_1, p_2) + f'_x(x_1, p_1), x_2 - x_1\rangle_{X^\star \times X} \\ &\leq \Big(\|f'_x(x_2, p_1) - f'_x(x_2, p_2)\| + \|f'_x(x_1, p_2) - f'_x(x_1, p_1)\|\Big)\|x_2 - x_1\|. \end{aligned}$$

Since $(x_1, p_1), (x_2, p_1), (x_1, p_2), (x_2, p_2) \in B_{\varepsilon_1}(\psi(p_0)) \times B_{\varepsilon_1}(p_0)$, we can use the constant α to further estimate

$$2\gamma\|x_2 - x_1\|^2 \leq 2\alpha\|p_1 - p_2\|\|x_2 - x_1\|$$

which results in

$$\|x_2 - x_1\| \leq \alpha\gamma^{-1}\|p_2 - p_1\|. \tag{3.4}$$

Now we choose $\varepsilon_2 > 0$ such that $\varepsilon_2 < \varepsilon_1$ and $\alpha\gamma^{-1}\varepsilon_2 < \varepsilon_1$. Let $p_3 \in B_{\varepsilon_2}(p_0) \subset B_{\varepsilon_1}(p_0)$ be given. In the same way as before, we denote the unique solutions of $f(\cdot, p_i)$ over $\hat{\Phi}$ by x_i for $i \in \{0, 3\}$. It is clear that $x_0 = \psi(p_0)$. Due to (3.4) we obtain

$$\|x_3 - x_0\| \leq \alpha\gamma^{-1}\|p_3 - p_0\| \leq \alpha\gamma^{-1}\varepsilon_2 < \varepsilon_1.$$

This means that x_3 is in the interior of $B_{\varepsilon_1}(\psi(p_0))$. Due to the construction of $\hat{\Phi}$ we obtain that x_3 is a local minimizer of $f(\cdot, p_3)$ on the set Φ. Because $f(\cdot, p_3)$ and Φ are convex it follows that $x_3 = \psi(p_3)$. Thus, if $p_1, p_2 \in B_{\varepsilon_2}(p_0)$ we can conclude

$$\|\psi(p_1) - \psi(p_2)\| \leq \alpha\gamma^{-1}\|p_1 - p_2\|$$

from (3.4). This shows that ψ is locally Lipschitz continuous at p_0 with Lipschitz constant $\alpha\gamma^{-1} > 0$.

For the case that f is quadratic, we can choose ε_1 and ε_2 arbitrarily large. This shows that ψ is globally Lipschitz continuous with Lipschitz constant $\alpha\gamma^{-1}$.

We would like to generalize Lemma 3.1.6 for cases where the feasible set can depend on the parameter p. Before we do that we provide a lemma that allows us to estimate Lagrange multipliers under a regularity condition.

Lemma 3.1.7. Let $\bar{p} \in V$ be given. Suppose that ψ exists in a neighborhood of $\bar{p}$ and is continuous at $\bar{p}$. We also assume that Φ is convex, f, g are continuously Fréchet differentiable, and that the partial derivatives f'_x and g'_x are locally Lipschitz continuous. As a regularity condition, we assume that $g'_x(\psi(\bar{p}), \bar{p})$ is surjective. Then there exists a constant $C > 0$ such that for all p_1, p_2 in a neighborhood of $\bar{p}$ there exist unique Lagrange multipliers λ_1, λ_2 for the problems (P(p_1)) and (P(p_2)) and the estimate

$$\|\lambda_1 - \lambda_2\|_{Y^\star} \leq C(\|p_1 - p_2\|_V + \|\psi(p_1) - \psi(p_2)\|_X)$$

holds.

Proof. Note that $p \mapsto g'_x(\psi(p), p)$ is continuous at $\bar{p} \in V$. Thus, according to Lemma 2.1.7 (b) the operators $g'_x(\psi(p), p)$ are surjective for all p in a neighborhood of $\bar{p}$. Then RZKCQ follows for these $p \in V$. Thus, for points p_1, p_2 in a sufficiently small neighborhood of $\bar{p}$ there exist Lagrange multipliers $\lambda_1, \lambda_2 \in Y^\star$ such that

$$f'_x(\psi(p_1), p_1) + g'_x(\psi(p_1), p_1)^\star\lambda_1 = f'_x(\psi(p_2), p_2) + g'_x(\psi(p_2), p_2)^\star\lambda_2 = 0$$

holds. The uniqueness of the Lagrange multipliers λ_1, λ_2 follows from the injectivity of $g'_x(\psi(p_1), p_1)^\star$ and $g'_x(\psi(p_2), p_2)^\star$. Let us first find a bound for $\|\lambda_1\|$. Due to Lemma 2.1.7 (a) we have

$$\|\lambda_1\| \leq C\|g'_x(\psi(p_1), p_1)^\star\lambda_1\| = C\|f'_x(\psi(p_1), p_1)\| < C^2$$

for a suitable constant $C > 0$. Then we can use the estimate Lemma 2.1.7 (c) to obtain

$$\begin{aligned}\|\lambda_1 - \lambda_2\| &\leq C(\|g_x'(\psi(p_1), p_1)^\star \lambda_1 - g_x'(\psi(p_2), p_2)^\star \lambda_2\| \\ &\quad + \|g_x'(\psi(p_1), p_1) - g_x'(\psi(p_2), p_2)\|\|\lambda_1\|) \\ &= C(\|f_x'(\psi(p_1), p_1) - f_x'(\psi(p_2), p_2)\| \\ &\quad + \|g_x'(\psi(p_1), p_1) - g_x'(\psi(p_2), p_2)\|\|\lambda_1\|) \\ &\leq (C + \|\lambda_1\|)(\|p_1 - p_2\| + \|\psi(p_1) - \psi(p_2)\|) \\ &\leq (C + C^2)(\|p_1 - p_2\| + \|\psi(p_1) - \psi(p_2)\|)\end{aligned}$$

for a suitable constant $C > 0$. This completes the proof.

A similar result for a Lipschitzian dependence of Lagrange multipliers on a parameter can be found in [Bonnans, Shapiro, 2000, Lemma 4.44].

Now we are in a position to generalize Lemma 3.1.6 for the case that the feasible set depends on the parameter p. A notable additional assumption is the surjectivity of $g_x'(\psi(\bar{p}), \bar{p})$, where $\bar{p} \in V$ is the point where we want to show continuity.

Lemma 3.1.8. Let $\bar{p} \in V$ be given and suppose that

(a) the sets Φ, $g(\cdot, p)^{-1}(\Phi)$ are convex and nonempty for all $p \in V$,

(b) the space X is reflexive,

(c) there exists a constant $\gamma > 0$ such that for all $p \in V$ the function $f(\cdot, p)$ is strongly convex with parameter γ on the feasible set $g(\cdot, p)^{-1}(\Phi)$,

(d) f, g are continuously Fréchet differentiable and the partial derivatives f_x', g_x' are locally Lipschitz continuous,

(e) the regularity condition that $g_x'(\psi(\bar{p}), \bar{p})$ is surjective holds.

Then the solution operator ψ exists and is locally Lipschitz continuous in a neighborhood of $\bar{p}$.

Proof. The existence of ψ follows from Lemma 2.3.2 (b).

In order to show that ψ is locally Lipschitz continuous in a neighborhood of $\bar{p}$, we first prove that ψ is continuous in $\bar{p}$. We aim to do this by using Lemma 3.1.4. The nontrivial assumptions of Lemma 3.1.4 are assumptions (c) and (f). Let $\{p_i\}_{i\in\mathbb{N}} \subset V$ be a sequence with $p_i \to \bar{p}$. Since $g_x'(\psi(\bar{p}), \bar{p})$ is surjective, we can apply Lemma 2.1.13 at the point $(\psi(\bar{p}), \bar{p})$. This yields the existence of a sequence $\{x_i\}_{i\in\mathbb{N}} \subset X$ such that $g(x_i, p_i) = g(\psi(\bar{p}), \bar{p}) \in \Phi$ and $\|x_i - \psi(\bar{p})\| \leq C\|p_i - \bar{p}\|$ hold for large $i \in \mathbb{N}$ and a constant $C > 0$. Thus, assumption (c) of Lemma 3.1.4 is shown. Next, we will show assumption (f) of Lemma 3.1.4. Again, let $\{p_i\}_{i\in\mathbb{N}} \subset V$ be a sequence with $p_i \to \bar{p}$ and let $\{x_i\}_{i\in\mathbb{N}} \subset B_\varepsilon(\psi(\bar{p}))$ be a sequence with $x_i \rightharpoonup x_0$ for some $x_0 \in B_\varepsilon(\psi(\bar{p}))$, where $\varepsilon > 0$ is sufficiently small. Since f is locally Lipschitz continuous, we can assume that there

exists a Lipschitz constant $C_L > 0$ for f on $B_\varepsilon(\psi(\bar{p})) \times B_\varepsilon(\bar{p})$. If we combine this with the convexity of $f(\cdot, \bar{p})$ we have

$$f(x_0, \bar{p}) \leq \liminf_{i\to\infty} f(x_i, \bar{p}) \leq \liminf_{i\to\infty} (f(x_i, p_i) + C_L\|p_i - \bar{p}\|) = \liminf_{i\to\infty} f(x_i, p_i).$$

Now suppose that $g(x_i, p_i) \in \Phi$ for all $i \in \mathbb{N}$. We want to show that this implies $g(x_0, \bar{p}) \in \Phi$. Since g'_x is continuous we know that there exists a constant $\alpha > 0$ such that $B_1(0) \subset g'_x(x, p)B_\alpha(0)$ holds for all $(x, p) \in B_{2\varepsilon}(\psi(\bar{p})) \times B_{2\varepsilon}(\bar{p})$ if $\varepsilon > 0$ is sufficiently small, see Lemma 2.1.7 (b). Without loss of generality we can also assume that g, g'_x, f'_x are Lipschitz continuous with Lipschitz constant C_L on $B_{2\varepsilon}(\psi(\bar{p})) \times B_{2\varepsilon}(\bar{p})$. Now we can apply Lemma 2.1.13 (b) at the point (x_i, p_i) with $\varepsilon_0 := \varepsilon$. This yields the existence of a sequence $\{\hat{x}_i\}_{i\in\mathbb{N}} \subset B_{2\varepsilon}(\psi(\bar{p}))$ such that

$$g(\hat{x}_i, \bar{p}) = g(x_i, p_i) \qquad \text{and} \qquad \|\hat{x}_i - x_i\| \leq 2\alpha C_L\|p_i - \bar{p}\|$$

holds if $\|p_i - \bar{p}\|$ is sufficiently small. The weak convergence $\hat{x}_i \rightharpoonup x_0$ follows. Since $g(\hat{x}_i, \bar{p}) \in \Phi$ for all $i \in \mathbb{N}$ and $g(\cdot, \bar{p})^{-1}(\Phi)$ is a convex and closed set we get $g(x_0, \bar{p}) \in \Phi$. Thus, assumption (f) of Lemma 3.1.4 is shown. Therefore, we can apply Lemma 3.1.4 which yields that ψ is continuous at $\bar{p}$.

Let $p_1, p_2 \in B_{\varepsilon_1/2}(\bar{p})$ be given, where $\varepsilon_1 > 0$ is sufficiently small. Our goal is to show a Lipschitz estimate for ψ at these points. We use the abbreviations $x_1 := \psi(p_1)$ and $x_2 := \psi(p_2)$. Note that due to the continuity of ψ we have $x_1, x_2 \in B_\varepsilon(\psi(\bar{p}))$ if $\varepsilon_1 > 0$ is sufficiently small. This allows us to apply Lemma 2.1.13 (b) again at the point (x_2, p_2). Because of $p_1 \in B_{\varepsilon_1}(p_2)$ this yields the existence of a point $x_3 \in X$ such that

$$g(x_3, p_1) = g(x_2, p_2) \in \Phi \qquad \text{and} \qquad \|x_3 - x_2\| \leq 2\alpha C_L\|p_1 - p_2\| \tag{3.5}$$

hold. Note that $\alpha, C_L, \varepsilon, \varepsilon_1$ do not depend on p_1, p_2 here.

Since ψ is continuous, we can apply Lemma 3.1.7 (if we make $\varepsilon_1 > 0$ smaller if necessary). Thus, for $i = 1, 2$ there are Lagrange multipliers $\lambda_i \in \mathcal{N}_\Phi(g(x_i, p_i))$, $\bar{\lambda} \in \mathcal{N}_\Phi(g(\psi(\bar{p}), \bar{p}))$ that correspond to the solutions $x_i, \psi(\bar{p}) \in X$ of the problems (P(p_i)) and (P($\bar{p}$)) such that

$$\|\lambda_i - \bar{\lambda}\|_{Y^\star} \leq C(\|p_i - \bar{p}\| + \|x_i - \psi(\bar{p})\|) \tag{3.6}$$

and

$$f'_x(x_i, p_i) + g'_x(x_i, p_i)^\star\lambda_i = 0 \tag{3.7}$$

hold. The estimates

$$\begin{aligned}
&\|-g'_x(x_3, p_1)^\star\lambda_2 - f'_x(x_3, p_1)\| \\
&\quad = \|-g'_x(x_3, p_1)^\star\lambda_2 + g'_x(x_2, p_2)^\star\lambda_2 + f'_x(x_2, p_2) - f'_x(x_3, p_1)\| \\
&\quad \leq (C_L\|\lambda_2\| + C_L)\|(x_3, p_1) - (x_2, p_2)\| \\
&\quad \leq C_L(\|\lambda_2\| + 1)(1 + 2\alpha C_L)\|p_1 - p_2\|
\end{aligned}$$

follow, where we used (3.5) and that $C_L > 0$ is a Lipschitz constant for f'_x and g'_x. Since λ_2 can be bounded by a constant due to (3.6), we have

$$\|-g'_x(x_3,p_1)^\star\lambda_2 - f'_x(x_3,p_1)\| \leq C\|p_1-p_2\|. \tag{3.8}$$

Because of $g(x_1,p_1) \in \Phi$ and $g(x_3,p_1) \in \Phi$ and the convexity of $g(\cdot,p_1)(\Phi)$ we know that $g(tx_1+(1-t)x_3,p_1) \in \Phi$ holds for every $t \in [0,1]$. Then the inequalities

$$\langle \lambda_1, g'_x(x_1,p_1)(x_3-x_1)\rangle_{Y^\star\times Y} \leq 0$$

and

$$\langle \lambda_2, g'_x(x_3,p_1)(x_1-x_3)\rangle_{Y^\star\times Y} \leq 0$$

follow from $\lambda_1 \in \mathcal{N}_\Phi(g(x_1,p_1))$ and $\lambda_2 \in \mathcal{N}_\Phi(g(x_2,p_2)) = \mathcal{N}_\Phi(g(x_3,p_1))$. By adding these inequalities we obtain

$$0 \leq \langle g'_x(x_1,p_1)^\star\lambda_1 - g'_x(x_3,p_1)^\star\lambda_2, x_1-x_3\rangle_{X^\star\times X}. \tag{3.9}$$

Since $f'_x(\cdot,p_1)$ is strongly convex on the feasible set $g(\cdot,p_1)^{-1}(\Phi)$ and $g(x_3,p_1) \in \Phi$ holds, we can use Lemma 2.1.30 (b), which results in

$$\begin{aligned}\gamma\|x_1-x_3\|^2 &\leq \langle f'_x(x_1,p_1) - f'_x(x_3,p_1), x_1-x_3\rangle\\ &= \langle -g'_x(x_1,p_1)^\star\lambda_1 - f'_x(x_3,p_1), x_1-x_3\rangle\\ &\leq \langle -g'_x(x_3,p_1)^\star\lambda_2 - f'_x(x_3,p_1), x_1-x_3\rangle\\ &\leq C\|p_1-p_2\|\|x_1-x_3\|,\end{aligned}$$

where we also used (3.7), (3.9), and (3.8). It follows that

$$\|x_1-x_3\| \leq C\gamma^{-1}\|p_1-p_2\|$$

holds. Then the claimed Lipschitz estimate follows from (3.5) via

$$\|\psi(p_1)-\psi(p_2)\| = \|x_1-x_2\| \leq \|x_1-x_3\| + \|x_3-x_2\| \leq (C\gamma^{-1}+2\alpha C_L)\|p_1-p_2\|.$$

We mention that a Lipschitzian stability result for the Lagrange multipliers and the solution operator can be found in [Bonnans, Shapiro, 2000, Theorem 4.51]. However, conditions involving the second derivatives are required in that theorem, which is not the case in our results Lemmas 3.1.7 and 3.1.8.

We give two examples to highlight the importance of strong convexity in order to obtain local Lipschitz continuity of the solution operator ψ. First, we provide an example where the solution operator ψ is continuous despite the lack of strong convexity of $f(\cdot,p)$, but not locally Hölder continuous (and therefore also not Lipschitz continuous).

Example 3.1.9. We consider the case that $X = V = Y := L^2(\Omega)$ with $\Omega := (0,1)$. We choose

$$\begin{aligned} f(x,p) &:= \int_0^1 \tfrac{1}{2}\exp(-\omega^{-1})x^2 - px \,\mathrm{d}\omega, \\ g(x,p) &:= x, \\ \Phi &:= \{y \in L^2(\Omega) \mid 0 \le y \le 1 \text{ a.e. in } \Omega\}. \end{aligned}$$

Then ψ exists and is continuous, but not locally Hölder continuous. Here, $f(\cdot,p)$ is strictly convex but not strongly convex for every $p \in V$.

Proof. Since the objective function is continuous and strictly convex and the feasible set is convex and bounded it follows that there exists a unique solution of (P(p)) for each $p \in V$. The KKT conditions can be written as

$$\exp(-\omega^{-1})x - p + \lambda = 0, \qquad \lambda \in \mathcal{N}_\Phi(x), \qquad x \in \Phi.$$

It can be calculated that x defined via

$$x(\omega) = \begin{cases} 0 & \text{if } p(\omega) \le 0, \\ \exp(\omega^{-1})p(\omega) & \text{if } 0 < p(\omega) < \exp(-\omega^{-1}), \\ 1 & \text{if } p(\omega) \ge \exp(-\omega^{-1}) \end{cases} \tag{3.10}$$

and $\lambda := p - \exp(-\omega^{-1})x$ satisfy these KKT conditions. This confirms that (3.10) defines a solution of (P(p)) for a given $p \in V$, see Proposition 2.3.8.

Next, we will argue that ψ is continuous. Let $\{p_i\}_{i\in\mathbb{N}}$ be a sequence in $L^2(\Omega)$ that converges to a function $p_0 \in L^2(\Omega)$. Then there exists a subsequence $\{p_{i_j}\}_{j\in\mathbb{N}}$ of $\{p_i\}_{i\in\mathbb{N}}$ that converges pointwise almost everywhere to p_0. From the solution formula (3.10) it follows that the sequence $\{\psi(p_{i_j})\}_{j\in\mathbb{N}}$ converges pointwise almost everywhere to $\psi(p_0)$. Since we have $|\psi(p_{i_j})| \le 1$ a.e. in Ω we can apply the dominated convergence theorem which implies that $\psi(p_{i_j}) \to \psi(p_0)$ in $L^2(\Omega)$ as $j \to \infty$. Since the limit $\psi(p_0)$ is independent of the subsequence $\{i_j\}_{j\in\mathbb{N}}$ it follows that the sequence $\{\psi(p_i)\}_{i\in\mathbb{N}}$ converges already to $\psi(p_0)$. Thus, $\psi : V \to X$ is a continuous function.

Finally, we show that ψ is not locally Hölder continuous at $0 \in V$. We define the sequence $\{p_i\}_{i\in\mathbb{N}}$ via $p_i(\omega) := \exp(-\omega^{-1})\chi_{(0,1/i)}$ for $i \in \mathbb{N}$. For a given $i \in \mathbb{N}$ we can calculate that $p_i \in L^2(\Omega)$ with

$$\|p_i\|_{L^2(\Omega)} = \Big(\int_0^{1/i} \exp(-\omega^{-1})^2 \,\mathrm{d}\omega\Big)^{1/2} \le (\tfrac{1}{i}\exp(-2i))^{1/2} \le \exp(-i)$$

and we obtain that the solution is given by $\psi(p_i) := \chi_{(0,1/i)}$ from (3.10). Thus, we obtain $\|\psi(p_i)\|_{L^2(\Omega)} = i^{-1/2}$. It follows that $p_i \to 0$ and $\psi(p_i) \to \psi(0) = 0$ as $i \to \infty$. Now let us

suppose that $\alpha > 0$ is a Hölder exponent of ψ on a neighborhood of $0 \in V$. Then we have

$$\frac{\|\psi(p_i) - \psi(0)\|_{L^2(\Omega)}}{\|p_i - 0\|^{\alpha}_{L^2(\Omega)}} \geq i^{-1/2} \exp(\alpha i) \to \infty \qquad (i \to \infty).$$

Since $p_i \to 0$, it follows that there can be no local Hölder exponent of ψ on a neighborhood of $0 \in V$.

If the objective function is not strongly convex then it can even happen that the solution operator is not even continuous. This is the case in the following example. In comparison to Example 3.1.9 we have a different feasible set. Again, the objective function $f(\cdot, p)$ is strictly convex but not strongly convex for every $p \in V$. With the exception of the strong convexity, all assumptions of Lemma 3.1.6 are satisfied. Although the solution operator ψ is well-defined, it is not continuous.

Example 3.1.10. We consider the case that $X = V = Y := L^2(\Omega)$ with $\Omega := (0, 1)$. We choose

$$\begin{aligned} f(x, p) &:= \int_0^1 \tfrac{1}{2}\omega x^2 - px \,\mathrm{d}\omega, \\ g(x, p) &:= x, \\ \Phi &:= B_1(0). \end{aligned}$$

Then there exists a unique solution of (P(p)) for each $p \in V$, but the solution operator ψ is not continuous.

Proof. In order to simplify the analysis of the KKT conditions, we use the equivalent problem with $Y = \mathbb{R}$, $g(x, p) = \|x\|^2_{L^2(\Omega)} - 1$, and $\Phi = (-\infty, 0] \subset \mathbb{R}$. Since the objective function is continuous and strictly convex and the feasible set is convex and bounded it follows that there exists a unique solution of (P(p)). Thus, the solution operator ψ exists. To show that ψ is not continuous, we define the sequence $\{p_i\}_{i\in\mathbb{N}}$ in V via $p_i := 1/i$ for $i \in \mathbb{N}$. Clearly, $p_i \to 0$ in $L^2(\Omega)$ as $i \to \infty$.

Let $i \in \mathbb{N}$ be given. We want to calculate $\psi(p_i)$. The KKT conditions corresponding to (P(p_i)) can be written as

$$\omega x - p_i + 2\alpha x = 0, \qquad \|x\|^2_{L^2(\Omega)} - 1 \leq 0, \qquad \alpha \geq 0, \qquad \alpha(\|x\|^2_{L^2(\Omega)} - 1) = 0, \quad (3.11)$$

where $\alpha \in \mathbb{R}$ is a multiplier and $x \in L^2(\Omega)$. Let us define

$$\alpha_i := i^{-2}(1 + \sqrt{1 + 4/i^2})^{-1} \qquad \text{and} \qquad x_i(\omega) := (\omega + 2\alpha_i)^{-1}/i \quad \forall \omega \in \Omega.$$

Then, after some calculations, it can be seen that $x_i \in L^2(\Omega)$ with

$$\|x_i\|^2_{L^2(\Omega)} = i^{-2}(2\alpha_i(1 + 2\alpha_i))^{-1} = 1$$

and

$$\omega x_i(\omega) - 1/i + 2\alpha_i x_i(\omega) = 0 \quad \forall \omega \in \Omega.$$

Thus, $\alpha_i \geq 0$ and $x_i \in L^2(\Omega)$ satisfy the KKT conditions (3.11). We conclude from Proposition 2.3.8 that x_i is a solution of (P(p_i)), i.e. $\psi(p_i) = x_i$.

It is easy to see that $0 \in L^2(\Omega)$ is the solution of (P(0)), i.e. $\psi(0) = 0$. Since $p_i \to 0$ as $i \to \infty$ and $\|\psi(p_i)\| = 1$ for $i \in \mathbb{N}$ we conclude that ψ is not continuous.

We comment that although the sequence $\{\psi(p_i)\}_{i\in\mathbb{N}}$ in the proof of Example 3.1.10 does not converge to $\psi(0)$, it can be shown that it converges weakly to $\psi(0)$.

3.1.3 Differentiability properties of solution operators

In this section we assume that $X = Y$ and $g(x, p) = x$, i.e. the feasible set is independent of the parameters. In Section 3.1.2 we discussed continuity properties of the solution operator ψ. Now we want to investigate differentiability properties of ψ, such as Fréchet differentiability or directional differentiability. For the first result we need to introduce the notion of a Legendre-$\star$ form.

Definition 3.1.11. Let X be a normed space. We say that a quadratic form $Q : X^\star \to \mathbb{R}$ is a *Legendre-$\star$ form* if it is sequentially weakly-$\star$ lower semi-continuous and if the implication

$$x_i \overset{\star}{\rightharpoonup} x,\ Q(x_i) \to Q(x) \quad \Rightarrow \quad x_i \to x$$

holds for all sequences $\{x_i\}_{i\in\mathbb{N}} \subset X^\star$ and $x \in X^\star$.

For more information on Legendre-$\star$ forms we refer to Chapter 4. The following proposition together with its proof are influenced by [Bonnans, Shapiro, 2000, Theorem 5.5]. It also has similarities to [Christof, G. Wachsmuth, 2020, Corollary 3.1]. We formulate this result for reflexive spaces as well as nonreflexive spaces with a separable predual space.

Proposition 3.1.12. Let $p_0 \in V$ be a point and $h \in V$ be a direction. Suppose that

(a) the solution operator ψ exists and is continuous at p_0,

(b) there exists a Banach space $X^{-\star}$ that is separable or reflexive such that $X \cong (X^{-\star})^\star$,

(c) f and f'_x are Fréchet differentiable,

(d) $f'_x(\psi(p_0), p_0)$ and $f''_{xp}(\psi(p_0), p_0)h$ are contained in the linear subspace $X^{-\star} \subset X^\star$,

(e) Φ is convex and polyhedric,

(f) $\mathcal{T}_\Phi(\psi(p_0))$ is weakly-$\star$ closed,

(g) the quadratic form $y \mapsto \langle f''_{xx}(\psi(p_0), p_0)y, y\rangle$ is a Legendre-$\star$ form on X.

Let $\{t_i\}_{i\in\mathbb{N}} \subset (0,\infty)$ be a sequence with $t_i \to 0$. Then any weak-$\star$ limit point of the sequence $\{t_i^{-1}(\psi(p_0+t_ih)-\psi(p_0))\}_{i\in\mathbb{N}}$ is also a strong limit point and solves the variational inequality

$$\text{find } \hat{y}\in\mathcal{K} \;:\; \langle f''_{xx}(\cdot)\hat{y}+f''_{xp}(\cdot)h, y-\hat{y}\rangle_{X^\star\times X}\geq 0 \quad \forall y\in\mathcal{K}. \tag{3.12}$$

Here, $(\cdot)$ is an abbreviation for $(\psi(p_0),p_0)$ and $\mathcal{K}$ is an abbreviation for the critical cone $\mathcal{K}_\Phi(\psi(p_0), f'_x(\psi(p_0),p_0))$.

Moreover, if (3.12) has at most one solution and ψ is locally Lipschitz continuous, then ψ is directionally differentiable at p_0 in direction h.

Proof. Let $x^h\in X$ be a weak-$\star$ limit point of the sequence $\{t_i^{-1}(\psi(p_0+t_ih)-\psi(p_0))\}_{i\in\mathbb{N}}$. To shorten notation, we define $x_i^h := t_i^{-1}(\psi(p_0+t_ih)-\psi(p_0))$ for $i\in\mathbb{N}$. Without loss of generality we can assume that $x_i^h \overset{\star}{\rightharpoonup} x^h$. Our first goal is to show that x^h solves (3.12).

Because $\{x_i^h\}_{i\in\mathbb{N}}\subset\mathcal{R}_\Phi(\psi(p_0))\subset\mathcal{T}_\Phi(\psi(p_0))$ and $\mathcal{T}_\Phi(\psi(p_0))$ is weakly-$\star$ closed, the weak-$\star$ limit x^h is also contained in $\mathcal{T}_\Phi(\psi(p_0))$. Since $\psi(p_0+t_ih)$ and $\psi(p_0)$ are global minimizers of (P(p_0+t_ih)) and (P(p_0)) we obtain from the optimality conditions that the inequalities

$$\langle f'_x(\psi(p_0+t_ih),p_0+t_ih), x_i^h\rangle \leq 0, \tag{3.13a}$$
$$-\langle f'_x(\psi(p_0),p_0), x_i^h\rangle \leq 0 \tag{3.13b}$$

hold. If we consider that $f'_x(\psi(p_0),p_0)\in X^{-\star}$ as well as Lemma 2.1.4 we can take the limit for $i\to\infty$ in these inequalities which results in

$$\langle f'_x(\psi(p_0),p_0), x^h\rangle = 0.$$

Thus, we have shown that $x^h\in\mathcal{K}_\Phi(\psi(p_0), f'_x(\psi(p_0),p_0))$.

Now let $y\in\mathcal{R}_\Phi(\psi(p_0))\cap f'_x(\psi(p_0),p_0)^\perp$ be given. Then there exists $\alpha\geq 0$, $x\in\Phi$ such that $y=\alpha(x-\psi(p_0))$. From the optimality conditions of (P(p_0+t_ih)) we obtain that

$$\begin{aligned}&\langle f'_x(\psi(p_0+t_ih),p_0+t_ih), y-\alpha t_i x_i^h\rangle\\ &= \langle f'_x(\psi(p_0+t_ih),p_0+t_ih), \alpha(x-\psi(p_0+t_ih))\rangle \geq 0\end{aligned}$$

holds. Using $\langle f'_x(\psi(p_0),p_0), y\rangle = 0$ yields

$$\langle f'_x(\psi(p_0+t_ih),p_0+t_ih) - f'_x(\psi(p_0),p_0), y-\alpha t_i x_i^h\rangle \geq \langle f'_x(\psi(p_0),p_0), \alpha t_i x_i^h\rangle. \tag{3.14}$$

If we add the inequalities (3.13a), (3.13b), and (3.14) and multiply them by t_i^{-1} we get the inequality

$$t_i^{-1}\langle f'_x(\psi(p_0+t_ih),p_0+t_ih) - f'_x(\psi(p_0),p_0), (1+\alpha t_i)x_i^h - y\rangle \leq \langle f'_x(\psi(p_0),p_0), -\alpha x_i^h\rangle.$$

Because

$$\left\| f''_{xx}(\psi(p_0),p_0)x_i^h + f''_{xp}(\psi(p_0),p_0)h - t_i^{-1}(f'_x(\psi(p_0+t_ih),p_0+t_ih) - f'_x(\psi(p_0),p_0)) \right\|$$

converges to 0 as $i \to \infty$, it follows that

$$\begin{aligned} &\limsup_{i\to\infty} \langle f''_{xx}(\psi(p_0),p_0)x_i^h + f''_{xp}(\psi(p_0),p_0)h, (1+\alpha t_i)x_i^h - y\rangle \\ &\quad \le \limsup_{i\to\infty} \langle f'_x(\psi(p_0),p_0), -\alpha x_i^h\rangle = \langle f'_x(\psi(p_0),p_0), -\alpha x^h\rangle = 0 \end{aligned} \tag{3.15}$$

holds. Here, we used $f'_x(\psi(p_0),p_0) \in X^{-\star}$ again. Using the weak-$\star$ convergence $x_i^h \overset{\star}{\rightharpoonup} x^h$, $f''_{xp}(\psi(p_0),p_0)h \in X^{-\star}$, and the fact that the quadratic form in assumption (g) is sequentially weakly-$\star$ lower semi-continuous, we arrive at the inequality

$$\langle f''_{xx}(\psi(p_0),p_0)x^h + f''_{xp}(\psi(p_0),p_0)h, x^h - y\rangle \le 0.$$

Since y was arbitrary in $\mathcal{R}_\Phi(\psi(p_0)) \cap f'_x(\psi(p_0),p_0)^\perp$ and by assumption (e) this set is dense in $\mathcal{K}_\Phi(\psi(p_0), f'_x(\psi(p_0),p_0))$, it follows that x^h solves (3.12).

Our next goal is to show that $x_i^h \to x^h$ holds. If we set $y = 2x^h$ and $y = 0$ in (3.12) then the equality

$$\langle f''_{xx}(\psi(p_0),p_0)x^h + f''_{xp}(\psi(p_0),p_0)h, x^h\rangle = 0 \tag{3.16}$$

can be concluded. Again, using the weak-$\star$ convergence $x_i^h \overset{\star}{\rightharpoonup} x^h$ and the fact that the quadratic form in assumption (g) is sequentially weakly-$\star$ lower semi-continuous yields

$$0 \le \liminf_{i\to\infty} \langle f''_{xx}(\psi(p_0),p_0)x_i^h + f''_{xp}(\psi(p_0),p_0)h, x_i^h\rangle$$

and if we use $y = 0$ and $\alpha = 0$ in (3.15) we even obtain

$$\lim_{i\to\infty} \langle f''_{xx}(\psi(p_0),p_0)x_i^h + f''_{xp}(\psi(p_0),p_0)h, x_i^h\rangle = 0.$$

From (3.16) and $f''_{xp}(\psi(p_0),p_0)h \in X^{-\star}$ we obtain

$$\langle f''_{xx}(\psi(p_0),p_0)x_i^h, x_i^h\rangle \to \langle f''_{xx}(\psi(p_0),p_0)x^h, x^h\rangle \qquad (i\to\infty).$$

Then the strong convergence $x_i^h \to x^h$ follows directly from the Legendre-$\star$ property of the quadratic form in assumption (g), see also Definition 3.1.11.

Now let us assume that (3.12) has at most one solution and that ψ is locally Lipschitz continuous. Then the sequence $\{t_i^{-1}(\psi(p_0+t_ih) - \psi(p_0))\}_{i\in\mathbb{N}}$ is bounded and therefore has at least one weak-$\star$ limit point, see Theorem 2.3.1. Since every weak-$\star$ limit point is also a strong limit point that solves (3.12) and there can be at most one solution of (3.12), a unique (strong) limit point x^h of $\{t_i^{-1}(\psi(p_0+t_ih) - \psi(p_0))\}_{i\in\mathbb{N}}$ exists. Additionally, the limit x^h is even independent of the sequence $\{t_i\}_{i\in\mathbb{N}}$. Thus, $x^h = \lim_{t\downarrow 0} t^{-1}(\psi(p_0 + th) - \psi(p_0))$ is the directional derivative of ψ at p_0 in direction h.

We remark that assumptions (d) and (f) of Proposition 3.1.12 are automatically satisfied if X is a reflexive Banach space.

We also used the assumption that a certain quadratic form is a Legendre-$\star$ form. This raises the question in which spaces Legendre-$\star$ forms can actually exist so that we can apply Proposition 3.1.12. We will discuss this question in detail in Chapter 4. Although the technical arguments presented in the proof of Proposition 3.1.12 can be applied if the Banach space X is reflexive or has a separable predual space, it turns out that the assumptions (b) and (g) imply that X is already isomorphic to a Hilbert space. This result will be shown in Theorem 4.3.9. Thus, Proposition 3.1.12 is only important in Hilbert spaces. and can be considered as a motivation for Chapter 4.

If Φ is a closed convex set which is not a linear subspace, then it is very natural that ψ has kinks and is not Fréchet differentiable but only directionally differentiable. If, however, Φ is a closed linear subspace, then, under some assumptions, the solution operator ψ can be even Fréchet differentiable.

Proposition 3.1.13. Let $p_0 \in V$ be given. Suppose that

(a) X is a Hilbert space,

(b) Φ is a closed linear subspace,

(c) f and f'_x are Fréchet differentiable,

(d) the operator $f''_{xx}(\psi(p_0), p_0) \in \mathbb{L}(X, X^\star)$ is coercive,

(e) ψ exists and is locally Lipschitz continuous.

Then ψ is Fréchet differentiable at p_0 with Fréchet derivative

$$\psi'(p_0) = -I_\Phi (I_\Phi^\star f''_{xx}(\psi(p_0), p_0) I_\Phi)^{-1} I_\Phi^\star f''_{xp}(\psi(p_0), p_0),$$

where $I_\Phi : \Phi \to X$, $x \mapsto x$ is the canonical embedding operator.

Proof. Since Φ is a closed linear subspace we can without loss of generality assume that $\Phi = X$. Then it follows from the optimality conditions of (P(p)) that $f'_x(\psi(p), p) = 0$ for all $p \in V$. Thus, for the critical cone we have $\mathcal{K}_X(\psi(p), f'_x(\psi(p), p)) = X$ for all $p \in V$. Therefore, for $h \in V$ the variational inequality (3.12) simplifies to

$$\text{find } \hat{y} \in X \ : \ f''_{xx}(\psi(p_0), p_0)\hat{y} = -f''_{xp}(\psi(p_0), p_0)h \tag{3.17}$$

in the current setting. Due to Lemma 2.1.35 the operator on the left-hand side is continuously invertible. Thus, (3.17) has a unique solution for each $h \in V$. It can be seen that the assumptions for Proposition 3.1.12 are satisfied for all $h \in V$ (a coercive quadratic form in a Hilbert space is a Legendre-$\star$ form, see also Theorem 4.2.1). Therefore, ψ is directionally differentiable at p_0 in all directions. Since the directional derivative of ψ in direction h is given as the solution of (3.17) it follows that ψ is Gâteaux differentiable

at p_0 with Gâteaux derivative

$$\psi'(p_0) = -f''_{xx}(\psi(p_0, p_0))^{-1} f''_{xp}(\psi(p_0), p_0).$$

Due to the optimality conditions of (P(p)) and (P($p+h$)) we have

$$\begin{aligned} 0 &= f'_x(\psi(p_0 + h), p_0 + h) - f'_x(\psi(p_0), p_0) \\ &= f''_{xx}(\psi(p_0), p_0)(\psi(p_0 + h) - \psi(p_0)) + f''_{xp}(\psi(p_0), p_0)h + o(\|h\|) \end{aligned}$$

as $\|h\| \to 0$. By applying the bounded linear operator $f''_{xx}(\psi(p_0), p_0)^{-1}$ we obtain

$$\psi(p_0 + h) - \psi(p_0) - \psi'(p_0)h = o(\|h\|)$$

as $\|h\| \to 0$, which shows that ψ is indeed Fréchet differentiable at p_0.

3.2 The optimal value function

3.2.1 Basic properties

Throughout Section 3.2, we work with the setting established in Section 3.1.1. We are interested in properties of the optimal value function of a parametrized optimization problem. The optimal value function was defined in Definition 3.1.2.

Lemma 3.2.1. Let $p \in V$ be given and let x be a feasible point of (P(p)). Then we have

(a) $\varphi(p) \leq f(x, p)$,

(b) $\varphi(p) = f(x, p) \quad \Leftrightarrow \quad x \in \Psi(p)$,

(c) if ψ exists, then $\varphi(p) = f(\psi(p), p)$.

Proof. The statements follow directly from the definition of the optimal value function φ.

From Lemma 3.2.1 (c), a simple corollary for the continuity of the optimal value function can be obtained.

Corollary 3.2.2. If ψ exists and f and ψ are continuous, then φ is continuous.

Sometimes it is possible to show that φ is continuous even when there are multiple solutions for some $p \in V$, see [Bonnans, Shapiro, 2000, Theorem 4.4] for a more general result.

In many situations φ is continuous even when ψ does not exist or is not continuous. In the following example, we consider a situation where ψ does not exist and another situation where φ is not continuous.

Example 3.2.3. (a) We consider the case that $X = V = Y := \mathbb{R}$ and choose $f(x,p) := px$, $g(x,p) := x$, and $\Phi := [0,1]$. Then ψ does not exist, but the optimal value function is given by $\varphi(p) = p^-$ which is continuous.

(b) We consider the setting of Example 3.1.5. Then the optimal value function φ is not continuous.

Proof. For part (a) we can calculate that

$$\Psi(p) = \begin{cases} \{1\} & \text{if } p < 0, \\ [0,1] & \text{if } p = 0, \\ \{0\} & \text{if } p > 0, \end{cases}$$

and thus $\varphi(p) = p^-$. Since Ψ is not single-valued, ψ does not exist.

Let us consider part (b). We recall that the solution operator was already calculated in (3.2). Thus, we obtain

$$\varphi(p) = f(\psi(p), p) = \begin{cases} -\frac{1}{2}\int_0^1 \sin^2(p^{-1}\omega)\,\mathrm{d}\omega & \text{if } p > 0, \\ 0 & \text{if } p \leq 0. \end{cases}$$

For $p \in (0,1)$ it can be calculated that $\varphi(p) \leq -1/8$ holds. Due to $\varphi(0) = 0$ this shows that φ is not continuous.

An important observation is that if we modify the objective function f by adding a term $\hat{f}(p)$ that only depends on p, the parametrized optimization problem (P(p)) produces the same solutions for each $p \in V$, but the optimal value function changes and is given by $\varphi(p) + \hat{f}(p)$. Such modifications can be helpful when we want to obtain certain desirable properties of φ without changing the underlying parametrized optimization problem.

3.2.2 Convexity and concavity of the optimal value function

In many parametrized optimization problems the optimal value function φ turns out to be convex or concave. A thorough discussion of this issue for finite-dimensional problems can be found in [Fiacco, Kyparisis, 1986]. However, many results from that article can be transferred to an infinite-dimensional setting without major modifications. We restrict ourselves to two results in this section. In one result the optimal value function is convex, and in the other result it is concave.

We start with the case where the feasible set and the objective function are convex. The result of the next proposition can be found in a finite-dimensional setting in [Fiacco, Kyparisis, 1986, Proposition 2.1] and can also be considered to be a special case of [Rockafellar, 1970, Theorem 5.7]. In a more specialized, but infinite-dimensional case the result can also be found in [Dempe, Harder, et al., 2019, Lemma 4.2].

Proposition 3.2.4. If $f : X \times V \to \mathbb{R}$ is convex and the set $g^{-1}(\Phi) \subset X \times V$ is convex, then the optimal value function φ is convex.

Proof. We generalize the proof of [Dempe, Harder, et al., 2019, Lemma 4.2]. Let $p_1, p_2 \in V$ and $\alpha, \beta \in [0,1]$ with $\alpha + \beta = 1$ be arbitrary. Then we can use the convexity of the set $g^{-1}(\Phi)$ and the function f to obtain

$$\begin{aligned}
\varphi(\alpha p_1 + \beta p_2) &= \inf\{f(x, \alpha p_1 + \beta p_2) \mid (x, \alpha p_1 + \beta p_2) \in g^{-1}(\Phi)\} \\
&= \inf\{f(\alpha x_1 + \beta x_2, \alpha p_1 + \beta p_2) \mid (\alpha x_1 + \beta x_2, \alpha p_1 + \beta p_2) \in g^{-1}(\Phi)\} \\
&\leq \inf\{\alpha f(x_1, p_1) + \beta f(x_2, p_2) \mid (\alpha x_1 + \beta x_2, \alpha p_1 + \beta p_2) \in g^{-1}(\Phi)\} \\
&\leq \inf\{\alpha f(x_1, p_1) + \beta f(x_2, p_2) \mid (x_1, p_1), (x_2, p_2) \in g^{-1}(\Phi)\} \\
&= \alpha \inf\{f(x_1, p_1) \mid g(x_1, p_1) \in \Phi\} + \beta \inf\{f(x_2, p_2) \mid g(x_2, p_2) \in \Phi\} \\
&= \alpha\varphi(p_1) + \beta\varphi(p_2).
\end{aligned}$$

Thus, we have shown that φ is convex.

Often, the function f is called jointly convex (or fully convex) instead of convex in this situation. This emphasizes that f is not just convex in x and p separately. For example, if $X = V = \mathbb{R}$, the function $f(x, p) = xp$ is convex in x and convex in p, but not convex on $\mathbb{R}^2$.

We continue with a result where the optimal value function turns out to be concave. For our result, we need to restrict ourselves to the situation where the constraint does not depend on the parameter p, i.e. we have a constraint of the form $x \in \Phi$. Note that Φ does not need to be convex. The result can be found in a finite-dimensional setting in [Fiacco, Kyparisis, 1986, Proposition 3.5].

Proposition 3.2.5. We assume that $X = Y$, $g(x, p) = x$, and that the function $f(x, \cdot) : V \to \mathbb{R}$ is concave for each feasible point $x \in X$. Then the optimal value function φ is concave.

Proof. Let $p_1, p_2 \in V$ and $\alpha \in [0,1]$ be arbitrary. Then we can use the concavity of $f(x, \cdot)$ for x with $x \in \Phi$ to obtain

$$\begin{aligned}
\alpha\varphi(p_1) + (1-\alpha)\varphi(p_2) &= \inf\{\alpha f(x, p_1) \mid x \in \Phi\} + \inf\{(1-\alpha) f(x, p_2) \mid x \in \Phi\} \\
&\leq \inf\{\alpha f(x, p_1) + (1-\alpha) f(x, p_2) \mid x \in \Phi\} \\
&\leq \inf\{f(x, \alpha p_1 + (1-\alpha) p_2) \mid x \in \Phi\} \\
&= \varphi(\alpha p_1 + (1-\alpha) p_2).
\end{aligned}$$

Thus, we have shown that φ is concave.

Note that no assumptions regarding convexity or concavity of f in the x variable are made in this proposition. In many parametrized problems the function f depends affinely on the parameter p. If this is the case, Proposition 3.2.5 can be applied.

3.2.3 Fréchet differentiability of the optimal value function

We are interested in differentiability properties of the optimal value function φ. As before, we will focus on the case where the optimization problem has a unique solution for each parameter, i.e. ψ exists. Then it turns out that the optimal value function is often Fréchet differentiable. If, however, there are parameters for which the optimization problem has multiple solutions, then it can happen that the optimal value function is not Fréchet differentiable. This is the case in Example 3.2.3 (a).

First, we look at the case where the constraint does not depend on the parameter. This means that we have a fixed feasible set and we can replace the constraint of (P(p)) by the simpler condition $x \in \Phi$ (as we did previously in Proposition 3.2.5). Note that Φ is not required to be convex. Then, under some partial Gâteaux differentiability assumptions, we can show the Fréchet differentiability of φ in the following theorem. The case of a parameter-dependent constraint is discussed later in Theorem 3.2.8.

Theorem 3.2.6. Suppose that $X = Y$ and $g(x,p) = x$ hold, that $f(x,\cdot) : V \to \mathbb{R}$ is Gâteaux differentiable for every $x \in X$, the partial Gâteaux derivative $f'_p : X \times V \to V^\star$ is continuous, and that the solution operator ψ exists and is continuous. Then φ is continuously Fréchet differentiable for every $p \in V$ with Fréchet derivative

$$\varphi'(p) = f'_p(\psi(p), p).$$

Proof. Let $p \in V$ be a parameter and let $h \in V \setminus \{0\}$ be a direction. Using the optimality of $\psi(p+h)$ in (P($p+h$)) and the mean value theorem we have

$$\begin{aligned}\varphi(p+h) - \varphi(p) &= f(\psi(p+h), p+h) - f(\psi(p), p)\\ &\le f(\psi(p), p+h) - f(\psi(p), p)\\ &= \langle f'_p(\psi(p), p+t_1 h), h\rangle_{V^\star \times V}\end{aligned}$$

for some $t_1 \in (0,1)$. It follows that

$$\varphi(p+h) - \varphi(p) - \langle f'_p(\psi(p),p), h\rangle_{V^\star\times V} \le \|f'_p(\psi(p), p+t_1 h) - f'_p(\psi(p),p)\|\|h\|$$

and because of the continuity of f'_p we obtain the upper estimate

$$\limsup_{h\to 0} \frac{1}{\|h\|}\Big(\varphi(p+h) - \varphi(p) - \langle f'_p(\psi(p),p), h\rangle_{V^\star\times V}\Big) \le 0. \tag{3.18}$$

To obtain a lower estimate we use similar arguments as above. First, we have

$$\begin{aligned}\varphi(p+h)-\varphi(p) &= f(\psi(p+h),p+h)-f(\psi(p),p)\\ &\geq f(\psi(p+h),p+h)-f(\psi(p+h),p)\\ &= \langle f_p'(\psi(p+h),p+t_2h),h\rangle_{V^\star\times V}\end{aligned}$$

for some $t_2 \in (0,1)$. Then it follows that

$$\varphi(p+h)-\varphi(p)-\langle f_p'(\psi(p),p),h\rangle_{V^\star\times V} \geq -\|f_p'(\psi(p+h),p+t_2h)-f_p'(\psi(p),p)\|\|h\|$$

and because of the continuity of f_p' we obtain

$$\liminf_{h\to 0}\frac{1}{\|h\|}\Big(\varphi(p+h)-\varphi(p)-\langle f_p'(\psi(p),p),h\rangle_{V^\star\times V}\Big)\geq 0. \tag{3.19}$$

Combining the estimates (3.18) and (3.19) yields the Fréchet differentiability at $p\in V$ with the Fréchet derivative

$$\varphi'(p) = f_p'(\psi(p),p).$$

Clearly, this Fréchet derivative is also continuous.

Let us compare this result with other results in the literature that discuss differentiability properties of the optimal value function. One such result can be found in [Bonnans, Shapiro, 2000, Theorem 4.13]. This theorem implies that (under some assumptions) φ is Fréchet differentiable if ψ exists, see [Bonnans, Shapiro, 2000, Remark 4.14]. However, a so-called inf-compactness condition is needed, which holds at a point $p_0\in V$ if there exists $\alpha\in\mathbb{R}$ and a compact set $K\subset X$ such that

$$\emptyset \neq \{x\in\Phi \mid f(x,p)\leq\alpha\}\subset K \tag{3.20}$$

holds for every p near p_0. The compact set in this condition makes it unsuitable for many infinite-dimensional optimization problems. We give an example where the inf-compactness condition is not satisfied, but Theorem 3.2.6 can still be applied.

Example 3.2.7. Let $V=X=Y$ be infinite-dimensional Hilbert spaces, $f(x,p):=\|x-p\|^2$, $g(x,p):=x$, and $\Phi:=B_1(0)$.

Then it can be seen that the map $\psi: V\to X$ exists and is continuous. However, the inf-compactness condition does not hold at the point $p_0:=0$. If p is close to p_0, then the level set in (3.20) must contain p. Hence, K must contain an open set, which is a contradiction to the compactness of K.

It can also be seen that the assumptions of Theorem 3.2.6 are satisfied and thus we know that φ is Fréchet differentiable for this example.

We remark that the function ψ in Example 3.2.7 is a projection onto a closed convex set and that φ is the squared distance function. The differentiability of the squared distance

function to a convex set in Hilbert spaces is a known result in convex optimization, see, for example, [Rockafellar, Wets, 1998, Theorem 2.26] for a proof in a finite-dimensional setting.

It should be noted that [Bonnans, Shapiro, 2000, Theorem 4.13] discusses the situation where X is a Hausdorff topological space. Particularly, it is possible to apply this result exploiting the weak topology of a Banach space. While this would remove the problems with the compact set in the inf-compactness condition, it would require that the objective function is weakly continuous, which is rarely the case.

Another result with similarities to Theorem 3.2.6 can be found in [Delfour, Zolésio, 2011, Theorem 10.2.1]. There, a directional derivative of φ is calculated, but their setting allows also for situations with more than one solution of (P(p)) for some parameters $p \in V$.

We can also give a result for the Fréchet differentiability of the optimal value function for the more general case where the constraint depends on the parameter p. In comparison to Theorem 3.2.6 this will require stronger assumptions. The proof also uses first-order optimality conditions of the problem and the Lagrange function, whereas the proof of Theorem 3.2.6 used more elementary methods. One of the more restrictive assumptions is that the Lagrange multiplier has to depend continuously on the parameter.

Theorem 3.2.8. Let $V_0 \subset V$ be open. We assume that f and g are continuously Fréchet differentiable, that Φ is closed and convex, that the solution operator $\psi : V_0 \to X$ exists on V_0 and is locally Lipschitz continuous, and that there exists a continuous map $\lambda : V_0 \to Y^\star$ which maps a parameter $p \in V_0$ to a Lagrange multiplier $\lambda(p)$ that corresponds to the solution $\psi(p)$ of (P(p)).

Then φ is continuously Fréchet differentiable for every $p \in V_0$ with Fréchet derivative

$$\varphi'(p) = \mathcal{L}'_p(\psi(p), \lambda(p), p) = f'_p(\psi(p), p) + g'_p(\psi(p), p)^\star \lambda(p).$$

Proof. Let $p, q \in V_0$ be parameters. Since $\lambda(q) \in \mathcal{N}_\Phi(g(\psi(q), q))$ and $g(\psi(p), p) \in \Phi$ we have

$$\langle \lambda(q), g(\psi(q), q) - g(\psi(p), p) \rangle_{Y^\star \times Y} \geq 0.$$

It follows that

$$\begin{aligned}
\varphi(q) - \varphi(p) &= f(\psi(q), q) - f(\psi(p), p) \\
&\leq f(\psi(q), q) - f(\psi(p), p) + \langle \lambda(q), g(\psi(q), q) - g(\psi(p), p) \rangle_{Y^\star \times Y} \\
&= \mathcal{L}(\psi(q), \lambda(q), q) - \mathcal{L}(\psi(p), \lambda(q), p) \\
&= \mathcal{L}(\psi(q), \lambda(q), q) - \mathcal{L}(\psi(p), \lambda(q), q) + \mathcal{L}(\psi(p), \lambda(q), q) - \mathcal{L}(\psi(p), \lambda(q), p).
\end{aligned}$$

Now we can apply the mean value theorem twice which yields

$$\varphi(q) - \varphi(p) \leq \mathcal{L}'_x(\hat{x}_{q,p}, \lambda(q), q)(\psi(q) - \psi(p)) + \mathcal{L}'_p(\psi(p), \lambda(q), \hat{p}_{q,p})(q - p) \tag{3.21}$$

with intermediate points $\hat{x}_{q,p} = t_{q,p}\psi(p) + (1 - t_{q,p})\psi(q)$, $\hat{p}_{q,p} = s_{q,p}p + (1 - s_{q,p})q$ where $t_{q,p}, s_{q,p} \in (0, 1)$. By exchanging p and q in the above arguments we also obtain

$$\varphi(q) - \varphi(p) \geq \mathcal{L}'_x(\hat{x}_{p,q}, \lambda(p), p)(\psi(q) - \psi(p)) + \mathcal{L}'_p(\psi(q), \lambda(p), \hat{p}_{p,q})(q - p) \tag{3.22}$$

with intermediate points $\hat{x}_{p,q} = t_{p,q}\psi(p) + (1 - t_{p,q})\psi(q)$, $\hat{p}_{p,q} = s_{p,q}p + (1 - s_{p,q})q$ where $t_{p,q}, s_{p,q} \in (0, 1)$.

Recall that due to the first-order necessary optimality conditions of (P(p)) we have $\mathcal{L}'_x(\psi(p), \lambda(p), p) = 0$. This implies the limits

$$\lim_{q \to p} \mathcal{L}'_x(\hat{x}_{q,p}, \lambda(q), q) = \lim_{q \to p} \mathcal{L}'_x(\hat{x}_{p,q}, \lambda(p), p) = 0.$$

Due to the local Lipschitz continuity of ψ we further obtain the convergences

$$\lim_{q \to p} \frac{1}{\|q - p\|} \mathcal{L}'_x(\hat{x}_{q,p}, \lambda(q), q)(\psi(q) - \psi(p)) = 0$$

and

$$\lim_{q \to p} \frac{1}{\|q - p\|} \mathcal{L}'_x(\hat{x}_{p,q}, \lambda(p), p)(\psi(q) - \psi(p)) = 0.$$

Since both $\mathcal{L}'_p(\psi(q), \lambda(p), \hat{p}_{p,q})$ and $\mathcal{L}'_p(\psi(p), \lambda(q), \hat{p}_{q,p})$ converge to $\mathcal{L}'_p(\psi(p), \lambda(p), p)$ as $q \to p$, we obtain the estimate

$$\limsup_{q \to p} \frac{1}{\|q - p\|} \Big(\varphi(q) - \varphi(p) - \mathcal{L}'_p(\psi(p), \lambda(p), p)(q - p)\Big) \leq 0$$

from (3.21) and

$$\liminf_{q \to p} \frac{1}{\|q - p\|} \Big(\varphi(q) - \varphi(p) - \mathcal{L}'_p(\psi(p), \lambda(p), p)(q - p)\Big) \geq 0$$

from (3.22). Combining the last two inequalities yields the claimed Fréchet differentiability of φ with Fréchet derivative $\mathcal{L}'_p(\psi(p), \lambda(p), p)$, which is equal to $f'_p(\psi(p), p) + g'_p(\psi(p), p)^\star \lambda(p)$. It also follows from the assumptions that this Fréchet derivative is continuous.

3.3 Bilevel optimization problems in an abstract setting

3.3.1 Notation and setting

The notation in this section will be based on the notation in Section 3.1.

Again, let X, Y, V be abstract Banach spaces. For each $p \in V$ we consider the optimization problem

$$\begin{aligned} \min_x \quad & f(x,p) \\ \text{s.t.} \quad & g(x,p) \in \Phi, \end{aligned} \tag{LL(p)}$$

which will be called the *lower level optimization problem*. Here, $f : X \times V \to \mathbb{R}$ is the lower level objective function and the function $g : X \times V \to Y$ together with the closed convex set $\Phi \subset Y$ constitute the constraint. Based on this lower level optimization problem, we consider the *upper level optimization problem*

$$\begin{aligned} \min_{x,p} \quad & F(x,p) \\ \text{s.t.} \quad & x \text{ solves (LL}(p)\text{)}, \\ & p \in \Phi_{UL}, \end{aligned} \tag{UL}$$

where $F : X \times V \to \mathbb{R}$ is the upper level objective function and $\Phi_{UL} \subset V$ is a closed set. This upper level optimization problem is also called a *bilevel optimization problem*.

For the lower level optimization problem (LL(p)) we will use the mappings Ψ, ψ, φ as defined in Definitions 3.1.1 and 3.1.2.

One possible interpretation for bilevel optimization problems involves two players, called leader and follower. First, the leader chooses a point $p \in \Phi_{UL}$. Then the follower chooses a point $x \in X$ such that $g(x,p) \in \Phi$ with the goal of minimizing his objective function f. The leader wants to minimize his objective function F, and since this function can depend on the choice x of the follower, the leader should take the actions of the follower into account.

However, if $\Psi(p)$ contains more than one element, it is not clear in this interpretation which point $x \in \Psi(p)$ the leader should use in his calculations. In principle, there are two possible approaches, called the *optimistic* and *pessimistic* formulation. The optimistic scenario of the bilevel optimization problem can be formulated via

$$\begin{aligned} \min_p \quad & \min_x \ F(x,p) \\ \text{s.t.} \quad & x \text{ solves (LL}(p)\text{)}, \\ & p \in \Phi_{UL}. \end{aligned}$$

Here, the leader can expect the "best" x among the preferred choices $x \in \Psi(p)$ of the follower. This could be the case if the leader and follower are cooperative. On the other side, if the leader does not know the attitude of the follower, it could make sense to consider a worst-case scenario, i.e. the leader would be forced to choose the "worst" $x \in \Psi(p)$ in the upper level optimization problem. This would be called the pessimistic formulation of the bilevel optimization problem and can be stated via

$$\begin{aligned} \min_p \quad & \max_x \ F(x,p) \\ \text{s.t.} \quad & x \text{ solves (LL}(p)\text{)}, \\ & p \in \Phi_{UL}. \end{aligned}$$

We mention that the pessimistic formulation of the bilevel optimization problem is very difficult to handle (and formally even a trilevel optimization problem).

For the optimistic formulation of the bilevel optimization problem we remark that it is equivalent to our original formulation (UL) when it comes to global minimizers. However, these problems are not equivalent when it comes to local minimizers, see [Dempe, Mordukhovich, Zemkoho, 2012, Example 6.10].

Note that the pessimistic and optimistic bilevel optimization problem coincide with (UL) if $\Psi(p)$ is a singleton or empty for every $p \in V$. In this thesis we mostly consider only bilevel optimization problems where $\Psi(p)$ is a singleton for all $p \in V$, i.e. ψ exists.

3.3.2 Reformulations

There are a number of reformulations of the bilevel optimization problem (UL). A relatively simple reformulation can be achieved by using the solution operator ψ. If the lower level optimization problem has a unique solution for every parameter $p \in V$, i.e. ψ exists, then we can equivalently reformulate (UL) as

$$\begin{aligned} \min_{p} \quad & F(\psi(p), p) \\ \text{s.t.} \quad & p \in \Phi_{UL}. \end{aligned}$$

A drawback of this reformulation is that $\psi(p)$ is usually not explicitly given and is also often a nonsmooth operator.

Another reformulation utilizes the optimal value function φ. Suppose that the optimal value function φ has values in $\mathbb{R}$. Then we can introduce the *optimal value reformulation*

$$\begin{aligned} \min_{x,p} \quad & F(x,p) \\ \text{s.t.} \quad & f(x,p) - \varphi(p) \leq 0, \\ & g(x,p) \in \Phi, \\ & p \in \Phi_{UL}, \end{aligned} \tag{OVR}$$

of (UL). This reformulation is an equivalent reformulation due to Lemma 3.2.1. It has the advantage that it has smooth data if f, F, g, and the optimal value function φ are smooth. The reformulation (OVR) will be the basis of Section 3.4, where we will discuss applications and drawbacks of this reformulation and consider a relaxation of (OVR) that is obtained by relaxing the constraint $f(x,p) - \varphi(p) \leq 0$.

For some $p \in V$ suppose that (LL(p)) is a convex optimization problem (i.e. the objective function $f(\cdot, p)$ and the feasible set $g(\cdot, p)^{-1}(\Phi)$ are convex), that $f(\cdot, p), g(\cdot, p)$ are continuously Fréchet differentiable, and that RZKCQ is satisfied for (LL(p)). Then the (global) solutions of (LL(p)) can be characterized by their KKT conditions, see

Theorem 2.3.5 and Proposition 2.3.8. This gives rise to the so-called *MPCC reformulation* of (UL).

$$\begin{aligned} \min_{x,p,\lambda} \quad & F(x,p) \\ \text{s.t.} \quad & f'_x(x,p) + g'_x(x,p)^\star \lambda = 0, \\ & (g(x,p),\lambda) \in \operatorname{gph} \mathcal{N}_\Phi, \\ & p \in \Phi_{UL}. \end{aligned} \tag{MPCCR}$$

We call this the MPCC reformulation because the constraint $(g(x,p),\lambda) \in \operatorname{gph} \mathcal{N}_\Phi$ can be understood as an infinite-dimensional complementarity condition. Recall that $\mathcal{N}_\Phi$ can be interpreted as a set-valued mapping and that we have

$$\operatorname{gph} \mathcal{N}_\Phi = \{(\hat{y}, \hat{\lambda}) \in Y \times Y^\star \mid \hat{y} \in \Phi, \hat{\lambda} \in \mathcal{N}_\Phi(\hat{y})\}$$

according to our definition of "gph". A disadvantage of this MPCC reformulation is that it requires higher regularity for f and g because their partial derivatives appear in the constraints of (MPCCR).

The MPCC reformulation is equivalent to (UL) in the sense that a global minimizer of one optimization problem can be transformed to a global minimizer of the other optimization problem (by adding or removing the Lagrange multiplier λ). However, these optimization problems are not always equivalent when it comes to local minimizers, see [Dempe, Dutta, 2010, Example 3.1]. Informally speaking, this is due to the additional topology introduced for the Lagrange multiplier $\lambda \in Y^\star$.

3.3.3 Formal derivation of stationarity conditions

We give a formal derivation for optimality conditions for (UL) based on the MPCC reformulation (MPCCR). Here we assume that Φ_{UL} is convex and that X is reflexive. Then we define the Lagrange function $\mathcal{L} : X \times V \times Y^\star \times X \times Y^\star \times Y^{\star\star} \times V^\star \to \mathbb{R}$ for (MPCCR) via

$$\begin{aligned} \mathcal{L}(x,p,\lambda,w,\xi,\hat{w},\rho) := F(x,p) &+ \langle f'_x(x,p) + g'_x(x,p)^\star \lambda, w \rangle_{X^\star \times X} \\ &+ \langle \xi, g(x,p) \rangle_{Y^\star \times Y} + \langle \hat{w}, \lambda \rangle_{Y^{\star\star} \times Y^\star} + \langle \rho, p \rangle_{V^\star \times V}. \end{aligned}$$

The optimality conditions are then derived by setting the partial derivatives of $\mathcal{L}$ to zero. For the partial derivative with respect to the variable $\lambda \in Y^\star$ this yields

$$0 = \mathcal{L}'_\lambda(x,p,\lambda,w,\xi,\hat{w},\rho) = g'_x(x,p)w + \hat{w},$$

which allows us to replace the variable $\hat{w} \in Y^{\star\star}$ with $-g'_x(x,p)w \in Y \subset Y^{\star\star}$. By setting the partial derivatives of $\mathcal{L}'_x$, $\mathcal{L}'_p$, and $\mathcal{L}'_w$ to zero we arrive at the conditions

$$F'_x(x,p) + f''_{xx}(x,p)w + \big(g''_{xx}(x,p)w\big)^\star \lambda + g'_x(x,p)^\star \xi = 0, \tag{3.23a}$$

$$F'_p(x,p) + f''_{px}(x,p)w + \big(g''_{px}(x,p)w\big)^\star \lambda + g'_p(x,p)^\star \xi + \rho = 0, \tag{3.23b}$$

$$f'_x(x,p) + g'_x(x,p)^\star \lambda = 0. \tag{3.23c}$$

We also have the conditions

$$(g(x,p),\lambda) \in \operatorname{gph} \mathcal{N}_\Phi, \tag{3.23d}$$
$$(p,\rho) \in \operatorname{gph} \mathcal{N}_{\Phi_{UL}} \tag{3.23e}$$

that correspond to the constraints $g(x,p) \in \Phi$ and $p \in \Phi_{UL}$. Similarly, the constraint $(g(x,p),\lambda) \in \operatorname{gph} \mathcal{N}_\Phi$ gives rise to the condition

$$\big((g(x,p),\lambda),(\xi,-g'_x(x,p)w)\big) \in \operatorname{gph} \mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_\Phi}, \tag{3.23f}$$

where $\mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_\Phi} : Y \times Y^\star \to \mathcal{P}(Y^\star \times Y^{\star\star})$ is an unspecified set-valued mapping that can be interpreted as a generalization of a normal cone to the nonconvex set $\operatorname{gph}\mathcal{N}_\Phi$. For example, one possible choice would be the limiting normal cone from Definition 2.4.1 (c), i.e. $\mathcal{N}^\sharp = \mathcal{N}^{\lim}$. We will specify $\mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_\Phi}$ in later sections when we consider specific examples of bilevel optimization problems.

3.4 Relaxation using the optimal value function

We will use the setting from Section 3.3 in this section. Recall that the set Φ is closed and convex. This section will be based on the optimal value reformulation (OVR). If f, g, F and the optimal value function φ are continuously Fréchet differentiability (see Theorem 3.2.6 or Theorem 3.2.8) then the reformulation (OVR) has the advantage over (UL) in the sense that it can be formulated using continuously Fréchet differentiable data. However, there are still difficulties with this reformulation. The next lemma shows that under reasonable assumptions RZKCQ is violated at every feasible point. Note that this is a generalization of [Dempe, Harder, et al., 2019, Lemma 5.1], which discusses the same question for a more specific problem. We will also use a similar technique in our proof. The result can be found in [Mehlitz, 2017, Lemma 4.32], too.

Lemma 3.4.1. Suppose that the functions f, g, φ are continuously Fréchet differentiable and that Φ_{UL} is convex. Then RZKCQ for the problem (OVR) is violated at every feasible point $(\bar{x},\bar{p})$.

Proof. Let $(\bar{x},\bar{p})$ be a feasible point of (OVR). Then $f(\bar{x},\bar{p}) - \varphi(\bar{p}) = 0$ holds due to Lemma 3.2.1. This implies $\mathcal{R}_{(-\infty,0]}(f(\bar{x},\bar{p}) - \varphi(\bar{p})) = (-\infty, 0]$. Suppose that RZKCQ holds for (OVR), i.e. we have

$$M := \begin{bmatrix} f'(\bar{x},\bar{p}) - \begin{bmatrix} 0 & \varphi'(\bar{p}) \end{bmatrix} \\ g'(\bar{x},\bar{p}) \end{bmatrix} \begin{pmatrix} X \\ \mathcal{R}_{\Phi_{UL}}(\bar{p}) \end{pmatrix} - \begin{pmatrix} (-\infty,0] \\ \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \end{pmatrix} = \mathbb{R} \times Y. \tag{3.24}$$

We consider the optimization problem

$$\begin{aligned} \min_{x,p} \quad & f(x,p) - \varphi(p) \\ \text{s.t.} \quad & g(x,p) \in \Phi. \end{aligned} \tag{3.25}$$

Due to the assumptions this optimization problem has continuously Fréchet differentiable data and convex sets Φ, Φ_{UL}. Note that we can conclude from (3.24) that

$$Y \subset g'(\bar{x},\bar{p}) \begin{pmatrix} X \\ \mathcal{R}_{\Phi_{UL}}(\bar{p}) \end{pmatrix} - \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \subset g'(\bar{x},\bar{p}) \begin{pmatrix} X \\ V \end{pmatrix} - \mathcal{R}_\Phi(g(\bar{x},\bar{p}))$$

is true, i.e. RZKCQ holds at $(\bar{x},\bar{p})$ for the problem (3.25). Recall that due to Lemma 3.2.1 we know that $f(x,p) - \varphi(p) \geq 0 = f(\bar{x},\bar{p}) - \varphi(\bar{p})$ holds for all feasible points (x,p) of (3.25). Thus, we know that $(\bar{x},\bar{p})$ is a global minimizer of (3.25). Since RZKCQ is also satisfied, there exists a Lagrange multiplier $\lambda \in \mathcal{N}_\Phi(g(\bar{x},\bar{p}))$ such that the KKT conditions

$$f'(\bar{x},\bar{p}) - \begin{bmatrix} 0 & \varphi'(\bar{p}) \end{bmatrix} + g'(\bar{x},\bar{p})^\star \bar{\lambda} = 0, \tag{3.26a}$$

$$\bar{\lambda} \in \mathcal{N}_\Phi(g(\bar{x},\bar{p})), \qquad g(\bar{x},\bar{p}) \in \Phi \tag{3.26b}$$

hold. Now we apply the functional $\begin{bmatrix} 1 & \bar{\lambda} \end{bmatrix} \in \mathbb{R} \times Y^\star$ to the set equality (3.24). This yields

$$\begin{aligned} \mathbb{R} &= \begin{bmatrix} 1 & \bar{\lambda} \end{bmatrix} (\mathbb{R} \times Y) = \begin{bmatrix} 1 & \bar{\lambda} \end{bmatrix} M \\ &= \left(f'(\bar{x},\bar{p}) - \begin{bmatrix} 0 & \varphi'(\bar{p}) \end{bmatrix} + g'(\bar{x},\bar{p})^\star \bar{\lambda} \right) \begin{pmatrix} X \\ \mathcal{R}_{\Phi_{UL}}(\bar{p}) \end{pmatrix} - \begin{bmatrix} 1 & \bar{\lambda} \end{bmatrix} \begin{pmatrix} (-\infty, 0] \\ \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \end{pmatrix} \\ &= [0,\infty) - \langle \bar{\lambda}, \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \rangle_{Y^\star \times Y}, \end{aligned}$$

where we used (3.26a) to simplify the expression. If we also consider that $\langle \bar{\lambda}, y \rangle_{Y^\star \times Y} \leq 0$ holds for all $y \in \mathcal{R}_\Phi(g(\bar{x},\bar{p}))$ due to (3.26b), we obtain

$$\mathbb{R} = [0,\infty) - \langle \bar{\lambda}, \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \rangle_{Y^\star \times Y} = [0,\infty)$$

which is false. Thus, our initial assumption that RZKCQ holds at $(\bar{x},\bar{p})$ for the problem (OVR) must be wrong, which proves our claim.

Not only are constraint qualifications violated for (OVR), it can also happen that the KKT conditions are not satisfied for a local minimizer, see [Dempe, Harder, et al., 2019, Example 5.1]. Therefore, it is not promising to attempt to prove weaker constraint qualifications for (OVR). Informally, an issue with (OVR) is that all points (x,p) with $g(x,p) \in \Phi$ satisfy $f(x,p) - \varphi(p) \geq 0$, so that the inequality constraint $f(x,p) - \varphi(p) \leq 0$ cannot be satisfied strictly. One also cannot avoid this issue by using the equality constraint $f(x,p) - \varphi(p) = 0$ in (OVR) instead of the inequality. Therefore, we will relax

this optimal value constraint.

For $\varepsilon > 0$, we consider the optimization problem

$$\begin{aligned} \min_{x,p} \quad & F(x,p) \\ \text{s.t.} \quad & f(x,p) - \varphi(p) - \varepsilon \leq 0, \\ & g(x,p) \in \Phi, \\ & p \in \Phi_{UL}, \end{aligned} \tag{OVR(ε))}$$

which is a relaxation of (OVR). Obviously, this problem is the same as (OVR) if $\varepsilon = 0$.

The goal of this section is to derive necessary optimality conditions for local minimizers of (UL). The strategy for obtaining optimality conditions for local minimizers of (OVR) is to study the limiting behavior of a system of stationarity conditions for local minimizers of (OVR(ε)) as $\varepsilon \to 0$. For specific bilevel optimization problems this idea was already used in [Harder, 2016, Section 5.3], [Dempe, Harder, et al., 2019, Section 5.2], and [Mehlitz, G. Wachsmuth, 2019, Section 4.2]. Many results in this section are abstract generalizations of results from these articles. Due to the abstract problem formulation many proofs in this section will be quite technical. While we consider the abstract bilevel optimization problem (UL) in this section, we will apply this to more specific bilevel optimization problems later and also compare it with existing stationarity conditions.

3.4.1 Satisfaction of RZKCQ for the ε-relaxation

Our next goal is to show that RZKCQ is satisfied at every feasible point for the relaxation (OVR(ε)). First, we need a technical lemma.

Lemma 3.4.2. Let $\bar{p} \in V$ be given and let $\bar{x}$ be a feasible point of (LL($\bar{p}$)). Moreover, we assume that $g(\cdot, \bar{p}) : X \to Y$ is Gâteaux differentiable at $\bar{x}$ and that RZKCQ holds at $\bar{x}$ for the problem (LL($\bar{p}$)). Then

$$\Big(g'_x(\bar{x},\bar{p})^{-1}(\mathcal{R}_\Phi(g(\bar{x},\bar{p})))\Big)^\circ = g'_x(\bar{x},\bar{p})^\star \mathcal{N}_\Phi(g(\bar{x},\bar{p}))$$

holds.

Proof. We use the abbreviations $T := g'_x(\bar{x},\bar{p})$ and $y_1 := g(\bar{x},\bar{p})$. Our first goal will be to show that

$$\operatorname{cl}(T^{-1}(\mathcal{R}_\Phi(y_1))) = T^{-1}(\mathcal{T}_\Phi(y_1)) \tag{3.27}$$

holds. Since the right-hand side of (3.27) is closed and $T^{-1}(\mathcal{R}_\Phi(y_1)) \subset T^{-1}(\mathcal{T}_\Phi(y_1))$ holds, the inclusion "$\subset$" in (3.27) follows.

To show the other inclusion, let $x_1 \in T^{-1}(\mathcal{T}_\Phi(y_1))$ be given. Moreover, let $\varepsilon > 0$ be given. Since $\mathcal{R}_\Phi(y_1)$ is dense in $\mathcal{T}_\Phi(y_1)$ there exists a point $y_2 \in Y$ with $\|y_2\| < \varepsilon$ such that $Tx_1 + y_2 \in \mathcal{R}_\Phi(y_1)$. Next, we apply Theorem 2.3.7. Then RZKCQ implies the existence

of $\alpha > 0$ (independent of ε) such that

$$B_\alpha(0) \subset T(B_1(0)) - (\Phi - y_1) \cap B_1(0)$$

holds. After scaling this inclusion by ε/α we know that there exist $x_2 \in X, y_3 \in \mathcal{R}_\Phi(y_1)$ such that $y_2 = Tx_2 - y_3$ and $\|x_2\|, \|y_3\| < \varepsilon/\alpha$. It follows that

$$T(x_1 + x_2) = Tx_1 + y_2 + y_3 \in \mathcal{R}_\Phi(y_1)$$

and therefore $x_1 + x_2 \in T^{-1}(\mathcal{R}_\Phi(y_1))$. Since $\varepsilon > 0$ was arbitrary and $\|x_2\| \leq \varepsilon/\alpha$ it follows that x_1 is in the closure of $T^{-1}(\mathcal{R}_\Phi(y_1))$.

Now that we have shown (3.27) we continue with calculating the polar cone of $T^{-1}(\mathcal{T}_\Phi(y_1))$. For that purpose we intent to apply Lemma 2.1.18. The assumption

$$T(X) - \mathcal{T}_\Phi(y_1) = Y$$

for this lemma is satisfied because RZKCQ is valid at the point $\bar{x}$. Hence, we obtain

$$(T^{-1}(\mathcal{T}_\Phi(y_1)))^\circ = T^\star(\mathcal{T}_\Phi(y_1)^\circ) = T^\star(\mathcal{N}_\Phi(y_1)).$$

Finally, by (3.27) it follows that

$$(T^{-1}(\mathcal{R}_\Phi(y_1)))^\circ = (T^{-1}(\mathcal{T}_\Phi(y_1)))^\circ = T^\star(\mathcal{N}_\Phi(y_1))$$

which completes the proof.

Now we are ready to state our result that the constraint qualification RZKCQ is satisfied for (OVR(ε)) under some assumptions.

Theorem 3.4.3. Let $\varepsilon > 0$ be given and let $(\bar{x}, \bar{p})$ be a feasible point of (OVR(ε)). Suppose that

(a) the functions f, g are Gâteaux differentiable at $(\bar{x}, \bar{p})$,

(b) the function φ is Gâteaux differentiable at $\bar{p}$,

(c) the function $f(\cdot, \bar{p})$ and the feasible set $g(\cdot, \bar{p})^{-1}(\Phi)$ are convex,

(d) Φ_{UL} is convex,

(e) RZKCQ is satisfied at the point $\bar{x}$ for (LL($\bar{p}$)).

Then RZKCQ is satisfied at $(\bar{x}, \bar{p})$ for (OVR(ε)).

Proof. First we consider the case where $f(\bar{x}, \bar{p}) - \varphi(\bar{p}) - \varepsilon < 0$. In this case we have

$\mathcal{R}_{(-\infty,0]}(f(\bar{x},\bar{p})-\varphi(\bar{p})-\varepsilon)=\mathbb{R}$ and RZKCQ for (OVR(ε)) takes the form

$$\begin{bmatrix} f'_x(\bar{x},\bar{p}) & f'_p(\bar{x},\bar{p})-\varphi'(\bar{p}) \\ g'_x(\bar{x},\bar{p}) & g'_p(\bar{x},\bar{p}) \end{bmatrix} \begin{pmatrix} X \\ \mathcal{R}_{\Phi_{UL}}(\bar{p}) \end{pmatrix} - \begin{pmatrix} \mathbb{R} \\ \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \end{pmatrix} = \begin{pmatrix} \mathbb{R} \\ Y \end{pmatrix}.$$

This follows directly from RZKCQ of the lower level which states that

$$g'_x(\bar{x},\bar{p})X - \mathcal{R}_\Phi(g(\bar{x},\bar{p})) = Y$$

holds.

Next, we check the case where $f(\bar{x},\bar{p})-\varphi(\bar{p})-\varepsilon=0$. In this case we have $\mathcal{R}_{(-\infty,0]}(f(\bar{x},\bar{p})-\varphi(\bar{p})-\varepsilon)=(-\infty,0]$. Because of $\varepsilon>0$ we have $f(\bar{x},\bar{p})>\varphi(\bar{p})$, which implies that $\bar{x}$ is not a global minimizer of (LL($\bar{p}$)). Since (LL($\bar{p}$)) is a convex optimization problem it follows from Proposition 2.3.8 that the first-order necessary optimality conditions for (LL($\bar{p}$)) are not satisfied at $\bar{x}$, i.e. $f'_x(\bar{x},\bar{p})+g'_x(\bar{x},\bar{p})^\star\lambda\neq 0$ holds for all $\lambda\in\mathcal{N}_\Phi(g(\bar{x},\bar{p}))$. Thus, by Lemma 3.4.2, we have

$$-f'_x(\bar{x},\bar{p}) \notin g'_x(\bar{x},\bar{p})^\star\mathcal{N}_\Phi(g(\bar{x},\bar{p})) = \Big(g'_x(\bar{x},\bar{p})^{-1}(\mathcal{R}_\Phi(g(\bar{x},\bar{p})))\Big)^\circ.$$

Therefore, there exists $x_1\in X$ such that

$$\langle -f'_x(\bar{x},\bar{p}),x_1\rangle_{X^\star\times X}=1 \quad\text{and}\quad g'_x(\bar{x},\bar{p})x_1\in\mathcal{R}_\Phi(g(\bar{x},\bar{p})).$$

Now, let $r_1\in\mathbb{R}, y_1\in Y$ be given. Using RZKCQ at the point $\bar{x}$ for (LL($\bar{p}$)) yields the existence of $x_2\in X, y_2\in\mathcal{R}_\Phi(g(\bar{x},\bar{p}))$ such that $g'_x(\bar{x},\bar{p})x_2-y_2=y_1$. We define $\alpha:=\max(0,f'_x(\bar{x},\bar{p})x_2-r_1)\in[0,\infty)$ and $r_2:=\min(0,f'_x(\bar{x},\bar{p})x_2-r_1)\in(-\infty,0]$. Because of $f'_x(\bar{x},\bar{p})x_1=-1$ it follows that $f'_x(\bar{x},\bar{p})(x_2+\alpha x_1)-r_2=r_1$. Thus, we obtain

$$\begin{pmatrix} r_1 \\ y_1 \end{pmatrix} = \begin{bmatrix} f'_x(\bar{x},\bar{p}) & f'_p(\bar{x},\bar{p})-\varphi'(\bar{p}) \\ g'_x(\bar{x},\bar{p}) & g'_p(\bar{x},\bar{p}) \end{bmatrix} \begin{pmatrix} x_2+\alpha x_1 \\ 0 \end{pmatrix} - \begin{pmatrix} r_2 \\ y_2+\alpha g'_x(\bar{x},\bar{p})x_1 \end{pmatrix}.$$

Because (r_1,y_1) was arbitrary in $\mathbb{R}\times Y$ and $y_2+\alpha g'_x(\bar{x},\bar{p})x_1\in\mathcal{R}_\Phi(g(\bar{x},\bar{p}))$ this implies

$$\begin{pmatrix} \mathbb{R} \\ Y \end{pmatrix} \subset \begin{bmatrix} f'_x(\bar{x},\bar{p}) & f'_p(\bar{x},\bar{p})-\varphi'(\bar{p}) \\ g'_x(\bar{x},\bar{p}) & g'_p(\bar{x},\bar{p}) \end{bmatrix} \begin{pmatrix} X \\ \mathcal{R}_{\Phi_{UL}}(\bar{p}) \end{pmatrix} - \begin{pmatrix} (-\infty,0] \\ \mathcal{R}_\Phi(g(\bar{x},\bar{p})) \end{pmatrix}$$

which is RZKCQ for (OVR(ε)).

3.4.2 Existence of minimizers for the ε-relaxation

To show that the existence of minimizers for (OVR(ε)) is not always a simple issue, we provide the following example. This is an instance of a bilevel optimization problem with convex data in the lower and upper level. Moreover, the objective function of the lower level optimization problem is strongly convex. The nonconvex bilevel optimization

problem has a solution, but the relaxation (OVR(ε)) does not. This is due to the nonconvexity of the feasible set and the infinite dimensions in both the upper and the lower level.

Example 3.4.4. We consider the setting $X = V = Y := L^2(\Omega)$ with $\Omega := (0,1) \subset \mathbb{R}^1$. The data for the lower level optimization problem is given as $f(x,p) := \|x-p\|^2_{L^2(\Omega)}$, $g(x,p) := x$, and $\Phi := L^2(\Omega)_+$. Then the lower level optimization problem has a unique solution for each parameter $p \in L^2(\Omega)$, which is given by $\psi(p) = p^+$.

We continue with the setting of the upper level optimization problem. Let $v_{0,0}, v_{i,j} \in L^2(\Omega)$ for $i \in \mathbb{N}, j \in \{0, \dots, 2^{i-1}-1\}$ be the Haar system as defined in Definition 2.2.2. Then we can define the auxiliary function $q : L^2(\Omega) \to \mathbb{R}$ via

$$q(x) := (x, v_{0,0})^2_{L^2(\Omega)} + \sum_{i \in \mathbb{N}} \sum_{j=0}^{2^{i-1}-1} 4^{-i} (x, v_{i,j})^2_{L^2(\Omega)}.$$

Finally, we choose $F(x,p) := q(x-1/2) + q(p+1/2)$ and $\Phi_{UL} := \{w \in L^2(\Omega) \mid -2 \leq w \leq 0 \text{ a.e. in } \Omega\}$ for the data of the upper level optimization problem.

Then the bilevel optimization problem (UL) has a unique solution. However, for $\varepsilon = 1/2$, the relaxation (OVR(ε)) has no solution.

Proof. It can be seen that the lower level optimization problem has a unique solution for each parameter $p \in L^2(\Omega)$ and that the solution operator $\psi : V \to X$ is defined via $\psi(p) := p^+$. It follows that the optimal value function is given by $\varphi(p) := \|p^-\|^2_{L^2(\Omega)}$.

First, we address the existence of a solution of the overall bilevel optimization problem (UL). We claim that $(\bar{x}, \bar{p}) := (0, -1/2)$ is the unique solution. For a feasible point (x,p) of (UL) we have $p \leq 0$ a.e. in Ω and therefore $x = p^+ = 0$ holds. Furthermore, note that q is strictly convex because the Haar system is an orthonormal basis, see Lemma 2.2.3. Thus, F is minimized if and only if $p = -1/2$, i.e. $(\bar{x}, \bar{p}) := (0, 1/2)$ is the unique solution of (UL).

Next, we want to show that for $\varepsilon = 1/2$ the relaxation (OVR(ε)) has no solution. First, we observe that $F(x,p) \geq 0$ for all $x, p \in L^2(\Omega)$. Because the Haar system is an orthonormal basis according to Lemma 2.2.3 we have $q(x) = 0$ if and only if $x = 0$. It follows that $F(x,p) = 0$ if and only if $x = 1/2$ and $p = -1/2$ a.e. in Ω. Since $f(1/2, -1/2) - \varphi(-1/2) - 1/2 = 1/4 > 0$, the point $(1/2, -1/2) \in X \times V$ is not a feasible point of (OVR(1/2)). Thus, $F(x,p) > 0$ holds for all feasible points (x,p) of (OVR(1/2)). Therefore, in order to demonstrate that (OVR(1/2)) has no solution it suffices to show that there exists a sequence of feasible points $\{(x_i, p_i)\}_{i \in \mathbb{N}}$ such that $F(x_i, p_i) \to 0$ as $i \to \infty$.

We define these points via

$$x_i := \tfrac{1}{2} + \sum_{j=0}^{2^{i-1}-1} 2^{-(i+1)/2} v_{i,j}, \qquad p_i := x_i - 1$$

for $i \in \mathbb{N}$. Let $i \in \mathbb{N}$ be given. We mention that we can also express (x_i, p_i) via

$$x_i(\omega) = \tfrac{1}{2}(1 + \operatorname{sgn}(\sin(2^i \pi \omega))), \qquad p_i(\omega) = \tfrac{1}{2}(-1 + \operatorname{sgn}(\sin(2^i \pi \omega))) \qquad \forall \omega \in (0,1).$$

We note that $x_i(\omega)$ has values in $\{0,1\}$ for almost all $\omega \in \Omega$ whereas $p_i(\omega)$ has values in $\{0,-1\}$ for almost all $\omega \in \Omega$. Therefore, we have $x_i \in \Phi$ as well as $p_i^- = p_i$ and $p_i \in \Phi_{UL}$. Since $\operatorname{meas}(\{p_i = -1\}) = 1/2$ it follows that $\|p_i\|^2_{L^2(\Omega)} = 1/2$. From the calculation

$$f(x_i, p_i) - \varphi(p_i) - \varepsilon = \|x_i - p_i\|^2_{L^2(\Omega)} - \|p_i^-\|^2_{L^2(\Omega)} - \tfrac{1}{2} = \|1\|^2_{L^2(\Omega)} - \tfrac{1}{2} - \tfrac{1}{2} = 0$$

we conclude that (x_i, p_i) is a feasible point of (OVR(1/2)).

It remains to show that $F(x_i, p_i) \to 0$ as $i \to \infty$. Let $i \in \mathbb{N}$ be given. We recall that the Haar system is an orthonormal system, see Lemma 2.2.3. We can use this fact to calculate

$$\begin{aligned} q(x_i - 1/2) &= q\left(\sum_{j=0}^{2^{i-1}-1} 2^{-(i+1)/2} v_{i,j} \right) \\ &= \sum_{j=0}^{2^{i-1}-1} 4^{-i} 2^{-(i+1)} (v_{i,j}, v_{i,j})^2_{L^2(\Omega)} = 4^{-i-1}. \end{aligned}$$

Due to $x_i - 1/2 = p_i + 1/2$ we obtain

$$F(x_i, p_i) = q(x_i - 1/2) + q(p_i + 1/2) = 2q(x_i - 1/2) = 2 \cdot 4^{-i-1} \to 0.$$

This completes the proof.

We mention that the sequence $\{(x_i, p_i)\}_{i \in \mathbb{N}}$ of feasible points converges weakly to the infeasible point $(1/2, -1/2) \in X \times V$. Thus, the function $(x,p) \mapsto f(x,p) - \varphi(p)$ is not sequentially weakly lower semi-continuous and this can cause the feasible set of (OVR(ε)) to be not weakly (sequentially) closed. This example shows that in order to guarantee the existence of a solution of (OVR(ε)) we need to have stronger assumptions. For example, it would be desirable that the function $(x,p) \mapsto f(x,p) - \varphi(p)$ is sequentially weakly lower semi-continuous. Let us give a list of assumptions that will be utilized for obtaining stationarity conditions of (OVR).

Assumption 3.4.5. (a) The point $(\bar{x}, \bar{p}) \in X \times V$ is a local minimizer of (OVR).

(b) The space X is a Hilbert space, V is a reflexive Banach space, and Y is a Banach

space.

(c) The function F is continuously Fréchet differentiable and sequentially weakly lower semi-continuous.

(d) The functions f and g are continuously Fréchet differentiable with locally Lipschitz continuous derivatives.

(e) The functions f'_x and g'_x are Gâteaux differentiable and their Gâteaux derivatives are continuous in the strong operator topologies of $\mathbb{L}(X \times V, X^\star)$ and $\mathbb{L}(X \times V, \mathbb{L}(X, Y))$.

(f) The functions f'_p and g'_p are partially Gâteaux differentiable with respect to x and their partial Gâteaux derivatives are continuous in the strong operator topologies of $\mathbb{L}(X, V^\star)$ and $\mathbb{L}(X, \mathbb{L}(V, Y))$.

(g) There exists a constant $\gamma > 0$ such that for every $p \in V$ the function $f(\cdot, p)$ is strongly convex with parameter γ on the feasible set $g(\cdot, p)^{-1}(\Phi)$.

(h) The sets $g(\cdot, p)^{-1}(\Phi), \Phi, \Phi_{UL}$ are closed, convex, and nonempty for all $p \in V$.

(i) The set $g(\cdot, \cdot)^{-1}(\Phi) \cap (X \times \Phi_{UL}) \subset X \times V$ is sequentially weakly closed.

(j) The function $(x, p) \mapsto f(x, p) - \varphi(p)$ restricted to a sufficiently small convex neighborhood of $(\bar{x}, \bar{p})$ is sequentially weakly lower semi-continuous.

(k) The operator $g'_x(\bar{x}, \bar{p}) : X \to Y$ is surjective.

(l) There exists an $r > 1$ such that the function $p \mapsto \|p\|_V^r$ is continuously Fréchet differentiable.

We give some remarks on these assumptions. Most of these assumptions are required to guarantee the existence of solutions of (OVR(ε)), or the existence and continuity of ψ near $\bar{p}$, see Lemma 3.4.6. Note that Assumption 3.4.5 (l) is always satisfied with $r = 2$ if V is a Hilbert space, but is also often satisfied in other spaces. For example, if $V^\star$ is separable, then there is an equivalent norm on V such that its square is continuously Fréchet differentiable, see [Börgens et al., 2019, Lemma 5.7].

The condition that the Gâteaux derivatives are only continuous in the strong operator topology (and not necessarily continuous) allows for a wider class of functions. For example, the Nemytskii operator $h : L^2((0,1)) \to L^2((0,1))$, $x \mapsto \sin(x(\cdot)) \in L^2(\Omega)$ provides a counterexample, i.e. its Gâteaux derivative $h'(x) = \cos(x) \in \mathbb{L}(L^2((0,1)), L^2((0,1)))$ is not continuous but only continuous in the strong operator topology of $\mathbb{L}(L^2((0,1)), L^2((0,1)))$. Note that the continuity in the strong operator topology is also used for the symmetry of second derivatives, see Lemmas 2.1.11 and 2.1.12. We will show in later sections that these assumptions are not too restrictive and can be satisfied in the setting of naturally occurring infinite-dimensional bilevel optimization problems.

Let us start with some results that follow from Assumption 3.4.5, including the existence of minimizers of (OVR(ε)).

Lemma 3.4.6. Let Assumption 3.4.5 be satisfied.

(a) The lower level optimization problem (LL(p)) has a unique global minimizer for every $p \in V$. The solution operator $\psi : V \to X$ exists and is Lipschitz continuous in a neighborhood of $\bar{p}$.

(b) Let $(x, p) \in X \times V$ be a feasible point of (OVR(ε)). Then

$$\|x - \psi(p)\|^2 \leq \tfrac{2\varepsilon}{\gamma}$$

holds.

(c) There exist $\delta, \varepsilon_0 > 0$ (which only depend on the data of (LL(p))) such that if $\Phi_{UL} \subset B_{\varepsilon_0}(\bar{p})$ then the function $(x, p) \mapsto f(x, p) - \varphi(p)$ is sequentially weakly lower semi-continuous on $B_\delta((\bar{x}, \bar{p}))$ and the feasible set of (OVR(ε)) is included in $B_\delta((\bar{x}, \bar{p}))$ for all $\varepsilon \in (0, \varepsilon_0)$.

(d) There exists an $\varepsilon_0 > 0$ (which only depends on the data of (LL(p))) such that if $\Phi_{UL} \subset B_{\varepsilon_0}(\bar{p})$ then (OVR(ε)) has a global minimizer for every $\varepsilon \in (0, \varepsilon_0)$.

Proof. Part (a) follows directly from Lemma 3.1.8.

For part (b), we can use the quadratic growth condition Lemma 2.3.2 (c) for the lower level optimization problem to obtain

$$\tfrac{\gamma}{2}\|x - \psi(p)\|^2 \leq f(x, p) - f(\psi(p), p) = f(x, p) - \varphi(p) \leq \varepsilon,$$

which yields the claim.

We continue with part (c). We choose $\delta > 0$ such that the function $(x, p) \mapsto f(x, p) - \varphi(p)$ is sequentially weakly lower semi-continuous on $B_\delta((\bar{x}, \bar{p}))$, see Assumption 3.4.5 (j). Let $\delta_1, C_L > 0$ be such that C_L is a Lipschitz constant of ψ on $B_{\delta_1}(\bar{p})$, see part (a). Then we choose $\varepsilon_0 > 0$ such that $(2\varepsilon_0)^{1/2}\gamma^{-1/2} + (C_L + 1)\varepsilon_0 < \delta$ and $\varepsilon_0 < \delta_1$. Now, let $\varepsilon \in (0, \varepsilon_0)$ be given and assume that $\Phi_{UL} \subset B_{\varepsilon_0}(\bar{p})$ holds. Further, let $(x, p) \in X \times V$ be a feasible point of (OVR(ε)). Then, by part (b) we have

$$\begin{aligned}\|(x, p) - (\bar{x}, \bar{p})\| &= \|x - \bar{x}\| + \|p - \bar{p}\| \leq \|x - \psi(p)\| + \|\psi(p) - \psi(\bar{p})\| + \|p - \bar{p}\| \\ &\leq (2\varepsilon)^{1/2}\gamma^{-1/2} + (C_L + 1)\|p - \bar{p}\| \leq (2\varepsilon_0)^{1/2}\gamma^{-1/2} + (C_L + 1)\varepsilon_0 < \delta.\end{aligned}$$

Thus, the feasible set of (OVR(ε)) is included in $B_\delta((\bar{x}, \bar{p}))$ for all $\varepsilon \in (0, \varepsilon_0)$, which proves the claim.

Finally, for part (d) we choose $\varepsilon_0, \delta > 0$ according to part (c). Let $\varepsilon \in (0, \varepsilon_0)$ be given. Since the feasible set of (OVR(ε)) is included in $B_\delta((\bar{x}, \bar{p}))$ it follows that the feasible set of (OVR(ε)) is bounded. Moreover, the sets $\Phi_{UL} \subset V$ and $g^{-1}(\cdot, \cdot)(\Phi) \cap (X \times \Phi_{UL}) \subset X \times V$ are sequentially weakly closed due to Assumption 3.4.5 (h) and Assumption 3.4.5 (i). Because the function $(x, p) \mapsto f(x, p) - \varphi(p)$ is sequentially weakly lower semi-continuous

on $B_\delta((\bar{x},\bar{p}))$ this implies that the feasible set of (OVR(ε)) is bounded and sequentially weakly closed. Because the objective function F is sequentially weakly lower semi-continuous due to Assumption 3.4.5 (c), the claim follows from Lemma 2.3.2 (d).

If we denote the feasible set of the problem (OVR(ε)) by A_ε and the feasible set of (OVR) by A_0, then one can obtain from Lemma 3.4.6 (b) that $A_\varepsilon \subset A_0 + C\sqrt{\varepsilon}B_1(0)$ (with a constant $C > 0$), which is a property that one would like to have for a relaxation.

In Lemma 3.4.6 (d) the condition that Φ_{UL} is included in a small neighborhood of $\bar{p}$ is acceptable, because for stationarity conditions we are only interested in the local behavior. This condition will be removed with a localization argument in the proof of Theorem 3.4.17.

3.4.3 Convergence of multipliers

Let us state the KKT systems for the lower level optimization problem and for (OVR(ε)).

Lemma 3.4.7. Let Assumption 3.4.5 be satisfied. We consider a point p_i in a sufficiently small neighborhood of $\bar{p}$ and a point x_i in a sufficiently small neighborhood of $\bar{x}$.

(a) If x_i is a feasible point of (LL(p_i)) then RZKCQ is satisfied for this problem at the point x_i. In particular, for the solution $\psi(p_i)$ of (LL(p_i)) there exists a unique multiplier $\lambda_i \in Y^\star$ such that the KKT conditions

$$f'_x(\psi(p_i),p_i) + g'_x(\psi(p_i),p_i)^\star\lambda_i = 0, \tag{3.28a}$$
$$(g(\psi(p_i),p_i),\lambda_i) \in \operatorname{gph}\mathcal{N}_\Phi \tag{3.28b}$$

hold.

(b) The optimal value function φ is continuously Fréchet differentiable in a neighborhood of $\bar{p}$ and its Fréchet derivative is given by

$$\varphi'(p_i) = \mathcal{L}'_p(\psi(p_i),\lambda_i,p_i) = f'_p(\psi(p_i),p_i) + g'_p(\psi(p_i),p_i)^\star\lambda_i.$$

(c) Let $\varepsilon_i > 0$ be given. If (x_i,p_i) is a feasible point of (OVR(ε_i)) then RZKCQ is satisfied for this problem at the point (x_i,p_i). Moreover, if (x_i,p_i) is a local minimizer of (OVR(ε_i)) there are multipliers $\beta_i \in \mathbb{R}$, $\mu_i \in Y^\star$, $\rho_i \in V^\star$ such that the KKT conditions

$$F'_x(x_i,p_i) + \beta_i f'_x(x_i,p_i) + g'_x(x_i,p_i)^\star\mu_i = 0, \tag{3.29a}$$
$$F'_p(x_i,p_i) + \beta_i(f'_p(x_i,p_i) - \varphi'(p_i)) + g'_p(x_i,p_i)^\star\mu_i + \rho_i = 0, \tag{3.29b}$$
$$(f(x_i,p_i) - \varphi(p_i) - \varepsilon_i,\beta_i) \in \operatorname{gph}\mathcal{N}_{(-\infty,0]}, \tag{3.29c}$$
$$(g(x_i,p_i),\mu_i) \in \operatorname{gph}\mathcal{N}_\Phi, \tag{3.29d}$$

$$(p_i, \rho_i) \in \operatorname{gph} \mathcal{N}_{\Phi_{UL}} \tag{3.29e}$$

hold.

Proof. Since $g'_x(\bar{x}, \bar{p})$ is surjective then $g'_x(x_i, p_i)$ is surjective if (x_i, p_i) is sufficiently close to $(\bar{x}, \bar{p})$, see Lemma 2.1.7 (b). Then the surjectivity of $g'_x(x_i, p_i)$ implies RZKCQ for (LL(p_i)) at x_i. Due to Lemma 3.4.6 (a) we can assume that $\psi(p_i)$ is sufficiently close to $\bar{x} = \psi(\bar{p})$ and thus we can apply the above observation for the case that $x_i = \psi(p_i)$. Therefore, by Corollary 2.3.6 there exists a unique multiplier $\lambda_i \in Y^\star$ such that (3.28) holds.

For part (b) we intend to use Theorem 3.2.8. The assumption that ψ is locally Lipschitz continuous near $\bar{p}$ is true due to Lemma 3.4.6 (a). By part (a) we know that there exists a function λ that is defined in a neighborhood of $\bar{p}$ which maps p to a Lagrange multiplier that corresponds to the solution $\psi(p)$ of (LL(p)). By Lemma 3.1.7 we obtain that λ is continuous in a neighborhood of $\bar{p}$. Thus, the assumptions for Theorem 3.2.8 are satisfied and part (b) follows.

Part (c) follows from Theorems 2.3.5 and 3.4.3, whose assumptions are satisfied due to parts (a) and (b) as well as Assumption 3.4.5.

Now that we have these sequences of multipliers we would like to know what happens in the limit. As a preparation, we provide a technical lemma that will be used multiple times in future lemmas.

Lemma 3.4.8. Let Assumption 3.4.5 be satisfied. Let $\{x_i\}_{i\in\mathbb{N}} \subset X$, $\{p_i\}_{i\in\mathbb{N}} \subset V$, $\{\beta_i\}_{i\in\mathbb{N}} \subset \mathbb{R}$ be sequences such that $(x_i, p_i) \to (\bar{x}, \bar{p})$ and $\beta_i(x_i - \psi(p_i)) \rightharpoonup \bar{w}$ for some $\bar{w} \in X$. Furthermore, let $\hat{g} : X \times V \to \mathbb{R}$ be a function that is partially Gâteaux differentiable with respect to the first component and whose partial Gâteaux derivative $\hat{g}'_x : X \times V \to X^\star$ is continuous. Then the convergence

$$\beta_i(\hat{g}(x_i, p_i) - \hat{g}(\psi(p_i), p_i)) \to \langle \hat{g}'_x(\bar{x}, \bar{p}), \bar{w} \rangle_{X^\star \times X} \tag{3.30}$$

holds. The above can be applied for functions $\hat{g}$ that are given via

$$\begin{aligned}
(x, p) &\mapsto \langle f'_x(x, p), h_1 \rangle_{X^\star \times X}, \\
(x, p) &\mapsto \langle \lambda, g'_x(x, p) h_1 \rangle_{Y^\star \times Y}, \\
(x, p) &\mapsto \langle f'_p(x, p), h_2 \rangle_{V^\star \times V}, \\
(x, p) &\mapsto \langle \lambda, g'_p(x, p) h_2 \rangle_{Y^\star \times Y}, \\
(x, p) &\mapsto \langle \lambda, g(x, p) \rangle_{Y^\star \times Y},
\end{aligned}$$

where $h_1 \in X, h_2 \in V, \lambda \in Y^\star$ are fixed.

Proof. Since $\hat{g}$ is Gâteaux differentiable in every point there exists $t_i \in (0,1)$ such that

$$\langle \hat{g}'_x(t_i x_i + (1-t_i)\psi(p_i), p_i), \beta_i(x_i - \psi(p_i))\rangle_{X^\star \times X} = \beta_i(\hat{g}(x_i, p_i) - \hat{g}(\psi(p_i), p_i))$$

holds for each $i \in \mathbb{N}$. Recall that $\bar{x} = \psi(\bar{p})$ holds because $(\bar{x}, \bar{p})$ is a feasible point of (UL). Since $(t_i x_i + (1-t_i)\psi(p_i), p_i) \to (\bar{x}, \bar{p})$, $\beta_i(x_i - \psi(p_i)) \rightharpoonup \bar{w}$, and $\hat{g}'_x(\cdot,\cdot)$ is continuous we obtain (3.30).

By using Assumption 3.4.5 it can be shown that the suggested functions are partially Gâteaux differentiable and that the partial Gâteaux derivatives at (x,p) which are given by

$$\begin{aligned} f''_{xx}(x,p)h_1, \\ (g''_{xx}(x,p)h_1)^\star\lambda, \\ f''_{xp}(x,p)h_2, \\ (g''_{xp}(x,p)h_2)^\star\lambda, \\ g'_x(x,p)^\star\lambda \end{aligned}$$

are continuous (the symmetry of the second derivatives was also used here, which follows from the continuity of the second derivatives in the strong operator topology, see Lemmas 2.1.11 and 2.1.12).

Our next lemma is a very important step as it describes the key convergences for $\varepsilon \to 0$ that arise from the KKT conditions of (OVR(ε)). This is also a generalization of [Dempe, Harder, et al., 2019, Lemma 5.5].

Lemma 3.4.9. Let Assumption 3.4.5 be satisfied. Let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be a decreasing sequence such that $\varepsilon_i > 0$ for all $i \in \mathbb{N}$ and $\varepsilon_i \to 0$ as $i \to \infty$. For large $i \in \mathbb{N}$ let (x_i, p_i) be a global minimizer of (OVR(ε_i)) and let $\lambda_i, \mu_i, \beta_i, \rho_i$ be Lagrange multipliers that satisfy (3.28) and (3.29). Furthermore, we assume that $(x_i, p_i) \to (\bar{x}, \bar{p})$. Then there exist $\bar{w} \in X$, $\bar{\xi} \in Y^\star$, $\bar{\lambda} \in Y^\star$, $\bar{\rho} \in V^\star$ such that the convergences

$$\beta_i(x_i - \psi(p_i)) \rightharpoonup \bar{w}, \tag{3.31a}$$
$$\mu_i - \beta_i\lambda_i \overset{\star}{\rightharpoonup} \bar{\xi}, \tag{3.31b}$$
$$\lambda_i \to \bar{\lambda}, \tag{3.31c}$$
$$\rho_i \overset{\star}{\rightharpoonup} \bar{\rho} \tag{3.31d}$$

hold along a subsequence. The limits satisfy the system

$$F'_x(\bar{x},\bar{p}) + f''_{xx}(\bar{x},\bar{p})\bar{w} + (g''_{xx}(\bar{x},\bar{p})\bar{w})^\star\bar{\lambda} + g'_x(\bar{x},\bar{p})^\star\bar{\xi} = 0, \tag{3.32a}$$
$$F'_p(\bar{x},\bar{p}) + f''_{px}(\bar{x},\bar{p})\bar{w} + (g''_{px}(\bar{x},\bar{p})\bar{w})^\star\bar{\lambda} + g'_p(\bar{x},\bar{p})^\star\bar{\xi} + \bar{\rho} = 0, \tag{3.32b}$$
$$f'_x(\bar{x},\bar{p}) + g'_x(\bar{x},\bar{p})^\star\bar{\lambda} = 0, \tag{3.32c}$$

$$(g(\bar{x},\bar{p}),\bar{\lambda}) \in \operatorname{gph} \mathcal{N}_{\Phi}, \tag{3.32d}$$
$$(\bar{p},\bar{\rho}) \in \operatorname{gph} \mathcal{N}_{\Phi_{UL}}. \tag{3.32e}$$

Proof. Since $(x_i, p_i) \to (\bar{x}, \bar{p})$, Lemma 3.4.7 guarantees the existence of the multipliers $\lambda_i, \mu_i, \beta_i, \rho_i$ for large $i \in \mathbb{N}$. Without loss of generality we assume that this is the case for all $i \in \mathbb{N}$. Note that because of Lemma 3.4.6 (a) we also have $\psi(p_i) \to \bar{x}$.

Let us address the convergence (3.31c). From (3.28a) and the continuity of f'_x it follows that

$$g'_x(\psi(p_i), p_i)^\star \lambda_i \to -f'_x(\bar{x}, \bar{p}).$$

Then the existence of the limit $\bar{\lambda} \in Y^\star$ and the convergence $\lambda_i \to \bar{\lambda}$ follow directly from Lemma 2.1.8 (b). It is also clear that the limit satisfies (3.32c).

Our next goal is to show that $\{\beta_i(x_i - \psi(p_i))\}_{i\in\mathbb{N}}$ is a bounded sequence. If we multiply (3.28a) by β_i and subtract it from (3.29a) we get the equation

$$F'_x(x_i, p_i) + \beta_i(f'_x(x_i, p_i) - f'_x(\psi(p_i), p_i)) + g'_x(x_i, p_i)^\star \mu_i - \beta_i g'_x(\psi(p_i), p_i)^\star \lambda_i = 0. \tag{3.33}$$

By Lemma 2.1.19 we have

$$\langle g'_x(\psi(p_i), p_i)^\star \lambda_i, x_i - \psi(p_i)\rangle_{X^\star \times X} \leq 0$$

and

$$\langle g'_x(x_i, p_i)^\star \mu_i, x_i - \psi(p_i)\rangle_{X^\star \times X} \geq 0.$$

Thus, if we test (3.33) with $x_i - \psi(p_i)$ we obtain the inequality

$$\langle F'_x(x_i, p_i) + \beta_i(f'_x(x_i, p_i) - f'_x(\psi(p_i), p_i)), x_i - \psi(p_i)\rangle_{X^\star \times X} \leq 0.$$

Since $f(\cdot, p_i)$ is strongly convex with parameter $\gamma > 0$ on the feasible set $g(\cdot, p_i)^{-1}(\Phi)$ it follows from Lemma 2.1.30 (b) that the estimate

$$\begin{aligned}\beta_i \gamma \|x - \psi(p_i)\|^2 &\leq \langle \beta_i(f'_x(x_i, p_i) - f'_x(\psi(p_i), p_i)), x_i - \psi(p_i)\rangle_{X^\star \times X} \\ &\leq \langle -F'_x(x_i, p_i), x_i - \psi(p_i)\rangle_{X^\star \times X} \\ &\leq \|F'_x(x_i, p_i)\| \, \|x_i - \psi(p_i)\|\end{aligned}$$

holds. If we divide by $\|x_i - \psi(p_i)\|$ and consider that F'_x is continuous we can see that $\{\beta_i(x_i - \psi(p_i))\}_{i\in\mathbb{N}}$ is indeed bounded. Thus, this sequence has a weakly convergent subsequence. Without loss of generality we can write that $\beta_i(x_i - \psi(p_i)) \rightharpoonup \bar{w}$ for a suitable $\bar{w} \in X$, which was the claim in (3.31a).

Next, we want to show that $\{\mu_i - \beta_i\lambda_i\}_{i\in\mathbb{N}} \subset Y^\star$ is a weakly-$\star$ converging sequence. From Lemma 3.4.8 we obtain the weak-$\star$ convergence

$$\beta_i(f'_x(x_i, p_i) - f'_x(\psi(p_i), p_i)) \overset{\star}{\rightharpoonup} f''_{xx}(\bar{x}, \bar{p})\bar{w}. \tag{3.34}$$

Combined with (3.33) this implies that $g'_x(x_i,p_i)^\star\mu_i - \beta_i g'_x(\psi(p_i),p_i)^\star\lambda_i$ is weakly-$\star$ convergent. Due to Lemma 3.4.8 we have

$$\beta_i g'_x(x_i,p_i)^\star\lambda - \beta_i g'_x(\psi(p_i),p_i)^\star\lambda \overset{\star}{\rightharpoonup} \big(g''_{xx}(\bar{x},\bar{p})\bar{w}\big)^\star\lambda$$

for all $\lambda \in Y^\star$. Then Lemma 2.1.6 implies the weak-$\star$ convergence

$$\beta_i g'_x(x_i,p_i)^\star\lambda_i - \beta_i g'_x(\psi(p_i),p_i)^\star\lambda_i \overset{\star}{\rightharpoonup} \big(g''_{xx}(\bar{x},\bar{p})\bar{w}\big)^\star\bar{\lambda}. \tag{3.35}$$

Thus, it follows from (3.33) that $g'_x(x_i,p_i)^\star\mu_i - \beta_i g'_x(x_i,p_i)^\star\lambda_i$ is also weakly-$\star$ convergent. Because of Lemma 2.1.8 (c) this implies that there exists $\bar{\xi} \in Y^\star$ such that $\mu_i - \beta_i\lambda_i \overset{\star}{\rightharpoonup} \bar{\xi}$ and

$$g'_x(x_i,p_i)^\star\mu_i - \beta_i g'_x(x_i,p_i)^\star\lambda_i \overset{\star}{\rightharpoonup} g'_x(\bar{x},\bar{p})^\star\bar{\xi}$$

hold. With the help of the convergences (3.34) and (3.35) we obtain (3.32a) by taking the weak-$\star$ limit in (3.33).

Next, we want to study the behavior of (3.29b) as $i \to \infty$. Due to Lemma 3.4.7 (b) we know that φ is continuously Fréchet differentiable at p_i for large $i \in \mathbb{N}$ with Fréchet derivative $f'_p(\psi(p_i),p_i) + g'_p(\psi(p_i),p_i)^\star\lambda_i$. Thus, we obtain

$$F'_p(x_i,p_i)+\beta_i(f'_p(x_i,p_i)-f'_p(\psi(p_i),p_i))+g'_p(x_i,p_i)^\star\mu_i-\beta_i g'_p(\psi(p_i),p_i)^\star\lambda_i+\rho_i = 0. \tag{3.36}$$

Then we can use Lemmas 2.1.6 and 3.4.8 again to obtain the convergences

$$\beta_i\big(f'_p(x_i,p_i) - f'_p(\psi(p_i),p_i)\big) \overset{\star}{\rightharpoonup} f''_{px}(\bar{x},\bar{p})\bar{w}, \tag{3.37}$$

$$\beta_i g'_p(x_i,p_i)^\star\lambda_i - \beta_i g'_p(\psi(p_i),p_i)^\star\lambda_i \overset{\star}{\rightharpoonup} \big(g''_{px}(\bar{x},\bar{p})\bar{w}\big)^\star\bar{\lambda}. \tag{3.38}$$

Since g'_p is continuous we can apply Lemma 2.1.10 to obtain the convergence

$$g'_p(x_i,p_i)^\star\mu_i - \beta_i g'_p(x_i,p_i)^\star\lambda_i \overset{\star}{\rightharpoonup} g'_p(\bar{x},\bar{p})^\star\bar{\xi}. \tag{3.39}$$

If we apply the convergences (3.37), (3.38), and (3.39) to the equation (3.36) and also use that F'_p is continuous, then it follows that ρ_i is weakly-$\star$ convergent. If we denote its weak-$\star$ limit by $\bar{\rho}$ then (3.31d) and (3.32b) follow. Finally, the relations (3.32d) and (3.32e) follow from (3.31c) and (3.31d) and the definition of the normal cone.

We remark that the conditions in (3.32) are the same as in (3.23) with the exception of (3.23f). Thus, we should check whether further conditions are satisfied by the limit objects.

For some problems in the literature, the stationarity conditions contain coordinate-wise or pointwise conditions, see e.g. [Harder, G. Wachsmuth, 2018b, Definition 3.3]. It is not immediately clear how to generalize these properties for an abstract setting. The following abstract tool will help to obtain additional conditions in the stationarity system.

Definition 3.4.10. An operator $T \in \mathbb{L}(Y, Y)$ is called a *normal-cone-preserving operator* (with respect to Φ) if

$$T^\star \lambda \in \mathcal{N}_\Phi(y)$$

holds for all $y \in \Phi$, $\lambda \in \mathcal{N}_\Phi(y)$.

Note that $T^\star \lambda \in \mathcal{N}_\Phi(y)$ is equivalent to $\lambda \in \mathcal{N}_{T\Phi}(Ty)$ which justifies the naming of this concept. Furthermore, for a given set Φ the set of normal-cone-preserving operators is a closed and convex cone. Clearly, the identity operator is always a normal-cone-preserving operator. In Chapters 5 and 6 we will see how the concept of normal-cone-preserving operators will be used in Lebesgue or Sobolev spaces. For now, we provide two examples in finite dimensions.

Example 3.4.11. (a) Let $Y := \mathbb{R}^n$ for some $n \in \mathbb{N}$ and let $\Phi := \{x \in \mathbb{R}^n \mid x_i \geq 0 \ \forall i \in \{1, \dots, n\}\}$ be the n-dimensional nonnegative orthant. Then the set of normal-cone-preserving operators with respect to Φ is given by the set of diagonal matrices in $\mathbb{R}^{n \times n}$ with nonnegative entries.

(b) Let $Y := \mathbb{R}^2$ and let $\Phi := B_1(0)$ be the closed unit disc. Then the set of normal-cone-preserving operators is given by $\{\alpha \operatorname{id}_{\mathbb{R}^2} \mid \alpha \in [0, \infty)\}$.

Proof. We start with part (a). It can be seen that the normal cone to Φ at a point $x \in \Phi \subset \mathbb{R}^n$ can be expressed via

$$\mathcal{N}_\Phi(x) = \{\lambda \in \mathbb{R}^n \mid \lambda_i \leq 0, \ \lambda_i x_i = 0 \ \forall i \in \{1, \dots, n\}\}. \tag{3.40}$$

Let T be a diagonal matrix in $\mathbb{R}^{n \times n}$ with nonnegative entries. Then it follows by a simple calculating that $T^\top \lambda = T\lambda \in \mathcal{N}_\Phi(y)$ holds for all $y \in \Phi$, $\lambda \in \mathcal{N}_\Phi(y)$.

Now, let T be a normal-cone-preserving operator with respect to Φ. We want to show that T is a diagonal matrix with nonnegative entries. Let $i \in \{1, \dots, n\}$ be given. We define $y = \sum_{j \in \{1,\dots,n\} \setminus \{i\}} e_j \in \Phi \subset \mathbb{R}^n$. Then we have $-e_i \in \mathcal{N}_\Phi(y)$ due to (3.40). It follows that $-T^\top e_i \in \mathcal{N}_\Phi(y)$ holds. Applying (3.40) again yields $e_i^\top(-T^\top e_i) \leq 0$, i.e. the entries on the diagonal are nonnegative. For $j \in \{1, \dots, n\} \setminus \{i\}$ we also have $y_j = 1$ and therefore we obtain $e_j^\top(-T^\top e_i) = 0$ from (3.40). Thus, $T^\top$ is a diagonal matrix.

We continue with part (b). It is clear that the operator $\alpha \operatorname{id}_Y$ for $\alpha \geq 0$ is a normal-cone-preserving operator for all convex sets $\Phi \subset Y$.

Now, let T be a normal-cone-preserving operator with respect to $\Phi \subset \mathbb{R}^2$. We want to show that $T = \alpha \operatorname{id}_{\mathbb{R}^2}$ for some $\alpha \geq 0$. Here, the normal cone to $\Phi \subset \mathbb{R}^2$ at a point $x \in \Phi \subset \mathbb{R}^2$ can be expressed via

$$\mathcal{N}_\Phi(x) = \begin{cases} \{0\} & \text{if } \|x\| < 1, \\ \{\alpha x \mid \alpha \geq 0\} & \text{if } \|x\| = 1. \end{cases}$$

Thus, by the definition of normal-cone-preserving operators we have $T^\top x = \alpha_x x$ for all $x \in \mathbb{R}^2$ with $\|x\| = 1$, where $\alpha_x \geq 0$ are suitable real numbers. Since T is a linear operator it is easy to see that the α_x values must be independent of x, i.e. $\alpha_x = \alpha$ holds for all $x \in \mathbb{R}^2$ with $\|x\| = 1$, where $\alpha \geq 0$ is a suitable real number. Thus, $T = \alpha \, \mathrm{id}_{\mathbb{R}^2}$ follows.

We now turn to the first application of normal-cone-preserving operators.

Lemma 3.4.12. We consider the setting of Lemma 3.4.9.

(a) The condition
$$\langle \bar{\lambda}, T g'_x(\bar{x}, \bar{p})\bar{w} \rangle_{Y^\star \times Y} = 0$$
holds for all normal-cone-preserving operators $T \in \mathbb{L}(Y, Y)$ with respect to Φ.

(b) The condition
$$g'_x(\bar{x}, \bar{p})\bar{w} \in \mathrm{cl}(\mathrm{lin}\, \mathcal{T}_\Phi(g(\bar{x}, \bar{p})))$$
holds.

Proof. Without loss of generality we assume that the convergences from (3.31) hold for $i \to \infty$ (and not only along a subsequence). If $\{\beta_i\}_{i \in \mathbb{N}}$ has a bounded subsequence, then $\bar{w} = 0$ and the claims follow directly. Thus, we can (without loss of generality) assume that $\beta_i > 0$ for all $i \in \mathbb{N}$ and $\beta_i \to \infty$ as $i \to \infty$. Let $T \in \mathbb{L}(Y, Y)$ be a normal-cone-preserving operator. Because of $T^\star \mu_i \in \mathcal{N}_\Phi(g(x_i, p_i))$ we have

$$\langle \beta_i^{-1} T^\star \mu_i, \beta_i(g(x_i, p_i) - g(\psi(p_i), p_i)) \rangle_{Y^\star \times Y} = \langle T^\star \mu_i, g(x_i, p_i) - g(\psi(p_i), p_i) \rangle_{Y^\star \times Y} \geq 0 \tag{3.41}$$

for all $i \in \mathbb{N}$. We note that
$$\beta_i^{-1} \mu_i \to \bar{\lambda}$$
follows from the boundedness of $\beta_i \lambda_i - \mu_i$ and $\lambda_i \to \bar{\lambda}$. We also obtain the weak convergence
$$\beta_i(g(x_i, p_i) - g(\psi(p_i), p_i)) \rightharpoonup g'_x(\bar{x}, \bar{p})\bar{w} \tag{3.42}$$
from Lemma 3.4.8. Thus, it follows from (3.41) that $\langle \bar{\lambda}, T g'_x(\bar{x}, \bar{p})\bar{w} \rangle_{Y^\star \times Y} \geq 0$. Similarly, by using λ_i instead of $\beta_i^{-1}\mu_i$ in the previous argument, the other inequality $\langle \bar{\lambda}, T g'_x(\bar{x}, \bar{p})\bar{w} \rangle_{Y^\star \times Y} \leq 0$ can be shown, which completes the proof of part (a).

For part (b), let $\lambda \in \mathcal{N}_\Phi(g(\bar{x}, \bar{p})) \cap (-\mathcal{N}_\Phi(g(\bar{x}, \bar{p})))$ be arbitrary. Then we obtain
$$\langle \lambda, g(x_i, p_i) - g(\bar{x}, \bar{p}) \rangle_{Y^\star \times Y} = \langle \lambda, g(\psi(p_i), p_i) - g(\bar{x}, \bar{p}) \rangle_{Y^\star \times Y} = 0$$
for all $i \in \mathbb{N}$ from multiple applications of the definition of the normal cone. This implies
$$\begin{aligned} 0 &= \beta_i \langle \lambda, g(x_i, p_i) - g(\bar{x}, \bar{p}) \rangle_{Y^\star \times Y} - \beta_i \langle \lambda, g(\psi(p_i), p_i) - g(\bar{x}, \bar{p}) \rangle_{Y^\star \times Y} \\ &= \langle \lambda, \beta_i(g(x_i, p_i) - g(\psi(p_i), p_i)) \rangle_{Y^\star \times Y} \end{aligned}$$

for all $i \in \mathbb{N}$. Using the weak convergence (3.42) yields that $\langle \lambda, g'_x(\bar{x},\bar{p})\bar{w}\rangle_{Y^\star \times Y} = 0$ holds for all $\lambda \in \mathcal{N}_\Phi(g(\bar{x},\bar{p})) \cap (-\mathcal{N}_\Phi(g(\bar{x},\bar{p})))$. Then the claim follows from

$$\begin{aligned} g'_x(\bar{x},\bar{p})\bar{w} &\in \Big(\mathcal{N}_\Phi(g(\bar{x},\bar{p})) \cap (-\mathcal{N}_\Phi(g(\bar{x},\bar{p})))\Big)^\circ \\ &= \Big(\mathcal{T}_\Phi(g(\bar{x},\bar{p}))^\circ \cap (-\mathcal{T}_\Phi(g(\bar{x},\bar{p})))^\circ\Big)^\circ \\ &= \big(\mathcal{T}_\Phi(g(\bar{x},\bar{p})) - \mathcal{T}_\Phi(g(\bar{x},\bar{p}))\big)^{\circ\circ} \\ &= \big(\operatorname{lin} \mathcal{T}_\Phi(g(\bar{x},\bar{p}))\big)^{\circ\circ} \\ &= \operatorname{cl}\big(\operatorname{lin} \mathcal{T}_\Phi(g(\bar{x},\bar{p}))\big), \end{aligned}$$

where we used (2.7), (2.8), and Theorem 2.1.16 for the set equalities.

It is also possible to obtain a condition on $\bar{\xi}$ involving normal-cone-preserving operators under the assumption that Φ is a closed convex cone. Recall that we say that a closed convex cone K induces a lattice structure if it is pointed and the function $\max_K$ exists and is continuous, see Definition 2.1.20.

Lemma 3.4.13. We consider the setting of Lemma 3.4.9. Additionally, we assume that $\Phi \subset Y$ is a closed convex cone.

(a) The condition
$$\langle \bar{\xi}, Tg(\bar{x},\bar{p})\rangle_{Y^\star \times Y} = 0$$
holds for all normal-cone-preserving operators $T \in \mathbb{L}(Y,Y)$ with respect to Φ.

(b) The relation
$$T^\star \bar{\xi} \in \operatorname{cl}(\operatorname{lin}(\Phi^\circ))$$
holds for all normal-cone-preserving operators $T \in \mathbb{L}(Y,Y)$ with respect to Φ.

(c) If Φ induces a lattice structure on Y then the condition
$$\langle \bar{\xi}, y\rangle_{Y^\star \times Y} = 0$$
holds for all $y \in Y$ with $0 \leq_\Phi y \leq_\Phi g(\bar{x},\bar{p})$.

Proof. We can use Lemma 2.1.17 to observe that the conditions

$$\begin{aligned} &T^\star \mu_i \in \Phi^\circ, \qquad \langle T^\star \mu_i, g(x_i,p_i)\rangle_{Y^\star \times Y} = 0, \\ &T^\star \lambda_i \in \Phi^\circ, \quad \langle T^\star \lambda_i, g(\psi(p_i),p_i)\rangle_{Y^\star \times Y} = 0 \end{aligned} \tag{3.43}$$

hold for all $i \in \mathbb{N}$. Then it follows that the inequalities

$$\langle \mu_i - \beta_i \lambda_i, Tg(x_i,p_i)\rangle_{Y^\star \times Y} \geq 0 \quad \text{and} \quad \langle \mu_i - \beta_i \lambda_i, Tg(\psi(p_i),p_i)\rangle_{Y^\star \times Y} \leq 0$$

hold for all $i \in \mathbb{N}$. By using the known convergences from Lemma 3.4.9 we obtain the

claim of part (a).

For part (b) we note that $T^\star(\mu_i - \beta_i\lambda_i) \in \mathrm{cl}(\mathrm{lin}(\Phi^\circ))$ holds for all $i \in \mathbb{N}$. Since $\mathrm{cl}(\mathrm{lin}(\Phi^\circ))$ is a closed and convex set, the weak convergence $T^\star(\mu_i - \beta_i\lambda_i) \rightharpoonup T^\star\bar{\xi}$ implies the claim.

We continue with part (c). Let $y \in Y$ be given such that $0 \leq_\Phi y \leq_\Phi g(\bar{x}, \bar{p})$ holds. For each $i \in \mathbb{N}$, the properties of $\min_\Phi$ and $g(x_i, p_i), g(\psi(p_i), p_i) \in \Phi$ imply the relations

$$\begin{aligned} \min_\Phi(y, g(x_i, p_i)) &\in \Phi, & g(x_i, p_i) - \min_\Phi(y, g(x_i, p_i)) &\in \Phi, \\ \min_\Phi(y, g(\psi(p_i), p_i)) &\in \Phi, & g(\psi(p_i), p_i) - \min_\Phi(y, g(\psi(p_i), p_i)) &\in \Phi. \end{aligned}$$

Using (3.43) with the normal-cone-preserving operator $T = \mathrm{id}_Y$ yields the inequalities

$$\langle \lambda_i, g(x_i, p_i) - \min_\Phi(y, g(x_i, p_i)) \rangle_{Y^\star \times Y} \leq 0, \tag{3.44a}$$
$$\langle \mu_i, g(\psi(p_i), p_i) - \min_\Phi(y, g(\psi(p_i), p_i)) \rangle_{Y^\star \times Y} \leq 0, \tag{3.44b}$$

and from $\mu_i \in \mathcal{N}_\Phi(g(x_i, p_i))$, $\lambda_i \in \mathcal{N}_\Phi(g(\psi(p_i), p_i))$ we obtain the inequalities

$$\langle \mu_i, \min_\Phi(y, g(x_i, p_i)) - g(x_i, p_i) \rangle_{Y^\star \times Y} \leq 0, \tag{3.45a}$$
$$\langle \lambda_i, \min_\Phi(y, g(\psi(p_i), p_i)) - g(\psi(p_i), p_i) \rangle_{Y^\star \times Y} \leq 0. \tag{3.45b}$$

If we combine (3.44a) with (3.45a) and (3.45b) with (3.44b), the inequalities

$$\begin{aligned} \langle \mu_i - \beta_i\lambda_i, g(x_i, p_i) - \min_\Phi(y, g(x_i, p_i)) \rangle_{Y^\star \times Y} &\geq 0, \\ \langle \mu_i - \beta_i\lambda_i, g(\psi(p_i), p_i) - \min_\Phi(y, g(\psi(p_i), p_i)) \rangle_{Y^\star \times Y} &\leq 0, \end{aligned}$$

hold for all $i \in \mathbb{N}$. Using the convergence $\mu_i - \beta_i\lambda_i \overset{\star}{\rightharpoonup} \bar{\xi}$ and the fact that $\min_\Phi$ is continuous yields

$$\langle \bar{\xi}, g(\bar{x}, \bar{p}) - \min_\Phi(y, g(\bar{x}, \bar{p})) \rangle_{Y^\star \times Y} = 0.$$

Since $\langle \bar{\xi}, g(\bar{x}, \bar{p}) \rangle_{Y^\star \times Y} = 0$ holds due to part (a) and $\min_\Phi(y, g(\bar{x}, \bar{p})) = y$ holds due to $y \leq_\Phi g(\bar{x}, \bar{p})$, the claim follows.

The next lemma is a generalization of the previous lemma. We additionally cover the case where Φ is described by upper and lower bounds, which are given by some order induced by a lattice structure.

Lemma 3.4.14. We consider the setting of Lemma 3.4.9. We suppose that Φ has the structure

$$\Phi = \Phi_0 \times ((\hat{\Phi} + y_l) \cap (y_u - \hat{\Phi})),$$

where $\Phi_0 \subset Y_0$ is a closed convex cone, $\hat{\Phi} \subset \hat{Y}$ is a closed convex cone that induces a lattice structure on $\hat{Y}$, $y_l, y_u \in \hat{Y}$, and $Y_0, \hat{Y}$ are Banach spaces such that $Y = Y_0 \times \hat{Y}$. Then the conditions

$$\langle \bar{\xi}, T(P_{Y_0} g(\bar{x}, \bar{p}), 0) \rangle_{Y^\star \times Y} = 0, \tag{3.46a}$$

$$T^\star \bar{\xi} \in \operatorname{cl}(\operatorname{lin}(\Phi_0^\circ)) \times \hat{Y}, \tag{3.46b}$$

$$\langle \bar{\xi}, T(0, \min_{\hat{\Phi}}(P_{\hat{Y}} g(\bar{x}, \bar{p}) - y_l, y_u - P_{\hat{Y}} g(\bar{x}, \bar{p}))) \rangle_{Y^\star \times Y} = 0 \tag{3.46c}$$

hold for all $T \in \mathbb{L}(Y, Y)$ that are normal-cone-preserving operators with respect to Φ. Here, $P_{Y_0} : Y \to Y_0, P_{\hat{Y}} : Y \to \hat{Y}$ are the canonical projections onto the spaces $Y_0, \hat{Y}$.

Proof. The conditions (3.46a) and (3.46b) can be shown similar to Lemma 3.4.13. Thus, we focus on (3.46c) which requires different arguments.

Let $y_1 \in (\hat{\Phi} + y_l) \cap (y_u - \hat{\Phi})$, $y_2 \in \hat{\Phi}$, and $\lambda \in \mathcal{N}_\Phi((0, y_1))$ be given. Then we have

$$\begin{aligned} y_l &\leq_{\hat{\Phi}} -\min_{\hat{\Phi}}(-y_l, y_u - 2y_1, y_2 - y_1) = y_1 - \min_{\hat{\Phi}}(y_1 - y_l, y_u - y_1, y_2) \\ &\leq_{\hat{\Phi}} y_1 \leq_{\hat{\Phi}} y_1 + \min_{\hat{\Phi}}(y_1 - y_l, y_u - y_1, y_2) = \min_{\hat{\Phi}}(2y_1 - y_l, y_u, y_2 + y_1) \\ &\leq_{\hat{\Phi}} y_u, \end{aligned}$$

which shows that

$$(0, y_1 \pm \min_{\hat{\Phi}}(y_1 - y_l, y_u - y_1, y_2)) \in \Phi$$

holds. Due to $\lambda \in \mathcal{N}_\Phi((0, y_1))$ the equation

$$\langle \lambda, (0, \min_{\hat{\Phi}}(y_1 - y_l, y_u - y_1, y_2)) \rangle_{Y^\star \times Y} = 0 \tag{3.47}$$

follows. Next, we apply this little observation to our specific situation. Let $i \in \mathbb{N}$ be given. If we use $\lambda = T^\star \lambda_i$, $y_1 = P_{\hat{Y}} g(\psi(p_i), p_i)$, and $y_2 = \min_{\hat{\Phi}}(P_{\hat{Y}} g(x_i, p_i) - y_l, y_u - P_{\hat{Y}} g(x_i, p_i))$ in (3.47) we obtain

$$\langle T^\star \lambda_i, (0, \hat{y}_i) \rangle_{Y^\star \times Y} = 0,$$

where

$$\hat{y}_i := \min_{\hat{\Phi}}(P_{\hat{Y}} g(\psi(p_i), p_i) - y_l, y_u - P_{\hat{Y}} g(\psi(p_i), p_i), P_{\hat{Y}} g(x_i, p_i) - y_l, y_u - P_{\hat{Y}} g(x_i, p_i)) \in \hat{Y}.$$

Likewise, if we use $\lambda = T^\star \mu_i$, $y_1 = P_{\hat{Y}} g(x_i, p_i)$, and $y_2 = \min_{\hat{\Phi}}(P_{\hat{Y}} g(\psi(p_i), p_i) - y_l, y_u - P_{\hat{Y}} g(\psi(p_i), p_i))$ in (3.47) we obtain

$$\langle T^\star \mu_i, (0, \hat{y}_i) \rangle_{Y^\star \times Y} = 0.$$

Combined with the previous equation this yields

$$\langle \mu_i - \beta_i \lambda_i, T(0, \hat{y}_i) \rangle_{Y^\star \times Y} = 0.$$

Since $\min_{\hat{\Phi}}$ is continuous (see Definition 2.1.20), we know that $\hat{y}_i \to \min_{\hat{\Phi}}(P_{\hat{Y}} g(\bar{x}, \bar{p}) - y_l, y_u - P_{\hat{Y}} g(\bar{x}, \bar{p}))$ as $i \to \infty$. Then (3.46c) follows from (3.31b) and the above.

The next lemma can be understood as a generalization of [Dempe, Harder, et al., 2019, Lemma 5.7]. In some sense, the obtained condition corresponds to the condition that upgrades weak stationarity to C-stationarity.

Lemma 3.4.15. We consider the setting of Lemma 3.4.9. Let $T \in \mathbb{L}(Y, Y)$ be a normal-cone-preserving operator with respect to Φ. Furthermore, we assume that g is affine and that there exist operators $T_1, T_2, T_3 \in \mathbb{L}(X, X)$ such that

(a) $g'_x(\bar{x}, \bar{p})(T_1 + T_2 + T_3) = T g'_x(\bar{x}, \bar{p})$ holds,

(b) the inequality
$$\langle T_1^\star f''_{xx}(x, p)\hat{x}, \hat{x}\rangle_{X^\star \times X} \geq 0$$
holds for all $\hat{x} \in X$ and for all (x, p) in a neighborhood of $(\bar{x}, \bar{p})$,

(c) the map
$$\hat{x} \mapsto \langle T_2^\star f''_{xx}(\bar{x}, \bar{p})\hat{x}, \hat{x}\rangle_{X^\star \times X}$$
is sequentially weakly lower semi-continuous and $(x, p) \mapsto T_2^\star f''_{xx}(x, p)$ is continuous at $(\bar{x}, \bar{p})$,

(d) the operator $T_3 \in \mathbb{L}(X, X)$ is compact.

Then the inequality
$$\langle \bar{\xi}, T g'_x(\bar{x}, \bar{p})\bar{w}\rangle_{Y^\star \times Y} \geq 0$$
holds.

Proof. We define the bounded linear operator $\hat{T} \in \mathbb{L}(X, X)$ via $\hat{T} := T_1 + T_2 + T_3$. In particular, the operator $\hat{T}$ satisfies $T g'_x(\bar{x}, \bar{p}) = g'_x(\bar{x}, \bar{p})\hat{T}$. Without loss of generality we assume that the convergences from (3.31) hold for $i \to \infty$ (and not only along a subsequence). Let $i \in \mathbb{N}$ be given. Since T is a normal-cone-preserving operator we have

$$\beta_i \langle \mu_i, T g'_x(x_i, p_i)(\psi(p_i) - x_i)\rangle_{Y^\star \times Y} \leq 0$$

and

$$\beta_i \langle \beta_i \lambda_i, T g'_x(\psi(p_i), p_i)(x_i - \psi(p_i))\rangle_{Y^\star \times Y} \leq 0.$$

Since g is affine we have $g'_x(\psi(p_i), p_i) = g'_x(\bar{x}, \bar{p}) = g'_x(x_i, p_i)$. Thus, by adding the previous two inequalities we obtain

$$\begin{aligned} 0 &\leq \beta_i \langle \mu_i - \beta_i \lambda_i, T g'_x(\bar{x}, \bar{p})(x_i - \psi(p_i))\rangle_{Y^\star \times Y} \\ &= \beta_i \langle \mu_i - \beta_i \lambda_i, g'_x(\bar{x}, \bar{p})\hat{T}(x_i - \psi(p_i))\rangle_{Y^\star \times Y} \end{aligned}$$

and therefore

$$\beta_i \langle g'_x(x_i, p_i)^\star \mu_i - \beta_i g'_x(\psi(p_i), p_i)^\star \lambda_i, \hat{T}(x_i - \psi(p_i))\rangle_{Y^\star \times Y} \geq 0. \tag{3.48}$$

Let us test (3.33) with $\beta_i \hat{T}(x_i - \psi(p_i))$. Then, together with (3.48) we obtain

$$\beta_i \langle F'_x(x_i, p_i), \hat{T}(x_i - \psi(p_i))\rangle + \beta_i^2 \langle f'_x(x_i, p_i) - f'_x(\psi(p_i), p_i), \hat{T}(x_i - \psi(p_i))\rangle \leq 0.$$

If we rewrite the second term using the mean value theorem, we obtain

$$\beta_i\langle F_x'(x_i,p_i),\hat{T}(x_i-\psi(p_i))\rangle+\beta_i^2\langle f_{xx}''(\hat{x}_i,p_i)(x_i-\psi(p_i)),\hat{T}(x_i-\psi(p_i))\rangle\leq 0,$$

where $\hat{x}_i\in\operatorname{conv}\{x_i,\psi(p_i)\}\subset X$ is a suitable point between x_i and $\psi(p_i)$. Note that $\hat{x}_i\to\bar{x}$ holds as $i\to\infty$. If we use the abbreviation $w_i:=\beta_i(x_i-\psi(p_i))\in X$ for $i\in\mathbb{N}$ then the inequality becomes

$$\langle F_x'(x_i,p_i),\hat{T}w_i\rangle_{X^\star\times X}+\langle\hat{T}^\star f_{xx}''(\hat{x}_i,p_i)w_i,w_i\rangle_{X^\star\times X}\leq 0. \tag{3.49}$$

Recall that $w_i\rightharpoonup\bar{w}$ holds. By assumption (c) we have

$$\begin{aligned}0&\leq\liminf_{i\to\infty}\langle T_2^\star f_{xx}''(\bar{x},\bar{p})(w_i-\bar{w}),w_i-\bar{w}\rangle\\&\leq\liminf_{i\to\infty}\Big(\langle T_2^\star f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),w_i-\bar{w}\rangle+\big\|T_2^\star f_{xx}''(\bar{x},\bar{p})-T_2^\star f_{xx}''(\hat{x}_i,p_i)\big\|\,\|w_i-\bar{w}\|^2\Big)\\&=\liminf_{i\to\infty}\langle T_2^\star f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),w_i-\bar{w}\rangle.\end{aligned}$$

Using assumption (b) results in the inequality

$$0\leq\liminf_{i\to\infty}\langle(T_1^\star+T_2^\star)f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),w_i-\bar{w}\rangle_{X^\star\times X}. \tag{3.50}$$

From assumption (d) we obtain $T_3(w_i-\bar{w})\to 0$. Since the operator $f_{xx}''(\hat{x}_i,p_i)$ converges to $f_{xx}''(\bar{x},\bar{p})$ in the strong operator topology, we can apply Lemma 2.1.10 which yields $f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w})\overset{\star}{\rightharpoonup}0$ as $i\to\infty$. Therefore, the convergence

$$\langle T_3^\star f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),w_i-\bar{w}\rangle_{X^\star\times X}=\langle f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),T_3(w_i-\bar{w})\rangle_{X^\star\times X}\to 0$$

holds. If we combine this with (3.50) and use $\hat{T}=T_1+T_2+T_3$ the inequality

$$0\leq\liminf_{i\to\infty}\langle\hat{T}^\star f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),w_i-\bar{w}\rangle_{X^\star\times X}$$

follows. If we use the convergences $f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w})\overset{\star}{\rightharpoonup}0$, $f_{xx}''(\hat{x}_i,p_i)\bar{w}\to f_{xx}''(\bar{x},\bar{p})\bar{w}$, and $w_i\rightharpoonup\bar{w}$ again, we obtain

$$\begin{aligned}0&\leq\liminf_{i\to\infty}\langle\hat{T}^\star f_{xx}''(\hat{x}_i,p_i)(w_i-\bar{w}),w_i\rangle_{X^\star\times X}\\&=\liminf_{i\to\infty}\langle\hat{T}^\star f_{xx}''(\hat{x}_i,p_i)w_i,w_i\rangle_{X^\star\times X}-\langle\hat{T}^\star f_{xx}''(\bar{x},\bar{p})\bar{w},\bar{w}\rangle_{X^\star\times X}.\end{aligned}$$

Now we can combine this result with (3.49). If we also consider the convergences $F_x'(x_i,p_i)\to F_x'(\bar{x},\bar{p})$ and $\hat{T}w_i\rightharpoonup\hat{T}\bar{w}$ this yields

$$\langle F_x'(\bar{x},\bar{p}),\hat{T}\bar{w}\rangle_{X^\star\times X}+\langle f_{xx}''(\bar{x},\bar{p})\bar{w},\hat{T}\bar{w}\rangle_{X^\star\times X}\leq 0.$$

Finally, using the equality (3.32a) implies

$$\begin{aligned} 0 &\leq \langle (g''_{xx}(\bar{x},\bar{p})\bar{w})^\star \bar{\lambda} + g'_x(\bar{x},\bar{p})^\star \bar{\xi}, \hat{T}\bar{w}\rangle_{X^\star \times X} \\ &= \langle g'_x(\bar{x},\bar{p})^\star \bar{\xi}, \hat{T}\bar{w}\rangle_{X^\star \times X} = \langle \bar{\xi}, T g'_x(\bar{x},\bar{p})\bar{w}\rangle_{X^\star \times X}. \end{aligned}$$

This completes the proof.

While the assumptions in Lemma 3.4.15 look complicated, they allow us to apply the lemma in a variety of scenarios. For example, if f''_{xx} is continuous at $(\bar{x},\bar{p})$ then we can set $T_1 = T_3 = 0$ and we only need to verify that $\hat{x} \mapsto \langle T_2^\star f''_{xx}(\bar{x},\bar{p})\hat{x}, \hat{x}\rangle_{X^\star \times X}$ is sequentially weakly lower semi-continuous, where T_2 satisfies $g'_x(\bar{x},\bar{p})T_2 = T g'_x(\bar{x},\bar{p})$. Similarly, if we set $T_2 = 0$ then assumption (c) of Lemma 3.4.15 is automatically satisfied and we do not require the continuity of f''_{xx}.

Since $\mathrm{id}_Y \in \mathbb{L}(Y,Y)$ is always a normal-cone-preserving operator, we can obtain the condition

$$\langle \bar{\xi}, g'_x(\bar{x},\bar{p})\bar{w}\rangle_{Y^\star \times Y} \geq 0$$

from Lemma 3.4.15 if the relevant assumptions hold.

Note that in Lemma 3.4.9 we still have the assumption $(x_i,p_i) \to (\bar{x},\bar{p})$. As a first step we will show the weak convergence $(x_i,p_i) \rightharpoonup (\bar{x},\bar{p})$ by restricting Φ_{UL} to a sufficiently small neighborhood of $\bar{p}$. Using another trick this will then be extended to strong convergence in the proof of Theorem 3.4.17.

Lemma 3.4.16. Let Assumption 3.4.5 be satisfied. Additionally, we assume that $(\bar{x},\bar{p})$ is a strict local minimizer. Furthermore, let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be a decreasing sequence such that $\varepsilon_i > 0$ for all $i \in \mathbb{N}$ and $\varepsilon_i \to 0$ as $i \to \infty$. If Φ_{UL} is included in a sufficiently small neighborhood of $\bar{p}$, then a global minimizer (x_i,p_i) of (OVR(ε_i)) exists for large $i \in \mathbb{N}$ and these minimizers satisfy

$$F(x_i,p_i) \to F(\bar{x},\bar{p}) \qquad \text{and} \qquad p_i \rightharpoonup \bar{p} \qquad \text{and} \qquad x_i \rightharpoonup \bar{x}.$$

Proof. The existence of the global minimizers (x_i,p_i) of (OVR(ε_i)) for large $i \in \mathbb{N}$ follows directly from Lemma 3.4.6 (d).

By Lemma 3.4.6 (c) there exists $\delta > 0$ such that the function $(x,p) \mapsto f(x,p) - \varphi(p)$ is sequentially weakly lower semi-continuous on $B_\delta((\bar{x},\bar{p}))$ and the feasible set of (OVR(ε_i)) is included in $B_\delta((\bar{x},\bar{p}))$ for large $i \in \mathbb{N}$ if Φ_{UL} is included in a sufficiently small neighborhood of $\bar{p}$. Thus, we have $(x_i,p_i) \in B_\delta((\bar{x},\bar{p}))$ for large $i \in \mathbb{N}$. Since X and V are reflexive Banach spaces, there exists a weakly convergent subsequence of $\{(x_i,p_i)\}_{i\in\mathbb{N}}$. Without loss of generality we have $x_i \rightharpoonup x_0$ and $p_i \rightharpoonup p_0$ for some $(x_0,p_0) \in B_\delta((\bar{x},\bar{p}))$. Then we get

$$f(x_0,p_0) - \varphi(p_0) \leq \liminf_{i\to\infty} f(x_i,p_i) - \varphi(p_i) \leq \liminf_{i\to\infty} \varepsilon_i = 0$$

from the aforementioned sequential weak lower semi-continuity of the function $(x,p) \mapsto f(x,p) - \varphi(p)$ on $B_\delta((\bar{x},\bar{p}))$. We also know that $g(x_0,p_0) \in \Phi$ and $p_0 \in \Phi_{UL}$ due to Assumption 3.4.5 (i). Thus, (x_0,p_0) is a feasible point of (OVR) and we have $x_0 = \psi(p_0)$. Now we can use the sequential weak lower semi-continuity of F and the global optimality of (OVR(ε_i)) to obtain

$$F(x_0,p_0) \leq \liminf_{i\to\infty} F(x_i,p_i) \leq F(\bar{x},\bar{p}). \tag{3.51}$$

Since ψ exists and is continuous near $\bar{p}$ we can assume that $(\bar{x},\bar{p})$ is the unique global minimizer of (OVR) if Φ_{UL} is included in a sufficiently small neighborhood of $\bar{p}$. Thus, (3.51) implies $x_0 = \bar{x}$ and $p_0 = \bar{p}$. Since the weak limit (x_0,p_0) does not depend on the subsequence that we extracted earlier, the convergences $x_i \rightharpoonup \bar{x}$ and $p_i \rightharpoonup \bar{p}$ hold for all subsequences. We also obtain from (3.51) that $F(\bar{x},\bar{p}) = \liminf_{i\to\infty} F(x_i,p_i)$ holds for all subsequences and thus the convergence $F(x_i,p_i) \to F(\bar{x},\bar{p})$ follows.

Now we are in position to state the main theorem of this section. We will mostly collect the conditions obtained in Lemmas 3.4.9 and 3.4.12 to 3.4.15 for the local minimizer $(\bar{x},\bar{p})$.

Theorem 3.4.17. Let Assumption 3.4.5 be satisfied. Then there exist multipliers $\bar{w} \in X$, $\bar{\xi} \in Y^\star$, $\bar{\lambda} \in Y^\star$, $\bar{\rho} \in V^\star$ that satisfy the system (3.32). Furthermore, we have the following conditions.

(a) The equality
$$\langle \bar{\lambda}, T g_x'(\bar{x},\bar{p})\bar{w}\rangle_{Y^\star\times Y} = 0$$
holds for all normal-cone-preserving operators $T \in \mathbb{L}(Y,Y)$ with respect to Φ.

(b) The relation
$$g_x'(\bar{x},\bar{p})\bar{w} \in \operatorname{cl}\big(\operatorname{lin} \mathcal{T}_\Phi(g(\bar{x},\bar{p}))\big)$$
holds.

(c) Suppose that Φ is a closed convex cone and that it induces a lattice structure on Y. Then the condition
$$\langle \bar{\xi}, y\rangle_{Y^\star\times Y} = 0$$
holds for all $y \in Y$ with $0 \leq_\Phi y \leq_\Phi g(\bar{x},\bar{p})$.

(d) Suppose that Φ has the structure
$$\Phi = \Phi_0 \times \big((\hat{\Phi} + y_l) \cap (y_u - \hat{\Phi})\big),$$
where $\Phi_0 \subset Y_0$ is a closed convex cone, $\hat{\Phi} \subset \hat{Y}$ is a closed convex cone that induces a lattice structure on $\hat{Y}$, $y_l, y_u \in \hat{Y}$, and $Y_0, \hat{Y}$ are Banach spaces such that $Y = Y_0 \times \hat{Y}$. Then (3.46) holds for all normal-cone-preserving operators $T \in \mathbb{L}(Y,Y)$ with respect to Φ.

(e) Let $T \in \mathbb{L}(Y,Y)$, $T_1, T_2, T_3 \in \mathbb{L}(X,X)$ be given such that T is a normal-cone-preserving operator. Moreover, we assume that g is affine and that the assumptions (a) to (d) in Lemma 3.4.15 hold. Then the inequality

$$\langle \xi, T g'_x(\bar{x},\bar{p})\bar{w}\rangle_{Y^\star \times Y} \geq 0$$

holds.

Proof. Let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be an arbitrary decreasing sequence such that $\varepsilon_i > 0$ for all $i \in \mathbb{N}$ and $\varepsilon_i \to 0$ as $i \to \infty$. We want to prove the result by applying Lemma 3.4.9. Therefore, we need to show that global minimizers (x_i, p_i) of (OVR(ε_i)) exist and that we have $(x_i, p_i) \to (\bar{x}, \bar{p})$. We will do this by applying the results in this section to the modified upper level optimization problem

$$\begin{aligned} \min_{x,p} \quad & \hat{F}(x,p) := F(x,p) + \|p-\bar{p}\|_V^r \\ \text{s.t.} \quad & x \text{ solves (LL}(p)\text{)}, \\ & p \in \Phi_{UL} \cap B_{\varepsilon_0}(\bar{p}), \end{aligned} \tag{UL$_{\text{mod}}$}$$

where $r > 1$ is chosen according to Assumption 3.4.5 (l) and $\varepsilon_0 > 0$ can be chosen arbitrarily small. Because $p \mapsto \|p-\bar{p}\|^r$ is sequentially weakly lower semi-continuous and continuously Fréchet differentiable due to Assumption 3.4.5 (l), it follows that Assumption 3.4.5 also holds for (UL$_{\text{mod}}$). The corresponding relaxed optimization problem using the optimal value reformulation is given by

$$\begin{aligned} \min_{x,p} \quad & F(x,p) + \|p-\bar{p}\|_V^r \\ \text{s.t.} \quad & f(x,p) - \varphi(p) - \varepsilon \leq 0, \\ & g(x,p) \in \Phi, \\ & p \in \Phi_{UL} \cap B_{\varepsilon_0}(\bar{p}), \end{aligned} \tag{OVR$_{\text{mod}}(\varepsilon)$}$$

where $\varepsilon > 0$ is again the relaxation parameter. If we choose $\varepsilon_0 > 0$ small enough then $(\bar{x}, \bar{p})$ is not only a local minimizer but also a global minimizer of (UL$_{\text{mod}}$). Additionally, due to the addition of the term $\|p-\bar{p}\|^r$ we know that $(\bar{x}, \bar{p}) = (\psi(\bar{p}), \bar{p})$ is also a strict local (or global) minimizer of (UL$_{\text{mod}}$). Because we have restricted the feasible set of the upper level optimization problem to an arbitrarily small neighborhood of $\bar{p}$ we can apply Lemma 3.4.16. Thus, sequences $\{p_i\}_{i\in\mathbb{N}} \subset V$ and $\{x_i\}_{i\in\mathbb{N}} \subset X$ exist such that (x_i, p_i) is a global minimizer of (OVR$_{\text{mod}}(\varepsilon_i)$) for large $i \in \mathbb{N}$ and we have the convergences $p_i \rightharpoonup \bar{p}$, $x_i \rightharpoonup \bar{x}$ as well as

$$F(x_i,p_i) + \|p_i - \bar{p}\|_V^r = \hat{F}(x_i,p_i) \to \hat{F}(\bar{x},\bar{p}) = F(\bar{x},\bar{p}).$$

Since F is sequentially weakly lower semi-continuous we have

$$\begin{aligned} F(\bar{x},\bar{p}) &\leq \liminf_{i\to\infty} F(x_i,p_i) \\ &\leq \liminf_{i\to\infty} F(x_i,p_i) + \limsup_{i\to\infty} \|p_i-\bar{p}\|_V^r \\ &= \lim_{i\to\infty} (F(x_i,p_i) + \|p_i-\bar{p}\|_V^r) \\ &= F(\bar{x},\bar{p}). \end{aligned}$$

This implies $\limsup_{i\to\infty}\|p_i-\bar{p}\|^r = 0$ and hence $p_i \to \bar{p}$. Due to Lemma 3.4.6 (a) we know that ψ exists and is continuous near $\bar{p}$. Thus, we also have $\psi(p_i) \to \psi(\bar{p}) = \bar{x}$. By Lemma 3.4.6 (b) we also have $\|x_i - \psi(p_i)\|^2 \leq (2\varepsilon_i)^{1/2}\gamma^{-1/2}$ for all $i \in \mathbb{N}$ which implies $x_i \to \bar{x}$. Recall that the Lagrange multipliers $\lambda_i, \mu_i, \beta_i, \rho_i$ that satisfy (3.28) and (3.29) exist due to Lemma 3.4.7.

Now we are able to apply Lemma 3.4.9. Therefore, (3.32) holds for the setting of the problem ($\mathrm{UL}_{\mathrm{mod}}$). However, our claim is that (3.32) holds for the original setting of the problem (UL). It turns out that this does not make a difference because the equalities

$$\hat{F}'_p(\bar{x},\bar{p}) = F'_p(\bar{x},\bar{p}), \qquad \hat{F}'_x(\bar{x},\bar{p}) = F'_x(\bar{x},\bar{p}), \qquad \mathcal{N}_{\Phi_{UL}\cap B_{\varepsilon_0}(\bar{p})}(\bar{p}) = \mathcal{N}_{\Phi_{UL}}(\bar{p})$$

can be shown. Indeed, $\hat{F}'_p(\bar{x},\bar{p}) = F'_p(\bar{x},\bar{p})$ is true because of Assumption 3.4.5 (l) and the symmetry of the norm, $\hat{F}'_x(\bar{x},\bar{p}) = F'_x(\bar{x},\bar{p})$ is trivial, and $\mathcal{N}_{\Phi_{UL}}(\bar{p}) = \mathcal{N}_{\Phi_{UL}\cap B_{\varepsilon_0}(\bar{p})}(\bar{p})$ is true because the normal cone at $\bar{p}$ depends only on the local behavior of the convex set Φ_{UL} near $\bar{p}$.

The additional conditions outlined in parts (a) to (e) follow directly from Lemmas 3.4.12 to 3.4.15, respectively.

We will apply this result in subsequent chapters for examples in infinite-dimensional spaces.

One might wonder whether our approach can be used to obtain M-stationarity or a condition involving limiting normal cones. In finite dimensions, M-stationarity would require the additional condition

$$((g(\bar{x},\bar{p}),\bar{\lambda}),(\bar{\xi},-g'_x(\bar{x},\bar{p})\bar{w})) \in \operatorname{gph}\mathcal{N}^{\lim}_{\operatorname{gph}\mathcal{N}_\Phi}, \tag{3.52}$$

cf. (3.23f) and Remark 2.5.2 (b). An important ingredient of our approach was to study the limiting behavior of the KKT systems (3.28) and (3.29), which was done in Lemma 3.4.9. Thus, it would be interesting to know whether one could obtain a stationarity condition like (3.52) from the limiting behavior of the KKT systems for (OVR(ε_i)) and (LL(p_i)) with the convergences from (3.31). However, if we only rely on the KKT points of (OVR(ε_i)) this is not possible. This is shown by the following counterexample, which is taken from [Harder, 2016, Example 3.14].

In this counterexample, we construct a sequence of KKT points for (OVR(ε_i)) such that (3.52) cannot be obtained by taking the limit as in Lemma 3.4.9.

We mention that (x_i, p_i) are not local minimizers of (OVR(ε_i)) and that the limit $(\bar{x}, \bar{p})$ is not a local minimizer of (UL) or (OVR). As the counterexample is in finite-dimensional spaces and the data in the constraints of (MPCCR) is affine, we know from Theorem 2.5.3 that any local minimizer would have to be an M-stationary point (i.e. (3.52) would be satisfied).

Example 3.4.18. We use $X = Y = V := \mathbb{R}$, $\Phi := [0, \infty)$, $\Phi_{UL} := \mathbb{R}$, $F(x, p) := p - 2x$, $f(x, p) := \frac{1}{2}x^2 - px$, and $g(x, p) := x$. For $i \in \mathbb{N}$ we set $p_i := -1/i$, $x_i := 1/i$, $\beta_i := i$, $\varepsilon_i := 3/(2i^2)$, $\mu_i := 0$, $\lambda_i := -1/i$, $\rho_i := 0$. These choices satisfy (3.28) and (3.29). Furthermore, there exist $\bar{x}, \bar{p}, \bar{w}, \bar{\xi}, \bar{\lambda}, \bar{\rho} \in \mathbb{R}$ such that $(x_i, p_i) \to (\bar{x}, \bar{p})$, (3.31), and the stationarity system (3.32) are satisfied, but (3.52) is not satisfied.

Proof. We note that we have $\psi(p) = \max(p, 0)$. By direct calculation one can show that (3.28) and (3.29) are satisfied. Since (3.31) and $(x_i, p_i) \to (\bar{x}, \bar{p})$ should be satisfied we have to set $\bar{w} := 1$, $\bar{\xi} := 1$, $\bar{\lambda} := 0$, $\bar{\rho} := 0$, $\bar{x} := 0$, $\bar{p} := 0$. Again, it can be calculated that (3.32) is satisfied.

Note that in our setting the limiting normal cone is given via

$$\mathcal{N}^{\lim}_{\operatorname{gph}\mathcal{N}_\Phi}((0,0)) = \{(\xi, s) \in \mathbb{R}^2 \mid (\xi \leq 0 \wedge s \geq 0) \vee \xi s = 0\}.$$

Then

$$(\bar{\xi}, -g'_x(\bar{x}, \bar{p})\bar{w}) = (1, -1) \notin \mathcal{N}^{\lim}_{\operatorname{gph}\mathcal{N}_\Phi}((0,0)) = \mathcal{N}^{\lim}_{\operatorname{gph}\mathcal{N}_\Phi}((g(\bar{x}, \bar{p}), \bar{\lambda}))$$

follows.

We conclude that in order to obtain stronger stationarity conditions such as M-stationarity or (3.52) (if they can be shown at all in infinite-dimensional spaces) one would need to use different methods.

4 Legendre forms

In this chapter we will discuss the topic of Legendre forms and Legendre-$\star$ forms. Parts of this chapter cover results from [Harder, 2018], but we also extend these results to so-called Legendre-$\star$ forms and nonreflexive spaces. Legendre forms and Legendre-$\star$ forms are defined in Definitions 4.1.1 and 4.1.2.

Legendre forms (defined in Definition 4.1.1) are often discussed in a Hilbert space setting, see [Hestenes, 1951] or [Ioffe, Tikhomirov, 1979]. However, the definition of Legendre forms can also be used (without changes) in Banach spaces. For example, Legendre forms are defined in arbitrary Banach spaces in [Bonnans, Shapiro, 2000, Definition 3.73].

There are various results in the literature in which a Legendre form appears in the assumptions. It would be interesting to know in which spaces these results can be applied. For example, Legendre forms can be relevant for second-order sufficient optimality conditions. In reflexive spaces, if the quadratic form induced by the second derivative of an objective function is a Legendre form, then for second-order sufficient optimality conditions it suffices to show that this quadratic form is positive on some closed convex cone, whereas it is usually required to show coercivity on that convex cone, see [Bonnans, Shapiro, 2000, Lemma 3.75]. Other results in which a Legendre form that is defined on a reflexive Banach space appears in the assumptions are [Bonnans, Shapiro, 2000, Lemma 4.86, Theorem 5.5, Theorem 5.27], [Christof, G. Wachsmuth, 2018, Theorem 5.10], and [G. Wachsmuth, 2019, Theorem 5.7]. The reflexivity in these theorems is mostly used to obtain weakly convergent subsequences of bounded sequences. Legendre-$\star$ forms can also be relevant for the directional differentiability of parametric optimization problems in Banach spaces with a reflexive or separable predual space, see Proposition 3.1.12.

In [Harder, 2018] it was shown that if a Legendre forms exists on a reflexive Banach space then the space is isomorphic to a Hilbert space. We will prove an extension of this result in Theorem 4.3.9 which also covers the case of Legendre-$\star$ forms in (nonreflexive) Banach spaces with a separable predual space. We emphasize that this answers an open question raised in [Harder, 2018, Section 5.2]. Examples from the literature, where a Legendre-$\star$ form that is defined on a Banach space with a separable predual space appears, include [Christof, G. Wachsmuth, 2018, Lemma 5.1] and [Christof, G. Wachsmuth, 2020, Corollary 3.1]. The Banach space in these results does not need to be reflexive. Whether Legendre-$\star$ forms can exist in such spaces is therefore an interesting question.

4.1 Definitions and basic results

Let us state the definition of a Legendre form.

Definition 4.1.1. Let X be a normed space. We say that a quadratic form $Q : X \to \mathbb{R}$ is a *Legendre form* if it is sequentially weakly lower semi-continuous and if the implication

$$x_i \rightharpoonup x,\ Q(x_i) \to Q(x) \quad \Rightarrow \quad x_i \to x$$

holds for all sequences $\{x_i\}_{i\in\mathbb{N}} \subset X$ and $x \in X$.

If we have a quadratic form on a dual space of a normed space, the similar concept of a Legendre-$\star$ form, which is based on the weak-$\star$ sequential topology, could be interesting. Legendre-$\star$ forms were already defined in Definition 3.1.11 but we repeat the definition for the convenience of the reader.

Definition 4.1.2. Let X be a normed space. We say that a quadratic form $Q : X^\star \to \mathbb{R}$ is a *Legendre-$\star$ form* if it is sequentially weakly-$\star$ lower semi-continuous and if the implication

$$x_i \overset{\star}{\rightharpoonup} x,\ Q(x_i) \to Q(x) \quad \Rightarrow \quad x_i \to x$$

holds for all sequences $\{x_i\}_{i\in\mathbb{N}} \subset X^\star$ and $x \in X^\star$.

A Legendre-$\star$ form was used in [Christof, G. Wachsmuth, 2018, Lemma 5.1] but the object was simply called a Legendre form. However, in order to differentiate this object from the Legendre forms as defined in Definition 4.1.1, we assign a different name to it. We also mention that Definition 4.1.2 is slightly more general than the definition used in [Christof, G. Wachsmuth, 2018, Lemma 5.1], since we require that Q is sequentially weakly-$\star$ lower semi-continuous instead of weakly-$\star$ lower semi-continuous.

We describe the relationship between Legendre forms and Legendre-$\star$ forms in the following corollary. The results follow directly from the definition.

Corollary 4.1.3. Every Legendre-$\star$ form is a Legendre form. Every Legendre form on a reflexive Banach space X is a Legendre-$\star$ form on X (where we use $X^\star$ as the predual space of X).

The next lemma shows that if a quadratic form on a Banach space is lower semi-continuous it is already continuous. A similar result is known for convex functions, see [Bonnans, Shapiro, 2000, Proposition 2.111]. However, a quadratic form does not need to be convex, thus a different class of function is discussed. The lemma together with its proof is taken directly from [Harder, 2018, Lemma 4.6]. We are not aware of another source for this result.

Lemma 4.1.4. Let Q be a quadratic form on a Banach space X. If Q is lower semi-continuous then it is continuous.

Proof. Our first goal is to find $x \in X, \varepsilon > 0$ such that Q is bounded on $B_{2\varepsilon}(x)$. In order to do this, consider the sets $A_i := \{y \in X \mid Q(y) \leq i\}$ for $i \in \mathbb{N}$. Because Q is lower semi-continuous, these sets are closed. Since $X = \bigcup_{n\in\mathbb{N}} A_i$, by Baire's theorem one of the sets A_i has to contain a nonempty open set. Therefore, Q is bounded from above on some nonempty open set. Since Q is lower semi-continuous, it is also locally bounded from below, i.e., for a given $x \in X$ we find $\hat{\varepsilon} > 0, \hat{\alpha} > -\infty$ such that $Q(y) > \hat{\alpha}$ for all $y \in B_{\hat{\varepsilon}}(x)$.

Thus, we know that there exist $x \in X, \varepsilon > 0, \alpha > 0$ such that $|Q(\hat{x})| < \alpha$ for all $\hat{x} \in B_{2\varepsilon}(x)$. Using the parallelogram law from Lemma 2.1.25 (a), we have

$$|Q(y)| = \frac{1}{2}|Q(x+y) + Q(x-y) - 2Q(x)| \leq 2\alpha \quad \forall\, y \in B_{2\varepsilon}(0).$$

Now it follows from the definition of a quadratic form that there exists a bilinear function $B : X \times X \to \mathbb{R}$ such that $Q(x) = B(x,x)$ for all $x \in X$. Without loss of generality we can assume that B is symmetric. Moreover, if $y_1, y_2 \in B_\varepsilon(0)$, we have $|B(y_1,y_2)| = \frac{1}{4}|Q(y_1+y_2) - Q(y_1-y_2)| \leq \alpha$. Thus, the bilinear function B is bounded in a neighborhood of 0 and therefore continuous. The claim follows.

Because sequentially weakly lower semi-continuous functions and sequentially weakly-$\star$ lower semi-continuous functions are lower semi-continuous, the following corollary is a direct consequence of Lemma 4.1.4.

Corollary 4.1.5. Every Legendre-$\star$ form is continuous. Every Legendre form on a Banach space is continuous.

In the following, whenever we have a continuous quadratic form $Q : X \to \mathbb{R}$ on a normed space X, we denote by $T \in \mathbb{L}(X, X^\star)$ the unique operator from Lemma 2.1.26, i.e. T is symmetric and satisfies $Q(x) = \langle Tx, x\rangle_{X^\star \times X}$ for all $x \in X$.

If a quadratic form is continuous, it is possible to replace the limit x with 0 in the definition of Legendre forms.

Proposition 4.1.6. (a) Let X be a Banach space and $Q : X \to \mathbb{R}$ be a quadratic form that is sequentially weakly lower semi-continuous. Then Q is a Legendre form if and only if the implication

$$x_i \rightharpoonup 0,\ Q(x_i) \to Q(0) \quad \Rightarrow \quad x_i \to 0$$

holds for all sequences $\{x_i\}_{i\in\mathbb{N}} \subset X$.

(b) Let X be a normed space and $Q : X^\star \to \mathbb{R}$ be a quadratic form that is sequentially

weakly-$\star$ lower semi-continuous. Then Q is a Legendre-$\star$ form if and only if the implication

$$x_i \overset{\star}{\rightharpoonup} 0,\ Q(x_i) \to Q(0) \quad \Rightarrow \quad x_i \to 0 \tag{4.1}$$

holds for all sequences $\{x_i\}_{i\in\mathbb{N}} \subset X^\star$.

Proof. The function Q is continuous in both cases due to Lemma 4.1.4.

We only prove part (b). The proof for part (a) works in the same way.

We suppose that (4.1) holds for all sequences $\{x_i\}_{i\in\mathbb{N}} \subset X^\star$. Let $\{y_i\}_{i\in\mathbb{N}} \subset X^\star$ be a sequence such that $y_i \overset{\star}{\rightharpoonup} y$ for some $y \in X^\star$ and $Q(y_i) \to Q(y)$. We have to show that $y_i \to y$ holds. Using the sequentially weakly-$\star$ lower semi-continuity of Q and the parallelogram law results in

$$\begin{aligned} 4Q(y) = Q(2y) + Q(0) &\leq \liminf_{i\to\infty} (Q(y_i + y) + Q(y_i - y)) \\ &\leq \limsup_{i\to\infty} (Q(y_i + y) + Q(y_i - y)) \leq \lim_{i\to\infty} 2(Q(y) + Q(y_i)) = 4Q(y). \end{aligned}$$

This implies the convergence $Q(y_i - y) \to 0$. By (4.1) it follows that $y_i - y \to 0$ holds. Thus, Q is a Legendre-$\star$ form. The other direction follows directly from the definition.

4.2 Legendre forms in Hilbert spaces

This section is taken from [Harder, 2018, Section 3]. We intend to give some results from the literature that discuss Legendre forms in Hilbert spaces. Note that due to Corollary 4.1.3 the notions Legendre forms and Legendre-$\star$ forms coincide. We mention that once we have shown Theorem 4.3.9, the results in this section extend to Banach spaces that are reflexive or have a separable predual space.

The following theorem is due to [Hestenes, 1951, Theorem 11.6] and yields a good characterization of Legendre forms in Hilbert spaces. Moreover, it is useful for constructing examples of Legendre forms.

Theorem 4.2.1. A quadratic form Q on a Hilbert space X is a Legendre form if and only if it can be expressed as

$$Q(x) = Q_1(x) - Q_2(x),$$

where Q_1 is a coercive quadratic form and Q_2 is a sequentially weakly continuous quadratic form.

It can be shown that a quadratic form $Q(x) = \langle Tx, x\rangle$ in a Hilbert space is sequentially weakly continuous if and only if T is compact, see [Ioffe, Tikhomirov, 1979, Theorem 1 in Section 6.2.3]. A consequence is that in a finite-dimensional space all quadratic forms are

Legendre forms. We note that if $T \in \mathbb{L}(X, X^\star)$ is a compact operator, the quadratic form $x \mapsto \langle Tx, x\rangle$ does not need to be weakly continuous. For instance, the compact operator $T \in \mathbb{L}(\ell^2, \ell^2)$ given by $Te_i = \frac{1}{i}e_i \, \forall i \in \mathbb{N}$ yields a sequentially weakly continuous quadratic form Q that is not weakly continuous. Indeed, $Q(x) = 1$ for all $x \in A := \{\sqrt{i}e_i \mid i \in \mathbb{N}\}$ but $Q(0) = 0$ and 0 is in the weak closure (but not in the weak sequential closure) of A.

A simple combination of [Hestenes, 1951, Theorems 7.1, 11.3, and 11.4] yields a statement that gives us a good description of the structure of Legendre forms in Hilbert spaces.

Theorem 4.2.2. Let Q be a Legendre form on a Hilbert space X. Then there is an orthogonal and Q-orthogonal decomposition

$$X = Y_+ \dotplus Y_0 \dotplus Y_-$$

with closed linear subspaces $Y_+, Y_0, Y_- \subset X$ such that Q is coercive on Y_+, $-Q$ is coercive on Y_-, and $Q = 0$ on Y_0. Moreover, Y_0 and Y_- are finite-dimensional.

It should be noted that it is not possible to simply prove this result (without orthogonality) in reflexive Banach spaces in a way that is analogous to the original proof in Hilbert spaces. This is because the proof of [Hestenes, 1951, Theorem 7.1] uses that a quadratic form can be expressed as a difference of two nonnegative quadratic forms. However, this is not possible in the Banach spaces ℓ^p where $p \in (1, 2)$, see [Kalton, Konyagin, Veselý, 2008, Corollary 1.7]. Nonetheless, the structure obtained in Theorem 4.2.2 can give us ideas for gaining insight into the structure of Legendre forms or Legendre-⋆ forms in spaces that are not Hilbert spaces.

4.3 Hilbertizability of spaces with Legendre-⋆ forms

The goal of this section will be to prove Theorem 4.3.9, which states that if a Legendre-⋆ form exists on a reflexive space or a space with a separable predual space then the underlying space is already isomorphic to a Hilbert space. This provides a generalization of [Harder, 2018, Theorem 1.1]. One difference is that our generalization can also be applied to spaces that are not reflexive, provided they have a separable predual space. The other difference is that Theorem 4.3.9 is formulated for Legendre-⋆ forms and not Legendre forms.

However, we formulate the results in this section such that they also cover the situation of Legendre forms in reflexive Banach spaces. We will do this by using the assumption that a normed space X exists which is separable or reflexive. Then the Legendre-⋆ form is defined on $X^\star$. In the reflexive case Legendre-⋆ forms and Legendre forms coincide (see Corollary 4.1.3) so Theorem 4.3.9 also contains the result of [Harder, 2018, Theorem 1.1].

One might wonder if Theorem 4.3.9 is also true for Legendre forms in nonreflexive spaces. As Example 4.4.3 shows, this is not the case.

We mention that the condition that X is separable or reflexive is a condition that guarantees the existence of weakly-$\star$ converging subsequences of bounded sequences in $X^\star$, see Theorem 2.3.1. Since this property is usually required in applications, our assumption that X is reflexive or separable is a reasonable one.

We start with some technical preparations. While investigating Legendre-$\star$ forms it can be useful to restrict the Legendre-$\star$ form to a weakly-$\star$ closed linear subspace of $X^\star$. In order to apply previous results of Legendre-$\star$ forms to such a linear subspace we need to show that a Legendre-$\star$ form stays a Legendre-$\star$ form if restricted to such a linear subspace. Therefore, we need the following lemma.

Lemma 4.3.1. Let X be a normed space and let $\hat{Y}$ be a weakly-$\star$ closed linear subspace of $X^\star$. Then $\hat{Y}$ has a predual space, i.e. there is a space Y such that $\hat{Y}$ is isometrically isomorphic to $Y^\star$.

If X is separable then Y is also separable. Moreover, a sequence $\{y_i\}_{i\in\mathbb{N}} \subset \hat{Y} \cong Y^\star$ converges in the weak-$\star$ topology of $X^\star$ if and only if it converges in the weak-$\star$ topology of $Y^\star$.

Proof. Note that $\hat{Y}^\perp$ is a closed linear subspace of X. Thus, the space $X/\hat{Y}^\perp$ is a normed space. We define the linear operator

$$\iota : \hat{Y} \to (X/\hat{Y}^\perp)^\star, \quad y \mapsto (x + \hat{Y}^\perp \mapsto \langle y, x\rangle_{X^\star\times X}).$$

Then it can be seen that the operator ι is well-defined. In order to show that ι is isometric, let $y \in \hat{Y}$ be given. Due to $\|x + \hat{Y}^\perp\|_{X/\hat{Y}^\perp} \leq \|x\|_X$ we obtain

$$\begin{aligned}\|y\|_{X^\star} &= \sup\Big\{|\langle y, x\rangle| \;\Big|\; x \in X,\ \|x\|_X \leq 1\Big\} \\ &\leq \sup\Big\{|\langle y, x\rangle| \;\Big|\; x \in X,\ \|x + \hat{Y}^\perp\|_{X/\hat{Y}^\perp} \leq 1\Big\} = \|\iota y\|_{(X/\hat{Y}^\perp)^\star}.\end{aligned}$$

For the other inequality, let $\varepsilon > 0$ be given. Then for each $x \in X$ with $\|x+\hat{Y}^\perp\|_{X/\hat{Y}^\perp} \leq 1$ there exists an element $x_0 \in \hat{Y}^\perp$ such that $\|x + x_0\|_X \leq 1 + \varepsilon$. It follows that

$$\begin{aligned}\|\iota y\|_{(X/\hat{Y}^\perp)^\star} &= \sup\Big\{|\langle y, x\rangle| \;\Big|\; x \in X,\ \|x + \hat{Y}^\perp\|_{X/\hat{Y}^\perp} \leq 1\Big\} \\ &\leq \sup\Big\{|\langle y, x\rangle| \;\Big|\; x \in X,\ x_0 \in \hat{Y}^\perp, \|x + x_0\|_X \leq 1 + \varepsilon\Big\} \\ &= \sup\Big\{|\langle y, x + x_0\rangle| \;\Big|\; x \in X,\ x_0 \in \hat{Y}^\perp, \|x + x_0\|_X \leq 1 + \varepsilon\Big\} = (1 + \varepsilon)\|y\|_{X^\star}.\end{aligned}$$

Since $\varepsilon > 0$ and $y \in \hat{Y}$ were arbitrary, we have shown that ι is an isometry. Therefore, ι is also injective. Next, we will show that ι is surjective. Let $x^\star \in (X/\hat{Y}^\perp)^\star$ be given. Then we define the linear functional

$$y^\star : X \to \mathbb{R}, \quad x \mapsto \langle x^\star, x + \hat{Y}^\perp \rangle_{(X/\hat{Y}^\perp)^\star \times (X/\hat{Y}^\perp)}.$$

Clearly, $y^\star \in X^\star$. It is also easy to see that $y^\star \in \hat{Y}^{\perp\perp}$. Since $\hat{Y}$ is weakly-⋆ closed we obtain from Theorem 2.1.16 that $\hat{Y}^{\perp\perp} = \hat{Y}$. Thus, $y^\star \in \hat{Y}$ and we have

$$\iota y^\star = (x + \hat{Y}^\perp \mapsto \langle y^\star, x \rangle_{X^\star \times X}) = \left(x + \hat{Y}^\perp \mapsto \langle x^\star, x + \hat{Y}^\perp \rangle_{(X/\hat{Y}^\perp)^\star \times (X/\hat{Y}^\perp)}\right) = x^\star.$$

Since $x^\star \in (X/\hat{Y}^\perp)^\star$ was arbitrary this shows that ι is surjective.

Summarizing the results, we know that ι is an isometric isomorphism. Thus, we can define $Y := X/\hat{Y}^\perp$ and $\hat{Y}$ is isometrically isomorphic to $Y^\star$.

The separability of Y follows from the separability of X.

For $x \in X$ and $y \in \hat{Y}$ we have

$$\langle y, x \rangle_{X^\star \times X} = \langle \iota y, x + \hat{Y}^\perp \rangle_{(X/\hat{Y}^\perp)^\star \times (X/\hat{Y}^\perp)}.$$

This equation implies that a sequence $\{y_i\}_{i \in \mathbb{N}} \subset \hat{Y} \cong Y^\star$ converges in the weak-⋆ topology of $X^\star$ if and only if $\{\iota y_i\}_{i \in \mathbb{N}} \subset Y^\star \cong \hat{Y}$ it converges in the weak-⋆ topology of $Y^\star \cong \hat{Y}$ (with respect to the predual space Y).

Using this lemma we can conclude the following corollary.

Corollary 4.3.2. Let X be a normed space and let $\hat{Y}$ be a weakly-⋆ closed linear subspace of $X^\star$. If $Q : X^\star \to \mathbb{R}$ is a Legendre-⋆ form, then its restriction to $\hat{Y}$ is again a Legendre-⋆ form.

Here the weak-⋆ convergent sequences in $\hat{Y}$ are understood with respect to the predual space Y of $\hat{Y}$ that was described in Lemma 4.3.1.

The analogous result of Corollary 4.3.2 for Legendre forms can be found in [Harder, 2018, Lemma 4.5] without the assumption that the linear subspace is weakly-⋆ closed (or even closed).

Now that we have shown that Legendre-⋆ forms are still Legendre-⋆ forms if restricted to a weakly-⋆ closed linear subspace, we will do the same for the property that the predual space is separable or reflexive.

Lemma 4.3.3. Let X be a normed space that is separable or reflexive and let $\hat{Y}$ be a weakly-$\star$ closed linear subspace of $X^\star$. If X is separable or reflexive then $\hat{Y}$ has a predual space Y (i.e. $\hat{Y}$ is isometrically isomorphic to $Y^\star$) that is separable or reflexive.

Proof. The existence of Y is guaranteed by Lemma 4.3.1. If X is separable then Y is separable by Lemma 4.3.1. If X is reflexive then $X^\star$ is reflexive. It is known that closed linear subspaces of reflexive spaces are again reflexive, see [Conway, 1985, Corollary V.4.3]. Therefore, $\hat{Y} \cong Y^\star$ is reflexive and hence Y is reflexive.

A first Hilbertizability result was already given in Proposition 2.1.34, which states that a Banach space is Hilbertizable if there exists a continuous and coercive quadratic form on that Banach space. We can use this result to show that if a Legendre-$\star$ form is positive then the underlying space is Hilbertizable. This is a generalization of [Harder, 2018, Proposition 4.8]. In order to facilitate the application of the next proposition we formulate it for a sequentially weakly-$\star$ closed linear subspace.

Proposition 4.3.4. Let X be a normed space that is separable or reflexive. If Q is a Legendre-$\star$ form on $X^\star$ and Q is positive on a sequentially weakly-$\star$ closed linear subspace $Y_+ \subset X^\star$ then Y_+ is Hilbertizable.

Proof. We follow the proof of [Harder, 2018, Proposition 4.8] but adapt it to our setting. Assume that Q is not coercive on Y_+. Then there is a sequence $\{y_i\}_{i\in\mathbb{N}}$ in Y_+ with $Q(y_i) \to 0$ as $i \to \infty$ and $\|y_i\| = 1$ for all $i \in \mathbb{N}$.

Because $\{y_i\}_{i\in\mathbb{N}}$ is bounded and X is separable or reflexive, we obtain from Theorem 2.3.1 that there is a weakly-$\star$ convergent subsequence of $\{y_i\}_{i\in\mathbb{N}}$. Without loss of generality we can assume that $y_i \overset{\star}{\rightharpoonup} y$ as $i \to \infty$ for some $y \in X^\star$. Since Y_+ is sequentially weakly-$\star$ closed we also know that $y \in Y_+$. We also have $0 \leq Q(y) \leq \liminf_{i\to\infty} Q(y_i) = 0$. It follows that $Q(y) = 0$ and therefore $y_i \overset{\star}{\rightharpoonup} y = 0$. Applying the definition of a Legendre-$\star$ form yields $y_i \to 0$ which is a contradiction to $\|y_i\| = 1$ for $i \in \mathbb{N}$.

Thus, we have shown that Q is coercive on Y_+. Now we can apply Proposition 2.1.34 to the closed linear subspace Y_+. It follows that Y_+ is Hilbertizable.

We mention that it can be shown that sequentially weakly-$\star$ closed linear subspaces are weakly-$\star$ closed if the predual space is separable, see [Conway, 1985, Corollary V.12.7]. However, we do not rely on this result.

Having discussed positive Legendre-$\star$ forms, we move to nonpositive Legendre-$\star$ forms. The next lemma is a generalization of [Harder, 2018, Proposition 4.9]. However, our proof has to use a different strategy because we are interested in applying the result for linear subspaces that are not necessarily weakly-$\star$ closed.

Lemma 4.3.5. Let X be a normed space that is separable or reflexive. If Q is a Legendre-$\star$ form on $X^\star$ and $Y \subset X^\star$ is a linear subspace such that Q is nonpositive on Y then Y is finite-dimensional.

Proof. Due to the lower semi-continuity of Q it can be seen that Q is also nonpositive on the closure of Y. Thus, we can without loss of generality assume that Y is closed.

We will show that the unit ball in Y is sequentially compact (with respect to the norm topology). Let $\{y_i\}_{i\in\mathbb{N}}$ be a sequence in Y such that $\|y_i\| \leq 1$ for all $i \in \mathbb{N}$. Because X is separable or reflexive we obtain from Theorem 2.3.1 that there is a bounded subsequence that converges weakly-$\star$ to an element $y \in X^\star$ with $\|y\| \leq 1$. Without loss of generality we have $y_i \overset{\star}{\rightharpoonup} y$ as $i \to \infty$. Using the sequential weak-$\star$ lower semi-continuity of Q we first obtain

$$Q(y_i - y) \leq \liminf_{j\to\infty} Q(y_i - y_j) \leq 0 \qquad \forall i \in \mathbb{N}$$

and then

$$0 = Q(0) \leq \liminf_{i\to\infty} Q(y_i - y) \leq \limsup_{i\to\infty} Q(y_i - y) \leq 0.$$

The convergence $Q(y_i - y) \to Q(0)$ follows. Applying the definition of the Legendre-$\star$ form yields that $y_i - y \to 0$ as $i \to \infty$. Since Y is closed it follows that $y \in Y$. Therefore, the sequence $\{y_i\}_{i\in\mathbb{N}}$ has a convergent subsequence in Y. Thus, the unit ball in Y is sequentially compact and therefore compact. According to [Rudin, 1991, Theorem 1.22] this implies that the linear subspace Y is finite-dimensional.

We have seen that in Hilbert spaces we have the decomposition of the underlying space into Q-orthogonal linear subspaces Y_+, Y_-, Y_0 such that Q is positive on Y_+, negative on Y_-, and vanishes on Y_0, see Theorem 4.2.2. Thus, it might be a good idea to construct similar linear subspaces for our situation. The next lemma is a straightforward generalization of [Harder, 2018, Lemma 4.11] to Legendre-$\star$ forms. It constructs a maximal linear subspace $Y_- \subset X^\star$.

Lemma 4.3.6. Let X be a normed space that is separable or reflexive. Suppose that Q is a Legendre-$\star$ form on $X^\star$. Then there is a maximal linear subspace $Y_- \subset X^\star$ (with respect to set inclusion) with the property that Q is negative on Y_-.

Proof. We follow the proof of [Harder, 2018, Lemma 4.11]. Suppose there is no maximal linear subspace with the desired property. Then we can construct a sequence of linear subspaces $\{Y_i\}_{i\in\mathbb{N}}$ such that $Y_i \subsetneq Y_{i+1}$ and Q is negative on Y_i for all $i \in \mathbb{N}$. If we define the union $Y := \bigcup_{i\in\mathbb{N}} Y_i$, then we obtain that Y is an infinite-dimensional linear subspace such that Q is negative on Y. By Lemma 4.3.5 this is a contradiction.

Now that we have constructed a maximal linear subspace Y_- such that Q is negative on Y_-, we have to construct a linear subspace Y_+ such that Q is positive on Y_+ and $Y_- \perp^Q Y_+$. Recall that the operator $T \in \mathbb{L}(X^\star, X^{\star\star})$ is the unique symmetric operator that satisfies $\langle Tx, x\rangle = Q(x)$ for all $x \in X^\star$, see Lemma 2.1.26.

Lemma 4.3.7. Let X be a normed space that is separable or reflexive. Suppose that Q is a Legendre-$\star$ form on $X^\star$ and that T is injective. Then there exist closed linear subspaces $Y_-, Y_+ \subset X^\star$ with $X^\star = Y_- \dotplus Y_+$, $Y_- \perp^Q Y_+$ such that Q is positive on Y_+ and negative on Y_-.

Proof. We can use Lemma 4.3.6 to obtain a maximal linear subspace $Y_- \subset X^\star$ (with respect to set inclusion) with the property that Q is negative on Y_-. We define $Y_+ := \{x \in X^\star \mid x \perp^Q y \ \ \forall y \in Y_-\}$. We remark that Y_+ is a linear subspace and because Q is continuous, this linear subspace is also closed. It is also clear that $Y_- \perp^Q Y_+$ holds. Let $x \in Y_+ \cap Y_-$ be given. Then we can use $Y_- \perp^Q Y_+$ to obtain

$$4Q(x) = Q(x+x) = Q(x) + Q(x) = 2Q(x)$$

and therefore $Q(x) = 0$. Since $x \in Y_-$ and Q is negative on Y_- this implies $x = 0$. Thus, we have shown that $Y_+ \cap Y_- = \{0\}$ holds.

Next, we argue that $X^\star = Y_+ + Y_-$. Let $x \in X^\star$ be given. We recall that due to Lemma 4.3.5 the linear subspace Y_- has to be finite-dimensional, i.e. $n = \dim Y_-$ for some $n \in \mathbb{N}$. Thus, we can find a pairwise Q-orthogonal basis $\{y_i\}_{i=1}^n$ of Y_- with the property that $Q(y_i) = -1$ for all $i \in \{1, \dots, n\}$. We define $\hat{x} := x + \sum_{j=1}^n \langle Tx, y_j\rangle y_j$. Then for $i \in \{1, \dots, n\}$ we have $\langle T\hat{x}, y_i\rangle = \langle Tx, y_i\rangle + \langle Tx, y_i\rangle Q(y_i) = 0$ which is equivalent to $\hat{x} \perp^Q y_i$. Since $\{y_i\}_{i=1}^n$ is a basis of Y_-, it follows that $\hat{x} \perp^Q Y_-$. Thus, $\hat{x} \in Y_+$ and from the definition of $\hat{x}$ it can be seen that $x \in \hat{x} + Y_- \subset Y_- + Y_+$, completing the proof of $X^\star = Y_- \dotplus Y_+$.

It remains to show that Q is positive on Y_+. Let $y \in Y_+$ be given. We assume that $Q(y) < 0$. Then it follows from $y \perp^Q Y_-$ that Q is negative on the linear subspace $Y_- \dotplus \operatorname{lin}(y)$. This would be a contradiction to the maximality of Y_-. Thus, we know that Q is nonnegative on Y_+.

Now let $y \in Y_+$ be given such that $Q(y) = 0$. Then for $\alpha \in \mathbb{R}$ and $x \in X^\star$ we can use (2.11) to obtain

$$2\alpha\langle Ty, x\rangle_{X^{\star\star}\times X^\star} = Q(\alpha y + x) - Q(\alpha y) - Q(x) = Q(\alpha y + x) - Q(x).$$

Since we have shown $X^\star = Y_- + Y_+$ earlier, there are elements $x_- \in Y_-, x_+ \in Y_+$ such that $x = x_- + x_+$. Then we can conclude

$$2\alpha\langle Ty, x\rangle_{X^{\star\star}\times X^\star} = Q(\alpha y + x_- + x_+) - Q(x_- + x_+) = Q(\alpha y + x_+) - Q(x_+)$$

from the Q-orthogonality $x_- \perp^Q x_+$. Since $\alpha y + x_+ \in Y_+$, we know that $Q(\alpha y + x_+) \geq 0$ holds. Then the inequality

$$2\alpha\langle Ty, x\rangle_{X^{\star\star}\times X^\star} \geq -Q(x_+)$$

follows. However, since $\alpha \in \mathbb{R}$ was arbitrary, this is only possible if $\langle Ty, x\rangle_{X^{\star\star}\times X^\star} = 0$. Because $x \in X^\star$ was arbitrary, this means that $y \in \ker(T)$ and therefore $y = 0$ by assumption. Thus, Q is positive on Y_+.

We move to a proof of Hilbertizability under the assumption that T is injective. An analogous result for Legendre forms can be found in [Harder, 2018, Proposition 4.14]. However, our proof is very different.

Lemma 4.3.8. Let X be a normed space that is separable or reflexive. Suppose that Q is a Legendre-⋆ form on $X^\star$ and that T is injective. Then $X^\star$ is Hilbertizable.

Proof. Our first step is to show that we can without loss of generality assume that X is a Banach space. By Lemma 2.1.5 there is a Banach space $\hat{X}$ such that there is an isometric isomorphism $\iota \in \mathbb{L}(\hat{X}^\star, X^\star)$ and $\{\iota x_i\}_{i\in\mathbb{N}} \subset X^\star$ converges weakly-⋆ for all weakly-⋆ convergent sequences $\{x_i\}_{i\in\mathbb{N}} \subset \hat{X}^\star$. Then $\hat{Q} : \hat{X}^\star \to \mathbb{R}$ defined via $\hat{Q}(x) := Q(\iota x)$ for $x \in \hat{X}$ is a quadratic form. Since for all weakly-⋆ convergent sequences $\{x_i\}_{i\in\mathbb{N}} \subset \hat{X}^\star$ the image sequence $\{\iota x_i\}_{i\in\mathbb{N}}$ converges weakly-⋆, it follows that $\hat{Q}$ is a Legendre-⋆ form. It can also be seen that the unique symmetric operator $\hat{T} \in \mathbb{L}(\hat{X}^\star, \hat{X}^{\star\star})$ that satisfies $\hat{Q}(x) = \langle \hat{T}x, x\rangle$ for all $x \in \hat{X}$ is injective. Because the Hilbertizability of $\hat{X}^\star$ implies the Hilbertizability of $X^\star$ we can without loss of generality assume that X is a Banach space.

Using Lemma 4.3.7 yields a decomposition $X^\star = Y_- \dotplus Y_+$ with closed linear subspaces $Y_-, Y_+ \subset X^\star$ such that Q is negative on Y_-, positive on Y_+, and $Y_- \perp^Q Y_+$. Our goal is to apply Proposition 4.3.4 to the linear subspace Y_+. Thus, we have to show that Y_+ is sequentially weakly-⋆ closed.

Let $\{y_i\}_{i\in\mathbb{N}}$ be a sequence in Y_+ that converges weakly-⋆ to an element $x \in X^\star$. Due to the decomposition of $X^\star = Y_- \dotplus Y_+$ there are $x_- \in Y_-, x_+ \in Y_+$ such that $x = x_+ + x_-$. Let us assume that $x_- \neq 0$. Then we have $Q(x_-) < 0$. We define $a := \liminf_{i\to\infty} Q(y_i)$ and $\alpha := Q(x_-)^{-1}(a - Q(x_+))/2 - 1$. Since X is a Banach space, the sequence $\{y_i\}_{i\in\mathbb{N}}$ is bounded. Because Q is continuous and $\{Q(y_i)\}_{i\in\mathbb{N}}$ is bounded, we have $0 \leq a < \infty$. Using the sequential weak-⋆ lower semi-continuity of Q for the weak-⋆ convergence $y_i + \alpha x_- \overset{\star}{\rightharpoonup} x_+ + (1+\alpha)x_-$ results in

$$Q(x_+ + (1+\alpha)x_-) \leq \liminf_{i\to\infty} Q(y_i + \alpha x_-).$$

If we also use the Q-orthogonalities $y_i \perp^Q x_-$, $x_+ \perp^Q x_-$ we obtain

$$Q(x_+) + (1+\alpha)^2 Q(x_-) \leq \liminf_{i\to\infty} Q(y_i) + Q(\alpha x_-) = a + \alpha^2 Q(x_-).$$

Then the inequality

$$Q(x_+) - a \leq -(1+2\alpha)Q(x_-) = Q(x_+) - a + Q(x_-)$$

follows, which is a contradiction to the condition that $Q(x_-) < 0$. Hence, $x_- = 0$ and we have $y_i \overset{\star}{\rightharpoonup} x = x_+ \in Y_+$. Since $\{y_i\}_{i\in\mathbb{N}} \subset Y_+$ was an arbitrary weakly-$\star$ convergent sequence, we have shown that Y_+ is sequentially weakly-$\star$ closed.

Thus, we can apply Proposition 4.3.4 and obtain that Y_+ is Hilbertizable. It follows from Lemma 2.1.2 that $X^\star$ is Hilbertizable.

Now we are able to prove the main theorem of this chapter. It is a generalization of [Harder, 2018, Theorem 1.1] to nonreflexive Banach spaces with a separable predual space. The difference of this theorem to Lemma 4.3.8 is that we can drop the assumption that T is injective.

Theorem 4.3.9. Let X be a normed space that is separable or reflexive. Suppose that Q is a Legendre-$\star$ form on $X^\star$. Then $X^\star$ is Hilbertizable.

Proof. We follow the proof of [Harder, 2018, Theorem 1.1] but adapt it to our setting.

Since $Q(x) = 0$ for all $x \in \ker T$, it follows from Lemma 4.3.5 that $\ker T$ is a finite-dimensional linear subspace. By Lemma 2.1.3 (b) there exists a weakly-$\star$ closed linear subspace Y of $X^\star$ with $X^\star = Y \dotplus \ker T$.

Let us consider the restriction of Q to Y and the corresponding symmetric operator $T_Y : Y \to Y^\star$. We will show that $\ker T_Y = \{0\}$. Let $y \in \ker T_Y$ be given. Then for each $\hat{y} \in Y$, $x_0 \in \ker T$ we have

$$\begin{aligned} 0 = \langle T_Y y, \hat{y}\rangle_{Y^\star \times Y} &= \tfrac{1}{2}(Q(y+\hat{y}) - Q(y) - Q(\hat{y})) \\ &= \langle Ty, \hat{y}\rangle_{X^{\star\star}\times X^\star} + \langle Tx_0, y\rangle_{X^{\star\star}\times X^\star} = \langle Ty, \hat{y} + x_0\rangle_{X^{\star\star}\times X^\star}, \end{aligned}$$

where we used (2.11). Thus, $\langle Ty, x\rangle_{X^{\star\star}\times X^\star} = 0$ for all $x \in X^\star = Y \dotplus \ker T$ and therefore $y \in \ker T \cap Y = \{0\}$.

Since Y is weakly-$\star$ closed, Corollary 4.3.2 and Lemma 4.3.3 allow us to transfer the assumptions to the linear subspace Y. Thus, we can apply Lemma 4.3.8 for the linear subspace Y and the predual space of Y. It follows that Y is Hilbertizable. Finally, we can apply Lemma 2.1.2 to $X^\star = Y \dotplus \ker T$, which completes the proof.

An important implication of this theorem is that we should not try to apply results where a Legendre-$\star$ form appears in spaces that have a separable predual space but are not Hilbertizable. Likewise, we should not try to apply results where a Legendre form appears in reflexive spaces that are not Hilbertizable. Such spaces include ℓ^1, ℓ^∞, $L^\infty(\Omega)$ for the nonreflexive case as well as $L^p(\Omega)$ and ℓ^p for $p \in (1,\infty) \setminus \{2\}$ for the reflexive case.

There are still some cases where we do not know whether Legendre forms or Legendre-$\star$ forms can exist. For example it is open whether Legendre forms can exist in $L^1(\Omega)$. Another question is what happens to Legendre-$\star$ forms in nonreflexive spaces if we

remove the condition that the predual space is separable. In such a case it might be more reasonable to redefine Legendre-$\star$ forms using weakly-$\star$ converging nets instead of weakly-$\star$ converging sequences. It is an open question whether Theorem 4.3.9 can be generalized to this setting.

4.4 Counterexamples

In this section we provide two counterexamples. The first counterexample discusses the case of so-called extended Legendre forms. The second counterexample shows that Legendre forms can exist in nonreflexive Banach spaces. This section is mostly taken from [Harder, 2018, Section 5].

We start with defining extended Legendre forms. This concept was introduced in [Bonnans, Shapiro, 2000, Definition 3.73].

Definition 4.4.1. Let X be a normed space. An *extended Legendre form* is a continuous and sequentially weakly lower semi-continuous function $Q : X \to \mathbb{R}$ that has the properties $Q(\alpha x) = \alpha^2 Q(x)$ for all $x \in X$, $\alpha > 0$ and

$$(Q(x_i) \to Q(x) \wedge x_i \rightharpoonup x) \quad \Rightarrow \quad x_i \to x$$

for all sequences $\{x_i\}_{i\in\mathbb{N}}$ and points x in X.

Note that an extended Legendre form does not have to be a quadratic form. An application for extended Legendre forms that are not quadratic forms can be found in [Bonnans, Shapiro, 2000, Proposition 3.74]. The question arises whether a result similar to Theorem 4.3.9 holds for extended Legendre forms. The following example shows that this is not the case and that extended Legendre forms can exist on Banach spaces that are not Hilbertizable.

Example 4.4.2. Let $p \in (1, \infty) \setminus \{2\}$ be given. Then the function

$$Q : \ell^p \to \mathbb{R}, \quad x \mapsto \|x\|_{\ell^p}^2$$

is an extended Legendre form on a reflexive Banach space which is not Hilbertizable.

Proof. Clearly, Q is continuous, sequentially weakly lower semi-continuous, and has the property $Q(\alpha x) = \alpha^2 Q(x)$ for all $x \in \ell^p, \alpha \in \mathbb{R}$.

Let $\{x_i\}_{i\in\mathbb{N}}$ be a sequence and x an element in ℓ^p such that $x_i \rightharpoonup x \in \ell^p$ and $\|x_i\|_{\ell^p}^2 \to \|x\|_{\ell^p}^2$. Then according to [Brezis, 2011, Proposition 3.32] it follows that $x_i \to x$. Note that we can apply [Brezis, 2011, Proposition 3.32] because ℓ^p is a uniformly convex Banach space, see [Clarkson, 1936, Section 3].

It remains to argue that the space ℓ^p is not Hilbertizable. According to [Albiac, Kalton,

2016, Remark 6.2.11 (g)], for $p \in (1,2)$ the so-called type (or Rademacher type) of ℓ^p is at most $p < 2$, and for $p \in (2,\infty)$ the so-called cotype (or Rademacher cotype) of ℓ^p is at least $p > 2$. However, according to [Albiac, Kalton, 2016, Theorem 7.4.1] any Banach space that is Hilbertizable must have type 2 and cotype 2. Therefore, the Banach space ℓ^p is not Hilbertizable.

In fact this counterexample works in all uniformly convex spaces that are not Hilbertizable.

We continue with our second counterexample, which will show that we cannot replace Legendre-$\star$ forms with Legendre forms in Theorem 4.3.9. This will be done using the space ℓ^1.

Example 4.4.3. There exists a Legendre form on the Banach space ℓ^1, although ℓ^1 is not Hilbertizable.

Proof. The space ℓ^1 is not Hilbertizable because it is not reflexive. According to [Conway, 1985, Proposition V.5.2] every weakly convergent sequence in ℓ^1 converges in norm. As a consequence, every lower semi-continuous quadratic form on ℓ^1 is a Legendre form. It remains to show the existence of a lower semi-continuous quadratic form on ℓ^1. This is indeed the case, the simplest example being $Q(x) = 0$.

However, Legendre forms are rarely relevant in nonreflexive Banach spaces such as ℓ^1, because there can be bounded sequences without a weakly convergent subsequence. Thus, Legendre-$\star$ forms would be more interesting in ℓ^1 for applications since bounded sequences have weakly-$\star$ convergent subsequences. However, by Theorem 4.3.9 there exist no Legendre-$\star$ forms on ℓ^1.

5 Optimal control of the obstacle problem

5.1 Problem statement

In this section we will consider the optimal control of the obstacle problem, which is an instance of an infinite-dimensional bilevel optimization problem. The optimal control of the obstacle problem is well-known in the literature. While it is simple to state, this bilevel optimization problem provides some interesting challenges.

To keep our notation more in line with the typical notation of optimal control, we will often use different variable names than in the abstract setting of Chapter 3. A list of translations of mathematical objects and variable names from the setting in Chapter 3 to the setting in Chapter 5 can be found in Table 5.1.1 on page 162. We will also make use of the results on capacity theory from Section 2.6. Some of the more basic results of capacity theory will not always be referenced. As established in Section 2.6, all functions in $H_0^1(\Omega)$ that appear in this chapter are assumed to be quasi-continuous representatives. Recall that throughout this thesis Ω always denotes an open and bounded subset of $\mathbb{R}^d$.

Let us introduce the obstacle problem. For a parameter $u \in L^2(\Omega)$ this optimization problem is given by

$$\begin{aligned} \min_{y \in H_0^1(\Omega)} \quad & \int_\Omega \tfrac{1}{2}|\nabla y|^2 - yu \,\mathrm{d}\omega \\ \text{s.t.} \quad & y \geq y_a \text{ q.e. on } \Omega, \end{aligned} \tag{OP(u)}$$

where $y_a \in H_0^1(\Omega)$ is a fixed obstacle. We mention that the condition $y \geq y_a$ q.e. on Ω is equivalent to $y \geq y_a$ a.e. on Ω due to Lemma 2.6.12 (c). The optimal control of the obstacle problem is given by

$$\begin{aligned} \min_{y,u} \quad & F(y,u) \\ \text{s.t.} \quad & y \text{ solves (OP}(u)\text{)}, \\ & u \in U_{\mathrm{ad}}. \end{aligned} \tag{OCOP}$$

Here, $F : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is a continuously Fréchet differentiable function that is also sequentially weakly lower semi-continuous, and the admissible set $U_{\mathrm{ad}} \subset L^2(\Omega)$ is a closed, convex, and nonempty set. The problem (OCOP) constitutes a bilevel optimization problem, where (OP(u)) is the corresponding lower level optimization problem. Like

in Chapter 3 we will use the notation $\psi : L^2(\Omega) \to H_0^1(\Omega)$ for the solution operator and $\varphi : L^2(\Omega) \to \mathbb{R}$ for the optimal value function that correspond to the parametrized optimization problem (OP(u)).

We will give some examples for the components of (OCOP). For example, the constraint $u \in U_{\text{ad}}$ can be used to describe pointwise upper and lower bounds on the control u, i.e. the set $U_{\text{ad}} \subset L^2(\Omega)$ is given via

$$U_{\text{ad}} := \{u \in L^2(\Omega) \mid u_a \leq u \leq u_b \text{ a.e. in } \Omega\}, \tag{5.1}$$

where $u_a, u_b : \Omega \to \mathbb{R} \cup \{\infty, -\infty\}$ are measurable functions that act as control constraint and are chosen such that U_{ad} is nonempty. As an example for the upper level objective function F we consider

$$F(y,u) := \tfrac{1}{2}\|\mathcal{I}_{H_0^1(\Omega)\to L^2(\Omega)} y - y_d\|^2_{L^2(\Omega)} + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)}. \tag{5.2}$$

Here the weight $\sigma \geq 0$ and the desired state $y_d \in L^2(\Omega)$ are fixed. This is a special case of the objective function used in [G. Wachsmuth, 2014].

There is a physical interpretation of the obstacle problem for y and u with small norms. Suppose that $d = 2$, that $\Omega \subset \mathbb{R}^2$ is a thin elastic membrane, and that an external force $u \in L^2(\Omega)$ causes a vertical deflection y of the membrane. Moreover, the membrane is fixed at the boundary $\partial\Omega$ so that $y = 0$ on $\partial\Omega$ holds, and there is a lower obstacle y_a that restricts some deflections of the membrane. Then the vertical deflection y can be modeled as the solution of the problem (OP(u)), where the objective function describes the elastic energy under the external force u. For the bilevel optimization problem (OCOP), suppose that we have a desired state $y_d \in L^2(\Omega)$ for the deflection of the membrane. Then we want to find a force $u \in L^2(\Omega)$ such that the resulting deflection y of the membrane is as close as possible to y_d. If we add control constraints $u \in U_{\text{ad}}$ and a quadratic cost term $\frac{\sigma}{2}\|u\|^2_{L^2(\Omega)}$, this leads to the problem (OCOP) with F chosen as in (5.2).

The optimal control of the obstacle problem is frequently studied in the literature, and its stationarity conditions are frequently an object of interest, see, for example, [Mignot, 1976; Barbu, 1984; Jarušek, Outrata, 2007; Hintermüller, Kopacka, 2009; Hintermüller, Surowiec, 2011; Outrata, Jarušek, Stará, 2011; Schiela, D. Wachsmuth, 2013; Hintermüller, Mordukhovich, Surowiec, 2014; G. Wachsmuth, 2014; 2016; Harder, G. Wachsmuth, 2018a; c; G. Wachsmuth, 2018]. The problem (OCOP) is also sometimes described via its (equivalent) MPCC reformulation

$$\begin{aligned}
\min_{y,u,\lambda} \quad & F(y,u) \\
\text{s.t.} \quad & -\Delta y - u + \lambda = 0, \\
& y - y_a \geq 0 \text{ q.e. on } \Omega, \\
& \langle \lambda, y - y_a \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0, \\
& \lambda \in H^{-1}(\Omega)_-, \\
& u \in U_{\text{ad}},
\end{aligned} \tag{5.3}$$

or by reformulating the obstacle problem as a variational inequality. In general, a lot of the difficulty is caused by the complementarity-type condition in (5.3). Since this complementarity-type condition lives in the Sobolev space $H_0^1(\Omega)$ and its dual space $H^{-1}(\Omega)$, the problem (OCOP) might also be considered as a good representative of other problems which feature complementarity-type conditions of Sobolev space character. Despite the difficulties, the problem still allows for some pointwise interpretation in many places. Here, our study of capacity theory in Section 2.6 can be useful for obtaining pointwise quasi-everywhere conditions.

Let us summarize the assumptions for the optimal control of the obstacle problem.

Assumption 5.1.1. (a) The obstacle y_a belongs to $H_0^1(\Omega)$.

(b) The function $F : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is continuously Fréchet differentiable and sequentially weakly lower semi-continuous.

(c) The set $U_{\mathrm{ad}} \subset L^2(\Omega)$ is a closed, convex, and nonempty set.

For part (a) of Assumption 5.1.1 we remark that often only $y_a \in H^1(\Omega)$ and $\max(y_a, 0) \in H_0^1(\Omega)$ is required for the obstacle in the literature. However, we use $y_a \in H_0^1(\Omega)$ in order to simplify the application of the abstract theory from Section 3.4. There are also other possibilities to consider a more general setting, e.g. another energy functional could be used as an objective function of (OP(u)), but then the notation would become more complicated.

Let us emphasize that Assumption 5.1.1 (b) is satisfied if F is given via (5.2) and that Assumption 5.1.1 (c) is satisfied if U_{ad} is given via (5.1).

We remark that Assumption 5.1.1 is not sufficient to guarantee the existence of solutions of (OCOP), see also Corollary 5.2.2.

The setting in this chapter fits into the abstract setting that was described in Section 3.3. We provide Table 5.1.1 to list how the spaces, functions, sets, and variables that appeared in the abstract setting of Sections 3.3 and 3.4 are used in the current setting of the bilevel optimization problem (OCOP). This also includes variable names for Lagrange multipliers that we use for stationarity conditions in Section 5.2. We mention that the notation mostly coincides with the one used in [Harder, G. Wachsmuth, 2018a].

5.2 Stationarity conditions

We are interested in stationarity conditions for local minimizers of (OCOP). We will show C-stationarity of local minimizers under Assumption 5.1.1 in Theorem 5.2.8. As a preparation, we provide some preliminary observations in Section 5.2.1, and formally derive the system of stationarity conditions for (OCOP) in Section 5.2.2.

Chapter 3	Chapter 5
X	$H_0^1(\Omega)$
Y	$H_0^1(\Omega)$
V	$L^2(\Omega)$
x	y
p	u
$g(x,p)$	$g(y,u) := y - y_a$
$\Phi \subset Y$	$H_0^1(\Omega)_+$
Φ_{UL}	U_{ad}
$f(x,p)$	$f(y,u) := \int_\Omega \frac{1}{2}\lvert\nabla y\rvert^2 - yu \,\mathrm{d}\omega$
$F(x,p)$	$F(y,u)$
(LL(p))	(OP(u))
(UL)	(OCOP)
$\lambda \in Y^\star$	$\lambda \in H^{-1}(\Omega)$
$w \in X$	$p \in H_0^1(\Omega)$
$\xi \in Y^\star$	$\nu \in H^{-1}(\Omega)$
$\rho \in V^\star$	$\mu \in L^2(\Omega)$

Table 5.1.1: Translation of variables and objects.

5.2.1 Preliminary observations

We give some preliminary observations. We are interested in the existence and continuity of the solution operator ψ and the optimal value function φ. Let us first consider a variation of the obstacle problem (OP(u)). For a parameter $\xi \in H^{-1}(\Omega)$ this modified obstacle problem is given by

$$\begin{aligned} \min_{y \in H_0^1(\Omega)} \quad & \tfrac{1}{2}\int_\Omega \lvert\nabla y\rvert^2 \,\mathrm{d}\omega - \langle \xi, y\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \\ \text{s.t.} \quad & y \in \{v \in H_0^1(\Omega) \mid v \geq y_a \text{ q.e. on } \Omega\}. \end{aligned} \tag{$\widehat{\text{OP}}(\xi)$}$$

We denote the corresponding solution operator by $\hat{\psi}(\xi)$. Note that this modified obstacle problem is almost the same as (OP(u)), but the space for the parameter has changed. In particular, for each $u \in L^2(\Omega)$ the problems (OP(u)) and ($\widehat{\text{OP}}(\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)}u)$) are equivalent. As can be seen in the next lemma, this modified obstacle problem will be helpful for obtaining nicer properties for the solution operator ψ.

Lemma 5.2.1. Suppose Assumption 5.1.1 is satisfied. Then the following holds.

(a) The problem ($\widehat{\text{OP}}(\xi)$) has a unique solution for each $\xi \in H^{-1}(\Omega)$, i.e. the solution operator $\hat{\psi} : H^{-1}(\Omega) \to H_0^1(\Omega)$ exists. Additionally, $\hat{\psi}$ is globally Lipschitz continuous.

(b) The solution operator $\psi : L^2(\Omega) \to H_0^1(\Omega)$ exists and is globally Lipschitz continuous. Moreover, ψ is weak-to-strong continuous, i.e. $\psi(u_i) \to \psi(u_0)$ holds for all weakly convergent sequences $\{u_i\}_{i\in\mathbb{N}} \subset L^2(\Omega)$ with weak limit $u_0 \in L^2(\Omega)$.

(c) The optimal value function φ has values in $\mathbb{R}$ and is continuous. Moreover, the optimal value function φ is concave.

Proof. For part (a) we will apply Lemma 3.1.6 to the parametrized optimization problem $(\widehat{\mathrm{OP}}(\xi))$. Clearly, the objective function is strongly convex with parameter $\gamma = 1$. The other assumptions in Lemma 3.1.6 are also satisfied. Since the function $(y, \xi) \mapsto \frac{1}{2}\int_\Omega |\nabla y|^2 \,\mathrm{d}\omega - \langle \xi, y\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$ is a quadratic and continuous function, we obtain that $\hat{\psi} : H^{-1}(\Omega) \to H_0^1(\Omega)$ exists and that it is a globally Lipschitz continuous function from Lemma 3.1.6.

Let us continue with part (b). From part (a) and from the equivalence of the problems $(\mathrm{OP}(u))$ and $(\widehat{\mathrm{OP}}(\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)}u))$ we obtain the existence of $\psi(u)$ and the relation $\psi(u) = \hat{\psi}(\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)}u)$ for all $u \in L^2(\Omega)$. Then the weak-to-strong continuity of the solution operator ψ follows from the compactness of the embedding operator $\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)}$.

Finally, the continuity of the optimal value function φ follows directly from part (b). Since the function $u \mapsto f(y, u)$ is affine and therefore concave for all $y \in H_0^1(\Omega)$, the concavity of φ follows from Proposition 3.2.5.

As a corollary of Lemma 5.2.1 (b) we obtain the existence of minimizers of (OCOP). However, this result does not play a crucial role for our future results on stationarity conditions.

Corollary 5.2.2. Let Assumption 5.1.1 be satisfied and suppose that $U_{\mathrm{ad}} \subset L^2(\Omega)$ is a bounded set. Then there exists a (global) minimizer of (OCOP).

Proof. It follows from Lemma 5.2.1 (b) and Assumption 5.1.1 (b) that the reduced upper level objective function

$$\hat{F} : L^2(\Omega) \to \mathbb{R}, \qquad u \mapsto F(\psi(u), u)$$

is sequentially weakly lower semi-continuous. Therefore, by Lemma 2.3.2 (d) there exists a minimizer of the problem

$$\begin{aligned} \min_u \quad & \hat{F}(u) \\ \text{s.t.} \quad & u \in U_{\mathrm{ad}}. \end{aligned}$$

Because this problem is equivalent to (OCOP), the claim follows.

For the stationarity results in the abstract setting in Section 3.4 we relied on Assumption 3.4.5. Thus, in order to apply the results from Section 3.4 we need to show that these assumptions are satisfied in the current setting.

Lemma 5.2.3. Suppose Assumption 5.1.1 is satisfied and that $(\bar{y}, \bar{u})$ is a local minimizer of (OCOP). Then the list of assumptions in Assumption 3.4.5 is satisfied for the problem (OCOP).

Proof. Most of the assumptions listed in Assumption 3.4.5 follow directly from Assumption 5.1.1 or Table 5.1.1 on page 162.

The only difficult part is Assumption 3.4.5 (j). We will show that the function $(y, u) \mapsto f(y, u)$ and $u \mapsto -\varphi(u)$ are sequentially weakly lower semi-continuous. Let $\{y_i\}_{i\in\mathbb{N}} \subset H_0^1(\Omega)$ and $\{u_i\}_{i\in\mathbb{N}} \subset L^2(\Omega)$ be weakly convergent sequences with weak limits $y_0 \in H_0^1(\Omega)$, $u_0 \in L^2(\Omega)$. Because the canonical embedding $\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} : L^2(\Omega) \to H^{-1}(\Omega)$ is a compact operator we have

$$\begin{aligned} f(y_0, u_0) &= \tfrac{1}{2}\|y_0\|^2_{H_0^1(\Omega)} - \langle \mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} u_0, y_0\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \\ &\le \liminf_{i\to\infty} \big(\tfrac{1}{2}\|y_i\|^2_{H_0^1(\Omega)} - \langle \mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} u_i, y_i\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}\big) \\ &= \liminf_{i\to\infty} f(y_i, u_i) \end{aligned}$$

and thus f is sequentially weakly lower semi-continuous. From Lemma 5.2.1 (b) we obtain the convergence $\psi(u_i) \to \psi(u_0)$. Then

$$\begin{aligned} -\varphi(u_0) &= -f(\psi(u_0), u_0) \\ &= -\tfrac{1}{2}\|\psi(u_0)\|^2_{H_0^1(\Omega)} + \langle \mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} u_0, \psi(u_0)\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \\ &= \lim_{i\to\infty} \big(-\tfrac{1}{2}\|\psi(u_i)\|^2_{H_0^1(\Omega)} + \langle \mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} u_i, \psi(u_i)\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}\big) \\ &= \lim_{i\to\infty} \big(-f(\psi(u_i), u_i)\big) = \lim_{i\to\infty} \big(-\varphi(u_i)\big) \end{aligned}$$

follows and thus $u \mapsto -\varphi(u)$ is sequentially weakly lower semi-continuous. This shows that Assumption 3.4.5 (j) holds.

We mention that the sequential weak lower semi-continuity of the function $(y, u) \mapsto f(y, u) - \varphi(u)$ can also be found in [Harder, 2016, Lemma 5.12].

5.2.2 Formal derivation of stationarity conditions

In order to see what stationarity conditions would be desirable, we discuss the formal derivation of stationarity conditions. Since this was already done for the abstract setting in Section 3.3.3 we only need to translate the stationarity conditions for our setting using Table 5.1.1 on page 162. Additionally, we will give possibilities for the as of yet unspecified set-valued mapping $\mathcal{N}^{\sharp}_{\operatorname{gph}\mathcal{N}_\Phi}$ that appeared in (3.23f).

We recall that the normal cone to the set $\Phi = H_0^1(\Omega)_+$ at a point $v \in H_0^1(\Omega)_+$ was already calculated in Proposition 2.6.26 (a) and can be expressed by

$$\mathcal{N}_{H_0^1(\Omega)_+}(v) = \{\xi \in H^{-1}(\Omega)_- \mid v = 0 \text{ q.e. on } \text{q-supp}(\xi)\}.$$

Let us translate the system (3.23) to our current setting while using the above expression for the normal cone to $H_0^1(\Omega)_+$. This results in the system

$$F_y'(\bar{y},\bar{u}) - \Delta\bar{p} + \bar{\nu} = 0, \tag{5.4a}$$
$$F_u'(\bar{y},\bar{u}) - \bar{p} + \bar{\mu} = 0, \tag{5.4b}$$
$$-\Delta\bar{y} - \bar{u} + \bar{\lambda} = 0, \tag{5.4c}$$
$$\bar{y} - y_a \geq 0 \text{ q.e. on } \Omega, \tag{5.4d}$$
$$\bar{\lambda} \in H^{-1}(\Omega)_-, \quad \bar{y} - y_a = 0 \text{ q.e. on } \text{q-supp}(\bar{\lambda}), \tag{5.4e}$$
$$(\bar{u},\bar{\mu}) \in \operatorname{gph}\mathcal{N}_{U_{\text{ad}}}, \tag{5.4f}$$

and

$$((\bar{y} - y_a, \bar{\lambda}), (\bar{\nu}, -\bar{p})) \in \operatorname{gph}\mathcal{N}_{\mathbb{K}}^{\sharp}, \tag{5.5}$$

where $\bar{y},\bar{p} \in H_0^1(\Omega)$, $\bar{u},\bar{\mu} \in L^2(\Omega)$, $\bar{\lambda},\bar{\nu} \in H^{-1}(\Omega)$ are variables or multipliers and $\mathcal{N}_{\mathbb{K}}^{\sharp} : H_0^1(\Omega) \times H^{-1}(\Omega) \to \mathcal{P}(H^{-1}(\Omega) \times H_0^1(\Omega))$ is an unspecified set-valued mapping that can be interpreted as a generalization of a normal cone to the nonconvex set $\mathbb{K}$ which is defined via

$$\mathbb{K} := \operatorname{gph}\mathcal{N}_{H_0^1(\Omega)_+} = \{(\hat{y},\hat{\lambda}) \in H_0^1(\Omega)_+ \times H^{-1}(\Omega)_- \mid \langle\hat{\lambda},\hat{y}\rangle = 0\}.$$

In the following definition we define various stationarity conditions for (OCOP) and implicitly describe some possible choices for $\mathcal{N}_{\mathbb{K}}^{\sharp}$. These definitions are the same as in [Harder, G. Wachsmuth, 2018a].

Definition 5.2.4. Let $(\bar{y},\bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ be given. We say that $(\bar{y},\bar{u})$ is *weakly stationary* or *W-stationary* for (OCOP) if there exist multipliers $\bar{p} \in H_0^1(\Omega)$, $\bar{\mu} \in L^2(\Omega)$, $\bar{\lambda},\bar{\nu} \in H^{-1}(\Omega)$ such that (5.4) and the conditions

$$\bar{\nu} \in \{v \in H_0^1(\Omega) \mid v = 0 \text{ q.e. on } \{\bar{y} = y_a\}\}^\circ, \tag{5.6a}$$
$$-\bar{p} = 0 \text{ q.e. on } \text{q-supp}(\bar{\lambda}) \tag{5.6b}$$

are satisfied. We call the point $(\bar{y},\bar{u})$ *C-stationary* if it is W-stationary and additionally the condition

$$\langle\bar{\nu}, w\bar{p}\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \geq 0 \tag{5.7}$$

holds for all $w \in W^{1,\infty}(\Omega)_+$. The point $(\bar{y},\bar{u})$ is called *M-stationary* if there exist multipliers $\bar{p} \in H_0^1(\Omega)$, $\bar{\mu} \in L^2(\Omega)$, $\bar{\lambda},\bar{\nu} \in H^{-1}(\Omega)$ and quasi-closed sets $A, B \subset \Omega$ such that (5.4) and the conditions

$$\text{q-supp}(\bar{\lambda}) \subset_q A \subset_q B \subset_q \{\bar{y} = y_a\}, \tag{5.8a}$$

$$\bar{\nu} \in \{v \in H_0^1(\Omega) \mid v \geq 0 \text{ q.e. on } B,\ v = 0 \text{ q.e. on } A\}^\circ, \tag{5.8b}$$
$$-\bar{p} \in \{v \in H_0^1(\Omega) \mid v \geq 0 \text{ q.e. on } B,\ v = 0 \text{ q.e. on } A\} \tag{5.8c}$$

are satisfied. Finally, for *strong stationarity* or *S-stationarity* of the point $(\bar{y}, \bar{u})$ we require the existence of multipliers $\bar{p} \in H_0^1(\Omega)$, $\bar{\mu} \in L^2(\Omega)$, $\bar{\lambda}, \bar{\nu} \in H^{-1}(\Omega)$ such that (5.4) and the conditions

$$\bar{\nu} \in \{v \in H_0^1(\Omega) \mid v \geq 0 \text{ q.e. on } \{\bar{y} = y_a\},\ v = 0 \text{ q.e. on q-supp}(\bar{\lambda})\}^\circ, \tag{5.9a}$$
$$-\bar{p} \in \{v \in H_0^1(\Omega) \mid v \geq 0 \text{ q.e. on } \{\bar{y} = y_a\},\ v = 0 \text{ q.e. on q-supp}(\bar{\lambda})\} \tag{5.9b}$$

hold.

Some of the conditions for W-, C-, M-, or S-stationarity can be interpreted as pointwise q.e. analogues of the same concepts in finite dimensions, see Definition 2.5.1 and (2.29). However, for some functionals in $H^{-1}(\Omega)$ a pointwise evaluation is not possible and therefore we use polar cones or multiplication by a function in $W^{1,\infty}(\Omega)_+$.

We remark that if Ω is sufficiently regular (e.g. if it satisfies the cone condition), it follows from the Sobolev embedding theorem that the multiplication operator $w : H_0^1(\Omega) \to H_0^1(\Omega)$ is well-defined and bounded for any function $w \in W^{1,q}(\Omega)$ with $q = \max(2, d + \varepsilon)$ for some $\varepsilon > 0$. Since $W^{1,\infty}(\Omega)_+$ is dense in $W^{1,q}(\Omega)_+$ in this case, it can be shown that the C-stationarity condition (5.7) is equivalent to

$$\langle \bar{\nu}, w\bar{p} \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \geq 0 \qquad \forall w \in W^{1,q}(\Omega)_+,$$

i.e. we can allow a wider class of functions for w.

As for the relationship between these stationarity concepts, the hierarchy

$$\text{S-stationary} \quad \Rightarrow \quad \text{M-stationary} \quad \Rightarrow \quad \text{C-stationary} \quad \Rightarrow \quad \text{W-stationary}$$

can be shown. In the case that the so-called biactive set $\{\bar{y} = y_a\} \setminus \text{q-supp}(\bar{\lambda})$ has zero capacity, these four notions of stationarity coincide. Since this set often has zero capacity, it can already be useful to show weak stationarity for local minimizers. However, this biactive set can also be large in some instances, and the example from [G. Wachsmuth, 2014, Section 6.1] shows that strong stationarity does not need to be satisfied for local minimizers even if the problem data is nice.

Note that the additional conditions for W-, C-, M-, or S-stationarity that supplement (5.4) are equivalent to (5.5) if we choose of $\mathcal{N}_{\mathbb{K}}^\sharp$ appropriately for the respective stationarity system. For example, if we define

$$\begin{aligned} \mathcal{N}_{\mathbb{K}}^{\text{weak}}(\hat{y}, \hat{\lambda}) := \{z \in H_0^1(\Omega) \mid z = 0 \text{ q.e. on } \{\hat{y} = 0\}\}^\circ \\ \times \{w \in H_0^1(\Omega) \mid w = 0 \text{ q.e. on q-supp}(\hat{\lambda})\} \end{aligned} \tag{5.10}$$

for a given pair $(\hat{y}, \hat{\lambda}) \in \mathbb{K} \subset H_0^1(\Omega)_+ \times H^{-1}(\Omega)_-$ then (5.6) is equivalent to (5.5) if we use $\mathcal{N}_{\mathbb{K}}^\sharp = \mathcal{N}_{\mathbb{K}}^{\text{weak}}$. Similar cones can be defined for the other stationarity concepts. We comment on some of these cones and their relationship to the cones defined in Section 2.4.

Remark 5.2.5. Let $(\hat{y}, \hat{\lambda}) \in \mathbb{K} \subset H_0^1(\Omega)_+ \times H^{-1}(\Omega)_-$ be given.

(a) We have the equality

$$\mathcal{N}_{\mathbb{K}}^{\text{weak}}(\hat{y}, \hat{\lambda}) = \mathcal{N}_{\mathbb{K}}^{\text{Clarke}}(\hat{y}, \hat{\lambda}).$$

In particular, (5.6) is equivalent to (5.5) if we use $\mathcal{N}_{\mathbb{K}}^{\sharp} = \mathcal{N}_{\mathbb{K}}^{\text{Clarke}}$.

(b) If we use $\mathcal{N}_{\mathbb{K}}^{\sharp} = \mathcal{N}^{\text{s-lim}}$, then (5.8) is equivalent to (5.5), i.e. the strong limiting normal cone to the set $\mathbb{K}$ at $(\hat{y}, \hat{\lambda})$ has the representation

$$\mathcal{N}_{\mathbb{K}}^{\text{s-lim}}(\hat{y}, \hat{\lambda}) = \bigcup_{\substack{\text{q-supp}(\hat{\lambda}) \subset A \subset B \subset \{\hat{y}=0\} \\ A,B \subset \Omega \text{ quasi-closed}}} \hat{\mathcal{K}}(A, B)^\circ \times \hat{\mathcal{K}}(A, B),$$

where $\hat{\mathcal{K}}(A, B) := \{v \in H_0^1(\Omega) \mid v \geq 0 \text{ q.e. on } B,\ v = 0 \text{ q.e. on } A\}$.

(c) If we use $\mathcal{N}_{\mathbb{K}}^{\sharp} = \mathcal{N}^{\text{Fréchet}}$, then (5.9) is equivalent to (5.5).

(d) The Fréchet normal cone to the set $\mathbb{K}$ at $(\hat{y}, \hat{\lambda})$ can be described by

$$\mathcal{N}_{\mathbb{K}}^{\text{Fréchet}}(\hat{y}, \hat{\lambda}) = \mathcal{K}_{H_0^1(\Omega)_+}(\hat{y}, \hat{\lambda})^\circ \times \mathcal{K}_{H_0^1(\Omega)_+}(\hat{y}, \hat{\lambda}). \tag{5.11}$$

For a proof of the claims in parts (a) and (b), see [Harder, G. Wachsmuth, 2018a, Theorem 5.2, Section 5.4]. The claims in parts (c) and (d) follow from a combination of Proposition 2.6.26 (c) and [Harder, G. Wachsmuth, 2018a, (30)]. While the limiting normal cone $\mathcal{N}_{\mathbb{K}}^{\text{lim}}$ does not appear in Remark 5.2.5, it will be the subject of investigation in Section 5.3. Some of the claims in Remark 5.2.5 have analogues in the context of finite-dimensional MPCCs, see Remark 2.5.2.

For more details on the various stationarity systems and comparisons, we refer to [Harder, G. Wachsmuth, 2018a].

5.2.3 C-stationarity for local minimizers

We want to show that C-stationarity holds at a local minimizer by applying the abstract theory from Section 3.4. One concept that is used in the abstract setting is the so-called normal-cone-preserving operator, see Definition 3.4.10. Thus, it is of interest to get to know some normal-cone-preserving operators in the current setting.

Lemma 5.2.6. Let $w \in W^{1,\infty}(\Omega)_+$ be given. Then the multiplication operator

$$w : H_0^1(\Omega) \to H_0^1(\Omega), \qquad v \mapsto wv$$

is a normal-cone-preserving operator with respect to the set $H_0^1(\Omega)_+ \subset H_0^1(\Omega)$.

Proof. Let $y \in H_0^1(\Omega)_+$ and $\lambda \in \mathcal{N}_{H_0^1(\Omega)_+}(y)$ be given. Then by Proposition 2.6.26 (a) we obtain $\lambda \in H^{-1}(\Omega)_-$ and q-supp$(\lambda) \subset_q \{y = 0\}$. Since $|wy| \leq \|w\|_{L^\infty(\Omega)}|y|$ holds a.e. in Ω, it also holds q.e. in Ω, see Lemma 2.6.12. It follows that $wy = 0$ holds q.e. on $\{y = 0\}$, which implies q-supp$(\lambda) \subset_q \{wy = 0\}$.

Let $\hat{y} \in H_0^1(\Omega)_+$ be arbitrary. Then we have $w\hat{y} \in H_0^1(\Omega)_+$ and therefore $w\hat{y} \geq 0$ q.e. on Ω. Thus, $w\hat{y} - wy \geq 0$ holds q.e. on $\{wy = 0\}$ and therefore also q.e. on q-supp(λ). By Corollary 2.6.25 (a) we obtain

$$0 \geq \langle \lambda, w\hat{y} - wy \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle w^\star \lambda, \hat{y} - y \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)},$$

where $w^\star : H^{-1}(\Omega) \to H^{-1}(\Omega)$ denotes the adjoint of the multiplication operator $v \mapsto wv$. Thus, we have $w^\star\lambda \in \mathcal{N}_{H_0^1(\Omega)_+}(y)$, which completes the proof that w is a normal-cone-preserving operator with respect to $H_0^1(\Omega)_+$.

In preparation for our result for C-stationarity we provide another technical lemma. This result can also be found in [Harder, 2016, Lemma 5.9] or in the proof of [G. Wachsmuth, 2016, Lemma 4.5].

Lemma 5.2.7. Let $w \in W^{1,\infty}(\Omega)_+$ be given. Then the map

$$q : H_0^1(\Omega) \to \mathbb{R}, \qquad v \mapsto \langle -\Delta v, wv \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$$

is sequentially weakly lower semi-continuous.

Proof. We follow the proof of [Harder, 2016, Lemma 5.9]. Let $\{v_i\}_{i\in\mathbb{N}}$ be a sequence in $H_0^1(\Omega)$ such that $v_i \rightharpoonup v_0$ for some $v_0 \in H_0^1(\Omega)$. Then

$$\begin{aligned} q(v_i) &= \langle -\Delta v_i, wv_i \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \\ &= \int_\Omega \nabla v_i^\top \nabla(wv_i)\,\mathrm{d}\omega \\ &= \int_\Omega v_i \nabla v_i^\top \nabla w\,\mathrm{d}\omega + \int_\Omega w \nabla v_i^\top \nabla v_i\,\mathrm{d}\omega \end{aligned}$$

holds for all $i \in \mathbb{N} \cup \{0\}$. For the first integral we make use of the convergences $v_i \to v_0$ and $\nabla v_i^\top \nabla w \rightharpoonup \nabla v_0^\top \nabla w$ in $L^2(\Omega)$. This yields

$$\begin{aligned} \int_\Omega v_0 \nabla v_0^\top \nabla w\,\mathrm{d}\omega &= \langle \nabla v_0^\top \nabla w, v_0 \rangle_{L^2(\Omega)\times L^2(\Omega)} \\ &= \lim_{i\to\infty} \langle \nabla v_i^\top \nabla w, v_i \rangle_{L^2(\Omega)\times L^2(\Omega)} \\ &= \lim_{i\to\infty} \int_\Omega v_i \nabla v_i^\top \nabla w\,\mathrm{d}\omega. \end{aligned}$$

Let us address the second integral. Because w is nonnegative, the function $v \mapsto$

$\int_\Omega w\nabla v^\top \nabla v \,\mathrm{d}\omega$ is convex, continuous, and therefore sequentially weakly lower semi-continuous as a function from $H_0^1(\Omega)$ to $\mathbb{R}$. If we combine all of the above, we get

$$\begin{aligned} q(v_0) &= \int_\Omega v_0 \nabla v_0^\top \nabla w \,\mathrm{d}\omega + \int_\Omega w \nabla v_0^\top \nabla v_0 \,\mathrm{d}\omega \\ &\leq \lim_{i\to\infty} \int_\Omega v_i \nabla v_i^\top \nabla w \,\mathrm{d}\omega + \liminf_{i\to\infty} \int_\Omega w \nabla v_i^\top \nabla v_i \,\mathrm{d}\omega \\ &= \liminf_{i\to\infty} q(v_i) \end{aligned}$$

which concludes the proof.

We can proceed with our main result in this section. We show that a local minimizer $(\bar{y}, \bar{u})$ of (OCOP) is C-stationary by applying the abstract theory from Section 3.4. The result can also be found in [G. Wachsmuth, 2016, Section 4] or in [Harder, 2016, Theorem 5.15].

Theorem 5.2.8. Let Assumption 5.1.1 be satisfied and let $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ be a local minimizer of (OCOP). Then $(\bar{y}, \bar{u})$ is C-stationary.

Proof. We want to apply Theorem 3.4.17. The required Assumption 3.4.5 is satisfied due to Lemma 5.2.3. Using Proposition 2.6.26 (a) and Table 5.1.1 on page 162, it can be seen that the abstract system (3.32) translates to (5.4) in our setting. Thus, by Theorem 3.4.17 there exist multipliers $\bar{p} \in H_0^1(\Omega)$, $\bar{\mu} \in L^2(\Omega)$, $\bar{\lambda}, \bar{\nu} \in H^{-1}(\Omega)$ such that the system (5.4) is satisfied.

Next, we will show (5.6a). Recall that $H_0^1(\Omega)_+$ induces a lattice structure on $H_0^1(\Omega)$. Thus, by Theorem 3.4.17 (c) the condition $\langle \bar{\nu}, y\rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$ holds for all $y \in Y$ with $0 \leq y \leq \bar{y} - y_a$ q.e. on Ω. By continuity and linearity of $\bar{\nu}$ the condition $\bar{\nu} \in Y_{\bar{y}}^\circ$ follows, where $Y_{\bar{y}}$ is defined via

$$Y_{\bar{y}} := \mathrm{cl}\big(\mathrm{lin}\{y \in H_0^1(\Omega) \mid 0 \leq y \leq \bar{y} - y_a \text{ q.e. on } \Omega\}\big).$$

Note that $Y_{\bar{y}}$ is a closed lattice ideal, see Lemma 2.1.23 (c). Thus, by Theorem 2.6.21 there exists a quasi-closed set $A \subset \Omega$ such that

$$Y_{\bar{y}} = \{v \in H_0^1(\Omega) \mid v = 0 \text{ q.e. on } A\} \tag{5.12}$$

holds. Since $\bar{y} - y_a \in Y_{\bar{y}}$ we obtain $A \subset_q \{\bar{y} - y_a = 0\}$. By Lemma 2.6.17 there exists a function $\hat{v} \in H_0^1(\Omega)$ such that $A =_q \{\hat{v} = 0\}$ holds. Then we have $\hat{v} \in Y_{\bar{y}}$. Due to the definition of $Y_{\bar{y}}$, the condition $v = 0$ q.e. on $\{\bar{y} - y_a = 0\}$ holds for all $v \in Y_{\bar{y}}$. Therefore, $\{\bar{y} - y_a = 0\} \subset_q \{\hat{v} = 0\} =_q A$ holds. Combined with the above this yields $A =_q \{\bar{y} - y_a = 0\} = \{\bar{y} = y_a\}$. Then (5.6a) follows from (5.12) and $\bar{\nu} \in Y_{\bar{y}}^\circ$.

Let us continue with (5.6b). We will apply Theorem 3.4.17 (a). For the normal-cone-preserving operator T we choose the normal-cone-preserving operator given in

Lemma 5.2.6 where $w \in W^{1,\infty}(\Omega)_+$ is arbitrary. An application of Theorem 3.4.17 (a) yields

$$\langle \bar{\lambda}, w\bar{p} \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0.$$

In particular, this condition also holds for all $w \in C_c^\infty(\Omega)_+$. Then (5.6b) follows from Lemma 2.6.28.

For C-stationarity it remains to show that (5.7) holds. We will do this using Theorem 3.4.17 (e). Again, for the normal-cone-preserving operator T we choose the normal-cone-preserving operator given in Lemma 5.2.6 where $w \in W^{1,\infty}(\Omega)_+$ is arbitrary. We make the same choice for the operator T_2 and set $T_1 = T_3 = 0$. With these choices for the bounded linear operators T, T_1, T_2, T_3, the only nontrivial assumption in Theorem 3.4.17 (e) is the sequential weak lower semi-continuity of the map $v \mapsto \langle f''_{yy}(\bar{y}, \bar{u})v, wv \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$. However, this was already shown in Lemma 5.2.7. Thus, Theorem 3.4.17 (e) yields (5.7).

5.3 The limiting normal cone to a complementarity set in Sobolev spaces

In this section we will take a look at the limiting normal cone $\mathcal{N}_{\mathbb{K}}^{\lim}$ to the nonconvex set $\mathbb{K} = \operatorname{gph} \mathcal{N}_{H_0^1(\Omega)_+}$. Most of the results in this section were already covered in [Harder, G. Wachsmuth, 2018c]. We use the notation and setting from Sections 5.1 and 5.2. For simplicity we assume that $y_a = 0$ in this section. Furthermore, we fix $(\bar{y}, \bar{\lambda}) \in \mathbb{K}$.

Let us briefly discuss why the limiting normal cone is relevant for the optimal control of the obstacle problem. If we replace $\mathcal{N}_{\mathbb{K}}^{\sharp}$ with $\mathcal{N}_{\mathbb{K}}^{\lim}$ in (5.5), then one can show that the system (5.4) together with (5.5) constitute necessary optimality conditions for local minimizers of (OCOP) under reasonable assumptions, see, e.g., [G. Wachsmuth, 2016, Proof of Lemma 4.4]. From [G. Wachsmuth, 2016, Proof of Lemma 4.4] one can also obtain the upper estimate

$$\mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda}) \subset \mathcal{N}_{\mathbb{K}}^{\text{weak}}(\bar{y}, \bar{\lambda}) \tag{5.13}$$

for the limiting normal cone to the set $\mathbb{K}$. If $d = 1$, then we can obtain the description

$$\mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda}) = \mathcal{N}_{\mathbb{K}}^{\text{s-lim}}(\bar{y}, \bar{\lambda}),$$

where $\mathcal{N}_{\mathbb{K}}^{\text{s-lim}}$ can be expressed via Remark 5.2.5 (b) which describes M-stationarity, see [Harder, G. Wachsmuth, 2018a, Theorem 5.4].

The question arises whether any reasonable lower estimates for $\mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda})$ can be found if $d \geq 2$. To our knowledge this was first done in [Harder, G. Wachsmuth, 2018c]. The main result is presented in Theorem 5.3.10 and shows that

$$(\nu, w) \in \mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda}) \qquad \Leftrightarrow \qquad (\nu, w) \in \mathcal{N}_{\mathbb{K}}^{\text{weak}}(\bar{y}, \bar{\lambda}) \tag{5.14}$$

holds for all $\nu \in L^p(\Omega) \subset H^{-1}(\Omega)$, $w \in H_0^1(\Omega)$ for exponents $p \in (1,2)$ where the embedding $L^p(\Omega) \hookrightarrow H^{-1}(\Omega)$ holds. The proof of this result can be quite technical. If the reader is interested in a simpler result that already demonstrates some of the techniques for the proof, we refer to [Harder, G. Wachsmuth, 2018c, Section 3], where (5.14) is shown for all $\nu \in L^2(\Omega)$ with $(\bar{y}, \bar{\lambda}) = (0,0) \in \mathbb{K}$. However, we will only consider the more general case in this section. The rest of this section is mostly taken from [Harder, G. Wachsmuth, 2018c].

5.3.1 A result from homogenization theory

An important ingredient is the following theorem, which is a (slightly generalized) result from [Cioranescu, Murat, 1997, Theorem 1.2].

Theorem 5.3.1. Suppose that $d \geq 2$ and let $\{\Omega_n\}_{n\in\mathbb{N}}$ be a sequence of open subsets of Ω. Suppose there exist sequences $\{v_n\}_{n\in\mathbb{N}} \subset H^1(\Omega)$, $\{\gamma_n\}_{n\in\mathbb{N}}, \{\mu_n\}_{n\in\mathbb{N}} \subset H^{-1}(\Omega)$ and a functional $\mu \in H^{-1}(\Omega)$ such that

$$v_n \in H^1(\Omega), \tag{H.1}$$
$$v_n = 0 \text{ q.e. on } \Omega \setminus \Omega_n, \tag{H.2}$$
$$v_n \rightharpoonup 1 \text{ in } H^1(\Omega), \tag{H.3}$$
$$\mu \in \begin{cases} W^{-1,d}(\Omega) & \text{if } d \geq 3, \\ W^{-1,2+\varepsilon}(\Omega) & \text{if } d = 2, \text{ for some } \varepsilon > 0, \end{cases} \tag{H.4'}$$
$$\mu_n \to \mu, \gamma_n \rightharpoonup \mu \text{ in } H^{-1}(\Omega),\ -\Delta v_n = \mu_n - \gamma_n, \langle \gamma_n, z_n \rangle = 0 \,\forall z_n \in H_0^1(\Omega_n). \tag{H.5'}$$

Let $\xi \in H^{-1}(\Omega)$ be given. We denote by w_n the unique (weak) solution of

$$-\Delta w_n = \xi \quad \text{in } H^{-1}(\Omega_n), \quad w_n \in H_0^1(\Omega_n) \subset H_0^1(\Omega).$$

Then w_n converges weakly in $H_0^1(\Omega)$ towards the unique solution w of

$$-\Delta w + \mu w = \xi, \quad w \in H_0^1(\Omega), \tag{5.15}$$

where the multiplication $\mu w \in H^{-1}(\Omega)$ is implicitly defined via $\langle \mu w, f \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle \mu, wf \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}$ for all $f \in C_c^\infty(\Omega)$.

Proof. This theorem is only a slight generalization of [Cioranescu, Murat, 1997, Theorem 1.2] and we will not repeat the proof. Instead, we only discuss the differences. First, we use a right-hand side $\xi \in H^{-1}(\Omega)$ instead of $f \in L^2(\Omega)$. It can be seen from the proof, that the right-hand side is only used as a functional over $H_0^1(\Omega)$, so the proof extends to a right-hand side of $\xi \in H^{-1}(\Omega)$.

Next, we generalize the condition $\mu \in W^{-1,\infty}(\Omega)$, which is used in [Cioranescu, Murat,

1997]. Again, inspecting the proof of [Cioranescu, Murat, 1997, Proposition 1.1, Theorem 1.2] reveals that we only need the property $\mu z \in H^{-1}(\Omega)$ for all $z \in H_0^1(\Omega)$. Using the Sobolev embedding theorem we get $z\,v \in W_0^{1,d/(d-1)}(\Omega)$ for $z, v \in H_0^1(\Omega)$ and dimensions $d \geq 3$. Hence, $\mu z \in H^{-1}(\Omega)$ if $z \in H_0^1(\Omega)$ and $\mu \in W_0^{1,d/(d-1)}(\Omega)^\star = W^{-1,d}(\Omega)$. This regularity is guaranteed by (H.4'). The case $d = 2$ is similar.

Finally, instead of condition (H.5) we can use (H.5') as explained in [Cioranescu, Murat, 1997, Remark 1.6].

We remark that the assumptions (H.1) to (H.3) of Theorem 5.3.1 imply $\Omega_n = \Omega$ for large n in dimension $d = 1$ due to the compact embedding $H^1(\Omega) \hookrightarrow C(\operatorname{cl}\Omega)$. We further remark that it can be shown that the functional $\mu \in H^{-1}(\Omega)$ is a nonnegative functional if the assumptions (H.1) to (H.5') are satisfied, i.e. $\mu \in H^{-1}(\Omega)_+$ holds. This implies that the partial differential equation (5.15) is uniquely solvable. However, for our purposes it suffices to know that this partial differential equation has a unique solution.

Let us explain how Theorem 5.3.1 is applied later. Suppose that we have a sequence $\{\Omega_n\}_{n\in\mathbb{N}}$ of open subsets of Ω such that the assumptions (H.1) to (H.5') are satisfied for some sequences $\{v_n\}_{n\in\mathbb{N}}$, $\{\gamma_n\}_{n\in\mathbb{N}}$ and $\{\mu_n\}_{n\in\mathbb{N}}$. Further, let $w \in H_0^1(\Omega)$ be arbitrary. We define w_n as the unique weak solution of

$$-\Delta w_n = -\Delta w + \mu w \in H^{-1}(\Omega_n), \quad w_n \in H_0^1(\Omega_n).$$

Then, Theorem 5.3.1 implies that $w_n \rightharpoonup w$ in $H_0^1(\Omega)$. That is, every $w \in H_0^1(\Omega)$ can be approximated weakly by a sequence $\{w_n\}_{n\in\mathbb{N}} \subset H_0^1(\Omega_n) \subset H_0^1(\Omega)$. In particular, $w_n = 0$ q.e. on $\Omega \setminus \Omega_n$.

5.3.2 Weak approximation of multipliers

In order to show that a pair of multipliers (ν, w) belongs to the limiting normal cone we need to find certain sequences $\{\nu_n\}_{n\in\mathbb{N}} \subset H^{-1}(\Omega)$, $\{w_n\}_{n\in\mathbb{N}} \subset H_0^1(\Omega)$ that converge weakly to ν and w, see Definition 2.4.1. For the construction of $\{w_n\}_{n\in\mathbb{N}}$ we use Theorem 5.3.1, which means that we need to show all the conditions of that theorem. A certain type of pointwise orthogonality of w_n and ν_n is required for all $n \in \mathbb{N}$ that are large enough. Therefore, we will construct $\{\nu_n\}_{n\in\mathbb{N}}$ as a sequence of functions in $L^\infty(\Omega)$ such that $\nu_n = 0$ a.e. on Ω_n and $w_n = 0$ q.e. on $\Omega \setminus \Omega_n$, where $\{\Omega_n\}_{n\in\mathbb{N}}$ is a sequence of open subsets of Ω such that the assumptions of Theorem 5.3.1 can be satisfied. Although the functions ν_n, w_n are pointwise orthogonal for all $n \in \mathbb{N}$ that are large enough, it will turn out that the weak limits $\nu \in H^{-1}(\Omega)$, $w \in H_0^1(\Omega)$ are not necessarily pointwise orthogonal and that our construction works for arbitrary $\nu \in L^p(\Omega) \subset H^{-1}(\Omega)$, $w \in H_0^1(\Omega)$, where $p \in (1, 2)$ is such that $L^p(\Omega) \hookrightarrow H^{-1}(\Omega)$.

Throughout this section, let $p \in (1, 2) \cap [2d/(d + 2), 2)$, $\nu \in L^p(\Omega)$, and $w \in H_0^1(\Omega)$ be chosen arbitrarily but fixed. Additionally, we make the assumption that $d \geq 2$ holds. Note that according to the Sobolev embedding theorem, these possible values for p are

precisely those $p \in (1,2)$ such that the embedding $L^p(\Omega) \hookrightarrow H^{-1}(\Omega)$ holds, see [Adams, Fournier, 2003, Theorem 4.12].

Let us start with defining the sequence $\{\Omega_n\}_{n\in\mathbb{N}}$ of open subsets of Ω and some related objects that will be important in this section. In order to avoid some case distinctions between $d = 2$ and $d \geq 3$, we introduce the auxiliary function L^d via

$$L^d(a) := \begin{cases} -\log(a)^{-1} & \text{for } a \in (0,1) \text{ and } d = 2, \\ a^{d-2} & \text{for } a \in (0,\infty) \text{ and } d \geq 3. \end{cases}$$

In any case, L^d is monotonically increasing and the range is $(0,\infty)$.

We are going to cover Ω by closed cubes. Therefore, fix a number $n \in \mathbb{N}$ and let $\{\omega_i^n\}_{i\in\mathbb{N}} = \frac{2}{n}\mathbb{Z}^d$ be a regular grid, and we define the cubes $P_i^n := \omega_i^n + [-\frac{1}{n}, \frac{1}{n}]^d$. These cubes have edge length $\frac{2}{n}$ and their interiors are pairwise disjoint. We also define $O_i^n := \operatorname{int} B_{1/n}(\omega_i^n) \subset P_i^n$ to be the (open) ball with radius $1/n$ that is contained in P_i^n and I_n as the set of indices i with $P_i^n \cap \Omega \neq \emptyset$. Note that I_n is finite for all $n \in \mathbb{N}$ because Ω is bounded. Furthermore, each cube contains a closed hole $T_i^n := B_{a_{i,n}}(\omega_i^n)$ at the center, which is a closed ball with radius $a_{i,n} \geq 0$. For technical reasons, we introduce another index set J_n, defined as

$$J_n := \{i \in I_n \mid P_i^n \subset \Omega,\ 0 < \|\nu\|^p_{L^p(P_i^n)} < n^{1-d}\}.$$

Contrary to the approach in [Harder, G. Wachsmuth, 2018c, Section 3], the radii of the holes T_i^n are not uniform and depend on the local values of ν. For $i \in J_n$, we define the radius $a_{i,n} > 0$ of the hole T_i^n implicitly via

$$L^d(a_{i,n}) = \operatorname{meas}(P_i^n) \operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1}, \tag{5.16}$$

where

$$\operatorname{avg}(P_i^n, |\nu|^p) := \frac{1}{\operatorname{meas}(P_i^n)} \int_{P_i^n} |\nu|^p \,\mathrm{d}\omega$$

is the average of the function $|\nu|^p$ over the set P_i^n. In the case that $i \in I_n \setminus J_n$ we set $a_{i,n} := 0$. The radii $a_{i,n}$ are well-defined because L^d is injective and has range $(0,\infty)$ and because $\operatorname{avg}(P_i^n, |\nu|^p)$ is positive for $i \in J_n$. Finally, we define the perforated domain

$$\Omega_n := \Omega \setminus \bigcup_{i\in J_n} T_i^n.$$

The next lemma shows that we have $a_{i,n} \leq \frac{1}{n}$ for n large enough, i.e., $T_i^n \subset P_i^n$ holds. Afterwards, we will only consider these parameters $n \in \mathbb{N}$ which guarantee $a_{i,n} \leq 1/n$.

Lemma 5.3.2. (a) For large $n \in \mathbb{N}$ we have

$$a_{i,n} < \left(\frac{1}{2n}\right)^{1+\varepsilon} \quad \forall i \in J_n,$$

where ε depends on the dimension, but not on i and n. In the case of $d = 2$ we even have

$$a_{i,n} < \exp(-n/8) \ \forall i \in J_n$$

for large $n \in \mathbb{N}$.

(b) For every large $n \in \mathbb{N}$ there exists a constant $C_n > 1$ such that

$$L^d(a_{i,n}) \leq \frac{1}{L^d(a_{i,n})^{-1} - L^d(1/n)^{-1}} \leq C_n L^d(a_{i,n})$$

holds for all $i \in J_n$. Moreover, $C_n \to 1$ as $n \to \infty$.

(c) The convergence

$$\lim_{n\to\infty} \operatorname{meas}\Big(\{\nu \neq 0\} \setminus \bigcup_{i\in J_n} P_i^n\Big) = 0$$

holds.

Proof. We start with part (a). Using the definition of the index set J_n and $2/p-1 \in (0,1)$, we have

$$\begin{aligned} L^d(a_{i,n}) &= \operatorname{meas}(P_i^n) \operatorname{avg}(P_i^n, |\nu|^p)^{2/p-1} \leq \operatorname{meas}(P_i^n)(1 + \operatorname{avg}(P_i^n, |\nu|^p)) \\ &= \left(\frac{2}{n}\right)^d + \|\nu\|^p_{L^p(P_i^n)} < \left(\frac{2}{n}\right)^d + n^{1-d} < 2^{d+1} n^{1-d}. \end{aligned} \tag{5.17}$$

For $d \geq 3$ and large $n \in \mathbb{N}$ this implies

$$a_{i,n} < 2^{d+1} n^{(1-d)/(d-2)} < \left(\frac{1}{2n}\right)^{1+\varepsilon},$$

where we set $\varepsilon := 1/(2(d-2))$. For $d = 2$, the inequality (5.17) yields $a_{i,n} < \exp(-n/8)$.

For part (b), we note that by part (a) it follows that $L^d(a_{i,n})/L^d(1/n) \to 0$ as $n \to \infty$, uniformly in $i \in J_n$. This implies the claim.

It remains to prove part (c). If $i \in I_n$ does not belong to J_n then there are three possible reasons: $\|\nu\|^p_{L^p(P_i^n)} = 0$, $\|\nu\|^p_{L^p(P_i^n)} \geq n^{1-d}$, or $P_i^n \not\subset \Omega$. Therefore, we have

$$\begin{aligned} \operatorname{meas}\Big(\{\nu \neq 0\} \setminus \bigcup_{i\in J_n} P_i^n\Big) &= \sum_{i\in I_n\setminus J_n} \operatorname{meas}(\{\nu \neq 0\} \cap P_i^n) \\ &\leq \sum_{i:\|\nu\|^p_{L^p(P_i^n)} \geq n^{1-d}} \operatorname{meas}(P_i^n) + \sum_{i:P_i^n \not\subset \Omega} \operatorname{meas}(\Omega \cap P_i^n). \end{aligned}$$

For the first term we have

$$\begin{aligned}\sum_{i:\|\nu\|^p_{L^p(P_i^n)}\geq n^{1-d}} \operatorname{meas}(P_i^n) &= \sum_{i:\|\nu\|^p_{L^p(P_i^n)}\geq n^{1-d}} (2n)^{-d} \\ &\leq 2^{-d} n^{-1} \sum_{i:\|\nu\|^p_{L^p(P_i^n)}\geq n^{1-d}} \|\nu\|^p_{L^p(P_i^n)} \\ &\leq 2^{-d} n^{-1} \|\nu\|^p_{L^p(\Omega)} \to 0\end{aligned}$$

as $n \to \infty$. The convergence of the second term follows from $\Omega = \bigcup_{n\in\mathbb{N}} \bigcup_{i:P_i^n\subset\Omega} P_i^n$. This proves the claim.

For the next lemma, the adaptive choice of the radii $a_{i,n}$ in (5.16) is crucial.

Lemma 5.3.3. We define the measurable function $\tilde{\nu}_n$ via

$$\tilde{\nu}_n := \sum_{i\in J_n} \chi_{T_i^n} \beta_{i,n},$$

where the real-valued coefficients $\beta_{i,n}$ satisfy

$$|\beta_{i,n}| \leq \frac{1}{\operatorname{meas}(T_i^n)} \int_{P_i^n} |\nu| \,\mathrm{d}\omega.$$

Then there is a constant $C > 0$ (depending only on the domain Ω and p) such that

$$\|\tilde{\nu}_n\|_{H^{-1}(\Omega)} \leq C \,\|\nu\|^{p/2}_{L^p(\Omega)} + C \,\|\nu\|_{L^p(\Omega)}$$

holds for all $n \in \mathbb{N}$.

Proof. Let $n \in \mathbb{N}$ be fixed. We define $u_{i,n}$ as the solution of

$$-\Delta u_{i,n} = \chi_{T_i^n} - a_{i,n}^d n^d \chi_{O_i^n} \quad \text{in } \Omega, \quad u_{i,n} \in H_0^1(\Omega).$$

It follows that

$$\tilde{\nu}_n = -\Delta\Big(\sum_{i\in J_n} \beta_{i,n} u_{i,n}\Big) + \sum_{i\in J_n} a_{i,n}^d n^d \beta_{i,n} \chi_{O_i^n}. \tag{5.18}$$

From Lemma 2.2.14 (a) we know that $\{u_{i,n} \neq 0\} \subset \operatorname{cl} O_i^n \subset P_i^n$ and

$$\|u_{i,n}\|^2_{H_0^1(\Omega)} \leq C \, a_{i,n}^{2d} L^d (a_{i,n})^{-1},$$

where the constant $C > 0$ does not depend on n and i. We continue with the boundedness

of $\|\tilde{\nu}_n\|_{H^{-1}(\Omega)}$. Using the isometry of $-\Delta$, (5.18) yields

$$\|\tilde{\nu}_n\|_{H^{-1}(\Omega)} \leq \Big\| \sum_{i\in J_n} \beta_{i,n} u_{i,n} \Big\|_{H_0^1(\Omega)} + \Big\| \sum_{i\in J_n} a_{i,n}^d n^d \beta_{i,n} \chi_{O_i^n} \Big\|_{H^{-1}(\Omega)}.$$

Since the functions $u_{i,n}$ are orthogonal with respect to the inner product in $H_0^1(\Omega)$, we have

$$\begin{aligned}
\Big\| \sum_{i\in J_n} \beta_{i,n} u_{i,n} \Big\|_{H_0^1(\Omega)}^2 &= \sum_{i\in J_n} |\beta_{i,n}|^2 \|u_{i,n}\|_{H_0^1(\Omega)}^2 \\
&\leq C \sum_{i\in J_n} a_{i,n}^{2\,d} L^d(a_{i,n})^{-1} \frac{1}{\operatorname{meas}(T_i^n)^2} \Big(\int_{P_i^n} |\nu| \,\mathrm{d}\omega \Big)^2 \\
&\leq C \sum_{i\in J_n} L^d(a_{i,n})^{-1} \Big(\int_{P_i^n} |\nu| \,\mathrm{d}\omega \Big)^2 \\
&\leq C \sum_{i\in J_n} L^d(a_{i,n})^{-1} \operatorname{meas}(P_i^n)^{2-\frac{2}{p}} \|\nu\|_{L^p(P_i^n)}^2 \\
&= C \sum_{i\in J_n} L^d(a_{i,n})^{-1} \operatorname{meas}(P_i^n) \operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1} \|\nu\|_{L^p(P_i^n)}^p \\
&= C \sum_{i\in J_n} \|\nu\|_{L^p(P_i^n)}^p \leq C \|\nu\|_{L^p(\Omega)}^p.
\end{aligned}$$

For the other term we have

$$\begin{aligned}
\Big\| \sum_{i\in J_n} a_{i,n}^d n^d \beta_{i,n} \chi_{O_i^n} \Big\|_{H^{-1}(\Omega)}^p &\leq C \Big\| \sum_{i\in J_n} a_{i,n}^d n^d \beta_{i,n} \chi_{O_i^n} \Big\|_{L^p(\Omega)}^p \\
&= C \sum_{i\in J_n} (a_{i,n} n)^{dp} |\beta_{i,n}|^p \|\chi_{O_i^n}\|_{L^p(\Omega)}^p \\
&\leq C \sum_{i\in J_n} n^{dp-d} \Big(\int_{P_i^n} |\nu| \,\mathrm{d}\omega \Big)^p \\
&\leq C \sum_{i\in J_n} n^{dp-d} \operatorname{meas}(P_i^n)^{(1-\frac{1}{p})p} \int_{P_i^n} |\nu|^p \,\mathrm{d}\omega \\
&\leq C \sum_{i\in J_n} \int_{P_i^n} |\nu|^p \,\mathrm{d}\omega \leq C \|\nu\|_{L^p(\Omega)}^p.
\end{aligned}$$

This completes the proof.

After we have established this boundedness, we are in a position to prove that every function $\tilde{\nu}$ which is pointwise bounded by $|\nu|$ can be approximated weakly by functions living on the holes T_i^n.

Lemma 5.3.4. Let $\tilde{\nu}$ be a function in $L^p(\Omega) \subset H^{-1}(\Omega)$ such that $|\tilde{\nu}| \leq |\nu|$. If we define

$$\tilde{\nu}_n := \sum_{i \in J_n} \chi_{T_i^n} \frac{1}{\operatorname{meas}(T_i^n)} \int_{P_i^n} \tilde{\nu} \, \mathrm{d}\omega,$$

then $\tilde{\nu}_n \rightharpoonup \tilde{\nu}$ in $H^{-1}(\Omega)$.

Proof. Because $\tilde{\nu}_n$ satisfies the requirements for Lemma 5.3.3, we know that $\tilde{\nu}_n$ is bounded in $H^{-1}(\Omega)$. Hence, it suffices to show the convergence on the dense linear subspace $C_c^\infty(\Omega) \subset H_0^1(\Omega)$. Let $f \in C_c^\infty(\Omega)$ be given. We have

$$\begin{aligned}\left|\langle \tilde{\nu}_n - \tilde{\nu}, f \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)}\right| &= \left|\int_\Omega (\tilde{\nu}_n - \tilde{\nu}) f \, \mathrm{d}\omega\right| \\ &\leq \left|\sum_{i \in J_n} \int_{P_i^n} (\tilde{\nu}_n - \tilde{\nu}) f \, \mathrm{d}\omega\right| + \left|\int_{\{\tilde{\nu}\neq 0\}\setminus \bigcup_{i\in J_n} P_i^n} f\tilde{\nu} \, \mathrm{d}\omega\right|.\end{aligned}$$

The second term converges to 0 because of Lemma 5.3.2 (c). For the first term we can use that f is uniformly continuous. This means that for each $\varepsilon > 0$ we can find arbitrarily large $n \in \mathbb{N}$ such that $|f(\omega) - f(\hat{\omega})| < \varepsilon$ for all $\omega, \hat{\omega} \in P_i^n, i \in J_n$. Thus,

$$\begin{aligned}\left|\sum_{i \in J_n} \int_{P_i^n} (\tilde{\nu}_n - \tilde{\nu}) f \, \mathrm{d}\omega\right| &\leq \sum_{i \in J_n} \left|\frac{1}{\operatorname{meas}(T_i^n)} \int_{P_i^n} \tilde{\nu}(\omega) \, \mathrm{d}\omega \int_{T_i^n} f(\hat{\omega}) \, \mathrm{d}\hat{\omega} - \int_{P_i^n} f(\omega)\tilde{\nu}(\omega) \, \mathrm{d}\omega\right| \\ &\leq \sum_{i \in J_n} \int_{P_i^n} |\tilde{\nu}(\omega)| \left|\frac{1}{\operatorname{meas}(T_i^n)} \int_{T_i^n} |f(\hat{\omega}) - f(\omega)| \, \mathrm{d}\hat{\omega}\right| \mathrm{d}\omega \\ &\leq \varepsilon \sum_{i \in J_n} \int_{P_i^n} |\tilde{\nu}| \, \mathrm{d}\omega \leq \varepsilon \|\tilde{\nu}\|_{L^1(\Omega)}.\end{aligned}$$

Since ε can be arbitrarily small, this proves that $\langle \tilde{\nu}_n - \tilde{\nu}, f \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} \to 0$ as $n \to \infty$.

If we use the above lemma with $\tilde{\nu} = \nu$ we get an explicit description of the sequence $\{\nu_n\}_{n\in\mathbb{N}}$ that converges weakly to ν.

The next lemma provides some technical estimates, which will be used to verify the assumptions of Theorem 5.3.1 in our setting with differently sized holes. Again, the specific choice (5.16) is crucial for these estimates.

Lemma 5.3.5. Let the size of the holes T_i^n be chosen according to (5.16). Then, there exists a constant $C > 0$, such that

$$\sum_{i \in J_n} L^d(a_{i,n}) \leq C \, \|\nu\|_{L^p(\Omega)}^{2-p}, \tag{5.19a}$$

$$n^{d-2} \sum_{i \in J_n} L^d(a_{i,n})^2 \to 0, \tag{5.19b}$$

$$n^{d\,q-d} \sum_{i\in J_n} L^d(a_{i,n})^q \le C\,\|\nu\|^p_{L^p(\Omega)}, \tag{5.19c}$$

where $q = p/(2-p)$.

Proof. We start with (5.19a). Using the definition (5.16) and $\operatorname{meas}(P_i^n) = (2/n)^d$, we find

$$\begin{aligned}\sum_{i\in J_n} L^d(a_{i,n}) &= \sum_{i\in J_n} \operatorname{meas}(P_i^n)\,\operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1} = \sum_{i\in J_n} \operatorname{meas}(P_i^n)^{2-\frac{2}{p}} \Big(\int_{P_i^n} |\nu|^p \,\mathrm{d}\omega\Big)^{\frac{2}{p}-1} \\ &= 2^{2\,d\,(1-\frac{1}{p})}\, n^{d\,(\frac{2}{p}-2)} \sum_{i\in J_n} \Big(\int_{P_i^n} |\nu|^p \,\mathrm{d}\omega\Big)^{\frac{2}{p}-1}.\end{aligned}$$

Since $\frac{2}{p} - 1 \in (0,1)$, we can use Holder's inequality to obtain

$$\begin{aligned}\sum_{i\in J_n} L^d(a_{i,n}) &\le 2^{2\,d\,(1-\frac{1}{p})}\, n^{d\,(\frac{2}{p}-2)} \Big(\sum_{i\in J_n} \int_{P_i^n} |\nu|^p \,\mathrm{d}\omega\Big)^{\frac{2}{p}-1} \Big(\sum_{i\in J_n} 1\Big)^{2-\frac{2}{p}} \\ &\le C\, n^{d\,(\frac{2}{p}-2)+d\,(2-\frac{2}{p})}\, \|\nu\|^{p(\frac{2}{p}-1)}_{L^p(\Omega)} \le C\,\|\nu\|^{2-p}_{L^p(\Omega)}.\end{aligned}$$

This shows (5.19a).

Next, we verify (5.19b). We use (5.17) and (5.19a) and obtain

$$n^{d-2} \sum_{i\in J_n} L^d(a_{i,n})^2 \le n^{d-2} \sum_{i\in J_n} L^d(a_{i,n})\, 2^{d+1}\, n^{1-d} = 2^{d+1}\, n^{-1} \sum_{i\in J_n} L^d(a_{i,n}) \to 0.$$

Finally, we address (5.19c). Using (5.16) and $q\,(2/p-1) = 1$ we get

$$n^{d\,q-d} \sum_{i\in J_n} L^d(a_{i,n})^q = n^{d\,q-d} \sum_{i\in J_n} \operatorname{meas}(P_i^n)^{q-1} \int_{P_i^n} |\nu|^p \,\mathrm{d}\omega \le 2^{d\,(q-1)}\, \|\nu\|^p_{L^p(\Omega)}.$$

As a next step, we verify that the conditions (H.1) to (H.5') of Theorem 5.3.1 are satisfied for the above choice of the perforated domain Ω_n. We are following the strategy of [Cioranescu, Murat, 1997, Theorem 2.2]. However, due to the variable size of the holes, the analysis is more involved.

We start by defining an appropriate $v_n \in H^1(\Omega)$. For $i \in J_n$ let $v_{i,n} \in H^1_0(\Omega)$ be defined as the solution to

$$\begin{aligned} v_{i,n} &= 1 \quad \text{in } T_i^n, \\ -\Delta v_{i,n} &= 0 \quad \text{in } O_i^n \setminus T_i^n, \\ v_{i,n} &= 0 \quad \text{in } \Omega \setminus O_i^n. \end{aligned}$$

Functions of this type are discussed in Lemma 2.2.14 (b). Note that the requirements on $a_{i,n}$ in this lemma are satisfied by Lemma 5.3.2 for $n \in \mathbb{N}$ large enough. We then define

$$v_n := 1 - \sum_{i \in J_n} v_{i,n}.$$

The next two lemmas show that (H.1) to (H.5') are satisfied.

Lemma 5.3.6. The conditions (H.1) to (H.3) are satisfied by the above choice of $\{v_n\}_{n\in\mathbb{N}}$.

Proof. The conditions (H.1) and (H.2) follows directly from our choice of v_n.

Let us show that $\{v_n\}_{n\in\mathbb{N}}$ is a bounded sequence in $H^1(\Omega)$. Because of $0 \le v_n \le 1$ it suffices to calculate $\big\| |\nabla v_n| \big\|^2_{H^1_0(\Omega)}$. Due to Lemma 2.2.14 (b) and (5.19a) we have

$$\big\| |\nabla v_n| \big\|^2 \le \sum_{i \in J_n} \|v_{i,n}\|^2_{H^1_0(\Omega)} \le C \sum_{i \in J_n} L^d(a_{i,n}) \le C \, \|\nu\|^{2-p}_{L^p(\Omega)}, \tag{5.20}$$

which shows that the sequence $\{v_n\}_{n\in\mathbb{N}}$ is bounded in $H^1(\Omega)$. We check that the convergence $v_n \to 1$ holds in $L^1(\Omega)$. Indeed,

$$\|v_n - 1\|_{L^1(\Omega)} = \sum_{i \in J_n} \int_{O_i^n} |v_{i,n}| \, \mathrm{d}\omega \le C \sum_{i \in J_n} \frac{1}{n} L^d(a_{i,n}) \to 0,$$

where we used (2.22) and (5.19a). Together with the boundedness of $\{v_n\}_{n\in\mathbb{N}}$ in $H^1(\Omega) \hookrightarrow L^1(\Omega)$ and the reflexivity of $H^1(\Omega)$, this implies $v_n \rightharpoonup 1$ in $H^1(\Omega)$. This shows (H.3).

We remark that (5.20) shows that the capacity of the holes $\bigcup_{i\in J_n} T_i^n$ remains bounded. Indeed, the function $1 - v_n$ can be used in Lemma 2.6.16 and we obtain

$$\operatorname{cap}\Big(\bigcup_{i \in J_n} T_i^n \Big) \le \|1 - v_n\|^2_{H^1_0(\Omega)} \le C \, \|\nu\|^{2-p}_{L^p(\Omega)}$$

Lemma 5.3.7. The conditions (H.4') and (H.5') are satisfied by the above choice of $\{v_n\}_{n\in\mathbb{N}}$ and some sequences $\{\mu_n\}_{n\in\mathbb{N}}$ and $\{\gamma_n\}_{n\in\mathbb{N}}$. In particular, we have $\mu = C_d \, |\nu|^{2-p}$, where $C_d = \max(1, d-2) S_d$ and S_d denotes the surface measure of the boundary of the d-dimensional unit ball $B_1(0) \subset \mathbb{R}^d$.

Proof. First, we prove (H.5'). We note that Δv_n only acts on the boundaries ∂T_i^n and ∂O_i^n. We set $\gamma_n, \mu_n \in H^{-1}(\Omega)$ such that $-\Delta v_n = \mu_n - \gamma_n$ and μ_n only acts on ∂O_i^n whereas γ_n only acts on ∂T_i^n. Then it can be seen that the condition $\langle \gamma_n, z_n \rangle = 0$ is true for all $z_n \in H^1_0(\Omega_n)$. It is possible to explicitly calculate μ_n. We denote by $\delta_{i,n} \in H^{-1}(\Omega)$ the surface measure on ∂O_i^n, i.e.,

$$\langle \delta_{i,n}, f \rangle_{H^{-1}(\Omega) \times H^1_0(\Omega)} = \int_{\partial O_i^n} f(s) \, \mathrm{d}s \quad \forall f \in C_c^\infty(\Omega).$$

Then, using integration by parts and (2.23), it turns out that

$$\mu_n = \sum_{i\in J_n} \frac{\partial v_n}{\partial n}\Big|_{\partial O_i^n} \delta_{i,n} = -\sum_{i\in J_n} \frac{\partial v_{i,n}}{\partial n}\Big|_{\partial O_i^n} \delta_{i,n} = \sum_{i\in J_n} \frac{1}{n\,d}\alpha_{i,n}\delta_{i,n} \tag{5.21}$$

holds, where $\frac{\partial \hat{v}}{\partial n}\big|_{\partial O_i^n}$ denotes the outer normal derivative of a function $\hat{v}$ on ∂O_i^n and the coefficients $\alpha_{i,n}$ are given by

$$\alpha_{i,n} := \frac{\max(1, d-2)\, n^d\, d}{L^d(a_{i,n})^{-1} - L^d(1/n)^{-1}}. \tag{5.22}$$

For later use we note that Lemma 5.3.2 (b) implies the existence of a constant $C > 0$ independent of i and n such that

$$0 \le \alpha_{i,n} \le C n^d\, L^d(a_{i,n}). \tag{5.23}$$

Now we introduce the function $z_{i,n}$ for $i \in J_n$ as the solution of the equation

$$-\Delta z_{i,n} = \alpha_{i,n} \quad \text{in } O_i^n, \qquad z_{i,n} = 0 \quad \text{on } \Omega \setminus O_i^n.$$

This function can be calculated explicitly and we find

$$z_{i,n}(\omega) = \frac{\alpha_{i,n}}{2\,d}(n^{-2} - |\omega - \omega_i^n|^2) \quad \forall \omega \in O_i^n$$

and

$$-\Delta z_{i,n} = \alpha_{i,n}\chi_{O_i^n} - \frac{1}{n\,d}\alpha_{i,n}\delta_{i,n} \in H^{-1}(\Omega). \tag{5.24}$$

For the $H_0^1(\Omega)$-norm of $z_{i,n}$ we can calculate

$$\|z_{i,n}\|^2_{H_0^1(\Omega)} = C\,\alpha_{i,n}^2 n^{-d-2}$$

and because of the orthogonality we have

$$\Big\|\sum_{i\in J_n} z_{i,n}\Big\|^2_{H_0^1(\Omega)} = \sum_{i\in J_n} \|z_{i,n}\|^2_{H_0^1(\Omega)} = C \sum_{i\in J_n} \alpha_{i,n}^2 n^{-d-2} \le C \sum_{i\in J_n} L^d(a_{i,n})^2\, n^{d-2} \to 0$$

due to (5.23) and (5.19b). Hence, (5.21) and (5.24) imply

$$\mu_n - \sum_{i\in J_n} \alpha_{i,n}\chi_{O_i^n} = \Delta\Big(\sum_{i\in J_n} z_{i,n}\Big) \to 0 \quad \text{in } H^{-1}(\Omega) \quad (n\to\infty).$$

Using Lemma 5.3.8 below yields $\mu_n \to \mu$ in $H^{-1}(\Omega)$, where $\mu := C_d\,|\nu|^{2-p}$. Finally, $\gamma_n \rightharpoonup \mu$ follows from $-\Delta v_n \rightharpoonup 0$ and $\mu_n \to \mu$, which completes the proof of (H.5').

Now, $\mu = C_d\,|\nu|^{2-p}$, $\nu \in L^p(\Omega)$ and the bounds on p imply

$$\mu \in L^{p/(2-p)}(\Omega) \subset \begin{cases} W^{-1,2+\varepsilon}(\Omega) & \text{if } d = 2, \\ W^{-1,d}(\Omega) & \text{if } d \geq 3 \end{cases}$$

for some $\varepsilon > 0$. Thus, the remaining condition (H.4') follows.

It remains to check the announced convergence of μ_n towards $\mu = C_d\,|\nu|^{2-p}$.

Lemma 5.3.8. Let $\alpha_{i,n}$ be defined as in (5.22). Then we have the convergence

$$\sum_{i\in J_n} \alpha_{i,n}\chi_{O_i^n} \to C_d\,|\nu|^{2-p} \quad \text{in } H^{-1}(\Omega),$$

where C_d is the constant defined in Lemma 5.3.7.

Proof. We will prove this by showing the weak convergence in $L^q(\Omega)$, where $q = p/(2-p) \in (1,\infty)$. Indeed, the boundedness follows from

$$\Big\|\sum_{i\in J_n} \alpha_{i,n}\chi_{O_i^n}\Big\|_{L^q(\Omega)}^q = \sum_{i\in J_n} \alpha_{i,n}^q \operatorname{meas}(O_i^n) \leq C \sum_{i\in J_n} n^{q\,d} L^d(a_{i,n})^q\, n^{-d} \leq C\,\|\nu\|_{L^p(\Omega)}^p,$$

where the last two inequalities are due to (5.23) and (5.19c), respectively. Thus, it is sufficient to show the weak convergence on the dense subset $C_c^\infty(\Omega) \subset L^q(\Omega)^\star$. Due to the definition of $\alpha_{i,n}$ and Lemma 5.3.2 (b) we have

$$\begin{aligned} &\big|\alpha_{i,n} - \max(1,d-2)dn^d L^d(a_{i,n})\big| \\ &\qquad = \max(1,d-2)dn^d \Big|\frac{1}{L^d(a_{i,n})^{-1} - L^d(1/n)^{-1}} - L^d(a_{i,n})\Big| \\ &\qquad \leq (C_n - 1)\,\max(1,d-2)\,d\,n^d\,L^d(a_{i,n}), \end{aligned}$$

where $\{C_n\}_{n\in\mathbb{N}}$ is a sequence of constants such that $C_n \to 1$. It follows that

$$\begin{aligned} &\Big\|\sum_{i\in J_n} \alpha_{i,n}\chi_{O_i^n} - \max(1,d-2)d\,n^d \sum_{i\in J_n} L^d(a_{i,n})\chi_{O_i^n}\Big\|_{L^q(\Omega)}^q \\ &\qquad \leq C\,(C_n-1)^q \sum_{i\in J_n} n^{dq-d} L^d(a_{i,n})^q \leq C\,(C_n-1)^q \|\nu\|_{L^p(\Omega)}^p \to 0, \end{aligned} \tag{5.25}$$

where we used (5.19c) again. Now using Lemma 5.3.2 (c) we also have

$$\begin{aligned} \Big\|\sum_{i\in I_n\setminus J_n} \operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1}\chi_{O_i^n}\Big\|_{L^q(\Omega)}^q &= \sum_{i\in I_n\setminus J_n} \operatorname{avg}(P_i^n, |\nu|^p)^{(\frac{2}{p}-1)q} \operatorname{meas}(O_i^n) \\ &\leq \int_{\Omega\setminus\bigcup_{i\in J_n} P_i^n} |\nu|^p \,\mathrm{d}\omega \to 0 \end{aligned}$$

as $n \to \infty$. By combining this with (5.25) and the definition (5.16) of $a_{i,n}$ we arrive at

$$\sum_{i \in J_n} \alpha_{i,n} \chi_{O_i^n} - \max(1, d-2) d\, 2^d \sum_{i \in I_n} \operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1} \chi_{O_i^n} \to 0$$

in $L^q(\Omega)$. Let $f \in C_c^\infty(\Omega)$ be given. Using the uniform continuity of $f \in C_c^\infty(\Omega)$ (similar to the proof of Lemma 5.3.4) it is possible to replace $\chi_{O_i^n}$ with $\chi_{P_i^n}$, i.e.

$$\Big\langle \sum_{i \in J_n} \alpha_{i,n} \chi_{O_i^n} - \max(1, d-2) S_d \sum_{i \in I_n} \operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1} \chi_{P_i^n}, f \Big\rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \to 0,$$

where we used that $2^{-d} d^{-1} S_d = \operatorname{meas}(O_i^n) / \operatorname{meas}(P_i^n)$, and $d^{-1} S_d$ is the volume of the d-dimensional unit ball. Now we apply Lemma 2.2.4 (b) to $v = |\nu|^p$. As a consequence, we have

$$C_d \sum_{i \in I_n} \operatorname{avg}(P_i^n, |\nu|^p)^{\frac{2}{p}-1} \chi_{P_i^n} \to C_d |\nu|^{2-p}$$

in $L^q(\Omega)$ with the constant $C_d = \max(1, d-2) S_d$. Combined with the calculations above, we have

$$\Big\langle \sum_{i \in J_n} \alpha_{i,n} \chi_{O_i^n} - C_d |\nu|^{2-p}, f \Big\rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \to 0.$$

The boundedness in $L^q(\Omega)$ of $\sum_{i \in J_n} \alpha_{i,n} \chi_{O_i^n}$ and the compact embedding into $H^{-1}(\Omega)$ (which follows from $q > p$, see also [Adams, Fournier, 2003, Theorem 6.3]) completes the proof.

We note that choosing the constant functions

$$\nu := \left(2^d C_0\right)^{-1/(2-p)} \text{ if } d = 2, \qquad \nu := \left(2^d C_0^{2-d}\right)^{-1/(2-p)} \text{ if } d \geq 3$$

yields the same size of the holes as in [Cioranescu, Murat, 1997, (2.4)] and we obtain the same value of μ, cf. [Cioranescu, Murat, 1997, (2.3)].

Now, the assumptions of Theorem 5.3.1 are satisfied and therefore we can finally construct our sequence $\{w_n\}_{n \in \mathbb{N}} \subset H_0^1(\Omega)$ that converges weakly to w.

Lemma 5.3.9. There exists a sequence $\{w_n\}_{n \in \mathbb{N}}$ with $w_n \in H_0^1(\Omega_n) \subset H_0^1(\Omega)$ such that $w_n \rightharpoonup w$ in $H_0^1(\Omega)$.

Proof. We choose $w_n \in H_0^1(\Omega_n) \subset H_0^1(\Omega)$ as the solution of

$$-\Delta w_n = -\Delta w + \mu w \quad \text{in } H^{-1}(\Omega_n),$$

where $\mu = C_d |\nu|^{2-p}$ and the constant C_d is defined as in Lemma 5.3.7. Then we can apply Theorem 5.3.1, whose assumptions are satisfied due to Lemmas 5.3.6 and 5.3.7. This yields the weak convergence $w_n \rightharpoonup \hat{w}$ where $\hat{w} \in H_0^1(\Omega)$ is the unique solution of the

partial differential equation

$$-\Delta\hat{w} + \mu\hat{w} = -\Delta w + \mu w.$$

Then the claim follows from $\hat{w} = w$.

5.3.3 Lower estimates for the limiting normal cone

With the results from Section 5.3.2 we can now proceed with out main result, which can also be found in [Harder, G. Wachsmuth, 2018c, Theorem 4.9]. Since we consider the limiting normal cone at an arbitrary point $(\bar{y}, \bar{\lambda}) \in \mathbb{K}$, the proof is rather technical and requires multiple steps. Some tools from capacity theory will also be relevant. Recall that the cone $\mathcal{N}_{\mathbb{K}}^{\text{weak}}$ that corresponds to the system of weak stationarity was defined in (5.10).

Theorem 5.3.10. Suppose that $d \geq 2$ and $p \in (1, 2)$ are given such that $L^p(\Omega) \hookrightarrow H^{-1}(\Omega)$. Let $(\nu, w) \in L^p(\Omega) \times H_0^1(\Omega)$ be given. Then the equivalence

$$(\nu, w) \in \mathcal{N}_{\mathbb{K}}^{\text{lim}}(\bar{y}, \bar{\lambda}) \qquad \Leftrightarrow \qquad (\nu, w) \in \mathcal{N}_{\mathbb{K}}^{\text{weak}}(\bar{y}, \bar{\lambda})$$

holds for every $(\bar{y}, \bar{\lambda}) \in \mathbb{K}$.

Proof. The implication "$\Rightarrow$" follows directly from (5.13) and it remains to check "$\Leftarrow$". Therefore, let $(\bar{y}, \bar{\lambda}) \in \mathbb{K}$ and $(\nu, w) \in \mathcal{N}^{\text{weak}}(\bar{y}, \bar{\lambda})$ with $\nu \in L^p(\Omega)$ be given. According to the definition of $\mathcal{N}^{\text{weak}}(\bar{y}, \bar{\lambda})$ in (5.10) this means that $w = 0$ q.e. on q-supp$(\bar{\lambda})$ and $\langle \nu, z \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0$ for all $z \in H_0^1(\Omega)$ with $z = 0$ q.e. on $\{\bar{y} = 0\} =_q \Omega \setminus \{\bar{y} > 0\}$. It can be shown that $\nu = 0$ a.e. on $\{\bar{y} > 0\}$. In fact, Lemma 2.6.29 implies $\langle \nu, z \rangle_{L^p(\Omega)\times L^{p'}(\Omega)} = 0$ for all $z \in L^{p'}(\Omega)$ with $z = 0$ a.e. on $\Omega \setminus \{\bar{y} > 0\}$. Here, $p' \in (2, \infty)$ is the exponent conjugate to p, i.e., $1 = 1/p + 1/p'$.

It well be convenient to work with open sets. Therefore, let $\varepsilon > 0$ be given. Because $\{\bar{y} > 0\}$ is quasi-open, there exists an open set G_ε, such that $\{\bar{y} > 0\} \cup G_\varepsilon$ is open and $\text{cap}(G_\varepsilon) < \varepsilon$.

The remaining part of the proof is divided into several steps. In steps 1 and 2, we use Lemmas 5.3.4 and 5.3.9 to construct approximations w_n to w and $\nu_{n,\varepsilon}$ to ν. The functions w_n will vanish on the holes, whereas $\nu_{n,\varepsilon}$ is supported only on the holes. In step 3, we construct an approximation to $\bar{y}$, which vanishes on the support of $\nu_{n,\varepsilon}$. Afterwards, we find a point in $\mathbb{K}$ such that $(\nu_{n,\varepsilon}, w_n)$ belongs to the Fréchet normal cone in this point, cf. steps 4 and 5. Finally, we pick a diagonal sequence in step 6 and conclude.

Step 1 (Construction of w_n)**:** Applying Lemma 5.3.9 yields the existence of a sequence $\{\hat{w}_n\}_{n\in\mathbb{N}}$ with $\hat{w}_n \in H_0^1(\Omega_n) \subset H_0^1(\Omega)$ and $\hat{w}_n \rightharpoonup w$ in $H_0^1(\Omega)$. Next, we define $w_n \in H_0^1(\Omega)$ by $w_n := \max(\min(\hat{w}_n, w^+), -w^-)$. From Lemma 2.2.10 we know that max and min are sequentially weakly continuous from $H_0^1(\Omega) \times H_0^1(\Omega)$ to $H_0^1(\Omega)$. It follows that

$w_n \rightharpoonup w$. Moreover, we have $\{w_n \neq 0\} \subset_q \{\hat{w}_n \neq 0\} \subset_q \Omega_n$ and q-supp$(\bar{\lambda}) \subset_q \{w = 0\} \subset_q \{w_n = 0\}$. From Corollary 2.6.25 (b) it follows that

$$\langle \bar{\lambda}, w_n^+ \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle \bar{\lambda}, w_n^- \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0 \tag{5.26}$$

holds.

Step 2 (Construction of $\nu_{n,\varepsilon}$)**:** We define $\nu_\varepsilon := \nu\chi_{\Omega\setminus G_\varepsilon}$ and

$$\nu_{n,\varepsilon} := \sum_{i\in J_n} \chi_{T_i^n} \frac{1}{\operatorname{meas}(T_i^n)} \int_{P_i^n} \nu_\varepsilon \,\mathrm{d}\omega.$$

According to Lemma 5.3.4, $\nu_{n,\varepsilon} \rightharpoonup \nu_\varepsilon$ as $n \to \infty$ in $H^{-1}(\Omega)$. Moreover, we have

$$\|\nu_{n,\varepsilon}\|_{H^{-1}(\Omega)} + \|\nu_{n,\varepsilon}^+\|_{H^{-1}(\Omega)} \leq C\,\|\nu\|_{L^p(\Omega)}^{p/2} + C\,\|\nu\|_{L^p(\Omega)} \tag{5.27}$$

for a constant $C > 0$ by applying Lemma 5.3.3 twice.

Step 3 (Construction of $\bar{y}_{n,\varepsilon}$)**:** Now we will argue that we can choose a sequence $\{\bar{y}_{n,\varepsilon}\}_{n\in\mathbb{N}} \subset H_0^1(\Omega)$ such that

$$0 \leq \bar{y}_{n,\varepsilon} \leq \bar{y}, \tag{5.28a}$$

$$\lim_{n\to\infty} \bar{y}_{n,\varepsilon} = \bar{y}, \tag{5.28b}$$

$$\{\bar{y}_{n,\varepsilon} > 0\} \subset \bigcup_{i:P_i^n \subset \{\bar{y}>0\}\cup G_\varepsilon} P_i^n \tag{5.28c}$$

hold. Indeed, this is possible: Because of $\bar{y} \in H_0^1(\{\bar{y} > 0\} \cup G_\varepsilon)$ and the fact that $C_c^\infty(\{\bar{y} > 0\} \cup G_\varepsilon)$ is dense in $H_0^1(\{\bar{y} > 0\} \cup G_\varepsilon)$ there exists a sequence $\{\hat{y}_{n,\varepsilon}\}_{n\in\mathbb{N}}$ in $C_c^\infty(\{\bar{y} > 0\}\cup G_\varepsilon)$ such that $\lim_{n\to\infty} \hat{y}_{n,\varepsilon} = \bar{y}$ and $\{\hat{y}_{n,\varepsilon} > 0\} + B_{3\sqrt{d}/n}(0) \subset \{\bar{y} > 0\}\cup G_\varepsilon$. The last condition implies

$$\{\hat{y}_{n,\varepsilon} > 0\} \subset \bigcup_{i:P_i^n \subset \{\bar{y}>0\}\cup G_\varepsilon} P_i^n.$$

Then we define $\bar{y}_{n,\varepsilon} := \max(\min(\bar{y}, \hat{y}_{n,\varepsilon}), 0)$, and we get (5.28a). Because max and min are continuous in $H_0^1(\Omega)$, we also have $\lim_{n\to\infty} \bar{y}_{n,\varepsilon} = \bar{y}$. The remaining condition follows from $\{\bar{y}_{n,\varepsilon} > 0\} \subset \{\hat{y}_{n,\varepsilon} > 0\}$. This yields a sequence $\{\bar{y}_{n,\varepsilon}\}_{n\in\mathbb{N}}$ satisfying (5.28).

Step 4 (Construction of $(y_{n,\varepsilon}, \lambda_{n,\varepsilon}) \in \mathbb{K}$)**:** Now, we define $y_{n,\varepsilon} := \bar{y}_{n,\varepsilon} + \frac{1}{n} w_n^- \geq 0$ and $\lambda_{n,\varepsilon} := \bar{\lambda} - \frac{1}{n}\nu_{n,\varepsilon}^+ \leq 0$. In order to show that this pair belongs to $\mathbb{K}$, it remains to check

$$\langle \lambda_{n,\varepsilon}, y_{n,\varepsilon} \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle \bar{\lambda}, \bar{y}_{n,\varepsilon} \rangle + \frac{1}{n} \langle \bar{\lambda}, w_n^- \rangle - \frac{1}{n} \langle \nu_{n,\varepsilon}^+, \bar{y}_{n,\varepsilon} \rangle - \frac{1}{n^2} \langle \nu_{n,\varepsilon}^+, w_n^- \rangle \overset{!}{=} 0. \tag{5.29}$$

The first term vanishes due to $0 = \langle \bar{\lambda}, \bar{y} \rangle \leq \langle \bar{\lambda}, \bar{y}_{n,\varepsilon} \rangle \leq 0$, where we used $\bar{\lambda} \leq 0$ and (5.28a). The second terms is zero due to (5.26). We note that $\nu_\varepsilon = 0$ a.e. on $\{\bar{y} > 0\} \cup G_\varepsilon$. Therefore, the function $\nu_{n,\varepsilon}$ can only be nonzero on holes T_i^n that belong to cubes P_i^n with $P_i^n \not\subset \{\bar{y} > 0\} \cup G_\varepsilon$. Thus, using that $\bar{y}_{n,\varepsilon} = 0$ on these P_i^n, cf. (5.28c), the third term vanishes. Finally, the last term disappears since $\nu_{n,\varepsilon}^+$ only lives on the holes and w_n^- vanishes there. This shows (5.29). Together with the signs of $y_{n,\varepsilon}$ and $\lambda_{n,\varepsilon}$, we have $(y_{n,\varepsilon}, \lambda_{n,\varepsilon}) \in \mathbb{K}$.

Step 5 (Verification of $(\nu_{n,\varepsilon}, w_n) \in \mathcal{N}_{\mathbb{K}}^{\text{Fréchet}}(y_{n,\varepsilon}, \lambda_{n,\varepsilon})$)**:** In face of (5.11) we have to show $\nu_{n,\varepsilon} \in \mathcal{K}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}, \lambda_{n,\varepsilon})^\circ$ and $w_n \in \mathcal{K}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}, \lambda_{n,\varepsilon})$. By using arguments similar to those that led to (5.29) we find $\langle \lambda_{n,\varepsilon}, w_n \rangle = 0$. Together with $\bar{y}_{n,\varepsilon}, w_n^+ \geq 0$ this yields

$$w_n = n\,(\bar{y}_{n,\varepsilon} + \frac{1}{n} w_n^+ - y_{n,\varepsilon}) \in \mathcal{T}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}) \cap \lambda_{n,\varepsilon}^\perp = \mathcal{K}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}, \lambda_{n,\varepsilon}).$$

In order to show $\nu_{n,\varepsilon} \in \mathcal{K}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}, \lambda_{n,\varepsilon})^\circ$, let $z \in H_0^1(\Omega)_+ \cap \lambda_{n,\varepsilon}^\perp$ be given. Similar to the derivation of (5.29), we find $\langle \nu_{n,\varepsilon}, y_{n,\varepsilon} \rangle = 0$. From $z \in H_0^1(\Omega)_+ \cap \lambda_{n,\varepsilon}^\perp$, $\lambda_{n,\varepsilon} = \bar{\lambda} - \frac{1}{n}\,\nu_{n,\varepsilon}^+$, and $\bar{\lambda}, -\frac{1}{n}\,\nu_{n,\varepsilon}^+ \leq 0$ we obtain $\langle \nu_{n,\varepsilon}^+, z \rangle = 0$. Thus,

$$\langle \nu_{n,\varepsilon}, z - y_{n,\varepsilon} \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} = \langle \nu_{n,\varepsilon}, z \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} = \langle -\nu_{n,\varepsilon}^-, z \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \leq 0$$

holds, where we used $z \geq 0$ and $\nu_{n,\varepsilon}^- \geq 0$ in the last step. Since z was arbitrary, we find $\nu_{n,\varepsilon} \in (H_0^1(\Omega)_+ \cap \lambda_{n,\varepsilon}^\perp - y_{n,\varepsilon})^\circ = (\mathcal{R}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}) \cap \lambda_{n,\varepsilon}^\perp)^\circ$. Because $H_0^1(\Omega)_+$ is polyhedric due to Corollary 2.2.9, it follows that $\nu_{n,\varepsilon} \in \mathcal{K}_{H_0^1(\Omega)_+}(y_{n,\varepsilon}, \lambda_{n,\varepsilon})^\circ$.

Step 6 (Choice of a diagonal sequence)**:** Finally, we have to choose a sequence of indices $\{(n_k, \varepsilon_k)\}_{k \in \mathbb{N}}$ such that

$$y_k := y_{n_k,\varepsilon_k} \to \bar{y}, \qquad \lambda_k := \lambda_{n_k,\varepsilon_k} \to \bar{\lambda}, \qquad w_k := w_{n_k} \rightharpoonup w, \qquad \nu_k := \nu_{n_k,\varepsilon_k} \rightharpoonup \nu.$$

Let $\{\varepsilon_k\}_{k \in \mathbb{N}}$ be a sequence with $\varepsilon_k > 0$ and $\varepsilon_k \to 0$. Then, we have

$$\|\nu - \nu_{\varepsilon_k}\|_{H^{-1}(\Omega)} = \|\nu \chi_{G_{\varepsilon_k}}\|_{H^{-1}(\Omega)} \leq C\,\|\nu \chi_{G_{\varepsilon_k}}\|_{L^p(\Omega)} = C\left(\int_{G_{\varepsilon_k}} |\nu|^p \,\mathrm{d}\omega\right)^{1/p},$$

which converges to 0 as $\varepsilon \to 0$ since $\mathrm{meas}(G_{\varepsilon_k}) \to 0$, which follows from $\mathrm{cap}(G_{\varepsilon_k}) \to 0$, see Lemma 2.6.3 (f).

Because $H_0^1(\Omega)$ is separable, we can find a sequence $\{z_m\}_{m \in \mathbb{N}}$ that is dense in $H_0^1(\Omega)$. We have $\nu_{n,\varepsilon_k} \rightharpoonup \nu_{\varepsilon_k}$ and $\bar{y}_{n,\varepsilon_k} \to \bar{y}$ as $n \to \infty$ for fixed k by steps 2 and 3. Therefore, we can choose $n_k \geq k$ in such a way that the conditions

$$\|\bar{y}_{n_k,\varepsilon_k} - \bar{y}\|_{H_0^1(\Omega)} < \varepsilon_k \quad \text{and} \quad \left|\langle \nu_{n_k,\varepsilon_k} - \nu_{\varepsilon_k}, z_m \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)}\right| < \varepsilon_k \quad \forall m \leq k$$

are satisfied. From the boundedness of w_n^-, we conclude $y_{n_k,\varepsilon_k} = \bar{y}_{n_k,\varepsilon_k} + \frac{1}{n} w_{n_k}^- \to \bar{y}$. Further, we obtain that

$$\lim_{k\to\infty} \langle \nu_{n_k,\varepsilon_k} - \nu, z_m \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = 0 \qquad \forall m \in \mathbb{N}.$$

Since ν_{n_k,ε_k} is also bounded, cf. (5.27), and $\{z_m\}_{m\in\mathbb{N}}$ is dense in $H_0^1(\Omega)$, it follows that $\nu_{n_k,\varepsilon_k} \rightharpoonup \nu$. The convergence $\lambda_{n_k,\varepsilon_k} \to \bar{\lambda}$ follows from $n_k \geq k$ and the boundedness of $\|\nu_{n_k,\varepsilon_k}^+\|_{H^{-1}(\Omega)}$, cf. (5.27). Finally, $w_{n_k} \rightharpoonup w$ follows from step 1.

Step 7 (Conclusion)**:** From steps 4 to 6, we find

$$\begin{aligned} (y_k, \lambda_k) &\in \mathbb{K}, & (\nu_k, w_k) &\in \mathcal{N}_{\mathbb{K}}^{\text{Fréchet}}(y_k, \lambda_k), \\ y_k &\to y \text{ in } H_0^1(\Omega), & w_k &\rightharpoonup w \text{ in } H_0^1(\Omega), \\ \lambda_k &\to \lambda \text{ in } H^{-1}(\Omega), & \nu_k &\rightharpoonup \nu \text{ in } H^{-1}(\Omega). \end{aligned}$$

Hence, $(\nu, w) \in \mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda})$ according to Lemma 2.4.3.

Our result Theorem 5.3.10 gives us a lower estimate for the limiting normal cone of the set $\mathbb{K}$. In particular, we were able to characterize the intersection $\mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda}) \cap L^p(\Omega) \times H_0^1(\Omega)$ for all $(\bar{y}, \bar{\lambda}) \in \mathbb{K}$. Unfortunately, our lower estimate for $\mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda})$ is rather large. Thus, one cannot hope to obtain good stationarity conditions for (OCOP) using the limiting normal cone if $d \geq 2$. For example, C-stationarity cannot be reached using the limiting normal cone. In the case that $(\bar{y}, \bar{\lambda}) = (0, 0)$ we would obtain that the limiting normal cone is dense in $\mathcal{N}_{\mathbb{K}}^{\text{weak}}(0, 0)$. Note that the limiting normal cone does not need to be closed in infinite dimensions, see [Mordukhovich, 2006, Example 1.7].

Although successful for a large class of potential multipliers $(\nu, w) \in H^{-1}(\Omega) \times H_0^1(\Omega)$, our method could not be applied for multipliers $\nu \in H^{-1}(\Omega) \setminus L^p(\Omega)$ and therefore we were not able to give a full characterization of the limiting normal cone of the set $\mathbb{K}$. For example, if ν is a (positive) measure on the line $(-1, 1) \times \{0\} \subset \Omega = (-1, 1)^2$ then it is not in $L^p(\Omega)$ but it can be in $H^{-1}(\Omega)$. In [Harder, G. Wachsmuth, 2018c, Example 5.1] we show that for such a functional ν that is not in $L^p(\Omega)$ the pair $(\nu, 0) \in H^{-1}(\Omega) \times H_0^1(\Omega)$ can still be contained in the set $\mathcal{N}_{\mathbb{K}}^{\lim}(0, 0)$. We are not aware of a counterexample which would demonstrate that $\mathcal{N}_{\mathbb{K}}^{\lim}(\bar{y}, \bar{\lambda}) \neq \mathcal{N}_{\mathbb{K}}^{\text{weak}}(\bar{y}, \bar{\lambda})$ holds.

6 Inverse optimal control problems

6.1 Problem statement and examples

In this chapter we will consider bilevel optimization problems where Lebesgue spaces play an important role. This means, that the "interesting" complementarity-type constraints in the MPCC reformulation (MPCCR) live in Lebesgue spaces. In comparison, the complementarity-type constraint in (5.3) lives in the Sobolev space $H_0^1(\Omega)$. In general, this chapter will have some parallels to Chapter 5. In particular, Sections 6.1 and 6.2 share most of the structure and some phrases with Sections 5.1 and 5.2.

When choosing instances of bilevel optimization problems in Lebesgue spaces, we would like them to have the right amount of pointwise structure. If the lower level constraint and the lower level objective function have too much of a pointwise structure, it would allow us to solve the lower level optimization problem for each point $\omega \in \Omega$ separately, which would not make for an interesting lower level optimization problem. A good way to mix the information of different points in Ω is to introduce partial differential equations in the constraint. This typically leads to optimal control problems in the lower level optimization problem. In order to achieve complementarity-type constraints in Lebesgue spaces in the MPCC reformulation, we add control constraints to the lower level optimization problem. The partial differential equations that appear in this chapter are all elliptic partial differential equations.

To keep our notation more in line with the typical notation of optimal control, we will often use different variables than in the abstract setting of Chapter 3. A list of translations of mathematical objects and variable names from the setting in Chapter 3 to the setting in Chapter 6 can be found in Table 6.1.1 on page 193.

We consider the following bilevel optimization problem. The lower level optimization problem is a parametrized optimal control problem that is given by

$$\begin{aligned} \min_{y \in \mathcal{Y}, u \in L^2(\Omega)} \quad & f(y,u,\alpha) + \tfrac{\sigma}{2}\|u\|^2_{L^2(\Omega)} \\ \text{s.t.} \quad & Ay - Bu = 0, \\ & u \in U_{\text{ad}}, \end{aligned} \tag{OC(α)}$$

where $\alpha \in \mathbb{R}^n$ is a parameter for some $n \in \mathbb{N}$, $\sigma > 0$ is a constant, $\mathcal{Y}$ is a Banach space, $f : \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n \to \mathbb{R}$ is the parameter-dependent objective function, $A \in \mathbb{L}(\mathcal{Y}, \mathcal{Y}^\star)$, $B \in \mathbb{L}(L^2(\Omega), \mathcal{Y}^\star)$ are bounded linear operators that describe the relationship between

the control u and the state y, and the closed and convex set $U_{\text{ad}} \subset L^2(\Omega)$ of admissible controls is given by

$$U_{\text{ad}} := \{u \in L^2(\Omega) \mid u_a \leq u \leq u_b \text{ a.e. in } \Omega\},$$

where $u_a, u_b : \Omega \to \mathbb{R} \cup \{\infty, -\infty\}$ are measurable functions that act as control constraints. As always, $\Omega \subset \mathbb{R}^d$ is an open and bounded set with $d \in \mathbb{N}$.

The upper level optimization problem is then given by

$$\begin{aligned} \min_{\alpha,y,u} \quad & F(y,u,\alpha) \\ \text{s.t.} \quad & (y,u) \text{ solves (OC}(\alpha)\text{)}, \\ & \alpha \in \Phi_{UL}. \end{aligned} \tag{IOC}$$

Here, $F : \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ is the objective function and $\Phi_{UL} \subset \mathbb{R}^n$ is a closed and convex set. We call (IOC) an *inverse optimal control problem*. Like in Chapter 3 we will use the notation $\psi : \mathbb{R}^n \to \mathcal{Y} \times L^2(\Omega)$ for the solution operator and $\varphi : \mathbb{R}^n \to \mathbb{R}$ for the optimal value function that correspond to the parametrized optimization problem (OC(α)).

We will give some possible examples for components of (IOC) and (OC(α)). The constraint $Ay - Bu = 0$ can be used to model a partial differential equation. For a first example, we set $\mathcal{Y} := H_0^1(\Omega)$ and

$$A := -\Delta \in \mathbb{L}(H_0^1(\Omega), H^{-1}(\Omega)), \qquad B := \mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} \in \mathbb{L}(L^2(\Omega), H^{-1}(\Omega)), \tag{6.1}$$

where $\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)}$ is the natural embedding of $L^2(\Omega)$ into $H^{-1}(\Omega)$. Then the constraint $Ay - Bu = 0$ describes an elliptic partial differential equation, namely Poisson's equation with homogeneous Dirichlet boundary conditions. It is also possible to choose other elliptic operators for A. For example, if we set $\mathcal{Y} = H^1(\Omega)$ and assume that Ω is connected and has a Lipschitz boundary, then it is also possible to consider Poisson's equation with Robin boundary conditions, i.e.

$$\begin{aligned} -\Delta y &= u && \text{on } \Omega, \\ \frac{\partial y}{\partial n} + c_1\, y &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where $c_1 \in L^\infty(\partial\Omega)_+$ is a coefficient with $\|c_1\|_{L^\infty(\partial\Omega)} > 0$. Using the weak formulation, this partial differential equation is described by $Ay - Bu = 0$, where the operators $A \in \mathbb{L}(\mathcal{Y}, \mathcal{Y}^\star)$, $B \in \mathbb{L}(L^2(\Omega), \mathcal{Y}^\star)$ are given by

$$\langle Av_1, v_2\rangle_{H^1(\Omega)^\star \times H^1(\Omega)} = \int_\Omega \nabla v_1 \nabla v_2 \,\mathrm{d}\omega + \int_{\partial\Omega} c_1 v_1 v_2 \,\mathrm{d}\omega \qquad \forall v_1, v_2 \in H^1(\Omega), \tag{6.2a}$$

$$B := \mathcal{I}^\star_{H^1(\Omega)\to L^2(\Omega)}. \tag{6.2b}$$

see [Hinze et al., 2009, Section 1.3.1.1]. There are also alternative choices for the operator B instead of the choices made in (6.1) or (6.2b). For example, if one intents to restrict

the actions of the control u to a measurable subset $\Omega_c \subset \Omega$, then one could choose the operators $B = \mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)}\chi_{\Omega_c}$ or $B = \mathcal{I}^{\star}_{H^1(\Omega)\to L^2(\Omega)}\chi_{\Omega_c}$.

For the function f in the lower level optimization problem we provide several examples. First, we consider the case where the parameters α_i act as weights of various objective functions $h_i : \mathcal{Y} \times L^2(\Omega) \to \mathbb{R}$, i.e. f is given via

$$f(y, u, \alpha) := \sum_{i=1}^{n} \alpha_i h_i(y, u). \tag{6.3}$$

This formulation appears in the inverse optimal control problem discussed in [Harder, G. Wachsmuth, 2018b]. This formulation has the disadvantage that $f(\cdot, \cdot, \alpha)$ is not necessarily convex if $\alpha_i < 0$ for some $i \in \{1, \dots, n\}$ even if h_i is convex. If one is only interested in the case where $f(\cdot, \cdot, \alpha)$ is convex, one would require that h_i is convex for all $i \in \{1, \dots, n\}$ and choose Φ_{UL} such that $\alpha \geq 0$ for all $\alpha \in \Phi_{UL}$. For example, one could choose the unit simplex in $\mathbb{R}^n$ for Φ_{UL}, which was done in [Harder, G. Wachsmuth, 2018b]. If we want to make sure that $f(\cdot, \cdot, \alpha)$ is convex for all $\alpha \in \mathbb{R}^n$ if h_i is convex for all $i \in \{1, \dots, n\}$, we can use a function that is very similar to the one defined in (6.3). The function f that is given by

$$f(y, u, \alpha) := \sum_{i=1}^{n} \alpha_i^2 h_i(y, u) \tag{6.4}$$

has always nonnegative weights. To ensure that the sum of the weights does not exceed 1 we can choose the closed unit ball (with respect to the Euclidean distance) in $\mathbb{R}^n$ for Φ_{UL}. If we would try to convert (6.4) into a problem of type (6.3) using a substitution such as $\hat{\alpha}_i = \alpha_i^2$ then it can happen that the upper level function F is no longer differentiable after the substitution, which is another advantage of the formulation (6.4). On the other hand, it is possible in some situations that the formulation (6.4) leads to more stationary points for the bilevel optimization problem.

We can also use the parameters α_i to act as coefficients for the desired state of a tracking-type objective function. For this case we define f via

$$f(y, u, \alpha) := \tfrac{1}{2}\|Ry - P\alpha\|^2_{\mathcal{M}} - \sigma\langle u, Q\alpha\rangle_{L^2(\Omega)\times L^2(\Omega)} + \tfrac{\sigma}{2}\|Q\alpha\|^2_{L^2(\Omega)}. \tag{6.5}$$

Here, $\mathcal{M}$ is a Hilbert space and $R \in \mathbb{L}(\mathcal{Y}, \mathcal{M})$, $P \in \mathbb{L}(\mathbb{R}^n, \mathcal{M})$, $Q \in \mathbb{L}(\mathbb{R}^n, L^2(\Omega))$, are bounded linear operators. Clearly, f is a quadratic function and $f(\cdot, \cdot, \alpha)$ is convex function for all $\alpha \in \mathbb{R}^n$. If we choose f as in (6.5) then this leads to the inverse optimal control problem discussed in [Dempe, Harder, et al., 2019]. A possible choice for R would be the natural embedding $\mathcal{I}_{H^1_0(\Omega)\to L^2(\Omega)}$ where we chose $\mathcal{M} := L^2(\Omega)$. The feasible set Φ_{UL} can be chosen as a compact polyhedron, e.g. the unit simplex.

As examples for the upper level objective function F we consider

$$F(y, u, \alpha) := \tfrac{1}{2}\|\mathcal{I}_{H^1_0(\Omega)\to L^2(\Omega)}y - y_o\|^2_{L^2(\Omega)} + \tfrac{\hat{\sigma}}{2}\|u - u_o\|^2_{L^2(\Omega)}. \tag{6.6}$$

Here we use $\mathcal{Y} := H_0^1(\Omega)$, and the observed functions $u_o, y_o \in L^2(\Omega)$ and the weight $\hat{\sigma} \geq 0$ are fixed. For $\hat{\sigma} = 0$ this upper level function was used in [Harder, G. Wachsmuth, 2018b]. For $\hat{\sigma} = 1$ this choice for the function F was used in [Dempe, Harder, et al., 2019, Example 2.1].

We comment on the interpretation of the inverse optimal control problem if we use (6.6) for the function F with $\hat{\sigma} = 1$. Suppose that we observe functions $u_o, y_o \in L^2(\Omega)$ and we know that these functions are (possibly perturbed) measurements of the optimal state and control of an optimal control problem such as (OC(α)). We know some things about the structure of the optimal control problem such as Ω and U_{ad}, but we do not know finitely many parameters $\alpha_i \in \mathbb{R}$, $i \in \{1, \dots, n\}$ that influence the objective function of the optimal control problem. In order to identify the actual parameters α_i from the measurements u_o, y_o we choose α such that the error between the measurements and the solutions $(y, u) = \psi(\alpha)$ is minimal. This justifies the name of an inverse optimal control problem for (IOC).

We state some assumptions for the data that appears in (IOC) and (OC(α)).

Assumption 6.1.1. (a) The sets $\Omega_a := \{u_a > -\infty\}$, $\Omega_b := \{u_b < \infty\}$ are measurable and we have $u_a \in L^2(\Omega_a)$, $u_b \in L^2(\Omega_b)$, and $u_a \leq u_b$ a.e. on Ω.

(b) The space $\mathcal{Y}$ is a Hilbert space and the operator $A \in \mathbb{L}(\mathcal{Y}, \mathcal{Y}^\star)$ is invertible.

(c) The regularity constant σ is positive.

(d) The set Φ_{UL} is nonempty, convex, and closed.

(e) The function F is continuously Fréchet differentiable and sequentially weakly lower semi-continuous.

(f) The function $f(\cdot, \cdot, \alpha) : \mathcal{Y} \times L^2(\Omega) \to \mathbb{R}$ is convex for all $\alpha \in \mathbb{R}^n$.

(g) The function f is continuously Fréchet differentiable and its Fréchet derivative is locally Lipschitz continuous.

(h) The partial derivatives f'_y, f'_u are Gâteaux differentiable and their Gâteaux derivatives are continuous in the strong operator topologies of $\mathbb{L}(\mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n, \mathcal{Y}^\star)$ and $\mathbb{L}(\mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n, L^2(\Omega)^\star)$.

(i) The partial derivative f'_α is partially Gâteaux differentiable with respect to y and u and these partial Gâteaux derivatives are continuous in the strong operator topologies of $\mathbb{L}(\mathcal{Y}, \mathbb{R}^n)$ and $\mathbb{L}(L^2(\Omega), \mathbb{R}^n)$.

Note that in parts (h) and (i) of Assumption 6.1.1 we only require continuity in the strong operator topology for the derivatives. To show that this is a weaker condition than requiring that f is twice Fréchet differentiable or that the second partial Gâteaux derivatives f''_{uu}, f''_{yu}, f''_{yy} are continuous, we provide the following example.

Example 6.1.2. Let us define the real-valued function $\pi : \mathbb{R} \to \mathbb{R}$ via

$$\pi(s) := \begin{cases} 0 & s \leq 0, \\ 2s^3 - s^4 & 0 < s < 1, \\ 2s - 1 & s \geq 1. \end{cases}$$

It can be seen that π is convex, twice continuously differentiable, and that its second derivative is bounded. Now we choose the function f via

$$f(y, u, \alpha) := \alpha^\top \alpha \int_\Omega \pi(u(\omega)) \, \mathrm{d}\omega.$$

Then f'_u is not Fréchet differentiable and its Gâteaux derivative is discontinuous at all points $(y, u, \alpha) \in \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ with $\alpha \neq 0$, but the function f still satisfies parts (f) to (i) of Assumption 6.1.1.

Proof. The convexity condition Assumption 6.1.1 (f) follows from the convexity of π.

We define the (nonlinear) Nemytskii operator

$$f_0 : L^2(\Omega) \to L^1(\Omega), \qquad (f_0(u))(\omega) = \pi(u(\omega)) \quad \forall \omega \in \Omega, u \in L^2(\Omega).$$

According to [Krasnoselskii et al., 1976, Theorem 20.3] the function f_0 is Fréchet differentiable and the Fréchet derivative satisfies

$$f'_0 : L^2(\Omega) \to L^2(\Omega) \subset \mathbb{L}(L^2(\Omega), L^1(\Omega)), \qquad (f'_0(u))(\omega) = \pi'(u(\omega)) \quad \forall \omega \in \Omega, u \in L^2(\Omega),$$

where we interpret $f'_0(u) \in L^2(\Omega) \subset \mathbb{L}(L^2(\Omega), L^1(\Omega))$ as a multiplication operator from $L^2(\Omega)$ to $L^1(\Omega)$. Since $\pi' : \mathbb{R} \to \mathbb{R}$ is globally Lipschitz continuous with a Lipschitz constant $C_\pi > 0$ we have

$$|(f'_0(u_1))(\omega) - (f'_0(u_2))(\omega)| = |\pi'(u_1(\omega)) - \pi'(u_2(\omega))| \leq C_\pi |u_1(\omega) - u_2(\omega)|$$

for almost all $\omega \in \Omega$. Squaring and integrating of this inequality yields that $f'_0 : L^2(\Omega) \to L^2(\Omega)$ is (globally) Lipschitz continuous.

Since f can be written as $f(y, u, \alpha) = \alpha^\top \alpha \int_\Omega f_0(u) \, \mathrm{d}\omega$ we obtain from the Fréchet differentiability of f_0 that f is Fréchet differentiable. The partial Fréchet derivatives of f are given by $f'_y(y, u, \alpha) = 0 \in \mathcal{Y}^\star$, $f'_u(y, u, \alpha) = \alpha^\top \alpha f'_0(u) \in L^2(\Omega)$, and $f'_\alpha(y, u, \alpha) = 2 \int_\Omega f_0(u) \, \mathrm{d}\omega \, \alpha^\top \in \mathbb{R}^{1 \times n}$. Since these partial Fréchet derivatives are locally Lipschitz continuous, Assumption 6.1.1 (g) follows.

Let us consider the Nemytskii operator

$$f_2 : L^2(\Omega) \to L^\infty(\Omega), \qquad (f_2(u))(\omega) = \pi''(u(\omega)) \quad \forall \omega \in \Omega, u \in L^2(\Omega).$$

Since $0 \leq \pi''(s) \leq C_\pi$ holds for all $s \in \mathbb{R}$ we have $\|f_2(u)\|_{L^\infty(\Omega)} \leq C_\pi$ for all $u \in L^2(\Omega)$. Let $h \in L^2(\Omega)$ be given. By using Lebesgue's dominated convergence theorem one can show that $u \mapsto f_2(u)h$ is a continuous function from $L^2(\Omega)$ to $L^2(\Omega)$. If we interpret $f_2(u) \subset L^\infty(\Omega) \subset \mathbb{L}(L^2(\Omega), L^2(\Omega))$ as a multiplication operator from $L^2(\Omega)$ to $L^2(\Omega)$, this means that f_2 is continuous in the strong operator topology of $\mathbb{L}(L^2(\Omega), L^2(\Omega))$.

Due to $\|f_2(u)\|_{L^\infty(\Omega)} \leq C_\pi$ for all $u \in L^2(\Omega)$ we can also conclude that the function $(u, h) \mapsto f_2(u)h$ is continuous from $L^2(\Omega) \times L^2(\Omega)$ to $L^2(\Omega)$. Thus, by [Goldberg, Kampowsky, Tröltzsch, 1992, Theorem 8] the Nemytskii operator $f_0' : L^2(\Omega) \to L^2(\Omega)$ is Gâteaux differentiable and the Gâteaux derivative f_0'' satisfies $f_0'' = f_2$. Due to $f_u'(y, u, \alpha) = \alpha^\top \alpha f_0'(u)$ it can then be shown that f_u' is Gâteaux differentiable and that its Gâteaux derivative is continuous in the strong operator topology of $\mathbb{L}(\mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n, L^2(\Omega))$. The same statements also hold trivially for $f_y'(y, u, \alpha) = 0$ and therefore Assumption 6.1.1 (h) holds.

Because of $f_\alpha'(y, u, \alpha) = 2 \int_\Omega f_0(u)\, \mathrm{d}\omega\, \alpha^\top$ we can conclude from our differentiability results for f_0 that f_α' is continuously Fréchet differentiable and thus Assumption 6.1.1 (i) holds.

It remains to show that f_u' is not Fréchet differentiable and that its Gâteaux derivative is not continuous at all points $(y, u, \alpha) \in \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ with $\alpha \neq 0$. Let $(y, u, \alpha) \in \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ with $\alpha \neq 0$ be given. Since continuity of the Gâteaux derivative at (y, u, α) implies Fréchet differentiability at (y, u, α), it suffices to show that f_u' is not Fréchet differentiable at (y, u, α). Suppose that f_u' is Fréchet differentiable at (y, u, α). Due to $f_u'(y, u, \alpha) = \alpha^\top \alpha f_0'(u)$ and $\alpha \neq 0$ we obtain that f_0' is Fréchet differentiable at u. Then, according to [Krasnoselskii et al., 1976, Theorem 20.1] $f_0' : L^2(\Omega) \to L^2(\Omega)$ must be an affine function. Since this is not the case, f_u' is not Fréchet differentiable at (y, u, α).

Let us discuss whether the examples that we talked about previously satisfy Assumption 6.1.1.

Corollary 6.1.3. (a) Suppose that $\mathcal{Y} := H_0^1(\Omega)$ and A is given via (6.1). Then Assumption 6.1.1 (b) is satisfied.

(b) Suppose that $\mathcal{Y} := H^1(\Omega)$, the domain Ω is connected and has a Lipschitz boundary, and A is given via (6.2a), where $c_1 \in L^\infty(\partial\Omega)_+$ is a coefficient with $\|c_1\|_{L^\infty(\partial\Omega)} > 0$. Then Assumption 6.1.1 (b) is satisfied.

(c) Suppose that for each $i \in \{1, \ldots, n\}$ there is a function $h_i : \mathcal{Y} \times L^2(\Omega) \to \mathbb{R}$ that is convex, continuously Fréchet differentiable, twice Gâteaux differentiable, and whose second Gâteaux derivative is continuous in the strong operator topology. Then the function f that is given via (6.4) satisfies parts (f) to (i) of Assumption 6.1.1.

(d) Suppose that $\mathcal{M}$ is a Hilbert space and that $R \in \mathbb{L}(\mathcal{Y}, \mathcal{M})$, $P \in \mathbb{L}(\mathbb{R}^n, \mathcal{M})$, $Q \in \mathbb{L}(\mathbb{R}^n, L^2(\Omega))$, are bounded linear operators. Then the function f that is given via (6.5) satisfies parts (f) to (i) of Assumption 6.1.1.

Chapter 3	Chapter 6
X	$\mathcal{Y} \times L^2(\Omega)$
Y	$\mathcal{Y}^\star \times L^2(\Omega)$
V	$\mathbb{R}^n$
x	(y, u)
p	α
$g(x,p)$	$\begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} \begin{pmatrix} y \\ u \end{pmatrix}$
$\Phi \subset Y$	$\{0\} \times U_{\mathrm{ad}} \subset \mathcal{Y}^\star \times L^2(\Omega)$
Φ_{UL}	Φ_{UL}
$f(x,p)$	$f(y,u,\alpha) + \frac{\sigma}{2}\|u\|^2_{L^2(\Omega)}$
$F(x,p)$	$F(y,u,\alpha)$
(LL(p))	(OC(α))
(UL)	(IOC)
$\lambda \in Y^\star$	$(p, \lambda) \in \mathcal{Y} \times L^2(\Omega)$
$w \in X$	$(\mu, w) \in \mathcal{Y} \times L^2(\Omega)$
$\xi \in Y^\star$	$(\rho, \xi) \in \mathcal{Y} \times L^2(\Omega)$
$\rho \in V^\star$	$z \in \mathbb{R}^n$

Table 6.1.1: Translation of variables and objects.

(e) Suppose that $\mathcal{Y} := H^1_0(\Omega)$, $\hat{\sigma} \geq 0$, $y_o, u_o \in L^2(\Omega)$, and that F is given via (6.6). Then Assumption 6.1.1 (e) is satisfied.

We omit a proof as it is mostly straightforward. We remark that part (b) follows from a combination of [Hinze et al., 2009, Theorem 1.21] and Lemma 2.1.35.

We mention that in order to guarantee the existence of solutions of (IOC), we need the compactness of Φ_{UL} in addition to Assumption 6.1.1, see also Corollary 6.2.2.

There are also possibilities to generalize many results in this chapter further. For example, most results also hold if we use an arbitrary measure space instead of Ω equipped with the Lebesgue measure.

The setting in this chapter fits into the abstract setting that was described in Section 3.3. We provide Table 6.1.1 to list how the spaces, functions, sets, and variables that appeared in the abstract setting of Sections 3.3 and 3.4 are used in the current setting of the bilevel optimization problem (IOC). This also includes variable names for Lagrange multipliers that we use for stationarity conditions in Section 6.2. We mention that we equip the Cartesian product $\mathcal{Y} \times L^2(\Omega)$ of the Hilbert spaces $\mathcal{Y}$ and $L^2(\Omega)$ with the norm given by $\|(y,u)\|_{\mathcal{Y}\times L^2(\Omega)} := (\|y\|^2_{\mathcal{Y}} + \|u\|^2_{L^2(\Omega)})^{1/2}$ for all $y \in \mathcal{Y}$, $u \in L^2(\Omega)$, so that $\mathcal{Y} \times L^2(\Omega)$ is again a Hilbert space.

6.2 Stationarity conditions

We are interested in stationarity conditions for a local minimizer of (IOC). We will show weak stationarity of local minimizers under Assumption 6.1.1 in Theorem 6.2.9 and C-stationarity under additional assumptions in Theorem 6.2.10. As a preparation, we provide some preliminary observations in Section 6.2.1, and formally derive the system of stationarity conditions for (IOC) in Section 6.2.2.

6.2.1 Preliminary observations

We give some preliminary observations. First, we discuss various issues such as the existence of the solution operator ψ or the optimal value function φ. Later, we show that Assumption 3.4.5 from the abstract setting in Section 3.4 is satisfied in our current setting.

Lemma 6.2.1. Suppose that Assumption 6.1.1 is satisfied. Then the following holds.

(a) The objective function $(y,u) \mapsto f(y,u,\alpha) + \frac{\sigma}{2}\|u\|^2_{L^2(\Omega)}$ of (OC(α)) is strongly convex with parameter $\gamma := \sigma(\|A^{-1}B\|^2+1)^{-1} > 0$ on the feasible set of (OC(α)) for all $\alpha \in \mathbb{R}^n$.

(b) The bounded linear operator

$$\begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} : \mathcal{Y} \times L^2(\Omega) \to \mathcal{Y}^\star \times L^2(\Omega),$$

is invertible.

(c) The problem (OC(α)) has a unique solution for each $\alpha \in \mathbb{R}^n$, i.e. the solution operator $\psi : \mathbb{R}^n \to \mathcal{Y} \times L^2(\Omega)$ exists. Additionally, ψ is locally Lipschitz continuous.

(d) The optimal value function φ has values in $\mathbb{R}$ and is continuous.

Proof. For part (a), let $\alpha \in \mathbb{R}^n$ be given and let $(y_1,u_1),(y_2,u_2) \in \mathcal{Y} \times L^2(\Omega)$ be feasible points of (OC(α)). Using $Ay_1 - Bu_1 = Ay_2 - Bu_2 = 0$ we obtain

$$\begin{aligned}
\gamma\|(y_1,u_1)-(y_2,u_2)\|^2_{\mathcal{Y}\times L^2(\Omega)} &= \gamma(\|y_1-y_2\|^2_{\mathcal{Y}} + \|u_1-u_2\|^2_{L^2(\Omega)}) \\
&\le \gamma(\|A^{-1}B(u_1-u_2)\|^2_{\mathcal{Y}} + \|u_1-u_2\|^2_{L^2(\Omega)}) \\
&\le \gamma(\|A^{-1}B\|^2+1)\|u_1-u_2\|^2_{L^2(\Omega)} \\
&= \sigma\langle u_1-u_2, u_1-u_2\rangle_{L^2(\Omega)\times L^2(\Omega)}.
\end{aligned}$$

Thus, by Lemma 2.1.30 (b) the function $(y,u) \mapsto \frac{\sigma}{2}\|u\|^2_{L^2(\Omega)}$ is strongly convex with parameter γ on the feasible set of (OC(α)). Since the function $f(\cdot,\cdot,\alpha)$ is convex and

the sum of a convex and a strongly convex function is strongly convex with the same parameter, the claim follows.

For part (b) it can be easily verified that

$$\begin{bmatrix} A^{-1} & A^{-1}B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} : \mathcal{Y}^\star \times L^2(\Omega) \to \mathcal{Y} \times L^2(\Omega)$$

is the inverse operator, which implies the claim.

For part (c) we can use Lemma 3.1.8 to obtain existence and local Lipschitz continuity of ψ. The assumptions for that lemma are satisfied because of Assumption 6.1.1 as well as part (a) and part (b).

Finally, part (d) follows directly from part (c) and Corollary 3.2.2.

We mention that the operator in part (b) of this lemma corresponds to the operator $g'_x(x,p)$ from Chapter 3 for any $x \in X$, $p \in V$.

If f is given by (6.5), then the function $(y,u,\alpha) \mapsto f(y,u,\alpha) + \frac{\sigma}{2}\|u\|^2_{L^2(\Omega)}$ is (jointly) convex and therefore φ is convex by Proposition 3.2.4.

As a corollary of Lemma 6.2.1 (c) we obtain the existence of minimizers of (IOC) if Φ_{UL} is compact. However, this result does not play a crucial role for our future results on stationarity conditions.

Corollary 6.2.2. Let Assumption 6.1.1 be satisfied and suppose that $\Phi_{UL} \subset \mathbb{R}^n$ is a compact set. Then there exists a (global) minimizer of (IOC).

Proof. Since $\Phi_{UL} \subset \mathbb{R}^n$ is compact and ψ exists and is continuous due to Lemma 6.2.1 (c), there exists a minimizer of the problem

$$\begin{aligned} \min_u \quad & \hat{F}(\alpha) \\ \text{s.t.} \quad & \alpha \in \Phi_{UL}, \end{aligned}$$

where $\hat{F} : \mathbb{R}^n \to \mathbb{R}$ is the reduced upper level objective function that satisfies $\hat{F}(\alpha) = F(y,u,\alpha)$ for all $y \in \mathcal{Y}, u \in L^2(\Omega), \alpha \in \mathbb{R}^n$ with $(y,u) = \psi(\alpha)$. Because this problem is equivalent to (IOC), the claim follows.

For the stationarity results in the abstract setting in Section 3.4 we relied on Assumption 3.4.5. Thus, in order to apply the results from Section 3.4 we need to show that these assumptions are satisfied in the current setting.

Lemma 6.2.3. Suppose that Assumption 6.1.1 is satisfied and that $(\bar{y}, \bar{u}, \bar{\alpha})$ is a local

minimizer of (IOC). Then the list of assumptions in Assumption 3.4.5 is satisfied for the problem (IOC).

Proof. Most of the assumptions listed in Assumption 3.4.5 follow directly from Assumption 6.1.1. Let us discuss the other assumptions. We note that Assumption 3.4.5 (g) was already shown in Lemma 6.2.1 (a). From Assumption 6.1.1 (i) we can use the continuity in the strong operator topologies to obtain that f'_α is Gâteaux differentiable with respect to $(y,u) \in \mathcal{Y} \times L^2(\Omega)$ and this partial Gâteaux derivative is continuous in the strong operator topology. This implies Assumption 3.4.5 (f).

Next, we will show Assumption 3.4.5 (j). Since f is locally Lipschitz continuous we can assume that there are constants $C_L > 0, \varepsilon > 0$ such that C_L is a Lipschitz constant of f on $B_\varepsilon(\bar{y}) \times B_\varepsilon(\bar{u}) \times B_\varepsilon(\bar{\alpha})$. Let $\{y_i\}_{i\in\mathbb{N}} \subset B_\varepsilon(\bar{y}) \subset \mathcal{Y}$, $\{u_i\}_{i\in\mathbb{N}} \subset B_\varepsilon(\bar{u}) \subset L^2(\Omega)$, $\{\alpha_i\}_{i\in\mathbb{N}} \subset B_\varepsilon(\bar{\alpha}) \subset \mathbb{R}^n$ be sequences such that the convergences $y_i \rightharpoonup y_0$, $u_i \rightharpoonup u_0$, and $\alpha_i \to \alpha$ hold for some $y_0 \in B_\varepsilon(\bar{y})$, $u_0 \in B_\varepsilon(\bar{u})$, $\alpha_0 \in B_\varepsilon(\bar{\alpha})$. By the convexity of $f(\cdot,\cdot,\alpha_0)$ we have

$$\begin{aligned} f(y_0,u_0,\alpha_0) &\le \liminf_{i\to\infty} f(y_i,u_i,\alpha_0) \\ &\le \liminf_{i\to\infty} \big(f(y_i,u_i,\alpha_i) + C_L\|\alpha_i-\alpha_0\|\big) \\ &= \liminf_{i\to\infty} f(y_i,u_i,\alpha_i). \end{aligned}$$

If we combine this with the continuity of φ from Lemma 6.2.1 (d) and the convexity of $u \mapsto \|u\|^2_{L^2(\Omega)}$ we obtain

$$f(y_0,u_0,\alpha_0) + \tfrac{\sigma}{2}\|u_0\|^2_{L^2(\Omega)} - \varphi(\alpha_0) \le \liminf_{i\to\infty} \big(f(y_i,u_i,\alpha_i) + \tfrac{\sigma}{2}\|u_i\|^2_{L^2(\Omega)} - \varphi(\alpha_i)\big).$$

Thus, Assumption 3.4.5 (j) is satisfied.

For Assumption 3.4.5 (k) the operator in question is invertible by Lemma 6.2.1 (b) and thus surjective. Finally, Assumption 3.4.5 (l) is satisfied if we choose $r = 2$ (since $\mathbb{R}^n$ is naturally equipped with the Euclidean norm).

6.2.2 Formal derivation of stationarity conditions

In order to see what stationarity conditions would be desirable, we discuss the formal derivation of stationarity conditions. Since this was already done for the abstract setting in Section 3.3.3, we only need to translate the stationarity conditions to our setting. Additionally, we will give possibilities for the (as of yet unspecified) set-valued mapping $\mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}}$ that appeared in (3.23f).

Let us start with calculating the normal cone $\mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}((0,u)) \subset \mathcal{Y} \times L^2(\Omega)$ to the set

$\{0\} \times U_{\mathrm{ad}} \subset \mathcal{Y}^\star \times L^2(\Omega)$ for some function $u \in U_{\mathrm{ad}}$. Due to

$$\mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}((0,u)) = \mathcal{N}_{\{0\}}(0) \times \mathcal{N}_{U_{\mathrm{ad}}}(u) = \mathcal{Y} \times \mathcal{N}_{U_{\mathrm{ad}}}(u) \tag{6.7}$$

we only need to calculate the normal cone to U_{ad}. We do this in the following lemma.

Lemma 6.2.4. Suppose that Assumption 6.1.1 (a) holds and $u \in U_{\mathrm{ad}}$ is a function. Then the normal cone to U_{ad} at u can be described via

$$\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u) \qquad \Leftrightarrow \qquad \lambda \leq 0 \text{ a.e. on } \{u < u_b\} \quad \text{and} \quad \lambda \geq 0 \text{ a.e. on } \{u > u_a\}$$

for all $\lambda \in L^2(\Omega)$.

Proof. Let $\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u)$ be given. Suppose that $\lambda > 0$ and $u < u_b$ a.e. on some Lebesgue measurable set $\Omega_0 \subset \Omega$ with positive measure. Then the function

$$v : \Omega \to \mathbb{R}, \qquad \omega \mapsto \begin{cases} \min(u(\omega) + 1, u_b(\omega)) & \text{if } \omega \in \Omega_0 \\ u(\omega) & \text{if } \omega \notin \Omega_0 \end{cases}$$

satisfies $v \in U_{\mathrm{ad}}$ and $u < v$ on Ω_0. Therefore, $\langle \lambda, v - u \rangle_{L^2(\Omega)\times L^2(\Omega)} > 0$ follows which is a contradiction to $\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u)$. Thus, $\lambda \leq 0$ has to hold a.e on $\{u < u_b\}$. Similarly, one can show that $\lambda \geq 0$ has to hold a.e. on $\{u > u_a\}$. For the other direction, suppose $\lambda \in L^2(\Omega)$ satisfies $\lambda \leq 0$ a.e. on $\{u < u_b\}$ and $\lambda \geq 0$ a.e. on $\{u > u_a\}$. Let $v \in U_{\mathrm{ad}}$ be arbitrary. Because the inclusions $\{u < v\} \subset \{u < u_b\}$ and $\{u > v\} \subset \{u > u_a\}$ hold up to a set of measure zero we have $\lambda(\omega)(v(\omega) - u(\omega)) \leq 0$ a.e. on Ω. By integration, we obtain $\langle \lambda, v - u \rangle_{L^2(\Omega)\times L^2(\Omega)} \leq 0$ and therefore $\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u)$.

Based on the observation (6.7), we can describe the graph of $\mathcal{N}_{\{0\}\times U_{\mathrm{ad}}} \subset \mathcal{Y} \times L^2(\Omega)$ via

$$((\hat{y}, u), (p, \lambda)) \in \operatorname{gph} \mathcal{N}_{\{0\}\times U_{\mathrm{ad}}} \qquad \Leftrightarrow \qquad \hat{y} = 0,\ p \in \mathcal{Y},\ (u, \lambda) \in \operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}},$$

where $\hat{y} \in \mathcal{Y}^\star$, $p \in \mathcal{Y}$, $u, \lambda \in L^2(\Omega)$ are arbitrary. Informally speaking, one could say that $\operatorname{gph} \mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}$ is partially convex in the components $\hat{y} \in \mathcal{Y}^\star$ and $p \in \mathcal{Y}$. Recall that $\mathcal{N}^\sharp$ is supposed to be a generalization of the normal cone. For the components of $\mathcal{N}^\sharp_{\operatorname{gph} \mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}}((\hat{y}, u), (p, \lambda))$ that correspond to $\hat{y} \in \mathcal{Y}^\star$ and $p \in \mathcal{Y}$ it is therefore reasonable to choose sets that agree with the respective normal cones. Thus, we can require that

$$\begin{pmatrix} ((\hat{y}, u), (p, \lambda)) \\ ((\rho, \xi), (\hat{\mu}, \hat{w})) \end{pmatrix} \in \mathcal{N}^\sharp_{\operatorname{gph} \mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}} \quad \Leftrightarrow \quad \begin{array}{llll} \hat{y} = 0, & p \in \mathcal{Y}, & (u, \lambda) \in \operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}, \\ \rho \in \mathcal{Y}, & \hat{\mu} = 0, & (\xi, \hat{w}) \in \mathcal{N}^\sharp_{\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}}((u, \lambda)) \end{array} \tag{6.8}$$

holds for all $\hat{y}, \hat{\mu} \in \mathcal{Y}^\star$, $p, \rho \in \mathcal{Y}$, $u, \lambda, \xi, \hat{w} \in L^2(\Omega)$. Here, $\mathcal{N}^\sharp_{\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}} : L^2(\Omega)^2 \to \mathcal{P}(L^2(\Omega)^2)$ is an unspecified set-valued mapping that can be interpreted as a generalization of a normal cone to the nonconvex set $\operatorname{gph} \mathcal{N}_{U_{\mathrm{ad}}}$.

Let us translate the system (3.23) to our current setting using Table 6.1.1 on page 193. Incorporating the above choices and observations, this results in the system

$$F_y'(\bar{y},\bar{u},\bar{\alpha}) + f_{yy}''(\bar{y},\bar{u},\bar{\alpha})\bar{\mu} + f_{yu}''(\bar{y},\bar{u},\bar{\alpha})\bar{w} + A^\star\bar{\rho} = 0, \tag{6.9a}$$
$$F_u'(\bar{y},\bar{u},\bar{\alpha}) + f_{uy}''(\bar{y},\bar{u},\bar{\alpha})\bar{\mu} + f_{uu}''(\bar{y},\bar{u},\bar{\alpha})\bar{w} + \sigma\bar{w} - B^\star\bar{\rho} + \bar{\xi} = 0, \tag{6.9b}$$
$$F_\alpha'(\bar{y},\bar{u},\bar{\alpha}) + f_{\alpha y}''(\bar{y},\bar{u},\bar{\alpha})\bar{\mu} + f_{\alpha u}''(\bar{y},\bar{u},\bar{\alpha})\bar{w} + \bar{z} = 0, \tag{6.9c}$$
$$f_y'(\bar{y},\bar{u},\bar{\alpha}) + A^\star\bar{p} = 0, \tag{6.9d}$$
$$f_u'(\bar{y},\bar{u},\bar{\alpha}) + \sigma\bar{u} - B^\star\bar{p} + \bar{\lambda} = 0, \tag{6.9e}$$
$$A\bar{y} - B\bar{u} = 0, \tag{6.9f}$$
$$\bar{u} \in U_{\mathrm{ad}} \tag{6.9g}$$
$$\bar{\lambda} \le 0 \text{ a.e. on } \{\bar{u} < u_b\}, \tag{6.9h}$$
$$\bar{\lambda} \ge 0 \text{ a.e. on } \{\bar{u} > u_a\}, \tag{6.9i}$$
$$(\bar{\alpha},\bar{z}) \in \operatorname{gph}\mathcal{N}_{\Phi_{UL}}, \tag{6.9j}$$
$$A\bar{\mu} - B\bar{w} = 0, \tag{6.9k}$$

and

$$((\bar{u},\bar{\lambda}),(\bar{\xi},-\bar{w})) \in \operatorname{gph}\mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}, \tag{6.10}$$

where $\bar{y},\bar{\mu},\bar{p},\bar{\rho} \in \mathcal{Y}$, $\bar{u},\bar{\lambda},\bar{w},\bar{\xi} \in L^2(\Omega)$, and $\bar{\alpha},\bar{z} \in \mathbb{R}^n$. Here, the conditions (6.9a) to (6.9e) come from (3.23a) to (3.23c), the conditions (6.9f) to (6.9i) encode (3.23d) via (6.7) and Lemma 6.2.4, and (6.9j) is the same as (3.23e). However, the conditions (6.9k) and (6.10) are not a direct result of (3.23) but come from applying our choice (6.8) for $\mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_{\{0\}\times U_{\mathrm{ad}}}}$ to (3.23f) with

$$\begin{pmatrix}\hat{\mu}\\ \hat{w}\end{pmatrix} = -\begin{bmatrix}A & -B\\ 0 & \mathrm{id}_{L^2(\Omega)}\end{bmatrix}\begin{pmatrix}\bar{\mu}\\ \bar{w}\end{pmatrix}.$$

In the following definition we define various stationarity conditions for (IOC) and implicitly describe some possible choices for $\mathcal{N}^\sharp_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}$.

Definition 6.2.5. Let $(\bar{y},\bar{u},\bar{\alpha}) \in \mathcal{Y}\times L^2(\Omega)\times\mathbb{R}^n$ be given. We say that $(\bar{y},\bar{u},\bar{\alpha})$ is *weakly stationary* or *W-stationary* for (IOC) if there exist multipliers $\bar{\mu},\bar{p},\bar{\rho} \in \mathcal{Y}$, $\bar{\lambda},\bar{w},\bar{\xi} \in L^2(\Omega)$, $\bar{z} \in \mathbb{R}^n$ such that (6.9) and the conditions

$$\bar{\xi} = 0 \text{ a.e. on } \{u_a < \bar{u} < u_b\}, \tag{6.11a}$$
$$\bar{w} = 0 \text{ a.e. on } \{\bar{\lambda} \ne 0\} \tag{6.11b}$$

are satisfied. We call the point $(\bar{y},\bar{u},\bar{\alpha})$ *C-stationary* if it is W-stationary and additionally the condition

$$\bar{\xi}\bar{w} \ge 0 \text{ a.e. on } \Omega \tag{6.12}$$

holds. The point $(\bar{y},\bar{u},\bar{\alpha})$ is called *M-stationary* if it is W-stationary and the multipliers

$\bar{\xi}, \bar{w}$ satisfy the conditions

$$\bar{\xi}\bar{w} = 0 \ \vee \ (\bar{\xi} \leq 0 \wedge \bar{w} \leq 0) \text{ a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}, \tag{6.13a}$$

$$\bar{\xi}\bar{w} = 0 \ \vee \ (\bar{\xi} \geq 0 \wedge \bar{w} \geq 0) \text{ a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_b\}. \tag{6.13b}$$

Finally, for *strong stationarity* or *S-stationarity* of the point $(\bar{y}, \bar{u}, \bar{\alpha})$ we require that it is W-stationary and the multipliers $\bar{\xi}, \bar{w}$ satisfy the conditions

$$\bar{\xi} \leq 0, \ \bar{w} \leq 0 \text{ a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}, \tag{6.14a}$$

$$\bar{\xi} \geq 0, \ \bar{w} \geq 0 \text{ a.e. on } \{\bar{\lambda} = 0\} \cap \{\bar{u} = u_b\}. \tag{6.14b}$$

We can interpret W-, C-, M-, or S-stationarity as pointwise a.e. analogues of the same concepts in finite dimensions, see Definition 2.5.1. As for the relationship between these stationarity concepts we have the hierarchy

$$\text{S-stationary} \quad \Rightarrow \quad \text{M-stationary} \quad \Rightarrow \quad \text{C-stationary} \quad \Rightarrow \quad \text{W-stationary}.$$

In the case that the so-called biactive sets $\{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}$ and $\{\bar{\lambda} = 0\} \cap \{\bar{u} = u_b\}$ have measure zero, these four notions of stationarity coincide. Since these sets have measure zero in many situations, it can already be useful to show weak stationarity for local minimizers. However, these biactive sets can also be large, see Example 6.2.12, where it is shown that strong stationarity does not need to be satisfied for local minimizers even if the problem data is nice.

Note that the additional conditions for W-, C-, M-, or S-stationarity are equivalent to (6.10) if we choose of $\mathcal{N}^{\sharp}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}$ appropriately for the respective stationarity system. We comment briefly on the relationship to the cones defined in Section 2.4.

Remark 6.2.6. Suppose that $u_a < u_b$ a.e. in Ω.

(a) If we use $\mathcal{N}^{\sharp}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}} = \mathcal{N}^{\text{Clarke}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}$, then (6.10) is equivalent to (6.11).

(b) If we use $\mathcal{N}^{\sharp}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}} = \mathcal{N}^{\text{s-lim}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}$, then (6.10) is equivalent to (6.11) and (6.13).

(c) If we use $\mathcal{N}^{\sharp}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}} = \mathcal{N}^{\text{Fréchet}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}$, then (6.10) is equivalent to (6.11) and (6.14).

(d) We have $\mathcal{N}^{\text{lim}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda)) = \mathcal{N}^{\text{Clarke}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda))$ for all $(u,\lambda) \in \operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}$.

Proof. We only provide a sketch of a proof for these claims. Parts (a) to (c) can be obtained from [Mehlitz, G. Wachsmuth, 2018, Theorem 3.11, Lemma 3.9, and Corollary 3.7].

For part (d) and a fixed $(u,\lambda) \in \operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}$, we first obtain $\operatorname{conv}\mathcal{N}^{\text{s-lim}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda)) \subset \mathcal{N}^{\text{lim}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda)) \subset \mathcal{N}^{\text{Clarke}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda))$ from [Mehlitz, G. Wachsmuth, 2018, Theorem 3.11]. Thus, it suffices to show that $\mathcal{N}^{\text{Clarke}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda)) \subset \operatorname{conv}\mathcal{N}^{\text{s-lim}}_{\operatorname{gph}\mathcal{N}_{U_{\mathrm{ad}}}}((u,\lambda))$ holds. Indeed,

using parts (a) and (b) one can show that all pairs $(v_1, v_2) \in \mathcal{N}^{\text{Clarke}}_{\operatorname{gph}\mathcal{N}_{U_{\text{ad}}}}((u,\lambda))$ satisfy $(2v_1, 0) \in \mathcal{N}^{\text{s-lim}}_{\operatorname{gph}\mathcal{N}_{U_{\text{ad}}}}((u,\lambda))$ and $(0, 2v_2) \in \mathcal{N}^{\text{s-lim}}_{\operatorname{gph}\mathcal{N}_{U_{\text{ad}}}}((u,\lambda))$, which implies $(v_1, v_2) \in \operatorname{conv}\mathcal{N}^{\text{s-lim}}_{\operatorname{gph}\mathcal{N}_{U_{\text{ad}}}}((u,\lambda))$.

The claims in Remark 6.2.6 have analogues in the context of finite-dimensional MPCCs, see Remark 2.5.2.

6.2.3 Weak and C-stationarity for local minimizers

We want to show that C-stationarity holds at a local minimizer by applying the abstract theory from Section 3.4. One concept that is used in the abstract setting is the so-called normal-cone-preserving operator, see Definition 3.4.10. Thus, it is of interest which normal-cone-preserving operators are applicable in the current setting. Despite the complicated structure of U_{ad} we are able to give a full characterization of the set of normal-cone-preserving operators in the following proposition.

Proposition 6.2.7. Suppose that Assumption 6.1.1 (a) is satisfied. Let us define the subsets $\Omega_1 := (\Omega_a \cup \Omega_b) \setminus \{u_a = u_b\}$, $\Omega_2 := \Omega \setminus (\Omega_a \cup \Omega_b)$, $\Omega_3 := \{u_a = u_b\}$. Then the set of normal-cone-preserving operators with respect to U_{ad} is given by the set

$$I_1\, L^\infty(\Omega_1)_+\, I_1^\star + I_2\, \mathbb{L}(L^2(\Omega), L^2(\Omega_2)) + \mathbb{L}(L^2(\Omega_3), L^2(\Omega))\, I_3^\star, \tag{6.15}$$

where we interpret $L^\infty(\Omega_1)_+ \subset L^\infty(\Omega_1) \subset \mathbb{L}(L^2(\Omega_1), L^2(\Omega_1))$ as the set of nonnegative multiplication operators on $L^2(\Omega_1)$, and for $i \in \{1,2,3\}$ the operator $I_i \in \mathbb{L}(L^2(\Omega_i), L^2(\Omega))$ denotes the canonical extension operator that extends functions in $L^2(\Omega_i)$ with zero on $\Omega \setminus \Omega_i$.

In particular, if $\Omega_1 = \Omega$ up to a set of measure zero then the set of normal-cone-preserving operators with respect to U_{ad} is given by the set of nonnegative multiplication operators $L^\infty(\Omega)_+ \subset \mathbb{L}(L^2(\Omega), L^2(\Omega))$.

Proof. First, let us mention that the sets $\Omega_1, \Omega_2, \Omega_3$ constitute a partition of Ω up to sets of measure zero, and $I_i^\star \in \mathbb{L}(L^2(\Omega), L^2(\Omega_i))$ is the canonical restriction operator that restricts functions on Ω to functions Ω_i for $i \in \{1,2,3\}$.

Let $T \in \mathbb{L}(L^2(\Omega), L^2(\Omega))$ be a bounded linear operator from the set described in (6.15), i.e. there exist bounded linear operators $T_2 \in \mathbb{L}(L^2(\Omega), L^2(\Omega_2))$, $T_3 \in \mathbb{L}(L^2(\Omega_3), L^2(\Omega))$ and a nonnegative function $w \in L^\infty(\Omega_1)_+$ such that

$$T = I_1 w I_1^\star + I_2 T_2 + T_3 I_3^\star,$$

where we also interpret $w \in L^\infty(\Omega_1)_+ \subset \mathbb{L}(L^2(\Omega_1), L^2(\Omega_1))$ as a multiplication operator on $L^2(\Omega_1)$.

In order to show that T is a normal-cone-preserving operator, let $u \in U_{\text{ad}}$, $\lambda \in \mathcal{N}_{U_{\text{ad}}}(u)$ be given. According to Lemma 6.2.4 this implies that $\lambda \leq 0$ a.e. on $\{u < u_b\}$ and $\lambda \geq 0$ a.e. on $\{u > u_a\}$. Combining these two properties we obtain $\lambda = 0$ a.e. on $\{u_a < u < u_b\}$ and therefore $\lambda = 0$ a.e. on Ω_2. This implies that $T_2^\star I_2^\star \lambda = 0$ holds a.e. in Ω. Because of $w \geq 0$ a.e. on Ω_1, we also obtain

$$w\lambda \geq 0 \text{ a.e. on } \{u < u_b\} \cap \Omega_1 \qquad \text{and} \qquad w\lambda \leq 0 \text{ a.e. on } \{u > u_a\} \cap \Omega_1.$$

Using the properties of I_1 and the fact that multiplication operators on $L^2(\Omega_1)$ are self-adjoint implies

$$(I_1 w I_1^\star)^\star \lambda \geq 0 \text{ a.e. on } \{u < u_b\}$$

and

$$(I_1 w I_1^\star)^\star \lambda \leq 0 \text{ a.e. on } \{u > u_a\}.$$

Similarly, since $\operatorname{meas}(\{u < u_b\} \cap \Omega_3) = \operatorname{meas}(\{u > u_a\} \cap \Omega_3) = 0$ holds, we obtain

$$I_3 T_3^\star \lambda = 0 \text{ a.e. on } \{u < u_b\} \cup \{u > u_a\}.$$

Combining these results we can conclude that

$$T^\star \lambda \geq 0 \text{ a.e. on } \{u < u_b\} \qquad \text{and} \qquad T^\star \lambda \leq 0 \text{ a.e. on } \{u > u_a\}$$

hold, which implies $T^\star \lambda \in \mathcal{N}_{U_{\text{ad}}}(u)$ due to Lemma 6.2.4.

For the other direction, let $T \in \mathbb{L}(L^2(\Omega), L^2(\Omega))$ be a normal-cone-preserving operator with respect to U_{ad}. We define the function $w := I_1^\star T^\star \chi_{\Omega_1} \in L^2(\Omega_1)$. Our first goal is to show that $w \in L^\infty(\Omega_1)_+$ and that $I_1^\star T^\star I_1$ can be described as the multiplication operator $w \in \mathbb{L}(L^2(\Omega_1), L^2(\Omega_1))$.

Let $\lambda \in L^2(\Omega_1)_+$ be given. We define the functions $\lambda_1, \lambda_2 \in L^2(\Omega_1)_+$ via

$$\lambda_1 := \chi_{\Omega_a} \lambda, \qquad \lambda_2 := \chi_{\Omega_b \setminus \Omega_a} \lambda.$$

Note that $\lambda = \lambda_1 + \lambda_2$ holds. Then we can find functions $u_1, u_2 \in U_{\text{ad}}$ such that

$$\begin{aligned}
\{u_a < u_1 < u_b\} \cap \Omega_1 &= \{\lambda_1 = 0\} \cap \Omega_1, \\
\{u_a = u_1 < u_b\} \cap \Omega_1 &= \{\lambda_1 > 0\} \cap \Omega_1, \\
\{u_a < u_2 < u_b\} \cap \Omega_1 &= \{\lambda_2 = 0\} \cap \Omega_1, \\
\{u_a < u_2 = u_b\} \cap \Omega_1 &= \{\lambda_2 > 0\} \cap \Omega_1
\end{aligned}$$

hold. Thus, by Lemma 6.2.4 we have $-I_1 \lambda_1 \in \mathcal{N}_{U_{\text{ad}}}(u_1)$ and $I_1 \lambda_2 \in \mathcal{N}_{U_{\text{ad}}}(u_2)$. This implies that $-T^\star I_1 \lambda_1 \in \mathcal{N}_{U_{\text{ad}}}(u_1)$ and $T^\star I_1 \lambda_2 \in \mathcal{N}_{U_{\text{ad}}}(u_2)$ hold. Another application of Lemma 6.2.4 yields

$$-T^\star I_1 \lambda_1 \leq 0 \text{ a.e. on } \{u_1 < u_b\} \cap \Omega_1 = \Omega_1,$$

$$
\begin{aligned}
-T^\star I_1 \lambda_1 = 0 \text{ a.e. on } \{u_a < u_1 < u_b\} \cap \Omega_1 = \{\lambda_1 = 0\} \cap \Omega_1,\\
T^\star I_1 \lambda_2 \geq 0 \text{ a.e. on } \{u_a < u_2\} \cap \Omega_1 = \Omega_1,\\
T^\star I_1 \lambda_2 = 0 \text{ a.e. on } \{u_a < u_2 < u_b\} \cap \Omega_1 = \{\lambda_2 = 0\} \cap \Omega_1,
\end{aligned}
$$

where the set equalities hold up to a set of measure zero. Due to $\lambda = \lambda_1 + \lambda_2$ and $\{\lambda = 0\} \cap \Omega_1 = \{\lambda_1 = 0\} \cap \{\lambda_2 = 0\} \cap \Omega_1$ up to a set of measure zero and because $\lambda \in L^2(\Omega_1)_+$ was arbitrary, we obtain the conditions

$$
T^\star I_1 \lambda \geq 0 \text{ a.e. on } \Omega_1 \quad \text{and} \quad T^\star I_1 \lambda = 0 \text{ a.e. on } \{\lambda = 0\} \cap \Omega_1 \qquad \forall \lambda \in L^2(\Omega_1)_+. \tag{6.16}
$$

Let $J_3, J_4 \subset \Omega_1$ be measurable subsets. Then we can apply the above with $\lambda = \chi_{J_3} \in L^2(\Omega_1)_+$ which yields

$$
\operatorname{meas}(J_3 \cap J_4) = 0 \qquad \Rightarrow \qquad \langle T^\star \chi_{J_3}, \chi_{J_4} \rangle_{L^2(\Omega)\times L^2(\Omega)} = 0.
$$

If we use this observation for the choices $J_3 = J_1, J_4 = J_2 \backslash J_1$ and $J_3 = \Omega_1 \backslash J_1, J_4 = J_2 \cap J_1$, where $J_1, J_2 \subset \Omega_1$ are arbitrary measurable subsets, we obtain

$$
\begin{aligned}
\langle T^\star \chi_{J_1}, \chi_{J_2} \rangle_{L^2(\Omega)\times L^2(\Omega)} &= \langle T^\star \chi_{J_1}, \chi_{J_2 \backslash J_1} + \chi_{J_2 \cap J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)}\\
&= \langle T^\star \chi_{J_1}, \chi_{J_2 \cap J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)}\\
&= \langle T^\star (\chi_{\Omega_1} - \chi_{\Omega_1 \backslash J_1}), \chi_{J_2 \cap J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)}\\
&= \langle T^\star \chi_{\Omega_1}, \chi_{J_2 \cap J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)}\\
&= \langle I_1 w, \chi_{J_2 \cap J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)},
\end{aligned}
$$

and thus, by using the pointwise structure of $\langle \cdot, \cdot \rangle_{L^2(\Omega)\times L^2(\Omega)}$,

$$
\langle T^\star \chi_{J_1}, \chi_{J_2} \rangle_{L^2(\Omega)\times L^2(\Omega)} = \langle I_1 w \chi_{J_1}, \chi_{J_2} \rangle_{L^2(\Omega)\times L^2(\Omega)}. \tag{6.17}
$$

For the choices $J_1 = J_2 = \{w \geq \|T\| + 1\} \subset \Omega_1$ this yields

$$
\begin{aligned}
(\|T\| + 1) \operatorname{meas}(J_1) &\leq \int_{\Omega_1} w \chi_{J_1}^2 \, \mathrm{d}\omega = \langle I_1 w \chi_{J_1}, \chi_{J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)}\\
&= \langle T^\star \chi_{J_1}, \chi_{J_1} \rangle_{L^2(\Omega)\times L^2(\Omega)}\\
&\leq \|T\| \|\chi_{J_1}\|_{L^2(\Omega)}^2 = \|T\| \operatorname{meas}(J_1),
\end{aligned}
$$

which is only possible if $\operatorname{meas}(J_1) = 0$. In other words, $w \leq \|T\| + 1$ holds a.e. on Ω_1. If we use (6.16) for the choice $\lambda = \chi_{\Omega_1} \in L^2(\Omega_1)_+$ we get that $w = I_1^\star T^\star \chi_{\Omega_1} \geq 0$ holds a.e. on Ω_1. Thus, we have $0 \leq w \leq \|T\| + 1$ a.e. on Ω_1, which yields $w \in L^\infty(\Omega_1)_+$. Therefore, $I_1 w \in L^\infty(\Omega) \subset \mathbb{L}(L^2(\Omega), L^2(\Omega))$ can be interpreted as a multiplication operator on $L^2(\Omega)$. Since the linear hull of $\{\chi_J \mid J \subset \Omega_1, J \text{ measurable}\}$ is dense in $L^2(\Omega_1)$ and $T^\star, I_1 w \in \mathbb{L}(L^2(\Omega), L^2(\Omega))$ are continuous operators, we can conclude from (6.17) that

$$
\langle T^\star I_1 v_1, I_1 v_2 \rangle_{L^2(\Omega)\times L^2(\Omega)} = \langle (I_1 w)(I_1 v_1), I_1 v_2 \rangle_{L^2(\Omega)\times L^2(\Omega)}
$$

holds for all $v_1, v_2 \in L^2(\Omega_1)$. Therefore, the equality

$$I_1^\star T^\star I_1 = I_1^\star (I_1 w) I_1 = w \in \mathbb{L}(L^2(\Omega_1), L^2(\Omega_1)) \tag{6.18}$$

holds.

For our next step we are interested in properties of $T^\star$ that also involve the sets Ω_2 and Ω_3. Let $u_0 \in L^2(\Omega)$ be defined via

$$u_0(\omega) := \begin{cases} u_a(\omega) & \text{if } \omega \in \Omega_a, \\ u_b(\omega) & \text{if } \omega \in \Omega_b \setminus \Omega_a, \\ 0 & \text{if } \omega \in \Omega_2. \end{cases}$$

Then the normal cone to U_{ad} at u_0 is given by

$$\mathcal{N}_{U_{\mathrm{ad}}}(u_0) = \left\{ \lambda \in L^2(\Omega) \;\middle|\; \begin{array}{l} \lambda \le 0 \text{ a.e. on } \Omega_a \setminus \Omega_3, \\ \lambda \ge 0 \text{ a.e. on } \Omega_b \setminus (\Omega_a \cup \Omega_3), \\ \lambda = 0 \text{ a.e. on } \Omega_2 \end{array} \right\}. \tag{6.19}$$

Note that $\operatorname{lin} \mathcal{N}_{U_{\mathrm{ad}}}(u_0) = \{\lambda \in L^2(\Omega) \mid \lambda = 0 \text{ a.e. on } \Omega_2\}$ holds. Due to $T^\star \mathcal{N}_{U_{\mathrm{ad}}}(u_0) \subset \mathcal{N}_{U_{\mathrm{ad}}}(u_0)$ this implies that $T^\star \lambda = 0$ a.e. on Ω_2 for all $\lambda \in L^2(\Omega)$ with $\lambda = 0$ a.e. on Ω_2. In other words,

$$I_2^\star T^\star I_1 = 0 \qquad \text{and} \qquad I_2^\star T^\star I_3 = 0 \tag{6.20}$$

hold. Next, let $\lambda \in L^2(\Omega)$ be given with $\lambda = 0$ a.e. on $\Omega_1 \cup \Omega_2$. Due to (6.19) we have $\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u_0)$ and $-\lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u_0)$. From $T^\star \lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u_0)$, $-T^\star \lambda \in \mathcal{N}_{U_{\mathrm{ad}}}(u_0)$, and (6.19) we conclude that $T^\star \lambda = 0$ holds a.e. on $(\Omega_a \setminus \Omega_3) \cup (\Omega_b \setminus (\Omega_a \cup \Omega_3)) = \Omega_1$. Since $\lambda \in L^2(\Omega)$ was arbitrary with $\lambda = 0$ a.e. on $\Omega_1 \cup \Omega_2$, it follows that

$$I_1^\star T^\star I_3 = 0 \tag{6.21}$$

holds.

Finally, we define the operators $T_2 \in \mathbb{L}(L^2(\Omega), L^2(\Omega_2))$, $T_3 \in \mathbb{L}(L^2(\Omega_3), L^2(\Omega))$ via

$$T_2^\star = (I_1 I_1^\star + I_2 I_2^\star) T^\star I_2 \qquad \text{and} \qquad T_3^\star := I_3^\star T^\star.$$

We observe that $I_1 I_1^\star + I_2 I_2^\star + I_3 I_3^\star = \mathrm{id}_{L^2(\Omega)}$ holds. If we combine the conditions (6.18), (6.20), and (6.21) with the definition of T_2, T_3, we obtain

$$\begin{aligned} T^\star &= (I_1 I_1^\star + I_2 I_2^\star + I_3 I_3^\star) T^\star = (I_1 I_1^\star + I_2 I_2^\star) T^\star + I_3 T_3^\star \\ &= (I_1 I_1^\star + I_2 I_2^\star) T^\star (I_1 I_1^\star + I_2 I_2^\star + I_3 I_3^\star) + I_3 T_3^\star \\ &= (I_1 I_1^\star + I_2 I_2^\star) T^\star (I_1 I_1^\star + I_3 I_3^\star) + T_2^\star I_2^\star + I_3 T_3^\star \\ &= I_1 I_1^\star T^\star I_1 I_1^\star + T_2^\star I_2^\star + I_3 T_3^\star \end{aligned}$$

$$= I_1 w I_1^\star + T_2^\star I_2^\star + I_3 T_3^\star.$$

By considering the adjoint operator of both sides of this equation, we obtain

$$T = I_1 w I_1^\star + I_2 T_2 + T_3 I_3^\star \in I_1\, L^\infty(\Omega_1)_+\, I_1^\star + I_2\, \mathbb{L}(L^2(\Omega), L^2(\Omega_2)) + \mathbb{L}(L^2(\Omega_3), L^2(\Omega))\, I_3^\star,$$

which completes the proof.

As a direct consequence of this proposition we provide the following corollary, which describes some possible normal-cone-preserving operators to the feasible set $\{0\} \times U_{\text{ad}}$ of (OC(α)).

Corollary 6.2.8. Suppose that Assumption 6.1.1 (a) is satisfied and let $\Omega_0 \subset \Omega$ be an arbitrary measurable subset. Then the following holds.

(a) The multiplication operator

$$\chi_{\Omega_0} : L^2(\Omega) \to L^2(\Omega)$$

is a normal-cone-preserving operator with respect to U_{ad}.

(b) The operator

$$\begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} : \mathcal{Y}^\star \times L^2(\Omega) \to \mathcal{Y}^\star \times L^2(\Omega) \tag{6.22}$$

is a normal-cone-preserving operator with respect to $\{0\} \times U_{\text{ad}} \subset \mathcal{Y}^\star \times L^2(\Omega)$.

Proof. It can easily be checked that χ_{Ω_0} is contained in the set described in (6.15). This implies part (a). Part (b) follows by direct calculations.

Now that we know some normal-cone-preserving operators to the feasible set $\{0\} \times U_{\text{ad}}$, we can apply Theorem 3.4.17. Our first main theorem in this section states that a local minimizer $(\bar{y}, \bar{u}, \bar{\alpha})$ of (IOC) is weakly stationary. Later, this result will be upgraded to C-stationarity.

Theorem 6.2.9. Let Assumption 6.1.1 be satisfied and let $(\bar{y}, \bar{u}, \bar{\alpha}) \in \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ be a local minimizer of (IOC). Then $(\bar{y}, \bar{u}, \bar{\alpha})$ is a weakly stationary point.

Proof. We want to apply Theorem 3.4.17. The required Assumption 3.4.5 is satisfied due to Lemma 6.2.3. Using Lemma 6.2.4 and Table 6.1.1 on page 193, it can be seen that the abstract system (3.32) translates to (6.9) in our setting, with the exception of (6.9k). Thus, by Theorem 3.4.17 there exist multipliers $\bar{\mu}, \bar{p}, \bar{\rho} \in \mathcal{Y}$, $\bar{\lambda}, \bar{w}, \bar{\xi} \in L^2(\Omega)$, $\bar{z} \in \mathbb{R}^n$ such that the conditions (6.9a) to (6.9j) hold.

In order to show (6.9k), we use Theorem 3.4.17 (b). This yields

$$\begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} \begin{pmatrix} \bar{\mu} \\ \bar{w} \end{pmatrix} \in \mathrm{cl}\big(\mathrm{lin}\, \mathcal{T}_{\{0\}\times U_{\mathrm{ad}}}((0,\bar{u}))\big) = \begin{pmatrix} 0 \\ \mathrm{cl}\big(\mathrm{lin}\, \mathcal{T}_{U_{\mathrm{ad}}}(\bar{u})\big) \end{pmatrix}$$

and then (6.9k) follows directly from the first component of this relation.

Next, we will show that (6.11b) holds by using Theorem 3.4.17 (a). For the normal-cone-preserving operator T we choose the normal-cone-preserving operator given in (6.22), where $\Omega_0 \subset \Omega$ is an arbitrary measurable set. Then we have

$$\begin{aligned} 0 &= \left\langle \begin{pmatrix} \bar{p} \\ \bar{\lambda} \end{pmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} \begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} \begin{pmatrix} \bar{\mu} \\ \bar{w} \end{pmatrix} \right\rangle_{(\mathcal{Y}^\star \times L^2(\Omega))^\star \times (\mathcal{Y}^\star \times L^2(\Omega))} \\ &= \langle \bar{\lambda}, \chi_{\Omega_0} \bar{w} \rangle_{L^2(\Omega)\times L^2(\Omega)} \\ &= \int_{\Omega_0} \bar{\lambda}(\omega)\bar{w}(\omega)\, \mathrm{d}\omega. \end{aligned}$$

Since $\Omega_0 \subset \Omega$ is an arbitrary measurable subset of Ω it follows that $\bar{\lambda}\bar{w} = 0$ a.e. on Ω, which is the same as (6.11b).

As a next goal, we want to use Theorem 3.4.17 (d) to show (6.11a). However, since the bounds u_a, u_b can be infinite on a part of the domain Ω, this is more complicated than the proof of (6.11b). We therefore first prove the statement with the additional assumption

$$u_a = 0 \text{ a.e. on } \Omega_a \setminus \Omega_b \qquad \text{and} \qquad u_b = 0 \text{ a.e. on } \Omega_b \setminus \Omega_a. \tag{6.23}$$

In order for the feasible set $\Phi = \{0\} \times U_{\mathrm{ad}}$ to fit into the structure used in Theorem 3.4.17 (d), we define

$$\begin{aligned} Y_0 &:= \mathcal{Y}^\star \times L^2(\Omega \setminus (\Omega_a \cap \Omega_b)), \\ \hat{Y} &:= L^2(\Omega_a \cap \Omega_b), \\ \Phi_0 &:= \{(0,u) \in Y_0 \mid u \geq 0 \text{ a.e. on } \Omega_a \setminus \Omega_b,\ u \leq 0 \text{ a.e. on } \Omega_b \setminus \Omega_a\}, \\ \hat{\Phi} &:= L^2(\Omega_a \cap \Omega_b)_+, \\ y_l &:= u_a \in L^2(\Omega_a \cap \Omega_b), \\ y_u &:= u_b \in L^2(\Omega_a \cap \Omega_b). \end{aligned}$$

Then it can be seen that $\mathcal{Y}^\star \times L^2(\Omega) \cong Y_0 \times \hat{Y}$ and $\{0\} \times U_{\mathrm{ad}} = \Phi_0 \times ((\hat{\Phi} + y_l) \cap (y_u - \hat{\Phi}))$ holds. Clearly, Φ_0 is a closed convex cone and $\hat{\Phi}$ induces a lattice structure on $\hat{Y}$, see Lemma 2.2.1. Again, for the normal-cone-preserving operator T we choose the normal-cone-preserving operator given in (6.22), where $\Omega_0 \subset \Omega$ is an arbitrary measurable set. Let us translate the resulting condition (3.46) into our setting.

If we consider Ω_0 as an arbitrary subset of $\Omega \setminus (\Omega_a \cap \Omega_b)$ then we obtain

$$\begin{aligned} 0 &= \left\langle \begin{pmatrix} \bar{\rho} \\ \bar{\xi} \end{pmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} \begin{pmatrix} A\bar{y} - B\bar{u} \\ \bar{u}\chi_{\Omega \setminus (\Omega_a \cap \Omega_b)} \end{pmatrix} \right\rangle_{(\mathcal{Y}^\star \times L^2(\Omega))^\star \times (\mathcal{Y}^\star \times L^2(\Omega))} \\ &= \langle \bar{\xi}, \chi_{\Omega_0} \bar{u} \rangle_{L^2(\Omega) \times L^2(\Omega)} \\ &= \int_{\Omega_0} \bar{\xi}(\omega) \bar{u}(\omega) \, \mathrm{d}\omega \end{aligned}$$

from (3.46a). Since $\Omega_0 \subset \Omega \setminus (\Omega_a \cap \Omega_b)$ is an arbitrary measurable subset it follows that $\bar{\xi}\bar{u} = 0$ a.e. on $\Omega \setminus (\Omega_a \cap \Omega_b)$. Due to our assumption (6.23) we have $\bar{u} \neq 0$ a.e. on $\{u_a < \bar{u} < u_b\} \cap ((\Omega_a \setminus \Omega_b) \cup (\Omega_b \setminus \Omega_a))$ and thus

$$\bar{\xi} = 0 \text{ a.e. on } \{u_a < \bar{u} < u_b\} \cap ((\Omega_a \setminus \Omega_b) \cup (\Omega_b \setminus \Omega_a)) \tag{6.24}$$

follows. Similarly, if we consider Ω_0 as an arbitrary subset of $\Omega_a \cap \Omega_b$ then we obtain

$$\begin{aligned} 0 &= \left\langle \begin{pmatrix} \bar{\rho} \\ \bar{\xi} \end{pmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} \begin{pmatrix} 0 \\ \chi_{\Omega_a \cap \Omega_b} \min(\bar{u} - u_a, u_b - \bar{u}) \end{pmatrix} \right\rangle_{(\mathcal{Y}^\star \times L^2(\Omega))^\star \times (\mathcal{Y}^\star \times L^2(\Omega))} \\ &= \langle \bar{\xi}, \chi_{\Omega_0} \min(\bar{u} - u_a, u_b - \bar{u}) \rangle_{L^2(\Omega) \times L^2(\Omega)} \\ &= \int_{\Omega_0} \bar{\xi}(\omega) \min(\bar{u}(\omega) - u_a, u_b - \bar{u}(\omega)) \, \mathrm{d}\omega \end{aligned}$$

from (3.46c). Since $\Omega_0 \subset \Omega_a \cap \Omega_b$ is an arbitrary measurable subset it follows that $\bar{\xi} \min(\bar{u} - u_a, u_b - \bar{u}) = 0$ a.e. on $\Omega_a \cap \Omega_b$. This implies

$$\bar{\xi} = 0 \text{ a.e. on } \{u_a < \bar{u} < u_b\} \cap \Omega_a \cap \Omega_b. \tag{6.25}$$

Let us consider the condition (3.46b). It can be calculated that the inclusion

$$\Phi_0^\circ \subset \{(y, u) \in Y_0^\star \mid u = 0 \text{ a.e. on } \Omega \setminus (\Omega_a \cup \Omega_b)\}$$

is satisfied. By taking the closure of the linear hull we obtain

$$\mathrm{cl}(\mathrm{lin}(\Phi_0^\circ)) \subset \{(y, u) \in Y_0^\star \mid u = 0 \text{ a.e. on } \Omega \setminus (\Omega_a \cup \Omega_b)\}.$$

If we now consider the choice $\Omega_0 = \Omega \setminus (\Omega_a \cup \Omega_b)$ then we obtain

$$\begin{pmatrix} 0 \\ \chi_{\Omega_0} \bar{\xi} \end{pmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix}^\star \begin{pmatrix} \bar{\rho} \\ \bar{\xi} \end{pmatrix} \in \mathrm{cl}(\mathrm{lin}(\Phi_0)) \times \hat{Y} \subset \{(y, u) \in \mathcal{Y} \times L^2(\Omega) \mid u = 0 \text{ a.e. on } \Omega_0\}$$

from (3.46b). This implies

$$\bar{\xi} = 0 \text{ a.e. on } \Omega \setminus (\Omega_a \cup \Omega_b). \tag{6.26}$$

Then the condition (6.11a) follows from the combination of (6.24), (6.25), and (6.26).

It remains to remove the additional assumption (6.23). We do this by substituting variables in (IOC) and (OC(α)) in a suitable manner. Let us define the function

$$u_c := u_a \chi_{\Omega_a \setminus \Omega_b} + u_b \chi_{\Omega_b \setminus \Omega_a} \in L^2(\Omega).$$

We then consider the variable substitutions

$$u = \hat{u} + u_c, \qquad y = \hat{y} + A^{-1} B u_c, \tag{6.27}$$

where $\hat{y} \in \mathcal{Y}$ and $\hat{u} \in L^2(\Omega)$ are the new variables. Due to the variable substitution, we need to define modified versions of f, F, U_{ad} via

$$\begin{aligned}
\hat{F}(\hat{y}, \hat{u}, \alpha) &:= F(\hat{y} + A^{-1} B u_c, \hat{u} + u_c, \alpha), \\
\hat{f}(\hat{y}, \hat{u}, \alpha) &:= f(\hat{y} + A^{-1} B u_c, \hat{u} + u_c, \alpha) + \sigma \langle \hat{u}, u_c \rangle_{L^2(\Omega) \times L^2(\Omega)} + \tfrac{\sigma}{2} \|u_c\|^2_{L^2(\Omega)}, \\
\hat{U}_{\text{ad}} &:= U_{\text{ad}} - u_c, \quad \hat{u}_a := u_a - u_c, \quad \hat{u}_b := u_b - u_c.
\end{aligned}$$

These modifications satisfy

$$\begin{aligned}
\hat{F}(\hat{y}, \hat{u}, \alpha) &= F(y, u, \alpha), \\
\hat{f}(\hat{y}, \hat{u}, \alpha) + \tfrac{\sigma}{2} \|\hat{u}\|^2_{L^2(\Omega)} &= f(y, u, \alpha) + \tfrac{\sigma}{2} \|u\|^2_{L^2(\Omega)}, \\
A\hat{y} - B\hat{u} &= Ay - Bu, \\
\hat{u} \in \hat{U}_{\text{ad}} &\Leftrightarrow u \in U_{\text{ad}}
\end{aligned}$$

for all $\hat{y}, y \in \mathcal{Y}, \hat{u}, u \in L^2(\Omega), \alpha \in \mathbb{R}^n$ that satisfy (6.27) and thus the optimization problems (OC(α)) and (IOC) stay the same during the variable substitution. It is also easy to see that the new problem data $\hat{f}, \hat{F}, \hat{U}_{\text{ad}}$ still satisfies Assumption 6.1.1. Finally, due to $\hat{u}_a = 0$ a.e. on $\Omega_a \setminus \Omega_b$ and $\hat{u}_b = 0$ a.e. on $\Omega_b \setminus \Omega_a$ the modified problem setting with the variables $\hat{y}, \hat{u}$ satisfies the assumption (6.23). Thus, the above can be applied to the problem after the variable substitution. It can be checked that the stationarity system consisting of (6.9) and (6.11) also stays the same during the variable substitution. (Note that we do not require any substitutions for the Lagrange multipliers.) Thus, we obtain (6.11a) even without the assumption (6.23).

Most of the work in the proof consists of properly applying the results from the abstract setting. For example, the assumption (6.23) was required because the lower bound u_a is not in $L^2(\Omega)$ if $\Omega \setminus \Omega_a$ has positive measure, but the lower and upper bounds y_l, y_u in Theorem 3.4.17 (d) have to be contained in a Banach space.

So far we have shown weak stationarity. In the next theorem we want to update this result to C-stationarity. However, this requires some additional assumptions.

Theorem 6.2.10. Let Assumption 6.1.1 be satisfied and let $(\bar{y}, \bar{u}, \bar{\alpha}) \in \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ be a local minimizer. Suppose that

(a) the mixed partial second derivatives f''_{yu}, f''_{uy} vanish,

(b) the partial second derivative f''_{uu} satisfies $f''_{uu}(y, u, \alpha) \in L^\infty(\Omega)$ for all (y, u, α) in a neighborhood of $(\bar{y}, \bar{u}, \bar{\alpha})$, i.e. for each (y, u, α) in a neighborhood of $(\bar{y}, \bar{u}, \bar{\alpha})$ there exists a function $w \in L^\infty(\Omega)$ such that

$$\langle f''_{uu}(y, u, \alpha)v_1, v_2\rangle_{L^2(\Omega)\times L^2(\Omega)} = \langle wv_1, v_2\rangle_{L^2(\Omega)\times L^2(\Omega)} \tag{6.28}$$

holds for all $v_1, v_2 \in L^2(\Omega)$.

Furthermore, suppose that at least one of the conditions

(c) the operator $B \in \mathbb{L}(L^2(\Omega), \mathcal{Y}^\star)$ is compact,

(d) the operator $f''_{yy}(\bar{y}, \bar{u}, \bar{\alpha})$ is compact and f''_{yy} is continuous at $(\bar{y}, \bar{u}, \bar{\alpha})$

holds.

Then $(\bar{y}, \bar{u}, \bar{\alpha})$ is a C-stationary point.

Proof. We want to apply Theorem 3.4.17 again. Assumption 3.4.5 is satisfied due to Lemma 6.2.3. Since Theorem 6.2.9 shows already that $(\bar{y}, \bar{u}, \bar{\alpha})$ is a W-stationary point, it remains to show that the multipliers $\bar{\xi}$ and $\bar{w}$ satisfy (6.12). We will do this using Theorem 3.4.17 (e).

As in the proof of Theorem 6.2.9, we choose the operator given in (6.22) for the normal-cone-preserving operator T, where $\Omega_0 \subset \Omega$ is an arbitrary measurable set. We also choose

$$T_1 := \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} : \mathcal{Y} \times L^2(\Omega) \to \mathcal{Y} \times L^2(\Omega)$$

and define the bounded linear operator

$$T_{23} := \begin{bmatrix} 0 & A^{-1}B\chi_{\Omega_0} \\ 0 & 0 \end{bmatrix} : \mathcal{Y} \times L^2(\Omega) \to \mathcal{Y} \times L^2(\Omega).$$

In the case that assumption (c) holds we choose $T_2 := 0$ and $T_3 := T_{23}$, otherwise we choose $T_2 := T_{23}$ and $T_3 := 0$.

We need to check that the assumptions (a) to (d) in Lemma 3.4.15 are satisfied. Due to $T_2 + T_3 = T_{23}$ we have

$$\begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} (T_1 + T_2 + T_3) = \begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} \begin{bmatrix} 0 & A^{-1}B\chi_{\Omega_0} \\ 0 & \chi_{\Omega_0} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} = T \begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix}$$

and thus assumption (a) in Lemma 3.4.15 is satisfied. For assumption (b) in Lemma 3.4.15 we need to check that

$$\langle \chi_{\Omega_0}(f''_{uu}(y,u,\alpha)+\sigma)v, v\rangle_{L^2(\Omega)\times L^2(\Omega)} \geq 0 \tag{6.29}$$

holds for all $v \in L^2(\Omega)$ and all (y,u,α) in a neighborhood of $(\bar{y},\bar{u},\bar{\alpha})$. Indeed, let (y,u,α) in a neighborhood of $(\bar{y},\bar{u},\bar{\alpha})$ and functions $v \in L^2(\Omega)$, $w \in L^\infty(\Omega)$ be given such that (6.28) is satisfied for $v_1 = v_2 = v$. Because $f(y,\cdot,\alpha) : L^2(\Omega) \to \mathbb{R}$ is a convex function we know that w is nonnegative a.e. on Ω. Therefore, $\chi_{\Omega_0}(w+\sigma) \geq 0$ a.e. on Ω and thus (6.29) follows. We continue with checking assumption (c) of Lemma 3.4.15. If $T_2 = 0$ then this condition is trivially satisfied and therefore we consider the case that assumption (d) and $T_2 = T_{23}$ hold. We need to show that the map $q : \mathcal{Y} \times L^2(\Omega)$ which is given via

$$q(\hat{y},\hat{u}) = \left\langle (A^{-1}B\chi_{\Omega_0})^\star f''_{yy}(\bar{y},\bar{u},\bar{\alpha})\hat{y},\hat{u}\right\rangle_{L^2(\Omega)\times L^2(\Omega)}$$

is sequentially weakly lower semi-continuous and that the function

$$\mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n \ni (y,u,\alpha) \mapsto (A^{-1}B\chi_{\Omega_0})^\star f''_{yy}(y,u,\alpha) \in \mathbb{L}(\mathcal{Y}, L^2(\Omega))$$

is continuous at $(\bar{y},\bar{u},\bar{\alpha})$. While the latter follows from the continuity of f''_{yy} at $(\bar{y},\bar{u},\bar{\alpha})$, the function q is sequentially weakly lower semi-continuous because the compact operator $f''_{yy}(\bar{y},\bar{u},\bar{\alpha})$ maps weakly convergent sequences to strongly convergent sequences. For assumption (d) in Lemma 3.4.15 we note that it is trivially true if $T_3 = 0$ holds. Thus, it remains to consider the case that $T_3 = T_{23}$ and assumption (c) hold. Since the operator B is compact we obtain that T_{23} is also compact.

In summary, all the assumptions of Theorem 3.4.17 (e) are satisfied. Thus, an application of Theorem 3.4.17 (e) yields

$$\begin{aligned} 0 &\leq \left\langle \begin{pmatrix} \bar{\rho} \\ \bar{\xi} \end{pmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \chi_{\Omega_0} \end{bmatrix} \begin{bmatrix} A & -B \\ 0 & \mathrm{id}_{L^2(\Omega)} \end{bmatrix} \begin{pmatrix} \bar{\mu} \\ \bar{w} \end{pmatrix} \right\rangle_{(\mathcal{Y}^\star\times L^2(\Omega))^\star\times(\mathcal{Y}^\star\times L^2(\Omega))} \\ &= \langle \bar{\xi}, \chi_{\Omega_0}\bar{w}\rangle_{L^2(\Omega)\times L^2(\Omega)} \\ &= \int_{\Omega_0} \bar{\xi}(\omega)\bar{w}(\omega)\,\mathrm{d}\omega. \end{aligned}$$

Since $\Omega_0 \subset \Omega$ is an arbitrary measurable subset of Ω it follows that $\bar{\xi}\bar{w} \geq 0$ a.e. on Ω. This completes the proof.

We remark that the additional assumptions for Theorem 6.2.10 are satisfied if f is given via (6.5) and B or R are compact. On the other hand, if the function f is given via (6.4) then we would be required to state additional assumption on the functions h_i if

we want to apply Theorem 6.2.10 in this setting. Note that if B is chosen via (6.1) then it is a compact operator, and assumption (d) can be ignored. For example, if B is compact then the function f that was defined in Example 6.1.2 satisfies the assumptions of Theorem 6.2.10. We further remark that assumption (b) of Theorem 6.2.10 requires that $f''_{uu}(\bar{y}, \bar{u}, \bar{\alpha})$ has a pointwise structure. This assumption is needed in order to show that the nonnegativity condition (6.29) holds.

In summary, Theorems 6.2.9 and 6.2.10 demonstrate that the abstract results from Theorem 3.4.17 can be applied in infinite-dimensional bilevel optimization problems in Lebesgue spaces. We mention that Theorem 6.2.10 is also a generalization of [Dempe, Harder, et al., 2019, Theorem 5.2].

There are also some similarities to the C-stationarity result in [Harder, G. Wachsmuth, 2018b, Theorem 4.8]. However, the setting in this article is different from our setting here, and a penalty approach is used for the proof there instead of the regularization approach from Section 3.4.

6.2.4 Counterexample for strong stationarity

In this section we want to show that there is a difference between M-stationarity and strong stationarity and that local minimizers do not need to be strongly stationary. We start with a lemma that shows M-stationarity of some points in a special situation.

Lemma 6.2.11. Let $(\bar{y}, \bar{u}, \bar{\alpha})$ be a feasible point of (IOC) with $\bar{u}(\omega) \in \{u_a(\omega), u_b(\omega)\}$ for almost all $\omega \in \Omega$. Furthermore, we assume that $F'_\alpha(\bar{y}, \bar{u}, \bar{\alpha}) \in -\mathcal{N}_{\Phi_{UL}}(\bar{\alpha})$ holds. Then $(\bar{y}, \bar{u}, \bar{\alpha})$ is an M-stationary point.

Proof. Since $(\bar{y}, \bar{u}, \bar{\alpha})$ is a feasible point of (IOC), we have $\psi(\bar{\alpha}) = (\bar{y}, \bar{u})$. Then the KKT conditions for (OC($\bar{\alpha}$)) are satisfied, i.e. there exist $\bar{\lambda} \in L^2(\Omega)$, $\bar{p} \in \mathcal{Y}$ such that (6.9d) to (6.9i) are satisfied. We set $\bar{w} = 0$ and $\bar{\mu} = 0$. If we also choose $\bar{z} = -F'_\alpha(\bar{y}, \bar{u}, \bar{\alpha})$, then (6.9c) holds. This choice implies that (6.9j) is also satisfied. Since $A^\star$ is a continuously invertible operator, the other multipliers $\bar{\rho} \in \mathcal{Y}$ and $\bar{\xi} \in L^2(\Omega)$ can be uniquely chosen such that (6.9) is satisfied. The conditions (6.11) hold because $\bar{w} = 0$ a.e. on Ω and $\operatorname{meas}(\{u_a < \bar{u} < u_b\}) = 0$. It remains to show that (6.13) holds. Since $\bar{\xi}\bar{w} = 0$ a.e. on Ω this is indeed true.

Let us give an example to show that a local minimizer of (IOC) does not need to be strongly stationary. The data in this example satisfies Assumption 6.1.1 (see also Corollary 6.1.3) and the additional assumptions in Theorem 6.2.10. Therefore, our results imply already C-stationarity of a local minimizer. This example is taken from [Harder, G. Wachsmuth, 2018b, Example 3.4] (with minor modifications) and was constructed by the author of this thesis.

Example 6.2.12. We set $n := 2$, $d := 1$, $\Omega := (-1,1) \subset \mathbb{R}^1$, $\sigma := 1/10$, $u_a := 0$, $u_b := +\infty$, $\mathcal{Y} := H_0^1(\Omega)$, and $\Phi_{UL} := \operatorname{conv}\{(0,1),(1,0)\} \subset \mathbb{R}^2$. Suppose that A, B are given by (6.1), f is given by (6.4) with $h_1(y,u) := \|y+1\|^2_{L^2(\Omega)}$, $h_2(y,u) := \|y-1\|^2_{L^2(\Omega)}$, and F is given by (6.6) with $\hat{\sigma} := 0$ and $y_o := \chi_{(-1,0)} - \chi_{(0,1)} \in L^2(\Omega)$.

Then the point $(\bar{y}, \bar{u}, \bar{\alpha}) := (0, 0, (1/2, 1/2))$ is a global minimizer of (IOC), but it is not a strongly stationary point. However, $(\bar{y}, \bar{u}, \bar{\alpha})$ is an M-stationary point.

Proof. We follow the proof of [Harder, G. Wachsmuth, 2018b, Example 3.4]. First we note that $(\bar{y}, \bar{u})$ is indeed a solution of (OC($\bar{\alpha}$)). Because we also have $\bar{\alpha} \in \Phi_{UL}$ this implies that $(\bar{y}, \bar{u}, \bar{\alpha})$ is a feasible point of (IOC).

Next, we will show that $(\bar{y}, \bar{u}, \bar{\alpha})$ is indeed a (global) minimizer of (IOC). Let (y, u, α) be another feasible point of (IOC). It follows from the symmetry and convexity of the functions h_i and the choice for the operator A that solutions (y, u) of (OC(α)) are even functions, i.e., they satisfy $u(\omega) = u(-\omega), y(\omega) = y(-\omega)$ for a.a. $\omega \in \Omega$. Then, the objective function F of the upper level optimization problem satisfies

$$F(y,u,\alpha) = \frac{1}{2}\|y - y_o\|^2_{L^2(\Omega)} = \frac{1}{2}\int_0^1 |y(\omega)-1|^2 + |y(\omega)+1|^2 \,\mathrm{d}\omega = \int_0^1 y(\omega)^2 + 1 \,\mathrm{d}\omega.$$

This function is minimized if and only if $y = 0$. Therefore, $(\bar{y}, \bar{u}, \bar{\alpha}) = (0, 0, (1/2, 1/2))$ is a global minimizer of (IOC).

For the disproof of strong stationarity, let us assume that $(\bar{y}, \bar{u}, \bar{\alpha})$ is a strongly stationary point, i.e. there exist $\bar{\mu}, \bar{p}, \bar{\rho} \in H_0^1(\Omega)$, $\bar{\lambda}, \bar{w}, \bar{\xi} \in L^2(\Omega)$, $\bar{z} \in \mathbb{R}^2$ such that (6.9), (6.11) and (6.14) are satisfied. Because of $f'_y(\bar{y}, \bar{u}, \bar{\alpha}) = (1/4)(2(\bar{y}+1) + 2(\bar{y}-1)) = 0$ and the invertibility of $A^\star$ we obtain $\bar{p} = 0$ from (6.9d). Then $\bar{\lambda} = 0$ follows from (6.9e) and $\bar{u} = f'_u(\bar{y}, \bar{u}, \bar{\alpha}) = 0$. Therefore, the biactive set $\{\bar{\lambda} = 0\} \cap \{\bar{u} = u_a\}$ is equal to $\Omega = (-1,1)$. Next, we consider the condition (6.9c). Since $F'_\alpha(\bar{y}, \bar{u}, \bar{\alpha})$ and $f''_{\alpha u}(\bar{y}, \bar{u}, \bar{\alpha})$ vanish we obtain

$$-\bar{z} = f''_{\alpha y}(\bar{y}, \bar{u}, \bar{\alpha})\bar{\mu} = \begin{pmatrix} 4\bar{\alpha}_1 \langle \bar{y}+1, \bar{\mu}\rangle_{L^2(\Omega)\times L^2(\Omega)} \\ 4\bar{\alpha}_2 \langle \bar{y}-1, \bar{\mu}\rangle_{L^2(\Omega)\times L^2(\Omega)} \end{pmatrix} = \begin{pmatrix} \langle 2, \bar{\mu}\rangle_{L^2(\Omega)\times L^2(\Omega)} \\ \langle -2, \bar{\mu}\rangle_{L^2(\Omega)\times L^2(\Omega)} \end{pmatrix}. \tag{6.30}$$

It can be calculated that $\mathcal{N}_{\Phi_{UL}}(\bar{\alpha}) = \operatorname{lin}\{(1,1)\}$ and therefore we obtain $\bar{z}_1 = \bar{z}_2$ from (6.9j). Then (6.30) implies $\langle 1, \bar{\mu}\rangle_{L^2(\Omega)\times L^2(\Omega)} = 0$. According to the strong stationarity condition (6.14a) we have $\bar{w} \le 0$ a.e. on Ω. Due to our choices for A, B we obtain $\bar{\mu} \le 0$ a.e. on Ω from (6.9k) and Theorem 2.2.13. Therefore, $\langle 1, \bar{\mu}\rangle_{L^2(\Omega)\times L^2(\Omega)} = 0$ implies $\bar{\mu} = 0$ and $\bar{w} = 0$. Since we have $\bar{\xi} \le 0$ a.e. on Ω by (6.14a) and we can obtain $\bar{\xi} = B^\star\bar{\rho} = \mathcal{I}_{H_0^1(\Omega)\to L^2(\Omega)}\bar{\rho}$ from (6.9b), we know that $\bar{\rho} \le 0$ holds a.e. on Ω. Now it follows from (6.9a) that

$$-A^\star\bar{\rho} = F'_y(\bar{y}, \bar{u}, \bar{\alpha}) = -\mathcal{I}_{L^2(\Omega)\to H^{-1}(\Omega)} y_o = -\chi_{(-1,0)} + \chi_{(0,1)} \in H^{-1}(\Omega)$$

holds. Because of $A^\star \bar{\rho} = -\bar{\rho}''$ we can therefore directly calculate that $\bar{\rho}(\omega) = \omega(|\omega| - 1)/2$ holds for a.a. $\omega \in \Omega$. This is a contradiction to $\bar{\rho} \leq 0$ a.e. on Ω. Therefore, $(\bar{y}, \bar{u}, \bar{\alpha})$ is not strongly stationary.

Finally, since $\bar{u} = u_a$ a.e. in Ω and $F'_\alpha(\bar{y}, \bar{u}, \bar{\alpha}) = 0$ we obtain that $(\bar{y}, \bar{u}, \bar{\alpha})$ is an M-stationary point due to Lemma 6.2.11.

We were not able to find an example of a local minimizer of (IOC) that is also not M-stationary. Thus, it remains an open question whether M-stationarity can be shown under reasonable assumptions similar to our proof of C-stationarity, or whether a counterexample for M-stationarity can be found.

6.3 A discretized version of a bilevel optimal control problem

In this section, we will look at a discretized version of (IOC) and estimate the error between the solutions of the discretized and the original undiscretized version. We first investigate this theoretically in a more general setting in Section 6.3.1. Then we consider the situation for more specific examples in Section 6.3.2, and we also conduct some numerical experiments in Section 6.3.3.

6.3.1 General discretization error estimates

We use the setting and notation of Section 6.1. Let us consider a discretization of the bilevel optimization problem (IOC). For a discretization parameter $h \in (0, 1)$, the discretized lower level optimization problem is given by

$$\begin{aligned} \min_{y_h \in \mathcal{Y}_h, u_h \in \mathcal{U}_h} \quad & f_h(y_h, u_h, \alpha) + \tfrac{\sigma}{2}\|u_h\|^2_{L^2(\Omega)} \\ \text{s.t.} \quad & A_h y_h - B_h u_h = 0, \\ & u_h \in U_{\text{ad},h}. \end{aligned} \tag{OC(α, h)}$$

Here, $\mathcal{Y}_h \subset \mathcal{Y}$ and $\mathcal{U}_h \subset L^2(\Omega)$ are finite-dimensional spaces that are equipped with the norms of $\mathcal{Y}$ and $L^2(\Omega)$, $f_h : \mathcal{Y}_h \times \mathcal{U}_h \times \mathbb{R}^n \to \mathbb{R}$ is a discretization of the function f, the operators $A_h \in \mathbb{L}(\mathcal{Y}_h, \mathcal{Y}_h^\star)$, $B_h \in \mathbb{L}(\mathcal{U}_h, \mathcal{Y}_h^\star)$ are discretizations of the operators A, B, and $U_{\text{ad},h} \subset \mathcal{U}_h$ is a discretization of U_{ad}. The discretized upper level optimization problem is then given by

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^n,\, y_h \in \mathcal{Y}_h,\, u_h \in \mathcal{U}_h} \quad & F_h(y_h, u_h, \alpha) \\ \text{s.t.} \quad & (y_h, u_h) \text{ solves (OC}(\alpha, h)\text{)}, \\ & \alpha \in \Phi_{UL}. \end{aligned} \tag{IOC(h)}$$

Here, $F_h : \mathcal{Y}_h \times \mathcal{U}_h \times \mathbb{R}^n \to \mathbb{R}$ is a discretization of the objective function F. Typically, the parameter h refers to a mesh size. A possible choice for $U_{\text{ad},h}$ would be $U_{\text{ad}} \cap \mathcal{U}_h$.

For each $h \in (0,1)$, we denote the solution operator for the discretized optimization problem (OC(α, h)) by $\psi^h(\alpha)$. In this section, we also use the notation $\psi_y : \mathbb{R}^n \to \mathcal{Y}$, $\psi_u : \mathbb{R}^n \to L^2(\Omega)$, $\psi_y^h : \mathbb{R}^n \to \mathcal{Y}_h$, $\psi_u^h : \mathbb{R}^n \to \mathcal{U}_h$ to refer to the y- and u-components of the solution operators ψ and ψ^h.

Our approach is as follows. We first give some precise assumptions for the discretized bilevel optimization problem (IOC(h)). These also include assumptions for the discretization errors of the solution of the discretized optimal control problem (OC(α, h)). Depending on the type of discretization (and other aspects), various convergence orders for the solutions of (OC(α, h)) can be found in the literature. Therefore, we do not fully specify the order of the convergence in our assumptions, but use (unspecified) constants instead. This allows us to track the influence of the various exponents in our main result in Theorem 6.3.3. Later, in Section 6.3.2 we discuss the situation for more specific examples and state the resulting order of convergence of the solutions of (OC(α, h)) for these specific examples.

Assumption 6.3.1. (a) The set $U_{\mathrm{ad},h} \subset \mathcal{U}_h \subset L^2(\Omega)$ is nonempty, convex, and closed for every $h \in (0,1)$.

(b) The function $f_h(\cdot,\cdot,\alpha) : \mathcal{Y}_h \times \mathcal{U}_h \to \mathbb{R}$ is convex for all $\alpha \in \mathbb{R}^n$, $h \in (0,1)$.

(c) The set $\Phi_{UL} \subset \mathbb{R}^n$ is compact.

(d) There exist constants $c_1, c_2, c_3, C > 0$ and a constant $C_k > 0$ for every $k \in \mathbb{N}$ such that for all $h \in (0,1)$ and $\alpha \in \Phi_{UL}$, $y_h \in \mathcal{Y}_h$, $u_h \in \mathcal{U}_h$ with $\max(\|y_h\|, \|u_h\|) < k$ the estimates

$$\|\psi_y^h(\alpha) - \psi_y(\alpha)\|_{\mathcal{Y}} \leq C h^{c_1}, \tag{6.31}$$

$$\|\psi_u^h(\alpha) - \psi_u(\alpha)\|_{L^2(\Omega)} \leq C h^{c_2}, \tag{6.32}$$

$$|F_h(y_h, u_h, \alpha) - F(y_h, u_h, \alpha)| \leq C_k h^{c_3} \tag{6.33}$$

hold.

Note that we implicitly used the existence of $\psi_y^h : \mathbb{R}^n \to \mathcal{Y}_h$ and $\psi_h^h : \mathbb{R}^n \to \mathcal{U}_h$ in part (d). As the next corollary states, this follows already from parts (a) and (b) of Assumption 6.3.1. The following corollary is a direct consequence of Lemma 2.3.2 (b).

Corollary 6.3.2. Suppose that parts (a) and (b) of Assumption 6.3.1 are satisfied. Then the solution operator $\psi^h : \mathbb{R}^n \to \mathcal{Y}_h \times \mathcal{U}_h$ exists.

Now we can move on to the main result of this section. We give an error estimate for the solutions of the discretized bilevel optimization problem (IOC(h)) under a growth condition. The resulting convergence order depends on the exponents in Assumption 6.3.1 (d) and on the exponent in the growth condition (6.34). Because there might be multiple solutions of (IOC), we consider the distance to the set of optimal solutions of (IOC).

Theorem 6.3.3. Suppose that Assumptions 6.1.1 and 6.3.1 are satisfied. Let $K \subset \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$ be the set of optimal solutions of (IOC) and let $\bar{F}$ denote the optimal value of (IOC). We assume that the growth condition

$$F(y, u, \alpha) \geq \bar{F} + C \operatorname{dist}((x, y, u), K)^{c_4} \tag{6.34}$$

holds for all feasible α, y, u of (IOC), where $C > 0$, $c_4 \geq 1$ are constants. Then, for all sufficiently small $h > 0$, every global solution $(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h)$ of (IOC(h)) satisfies

$$\operatorname{dist}((\bar{y}_h, \bar{u}_h, \bar{\alpha}_h), K) \leq \hat{C} h^{\min(c_1, c_2, c_3)/c_4} \tag{6.35}$$

with a suitable constant $\hat{C} > 0$. Moreover, if F is quadratic, then (6.35) holds for all $h \in (0, 1)$.

Proof. Let

$$K_1 := \{(\psi_y(\alpha), \psi_u(\alpha), \alpha) \mid \alpha \in \Phi_{UL}\} \subset \mathcal{Y} \times L^2(\Omega) \times \mathbb{R}^n$$

be the feasible set of (IOC). Since Φ_{UL} is compact and ψ is continuous due to Lemma 6.2.1 (c), the set K_1 is also compact. Because F is locally Lipschitz continuous and K_1 is compact, there exists an $\varepsilon_0 > 0$ such that F is Lipschitz continuous on $K_1 + B_{\varepsilon_0}(0)$ with Lipschitz constant C_F, see Lemma 2.1.1. Due to (6.31) and (6.32) we know that $(\psi_y^h(\alpha), \psi_u^h(\alpha), \alpha) \in K_1 + B_{\varepsilon_0}(0)$ holds for all $\alpha \in \Phi_{UL}$ and sufficiently small $h > 0$. Therefore, we obtain the estimate

$$\begin{aligned} |F(\psi_y^h(\alpha), \psi_u^h(\alpha), \alpha) - F(\psi_y(\alpha), \psi_u(\alpha), \alpha)| &\leq C_F(\|\psi_y^h(\alpha) - \psi_y(\alpha)\| + \|\psi_u^h(\alpha) - \psi_u(\alpha)\|) \\ &\leq C_F C(h^{c_1} + h^{c_2}) \leq 2C_F C h^{\min(c_1, c_2)} \end{aligned} \tag{6.36}$$

for all $\alpha \in \Phi_{UL}$, $h \in (0, h_0)$, where $h_0 \in (0, 1]$ is sufficiently small. Moreover, if F is quadratic, then it is globally Lipschitz continuous on each bounded set and therefore (6.36) holds for all $h \in (0, 1)$, i.e. we can choose $h_0 = 1$. In both cases, there is a constant $k \in \mathbb{N}$ such that $\|\psi^h(\alpha)\| < k$ holds for all $\alpha \in \Phi_{UL}$, $h \in (0, h_0)$. Therefore, by combining (6.33) and (6.36), we obtain the inequality

$$|F_h(\psi_y^h(\alpha), \psi_u^h(\alpha), \alpha) - F(\psi_y(\alpha), \psi_u(\alpha), \alpha)| \leq (C_k + 2C_F C) h^{\min(c_1, c_2, c_3)} \tag{6.37}$$

for all $\alpha \in \Phi_{UL}$, $h \in (0, h_0)$. Now, let $h \in (0, h_0)$, $\varepsilon > 0$ be fixed and let $(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h)$ be a global solution of (IOC(h)). Furthermore, let $(\bar{y}, \bar{u}, \bar{\alpha}) \in K$ be a global solution of (IOC) such that

$$\|(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h) - (\bar{y}, \bar{u}, \bar{\alpha})\| \leq \operatorname{dist}((\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h), K) + \varepsilon$$

holds. Because $(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h)$ is a feasible point of (IOC), the growth condition (6.34) yields

$$C(\|(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h) - (\bar{y}, \bar{u}, \bar{\alpha})\| - \varepsilon)^{c_4} \leq F(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h) - \bar{F}. \tag{6.38}$$

By using the optimality of $(\bar{y}, \bar{u}, \bar{\alpha})$ and $(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h)$ and by applying the inequality (6.37) twice we obtain the chain of inequalities

$$\begin{aligned} F(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h) &\leq F_h(\psi_y^h(\bar{\alpha}_h), \psi_u^h(\bar{\alpha}_h), \bar{\alpha}_h) + (C_k + 2C_F C)h^{\min(c_1,c_2,c_3)} \\ &= F_h(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h) + (C_k + 2C_F C)h^{\min(c_1,c_2,c_3)} \\ &\leq F_h(\psi_y^h(\bar{\alpha}), \psi_u^h(\bar{\alpha}), \bar{\alpha}) + (C_k + 2C_F C)h^{\min(c_1,c_2,c_3)} \\ &\leq F(\psi_y(\bar{\alpha}), \psi_u(\bar{\alpha}), \bar{\alpha}) + 2(C_k + 2C_F C)h^{\min(c_1,c_2,c_3)} \\ &= \bar{F} + 2(C_k + 2C_F C)h^{\min(c_1,c_2,c_3)}. \end{aligned}$$

If we combine this with (6.38) we obtain

$$\|(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h) - (\bar{y}, \bar{u}, \bar{\alpha})\| - \varepsilon \leq C_0 h^{\min(c_1,c_2,c_3)/c_4}$$

for a suitable constant $C_0 > 0$. By using (6.31) and (6.32) this implies

$$\begin{aligned} \operatorname{dist}((\bar{y}_h, \bar{u}_h, \bar{\alpha}_h), K) &\leq \|(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h) - (\bar{y}, \bar{u}, \bar{\alpha})\| \\ &= \|(\psi_y^h(\bar{\alpha}_h), \psi_u^h(\bar{\alpha}_h), \bar{\alpha}_h) - (\bar{y}, \bar{u}, \bar{\alpha})\| \\ &\leq \|(\psi_y(\bar{\alpha}_h), \psi_u(\bar{\alpha}_h), \bar{\alpha}_h) - (\bar{y}, \bar{u}, \bar{\alpha})\| + Ch^{c_1} + Ch^{c_2} \\ &\leq C_0 h^{\min(c_1,c_2,c_3)/c_4} + \varepsilon + Ch^{c_1} + Ch^{c_2} \\ &\leq \hat{C} h^{\min(c_1,c_2,c_3)/c_4} + \varepsilon, \end{aligned}$$

where $\hat{C} > 0$ is a suitable constant. Finally, since $\varepsilon > 0$ can be chosen arbitrarily small, we obtain

$$\operatorname{dist}((\bar{y}_h, \bar{u}_h, \bar{\alpha}_h), K) \leq \hat{C} h^{\min(c_1,c_2,c_3)/c_4}.$$

6.3.2 Discretization error estimates for PDE-based examples

Let us discuss possible values for the convergence orders in Assumption 6.3.1 (d) for some more specific examples. If the lower level optimization problem is a typical PDE-constrained optimal control problem, then we can use some results from the literature to estimate the convergence order in (6.31) and (6.32). These estimates can be used to apply Theorem 6.3.3 to some instances of the bilevel optimization problem (IOC).

We consider the case that $\mathcal{Y} = H_0^1(\Omega)$ and that $A : H_0^1(\Omega) \to H^{-1}(\Omega)$, $B : L^2(\Omega) \to H^{-1}(\Omega)$ describe Poisson's equation with homogeneous Dirichlet boundary conditions as in (6.1). We discretize the problems (OC(α)) and (IOC) using the finite element method. There are various possibilities for the choice of finite elements. We mainly focus on the case where we discretize the control $u \in L^2(\Omega)$ using piecewise constant functions, and later give some brief remarks for the case of piecewise linear controls. The state $y \in H_0^1(\Omega)$ will be discretized using piecewise linear and continuous functions. Note that we will not always provide all the technical details of the finite element method in

this section. We start with providing error estimates for an example of the lower level optimization problem.

Example 6.3.4. Suppose that $d := 2$, $\Omega \subset \mathbb{R}^2$ is a convex and polygonal domain, $\mathcal{Y} := H_0^1(\Omega)$, the operators A, B are given by (6.1), the function f is given via (6.5) with $\mathcal{M} := L^2(\Omega)$, $R := \mathcal{I}_{H_0^1(\Omega)\to L^2(\Omega)}$, $Q := 0$, image $P \subset C^1(\operatorname{cl}\Omega)$, and that the control constraints u_a, u_b are constant functions with $-\infty < u_a < u_b < \infty$.

We create a discretization (OC(α, h)) of the optimal control problem (OC(α)) using finite elements with mesh size $h \in (0,1)$, i.e. the domain Ω is triangulated. We assume that the family of triangulations is sufficiently regular, see [Ciarlet, 1991, (H1)]. For the space $\mathcal{U}_h \subset L^2(\Omega)$ we use the space of functions that are constant on each triangle, and for the space $\mathcal{Y}_h \subset H_0^1(\Omega)$ we use the space of function that are affine on each triangle, continuous on $\operatorname{cl}\Omega$, and obey the homogeneous Dirichlet boundary conditions. The operators A_h, B_h are created using the usual finite element discretization of Poisson's equation. Further, we use $U_{\mathrm{ad},h} := U_{\mathrm{ad}} \cap \mathcal{U}_h$ for the discrete admissible set. The function f is discretized using

$$f_h(y_h, u_h, \alpha) := \tfrac{1}{2}\|y_h - \Pi_h P\alpha\|^2_{L^2(\Omega)},$$

where $\Pi_h : C^1(\operatorname{cl}\Omega) \to \mathcal{Y}_h \subset L^2(\Omega)$ is the interpolation operator that preserves the function value on all nodes of the finite element mesh.

Moreover, suppose that Φ_{UL} is compact. Then there is a constant $C > 0$ such that

$$\|\psi_y^h(\alpha) - \psi_y(\alpha)\|_{\mathcal{Y}} + \|\psi_u^h(\alpha) - \psi_u(\alpha)\|_{L^2(\Omega)} \le Ch \tag{6.39}$$

holds for all $h \in (0,1)$, $\alpha \in \Phi_{UL}$.

Proof. In [Tröltzsch, 2010, Section 2.2], the problem is studied with $f(y,u,\alpha) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)}$ and $f_h(y_h,u_h,\alpha) = \frac{1}{2}\|y_h - y_d\|^2_{L^2(\Omega)}$, where $y_d \in L^2(\Omega)$ is fixed. Then the estimate $\|\bar{u} - \bar{u}_h\| \le Ch$ is shown, where $\bar{u} \in L^2(\Omega)$, $\bar{u}_h \in \mathcal{U}_h$ are the u-components of the respective solutions, see [Tröltzsch, 2010, (17)]. Moreover, the constant $C > 0$ does not depend on the precise value of y_d, but only depends on an upper bound of $\|y_d\|_{L^2(\Omega)}$.

However, for our purposes we also want to discretize y_d. Suppose that $y_d \in C^1(\operatorname{cl}\Omega) \subset L^2(\Omega)$. Then the estimate

$$\|y_d - \Pi_h y_d\|_{L^2(\Omega)} \le Ch \tag{6.40}$$

can be shown, where $C > 0$ can depend on an upper bound of $\||\nabla y_d|\|_{L^\infty(\Omega)}$ but does not depend on h. Since f_h is quadratic, we can use Lemma 3.1.6 to show that the solution $\bar{u}_h$ has a (global) Lipschitz dependence on y_d. Thus, the solution $\bar{u}_h$ still satisfies the estimate

$$\|\bar{u} - \bar{u}_h\| \le Ch$$

even if we use $f_h(y_h,u_h,\alpha) = \frac{1}{2}\|y_h - \Pi_h y_d\|^2_{L^2(\Omega)}$ instead of $f_h(y_h,u_h,\alpha) = \frac{1}{2}\|y_h - y_d\|^2_{L^2(\Omega)}$. We can use these observations for $y_d := P\alpha \in C^1(\operatorname{cl}\Omega)$, where $\alpha \in \Phi_{UL}$. Since Φ_{UL} is

compact, the constant $C > 0$ can be chosen such that the error does not depend on α. Therefore, we obtain the estimate

$$\|\psi_u^h(\alpha) - \psi_u(\alpha)\|_{L^2(\Omega)} \leq Ch \qquad \forall \alpha \in \Phi_{UL} \tag{6.41}$$

for a suitable constant $C > 0$. Note that (under a regularity assumption for the family of triangulations) the estimate $\|A_h^{-1}B_h - A^{-1}B\|_{\mathbb{L}(\mathcal{U}_h,\mathcal{Y})} \leq Ch$ can be obtained from the literature, see, e.g., [Ciarlet, 1991, Theorem 18.1]. Therefore, we obtain

$$\begin{aligned}
\|\psi_y^h(\alpha) - \psi_y(\alpha)\|_{\mathcal{Y}} &= \|A_h^{-1}B_h\psi_u^h(\alpha) - A^{-1}B\psi_u(\alpha)\|_{\mathcal{Y}} \\
&\leq \|A_h^{-1}B_h\psi_u^h(\alpha) - A^{-1}B\psi_u^h(\alpha)\|_{\mathcal{Y}} + \|A^{-1}B\psi_u^h(\alpha) - A^{-1}B\psi_u(\alpha)\|_{\mathcal{Y}} \\
&\leq \|A_h^{-1}B_h - A^{-1}B\|_{\mathbb{L}(\mathcal{U}_h,\mathcal{Y})}\|\psi_u^h(\alpha)\|_{\mathcal{U}_h} + Ch \\
&\leq Ch \sup\{\|\psi_u^h(\alpha)\|_{L^2(\Omega)} \mid \alpha \in \Phi_{UL}\} + Ch.
\end{aligned}$$

Then the claim follows by adding the inequality (6.41).

Let us continue the example with the discretization error estimate for the upper level optimization problem. This consists mostly of an application of Theorem 6.3.3. An important assumption for this theorem is the growth condition (6.34). Although linear growth can happen in some cases, it is more realistic to only expect quadratic growth.

Example 6.3.5. We consider the setting of Example 6.3.4. Additionally, we assume that the upper level objective function F is given by

$$F(y, u, \alpha) := \tfrac{1}{2}\|y - y_o\|^2_{L^2(\Omega)} + z^\top \alpha,$$

where $y_o \in C^1(\operatorname{cl}\Omega)$ and $z \in \mathbb{R}^n$ are fixed. The discretization is then given by

$$F_h(y_h, u_h, \alpha) := \tfrac{1}{2}\|y_h - \Pi_h y_o\|^2_{L^2(\Omega)} + z^\top \alpha,$$

where $\Pi_h : C^1(\operatorname{cl}\Omega) \to \mathcal{Y}_h \subset L^2(\Omega)$ is the same interpolation operator as in Example 6.3.4.

Let $K \subset H_0^1(\Omega) \times L^2(\Omega) \times \mathbb{R}^n$ be the set of optimal solutions of (IOC). We assume that a quadratic growth condition holds for (IOC), i.e. that (6.34) holds with $c_4 = 2$. Then, for all $h \in (0, 1)$, the error estimate

$$\operatorname{dist}((\bar{y}_h, \bar{u}_h, \bar{\alpha}_h), K) \leq \hat{C}h^{1/2}$$

holds for every global solution $(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h)$ of (IOC(h)), where $\hat{C}$ is a suitable constant.

Proof. Since the interpolation error estimate (6.40) can also be used for y_o instead of y_d, the estimate (6.33) holds with $c_3 = 1$. Together with the observations in Example 6.3.4

we therefore know that Assumption 6.3.1 is satisfied. Hence, we can apply Theorem 6.3.3 with $c_1 = c_2 = c_3 = 1$ and $c_4 = 2$, which yields the claim.

Thus, we obtain a convergence order of $h^{1/2}$ for a PDE-based example of (IOC), where we discretized the controls using piecewise constant functions. If one is able to show a better convergence order for the error estimates in Assumption 6.3.1 (d), then a better convergence order can be obtained for solutions of the upper level optimization problem.

If one uses piecewise linear controls, then one can hope to obtain a better convergence order for the controls in (6.32). For example, in [Rösch, 2006, Theorem 1] the error estimate

$$\|\psi_u^h(\alpha) - \psi_u(\alpha)\|_{L^2(\Omega)} \leq Ch^{3/2} \tag{6.42}$$

is shown (under some assumptions) for the discretization with piecewise linear controls if $d = 1$. However, for the state y we would still have the convergence order h if we measure the error in the $H_0^1(\Omega)$-norm. A possible way to avoid this problem would be to instead use the $L^2(\Omega)$-error for the state y, which would result in the estimate

$$\|\psi_y^h(\alpha) - \psi_y(\alpha)\|_{L^2(\Omega)} \leq Ch^{3/2}$$

if (6.42) holds. If the convergence order in the estimate (6.33) is improved too, which can be done by increasing the regularity of y_o in Example 6.3.5, this could result in the improved convergence order $h^{3/4}$ for the discretization of (IOC). However, this would require modifying the proof of Theorem 6.3.3 to use the $L^2(\Omega)$-norm instead of the $\mathcal{Y}$-norm for the state y in some places.

6.3.3 Numerical examples

In this section we investigate the discretization error numerically. We consider the inverse optimal control problem and its discretization as specified in Examples 6.3.4 and 6.3.5. Moreover, we use $n := 1$, $d := 2$, $\Omega := (-1, 1)^2$, $\Phi_{UL} := [0, 1]$, $u_a := -8$, $u_b := 8$, $\sigma := 1/100$, and $z := 1$. We further define the functions $\hat{p}, y_o \in C^2(\operatorname{cl}\Omega) \cap L^2(\Omega)$ via

$$\begin{aligned}\hat{p}(\omega) &:= 10\, \exp(-10(\omega_1 + 0.4)^2 - 10\,(\omega_2 - 0.5)^2),\\ y_o(\omega) &:= 0.4\,\hat{p}(\omega) + \exp(-5(\omega_1 - 0.7)^2 - 5\,(\omega_2 - 0.3)^2),\end{aligned}$$

and set $P \in \mathbb{L}(\mathbb{R}^1, L^2(\Omega))$ as the linear operator that satisfies $P\alpha = \alpha\hat{p}$ for all $\alpha \in \mathbb{R}$.

We solve the discretized problem (IOC(h)) for various $h > 0$ in the following way. For the solution $\psi^h(\alpha)$ of the discretized optimal control problem (OC(α, h)) we use a semismooth Newton algorithm in combination with a backtracking line search for globalization. For an explanation of the semismooth Newton algorithm we refer to [Ulbrich, 2002]. Since we can calculate $\psi^h(\alpha)$, we can also calculate the discretized reduced upper level objective function $\hat{F}_h(\alpha) := F_h(\psi_y^h(\alpha), \psi_u^h(\alpha), \alpha)$ for a given $\alpha \in \mathbb{R}$. In order to solve the discretized upper level (IOC(h)) we need to minimize $\hat{F}_h$ over the interval $\Phi_{UL} = [0, 1]$. We do this

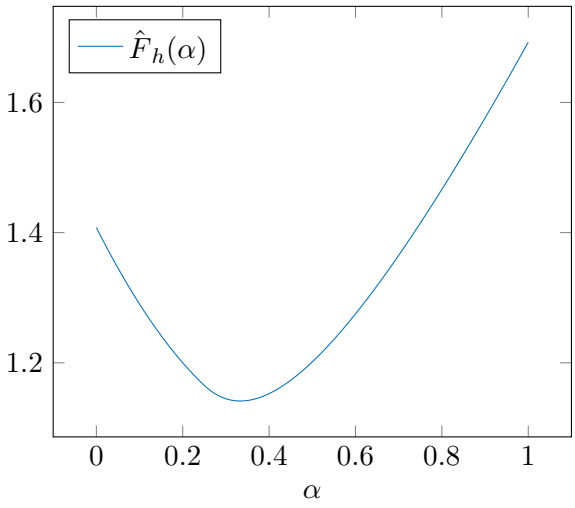

Figure 6.3.1: Discretized reduced upper level function $\hat{F}_h$ for $h = 1/32$

using the golden-section search, see [Kiefer, 1953]. Note that the golden-section search is only useful for finding the global minimizer if $\hat{F}_h$ is an unimodal (or quasi-convex) function, i.e. there is an $\bar{\alpha} \in [0,1]$ such that $\hat{F}_h$ is monotonically decreasing on $[0,\bar{\alpha}]$ and monotonically increasing on $[\bar{\alpha},1]$. However, a plot of the function $\hat{F}_h$ for $h = 1/32$ in Figure 6.3.1 suggests that this is the case (and it is also very likely to be true for smaller $h > 0$). Furthermore, the plot indicates that a quadratic growth condition holds, which was an assumption in Example 6.3.5.

We terminate the golden-section search as soon as we find an interval whose length is smaller than 10^{-12} and which contains the minimizer. In order to reach this termination condition we need 59 function evaluations of $\hat{F}_h$. This low number of function evaluations allows us to use finer discretizations in our numerical experiments, because calculating $\psi^h(\alpha)$ can be quite expensive for small $h > 0$. For example, if $h = 2^{-9}$ then we have $\dim \mathcal{Y}_h = 2^{21} - 2^{11} + 1 = 2095105$ and $\dim \mathcal{U}_h = 2^{22} = 4194304$.

We solved the problem (IOC(h)) for the mesh sizes $h \in \{2^1, 2^0, \dots, 2^{-9}\}$. In Figure 6.3.2 the errors $|\bar{\alpha}_{\bar{h}} - \bar{\alpha}_h|$, $\|\bar{y}_{\bar{h}} - \bar{y}_h\|_{H_0^1(\Omega)}$, $\|\bar{y}_{\bar{h}} - \bar{y}_h\|_{L^2(\Omega)}$, and $\|\bar{u}_{\bar{h}} - \bar{u}_h\|_{L^2(\Omega)}$ are plotted, where $(\bar{y}_h, \bar{u}_h, \bar{\alpha}_h)$ denotes a solution of (IOC(h)) and $\bar{h} := 2^{-9}$ is the smallest available mesh size. Using a linear regression for the logarithm of these data points and assuming that the real solution $(\bar{y}, \bar{u}, \bar{\alpha})$ of (IOC) is unique and very close to $(\bar{y}_{\bar{h}}, \bar{u}_{\bar{h}}, \bar{\alpha}_{\bar{h}})$ results in the experimental order of convergences

$$\begin{aligned} |\bar{\alpha} - \bar{\alpha}_h| &\approx C_\alpha h^{1.898}, \\ \|\bar{y} - \bar{y}_h\|_{H_0^1(\Omega)} &\approx C_{y,1} h^{1.022}, \\ \|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} &\approx C_{y,0} h^{1.892}, \\ \|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} &\approx C_u h^{1.039}, \end{aligned}$$

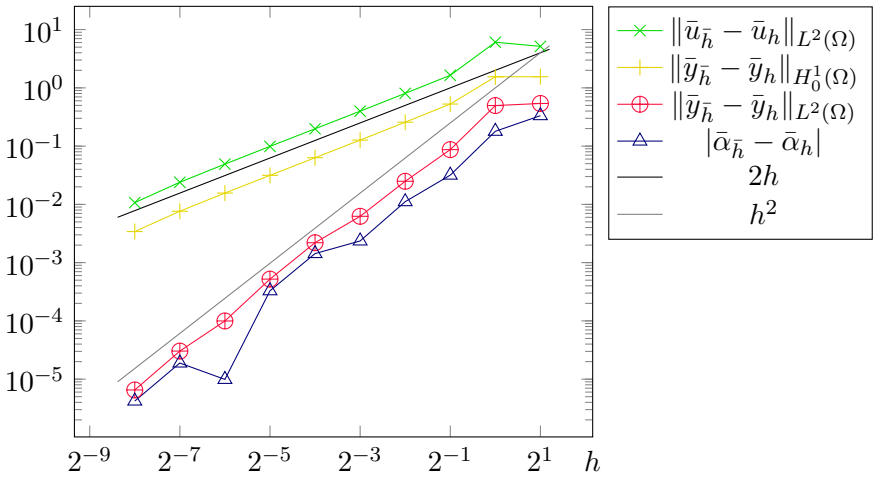

Figure 6.3.2: Errors of discretized solution of (IOC(h))

where $C_\alpha, C_u, C_{y,1}, C_{y,0} > 0$ are constants. Thus, we conjecture that the error estimate

$$\|\bar{y} - \bar{y}_h\|_{H_0^1(\Omega)} + \|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq Ch$$

holds for the discretization with piecewise constant controls in the setting of Examples 6.3.4 and 6.3.5. For $\bar{\alpha}_h$ and the $L^2(\Omega)$-error of $\bar{y}_h$ we conjecture that a convergence order of h or better holds.

A possible explanation for the better convergence order of $|\bar{\alpha} - \bar{\alpha}_h|$ could be that the upper level objective function F is locally Lipschitz continuous with respect to the $L^2(\Omega)$-norm of y (and not only with respect to the $H_0^1(\Omega)$-norm of y). This observation and the similar experimental convergence order of $|\bar{\alpha} - \bar{\alpha}_h|$ and $\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)}$ suggest that the good convergence order of $\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)}$ allows for the good convergence order of $|\bar{\alpha} - \bar{\alpha}_h|$.

In summary, our numerical experiments leads to the conjecture that the estimate

$$\operatorname{dist}((\bar{y}_h, \bar{u}_h, \bar{\alpha}_h), K) \leq \hat{C}h$$

holds in the setting of Example 6.3.5.

One might ask why the application of Theorem 6.3.3 only yields the order of convergence $h^{1/2}$ and not h as observed in the numerical experiments. A possible reason for this mismatch between theory and numerical results is that we did not use first-order stationarity conditions for the proof of Theorem 6.3.3 and instead relied on the assumption of a growth condition. If one were to use a similar method (without stationarity conditions but with a quadratic growth condition) for the discretization error estimates of the lower level optimization problem, then one would only obtain the error estimate

$$\|\psi_y^h(\alpha) - \psi_y(\alpha)\|_{\mathcal{Y}} + \|\psi_u^h(\alpha) - \psi_u(\alpha)\|_{L^2(\Omega)} \leq Ch^{1/2}$$

instead of (6.39). To obtain the convergence order h in Example 6.3.4 one usually has to use the stationarity conditions for the lower level optimization problem, see, e.g., [Tröltzsch, 2010]. Thus, it would be interesting to know whether any of the stationarity conditions for (IOC) discussed in Section 6.2 could be useful for deriving error estimates. However, since these stationarity conditions for (IOC) have a complicated structure, using them for discretization error estimates will not be a trivial matter.

7 Conclusion

In this thesis we proved necessary optimality conditions for local minimizers of bilevel optimization problems in an abstract setting. We applied these results for more specific instances of bilevel optimization problems, namely for the optimal control of the obstacle problem and for a class of inverse optimal control problems in Lebesgue spaces. The resulting stationarity systems are of C-stationarity type. Apart from these stationarity results we also obtained a variety of other results in this thesis. We discuss our results briefly.

In Section 3.4 we used the relaxation of the optimal value reformulation to derive necessary optimality conditions, which is an approach that was previously used in [Dempe, Harder, et al., 2019] for a class of inverse optimal control problems. A possible subject for future research could be to investigate how the abstract results in Section 3.4 can be applied in further bilevel optimization problems (beyond the applications presented in Chapters 5 and 6).

We also introduced the new concept of so-called normal-cone-preserving operators. This concept was subsequently successfully applied to obtain pointwise conditions for the stationarity systems of bilevel optimization problems in Lebesgue and Sobolev spaces. Additionally, we were able to fully describe the set of possible normal-cone-preserving operators with respect to a convex subset in $L^2(\Omega)$ with a complicated structure, see Proposition 6.2.7.

In our investigation of Legendre forms and Legendre-$\star$ forms we found out that they cannot exist in reflexive Banach spaces or Banach spaces with a separable predual space that are not Hilbertizable, see Theorem 4.3.9. Thus, we were able to generalize the results from [Harder, 2018] to nonreflexive spaces. For practical purposes this limits the application of the concept of Legendre forms and Legendre-$\star$ forms to Hilbert spaces. However, our results do not address the question, whether Legendre forms can exist on $L^1(\Omega)$, which is an open question to the best of our knowledge.

In Section 5.3 we provided lower estimates for the limiting normal cone to the complementarity set $\mathbb{K} = \operatorname{gph} \mathcal{N}_{H_0^1(\Omega)_+} \subset H_0^1(\Omega) \times H^{-1}(\Omega)$. We mention that these results were already published in [Harder, G. Wachsmuth, 2018c]. Since these lower estimates are rather large, this is further evidence that some of the tools of variational analysis such as the limiting normal cone might not be as useful in infinite-dimensional spaces as in finite-dimensional spaces. Similar observations have been made in [Mehlitz, 2017; Mehlitz, G. Wachsmuth, 2018].

For the inverse optimal control problems that we discussed in Sections 6.1 and 6.2 we also considered its discretization in Section 6.3. Under some assumptions, most prominently a growth condition, we derived discretization error estimates (that depend on the constants present in the assumptions) in Theorem 6.3.3. In Example 6.3.5 we applied this result to a particular example, where we discretized the control using piecewise constant functions. This resulted in a convergence order of $h^{1/2}$. However, our numerical experiments indicated that a better convergence order of h could be possible. For future research it would be interesting to close that gap and provide improved error estimates for discretizations of this infinite-dimensional bilevel optimization problem, possibly by using the stationarity conditions provided in Section 6.2.

Throughout this thesis, we provided several interesting examples and counterexamples, that are novel to the best of our knowledge. These include Examples 3.1.5, 3.1.9, and 3.1.10, which consider the continuity (or lack thereof) of the solution operator ψ, and Example 3.4.4 to show that the ε-relaxation of the optimal value reformulation does not need to have solutions in the absence of certain assumptions. We also provide Example 6.2.12 to show that strong stationarity does not need to hold for local minimizers of (IOC). This counterexample already appeared in [Harder, G. Wachsmuth, 2018b].

Finally, we remark that (to the best of our knowledge) it is an open question whether M-stationarity can be shown for local minimizers of the infinite-dimensional bilevel optimization problems presented in Chapters 5 and 6. It would be interesting to know whether M-stationarity of local minimizers can be shown under reasonable assumptions or whether a counterexample can be found.

Notation

Function spaces

$C^k(\Omega)$	space of k-times continuously differentiable functions $f : \Omega \to \mathbb{R}$
$C^k(\operatorname{cl}\Omega)$	space of functions in $C^k(\Omega)$ that are continuously extensible to $\operatorname{cl}\Omega$
$C(\operatorname{cl}\Omega)$	space of continuous functions on $\operatorname{cl}\Omega$
$C_c(\Omega)$	space of continuous functions with compact support in Ω
$C_c^\infty(\Omega)$	space of infinitely often differentiable functions with compact support in Ω
$H^1(\Omega)$	Sobolev space $W^{1,2}(\Omega)$
$H_0^1(\Omega)$	subspace of $H^1(\Omega)$ of functions with zero boundary conditions
$H^{-1}(\Omega)$	dual space of $H_0^1(\Omega)$
$L^p(\Omega)$	Lebesgue space of p-integrable Lebesgue measurable functions on Ω
$L_\mu^p(\Omega)$	Lebesgue space with exponent p and measure μ
ℓ^p	space of p-summable sequences
$W^{1,p}(\Omega)$	Sobolev space of p-integrable functions on Ω with p-integrable weak derivatives
$W_0^{1,p}(\Omega)$	subspace of $W^{1,p}(\Omega)$ of functions with zero boundary conditions
$W^{-1,p}(\Omega)$	dual space of $W_0^{1,q}(\Omega)$ where $1/p + 1/q = 1$

Functional analysis

∂	boundary of a set
cl	closure of a set
conv	convex hull of a set
id_X	identity operator on a normed space X
image	image of an operator or function
int	interior of a set
ker	kernel of an operator

lin	linear hull of a set
$B_\alpha(x)$	closed ball with radius $\alpha \geq 0$ and center x
$T^\star$	adjoint operator of an operator T
$X^\star$	dual space of a normed space X
$\mathbb{L}(X, Y)$	space of bounded linear operators between normed spaces
$\lvert\cdot\rvert$	absolute value or Euclidean norm in $\mathbb{R}^n$
$\lVert\cdot\rVert_X$, $\lVert\cdot\rVert$	norm on a normed space X
$\langle\cdot,\cdot\rangle_{X^\star\times X}$, $\langle\cdot,\cdot\rangle$	dual pairing for a normed space X
$X \hookrightarrow Y$	a normed space X embeds continuously in a normed space Y
$A^\perp$, $x^\perp$	annihilator of a set A or a point x
$\perp^Q$	Q-orthogonality for a quadratic form Q
$Y_1 \dot{+} Y_2$	direct sum of linear subspaces

Cones

A°	polar cone of a set A
$\mathcal{K}_A(x, x^\star)$	critical cone to a set A at a point x and functional $x^\star$
$\mathcal{N}_A(x)$	normal cone to a set A at a point x
$\mathcal{N}_A^{\lim}(x)$	limiting normal cone to a set A at a point x
$\mathcal{N}_A^{\text{s-lim}}(x)$	strong limiting normal cone to a set A at a point x
$\mathcal{N}_A^{\text{Fréchet}}(x)$	Fréchet normal cone to a set A at a point x
$\mathcal{N}_A^{\text{Clarke}}(x)$	Clarke normal cone to a set A at a point x
$\mathcal{R}_A(x)$	radial cone to a set A at a point x
$\mathcal{T}_A(x)$	tangent cone to a set A at a point x
$C_c^\infty(\Omega)_+$	cone of nonnegative functions in $C_c^\infty(\Omega)$
$C_c(\Omega)_+$	cone of nonnegative functions in $C_c(\Omega)$
$H_0^1(\Omega)_+$	cone of a.e. nonnegative functions in $H_0^1(\Omega)$
$H_0^1(\Omega)_-$	cone of a.e. nonpositive functions in $H_0^1(\Omega)$
$H^{-1}(\Omega)_+$	cone of nonnegative functionals in $H^{-1}(\Omega)$, polar cone of $H_0^1(\Omega)_-$
$H^{-1}(\Omega)_-$	cone of nonpositive functionals in $H^{-1}(\Omega)$, polar cone of $H_0^1(\Omega)_+$
$L^p(\Omega)_+$	cone of a.e. nonnegative functions in $L^p(\Omega)$
$W^{1,p}(\Omega)_+$	cone of a.e. nonnegative functions in $W^{1,p}(\Omega)$

$W_0^{1,p}(\Omega)_+$	cone of a.e. nonnegative functions in $W_0^{1,p}(\Omega)$
$W^{-1,p}(\Omega)_+$	cone of nonnegative functionals in $W^{-1,p}(\Omega)$

Functions and sets

v^+	positive part of v, i.e. $\max(v, 0)$
v^-	negative part of v, i.e. $-\min(0, v)$
$\{f \geq g\}$	abbreviation for $\{\omega \in \Omega \mid f(\omega) \geq g(\omega)\}$
$\{f \leq g\}$	abbreviation for $\{\omega \in \Omega \mid f(\omega) \leq g(\omega)\}$
$\{f \neq g\}$	abbreviation for $\{\omega \in \Omega \mid f(\omega) \neq g(\omega)\}$
$\{f = g\}$	abbreviation for $\{\omega \in \Omega \mid f(\omega) = g(\omega)\}$
χ_A	indicator function for a set A
∇	gradient
Δ	Laplace operator
$\frac{\partial v}{\partial n}\big\vert_{\partial O}$	outer normal derivative on ∂O
$\operatorname{avg}(A, f)$	average of a function f on a set A
$\operatorname{cap}(A)$	capacity of a set A
$\operatorname{dist}(x, A)$	distance of a point x to a set A
exp	exponential function
gph	graph of a set-valued mapping
$\operatorname{meas}(A)$	Lebesgue measure of a set A
$\text{q-supp}(\xi)$	quasi-support of a functional $\xi \in H^{-1}(\Omega)_+ \cup H^{-1}(\Omega)_-$
sgn	sign function
$\operatorname{supp}(f)$	support of a function f
$A \subset_q B$	abbreviation for $\operatorname{cap}(A \setminus B) = 0$
$A =_q B$	abbreviation for $\operatorname{cap}\big((A \setminus B) \cup (B \setminus A)\big) = 0$
$\leq_\Phi$	generalized inequality with respect to a convex cone Φ
$\mathcal{P}(A)$	power set of a set A
$\widehat{\mathcal{N}}_A^\varepsilon(x)$	set of ε-normals to a set A at a point x

Commonly used variables and symbols

C	a constant used for estimates
e_i	unit vector in $\mathbb{R}^n$ or ℓ^p
$\mathcal{I}_{X(\Omega)\to Y(\Omega)}$	an embedding between function spaces
$\mathcal{L}$	Lagrange function

$o(\cdot)$	Landau symbol
u	control variable
U_{ad}	set of admissible controls
y	state variable
$\mathbb{K}$	abbreviation for $\operatorname{gph} \mathcal{N}_{H_0^1(\Omega)_+}$
$\mathbb{N}$	set of natural numbers, starting with 1
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{Z}$	set of integers
φ	optimal value function of a parametrized optimization problem
Ψ	set-valued solution mapping of a parametrized optimization problem
ψ	solution operator of a parametrized optimization problem
Ω	domain, usually open and bounded subset of $\mathbb{R}^d$
ω	a point in Ω

Bibliography

Adams, Robert; Fournier, John (2003). *Sobolev Spaces*. 2nd ed. New York: Academic Press (cit. on pp. 45, 49, 173, 182).

Albiac, Fernando; Kalton, Nigel J. (2016). *Topics in Banach Space Theory*. 2nd ed. Springer International Publishing. DOI: 10.1007/978-3-319-31557-7 (cit. on pp. 43, 157, 158).

Albrecht, Sebastian; Leibold, Marion; Ulbrich, Michael (2012). "A bilevel optimization approach to obtain optimal cost functions for human arm movements". *Numerical Algebra, Control and Optimization* 2.1, pp. 105–127. DOI: 10.3934/naco.2012.2.105 (cit. on p. 14).

Albrecht, Sebastian; Ulbrich, Michael (2017). "Mathematical programs with complementarity constraints in the context of inverse optimal control for locomotion". *Optimization Methods & Software* 32.4, pp. 670–698. DOI: 10.1080/10556788.2016.1225212 (cit. on p. 14).

Attouch, Hedy; Buttazzo, Giuseppe; Michaille, Gérard (2014). *Variational Analysis in Sobolev and BV Spaces*. 2nd ed. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611973488 (cit. on p. 69).

Ball, Keith; Carlen, Eric A.; Lieb, Elliott H. (Dec. 1994). "Sharp uniform convexity and smoothness inequalities for trace norms". *Inventiones Mathematicae* 115.1, pp. 463–482. DOI: 10.1007/bf01231769 (cit. on p. 40).

Barbu, Viorel (1984). *Optimal Control of Variational Inequalities*. Vol. 100. Research Notes in Mathematics. Boston: Pitman (cit. on p. 160).

Bonnans, J. Frédéric; Shapiro, Alexander (2000). *Perturbation Analysis of Optimization Problems*. Berlin: Springer (cit. on pp. 16, 31–34, 42, 50, 68, 69, 74, 77, 78, 81, 87, 94, 96, 99, 101, 104, 108, 112, 113, 145, 146, 157).

Börgens, Eike; Kanzow, Christian; Mehlitz, Patrick; Wachsmuth, Gerd (2019). *New Constraint Qualifications for Optimization Problems in Banach Spaces based on Cone Continuity Properties*. To appear. arXiv: 1912.06531 (cit. on p. 125).

Borwein, Jonathan; Guirao, Antonio J.; Hájek, Petr; Vanderwerff, Jon (Oct. 2008). "Uniformly convex functions on Banach spaces". *Proceedings of the American Mathematical Society* 137.03, pp. 1081–1091. DOI: 10.1090/s0002-9939-08-09630-5 (cit. on p. 43).

Brezis, Haim (2011). *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York (cit. on p. 157).

Christof, Constantin; Wachsmuth, Gerd (2018). "No-Gap Second-Order Conditions via a Directional Curvature Functional". *SIAM Journal on Optimization* 28.3, pp. 2097–2130. DOI: 10.1137/17M1140418 (cit. on pp. 145, 146).

Christof, Constantin; Wachsmuth, Gerd (2020). "Differential Sensitivity Analysis of Variational Inequalities with Locally Lipschitz Continuous Solution Operators". *Applied Mathematics & Optimization* 81.1, pp. 23–62. DOI: 10.1007/s00245-018-09553-y (cit. on pp. 104, 145).

Ciarlet, Philippe G. (1991). "Basic error estimates for elliptic problems". *Handbook of numerical analysis, Vol. II*. Handb. Numer. Anal., II. North-Holland, Amsterdam, pp. 17–351 (cit. on pp. 216, 217).

Cioranescu, Doina; Murat, François (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 (cit. on pp. 171, 172, 178, 182).

Clarkson, James A. (1936). "Uniformly convex spaces". *Transactions of the American Mathematical Society* 40.3, pp. 396–414. DOI: 10.2307/1989630 (cit. on p. 157).

Conway, John B. (1985). *A Course in Functional Analysis*. Springer New York. DOI: 10.1007/978-1-4757-3828-5 (cit. on pp. 58, 152, 158).

Delfour, Michael; Zolésio, Jean-Paul (2011). *Shapes and Geometries*. 2nd ed. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9780898719826 (cit. on pp. 69, 113).

Dempe, Stephan (2002). *Foundations of bilevel programming*. Dordrecht: Kluwer Academic Publishers (cit. on p. 13).

Dempe, Stephan (2018). *Bilevel optimization: theory, algorithms and application*. Tech. rep. Preprint. URL: https://tu-freiberg.de/sites/default/files/media/fakultaet-fuer-mathematik-und-informatik-fakultaet-1-9277/prep/preprint_2018_11_dempe.pdf (cit. on p. 13).

Dempe, Stephan; Dutta, Joydeep (Feb. 2010). "Is bilevel programming a special case of a mathematical program with complementarity constraints?" *Mathematical Programming* 131.1-2, pp. 37–48. DOI: 10.1007/s10107-010-0342-1 (cit. on p. 117).

Dempe, Stephan; Harder, Felix; Mehlitz, Patrick; Wachsmuth, Gerd (2019). "Solving inverse optimal control problems via value functions to global optimality". *Journal of Global Optimization* 74.2, pp. 297–325. DOI: 10.1007/s10898-019-00758-1 (cit. on pp. 17, 109, 110, 118–120, 129, 136, 189, 190, 210, 223).

Dempe, Stephan; Mordukhovich, Boris S.; Zemkoho, Alain B. (Jan. 2012). "Sensitivity Analysis for Two-Level Value Functions with Applications to Bilevel Programming". *SIAM Journal on Optimization* 22.4, pp. 1309–1343. DOI: 10.1137/110845197 (cit. on p. 116).

Dobrowolski, Manfred (2010). *Angewandte Funktionalanalysis*. Springer Nature. DOI: 10.1007/978-3-642-15269-6 (cit. on p. 49).

Fabian, Marián; Habala, Petr; Hájek, Petr; Santalucía, Vicente Montesinos; Pelant, Jan; Zizler, Václav (2001). *Functional Analysis and Infinite-Dimensional Geometry*. Springer New York. DOI: 10.1007/978-1-4757-3480-5 (cit. on p. 31).

Fiacco, Anthony V.; Kyparisis, Jerzy (1986). "Convexity and concavity properties of the optimal value function in parametric nonlinear programming". *Journal of Optimization Theory and Applications* 48.1, pp. 95–126. DOI: 10.1007/BF00938592 (cit. on pp. 109, 110).

Flegel, Michael L.; Kanzow, Christian (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 (cit. on p. 68).

Folland, Gerald B. (1999). *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley (cit. on p. 57).

Fuglede, Bent (1971). "The quasi topology associated with a countably subadditive set function". en. *Annales de l'Institut Fourier* 21.1, pp. 123–169. DOI: 10.5802/aif.364 (cit. on p. 74).

Fukushima, Masatoshi; Oshima, Yoichi; Takeda, Masayoshi (2010). *Dirichlet Forms and Symmetric Markov Processes*. 2nd ed. De Gruyter. DOI: 10.1515/9783110218091 (cit. on pp. 69, 84).

Goldberg, Hyman; Kampowsky, Winfried; Tröltzsch, Fredi (1992). "On Nemytskij operators in L_p-spaces of abstract functions". *Mathematische Nachrichten* 155, pp. 127–140. DOI: 10.1002/mana.19921550110 (cit. on p. 192).

Haar, Alfred (1910). "Zur Theorie der orthogonalen Funktionensysteme". *Mathematische Annalen* 69.3, pp. 331–371. DOI: 10.1007/bf01456326 (cit. on p. 47).

Harder, Felix (2016). "Optimal Control of the Obstacle Problem Using the Value Function". Master's thesis. Technische Universität Chemnitz, Germany (cit. on pp. 68, 120, 142, 164, 168, 169).

Harder, Felix (2018). "Legendre Forms in Reflexive Banach Spaces". *Zeitschrift für Analysis und ihre Anwendungen* 37.4, pp. 377–388. DOI: 10.4171/zaa/1619 (cit. on pp. 17, 18, 23, 42, 145, 146, 148, 149, 151–153, 155–157, 223).

Harder, Felix; Mehlitz, Patrick; Wachsmuth, Gerd (2020). *Reformulation of the M-stationarity conditions as a system of discontinuous equations and its solution by a semismooth Newton method*. Preprint. arXiv: 2002.10124. URL: https://spp1962.wias-berlin.de/preprints/135.pdf (cit. on p. 68).

Harder, Felix; Wachsmuth, Gerd (2017). *The limiting normal cone of a complementarity set in Sobolev spaces*. Preprint. URL: https://spp1962.wias-berlin.de/preprints/023.pdf (cit. on pp. 18, 55).

Harder, Felix; Wachsmuth, Gerd (2018a). "Comparison of Optimality Systems for the Optimal Control of the Obstacle Problem". *GAMM-Mitteilungen* 40.4, pp. 312–338. DOI: 10.1002/gamm.201740004 (cit. on pp. 14, 16, 18, 68, 79, 83, 84, 87, 160, 161, 165, 167, 170).

Harder, Felix; Wachsmuth, Gerd (2018b). "Optimality conditions for a class of inverse optimal control problems with partial differential equations". *Optimization* 68.2-3, pp. 615–643. DOI: 10.1080/02331934.2018.1495205 (cit. on pp. 17, 18, 131, 189, 190, 210, 211, 224).

Harder, Felix; Wachsmuth, Gerd (2018c). "The limiting normal cone of a complementarity set in Sobolev spaces". *Optimization* 67.10, pp. 1579–1603. DOI: 10.1080/02331934.2018.1484467 (cit. on pp. 17, 18, 48, 54, 68, 90, 160, 170, 171, 173, 183, 186, 223).

Hatz, Kathrin; Schlöder, Johannes P.; Bock, Hans Georg (2012). "Estimating parameters in optimal control problems". *SIAM Journal on Scientific Computing* 34.3, A1707–A1728. DOI: 10.1137/110823390 (cit. on p. 14).

Heinonen, Juha; Kilpeläinen, Tero; Martio, Olli (1993). *Nonlinear potential theory of degenerate elliptic equations*. Oxford Mathematical Monographs. Oxford Science Publications. New York: The Clarendon Press Oxford University Press (cit. on pp. 76, 77, 84).

Hestenes, Magnus R. (1951). "Applications of the theory of quadratic forms in Hilbert space to the calculus of variations". *Pacific Journal of Mathematics* 1, pp. 525–581 (cit. on pp. 145, 148, 149).

Hintermüller, Michael; Kopacka, Ian (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. DOI: 10.1137/080720681 (cit. on p. 160).

Hintermüller, Michael; Mordukhovich, Boris S.; Surowiec, Thomas (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. DOI: 10.1007/s10107-013-0704-6 (cit. on p. 160).

Hintermüller, Michael; Surowiec, Thomas (2011). "First-order optimality conditions for elliptic mathematical programs with equilibrium constraints via variational analysis". *SIAM Journal on Optimization* 21.4, pp. 1561–1593. DOI: 10.1137/100802396 (cit. on p. 160).

Hinze, Michael; Pinnau, Rene; Ulbrich, Michael; Ulbrich, Stefan (2009). *Optimization with PDE Constraints*. Springer Netherlands. DOI: 10.1007/978-1-4020-8839-1 (cit. on pp. 42, 188, 193).

Holler, Gernot; Kunisch, Karl; Barnard, Richard C (2018). "A bilevel approach for parameter learning in inverse problems". *Inverse Problems* 34.11, p. 115012. DOI: 10.1088/1361-6420/aade77 (cit. on p. 14).

Ioffe, Alexander D.; Tikhomirov, Vladimir M. (1979). *Theory of Extremal Problems*. Studies in Logic and the Foundations of Mathematics. North-Holland Publishing Company (cit. on pp. 145, 148).

Jarušek, Jiří; Outrata, Jiří 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 (cit. on pp. 14, 160).

Kalton, Nigel J.; Konyagin, Sergei V.; Veselý, Libor (Jan. 2008). "Delta-semidefinite and Delta-convex Quadratic Forms in Banach Spaces". *Positivity* 12.2, pp. 221–240. DOI: 10.1007/s11117-007-2106-6 (cit. on p. 149).

Kiefer, Jack (Mar. 1953). "Sequential minimax search for a maximum". *Proceedings of the American Mathematical Society* 4.3, pp. 502–506. DOI: 10.1090/s0002-9939-1953-0055639-3 (cit. on p. 219).

Kilpeläinen, Tero; Malý, Jan (1992). "Supersolutions to degenerate elliptic equation on quasi open sets". *Communications in Partial Differential Equations* 17.3-4, pp. 371–405. DOI: 10.1080/03605309208820847 (cit. on p. 75).

Kinderlehrer, David; Stampacchia, Guido (1980). *An Introduction to Variational Inequalities and Their Applications*. New York: Academic Press (cit. on p. 50).

Krasnoselskii, Mark Alexandrovich; Zabreiko, Petr Petrovich; Pustylnik, Evgenii Izievic; Sobolevskii, Pavel Evseevich (1976). *Integral Operators in Spaces of Summable Functions*. Leyden: Noordhoff (cit. on pp. 191, 192).

Luo, Zhi-Quan; Pang, Jong-Shi; Ralph, Daniel (1996). *Mathematical Programs with Equilibrium Constraints*. Cambridge: Cambridge University Press (cit. on pp. 14, 65).

Mehlitz, Patrick (2017). "Contributions to complementarity and bilevel programming in Banach spaces". PhD thesis. Technische Universität Bergakademie Freiberg. URN: urn:nbn:de:bsz:105-qucosa-227091 (cit. on pp. 14, 118, 223).

Mehlitz, Patrick; Wachsmuth, Gerd (2016). "Weak and strong stationarity in generalized bilevel programming and bilevel optimal control". *Optimization* 65.5, pp. 907–935. DOI: 10.1080/02331934.2015.1122007 (cit. on p. 14).

Mehlitz, Patrick; Wachsmuth, Gerd (2018). "The Limiting Normal Cone to Pointwise Defined Sets in Lebesgue Spaces". *Set-Valued and Variational Analysis* 26.3, pp. 449–467. DOI: 10.1007/s11228-016-0393-4 (cit. on pp. 64, 199, 223).

Mehlitz, Patrick; Wachsmuth, Gerd (2019). *Bilevel optimal control: existence results and stationarity conditions*. Preprint. arXiv: 1906.08026 (cit. on p. 120).

Mignot, Fulbert (1976). "Contrôle dans les inéquations variationelles elliptiques". *Journal of Functional Analysis* 22.2, pp. 130–185. DOI: 10.1016/0022-1236(76)90017-3 (cit. on pp. 14, 160).

Mombaur, Katja; Truong, Anh; Laumond, Jean-Paul (2010). "From human to humanoid locomotion—an inverse optimal control approach". *Autonomous Robots* 28.3, pp. 369–383. DOI: 10.1007/s10514-009-9170-7 (cit. on p. 14).

Mordukhovich, Boris S. (2006). *Variational Analysis and Generalized Differentiation. Volume 1: Basic Theory*. Berlin: Springer (cit. on pp. 64, 65, 186).

Outrata, Jiří V.; Jarušek, Jiří; Stará, Jana (2011). "On optimality conditions in control of elliptic variational inequalities". *Set-Valued and Variational Analysis* 19.1, pp. 23–42. DOI: 10.1007/s11228-010-0158-4 (cit. on p. 160).

Outrata, Jiří V.; Kočvara, Michael; Zowe, Jochem (1998). *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. Dordrecht: Kluwer Academic (cit. on pp. 14, 65).

Pettis, Billy James (June 1939). "A proof that every uniformly convex space is reflexive". *Duke Mathematical Journal* 5.2, pp. 249–253. DOI: 10.1215/s0012-7094-39-00522-3 (cit. on p. 43).

Robinson, Stephen M. (1976). "Stability theory for systems of inequalities. II. Differentiable nonlinear systems". *SIAM Journal on Numerical Analysis* 13.4, pp. 497–513. DOI: 10.1137/0713043 (cit. on p. 61).

Rockafellar, Ralph Tyrrell (1970). *Convex Analysis*. Princeton University Press (cit. on p. 109).

Rockafellar, Ralph Tyrrell; Wets, Roger J. B. (1998). *Variational Analysis*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-02431-3 (cit. on p. 113).

Rösch, Arnd (2006). "Error estimates for Linear-Quadratic Optimal Control Problems with Control Constraints". *Optimization Methods and Software* 21.1, pp. 121–134. DOI: 10.1080/10556780500094945 (cit. on p. 218).

Rudin, Walter (1976). *Principles of Mathematical Analysis (International Series in Pure and Applied Mathematics)*. 3rd ed. McGraw-Hill Education (cit. on p. 28).

Rudin, Walter (1987). *Real and Complex Analysis*. 3rd ed. McGraw–Hill, New York (cit. on pp. 57, 58).

Rudin, Walter (1991). *Functional analysis*. 2nd ed. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York (cit. on pp. 23, 25, 58, 153).

Schiela, Anton; Wachsmuth, Daniel (2013). "Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints". *ESAIM Math. Model. Numer. Anal.* 47.3, pp. 771–787. DOI: 10.1051/m2an/2012049 (cit. on pp. 14, 160).

Schirotzek, Winfried (2007). *Nonsmooth analysis*. Universitext. Springer, Berlin. DOI: 10.1007/978-3-540-71333-3 (cit. on p. 32).

Stampacchia, Guido (1963-1964). "Équations elliptiques du second ordre à coefficients discontinus". fr. *Séminaire Jean Leray* 3, pp. 1–77 (cit. on p. 50).

Stollmann, Peter (1993). "Closed ideals in Dirichlet spaces". *Potential Analysis* 2.3, pp. 263–268. DOI: 10.1007/bf01048510 (cit. on p. 84).

Tröltzsch, Fredi (2010). "On Finite Element Error Estimates for Optimal Control Problems with Elliptic PDEs". *Large-Scale Scientific Computing*. Springer Berlin Heidelberg, pp. 40–53. DOI: 10.1007/978-3-642-12535-5_4 (cit. on pp. 216, 221).

Ulbrich, Michael (2002). "Semismooth Newton Methods for Operator Equations in Function Spaces". *SIAM Journal on Optimization* 13.3, pp. 805–841. DOI: 10.1137/s1052623400371569 (cit. on p. 218).

Velichkov, Bozhidar (2013). "Existence and regularity results for some shape optimization problems". PhD thesis. Scuola Normale Superiore (cit. on p. 80).

Wachsmuth, Gerd (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 (cit. on pp. 84, 160, 166).

Wachsmuth, Gerd (2015). "Mathematical Programs with Complementarity Constraints in Banach Spaces". *Journal of Optimization Theory and Applications* 166.2, pp. 480–507. DOI: 10.1007/s10957-014-0695-3 (cit. on p. 14).

Wachsmuth, Gerd (2016). "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 (cit. on pp. 14, 53, 160, 168–170).

Wachsmuth, Gerd (2018). "Pointwise Constraints in Vector-Valued Sobolev Spaces. With Applications in Optimal Control". *Applied Mathematics & Optimization* 77.3, pp. 463–497. DOI: 10.1007/s00245-016-9381-1 (cit. on p. 160).

Wachsmuth, Gerd (2019). "A guided tour of polyhedric sets". *Journal of Convex Analysis* 26.1, pp. 153–188 (cit. on p. 145).

Zowe, Jochem; Kurcyusz, Stanisław (1979). "Regularity and stability for the mathematical programming problem in Banach spaces". *Applied Mathematics and Optimization* 5.1, pp. 49–62. DOI: 10.1007/BF01442543 (cit. on pp. 61, 62).