Technische Universität München
Fakultät für Mathematik

Inexact bundle methods in Hilbert space with applications to optimal control problems governed by variational inequalities

Lukas Alexander Hertlein

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Dr. Jürgen Richter-Gebert
Prüfer der Dissertation:	1. Prof. Dr. Michael Ulbrich
	2. Prof. Dr. Dominikus Noll
	3. Prof. Dr. Anton Schiela

Die Dissertation wurde am 15.12.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 06.05.2022 angenommen.

Abstract

This doctoral thesis investigates nonsmooth, nonconvex optimization problems in Hilbert spaces. We develop a novel inexact bundle method which can be used to minimize arbitrary locally Lipschitz continuous functions as long as the user can provide sufficiently steep subgradient-based linearizations. The method is specially designed to allow for inexact function value and subgradient evaluations. We demonstrate how to efficiently solve the subproblem of the bundle method which is a piecewise quadratic optimization problem in Hilbert space. As a primary application, we consider optimal control problems governed by variational inequalities. For the numerical realization of the algorithm, we develop error estimates which bound the error of an approximate solution of the subproblem. We perform numerical verifications for examples from the problem subclass of optimal control problems governed by the obstacle problem.

Zusammenfassung

Diese Doktorarbeit befasst sich mit nichtglatten, nichtkonvexen Optimierungsproblemen in Hilberträumen. Es wird eine neuartige Bundlemethode entwickelt. Diese kann beliebige lokal Lipschitz-stetige Funktionen minimieren, solange genügend steile subgradientenbasierte Linearisierungen verfügbar sind. Die Methode benötigt lediglich inexakte Funktionswerte und Subgradienten. Besonderes Augenmerk wird darauf gerichtet, wie das Teilproblem der Bundlemethode, welches ein stückweise quadratisches Optimierungsproblem im Hilbertraum ist, effizient gelöst werden kann. Als Hauptanwendung werden Optimalsteuerungsprobleme mit Variationsungleichungsnebenbedingungen betrachtet. Für die numerische Umsetzung des Algorithmus entwickeln wir Fehlerschätzer, welche die Güte der approximativen Lösung des Teilproblems beschreiben. An Beispielen aus der Teilproblemklasse der Optimalsteuerungsprobleme mit Hindernisnebenbedingungen werden numerische Tests durchgeführt.

Notation

The following notation is used in this thesis:

$\emptyset$	empty set
$\lvert S\rvert$	cardinality of the set S
2^S	power set of the set S
$S \times T$	product set of the sets S and T
$\mathbb{N}$	the natural numbers: $\mathbb{N} = \{0, 1, 2, \dots\}$
$\mathbb{N}_+$	the positive natural numbers: $\mathbb{N}_+ = \{1, 2, 3, \dots\}$
$\mathbb{R}$	the real numbers
$\mathbb{C}$	the complex numbers
$\bar{\mathbb{R}}$	extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\} = [-\infty, \infty]$
$(0,\infty)$, $[0,\infty)$	set of positive and nonnegative real numbers, respectively
$[x]_+$	the positive part, $[x]_+ = \max\{x, 0\}$ for $x \in \mathbb{R}$
(a,b), $[a,b]$, $[a,b)$, $(a,b]$	open, closed, half-open line segments, respectively, with endpoints $a, b \in V$, V vector space, e.g., $[a,b] = \{ta + (1-t)b : 0 \leq t \leq 1\}$
$\mathbb{R}^n$	Euclidean space equipped with the inner product $(x,y) = x^\top y$ and norm $\lVert \cdot \rVert \equiv \lVert \cdot \rVert_{\mathbb{R}^n}$
$C^k(S,\mathbb{R}^m), C^k(S)$	Space of n-times continuously differentiable functions from $S \subset \mathbb{R}^n$ to $\mathbb{R}^m$, $C^k(S) := C^k(S,\mathbb{R})$
$C_c^\infty(S)$	$C_c^\infty(S) := \{v \in \cap_{n=0}^\infty C^n(S) :$ the support of v is a compact subset of $S\}$
$L^p(\Omega)$	Lebesgue space of p-integrable functions on the domain $\Omega \subset \mathbb{R}^n$
$H^k(\Omega)$	Sobolev space of k-times weakly differentiable square integrable functions on the domain $\Omega \subset \mathbb{R}^n$
$H_0^1(\Omega)$	Subspace of $H^1(\Omega)$ containing the functions with zero boundary in the sense of traces
$x^\top$	transpose of the vector or matrix x
I_n	identity matrix in $\mathbb{R}^{n\times n}$, $n \in \mathbb{N}_+$
$\lVert \cdot \rVert_{\mathrm{op}}$	the operator norm, i.e., $\lVert M\rVert_{\mathrm{op}} := \max_{\lVert x\rVert_{\mathbb{R}^n}=1} \lVert Mx\rVert_{\mathbb{R}^n}$
$\lvert M\rvert$	the absolute value (componentwise) $\lvert M\rvert_{i,j} := \lvert M_{i,j}\rvert$, $M \in \mathbb{R}^{n\times m}$
$\lVert \cdot \rVert_X$	induced norm on the Hilbert space X with inner product $(\cdot,\cdot)_X$: $\lVert \cdot \rVert_X^2 := (\cdot,\cdot)_X$
$\bar{B}_X(x,r)$	the closed unit ball of the space X centered at $x \in X$ with radius $r \geq 0$

$\delta_{\mathscr{A}}$	the indicator function of a set $\mathscr{A} \subset X$, $\delta_{\mathscr{A}}(y) = 0$ for $y \in \mathscr{A}$ and $\delta_{\mathscr{A}}(y) = 1$ for $y \notin \mathscr{A}$
$\chi_{\mathscr{A}}$	the characteristic function of a set $\mathscr{A} \subset X$, $\chi_{\mathscr{A}}(y) = 0$ for $y \in \mathscr{A}$ and $\chi_{\mathscr{A}}(y) = \infty$ for $y \notin \mathscr{A}$
$N_{\mathscr{A}}(y)$	the normal cone of a convex set $\mathscr{A} \subset X$, $N_{\mathscr{A}}(y) = \partial \chi_{\mathscr{A}}(y)$
$\operatorname{dom} f$	the effective domain of a function $f : X \to \mathbb{R} \cup \{\infty\}$, $\operatorname{dom} f := \{x \in X : f(x) < \infty\}$
∂	the convex subdifferential
∂_C	Clarke's subdifferential
$\mathrm{cl}(S)$	closure of the set $S \subset X$ in the space X
$\mathrm{co}(S)$	convex hull of the set $S \subset X$
$x_n \to \bar{x}$	convergence of the sequence $(x_n)_{n\in\mathbb{N}} \subset X$ in the Banach space X to the point $\bar{x} \in X$
$x_n \rightharpoonup \bar{x}$	weak convergence of the sequence $(x_n)_{n\in\mathbb{N}} \subset X$ in the Banach space X to the point $\bar{x} \in X$
Id_X	identity function on the space X, $\mathrm{Id}_X(x) = x$ for all $x \in X$
R_H	the Riesz map $R_H : H \to H^*$ in the Hilbert space H, cf. Section 2.1
P_M	orthogonal projection onto the closed linear subspace M of a Hilbert space H, cf. Section 2.1.2
$\mathscr{L}(X,Y)$	the space of linear and bounded operators from the Banach space X to the Banach space Y, cf. Section 2.1

A selection of frequently used variables:

ι	embedding from the space X to the space Y
p	nonsmooth, outer part of the objective function, $p : Y \to \mathbb{R}$ cf. Chapter 3
f	nonsmooth part of the objective function, $f = p \circ \iota$, cf. Chapter 3
w	smooth part of the objective function, $w : X \to \mathbb{R}$, cf. Chapter 3
G	subgradient multifunction, $G : Y \rightrightarrows Y^*$, cf. Section 3.1.2
τ_i	proximity parameter in iteration i, cf. Section 3.1.6
Q_i	curvature operator in iteration i, cf. Section 3.1.6
$\frac{1}{2}\|\iota(\,\cdot\, - x_i)\|^2_{Q_i+\tau_i R_Y}$	proximity term in iteration i, cf. Section 3.1.6
ϕ_i	cutting plane model in iteration i, cf. Section 3.1.4
Φ_i	local model in iteration i, $\Phi_i = \phi_i + w + \delta_{\mathscr{F}}$, cf. Section 3.1.1
Ψ_i	piecewise quadratic model in iteration i, $\Psi_i = \Phi_i + \frac{1}{2}\|\iota(\,\cdot\, - x_i)\|^2_{Q_i+\tau_i R_Y}$, cf. (3.1.2)
α	Tikhonov regularization parameter, cf. Section 4.2
$\tilde{E}$	$\tilde{E} := R_Y \iota R_X^{-1} \iota^* \in \mathscr{L}(Y^*)$, cf. Section 4.3.1
$\tilde{D}_\tau$	$\tilde{D}_\tau := \alpha\,\mathrm{Id}_{Y^*} + \tau\tilde{E} \in \mathscr{L}(Y^*)$, cf. Section 4.3.1
D_τ	$D_\tau := \alpha R_X + \tau \iota^* R_Y \iota \in \mathscr{L}(X, X^*)$, cf. Section 4.4.3

Contents

1. Introduction

We are interested in optimal control problems of the form

$$\begin{aligned} &\underset{(y,u)\in H\times U}{\text{minimize}} && J(y)+\tfrac{\alpha}{2}\|u\|_U^2 \\ &\text{subject to} && \langle Ay-\mathring{F}(\iota u), v-y\rangle_{H^*,H}\geq 0 \qquad \forall v\in K \\ & && y\in K, u\in U_{\text{ad}}, \end{aligned} \tag{P}$$

where U,H are Hilbert spaces such that $\iota \in \mathscr{L}(U,H)$ is a compact embedding, $J : H \to \mathbb{R}$, $\alpha > 0$, $\mathring{F} : H \to H^*$, $A \in \mathscr{L}(H,H^*)$, and $K \subset H$ and $U_{\text{ad}} \subset U$ are nonempty, closed and convex. The governing variational inequality

$$\text{Find } y\in K: \qquad \langle Ay-b, v-y\rangle_{H^*,H}\geq 0 \qquad \forall v\in K \tag{VI}$$

arises naturally in many physics and engineering applications such as the contact of an elastic body with a rigid foundation (Signorini's problem), the filtration of a liquid through a porous medium and lubrication [66, 67, 40]. Under appropriate assumptions on the data, the variational inequality (VI) has a unique solution and the solution map $S : H^* \to H$, $b \mapsto y$, is Lipschitz continuous but not necessarily Gâteaux-differentiable. Inserting the solution operator into the objective function gives rise to the reduced problem

$$\underset{u\in U_{\text{ad}}}{\text{minimize}} \quad J(S(\mathring{F}(\iota u)))+\tfrac{\alpha}{2}\|u\|_U^2. \tag{P'}$$

This is a nonconvex and nondifferentiable optimization problem posed in Hilbert space and thus quite challenging to solve. Furthermore, any solution procedure that can be executed on a computer has to deal with discretization issues which occur when elements of an infinite dimensional Hilbert space cannot be fully described by a finite dimensional vector. Therefore, the need arises to incorporate inexact function value and subgradient evaluation into the solution method.

Based on the work of [99], we develop a novel bundle method in Hilbert space setting which is tailored for the structure of Problem (P'). It aims at solving problems of the form

$$\min_{x\in X}\; p(\iota x)+w(x), \quad x\in\mathscr{F}, \tag{1.0.1}$$

where X and Y are Hilbert spaces such that $\iota \in \mathscr{L}(X,Y)$ is a compact embedding, $p : Y \to \mathbb{R}$ and $w : X \to \mathbb{R}$ are Lipschitz on bounded sets, w is additionally continuously differentiable and strongly convex and $\mathscr{F}$ is a nonempty, closed and convex set. As it is common for bundle methods, the algorithm collects function values and subgradient information to build a piecewise quadratic model of the objective function. Our method requires only inexact evaluations of function values $\tilde{f}$ and subgradients g. In particular, at the point $x \in X$, the error of the function value $|\tilde{f} - p(\iota x)|$ has to be bounded and an inexact subgradient g has to be drawn from the set $G(\iota x)$, where the multifunction $G : Y \rightrightarrows Y^*$ acts as a subdifferential approximation. We carve out the minimal assumptions on G required to show convergence of the bundle method. A possible choice of G is $G := \partial_C p(\cdot) + \bar{B}_{Y^*}(0,\delta)$, where ∂_C denotes Clarke's subdifferential and $\delta \geq 0$. These possible choices for the function value and the subgradient yield a very flexible algorithm. The model of the objective function can also incorporate curvature information of the nonsmooth part p, which can be obtained, for example, using the BFGS-method. The model has a unique minimizer called minimizing iterate. Depending on the function value at the minimizing iterate, either the current model is improved or a new model is constructed. The minimizer of the model does not have to be computed exactly but rather an approximation thereof, called trial iterate, can be used. The convergence analysis of the bundle method is presented in an abstract framework for which we only require that the model value at the trial iterate converges to the minimum of the model as the algorithm progresses. We proof that any weak subsequence of iterates is η-stationary, where the radius $\eta \geq 0$ depends on how well the model captures the behavior of the objective function. This is a novel result which, to the best of the author's knowledge, is not even available in the finite dimensional setting.

We then investigate conditions which ensure a bound on the size of η. On the one hand, if one further assumes additional regularity of the objective function beyond Lipschitz continuity (such as weak or approximate convexity) and one uses sufficiently exact function values and subgradients, then one can guarantee a priori that η is below a given threshold. On the other hand, if no further regularity beyond Lipschitz continuity of the objective function is given, one can track a simple indicator of the quality of the model and improve the model whenever this indicator suggests to do so. Improving the model is done by adding new inexact function values and subgradients at appropriate points. If one can always find a sufficiently good model (in the sense of this indicator), then η also can be guaranteed to be below a given threshold. A sufficiently good model always exists, however, how to find such a model is problem dependent.

In order to execute the bundle method, one needs a way to solve the piecewise quadratic subproblem efficiently. For the case of $w := \frac{\alpha}{2}\|\cdot\|_X^2$, we investigate a dual reformulation. This dual reformulation results in a n_p-dimensional piecewise quadratic optimization problem, where n_p is the number of quadratic regions of the bundle subproblem. As n_p is typically small, the cost to solve this problem can be neglected. However, to set up the dual problem, the inverse of an operator $F : X \to X^*$ has to be evaluated. In the applications considered here, the evaluation of F^{-1} corresponds to the solution of a linear elliptic partial differential equation (PDE) which dominates the costs of solving the subproblem. As it is not possible to evaluate F^{-1} exactly, we provide error estimates for the accuracy of the solution if approximations of F^{-1} are used. These error estimates are then used to enforce that the computed approximate solution of the subproblem meets the accuracy requirements of the bundle method.

We then proceed by applying the bundle method to the optimal control of the obstacle problem, which is a problem of the form (P). Given a domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$, the obstacle $\psi \in H^1(\Omega)$ defines the set K via $K := \{v \in H_0^1(\Omega) : v \geq \psi \text{ a.e. on } \Omega\}$. We use a finite element discretization to numerically approximate the function value of the reduced objective function. The recent paper [110] provides a formula to compute an element $g \in \partial_C J(S(\mathring{F}(\cdot)))(w)$ of the Clarke subdifferential of the reduced objective function at an arbitrary point $w \in H^{-1}(\Omega)$. We apply both a priori as well as a posteriori error estimates to balance the error contributions from inexact objective function evaluation and inexact subproblem solves. Numerical results for a set of test problems are provided.

Finally, we are interested to obtain a robust solution of Problem (P). To do so, let $(\Xi, \mathscr{A}, P)$ be a probability space, $\xi \in \Xi$ a parameter and consider the parametric obstacle problem

$$\text{Find } y_\xi \in K_\xi : \quad \langle A_\xi y_\xi - b_\xi, v_\xi - y_\xi \rangle_{Z^*,Z} \geq 0 \quad \forall v_\xi \in K_\xi, \tag{VI$_\xi$}$$

with the data A_ξ, b_ξ, ψ_ξ and $K_\xi := \{v_\xi \in H_0^1(\Omega) : v_\xi \geq \psi_\xi \text{ a.e. on } \Omega\}$. As for the deterministic problem, (VI$_\xi$) defines a solution operator S_ξ for P-a.a. $\xi \in \Xi$. This leads to the optimal control problem

$$\min_{u \in U_{\text{ad}}} \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(\iota u)))\right] + \tfrac{\alpha}{2}\|u\|_U^2, \tag{P$_\xi$}$$

where $J_\xi : H \to \mathbb{R}$, $\mathring{F}_\xi : H \to H^*$ and $\mathbb{E}$ denotes the expectation with respect to ξ. Under appropriate assumptions, the family of variational inequalities (VI$_\xi$) can be equivalently expressed via the stochastic obstacle problem

$$\text{Find } \mathbf{y} \in \mathbf{K}, \quad \langle \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{v} - \mathbf{y} \rangle_{\mathbf{H}^*,\mathbf{H}} \geq 0 \quad \text{for all } \mathbf{v} \in \mathbf{K}, \tag{\textbf{VI}}$$

which is formulated in the Bochner space $\mathbf{H} := L^2(\Xi, H_0^1(\Omega))$ and the set $\mathbf{K} \subset \mathbf{H}$ is given via

$$\mathbf{K} := \{\mathbf{v} \in \mathbf{H} : \mathbf{v}(\xi) \in K_\xi \text{ for } P\text{-a.a. } \xi \in \Xi\}.$$

Again, **(VI)** has a unique solution with Lipschitz continuous solution operator $\mathbf{S} : \mathbf{H}^* \to \mathbf{H}$ and (P$_\xi$) can be reformulated as the stochastic optimal control problem

$$\min_{u \in U_{\text{ad}}} \mathbf{J}(\mathbf{S}(\mathring{\mathbf{F}}(\iota u)) + \tfrac{\alpha}{2}\|u\|_U^2, \tag{\textbf{P}}$$

where $\mathbf{J} : \mathbf{H} \to \mathbb{R}$ is defined via $\mathbf{J}(\mathbf{y}) := \mathbb{E}\left[J_\xi(\mathbf{y}(\xi))\right]$ and $\mathring{\mathbf{F}} : H \to \mathbf{H}^*$ is given via $\mathring{\mathbf{F}}(w)(\xi) := \mathring{F}_\xi(w)$. To apply the bundle method to **(P)**, one would like to use $G = \partial_C \mathbf{J}(\mathbf{S}(\mathring{\mathbf{F}}(\cdot))) + \bar{B}_{\mathbf{H}^*}(0, \delta)$ as the approximate subdifferential. However, the calculus rules for the Clarke subdifferential, which often take the form of inclusions, make it difficult to compute an element of the subdifferential of the function $\mathbf{J}(\mathbf{S}(\mathring{\mathbf{F}}(\cdot))$. As we can compute a subgradient $g(w, \xi) \in \partial_C J_\xi(S_\xi(\mathring{F}_\xi(\cdot)))(w)$ of the reduced parametric objective function for P-a.a. $\xi \in \Xi$, we consider the enlarged subdifferential

$$G(w) := \{\mathbb{E}[g(w,\xi)] : \xi \mapsto g(w,\xi) \in L^1(\Xi, H^{-1}(\Omega)), g(w,\xi) \in \partial_w J_\xi(S_\xi(\mathring{F}_\xi(w)))\ P\text{-a.e. }\}$$

and show that G can be used as a subgradient approximation in the bundle method.

Overview of existing literature

The variational inequality (VI) can equivalently be rewritten as an optimization problem, as the equality $y = S(u)$ (cf. Section 2.4) or, in the case of the obstacle problem, as the complementarity system

$$\text{Find } (y,\xi) \in H_0^1(\Omega) \times H^{-1}(\Omega) : \quad \xi = Ay - b,\ \ y \geq \psi,\ \xi \geq 0,\ \langle \xi, y - \psi \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = 0,$$

where $\xi \in H^{-1}(\Omega)$ is a slack variable (cf. Section 5.1). Depending on the reformulation, Problem (P) can be classified as a bilevel optimization problem, a mathematical problem with equilibrium constraints (MPEC) or a mathematical problem with complementarity constraints (MPCC), respectively. In finite dimensions, an overview on MPECs is given in [103] and [116] presents stationarity concepts for MPCCs. Due to a lack of constraint qualifications, deriving necessary optimality conditions is difficult already in finite dimensions and several distinct stationarity concepts have been established ([92, 131, 75, 57, 117, 132]). Solution algorithms including smoothing [63, 56, 75, 117], regularization [55, 58] and penalization [81] have been proposed. These approaches have in common that a sequence of large scale optimization problems has to be solved. Using the reformulation (P') avoids this, but the resulting reduced problem is nonconvex and nondifferentiable. To solve it, we develop a bundle method in function space.

Bundle methods were introduced in 1975 with the works of [77, 134] in the finite dimensional context with exact function and subgradient information. Since then, a vast body of literature was published and it is out of the scope of this work to give the numerous references to papers addressing the convex case. Excellent reviews can be found in [84, 26]. Bundle methods for nonconvex function include, e.g., [5, 46, 69, 71, 65, 78, 82, 85, 89, 90, 100] and this is still subject to current research. Bundle methods for convex objective functions which use inexact data can, for example, be found in [25, 53, 70, 121, 72, 86]. [70] requires eventually exact data and [53] works with exact function values but inexact subgradients. The convergence theory of bundle methods for nonconvex problems with inexact function and subgradient evaluations is quite involved already in finite dimensions and only few papers such as [25, 47, 48, 83, 119, 99] cover this subject.

The literature for infinite-dimensional convergence theory of bundle methods is scarce and we are only aware of papers where the convex case with exact function and subgradient values is addressed, such as [22, 126]. Our approach is inspired by [99] and provides an extension of the method in [99] to infinite-dimensional Hilbert spaces. First results were already published in [50, 49]. The theory we develop is more general as in [50] and includes the results [50, Thm. 5.5] and [49, Thm. 2.6] as special cases. To the best of the author's knowledge, this work provides the first convergence analysis for bundle methods in general Hilbert spaces for the optimization of nonconvex problems with inexact function and subgradient evaluations. In the case that X and Y are finite dimensional and the objective function is approximately convex, the given convergence theorems recovers the state of the art of the theory available. However, to the best of the author's knowledge, the theory for general Lipschitz continuous objective functions is new, even for finite dimensional spaces.

The works on variational inequalities and in particular on the obstacle problem are vast, see, e.g., [91, 115, 40, 67, 38, 39, 120]. A priori error estimates for Finite Element Method (FEM) discretizations of the obstacle problem have been obtained already in [16, 17]. In recent years, a posteriori error estimates have been developed. In [98], residual based a posteriori error estimates for the obstacle problem were derived. Constant free a posteriori error estimates were derived in [95, 111]. To the best of the author's knowledge there are no systematic approaches to use these estimators in an adaptive inexact algorithm to solve optimization problems with variational inequality constraints.

The stochastic obstacle problem has been considered, e.g., in [44, 45]. Compared with the optimal control of deterministic obstacle problems, the literature concerning the optimal control of stochastic obstacle problems is less developed. Recent publications such as [74, 36, 43] address optimization problems with stochastic PDE constraints. The first uses stochastic collocation in combination with sparse grids to approximate the expectation in the objective function, whereas the second uses low-rank tensors. In [43], error estimates for the finite element discretization of the optimality system are provided. Stochastic MPECs with finite dimensional control space have been considered, for example, in [80, 118, 135]. Here, quasi-Monte Carlo or sample average approximation techniques have been used.

List of prior publications and manuscripts

[49] L. HERTLEIN, A.-T. RAULS, M. ULBRICH, AND S. ULBRICH, *An inexact bundle method and subgradient computations for optimal control of deterministic and stochastic obstacle problems*. Priprint, accepted for publication in SPP1962 Special Issue, Birkhäuser, 2019.

[50] L. HERTLEIN AND M. ULBRICH, *An inexact bundle algorithm for nonconvex nonsmooth minimization in Hilbert space*, SIAM J. Control Optim., 57 (2019), pp. 3137–3165.

Structure of the dissertation

The doctoral thesis is divided into seven chapters. Chapter 2 starts off with mathematical preliminaries. In Chapter 3, the bundle method is presented and the convergence theory is developed. Earlier versions of the bundle method were already communicated in [50, 49]. Chapter 4 is contend with the efficient solution of the bundle subproblem and provides error estimates for the accuracy of the solution. Chapter 5 is devoted to the optimal control problem governed by the obstacle problem. Chapter 6 includes implementation details and numerical results. In Chapter 7 we consider the optimal control problem governed by the stochastic obstacle problem. This chapter is based on the article [49].

2. Preliminaries

In this dissertation, whenever there is a citation mark at a lemma or a theorem and no proof is given, then the statement of this lemma or theorem can be found in the given source. It may be quoted literally or with minor changes in notation to accommodate for the present use of notation.

2.1. Banach and Hilbert spaces

In this section, we review some basic facts about Banach and Hilbert spaces. We choose to use [3] as a primary reference but there are many more excellent books which cover this topic.

A tuple $(X, \|\cdot\|_X)$ is called a (real) Banach space if X is a $\mathbb{R}$-vector space, $\|\cdot\|_X$ is a norm on X and X is complete with respect to $\|\cdot\|_X$ (cf. [3, Chap. 2.22]). If no confusion is possible, then we also say that X is a Banach space. Let X, Y be real Banach spaces. Denote by $(\mathscr{L}(X,Y), \|\cdot\|_{\mathscr{L}(X,Y)})$ the space of linear and bounded operators with the operator norm $\|A\|_{\mathscr{L}(X,Y)} := \sup_{\|x\|_X=1} \|Ax\|_Y$. By [3, Thm. 5.3], $\mathscr{L}(X,Y)$ is a Banach space. The dual space of X is defined by $X^* := \mathscr{L}(X,\mathbb{R})$. We denote $\mathscr{L}(X) := \mathscr{L}(X,X)$. For an operator $A \in \mathscr{L}(X,Y)$, we define the Banach space adjoint $A^* \in \mathscr{L}(Y^*,X^*)$ by $\langle A^*y', x\rangle_{X^*,X} := \langle y', Ax\rangle_{Y^*,Y}$ for all $x \in X, y' \in Y^*$. An operator $A \in \mathscr{L}(X,X^*)$ is called symmetric if $\langle Ax_1, x_2\rangle_{X^*,X} = \langle Ax_2, x_1\rangle_{X^*,X}$ for all $x_1, x_2 \in X$.

DEFINITION 2.1.1. *An operator $A \in \mathscr{L}(X,Y)$ is called invertible if, for all $y \in Y$, there exists a unique $x \in X$ such that $Ax = y$ and the corresponding map $A^{-1} : y \mapsto x$ is a bounded and linear operator.*

For a $\mathbb{R}$-vector space H, a positive definite symmetric bilinear form $(\cdot,\cdot)_H : H \times H \to \mathbb{R}$ is called an inner product on H. A tuple $(H, (\cdot,\cdot)_H)$ is called a real Hilbert space if H is a $\mathbb{R}$-vector space, $(\cdot,\cdot)_H$ is an inner product on H and H is complete with respect to the norm $\|\cdot\|_H : x \mapsto \sqrt{(x,x)_H}$ (cf. [3, Chap. 2.22]). We denote by $R_H \in \mathscr{L}(H,H^*)$ the Riesz map defined by

$$\langle R_H x, y\rangle_{H^*,H} := (y,x)_H \qquad \text{for all } x, y \in H.$$

By the Riesz representation theorem (cf. [3, Thm. 6.1]), R_H is an isometric linear isomorphism, i.e., R_H is invertible (in particular $R_H^{-1} \in \mathscr{L}(H^*,H)$) and $\|R_H x\|_{H^*} = \|x\|_H$ for all $x \in H$. Note that $\|R_H^{-1}x'\|_H = \|x'\|_{H^*}$ for all $x' \in H^*$ and

$$(R_H^{-1}x', x)_H = \langle R_H R_H^{-1} x', x\rangle_{H^*,H} = \langle x', x\rangle_{H^*,H} \qquad \text{for all } x \in H, x' \in H^*.$$

Furthermore, R_H is symmetric, i.e.,

$$\langle R_H x_1, x_2\rangle_{H^*,H} = (x_1,x_2)_H = (x_2,x_1)_H = \langle R_H x_2, x_1\rangle_{H^*,H} \qquad \text{for all } x_1, x_2 \in H.$$

The inner product $(\cdot,\cdot)_{H^*} := (R_H^{-1}\cdot, R_H^{-1}\cdot)_H$ induces the norm $\|\cdot\|_{H^*}$ because

$$\|x'\|_{H^*} = \|R_H^{-1}x'\|_H = \sqrt{(R_H^{-1}x', R_H^{-1}x')_H} \qquad \text{for all } x' \in H^*.$$

In particular, $(H^*, (R_H^{-1}\cdot, R_H^{-1}\cdot)_H^{1/2}) = (H^*, \|\cdot\|_{H^*})$ is the dual space of $(H, \|\cdot\|_H)$. By [3, Thm. 5.3], $(H^*, \|\cdot\|_{H^*})$ is complete. Therefore, $(H^*, (\cdot,\cdot)_{H^*})$ is a Hilbert space. It holds

$$(R_H x, x')_{H^*} = (x, R_H^{-1}x')_H = (R_H^{-1}x', x)_H = \langle x', x\rangle_{H^*,H} \qquad \text{for all } x \in H, x' \in H^*.$$

DEFINITION 2.1.2. *If X,Y are Hilbert spaces, we define the Hilbert space adjoint $A^{\circledast} \in \mathscr{L}(Y,X)$ of an operator $A \in \mathscr{L}(X,Y)$ by $A^{\circledast} := R_X^{-1}A^*R_Y$. An operator $A \in \mathscr{L}(X)$ is called self-adjoint if $A = A^{\circledast}$.*

From the definition of the Hilbert space adjoint it directly follows that

$$(A^{\circledast}y, x)_X = (y, Ax)_Y \qquad \text{for all } x \in X, y \in Y.$$

LEMMA 2.1.3. *For an operator $A \in \mathscr{L}(H)$ it is equivalent:*
1) A is self-adjoint.
2) $(Ax,y)_H = (Ay,x)_H$ for all $x,y \in H$.
3) $R_H A$ is symmetric.

Proof. The equivalence of 1) and 2) can be seen by

$$(Ax,y)_H = (y,Ax)_H = \langle R_H y, Ax\rangle_{H^*,H} = \langle A^* R_H y, x\rangle_{H^*,H} = (R_H^{-1}A^*R_H y, x)_H \qquad \text{for all } x,y \in H.$$

Furthermore, 2) is equivalent to 3) because $\langle R_H Ax, y\rangle_{H^*,H} = (Ax,y)_H$ and $(Ay,x)_H = \langle R_H Ay, x\rangle_{H^*,H}$ for all $x,y \in H$. ∎

2.1.1. Embeddings

DEFINITION 2.1.4. *An injective operator $\iota : X \to Y$ from one Banach space X to another Banach space Y is called an embedding.*

LEMMA 2.1.5. *Let $\iota \in \mathscr{L}(X,Y)$ be an embedding of the Hilbert space X into the Hilbert space Y, let $(x_i)_{i\in\mathbb{N}} \subset X$ be a bounded sequence and let $\bar{x} \in X$ be a point. If $\iota x_i \to \iota\bar{x}$ as $i \to \infty$, then $x_i \rightharpoonup \bar{x}$ as $i \to \infty$.*

Proof. The injectivity of ι implies that $\ker\iota := \{x \in X : \iota x = 0\} = \{0\}$. By [12, Fact 2.25(iv)], the set $\iota^{\circledast}(Y) = R_X^{-1}\iota^* R_Y(Y)$ is dense in $(\ker\iota)^\perp = \{0\}^\perp = X$. Consequently, by the Riesz representation theorem, also $\iota^*(Y^*)$ is dense in X^*. For arbitrary $y' \in Y^*$, it holds that

$$\langle \iota^* y', x_i\rangle_{X^*,X} = \langle y', \iota x_i\rangle_{Y^*,Y} \to \langle y', \iota\bar{x}\rangle_{Y^*,Y} = \langle \iota^* y', \bar{x}\rangle_{X^*,X} \qquad \text{as } i \to \infty.$$

Since $(x_i)_{i\in\mathbb{N}}$ is bounded in X, [137, Prop. 21.23(g)] shows that $x_i \rightharpoonup \bar{x}$ as $i \to \infty$. ∎

2.1.2. Projections

Let M be a nonempty closed convex subset of a Hilbert space H. The metric projection $P_M : H \to H$ onto M is defined via

$$P_M(x) \in M, \qquad \|P_M(x) - x\|_H = \min_{v \in H} \|v - x\|_H \qquad \text{for all } x \in H.$$

LEMMA 2.1.6 ([59, Lem. 1.10]). *It holds:*

1) *P_M is well-defined.*
2) *For all $x, y \in H$ there holds:*

$$y = P_M(x) \qquad \Leftrightarrow \qquad y \in M, \quad (x - y, v - y)_H \leq 0 \quad \forall v \in M.$$

3) *P_M is nonexpansive, i.e.,*

$$\|P_M(x) - P_M(y)\|_H \leq \|x - y\|_H \qquad \forall x, y \in H.$$

4) *P_M is monotone, i.e.,*

$$(P_M(x) - P_M(y), x - y)_H \geq 0 \qquad \forall x, y \in H.$$

Furthermore, equality holds if and only if $P_M(x) = P_M(y)$.

Now let M be a closed linear subspace of the Hilbert space $H \neq \{0\}$. Denote the orthogonal complement of M by

$$M^\perp := \{x \in H : (x, v)_H = 0 \qquad \text{for all } v \in M\}.$$

According to [129, Thm. 2.10.11], the Hilbert space H can be decomposed into the direct sum of M and $M^\perp$, i.e., $H = M \oplus M^\perp$. Thus, every element $x \in H$ can be uniquely written as $x = y + z$ with $y \in M$ and $z \in M^\perp$. We call the mapping $x \mapsto y$ the orthogonal projection onto the subspace M. Since M is a nonempty closed convex subset of H, P_M is well-defined and by [137, Prop. 21.44(a), (c)], the orthogonal projection $y \in M$ of $x \in H$ onto M is given via $y = P_M(x)$. Therefore, we also denote the orthogonal projection by P_M.

LEMMA 2.1.7. *The orthogonal projection $P_M : H \to H$ has the following properties:*

1) *$P_M x \in M$ for all $x \in H$.*
2) *$P_M \in \mathscr{L}(H)$ with $\|P_M\|_{\mathscr{L}(H)} = 1$.*
3) *If $x \in M$, then $P_M x = x$.*
4) *If $x \in M^\perp$, then $P_M x = 0$.*
5) *P_M is self-adjoint, i.e., $P_M^{\circledast} = P_M$.*
6) *$\|x - P_M x\|_H = \inf_{y \in M} \|x - y\|_H$ for all $x \in H$.*
7) *The projection $y = P_M x$ of an arbitrary element $x \in H$ is characterized by the variational equation*

$$\text{Find } y \in M, \qquad (y, v)_H = (x, v)_H \qquad \text{for all } v \in M. \tag{2.1.1}$$

Proof. Can be found in [129, Thm. 2.10.15], [129, Rem. 2.10.17 (ii)], [129, Ex. 3.5.10 (iv)], [137, Prop. 21.44(a), (c)] and Lemma 2.1.6. ☐

2.2. Selected regularity concepts

We start with the definition of lower semicontinuity.

DEFINITION 2.2.1. *Let X be a Banach space and $x \in X$. A function $f : X \to \mathbb{R} \cup \{\infty\}$ is called* lower semicontinuous *at x if for every sequence $(x_n)_{n\in\mathbb{N}} \subset X$,*

$$x_n \to x \quad \Rightarrow \quad f(x) \leq \liminf_{n\to\infty} f(x_n).$$

Remark 2.2.2. We do not have to distinguish between the notions "sequential lower semicontinuity" ([12, Def. 1.33]) and "lower semicontinuity" ([12, Def. 1.21]) because every Banach space X is a sequential topological space and thus these notions coincide, cf. ([12, Rem. 1.37]).

2.2.1. Convex functions

Next we give the basic definitions of convexity. Convexity is a corner stone of modern analysis. We use the precise definitions of the excellent book [12].

DEFINITION 2.2.3 ([12, Prop. 8.4, Def. 10.7]). *Let X be a Banach space. A function $f : X \to \mathbb{R} \cup \{\infty\}$ is called proper if* $\operatorname{dom} f \neq \emptyset$. *A proper function $f : X \to \mathbb{R} \cup \{\infty\}$ is called convex if*

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) \qquad \textit{for all } t \in (0,1), x, y \in \operatorname{dom} f.$$

The function f is called μ-strongly convex, $\mu > 0$, if

$$f(tx + (1-t)y) + t(1-t)\tfrac{\mu}{2}\|x-y\|_X^2 \leq t f(x) + (1-t) f(y) \qquad \textit{for all } t \in (0,1), x, y \in \operatorname{dom} f.$$

DEFINITION 2.2.4 (convex subdifferential, [12, Def. 16.1, Def. 17.1, Prop. 17.14]). *Let $f : X \to \mathbb{R} \cup \{\infty\}$ be proper and convex. For $x \in \operatorname{dom} X$ and $v \in X$, the* directional derivative *of f at x in the direction v is defined by*

$$f'(x;v) := \lim_{t\to+0} \frac{f(y+tv) - f(y)}{t}$$

and the convex subdifferential *of f at x is given by*

$$\partial f(x) := \{x' \in X^* : f'(x;v) \geq \langle x', v\rangle_{X^*,X} \textit{ for all } v \in X\}.$$

PROPOSITION 2.2.5 ([12, Prop. 17.14]). *If $f : X \to \mathbb{R} \cup \{\infty\}$ is proper and convex, then*

$$f(y) - f(x) \geq \langle g, y-x\rangle_{X^*,X} \quad \forall g \in \partial f(x), \ \forall x, y \in \operatorname{dom} f.$$

PROPOSITION 2.2.6 ([12, Prop. 17.26]). *If $f : X \to \mathbb{R} \cup \{\infty\}$ is proper and μ-strongly convex, $\mu > 0$, then*

$$f(y) - f(x) \geq \langle g, y-x\rangle_{X^*,X} + \tfrac{\mu}{2}\|y-x\|_X^2 \quad \forall g \in \partial f(x), \ \forall x, y \in \operatorname{dom} f.$$

2.2.2. Lipschitz functions

Let X be a Banach space and M be a subset of X. A function $f : X \to \mathbb{R}$ is said to be *Lipschitz* on M if there exist a constant $L \geq 0$ such that

$$|f(y) - f(x)| \leq L\|y - x\|_X \qquad \text{for all } x, y \in M.$$

The function f is called Lipschitz *near* $\bar{x} \in X$ if there exist $\delta > 0$ such that f is Lipschitz on $\bar{B}_X(\bar{x}, \delta)$.

DEFINITION 2.2.7 (Clarke's subdifferential). *Let $f : X \to \mathbb{R}$ be Lipschitz near $x \in X$. For $v \in X$, the* generalized directional derivative *of f at x in the direction v is defined by*

$$f^\circ(x; v) := \limsup_{\substack{y \to x, \\ t \downarrow 0}} \frac{f(y + tv) - f(y)}{t}$$

and Clarke's subdifferential *of f at x is given by*

$$\partial_C f(x) := \{x' \in X^* : f^\circ(x; v) \geq \langle x', v\rangle_{X^*, X} \textit{ for all } v \in X\}.$$

LEMMA 2.2.8. *Let X be a Banach space, $V \subset X$ be a bounded set and $F : X \to \mathbb{R}$ be a function which is Lipschitz on bounded sets. Then the set $\bigcup_{v \in V} \partial_C F(v)$ is bounded in X^*.*

Proof. Let $\tilde{V}$ be an open and bounded set such that $V \subset \tilde{V}$ and denote the Lipschitz constant of F on $\tilde{V}$ by L. By [21, Prop. 2.1.2], there holds $\partial_C F(v) \subset \bar{B}_{X^*}(0, L)$ for all $v \in \tilde{V}$. ∎

2.2.3. ε-convex and approximately convex functions

DEFINITION 2.2.9 ([64, Def. 3.3]). *Let X be a Banach space. A function $f : X \to \mathbb{R} \cup \{\infty\}$ is called ε-convex, $\varepsilon \geq 0$, if for each $x, y \in X$ and $t \in [0, 1]$ there holds*

$$f(tx + (1 - t)y) \leq t f(x) + (1 - t) f(y) + \varepsilon t(1 - t)\|x - y\|_X.$$

LEMMA 2.2.10. *If the function $f : X \to \mathbb{R} \cup \{\infty\}$ is ε-convex, then it holds*

$$f(x + u) - f(x) \geq \langle x', u\rangle_{X^*, X} - \varepsilon\|u\|_X \qquad \textit{for all } x, u \in X, x' \in \partial_C f(x).$$

Proof. Let $x, u \in X$ and $x' \in \partial_C f(x)$ be arbitrary. The definition of Clarke's subdifferential yields

$$\langle x', u\rangle_{X^*, X} \leq \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + tu) - f(y)}{t} \leq \limsup_{\substack{y \to x \\ t \downarrow 0}} \Big(f(u + y) - f(y) + \varepsilon(1 - t)\|u\|_X\Big).$$ ∎

First introduced by [128], the class of approximately convex function plays a vital role in nonconvex optimization.

DEFINITION 2.2.11 ([24, Def. 1]). *Let X be a Banach space. A function $f : X \to \mathbb{R} \cup \{\infty\}$ is called approximately convex at $x_0 \in X$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in \bar{B}_X(x_0, \delta)$ and $t \in (0, 1)$*

$$f(tx + (1 - t)y) \leq t f(x) + (1 - t) f(y) + \varepsilon t(1 - t)\|x - y\|_X.$$

THEOREM 2.2.12 ([24, Thm. 2]). *Let f be locally Lipschitz on X and $x_0 \in X$. The function f is approximately convex at x_0 if and only if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x \in \bar{B}_X(x_0, \delta)$ and $x' \in \partial_C f(x)$*

$$f(x+u) - f(x) \geq \langle x', u \rangle_{X^*,X} - \varepsilon \|u\|_X,$$

whenever $\|u\|_X < \delta$ is such that $x + u \in \bar{B}_X(x_0, \delta)$.

Lower-C^k functions

We start with the basic definition of a Gâteaux derivative.

DEFINITION 2.2.13 (Gâteaux derivative). *Let X, Y be Banach spaces and $V \subset X$ be an open set. An operator $f : V \to Y$ is called Gâteaux differentiable at $x \in V$ if for each $h \in X$ the limit*

$$df(x,h) := \lim_{t \to 0^+} \frac{f(x+th) - f(x)}{t}$$

exists and the mapping $df(x,\cdot)$ is bounded and linear. If f is Gâteaux differentiable at every $x \in V$, then f is called Gâteaux differentiable and, in this case, the operator $d_G f : V \to \mathscr{L}(X,Y)$, $x \mapsto df(x,\cdot)$ is called Gâteaux derivative of f. For $k \geq 1$, we denote by $C^k_G(V,Y)$ the set of all functions f for which the k-th Gâteaux derivative $\partial^k f$ exists.

The following concept was first introduced by Rockafellar in [114] and was later transferred to the Hilbert space setting by Penot [107].

DEFINITION 2.2.14 (Lower-C^k). *Let X be a Hilbert space, $k \in \mathbb{N}_+$, and $W \subset X$ be an open set. A function $f : W \to \mathbb{R}$ is called lower-C^k at $w \in W$ if there exists an open neighborhood V of w in W, a compact topological space S and a function $F : S \times V \to \mathbb{R}$ with the property that, for all $s \in S$ and $1 \leq j \leq k$, it holds $F(s,\cdot) \in C^k_G(V,\mathbb{R})$, $(s,x) \mapsto \partial^j_x F(s,x)$ is continuous and $f(x) = \sup_{s \in S} F(s,x)$ for all $x \in V$. If f is lower-C^k at w for all $w \in W$, then f is called lower-C^k.*

Rockafellar proved in the finite dimensional setting ([114, Cor. 6]) that the classes of lower-C^k functions coincide for $2 \leq k \leq \infty$.

LEMMA 2.2.15 ([24, Cor. 3]). *In finite dimensions a locally Lipschitz function is approximately convex if and only if it is lower-C^1.*

2.3. The Lax-Milgram theorem

DEFINITION 2.3.1. *Let H be a real Hilbert space. A bilinear form $a : H \times H \to \mathbb{R}$ is called bounded if there exists a constant $M > 0$ such that*

$$|a(x,y)| \leq M \|x\|_H \|y\|_H \qquad \text{for all } x, y \in H.$$

If there exists a constant $m > 0$ such that

$$a(x,x) \geq m \|x\|_H^2 \qquad \text{for all } x \in H,$$

then $a : H \times H \to \mathbb{R}$ is called coercive.

THEOREM 2.3.2 (Lax-Milgram Theorem). *Let H be a real Hilbert space, let $a : H \times H \to \mathbb{R}$ be a bounded and coercive (with constant m) bilinear form and let $f \in H^*$. Then the variational equation*

$$\textit{Find } z \in H: \qquad a(z,w) = \langle f, w\rangle_{H^*,H} \qquad \textit{for all } w \in H$$

has a unique solution and it holds $\|z\|_H \le m^{-1}\|f\|_{H^*}$.

Proof. The proof of the first part can be found for example in [32, §6.2 Thm. 1]. For the second part, the coercivity of a yields

$$m\|z\|_H^2 \le a(z,z) = \langle f, z\rangle_{H^*,H} \le \|f\|_{H^*}\|z\|_H. \qquad \square$$

COROLLARY 2.3.3. *Let $A \in \mathscr{L}(H)$ be such that the bilinear form $(A\,\cdot,\cdot)_H$ is coercive. Then A is invertible in the sense of Definition 2.1.1.*

Proof. Since $(A\,\cdot,\cdot)_H$ is a bounded coercive bilinear form, Theorem 2.3.2 yields that the variational equation

$$\text{Find } z \in H: \qquad a(z,w) = (f,w)_H \qquad \text{for all } w \in H$$

is uniquely solvable. Denote by $A^{-1} : f \mapsto z$ the solution operator. Obviously, A^{-1} is linear. Denote the coercivity constant of a by m. From $\|A^{-1}f\|_H = \|z\|_H \le m^{-1}\|(f,\cdot)_H\|_{H^*} = m^{-1}\|f\|_H$ we deduce that A^{-1} is a bounded operator. Therefore, $A^{-1} \in \mathscr{L}(H)$ and A is invertible. $\square$

2.4. Variational inequalities

Let H be a Hilbert space, $A \in \mathscr{L}(H,H^*)$ be a symmetric linear operator, $b \in H^*$ and $K \subset H$ be a nonempty closed convex subset of H. We consider the variational inequality

$$\text{Find } y \in K: \qquad \langle Ay - b, v - y\rangle_{H^*,H} \ge 0 \qquad \text{for all } v \in K. \tag{2.4.1}$$

DEFINITION 2.4.1. *An operator $A \in \mathscr{L}(H,H^*)$ is called coercive if there exists a number $C_L > 0$ such that $\langle Ay, y\rangle_{H^*,H} \ge C_L\|y\|_H^2$ for all $y \in H$.*

THEOREM 2.4.2 (e.g., [67, Thm. 2.1]). *If A is coercive with constant C_L, then, for all $b \in H^*$, (2.4.1) has a unique solution $y \in H$ and the solution operator $\mathscr{S} : H^* \to H$, $\mathscr{S}(b) := y$, is Lipschitz with modulus $1/C_L$.*

Now we turn to investigating equivalent reformulations of a variational inequality. To do so, we recall the following definitions from convex analysis. The tangent cone of a closed convex set $K \subset H$ at $y \in K$ is defined by

$$T_K(y) := \{w \in H : \exists t_n \to 0^+, w_n \xrightarrow{H} w : y + t_n w_n \in K\},$$

i.e., $T_K(y)$ is the closed conic hull of $K - y$. Further, the polar cone of a set $M \subset H$ is defined by

$$M^\circ := \{v' \in H^* : \langle v', v\rangle_{H^*,H} \le 0 \text{ for all } v \in M\}.$$

LEMMA 2.4.3. *Let $A \in \mathscr{L}(H,H^*)$ be symmetric, $b \in H^*$ and $\gamma > 0$ be arbitrary. The following formulations are equivalent:*

1) *Find $y \in K$:* $\quad \langle Ay - b, v - y\rangle_{H^*,H} \geq 0 \quad \forall v \in K$.
2) *Find $y \in K$:* $\quad Ay - b \in -T_K(y)^\circ$.
3) *Find $y \in K$:* $\quad y \in \arg\min_{v\in K} \frac{1}{2}\langle Av - b, v\rangle_{H^*,H}$.
4) *Find $y \in K$:* $\quad y = P_K(\, y - \gamma R_H(Ay - b)\,)$.

Proof. $1 \Rightarrow 2$). Let $y \in K$ be a solution of the variational inequality. Let $w \in T_K(y)$ be arbitrary and choose $t_n > 0$ and $w_n \in H$ for all $n \in \mathbb{N}$ such that $v_n := y + t_n w_n \in K$ and $w_n \to w$ in H as $n \to \infty$. Then

$$0 \leq \langle Ay - b, v_n - y\rangle_{H^*,H} = t_n \langle Ay - b, w_n\rangle_{H^*,H} \qquad \text{for all } n \in \mathbb{N}.$$

Taking the limit $n \to \infty$ gives $\langle Ay - b, w\rangle_{H^*,H} \geq 0$ for all $w \in T_K(y)$, i.e., $Ay - b \in -T_K(y)^\circ$.
$2) \Rightarrow 1)$. Let $y \in K$ be such that $Ay - b \in -T_K(y)^\circ$. By the definition of the polar cone, there holds $\langle Ay - b, w\rangle_{H^*,H} \geq 0$ for all $w \in T_K(y)$. As K is convex, $v - y \in T_K(y)$ for arbitrary $v \in K$ and we find $\langle Ay - b, v - y\rangle_{H^*,H} \geq 0$ for all $v \in K$.
$1) \Rightarrow 3)$. Define $g : H \to \mathbb{R}$ by $g(y) := \frac{1}{2}\langle Ay - b, y\rangle_{H^*,H}$. For all $v \in K$ it holds

$$\begin{aligned}
g(v) - g(y) &= \frac{1}{2}\langle Av - b, v\rangle_{H^*,H} - \frac{1}{2}\langle Ay - b, y\rangle_{H^*,H} \\
&\geq \frac{1}{2}\langle Av, v\rangle_{H^*,H} - \frac{1}{2}\langle A(y - v), y - v\rangle_{H^*,H} - \frac{1}{2}\langle Ay, y\rangle_{H^*,H} - \frac{1}{2}\langle b, v - y\rangle_{H^*,H} \\
&= \langle Ay - b, v - y\rangle_{H^*,H} \geq 0.
\end{aligned}$$

$3) \Rightarrow 1)$. [59, Thm. 1.46] implies that the necessary optimality condition $\langle g'(y), v - y\rangle_{H^*,H} \geq 0 \ \forall v \in K$ holds. Since $A \in \mathscr{L}(H,H^*)$ is symmetric, $g'(y) = Ay - b$, i.e., y solves 1).
$1) \Rightarrow 4)$. Can be found in [59, Lem. 1.11]. ∎

2.5. Nonlinear optimization problems

We consider the optimization problem

$$\begin{array}{cl} \text{minimize} & J(u) := j(S(\iota u), u) \\ \text{s.t.} & u \in U_{\text{ad}}. \end{array} \tag{2.5.1}$$

Here, U is a reflexive Banach space, V,W are a Banach spaces, $\iota \in \mathscr{L}(U,V)$ is compact and $S : V \to W$ is continuous. Further, $U_{\text{ad}} \subset U$ is a nonempty, closed and convex set. Denote $W_{\text{ad}} := S(\iota(U_{\text{ad}}))$. The objective function $j : W_{\text{ad}} \times U_{\text{ad}} \to \mathbb{R} \cup \{+\infty\}$ is bounded below and strong×weak sequentially lower semicontinuous, i.e., for each strongly convergent sequence $y_n \to y^*$ in W_{ad} and weakly convergent sequence $u_n \rightharpoonup u^*$ in U_{ad} it holds $\liminf_{n\to\infty} j(y_n, u_n) \geq j(y^*, u^*)$.

THEOREM 2.5.1. *If the reduced objective function $J : U_{ad} \to \mathbb{R} \cup \{+\infty\}$ is coercive, then there exists a solution to problem* (2.5.1).

Proof. As j is bounded below on $W_{\text{ad}} \times U_{\text{ad}}$, it follows that $J(U_{\text{ad}}) \subset j(Y_{\text{ad}}, U_{\text{ad}})$ is bounded below. Let $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}}$ denote an infimizing sequence, i.e., $\lim_{n\to\infty} J(u_n) = \inf_{u\in U_{\text{ad}}} J(u)$. There exists a $\gamma \in \mathbb{R}$

such that, for all $n \in \mathbb{N}$ sufficiently large, the control u_n is in the level set $\mathrm{lev}_\gamma J$, i.e., it holds

$$u_n \in \mathrm{lev}_\gamma J := \{u \in U_{\mathrm{ad}} : J(u) \leq \gamma\}.$$

As J is coercive, the level set $\mathrm{lev}_\gamma J$ is bounded (cf. [12, Prop. 11.12]). This shows that the sequence $(u_n)_{n\in\mathbb{N}} \subset U$ is bounded. As U is a reflexive Banach space, there exists a convergent subsequence, also denoted by $(u_n)_{n\in\mathbb{N}}$, and an element $u^* \in U$ such that $u_n \rightharpoonup u^*$. As $\imath : U \to V$ is compact, it follows that $\imath u_n \to \imath u^*$ strongly in V. Since $S : V \to W$ is continuous, we get $y_n := S(\imath u_n) \to S(\imath u^*) =: y^*$. As U_{ad} is closed and convex, it is weakly closed. Therefore $u^* \in U_{\mathrm{ad}}$ and $y^* \in W_{\mathrm{ad}}$. This gives

$$\inf_{u\in U_{\mathrm{ad}}} J(u) \leq J(u^*) = j(y^*,u^*) \leq \liminf_{n\to\infty} j(y_n,u_n) = \lim_{n\to\infty} J(u_n) = \inf_{u\in U_{\mathrm{ad}}} J(u),$$

showing that (y^*,u^*) is a solution of the problem. □

Remark 2.5.2. If $j : W_{\mathrm{ad}} \times U_{\mathrm{ad}}$ is given via $j(w,u) := j_1(w) + j_2(u)$, where $j_1 : W_{\mathrm{ad}} \to \mathbb{R}$ is bounded below and $j_2 : U_{\mathrm{ad}} \to \mathbb{R}$ is coercive, then J is coercive. This can be seen via

$$J(u) = j_1(S(\imath u))) + j_2(u) \geq c + j_2(u) \to \infty \qquad \text{as } \|u\|_U \to \infty,$$

where $c \in \mathbb{R}$ is the lower bound of j_1.

COROLLARY 2.5.3. *If U_{ad} is bounded, then there exists a solution to problem* (2.5.1).

Proof. Apply Theorem 2.5.1 to the function $\tilde{j} : W \times U \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\tilde{j}(y,u) := \begin{cases} j(y,u) & \text{if } u \in U_{\mathrm{ad}}, \\ \infty & \text{else.} \end{cases}$$

□

2.6. Sobolev spaces

A set $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is called a domain if it is a Lebesgue-measurable set with nonempty interior. For $1 \leq p \leq \infty$, denote by $L^p(\Omega)$ the spaces of (equivalence classes of) Lebesgue integrable functions on an open domain $\Omega \subset \mathbb{R}^d$. We denote by $W^{k,p}(\Omega)$, $k \in \mathbb{N}$, $1 \leq p < \infty$, the Sobolev spaces with corresponding norms

$$\|f\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha|\leq k} \|D_w^\alpha f\|_{L^p(\Omega)}^p \right)^{1/p},$$

where $D_w^\alpha f$ denotes the weak derivative of f corresponding to the multi-index α. We further define the semi-norms $|\cdot|_{W^{k,p}(\Omega)}$ via

$$|f|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha|= k} \|D_w^\alpha f\|_{L^p(\Omega)}^p \right)^{1/p}.$$

For a bounded, open domain $T \subset \mathbb{R}^d$ with Lipschitz boundary, by [31, Thm. 4.6] there exists a linear and bounded operator $\mathrm{tr}_T : W^{1,p}(T) \to L^p(\partial T)$, called the trace, such that $\mathrm{tr}_T f = f$ on ∂T for all $f \in W^{1,p}(T) \cap C^0(\mathrm{cl}(T))$. We denote $H^k(\Omega) := W^{k,2}(\Omega)$, $k \geq 1$, and $H_0^1(\Omega)$ denotes the space of all $H^1(\Omega)$ functions with zero boundary in the sense of traces. We equip $H_0^1(\Omega)$ with the inner product

$$(u,v)_{H_0^1(\Omega)} := \int_\Omega \nabla u(\omega)^\top \nabla v(\omega)\,\mathrm{d}\lambda(\omega) \qquad \text{for all } u,v \in H_0^1(\Omega).$$

By the Poincaré-Friedrich inequality (cf. [15, Cor. 9.19]), there exists a constant $C_{F,\Omega} > 0$ (which depends only on Ω) such that

$$\|w\|_{L^2(\Omega)} \leq C_{F,\Omega} \|\nabla w\|_{L^2(\Omega)^2} = C_{F,\Omega} \|w\|_{H_0^1(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega). \tag{2.6.1}$$

Therefore, the norm $\|\cdot\|_{H_0^1(\Omega)}$ induced by $(\cdot,\cdot)_{H_0^1(\Omega)}$ is equivalent to the norm on $H^1(\Omega)$. The spaces $H_0^1(\Omega)$ and $H^k(\Omega)$, $k \geq 1$, are Hilbert spaces. We denote $H^{-1}(\Omega) := (H_0^1(\Omega))^*$. By the compactness theorem of Rellich (cf. [137, Prop. 19.25]), the embedding $\tilde{\iota} : H_0^1(\Omega) \to L^2(\Omega)$, $\tilde{\iota}(x)(\omega) := x(\omega)$ is compact. Therefore, the (Banach space) adjoint operator $\iota := \tilde{\iota}^*$ is also compact. The operator $\iota : L^2(\Omega) \cong L^2(\Omega)^* \to H^{-1}(\Omega)$ is given via

$$\langle \iota x, y \rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = (x, \tilde{\iota}(y))_{L^2(\Omega)} = \int_\Omega x(\omega) y(\omega)\,\mathrm{d}\lambda(\omega) \qquad \text{for all } x \in L^2(\Omega), y \in H_0^1(\Omega).$$

For a vector valued function $u \in L^2(\Omega,\mathbb{R}^2) = L^2(\Omega)^2$, we denote by ∂_i, $i \in \{1,2\}$ the partial weak derivative with respect to dimension i, whenever this is well-defined. The divergence $\mathrm{div}\,u$ of u is defined by $\mathrm{div}\,u := (\partial_1 + \partial_2)u$. The space $H(\Omega,\mathrm{div}) := \{w \in L^2(\Omega,\mathbb{R}^2) : \mathrm{div}\,w \in L^2(\Omega)\}$ of vector valued square-integrable functions with square-integrable divergence is a Hilbert space when endowed with the inner product $(u,v)_{H(\Omega,\mathrm{div})} := \int_\Omega uv + \mathrm{div}\,u^\top \mathrm{div}\,v\,\mathrm{d}\lambda$. For further information regarding Sobolev spaces we refer to [15, 14, 137].

The Poincaré-Friedrich constant

The smallest possible constant $C_{F,\Omega}$ in (2.6.1) is called the Poincaré-Friedrich constant and plays an important role in practical implementations (cf. Lemma 5.4.2 and Theorem 5.3.3). Thus, we present some known values of $C_{F,\Omega}$ for several domains Ω. Let $\Pi \subset \mathbb{R}^2$ be a rectangle with side lengths a and b. By [76, §2.2], the choice $C_{F,\Pi} := ((\frac{\pi}{a})^2 + (\frac{\pi}{b})^2)^{-1/2}$ is sharp, i.e., $C_{F,\Omega} = C_{F,\Pi}$ is the smallest possible constant in (2.6.1). Whenever $\Omega \subset \hat{\Omega}$, the Friedrich constant $C_{F,\Omega}$ corresponding to $H_0^1(\Omega)$ is smaller or equal to the Friedrich constant $C_{F,\hat{\Omega}}$, cf. [111, Chap. 1.4.3]. Thus, whenever Ω is contained in the rectangle Π, we can estimate $C_{F,\Omega} \leq C_{F,\Pi} = ((\frac{\pi}{a})^2 + (\frac{\pi}{b})^2)^{-1/2}$. Furthermore, a simple scaling argument shows $C_{F,a\Omega+b} = aC_{F,\Omega}$ for all $a > 0$ and $b \in \mathbb{R}^2$. In [106, Tab.3], it is numerically verified that $C_{F,\Omega_L} < 0.162$, where $\Omega_L := (0,1)^2 \setminus [1/2,1)^2$ is the L-shaped domain.

2.7. The finite element method

We give a short review of the finite element method and introduce our notation. More information can be found, e.g., in [14]. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with polygonal boundary. A triangle $T \subset \mathbb{R}^2$

is an open polygonal set with three vertices. A triangulation $\mathscr{T}$ of Ω is a collection of triangles $T \subset \mathbb{R}^2$ such that

1) $\mathrm{cl}(\Omega) = \cup_{T\in\mathscr{T}} \mathrm{cl}(T)$.
2) The intersection of each pair of distinct triangles is either empty, a single vertex or a single edge of both triangles.

Denote by h_T the diameter of $T \in \mathscr{T}$ and by ρ_T the diameter of the largest ball inscribed in T. A family of triangulations of Ω, $(\mathscr{T}^h)_h$, $0 < h \leq 1$, is called regular with parameter $\sigma > 0$, if it holds

$$\max_{T\in\mathscr{T}^h} \frac{h_T}{\rho_T} \leq \sigma \qquad \text{for all } 0 < h \leq 1. \tag{2.7.1}$$

A family of triangulations $(\mathscr{T}^h)_h$, $0 < h \leq 1$, is called quasi-uniform if there exist constants $c_1 > 0$ and $c_2 > 0$ such that

$$\min\{\rho_T : T \in \mathscr{T}^h\} \geq c_1 h \quad \text{and} \quad \max\{h_T : T \in \mathscr{T}^h\} \leq c_2 h \qquad \text{for all } 0 < h \leq 1. \tag{2.7.2}$$

Note that a quasi-uniform family of triangulations is always regular with parameter $\sigma := c_2/c_1$. Now define the piecewise-linear finite element spaces

$$U^h := \{u^h \in C(\bar\Omega) : u^h|_T \in P^1(T) \text{ for all } T \in \mathscr{T}^h\} \quad \text{and} \quad V^h := U^h \cap H^1_0(\Omega). \tag{2.7.3}$$

Here, $P^1(T)$ denotes the set of affine linear functions on the set T. If desired, one could alternatively use piecewise constant finite elements for the space U^h, but this is not carried out here. Denote by n_U the number of nodes of the mesh and by n_V the number of interior nodes. Denote by $n_i \in \mathbb{R}^2$, $1 \leq i \leq n_V$, the interior nodes and by $n_i \in \mathbb{R}^2$, $n_V + 1 \leq i \leq n_U$ the boundary nodes. Define by $\phi_i \in H^1(\Omega)$, $i = 1,\dots,n_U$, the uniquely defined function with $\phi_i(n_i) = 1$, $\phi_i(n_j) = 0$ for $j \neq i$ and $\phi_i|_T \in P^1(T)$. Then $\{\phi_i, 1 \leq i \leq n_U\}$ and $\{\phi_i, 1 \leq i \leq n_V\}$ form a basis of U^h and V^h. For $u^h \in U^h$ and $v^h \in V^h$, we denote by $\mathsf{u} \in \mathbb{R}^{n_U}$ and $\mathsf{v} \in \mathbb{R}^{n_V}$ the coordinates of u^h and v^h, respectively, if it holds $u^h = \sum_{i=1}^{n_U} \mathsf{u}_i \phi_i$ and $v^h = \sum_{j=1}^{n_V} \mathsf{v}_j \phi_j$. Denote by $\mathsf{I} \in \mathbb{R}^{n_U \times n_V}$ the prolongation matrix with $\mathsf{I}_{i,j} = 1$ if $i = j$ and $\mathsf{I}_{i,j} = 0$ otherwise. If we view $v^h \in V^h$ as an element of $H^1_0(\Omega)$, then $u^h := \iota^* v^h \in L^2(\Omega)$ is an element of U^h and the coordinates $\mathsf{u} \in \mathbb{R}^{n_U}$ of u^h are given via $\mathsf{u} = \mathsf{I}\mathsf{v}$, where $\mathsf{v} \in \mathbb{R}^{n_V}$ are the coordinates of v^h. Define the mass matrix $\mathsf{M} \in \mathbb{R}^{n_U \times n_U}$ and the stiffness matrix $\mathsf{K} \in \mathbb{R}^{n_V \times n_V}$ by

$$\mathsf{M} := \left(\int_\Omega \phi_i \phi_j \,\mathrm{d}\lambda\right)_{ij}, \qquad \mathsf{K} := \left(\int_\Omega \nabla\phi_i^\top \nabla\phi_j \,\mathrm{d}\lambda\right)_{ij}.$$

Since the functions ϕ_i are affine linear on each triangle, the mass and stiffness matrices can be assembled efficiently, cf. [34]. Let $u^h, \hat u^h \in U^h$ and $v^h, \hat v^h \in V^h$ be arbitrary with corresponding coordinates $\mathsf{u}, \hat{\mathsf{u}} \in \mathbb{R}^{n_U}$ and $\mathsf{v}, \hat{\mathsf{v}} \in \mathbb{R}^{n_V}$. Then the quantities $(u^h, \hat u^h)_{L^2(\Omega)}$ and $(v^h, \hat v^h)_{H^1_0(\Omega)}$ can be computed exactly via

$$(u^h, \hat u^h)_{L^2(\Omega)} = \int_\Omega u^h \hat u^h \,\mathrm{d}\lambda = \mathsf{u}^\top \mathsf{M} \hat{\mathsf{u}} \quad \text{and} \quad (v^h, \hat v^h)_{H^1_0(\Omega)} = \int_\Omega \nabla {v^h}^\top \nabla \hat v^h \,\mathrm{d}\lambda = \mathsf{v}^\top \mathsf{K} \hat{\mathsf{v}}. \tag{2.7.4}$$

Linear Interpolation

Let $T \subset \mathbb{R}^2$ be a triangle with vertices $n_1, n_2, n_3 \in \mathbb{R}^2$. The (local) Lagrange interpolation operator $I^h_T : C^0(\bar T) \to P^1(T)$ maps any continuous function x to the linear function $I^h_T x$ which fulfills $(I^h_T x)(n_i) =$

$x(n_i)$ for $i = 1,2,3$. For a given triangulation $\mathscr{T}^h$, the (global) Lagrange interpolation operator $I^h : C^0(\mathrm{cl}(\Omega)) \to U^h$ is defined via $I^h x := \sum_{i=1}^{n_U} x(n_i)\phi_i$. Note that $I^h|_T = I_T^h$ for all $T \in \mathscr{T}^h$. We now apply the interpolation estimate [20, Thm. 3.1.6] to our situation.

THEOREM 2.7.1. *Let $(\mathscr{T}^h)_h$, $0 < h \leq 1$, be a regular family of triangulations of a polygonal domain $\Omega \subset \mathbb{R}^2$ and suppose the numbers $p, q \in \mathbb{R}$ fulfill $1 < p < 2$ and $p \leq q \leq 2p/(2-p)$ or $p = 2$ and $q \geq 2$. Then the Lagrange interpolation operator $I_T^h : W^{2,p}(T) \to P^1(T)$ is well-defined and there exists a constant C (depending only on p, q and the regularity parameter σ in* (2.7.1)*) such that for $s \in \{0,1\}$ it holds*

$$|v - I_T^h v|_{W^{s,q}(T)} \leq C h_T^{2-s+2/q-2/p} |v|_{W^{2,p}(T)} \qquad \textit{for all } v \in W^{2,p}(T), T \in \mathscr{T}^h, 0 < h \leq 1.$$

Proof. Let $T \subset \Omega$ be an arbitrary triangle and $s \in \{0,1\}$. By [1, Thm. 4.12], the continuous embedding $W^{2,p}(T) \subset C^0(\bar{T})$ holds true and thus the Lagrange interpolation operator I_T^h is well-defined. Denote by $\hat{T} \subset \mathbb{R}^2$ the reference triangle of the finite element method, i.e., the triangle with vertices $(0,0)$, $(1,0)$ and $(0,1)$. For the given values of s, p, q, [1, Thm. 4.12] implies that the continuous embedding $W^{2,p}(\hat{T}) \subset W^{s,q}(\hat{T})$ holds. Therefore, [20, Thm. 3.1.6] yields

$$|v - I_T^h v|_{W^{s,q}(T)} \leq C h_T^{2-s+2/q-2/p} |v|_{W^{2,p}(T)} \qquad \text{for all } v \in W^{2,p}(T), T \in \mathscr{T}^h, 0 < h \leq 1. \qquad \square$$

COROLLARY 2.7.2. *Let $(\mathscr{T}^h)_h$, $0 < h \leq 1$, be a quasi-uniform family of triangulations of a polygonal domain $\Omega \subset \mathbb{R}^2$ and numbers $p, q \in \mathbb{R}$ with $1 < p < 2$ and $p \leq q \leq 2p/(2-p)$ or $p = 2$ and $q \geq 2$. Then the Lagrange interpolation operator $I^h : W^{2,p}(\Omega) \to U^h$ is well-defined and there exists a constant C (depending only on p, q and the quasi-uniformity constants c_1, c_2 in* (2.7.2)*) such that for $s \in \{0,1\}$ it holds*

$$|v - I^h v|_{W^{s,q}(\Omega)} \leq C h^{2-s+2/q-2/p} |v|_{W^{2,p}(\Omega)} \qquad \textit{for all } v \in W^{2,p}(\Omega), 0 < h \leq 1.$$

Proof. By [1, Thm. 4.12], the continuous embedding $W^{2,p}(\Omega) \subset C^0(\mathrm{cl}(\Omega))$ holds true and thus the Lagrange interpolation operator I_T^h is well-defined. The definition of the Sobolev semi-norm $|\cdot|_{W^{s,q}(\Omega)}$ yields

$$|v|_{W^{s,q}(\Omega)}^q = \sum_{|\alpha|=s} \int_\Omega |D_w^\alpha(v)|^q \,\mathrm{d}\lambda = \sum_{T \in \mathscr{T}^h} \sum_{|\alpha|=s} \int_T |D_w^\alpha(v)|^q \,\mathrm{d}\lambda = \sum_{T \in \mathscr{T}^h} |v|_{W^{s,q}(T)}^q \quad \text{for all } v \in W^{s,q}(\Omega).$$

As $(\mathscr{T}^h)_h$ is a quasi-uniform family of triangulations, there exist constants $c_1, c_2 > 0$ such that

$$\min\{\rho_T : T \in \mathscr{T}^h\} \geq c_1 h \quad \text{and} \quad \max\{h_T : T \in \mathscr{T}^h\} \leq c_2 h \qquad \text{for all } 0 < h \leq 1.$$

Thus, $(\mathscr{T}^h)_h$ is a regular family of triangulations with regularity parameter $\sigma := c_2/c_1$ and it holds $h_T \leq c_2 h$ for all $T \in \mathscr{T}^h$. By Theorem 2.7.1, there exists a constant C, depending only on p, q and c_2/c_1, such that

$$|v - I^h v|_{W^{s,q}(\Omega)} = \left(\sum_{T \in \mathscr{T}^h} |v - I^h v|_{W^{s,q}(T)}^q \right)^{1/q} \leq C c_2 h^{2-s+2/q-2/p} \left(\sum_{T \in \mathscr{T}^h} |v|_{W^{2,p}(T)}^q \right)^{1/q}.$$

Denote by ℓ^r, $1 \le r < \infty$, the space of sequences for which the norm $\|x\|_{\ell^r} := (\sum_{i\in\mathbb{N}} |x_i|^r)^{1/r}$ is finite. By [20, Chap. 2.27], it holds $\|x\|_{\ell^q} \le \|x\|_{\ell^p}$ whenever $1 \le p \le q < \infty$. Thus, we conclude the proof via

$$\left(\sum_{T\in\mathscr{T}^h} |v|^q_{W^{2,p}(T)}\right)^{1/q} \le \left(\sum_{T\in\mathscr{T}^h} |v|^p_{W^{2,p}(T)}\right)^{1/p} = |v|_{W^{2,p}(\Omega)}. \qquad \square$$

LEMMA 2.7.3. *Let $(\mathscr{T}^h)_h$, $0 < h \le 1$, be a regular family of triangulations of Ω. Then there exists a constant $C_\sigma > 0$, depending only on the regularity parameter σ of the triangulations, such that*

$$\|\operatorname{tr}_T v\|^2_{L^2(\partial T)} \le C_\sigma h_T^{-1} \|v\|^2_{L^2(T)} + C h_T |v|^2_{H^1(T)} \qquad \text{for all } v \in H^1(T), T \in \mathscr{T}^h, 0 < h \le 1.$$

Proof. This follows from the fact that $\operatorname{tr}_{\hat{T}} : H^1(\hat{T}) \to L^2(\partial\hat{T})$ is a linear and bounded operator on the reference triangle $\hat{T}$ and a simple scaling argument (cf. [2, Cor. 1.2]). $\square$

2.8. Capacity theory

The concept of capacity plays an important role for the mathematical analysis of the obstacle problem, cf. Chapter 5.

DEFINITION 2.8.1 ([110, Def. 2.1]). *a) Let $d \in \mathbb{N}_+$ and $\Omega \subset \mathbb{R}^d$ be an open set. The capacity of a set $E \subset \Omega$ is defined by*

$$\operatorname{cap}(E) := \inf\left\{ \int_\Omega |\nabla v|^2 \, d\lambda : v \in H^1_0(\Omega), v \ge 1 \text{ a.e. in a neighborhood of } E \right\}.$$

b) A set $O \subset \Omega$ is called quasi-open if for all $\varepsilon > 0$ there exists an open set Ω_ε such that $O \cup \Omega_\varepsilon$ is open. The complement of a quasi-open set is called quasi-closed.

c) A function $v : \Omega \to \mathbb{R}$ is called quasi-continuous (quasi–upper-semicontinuous, quasi-lower-semicontinuous, respectively) if for all $\varepsilon > 0$ there exists an open set $\Omega_\varepsilon \subset \Omega$ with $\operatorname{cap}(\Omega_\varepsilon) < \varepsilon$ such that $v|_{\Omega\setminus\Omega_\varepsilon}$ is continuous (quasi–upper-semicontinuous, lower-semicontinuous, respectively).

If a set has zero capacity, then it has measure zero, but there exist sets with measure zero and positive capacity (cf. [7, Prop. 5.8.5]). We say that a property $P(\omega)$ holds quasi everywhere (q.e.) on $O \subset \Omega$ if it holds for all $\omega \in O \setminus Z$, where $Z \subset O$ is a set with $\operatorname{cap}(Z) = 0$.

LEMMA 2.8.2 ([27, Chap. 8, Thm. 6.1]). *Every $v \in H^1_0(\Omega)$ has a quasi-continuous representative. Any two quasi-continuous representatives are equal quasi-everywhere on Ω.*

In view of this lemma, when we speak about a function in $H^1_0(\Omega)$, we always mean the quasi-continuous representative.

LEMMA 2.8.3 ([131, Lem. 2.3]). *Let $O \subset \Omega$ be a quasi-open subset and $v : \Omega \to \mathbb{R}$ a quasi-continuous function. Then, $v \ge 0$ a.e. on O implies $v \ge 0$ on O.*

In particular, this shows that for two functions $v, w \in H^1_0(\Omega)$, the statement $v \ge w$ q.e. on Ω is equivalent to $v \ge w$ a.e. on Ω. This relation $\ge$ (a.e. or q.e.) defines a partial order on $H^1_0(\Omega)$ and $(H^1_0(\Omega), \ge)$

forms a vector lattice [115, Chap. 4.5]. Therefore, the pointwise maximum $(v)_+ := \max(0,v) \in H_0^1(\Omega)$ exists for all $v \in H_0^1(\Omega)$.

LEMMA 2.8.4 ([110, Lem. 2.3]). *Suppose $v : \Omega \to \mathbb{R}$ is a function. The following assertions are equivalent:*

a) v is quasi-lower-semicontinuous;
b) the sets $\{v > c\}$ are quasi-open for all $c \in \mathbb{R}$;
c) $-v$ is quasi-upper-semicontinuous.

This shows that the set $\{\omega \in \Omega : v(\omega) < 0\}$ is quasi-open and the set $\{\omega \in \Omega : v(\omega) \geq 0\}$ is quasi-closed for any quasi-upper-semicontinuous function v. For a quasi-open set $O \subset \Omega$, the Sobolev space $H_0^1(O)$ is defined via

$$H_0^1(O) := \{v \in H_0^1(\Omega) : v = 0 \text{ q.e. on } \Omega \setminus O\}.$$

As $H_0^1(O)$ is a closed subspace of the Hilbert space $H^1(\mathbb{R}^d)$, it is a Hilbert space.

2.9. Bochner spaces

This section recalls important concepts of Bochner spaces. We base this on the excellent book [62]. Let $(\Xi, \mathscr{A}, P)$ be a measure space and B be a Banach space. A closed valued multifunction $G : \Xi \rightrightarrows B$ is called measurable if the pre-image $G^{-1}(\mathscr{O}) := \{\xi \in \Xi : G(\xi) \cap \mathscr{O} \neq \emptyset\}$ of every open set $\mathscr{O}$ is a measurable set (cf. [8, Def. 8.1.1]). In particular, a function $f : \Xi \to B$ is measurable if the pre-image $f^{-1}(\mathscr{O}) := \{\xi \in \Xi : f(\xi) \in \mathscr{O}\}$ of every open set $\mathscr{O}$ is a measurable set. A function $f : \Xi \to B$ is called P-simple if it is of the form $f(\xi) = \sum_{i=1}^N \delta_{A_n}(\xi) b_n$ where δ_{A_n} is the indicator function of the set $A_n \in \mathscr{A}$ with $P(A_n) < \infty$ and $b_n \in B$ for all $1 \leq n \leq N$ (cf. [62, Def. 1.1.13]). A function $f : \Xi \to B$ is called strongly P-measurable if there exists a sequence of simple functions $f_n : \Xi \to B$ converging to f P-almost everywhere (cf. [62, Def. 1.1.14]).

THEOREM 2.9.1 ([62, Thm. 1.2.20, Prop. 1.1.16, Cor. 1.1.10]). *If B is separable and P is σ-finite, then, for a function $f : \Xi \to B$, the following assertions are equivalent:*

1) f is strongly P-measurable;
2) $\xi \mapsto \langle b', f(\xi)\rangle_{B^*,B}$ is measurable for all $b' \in B^*$;
3) $\tilde{f}$ is measurable and $f = \tilde{f}$ P-a.e.

LEMMA 2.9.2 ([62, Cor. 1.1.23]). *The P-almost everywhere limit $f : \Xi \to B$ of a sequence of strongly P-measurable functions $f_n : \Xi \to B$ is strongly P-measurable.*

Let $\tilde{B}$ be a Banach space. A function $f : \Xi \to \mathscr{L}(B, \tilde{B})$ is called strongly P-measurable if, for all $b \in B$, the $\tilde{B}$-valued function $\xi \mapsto f(\xi)b$ is strongly P-measurable (cf. [62, Def. 1.1.27]).

LEMMA 2.9.3 ([62, Prop. 1.1.28]). *If $f : \Xi \to B$ and $g : \Xi \to \mathscr{L}(B, \tilde{B})$ are strongly P-measurable, then $gf : \Xi \to \tilde{B}$ is strongly P-measurable.*

A strongly P-measurable function $f : \Xi \to B$ is called Bochner integrable with respect to P if there exists a sequence of P-simple functions $f_n : \Xi \to B$ such that

$$\lim_{n\to\infty} \int_\Xi \|f(\xi) - f_n(\xi)\|_B \, dP(\xi) = 0.$$

DEFINITION 2.9.4 ([62, Def. 1.2.15]). *For $1 \leq p < \infty$, we define $L^p(\Xi, B)$ as the linear space of all equivalence classes of strongly P-measurable functions $f : \Xi \to B$ for which*

$$\int_{\Xi} \|f(\xi)\|_B^p \, dP(\xi) < \infty.$$

We define $L^\infty(\Xi, B)$ as the linear space of all equivalence classes of strongly P-measurable functions $f : \Xi \to B$ for which there exits a $r \geq 0$ such that $P(\{\xi \in \Xi : \|f(\xi)\|_B > r\}) = 0$.

Endowed with the norms

$$\|f\|_{L^p(\Xi,B)} := \left(\int_{\Xi} \|f(\xi)\|_B^p \, dP(\xi) \right)^{1/p}$$

and

$$\|f\|_{L^\infty(\Xi,B)} := \inf\{r \geq 0 : P(\{\xi \in \Xi : \|f(\xi)\|_B > r\}) = 0\},$$

the spaces $L^p(\Xi, B)$, $1 \leq p \leq \infty$, are Banach spaces.

THEOREM 2.9.5 ([62, Thm. 1.3.10, Thm. 1.3.21]). *Let $(\Xi, \mathscr{A}, P)$ be a σ-finite measure space, B be a reflexive Banach space, and let $1 \leq p < \infty$ and $1/p + 1/q = 1$. Every $g \in L^q(\Xi, B^*)$ defines an element $\phi_g \in (L^p(\Xi, B))^*$ via*

$$\langle \phi_g, f \rangle_{L^p(\Xi,B)^*, L^p(\Xi,B)} := \int_{\Xi} \langle g(\xi), f(\xi) \rangle_{B^*,B} \, dP(\xi).$$

Furthermore, the mapping $g \mapsto \phi_g$ establishes an isometric isomorphism of Banach spaces

$$L^q(\Xi, B^*) \simeq (L^p(\Xi, B))^*.$$

If H is a Hilbert space, [18, Thm. 3.1] implies that $L^2(\Xi, H)$ is a Hilbert space. Furthermore, if $(\Xi, \mathscr{A}, P)$ is a σ-finite measure space, then Theorem 2.9.5 yields $L^q(\Xi, H^*) \simeq L^p(\Xi, H)^*$ with $1 \leq p < \infty$, $1/p + 1/q = 1$.

3. The bundle method

In this section we present a bundle method for nonsmooth, nonconvex minimization in Hilbert spaces. First results were already published in

[50] L. HERTLEIN AND M. ULBRICH, *An inexact bundle algorithm for nonconvex nonsmooth minimization in Hilbert space*, SIAM J. Control Optim., 57 (2019), pp. 3137–3165.

The results of [50] were extend to the more general setting of locally Lipschitz objective functions in

[49] L. HERTLEIN, A.-T. RAULS, M. ULBRICH, AND S. ULBRICH, *An inexact bundle method and subgradient computations for optimal control of deterministic and stochastic obstacle problems*. Priprint, accepted for publication in SPP1962 Special Issue, Birkhäuser, 2019.

Due to space limitations, most proofs could not be presented in [49]. Instead, the bundle method with a full convergence theory is outlined in this chapter.

3.1. Algorithm

Convergence proofs for bundle methods, especially in a finite-dimensional setting, usually use strongly convergent subsequences of compact sets. In finite dimensions, bounded and closed sets are compact. In infinite dimensions, this is no longer true and in general Hilbert spaces, compactness is a rather strong property. Therefore, we propose the problem setting below. This setting arises naturally in applications such as the optimal control of variational inequalities and enables us to proof convergence also in infinite dimensions. We consider the problem class

$$\min_{x\in X}\; p(\iota x)+w(x), \quad x\in\mathscr{F}. \tag{3.1.1}$$

Here, X and Y are Hilbert spaces and $\iota\in\mathscr{L}(X,Y)$ is a linear, injective and compact operator, i.e., a compact embedding. The constraint set $\mathscr{F}\subset X$ is a nonempty, closed and convex set, $\mathscr{F}_X\supset\mathscr{F}$ is an open convex X-neighborhood of $\mathscr{F}$ and $\mathscr{F}_Y\subset Y$ is an open convex Y-neighborhood of $\iota(\mathscr{F})$ with $\iota(\mathscr{F}_X)\subset\mathscr{F}_Y$. We assume that $p:\mathscr{F}_Y\to\mathbb{R}$ is Lipschitz on bounded sets and define $f:\mathscr{F}_X\to\mathbb{R}$ by $f=p\circ\iota$. Further, we assume that $w:\mathscr{F}_X\to\mathbb{R}$ is continuously differentiable, Lipschitz continuous on bounded sets and μ-strongly convex, cf. Definition 2.2.3. Note that these assumptions imply that w is also weakly sequentially lower semicontinuous. In this section, we use $\partial f(x)$ to denote Clarke's subdifferential if f is Lipschitz near x. We use the same notation for the convex subdifferential if f is a real- or extended real-valued convex function. This does not generate any ambiguity since these subdifferentials coincide for a convex and locally Lipschitz continuous function defined on an open convex set [21, Prop. 2.2.7].

Bundle methods progressively build a local model which approximates the function f around a point x. To do so, in iteration i, a finite set $\mathscr{M}_i$ of affine linear functions, called cutting planes, is selected. The convex function $\phi_i := \max\{m \in \mathscr{M}_i\}$ is chosen as the local model of f at the serious iterate x_i. The bundle method's subproblem is given by

$$\min_{y \in \mathscr{F}} \phi_i(y) + w(y) + \tfrac{1}{2}\|\iota(y - x_i)\|^2_{Q_i + \tau_i R_Y}.$$

Here, τ_i denotes the proximity parameter and Q_i may represent curvature information of p at x_i. A trial iterate $\tilde{y}_i$ is computed as an approximate minimizer of the bundle subproblem such that an approximation of the function value and of the subgradient at $\tilde{y}_i$ can be computed. If this trial iterate $\tilde{y}_i$ fulfills a certain decrease condition, it is accepted as new serious iterate x_{i+1} and we start building a new model around x_{i+1}. Otherwise, a new cutting plane is selected which enriches the old model. If the new model is not substantially improved, the proximity parameter is increased to gather more cutting plane information close to the serious iterate x_i.

3.1.1. The subproblem of the bundle method

In Sections 3.1.4 and 3.1.6, we define the local model ϕ_i and the proximity term $\frac{1}{2}\|\iota(\,\cdot\, - x_i)\|^2_{Q_i + \tau_i R_Y}$ properly. Here, however, it is sufficient to know that ϕ_i is convex and finite on X and that $\|\cdot\|^2_{Q_i + \tau_i R_Y} = \langle (Q_i + \tau_i R_Y)\,\cdot, \cdot\rangle_{Y^*,Y}$ is a norm on Y. The *subproblem of the bundle method* in iteration i is given by

$$\min_{y \in X} \Psi_i(y) := \phi_i(y) + w(y) + \delta_{\mathscr{F}}(y) + \tfrac{1}{2}\|\iota(y - x_i)\|^2_{Q_i + \tau_i R_Y}. \tag{3.1.2}$$

On $\mathscr{F}$, the piecewise quadratic model Ψ_i is the sum of two convex functions, ϕ_i and $\|\iota(\,\cdot\, - x_i)\|^2_{Q_i + \tau_i R_Y}$, and the strongly convex function w. Thus, Ψ_i is strongly convex on $\mathscr{F}$ and the subproblem has a unique minimum $y_i \in \mathscr{F}$. We define the *local model* $\Phi_i : X \to \mathbb{R} \cup \{\infty\}$ by

$$\Phi_i := \begin{Bmatrix} \phi_i + w & \text{on } \mathscr{F} \\ \infty & \text{else} \end{Bmatrix} = \phi_i + w + \delta_{\mathscr{F}}.$$

As w is defined on the open X-neighborhood $\mathscr{F}_X$ of $\mathscr{F}$ and w is finite on $\mathscr{F}_X$, the interior of the effective domain of w is $\mathscr{F}_X$. Furthermore, $\operatorname{int}\operatorname{dom}\phi_i = X$ yields $\operatorname{dom}\delta_{\mathscr{F}} \cap \operatorname{int}\operatorname{dom} w \cap \operatorname{int}\operatorname{dom}\phi_i = \mathscr{F} \neq \emptyset$. Consequently, the sum rule of the convex subdifferential [12, Cor. 16.50] can be applied and yields $\partial\Phi_i = \partial\phi_i + w' + N_{\mathscr{F}}$. The fact that y_i minimizes the subproblem of the bundle method can equivalently be expressed by

$$0 \in \partial\big(\Phi_i + \tfrac{1}{2}\|\iota(\cdot - x_i)\|^2_{Q_i + \tau_i R_Y}\big)(y_i) = \partial\phi_i(y_i) + w'(y_i) + N_{\mathscr{F}}(y_i) + \iota^*(Q_i + \tau_i R_Y)\iota(y_i - x_i),$$

where the equality results from the sum rule of the convex subdifferential [12, Cor. 16.50]. Therefore, there exist elements $g_i^* \in \partial\phi_i(y_i)$ and $n_i^* \in N_{\mathscr{F}}(y_i)$ such that

$$e_i := \iota^*(Q_i + \tau_i R_Y)\iota(x_i - y_i) = g_i^* + w'(y_i) + n_i^* \in \partial\Phi_i(y_i) \tag{3.1.3}$$

and the subgradient inequality gives

$$\langle e_i, y - y_i\rangle_{X^*,X} \leq \Phi_i(y) - \Phi_i(y_i) \qquad \text{for all } y \in X. \tag{3.1.4}$$

Using that w is even μ-strongly convex, we get for all $y \in X$ the improved inequality

$$\langle e_i, y - y_i \rangle_{X^*,X} + \frac{\mu}{2}\|y - y_i\|_X^2 \le \Phi_i(y) - \Phi_i(y_i). \tag{3.1.5}$$

3.1.2. Subdifferential approximation

Most bundle methods use subgradients which are drawn from the convex subdifferential, from Clake's subdifferential or approximations thereof. In order to cover these cases and to facilitate an abstract theory, we define a multifunction G which acts as an abstract subdifferential. We only require properties which are actually needed in the convergence proof. Let $\hat{\mathscr{V}} \subset Y$ be a closed set. We require that $G : \hat{\mathscr{V}} \rightrightarrows Y^*$ fulfills the following

Assumption 3.1.1. The multifunction $G : \hat{\mathscr{V}} \rightrightarrows Y^*$ has the following properties:

1) For all $\hat{v} \in \hat{\mathscr{V}}$, the image $G(\hat{v})$ is nonempty and convex.
2) For all bounded sets $B \subset Y$, the set $G(B \cap \hat{\mathscr{V}}) := \cup_{\hat{v} \in B \cap \hat{\mathscr{V}}} G(\hat{v})$ is bounded in Y^*.
3) G has a weakly closed graph, i.e., for all sequences $(v_n)_{n\in\mathbb{N}} \subset \hat{\mathscr{V}}$ and $(g_n)_{n\in\mathbb{N}} \subset Y^*$ such that $v_n \to \bar{v}$ in Y, $g_n \rightharpoonup g$ in Y^* and $g_n \in G(v_n)$ for all $n \in \mathbb{N}$, it holds $g \in G(\bar{v})$.

We note that property 3) implies that $G(\hat{v})$ is closed for all $\hat{v} \in \hat{\mathscr{V}}$ and property 1) implies that $G(\hat{v})$ is weakly sequentially compact in Y^* for all $\hat{v} \in \hat{\mathscr{V}}$. Both the Clarke subdifferential and the convex subdifferential, i.e., $G := \partial p$, fulfill this assumption (cf. [21, Prop. 2.1.2 (a)], Lemma 2.2.8,[21, Prop. 2.1.5 (b)]). Furthermore, inexact subgradients can be allowed by choosing $G := \partial p + \bar{B}_{Y^*}(0, \varepsilon_G)$, $\varepsilon_G > 0$, or, more general, by choosing $G := \partial p + C$ where $C \subset Y^*$ is a nonempty closed convex set. Another example can be found in Section 7.3.

Due to the inexact nature of the algorithm, we can only expect to converge to approximately stationary points, which we make precise in the following definition:

DEFINITION 3.1.2. *Let the multifunction G fulfill Assumption 3.1.1. A point $\bar{x} \in X$ with $\iota\bar{x} \in \hat{\mathscr{V}}$ is called η-G-stationary, $\eta \ge 0$, if*

$$0 \in w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^*(G(\iota\bar{x}) + \bar{B}_{Y^*}(0, \eta)).$$

A point which is 0-G-stationary is called G-stationary.

Remark 3.1.3. If $G := \partial_C p$ is the Clarke subdifferential, the chain rule [21, Thm. 2.3.10] implies that $\partial_C f(y) \subset \iota^*\partial_C p(\iota y)$ for all $y \in \mathscr{F}_X$. If p or $-p$ is regular at ιy in the sense of Clarke (cf. [21, Def. 2.3.4]), then equality holds at this point. Thus, if p or $-p$ is regular at $\iota\bar{x}$, $\partial_C p$-stationarity is equivalent to $0 \in \partial_C f(\bar{x}) + w'(\bar{x}) + N_{\mathscr{F}}(\bar{x})$. If p is convex and $G := \partial p$, then the sum rule of the convex subdifferential [12, Cor. 16.50] yields $\partial f(\bar{x}) + w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) = \partial J(\bar{x})$, i.e., stationarity is equivalent to $0 \in \partial J(\bar{x})$. Note that if p is convex and $G := \partial p + \bar{B}_{Y^*}(0, \varepsilon_G)$, i.e., one uses subgradients which are at most ε_G away of the true subgradient, then G-stationarity as in Definition 3.1.2 corresponds to $0 \in \partial J(\bar{x}) + \iota^*\bar{B}_{Y^*}(0, \varepsilon_G)$.

3.1.3. Bundle information and trial iterates

For the algorithm, we use points $\tilde{y} \in \mathscr{F}$ at which we compute an approximation $\tilde{f}$ of the exact function value $f(\tilde{y})$. Furthermore, we use points $v \in \mathscr{F}$ at which we compute an approximation $\tilde{g}$ of an element of the subdifferential of p at $\tilde{y}$. The decoupling of points at which function value approximations are drawn from points at which subgradient approximations are drawn might be helpful to find new subgradients which improve the local model. However, although it is possible to draw subgradients at points v, which are far away from the trial iterate $\tilde{y}$, these subgradients might not be useful. We only impose very weak restrictions on $\tilde{y}$, $\tilde{f}$ and v and derive the relevant behavior of Algorithm 3.1 based on this. To obtain concrete optimality statements, one needs to add further exactness requirements on the data $\tilde{y}$, $\tilde{f}$ and v. This can be done in several ways and we carry this out in Section 3.4.

We assume that there is a set $\tilde{\mathscr{T}} \subset \mathscr{F}$ such that for every point $\tilde{y} \in \tilde{\mathscr{T}}$ we can compute a *function value approximation* $\tilde{f} \in \mathbb{R}$ which is assumed to fulfill

$$|\tilde{f} - f(\tilde{y})| \leq \Delta, \tag{3.1.6}$$

where $\Delta \geq 0$ is a constant. Further, we assume that there exists a set $\mathscr{V} \subset \mathscr{F}$ such that, for every $v \in \mathscr{V}$, we can compute an approximation of an element of the subdifferential of the function f at v, i.e., $\iota v \in \hat{\mathscr{V}}$ and we can compute an element of $G(\iota v)$. We call any element $\tilde{g} \in G(\iota v) \subset Y^*$ an *approximate subgradient* at the *subgradient base point* v. Note that for technical reasons we call the element $\tilde{g} \in Y^*$ approximate subgradient and not the canonical choice $\iota^* \tilde{g} \in X^*$. In iteration i of the algorithm, we assume that we can compute an approximation $\tilde{y}_i \in \tilde{\mathscr{T}} \cap \mathscr{V}$ of the minimizer of the bundle subproblem $y_i \in X$ with arbitrary precision, i.e., such that

$$\Psi_i(\tilde{y}_i) - \Psi_i(y_i) \to 0 \qquad \text{as } i \to \infty. \tag{3.1.7}$$

We call $\tilde{y}_i$ a *trial iterate*. Since $0 \in \Psi_i(y_i)$ and Ψ_i is μ-strongly convex, this assumption yields

$$0 \leq \langle 0, \tilde{y}_i - y_i \rangle_{X^*,X} + \tfrac{\mu}{2} \|\tilde{y}_i - y_i\|_X^2 \leq \Psi_i(\tilde{y}_i) - \Psi_i(y_i) \to 0, \tag{3.1.8}$$

i.e., $\tilde{y}_i - y_i \to 0$ in X as $i \to \infty$. Therefore, (3.1.7) implicitly is a richness assumption on $\tilde{\mathscr{T}} \cap \mathscr{V}$ in the sense that $\iota(\tilde{\mathscr{T}} \cap \mathscr{V})$ has to contain a point which is close to ιy_i. Since we do not know the locations of the minimizers y_i in advance, the set $\tilde{\mathscr{T}} \cap \mathscr{V}$ actually has to be dense in $\mathscr{F}$.

For now, we do not impose any further restrictions on $\tilde{\mathscr{T}}$, $\tilde{f}$, $\tilde{y}$, or $\mathscr{V}$. In particular, this means that the function value approximation $\tilde{f}$ might be far away from the exact function value $f(\tilde{y})$ and approximate subgradients $\tilde{g}$ can be drawn at points $v \in \mathscr{V}$ far away from the trial iterate $\tilde{y}$. During the convergence proof we will discover which additional properties are needed to guarantee convergence of the bundle method. In Section 3.5 we discuss several strategies which ensure these properties.

3.1.4. The cutting plane model ϕ_i

Fix the downshift parameter $c > 0$ once and for all. For $x \in X$, $\tilde{f}^x \in \mathbb{R}$, $(\tilde{y}, \tilde{f}, v, \tilde{g}) \in X \times \mathbb{R} \times X \times Y^*$, define the *tangent* $t_{\tilde{y},\tilde{f},\tilde{g}} : X \to \mathbb{R}$, the *downshift* $s_{\tilde{y},\tilde{f},v,\tilde{g},x} \in \mathbb{R}$ and the *downshifted tangent* $m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot,x)$:

$X \to \mathbb{R}$ by

$$\begin{aligned} t_{\tilde{y},\tilde{f},\tilde{g}}(\cdot) &:= \tilde{f} + \langle \tilde{g}, \iota(\cdot - \tilde{y})\rangle_{Y^*,Y}, \\ s_{\tilde{y},\tilde{f},v,\tilde{g},x} &:= [\tilde{f} + \langle \tilde{g}, \iota(x - \tilde{y})\rangle_{Y^*,Y} - \tilde{f}^x]_+ + c\|\iota(v-x)\|_Y^2, \\ m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot, x) &:= t_{\tilde{y},\tilde{f},\tilde{g}}(\cdot) - s_{\tilde{y},\tilde{f},v,\tilde{g},x}. \end{aligned} \tag{3.1.9}$$

We immediately note

LEMMA 3.1.4. *For all* $(\tilde{y}, \tilde{f}, v, \tilde{g}) \in X \times \mathbb{R} \times X \times Y^*$ *and* $x \in X$ *it holds*

$$m_{\tilde{y},\tilde{f},v,\tilde{g}}(y,x) \le \tilde{f}^x + \langle \tilde{g}, \iota(y-x)\rangle_{Y^*,Y} - c\|\iota(v-x)\|_Y^2 \qquad \textit{for all } y \in Y.$$

Proof. For arbitrary $\lambda \in \mathbb{R}$ there holds $-[\lambda]_+ \le -\lambda$. The estimate readily follows from the definition of the downshifted tangent (3.1.9). ∎

Denote by

$$\mathscr{B}_a := \{(\tilde{y}, \tilde{f}, v, \tilde{g}) : \tilde{y} \in \hat{\mathscr{T}}, \tilde{f} \in f(\tilde{y}) + \bar{B}(0,\Delta), v \in \hat{\mathscr{V}}, \tilde{g} \in \mathscr{G}(\iota v)\}$$

the set of all bundle information which can possibly be chosen. In iteration i, the algorithm uses a nonempty, finite set of *bundle information* $\mathscr{B}_i \subset \mathscr{B}_a$. Let $\mathscr{D}_i$ denote the corresponding set of *downshifted tangents*:

$$\mathscr{D}_i := \{m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot, x_i) : (\tilde{y}, \tilde{f}, v, \tilde{g}) \in \mathscr{B}_i\}. \tag{3.1.10}$$

We choose a nonempty, finite subset $\mathscr{M}_i$ of $\mathrm{co}(\mathscr{D}_i)$ to build the *cutting plane model* $\phi_i : X \to \mathbb{R}$ by

$$\phi_i(y) := \max\{m(y) : m \in \mathscr{M}_i\}. \tag{3.1.11}$$

Whenever the serious iterate is not updated, the set $\mathscr{M}_{i+1}$ of the next iteration needs to retain information from the old bundle $\mathscr{M}_i$. Details on how to choose $\mathscr{M}_{i+1}$ are given in the next section.

3.1.5. Aggregation of cutting planes

The canonical choice for the cutting plane model is to include all computed bundle information $\mathscr{B}_i$ by setting $\mathscr{M}_i = \mathscr{D}_i$. This yields the full model

$$\phi_i^{\text{full}} := \max\{m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot, x_i) : (\tilde{y}, \tilde{f}, v, \tilde{g}) \in \mathscr{B}_i\}.$$

However, if the number of elements in $\mathscr{B}_i$ is large, it might be difficult to solve the subproblem of the bundle method. To overcome this problem, one can aggregate or delete cutting planes, i.e., one can choose a (possibly small) subset $\mathscr{M}_i \subset \mathrm{co}(\mathscr{D}_i)$ for the definition of the model ϕ_i. It is important not to delete cutting planes which incorporate relevant information. To make this precise, we need to define the aggregate, exactness and trial cutting plane.

For $m \in \mathscr{M}_i$, denote by $g_m := m'(0) \in X^*$ the gradient of m in X and let $\hat{g}_m \in Y^*$ be the element which fulfills $\iota^* \hat{g}_m = g_m$. As the set $\mathscr{M}_i$ is finite, by [21, Prop. 2.3.12] it holds for all $y \in X$ that

$$\partial \phi_i(y) = \operatorname{co}(\{\iota^* \hat{g}_m : m \in \mathscr{M}_i,\ m(y) = \phi_i(y)\}). \tag{3.1.12}$$

Recall (cf. (3.1.3)) that $g_i^* \in \partial \phi_i(y_i)$. Therefore, there exist numbers $\lambda_m \geq 0$ with $\sum_{m \in \mathscr{M}_i} \lambda_m = 1$ and $g_i^* = \sum_{m \in \mathscr{M}_i} \lambda_m \iota^* \hat{g}_m$. In view of this, we call

$$\hat{g}_i^* := \sum_{m \in \mathscr{M}_i} \lambda_m \hat{g}_m \in Y^* \tag{3.1.13}$$

the *aggregate subgradient*. For $i \in \mathbb{N}$, define the aggregate cutting plane by

$$m_i^*(\cdot) := \sum_{m \in \mathscr{M}_i} \lambda_m m(\cdot). \tag{3.1.14}$$

Furthermore, note that $\lambda_m = 0$ whenever $m(y_i) \neq \phi_i(y_i)$ for all $m \in \mathscr{M}_i$. This implies

$$m_i^*(y_i) = \phi_i(y_i). \tag{3.1.15}$$

The algorithm ensures that only points $x_i \in X$, where at least one approximate subgradient $\tilde{g}_i^x \in G(\iota x_i)$ can be computed, are chosen as the serious iterate. We call $\tilde{g}_i^x$ the *exactness subgradient* and define the *exactness plane* $m_i^x : X \to \mathbb{R}$ by

$$m_i^x(\cdot) := m_{x_i, \tilde{f}_i^x, x_i, \tilde{g}_i^x}(\cdot, x_i) = \tilde{f}_i^x + \langle \tilde{g}_i^x, \iota(\cdot - x_i) \rangle_{Y^*,Y}.$$

We require that every model ϕ_i majorizes the exactness plane model. Furthermore, whenever the current iteration was not successful, we require that the next model majorizes the aggregate cutting plane

Assumption 3.1.5. For each $i \in \mathbb{N}$, the set $\mathscr{M}_i \subset \operatorname{co}(\mathscr{D}_i)$ is chosen such that $\phi_i \geq m_i^x$. If iteration $i \in \mathbb{N}$ is not successful, then $\mathscr{M}_{i+1}$ is chosen such that additionally $\phi_{i+1} \geq m_i^*$ holds.

At the subgradient base point $v_i \in \mathscr{V}$, cf. Section 3.1.3, we compute an approximate subgradient $\tilde{g}_i \in G(\iota v_i)$ and construct the *trial plane* $m_i := m_{\tilde{y}_i, \tilde{f}_i, v_i, \tilde{g}_i} : X \to \mathbb{R}$. This gives

$$m_i(\cdot, x_i) = \tilde{f}_i + \langle \tilde{g}_i, \iota(\cdot - \tilde{y}_i) \rangle_{Y^*,Y} - [\tilde{f}_i + \langle \tilde{g}_i, \iota(x_i - \tilde{y}_i) \rangle_{Y^*,Y} - \tilde{f}_i^x]_+ - c\|\iota(v_i - x_i)\|_Y^2. \tag{3.1.16}$$

Aggregating all cutting planes except the trial plane leads to the aggregate model

$$\phi_{i+1}^{\mathrm{agg}} := \max\{m_i^x(\cdot), m_i(\cdot, x_i), m_i^*(\cdot)\}.$$

However, choosing the aggregate model may lead to a situation where one has to recompute subgradients. Thus it can be beneficial to aggregate only cutting planes with small multiplier λ_m or, alternatively, to delete inactive cutting planes. Further, it may be beneficial to recycle some cutting planes from previous iterations.

LEMMA 3.1.6. *Let $x_i \in X$ be a serious iterate. For all $i \in \mathbb{N}$ it holds $\phi_i(x_i) = \tilde{f}_i^x$.*

Proof. By Assumption 3.1.5 and Lemma 3.1.4 we get for all $i \in \mathbb{N}$

$$\tilde{f}_i^x = m_i^x(x_i) \leq \phi_i(x_i) = \max\{m(x_i) : m \in \mathscr{M}_i\} \leq \max\{m(x_i) : m \in \mathscr{D}_i\} \leq \tilde{f}_i^x.$$

□

3.1.6. Curvature information and proximity control

If there is curvature information of p around ιx_i available, we want to incorporate this into the model. Fix the constants $\bar{q} > 0$, $\xi > 0$ and $T \geq \bar{q} + \xi$. Let $Q_i \in \mathscr{L}(Y, Y^*)$ be symmetric, i.e., $\langle Q_i x, y \rangle_{Y^*,Y} = \langle Q_i y, x \rangle_{Y^*,Y}$ for all $x, y \in Y$. We assume that there exists a constant $0 \leq q_i \leq \bar{q}$ such that

$$\langle Q_i w, w \rangle_{Y^*,Y} \geq -q_i \|w\|_Y^2 \quad \text{for all } w \in Y \qquad \text{and} \qquad \|Q_i\|_{\mathscr{L}(Y,Y^*)} \leq \bar{q}. \tag{3.1.17}$$

Denote by $R_Y : Y \to Y^*$ the Riesz map in Y, i.e., the continuous linear operator which maps $w \in Y$ to $(w, \cdot)_Y$. For any *proximity parameter* $\tau_i \geq q_i + \xi$, the positive definite symmetric bilinear form $\langle (Q_i + \tau_i R_Y) \cdot, \cdot \rangle_{Y^*,Y}$ defines a norm on Y via $\|\cdot\|^2_{Q_i+\tau_i R_Y} := \langle (Q_i + \tau_i R_Y) \cdot, \cdot \rangle_{Y^*,Y}$. Using the relations above, we get the estimate

$$\|\cdot\|^2_{Q_i+\tau_i R_Y} \geq (-q_i + \tau_i)\|\cdot\|_Y^2 \geq \xi \|\cdot\|_Y^2, \tag{3.1.18a}$$

and for large τ_i we can make use of

$$\|\cdot\|^2_{Q_i+\tau_i R_Y} \geq (-q_i + \tau_i)\|\cdot\|_Y^2 \geq (\tau_i - \bar{q})\|\cdot\|_Y^2. \tag{3.1.18b}$$

Remark 3.1.7. In particular, this setting allows for the use of no curvature information, i.e., $Q_i = 0$ for all i. Furthermore, it is possible to use low rank schemes such as the BFGS formula as curvature information, cf. Section 4.4.

3.1.7. Full algorithm and preparation of analysis

In Algorithm 3.1 the inexact bundle method is presented. First we show that in every iteration i the variables ρ_i and $\tilde{\rho}_i$ can be computed. Since $\Psi_i(\tilde{y}_i) < \Psi_i(x_i)$, we find

$$\Phi_i(x_i) - \Phi_i(\tilde{y}_i) = \Psi_i(x_i) - \Psi_i(\tilde{y}_i) + \tfrac{1}{2}\|\tilde{y}_i - x_i\|^2_{Q_i+\tau_i R_Y} > 0. \tag{3.1.19}$$

Consequently both ρ_i and $\tilde{\rho}_i$ are well-defined. Therefore, if the assumptions of Section 3.1.3 are met (in particular if we can compute a trial iterate $\tilde{y}_i \in \tilde{\mathscr{T}} \cap \mathscr{V}$ such that $\Psi_i(\tilde{y}_i) - \Psi_i(y_i) \to 0$ as $i \to \infty$), then every step of Algorithm 3.1 can be executed.

LEMMA 3.1.8. *The sequence of approximate function values at serious iterates, $(\tilde{J}_i^x)_{i\in\mathbb{N}}$, is non-increasing.*

Proof. Since $x_i \in \mathscr{F}$ is feasible, Lemma 3.1.6 gives $\tilde{J}_i^x := \tilde{f}_i^x + w(x) = \Phi_i(x_i)$. For all $i \in \mathbb{N}$ with $\rho_i \geq \gamma$, the estimate (3.1.19) yields

$$\tilde{J}_i^x - \tilde{J}_{i+1}^x = \tilde{J}_i^x - \tilde{J}_i \geq \gamma(\Phi_i(x_i) - \Phi_i(\tilde{y}_i)) > 0, \tag{3.1.20}$$

and $\tilde{J}_{i+1}^x = \tilde{J}_i^x$ for all $i \in \mathbb{N}$ with $\rho_i < \gamma$. Combining this yields $\tilde{J}_{i+1}^x \leq \tilde{J}_i^x$ for all $i \in \mathbb{N}$. ∎

The following lemma exploits the structure of the objective function of the bundle subproblem $\Psi_i = \phi_i + w + \frac{1}{2}\|\cdot - x_i\|^2_{Q_i+\tau_i R_Y}$ to estimate the distance between an arbitrary point $z \in X$ and the minimizer of the bundle subproblem.

Algorithm 3.1: Inexact bundle method

Parameters : $0 < \gamma < \tilde{\gamma} < 1$, $\Delta > 0$, $0 < \bar{q} < \bar{q} + \xi \leq T$. Gradient approximation multifunction $G : \mathscr{F} \rightrightarrows Y^*$ fulfilling Assumption 3.1.1.

Initialization: Choose a start iterate $x_0 \in \tilde{\mathscr{T}} \cap \mathscr{V}$. Compute $\tilde{f}_0^x \in \bar{B}(f(x_0), \Delta)$ and $\tilde{g}_0^x \in G(\iota x_0)$. Set $\tilde{J}_0^x = \tilde{f}_0^x + w(x_0)$ and choose a symmetric operator $Q_0 \in \mathscr{L}(Y, Y^*)$ and $q_0 \leq \bar{q}$ satisfying (3.1.17). Choose $\tau_0 \in [q_0 + \xi, T]$, $\mathscr{B}_0 \supset \{(x_0, \tilde{f}_0^x, x_0, \tilde{g}_0^x)\}$ and $\mathscr{M}_0$ according to Assumption 3.1.5.

for $i = 0, 1, \ldots$ **do**

  Set $\Phi_i = \max\{m : m \in \mathscr{M}_i\} + w + \delta_{\mathscr{F}}$ and $\Psi_i = \Phi_i + \frac{1}{2}\|\iota(\,\cdot\, - x_i)\|^2_{Q_i + \tau_i R_Y}$.

  Trial iterate generation. Compute a trial iterate $\tilde{y}_i \in \tilde{\mathscr{T}} \cap \mathscr{V}$ which fulfills (3.1.7).

  if $\Psi_i(\tilde{y}_i) \geq \Psi_i(x_i)$ **then** (subproblem iteration)

    Set $x_{i+1} = x_i$, $\tilde{f}_{i+1}^x = \tilde{f}_i^x$, $\tilde{J}_{i+1}^x = \tilde{J}_i^x$, $\tilde{g}_{i+1}^x = \tilde{g}_i^x$, $Q_{i+1} = Q_i$, $q_{i+1} = q_i$, $\tau_{i+1} = \tau_i$, $\mathscr{B}_{i+1} = \mathscr{B}_i$, $\Phi_{i+1} = \Phi_i$. Continue to the next iteration.

  else

    Compute $\tilde{f}_i \in \bar{B}(f(\tilde{y}_i), \Delta)$ and set $\tilde{J}_i = \tilde{f}_i + w(\tilde{y}_i)$.

  end

  Acceptance test. Set

$$\rho_i = \frac{\tilde{J}_i^x - \tilde{J}_i}{\Phi_i(x_i) - \Phi_i(\tilde{y}_i)}.$$

  if $\rho_i \geq \gamma$ **then** (successful iteration)

    Set $x_{i+1} = \tilde{y}_i$, $\tilde{f}_{i+1}^x = \tilde{f}_i$ and $\tilde{J}_{i+1}^x = \tilde{J}_i$. Compute $\tilde{g}_{i+1}^x \in G(\iota x_{i+1})$ and choose a symmetric operator $Q_{i+1} \in \mathscr{L}(Y, Y^*)$ and $q_{i+1} \leq \bar{q}$ satisfying (3.1.17). Choose $\tau_{i+1} \in [q_{i+1} + \xi, T]$, $\mathscr{B}_{i+1} \supset \{(x_{i+1}, \tilde{f}_{i+1}^x, x_{i+1}, \tilde{g}_{i+1}^x)\}$ and $\mathscr{M}_{i+1}$ according to Assumption 3.1.5. Continue to the next iteration.

  else

    Set $x_{i+1} = x_i$, $\tilde{f}_{i+1}^x = \tilde{f}_i^x$, $\tilde{g}_{i+1}^x = \tilde{g}_i^x$, $\tilde{J}_{i+1}^x = \tilde{J}_i^x$, $Q_{i+1} = Q_i$, $q_{i+1} = q_i$.

  end

  Update local model. Enrich the set of bundle information by choosing $\mathscr{B}_{i+1}$ such that $\mathscr{B}_i \subset \mathscr{B}_{i+1} \subset \mathscr{B}_a$, choose a set of cutting planes $\mathscr{M}_{i+1}$ according to Assumption 3.1.5 and set $\Phi_{i+1} = \max\{m : m \in \mathscr{M}_{i+1}\} + w + \delta_{\mathscr{F}}$.

  Update proximity parameter.

$$\text{Set } \tilde{\rho}_i = \frac{\tilde{J}_i^x - \Phi_{i+1}(\tilde{y}_i)}{\Phi_i(x_i) - \Phi_i(\tilde{y}_i)} \text{ and update } \tau_{i+1} = \begin{cases} 2\tau_i & \text{if } \tilde{\rho}_i \geq \tilde{\gamma} \text{ (proximity iteration)} \\ \tau_i & \text{if } \tilde{\rho}_i < \tilde{\gamma} \text{ (model iteration)} \end{cases}.$$

end

LEMMA 3.1.9. *For arbitrary $z \in X$, it holds*

$$\tfrac{1}{2}\|\iota(z - y_i)\|^2_{Q_i + \tau_i R_Y} + \tfrac{\mu}{2}\|z - y_i\|^2_X \leq \Psi_i(z) - \Psi_i(y_i).$$

Proof. Let $z \in X$ be arbitrary. Recall from Section 3.1.6 that $\langle Q_i + \tau_i R_Y \,\cdot\,, \cdot\rangle_{Y^*,Y}$ is a symmetric and positive definite bilinear form which defines the inner product $(\cdot,\cdot)_{Q_i + \tau_i R_Y}$ on Y. By the polarization identity

$$\|a\|^2_{Q_i + \tau_i R_Y} = \|a + b\|^2_{Q_i + \tau_i R_Y} - \|b\|^2_{Q_i + \tau_i R_Y} - 2(b, a)_{Q_i + \tau_i R_Y} \qquad \text{for all } a, b \in Y,$$

setting $a = \iota(z-y_i)$ and $b = \iota(y_i - x_i)$, we obtain

$$\tfrac{1}{2}\|\iota(z-y_i)\|^2_{Q_i+\tau_i R_Y} = \tfrac{1}{2}\|\iota(z-x_i)\|^2_{Q_i+\tau_i R_Y} - \tfrac{1}{2}\|\iota(y_i-x_i)\|^2_{Q_i+\tau_i R_Y} + (\iota(x_i-y_i),\iota(z-y_i))_{Q_i+\tau_i R_Y}.$$

Therefore, (3.1.5) implies the claim via

$$\begin{aligned}(\iota(x_i-y_i),\iota(z-y_i))_{Q_i+\tau_i R_Y} + \tfrac{\mu}{2}\|z-y_i\|^2_X &\\ &\le \Phi_i(z) - \Phi_i(y_i)\\ &= \Psi_i(z) - \tfrac{1}{2}\|\iota(z-x_i)\|^2_{Q_i+\tau_i R_Y} - \Psi_i(y_i) + \tfrac{1}{2}\|\iota(y_i-x_i)\|^2_{Q_i+\tau_i R_Y}.\end{aligned}$$ □

LEMMA 3.1.10. *If, from iteration i_0 onwards, Algorithm 3.1 produces only subproblem iterations, then the serious iterate x_{i_0} is G-stationary (in the sense of Definition 3.1.2).*

Proof. Let i_0 be an iteration index such that iteration i is a subproblem iteration for all $i \ge i_0$. Then $\Psi_i = \Psi_{i_0}$, $x_i = x_{i_0}$, $y_i = y_{i_0}$ and $\Psi_{i_0}(\tilde{y}_i) \ge \Psi_{i_0}(x_{i_0})$ for all $i \ge i_0$. By (3.1.7), $\Psi_{i_0}(\tilde{y}_i) \to \Psi_{i_0}(y_{i_0})$ as $i \to \infty$ which yields

$$0 \le \Psi_{i_0}(x_{i_0}) - \Psi_{i_0}(y_{i_0}) = \Psi_{i_0}(x_{i_0}) - \Psi_{i_0}(\tilde{y}_i) + (\Psi_{i_0}(\tilde{y}_i) - \Psi_{i_0}(y_{i_0})) \le \Psi_{i_0}(\tilde{y}_i) - \Psi_{i_0}(y_{i_0}) \to 0,$$

i.e., $\Psi_{i_0}(x_{i_0}) = \Psi_{i_0}(y_{i_0})$. Thus, Lemma 3.1.9 gives $x_{i_0} = y_{i_0}$. Plugging this into (3.1.4) shows

$$0 \le \Phi_{i_0}(y) - \Phi_{i_0}(x_{i_0}) \qquad \text{for all } y \in X.$$

Since Φ_i is convex, this shows that $0 \in \partial\Phi_{i_0}(x_{i_0}) = \partial\phi_{i_0}(x_{i_0}) + w'(x_{i_0}) + N_{\mathscr{F}}(x_{i_0})$. Combining (3.1.12) and Lemma 3.1.6 yields $\partial\phi_{i_0}(x_{i_0}) = \iota^* \operatorname{co}(\{\hat{g}_m : m \in \mathscr{M}_i,\ m(x_{i_0}) = \tilde{f}^x_{i_0}\})$, where $\hat{g}_m \in Y^*$ is defined by $\iota^*\hat{g}_m := m'(0)$ for all $m \in \mathscr{M}_i$. Whenever a subgradient $\tilde{g}$ is drawn at a point $v \ne x_{i_0}$, Lemma 3.1.4 yields for the corresponding downshifted tangent $m = m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot, x_{i_0})$ that $m(x_{i_0}) \ne \tilde{f}^x_{i_0}$ and this downshifted tangent cannot contribute to the subdifferential $\partial\phi_{i_0}(x_{i_0})$. Therefore, only subgradients which are drawn at the serious iterate contribute to $\partial\phi_{i_0}(x_{i_0})$, which results in $\partial\phi_{i_0}(x_{i_0}) \subset \iota^* \operatorname{co}(G(\iota x_{i_0})) = \iota^* G(\iota x_{i_0})$ and $0 \in w'(x_{i_0}) + N_{\mathscr{F}}(x_{i_0}) + \iota^* G(\iota x_{i_0})$. □

Let us define the following sets:

DEFINITION 3.1.11. *We define the set of serious iterates produced by the algorithm by*

$$\mathscr{S} := \{x_i : i \in \mathbb{N}\},$$

the set of all minimizers of the bundle subproblems by

$$\mathscr{T} := \{y_i : i \in \mathbb{N}\},$$

the set of bundle information which is used by the algorithm by

$$\mathscr{B}_u := \cup_{i\in\mathbb{N}} \mathscr{B}_i,$$

and the set of all subgradients used by the algorithm by

$$\tilde{\mathscr{G}} := \{\tilde{g} \in Y^* : \exists \tilde{y}, v \in X, \tilde{f} \in \mathbb{R}, (\tilde{y}, \tilde{f}, v, \tilde{g}) \in \mathscr{B}_u\}.$$

Some of the following lemmas and theorems use the assumption that the set of serious iterates, $\mathscr{S}$, is bounded. This assumption can be guaranteed to hold if further assumptions on the problem data are met. For example, if the constraint set $\mathscr{F}$ is bounded, the set $\mathscr{S} \subset \mathscr{F}$ is automatically bounded. Furthermore, a boundedness assumption on the level set yields boundedness of the set of serious iterates.

LEMMA 3.1.12. *If the initial point $x_0 \in \mathscr{F}$ is such that $\mathscr{F}_0 := \{x \in \mathscr{F} : J(x) \leq J(x_0) + 2\Delta\}$ is bounded in X, then the set of serious iterates $\mathscr{S}$ is bounded in X.*

Proof. Lemma 3.1.8 and the function value boundedness condition (3.1.6) imply

$$J(x_i) - \Delta \leq \tilde{J}_i^x \leq \tilde{J}_0^x \leq J(x_0) + \Delta \qquad \text{for all } i \in \mathbb{N}$$

and thus $\mathscr{S} \subset \mathscr{F}_0$. The boundedness assumption on the level set $\mathscr{F}_0$ implies that the set of serious iterates $\mathscr{S}$ is bounded in X. ∎

LEMMA 3.1.13. *If the set of serious iterates $\mathscr{S}$ is bounded in X, then the sets of minimizers of the bundle subproblems $\mathscr{T}$ is bounded in X and $\iota(\mathscr{T})$ is bounded in Y.*

Proof. Denote by ϕ_i the cutting plane model in iteration i and the exactness subgradient by $\tilde{g}_i^x \in G(\iota x_i)$. By Lemma 3.1.6 it holds $m_i^x(x_i) = \tilde{f}_i^x = \phi_i(x_i)$. Consequently, the definition of the exactness plane m_i^x and Assumption 3.1.5 yield

$$\langle \iota^* \tilde{g}_i^x, y - x_i \rangle_{X^*,X} = m_i^x(y) - m_i^x(x_i) \leq \phi_i(y) - \phi_i(x_i) \qquad \text{for all } y \in X,$$

which shows that $\iota^* \tilde{g}_i^x \in \partial \phi_i(x_i)$. The fact $x_i \in \mathscr{F}$ implies $0 \in N_{\mathscr{F}}(x_i)$. Hence $\iota^* \tilde{g}_i^x + w'(x_i) \in \partial \Phi_i(x_i)$ for all $i \in \mathbb{N}$. Since Φ_i is μ-strongly convex, we get

$$\langle \iota^* \tilde{g}_i^x + w'(x_i), y - x_i \rangle_{X^*,X} + \tfrac{\mu}{2} \|y - x_i\|_X^2 \leq \Phi_i(y) - \Phi_i(x_i) \quad \text{for all } y \in X. \tag{3.1.21}$$

Choosing $y = x_i$ in (3.1.5) yields

$$\|\iota(y_i - x_i)\|_{Q_i + \tau_i R_Y}^2 + \tfrac{\mu}{2} \|y_i - x_i\|_X^2 \leq \Phi_i(x_i) - \Phi_i(y_i). \tag{3.1.22}$$

Thus we get

$$\|\iota(y_i - x_i)\|_{Q_i + \tau_i R_Y}^2 + \mu \|y_i - x_i\|_X^2 \leq \|\iota^* \tilde{g}_i^x + w'(x_i)\|_{X^*} \|y_i - x_i\|_X. \tag{3.1.23}$$

We immediately obtain

$$\|y_i - x_i\|_X \leq \tfrac{1}{\mu} \|\iota^* \tilde{g}_i^x + w'(x_i)\|_{X^*}. \tag{3.1.24}$$

Young's inequality for products gives

$$\|\iota^* \tilde{g}_i^x + w'(x_i)\|_{X^*} \|y_i - x_i\|_X \leq \tfrac{1}{4\mu} \|\iota^* \tilde{g}_i^x + w'(x_i)\|_{X^*}^2 + \mu \|y_i - x_i\|_X^2$$

which shows together with (3.1.23) that

$$\|\iota(y_i - x_i)\|_{Q_i + \tau_i R_Y} \leq \tfrac{1}{2\sqrt{\mu}} \|\iota^* \tilde{g}_i^x + w'(x_i)\|_{X^*}. \tag{3.1.25}$$

From (3.1.18a) we get

$$\|\iota(y_i - x_i)\|_Y \leq \frac{1}{2\sqrt{\mu\xi}}\|\iota^* \tilde{g}_i^x + w'(x_i)\|_{X^*}. \tag{3.1.26}$$

Since the set of serious iterates $\mathscr{S}$ is bounded in X, Assumption 3.1.1 implies that the sequence of exactness gradients $(\iota^*\tilde{g}_i^x)_{i\in\mathbb{N}} \subset \iota^* G(\iota\mathscr{S})$ is bounded and Lemma 2.2.8 yields the boundedness of $w'(\mathscr{S})$. Hence, (3.1.24) shows that the set of minimizers of the bundle subproblem $\mathscr{T}$ is bounded in X and (3.1.26) shows that $\iota^*(\mathscr{T})$ is bounded in Y. ∎

If the set of serious iterates and the set of subgradient base points are bounded, then the Lipschitz continuity of w and the properties of G imply that the set of used bundle information $\mathscr{B}_u$ is bounded. This can be ensured in the following way.

LEMMA 3.1.14. *If the set of serious iterates $\mathscr{S}$ is bounded in X, then the set of trial iterates $\tilde{\mathscr{Y}} := \{\tilde{y}_i : i \in \mathbb{N}\}$ is bounded in X. If additionally the set of all subgradient base points $\mathscr{V}$ used in the algorithm is bounded, then the set of bundle information used by the algorithm, $\mathscr{B}_u$, is bounded in $X \times \mathbb{R} \times X \times Y^*$.*

Proof. If $\mathscr{S}$ is bounded in X, then Lemma 3.1.13 implies that the set of minimizers of the bundle subproblem $\mathscr{T} = \{y_i : i \in \mathbb{N}\}$ is bounded in X. Equation (3.1.8) yields $y_i - \tilde{y}_i \to 0$ in X as $i \to \infty$ which shows that the set of trial iterates $\tilde{\mathscr{Y}} := \{\tilde{y}_i : i \in \mathbb{N}\}$ is bounded in X. Now let additionally $\mathscr{V}$ be bounded in X. This implies that $\iota(\mathscr{V})$ is bounded in Y and Assumption 3.1.1 yields that the set of subgradients $\tilde{\mathscr{G}} \subset G(\iota(\mathscr{V}))$ is bounded in Y^*. The function value boundedness condition (3.1.6) implies the inclusion $\{\tilde{f}_i : i \in \mathbb{N}\} \subset f(\tilde{\mathscr{Y}}) + \bar{B}(0,\Delta)$ holds. Since $\tilde{\mathscr{Y}}$ is bounded and f is Lipschitz on bounded sets, this shows that the set of function value approximations produced by the algorithm, $\{\tilde{f}_i : i \in \mathbb{N}\}$, is bounded. The boundedness of $\mathscr{B}_u$ follows from $\mathscr{B}_u \subset \tilde{\mathscr{Y}} \times \{\tilde{f}_i : i \in \mathbb{N}\} \times \mathscr{V} \times \tilde{\mathscr{G}}$. ∎

The next lemma shows that if the proximity parameter τ_i goes to infinity, then the distance of the minimizer of the subproblem and the serious iterate goes to zero. This fact motivates the name "proximity parameter".

LEMMA 3.1.15. *Assume that there exists a subsequence of iterates $\mathscr{J} \subset \mathbb{N}$ such that $\tau_i \to \infty$ as $\mathscr{J} \ni i \to \infty$. If $\mathscr{S}$ is bounded, then $\iota(y_i - x_i) \to 0$ in Y as $\mathscr{J} \ni i \to \infty$.*

Proof. Since $\mathscr{S}$ is bounded, the sequence $(\|\iota^*\tilde{g}_i^x + w'(x_i)\|_{X^*})_{i\in\mathscr{J}}$ is bounded (cf. Lemma 3.1.13). For all $i \in \mathscr{J}$ sufficiently large such that $\tau_i > \bar{q}$, (3.1.25) and (3.1.18b) yield

$$\|\iota(y_i - x_i)\|_Y \leq \frac{1}{2\sqrt{\mu(\tau_i - \bar{q})}}\|\iota^*\tilde{g}_i^x + w'(x_i)\|_{X^*}.$$

Due to $\tau_i \to \infty$, this shows that $\iota(y_i - x_i) \to 0$ in Y as $\mathscr{J} \ni i \to \infty$. ∎

3.2. The upper envelope function ϕ

Notation 3.2.1. Let $(i_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ be a subsequence of bundle iterations. In the following, we often do not distinguish between the set of indices of the subsequence $\mathscr{J} := \{i_n : n \in \mathbb{N}\}$ and the subsequence itself.

DEFINITION 3.2.2. *Let $\mathscr{I}$ be a subsequence of the sequence of indices produced by Algorithm 3.1 which fulfills $x_i \rightharpoonup \bar{x}$ in X as $\mathscr{I} \ni i \to \infty$ and that $(\tilde{f}_i^x)_{i\in\mathscr{I}}$ converges with $\bar{f} := \lim_{\mathscr{I}\ni i\to\infty} \tilde{f}_i^x < \infty$. Define the upper envelope function $\phi = \phi_{\mathscr{I}} : X \to \mathbb{R}$ by*

$$\phi(y) := \sup\{m_{\tilde{y},\tilde{f},v,\tilde{g}}(y,\bar{x}) : (\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_u\}.$$

LEMMA 3.2.3. *Let $\phi = \phi_{\mathscr{I}}$ be the upper envelope function with corresponding subsequence $\mathscr{I}$ according to Definition 3.2.2. If the set of used bundle information $\mathscr{B}_u$ is bounded (in $X \times \mathbb{R} \times X \times Y^*$), then the following holds:*

1) *$\phi(y) < \infty$ for all $y \in X$.*
2) *$\phi(\bar{x}) = \bar{f}$.*
3) *The functions ϕ and ϕ_i are convex and Lipschitz continuous on X for all $i \in \mathbb{N}$. There exists a constant $L \geq 0$ such that for all $i \in \mathbb{N}$ it holds*

$$|\phi_i(y) - \phi_i(z)| \leq L\|\iota(y-z)\|_Y \qquad \textit{for all } y,z \in X.$$

4) *$\partial\phi(\bar{x}) \subset \iota^* G(\iota\bar{x})$.*

Proof. 1) The definition of the downshifted tangent (3.1.9) yields for all $(\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_u$ and arbitrary but fixed $\bar{x}, y \in X$ that

$$|m_{\tilde{y},\tilde{f},v,\tilde{g}}(y,\bar{x})| \leq |\tilde{f}| + \|\tilde{g}\|_{Y^*}\|\iota(y-\tilde{y})\|_Y + |\tilde{f}-\bar{f}| + \|\tilde{g}\|_{Y^*}\|\iota(\bar{x}-\tilde{y})\|_Y + c\|\iota(v-\bar{x})\|_Y^2.$$

Since $\mathscr{B}_u$ is bounded, the set $\{m_{\tilde{y},\tilde{f},v,\tilde{g}}(y,\bar{x}) : (\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_u\}$ is bounded, too. This guarantees that $\phi(y) < \infty$ for all $y \in X$.

2) For $i \in \mathscr{I}$, denote by $\tilde{g}_i^x \in G(\iota x_i)$ the exactness subgradient at $x_i \in \mathscr{S}$. Consider the scalars

$$\bar{m}_i := m_{x_i,\tilde{f}_i^x,x_i,\tilde{g}_i^x}(\bar{x},\bar{x}) = \tilde{f}_i^x + \langle\tilde{g}_i^x, \iota(\bar{x}-x_i)\rangle_{Y^*,Y} - [\tilde{f}_i^x + \langle\tilde{g}_i^x, \iota(\bar{x}-x_i)\rangle_{Y^*,Y} - \bar{f}]_+ - c\|\iota(x_i-\bar{x})\|_Y^2.$$

Since ι is compact and $x_i \rightharpoonup \bar{x}$ as $\mathscr{I} \ni i \to \infty$, the last term converges to zero as $\mathscr{I} \ni i \to \infty$. Since $(x_i,\tilde{f}_i^x,x_i,\tilde{g}_i^x) \in \mathscr{B}_u$, the set $\{\tilde{g}_i^x : i \in \mathscr{I}\}$ is bounded in Y^*. This shows, together with $\iota x_i \to \iota\bar{x}$ in Y, $\tilde{f}_i^x \to \bar{f}$ and the continuity of the function $[\cdot]_+$ that the right hand side converges to $\bar{f}$, i.e., $\bar{m}_i \to \bar{f}$ as $\mathscr{I} \ni i \to \infty$. By Lemma 3.1.4, it holds $\bar{f} \geq m_{\tilde{y},\tilde{f},v,\tilde{g}}(\bar{x},\bar{x})$ for all $(\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_u$. Therefore we conclude

$$\bar{f} \geq \sup\{m_{\tilde{y},\tilde{f},v,\tilde{g}}(\bar{x},\bar{x}) : (\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_u\} = \phi(\bar{x}) \geq \lim_{i\to\infty} \bar{m}_i = \bar{f}.$$

3) Since $\mathscr{B}_u$ is bounded, also the set of used subgradients, $\tilde{\mathscr{G}}$, is bounded which implies that the supremum $L := \sup_{\tilde{g}\in\mathrm{co}(\tilde{\mathscr{G}})} \|\tilde{g}\|_{Y^*}$ is finite. Fix $i \in \mathbb{N}$ and let $m \in \mathscr{M}_i \subset \mathrm{co}(\mathscr{D}_i)$ be an arbitrary cutting plane. Then $m : X \to \mathbb{R}$ has the form $m(\cdot) = \langle g, \iota\,\cdot\rangle_{Y^*,Y} + b$ with $g \in \mathrm{co}(\tilde{\mathscr{G}})$ and $b \in \mathbb{R}$. Because ϕ and ϕ_i are pointwise suprema of convex and Lipschitz continuous functions on X with modulus $\|\iota^*\|_{L(Y^*,X^*)}L$, they are convex and Lipschitz continuous (see, e.g., [19, Prop. 2.16.5 and Prop. 2.6.3]). Furthermore, the function $\tilde{m} : Y \to \mathbb{R}$ defined by $m(\cdot) := \langle g, \cdot\rangle_{Y^*,Y} + b$ is Lipschitz continuous on Y with constant L. Since $m \in \mathscr{M}_i$ was arbitrary, this implies that $\tilde{\phi}_i(\cdot) := \max_{m\in\mathscr{M}_i} \tilde{m}(\cdot)$ is Lipschitz on Y with constant L and $\phi_i = \tilde{\phi}_i \circ \iota$. Therefore, $|\phi_i(y) - \phi_i(z)| \leq L\|\iota(y-z)\|_Y$ for all $y,z \in X$.

4) Fix a subgradient $g \in \partial\phi(\bar{x}) \subset X^*$ and a direction $w \in X$. For $n \in \mathbb{N}$, choose $t_n > 0$ and $(\tilde{y}_n, \tilde{f}_n, v_n, \tilde{g}_n) \in \mathscr{B}_u$ such that $t_n \to 0$ and $\phi(\bar{x}+t_n w) = m_{\tilde{y}_n,\tilde{f}_n,v_n,\tilde{g}_n}(\bar{x}+t_n w,\bar{x}) + \alpha_n$ with $|\alpha_n| \le t_n^2$. We claim that $\iota v_n \to \iota\bar{x}$ in Y as $n \to \infty$. Lemma 3.1.4 gives

$$c\|\iota(v_n - \bar{x})\|_Y^2 \le \bar{f} - m_{\tilde{y}_n,\tilde{f}_n,v_n,\tilde{g}_n}(\bar{x}+t_n w,\bar{x}) + \langle \tilde{g}_n, t_n \iota w\rangle_{Y^*,Y}$$

which shows that

$$c\|\iota(v_n - \bar{x})\|_Y^2 \le \phi(\bar{x}) - \phi(\bar{x}+t_n w) + \alpha_n + t_n\|\tilde{g}_n\|_{Y^*}\|\iota w\|_Y.$$

By part 3), ϕ is continuous which implies $\phi(\bar{x}+t_n w) \to \phi(\bar{x})$ as $n \to \infty$. Since $\tilde{g}_n \in \tilde{\mathscr{G}}$ and $\tilde{\mathscr{G}}$ is bounded in Y^*, the right hand side of the last inequality converges to zero as $n \to \infty$. This shows $\iota v_n \to \iota\bar{x}$ in Y as $n \to \infty$. Further, for all $n \in \mathbb{N}$, define the affine linear functions $\hat{m}_n : \mathbb{R} \to \mathbb{R}$ by

$$\hat{m}_n(\kappa) := \bar{f} + \kappa\langle g, w\rangle_{X^*,X} - m_{\tilde{y}_n,\tilde{f}_n,v_n,\tilde{g}_n}(\bar{x}+\kappa w,\bar{x}).$$

We calculate the slope of $\hat{m}_n$. Lemma 3.1.4 yields $m_{\tilde{y}_n,\tilde{f}_n,v_n,\tilde{g}_n}(\bar{x},\bar{x}) \le \bar{f}$ or equivalently $\hat{m}_n(0) \ge 0$ for all $n \in \mathbb{N}$. Since ϕ is convex and $g \in \partial\phi(\bar{x})$ it holds

$$\bar{f} + \langle g, y-\bar{x}\rangle_{X^*,X} = \phi(\bar{x}) + \langle g, y-\bar{x}\rangle_{X^*,X} \le \phi(y) \qquad \text{for all } y \in X$$

which gives

$$\hat{m}_n(t_n) = \bar{f} + t_n\langle g, w\rangle_{X^*,X} - \phi(\bar{x}+t_n w) + \alpha_n \le \alpha_n \qquad \text{for all } n \in \mathbb{N}.$$

This yields for the slope of $\hat{m}_n$

$$\langle g - \iota^*\tilde{g}_n, w\rangle_{X^*,X} = \hat{m}_n{}'(0) = \frac{\hat{m}_n(t_n) - \hat{m}_n(0)}{t_n} \le \frac{\alpha_n}{t_n} \le t_n \qquad \text{for all } n \in \mathbb{N}.$$

Since $\tilde{\mathscr{G}}$ is bounded in Y^*, there exists a subsequence $(\tilde{g}_n)_{i\in\mathbb{N}}$ (further denoted with the same indices) and $\tilde{g} \in Y$ with $\tilde{g}_n \rightharpoonup \tilde{g}$ in Y^*. This shows

$$\langle g, w\rangle_{X^*,X} \le \lim_{n\to\infty}\langle \iota^*\tilde{g}_n, w\rangle_{X^*,X} + t_n = \langle \iota^*\tilde{g}, w\rangle_{X^*,X}.$$

Now we conclude the proof. By Assumption 3.1.1, the set $G(\iota\bar{x})$ is nonempty, convex, and weakly sequentially compact in Y^*. Consequently $K := \iota^*G(\iota\bar{x})$ is a nonempty, convex, and closed subset of X^*. Since $\tilde{g}_n \in G(\iota v_n)$, $\iota v_n \to \iota\bar{x}$ in Y, $\tilde{g}_n \rightharpoonup \tilde{g}$ in Y^* and the multifunction $G(\cdot)$ has a weakly closed graph, it holds that $\tilde{g} \in G(\iota\bar{x})$. This shows $\iota^*\tilde{g} \in K$ and

$$\langle g, w\rangle_{X^*,X} \le \langle \iota^*\tilde{g}, w\rangle_{X^*,X} \le \max\{\langle \hat{g}, w\rangle_{X^*,X} : \hat{g} \in K\}.$$

Since $w \in X$ was arbitrary and K is nonempty, convex, and closed in X^*, this implies $g \in K$ by the Hahn-Banach theorem. ∎

THEOREM 3.2.4. *Let $\mathscr{I} \subset \mathbb{N}$ be a subsequence of iterates such that $x_i \rightharpoonup \bar{x}$ in X and $\iota y_i \to \iota\bar{x}$ in Y as $\mathscr{I} \ni i \to \infty$, denote $e_i := \iota^*(Q_i + \tau_i R_Y)\iota(x_i - y_i)$ for $i \in \mathscr{I}$ and let $E \subset X^*$ be the set of strong limit* *points of $(e_i)_{i\in\mathscr{I}}$ in X^*. If $\mathscr{B}_u$ is bounded, then $E \subset \partial w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^* G(\iota\bar{x})$.*

Proof. If $E = \emptyset$, there is nothing to show. Otherwise, let $\bar{e} \in E$ be arbitrary and switch to a subsequence (also denoted by $\mathscr{I}$) to get $e_i \to \bar{e}$ in X^* as $\mathscr{I} \ni i \to \infty$. Since the set of used bundle information $\mathscr{B}_u$ is bounded, also the set of serious iterates $\mathscr{S}$ and the set of subgradient base points $\{\hat{v}_i : i \in \mathscr{I}\} \subset X$ are bounded. The Lipschitz continuity of the function $f : X \to \mathbb{R}$ on the bounded set $\mathscr{S}$ implies that the set of approximate function values, $\{\tilde{f}_i^x : x_i \in \mathscr{S}\} \subset f(\mathscr{S}) + \bar{B}(0,\Delta)$, is bounded. Therefore, there exist elements $\bar{f} \in \mathbb{R}$ and $\bar{v} \in X$ and a further subsequence (again denoted by $\mathscr{I}$) such that $\tilde{f}_i^x \to \bar{f}$, $x_i \rightharpoonup \bar{x}$ in X, $\iota y_i \to \iota\bar{x}$ in Y, $\hat{v}_i \rightharpoonup \bar{v}$ in X and $e_i \to \bar{e}$ in X^* as $\mathscr{I} \ni i \to \infty$. Define the upper envelope function $\phi = \phi_{\mathscr{I}}$ corresponding to $\mathscr{I}$ according to Definition 3.2.2. For all $i \in \mathscr{I}$, denote

$$\Phi_i := \begin{cases} \phi_i + w & \text{on } \mathscr{F} \\ \infty & \text{else} \end{cases} \quad \text{and} \quad \Phi := \begin{cases} \phi + w & \text{on } \mathscr{F} \\ \infty & \text{else} \end{cases}.$$

We divide the proof into three parts.

(i) $\limsup_{\mathscr{I}\ni i\to\infty} \Phi_i(y) \le \Phi(y)$ for all $y \in \mathscr{F}$. Fix $y \in X$. For all $i \in \mathscr{I}$, denote by $\mathscr{M}_i$ the set of cutting planes of the cutting plane model ϕ_i and by $\mathscr{D}_i$ the set of downshifted tangents; cf. (3.1.10). Furthermore, define $\mathscr{E}_i := \{m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot,x_i) : (\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_u\}$. We have $\mathscr{M}_i \subset \operatorname{co}(\mathscr{D}_i) \subset \operatorname{co}(\mathscr{E}_i)$. For all $i \in \mathscr{I}$, choose $\alpha_i > 0$ such that $\alpha_i \to 0$ as $\mathscr{I} \ni i \to \infty$. Then there exist tuples $(\hat{t}_i, \hat{f}_i, \hat{v}_i, \hat{g}_i) \in \mathscr{B}_u$ such that

$$\phi_i(y) = \max_{m\in\mathscr{M}_i} m_i(y) \le \sup_{m\in\operatorname{co}(\mathscr{E}_i)} m(y) = \sup_{m\in\mathscr{E}_i} m(y) \le m_{\hat{t}_i,\hat{f}_i,\hat{v}_i,\hat{g}_i}(y,x_i) + \alpha_i \qquad \text{for all } i \in \mathscr{I}.$$

Define $\hat{m}_i := m_{\hat{t}_i,\hat{f}_i,\hat{v}_i,\hat{g}_i}$ for all $i \in \mathscr{I}$. By definition of the upper envelope function ϕ we get

$$\phi_i(y) - \phi(y) \le \hat{m}_i(y,x_i) + \alpha_i - \hat{m}_i(y,\bar{x}). \tag{3.2.1}$$

For all $a,b \in \mathbb{R}$ it holds: $-[a]_+ + [b]_+ \le [b-a]_+$. Combining this with (3.1.9) shows

$$\hat{m}_i(y,x_i) - \hat{m}_i(y,\bar{x}) \le [\tilde{f}_i^x - \bar{f} + \langle \hat{g}_i, \iota(\bar{x}-x_i)\rangle_{Y^*,Y}]_+ - c\|\iota(\hat{v}_i - x_i)\|_Y^2 + c\|\iota(\hat{v}_i - \bar{x})\|_Y^2.$$

As $\tilde{f}_i^x \to \bar{f}$, $x_i \rightharpoonup \bar{x}$, $\mathscr{B}_u$ is bounded and $[\cdot]_+$ is continuous, the first term of the right hand side converges to zero as $\mathscr{I} \ni i \to \infty$. By construction, we have $\hat{v}_i \rightharpoonup \bar{v}$ in X, which yields

$$\lim_{\mathscr{I}\ni i\to\infty} \|\iota(\hat{v}_i - x_i)\|_Y^2 = \|\iota(\bar{v}-\bar{x})\|_Y^2 = \lim_{\mathscr{I}\ni i\to\infty} \|\iota(\hat{v}_i - \bar{x})\|_Y^2.$$

Since $\alpha_i \to 0$ as $\mathscr{I} \ni i \to \infty$, this shows that the right hand side of (3.2.1) converges to zero. Therefore we obtain for all $y \in X$ that $\limsup_{\mathscr{I}\ni i\to\infty} \phi_i(y) \le \phi(y)$. For $y \in \mathscr{F}$, the value $\delta_{\mathscr{F}}(y) + w(y)$ is finite. This shows the desired result.

(ii) $\liminf_{\mathscr{I}\ni i\to\infty} \Phi_i(y_i) \ge \Phi(\bar{x})$. Denote the exactness plane of iteration i by $m_i^x := m_{x_i,\tilde{f}_i^x,x_i,\tilde{g}_i^x}(\cdot,x_i)$ with the exactness subgradient $\tilde{g}_i^x \in G(\iota x_i) \subset \tilde{\mathscr{G}}$. Since the set of used subgradients $\tilde{\mathscr{G}}$ is bounded in Y^* and

$\iota(x_i - y_i) \to 0$ strongly in Y, the definition of the exactness plane m_i^x gives

$$m_i^x(y_i) = \tilde{f}_i^x + \langle \tilde{g}_i^x, \iota(y_i - x_i)\rangle_{Y^*,Y} \to \bar{f} \qquad \text{as } \mathscr{I} \ni i \to \infty.$$

Denote by $\hat{g}_i^* \in Y^*$ the aggregate subgradient defined in (3.1.13) which fulfills $\iota^*\hat{g}_i^* \in \partial\phi_i(y_i)$. In virtue of (3.1.12) there holds $\hat{g}_i^* \in \mathrm{co}(\tilde{\mathscr{G}})$. Thus we get $\langle \hat{g}_i^*, \iota(x_i - y_i)\rangle_{Y^*,Y} \to 0$. Assumption 3.1.5, the subgradient inequality for $\iota^*\hat{g}_i^* \in \partial\phi_i(y_i)$ and $\phi_i(x_i) = \tilde{f}_i^x$, see Lemma 3.1.6, yield

$$m_i^x(y_i) \le \phi_i(y_i) \le \phi_i(x_i) - \langle \hat{g}_i^*, \iota(x_i - y_i)\rangle_{Y^*,Y} \to \bar{f} \qquad \text{as } \mathscr{I} \ni i \to \infty.$$

Since both left and right hand side converge to $\bar{f}$, we deduce $\phi_i(y_i) \to \bar{f}$. Furthermore, since ι is injective and $\iota y_i \to \iota\bar{x}$, Lemma 2.1.5 implies $y_i \rightharpoonup \bar{x}$ as $\mathscr{I} \ni i \to \infty$. The weak lower semicontinuity of the function w in X and $y_i \rightharpoonup \bar{x}$ in X imply that there holds $\liminf_{\mathscr{I}\ni i\to\infty} w(y_i) \ge w(\bar{x})$. The constraint set $\mathscr{F}$ is assumed to be closed and convex in X. This implies that $\mathscr{F}$ is weakly closed in X and $\bar{x} \in \mathscr{F}$. As $y_i \in \mathscr{F}$, it holds $\delta_{\mathscr{F}}(y_i) = \delta_{\mathscr{F}}(\bar{x}) = 0$ for all $i \in \mathscr{I}$. The second part of Lemma 3.2.3 states that $\bar{f} = \phi(\bar{x})$. Combining this, we obtain

$$\liminf_{\mathscr{I}\ni i\to\infty} \Phi_i(y_i) = \lim_{\mathscr{I}\ni i\to\infty} \phi_i(y_i) + \liminf_{\mathscr{I}\ni i\to\infty} w(y_i) \ge \bar{f} + w(\bar{x}) + \delta_{\mathscr{F}}(\bar{x}) = \Phi(\bar{x}).$$

(iii) $E \subset \partial w'(\bar{x}) + \iota^*G(\iota\bar{x}) + N_{\mathscr{F}}(\bar{x})$. By construction, we have $e_i \to \bar{e}$ in X^* as $\mathscr{I} \ni i \to \infty$. Passing to the limit superior $\mathscr{I} \ni i \to \infty$ for the inequality (3.1.4) gives for all $y \in X$ that

$$\langle \bar{e}, y - \bar{x}\rangle_{X^*,X} = \lim_{\mathscr{I}\ni i\to\infty} \langle e_i, y - y_i\rangle_{X^*,X} \le \limsup_{\mathscr{I}\ni i\to\infty} \Phi_i(y) - \liminf_{\mathscr{I}\ni i\to\infty} \Phi_i(y_i) \le \Phi(y) - \Phi(\bar{x}).$$

This shows that $\bar{e} \in \partial\Phi(\bar{x})$. Recall that w is defined on the open X-neighborhood $\mathscr{F}_X$ of $\mathscr{F}$. As w is finite on $\mathscr{F}_X$, the interior of the effective domain of w is $\mathscr{F}_X$. By Lemma 3.2.3, $\operatorname{int}\operatorname{dom}\phi = X$. Therefore $\operatorname{dom}\delta_{\mathscr{F}} \cap \operatorname{int}\operatorname{dom} w \cap \operatorname{int}\operatorname{dom}\phi = \mathscr{F} \neq \emptyset$. Consequently the sum rule of the convex subdifferential [12, Cor. 16.50] can be applied to $\Phi = \phi + w + \delta_{\mathscr{F}}$ on $\mathscr{F}_X$ and yields $\partial\Phi = \partial\phi + w' + N_{\mathscr{F}}$. Lemma 3.2.3 gives

$$\bar{e} \in \partial\Phi(\bar{x}) \subset w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^*G(\iota\bar{x}).$$

As $\bar{e} \in E$ was arbitrary, we conclude $E \subset \partial\Phi(\bar{x}) \subset w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^*G(\iota\bar{x})$. ☐

COROLLARY 3.2.5. *Let $\mathscr{I}$ be a subsequence of iterates such that $x_i \rightharpoonup \bar{x}$ in X and $\iota y_i \to \iota\bar{x}$ in Y as $\mathscr{I} \ni i \to \infty$. If $\mathscr{B}_u$ is bounded, then $\bar{x}$ is η-G-stationary with $\eta = \liminf_{i\in\mathscr{I}} \|(Q_i + \tau_i R_Y)\iota(x_i - y_i)\|_{Y^*}$.*

Proof. We only consider the case where $\eta < \infty$ since otherwise there is nothing to show. Define $\hat{e}_i := (Q_i + \tau_i R_Y)\iota(x_i - y_i)$ and switch to a subsequence such that $\|\hat{e}_i\|_{Y^*} \to \eta$. Then the sequence $(\hat{e}_i)_{i\in\mathbb{N}}$ is bounded in Y^*. By the Banach-Alaoglu theorem we can switch to a weakly convergent subsequence (also denoted by $(\hat{e}_i)_{i\in\mathbb{N}}$) such that $\hat{e}_i \rightharpoonup \hat{e}$ with $\|\hat{e}\|_{Y^*} \le \eta$. Schauder's Theorem (cf., e.g., [3, Thm. 12.6]) implies that the adjoint $\iota^* \in \mathscr{L}(Y^*, X^*)$ of the compact operator ι is compact which shows $\iota^*\hat{e}_i \to \iota^*\hat{e}$ in X^*. Consequently, Theorem 3.2.4 can be applied which gives $\iota^*\hat{e} \in w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^*G(\iota\bar{x})$. As $\hat{e} \in \bar{B}_{Y^*}(0,\eta)$, this shows that $\bar{x}$ is η-G-stationary. ☐

COROLLARY 3.2.6. *Let $\mathscr{I}$ be a subsequence of iterates such that $x_i \rightharpoonup \bar{x}$ in X and $\tau_i \to \infty$. If $\mathscr{B}_u$ is bounded, then $\bar{x}$ is η-G-stationary with $\eta = \liminf_{i\in\mathscr{I}} \|(Q_i + \tau_i R_Y)\iota(x_i - y_i)\|_{Y^*}$.*

Proof. Lemma 3.1.15 implies that $\iota(y_i - x_i) \to 0$ as $\mathscr{I} \ni i \to \infty$. Therefore, Corollary 3.2.5 yields the desired statement. ∎

3.3. Proof of convergence of the algorithm

LEMMA 3.3.1. *Suppose that from iteration i_0 onward there are only model iterations. If $\mathscr{B}_u$ is bounded, then $\tilde{y}_i \to x_{i_0}$ in X and x_{i_0} is G-stationary.*

Proof. In the given situation, the serious iterate x_i and the proximity parameter τ_i do not change anymore, i.e., $x_i = x_{i_0}$ and $\tau_i = \tau_{i_0}$ for all $i \geq i_0$. Denote $\bar{x} := x_{i_0}$ and $\bar{Q} := Q_{i_0} + \tau_{i_0} R_Y$. We first prove $y_i \to \bar{x}$ in X. Recall from Section 3.1.6 that $\langle \bar{Q}\,\cdot\,,\cdot\rangle_{Y^*,Y}$ is a symmetric and positive definite bilinear form, which implies that $(Y, \langle \bar{Q}\,\cdot\,,\cdot\rangle_{Y^*,Y})$ is an inner product space. By the polarization identity

$$\|a-b\|_{\bar{Q}}^2 = \|a\|_{\bar{Q}}^2 + \|b\|_{\bar{Q}}^2 - 2\langle \bar{Q}b, a\rangle_{Y^*,Y} \qquad \text{for all } a,b \in Y, \tag{3.3.1}$$

setting $a = \iota y_{i+1} - \iota y_i$ and $b = \iota\bar{x} - \iota y_i$ we obtain the equation

$$\|\iota(y_{i+1} - \bar{x})\|_{\bar{Q}}^2 = \|\iota(y_{i+1} - y_i)\|_{\bar{Q}}^2 + \|\iota(y_i - \bar{x})\|_{\bar{Q}}^2 - 2\langle \iota^*\bar{Q}\iota(\bar{x} - y_i), y_{i+1} - y_i\rangle_{Y^*,Y}.$$

Denote by m_i^* the aggregate cutting plane defined in (3.1.14) and set $M_i^* := m_i^* + w + \delta_{\mathscr{F}}$. Equation (3.1.15) implies for $i \geq i_0$ that $y_i \in \mathscr{F}$ and

$$\Psi_i(y_i) = \phi_i(y_i) + w(y_i) + \tfrac{1}{2}\|\iota(y_i - \bar{x})\|_{\bar{Q}}^2 = M_i^*(y_i) + \tfrac{1}{2}\|\iota(y_i - \bar{x})\|_{\bar{Q}}^2.$$

By (3.1.3) it holds for all $i \geq i_0$ that

$$\iota^*\bar{Q}\iota(\bar{x} - y_i) = g_i^* + w'(y_i) + n_i^* \in \partial_1 m_i^*(y_i) + w'(y_i) + N_{\mathscr{F}}(y_i) = \partial M_i^*(y_i).$$

Since M_i^* is μ-strongly convex on $\mathscr{F}$, it follows that

$$M_i^*(y_i) + \tfrac{\mu}{2}\|y_{i+1} - y_i\|_X^2 \leq M_i^*(y_{i+1}) - \langle \iota^*\bar{Q}\iota(\bar{x} - y_i), y_{i+1} - y_i\rangle_{X^*,X}.$$

By Assumption 3.1.5 we have $M_i^*(y_{i+1}) \leq \Phi_{i+1}(y_{i+1})$. Combining all this yields

$$\begin{aligned}
\Psi_i(y_i) &\leq \Psi_i(y_i) + \tfrac{\mu}{2}\|y_{i+1} - y_i\|_X^2 + \tfrac{1}{2}\|\iota(y_i - y_{i+1})\|_{\bar{Q}}^2 \\
&= M_i^*(y_i) + \tfrac{\mu}{2}\|y_{i+1} - y_i\|_X^2 + \tfrac{1}{2}\|\iota(y_i - y_{i+1})\|_{\bar{Q}}^2 + \tfrac{1}{2}\|\iota(y_i - \bar{x})\|_{\bar{Q}}^2 \\
&\leq M_i^*(y_{i+1}) + \tfrac{1}{2}\|\iota(y_{i+1} - \bar{x})\|_{\bar{Q}}^2 \leq \Psi_{i+1}(y_{i+1}) \leq \Psi_{i+1}(\bar{x}) = \bar{f} + w(\bar{x}).
\end{aligned}$$

As the sequence $(\Psi_i(y_i))_{i\in\mathbb{N}}$ is monotonically increasing and bounded from above, it converges. The last chain of estimates shows that

$$0 \leq \tfrac{1}{2}\|\iota(y_{i+1} - y_i)\|_{\bar{Q}}^2 + \tfrac{\mu}{2}\|y_{i+1} - y_i\|_X^2 \leq \Psi_{i+1}(y_{i+1}) - \Psi_i(y_i) \to 0 \qquad \text{as } i \to \infty,$$

implying that $(\|\iota(y_{i+1} - y_i)\|_{\bar{Q}})_{i\in\mathbb{N}}$ and $(\|y_{i+1} - y_i\|_X)_{i\in\mathbb{N}}$ converge to zero. By Lemma 3.1.13, the sequence $(\|\iota(y_i - \bar{x})\|_{\bar{Q}})_{i\in\mathbb{N}}$ is bounded. The Cauchy-Schwarz inequality on the inner product space

$(Y, \langle \bar{Q}\cdot,\cdot\rangle_{Y^*,Y})$ yields

$$\begin{aligned}\left| \|\iota(y_i-\bar{x})\|_{\bar{Q}}^2 - \|\iota(y_{i+1}-\bar{x})\|_{\bar{Q}}^2 \right| &= \left|\langle \bar{Q}\iota(y_i-y_{i+1}), \iota(y_i+y_{i+1}-2\bar{x})\rangle_{Y^*,Y}\right| \\ &\le \|\iota(y_i-y_{i+1})\|_{\bar{Q}} \left(\|\iota(y_i-\bar{x})\|_{\bar{Q}} + \|\iota(y_{i+1}-\bar{x})\|_{\bar{Q}}\right) \to 0.\end{aligned}$$

Combining this with $\Phi_i(y_i) = \Psi_i(y_i) - \frac{1}{2}\|\iota(y_i-\bar{x})\|_{\bar{Q}}^2$, we obtain $\Phi_{i+1}(y_{i+1}) - \Phi_i(y_i) \to 0$. Lemma 3.1.13 states that the sequence minimizers of the bundle subproblem, $(y_i)_{i\in\mathbb{N}}$, is bounded in X. Thus, by the Lipschitz continuity of w on bounded sets,

$$|\phi_{i+1}(y_{i+1}) - \phi_i(y_i)| \le |\Phi_{i+1}(y_{i+1}) - \Phi_i(y_i)| + |w(y_{i+1}) - w(y_i)| \to 0.$$

By Lemma 3.2.3, there exists a constant L such that for all $i \in \mathbb{N}$ and for all $x, y \in X$ the Lipschitz like estimate $|\phi_i(x) - \phi_i(y)| \le L\|\iota(x-y)\|_Y$ holds. From $\|\iota(y_{i+1}-y_i)\|_Y^2 \le \xi^{-1}\|\iota(y_{i+1}-y_i)\|_{\bar{Q}}^2 \to 0$ (cf. 3.1.18a) and $\|\iota(\tilde{y}_i - y_i)\|_Y \to 0$ (cf. (3.1.8)) we infer

$$\begin{aligned}|\Phi_{i+1}(\tilde{y}_i) - \Phi_i(\tilde{y}_i)| &= |\phi_{i+1}(\tilde{y}_i) - \phi_i(\tilde{y}_i)| \\ &\le |\phi_{i+1}(\tilde{y}_i) - \phi_{i+1}(y_{i+1})| + |\phi_{i+1}(y_{i+1}) - \phi_i(y_i)| + |\phi_i(y_i) - \phi_i(\tilde{y}_i)| \\ &\le L\|\iota(\tilde{y}_i - y_{i+1})\|_Y + |\phi_{i+1}(y_{i+1}) - \phi_i(y_i)| + L\|\iota(y_i-\tilde{y}_i)\|_Y \to 0.\end{aligned}$$

Next, we show that $\Phi_i(\tilde{y}_i) \to \bar{J} := \bar{f} + w(\bar{x})$. Since there are only model iterations from iteration i_0 onward, there holds $\tilde{\gamma} > \tilde{\rho}_i$ for all $i \ge i_0$. This gives

$$0 < 1 - \tilde{\gamma} < 1 - \tilde{\rho}_i = 1 - \frac{\bar{J} - \Phi_{i+1}(\tilde{y}_i)}{\bar{J} - \Phi_i(\tilde{y}_i)} = \frac{\Phi_i(\tilde{y}_i) - \Phi_{i+1}(\tilde{y}_i)}{\bar{J} - \Phi_i(\tilde{y}_i)}.$$

Due to $\Phi_i(\tilde{y}_i) - \Phi_{i+1}(\tilde{y}_i) \to 0$ we find $\Phi_i(\tilde{y}_i) \to \bar{J}$ as $i \to \infty$. As a next step we want to show that $\Phi_i(y_i) \to \bar{J}$. From $\iota(\tilde{y}_i - y_i) \to 0$ we deduce

$$\|\iota(y_i - \tilde{y}_i)\|_{\bar{Q}}^2 \le \|\bar{Q}\|_{\mathscr{L}(Y,Y^*)}\|\iota(y_i-\tilde{y}_i)\|_Y^2 \to 0 \qquad \text{as } i \to \infty.$$

Using $a = \iota(y_i - \tilde{y}_i)$ and $b = \iota(\bar{x} - \tilde{y}_i)$ in the polarization identity (3.3.1) gives

$$\|\iota(y_i-\bar{x})\|_{\bar{Q}}^2 - \|\iota(\tilde{y}_i-\bar{x})\|_{\bar{Q}}^2 = \|\iota(y_i-\tilde{y}_i)\|_{\bar{Q}}^2 - 2\langle \bar{Q}\iota(\bar{x}-\tilde{y}_i), \iota(y_i-\tilde{y}_i)\rangle_{Y^*,Y} \to 0 \qquad \text{as } i \to \infty.$$

Thus, (3.1.7) and $\Phi_i(\tilde{y}_i) \to \bar{J}$ as $i \to \infty$ imply

$$\Phi_i(y_i) = \Psi_i(y_i) - \tfrac{1}{2}\|\iota(y_i-\bar{x})\|_{\bar{Q}}^2 - \Psi_i(\tilde{y}_i) + \Phi_i(\tilde{y}_i) + \tfrac{1}{2}\|\iota(\tilde{y}_i-\bar{x})\|_{\bar{Q}}^2 \to \bar{J} \qquad \text{as } i \to \infty.$$

Now we complete the proof. Using (3.1.22) implies $y_i \to \bar{x}$ and $\tilde{y}_i \to \bar{x}$ in X by

$$0 \le \tfrac{\mu}{2}\|y_i - \bar{x}\|_X^2 \le \bar{J} - \Phi_i(y_i) \to 0 \qquad \text{as } i \to \infty. \tag{3.3.2}$$

Consequently, as $e_i := \iota^*\bar{Q}\iota(\bar{x} - y_i) \to 0 \in X^*$ for $i \to \infty$, Theorem 3.2.4 yields the desired result $0 \in w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^* G(\iota\bar{x})$. ∎

Recall that an iteration $i \in \mathbb{N}$ is called proximity iteration if $\rho_i < \gamma$ and $\tilde{\rho}_i \ge \tilde{\gamma}$, cf. Algorithm 3.1. To proof convergence of the bundle method to η-G-stationary points, it is critical to consider the case when there exists an infinite subsequence $\mathscr{I}$ of proximity iterations for which it holds that $\tau_i \to \infty$ and $x_i \rightharpoonup \bar{x}$ as $\mathscr{I} \ni i \to \infty$. In this case, Corollary 3.2.6 yields that $\bar{x}$ is η-G-stationary with $\eta = \liminf_i \|(Q_i + \tau_i R_Y)\iota(x_i - y_i)\|_{Y^*}$. As this statement is valid for any sequence $\mathscr{I}$ which fulfills the above properties and we are interested in the smallest possible stationarity radius η, we proceed as follows. Denote $\hat{e}_i := (Q_i + \tau_i R_Y)\iota(x_i - y_i)$, let $\bar{x} \in X$ be fixed and define the set

$$\mathscr{E}_{\bar{x}} := \big\{\bar{\varepsilon} \in [0,\infty] : \text{ there exists a subsequence of iterations } \mathscr{I} \text{ such that } \rho_i < \gamma, \tilde{\rho}_i \geq \tilde{\gamma}, \tau_i \to \infty, x_i \rightharpoonup \bar{x}, \|\hat{e}_i\|_{Y^*} \to \bar{\varepsilon} \text{ as } \mathscr{I} \ni i \to \infty\big\}. \tag{3.3.3}$$

LEMMA 3.3.2. *If $\mathscr{E}_{\bar{x}} \neq \emptyset$ and $\mathscr{B}_u$ is bounded, then $\bar{x}$ is $\inf \mathscr{E}_{\bar{x}}$-$G$-stationary.*

Proof. Let $(\eta_n)_{n\in\mathbb{N}} \subset \mathscr{E}_{\bar{x}}$ be a sequence such that $\eta_n \to \inf \mathscr{E}_{\bar{x}}$. From Corollary 3.2.6, we infer that $\bar{x}$ is η_n-G-stationary for all $n \in \mathbb{N}$. Therefore there exist elements $z_n \in \bar{B}_{Y^*}(0,\eta_n)$ with $\iota^* z_n \in w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^* G(\iota\bar{x})$. By the Banach-Alaoglu theorem we can switch to a weakly convergent subsequence such that $z_n \rightharpoonup z$ in Y^*, where $z \in Y^*$. As the norm $\|\cdot\|_{Y^*}$ is weakly lower semicontinuous, we get $\|z\|_{Y^*} \leq \liminf_n \|z_n\|_{Y^*} = \eta$. Since $w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^* G(\iota\bar{x})$ is a closed set, also $\iota^* z \in w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^* G(\iota\bar{x})$, i.e., $\bar{x}$ is η-G-stationary. ∎

THEOREM 3.3.3 (Convergence of the bundle method). *If the set $\mathscr{B}_u$ is bounded, then any weak limit point $\bar{x}$ of the sequence $(x_i)_{i\in\mathbb{N}}$ of serious iterates of Algorithm 3.1 is η-G-stationary, where $\eta := \inf \mathscr{E}_{\bar{x}}$ if $\mathscr{E}_{\bar{x}} \neq \emptyset$ and $\eta = 0$ otherwise.*

Proof. Since the set $\mathscr{B}_u$ is bounded, the Banach-Alaoglu theorem ensures the existence of a weak accumulation point of the sequence of serious iterates $(x_i)_{i\in\mathbb{N}}$. Let $\bar{x}$ be an arbitrary weak accumulation point. If there are only subproblem iterations from iteration i_0 onward, then Lemma 3.1.10 yields that the serious iterate x_{i_0} is G-stationary. Since $x_i = x_{i_0}$ for all $i \geq i_0$, the claim follows as obviously $x_i \rightharpoonup \bar{x} = x_{i_0}$. Now assume that there exists a subsequence of iterations which are not subproblem iterations. By renumbering, we can assume without loss of generality that every iteration is not a subproblem iteration. We divide the proof into two cases (1) and 2)) and two subcases each (1a), 1b), 2a), 2b)):

Case 1) First assume that from iteration i_0 onward every iteration is not successful, i.e., $\rho_i < \gamma$ for all $i \geq i_0$. Then $x_i = x_{i_0}$ for all $i \geq i_0$ and we set $\bar{x} = x_{i_0}$. We again consider two cases:

Case 1a) We assume that from iteration $\hat{i}_0 \geq i_0$ onward, there are only model iterations, i.e., $\tilde{\rho}_i < \tilde{\gamma}$ for all $i \geq \hat{i}_0$. In this case, $\mathscr{E}_{\bar{x}} = \emptyset$, i.e., we have to prove that $\bar{x}$ is G-stationary. This is exactly the statement of Lemma 3.3.1.

Case 1b) Now assume that Case 1) holds true but Case 1a) does not hold. Then there exist infinitely many proximity iterations. Denote by $\mathscr{I} \subset \{\hat{i}_0, \hat{i}_0 + 1, \ldots\}$ a subsequence of proximity iterations. Since every iteration from i_0 onward is not successful, the proximity parameter τ_i is never decreased. In each proximity iteration $i \in \mathscr{I}$, the proximity parameter is doubled. Therefore, $\tau_i \to \infty$ as $\mathscr{I} \ni i \to \infty$ and $\mathscr{E}_{\bar{x}} \neq \emptyset$. Thus, Corollary 3.2.6 implies that $\bar{x}$ is η-G-stationary.

Case 2) We now assume that Case 1) does not hold, i.e., for every iteration i, there exists a successful iteration j with $j > i$. Then there exists infinitely many successful iterations and we denote by $i_n \in \mathbb{N}$

the n-th successful iteration for $n \in \mathbb{N}$. This means that $\rho_{i_n} \geq \gamma$ and that the trial iterate $\tilde{y}_{i_n}$ becomes the new serious iterate x_{i_n+1}. The serious iterates are changed only in successful iterations. Therefore, we can choose a subsequence of successful iterations such that $x_{i_n} \rightharpoonup \bar{x}$ as $n \to \infty$ (we keep the same index). Since i_n is not a subproblem iteration (i.e., $\Psi_{i_n}(\tilde{y}_{i_n}) < \Psi_{i_n}(x_{i_n})$), the definition of ρ_i and (3.1.18a) yield

$$\tilde{J}^x_{i_n} - \tilde{J}_{i_n} \geq \gamma\big(\Phi_{i_n}(x_{i_n}) - \Phi_{i_n}(\tilde{y}_{i_n})\big) \geq \gamma\big(\Psi_{i_n}(x_{i_n}) - \Psi_{i_n}(\tilde{y}_{i_n}) + \|\tilde{y}_{i_n} - x_{i_n}\|^2_{Q_{i_n}+\tau_{i_n}R_Y}\big) > \tfrac{\gamma\xi}{2}\|\iota(\tilde{y}_{i_n} - x_{i_n})\|^2_Y.$$

The Lipschitz continuity of the objective function $J = f + w$ on the bounded set $\mathscr{S}$ implies that the set $\{\tilde{J}^x_i : x_i \in \mathscr{S}\} \subset J(\mathscr{S}) + \bar{B}_X(0,\Delta)$ is bounded. Since i_n is a successful iteration, it holds $\tilde{J}^x_{i_n+1} = \tilde{J}_{i_n}$. As $i_{n+1} \geq i_n + 1$ for all $n \in \mathbb{N}$, Lemma 3.1.8 yields that $\tilde{J}^x_{i_{(n+1)}} \leq \tilde{J}^x_{(i_n)+1} = \tilde{J}_{i_n}$. Consequently, for all $N \in \mathbb{N}$, we find

$$\operatorname{diam}\{\tilde{J}^x_i : x_i \in \mathscr{S}\} \geq \tilde{J}^x_{i_0} - \tilde{J}^x_{i_N} = \sum_{n=0}^{N-1} \tilde{J}^x_{i_n} - \tilde{J}^x_{i_{(n+1)}} \geq \sum_{n=0}^{N-1} \tilde{J}^x_{i_n} - \tilde{J}_{i_n} > \frac{\gamma\xi}{2}\sum_{n=0}^{N-1} \|\iota(\tilde{y}_{i_n} - x_{i_n})\|^2_Y,$$

which shows that $\|\iota(\tilde{y}_{i_n} - x_{i_n})\|_Y \to 0$ as $n \to \infty$. Since $x_i \rightharpoonup \bar{x}$ in X as $i \to \infty$ and the operator $\iota \in \mathscr{L}(X,Y)$ is compact, we conclude $x_{i_n} \rightharpoonup \bar{x}$, $\iota x_{i_n} \to \iota\bar{x}$, $\iota\tilde{y}_{i_n} \to \iota\bar{x}$ and (3.1.8) gives $\iota y_{i_n} \to \iota\bar{x}$ in Y as $n \to \infty$. In the following denote $\hat{e}_{i_n} := (Q_{i_n} + \tau_{i_n}R_Y)\iota(x_{i_n} - y_{i_n}) \in Y^*$ as usual.

Case 2a) In the case that $\liminf_n \|\hat{e}_{i_n}\|_{Y^*} = 0$, Corollary 3.2.5 yields that $\bar{x}$ is 0-G-stationary. If $\mathscr{E}_{\bar{x}} \neq \emptyset$, then $\eta = \inf \mathscr{E}_{\bar{x}} \in [0,\infty]$ and if $\mathscr{E}_{\bar{x}} = \emptyset$, then $\eta = 0$. Therefore, $\eta \geq 0$. In particular, this shows that $\bar{x}$ is η-G-stationary.

Case 2b) We now assume that $\liminf_n \|\hat{e}_{i_n}\|_{Y^*} > 0$. We want to show that $\mathscr{E}_{\bar{x}} \neq \emptyset$. To do so, we construct a subsequence of proximity iterates $\mathscr{J}_p$ such that $x_j \rightharpoonup \bar{x}$ and $\tau_j \to \infty$ as $\mathscr{J}_p \ni j \to \infty$. As $\liminf_n \|\hat{e}_{i_n}\|_{Y^*} > 0$, there exists a number $\varepsilon > 0$ and an index $n_0 \in \mathbb{N}$ such that $\|\hat{e}_{i_n}\|_{Y^*} \geq \varepsilon > 0$ for every $n \geq n_0$. By (3.1.17) it holds for every $n \geq n_0$ that

$$\varepsilon \leq \|\hat{e}_{i_n}\|_{Y^*} \leq \|Q_{i_n} + \tau_{i_n}R_Y\|_{\mathscr{L}(Y,Y^*)}\|\iota(y_{i_n} - x_{i_n})\|_Y \leq (\bar{q} + \tau_{i_n})\|\iota(y_{i_n} - x_{i_n})\|_Y.$$

As $\iota(y_{i_n} - x_{i_n}) \to 0$ in Y, we find $\tau_{i_n} \to \infty$ as $n \to \infty$. Now we conclude the proof. Choose $n_T \in \mathbb{N}$ such that $\tau_{i_n} > T$ for all $n \geq n_T$. Denote by i^*_n the last successful iteration before iteration i_n. Line 11 of Algorithm 3.1 ensures that the first proximity parameter after a successful iteration is smaller or equal to T, in particular $\tau_{i^*_n+1} \leq T$ for all $n \in \mathbb{N}$. During an unsuccessful iteration, the proximity parameter is either doubled or remains the same. As $\tau_{i^*_n+1} \leq T$ and $\tau_{i_n} > T$, for all $n \geq n_T$, there exists an index $j_n \in \{i^*_n + 1, \ldots, i_n\}$ such that $2\tau_{j_n} = \tau_{j_n+1} = \tau_{i_n}$. This implies $\tilde{\rho}_{j_n} \geq \tilde{\gamma}$ and $\rho_{j_n} < \gamma$ for all $n \geq n_T$, i.e., j_n is a proximity iteration. Furthermore we have $\tau_{j_n} = \frac{1}{2}\tau_{i_n} \to \infty$ and $x_{j_n} = x_{i_n} \rightharpoonup \bar{x}$ as $n \to \infty$. Therefore $\mathscr{E}_{\bar{x}} \neq \emptyset$ and Corollary 3.2.6 implies that $\bar{x}$ is η-G-stationary. □

Remark 3.3.4. To obtain a meaningful stationarity statement from Theorem 3.3.3, we need to bound the size of $\eta = \inf \mathscr{E}_{\bar{x}}$. This issue is addressed in the next section.

Remark 3.3.5. In practice, it is difficult to select appropriate cutting planes for aggregation. In order to fulfill Assumption 3.1.5, in particular the assumption $\phi_{i+1} \geq m^*_i$ for not successful iterations i, one needs knowledge of the exact dual weights λ_m, $m \in \mathscr{M}_i$ of the aggregate cutting plane m^*_i. However, to obtain these weights, the exact value of $g^*_i = -\iota^*(Q_i + \tau_i R_Y)\iota(x_i - y_i) - w'(y_i) - n^*_i$ has to be computed, which basically requires to solve the bundle subproblem exactly. By checking the proof of Theorem 3.3.3, we find that the assumption $\phi_{i+1} \geq m_i^*$ is used only in Lemma 3.3.1. Careful inspection shows that this condition can be weakened to $\phi_{i+1}(y_{i+1}) \geq m_i^*(y_{i+1})$. Note that the exact minimizer

$$y_{i+1} = \underset{y \in X}{\arg\min}\, \phi_{i+1}(y) + w(y) + \delta_{\mathscr{F}}(y) + \tfrac{1}{2}\|\iota(y - x_i)\|^2_{Q_i + \tau_i R_Y}$$

depends on ϕ_{i+1} which makes it difficult to select a model ϕ_{i+1} that fulfills $\phi_{i+1}(y_{i+1}) \geq m_i^*(y_{i+1})$. As a heuristic, one still can compute an approximate aggregate cutting plane $\tilde{m}_i^*$ and exclude all cutting planes $m \in \mathscr{M}_i$ from the next model which fulfill $m < \tilde{m}_i^*$. However, in this case one can not use the convergence statement Theorem 3.3.3. To circumvent this difficulty, we always use the full model which uses all computed bundle information by setting $\mathscr{M}_i = \mathscr{D}_i$. Then Assumption 3.1.5 is always true. In Chapter 4 we develop an algorithm which efficiently computes an approximation of the bundle subproblem. This algorithm adaptively selects which cutting planes should be included into the model and thus can be viewed as an automated aggregation strategy.

The case $\mathbf{w \equiv 0}$ and $\mathscr{F}$ bounded

If the function w is set to zero ($w \equiv 0$) and the feasible set $\mathscr{F}$ is bounded in X, similar convergence statements hold true. In this case, we cannot use (3.1.5) anymore and we have to work with (3.1.4). Therefore, Lemma 3.1.9 has to be changed to

LEMMA 3.3.6. *For arbitrary $z \in X$, it holds $\frac{1}{2}\|\iota(z - y_i)\|^2_{Q_i + \tau_i R_Y} \leq \Psi_i(z) - \Psi_i(y_i)$.*

Combining Lemma 3.3.6 with (3.1.18a) yields

$$\tfrac{\xi}{2}\|\iota(z - y_i)\|_Y^2 \leq \tfrac{1}{2}\|\iota(z - y_i)\|^2_{Q_i + \tau_i R_Y} \leq \Psi_i(z) - \Psi_i(y_i) \qquad \text{for all } z \in X. \tag{3.3.4}$$

Lemma 3.1.10 still holds true, but we have to argue differently to show $x_{i_0} = y_{i_0}$. We deduce from $\Psi_{i_0}(x_{i_0}) = \Psi_{i_0}(y_{i_0})$ and (3.3.4) that $\iota x_{i_0} = \iota y_{i_0}$ and the injectivity of ι yields $x_{i_0} = y_{i_0}$. Since (3.1.5) does not hold, (3.1.8) cannot be used to infer $\tilde{y}_i - x_i \to 0$ strongly in X as $i \to \infty$. However, since $\mathscr{F}$ is bounded in X, also the set of minimizers of the bundle subproblem $\mathscr{T} \subset \mathscr{F}$ and the set of trial iterates $\{\tilde{y}_i : i \in \mathbb{N}\} \subset \mathscr{F}$ are bounded which implies that both Lemmas 3.1.13 and 3.1.14 hold true. The statement of Lemma 3.1.15 remains the same, but the proof has to be changed to

Proof. Since $\mathscr{S}$ is bounded, Assumption 3.1.1 implies that the sequence $(\|\tilde{g}_i^x\|_{Y^*})_{i \in \mathscr{J}}$ is bounded. For all $i \in \mathscr{I}$ sufficiently large such that $\tau_i > \bar{q}$, (3.1.18b), (3.1.22) and (3.1.21) yield

$$(\tau_i - \bar{q})\|\iota(y_i - x_i)\|_Y^2 \leq \|\iota(y_i - x_i)\|^2_{Q_i + \tau_i R_Y} \leq \Phi_i(x_i) - \Phi_i(y_i) \leq \|\tilde{g}_i^x\|_{Y^*}\|\iota(y_i - x_i)\|_Y$$

Due to $\tau_i \to \infty$, this shows that $\iota(y_i - x_i) \to 0$ in Y as $\mathscr{J} \ni i \to \infty$. □

Furthermore, Lemma 3.3.1 has to be changed to

LEMMA 3.3.7. *Suppose that from iteration i_0 onward there are only model iterations. If $\mathscr{B}_u$ is bounded, then $\iota\tilde{y}_i \to \iota x_{i_0}$ in Y and x_{i_0} is G-stationary.*

The proof of Lemma 3.3.7 is the same as the proof of Lemma 3.3.1 except that we do not need to show that $w(y_{i+1}) - w(y_i) \to 0$ and thus we do not need the μ-strong convexity of w to show $\|y_{i+1} - y_i\|_X \to 0$

as $i \to \infty$. Also (3.3.2) has to be changed to

$$0 \le \xi \|\iota(y_i - x_i)\|_Y^2 \le \|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y}^2 \le \bar{J} - \Phi_i(y_i) \to 0,$$

which follows from (3.1.22) and (3.1.18a). Combining (3.3.4) with (3.1.7) yields $\iota(\tilde{y}_i - x_i) \to 0$ in Y as $i \to \infty$. With these changes, the proof for Theorem 3.3.3 is still valid with $w \equiv 0$ and $\mathscr{F}$ bounded in X.

3.4. Inexactness schemes

In this chapter, we turn on developing tangible estimates which imply η-G-stationarity of any weak limit point of Algorithm 3.1. First we consider the situation of [50, Thm. 5.5]. There, the following exactness conditions on the function value and the trial iterate are employed for all $i \in \mathbb{N}$:

$$\|\iota(\tilde{y}_i - y_i)\|_Y \le M \|\iota(y_i - x_i)\|_Y, \qquad M \ge 0, \tag{3.4.1}$$

$$\theta(\Phi_i(x_i) - \Phi_i(y_i)) \le \Phi_i(x_i) - \Phi_i(\tilde{y}_i), \qquad 0 < \theta < 1, \tag{3.4.2}$$

$$\tilde{f}_i - f(\tilde{y}_i) \le \tilde{f}_i^x - f(x_i) + \mathring{\varepsilon}_{1,i} \|\iota(\tilde{y}_i - x_i)\|_Y, \qquad \mathring{\varepsilon}_{1,i} \to \mathring{\varepsilon}_1,\ \mathring{\varepsilon}_1 \ge 0. \tag{3.4.3}$$

In successful iterations the new set of cutting planes $\mathscr{M}_{i+1}$ is chosen as the set containing only the exactness plane, i.e., $\mathscr{M}_{i+1} := \{m_{x_{i+1}, \tilde{f}_{i+1}^x, x_{i+1}, \tilde{g}_{i+1}^x}\}$. In unsuccessful iterations it is assumed that a subgradient $\tilde{g}_i \in Y^*$ can be computed which fulfills

$$\tilde{g}_i \in G(\iota\tilde{y}_i), \ \ G = \partial_C p + \bar{B}_{Y^*}(0, \Delta_2), \qquad \Delta_2 \ge 0, \tag{3.4.4}$$

$$\operatorname{dist}(\tilde{g}_i, \partial_C p(\iota\tilde{y}_i)) \le \mathring{\varepsilon}_{2,i}, \qquad \mathring{\varepsilon}_{2,i} \to \mathring{\varepsilon}_2,\ \mathring{\varepsilon}_2 \ge 0, \tag{3.4.5}$$

and the new set of cutting planes $\mathscr{M}_{i+1}$ is chosen as $\mathscr{M}_{i+1} := \mathscr{M}_i \cup \{m_{\tilde{y}_i, \tilde{f}_i, \tilde{y}_i, \tilde{g}_i}\}$, i.e., only the trial plane is added to the set of cutting planes. The following lemma reproduces [50, Thm. 5.5] in a slightly more general form.

LEMMA 3.4.1. *Assume that the initial point $x_0 \in \mathscr{F}$ is such that the level set $\mathscr{F}_0 := \{x \in \mathscr{F} : J(x) \le J(x_0) + 2\Delta\}$ is bounded in X. If p is approximately convex on $\iota\mathscr{F}$, $\tilde{y}_i$ fulfills* (3.4.1) *and* (3.4.2)*, $\tilde{f}_i$ fulfills* (3.4.3)*, $\tilde{g}_i$ fulfills* (3.4.4) *and* (3.4.5) *and $v_i = \tilde{y}_i$ for all $i \in \mathbb{N}$, then every weak limit point of the sequence of serious iterates is $(\frac{M+1}{\theta(\tilde{\gamma}-\gamma)}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2) + \Delta_2)$-$\partial_C p$-stationary.*

Proof. First we show that the set of used bundle information, $\mathscr{B}_u$, is bounded. Since the level set $\mathscr{F}_0$ is bounded in X, Lemma 3.1.12 implies that the set of serious iterates $\mathscr{S}$ is bounded in X. Since all subgradient base points are trial iterates, Lemma 3.1.14 shows that the set of bundle information $\mathscr{B}_u$ is bounded in $X \times \mathbb{R} \times X \times Y^*$. Let $\bar{x}$ be a weak limit point of the sequence of serious iterates. By Theorem 3.3.3, $\bar{x}$ is η-G-stationary, where $\eta = \inf \mathscr{E}_{\bar{x}}$ if $\mathscr{E}_{\bar{x}} \neq \emptyset$ and $\eta = 0$ otherwise. By the definition of η-G-stationarity, Definition 3.1.2, $\bar{x}$ is η-G-stationary with $G = \partial_C p + \bar{B}_{Y^*}(0, \Delta_2)$ if and only if $\bar{x}$ is $(\eta + \Delta_2)$-$\partial_C p$-stationary. So, if $\mathscr{E}_{\bar{x}} = \emptyset$, then $\eta = 0$ and we are done. Thus assume that $\mathscr{E}_{\bar{x}} \neq \emptyset$, i.e., that there exists $\bar{\varepsilon} \in \mathscr{E}_{\bar{x}}$ and a subsequence of proximity iterates $\mathscr{I} \subset \mathbb{N}$ such that $\tau_i \to \infty$, $x_i \to \bar{x}$ and $\|\hat{e}_i\|_{Y^*} \to \bar{\varepsilon}$ as $\mathscr{I} \ni i \to \infty$. We aim at bounding $\bar{\varepsilon}$. By definition of Q_i, cf. (3.1.17), it holds $\|Q_i\|_{\mathscr{L}(Y,Y^*)} \le \bar{q}$ and thus $\|Q_i + \tau_i R_Y\|_{\mathscr{L}(Y,Y^*)} \le \tau_i + \bar{q}$. Combining this with (3.1.18b) yields

$$\|\hat{e}_i\|_{Y^*} = \|(Q_i + \tau_i R_Y)\iota(x_i - y_i)\|_{Y^*} \le (\tau_i + \bar{q})\|\iota(y_i - x_i)\|_Y \le \frac{\tau_i + \bar{q}}{\tau_i - \bar{q}} \frac{\|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y}^2}{\|\iota(y_i - x_i)\|_Y}. \tag{3.4.6}$$

Furthermore, (3.1.4) and (3.4.2) give

$$\|\iota(y_i-x_i)\|^2_{Q_i+\tau_i R_Y} \le \Phi_i(x_i)-\Phi_i(y_i) \le \theta^{-1}(\Phi_i(x_i)-\Phi_i(\tilde{y}_i)). \tag{3.4.7}$$

For every proximity iteration $i \in \mathscr{I}$, lines 9 and 16 of Algorithm 3.1 yield

$$\frac{\tilde{J}_i - \Phi_{i+1}(\tilde{y}_i)}{\Phi_i(x_i)-\Phi_i(\tilde{y}_i)} = \tilde{\rho}_i - \rho_i \ge \tilde{\gamma} - \gamma,$$

i.e.,

$$\Phi_i(x_i)-\Phi_i(\tilde{y}_i) \le \tfrac{1}{\tilde{\gamma}-\gamma}\left(\tilde{J}_i - \Phi_{i+1}(\tilde{y}_i)\right) = \tfrac{1}{\tilde{\gamma}-\gamma}\left(\tilde{f}_i - \phi_{i+1}(\tilde{y}_i)\right). \tag{3.4.8}$$

Since the trial plane

$$m_i(\cdot,x_i) = \tilde{f}_i + \langle \tilde{g}_i, \iota(\cdot - \tilde{y}_i)\rangle_{Y^*,Y} - [\tilde{f}_i + \langle \tilde{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}^x_i]_+ - c\|\iota(\tilde{y}_i - x_i)\|^2_Y$$

is included into the model, i.e., $\phi_{i+1} \ge m_i(\,\cdot\,,x_i)$, we find

$$\tilde{f}_i - \phi_{i+1}(\tilde{y}_i) \le \tilde{f}_i - m_i(\tilde{y}_i,x_i) = [\tilde{f}_i + \langle \tilde{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}^x_i]_+ + c\|\iota(\tilde{y}_i-x_i)\|^2_Y. \tag{3.4.9}$$

Next we estimate the linearization error $e_{\text{lin}} := \tilde{f}_i + \langle \tilde{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}^x_i$. Since $\mathscr{B}_u$ is bounded, Lemma 3.1.15, $\iota x_i \to \iota\bar{x}$ and (3.4.1) imply $\iota y_i \to \iota\bar{x}$ and $\iota\tilde{y}_i \to \iota\bar{x}$ as $\mathscr{I} \ni i \to \infty$. Let $\varepsilon' > 0$ be arbitrary. Since p is approximately convex at $\iota\bar{x} \in \iota\mathscr{F}$, [24, Thm. 2] implies that there exists a $\delta' > 0$ such that for all $\mathring{y} \in \bar{B}_Y(\iota\bar{x},\delta')$ and $\mathring{g} \in \partial_C p(\mathring{y})$ it holds that

$$p(\mathring{y}) + \langle \mathring{g}, \mathring{x}-\mathring{y}\rangle_{Y^*,Y} - p(\mathring{x}) \le \varepsilon'\|\mathring{y}-\mathring{x}\|_Y \qquad \text{for all } \mathring{x} \in \bar{B}_Y(\mathring{y},\delta') \cap \bar{B}_Y(\iota\bar{x},\delta').$$

Now choose $i_{\varepsilon'} \in \mathbb{N}$ sufficiently large to ensure $\iota x_i \in \bar{B}_Y(\iota\bar{x},\frac{\delta'}{2})$ and $\iota\tilde{y}_i \in \bar{B}_Y(\iota\bar{x},\frac{\delta'}{2})$ for all $i \ge i_{\varepsilon'}$. This is possible since both sequences $(\iota x_i)_{i\in\mathbb{N}}$ and $(\iota\tilde{y}_i)_{i\in\mathbb{N}}$ converge to $\iota\bar{x}$ in Y. Denote by $g_i \in \partial_C p(\iota\tilde{y}_i)$ an exact subgradient such that $\|\tilde{g}_i - g_i\|_{Y^*} \le \mathring{\varepsilon}_{2,i}$. Setting $\mathring{x} = \iota x_i$, $\mathring{y} = \iota\tilde{y}_i$ and $\mathring{g} = \tilde{g}_i$ and using $f = p\circ\iota$, we find for arbitrary $i \in \mathscr{I}$ that

$$f(\tilde{y}_i) + \langle g_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - f(x_i) = p(\iota\tilde{y}_i) + \langle g_i, \iota x_i - \iota\tilde{y}_i\rangle_{Y^*,Y} - p(\iota x_i) \le \varepsilon'\|\iota(\tilde{y}_i-x_i)\|_Y.$$

Therefore, the linearization error e_{lin} can be bounded by

$$\begin{aligned} e_{\text{lin}} &= \tilde{f}_i + \langle \tilde{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}^x_i \\ &\le f(\tilde{y}_i) + \langle g_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - f(x_i) + \|\tilde{g}_i - g_i\|_{Y^*}\|\iota(\tilde{y}_i-x_i)\|_Y + \tilde{f}_i - f(\tilde{y}_i) - \tilde{f}^x_i + f(x_i) \\ &\le (\varepsilon' + \mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_{2,i})\|\iota(\tilde{y}_i-x_i)\|_Y. \end{aligned}$$

Because $[\cdot]_+$ is monotone, combining (3.4.1) and (3.4.9) with the last inequality gives

$$\tilde{f}_i - \phi_{i+1}(\tilde{y}_i) \le [e_{\text{lin}}]_+ + c\|\iota(\tilde{y}_i-x_i)\|^2_Y \le (\varepsilon' + \mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_{2,i} + c\|\iota(\tilde{y}_i-x_i)\|_Y)(M+1)\|\iota(y_i-x_i)\|_Y.$$

Consequently, (3.4.6)–(3.4.8) give for all $i \in \mathscr{I}$

$$\begin{aligned}
\|\hat{e}_i\|_{Y^*} &\leq \frac{\tau_i + \bar{q}}{\tau_i - \bar{q}} \frac{\|\iota(y_i - x_i)\|^2_{Q_i + \tau_i R_Y}}{\|\iota(y_i - x_i)\|_Y} \\
&\leq \frac{\tau_i + \bar{q}}{\tau_i - \bar{q}} \frac{1}{\theta(\tilde{\gamma} - \gamma)} \frac{\tilde{f}_i - \phi_{i+1}(\tilde{y}_i)}{\|\iota(y_i - x_i)\|_Y} \\
&\leq \frac{\tau_i + \bar{q}}{\tau_i - \bar{q}} \frac{M+1}{\theta(\tilde{\gamma} - \gamma)} (\varepsilon' + \mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_{2,i} + c\|\iota(\tilde{y}_i - x_i)\|_Y).
\end{aligned}$$

Therefore,

$$\bar{\varepsilon} = \lim_{i\to\infty} \|\hat{e}_i\|_{Y^*} \leq \lim_{i\to\infty} \frac{\tau_i + \bar{q}}{\tau_i - \bar{q}} \frac{M+1}{\theta(\tilde{\gamma} - \gamma)} (\varepsilon' + \mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_{2,i} + c\|\iota(\tilde{y}_i - x_i)\|_Y) = \frac{M+1}{\theta(\tilde{\gamma} - \gamma)} (\varepsilon' + \mathring{\varepsilon}_1 + \mathring{\varepsilon}_2).$$

As $\varepsilon' > 0$ was arbitrary, we further deduce $\eta = \inf \mathscr{E}_{\bar{x}} \leq \frac{M+1}{\theta(\tilde{\gamma}-\gamma)}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2)$. ∎

Remark 3.4.2. The approximate convexity of p is really only needed at the weak limit point $\bar{x}$. But since $\bar{x}$ is obviously not known before running the algorithm, we assume that p is approximately convex on the whole of $\iota\mathscr{F}$.

While the conditions (3.4.1)–(3.4.5) certainly lead to an implementable algorithm which guarantees convergence to η-$\partial_C p$-stationary points, conditions (3.4.1) and (3.4.2) might be difficult to fulfill. In both conditions the exact minimizer y_i of the bundle subproblem appears. Since y_i is not known exactly, one needs to use the property that y_i minimizes Ψ_i and use error estimates for this optimization problem to guarantee that (3.4.1) and (3.4.2) are fulfilled. This was carried out in [50, Chap. 6]. However, in our numerical experiments, this approach leads to premature refinement in order to fulfill (3.4.1) and (3.4.2). To avoid this difficulty, we propose new inexactness schemes (Theorems 3.4.9 and 3.4.11) which control the error $\Psi_i(\tilde{y}_i) - \Psi_i(y_i)$ directly. First we need some additional lemmas.

LEMMA 3.4.3. *For all $i \in \mathbb{N}$ it holds that $\|y_i - x_i\|_Y \leq \frac{\|\hat{e}_i\|_{Y^*}}{\tau_i - \bar{q}}$.*

Proof. By (3.1.17) and (3.1.18b), the bilinear form $\langle (Q_i + \tau_i R_Y)\cdot, \cdot\rangle_{Y^*,Y}$ is bounded and coercive with parameter $\tau_i - \bar{q}$, since it fulfills $\langle (Q_i + \tau_i R_Y)v, v\rangle_{Y^*,Y} = \|v\|^2_{Q_i+\tau_i R_Y} \geq (\tau_i - \bar{q})\|v\|^2_Y$. Consequently, the Lax-Milgram theorem implies that the operator $Q_i + \tau_i R_Y \in \mathscr{L}(Y, Y^*)$ is invertible and $\|(Q_i + \tau_i R_Y)^{-1}\|_{\mathscr{L}(Y^*,Y)} \leq 1/(\tau_i - \bar{q})$. Therefore, we conclude

$$\begin{aligned}
\|\iota(\tilde{y}_i - x_i)\|_Y &= \|(Q_i + \tau_i R_Y)^{-1}(Q_i + \tau_i R_Y)\iota(\tilde{y}_i - x_i)\|_Y \\
&\leq \|(Q_i + \tau_i R_Y)^{-1}\|_{\mathscr{L}(Y^*,Y)} \|(Q_i + \tau_i R_Y)\iota(y_i - x_i)\|_{Y^*} \\
&\leq \frac{\|\hat{e}_i\|_{Y^*}}{\tau_i - \bar{q}}.
\end{aligned}$$

∎

Assumption 3.4.4. For any subsequence of proximity iterations $\mathscr{I} \subset \mathbb{N}$ with $\tau_i \to \infty$ as $\mathscr{I} \ni i \to \infty$ there exists a further subsequence $\mathscr{I}' \subset \mathscr{I}$ and numbers $a, b \geq 0$ and $a_i, b_i, c_i, d_i \in \mathbb{R}$, $i \in \mathscr{I}'$, with $a_i \to a$, $b_i \to b$, $c_i \to 0$ and $d_i \to 0$ as $\mathscr{I}' \ni i \to \infty$ such that

$$\Psi_i(x_i) - \Psi_i(y_i) \leq \max\left\{a_i \tau_i^{-1},\ b_i\|y_i - x_i\|_Y\right\} + c_i \tau_i^{-1} + d_i\|y_i - x_i\|_Y \quad \text{for all } i \in \mathscr{I}'.$$

LEMMA 3.4.5. *If the set $\mathscr{B}_u$ is bounded and Assumption 3.4.4 holds true, then every weak limit point of the sequence of serious iterates is* $\max\{\sqrt{2a}, 2b\}$*-G-stationary in the sense of Definition 3.1.2.*

Proof. Let $\bar{x} \in X$ be an arbitrary weak limit point of the sequence of serious iterates. As the set of serious iterates is bounded, such a weak limit point exists. First consider the case that $\mathscr{E}_{\bar{x}} = \emptyset$, where $\mathscr{E}_{\bar{x}}$ is defined in (3.3.3), i.e.,

$$\mathscr{E}_{\bar{x}} = \big\{\bar{\varepsilon} \in [0,\infty] : \text{ there exists a subsequence of iterations } \mathscr{I} \text{ such that} \\ \rho_i < \gamma, \tilde{\rho}_i \geq \tilde{\gamma}, \tau_i \to \infty, x_i \rightharpoonup \bar{x}, \|\hat{e}_i\|_{Y^*} \to \bar{\varepsilon} \text{ as } \mathscr{I} \ni i \to \infty\big\}.$$

Then Theorem 3.3.3 implies that $\bar{x}$ is 0-G-stationarity which yields $\max\{\sqrt{2a}, 2b\}$-G-stationarity. Now consider the case that $\mathscr{E}_{\bar{x}} \neq \emptyset$. Then Theorem 3.3.3 yields $\inf \mathscr{E}_{\bar{x}}$-$G$-stationarity of $\bar{x}$. In the rest of the proof we show that $\inf \mathscr{E}_{\bar{x}} \leq \max\{\sqrt{2a}, 2b\}$. To do so, let $\bar{\varepsilon} \in \mathscr{E}_{\bar{x}}$ be arbitrary and let $\mathscr{I} \subset \mathbb{N}$ be a subsequence of proximity iterates such that $\tau_i \to \infty$ and $\|e_i\|_{Y^*} \to \bar{\varepsilon}$ as $\mathscr{I} \ni i \to \infty$. By definition of Q_i, cf. (3.1.17), it holds $\|Q_i\|_{\mathscr{L}(Y,Y^*)} \leq \bar{q}$ and thus $\|Q_i + \tau_i R_Y\|_{\mathscr{L}(Y,Y^*)} \leq \tau_i + \bar{q}$. Combining this with (3.1.18b) and Lemma 3.1.9 results in

$$\|\hat{e}_i\|_{Y^*}^2 \leq (\tau_i + \bar{q})^2 \|\iota(y_i - x_i)\|_Y^2 \leq \frac{(\tau_i + \bar{q})^2}{\tau_i - \bar{q}} \|\iota(y_i - x_i)\|_{Q_i + \tau_i R_Y}^2 \leq 2\frac{(\tau_i + \bar{q})^2}{\tau_i - \bar{q}} \big(\Psi_i(x_i) - \Psi_i(y_i)\big).$$

Switching to a further subsequence $\mathscr{I}'$ for which Assumption 3.4.4 holds, yields

$$\|\hat{e}_i\|_{Y^*}^2 \leq 2\max\left\{\frac{(\tau_i + \bar{q})^2}{\tau_i(\tau_i - \bar{q})} a_i, \frac{(\tau_i + \bar{q})^2}{\tau_i - \bar{q}} \|y_i - x_i\|_Y b_i\right\} + 2\frac{(\tau_i + \bar{q})^2}{\tau_i(\tau_i - \bar{q})} c_i + 2\frac{(\tau_i + \bar{q})^2}{\tau_i - \bar{q}} \|y_i - x_i\|_Y d_i,$$

and Lemma 3.4.3 gives

$$\|\hat{e}_i\|_{Y^*}^2 \leq 2\max\left\{\frac{(\tau_i + \bar{q})^2}{\tau_i(\tau_i - \bar{q})} a_i, \frac{(\tau_i + \bar{q})^2}{(\tau_i - \bar{q})^2} \|\hat{e}_i\|_{Y^*} b_i\right\} + 2\frac{(\tau_i + \bar{q})^2}{\tau_i(\tau_i - \bar{q})} c_i + 2\frac{(\tau_i + \bar{q})^2}{(\tau_i - \bar{q})^2} \|\hat{e}_i\|_{Y^*} d_i.$$

Since $\tau_i \to \infty$ as $\mathscr{I}' \ni i \to \infty$, taking the limit to infinity leads to

$$\bar{\varepsilon}^2 = \lim_{\substack{i\to\infty \\ i\in\mathscr{I}'}} \|\hat{e}_i\|_{Y^*}^2 \leq 2\max\left\{\lim_{\substack{i\to\infty \\ i\in\mathscr{I}'}} a_i,\ \lim_{\substack{i\to\infty \\ i\in\mathscr{I}'}} \|\hat{e}_i\|_{Y^*} \lim_{\substack{i\to\infty, \\ i\in\mathscr{I}'}} b_i\right\} + 2\lim_{\substack{i\to\infty, \\ i\in\mathscr{I}'}} c_i + 2\bar{\varepsilon} \lim_{\substack{i\to\infty, \\ i\in\mathscr{I}'}} d_i = 2\max\{a, \bar{\varepsilon} b\}.$$

In the case that $a \geq \bar{\varepsilon} b$, we get $\bar{\varepsilon}^2 \leq 2\max\{a, \bar{\varepsilon} b\} = 2a$, i.e., $\bar{\varepsilon} \leq \sqrt{2a}$. In the case that $a < \bar{\varepsilon} b$, we get $\bar{\varepsilon}^2 \leq 2\max\{a^2, \bar{\varepsilon} a\} = 2\bar{\varepsilon} b$, i.e., $\bar{\varepsilon} \leq 2b$. Therefore, $\inf \mathscr{E}_{\bar{x}} \leq \bar{\varepsilon} \leq \max\{\sqrt{2a}, 2b\}$. ∎

The last lemma shows that, if we can bound the quantity $\Psi_i(x_i) - \Psi_i(y_i)$ according to Assumption 3.4.4, then we immediately get a corresponding stationarity statement. If in each proximity iteration $i \in \mathbb{N}$ the cutting plane $m_{\tilde{y}_i, \tilde{f}_i, v, \tilde{g}}(\,\cdot\,, x_i)$ with subgradient base point $v \in X$ and subgradient $\tilde{g} \in Y^*$ is included into the next model ϕ_{i+1}, then $\Psi_i(x_i) - \Psi_i(y_i)$ can be bounded by the subproblem error $\Psi_i(\tilde{y}_i) - \Psi_i(y_i)$, the linearization error $[\tilde{f}_i + \langle \tilde{g}, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+$ and the downshift error $c\|\iota(v - x_i)\|_Y^2$.

LEMMA 3.4.6. *Let $v \in X$ and $\tilde{g} \in Y^*$ be arbitrary. For every proximity iteration with index $i \in \mathbb{N}$ with*

$\phi_{i+1}(\cdot) \geq m_{\tilde{y}_i,\tilde{f}_i,v,\tilde{g}}(\,\cdot\,,x_i)$*, it holds*

$$\Psi_i(x_i) - \Psi_i(y_i) \leq \Psi_i(\tilde{y}_i) - \Psi_i(y_i) + \tfrac{1}{\tilde{\gamma}-\gamma}[\tilde{f}_i + \langle \tilde{g}, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + \tfrac{c}{\tilde{\gamma}-\gamma}\|\iota(v - x_i)\|_Y^2.$$

Proof. Let $v \in X$ and $\tilde{g} \in Y^*$ be arbitrary and let $i \in \mathbb{N}$ be the index of a proximity iteration with $\phi_{i+1}(\cdot) \geq m_{\tilde{y}_i,\tilde{f}_i,v,\tilde{g}}(\,\cdot\,,x_i)$. Then it holds that $\rho_i < \gamma$ and $\tilde{\rho}_i \geq \tilde{\gamma}$, which gives $\tilde{\gamma} - \gamma < \tilde{\rho}_i - \rho_i$. Using the definitions of ρ_i and $\tilde{\rho}_i$ in Algorithm 3.1 we obtain

$$\tilde{J}_i^x - \Phi_i(\tilde{y}_i) < \tfrac{1}{\tilde{\gamma}-\gamma}\left(\tilde{J}_i - \Phi_{i+1}(\tilde{y}_i)\right).$$

Together with $\Psi_i(x_i) = \tilde{J}_i^x$ and $\Psi_i(\tilde{y}_i) \geq \Phi_i(\tilde{y}_i)$, we get

$$\Psi_i(x_i) - \Psi_i(y_i) \leq \tilde{J}_i^x - \Phi_i(\tilde{y}_i) + \Psi_i(\tilde{y}_i) - \Psi_i(y_i) \leq \tfrac{1}{\tilde{\gamma}-\gamma}\left(\tilde{J}_i - \Phi_{i+1}(\tilde{y}_i)\right) + \Psi_i(\tilde{y}_i) - \Psi_i(y_i).$$

By definition, the cutting plane $\hat{m}_i := m_{\tilde{y}_i,\tilde{f}_i,v,\tilde{g}}(\,\cdot\,,x_i)$ is given by

$$\hat{m}_i(\cdot) = \tilde{f}_i + \langle \hat{g}_i, \iota(\cdot - \tilde{y}_i)\rangle_{Y^*,Y} - [\tilde{f}_i + \langle \tilde{g}, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ - c\|\iota(v - x_i)\|_Y^2.$$

From $\Phi_{i+1}(\tilde{y}_i) = \phi_{i+1}(\tilde{y}_i) + w(\tilde{y}_i) \geq \hat{m}_i(\tilde{y}_i) + w(\tilde{y}_i)$, we deduce

$$\tilde{J}_i - \Phi_{i+1}(\tilde{y}_i) \leq \tilde{f}_i - \hat{m}_i(\tilde{y}_i) = [\tilde{f}_i + \langle \tilde{g}, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + c\|\iota(v - x_i)\|_Y^2.$$

Combining all this yields

$$\Psi_i(x_i) - \Psi_i(y_i) \leq \Psi_i(\tilde{y}_i) - \Psi_i(y_i) + \tfrac{1}{\tilde{\gamma}-\gamma}[\tilde{f}_i + \langle \tilde{g}, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + \tfrac{c}{\tilde{\gamma}-\gamma}\|\iota(v - x_i)\|_Y^2. \qquad \square$$

In order to estimate the downshift error $c\|\iota(v - x_i)\|_Y^2$, we employ the following results.

LEMMA 3.4.7. *Assume that there exists a subsequence of iterates $\mathscr{J} \subset \mathbb{N}$ such that $\iota(y_i - x_i) \to 0$ in Y as $\mathscr{J} \ni i \to \infty$. If $\mathscr{S}$ is bounded in X, then $\|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y} \to 0$ and $\|\iota(\tilde{y}_i - x_i)\|_{Q_i+\tau_i R_Y} \to 0$ as $\mathscr{I} \ni i \to \infty$.*

Proof. Recall that the exactness plane m_i^x is defined via

$$m_i^x(\cdot) := m_{x_i,\tilde{f}_i^x,x_i,\tilde{g}_i^x}(\cdot,x_i) = \tilde{f}_i^x + \langle \tilde{g}_i^x, \iota(\cdot - x_i)\rangle_{Y^*,Y}.$$

By Assumption 3.1.5, in any iteration i, the exactness plane is included into the model, i.e., $m_i^x \leq \phi_i$. Therefore,

$$\tilde{J}_i^x - \Phi_i(y_i) = \tilde{f}_i^x - \phi_i(y_i) \leq \tilde{f}_i^x - m_i^x(y_i) = \langle \tilde{g}_i^x, \iota(x_i - y_i)\rangle_{Y^*,Y}.$$

As $\iota(y_i - x_i) \to 0$ in Y and $(\tilde{g}_i^x)_{i\in\mathbb{N}} \subset G(\iota\mathscr{S})$ is bounded in Y^*, we find $\langle \tilde{g}_i^x, \iota(x_i - y_i)\rangle_{Y^*,Y} \to 0$. From $\Phi_i(y_i) + \frac{1}{2}\|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y}^2 = \Psi_i(y_i) \leq \Psi_i(x_i) = \tilde{J}_i^x$ we get

$$\tfrac{1}{2}\|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y}^2 \leq \tilde{J}_i^x - \Phi_i(y_i) \leq \langle \tilde{g}_i^x, \iota(x_i - y_i)\rangle_{Y^*,Y} \to 0,$$

i.e., $\|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y} \to 0$. From Lemma 3.1.9 and (3.1.7) we infer

$$\tfrac{1}{2}\|\iota(\tilde{y}_i - y_i)\|^2_{Q_i+\tau_i R_Y} \le \Psi_i(\tilde{y}_i) - \Psi_i(y_i) \to 0,$$

which yields $\|\iota(\tilde{y}_i - x_i)\|_{Q_i+\tau_i R_Y} \le \|\iota(\tilde{y}_i - y_i)\|_{Q_i+\tau_i R_Y} + \|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y} \to 0$. ∎

COROLLARY 3.4.8. *Assume that there exists a subsequence of iterates $\mathscr{J} \subset \mathbb{N}$ such that $\tau_i \to \infty$ as $\mathscr{J} \ni i \to \infty$. If $\mathscr{S}$ is bounded in X, then $\tau_i\|\iota(y_i - x_i)\|^2_Y \to 0$ and $\tau_i\|\iota(\tilde{y}_i - x_i)\|^2_Y \to 0$ as $\mathscr{I} \ni i \to \infty$.*

Proof. By Lemma 3.1.15 it holds $\iota(y_i - x_i) \to 0$ in Y as $\mathscr{J} \ni i \to \infty$. Thus, Lemma 3.4.7 yields that $\|\iota(y_i - x_i)\|_{Q_i+\tau_i R_Y} \to 0$ and $\|\iota(\tilde{y}_i - x_i)\|_{Q_i+\tau_i R_Y} \to 0$ as $\mathscr{I} \ni i \to \infty$. For all $i \in \mathscr{I}$ sufficiently large, it holds $\tau_i \ge 2\bar{q}$. Consequently, (3.1.18b) gives

$$\begin{aligned}
&\tfrac{1}{2}\tau_i\|\iota(y_i - x_i)\|^2_Y \le (\tau_i - \bar{q})\|\iota(y_i - x_i)\|^2_Y \le \|\iota(y_i - x_i)\|^2_{Q_i+\tau_i R_Y} \to 0 \qquad \text{as } \mathscr{I} \ni i \to \infty,\\
&\tfrac{1}{2}\tau_i\|\iota(\tilde{y}_i - x_i)\|^2_Y \le (\tau_i - \bar{q})\|\iota(\tilde{y}_i - x_i)\|^2_Y \le \|\iota(\tilde{y}_i - x_i)\|^2_{Q_i+\tau_i R_Y} \to 0 \qquad \text{as } \mathscr{I} \ni i \to \infty.
\end{aligned}$$ ∎

Now we are able to prove the convergence statement for general Lipschitz continuous functions.

THEOREM 3.4.9. *Assume that the initial point $x_0 \in \mathscr{F}$ is such that the level set $\mathscr{F}_0 := \{x \in \mathscr{F} : J(x) \le J(x_0) + 2\Delta\}$ is bounded in X. Let $M_v \ge 0$, $\alpha \ge 0$ and $\beta \ge 0$ be constants and let $(\alpha_i)_{i\in\mathbb{N}} \subset [0,\infty)$ and $(\beta_i)_{i\in\mathbb{N}} \subset [0,\infty)$ be forcing sequences with $\alpha_i \to \alpha$ and $\beta_i \to \beta$ as $i \to \infty$. Assume that in every iteration of the bundle method a trial iterate can be computed which fulfills*

$$\Psi_i(\tilde{y}_i) - \Psi_i(y_i) \le \alpha_i \tau_i^{-1}, \tag{3.4.10}$$

and that in every unsuccessful iteration a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \hat{g}_i) \in \mathscr{B}_a$ of bundle information can be computed which fulfills the conditions

$$\|\iota(v_i - \tilde{y}_i)\|_Y \le M_v\|\iota(\tilde{y}_i - x_i)\|_Y, \tag{3.4.11}$$

$$[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ \le \beta_i \tau_i^{-1}. \tag{3.4.12}$$

Assume that the cutting plane $m_{\tilde{y}_i,\tilde{f}_i,v_i,\hat{g}_i}(\,\cdot\,, x_i)$ is included into the new model ϕ_{i+1} whenever iteration i was not successful and assume that the set of subgradient base points $\mathscr{V}$ is bounded. Then every weak limit point of the sequence of serious iterates is η-G-stationary in the sense of Definition 3.1.2 with $\eta := \sqrt{2}\sqrt{\alpha + \beta/(\tilde{\gamma} - \gamma)}$.

Proof. First we show that the set of used bundle information, $\mathscr{B}_u$, is bounded. Since the level set $\mathscr{F}_0$ is bounded in X, Lemma 3.1.12 implies that the set of serious iterates $\mathscr{S}$ is bounded in X. Thus, Lemma 3.1.14 shows that the set of bundle information $\mathscr{B}_u$ is bounded in $X \times \mathbb{R} \times X \times Y^*$. In order to apply Lemma 3.4.5, we need to check if Assumption 3.4.4 is fulfilled. To do so, let $\mathscr{I} \subset \mathbb{N}$ be an arbitrary subsequence of proximity iterations with $\tau_i \to \infty$ as $\mathscr{I} \ni i \to \infty$. Since the cutting plane $m_{\tilde{y}_i,\tilde{f}_i,\hat{v}_i,\hat{g}_i}(\,\cdot\,, x_i)$ is included into the new model ϕ_{i+1}, Lemma 3.4.6 implies that

$$\Psi_i(x_i) - \Psi_i(y_i) \le \Psi_i(\tilde{y}_i) - \Psi_i(y_i) + \tfrac{1}{\tilde{\gamma}-\gamma}[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + \tfrac{c}{\tilde{\gamma}-\gamma}\|\iota(\hat{v}_i - x_i)\|^2_Y.$$

Using (3.4.11), we estimate

$$\|\iota(\hat{v}_i - x_i)\|_Y \le \|\iota(\hat{v}_i - \tilde{y}_i)\|_Y + \|\iota(\tilde{y}_i - x_i)\|_Y \le (M_v + 1)\|\iota(\tilde{y}_i - x_i)\|_Y.$$

Combining this with (3.4.10) and (3.4.12) yields

$$\Psi_i(x_i) - \Psi_i(y_i) \le \Psi_i(\tilde{y}_i) - \Psi_i(y_i) + \frac{1}{\tilde{\gamma}-\gamma}[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + \frac{c(M_v+1)^2}{\tilde{\gamma}-\gamma}\|\iota(\tilde{y}_i - x_i)\|_Y^2$$
$$\le \left(\alpha_i + \frac{\beta_i}{\tilde{\gamma}-\gamma} + \frac{c(M_v+1)^2}{\tilde{\gamma}-\gamma}\tau_i\|\iota(\tilde{y}_i - x_i)\|_Y^2\right)\tau_i^{-1}.$$

Consequently, Assumption 3.4.4 is fulfilled with

$$a_i := \alpha_i + \frac{\beta_i}{\tilde{\gamma}-\gamma} + \frac{c(M_v+1)^2}{\tilde{\gamma}-\gamma}\tau_i\|\iota(\tilde{y}_i - x_i)\|_Y^2, \quad b_i := 0, \quad c_i := 0 \quad \text{and} \quad d_i := 0.$$

Corollary 3.4.8 implies that $\tau_i\|\iota(\tilde{y}_i - x_i)\|_Y^2 \to 0$ as $\mathscr{I} \ni i \to \infty$. Therefore, the sequence $(a_i)_{i\in\mathscr{I}}$ converges to $a := \alpha + \beta/(\tilde{\gamma}-\gamma)$ as $\mathscr{I} \ni i \to \infty$ and Lemma 3.4.5 implies that every weak limit point of the sequence of serious iterates is $\sqrt{2a}$-G-stationary in the sense of Definition 3.1.2. □

In the previous theorem it is assumed that in every unsuccessful iteration a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \hat{g}_i) \in \mathscr{B}_a$ of bundle information can be computed which fulfills (3.4.11) and (3.4.12). While it is easy to check if each condition is satisfied, in general it is not clear how to find a subgradient $\hat{g}_i$ which fulfills the linearization condition (3.4.12). If the objective function p has additional regularity properties, we can guarantee that any choice of $\hat{g}_i \in G(\iota\tilde{y}_i)$ leads to convergence of the bundle method. We generalize the concept of ε-convexity, cf. Section 2.2.3.

DEFINITION 3.4.10. *A function $p : Y \to \mathbb{R}\cup\infty$ is called ε-G-convex at $\bar{v} \in Y$ if there exists a $\delta > 0$ such that for all $v \in \bar{B}_Y(\bar{v},\delta)$ and $s \in \bar{B}_Y(0,\delta)$ which fulfill $v+s \in \bar{B}_Y(\bar{v},\delta)$ it holds*

$$p(v) + \langle g, s\rangle_{Y^*,Y} - p(v+s) \le \varepsilon\|s\|_Y \qquad \forall g \in G(v).$$

Note that every ε-convex function in the sense of Definition 2.2.9 is ε-∂_C-convex (cf. Lemma 2.2.10). Furthermore, by Theorem 2.2.12, a locally Lipschitz function $f : Y \to \mathbb{R}\cup\infty$ is approximately convex at $\bar{v} \in Y$ (cf. Definition 2.2.11) if and only if f is ε-∂_C-convex at $\bar{v}$ for all $\varepsilon > 0$.

THEOREM 3.4.11. *Assume that the initial point $x_0 \in \mathscr{F}$ is such that the level set $\mathscr{F}_0 := \{x \in \mathscr{F} : J(x) \le J(x_0) + 2\Delta\}$ is bounded in X. Let $\mathring{\varepsilon}_1 \ge 0$ be a constant and let $(\alpha_i)_{i\in\mathbb{N}} \subset [0,\infty)$ and $(\mathring{\varepsilon}_{1,i})_{i\in\mathbb{N}} \subset [0,\infty)$ be forcing sequences with $\alpha_i \to 0$ and $\mathring{\varepsilon}_{1,i} \to \mathring{\varepsilon}_1$ as $i \to \infty$. Assume that in every iteration of the bundle method a trial iterate can be computed which fulfills*

$$\Psi_i(\tilde{y}_i) - \Psi_i(y_i) \le \alpha_i\tau_i^{-1}, \tag{3.4.13}$$

and that in every unsuccessful iteration a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \hat{g}_i) \in \mathscr{B}_a$ of bundle information can be computed which fulfills the condition

$$[\tilde{f}_i - f(\tilde{y}_i) + f(x_i) - \tilde{f}_i^x]_+ \le \mathring{\varepsilon}_{1,i}\|\iota(\tilde{y}_i - x_i)\|_Y. \tag{3.4.14}$$

Assume that the cutting plane $m_{\tilde{y}_i,\tilde{f}_i,\tilde{y}_i,\hat{g}_i}(\cdot, x_i)$ with $\hat{g}_i \in G(\iota\tilde{y}_i)$ is included into the new model ϕ_{i+1} whenever iteration i is not successful and assume the set of all subgradient base points $\mathscr{V}$ is bounded. If p is $\mathring{\varepsilon}_2$-G-convex on $\iota\mathscr{F}$, $\mathring{\varepsilon}_2 \ge 0$, then every weak limit point of the sequence of serious iterates is η-G-stationary in the sense of Definition 3.1.2 with $\eta := \frac{2}{\tilde{\gamma}-\gamma}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2)$.

Proof. First we show that the set of used bundle information, $\mathscr{B}_u$, is bounded. Since the level set $\mathscr{F}_0$ is bounded in X, Lemma 3.1.12 implies that the set of serious iterates $\mathscr{S}$ is bounded in X. Therefore, Lemma 3.1.14 shows that the set of bundle information $\mathscr{B}_u$ is bounded in $X \times \mathbb{R} \times X \times Y^*$.
In order to apply Lemma 3.4.5, we need to check if Assumption 3.4.4 is fulfilled. To do so, let $\mathscr{I} \subset \mathbb{N}$ be an arbitrary subsequence of proximity iterations with $\tau_i \to \infty$ as $\mathscr{I} \ni i \to \infty$. Let $\bar{x} \in X$ be a weak accumulation point of the bounded sequence $(x_i)_{i\in\mathscr{I}}$ and choose a subsequence $\mathscr{I}' \subset \mathscr{I}$ such that $x_i \rightharpoonup \bar{x}$ as $\mathscr{I}' \ni i \to \infty$. Since p is $\mathring{\varepsilon}_2$-G-convex at $\iota\bar{x} \in \iota\mathscr{F}$, there exists a $\delta > 0$ such that for all $x, \tilde{y} \in \bar{B}_Y(\iota\bar{x}, \delta)$ and all $g \in G(\iota\tilde{y})$ it holds

$$f(\tilde{y}) + \langle g, \iota(x-\tilde{y})\rangle_{Y^*,Y} - f(x) = p(\iota\tilde{y}) + \langle g, \iota x - \iota\tilde{y}\rangle_{Y^*,Y} - p(\iota x) \le \mathring{\varepsilon}_2 \|\iota(\tilde{y}-x)\|_Y.$$

By $x_i \rightharpoonup \bar{x}$ as $\mathscr{I}' \ni i \to \infty$ and since $\iota \in \mathscr{L}(X,Y)$ is a compact operator, Lemma 3.1.15 implies that $\iota x_i \to \iota\bar{x}$ and $\iota\tilde{y}_i \to \iota\bar{x}$ in Y as $\mathscr{I}' \ni i \to \infty$. Choose a subsequence $\mathscr{I}'' \subset \mathscr{I}'$ such that $\iota x_i, \iota\tilde{y}_i \in \bar{B}_Y(\iota\bar{x}, \delta)$ for all $i \in \mathscr{I}''$. From $\hat{g}_i \in G(\iota\tilde{y}_i)$ and the $\mathring{\varepsilon}_2$-G-convexity of p at $\iota\bar{x} \in \iota\mathscr{F}$, we obtain

$$f(\tilde{y}_i) + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - f(x_i) \le \mathring{\varepsilon}_2 \|\iota(\tilde{y}_i - x_i)\|_Y \qquad \text{for all } i \in \mathscr{I}''.$$

Because $[\cdot]_+$ is monotone and $[a+b]_+ \le [a]_+ + [b]_+$ for all $a, b \in \mathbb{R}$, we infer using (3.4.14) that

$$\begin{aligned}
[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ &\le [f(\tilde{y}_i) + \langle \hat{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - f(x_i)]_+ + [\tilde{f}_i - f(\tilde{y}_i) + f(x_i) - \tilde{f}_i^x]_+ \\
&\le (\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)\|\iota(\tilde{y}_i - x_i)\|_Y.
\end{aligned}$$

By (3.1.18b) and Lemma 3.1.9 we get

$$\tfrac{1}{2}(\tau_i - \bar{q})\|\tilde{y}_i - y_i\|_Y^2 \le \tfrac{1}{2}\|\tilde{y}_i - y_i\|^2_{Q_i + \tau_i R_Y} \le \Psi_i(\tilde{y}_i) - \Psi_i(y_i) \le \alpha_i \tau_i^{-1},$$

and abbreviating $c_i' := (\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)(2\alpha_i\tau_i)^{1/2}(\tau_i - \bar{q})^{-1/2}$ yields

$$(\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)\|\iota(\tilde{y}_i - y_i)\|_Y \le (\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)\sqrt{\frac{2\alpha_i}{\tau_i(\tau_i - \bar{q})}} = (\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)\sqrt{\frac{2\alpha_i\tau_i}{\tau_i - \bar{q}}}\,\tau_i^{-1} = c_i'\tau_i^{-1}.$$

We note that $c_i' \to 0$ as $\mathscr{I}'' \ni i \to \infty$. Combing the above results gives

$$\begin{aligned}
[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ &\le (\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)(\|\iota(y_i - x_i)\|_Y + \|\iota(\tilde{y}_i - y_i)\|_Y) \\
&\le (\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2)\|\iota(y_i - x_i)\|_Y + c_i'\tau_i^{-1}.
\end{aligned}$$

Since the cutting plane $m_{\tilde{y}_i,\tilde{f}_i,\tilde{y}_i,\hat{g}_i}(\,\cdot\,, x_i)$ is included into the new model ϕ_{i+1}, Lemma 3.4.6 implies

$$\Psi_i(x_i) - \Psi_i(y_i) \le \Psi_i(\tilde{y}_i) - \Psi_i(y_i) + \tfrac{1}{\tilde{\gamma}-\gamma}[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + \tfrac{c}{\tilde{\gamma}-\gamma}\|\iota(\tilde{y}_i - x_i)\|_Y^2$$

for all $i \in \mathscr{I}''$. Combining this with (3.4.13) yields

$$\Psi_i(x_i) - \Psi_i(y_i) \le \frac{\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2}{\tilde{\gamma} - \gamma}\|\iota(y_i - x_i)\|_Y + \left(\alpha_i + \frac{c_i'}{\tilde{\gamma}-\gamma} + \frac{c}{\tilde{\gamma}-\gamma}\tau_i\|\iota(\tilde{y}_i - x_i)\|_Y^2\right)\tau_i^{-1}.$$

Corollary 3.4.8 implies that $\tau_i\|\iota(\tilde{y}_i - x_i)\|_Y^2 \to 0$ as $\mathscr{I}' \ni i \to \infty$. Consequently, Assumption 3.4.4 is fulfilled with the subsequence $\mathscr{I}''$,

$$a_i := 0, \quad b_i := \frac{\mathring{\varepsilon}_{1,i} + \mathring{\varepsilon}_2}{\tilde{\gamma} - \gamma}, \quad c_i := \alpha_i + \frac{c_i'}{\tilde{\gamma} - \gamma} + \frac{c}{\tilde{\gamma} - \gamma} \tau_i \|\iota(\tilde{y}_i - x_i)\|_Y^2 \quad \text{and} \quad d_i := 0.$$

As the sequence $(b_i)_{i \in \mathscr{I}''}$ converges to $b := (\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2)/(\tilde{\gamma} - \gamma)$ as $\mathscr{I}'' \ni i \to \infty$, Lemma 3.4.5 implies that every weak limit point of the sequence of serious iterates is $2b$-G-stationary in the sense of Definition 3.1.2. ∎

Applying Theorem 3.4.11 to the setting of Lemma 3.4.1 yields the following result.

COROLLARY 3.4.12. *Assume that the initial point $x_0 \in \mathscr{F}$ is such that the level set $\mathscr{F}_0 := \{x \in \mathscr{F} : J(x) \le J(x_0) + 2\Delta\}$ is bounded in X. Let the multifunction $G : Y \rightrightarrows Y^*$ be defined via $G := \partial_C p + \bar{B}_{Y^*}(0, \Delta_2)$, where $\Delta_2 \ge 0$. Assume that in every iteration of the bundle method a trial iterate can be computed which fulfills*

$$\Psi_i(\tilde{y}_i) - \Psi_i(y_i) \le \alpha_i \tau_i^{-1}, \qquad \alpha_i \to 0 \qquad (3.4.15)$$

and that in every unsuccessful iteration a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \hat{g}_i) \in \mathscr{B}_a$ of bundle information can be computed which fulfills the conditions

$$[\tilde{f}_i - f(\tilde{y}_i) + f(x_i) - \tilde{f}_i^x]_+ \le \mathring{\varepsilon}_{1,i} \|\iota(\tilde{y}_i - x_i)\|_Y, \qquad \mathring{\varepsilon}_{1,i} \to \mathring{\varepsilon}_1,\ \mathring{\varepsilon}_1 \ge 0, \qquad (3.4.16)$$

$$\operatorname{dist}(\hat{g}_i, \partial_C p(\iota \tilde{y}_i)) \le \mathring{\varepsilon}_{2,i}, \qquad \mathring{\varepsilon}_{2,i} \to \mathring{\varepsilon}_2,\ \mathring{\varepsilon}_2 \ge 0. \qquad (3.4.17)$$

Assume that the cutting plane $m_{\tilde{y}_i, \tilde{f}_i, \tilde{y}_i, \tilde{g}_i}(\cdot, x_i)$ is included into the new model ϕ_{i+1} whenever iteration i is not successful and assume that the set of subgradient base points $\mathscr{V}$ is bounded. If p is approximately convex on $\iota\mathscr{F}$, then every weak limit point of the sequence of serious iterates is η-$\partial_C p$-stationary in the sense of Definition 3.1.2 with $\eta := \frac{2}{\tilde{\gamma} - \gamma}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2) + \Delta_2$.

Proof. Assume that p is approximately convex on $\iota\mathscr{F}$, let $\bar{x}$ be any weak limit point of the sequence of serious iterates and let $\varepsilon > 0$ be arbitrary. Then, p is ε-∂_C-convex on $\iota\mathscr{F}$. Consequently, p is $(\varepsilon + \Delta_2)$-G-convex on $\iota\mathscr{F}$ and Theorem 3.4.11 shows that $\bar{x}$ is $\frac{2}{\tilde{\gamma} - \gamma}(\mathring{\varepsilon}_1 + \Delta_2 + \varepsilon)$-$G$-stationary. However, using (3.4.17) and that p is ε-∂_C-convex on $\iota\mathscr{F}$, a slight modification of Theorem 3.4.11 shows that $\bar{x}$ is $\frac{2}{\tilde{\gamma} - \gamma}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2 + \varepsilon)$-$G$-stationary, i.e.,

$$\begin{aligned} 0 \ &\in \ w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^*(G(\iota\bar{x}) + \bar{B}_{Y^*}(0, \tfrac{2}{\tilde{\gamma} - \gamma}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2 + \varepsilon))) \\ &= w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^*(\partial_C p(\iota\bar{x}) + \bar{B}_{Y^*}(0, \tfrac{2}{\tilde{\gamma} - \gamma}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2 + \varepsilon) + \Delta_2)). \end{aligned}$$

Since this holds for all $\varepsilon > 0$, the norm $\|\cdot\|_{Y^*}$ is weakly lower semicontinuous, and $w'(\bar{x}) + N_{\mathscr{F}}(\bar{x}) + \iota^* \partial_C p(\iota\bar{x})$ is a closed set, we find that $\bar{x}$ is η-∂_C-stationary with $\eta = \frac{2}{\tilde{\gamma} - \gamma}(\mathring{\varepsilon}_1 + \mathring{\varepsilon}_2) + \Delta_2$ (cf. the proof of Lemma 3.3.2 for details). ∎

Comparing Corollary 3.4.12 to Lemma 3.4.1 shows that the conditions (3.4.1) and (3.4.2) were replaced by (3.4.15). Using error estimates for the solution of the obstacle problem (cf. Chapter 4), the condition (3.4.15) can be enforced straightforwardly. In this sense, Corollary 3.4.12 is an improvement over Lemma 3.4.1 and [50, Thm. 5.5].

3.5. Practical implementation

The bundle method Algorithm 3.1 is quite abstract and thus provides a lot of flexibility, for example in the choice of the cutting plane model, the function value approximation, the trial iterate and the proximity parameter. In this section, we aim at developing a concrete version of the bundle method which can directly be implemented. This version of the bundle method should work both for approximately convex objective functions and, if a sufficiently steep subgradient-based linearization is provided, for general locally Lipschitz objective functions. Therefore, we need to construct function value approximations and subgradients which fulfill both the error bounds in Theorem 3.4.9 and Corollary 3.4.12. In order to do so, we assume that there are three computable oracles available, namely a function value oracle, a subgradient oracle and a trial iterate oracle.

3.5.1. The function value oracle

This section regards the question on how to choose the function value approximation $\tilde{f}_i$ which can be used to find a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \tilde{g}_i) \in \mathscr{B}_a$ of bundle information which fulfills (3.4.12) and (3.4.16). Assume that we have access to a computable oracle $O_f : X \times (0,1] \to \mathbb{R}$ and a computable error bound $\hat{\varepsilon}_f : X \times (0,1] \to (0,\infty)$ such that

$$|f(x) - O_f(x,h)| \leq \hat{\varepsilon}_f(x,h) \qquad \text{and} \qquad \hat{\varepsilon}_f(x,h) \to 0 \text{ as } h \to 0. \tag{3.5.1}$$

If one simply chooses $\tilde{f}_i := O_f(\tilde{y}_i, h_i)$ for a given accuracy level h_i, then one might run into trouble. It might happen that $\tilde{f}_i^x = O_f(x_i, h_i^x) = f(x_i) - \hat{\varepsilon}_f(x_i, h_i^x)$ while $\tilde{f}_i = O_f(\tilde{y}_i, h_i) = f(\tilde{y}_i) + \hat{\varepsilon}_f(\tilde{y}_i, h_i)$. In this case, the exactness condition (3.4.12),

$$[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ \leq \beta_i \tau_i^{-1} \tag{3.4.12}$$

might not be satisfiable, even if $f(\tilde{y}_i) < f(x_i)$, $\hat{g}_i = 0$ and $\hat{\varepsilon}_f(\tilde{y}_i, h_i) = 0$. Indeed, assume that $\tilde{f}_i = f(\tilde{y}_i)$, $\tilde{f}_i^x = f(x_i) - \hat{\varepsilon}_f(x_i, h_i^x)$, $\hat{g}_i = 0$, $\hat{\varepsilon}_f(x_i, h_i^x) > \beta_i \tau_i^{-1}$ and $0 \leq f(x_i) - f(\tilde{y}_i) < \hat{\varepsilon}_f(x_i, h_i^x) - \beta_i \tau_i^{-1}$. In this case, it holds

$$[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ = f(\tilde{y}_i) - f(x_i) + \hat{\varepsilon}_f(x_i, h_i^x) > \beta_i \tau_i^{-1}.$$

To circumvent this issue, we aim at choosing $\tilde{f}_i$ such that $\tilde{f}_i \geq f(\tilde{y}_i)$ for all $i \in \mathbb{N}$. To do so, we choose an appropriate accuracy level $h_i \in (0,\infty)$, a lift term $l_i \geq \hat{\varepsilon}_f(\tilde{y}_i, h_i)$ and set

$$\tilde{f}_i := O_f(\tilde{y}_i, h_i) + l_i.$$

This implies that $\tilde{f}_i$ is an upper approximation of the exact function value $f(\tilde{y}_i)$, i.e.,

$$f(\tilde{y}_i) \leq |f(\tilde{y}_i) - O_f(\tilde{y}_i, h_i)| + O_f(\tilde{y}_i, h_i) \leq O_f(\tilde{y}_i, h_i) + l_i = \tilde{f}_i \qquad \text{for all } i \in \mathbb{N}. \tag{3.5.2}$$

In particular, this gives $f(x_i) \leq \tilde{f}_i^x$ for all $i \in \mathbb{N}$ and

$$\tilde{f}_i - f(\tilde{y}_i) - \tilde{f}_i^x + f(x_i) \leq |O_f(\tilde{y}_i, h_i) + l_i - f(\tilde{y}_i)| \leq l_i + \hat{\varepsilon}_f(\tilde{y}_i, h_i) \leq 2l_i \qquad \text{for all } i \in \mathbb{N}. \tag{3.5.3}$$

Therefore, if the linearization error $[f(\tilde{y}_i)+\langle \hat{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - f(x_i)]_+$ and the lift term l_i are sufficiently small, then the estimate

$$[\tilde{f}_i+\langle \hat{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ \leq [f(\tilde{y}_i)+\langle \hat{g}_i, \iota(x_i-\tilde{y}_i)\rangle_{Y^*,Y} - f(x_i)]_+ + [\tilde{f}_i - f(\tilde{y}_i) - \tilde{f}_i^x + f(x_i)]_+$$

shows that the exactness condition (3.4.12) is fulfilled. Also, if $2l_i \leq \mathring{\varepsilon}_{1,i}\|\iota(\tilde{y}_i - x_i)\|_Y$, then (3.4.16) is fulfilled. In many applications, the term $\|\iota(\tilde{y}_i - x_i)\|_Y$ cannot be computed exactly, e.g., if $X = L^2(\Omega)$ and $Y = H^{-1}(\Omega)$, cf. Chapter 5. We thus assume that we have access to a computable oracle $\underline{O}_{\|\cdot\|_Y}$: $X \setminus \{0\} \times (0,1] \to (0,\infty)$ and that there exists a (possibly unknown) constant $C_{\|\cdot\|_Y} > 0$ such that for a point $x \in X \setminus \{0\}$ and an accuracy level h it holds

$$0 < C_{\|\cdot\|_Y}\underline{O}_{\|\cdot\|_Y}(x,h) \leq \|\iota x\|_Y. \tag{3.5.4}$$

Now, if we require that $l_i \leq \mathring{\varepsilon}_{1,i}\underline{O}_{\|\cdot\|_Y}(\tilde{y}_i - x_i)$, then we can guarantee

$$[\tilde{f}_i - f(\tilde{y}_i) - \tilde{f}_i^x + f(x_i)]_+ \leq 2l_i \leq 2\mathring{\varepsilon}_{1,i}\underline{O}_{\|\cdot\|_Y}(\tilde{y}_i - x_i) \leq 2C_{\|\cdot\|_Y}^{-1}\mathring{\varepsilon}_{1,i}\|\iota x\|_Y \qquad \text{for all } i \in \mathbb{N}.$$

Next, we address the question on how to choose the lift term l_i. Simply using $l_i := \hat{\varepsilon}_f(\tilde{y}_i, h_i)$ yields a new issue. Assume that the error bound $\hat{\varepsilon}_f(x,h)$ depends on the point x, i.e., that the function $\hat{\varepsilon}_f(\cdot, h_i)$ is not constant. Then the approximate function values $\tilde{f}_i = O_f(\tilde{y}_i, h_i) + l_i = O_f(\tilde{y}_i, h_i) + \hat{\varepsilon}_f(\tilde{y}_i, h_i)$ can be interpreted as approximations of the function value of the function $f(\cdot) + \hat{\varepsilon}_f(\cdot, h_i)$ at the point $\tilde{y}_i$. Furthermore, the subgradients $\tilde{g}_i \in G(\iota\tilde{y}_i)$ are an approximation of the subgradients of f. Thus, the algorithm works with approximate function values of the function $f(\cdot)+\hat{\varepsilon}_f(\cdot, h_i)$ and with approximate subgradients of the function f. While the choice $\tilde{f}_i = O_f(\tilde{y}_i, h_i) + \hat{\varepsilon}_f(\tilde{y}_i, h_i)$ is covered by the convergence theory, our numerical results suggest that this leads to a slow progress of the algorithm. Instead of choosing $l_i = \hat{\varepsilon}_f(\tilde{y}_i, h_i)$, we use the same lift term l_i for all trial iterates with $\hat{\varepsilon}_f(\tilde{y}_i, h_i) \leq l_i$. If a trial iterate $\tilde{y}_i$ is encountered with $\hat{\varepsilon}_f(\tilde{y}_i, h_i) > l_i$, we decrease h_i until $\hat{\varepsilon}_f(\tilde{y}_i, h_i) \leq l_i$ is fulfilled. Whenever a better approximation of the function value is needed, the lift term l_i is reduced (which eventually leads to a decrease of h_i to fulfill $\hat{\varepsilon}_f(\tilde{y}_i, h_i) \leq l_i$). With this strategy, the approximate function values $\tilde{f}_i$ are approximations of $f(\cdot) + l_i$ with l_i constant. As the functions $f(\cdot)+l_i$ and $f(\cdot)$ have the same subgradients, the function value approximation $\tilde{f}_i = O_f(\tilde{y}_i, h_i) + l_i$ and the approximate subgradients $\tilde{g}_i \in G(\iota\cdot)$ fit together nicely. Our numerical experiments suggest that, while both approaches $\tilde{f}_i := O_f(\tilde{y}_i, h_i) + \hat{\varepsilon}_f(\tilde{y}_i, h_i)$ and $\tilde{f}_i := O_f(\tilde{y}_i, h_i) + l_i$ converge, the latter approach works better. In Algorithm 3.2 we present the discussed strategy to compute $\tilde{f}_i := O_f(\tilde{y}_i, h_i) + l_i$.

Algorithm 3.2: Function value approximation $[(\tilde{f}, \hat{h}) :=$`FunctionValue`$(\tilde{y}, h, l)]$

Parameters : Function value oracle $O_f : X \times (0,1] \to Y^*$ with error bound $\hat{\varepsilon}_f : X \times (0,1] \to (0,\infty)$
Input : Trial iterate $\tilde{y}$, current accuracy level h and desired lift term l.
Output : Function value approximation $\tilde{f}$ and new accuracy level $\hat{h}$.

while $\hat{\varepsilon}_f(\tilde{y}, h) > l$ **do**
  Set $h := h/2$.
end
Return $\tilde{f} := O_f(\tilde{y}, h) + l$ and $\hat{h} := h$.

3.5.2. The subgradient oracle

The goal of this section is to construct a way of computing a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \tilde{g}_i) \in \mathscr{B}_a$ of bundle information which fulfills (3.4.12). In many practical applications, one only has access to a subgradient oracle $O_g : X \times (0,1] \to Y^*$ such that for any subgradient base point $v \in X$ and accuracy level h a subgradient $\tilde{g}_i = O_g(v,h)$ can be computed which fulfills $\tilde{g}_i \in G(\iota v_i)$. However, for general locally Lipschitz functions, it is not sufficient to find a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \tilde{g}_i) \in \mathscr{B}_a$ which fulfills (3.4.12). Indeed, consider the one dimensional case $X = Y := \mathbb{R}$ with exact function value approximations $\tilde{f}_i := f(\tilde{y}_i)$ and $\tilde{f}_i^x := f(x_i)$, $\tilde{y}_i := 0$, $x_i := 1$, $\beta_i \tau_i^{-1} < 1$ and p equals to the lightning function $L : \mathbb{R} \to \mathbb{R}$ (see, e.g., [73, Ex. BE.0]). The lightning function L is Lipschitz with constant 1, fulfills $L(0) = L(1) = 0$ and enjoys the property $\partial_C L(t) = [-1,1]$ for all $t \in \mathbb{R}$. In this case, for any $t \in \mathbb{R}$, the oracle might produce the subgradient $\hat{g}_i = 1 \in \partial_C L(t)$ which leads to

$$[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ = L(0) + \langle 1, \iota(1-0)\rangle_{\mathbb{R}^*,\mathbb{R}} - L(1) = 1 > \beta_i \tau_i^{-1}. \tag{3.5.5}$$

This shows that condition (3.4.12) cannot be fulfilled for any choice of subgradient base point $t \in \mathbb{R}$. However, in this example, if the subgradient oracle returns a subgradient in the interval $[-1, \beta_i \tau_i^{-1}]$, then every subgradient base point $t \in \mathbb{R}$ can be used to find a tuple $(\tilde{y}_i, \tilde{f}_i, v_i, \tilde{g}_i) \in \mathscr{B}_a$ which fulfills (3.4.12). In the general case, we have the following statement.

LEMMA 3.5.1. *For every $\varepsilon > 0$ there exists a set $D \subset [0,1]$ with positive Lebesgue measure such that for all $t \in D$ there exists a subgradient $\tilde{g}_i \in \partial p(\iota(tx_i + (1-t)\tilde{y}_i))$ which fulfills*

$$f(\tilde{y}_i) + \langle \tilde{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - f(x_i) \le \varepsilon.$$

Proof. Let $\varepsilon > 0$ be arbitrary. Define the locally Lipschitz continuous function $r : \mathbb{R} \to \mathbb{R}$ via $r(t) := p(\iota(t\tilde{y}_i + (1-t)x_i))$. By [93, Cor. 2.2], there exists a set $D \subset [0,1]$ with positive Lebesgue measure such that, for each $t \in D$, $r'(t)$ exists and $r(1) - r'(t) - r(0) \le \varepsilon$. From [21, Prop. 2.2.2] and [21, Thm. 2.3.10], we obtain for all $t \in \tilde{D}$

$$r'(t) \in \partial_C r(t) \subset \{\langle g, \iota(\tilde{y}_i - x_i)\rangle_{Y^*,Y} : g \in \partial_C p(\iota(t\tilde{y}_i + (1-t)x_i))\}.$$

This shows that there exists a subgradient $\tilde{g}_i \in \partial p(\iota(tx_i + (1-t)\tilde{y}_i))$ which fulfills

$$f(\tilde{y}_i) + \langle \tilde{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - f(x_i) = r(1) - r'(t) - r(0) \le \varepsilon.$$ □

In general, it is not clear how to find a subgradient which fulfills (3.4.12). Therefore, we use Algorithm 3.3 as a heuristic to search for a suitable subgradient. Note that we cannot guarantee that Algorithm 3.3 actually returns a subgradient which fulfills (3.4.12). However, we motivate Algorithm 3.3 by the fact that

$$[\tilde{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ \le [O_f(\tilde{y}_i, h_i) - l_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ + 2l_i,$$

i.e., we search for a new subgradient whenever the computed linearization error

$$e_{\text{comp}} := [O_f(\tilde{y}_i, h_i) - l_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+$$

is larger than the computed lifting error $2l_i$. In the case that the computed linearization error should be reduced, we choose a new subgradient base point within the interval $[L,R] \subset [x_i, \tilde{y}_i]$ which is constructed in such a way that $t \mapsto f(x_i + t(\tilde{y}_i - x_i))$ can be expected to be flat in $[L,R]$. Note that, by (3.5.2), the computed linearization error e_{comp} always fulfills

$$e_{\text{comp}} = [O_f(\tilde{y}_i, h_i) - l_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ \leq [f(\tilde{y}_i) + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - f(x_i)]_+$$

and thus Lemma 3.5.1 is applicable.

Algorithm 3.3: Subgradient search $[(\tilde{g}, \hat{v}, \hat{g}, \hat{f}, h, l) :=$ `Subgradient`$(e,h,l)]$

Parameters : Subgradient oracle $O_g : X \times (0,1] \to Y^*$, function value oracle $O_f : X \times (0,1] \to \mathbb{R}$, trial iterate $\tilde{y}$, serious iterate x, approximation of the function value at the serious iterate $\tilde{f}^x$, number of iterations $n_{it} \in \mathbb{N}_+$

Input : Desired error tolerance e, current accuracy level h and lift term l.

Output : Subgradient $\tilde{g}$ at $\tilde{y}_i$, subgradient base point $\hat{v}$, subgradient $\hat{g}$, function value approximation $\hat{f}$, new lift term l and new accuracy level $\hat{h}$.

Initialization: Set $M := \tilde{y}, L := x, R := \tilde{y}$, $f_L := \tilde{f}^x$, $f_R := O_f(\tilde{y},h)$. Compute $(\hat{f},h) :=$ `FunctionValue`$(\tilde{y},l,h)$ and $\hat{g} = O_g(\tilde{y},h)$. Set $\tilde{g} := \hat{g}$.

```
for j = 1,...,n_it do
    if [f̂ + ⟨ĝ, ι(x − ỹ)⟩_{Y*,Y} − f̃^x]_+ ≤ e then
        return g̃, v̂ := M, ĝ, f̂, l and h.
    end

    if [O_f(ỹ,h) − l + ⟨ĝ, ι(x − ỹ)⟩_{Y*,Y} − f̃^x]_+ < 2l then
        Set l := l/2 and compute (f̂,h) := FunctionValue(ỹ,l,h).
        Continue to the next iteration.
    else
        Set M := (L+R)/2, compute f_M := O_f(M,h) and ĝ := O_g(M).
        if f_L − f_M < f_M − f_R then
            Set R := M and f_R := f_M.
        else
            Set L := M and f_L := f_M.
        end
    end
end
return g̃, v̂ := M, ĝ, f̂, l and h.
```

3.5.3. The trial iterate oracle

To execute the bundle method, we need to compute an approximate solution of the bundle subproblem, i.e., a trial iterate. We postpone the task of developing a practical algorithm to compute such a trial iterate to Chapter 4. For now, we assume that we have access to a computable oracle $O_{\tilde{y}} : (0,\infty)^2 \to X \times (0,\infty)$ such that, for a desired accuracy $\varepsilon > 0$ and at a given accuracy level h, the oracle returns an approximate solution $\tilde{y}$ and a new accuracy level $\tilde{h} \le h$ via $(\tilde{y},\tilde{h}) = O_{\tilde{y}}(\varepsilon,h)$, such that

$$\Psi(\tilde{y}) < \Psi(x) \qquad \text{and} \qquad \Psi(\tilde{y}) \le \min_{y\in X} \Psi(y) + C_{\tilde{y}}\varepsilon, \tag{3.5.6}$$

where the constant $C_{\tilde{y}} > 0$ does not depend on Ψ, $\tilde{y}$ and h and does not have to be known. Note that the oracle is allowed to increase the accuracy level. The additional flexibility gained by introducing the unknown constant $C_{\tilde{y}}$ will be useful later on. As an illustrative example, if the evaluation of $O_{\tilde{y}}(\varepsilon,h)$ involves solving a PDE, FEM error estimates can be employed. These FEM error estimates often have the form $\|u-u^h\|_{H_0^1(\Omega)} \le Ch^\alpha \|f\|_{L^2(\Omega)}$, where $u \in H_0^1(\Omega)$ is the solution to the PDE on the domain Ω with right hand side $f \in L^2(\Omega)$, u^h is an approximation of u, $\alpha > 0$ is a known constant and $C > 0$ is a constant which only depends on problem data (such as Ω) and the construction of the approximation u^h but not on u^h itself. Often, the constant C is not known and it is very difficult to compute a suitable estimate of C. In this case, we can include C into $C_{\tilde{y}}$ and we do not have to know C at all.

3.5.4. Practical algorithm

At the end of a successful iteration of Algorithm 3.1, one is allowed to choose the new proximity parameter τ_{i+1} arbitrarily in the interval $[q_{i+1}+\xi, T]$. Our numerical experiments suggest that if the ratio ρ_i of computed reduction $\tilde{J}_i^x - \tilde{J}_i$ to predicted reduction $\Phi_i(x_i) - \Phi_i(\tilde{y}_i)$ is large, then it is beneficial to decrease the value of the proximity parameter. We thus introduce a threshold $\Gamma \in (\gamma,1)$ and use $\tau_{i+1} := P_{[q_{i+1}+\xi,T]}(\frac{1}{2}\tau_i)$ if $\rho_i \ge \Gamma$ and $\tau_{i+1} := P_{[q_{i+1}+\xi,T]}(\tau_i)$ if $\rho_i < \Gamma$. Furthermore, let $(\nu_i^\Psi)_{i\in\mathbb{N}} \subset (0,\infty)$, $(\nu_i^{\text{lin}})_{i\in\mathbb{N}} \subset (0,\infty)$ and $(\nu_i^f)_{i\in\mathbb{N}} \subset (0,\infty)$ be forcing sequences with the property that for any subsequence $\mathscr{I} \subset \mathbb{N}$ with $\tau_i \to \infty$ as $\mathscr{I} \ni i \to \infty$ it holds

$$\tau_i \nu_i^\Psi \to 0, \qquad \tau_i \nu_i^{\text{lin}} \to 0 \qquad \text{and} \qquad \nu_i^f \to 0 \qquad\qquad \text{as } \mathscr{I} \ni i \to \infty. \tag{3.5.7}$$

Now we are able to present the practical algorithm, Algorithm 3.4. In order to present a convergence theorem for Algorithm 3.4, we use the following nomenclature. Denote by $(\tilde{g},\hat{v},\hat{g},\hat{f},h,l) :=$ `Subgradient`(e,h,l) the output of Algorithm 3.3. We call the subgradient $\hat{g}$ valid, if

$$[\hat{f} + \langle \hat{g}, \iota(x-\tilde{y})\rangle_{Y^*,Y} - \tilde{f}^x]_+ \le e.$$

THEOREM 3.5.2. *Let the initial point $x_0 \in \mathscr{F}$ be such that the level set $\{x \in X : J(x) \le J(x_0) + 2\Delta\}$ is bounded in X and assume that one of the following holds true:*

- *Algorithm 3.3 always returns a valid subgradient.*
- *$G := \partial_C p + \bar{B}_{Y^*}(0,\Delta_2)$, $\Delta_2 \ge 0$, $\operatorname{dist}(\tilde{g}_i, \partial_C p(\iota\tilde{y}_i)) \to 0$ and p is approximately convex.*

Then every weak limit point of the sequence of serious iterates produced by Algorithm 3.4 is G-stationary in the sense of Definition 3.1.2.

Proof. First note that every trial iterate $\tilde{y}_i$ fulfills $\Psi_i(\tilde{y}_i) - \Psi_i(y_i) \le \Psi(\tilde{y}) \le C_{\tilde{y}}\nu_i^\Psi$. Defining $\alpha_i := C_{\tilde{y}}\nu_i^\Psi \tau_i$, we find that (3.4.10) and (3.4.15) are fulfilled and it holds $\alpha_i \to 0 =: \alpha$. From Algorithm 3.3, one easily sees that v_i is contained in the line segment $[x_i,\tilde{y}_i]$. Setting $M_v = 1$ shows that (3.4.11) is fulfilled. In successful iterations, only the subgradient base point x_{i+1} is added. In unsuccessful iterations, two subgradient base points, $\tilde{y}_i$ and $v_i \in [x_i,\tilde{y}_i]$, are added. Since $\mathscr{S} \subset \mathscr{F}$ is bounded, Lemma 3.1.14

Algorithm 3.4: Inexact bundle method for optimal control problems

Parameters : $0 < \gamma < \tilde{\gamma} < 1$, $\gamma < \Gamma < 1$, $0 < \bar{q} < \bar{q} + \xi \leq T$. Forcing sequences $(\nu_i^{\Psi})_{i\in\mathbb{N}}$, $(\nu_i^{\mathrm{lin}})_{i\in\mathbb{N}}$ and $(\nu_i^{f})_{i\in\mathbb{N}}$ fulfilling (3.5.7). Oracles O_f, O_g, $O_{\tilde{y}}$, $\underline{O}_{\|\cdot\|_Y}$.

Initialization: Choose a start iterate $x_0 \in \tilde{\mathscr{T}} \cap \mathscr{V}$, an initial lift term l_0 and set $h_0 = 1$. Compute $\tilde{f}_0^x = O_f(x_0,h_0)$ and $\tilde{g}_0^x = O_g(x_0,h_0)$. Set $\tilde{J}_0^x = \tilde{f}_0^x + w(x_0)$ and choose a symmetric operator $Q_0 \in \mathscr{L}(Y,Y^*)$ and $q_0 \leq \bar{q}$ satisfying (3.1.17). Choose $\tau_0 \in [q_0+\xi, T]$ and set $\mathscr{B}_0 = \{(x_0, \tilde{f}_0^x, x_0, \tilde{g}_0^x)\}$ and $\mathscr{M}_0 = \{m_0^x\}$.

for $i = 0,1,\dots$ **do**

Set $\Phi_i = \max\{m : m \in \mathscr{M}_i\} + w + \delta_{\mathscr{F}}$ and $\Psi_i = \Phi_i + \frac{1}{2}\|\iota(\,\cdot\, - x_i)\|^2_{Q_i+\tau_i R_Y}$.

Trial iterate generation. Compute a new trial iterate and accuracy level via $(\tilde{y}_i, \tilde{h}_i) := O_{\tilde{y}}(\nu_i^{\Psi}, h_i)$.

Function value refinement. Choose a new lift term $\hat{l}_i \in (0, \max\{l_i, \nu_i^f \underline{O}_{\|\cdot\|_Y}(\tilde{y}_i - x_i)\}]$, compute a new function value and accuracy level via $(\tilde{f}_i, \hat{h}_i) :=$ `FunctionValue`$(\tilde{y}, \tilde{h}_i, \hat{l}_i)$ and set $\tilde{J}_i = \tilde{f}_i + w(\tilde{y}_i)$.

Acceptance test. Set

$$\rho_i = \frac{\tilde{J}_i^x - \tilde{J}_i}{\Phi_i(x_i) - \Phi_i(\tilde{y}_i)}.$$

if $\rho_i \geq \gamma$ **then** (successful iteration)

Set $x_{i+1} := \tilde{y}_i$, $\tilde{f}_{i+1}^x := \tilde{f}_i$, $\tilde{J}_{i+1}^x := \tilde{J}_i$, $h_{i+1} := \hat{h}_i$, $l_{i+1} := \hat{l}_i$ and compute an exactness subgradient $\tilde{g}_{i+1}^x = O_g(x_{i+1}, h_{i+1})$. Choose a symmetric operator $Q_{i+1} \in \mathscr{L}(Y,Y^*)$ with curvature bound $q_{i+1} \leq \bar{q}$ satisfying (3.1.17). Set the proximity parameter to

$$\tau_{i+1} := \begin{cases} P_{[q_{i+1}+\xi,T]}(\tau_i) & \rho_i < \Gamma \\ P_{[q_{i+1}+\xi,T]}(\frac{1}{2}\tau_i) & \rho_i \geq \Gamma \end{cases}$$

and set the bundle information to $\mathscr{B}_{i+1} = \{(x_{i+1}, \tilde{f}_{i+1}^x, x_{i+1}, \tilde{g}_{i+1}^x)\}$. Use $\mathscr{M}_{i+1} = \{m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot, x_i) : (\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_{i+1}\}$ and continue to the next iteration.

else

Set $x_{i+1} := x_i$, $\tilde{f}_{i+1}^x := \tilde{f}_i^x$, $\tilde{g}_{i+1}^x := \tilde{g}_i^x$, $\tilde{J}_{i+1}^x := \tilde{J}_i^x$, $Q_{i+1} := Q_i$, $q_{i+1} := q_i$.

end

Update local model. Compute $(\tilde{g}_i, v_i, \hat{g}_i, \hat{f}_i, h_i', l_i') :=$ `Subgradient`$(\nu_i^{\mathrm{lin}}, \hat{h}_i, \hat{l}_i)$. Set the bundle information to $\mathscr{B}_{i+1} := \mathscr{B}_i \cup \{(\tilde{y}_i, \tilde{f}_i, \tilde{y}_i, \tilde{g}_i), (\tilde{y}_i, \hat{f}_i, v_i, \hat{g}_i)\}$. Set $h_{i+1} := h_i'$, $l_{i+1} := l_i'$, $\mathscr{M}_{i+1} := \{m_{\tilde{y},\tilde{f},v,\tilde{g}}(\cdot, x_i) : (\tilde{y},\tilde{f},v,\tilde{g}) \in \mathscr{B}_{i+1}\}$, and $\Phi_{i+1} := \max\{m : m \in \mathscr{M}_{i+1}\} + w + \delta_{\mathscr{F}}$.

Update proximity parameter.

$$\text{Set } \tilde{\rho}_i = \frac{\tilde{J}_i^x - \Phi_{i+1}(\tilde{y}_i)}{\Phi_i(x_i) - \Phi_i(\tilde{y}_i)} \text{ and update } \tau_{i+1} = \begin{cases} 2\tau_i & \text{if } \tilde{\rho}_i \geq \tilde{\gamma} \text{ (proximity iteration)} \\ \tau_i & \text{if } \tilde{\rho}_i < \tilde{\gamma} \text{ (model iteration)} \end{cases}.$$

end

shows that the set of trial iterates $\tilde{\mathscr{Y}} = \{\tilde{y}_i : i \in \mathbb{N}\}$ is bounded in X. Thus, the set of subgradient base points $\mathscr{V} \subset \mathrm{co}(\mathscr{S} \cup \tilde{\mathscr{Y}})$ is bounded and Lemma 3.1.14 yields the boundedness of the set of bundle information $\mathscr{B}_u$.

Now assume that Algorithm 3.3 always returns a valid subgradient, i.e., assume that in every unsuccessful iteration Algorithm 3.3 returns a tuple $(\tilde{g}_i, v_i, \hat{g}_i, \hat{f}_i, h'_i, l'_i)$ such that

$$[\hat{f}_i + \langle \hat{g}_i, \iota(x_i - \tilde{y}_i)\rangle_{Y^*,Y} - \tilde{f}_i^x]_+ \leq \nu_i^{\mathrm{lin}}.$$

Defining $\beta_i := \nu_i^{\mathrm{lin}} \tau_i$ shows that β_i fulfills (3.4.12) and it holds $\beta_i \to 0 =: \beta$. Since the cutting plane $m_{\tilde{y}_i, \hat{f}_i, v_i, \hat{g}_i}(\cdot, x_i)$ is included into the new model ϕ_{i+1} whenever iteration i was not successful, Theorem 3.4.9 shows that every weak limit point of the sequence of serious iterates produced by Algorithm 3.4 is η-G-stationary with $\eta = 0$.

Finally assume that $G := \partial_C p + \bar{B}_{Y^*}(0, \Delta_2)$, $\Delta_2 \geq 0$, $\operatorname{dist}(\tilde{g}_i, \partial_C p(\iota \tilde{y}_i)) \to 0$ and p is approximately convex. By (3.5.3), we find

$$[\tilde{f}_i - f(\tilde{y}_i) - \tilde{f}_i^x + f(x_i)]_+ \leq 2l_i \leq \nu_i^f \underline{Q}_{\|\cdot\|_Y}(\tilde{y}_i - x_i) \leq \nu_i^f C_{\|\cdot\|_Y}^{-1} \|\iota(\tilde{y}_i - x_i)\|_Y.$$

Since $\mathring{\varepsilon}_{1,i} := \nu_i^f C_{\|\cdot\|_Y}^{-1} \to 0$, (3.4.16) is fulfilled with $\mathring{\varepsilon}_1 := 0$. Therefore, Corollary 3.4.12 shows that every weak limit point of the sequence of serious iterates produced by Algorithm 3.4 is G-stationary. □

4. The bundle subproblem

The goal of this chapter is to provide an efficient procedure to compute an approximate solution $\tilde{y}_i$ of the subroblem of iteration i of the bundle method and to develop corresponding error estimates. In particular, we are interested in upper bounds on the quantity $\Psi_i(\tilde{y}_i) - \min_{y\in X} \Psi_i(y)$ to construct an oracle which fulfills (3.5.6). In this chapter, the iteration number i is always the same. Thus, we suppress the index i.

4.1. Automated aggregation

The bundle subproblem (3.1.2) has the form

$$\min_{y\in X} \Psi(y) := \max\{m(y) : m \in \mathscr{M}\} + w(y) + \delta_{\mathscr{F}}(y) + \tfrac{1}{2}\|\iota(y - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y}. \tag{4.1.1}$$

Here, $x_{\mathrm{SI}} \in X$ is the serious iterate, $\mathscr{M} = \{m_j, 1 \le j \le n_p\}$, $n_p \in \mathbb{N}_+$, is a finite set of affine linear cutting planes and the curvature approximation $Q : Y \to Y^*$ and the proximity parameter $\tau > 0$ are chosen such that the bilinear form $\langle (Q+\tau R_Y)\cdot,\cdot\rangle_{Y^*,Y}$ induces the norm $\|\cdot\|^2_{Q+\tau R_Y} := \langle (Q+\tau R_Y)\cdot,\cdot\rangle_{Y^*,Y}$ on Y, cf. Section 3.1.6. Two difficulties arise in solving (4.1.1). First, a large number of cutting planes n_p may lead to high computational costs. Although aggregation of cutting planes can reduce the number n_p, in practice it is difficult to develop a sound aggregation strategy, cf. Remark 3.3.5. Second, in many practical applications, the curvature term $\frac{1}{2}\|\iota(y - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y}$ may not be computable exactly. For example, if $X^* = L^2(\Omega)$, $Y^* = H^1_0(\Omega)$ and $Q = 0$, it holds

$$\|\iota(y - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y} = \tau\langle R_Y \iota(y - x_{\mathrm{SI}}), \iota(y - x_{\mathrm{SI}})\rangle_{Y^*,Y} = \tau\|\iota(y - x_{\mathrm{SI}})\|^2_{H^{-1}(\Omega)}.$$

To compute the $H^{-1}(\Omega)$-norm of $y - x_{\mathrm{SI}} \in L^2(\Omega)$, one usually solves the Dirichlet problem $-\Delta u = y - x_{\mathrm{SI}}$ on Ω and $u = 0$ on ∂D. However, this PDE cannot be solved analytically and thus numerical methods to approximate the solution have to be employed.

Our strategy to solve (4.1.1) up to a desired accuracy is as follows. We approximately solve the reduced problem

$$\min_{y\in X} \hat{\Psi}(y) := \max\{m(y) : m \in \hat{\mathscr{M}}\} + w(y) + \delta_{\mathscr{F}}(y) + \tfrac{1}{2}\|\iota(y - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y}, \tag{4.1.2}$$

where $\hat{\mathscr{M}}$ is an appropriate subset of $\mathscr{M}$. If the approximate solution fulfills (3.5.6), we are done. Otherwise, we either increase the accuracy of the solution to (4.1.2) or we enlarge the set $\hat{\mathscr{M}} \subset \mathscr{M}$. In Section 4.2 we discuss how to solve the reduced problem (4.1.2). For now, let us assume that we have access to a subproblem oracle O_s which, given a piecewise quadratic strongly convex function $\hat{\Psi} : X \to \mathbb{R}$ and an accuracy level h, returns an approximate solution $\hat{z} = O_s(\hat{\Psi}, h)$ of the problem $\min_{y\in X} \hat{\Psi}(y)$.

Furthermore, we assume that there exists a corresponding computable error bound $\hat{\varepsilon}_s = \hat{\varepsilon}_s(\hat{\Psi}, h)$ such that

$$|\hat{\Psi}(\hat{z}) - \min_{y \in X} \hat{\Psi}(y)| \leq C_{\hat{\Psi}} \hat{\varepsilon}_s(\hat{\Psi}, h) \qquad \text{and} \qquad \hat{\varepsilon}_s(\hat{\Psi}, h) \to 0 \text{ as } h \to 0. \tag{4.1.3}$$

Here, the constant $C_{\hat{\Psi}} > 0$ does not depend on $\hat{\Psi}$, $\hat{z}$ and h and does not have to be computable.

Furthermore, to be able to verify if a computed trial iterate $\tilde{y}$ fulfills $\Psi(\tilde{y}) < \Psi(x_{\mathrm{SI}})$, we assume that we have access to model value oracles which yield lower and upper bounds for the value of $\Psi(\tilde{y})$ for a given accuracy level h. The lower bound $\underline{O}_{\Psi}(h)$ and the upper bound $\overline{O}_{\Psi}(h)$ are supposed to fulfill

$$\underline{O}_{\Psi}(h) \leq \Psi(\tilde{y}) \leq \overline{O}_{\Psi}(h), \qquad \text{and} \qquad \overline{O}_{\Psi}(h) - \underline{O}_{\Psi}(h) \to 0 \text{ as } h \to 0. \tag{4.1.4}$$

In order to efficiently solve the bundle subproblem, the algorithm should balance the number of cutting planes $|\hat{\mathscr{M}}|$ for the reduced model, the accuracy of solving the reduced problem, and the accuracy of the model value oracle. Ideally, $\hat{\mathscr{M}}$ only contains cutting planes which are active at the exact solution, i.e., cutting planes $m \in \mathscr{M}$ such that $m(y^*) + w(y) + \frac{1}{2}\|\iota(y - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y} = \Psi(y^*)$ where y^* is the exact solution of (4.1.1). Since obviously y^* is not known prior to solving (4.1.1), we develop a strategy which successively builds up $\hat{\mathscr{M}}$ until $|\Psi - \hat{\Psi}|$ is sufficiently small in a neighborhood of y^*. We start by consulting the subproblem oracle to obtain an approximate solution to (4.1.2), i.e., $\hat{z} := O_s(\hat{\Psi}, h)$ and use this as a candidate for a trial iterate. If the error estimate $\hat{\varepsilon}_s(\hat{\Psi}, h)$ for the solution of the aggregated bundle subproblem (4.1.2) is larger than the constant $f^1 > 0$, we increase the accuracy level $h_{\mathrm{new}} = h/2$ and recompute $\hat{z}_{\mathrm{new}} = O_s(\hat{\Psi}, h_{\mathrm{new}})$. Otherwise, we compute the value $\phi(\hat{z}) = \max_{m \in \mathscr{M}} m(\hat{z})$ of the local model without aggregation. Let $f^2 > 0$ be a fixed constant. If $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) = \phi(\hat{z}) - \hat{\phi}(\hat{z}) > f^2$, then we add the cutting plane $m \in \mathscr{M}$ with largest value $m(\hat{z})$ to the aggregate model $\hat{\phi}$, i.e., the new aggregate model is $\hat{\phi}_{\mathrm{new}} = \max(\hat{\phi}(\cdot), m(\cdot))$. We then update the reduced model $\hat{\Psi}_{\mathrm{new}} = \hat{\phi}_{\mathrm{new}}(y) + w(y) + \delta_{\mathscr{F}}(y) + \frac{1}{2}\|\iota(y - x_{\mathrm{SI}})\|^2_{Q_i + \tau_i R_Y}$ and recompute $\hat{z}_{\mathrm{new}} = O_s(\hat{\Psi}_{\mathrm{new}}, h)$. If both conditions $\hat{\varepsilon}_s(\Psi, h) \leq f^1$ and $\phi(\hat{z}) - \hat{\phi}(\hat{z}) \leq f^2$ are fulfilled, we use $\hat{z}$ as the trial iterate $\tilde{y}$. This procedure is presented in Algorithm 4.1.

If the serious iterate x_{SI} minimizes the full model Ψ, then no trial iterate $\tilde{y}$ can be found which fulfills $\Psi(\tilde{y}) < \Psi(x_{\mathrm{SI}}) = \min_{y \in X} \Psi(y)$. Unfortunately, since we work with inexact data, there is no way of determining if x_{SI} minimizes Ψ. In this case, Algorithm 4.1 keeps refining indefinitely. However, in the case that the serious iterate x_{SI} does not minimize the full model Ψ, we have the following convergence statement.

THEOREM 4.1.1. *If $\Psi(x_{SI}) > \min_{y \in X} \Psi(y)$, then Algorithm 4.1 terminates after finitely many steps and the computed trial iterate $\tilde{y}$ fulfills $\Psi(\tilde{y}) \leq \min_{y \in X} \Psi(y) + C_{\hat{\Psi}} f^1 + f^2$ and $\Psi(\tilde{y}) < \Psi(x_{SI})$.*

Proof. First observe that $\hat{\mathscr{M}} \subset \mathscr{M}$ in every step of the algorithm. Therefore,

$$\max\{m(y) : m \in \hat{\mathscr{M}}\} \leq \max\{m(y) : m \in \mathscr{M}\} \qquad \text{for all } y \in X.$$

This shows that $\hat{\Psi}(y) \leq \Psi(y)$ for all $y \in X$ and thus $\min_{y \in X} \hat{\Psi}(y) \leq \min_{y \in X} \Psi(y)$. Consequently, (4.1.3)

Algorithm 4.1: Computing a trial iterate $[(\tilde{y},h) :=$ `TrialIterate`$(h, f^1, f^2)]$

Parameters : Subproblem oracle O_s with error bound $\hat{\varepsilon}_s$, model value oracles $\underline{O}_\Psi$ and $\overline{O}_\Psi$.
Input : Initial accuracy level h, desired error bounds $f^1 > 0$, $f^2 > 0$.
Output : Trial iterate $\tilde{y}$ and new accuracy level h.
Initialization: Choose a subset of cutting planes $\hat{\mathscr{M}}$ such that $\{m^x\} \subset \hat{\mathscr{M}} \subset \mathscr{M}$. Set the full model to $\Psi(\cdot) = \max_{m\in\mathscr{M}} m(\cdot) + w(\cdot) + \delta_{\mathscr{F}}(\cdot) + \frac{1}{2}\|\iota(\cdot - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y}$.

1 **Loop**
2 Set the aggregate model to $\hat{\Psi}(\cdot) = \max_{m\in\hat{\mathscr{M}}} m(\cdot) + w(\cdot) + \delta_{\mathscr{F}}(\cdot) + \frac{1}{2}\|\iota(\cdot - x_{\mathrm{SI}})\|^2_{Q+\tau R_Y}$.
3 Compute approximate solution of aggregate bundle subproblem $\hat{z} = O_s(\hat{\Psi}, h)$.
4 **if** $\hat{\varepsilon}_s(\hat{\Psi}, h) > f^1$ **then**
5 Set $h = h/2$. Continue to the next iteration.
6 **end**
7 **if** $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) > f^2$ **then**
8 Set $\hat{\mathscr{M}} := \hat{\mathscr{M}} \cup \arg\max_{m\in\mathscr{M}} m(\hat{z})$. Continue to the next iteration.
9 **end**
10 Set $h_c := h$.
11 **if** $\overline{O}_\Psi(h_c) < \Psi(x_{SI})$ **then**
12 **return** $\tilde{y} := \hat{z}$ and h.
13 **end**
14 **if** $\underline{O}_\Psi(h_c) > \Psi(x_{SI})$ **then**
15 **if** $\hat{\varepsilon}_s(\hat{\Psi}, h) \geq \Psi(\hat{z}) - \hat{\Psi}(\hat{z})$ **then**
16 Set $h = h/2$. Continue to the next iteration.
17 **else**
18 Set $\hat{\mathscr{M}} := \hat{\mathscr{M}} \cup \arg\max_{m\in\mathscr{M}} m(\hat{z})$. Continue to the next iteration.
19 **end**
20 **end**
21 **if** $\hat{\varepsilon}_s(\hat{\Psi}, h) > \max\{\Psi(\hat{z}) - \hat{\Psi}(\hat{z}), \overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c)\}$ **then**
22 Set $h = h/2$. Continue to the next iteration.
23 **end**
24 **if** $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) > \overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c)$ **then**
25 Set $\hat{\mathscr{M}} := \hat{\mathscr{M}} \cup \arg\max_{m\in\mathscr{M}} m(\hat{z})$. Continue to the next iteration.
26 **end**
27 Set $h_c := h_c/2$. Go to line 11.
28 **end**

yields

$$\begin{aligned}\Psi(\hat{z}) - \min_{y\in X}\Psi(y) &= \Psi(\hat{z}) - \hat{\Psi}(\hat{z}) + \hat{\Psi}(\hat{z}) - \min_{y\in X}\hat{\Psi}(y) + \min_{y\in X}\hat{\Psi}(y) - \min_{y\in X}\Psi(y) \\ &\leq \Psi(\hat{z}) - \hat{\Psi}(\hat{z}) + C_{\hat{\Psi}}\hat{\varepsilon}_s(\hat{\Psi}, h).\end{aligned} \tag{4.1.5}$$

Whenever Algorithm 4.1 terminates, it holds $\hat{\varepsilon}_s(\hat{\Psi}, h) \leq f^1$, $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) \leq f^2$ and $\overline{O}_\Psi(h_c) < \Psi(x_{\mathrm{SI}})$.

Thus, whenever a trial iterate $\tilde{y} = \hat{z}$ is returned, it holds $\Psi(\tilde{y}) \leq \min_{y\in X} \Psi(y) + C_{\hat{\Psi}} f^1 + f^2$ and $\Psi(\tilde{y}) \leq \overline{O}_\Psi(h_c) < \Psi(x_{\mathrm{SI}})$. This shows the second part of the theorem.

Now we show that Algorithm 4.1 terminates after finitely many steps. First consider the case that, after finitely many steps, it holds $\hat{\varepsilon}_s(\hat{\Psi}, h) = 0$ and $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) = 0$. In this case, (4.1.3) yields $\hat{\Psi}(\hat{z}) = \min_{y\in X} \hat{\Psi}(y)$ and we obtain

$$\min_{y\in X} \Psi(y) \leq \Psi(\hat{z}) = \hat{\Psi}(\hat{z}) = \min_{y\in X} \hat{\Psi}(y) \leq \min_{y\in X} \Psi(y).$$

Thus, $\Psi(\hat{z}) = \min_{y\in X} \Psi(y)$ and (4.1.4) implies $\underline{O}_\Psi(h_c) \leq \Psi(\hat{z}) = \min_{y\in X} \Psi(y) < \Psi(x_{\mathrm{SI}})$. Therefore, the **if**-clause in line 14 is not entered. Also the **if**-clauses in lines 4, 7, 21, and 24 are not entered. Thus, h and $\hat{\mathscr{M}}$ are not changed anymore, but h_c is halved whenever line 27 is executed. As $\overline{O}_\Psi(h_c) \to \Psi(\hat{z})$ for $h_c \to 0$ and $\Psi(\hat{z}) = \min_{y\in X} \Psi(y) < \Psi(x_{\mathrm{SI}})$, after finitely many steps it holds that $\overline{O}_\Psi(h_c) < \Psi(x_{\mathrm{SI}})$. Consequently line 11 ensures that Algorithm 4.1 stops.

From now on we consider the case that, throughout the algorithm, either $\hat{\varepsilon}_s(\hat{\Psi}, h) \neq 0$ or $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) \neq 0$, i.e., that $\max\{\hat{\varepsilon}_s(\hat{\Psi}, h), \Psi(\hat{z}) - \hat{\Psi}(\hat{z})\} > 0$. We continue by proofing that every iteration is completed after a finite number of steps by showing that line 27 can only be executed a finite number of times in each iteration. Note that h, $\hat{\mathscr{M}}$ and $\hat{z}$ are not changed within any iteration. Thus, also $\hat{\varepsilon}_s(\hat{\Psi}, h)$ and $\Psi(\hat{z}) - \hat{\Psi}(\hat{z})$ are not changed within any iteration. Whenever line 27 is executed, h_c is halved and $\overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c) \to 0$ as $h_c \to 0$. Since $\max\{\hat{\varepsilon}_s(\hat{\Psi}, h), \Psi(\hat{z}) - \hat{\Psi}(\hat{z})\} > 0$, after finitely many steps, it holds that $\overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c) < \max\{\hat{\varepsilon}_s(\hat{\Psi}, h), \Psi(\hat{z}) - \hat{\Psi}(\hat{z})\}$. If $\overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c) < \Psi(\hat{z}) - \hat{\Psi}(\hat{z})$, then line 24 ensures that the current iteration is terminated. Otherwise, $\overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c) \geq \Psi(\tilde{y}) - \hat{\Psi}(\tilde{y})$,

$$\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) \leq \overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c) < \max\{\hat{\varepsilon}_s(\hat{\Psi}, h), \Psi(\hat{z}) - \hat{\Psi}(\hat{z})\} = \hat{\varepsilon}_s(\hat{\Psi}, h)$$

and line 21 ensures that the current iteration is terminated. This shows that every iteration is terminated after a finite number of steps.

Next, we show that Algorithm 4.1 terminates after a finite number of iterations. At the end of every iteration, either h is decreased or $\hat{\mathscr{M}}$ is set to $\hat{\mathscr{M}} \cup \arg\max_{m\in\mathscr{M}} m(\hat{z})$. Whenever $\hat{\mathscr{M}}$ is updated, it holds $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) > 0$. As

$$\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) = \max_{m\in\mathscr{M}} m(\hat{z}) - \max_{m\in\hat{\mathscr{M}}} m(\hat{z}) > 0,$$

the nonempty set $\arg\max_{m\in\mathscr{M}} m(\hat{z})$ and the set $\hat{\mathscr{M}}$ are disjoint. Thus, whenever $\hat{\mathscr{M}}$ is updated, the number of elements in $\hat{\mathscr{M}}$ increases. As $\mathscr{M}$ is a finite set, this can happen only a finite number of times. Therefore, after a finite number of iterations, $\hat{\mathscr{M}}$ does not change anymore. Consequently, $\hat{\Psi}$ does not change and h is halved in every subsequent iteration which yields $\hat{\varepsilon}_s(\hat{\Psi}, h) \to 0$. Therefore, if Algorithm 4.1 did not stop earlier, there is an iteration with $\hat{\varepsilon}_s(\hat{\Psi}, h) < \min\{f^1, (\Psi(x_{\mathrm{SI}}) - \min_{y\in X} \Psi(y))/(2 + C_{\hat{\Psi}})\}$. As argued before, this iteration has to be completed after finitely many steps. We now argue that the algorithm has to stop at the end of this iteration. Since $\hat{\mathscr{M}}$ is not changed anymore, this iteration can only end at lines 5, 12, 16 or 22. Since $\hat{\varepsilon}_s(\hat{\Psi}, h) < f^1$, line 5 cannot be executed. Now assume that line 16 is executed. This can only happen if $\underline{O}_\Psi(h_c) > \Psi(x_{\mathrm{SI}})$ and $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) \leq \hat{\varepsilon}_s(\hat{\Psi}, h)$ hold true.

However, this can not be true, since (4.1.4) and (4.1.5) imply

$$\begin{aligned}\underline{O}_\Psi(h_c) &\le \Psi(\hat{z}) = \Big(\Psi(\hat{z}) - \min_{y\in X}\Psi(y)\Big) + \Big(\min_{y\in X}\Psi(y) - \Psi(x_{\mathrm{SI}})\Big) + \Psi(x_{\mathrm{SI}}) \\ &\le \Psi(\hat{z}) - \hat{\Psi}(\hat{z}) + C_{\hat{\Psi}}\hat{\varepsilon}_s(\hat{\Psi},h) - (2+C_{\hat{\Psi}})\hat{\varepsilon}_s(\hat{\Psi},h) + \Psi(x_{\mathrm{SI}}) < \Psi(x_{\mathrm{SI}}).\end{aligned}$$

Thus, line 16 is not executed. Finally, we assume that line 22 is executed. In this case, $\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) < \hat{\varepsilon}_s(\hat{\Psi},h)$ and $\overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c) < \hat{\varepsilon}_s(\hat{\Psi},h)$. Combining this with (4.1.4) and (4.1.5) implies

$$\begin{aligned}\overline{O}_\Psi(h_c) &\le \Big(\overline{O}_\Psi(h_c) - \underline{O}_\Psi(h_c)\Big) + \Big(\Psi(\hat{z}) - \min_{y\in X}\Psi(y)\Big) + \Big(\min_{y\in X}\Psi(y) - \Psi(x_{\mathrm{SI}})\Big) + \Psi(x_{\mathrm{SI}}) \\ &\le \hat{\varepsilon}_s(\hat{\Psi},h) + \Big(\Psi(\hat{z}) - \hat{\Psi}(\hat{z})\Big) + C_{\hat{\Psi}}\hat{\varepsilon}_s(\hat{\Psi},h) - (2+C_{\hat{\Psi}})\hat{\varepsilon}_s(\hat{\Psi},h) + \Psi(x_{\mathrm{SI}}) < \Psi(x_{\mathrm{SI}}).\end{aligned}$$

But this shows that line 22 cannot be executed, because the **if**-statement in line 11 would lead to the termination of Algorithm 4.1 in line 12. As lines 5, 16, and 22 cannot be executed, line 12 has to be executed in this iteration, i.e., the algorithm stops. Because every iteration is executed in finitely many steps, the algorithm terminates after finitely many steps. ∎

Remark 4.1.2. Algorithm 4.1 can be executed if the term

$$\Psi(\hat{z}) - \hat{\Psi}(\hat{z}) = \max_{m\in\mathscr{M}} m(\hat{z}) - \max_{m\in\hat{\mathscr{M}}} m(\hat{z})$$

can be computed exactly.

4.2. The dual of the bundle subproblem

In this section, we want to solve the reduced bundle problem (4.1.2) in order to a subproblem oracle which fulfills (4.1.3). Our primary interest concerns the case where $w : X \to \mathbb{R}$ is given as a Tikhonov regularization term, i.e., $w := \frac{\alpha}{2}\|\cdot\|_X^2$, $\alpha > 0$, and no control constraints are given, i.e., $\mathscr{F} = X$. In this case, the bundle subproblem (4.1.1) is an unconstrained, piecewise quadratic optimization problem in the Hilbert space X. Whenever the dimension of X is large (or when X is infinite dimensional), it is beneficial to solve the dual problem to (4.1.1) and compute the primal solution from the solution of the dual problem. The dual problem is a low dimensional optimization problem and can be solved efficiently via the method [68] which is tailored for problems with this structure. However, to set up the dual problem, operator equations in the space X have to be solved which usually cannot be done exactly. Thus, we include an error analysis to bound the error between the primal solution and the solution to the perturbed dual problem.

In order to ease presentation, from now on we do not discriminate between $\hat{\Psi}$ and Ψ. For the rest of this section, the prime symbol $'$ indicates that a variable is an element of a dual space, e.g., $g' \in X^*$. Now choose $g'_j \in X^*$ and $s_j \in \mathbb{R}$ such that the cutting plane $m_j \in \hat{\mathscr{M}}$ can be represented via $m_j(\cdot) = \langle g'_j, \cdot\rangle_{X^*,X} + s_j$, $j \in I := \{1,\dots,n_p\}$. Then we can write the reduced bundle subproblem (4.1.2) in the form

$$\underset{y\in X}{\text{minimize}} \quad \Psi(y) = \max_{j\in I}\left\{\langle g'_j, y\rangle_{X^*,X} + s_j\right\} + \frac{\alpha}{2}\|y\|_X^2 + \frac{1}{2}\|\iota(y - x_{\mathrm{SI}})\|_{Q+\tau R_Y}^2. \tag{4.2.1}$$

Rearranging terms and substituting $d := y - x_{\mathrm{SI}} \in X$ gives the equivalent problem

$$\underset{d\in X}{\text{minimize}} \quad \Psi(x_{\mathrm{SI}}+d) = \tfrac{1}{2}\langle Fd, d\rangle_{X^*,X} + \max_{j\in I}\left\{\langle p'_j, d\rangle_{X^*,X} + q_j\right\}, \tag{4.2.2}$$

where $F \in \mathscr{L}(X,X^*)$ is defined by $F := \alpha R_X + \iota^*(Q+\tau R_Y)\iota$ and $p'_j := g'_j + \alpha R_X x_{\mathrm{SI}} \in X^*$, $q \in \mathbb{R}^{n_p}$, $q_j := s_j + \langle g'_j, x_{\mathrm{SI}}\rangle_{X^*,X} + \frac{\alpha}{2}\|x_{\mathrm{SI}}\|_X^2$, $j \in I$.

LEMMA 4.2.1. *If Q fulfills the assumptions of Section 3.1.6, then the operator $F \in \mathscr{L}(X,X^*)$ is invertible.*

Proof. Since $\langle F\cdot,\cdot\rangle_{X^*,X} : X\times X \to \mathbb{R}$ is a bounded and coercive bilinear form, the Lax-Milgram theorem (Theorem 2.3.2) implies that F is invertible. ∎

Define $\Psi_\lambda : \mathbb{R}^{n_p} \to \mathbb{R}$ by $\Psi_\lambda(\lambda) := \Psi(x_{\mathrm{SI}} - \sum_{i\in I}\lambda_i F^{-1}p_i)$ and consider the dual problem

$$\begin{aligned} &\underset{\lambda\in\mathbb{R}^{n_p}}{\text{minimize}} \quad \Psi_\lambda(\lambda) = \tfrac{1}{2}\lambda^\top \mathsf{H}\lambda + \max_{j\in I}(q-\mathsf{H}\lambda)_j \\ &\text{subject to} \quad \lambda \in \Lambda, \end{aligned} \tag{4.2.3}$$

where we define $\Lambda := \{\lambda \in \mathbb{R}^{n_p} : \lambda_i \geq 0, i \in I, \sum_{i\in I}\lambda_i = 1\}$ and $\mathsf{H} \in \mathbb{R}^{n_p\times n_p}$ by $\mathsf{H}_{i,j} := \langle p'_i, F^{-1}p_j\rangle_{X^*,X}$, $i,j \in I$. Note that the matrix H is positive semidefinite because $\langle F\cdot,\cdot\rangle_{X^*,X}$ is positive semidefinite and

$$\lambda^\top\mathsf{H}\lambda = \langle F(\sum_{i\in I}\lambda_i F^{-1}p'_i), \sum_{j\in I}\lambda_j F^{-1}p'_j\rangle_{X^*,X} \geq 0 \qquad \text{for all } \lambda \in \mathbb{R}^{n_p}. \tag{4.2.4}$$

Problems (4.2.1) and (4.2.3) are equivalent in the following sense:

LEMMA 4.2.2. *The problem* (4.2.1) *has a unique solution $y^* \in X$ and there exists a $\hat\lambda \in \Lambda$ such that $y^* = x_{SI} - \sum_{i\in I}\hat\lambda_i F^{-1}p'_i$. Furthermore, $\lambda^* \in \Lambda$ solves problem* (4.2.3) *if and only if $x_{SI} - \sum_{i\in I}\lambda^*_i F^{-1}p'_i$ solves problem* (4.2.1).

Proof. Since $\alpha > 0$, the objective function Ψ is strongly convex and (4.2.1) has a unique solution. Denote this solution by $y^* \in X$. Then $d^* := y^* - x_{\mathrm{SI}}$ is the solution of problem (4.2.2). The optimality condition is given by $0 \in \partial\Psi(\cdot + x_{\mathrm{SI}})(d^*)$. Since F is self-adjoint, the derivative of $\frac{1}{2}\langle F\cdot,\cdot\rangle_{X^*,X}$ at d^* is Fd^* and we find $0 \in Fd^* + \partial(\max_{j\in I}\{\langle p'_j,\cdot\rangle_{X^*,X} + q_j\})(d^*)$. Thus, there exists $\hat\lambda \in \Lambda$ such that $0 = Fd^* + \sum_{i\in I}\hat\lambda_i p'_i$. Since $F \in \mathscr{L}(X,X^*)$ is invertible it holds $y^* = d^* + x_{\mathrm{SI}} = x_{\mathrm{SI}} - \sum_{i\in I}\hat\lambda_i F^{-1}p'_i$. Now let $\lambda^* \in \arg\min_{\lambda\in\Lambda}\Psi_\lambda(\lambda)$. As just shown, the solution of (4.2.1) can be written as $y^* = x_{\mathrm{SI}} - \sum_{i\in I}\hat\lambda_i F^{-1}p'_i$ with $\hat\lambda \in \Lambda$. The definition of Ψ_λ yields

$$\Psi(x_{\mathrm{SI}} - \sum_{i\in I}\lambda^*_i F^{-1}p'_i) = \Psi_\lambda(\lambda^*) \leq \Psi_\lambda(\hat\lambda) = \Psi(y^*) = \min_{y\in X}\Psi(y),$$

i.e., $x_{\mathrm{SI}} - \sum_{i\in I}\lambda^*_i F^{-1}p'_i$ solves problem (4.2.1). Finally let $x_{\mathrm{SI}} - \sum_{i\in I}\lambda^*_i F^{-1}p'_i$ be the solution to (4.2.1). Then

$$\Psi_\lambda(\lambda^*) = \Psi(x_{\mathrm{SI}} - \sum_{i\in I}\lambda^*_i F^{-1}p'_i) \leq \Psi(x_{\mathrm{SI}} - \sum_{i\in I}\lambda_i F^{-1}p'_i) = \Psi_\lambda(\lambda) \qquad \text{for all } \lambda \in \mathbb{R}^{n_p},$$

i.e., λ^* solves problem (4.2.3). ∎

4.2.1. Approximation of the dual problem

In many cases, the matrix H cannot be computed exactly since this involves evaluating the operator F^{-1}. Thus, we now study how to approximate the dual problem (4.2.3) in such a way, that the solution to the approximated dual problem yields an approximation of the solution to the bundle subproblem (4.2.1). Let $p_j^{F^{-1}} \in X$, $j \in I$, be arbitrary and define $\tilde{\mathsf{H}} \in \mathbb{R}^{n_p \times n_p}$ by $\tilde{\mathsf{H}}_{i,j} := \langle p_i', p_j^{F^{-1}} \rangle_{X^*,X}$, $i, j \in I$. Later we use $p_j^{F^{-1}}$ as an approximation of $F^{-1}p_j'$ such that the matrix $\tilde{\mathsf{H}}$ can be computed exactly. For $\varepsilon_H \geq 0$, we formulate the approximated problem

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^{n_p}}{\text{minimize}} \quad & \tilde{\Psi}_\lambda(\lambda) := \tfrac{1}{2}\lambda^\top \tilde{\mathsf{H}}\lambda + \tfrac{\varepsilon_H}{2}\|\lambda\|_2^2 + \max_{j \in I}(q - \tilde{\mathsf{H}}\lambda)_j \\ \text{subject to} \quad & \lambda \in \Lambda. \end{aligned} \tag{4.2.5}$$

Although H is positive semidefinite (cf. (4.2.4)), $\tilde{\mathsf{H}}$ may lack this property. To be able to solve (4.2.5) efficiently, the term $\frac{\varepsilon_H}{2}\|\lambda\|_2^2$ is added. For $\varepsilon_H \geq 0$ sufficiently large, the objective function $\tilde{\Psi}_\lambda$ is strongly convex (cf. Lemma 4.2.6). Problem (4.2.5) can be reformulated into the equivalent linear-quadratic optimization problem

$$\begin{aligned} \underset{(\lambda,m) \in \mathbb{R}^{n_p+1}}{\text{minimize}} \quad & \tfrac{1}{2}\lambda^\top \hat{\mathsf{H}}\lambda + m \\ \text{subject to} \quad & (q - \tilde{\mathsf{H}}\lambda)_j \leq m, \qquad j \in I, \\ & \lambda \in \Lambda, \end{aligned} \tag{4.2.6}$$

where $\hat{\mathsf{H}} := \tilde{\mathsf{H}}_\Delta + \varepsilon_H I_{n_p}$ with $\tilde{\mathsf{H}}_\Delta := \frac{1}{2}(\tilde{\mathsf{H}} + \tilde{\mathsf{H}}^\top)$ and $I_{n_p} \in \mathbb{R}^{n_p \times n_p}$ is the identity matrix. Let $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue of a symmetric matrix. If $\varepsilon_H \geq \max\{0, -\lambda_{\min}(\tilde{\mathsf{H}}_\Delta)\}$, then the objective function is convex and every local solution of this problem is also a global solution. This problem can be solved using standard methods such as active-set and interior-point algorithms, see, e.g., [97, Chap. 16]. If $\varepsilon_H > \max\{0, -\lambda_{\min}(\tilde{\mathsf{H}}_\Delta)\}$, the matrix $\hat{\mathsf{H}}$ is positive definite and one can apply the active set method developed in [68] which is specially tailored for problems of this type.

To obtain an approximation $\tilde{y}^*$ of the solution y^* to problem (4.2.1), we choose $p_j^{F^{-1}} \in X$ and $\varepsilon_H \geq 0$, solve problem (4.2.5) which yields the solution $\tilde{\lambda}^*$ and set $\tilde{y}^* = x_{\mathrm{SI}} - \sum_{i \in I} \tilde{\lambda}_i^* p_i^{F^{-1}}$. The following theorem gives an error bound on $\Psi(\tilde{y}^*) - \Psi(y^*)$.

LEMMA 4.2.3. *Let $p_j^{F^{-1}} \in X$, $j \in I$, be arbitrary. Let y^* be the solution of problem* (4.2.1)*, let $\tilde{\lambda}^*$ be the solution of problem* (4.2.5) *and define $\tilde{d}^* := \sum_{j \in I} \tilde{\lambda}_j^* p_j^{F^{-1}} \in X$, $\tilde{y}^* := x_{SI} - \tilde{d}^* \in X$ and $\mathring{d}' := \sum_{i \in I} \tilde{\lambda}_i^* p_i' \in X^*$. Then it holds*

$$\begin{aligned} \Psi(\tilde{y}^*) - \Psi(y^*) \leq \tfrac{1}{2}\langle F\tilde{d}^* - \mathring{d}', \tilde{d}^* \rangle_{X^*,X} + \tfrac{\varepsilon_H}{2}(1 - \|\tilde{\lambda}^*\|_2^2) \\ + \tfrac{1}{2}\max_{i,j \in I}\langle p_i', p_j^{F^{-1}} - F^{-1}p_j' \rangle_{X^*,X} + \max_{i,j \in I}\langle p_i', F^{-1}p_j' - p_j^{F^{-1}} \rangle_{X^*,X}. \end{aligned}$$

Proof. We split $\Psi(\tilde{y}^*) - \Psi(y^*) = \big(\Psi(\tilde{y}^*) - \tilde{\Psi}_\lambda(\tilde{\lambda}^*)\big) + \big(\tilde{\Psi}_\lambda(\tilde{\lambda}^*) - \Psi_\lambda(\lambda^*)\big)$ and estimate both differences. Denoting $\mathring{H} := \tilde{\mathsf{H}} + \varepsilon_H I_{n_p} - \mathsf{H}$, we find

$$\tilde{\Psi}_\lambda(\tilde{\lambda}^*) - \Psi_\lambda(\lambda^*) \le \tilde{\Psi}_\lambda(\lambda^*) - \Psi_\lambda(\lambda^*) = \tfrac{1}{2}\lambda^{*\top}\mathring{H}\lambda^* + \max_{j\in I}(q - \tilde{\mathsf{H}}\lambda^*)_j - \max_{j\in I}(q - \mathsf{H}\lambda^*)_j.$$

Since $\lambda^* \in \Lambda$ it holds

$$\lambda_m^* \mathring{H}_{m,n} \lambda_n^* \le \lambda_m^* \lambda_n^* \max_{i,j\in I} \mathring{H}_{i,j} \qquad \text{for all } m,n \in I,$$

which shows that

$$\lambda^{*\top}\mathring{H}\lambda^* = \sum_{m,n\in I} \lambda_m^* \mathring{H}_{m,n}\lambda_n^* \le \max_{i,j\in I} \mathring{H}_{i,j} \sum_{m\in I}\lambda_m^* \sum_{n\in I}\lambda_n^* = \max_{i,j\in I}\mathring{H}_{i,j}.$$

Furthermore,

$$\max_{j\in I}(q - \tilde{\mathsf{H}}\lambda^*)_j = \max_{j\in I}\big(((\mathsf{H}-\tilde{\mathsf{H}})\lambda^*)_j + (q - \mathsf{H}\lambda^*)_j\big) \le \max_{j\in I}((\mathsf{H}-\tilde{\mathsf{H}})\lambda^*)_j + \max_{j\in I}(q - \mathsf{H}\lambda^*)_j.$$

Combining this yields

$$\tilde{\Psi}_\lambda(\tilde{\lambda}^*) - \Psi_\lambda(\lambda^*) \le \tfrac{1}{2}\max_{i,j\in I}\mathring{H}_{i,j} + \max_{j\in I}((\mathsf{H}-\tilde{\mathsf{H}})\lambda^*)_j \le \tfrac{1}{2}\max_{i,j\in I}(\tilde{\mathsf{H}}-\mathsf{H})_{i,j} + \tfrac{\varepsilon_H}{2} + \max_{i,j\in I}(\mathsf{H}-\tilde{\mathsf{H}})_{i,j},$$

and it holds $(\mathsf{H}-\tilde{\mathsf{H}})_{i,j} = \langle p_i', F^{-1}p_j' - p_j^{F^{-1}}\rangle_{X^*,X}$ for $i,j \in I$. Furthermore, the definitions of Ψ and $\tilde{\Psi}_\lambda$ give

$$\Psi(\tilde{y}^*) = \Psi(x_{\mathrm{SI}} - \sum_{j\in I}\tilde{\lambda}_j^* p_j^{F^{-1}}) = \tfrac{1}{2}\sum_{i,j\in I}\tilde{\lambda}_i^*\langle F p_i^{F^{-1}}, p_j^{F^{-1}}\rangle_{X^*,X}\tilde{\lambda}_j^* + \max_{j\in I}(q - \tilde{\mathsf{H}}\tilde{\lambda}^*)_j$$

and

$$\tilde{\Psi}_\lambda(\tilde{\lambda}^*) = \tfrac{1}{2}\sum_{i,j\in I}\tilde{\lambda}_i^*\langle p_i', p_j^{F^{-1}}\rangle_{X^*,X}\tilde{\lambda}_j^* + \tfrac{\varepsilon_H}{2}\|\tilde{\lambda}^*\|_2^2 + \max_{j\in I}(q - \tilde{\mathsf{H}}\tilde{\lambda}^*)_j.$$

Therefore, we find

$$\begin{aligned}\Psi(\tilde{y}^*) - \tilde{\Psi}_\lambda(\tilde{\lambda}^*) &= \tfrac{1}{2}\sum_{i,j\in I}\tilde{\lambda}_i^*\langle F p_i^{F^{-1}} - p_i', p_j^{F^{-1}}\rangle_{X^*,X}\tilde{\lambda}_j^* - \tfrac{\varepsilon_H}{2}\|\tilde{\lambda}^*\|_2^2 \\ &= \tfrac{1}{2}\langle F\tilde{d}^* - \mathring{d}', \tilde{d}^*\rangle_{X^*,X} - \tfrac{\varepsilon_H}{2}\|\tilde{\lambda}^*\|_2^2.\end{aligned}$$

□

To turn this theorem into a computable error bound, one has to choose $p_j^{F^{-1}} \in X$, $j \in I$, such that bounds on the errors $\|F^{-1}p_j' - p_j^{F^{-1}}\|_X$ and $\|F\tilde{d}^* - \mathring{d}'\|_{X^*}$ can be computed.

Assumption 4.2.4. For the approximations $p_j^{F^{-1}} \in X,\ j \in I$, there exist corresponding error estimates $e_{i,j,F^{-1}} \in \mathbb{R}$, $i,j \in I$, and $e_F \in \mathbb{R}$ fulfilling

$$|\langle p_i', F^{-1}p_j' - p_j^{F^{-1}}\rangle_{X^*,X}| \le e_{i,j,F^{-1}} \qquad \text{and} \qquad \langle F\tilde{d}^* - \mathring{d}', \tilde{d}^*\rangle_{X^*,X} \le e_F.$$

Remark 4.2.5. As $\tilde{d}^* = \sum_{j\in I} \tilde{\lambda}_j^* p_j^{F^{-1}} \in X$ and $\mathring{d}' := \sum_{i\in I} \tilde{\lambda}_i^* p_i' \in X^*$, we find

$$\langle F\tilde{d}^* - \mathring{d}', \tilde{d}^* \rangle_{X^*,X} \le \|F\tilde{d}^* - \mathring{d}'\|_{X^*} \|\tilde{d}^*\|_X \le \|\tilde{d}^*\|_X \sum_{j\in I} \tilde{\lambda}_j^* \|F p_j^{F^{-1}} - p_j'\|_{X^*}.$$

Thus, if $p_j^{F^{-1}}$ is a sufficiently accurate approximation of $F^{-1}p_j'$, then the values of e_F and $e_{i,j,F^{-1}}$ can simultaneously be chosen small.

In the following, a convex optimization problem with a μ-strongly convex objective function, $\mu > 0$, is called a μ-strongly convex optimization problem.

LEMMA 4.2.6. *Choose a safeguarding parameter $\varepsilon_s > 0$. Then the approximated bundle subproblem* (4.2.6) *with $\varepsilon_H := \max\{0, \varepsilon_s - \lambda_{min}(\tilde{\mathsf{H}}_\Delta)\}$ is ε_s-strongly convex. Let Assumption 4.2.4 hold true and define*

$$e_\lambda := \begin{cases} 0 & \text{if } \lambda_{min}(\tilde{\mathsf{H}}_\Delta) \ge \varepsilon_s, \\ \frac{1}{2}\big(\varepsilon_s + \frac{1}{2}\max_{i\in I}\sum_{j\in I}(e_{i,j,F^{-1}} + e_{j,i,F^{-1}})\big)(1 - \|\tilde{\lambda}^*\|_2^2) & \text{else.} \end{cases} \tag{4.2.7}$$

Then it holds $\Psi(\tilde{y}^*) - \min_{y\in X}\Psi(y) \le \frac{1}{2}e_F + \frac{3}{2}\max_{i,j\in I} e_{i,j,F^{-1}} + e_\lambda$.

Proof. Since $\hat{\mathsf{H}}$ is defined by $\hat{\mathsf{H}} = \tilde{\mathsf{H}}_\Delta + \varepsilon_H I_{n_p}$, where $I_{n_p} \in \mathbb{R}^{n_p\times n_p}$ is the identity matrix, it is easy to see that the given choice of ε_H implies that $\hat{\mathsf{H}}$ has smallest eigenvalue ε_s. Therefore, the problem (4.2.6) is ε_s-strongly convex. In the case that $\varepsilon_H = 0$, Lemma 4.2.3 shows

$$\Psi(\tilde{y}^*) - \Psi(y^*) \le \tfrac{1}{2}e_F + \tfrac{3}{2}\max_{i,j\in I} e_{i,j,F^{-1}} \le \tfrac{1}{2}e_F + \tfrac{3}{2}\max_{i,j\in I} e_{i,j,F^{-1}} + e_\lambda.$$

In the case that $\varepsilon_H > 0$, i.e., $\varepsilon_H = \varepsilon_s - \lambda_{\min}(\tilde{\mathsf{H}}_\Delta)$, we proceed as follows. Since the matrix $\mathsf{H} = (\langle p_i', F^{-1}p_j'\rangle_{X^*,X})_{i,j}$ is positive semidefinite (cf. (4.2.4)), we find both

$$\lambda_{\min}(\tilde{\mathsf{H}}_\Delta - \mathsf{H}) = \min_{\|\lambda\|_2=1} \lambda^\top(\tilde{\mathsf{H}}_\Delta - \mathsf{H})\lambda \le \min_{\|\lambda\|_2=1} \lambda^\top \tilde{\mathsf{H}}_\Delta \lambda = \lambda_{\min}(\tilde{\mathsf{H}}_\Delta)$$

and

$$\varepsilon_H = \varepsilon_s - \lambda_{\min}(\tilde{\mathsf{H}}_\Delta) \le \varepsilon_s - \lambda_{\min}(\tilde{\mathsf{H}}_\Delta - \mathsf{H}). \tag{4.2.8}$$

By the Gershgorin circle theorem [37, Thm. 2], for any symmetric matrix $A \in \mathbb{R}^{n_p\times n_p}$ it holds that

$$-\lambda_{\min}(A) = \lambda_{\max}(-A) \in \bigcup_{i\in I}\Big[-A_{i,i} - \sum_{j\in I, j\ne i}|A_{i,j}|, -A_{i,i} + \sum_{j\in I, j\ne i}|A_{i,j}|\Big],$$

which shows $-\lambda_{\min}(A) \le \max_{i\in I}\sum_{j\in I}|A_{i,j}|$. Consequently, by Assumption 4.2.4,

$$\begin{aligned} -\lambda_{\min}(\tilde{\mathsf{H}}_\Delta - \mathsf{H}) &\le \max_{i\in I}\sum_{j\in I}\tfrac{1}{2}\big|\langle p_i', F^{-1}p_j' - p_j^{F^{-1}}\rangle_{X^*,X} + \langle p_j', F^{-1}p_i' - p_i^{F^{-1}}\rangle_{X^*,X}\big| \\ &\le \tfrac{1}{2}\max_{i\in I}\sum_{j\in I} e_{i,j,F^{-1}} + e_{j,i,F^{-1}}. \end{aligned}$$

Combining this with Lemma 4.2.3 and (4.2.8) shows

$$\Psi(\tilde{y}^*) - \Psi(y^*) \leq \tfrac{1}{2} e_F + \tfrac{3}{2} \max_{i,j \in I} e_{i,j,F^{-1}} + \tfrac{1}{2}(\varepsilon_s + \tfrac{1}{2} \max_{i \in I} \sum_{j \in I} e_{i,j,F^{-1}} + e_{j,i,F^{-1}})(1 - \|\tilde{\lambda}^*\|_2^2). \qquad \square$$

In order to construct an executable algorithm, we need the following assumption.

Assumption 4.2.7. Define the set $P := \{p_i', i \in I\} \subset X^*$. We assume that there is a computable oracle $O_{F^{-1}} : P \times (0,1] \to X$, which, given $p_i' \in P$ and an accuracy level $h \in (0,1]$, returns an approximation $p_i^{F^{-1}} := O_{F^{-1}}(p_i', h)$ of $F^{-1} p_i'$ for all $i \in I$. We assume that there exist corresponding computable error bounds $\hat{\varepsilon}_{F^{-1}} : P \times P \times (0,1] \to (0,\infty)$ and $\hat{\varepsilon}_F : \mathrm{co}(\{p_i^{F^{-1}}, i \in I\}) \times \mathrm{co}(P) \times (0,1] \to (0,\infty)$ and that there exist constants $C_F > 0$ and $C_{F^{-1}} > 0$ such that for all $p_i, p_j \in P$ it holds

$$|\langle p_i', F^{-1} p_j' - p_j^{F^{-1}} \rangle_{X^*,X}| \leq C_{F^{-1}} \hat{\varepsilon}_{F^{-1}}(p_i', p_j', h), \qquad \hat{\varepsilon}_{F^{-1}}(p_i', p_j', h) \to 0 \text{ as } h \to 0$$

and for arbitrary $\lambda \in \Lambda := \{\lambda \in [0,\infty)^{n_p} : \sum_{j=1}^{n_p} \lambda_j = 1\}$ with $\tilde{d}^* := \sum_j \lambda_j p_j^{F^{-1}}$, $\mathring{d}' := \sum_j \lambda_j p_j'$ it holds

$$\langle F \tilde{d}^* - \mathring{d}', \tilde{d}^* \rangle_{X^*,X} \leq C_F \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h), \qquad \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) \to 0 \text{ as } h \to 0.$$

We further assume that the constants C_F and $C_{F^{-1}}$ do not depend on F, p_i', $p_i^{F^{-1}}$, $i \in I$, and h.

Now, Lemma 4.2.6 motivates Algorithm 4.2.

THEOREM 4.2.8. *Assume that the oracle $O_{F^{-1}}$ and the error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ fulfill Assumption 4.2.7. Further suppose that the terms $\langle p_i', p_j^{F^{-1}} \rangle_{X^*,X}$, $i, j \in I$ in line 3 of Algorithm 4.2 can be computed exactly. Then Algorithm 4.2 can be executed. Denote by $\tilde{y}^{*h}$ and e_y^h the output of Algorithm 4.2 for accuracy $h \in (0,1]$ and safeguarding parameter ε_s^h. Then it holds*

$$\Psi(\tilde{y}^{*h}) - \min_{y \in X} \Psi(y) \leq \max\{C_F, C_{F^{-1}}, 1\} e_y^h.$$

If the safeguarding parameters $(\varepsilon_s^h)_h \subset (0,\infty)$ are chosen such that $\varepsilon_s^h \to 0$ as $h \to 0$, then $e_y^h \to 0$ as $h \to 0$. In particular, $O_s(\hat{\Psi}, h) := \tilde{y}^{*h}$, $\hat{\varepsilon}_s(\hat{\Psi}, h) := e_y^h$ and $C_{\hat{\Psi}} := \max\{C_F, C_{F^{-1}}, 1\}$ fulfill (4.1.3) *and thus can be used as a subproblem oracle with corresponding error estimates.*

Proof. By Lemma 4.2.6, we find

$$\Psi(\tilde{y}^{*h}) - \min_{y \in X} \Psi(y) \leq \tfrac{1}{2} e_F + \tfrac{3}{2} \max_{i,j \in I} e_{i,j,F^{-1}} + e_\lambda = \tfrac{1}{2} C_F \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) + \tfrac{3}{2} C_{F^{-1}} \max_{i,j \in I} \hat{\varepsilon}_{F^{-1}}(p_i', p_j', h) + \tilde{e}_\lambda,$$

where

$$\tilde{e}_\lambda = \begin{cases} 0 & \text{if } \lambda_{\min}(\tilde{\mathsf{H}}_\Delta) \geq \varepsilon_s^h, \\ \tfrac{1}{2}\big(\varepsilon_s^h + C_{F^{-1}} \tfrac{1}{2} \max_{i \in I} \sum_{j \in I} (\hat{\varepsilon}_{F^{-1}}(p_i', p_j', h) + \hat{\varepsilon}_{F^{-1}}(p_j', p_i', h))\big)(1 - \|\tilde{\lambda}^*\|_2^2) & \text{else.} \end{cases}$$

This readily shows $\Psi(\tilde{y}^{*h}) - \min_{y \in X} \Psi(y) \leq \max\{C_F, C_{F^{-1}}, 1\} e_y^h$. $\square$

Algorithm 4.2: Solving the reduced bundle subproblem $[(\tilde{y}^{*h}, e_y^h) := \texttt{SolveSubproblem}(h, \varepsilon_s^h)]$

Parameters: Oracle $O_{F^{-1}}$ and error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7. Serious iterate $x_{\mathrm{SI}} \in X$, subgradients $g'_j \in X^*$, $j \in I = \{1, \dots, n_p\}$, $n_p \in \mathbb{N}_+$.
Input : Accuracy parameter h. Safeguarding parameter $\varepsilon_s^h > 0$
Output : Approximate solution of the reduced bundle subproblem $\tilde{y}^{*h}$ with error estimate e_y^h.

Set $p'_j := g'_j + \alpha R_X x_{\mathrm{SI}} \in X^*$, $j \in I$.
Compute $p_j^{F^{-1}} := O_{F^{-1}}(p'_j, h)$ and $e_{i,j,F^{-1}} := \hat{\varepsilon}_{F^{-1}}(p'_i, p'_j, h)$ for $i, j \in I$.
Set $\tilde{\mathsf{H}}_{i,j} := \langle p'_i, p_j^{F^{-1}} \rangle_{X^*,X}$, $i, j \in I$, $\tilde{\mathsf{H}}_\Delta := \frac{1}{2}(\tilde{\mathsf{H}} + \tilde{\mathsf{H}}^\top)$, $\varepsilon_H := \max\{0, \varepsilon_s^h - \lambda_{\min}(\tilde{\mathsf{H}}_\Delta)\}$ and $\hat{\mathsf{H}} := \tilde{\mathsf{H}}_\Delta + \varepsilon_H I_{n_p}$.
Compute $(\tilde{\lambda}^*, m^*)$ via

$$\begin{aligned} (\tilde{\lambda}^*, m^*) := \operatorname*{arg\,min}_{\lambda, m} \quad & \tfrac{1}{2}\lambda^\top \hat{\mathsf{H}} \lambda + m \\ \text{subject to} \quad & (q - \tilde{\mathsf{H}}\lambda)_j \le m, \qquad j \in I \\ & \lambda \in \Lambda, m \in \mathbb{R}. \end{aligned}$$

Set $\tilde{d}^* := \sum_{j \in I} \tilde{\lambda}_j^* p_j^{F^{-1}}$ and $\mathring{d}' := \sum_{i \in I} \tilde{\lambda}_i^* p'_i \in X^*$ and compute $e_F := \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h)$.
if $\lambda_{min}(\tilde{\mathsf{H}}_\Delta) \ge \varepsilon_s^h$ **then**
  Set $e_\lambda := 0$.
else
  Set $e_\lambda := \frac{1}{2}\big(\varepsilon_s^h + \frac{1}{2}\max_{i \in I} \sum_{j \in I}(e_{i,j,F^{-1}} + e_{j,i,F^{-1}})\big)(1 - \|\tilde{\lambda}^*\|_2^2)$.
end
Set $\tilde{y}^{*h} := x_{\mathrm{SI}} - \tilde{d}^*$ and $e_y^h := \frac{1}{2} e_F + \frac{3}{2}\max_{i,j \in I} e_{i,j,F^{-1}} + e_\lambda$.

4.3. Discretization

In the case that X and Y are infinite dimensional Hilbert spaces, we cannot work with arbitrary elements $x \in X$ because this would need an infinite amount of storage. In this section we employ a discretization which facilitates the development of computable oracles $O_{F^{-1}}$ and error estimates $\hat{\varepsilon}_F$, $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7. To do so, let $Y^{*h} \subset Y^*$, $h > 0$, be a finite dimensional linear subspace of Y^* equipped with the same inner product as Y^*. Then by [129, Cor. 5.25.10], Y^{*h} is a Hilbert space. Since Y^{*h} is closed, the projection operator $P_{Y^{*h}} \in \mathscr{L}(Y^*)$ (cf. Section 2.1.2) is well-defined. Further, we define the finite dimensional Hilbert spaces $X^{*h} := (\iota^* Y^{*h}, (\cdot,\cdot)_{X^*})$, $X^h := (R_X^{-1}\iota^* Y^{*h}, (\cdot,\cdot)_X)$ and $Y^h := (\iota R_X^{-1} \iota^* Y^{*h}, (\cdot,\cdot)_Y)$. We work under the following assumption:

Assumption 4.3.1. The quantities $(x^h, y^h)_{Y^*}$ and $(\iota^* x^h, \iota^* y^h)_{X^*}$ can be computed exactly for all $x^h, y^h \in Y^{*h}$.

Remark 4.3.2. In the case that $Y^* = H_0^1(\Omega)$ and $X^* = L^2(\Omega)$, one can choose Y^{*h} as a space of finite element functions, cf. Section 5.3. Then, the quantities $(x^h, y^h)_{Y^*} = (x^h, y^h)_{H_0^1(\Omega)}$ and $(\iota^* x^h, \iota^* y^h)_{X^*} = (\iota^* x^h, \iota^* y^h)_{L^2(\Omega)}$ can be computed exactly for arbitrary finite element functions $x^h, y^h \in Y^{*h}$.

Remark 4.3.3. Note that the space $X^{*h} := (\iota^* Y^{*h}, (\cdot,\cdot)_{X^*})$ might not be a natural discretization of the space X^*. For example, in the case $Y^* = H_0^1(\Omega)$ and $X^* = L^2(\Omega)$, every function in $Y^* = H_0^1(\Omega)$ is equals to zero on the boundary of Ω (in the sense of traces). Therefore, X^{*h} only contains functions which are zero on the boundary, whereas $X^* = L^2(\Omega)$ contains functions which are not equals to zero on the boundary of Ω. However, for our purposes it suffices to calculate an approximation of F and F^{-1}. This can be done with the given discretization because it holds $Fx^h \in X^{*h}$ and $F^{-1}R_X^{-1}x^h \in X^h$ for all $x^h \in X^h$, cf. (4.3.3) and (4.3.4).

Remark 4.3.4. The choice of the subspace Y^{*h} plays a vital role in the quality of the approximation. In Sections 5.4 and 5.5 we discuss different strategies for choosing Y^{*h}.

4.3.1. No curvature information $Q = 0$

In this subsection we develop suitable approximations and error estimates to fulfill Assumption 4.2.4 for the case of no curvature $Q = Q_0 = 0$. As $Q_0 = 0$ fulfills the assumptions of Section 3.1.6 (cf. Remark 3.1.7), Lemma 4.2.1 implies that the operator $F_0 := \alpha R_X + \iota^*(Q_0 + \tau R_Y)\iota = \alpha R_X + \tau\iota^* R_Y \iota$ is invertible. We define the operators $\tilde{E}, \tilde{D}_\tau \in \mathscr{L}(Y^*)$ by $\tilde{E} := R_Y \iota R_X^{-1}\iota^*$ and $\tilde{D}_\tau := \alpha \operatorname{Id}_{Y^*} + \tau\tilde{E}$, respectively. Here and in the following, the tilde symbol over an operator indicates that the operator is an element of $\mathscr{L}(Y^*)$. Using the identities of Section 2.1, we find for all $x', y' \in Y^*$ that

$$(\tilde{E}x', y')_{Y^*} = (R_Y \iota R_X^{-1}\iota^* x', y')_{Y^*} = \langle y', \iota R_X^{-1}\iota^* x'\rangle_{Y^*,Y} = \langle \iota^* y', R_X^{-1}\iota^* x'\rangle_{X^*,X} = (\iota^* y', \iota^* x')_{X^*} \tag{4.3.1}$$

and

$$(\tilde{D}_\tau x', y')_{Y^*} = \alpha(x', y')_{Y^*} + \tau(\tilde{E}x', y')_{Y^*} = \alpha(x', y')_{Y^*} + \tau(\iota^* x', \iota^* y')_{X^*}. \tag{4.3.2}$$

In particular, this shows that $\tilde{E}$ and $\tilde{D}_\tau$ are Hilbert space self-adjoint operators according to Definition 2.1.2. Furthermore,

$$F_0 R_X^{-1}\iota^* = \alpha\iota^* + \tau\iota^* R_Y \iota R_X^{-1}\iota^* = \iota^*(\alpha \operatorname{Id}_{Y^*} + \tau R_Y \iota R_X^{-1}\iota^*) = \iota^*\tilde{D}_\tau. \tag{4.3.3}$$

The definition of $\tilde{D}_\tau$ and (4.3.1) yield $(\tilde{D}_\tau x', x')_{Y^*} = \alpha\|x'\|_{Y^*}^2 + \tau\|\iota^* x'\|_{X^*}^2 \geq \alpha\|x'\|_{Y^*}^2$ for all $x' \in Y^*$. Therefore, the bilinear form $(\tilde{D}_\tau \cdot, \cdot)_{Y^*}$ is coercive and Corollary 2.3.3 yields that $\tilde{D}_\tau \in \mathscr{L}(Y^*)$ is invertible. Thus, we find

$$F_0^{-1}\iota^* = F_0^{-1}\iota^*\tilde{D}_\tau\tilde{D}_\tau^{-1} = R_X^{-1}\iota^*\tilde{D}_\tau^{-1}. \tag{4.3.4}$$

For arbitrary $y' \in Y^*$, $z' := \tilde{D}_\tau^{-1} y' \in Y^*$ is characterized by the variational equation

$$\text{Find } z' \in Y^*: \qquad \alpha(z', w')_{Y^*} + \tau(\iota^* z', \iota^* w')_{X^*} = (y', w')_{Y^*} \qquad \text{for all } w' \in Y^*. \tag{4.3.5}$$

Consequently, for $v, w \in X^h$, i.e., $v = R_X^{-1}\iota^* v^h$, $w = R_X^{-1}\iota^* w^h$ with $v^h, w^h \in Y^{*h}$, we compute

$$\begin{aligned}
(\iota v, \iota w)_{Q+\tau R_Y} &= (\iota v, \iota w)_{\tau R_Y} = \tau\langle R_Y \iota R_X^{-1}\iota^* v^h, \iota R_X^{-1}\iota^* w^h\rangle_{Y^*,Y} = \tau(\iota^*\tilde{E}v^h, \iota^* w^h)_{X^*}, \\
\langle F_0 v, w\rangle_{X^*,X} &= (F_0 R_X^{-1}\iota^* v^h, \iota^* w^h)_{X^*} = (\iota^*\tilde{D}_\tau v^h, \iota^* w^h)_{X^*},
\end{aligned}$$

and for $v' := \iota^* v^h \in X^{*h}$, $w' := \iota^* w^h \in X^{*h}$ we get

$$\langle v', F_0^{-1} w' \rangle_{X^*,X} = \langle \iota^* v^h, F_0^{-1} \iota^* w^h \rangle_{X^*,X} = (\iota^* v^h, \iota^* \tilde{D}_\tau^{-1} w^h)_{X^*}.$$

In the following we study how to approximate the operators $\tilde{E}$ and $\tilde{D}_\tau^{-1} \in \mathscr{L}(Y^*)$.

The operator $\tilde{E}^h$

We approximate $\tilde{E}$ by $\tilde{E}^h := P_{Y^{*h}} \tilde{E}|_{Y^{*h}} \in \mathscr{L}(Y^{*h})$, where $P_{Y^{*h}} \in \mathscr{L}(Y^*)$ is the orthogonal projection onto the closed linear subspace Y^{*h}, cf. Section 2.1.2. By Lemma 2.1.7 and (4.3.1), for arbitrary $z^h \in Y^{*h}$, $\tilde{E}^h z^h \in Y^{*h}$ is characterized via the variational equation

$$(\tilde{E}^h z^h, w^h)_{Y^*} = (P_{Y^{*h}} \tilde{E} z^h, w^h)_{Y^*} = (\tilde{E} z^h, w^h)_{Y^*} = (\iota^* z^h, \iota^* w^h)_{X^*} \quad \text{for all } w^h \in Y^{*h}. \tag{4.3.6}$$

By Assumption 4.3.1, we can compute $(\cdot,\cdot)_{Y^*}$ and $(\iota^* \cdot, \iota^* \cdot)_{X^*}$ exactly for arguments from Y^{*h}. Thus, we can evaluate the operator $\tilde{E}^h$ exactly by solving a finite dimensional system of linear equations.

The operator $\tilde{D}_\tau^h$

We approximate $\tilde{D}_\tau$ by $\tilde{D}_\tau^h := \alpha \operatorname{Id}_{Y^{*h}} + \tau \tilde{E}^h \in \mathscr{L}(Y^{*h})$. Contrary to $\tilde{E}$, we cannot approximate $\tilde{D}_\tau^{-1}$ via $P_{Y^{*h}} \tilde{D}_\tau^{-1}$ since $P_{Y^{*h}} \tilde{D}_\tau^{-1}$ cannot be evaluated in the same manner as $P_{Y^{*h}} \tilde{E}$. Instead, we proceed as follows: From the fact $(\tilde{D}_\tau^h x^h, x^h)_{Y^*} = \alpha \|x^h\|_{Y^*}^2 + \tau \|\iota^* x^h\|_{X^*}^2 \geq \alpha \|x^h\|_{Y^*}^2$ for all $x^h \in Y^*$, we infer that the bilinear form $(\tilde{D}_\tau^h \cdot, \cdot)_{Y^*} : Y^{*h} \times Y^{*h} \to \mathbb{R}$ is coercive (in Y^{*h}). Thus, Corollary 2.3.3 implies that $\tilde{D}_\tau^h \in \mathscr{L}(Y^{*h})$ is invertible. We approximate $\tilde{D}_\tau^{-1}$ via $\tilde{D}_\tau^{-h} := (\tilde{D}_\tau^h)^{-1}$. Note that for arbitrary $y^h \in Y^{*h}$, $z^h := \tilde{D}_\tau^{-h} y^h \in Y^{*h}$ is characterized by the finite dimensional variational equation

$$\text{Find } z^h \in Y^{*h} : \qquad \alpha (z^h, w^h)_{Y^*} + \tau (\iota^* z^h, \iota^* w^h)_{X^*} = (y^h, w^h)_{Y^*} \qquad \text{for all } w^h \in Y^{*h} \tag{4.3.7}$$

and, under Assumption 4.3.1, we can compute $\tilde{D}_\tau^{-h} y^h$ exactly.

Error estimates for $Q = 0$

In the following, we work under the assumptions:

Assumption 4.3.5. The serious iterate x_{SI} is an element of X^h and the subgradients g'_j are elements of X^{*h} for all $j \in I$.

Assumption 4.3.6. For arbitrary $x^h, y^h \in Y^{*h}$, there exists an error estimate $e^E_{x^h,y^h} \geq 0$ fulfilling

$$(\iota^* (\tilde{E} - \tilde{E}^h) x^h, \iota^* y^h)_{X^*} \leq e^E_{x^h,y^h}.$$

Assumption 4.3.7. For arbitrary $x^h, y^h \in Y^{*h}$, there exists an error estimate $e^{D^{-1}}_{x^h,y^h} \geq 0$ fulfilling

$$|(\iota^* x^h, \iota^* (\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h}) y^h)_{X^*}| \leq e^{D^{-1}}_{x^h,y^h}. \tag{4.3.8}$$

Remark 4.3.8. Since both $\tilde{E}$ and $\tilde{D}_\tau^{-1}$ can be characterized by variational equations, several techniques for error estimates which fulfill the previous assumption are available, for example a priori or a posteriori error estimates (cf. Theorems 4.6.7 and 5.5.3).

LEMMA 4.3.9. *Let Assumption 4.3.5 hold such that $x_{SI} = R_X^{-1}\iota^* x_{SI}^h$, $g_j' = \iota^* g_j^h$ with $x_{SI}^h, g_j^h \in Y^{*h}$ and set $p_j^h := g_j^h + \alpha x_{SI}^h \in Y^{*h}$, $j \in I$. Further let Assumptions 4.3.6 and 4.3.7 hold and define $\tilde{d}^{*h} := \sum_{j\in I} \tilde{\lambda}_j^* \tilde{D}_\tau^{-h} p_j^h \in Y^{*h}$. Then the approximation $p_j^{F^{-1}} := R_X^{-1}\iota^*\tilde{D}_\tau^{-h} p_j^h$ and the error estimates $e_{i,j,F^{-1}} := e^{D^{-1}}_{p_i^h, p_j^h}$ and $e_F := \tau e^E_{\tilde{d}^{*h},\tilde{d}^{*h}}$, $i, j \in I$, fulfill Assumption 4.2.4.*

Proof. First note that $p_j' = g_j' + \alpha R_X x_{\mathrm{SI}} = \iota^* p_j^h$, $j \in I$. From (4.3.4), we deduce

$$|\langle p_i', F_0^{-1} p_j' - p_j^{F^{-1}} \rangle_{X^*,X}| = |(\iota^* p_i^h, \iota^*(\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h}) p_j^h)_{X^*}| \leq e^{D^{-1}}_{p_i^h,p_j^h} = e_{i,j,F^{-1}}.$$

Furthermore, the definition of $\tilde{d}^*$ (cf. Lemma 4.2.3) and (4.3.3) yield $\tilde{d}^* = R_X^{-1}\iota^*\tilde{d}^{*h} \in X^h$ and

$$F_0\tilde{d}^* = F_0 R_X^{-1}\iota^*\tilde{d}^{*h} = \iota^*\tilde{D}_\tau\tilde{d}^{*h}.$$

The definition of $\mathring{d}' \in X^*$ and $\tilde{D}_\tau^{-h} = (\tilde{D}_\tau^h)^{-1}$ give

$$\mathring{d}' = \sum_{j\in I} \tilde{\lambda}_j^* \iota^* \tilde{D}_\tau^h \tilde{D}_\tau^{-h} p_j^h = \iota^*\tilde{D}_\tau^h \tilde{d}^{*h}.$$

Therefore, we estimate

$$\langle F_0\tilde{d}^* - \mathring{d}', \tilde{d}^* \rangle_{X^*,X} = (\iota^*(\tilde{D}_\tau - \tilde{D}_\tau^h)\tilde{d}^{*h}, \iota^*\tilde{d}^{*h})_{X^*} = \tau(\iota^*(\tilde{E} - \tilde{E}^h)\tilde{d}^{*h}, \iota^*\tilde{d}^{*h})_{X^*} \leq \tau e^E_{\tilde{d}^{*h},\tilde{d}^{*h}} = e_F. \quad \square$$

4.4. L-BFGS curvature in Hilbert space

In smooth optimization, the BFGS formula is among the most successful ways to approximate the Hessian of a given function using only first-order derivative information. Also for nonsmooth optimization, it was successfully implemented, cf. [79]. For large-scale optimization problems however, it is not feasible to store the full BFGS curvature in memory. Instead, one can use the limited memory BFGS (L-BFGS) formula which was introduced in the finite dimensional setting in [96]. In this section we present a curvature operator Q_{BFGS} based on the L-BFGS formula in Hilbert space and develop error estimates for the corresponding operator $F_{\mathrm{BFGS}} := \alpha\,\mathrm{Id}_X + R_X^{-1}\iota^*(Q_{\mathrm{BFGS}} + \tau R_Y)\iota$ which fulfill Assumption 4.2.4.

Let Y be a Hilbert space, let $H_0 \in \mathscr{L}(Y,Y^*)$ be an operator and let $L \in \mathbb{N}_+$ be a given number. Denote by $L_{\mathrm{it}} \in \mathbb{N}_+$ the number of previous calculated trial iterates and set $\tilde{L} := \min\{L, L_{\mathrm{it}}\}$. Further denote by $y_l \in Y$ the $(\tilde{L}-l)^{\mathrm{th}}$ previous trial iterate and by $g_l' \in Y^*$ the corresponding subgradient, where $0 \leq l \leq \tilde{L}$. We define $d_l := y_{l+1} - y_l \in Y$ and $r_l' := g_{l+1}' - g_l' \in Y^*$. The BFGS update formula to determine $H_{l+1} \in \mathscr{L}(Y,Y^*)$, $0 \leq l < \tilde{L}$, is given by

$$H_{l+1} := H_l + \frac{\langle r_l', \cdot\rangle_{Y^*,Y}}{\langle r_l', d_l\rangle_{Y^*,Y}} r_l' - \frac{\langle H_l d_l, \cdot\rangle_{Y^*,Y}}{\langle H_l d_l, d_l\rangle_{Y^*,Y}} H_l d_l. \tag{4.4.1}$$

This gives rise to the L-BFGS curvature operator, which is given by $H_{\tilde{L}}$. Using this formula yields curvature operators with positive curvature:

LEMMA 4.4.1. *If $\langle H_0 \cdot, \cdot \rangle_{Y^*,Y}$ is a symmetric and positive bilinear form on Y and $\langle r_l', d_l \rangle_{Y^*,Y} > 0$ for all $l \in \mathbb{N}$, then $\langle H_l \cdot, \cdot \rangle_{Y^*,Y}$ is a symmetric and positive definite bilinear form for all $l \in \mathbb{N}$.*

Proof. The proof of this classical result follows the lines of [125, Thm. 13.4]. Proof by induction. The start for $l = 0$ is already given. Now assume this holds for $l \in \mathbb{N}$. Then the symmetric and positive definite bilinear form $\langle H_l \cdot, \cdot \rangle_{Y^*,Y}$ defines the inner product $(\cdot, \cdot)_{H_l}$ on Y. For arbitrary $y \in Y$, the Cauchy-Schwarz inequality yields

$$\frac{(d_l, y)_{H_l}^2}{(d_l, d_l)_{H_l}} \leq \frac{\|d_l\|_{H_l}^2 \|y\|_{H_l}^2}{\|d_l\|_{H_l}^2} = \|y\|_{H_l}^2,$$

and this inequality is strict if d_l and y are linearly independent. In this case we see

$$\langle H_{l+1} y, y \rangle_{Y^*,Y} = (y,y)_{H_l} + \frac{\langle r_l', y \rangle_{Y^*,Y}^2}{\langle r_l', d_l \rangle_{Y^*,Y}} - \frac{(d_l, y)_{H_l}^2}{(d_l, d_l)_{H_l}} > \frac{\langle r_l', y \rangle_{Y^*,Y}^2}{\langle r_l', d_l \rangle_{Y^*,Y}} \geq 0.$$

If d_l and y are not linearly independent and $y \neq 0$ there exists $t \in \mathbb{R} \setminus \{0\}$ such that $y = t d_l$ and we get

$$\langle H_{l+1} y, y \rangle_{Y^*,Y} = \frac{\langle r_l', t d_l \rangle_{Y^*,Y}^2}{\langle r_l', d_l \rangle_{Y^*,Y}} = t^2 \langle r_l', d_l \rangle_{Y^*,Y} > 0.$$

Since H_{l+1} is obviously symmetric, this shows that H_{l+1} is a symmetric and positive definite bilinear form. □

4.4.1. Approximate L-BFGS curvature

From now on we work under the assumption that the $(\tilde{L} - l)^{\text{th}}$ previous trial iterate y_l is an element of $Y^h = \iota R_X^{-1} \iota^* Y^{*h}$ and that the corresponding subgradient g_l^h is an element of Y^{*h} for $0 \leq l \leq \tilde{L}$ (cf. Section 4.3 for the definition of Y^h and Y^{*h}). Therefore, there exist elements $d_l^h \in Y^{*h}$ and $r_l^h \in Y^{*h}$ such that $d_l = y_{l+1} - y_l = \iota R_X^{-1} \iota^* d_l^h$ and $r_l^h = g_{l+1}^h - g_l^h$. Naturally, we choose $H_0 := \mu R_Y$, $\mu > 0$. If we use the exact L-BFGS formula (4.4.1) to assemble the curvature operator, then the evaluation of $H_{\tilde{L}}$ requires the exact computation of, e.g., $\langle H_0 d_0, d_0 \rangle_{Y^*,Y} = \mu \langle R_Y d_0, d_0 \rangle_{Y^*,Y} = \mu \|d_0\|_Y^2$. However, this might not be possible; under Assumption 4.3.1 only the approximation $\mu (\tilde{E}^h d_0^h, d_0^h)_{Y^*}$ can be computed exactly. Since the term $\langle H_0 d_0, d_0 \rangle_{Y^*,Y}$ appears in the denominator in the L-BFGS formula (4.4.1), error estimation may lead to a large error estimator even if the true error is small. Instead, we construct a curvature operator $Q_{\text{BFGS}} : Y \to Y^*$ such that we can compute numbers q_{BFGS} and $\bar{q}_{\text{BFGS}}$ which bound the curvature of Q_{BFGS} according to (3.1.17), the operator $F_{\text{BFGS}} := \alpha \operatorname{Id}_X + R_X^{-1} \iota^* (Q_{\text{BFGS}} + \tau R_Y) \iota$ is invertible with inverse F_{BFGS}^{-1} and both $(F_{\text{BFGS}} x, y)_X$ and $(x, F_{\text{BFGS}}^{-1} y)_X$ can be approximated efficiently for all $x, y \in X^h$. To do so, we define $H_0^h := \mu \tilde{E}^h \in \mathscr{L}(Y^{*h})$ and $H_{l+1}^h \in \mathscr{L}(Y^{*h})$, $0 \leq l < \tilde{L}$, by

$$H_{l+1}^h = H_l^h + \frac{(\iota^* r_l^h, \iota^* \cdot)_{X^*}}{(\iota^* r_l^h, \iota^* d_l^h)_{X^*}} r_l^h - \frac{(\iota^* H_l^h d_l^h, \iota^* \cdot)_{X^*}}{(\iota^* H_l^h d_l^h, \iota^* d_l^h)_{X^*}} H_l^h d_l^h. \tag{4.4.2}$$

In the case that $(\iota^* r_l^h, \iota^* d_l^h)_{X^*} = 0$ or $(\iota^* H_l^h d_l^h, \iota^* d_l^h)_{X^*} = 0$, we exclude r_l^h and d_l^h from the L-BFGS formula. Inspired by Lemma 4.4.1, if $(\iota^* r_l^h, \iota^* d_l^h)_{X^*} < 0$ we also exclude r_l^h and d_l^h from the L-BFGS formula. Define $u_l^h, v_l^h \in Y^{*h}$, $1 \leq l \leq 2\tilde{L}$, via

$$\begin{aligned} u^h_{2l+1} &:= r^h_l, & v^h_{2l+1} &:= \frac{r^h_l}{(\iota^* r^h_l, \iota^* d^h_l)_{X^*}}, & &\text{for } 0 \le l < \tilde{L}, \\ u^h_{2l+2} &:= H^h_l d^h_l, & v^h_{2l+2} &:= -\frac{H^h_l d_l}{(\iota^* H^h_l d^h_l, \iota^* d^h_l)_{X^*}} & &\text{for } 0 \le l < \tilde{L}. \end{aligned} \tag{4.4.3}$$

Note that (under Assumption 4.3.1) we can compute u^h_l and v^h_l exactly. Further define

$$V^h : Y \to \mathbb{R}^{2\tilde{L}},\ (V^h y)_l := \langle v^h_l, y\rangle_{Y^*,Y} \qquad \text{and} \qquad U^h : \mathbb{R}^{2\tilde{L}} \to Y^*,\ U^h\beta := \sum_{l=1}^{2\tilde{L}} u^h_l \beta_l. \tag{4.4.4}$$

Then the approximate L-BFGS operator is given by

$$Q_{\text{BFGS}} := \mu R_Y + \sum_{l=1}^{2\tilde{L}} u^h_l \langle v^h_l, \cdot\rangle_{Y^*,Y} = H_0 + U^h V^h. \tag{4.4.5}$$

The sole difference between the approximate L-BFGS operator Q_{BFGS} and the exact L-BFGS operator $H_{\tilde{L}}$ is that $H_l \iota R_X^{-1} \iota^* d^h_l$ and $\langle H_l \iota R_X^{-1} \iota^* d^h_l, \iota R_X^{-1} \iota^* d^h_l\rangle_{Y^*,Y}$ are replaced by $H^h_l d^h_l$ and $(\iota^* H^h_l d^h_l, \iota^* d^h_l)_{X^*}$, $0 \le l < \tilde{L}$, respectively.

4.4.2. The spectrum of Q_{BFGS}

To be able to use Q_{BFGS} as a curvature operator for the bundle method, we need to compute numbers $0 \le q_{\text{BFGS}} \le \bar{q}_{\text{BFGS}}$ such that (3.1.17) is fulfilled, i.e.,

$$\langle Q_{\text{BFGS}}\, v, v\rangle_{Y^*,Y} \ge -q_{\text{BFGS}} \|v\|_Y^2 \quad \text{for all } v \in Y \quad \text{and} \quad \|Q_{\text{BFGS}}\|_{\mathscr{L}(Y,Y^*)} \le \bar{q}_{\text{BFGS}}.$$

Sharp estimates for q_{BFGS} and $\bar{q}_{\text{BFGS}}$ can be obtained by computing the spectrum of $R_Y^{-1} U^h V^h \in \mathscr{L}(Y)$. As $R_Y^{-1} U^h V^h$ is an operator with low rank, it is more efficient to compute the spectrum of the operator $V^h R_Y^{-1} U^h \in \mathscr{L}(\mathbb{R}^{2\tilde{L}}, \mathbb{R}^{2\tilde{L}})$. Define the two numbers

$$q_{\text{BFGS}} := -\min\{0, \mu + \min \sigma(A)\}, \qquad \bar{q}_{\text{BFGS}} := \mu + \max |\sigma(A)|, \tag{4.4.6}$$

where $A \in \mathbb{R}^{2\tilde{L}\times 2\tilde{L}}$ is defined by $A_{k,l} := (u^h_k, v^h_l)_{Y^*}$ and $\sigma(A)$ defines the set of eigenvalues of A. Note that we can compute A and thus also q_{BFGS} and $\bar{q}_{\text{BFGS}}$ exactly.

THEOREM 4.4.2. *The L-BFGS operator $Q_{BFGS} \in \mathscr{L}(Y, Y^*)$ is symmetric and the numbers q_{BFGS} and $\bar{q}_{BFGS}$ fulfill* (3.1.17).

Proof. For all $x,y \in Y$ it holds

$$\begin{aligned}
\langle Q_{\mathrm{BFGS}}\, x,y\rangle_{Y^*,Y} &= \mu\langle R_Y x,y\rangle_{Y^*,Y} + \sum_{l=1}^{2\tilde{L}} \langle u_l^h,y\rangle_{Y^*,Y}\langle v_l^h,x\rangle_{Y^*,Y} \\
&= \mu(x,y)_Y + \sum_{l=1}^{2\tilde{L}} \frac{\langle r_l^h,y\rangle_{Y^*,Y}\langle r_l^h,x\rangle_{Y^*,Y}}{(\iota^* r_l^h,\iota^* d_l^h)_{X^*}} - \sum_{l=1}^{2\tilde{L}} \frac{\langle H_l^h d_l^h,y\rangle_{Y^*,Y}\langle H_l^h d_l^h,x\rangle_{Y^*,Y}}{(\iota^* H_l^h d_l^h,\iota^* d_l^h)_{X^*}} \\
&= \langle Q_{\mathrm{BFGS}}\, y,x\rangle_{Y^*,Y},
\end{aligned}$$

i.e., Q_{BFGS} is symmetric and $R_Y^{-1}Q_{\mathrm{BFGS}} \in \mathscr{L}(Y)$ is self-adjoint (cf. Lemma 2.1.3). Consequently, Theorem A.3 implies

$$\langle Q_{\mathrm{BFGS}}\, v,v\rangle_{Y^*,Y} = (R_Y^{-1}Q_{\mathrm{BFGS}}\, v,v)_Y \geq \Big(\mu + \min\sigma(V^h R_Y^{-1} U^h)\cup\{0\}\Big)\|v\|_Y^2 \quad \text{for all } v \in Y \tag{4.4.7}$$

and

$$\|Q_{\mathrm{BFGS}}\|_{\mathscr{L}(Y,Y^*)} = \|R_Y^{-1}Q_{\mathrm{BFGS}}\|_{\mathscr{L}(Y)} \leq \mu + \max|\sigma(V^h R_Y^{-1}U^h)|. \tag{4.4.8}$$

Since for arbitrary $\alpha,\beta \in \mathbb{R}^{2\tilde{L}}$ it holds

$$(V^h R_Y^{-1}U^h\alpha,\beta)_{\mathbb{R}^{2\tilde{L}}} = \sum_{l=1}^{2\tilde{L}}\langle v_l^h, R_Y^{-1}U^h\alpha\rangle_{Y^*,Y}\beta_l = \sum_{l=1}^{2\tilde{L}}\sum_{k=1}^{2\tilde{L}} \alpha_k (v_l^h,u_k^h)_{Y^*}\beta_l = \alpha^\top A b,$$

we deduce $\sigma(V^h R_Y^{-1}U^h) = \sigma(A)$. Discriminating the three cases $\min\sigma(A) \leq -\mu$, $-\mu < \min\sigma(A) \leq 0$ and $\min\sigma(A) > 0$, one finds that

$$\mu + \min\big\{\sigma(V^h R_Y^{-1}U^h)\cup\{0\}\big\} \geq \min\big\{0, \mu + \min\sigma(A)\cup\{0\}\big\} = \min\big\{0,\mu+\min\sigma(A)\big\}. \tag{4.4.9}$$

Combining (4.4.7) and (4.4.9) yields $\langle Q_{\mathrm{BFGS}}\, v,v\rangle_{Y^*,Y} \geq -q_{\mathrm{BFGS}}\|v\|_Y^2$ for all $v \in Y$ and combining (4.4.8) with $\sigma(V^h R_Y^{-1}U^h) = \sigma(A)$ yields $\|Q_{\mathrm{BFGS}}\|_{\mathscr{L}(Y,Y^*)} \leq \bar{q}_{\mathrm{BFGS}}$. □

For a given L-BFGS operator Q_{BFGS}, Theorem 4.4.2 enables us to find numbers $0 \leq q_{\mathrm{BFGS}} \leq \bar{q}_{\mathrm{BFGS}}$ which fulfill (3.1.17). To fulfill all requirements of Section 3.1.6, we need to make sure that $\bar{q}_{\mathrm{BFGS}} \leq \bar{q}$, where $\bar{q} \in (0,\infty)$ is a given constant. To ensure this, we change the L-BFGS formula whenever $\bar{q}_{\mathrm{BFGS}} > \bar{q}$, possibly by using $Q_{\mathrm{BFGS}} := H_{\tilde{L}-1}^h$ or by excluding r_0^h and d_0^h from the L-BFGS formula.

4.4.3. The operator F_{BFGS} and its inverse

Similar to F_0 in Section 4.3.1, in this section we define the operator $F = F_{\mathrm{BFGS}}$ for the case of BFGS curvature $Q = Q_{\mathrm{BFGS}}$. Recall that in Section 4.3.1 the operator $\tilde{D}_\tau = \alpha\,\mathrm{Id}_{Y^*} + \tau\tilde{E} = \alpha\,\mathrm{Id}_{Y^*} + \tau R_Y \iota R_X^{-1}\iota^* \in \mathscr{L}(Y^*)$ is introduced. We now define the operator

$$D_{\tau+\mu} := \alpha R_X + (\tau+\mu)\iota^* R_Y \iota \in \mathscr{L}(X,X^*).$$

If $\tau+\mu \geq 0$, it can be shown (cf. Section 4.3.1) that $D_{\tau+\mu}$ is invertible and it holds

$$\begin{aligned} D_{\tau+\mu}R_X^{-1}\iota^* &= \iota^*\tilde{D}_{\tau+\mu}, \\ D_{\tau+\mu}^{-1}\iota^* &= R_X^{-1}\iota^*\tilde{D}_{\tau+\mu}^{-1}. \end{aligned}$$

Also recall that the BFGS curvature is defined as $Q_{\text{BFGS}} = \mu R_Y + U^h V^h \in \mathscr{L}(Y,Y^*)$. We define

$$F_{\text{BFGS}} := \alpha R_X + \iota^*(Q_{\text{BFGS}} + \tau R_Y)\iota = D_{\tau+\mu} + \iota^* U^h V^h \iota \in \mathscr{L}(X,X^*)$$

for $\tau \geq q_{\text{BFGS}}$, where q_{BFGS} is defined in (4.4.6). By Theorem 4.4.2, it holds for all $x\in X$ that

$$\langle F_{\text{BFGS}}\, x,x\rangle_{X^*,X} = \alpha\|x\|_X^2 + \langle (Q+\tau R_Y)\iota x,\iota x\rangle_{Y^*,Y} \geq \alpha\|x\|_X^2 + (\tau - q_{\text{BFGS}})\|\iota x\|_Y^2 \geq \alpha\|x\|_X^2,$$

i.e., the bilinear form $\langle F_{\text{BFGS}}\,\cdot,\cdot\rangle_{X^*,X} : X\times X \to \mathbb{R}$ is coercive. Therefore, Corollary 2.3.3 shows that $F_{\text{BFGS}} \in \mathscr{L}(X,X^*)$ is invertible. The inverse of F_{BFGS} can be calculated using the following theorem:

THEOREM 4.4.3 (Sherman-Morrison-Woodbury formula). *Let X,Y be Hilbert spaces. Let $D\in\mathscr{L}(X)$ and $Z\in\mathscr{L}(Y)$ both be invertible and let $U\in\mathscr{L}(Y,X)$ and $V\in\mathscr{L}(X,Y)$. Then $D+UZV$ is invertible if and only if $Z^{-1}+VD^{-1}U$ is invertible and in this case it holds:*

$$(D+UZV)^{-1} = D^{-1} - D^{-1}U(Z^{-1}+VD^{-1}U)^{-1}VD^{-1}.$$

Proof. Simple calculation, see also [28]. □

COROLLARY 4.4.4. *Let X,Y be Hilbert spaces. Let $D\in\mathscr{L}(X,X^*)$ and $Z\in\mathscr{L}(Y)$ both be invertible and let $U\in\mathscr{L}(Y,X^*)$ and $V\in\mathscr{L}(X,Y)$. Then $D+UZV$ is invertible if and only if $Z^{-1}+VD^{-1}U$ is invertible and in this case it holds:*

$$(D+UZV)^{-1} = D^{-1} - D^{-1}U(Z^{-1}+VD^{-1}U)^{-1}VD^{-1}.$$

Proof. Apply Theorem 4.4.3 to $R_X^{-1}D + (R_X^{-1}U)ZV$. □

Corollary 4.4.4 yields that the inverse of $F_{\text{BFGS}} = D_{\tau+\mu} + \iota^*U^hV^h\iota \in \mathscr{L}(X,X^*)$ is given by

$$F_{\text{BFGS}}^{-1} = D_{\tau+\mu}^{-1} - D_{\tau+\mu}^{-1}\iota^*U^h(\text{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h\iota D_{\tau+\mu}^{-1}\iota^*U^h)^{-1}V^h\iota D_{\tau+\mu}^{-1} \tag{4.4.10}$$

and that the operator $\text{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h\iota D_{\tau+\mu}^{-1}\iota^*U^h \in \mathscr{L}(\mathbb{R}^{2\tilde{L}})$ is invertible for all $\tau \geq q_{\text{BFGS}}$.

4.4.4. Approximation of the inverse of F_{BFGS}

Note that we cannot compute $D_{\tau+\mu}^{-1}$ exactly. Consequently, $F_{\text{BFGS}}^{-1}p_j'$ cannot be evaluated exactly via (4.4.10) for arbitrary $p_j' \in X^*$. The goal of this section is to provide a computable approximation $p_j^{F_{\text{BFGS}}^{-1}} \in X$ of $F_{\text{BFGS}}^{-1}p_j'$ which can be used in Assumption 4.2.4. Recall that $p_j' = g_j' + \alpha R_X x_{\text{SI}} \in X^*$, $j\in I$, where $g_j'\in X^*$ are the approximated subgradients, and $x_{\text{SI}}\in X$ is the serious iterate (cf. Section 4.2). In the following, let Assumption 4.3.5 be fulfilled, i.e., there exist $p_i^h \in Y^{*h}$ such that $p_i' = \iota^* p_i^h$ for all $i\in I$. Further, define the operator

$$\tilde{F}_{\text{BFGS}} := \alpha\,\text{Id}_{Y^*} + (Q_{\text{BFGS}} + \tau R_Y)\iota R_X^{-1}\iota^* = \tilde{D}_{\tau+\mu} + U^hV^h\iota R_X^{-1}\iota^* \in \mathscr{L}(Y^*). \tag{4.4.11}$$

Note that although F_{BFGS} is symmetric, the operator $\tilde{F}_{\mathrm{BFGS}}$ might not be self-adjoint. Since the matrix $\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota D_{\tau+\mu}^{-1} \iota^* U^h$ is invertible (cf. (4.4.10)), also

$$\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-1} U^h = \mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota D_{\tau+\mu}^{-1} \iota^* U^h$$

is invertible. Therefore, the Sherman Morrison Woodbury formula (Theorem 4.4.3) implies that $\tilde{F}_{\mathrm{BFGS}}$ is invertible and the inverse is given by

$$\tilde{F}_{\mathrm{BFGS}}^{-1} = \tilde{D}_{\tau+\mu}^{-1} - \tilde{D}_{\tau+\mu}^{-1} U^h (\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-1} U^h)^{-1} V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-1}. \tag{4.4.12}$$

It is easy to see (cf. Section 4.3.1) that $F_{\mathrm{BFGS}} R_X^{-1} \iota^* = \iota^* \tilde{F}_{\mathrm{BFGS}}$ and that $F_{\mathrm{BFGS}}^{-1} \iota^* = R_X^{-1} \iota^* \tilde{F}_{\mathrm{BFGS}}^{-1}$. Therefore, it holds that $F_{\mathrm{BFGS}} R_X^{-1} p'_j = F_{\mathrm{BFGS}} R_X^{-1} \iota^* p_j^h = \iota^* \tilde{F}_{\mathrm{BFGS}} p_j^h$ and

$$F_{\mathrm{BFGS}}^{-1} p'_j = F_{\mathrm{BFGS}}^{-1} \iota^* p_j^h = R_X^{-1} \iota^* \tilde{F}_{\mathrm{BFGS}}^{-1} p_j^h \qquad \text{for all } j \in I. \tag{4.4.13}$$

Consequently, approximating $\tilde{F}_{\mathrm{BFGS}} p_j^h$ and $\tilde{F}_{\mathrm{BFGS}}^{-1} p_j^h$ readily yields approximations for $F_{\mathrm{BFGS}} R_X^{-1} p'_j$ and $F_{\mathrm{BFGS}}^{-1} p'_j$. For $x^h \in Y^{*h}$, we can evaluate $U^h V^h \iota R_X^{-1} \iota^* x^h$ exactly, but not $\tilde{D}_{\tau+\mu} x^h$. Thus, we approximate $\tilde{F}_{\mathrm{BFGS}} = \tilde{D}_{\tau+\mu} + U^h V^h \iota R_X^{-1} \iota^*$ via

$$\tilde{F}_{\mathrm{BFGS}}^h := P_{Y^{*h}} \tilde{F}_{\mathrm{BFGS}}|_{Y^{*h}} \in \mathscr{L}(Y^{*h}).$$

Equation (4.4.4) yields for arbitrary $x^h, y^h \in Y^{*h}$ that

$$\begin{aligned}(\tilde{F}_{\mathrm{BFGS}}^h x^h, y^h)_{Y^*} &= (P_{Y^{*h}} \tilde{F}_{\mathrm{BFGS}} x^h, y^h)_{Y^*} = (\tilde{F}_{\mathrm{BFGS}} x^h, y^h)_{Y^*} \\ &= \alpha (x^h, y^h)_{Y^*} + (\tau+\mu)(\iota^* x^h, \iota^* y^h)_{X^*} + \sum_{l=1}^{2\tilde{L}} (u_l^h, y^h)_{Y^*} (\iota^* v_l^h, \iota^* x^h)_{X^*}.\end{aligned} \tag{4.4.14}$$

Thus, under Assumption 4.3.1, $(\tilde{F}_{\mathrm{BFGS}}^h x^h, y^h)_{Y^*}$ can be computed exactly. Since $u_l^h \in Y^{*h}$, $1 \le l \le 2\tilde{L}$, the operator U^h maps to Y^{*h} and we find

$$\tilde{F}_{\mathrm{BFGS}}^h = P_{Y^{*h}} \tilde{F}_{\mathrm{BFGS}}|_{Y^{*h}} = \tilde{D}_{\tau+\mu}^h + U^h V^h \iota R_X^{-1} \iota^*|_{Y^{*h}}.$$

Because $\tilde{D}_{\tau+\mu}^h \in \mathscr{L}(Y^{*h})$, the Sherman Morrison Woodbury formula (Theorem 4.4.3) implies that the operator $\tilde{F}_{\mathrm{BFGS}}^h \in \mathscr{L}(Y^{*h})$ is invertible if and only if $\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} U^h$ is invertible and in this case it holds

$$\tilde{F}_{\mathrm{BFGS}}^{-h} = \tilde{D}_{\tau+\mu}^{-h} - \tilde{D}_{\tau+\mu}^{-h} U^h (\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} U^h)^{-1} V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h}. \tag{4.4.15}$$

Define the matrix $\tilde{A}^h \in \mathbb{R}^{2\tilde{L} \times 2\tilde{L}}$ by

$$\tilde{A}_{k,l}^h := (\iota^* \tilde{D}_{\tau+\mu}^{-h} u_k^h, \iota^* v_l^h)_{X^*}. \tag{4.4.16}$$

For arbitrary $\alpha,\beta \in \mathbb{R}^{2\tilde{L}}$ it holds that

$$(V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} U^h \alpha, \beta)_{\mathbb{R}^{2\tilde{L}}} = \sum_{k=1}^{2\tilde{L}} \alpha_k (V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} u_k^h, \beta)_{\mathbb{R}^{2\tilde{L}}} = \sum_{k,l=1}^{2\tilde{L}} \alpha_k \langle v_l^h, \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} u_k^h \rangle_{Y^*,Y} \beta_l$$
$$= \sum_{k,l=1}^{2\tilde{L}} \alpha_k (\iota^* v_l^h, \iota^* \tilde{D}_{\tau+\mu}^{-h} u_k^h)_{X^*} \beta_l = \alpha^\top \tilde{A}^h \beta,$$

i.e., the operator $\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} U^h \in \mathscr{L}(\mathbb{R}^{2\tilde{L}})$ is induced by the matrix $I_{2\tilde{L}} + \tilde{A}^h \in \mathbb{R}^{2\tilde{L}\times 2\tilde{L}}$. Therefore, the operator $\tilde{F}_{\mathrm{BFGS}}^h \in \mathscr{L}(Y^{*h})$ is invertible if and only if the matrix $I_{2\tilde{L}} + \tilde{A}^h$ is invertible. Note that we can compute $\tilde{A}^h$ exactly. Therefore, we can check for a given discretization space Y^{*h} if the operator $\tilde{F}_{\mathrm{BFGS}}^h$ is invertible. The next lemma answers the question if $\tilde{F}_{\mathrm{BFGS}}^h$ is invertible for a sufficiently accurate discretization space Y^{*h}.

LEMMA 4.4.5. *There exists a constant $\delta > 0$ such that the condition*

$$|(\iota^*(\tilde{D}_{\tau+\mu}^{-1} - \tilde{D}_{\tau+\mu}^{-h}) u_k^h, \iota^* v_l^h)_{X^*}| \le \delta \qquad \text{for all } 1 \le k,l \le 2\tilde{L}$$

implies that $\tilde{F}_{BFGS}^h$ is invertible.

Proof. Similar to $\tilde{A}^h$, define the matrix $\tilde{A} \in \mathbb{R}^{2\tilde{L}\times 2\tilde{L}}$ by

$$\tilde{A}_{k,l} := (\iota^* \tilde{D}_{\tau+\mu}^{-1} u_k^h, \iota^* v_l^h)_{X^*}. \tag{4.4.17}$$

This yields

$$(V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-1} U^h \alpha, \beta)_{\mathbb{R}^{2\tilde{L}}} = \alpha^\top \tilde{A} \beta \qquad \text{for all } \alpha,\beta \in \mathbb{R}^{2\tilde{L}},$$

i.e., the operator $\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-1} U^h \in \mathscr{L}(\mathbb{R}^{2\tilde{L}})$ is induced by the matrix $I_{2\tilde{L}} + \tilde{A} \in \mathbb{R}^{2\tilde{L}\times 2\tilde{L}}$. Since $\tilde{F}_{\mathrm{BFGS}}$ is invertible, $\mathrm{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-1} U^h$ and $I_{2\tilde{L}} + \tilde{A}$ are invertible. From

$$1 = \|I_{2\tilde{L}}\|_{\mathrm{op}} = \|(I_{2\tilde{L}} + \tilde{A})(I_{2\tilde{L}} + \tilde{A})^{-1}\|_{\mathrm{op}} \le \|I_{2\tilde{L}} + \tilde{A}\|_{\mathrm{op}} \|(I_{2\tilde{L}} + \tilde{A})^{-1}\|_{\mathrm{op}}$$

we infer that $\|(I_{2\tilde{L}} + \tilde{A})^{-1}\|_{\mathrm{op}} \ge \|I_{2\tilde{L}} + \tilde{A}\|_{\mathrm{op}}^{-1} > 0$ and

$$\delta := \|(I_{2\tilde{L}} + \tilde{A})^{-1}\|_{\mathrm{op}}^{-1} > 0.$$

If $|(\iota^*(\tilde{D}_{\tau+\mu}^{-1} - \tilde{D}_{\tau+\mu}^{-h}) u_k^h, \iota^* v_l^h)_{X^*}| \le \delta$ for all $1 \le k,l \le 2\tilde{L}$, then

$$\|\tilde{A} - \tilde{A}^h\|_{\mathrm{op}} = \sup_{\|\alpha\|_{\mathbb{R}^{2\tilde{L}}}=1, \|\beta\|_{\mathbb{R}^{2\tilde{L}}}=1} \sum_{1\le k,l \le 2\tilde{L}} |\alpha_k (\iota^*(\tilde{D}_{\tau+\mu}^{-1} - \tilde{D}_{\tau+\mu}^{-h}) u_k^h, \iota^* v_l^h)_{X^*} \beta_l| \le \delta = \|(I_{2\tilde{L}} + \tilde{A})^{-1}\|_{\mathrm{op}}^{-1}.$$

Under the given conditions, [129, Prop. 3.3.9] implies that $I_{2\tilde{L}} + \tilde{A}^h$ is invertible. By Theorem 4.4.3, this yields the invertibility of $\tilde{F}_{\mathrm{BFGS}}^h$. ∎

To find an approximation $p_j^{F_{\mathrm{BFGS}}^{-1}} \in X$ of $F_{\mathrm{BFGS}}^{-1} p_j' = R_X^{-1} \iota^* \tilde{F}_{\mathrm{BFGS}}^{-1} p_j^h$, we start with a given discretization space Y^{*h} and check if the matrix $\tilde{A}^h$ is invertible. If this is the case, $\tilde{F}_{\mathrm{BFGS}}^h \in \mathscr{L}(Y^{*h})$ is invertible and we set $p_j^{F_{\text{BFGS}}^{-1}} := R_X^{-1}\iota^*\tilde{F}_{\text{BFGS}}^{-h} p_j^h$. Otherwise, we refine the approximation space Y^{*h} such that the error $\max_{1\le k,l\le 2\tilde{L}} |(\iota^*(\tilde{D}_{\tau+\mu}^{-1} - \tilde{D}_{\tau+\mu}^{-h})u_k^h, \iota^* v_l^h)_{X^*}|$ is reduced. Lemma 4.4.5 implies that this leads to an operator $\tilde{F}_{\text{BFGS}}^h$ which is invertible. In the following, we always assume that Y^{*h} is chosen such that $\tilde{F}_{\text{BFGS}}^h$ is invertible and we denote the inverse of $\tilde{F}_{\text{BFGS}}^h$ by $\tilde{F}_{\text{BFGS}}^{-h} := (\tilde{F}_{\text{BFGS}}^h)^{-1} \in \mathscr{L}(Y^{*h})$. There are two ways to compute the value $z^h := \tilde{F}_{\text{BFGS}}^{-h} y^h \in Y^{*h}$. Either via (4.4.15) or via the finite dimensional linear system

$$\text{Find } z^h \in Y^{*h}: \qquad (\tilde{F}_{\text{BFGS}}^h z^h, w^h)_{Y^*} = (y^h, w^h)_{Y^*} \qquad \text{for all } w^h \in Y^{*h}. \tag{4.4.18}$$

Since the operator $\tilde{F}_{\text{BFGS}}^h$ is not self-adjoint, (4.4.18) cannot be solved via the conjugate gradient method, see, e.g., [51, 42]. Instead, other Krylov subspace methods such as CGS [122] or BiCGSTAB [127] have to be used. However, in every iteration of the Krylov subspace method one needs to evaluate $\tilde{F}_{\text{BFGS}}^h = \tilde{D}_{\tau+\mu} + U^h V^h \iota R_X^{-1}\iota^*$. If the evaluation of $\tilde{D}_{\tau+\mu}$ is costly (for example if $\tilde{D}_{\tau+\mu}$ is evaluated by solving the linear system (4.3.6)), it is advantageous to use the formula (4.4.15). Evaluation of $\tilde{F}_{\text{BFGS}}^{-h} y^h$ with this formula requires the evaluation of $\tilde{D}_{\tau+\mu}^{-h} u_l^h$, $1 \le l \le 2\tilde{L}$ and $\tilde{D}_{\tau+\mu}^{-h} y^h$, i.e., $2\tilde{L}+1$ solves of the linear system (4.3.7). If the values $\tilde{D}_{\tau+\mu}^{-h} u_l^h \in Y^{*h}$, $1 \le l \le 2\tilde{L}$, are stored, this reduces to one evaluation of $\tilde{D}_{\tau+\mu}^{-h}\tilde{y}^h$ for each consecutive evaluation of $\tilde{F}_{\text{BFGS}}^{-h}\tilde{y}^h$ with a different $\tilde{y}^h \in Y^{*h}$.

4.4.5. A first error estimate for $\tilde{F}_{\text{BFGS}}$

Given the approximation $p_j^{F_{\text{BFGS}}^{-1}} = R_X^{-1}\iota^*\tilde{F}_{\text{BFGS}}^{-h} p_j^h$ developed in the previous section, the goal of this section is to develop error estimates $e_{i,j,F_{\text{BFGS}}^{-1}} \in \mathbb{R}$, $i, j \in I$, and $e_{F_{\text{BFGS}}} \in \mathbb{R}$ which fulfill Assumption 4.2.4, i.e.,

$$|\langle p_i', F_{\text{BFGS}}^{-1} p_j' - p_j^{F_{\text{BFGS}}^{-1}} \rangle_{X^*,X}| \le e_{i,j,F_{\text{BFGS}}^{-1}} \qquad \text{and} \qquad \langle F_{\text{BFGS}}\tilde{d}^* - \mathring{d}', \tilde{d}^* \rangle_{X^*,X} \le e_{F_{\text{BFGS}}}.$$

In Theorems 4.4.8, 4.6.17, 5.5.8, and 5.5.14 we develop several distinct error estimates for the term $|(\iota^* x^h, \iota^*(\tilde{F}_{\text{BFGS}}^{-1} - \tilde{F}_{\text{BFGS}}^{-h})y^h)_{X^*}|$ with arbitrary $x^h, y^h \in Y^{*h}$. For now, let us introduce the following assumption.

Assumption 4.4.6. For arbitrary $x^h, y^h \in Y^{*h}$ there exists an error estimate $e_{x^h,y^h}^{F_{\text{BFGS}}^{-1}} \ge 0$ fulfilling

$$|(\iota^* x^h, \iota^*(\tilde{F}_{\text{BFGS}}^{-1} - \tilde{F}_{\text{BFGS}}^{-h})y^h)_{X^*}| \le e_{x^h,y^h}^{F_{\text{BFGS}}^{-1}}.$$

LEMMA 4.4.7. *Let Assumption 4.3.5 hold such that $p_j' = R_X^{-1}\iota^* p_j^h$ with $p_j^h \in Y^{*h}$ for all $j \in I$. Under Assumptions 4.3.6 and 4.4.6, the approximation and the error estimates*

$$p_j^{F_{BFGS}^{-1}} := R_X^{-1}\iota^*\tilde{F}_{BFGS}^{-h} p_j^h, \qquad e_{i,j,F_{BFGS}^{-1}} := e_{p_i^h,p_j^h}^{F_{BFGS}^{-1}}, \qquad e_{F_{BFGS}} := (\tau+\mu)e_{\tilde{d}^{*h},\tilde{d}^{*h}}^E,$$

with $\tilde{d}^{*h} := \sum_{j\in I}\tilde{\lambda}_j^*\tilde{F}_{BFGS}^{-h} p_j^h$, fulfill Assumption 4.2.4.

Proof. First we calculate

$$\begin{aligned}|\langle p_i', F_{\mathrm{BFGS}}^{-1} p_j' - p_j^{F^{-1}} \rangle_{X^*,X}| &= |\langle \iota^* p_i^h, F_{\mathrm{BFGS}}^{-1} \iota^* p_j^h - R_X^{-1} \iota^* \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h \rangle_{X^*,X}| \\ &= |(\iota^* p_i^h, \iota^* (\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h}) p_j^h)_{X^*}| \le e_{p_i^h, p_j^h}^{F_{\mathrm{BFGS}}^{-1}}.\end{aligned}$$

The definition of $\mathring{d}'$ (cf. Lemma 4.2.3) and $\tilde{F}_{\mathrm{BFGS}}^{-h} = (\tilde{F}_{\mathrm{BFGS}}^{h})^{-1}$ yield

$$\mathring{d}' = -\sum_{j \in I} \tilde{\lambda}_j^* \iota^* \tilde{F}_{\mathrm{BFGS}}^{h} \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h = \iota^* \tilde{F}_{\mathrm{BFGS}}^{h} \tilde{d}^{*h}$$

and $\tilde{d}^* = R_X^{-1} \iota^* \tilde{d}^{*h} \in X^h$ yields

$$F_{\mathrm{BFGS}} \tilde{d}^* = F_{\mathrm{BFGS}} R_X^{-1} \iota^* \tilde{d}^{*h} = \iota^* \tilde{F}_{\mathrm{BFGS}} \tilde{d}^{*h}.$$

Therefore we estimate

$$\begin{aligned}\langle F_{\mathrm{BFGS}} \tilde{d}^* - \mathring{d}', \tilde{d}^* \rangle_{X^*,X} &= (\iota^* (\tilde{F}_{\mathrm{BFGS}} - \tilde{F}_{\mathrm{BFGS}}^{h}) \tilde{d}^{*h}, \iota^* \tilde{d}^{*h})_{X^*} \\ &= (\tau + \mu)(\iota^* (\tilde{E} - \tilde{E}^h) \tilde{d}^{*h}, \iota^* \tilde{d}^{*h})_{X^*} \\ &\le (\tau + \mu) e_{\tilde{d}^{*h}, \tilde{d}^{*h}}^{E}.\end{aligned}$$

□

In the following, let Assumption 4.3.7 be fulfilled, i.e., for all $x^h, y^h \in Y^{*h}$ there exists an error estimate $e_{x^h,y^h}^{D^{-1}} \in \mathbb{R}$ such that $|(\iota^* x^h, \iota^* (\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h}) y^h)_{X^*}| \le e_{x^h,y^h}^{D^{-1}}$. Then (4.4.12) and (4.4.15) directly yield an error bound $e_{x_i^h, y_j^h}^{F_{\mathrm{BFGS}}^{-1}}$ for Assumption 4.4.6. To shorten the notation, we set

$$\begin{aligned} &e_{x^h,u^h_\cdot}^{D^{-1}} \in \mathbb{R}^{2\tilde{L}}, && (e_{x^h,u^h_\cdot}^{D^{-1}})_k := e_{x^h,u^h_k}^{D^{-1}}, \\ &e_{v^h_\cdot,y^h}^{D^{-1}} \in \mathbb{R}^{2\tilde{L}}, && (e_{v^h_\cdot,y^h}^{D^{-1}})_l := e_{v^h_l,y^h}^{D^{-1}}, \\ &e_{u^h_\cdot,v^h_\cdot}^{D^{-1}} \in \mathbb{R}^{2\tilde{L} \times 2\tilde{L}}, && (e_{u^h_\cdot,v^h_\cdot}^{D^{-1}})_{k,l} := e_{u^h_k,v^h_l}^{D^{-1}}. \end{aligned}$$

We also use the notation $|m| \in \mathbb{R}^n$, $|M| \in \mathbb{R}^{n \times n}$ for the elementwise absolute value of a vector $m \in \mathbb{R}^n$ and a matrix $M \in \mathbb{R}^{n \times n}$, $n \in \mathbb{N}_+$, cf. the section "Notation".

THEOREM 4.4.8. *Let $x^h, y^h \in Y^{*h}$ are arbitrary and suppose that Assumption 4.3.7 is fulfilled and define $\tilde{A}^h$ according to* (4.4.16). *If $\||(I_{2\tilde{L}} + \tilde{A}^h)^{-1}| e_{u^h_\cdot,v^h_\cdot}^{D^{-1}}\|_{op} < 1$, then it holds*

$$|(\iota^* x^h, \iota^* (\tilde{F}_{BFGS}^{-1} - \tilde{F}_{BFGS}^{-h}) y^h)_{X^*}| \le e_{x^h,y^h}^{D^{-1}} + (\|b^h\|_{\mathbb{R}^{2\tilde{L}}} + e_b) \|e_{x^h,u^h_\cdot}^{D^{-1}}\|_{\mathbb{R}^{2\tilde{L}}} + \|c^h\|_{\mathbb{R}^{2\tilde{L}}} e_b,$$

where $b^h, c^h \in \mathbb{R}^{2\tilde{L}}$, and $e_b \in \mathbb{R}$, are defined by

$$b^h := (I_{2\tilde{L}} + \tilde{A}^h)^{-1} V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} y^h, \qquad c_l^h := (\iota^* x^h, \iota^* \tilde{D}_{\tau+\mu}^{-h} u_l^h)_{X^*},\ 1 \le l \le 2\tilde{L},$$

and

$$e_b := \frac{\| |(I_{2\tilde{L}} + \tilde{A}^h)^{-1}| (e^{D^{-1}}_{u^h_\cdot, v^h_\cdot} |b^h| + e^{D^{-1}}_{v^h_\cdot, y^h}) \|_{\mathbb{R}^{2\tilde{L}}}}{1 - \| |(I_{2\tilde{L}} + \tilde{A}^h)^{-1}| e^{D^{-1}}_{u^h_\cdot, v^h_\cdot} \|_{op}}.$$

Proof. Equation (4.4.15) yields

$$\begin{aligned}
(\iota^* x^h, \iota^* \tilde{F}^{-h}_{\mathrm{BFGS}} y^h)_{X^*} &= (\iota^* x^h, \iota^* \tilde{D}^{-h}_{\tau+\mu} y^h)_{X^*} - (\iota^* x^h, \iota^* \tilde{D}^{-h}_{\tau+\mu} U^h (I_{2\tilde{L}} + \tilde{A}^h)^{-1} V^h \iota R_X^{-1} \iota^* \tilde{D}^{-h}_{\tau+\mu} y^h)_{X^*} \\
&= (\iota^* x^h, \iota^* \tilde{D}^{-h}_{\tau+\mu} y^h)_{X^*} - (\iota^* x^h, \iota^* \tilde{D}^{-h}_{\tau+\mu} U^h b^h)_{X^*} \\
&= (\iota^* x^h, \iota^* \tilde{D}^{-h}_{\tau+\mu} y^h)_{X^*} - {b^h}^\top c^h.
\end{aligned}$$

Similarly, Equation (4.4.12) yields

$$(\iota^* x^h, \iota^* \tilde{F}^{-1}_{\mathrm{BFGS}} y^h)_{X^*} = (\iota^* x^h, \iota^* \tilde{D}^{-1}_{\tau+\mu} y^h)_{X^*} - b^\top c,$$

where we define $\tilde{A}$ as in (4.4.17),

$$b := (I_{2\tilde{L}} + \tilde{A})^{-1} V^h \iota R_X^{-1} \iota^* \tilde{D}^{-1}_{\tau+\mu} y^h, \qquad c_l := (\iota^* x^h, \iota^* \tilde{D}^{-1}_{\tau+\mu} u^h_l)_{X^*},\ 1 \le l \le 2\tilde{L}.$$

Applying the triangle inequality leads to

$$\begin{aligned}
&|(\iota^* x^h, \iota^* (\tilde{F}^{-1}_{\mathrm{BFGS}} - \tilde{F}^{-h}_{\mathrm{BFGS}}) y^h)_{X^*}| \\
&\le |(\iota^* x^h, \iota^* (\tilde{D}^{-1}_{\tau+\mu} - \tilde{D}^{-h}_{\tau+\mu}) y^h)_{X^*}| + |b^\top c - {b^h}^\top c^h| \\
&\le e^{D^{-1}}_{x^h, y^h} + |b^\top (c - c^h)| + |(b - b^h)^\top c^h| \\
&\le e^{D^{-1}}_{x^h, y^h} + (\|b^h\|_{\mathbb{R}^{2\tilde{L}}} + \|b - b^h\|_{\mathbb{R}^{2\tilde{L}}}) \|c - c^h\|_{\mathbb{R}^{2\tilde{L}}} + \|b - b^h\|_{\mathbb{R}^{2\tilde{L}}} \|c^h\|_{\mathbb{R}^{2\tilde{L}}}.
\end{aligned}$$

From $|c_l - c^h_l| \le e^{D^{-1}}_{x^h, u^h_l}$ we deduce $\|c - c^h\|_{\mathbb{R}^{2\tilde{L}}} \le \|e^{D^{-1}}_{x^h, u^h_\cdot}\|_{\mathbb{R}^{2\tilde{L}}}$. Furthermore, b and b^h are the solutions to the linear equations

$$\begin{aligned}
(I_{2\tilde{L}} + \tilde{A})\ b\ &= V^h \iota R_X^{-1} \iota^* \tilde{D}^{-1}_{\tau+\mu} y^h, \\
(I_{2\tilde{L}} + \tilde{A}^h) b^h &= V^h \iota R_X^{-1} \iota^* \tilde{D}^{-h}_{\tau+\mu} y^h,
\end{aligned}$$

respectively. Note that

$$|(I_{2\tilde{L}} + \tilde{A}) - (I_{2\tilde{L}} + \tilde{A}^h)| = |\tilde{A} - \tilde{A}^h| \le e^{D^{-1}}_{u^h_\cdot, v^h_\cdot},$$

where "$\le$" is to be understood in the componentwise sense. Furthermore,

$$|V^h \iota R_X^{-1} \iota^* \tilde{D}^{-1}_{\tau+\mu} y^h - V^h \iota R_X^{-1} \iota^* \tilde{D}^{-h}_{\tau+\mu} y^h|_l = |(\iota^* v^h_l, \iota^* (\tilde{D}^{-1}_{\tau+\mu} - \tilde{D}^{-h}_{\tau+\mu}) y^h)_{X^*}| \le e^{D^{-1}}_{v^h_l, y^h} \quad \text{for all } 1 \le l \le 2\tilde{L}.$$

If $\| |(I_{2\tilde{L}} + \tilde{A}^h)^{-1}| e^{D^{-1}}_{u^h_\cdot, v^h_\cdot} \|_{\mathrm{op}} < 1$, then [52, Thm. 7.4] yields the estimate

$$\|b-b^h\|_{\mathbb{R}^{2\tilde{L}}} \leq \frac{\||(I_{2\tilde{L}}+\tilde{A}^h)^{-1}|(e^{D^{-1}}_{u^h_\cdot,v^h_\cdot}|b^h|+e^{D^{-1}}_{v^h_\cdot,y^h})\|_{\mathbb{R}^{2\tilde{L}}}}{1-\||(I_{2\tilde{L}}+\tilde{A}^h)^{-1}|e^{D^{-1}}_{u^h_\cdot,v^h_\cdot}\|_{\mathrm{op}}} = e_b. \qquad \square$$

Although the error bound of Theorem 4.4.8 fulfills Assumption 4.4.6, this bound might be very large if $I_{2\tilde{L}}+\tilde{A}^h$ is close to singular. Then one might use a bound which is derived from the fact that $\tilde{F}^h_{\mathrm{BFGS}}y^h$ is the solution of the linear system (4.4.18), cf. Theorems 4.6.17, 5.5.8, and 5.5.14. However, in many situations the operator $\tilde{D}^{-h}_{\tau+\mu}$ and an error estimate $e^{D^{-1}}_{x^h,y^h}$ can be evaluated using preexisting methods or with external program code, whereas it requires extra work to compute the error estimates of Theorems 4.6.17, 5.5.8, and 5.5.14.

4.5. Error estimates for the objective function of the bundle method

In this section, we address the issue of finding computable lower and upper bounds for the objective function of the bundle method which fulfill (4.1.4). Recall that the objective function $\Psi(\tilde{y})$ at the trial iterate $\tilde{y}\in X$ is given (cf. (4.2.1)) via

$$\Psi(y) = \max_{j\in I}\left\{\langle g'_j,y\rangle_{X^*,X}+s_j\right\}+\tfrac{\alpha}{2}\|y\|_X^2+\tfrac{1}{2}\|\iota(y-x_{\mathrm{SI}})\|^2_{Q+\tau R_Y},$$

where $g'_j\in X^*$, $s_j\in\mathbb{R}$, $j\in I:=\{1,\dots,n_p\}$, $\alpha>0$ and $x_{\mathrm{SI}}\in X$. We now assume that $x_{\mathrm{SI}}\in X^h$ and $g'_j\in X^{*h}$, $j\in I$, i.e., there exist $x^h_{\mathrm{SI}}\in Y^{*h}$ and $g^h_j\in Y^{*h}$ such that $x_{\mathrm{SI}}=R_X^{-1}\iota^*x^h_{\mathrm{SI}}$ and $g'_j=\iota^*g^h_j$, cf. Assumption 4.3.5. Whenever $\tilde{y}\in X^h$, i.e., whenever there exists a $y^h\in Y^{*h}$ such that $\tilde{y}=R_X^{-1}\iota^*y^h$, then we find

$$\Psi(\tilde{y}) = \max_{j\in I}\left\{(\iota^*g^h_j,\iota^*y^h)_{X^*}+s_j\right\}+\tfrac{\alpha}{2}\|\iota^*y^h\|^2_{X^*}+\tfrac{1}{2}\|\iota R_X^{-1}\iota^*(y^h-x^h_{\mathrm{SI}})\|^2_{Q_i+\tau_i R_Y}, \tag{4.5.1}$$

Under Assumption 4.3.1, the quantities $(x^h,y^h)_{Y^*}$ and $(\iota^*x^h,\iota^*y^h)_{X^*}$ can be computed exactly for all $x^h,y^h\in Y^{*h}$. Thus, the first two terms of $\Psi(\tilde{y})$ can be computed exactly. However, the last term is not of the form $(\cdot,\cdot)_{Y^*}$ or $(\iota^*\cdot,\iota^*\cdot)_{X^*}$. Consequently, lower and upper approximations of $\|\iota f\|_{Q+\tau R_Y}$ for $f\in X^h$ are needed. We start by providing lower and upper bounds for the term $\|\iota f\|_{R_Y}=\langle R_Y\iota f,\iota f\rangle_{Y^*,Y}=\|\iota f\|_Y$.

LEMMA 4.5.1. *For $f^h\in Y^{*h}$ and $f:=R_X^{-1}\iota^*f^h\in X^h$ it holds*

$$\|\iota f\|_Y^2=\|\tilde{E}f^h\|^2_{Y^*}=\|\tilde{E}^hf^h\|^2_{Y^*}+\|(\tilde{E}-\tilde{E}^h)f^h\|^2_{Y^*}.$$

Proof. By definition, $\tilde{E}=R_Y\iota R_X^{-1}\iota^*\in\mathscr{L}(Y^*)$ and $\tilde{E}^h=P_{Y^{*h}}\tilde{E}|_{Y^{*h}}\in Y^{*h}$. The Riesz representation theorem (cf. [3, Thm. 6.1]) yields

$$\|\iota f\|_Y=\|\iota R_X^{-1}\iota^*f^h\|_Y=\|R_Y\iota R_X^{-1}\iota^*f^h\|_{Y^*}=\|\tilde{E}f^h\|_{Y^*}.$$

Since $\tilde{E}^hf^h=P_{Y^{*h}}\tilde{E}f^h$ and $(\tilde{E}-\tilde{E}^h)f^h=(\mathrm{Id}_{Y^*}-P_{Y^*})\tilde{E}f^h$ are orthogonal in Y^* (cf. Section 2.1.2), we obtain

$$\begin{aligned}\|(\tilde{E}-\tilde{E}^h)f^h\|_{Y^*}^2 &= ((\tilde{E}-\tilde{E}^h)f^h,(\tilde{E}-\tilde{E}^h)f^h)_{Y^*}\\ &= ((\tilde{E}-\tilde{E}^h)f^h,(\tilde{E}+\tilde{E}^h)f^h)_{Y^*} = \|\tilde{E}f^h\|_{Y^*}^2 - \|\tilde{E}^hf^h\|_{Y^*}^2.\end{aligned}$$ ∎

THEOREM 4.5.2. *Assume that $Q=0$ and that there exists a computable error estimator $e_{\tilde{E}}(h)\geq 0$ such that $\|(\tilde{E}-\tilde{E}^h)(y^h-x^h_{SI})\|_{Y^*}\leq e_{\tilde{E}}(h)$ and $e_{\tilde{E}}(h)\to 0$ as $h\to 0$. Then the error estimators*

$$\begin{aligned}\underline{O}_\Psi(h) &:= \max_{j\in I}\Big\{(\iota^*g_j^h,\iota^*y^h)_{X^*}+s_j\Big\}+\tfrac{\alpha}{2}\|\iota^*y^h\|_{X^*}^2+\tfrac{\tau}{2}\|\tilde{E}^h(y^h-x^h_{SI})\|_{Y^*}^2,\\ \overline{O}_\Psi(h) &:= \max_{j\in I}\Big\{(\iota^*g_j^h,\iota^*y^h)_{X^*}+s_j\Big\}+\tfrac{\alpha}{2}\|\iota^*y^h\|_{X^*}^2+\tfrac{\tau}{2}\|\tilde{E}^h(y^h-x^h_{SI})\|_{Y^*}^2+\tfrac{\tau}{2}e_{\tilde{E}}^2(h)\end{aligned}$$

fulfill (4.1.4)*, i.e., it holds*

$$\underline{O}_\Psi(h)\leq\Psi(\tilde{y})\leq\overline{O}_\Psi(h),\qquad \textit{and}\qquad \overline{O}_\Psi(h)-\underline{O}_\Psi(h)\to 0 \textit{ as } h\to 0.$$

Proof. Let $x^h\in Y^{*h}$ be arbitrary and set $x=R_X^{-1}\iota^*x^h$. Since $Q=0$ and $\tilde{E}=R_Y\iota R_X^{-1}\iota^*$, there holds

$$\|\iota x\|_{Q+\tau R_Y}^2=\langle(0+\tau R_Y)\iota R_X^{-1}\iota^*x^h,\iota R_X^{-1}\iota^*x^h\rangle_{Y^*,Y}=\tau\langle\tilde{E}x^h,\iota R_X^{-1}\iota^*x^h\rangle_{Y^*,Y}=\tau\|\tilde{E}x^h\|_{Y^*}^2.$$

Choosing $x^h=y^h+x^h_{\mathrm{SI}}$, (4.5.1) and Lemma 4.5.1 yield the desired result. ∎

THEOREM 4.5.3. *Assume that $Q=Q_{BFGS}$ is given as the L-BFGS curvature operator* (4.4.5) *and that there exists a computable error estimator $e_{\tilde{E}}(h)\geq 0$ such that $\|(\tilde{E}-\tilde{E}^h)(y^h-x^h_{SI})\|_{Y^*}\leq e_{\tilde{E}}$ and $e_{\tilde{E}}(h)\to 0$ as $h\to 0$. Then the error estimators*

$$\begin{aligned}\underline{O}_\Psi(h) &:= \max_{j\in I}\Big\{(\iota^*g_j^h,\iota^*y^h)_{X^*}+s_j\Big\}+\tfrac{\alpha}{2}\|\iota^*y^h\|_{X^*}^2\\ &\qquad\qquad +\sum_{l=1}^{2\tilde{L}}(\iota^*v_l^h,\iota^*x^h)_{X^*}(\iota^*u_l^h,\iota^*x^h)_{X^*}+\tfrac{\tau+\mu}{2}\|\tilde{E}^h(y^h-x^h_{SI})\|_{Y^*}^2,\\ \overline{O}_\Psi(h) &:= \underline{O}_\Psi(h)+\tfrac{\tau+\mu}{2}e_{\tilde{E}}^2(h),\end{aligned}$$

fulfill (4.1.4)*, i.e., it holds $\underline{O}_\Psi(h)\leq\Psi(\tilde{y})\leq\overline{O}_\Psi(h)$, and $\overline{O}_\Psi(h)-\underline{O}_\Psi(h)\to 0$ as $h\to 0$.*

Proof. Let $x^h\in Y^{*h}$ be arbitrary and set $x=R_X^{-1}\iota^*x^h$. Since $Q_{\mathrm{BFGS}}=\mu R_Y+\sum_{l=1}^{2\tilde{L}}u_l^h\langle v_l^h,\cdot\rangle_{Y^*,Y}$ and $\tilde{E}=R_Y\iota R_X^{-1}\iota^*$, it holds

$$\begin{aligned}\|x\|_{Q+\tau R_Y}^2 &= \langle(Q_{\mathrm{BFGS}}+\tau R_Y)\iota R_X^{-1}\iota^*x^h,\iota R_X^{-1}\iota^*x^h\rangle_{Y^*,Y}\\ &= \langle(\tau+\mu)R_Y\iota R_X^{-1}\iota^*x^h+\sum_{l=1}^{2\tilde{L}}u_l^h\langle v_l^h,\iota R_X^{-1}\iota^*x^h\rangle_{Y^*,Y},\iota R_X^{-1}\iota^*x^h\rangle_{Y^*,Y}\\ &= (\tau+\mu)\|\tilde{E}x^h\|_{Y^*}^2+\sum_{l=1}^{2\tilde{L}}(\iota^*v_l^h,\iota^*x^h)_{X^*}(\iota^*u_l^h,\iota^*x^h)_{X^*}.\end{aligned}$$

Choosing $x^h=y^h+x^h_{\mathrm{SI}}$, (4.5.1) and Lemma 4.5.1 yield the desired result. ∎

4.6. A priori error estimates

In order to fulfill Assumptions 4.3.6, 4.3.7, and 4.4.6, we need error bounds for the quantities $(\iota^*(\tilde{E} - \tilde{E}^h)x^h, \iota^* y^h)_{X^*}$, $|(\iota^* x^h, \iota^*(\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h})y^h)_{X^*}|$ and $|(\iota^* x^h, \iota^*(\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h})y^h)_{X^*}|$ for $x^h, y^h \in Y^{*h}$. As $z' := \tilde{D}_\tau^{-1} y^h$ and $z'_{\mathrm{BFGS}} := \tilde{F}_{\mathrm{BFGS}}^{-1} y^h$ are defined via the variational equations

$$\text{Find } z' \in Y^* : \qquad (\tilde{D}_\tau z', w')_{Y^*} = (y^h, w')_{Y^*}, \qquad \text{for all } w' \in Y^*$$

and

$$\text{Find } z'_{\mathrm{BFGS}} \in Y^* : \qquad (\tilde{F}_{\mathrm{BFGS}} z'_{\mathrm{BFGS}}, w')_{Y^*} = (y^h, w')_{Y^*}, \qquad \text{for all } w' \in Y^*,$$

respectively, we consider the following abstract setting:

Let $a : Y^* \times Y^* \to \mathbb{R}$ be a bounded bilinear form. For $x' \in X^*$ and $y' \in Y^*$, define the variational equations

$$\text{Find } z' \in Y^* : \qquad a(z', w') = (y', w')_{Y^*}, \qquad \text{for all } w' \in Y^*, \tag{4.6.1}$$

$$\text{Find } z^h \in Y^{*h} : \qquad a(z^h, w^h) = (y', w^h)_{Y^*}, \qquad \text{for all } w^h \in Y^{*h}, \tag{4.6.2}$$

and the adjoint problem

$$\text{Find } \Phi'_x \in Y^* : \qquad a(w', \Phi'_x) = (x', \iota^* w')_{X^*}, \qquad \text{for all } w' \in Y^*. \tag{4.6.3}$$

Note that, at this point, we do not assume that the bilinear form a is coercive and thus the Lax-Milgram theorem cannot be used to infer unique solvability of the VEs (4.6.1)–(4.6.3). Instead, assume that (4.6.1)–(4.6.3) have unique solutions. If we are interested in estimating the quantity $|(x', \iota^*(z' - z^h))_{X^*}|$, then the following abstract result, which is often called the Aubin-Nietsche trick (cf. [20, Thm. 3.2.4]), helps to utilize that we only need error estimates in the weaker space X^*.

THEOREM 4.6.1 (Aubin-Nietsche trick). *Let $a : Y^* \times Y^* \to \mathbb{R}$ be a bounded (with constant M) bilinear form and $x' \in X^*$, $y' \in Y^*$. Let* (4.6.1)–(4.6.3) *be uniquely solvable and denote by $z' \in Y^*$, $z^h \in Y^{*h}$ and $\Phi'_x \in Y^*$ the respective solutions. Let $\Phi^h_x \in Y^{*h}$ be arbitrary. Then it holds that*

$$|(x', \iota^*(z' - z^h))_{X^*}| \le M \|z' - z^h\|_{Y^*} \|\Phi'_x - \Phi^h_x\|_{Y^*}.$$

Proof. Abbreviate the error $e' := z' - z^h \in Y^*$. First, note that Galerkin orthogonality holds, i.e.,

$$a(e', w^h) = a(z' - z^h, w^h) = (y^h, w^h)_{Y^*} - a(z^h, w^h) = 0 \quad \text{for all } w^h \in Y^{*h}. \tag{4.6.4}$$

The definition of Φ'_x yields for arbitrary $\Phi^h_x \in Y^{*h}$ that

$$(x', \iota^* e')_{X^*} = a(e', \Phi'_x) = a(e', \Phi'_x - \Phi^h_x).$$

Since the bilinear form a is bounded with constant M, we find

$$|(x', \iota^* e')_{X^*}| = |a(z' - z^h, \Phi'_x - \Phi^h_x)| \le M \|z' - z^h\|_{Y^*} \|\Phi'_x - \Phi^h_x\|_{Y^*}. \qquad \square$$

In the case that the bilinear form $a : Y^* \times Y^* \to \mathbb{R}$ is coercive, the error $\|z - z^h\|_{Y^*}$ can be bounded via the following lemma.

LEMMA 4.6.2. *Let $a : Y^* \times Y^* \to \mathbb{R}$ be a bounded coercive symmetric bilinear form with coecivity constant m and denote by $z' \in Y^*$ and $z^h \in Y^{*h}$ the solution to* (4.6.1) *and* (4.6.2)*, respectively. Then*

$$\|z' - z^h\|_{Y^*} \leq \frac{1}{m}\|y'\|_{Y^*}.$$

Proof. By the Lax-Milgram theorem (Theorem 2.3.2), (4.6.1) and (4.6.2) have unique solutions. Using the coercivity of a and the Galerkin orthogonality (4.6.4) results in

$$m\|z'-z^h\|_{Y^*}^2 \leq a(z'-z^h, z'-z^h) = a(z'-z^h, z') = a(z', z'-z^h) = (y', z'-z^h)_{Y^*} \leq \|z'-z^h\|_{Y^*}\|y'\|_{Y^*}. \square$$

Next we want to find a suitable $\Phi_x^h \in Y^{*h}$ such that the term $\|\Phi_x' - \Phi_x^h\|_{Y^*}$ becomes small. If a is a coercive symmetric bilinear form with coercivity constant m and Φ_x^h is chosen to be the solution of the discretized adjoint problem

$$\text{Find } \Phi_x^h \in Y^{*h}: \qquad a(w^h, \Phi_x^h) = (x', \iota^* w^h)_{X^*}, \qquad \text{for all } w^h \in Y^{*h},$$

then Lemma 4.6.2 implies $\|\Phi_x' - \Phi_x^h\|_{Y^*} \leq m^{-1}\|R_Y \iota R_X^{-1} x'\|_{Y^*}$. However, this estimate can be improved if it is known a priori that the solution of the adjoint problem (4.6.3) has additional regularity.

DEFINITION 4.6.3. *Let $(Y^{*h})_{h\in(0,1]}$ be a family of discretization spaces, i.e. let Y^{*h} be a finite dimensional linear subspace of Y^* for all $h \in (0,1]$. Denote by $\Phi_x' \in Y^*$ the solution to* (4.6.3)*. The adjoint problem* (4.6.3) *is called $\mathring{\gamma}$-regular relative to $(Y^{*h})_h$ if there exists a number $\mathring{\gamma} \geq 0$ and a constant $C_{reg} \geq 0$ such that for all $0 < h \leq 1$ and all $x^h \in Y^{*h}$ there exists an element $\Phi_x^h \in Y^{*h}$ which fulfills*

$$\|\Phi_x' - \Phi_x^h\|_{Y^*} \leq C_{reg} h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*}.$$

THEOREM 4.6.4. *If $a : Y^* \times Y^* \to \mathbb{R}$ is a coercive (with constant m) and bounded (with constant M) symmetric bilinear form and the adjoint problem* (4.6.3) *is $\mathring{\gamma}$-regular relative to $(Y^{*h})_h$ (with constant C_{reg}), then there holds*

$$|(\iota^* x^h, \iota^*(z'-z^h))_{X^*}| \leq \frac{C_{reg}M}{m}\, h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*}\|y'\|_{Y^*} \qquad \text{for all } x^h \in Y^{*h}, y' \in Y^*,$$

where $z' \in Y^*$ and $z^h \in Y^{*h}$ are the solutions to (4.6.1) *and* (4.6.1)*, respectively.*

Proof. Let $x^h \in Y^{*h}$, $y' \in Y^*$ be arbitrary and denote by $\Phi_x^h \in Y^{*h}$ the element from Definition 4.6.3. By the Aubin-Nietsche trick (Theorem 4.6.1) we find

$$|(\iota^* x^h, \iota^*(z'-z^h))_{X^*}| \leq M\|z'-z^h\|_{Y^*}\|\Phi_x' - \Phi_x^h\|_{Y^*}.$$

Lemma 4.6.2 yields $\|z'-z^h\|_{Y^*} \leq m^{-1}\|y'\|_{Y^*}$ As the adjoint problem (4.6.3) is $\mathring{\gamma}$-regular, we conclude

$$|(\iota^* x^h, \iota^*(z'-z^h))_{X^*}| \leq \frac{C_{\mathrm{reg}}M}{m} h^{\mathring{\gamma}}\|\iota^* x^h\|_{X^*}\|y'\|_{Y^*}. \qquad \square$$

Using the same tools as in the proof of Theorem 4.6.4, we now develop an upper bound for the quantity $(\iota^*(\tilde{E}-\tilde{E}^h)f^h, \iota^* f^h)_{X^*}$.

LEMMA 4.6.5. *For arbitrary $x^h, y^h \in Y^{*h}$, it holds*

$$|(\iota^* x^h, \iota^*(\tilde{E} - \tilde{E}^h) y^h)_{X^*}| \leq \|(\tilde{E} - \tilde{E}^h) x^h\|_{Y^*} \|(\tilde{E} - \tilde{E}^h) y^h\|_{Y^*}.$$

Proof. Let $x^h, y^h \in Y^{*h}$ be arbitrary. Then $z' := \tilde{E} y^h \in Y^*$ and $z^h := \tilde{E}^h y^h \in Y^{*h}$ are characterized via the variational equations

$$\text{Find } z' \in Y^* : \qquad (z', w')_{Y^*} = (\tilde{E} y^h, w')_{Y^*} \qquad \text{for all } w' \in Y^*$$

and

$$\text{Find } z^h \in Y^{*h} : \qquad (z^h, w^h)_{Y^*} = (\tilde{E} y^h, w^h)_{Y^*} \qquad \text{for all } w^h \in Y^{*h},$$

i.e., z' and z^h are the solutions to (4.6.1) and (4.6.2) with $a(\cdot,\cdot) := (\cdot,\cdot)_{Y^*}$ and right hand side $\tilde{E} y^h$. Since $a(\cdot,\cdot) := (\cdot,\cdot)_{Y^*}$ is a bounded bilinear form with continuity constant $M = 1$, the Aubin-Nietsche trick (Theorem 4.6.1) yields for arbitrary $\Phi_x^h \in Y^{*h}$ that

$$|(\iota^* x^h, \iota^*(z' - z^h))_{X^*}| \leq \|z' - z^h\|_{Y^*} \|\Phi_x' - \Phi_x^h\|_{Y^*}, \tag{4.6.5}$$

where $\Phi_x' \in Y^*$ is the solution to the adjoint equation

$$\text{Find } \Phi_x' \in Y^* : \qquad a(w', \Phi_x') = (\iota^* x^h, \iota^* w')_{X^*}, \qquad \text{for all } w' \in Y^*.$$

Note that the definition of $\tilde{E} \in \mathscr{L}(Y^*)$, $\tilde{E} = R_Y \iota R_X^{-1} \iota^*$, yields $(\tilde{E} y^h, w')_{Y^*} = (\iota^* y^h, \iota^* w')_{X^*}$ for all $w' \in Y^*$. Therefore, $\Phi_x' = \tilde{E} x^h$ and we can choose $\Phi_x^h := \tilde{E}^h x^h$ in (4.6.5). ∎

COROLLARY 4.6.6 (A priori estimate for $\tilde{E}$). *Let $v^h \in Y^{*h}$ be arbitrary. If the problem*

$$\text{Find } z' \in Y^* : \qquad (z', w')_{Y^*} = (\iota^* v^h, \iota^* w')_{X^*} \qquad \text{for all } w' \in Y^* \tag{4.6.6}$$

is $\mathring{\gamma}$-regular with respect to Y^{*h} (with constant C_{reg}), then it holds

$$|(\iota^* x^h, \iota^*(\tilde{E} - \tilde{E}^h) y^h)_{X^*}| \leq C_{reg}^2 h^{2\mathring{\gamma}} \|\iota^* x^h\|_{X^*} \|\iota^* y^h\|_{X^*} \qquad \text{for all } x^h, y^h \in Y^{*h}.$$

In particular, $e^E_{x^h,y^h} := C_{reg}^2 h^{2\mathring{\gamma}} \|\iota^* x^h\|_{X^*} \|\iota^* y^h\|_{X^*}$ fulfills Assumption 4.3.6.

Proof. Let $x^h, y^h \in Y^{*h}$ be arbitrary. By Lemma 4.6.5, it holds

$$|(\iota^* x^h, \iota^*(\tilde{E} - \tilde{E}^h) y^h)_{X^*}| \leq \|(\tilde{E} - \tilde{E}^h) x^h\|_{Y^*} \|(\tilde{E} - \tilde{E}^h) y^h\|_{Y^*}.$$

By the definition of $\tilde{E} \in \mathscr{L}(Y^*)$, $\tilde{E} = R_Y \iota R_X^{-1} \iota^*$, both $\tilde{E} x^h$ and $\tilde{E} y^h$ solve (4.6.6) with the right hand sides $v^h = x^h$ and $v^h = y^h$, respectively. As (4.6.6) is $\mathring{\gamma}$-regular, there exists a constant $C_{\text{reg}} \geq 0$ and elements $\Phi_x^h, \Phi_y^h \in Y^{*h}$ with

$$\|\tilde{E} x^h - \Phi_x^h\|_{Y^*} \leq C_{\text{reg}} h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*} \quad \text{and} \quad \|\tilde{E} y^h - \Phi_y^h\|_{Y^*} \leq C_{\text{reg}} h^{\mathring{\gamma}} \|\iota^* y^h\|_{X^*}.$$

By definition, $\tilde{E}^h x^h$ is the metric projection of $\tilde{E} x^h$ onto the closed subspace Y^{*h}. The definition of the metric projection (cf. Section 2.1.2) yields $\|\tilde{E} x^h - \tilde{E}^h x^h\|_{Y^*} = \|\tilde{E} x^h - P_{Y^{*h}} \tilde{E} x^h\|_{Y^*} = \min_{v^h \in Y^{*h}} \|\tilde{E} x^h -$

$v^h\|_{Y^{*h}} \leq \|\tilde{E}x^h - \Phi_x^h\|_{Y^*}$. Since this also holds for y^h, we conclude

$$|(\iota^* x^h, \iota^*(\tilde{E}y^h - \tilde{E}^h y^h))_{X^*}| \leq \|(\tilde{E} - \tilde{E}^h)x^h\|_{Y^*}\|(\tilde{E} - \tilde{E}^h)y^h\|_{Y^*} \leq C_{\mathrm{reg}}^2 h^{2\mathring{\gamma}}\|\iota^* x^h\|_{X^*}\|\iota^* y^h\|_{X^*}. \qquad \square$$

4.6.1. A priori error estimates for $Q = 0$

Let $x^h \in Y^{*h}$. We consider the adjoint problem

$$\text{Find } \Phi_x' \in Y^* : \qquad (\tilde{D}_{\tau+\mu}\, w', \Phi_x')_{Y^*} = (\iota^* x^h, \iota^* w')_{X^*} \qquad \text{for all } w' \in Y^*. \tag{4.6.7}$$

THEOREM 4.6.7 (A priori estimate for $\tilde{D}_\tau^{-1}$). *Let the adjoint problem* (4.6.7) *(with* $\mu = 0$*) be* $\mathring{\gamma}$*-regular (with constant* C_{reg}*) and define* $C := (1 + \frac{\tau}{\alpha}\|\iota^*\|^2_{\mathscr{L}(Y^*,X^*)})C_{reg}$*. Then it holds*

$$|(\iota^* x^h, \iota^*(\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h})y^h)_{X^*}| \leq C h^{\mathring{\gamma}}\|\iota^* x^h\|_{X^*}\|y^h\|_{Y^*} \qquad \textit{for all } x^h, y^h \in Y^{*h}.$$

In particular, $e^{D^{-1}}_{x^h,y^h} := C h^{\mathring{\gamma}}\|\iota^* x^h\|_{X^*}\|y^h\|_{Y^*}$ *fulfills Assumption 4.3.7.*

Proof. Note that $a(\cdot,\cdot) := (\tilde{D}_\tau \cdot, \cdot)_{Y^*}$ is a coercive symmetric bounded bilinear form because

$$(\tilde{D}_\tau v', w')_{Y^*} \leq \alpha\|v'\|_{Y^*}\|w'\|_{Y^*} + \tau|(\iota^* v', \iota^* w')_{X^*}| \leq M\|v'\|_{Y^*}\|w'\|_{Y^*} \qquad \text{for all } v', w' \in Y^*$$

with $M := \alpha + \tau\|\iota^*\|^2_{\mathscr{L}(Y^*,X^*)}$. Furthermore, $\tilde{E} = R_Y \iota R_X^{-1}\iota^*$ shows

$$(\tilde{D}_\tau v', v')_{Y^*} = \alpha\|v'\|^2_{Y^*} + \tau(\tilde{E}v', v')_{Y^*} = \alpha\|v'\|^2_{Y^*} + \tau\|\iota^* v'\|^2_{X^*} \geq \alpha\|v'\|^2_{Y^*} \qquad \text{for all } v' \in Y^*.$$

Since $(\tilde{D}_\tau^h v^h, w^h)_{Y^*} = (\tilde{D}_\tau v^h, w^h)_{Y^*}$ for all $v^h, w^h \in Y^{*h}$, we find that $z^h = \tilde{D}_\tau^{-h} y^h$ and $z' = \tilde{D}_\tau^{-1} y^h$, where $z' \in Y^*$, $z^h \in Y^{*h}$ are the solutions to (4.6.1) and (4.6.2), respectively. Therefore, Theorem 4.6.4 yields for arbitrary $x^h, y^h \in Y^{*h}$ that

$$\begin{aligned} |(\iota^* x^h, \iota^*(\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h})y^h)_{X^*}| &= |(\iota^* x^h, \iota^*(z' - z^h))_{X^*}| \\ &\leq C_{\mathrm{reg}}(1 + \tfrac{\tau}{\alpha}\|\iota^*\|^2_{\mathscr{L}(Y^*,X^*)})h^{\mathring{\gamma}}\|\iota^* x^h\|_{X^*}\|y^h\|_{Y^*}. \qquad \square \end{aligned}$$

4.6.2. A priori error estimates for the L-BFGS operator

Now we are interested in error estimates for the inverse of the L-BFGS operator $\tilde{F}_{\mathrm{BFGS}}$. In contrast to $(\tilde{D}_\tau \cdot, \cdot)_{Y^*}$, the bilinear form $(\tilde{F}_{\mathrm{BFGS}} \cdot, \cdot)_{Y^*}$ is not necessarily coercive. Therefore, we cannot use the Lax-Milgram theorem and Lemma 4.6.2. Since we already know that $\tilde{F}_{\mathrm{BFGS}} \in \mathscr{L}(Y^*)$ is invertible, we do not need the Lax Milgram theorem to show that (4.6.1) is uniquely solvable. However, we need to replace Lemma 4.6.2 with a similar statement. To apply a result of [6], we need the following definition.

DEFINITION 4.6.8 ([6, Def. 4.2]). *A continuous bilinear form* $a : Y^* \times Y^* \to \mathbb{R}$ *is called essentially coercive if, for each sequence* $(u_n)_{n\in\mathbb{N}} \subset Y^*$ *which fulfills* $u_n \rightharpoonup 0$ *and* $\lim_{n\to\infty} a(u_n, u_n) = 0$*, one has* $\lim_{n\to\infty}\|u_n\|_{Y^*} = 0$.

For our purposes we need a slight generalization of [6, Thm. 5.2].

THEOREM 4.6.9. *Let $\mathscr{V}$ be a separable, infinite-dimensional Hilbert space and let $\mathscr{W} \subset \mathscr{V}$ be a dense subset. Let $\mathscr{V}_n$, $n \in \mathbb{N}$, be a finite-dimensional subspaces of $\mathscr{V}$ such that for all $u \in \mathscr{W}$*

$$\operatorname{dist}(u, V_n) = \|u - P_{\mathscr{V}_n} u\|_{\mathscr{V}} \to 0 \qquad \text{as } n \to \infty.$$

Further, let $a : \mathscr{V} \times \mathscr{V} \to \mathbb{R}$ be a continuous, essentially coercive bilinear form which fulfills uniqueness, i.e., for all $u \in \mathscr{V}$ it holds

$$a(u,v) = 0 \ (\forall v \in \mathscr{V}) \quad \Rightarrow \quad u = 0.$$

Then, for a given $L \in \mathscr{V}^*$ and for all $n \in \mathbb{N}$, the variational equations

$$\begin{aligned} &\text{Find } u \in \mathscr{V}: \qquad && a(u,v) = \langle L, v\rangle_{\mathscr{V}^*,\mathscr{V}} \qquad && \text{for all } v \in \mathscr{V} \\ &\text{Find } u_n \in \mathscr{V}_n: \qquad && a(u_n,v) = \langle L, v\rangle_{\mathscr{V}^*,\mathscr{V}} \qquad && \text{for all } v_n \in \mathscr{V} \end{aligned}$$

have unique solutions (denoted by u and u_n, respectively). Furthermore, there exists $n_0 \in \mathbb{N}$ and $C > 0$ such that for each $L \in \mathscr{V}^*$ and each $n \geq n_0$ it holds

$$\|u - u_n\|_{\mathscr{V}} \leq C \|u - P_{\mathscr{V}_n} u\|_{\mathscr{V}} \qquad \text{for all } n \geq n_0.$$

Proof. By adopting [6, Thm. 5.2 (i) $\Rightarrow$(ii)] and [6, Prop. 2.5] using a density argument. ∎

Remark 4.6.10. The case $\mathscr{W} := \mathscr{V}$ is covered in [6, Thm. 5.2].

LEMMA 4.6.11. *The continuous bilinear form $a : Y^* \times Y^* \to \mathbb{R}$, $a(\cdot,\cdot) := (\tilde{F}_{BFGS}\, \cdot, \cdot)_{Y^*}$ is essentially coercive.*

Proof. The definition of $\tilde{F}_{\mathrm{BFGS}}$ yields for arbitrary $u' \in Y^*$ that

$$\begin{aligned} |a(u',u')| &= |(\tilde{F}_{\mathrm{BFGS}}\, u', u')_{Y^*}| \\ &= |\alpha(u',u')_{Y^*} + ((Q_{\mathrm{BFGS}} + \tau R_Y)\iota R_X^{-1}\iota^* u', u')_{Y^*}| \\ &\geq |\alpha(u',u')_{Y^*}| - |((Q_{\mathrm{BFGS}} + \tau R_Y)\iota R_X^{-1}\iota^* u', u')_{Y^*}| \\ &= \alpha\|u'\|_{Y^*}^2 - |\langle \mathscr{K} u', u'\rangle_{Y^{**},Y^*}|, \end{aligned}$$

where the operator $\mathscr{K} := R_{Y^*}(Q_{\mathrm{BFGS}} + \tau R_Y)\iota R_X^{-1}\iota^* \in \mathscr{L}(Y^*, Y^{**})$ is compact. Thus, [6, Thm. 4.3] shows that a is essentially coercive. ∎

In the rest of this section, we work under the assumption that every element of Y^* can be approximated arbitrarily well by elements from Y^{*h} for $h \to 0$.

Assumption 4.6.12. Let Y^* be a separable and infinite dimensional Hilbert space and let Y^{*h} be a finite dimensional subspace of Y^* for $0 < h \leq 1$. We assume that there exists a dense subset $\mathscr{W} \subset Y^*$ such that that $\operatorname{dist}(v', Y^{*h}) \to 0$ as $h \to 0$ for all $v' \in \mathscr{W}$, where $\operatorname{dist}(v', Y^{*h}) := \inf\{\|v' - v^h\|_{Y^{*h}} : v^h \in Y^{*h}\} = \|v' - P_{Y^{*h}} v'\|_{Y^*}$.

Remark 4.6.13. By [87, Cor. 1.12.12], the reflexive space Y^* is separable if and only if Y is separable. Furthermore, by [3, Lem. 9.1], an infinite-dimensional normed space H is separable, if and only if there exist finite-dimensional subspaces H_n, $n \in \mathbb{N}$, such that $H_n \subset H_{n+1}$ for all $n \in \mathbb{N}$ and $\cup_{n\in\mathbb{N}} H_n$ is dense in H. Therefore, the assumption that Y^* is separable is crucial for the existence of a sequence of subspaces $(Y^{*h})_{h\in H}$ such that $\operatorname{dist}(v', Y^{*h}) \to 0$ as $h \to 0$ for all $v' \in Y^*$.

THEOREM 4.6.14. *Under Assumption 4.6.12, there exists a $\mathring{h} > 0$ and a constant $m > 0$ such that, for all $h \in (0, \mathring{h}]$, the operator $\tilde{F}^h_{BFGS} \in \mathscr{L}(Y^{*h})$ is invertible and it holds*

$$\|(\tilde{F}^{-h}_{BFGS} - \tilde{F}^{-1}_{BFGS})y^h\|_{Y^*} \le m\|y^h\|_{Y^*} \qquad \textit{for all } y^h \in Y^{*h}.$$

Proof. Since $\tilde{F}_{\mathrm{BFGS}}$ is invertible, the bilinear form a fulfills uniqueness, i.e., for all $u' \in Y^*$ it holds

$$a(u', v') = 0 \ (\forall v' \in Y^*) \quad \Rightarrow \quad \tilde{F}_{\mathrm{BFGS}}\, u' = 0 \quad \Rightarrow \quad u' = 0.$$

Furthermore, by Lemma 4.6.11, $a : Y^* \times Y^* \to \mathbb{R}$ is essentially coercive and Assumption 4.6.12 holds. Therefore, Theorem 4.6.9 implies that there exists a $\mathring{h} > 0$ and a constant $m' > 0$ such that for all $L'' \in Y^{**}$ and all $h \in (0, \mathring{h}]$ the variational equation

$$\text{Find } z^h \in Y^{*h}: \qquad (\tilde{F}_{\mathrm{BFGS}}\, z^h, w^h)_{Y^*} = \langle L'', w^h\rangle_{Y^{**},Y^*} \qquad \text{for all } w^h \in Y^{*h}$$

has a unique solution $z^h \in Y^{*h}$ and it holds

$$\|z^h - z'\|_{Y^*} \le m'\|z' - P_{Y^{*h}} z'\|_{Y^*}$$

where $z' \in Y^{*h}$ is the unique solution to

$$\text{Find } z' \in Y^*: \qquad (\tilde{F}_{\mathrm{BFGS}}\, z', w')_{Y^*} = \langle L'', w'\rangle_{Y^{**},Y^*} \qquad \text{for all } w' \in Y^*.$$

Let $h \in (0, \mathring{h}]$ and $y^h \in Y^{*h}$ be arbitrary. Choosing $L'' = R_{Y^*} y^h$ yields

$$(\tilde{F}^h_{\mathrm{BFGS}}\, z^h, w^h)_{Y^*} = (\tilde{F}_{\mathrm{BFGS}}\, z^h, w^h)_{Y^*} = \langle L'', w^h\rangle_{Y^{**},Y^*} = (y^h, w^h)_{Y^*} \qquad \text{for all } w^h \in Y^{*h}.$$

Since $y^h \in Y^{*h}$ was arbitrary, $\tilde{F}^h_{\mathrm{BFGS}} \in \mathscr{L}(Y^{*h})$ is invertible and $z^h = \tilde{F}^{-h}_{\mathrm{BFGS}} y^h$. Analogously, $z' = \tilde{F}^{-1}_{\mathrm{BFGS}} y^h$ and Lemma 2.1.7 yields for all $y^h \in Y^{*h}$ that

$$\|(\tilde{F}^{-h}_{\mathrm{BFGS}} - \tilde{F}^{-1}_{\mathrm{BFGS}})y^h\|_{Y^*} = \|z^h - z'\|_{Y^*} \le m'\|z' - P_{Y^{*h}} z'\|_{Y^*} \le m'\|z'\|_{Y^*} \le m'\|\tilde{F}^{-1}_{\mathrm{BFGS}}\|_{\mathscr{L}(Y^{*h})}\|y^h\|_{Y^*}. \quad \square$$

Remark 4.6.15. Since the existence of the constant m' in Theorem 4.6.9 is derived via a proof by contradiction, we cannot say anything about the dependence of τ on the constant m'.

Now, let us show the following regularity result:

LEMMA 4.6.16. *If the adjoint problem* (4.6.7) *is $\mathring{\gamma}$-regular relative to $(Y^{*h})_h$, then the problem*

$$\textit{Find } \Phi'_x \in Y^*: \qquad (\Phi'_x, \tilde{F}_{BFGS}\, w')_{Y^*} = (\iota^* x^h, \iota^* w')_{X^*} \qquad \textit{for all } w' \in Y^* \tag{4.6.8}$$

is $\mathring{\gamma}$-regular relative to $(Y^{*h})_h$.

Proof. Assume that (4.6.7) is $\mathring{\gamma}$-regular and that Φ'_x solves (4.6.8). The $\mathring{\gamma}$-regularity of (4.6.7) implies there exists a constant $C_{\mathrm{reg}} \ge 0$ such that for all $0 < h \le 1$ and $x^h \in Y^{*h}$ there exists an element $\Psi^h_x \in Y^{*h}$

which fulfills

$$\|\Psi'_x - \Psi^h_x\|_{Y^*} \le C_{\text{reg}} h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*},$$

where Ψ'_x solves (4.6.7) with right hand side x^h. Now let $0 < h \le 1$ and $x^h \in Y^{*h}$ be arbitrary but fixed. By (4.4.11), the operator $\tilde{F}_{\text{BFGS}}$ has the form $\tilde{F}_{\text{BFGS}} = \tilde{D}_{\tau+\mu} + U^h V^h \iota R_X^{-1} \iota^*$. As the operator $\tilde{D}_{\tau+\mu}$ is Hilbert space self-adjoint and Φ'_x solves the adjoint problem (4.6.8), we find for all $w' \in Y^*$ that

$$\begin{aligned}
(\tilde{D}_{\tau+\mu}\Phi'_x, w')_{Y^*} &= (\Phi'_x, \tilde{D}_{\tau+\mu} w')_{Y^*} \\
&= (\Phi'_x, \tilde{F}_{\text{BFGS}}\, w')_{Y^*} - (\Phi'_x, U^h V^h \iota R_X^{-1} \iota^* w')_{Y^*} \\
&= (\iota^* x^h, \iota^* w')_{X^*} - \sum_{l=1}^{2\tilde{L}} (u^h_l, \Phi'_x)_{Y^*} (\iota^* v^h_l, \iota^* w')_{X^*} \\
&= (\iota^* x^h, \iota^* w')_{X^*} - \sum_{l=1}^{2\tilde{L}} (\iota^* \tilde{F}^{-1}_{\text{BFGS}}\, u^h_l, \iota^* x^h)_{X^*} (\iota^* v^h_l, \iota^* w')_{X^*} \\
&= (\iota^* c^h, \iota^* w')_{X^*},
\end{aligned}$$

where $c^h := x^h - \sum_{l=1}^{2\tilde{L}} (\iota^* \tilde{F}^{-1}_{\text{BFGS}}\, u^h_l, \iota^* x^h)_{X^*} v^h_l \in Y^{*h}$. This shows that Φ'_x solves the problem (4.6.7) with right hand side c^h. Therefore, the $\mathring{\gamma}$-regularity of (4.6.7) implies that there exists an element $\Psi^h_c \in Y^{*h}$ with fulfills

$$\|\Phi'_x - \Psi^h_c\|_{Y^*} \le C_{\text{reg}} h^{\mathring{\gamma}} \|\iota^* c^h\|_{X^*}.$$

Consequently, we conclude

$$\begin{aligned}
\|\Phi'_x - \Psi^h_c\|_{Y^*} &\le C_{\text{reg}} h^{\mathring{\gamma}} \left\| \iota^* x^h - \sum_{l=1}^{2\tilde{L}} (\iota^* \tilde{F}^{-1}_{\text{BFGS}}\, u^h_l, \iota^* x^h)_{X^*} \iota^* v^h_l \right\|_{X^*} \\
&\le C_{\text{reg}} \left(1 + \sum_{l=1}^{2\tilde{L}} \|\iota^* \tilde{F}^{-1}_{\text{BFGS}}\, u^h_l\|_{X^*} \|\iota^* v^h_l\|_{X^*} \right) h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*}. \qquad \square
\end{aligned}$$

THEOREM 4.6.17 (A priori estimate for $\tilde{F}^{-1}_{\text{BFGS}}$). *If Assumption 4.6.12 holds and the adjoint problem (4.6.7) is $\mathring{\gamma}$-regular, then there exists a $\mathring{h} \in (0,1]$ and a constant $C \ge 0$ such that for all $h \in (0, \mathring{h}]$ the operator $\tilde{F}^h_{BFGS}$ is invertible and it holds*

$$|(\iota^* x^h, \iota^* (\tilde{F}^{-1}_{BFGS} - \tilde{F}^{-h}_{BFGS}) y^h)_{X^*}| \le C h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*} \|y^h\|_{Y^*} \qquad \text{for all } x^h, y^h \in Y^{*h}.$$

In particular, $e^{F^{-1}_{BFGS}}_{x^h,y^h} := C h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*} \|y^h\|_{Y^*}$ *fulfills Assumption 4.4.6 for all* $h \in (0, \mathring{h}]$.

Proof. Let $x^h, y^h \in Y^{*h}$ be fixed. By Theorem 4.6.14, there exists a $\mathring{h} > 0$ and a constant $m > 0$ such that, for all $h \le \mathring{h}$, the operator $\tilde{F}^h_{\text{BFGS}} \in \mathscr{L}(Y^{*h})$ is invertible and it holds

$$\|z' - z^h\|_{Y^*} = \|(\tilde{F}^{-h}_{\text{BFGS}} - \tilde{F}^{-1}_{\text{BFGS}}) y^h\|_{Y^*} \le m \|y^h\|_{Y^*},$$

where $z' := \tilde{F}^{-1}_{\text{BFGS}} y^h \in Y^*$ and $z^h := \tilde{F}^{-h}_{\text{BFGS}} y^h \in Y^{*h}$. Observe that $a : Y^* \times Y^* \to \mathbb{R}$, $a(v', w') :=$

$(\tilde{F}_{\mathrm{BFGS}}\, v', w')_{Y^*}$ is a bounded bilinear form which is bounded by the constant $M := \|\tilde{F}_{\mathrm{BFGS}}\|_{\mathscr{L}(Y^*)}$ and $z' \in Y^*$, $z^h \in Y^{*h}$ are the solutions to (4.6.1) and (4.6.2), respectively. Let $\Phi'_x \in Y^*$ be the solution to the adjoint equation

$$(w', \tilde{F}_{\mathrm{BFGS}}\, \Phi'_x)_{Y^*} = (\iota^* x^h, \iota^* w')_{X^*} \qquad \text{for all } w' \in Y^*.$$

By Lemma 4.6.16, the adjoint equation is $\mathring{\gamma}$-regular, i.e., there exists a constant $C_{\mathrm{reg}} \geq 0$ such that for all $0 < h \leq 1$ and $x^h \in Y^{*h}$ there exists an element $\Phi^h_x \in Y^{*h}$ such that

$$\|\Phi'_x - \Phi^h_x\|_{Y^*} \leq C_{\mathrm{reg}} h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*}.$$

Combining this with the Aubin-Nietsche trick (Theorem 4.6.1) yields for all $h \in (0, \mathring{h}]$, that

$$\begin{aligned}
|(\iota^* x^h, \iota^* (\tilde{F}^{-1}_{\mathrm{BFGS}} - \tilde{F}^{-1}_{\mathrm{BFGS}}) y^h)_{X^*}| &= |(\iota^* x^h, \iota^* (z' - z^h))_{X^*}| \\
&= |(z' - z^h, \tilde{F}_{\mathrm{BFGS}}\, \Phi'_x)_{Y^*}| \\
&\leq \|\tilde{F}_{\mathrm{BFGS}}\|_{\mathscr{L}(Y^*)} \|z' - z^h\|_{Y^*} \|\Phi'_x - \Phi^h_x\|_{Y^*} \\
&\leq m C_{\mathrm{reg}} \|\tilde{F}_{\mathrm{BFGS}}\|_{\mathscr{L}(Y^*)}\, h^{\mathring{\gamma}} \|\iota^* x^h\|_{X^*} \|y^h\|_{Y^*}.
\end{aligned}$$

□

5. Optimal control of the obstacle problem

5.1. The obstacle problem

Let $\Omega \subset \mathbb{R}^2$ be open and bounded. Let $\psi : \Omega \to \mathbb{R}$ be a quasi-upper-semicontinuous obstacle (cf. Section 2.8) which defines the set

$$K := \{y \in H_0^1(\Omega) : y \geq \psi \text{ a.e. on } \Omega\}. \tag{5.1.1}$$

We assume that ψ is such that $K \neq \emptyset$. The set K is a closed and convex subset of the Hilbert space $H_0^1(\Omega)$. The obstacle problem is given by the variational inequality

$$\text{Find } y \in K : \qquad \langle Ay - b, v - y\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \qquad \forall v \in K, \tag{5.1.2}$$

where $A \in \mathscr{L}(H_0^1(\Omega), H^{-1}(\Omega))$ is assumed to be coercive and $b \in H^{-1}(\Omega)$. By Theorem 2.4.2, this VI has a unique solution and the solution operator $S : H^{-1}(\Omega) \to H_0^1(\Omega)$, $b \mapsto y$, is Lipschitz continuous. Under a regularity assumption on the obstacle, $H^2(\Omega)$-regularity of the solution of the obstacle problem can be inferred.

LEMMA 5.1.1 ([67, Chap. IV, Thm. 2.3]). *If $u \in L^2(\Omega)$ and the obstacle $\psi \in H^1(\Omega)$ is such that* $\max\{-\Delta\psi - u, 0\} \in L^2(\Omega)$*, then it holds that $S(\iota u) \in H^2(\Omega)$, where ι is the embedding from $L^2(\Omega)$ to $H^{-1}(\Omega)$.*

If one wishes to consider an upper obstacle $\tilde{\psi}$ and search for functions $\tilde{y} \in H_0^1(\Omega)$ with $\tilde{y} \leq \tilde{\psi}$ instead of the lower obstacle (5.1.1), a simple transform will do the trick.

LEMMA 5.1.2. *If $y \in H_0^1(\Omega)$ is a solution to* (5.1.2)*, then $\tilde{y} := -\bar{y}$ is a solution to the obstacle problem*

$$\textit{Find } \tilde{y} \in \tilde{K} : \qquad \langle A\tilde{y} - \tilde{b}, \tilde{v} - \tilde{y}\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \qquad \forall \tilde{v} \in \tilde{K},$$

where $\tilde{b} := -b$, $\tilde{\psi} := -\psi$ and $\tilde{K} := \{\tilde{y} \in H_0^1(\Omega) : \tilde{y} \leq \tilde{\psi}$ a.e. on $\Omega\}$.

Proof. Can be easily verified by observing that $\tilde{K} = -K$. ∎

LEMMA 5.1.3 ([115, Chap. 4 Prop. 5.6]). *The obstacle problem* (5.1.2) *is equivalent to the complementary system*

$$\textit{Find } (y,\xi) \in H_0^1(\Omega) \times H^{-1}(\Omega) : \qquad \xi = Ay - b,\ y \geq \psi,\ \xi \geq 0,\ \langle \xi, y - \psi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0,$$

where the inequality $\xi \geq 0$ is to be understood in the dual space, i.e.,

$$\langle \xi, v^+\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \quad \textit{for all } v^+ \in H_0^1(\Omega)^+ := \{v \in H_0^1(\Omega) : v \geq 0 \textit{ a.e. on } \Omega\}.$$

5.2. The optimal control problem

We want to apply the bundle method to optimal control of the obstacle problem, i.e., the problem

$$\begin{aligned} \underset{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)}{\text{minimize}} \quad & J(y) + \tfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \\ \text{subject to} \quad & y = S(\mathring{F}(\iota u)) \end{aligned} \tag{5.2.1}$$

where $J : H_0^1(\Omega) \to \mathbb{R}$ is the Fréchet-differentiable objective function (corresponding to the state), the parameter $\alpha > 0$ defines the Tikhonov regularization term $\frac{\alpha}{2}\|\cdot\|^2_{L^2(\Omega)}$, $\iota \in \mathscr{L}(L^2(\Omega), H^{-1}(\Omega))$ is a compact embedding from $L^2(\Omega)$ to $H^{-1}(\Omega)$, $\mathring{F} : H^{-1}(\Omega) \to H^{-1}(\Omega)$ maps the control to the force term and $S : H^{-1}(\Omega) \to H_0^1(\Omega)$ is the solution operator of the obstacle problem (5.1.2). Further requirements on the objective function J and on $\mathring{F}$ are stated below as needed. Plugging the solution operator into the objective function yields the reduced problem

$$\underset{u\in L^2(\Omega)}{\text{minimize}} \quad J(S(\mathring{F}(\iota u))) + \tfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)}. \tag{5.2.2}$$

Obviously, Problem (5.2.1) and Problem (5.2.2) are equivalent.

THEOREM 5.2.1. *If $\mathring{F}$ is given as $\mathring{F}(w) := w + \mathring{f}$ with arbitrary $\mathring{f} \in H^{-1}(\Omega)$ and $J : H_0^1(\Omega) \to \mathbb{R}$ is bounded below and lower semicontinuous, then* Problem (5.2.1) *has a solution.*

Proof. The joint objective functional $j : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$, $j(y,u) := J(y) + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}$ is bounded below. As the Tikhonov regularization term $\frac{\alpha}{2}\|\cdot\|^2_{L^2(\Omega)}$ is convex and continuous, it is weakly lower semicontinuous. Thus, j is strong$\times$weak sequentially lower semicontinuous. As $\frac{\alpha}{2}\|\cdot\|^2_{L^2(\Omega)}$ is coercive, Remark 2.5.2 implies that the reduced objective function $J : u \mapsto j(S(\iota u + \mathring{f}), u)$ is coercive. Therefore, Theorem 2.5.1 yields the existence of a solution to Problem (5.2.1). ∎

Example 5.2.2 (Tracking type objective function). By choosing the tracking type objective function $J : H_0^1(\Omega) \to \mathbb{R}$, $J(w) := \frac{1}{2}\|Ow - y_d\|^2_H$, where the observation operator $O \in \mathscr{L}(H_0^1(\Omega), H)$ and the desired state $y_d \in H$ are defined on Hilbert space H, we obtain the tracking type optimal control problem

$$\underset{u\in L^2(\Omega)}{\text{minimize}} \quad \tfrac{1}{2}\|OS(\iota u + \mathring{f}) - y_d\|^2_H + \tfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)}. \tag{5.2.3}$$

Here, we chose $\mathring{F} : H^{-1}(\Omega) \to H^{-1}(\Omega)$ to be $\mathring{F}(w) := w + \mathring{f}$, where $\mathring{f} \in H^{-1}(\Omega)$ is an arbitrary external force. As J is obviously bounded below and continuous, Theorem 5.2.1 implies that the optimal control problem governed by the obstacle problem with tracking type objective function has a solution.

For the subsequent analysis, we define the active set $\mathscr{A}(w) \subset \Omega$ at a point $w \in H^{-1}(\Omega)$ to be the set at which the solution to the obstacle problem coincides with the obstacle, i.e.,

$$\mathscr{A} = \mathscr{A}(w) := \{\omega \in \Omega : S(\mathring{F}(w))(\omega) = \psi(\omega)\}. \tag{5.2.4}$$

Further, we define the inactive set $\mathscr{I} \subset \Omega$ to be the complement of the active set, i.e.,

$$\mathscr{I} = \mathscr{I}(w) := \{\omega \in \Omega : S(\mathring{F}(w))(\omega) > \psi(\omega)\}. \tag{5.2.5}$$

As ψ is assumed to be quasi-upper-semicontinuous and $S(\mathring{F}(w)) \in H_0^1(\Omega)$, $\psi - S(\mathring{F}(w))$ is quasi-upper-semicontinuous and Lemma 2.8.4 implies that $\mathscr{I}(w)$ is a quasi-open set.

In [131], strong stationarity conditions are derived which do not require any regularity of Ω. As [131] uses the setting of a lower obstacle, we transfer the result [131, Thm. 5.2, Prop. 2.5] to the given upper obstacle setting.

THEOREM 5.2.3. *Let $(\bar{y},\bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ be a local solution to* Problem (5.2.1) *with $\mathring{F} \in \mathscr{L}(H^{-1}(\Omega))$ defined via $\mathring{F}(w) := w + \mathring{f}$, $\mathring{f} \in H^{-1}(\Omega)$. If $\bar{u} \in H_0^1(\Omega)$, then there exist multipliers $(\bar{\xi}, \bar{p}, \bar{\mu}) \in H^{-1}(\Omega) \times H_0^1(\Omega) \times H^{-1}(\Omega)$ which satisfy the strong optimality conditions*

$$A\bar{y} = \bar{u} + \mathring{f} + \bar{\xi} \qquad \text{in } H^{-1}(\Omega), \tag{5.2.6a}$$
$$A^*\bar{p} = J'(\bar{y}) - \bar{\mu} \qquad \text{in } H^{-1}(\Omega), \tag{5.2.6b}$$
$$\alpha\bar{u} + \bar{p} = 0 \qquad \text{in } L^2(\Omega), \tag{5.2.6c}$$
$$\bar{p} \in \mathscr{K}(\bar{u},\bar{\xi}), \tag{5.2.6d}$$
$$\langle \bar{\mu}, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \qquad \forall v \in \mathscr{K}(\bar{u},\bar{\xi}), \tag{5.2.6e}$$

where the critical cone $\mathscr{K}(\bar{u},\bar{\xi}) \subset H_0^1(\Omega)$ is defined via

$$\mathscr{K}(\bar{u},\bar{\xi}) := \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. in } \mathscr{A}(\bar{u}), \langle \bar{\xi}, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0\}. \tag{5.2.7}$$

Proof. Let $(\bar{y},\bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ be a local solution to Problem (5.2.1) with $\mathring{F} \in \mathscr{L}(H^{-1}(\Omega))$ defined via $\mathring{F}(w) = w + \mathring{f}$, $\mathring{f} \in H^{-1}(\Omega)$. By Lemma 5.1.3, the multiplier $\bar{\xi} := A\bar{y} - \bar{u} - \mathring{f} \in H^{-1}(\Omega)$ fulfills $\bar{\xi} \geq 0$ and $\langle \bar{\xi}, \bar{y} - \psi\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0$. Thus, [131, Prop. 2.5] implies $\bar{\xi} \in T_{\tilde{K}}(-\bar{y})^\circ$ where $T_{\tilde{K}}(-\bar{y})^\circ$ is the polar of the tangent cone (cf. Section 2.4) of the set $\tilde{K} := \{\tilde{y} \in H_0^1(\Omega) : \tilde{y} \leq -\psi \text{ a.e. on } \Omega\}$ at the point $-\bar{y}$. As $-\bar{y} \in \tilde{K}$, the tuple $(-\bar{y}, -\bar{u}, \bar{\xi})$ is feasible for the optimization problem

$$\begin{aligned}
&\text{minimize} && \tilde{j}(\tilde{y}) + \tfrac{\alpha}{2}\|\tilde{u}\|^2_{L^2(\Omega)} \\
&\text{subject to} && (\tilde{y},\tilde{u},\tilde{\xi}) \in H_0^1(\Omega) \times L^2(\Omega) \times H^{-1}(\Omega), \\
& && A\tilde{y} = \tilde{u} - \tilde{\xi} - \mathring{f}, \\
& && \tilde{y} \in \tilde{K}, \quad \tilde{\xi} \in T_{\tilde{K}}(\tilde{y})^\circ,
\end{aligned} \tag{5.2.8}$$

where $\tilde{j} : H_0^1(\Omega) \to \mathbb{R}$ is defined via $\tilde{j}(\tilde{y}) := J(-\tilde{y})$. Furthermore, whenever $(\tilde{y},\tilde{u},\tilde{\xi})$ is feasible for Problem (5.2.8), Lemma 5.1.2 shows that $(-\tilde{y},-\tilde{u})$ is feasible for Problem (5.2.1). Consequently, as $(\bar{y},\bar{u})$ is a local solution to Problem (5.2.1) and $J(\bar{y}) = \tilde{j}(-\bar{y})$, the tuple $(-\bar{y},-\bar{u},\bar{\xi})$ is a local solution to Problem (5.2.8). As Problem (5.2.8) has the form of problem (P) in [131], [131, Thm. 5.2, Prop. 2.5] yields the existence of multipliers $(\bar{p},\bar{\mu}) \in H_0^1(\Omega) \times H^{-1}(\Omega)$ such that

$$\begin{aligned}
A^*\bar{p} + \tilde{j}'(-\bar{y}) + \bar{\mu} &= 0 \quad \text{in } H^{-1}(\Omega), \\
\alpha(-\bar{u}) - \bar{p} &= 0 \quad \text{in } L^2(\Omega), \\
-\bar{p} &\in T_{\tilde{K}}(-\bar{y}) \cap \bar{\xi}^\perp, \\
\bar{\mu} &\in (T_{\tilde{K}}(-\bar{y}) \cap \bar{\xi}^\perp)^\circ.
\end{aligned}$$

Note that $\tilde{J}'(-\bar{y}) = -J(\bar{y})$. We conclude the proof by observing that [131, Prop. 2.5] implies

$$\begin{aligned} T_{\tilde{K}}(-\bar{y}) \cap \bar{\xi}^{\perp} &= \{z \in H_0^1(\Omega) : z \leq \text{q.e. on } -\bar{y} = -\psi, \langle \bar{\xi}, z\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0\} \\ &= -\{z \in H_0^1(\Omega) : z \geq \text{q.e. on } \mathscr{A}(\bar{u}), \langle \bar{\xi}, z\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0\} \\ &= -\mathscr{K}(\bar{u}, \bar{\xi}). \end{aligned}$$

□

The critical cone $\mathscr{K}(\bar{u}, \bar{\xi})$ can be characterized in the following way. By [109, Thm. 3.9] there exists a quasi-closed set $\mathscr{A}_s(\bar{u}) \subset \mathscr{A}(\bar{u})$ which is unique up to capacity zero such that

$$\begin{aligned} \mathscr{K}(\bar{u}, \bar{\xi}) &= \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. in } \mathscr{A}(\bar{u}), \langle \xi, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = 0\} \\ &= \{v \in H_0^1(\Omega) : v \geq 0 \text{ q.e. on } \mathscr{A}(\bar{u}), v = 0 \text{ q.e. on } \mathscr{A}_s(\bar{u})\}. \end{aligned} \tag{5.2.10}$$

The set $\mathscr{A}_s(\bar{u})$ is called the strictly active set and plays an important role in the computation of a subgradient, cf. Lemma 5.2.6 and Section 5.3.4 below.

We apply the bundle method to solve Problem (5.2.2); the problem is of the form (3.1.1) with $X^* = X := L^2(\Omega)$, $Y := H^{-1}(\Omega)$, $Y^* = H_0^1(\Omega)$, $p : Y \to \mathbb{R}$, $p := J(S(\mathring{F}(\cdot)))$, $f : U \to \mathbb{R}$, $f = p \circ \iota$, $w : X \to \mathbb{R}$, $w := \frac{\alpha}{2}\|\cdot\|^2_{L^2(\Omega)}$ and $\mathscr{F} := L^2(\Omega)$. To execute the bundle method, we need to compute an approximation of a subgradient of $p = J(S(\mathring{F}(\cdot)))$. In the following we study how to obtain an exact subgradient. Recall that $(\cdot)_+ : H_0^1(\Omega) \to H_0^1(\Omega)$ denotes the pointwise maximum, i.e., $(v)_+(\omega) := \max(0, v(\omega))$.

DEFINITION 5.2.4 ([115, Chap. 4.5]). *An operator $A : H^{-1}(\Omega) \to H_0^1(\Omega)$ is called strictly T-monotone if A satisfies*

$$\langle Av, (v)_+\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} > 0 \qquad \textit{for all } v \in H_0^1(\Omega) \textit{ with } (v)_+ \neq 0.$$

Example 5.2.5. The operator $A = -\Delta$ induced by the negative Laplace operator via

$$\langle Ay, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} := \int_\Omega \nabla y^\top \nabla v \, \mathrm{d}\lambda \qquad \text{for all } y, v \in H_0^1(\Omega)$$

is strictly T-monotone and coercive.

LEMMA 5.2.6. *Let $J : H_0^1(\Omega) \to \mathbb{R}$ be a continuously differentiable function and let $\mathring{F} : H^{-1}(\Omega) \to H^{-1}(\Omega)$ be a Lipschitz continuous, continuously differentiable, monotone function. Furthermore, let $A \in \mathscr{L}(H_0^1(\Omega), H^{-1}(\Omega))$ be a coercive and strictly T-monotone operator and let $w \in H^{-1}(\Omega)$ be arbitrary. Denote by $q \in H_0^1(\Omega)$ the unique solution of the variational equation*

$$\textit{Find } q \in H_0^1(D): \quad \langle A^* q, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = \langle J'(S(\mathring{F}(w))), v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}, \quad \forall v \in H_0^1(D), \tag{5.2.11}$$

where $D = \mathscr{I}(w)$ (cf. (5.2.5)) or $D = \Omega \setminus \mathscr{A}_s$ (cf. (5.2.10)). Then it holds that $\mathring{F}'(w)^* q \in \partial_C p(w)$.

Proof. This is an application of [109, Thm. 4.21]. □

5.3. Discretization

In the rest of this chapter, we consider the optimal control problem with tracking type objective

$$\begin{aligned}
&\underset{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)}{\text{minimize}} && \tfrac{1}{2}\|y-y_d\|_{L^2(\Omega)}^2 + \tfrac{\alpha}{2}\|u\|_{L^2(\Omega)}^2 \\
&\text{subject to} && y\in K, \quad \int_\Omega \nabla y^T \nabla(v-y)\,\mathrm{d}\lambda \geq \int_D (u+\mathring{f})(v-y)\,\mathrm{d}\lambda \quad \forall v\in K,
\end{aligned} \tag{5.3.1}$$

where $K := \{y\in H_0^1(\Omega) : y\geq \psi \text{ a.e. on } \Omega\}$, $\psi\in H^1(\Omega)\cap C^0(\mathrm{cl}(\Omega))$, $\mathring{f}\in L^2(\Omega)$ and $y_d\in L^2(\Omega)$. This problem is of the form Problem (5.2.1) with $J := \frac{1}{2}\|\iota^*(\cdot)-y_d\|_{L^2(\Omega)}$ and $\mathring{F}(w) := w + \iota\mathring{f}$ for all $w\in H^{-1}(\Omega)$, cf. Examples 5.2.2 and 5.2.5. We set $Y := H^{-1}(\Omega)$, $Y^* = H_0^1(\Omega)$, $X^* := L^2(\Omega)$ and $X := L^2(\Omega)^*$. In the following, we identify $L^2(\Omega)^*$ with $L^2(\Omega)$ and the Riesz map $R_X : L^2(\Omega)\to L^2(\Omega)$, $R_X(x) = x$, is explicitly written only if the connection to previous results is to be pointed out. Furthermore, the embedding $\iota^* : H_0^1(\Omega)\to L^2(\Omega)$, $\iota^*(x)(\omega) := x(\omega)$, is explicitly written only if it is necessary to do so.

In order to apply the bundle method (Algorithm 3.4) to this problem, we need approximations of the function value, a subgradient, and the minimizing iterate. To do so, we employ the finite element method, cf. Section 2.7. We start at an initial triangulation $\mathscr{T}^{h_0}$ of Ω with initial mesh width $h_0\in(0,1]$. We use the corresponding finite element space $V^{h_0}\subset H_0^1(\Omega)$ to discretize $Y^* = H_0^1(\Omega)$. Using this discretization, function values, subgradients and minimizing iterates with corresponding error estimates are constructed as described below. Whenever a computed error estimate exceeds the bounds needed for Algorithm 3.4, we refine the triangulation of Ω in such a way that this error estimate is reduced. Each refinement is done in such a way that the constructed sequence of triangulations $(\mathscr{T}^h)_h$ is regular (cf. Section 2.7). This leads to a sequence of finite element spaces $(V^h)_h$ which are nested, i.e.,

$$V^h \subset V^{\tilde{h}} \qquad \text{for all } \tilde{h}\leq h. \tag{5.3.2}$$

5.3.1. Discretization of the obstacle problem

To obtain an approximation of the objective function $f = J(S(\mathring{F}(\iota(\cdot))))$, we first approximate the solution operator S of the obstacle problem. Any triangulation $\mathscr{T}^h$ yields a corresponding finite element space $V^h\subset H_0^1(\Omega)$. Define

$$K^h := \{v^h\in V^h : v^h \geq I^h\psi \ \text{ a.e. on } \Omega\},$$

where $I^h : C^0(\mathrm{cl}(\Omega))\to V^h$ denotes the Lagrange interpolation operator. We discretize the obstacle problem (5.1.2) by

$$\text{Find } y^h\in K^h : \qquad \langle Ay^h - b, v^h - y^h\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \geq 0 \qquad \forall v^h\in K^h. \tag{5.3.3}$$

As (5.3.3) is a variational inequality, Theorem 2.4.2 shows the following.

LEMMA 5.3.1. *For all $b\in H^{-1}(\Omega)$ the variational inequality* (5.3.3) *has a unique solution $y^h\in K^h$. The solution operator $S^h : H^{-1}(\Omega)\to V^h$ is globally Lipschitz continuous and the Lipschitz modulus does not depend on the space V^h.*

Denote by $\mathsf{y},\mathsf{v},\Psi \in \mathbb{R}^{n_V}$ the coordinates of $y^h, v^h, I^h\psi \in V^h$ and define the set

$$\mathsf{K}_\psi := \{\mathsf{v} \in \mathbb{R}^{n_V} : \mathsf{v} \geq \Psi \text{ componentwise }\}.$$

As A is induced by the negative Laplace operator, it holds

$$\langle Ay^h, v^h - y^h\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = (y^h, v^h - y^h)_{H_0^1(\Omega)} = \mathsf{y}^\top \mathsf{K}(\mathsf{v}-\mathsf{y}),$$

where K is the stiffness matrix introduced in Section 2.7. We define the vector $\hat{\mathsf{b}} \in \mathbb{R}^{n_V}$ by $\hat{\mathsf{b}}_i := \langle b, \phi_i\rangle_{H_0^1(\Omega)^*,H_0^1(\Omega)}$. This gives

$$\langle b, v^h - y^h\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = \hat{\mathsf{b}}^\top(\mathsf{v}-\mathsf{y}).$$

If $b \in H^{-1}(\Omega)$ has the form $b = (b^h, \cdot)_{L^2(\Omega)}$ with $b^h \in U^h$ and coordinates $\mathsf{b} \in \mathbb{R}^{n_U}$, then

$$\hat{\mathsf{b}}^\top \mathsf{v} = \langle b, v^h\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = (b^h, v^h)_{L^2(\Omega)} = \mathsf{b}^\top \mathsf{M}\mathsf{I}\mathsf{v} \quad \text{for all } v^h \in V^h \text{ with coordinates } \mathsf{v} \in \mathbb{R}^{n_V},$$

i.e., $\hat{\mathsf{b}} = \mathsf{I}^\top \mathsf{M}\mathsf{b}$, where the mass matrix $\mathsf{M} \in \mathbb{R}^{n_U \times n_U}$ and the prolongation matrix $\mathsf{I} \in \mathbb{R}^{n_U \times n_V}$ were defined in Section 2.7. For arbitrary $b^h \in L^2(\Omega)$, i.e., if $b \in H^{-1}(\Omega)$ has the form $b = (b^h, \cdot)_{L^2(\Omega)}$, we obtain

$$\hat{\mathsf{b}}_i = \langle b, \phi_i\rangle_{H_0^1(\Omega)^*,H_0^1(\Omega)} = (b^h, \phi_i)_{L^2(\Omega)} = \int_\Omega b^h \phi_i \,\mathrm{d}\lambda \tag{5.3.4}$$

and numerical integration techniques (cf. Section 6.1) can be used to compute the vector $\hat{\mathsf{b}} \in \mathbb{R}^{n_V}$. With these definitions, (5.3.3) is equivalent to

$$\text{Find } \mathsf{y} \in \mathsf{K}_\psi : \qquad (\mathsf{K}\mathsf{y} - \hat{\mathsf{b}})^\top(\mathsf{v}-\mathsf{y}) \geq 0 \qquad \forall \mathsf{v} \in \mathsf{K}_\psi. \tag{5.3.5}$$

By Lemma 2.4.3, this finite dimensional VI is equivalent to finding $\mathsf{y} \in \mathsf{K}_\psi$ such that

$$\mathsf{y} = P_{\mathsf{K}_\psi}(\mathsf{y} - \gamma(\mathsf{K}\mathsf{y} - \hat{\mathsf{b}})). \tag{5.3.6}$$

In contrast to $P_{K^h} : V^h \to V^h \subset H_0^1(\Omega)$, the finite dimensional projection $P_{\mathsf{K}_\psi} : \mathbb{R}^{n_V} \to \mathbb{R}^{n_V}$ can be efficiently computed via

$$P_{\mathsf{K}_\psi}(\mathsf{v}) = \max\{\mathsf{v}, \Psi\} \qquad \text{for all } \mathsf{v} \in \mathbb{R}^{n_V},$$

where the max operator is to be understood coordinatewise. As $P_{\mathsf{K}_\psi} : \mathbb{R}^{n_V} \to \mathbb{R}^{n_V}$ is piecewise linear in every coordinate, it is semismooth (cf. [124, Def. 2.5, Prop. 2.26]). Thus, (5.3.6) can be solved via a semismooth Newton method (cf. [124, Alg. 2.11]) which inhibits a locally q-superlinear convergence behavior (cf. [124, Prop. 2.12]). However, since this algorithm is based on the finite dimensional reformulation (5.3.6), it might not be mesh independent.

Another approach to compute a solution of the obstacle problem (5.1.2) is to use the semismooth Newton method in Hilbert space proposed in [124, Chap. 9.2]. There, a regularized dual problem is solved. However, this regularization leads to additional error terms. We thus use a hybrid algorithm. The last trial iterate of the bundle method is used as a starting iterate y_0. Whenever the norm of the residual $\mathsf{y}_0 - P_{\mathsf{K}_\psi}(\mathsf{y}_0 - \gamma(\mathsf{K}\mathsf{y}_0 - \hat{\mathsf{b}}))$ is larger than a prescribed bound, we use the semismooth Newton method in Hilbert space with a large regularization parameter to obtain an approximation of the solution $\tilde{\mathsf{y}}_0$. We then use the semismooth Newton method in finite dimensions with the starting point $\tilde{\mathsf{y}}_0$ to gain an accurate solution of (5.3.6).

5.3.2. Computation of function value approximations

We now construct a computable function value oracle $O_f : V^h \times (0,1] \to \mathbb{R}$ which can be used in Algorithms 3.2 and 3.3 to compute a function value approximation and a subgradient, respectively. A computable error bound $\hat{\varepsilon}_f : V^h \times (0,1] \to (0,\infty)$ which fulfills (3.5.1) is constructed in Section 5.4.1 for the case of uniform mesh refinement and in Section 5.5.1 for the case of adaptive mesh refinement. Recall that $f(u) = \frac{1}{2}\|S(u+\mathring{f}) - y_d\|^2_{L^2(\Omega)}$. We approximate $S(u^h+\mathring{f})$ via $S^h(u^h+\mathring{f})$ (cf. Section 5.3.1). This yields the function value oracle

$$\begin{aligned} O_f(u^h,h) &:= \tfrac{1}{2}\|S^h(u^h+\mathring{f}) - y_d\|^2_{L^2(\Omega)} \\ &= \tfrac{1}{2}\|S^h(u^h+\mathring{f})\|^2_{L^2(\Omega)} - (S^h(u^h+\mathring{f}), y_d)_{L^2(\Omega)} + \tfrac{1}{2}\|y_d\|^2_{L^2(\Omega)}. \end{aligned} \tag{5.3.7}$$

Since $S^h(u^h+\mathring{f}) \in V^h$, the term $\frac{1}{2}\|S^h(u^h+\mathring{f})\|^2_{L^2(\Omega)}$ can be computed exactly. Furthermore, the term $\frac{1}{2}\|y_d\|^2_{L^2(\Omega)}$ can be computed analytically prior to execution of the algorithm. In order to evaluate the term $(S^h(u^h+\mathring{f}), y_d)_{L^2(\Omega)}$, we use numerical integration (cf. Section 6.1) to compute

$$\mathsf{y_d} \in \mathbb{R}^{n_U}, \qquad (\mathsf{y_d})_i := \int_\Omega \phi_i\, y_d \,\mathrm{d}\lambda, \tag{5.3.8}$$

where ϕ_i is the nodal basis function corresponding to node n_i, $1 \le i \le n_U$. If $\mathsf{y} \in \mathbb{R}^{n_U}$ denotes the nodal values of $y^h := S^h(u^h+\mathring{f}) \in V^h$, then we obtain

$$(S^h(u^h+\mathring{f}), y_d)_{L^2(\Omega)} = \sum_i (\mathsf{y}_i\phi_i, y_d)_{L^2(\Omega)} = \mathsf{y}^\top \mathsf{y_d},$$

i.e., the term $(S^h(u^h+\mathring{f}), y_d)_{L^2(\Omega)}$ can be evaluated exactly. Therefore, the function value oracle $O_f(u^h,h)$ is computable. The integrals in (5.3.8) are to be computed numerically in such a way that the integration error is negligible, cf. Section 6.1.

Computation of an error estimate for the function value approximation

In Sections 5.4.1 and 5.5.1, we develop a priori and a posterior error estimates for the $H^1_0(\Omega)$-error of the state, i.e., for $\|S(u+\mathring{f}) - S^h(u+\mathring{f})\|_{H^1_0(\Omega)}$. Here, we show how this leads to an error estimate for the function value.

LEMMA 5.3.2. *For any $e_S(u) \in \mathbb{R}$ which fulfills $\|S(u+\mathring{f}) - S^h(u+\mathring{f})\|_{H^1_0(\Omega)} \le e_S(u)$, it holds*

$$|O_f(u,h) - f(u^h)| \le \hat{\varepsilon}_f(u,h) := C_{F,\Omega}\|S^h(u+\mathring{f}) - y_d\|_{L^2(\Omega)} e_S(u) + \tfrac{1}{2}C^2_{F,\Omega} e_S(u)^2.$$

Proof. Using the abbreviations $a := S(u+\mathring{f}) - y_d \in H_0^1(\Omega)$ and $a^h := S^h(u+\mathring{f}) - y_d \in H_0^1(\Omega)$ we get

$$|O_f^{\mathrm{uf}}(u,h) - f(u^h)| = \tfrac{1}{2}|(a+a^h, a-a^h)_{L^2(\Omega)}| \leq \|a^h\|_{L^2(\Omega)}\|a-a^h\|_{L^2(\Omega)} + \tfrac{1}{2}\|a-a^h\|^2_{L^2(\Omega)}.$$

Recall that the Friedrichs constant fulfills $C_{F,\Omega} = \|\iota^*\|_{\mathscr{L}(H_0^1(\Omega),L^2(\Omega))}$, cf. Section 2.6. Therefore,

$$\|a-a^h\|_{L^2(\Omega)} \leq \|\iota^*\|_{\mathscr{L}(H_0^1(\Omega),L^2(\Omega))}\|S(u+\mathring{f}) - S^h(u+\mathring{f})\|_{H_0^1(\Omega)} \leq C_{F,\Omega} e_S(u). \qquad \square$$

Note that we can compute $\|S^h(u+\mathring{f}) - y_d\|_{L^2(\Omega)}$ exactly via

$$\begin{aligned}\|S^h(u+\mathring{f}) - y_d\|^2_{L^2(\Omega)} &= \|S^h(u+\mathring{f})\|^2_{L^2(\Omega)} - 2(S^h(u+\mathring{f}), y_d)_{L^2(\Omega)} + \|y_d\|^2_{L^2(\Omega)}\\ &= \mathsf{y}^\top\mathsf{M}\mathsf{y} - 2\mathsf{y}^\top\mathsf{y_d} + \|y_d\|^2_{L^2(\Omega)},\end{aligned}$$

cf. (5.3.8). Therefore, we can compute $\hat{\varepsilon}_f(u,h)$ exactly if $e_S(u)$ can be computed exactly. Furthermore, Lemma 5.3.2 implies that O_f and $\hat{\varepsilon}_f$ fulfill (3.5.1) and thus they can be used in Algorithms 3.2 and 3.3 to compute a function value approximation and a subgradient, respectively.

Computation of a lower bound for the $H^{-1}(\Omega)$-norm

In order to control the size of the lift term of the function value approximation, we need an oracle $\underline{O}_{\|\cdot\|_Y}$ for a lower bound of the $H^{-1}(\Omega)$-norm of an element $x \in H^{-1}(\Omega)$, cf. Section 3.5.1. For $x = R_X^{-1}\iota^* x^h$ with $x^h \in V^h$, we define $\underline{O}_{\|\cdot\|_Y}(x,h) := \|\tilde{E}^h x^h\|_{Y^*}$. The oracle $\underline{O}_{\|\cdot\|_Y}(x,h)$ can be computed exactly by computing the coordinates $\mathsf{z} \in \mathbb{R}^{n_V}$ of $z^h := \tilde{E}^h x^h = P_{Y^{*h}}\tilde{E}x^h$ via the linear system

$$\mathsf{z}^\top\mathsf{K}\mathsf{w} = (z^h, w^h)_{Y^*} = (\iota^* x^h, \iota^* w^h)_{X^*} = \mathsf{x}^\top\mathsf{M}\mathsf{w} \qquad \text{for all } \mathsf{w} \in \mathbb{R}^{n_V}, \tag{5.3.9}$$

cf. (4.3.6). By Lemma 4.5.1, we find $\underline{O}_{\|\cdot\|_Y}(x,h) \leq \|\iota x\|_{H^{-1}(\Omega)}$. Further, for $x = R_X^{-1}\iota^* x^h \neq 0$, the Lax-Milgram theorem, Theorem 2.3.2, shows that $\tilde{E}^h x^h \neq 0$ which implies $\underline{O}_{\|\cdot\|_Y}(x,h) > 0$. Consequently, $\underline{O}_{\|\cdot\|_Y}$ and $C_{\|\cdot\|_Y} := 1$ fulfill (3.5.4) and thus can be used in Algorithm 3.4 as an oracle for the lower bound of the Y-norm.

5.3.3. Constant free error estimates for the Dirichlet problem

In the following sections, we need an error estimate for the solution of the Dirichlet problem. Most error bounds for finite element discretizations involve an unknown constant. Since we need a concrete upper bound for the error, we use the following error bound developed in [112, 95, 111].

THEOREM 5.3.3 ([112, Cor. 3.1]). *Let $D \subset \mathbb{R}^2$ be a bounded domain with Lipschitz continuous boundary ∂D and let $f \in L^2(D)$ be arbitrary. Denote by $u \in H_0^1(D)$ the solution to the Dirichlet problem*

$$\begin{aligned}-\Delta u &= f \qquad \text{in } D,\\ u &= 0 \qquad \text{on } \partial D.\end{aligned}$$

For arbitrary $v \in H_0^1(D)$ there holds

$$\|u-v\|_{H_0^1(D)} \leq \|\nabla v - w\|_{L^2(D)^2} + C_{F,D}\|\operatorname{div} w + f\|_{L^2(D)} \qquad \text{for all } w \in H(D,\operatorname{div}).$$

Here, $C_{F,D} := \sup_{w \in H_0^1(D)} \|w\|_{L^2(D)}/\|w\|_{H_0^1(D)}$ is the Friedrich constant on the domain D.

Known values of $C_{F,D}$ for several domains D can be found in Section 2.6. For a given domain $D \subset \mathbb{R}^2$, we define the majorant M_D via

$$M_D : H_0^1(D) \times H(D,\operatorname{div}) \to \mathbb{R}, \qquad M_D(v,w) := \|\nabla v - w\|_{L^2(D)^2} + C_{F,D}\|\operatorname{div} w + f\|_{L^2(D)}. \tag{5.3.10}$$

As pointed out in [112], there are several ways to find a suitable $w \in H(\Omega,\operatorname{div})$ such that $M_D(v,w)$ is small. For $w = \nabla u$, it holds $M_D(v,w) = \|\nabla(v-u)\|_{L^2(D)^2} = \|u-v\|_{H_0^1(D)}$, i.e., the majorant $M_D(v,\nabla u)$ recovers the error exactly. Since we cannot compute ∇u exactly, one strategy is to construct a $w^h \in H(\Omega,\operatorname{div})$ based on $\nabla P_{V^h} u \in L^2(\Omega,\mathbb{R}^2)$ such that $M(v,w^h)$ can be computed exactly.

A simple approach is given by local post-processing, cf. [111, Chap. 2.6.3]. Let $O_i \subset \mathbb{R}^2$ be the patch around interior node n_i, $1 \leq i \leq n_V$, i.e., $O_i := \cup_{j=1,\dots,m_i} T_{i_j}$, where T_{i_j}, $j = 1,\dots,m_i$, are the triangles adjacent to interior node n_i. Define the gradient averaging operator $G^h : V^h \to V^h \times V^h$ by

$$G^h v^h(x_i) := \sum_{j=1}^{m_i} \frac{|T_{i_j}|}{|O_i|} (\nabla v^h)_{i_j} \qquad \text{for all } v^h \in V^h,$$

where $(\nabla v^h)_{i_j} \in \mathbb{R}^2$ denotes the value of ∇v^h on the triangle T_{i_j}. This leads to the error estimate

$$\|u-v^h\|_{H_0^1(D)} \leq M_D(v^h, G^h v^h) = \|\nabla v^h - G^h v^h\|_{L^2(D)^2} + C_{F,D}\|\operatorname{div} G^h v^h + f\|_{L^2(D)}. \tag{5.3.11}$$

Another strategy to compute a suitable $w \in H(\Omega,\operatorname{div})$ is to solve the problem

$$w_\beta^{\tilde{h}} := \operatorname*{arg\,min}_{w^{\tilde{h}} \in W^{\tilde{h}}} (1+\beta)\|\nabla v^h - w^{\tilde{h}}\|^2_{L^2(\Omega,\mathbb{R}^2)} + (1+\tfrac{1}{\beta})C_{F,\Omega}^2\|\operatorname{div} w^{\tilde{h}} + f\|^2_{L^2(\Omega)} \tag{5.3.12}$$

with an appropriate finite dimensional subspace $W^{\tilde{h}} \subset H(\Omega,\operatorname{div})$ and $\beta > 0$ fixed. Since (5.3.12) is a quadratic finite dimensional problem, it can be solved efficiently. If $(V^{\tilde{h}_i})_{i\in\mathbb{N}}$ is a sequence of linear finite element spaces generated by quasi-uniform meshes with mesh diameter $\tilde{h}_i$ with $\tilde{h}_i \to 0$ as $i \to 0$, then $W^{\tilde{h}_i} := (V^{\tilde{h}_i} \times V^{\tilde{h}_i})_{i\in\mathbb{N}}$ is limit dense in $H(\Omega,\operatorname{div})$ in the sense that for any $\varepsilon > 0$ and any $w \in H(\Omega,\operatorname{div})$ there exists an $i \in \mathbb{N}$ such that

$$\inf_{w^{\tilde{h}_j} \in V^{\tilde{h}_j} \times V^{\tilde{h}_j}} \|w - w^{\tilde{h}_j}\|_{H(\Omega,\operatorname{div})} \leq \varepsilon \qquad \forall j \geq i.$$

This follows from the density of the set of vector valued polynomials in $H(\Omega,\operatorname{div})$ and the interpolation estimate Corollary 2.7.2. Let u_{h_i} be the finite element approximation of u and $\varepsilon > 0$ be arbitrary. Combining the limit density of $W^{\tilde{h}_i}$ with [112, Thm. 3.1] and [112, Eq.(3.17)] shows that there exist numbers $\hat{\beta} > 0$ and $\hat{h} > 0$ such that

$$\|u - u_{h_i}\|_{H_0^1(\Omega)} \leq M_D(u_{h_i}, w_\beta^{\tilde{h}_i}) \leq \|u - u_{h_i}\|_{H_0^1(\Omega)} + \varepsilon \qquad \text{for all } \beta \in (0, \hat{\beta}),\ \tilde{h}_i \in (0, \hat{h}). \tag{5.3.13}$$

This means that $M_D(u_{h_i}, w_\beta^{\tilde{h}_i})$ approximates $\|u - u_{h_i}\|_{H_0^1(\Omega)}$ arbitrarily well for sufficiently small values of β and $\tilde{h}_i$.

5.3.4. Computation of a subgradient approximation

We now consider the issue of constructing a subgradient oracle which can be used in Algorithm 3.3 to find an appropriate subgradient, cf. Section 3.5.2. Let either $D = \mathscr{I}$ or $D = \Omega \setminus \mathscr{A}_s$, where $\mathscr{I}$ and $\mathscr{A}_s$ are the inactive set and the strictly active set defined in (5.2.5) and (5.2.10), respectively. Recall that $\mathring{F}(w) = w + \iota \mathring{f}$ for all $w \in H^{-1}(\Omega)$. Therefore, $\mathring{F}'(w) = \mathrm{Id}_{H^{-1}(\Omega)}$ for all $w \in H^{-1}(\Omega)$. By Lemma 5.2.6, $g := \mathring{F}'(w)^* q = q$ is a Clarke subgradient of $p(\cdot) = J(S(\mathring{F}(\cdot)))$ at $w \in H^{-1}(\Omega)$, where $q \in H_0^1(\Omega)$ is the solution to the variational equation

$$\text{Find } q \in H_0^1(D): \quad \langle -\Delta q, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = \langle J'(S(\mathring{F}(w))), v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \quad \forall v \in H_0^1(D).$$

Since the objective function $J : H_0^1(\Omega) \to \mathbb{R}$ is given by $J(y) := \frac{1}{2}\|\iota^* y - y_d\|_{L^2(\Omega)}^2$, we find

$$\langle J'(y), v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = (y - y_d, v)_{L^2(\Omega)} \qquad \text{for all } y, v \in H_0^1(\Omega), \tag{5.3.14}$$

which leads to the variational equation

$$\text{Find } q \in H_0^1(D): \quad \langle -\Delta q, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = (S(\mathring{F}(w)) - y_d, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(D). \tag{5.3.15}$$

Note that the inactive set $\mathscr{I}$ and the strictly active set $\mathscr{A}_s$ depend on the exact solution y of the obstacle problem. Therefore, we cannot compute D exactly. Thus, we present two strategies to approximate the subgradient $g = q$. For the first strategy, we compute an approximation of the state y^h via (5.3.3). Then we compute $q^h \in V^h \cap H_0^1(D^h)$ via the variational equation

$$\text{Find } q^h \in V^h \cap H_0^1(D^h): \quad \langle -\Delta q^h, v^h \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} = (y^h - y_d, v^h)_{L^2(\Omega)} \quad \forall v^h \in V^h \cap H_0^1(D^h).$$

Here, the D^h is given as the discrete inactive set $D^h := \{\omega \in \Omega : y^h(\omega) > I^h \psi(\omega)\}$. This leads to the subgradient oracle

$$O_g^1 : X^h \times (0,1] \to X^{*h}, \qquad O_g^1(y^h, h) := \iota^* q^h, \tag{5.3.16}$$

where the linear subspaces $X^{*h} = X^h \subset X^*$ correspond to the finite element space $V^h \subset L^2(\Omega)$. Note that we cannot guarantee that $O_g^1(y^h, h)$ is actually a Clarke subgradient or even that it is close to a Clarke subgradient. Consequently, it is not clear if Assumption 3.1.1 is fulfilled for any meaningful subdifferential G. However, in our numerical tests, O_g^1 performed very well, cf. Section 6.2.

In order to compute subgradients which fulfill Assumption 3.1.1 with $G := \partial_C p + \bar{B}_{H_0^1(\Omega)}(0, \varepsilon_G)$, $\varepsilon_G > 0$, we proceed as suggested in [109]. There, the solution to (5.3.15) is approximated by the solutions q_n

of

$$\text{Find } q_n \in H_0^1(D_n): \quad \langle -\Delta q_n, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = (y_n - y_d, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(D_n). \tag{5.3.17}$$

For each $n \in \mathbb{N}$, D_n is assumed to be a quasi-open subset of D and the approximation $y_n \in H_0^1(\Omega)$ of $S(\mathring{F}(w))$ is to be chosen later. By [109, Lem. 7.1], the following a posteriori error estimate holds true:

$$\|q - q_n\|_{H_0^1(\Omega)} \leq \| -\Delta q_n - J'(y_n)\|_{H^{-1}(D)} + \|J'(y_n) - J'(S(\mathring{F}(w)))\|_{H^{-1}(\Omega)}.$$

If $y_n \to S(\mathring{F}(w))$ in $H_0^1(\Omega)$ and $H_0^1(D_n) \to H_0^1(D)$ in the sense of Mosco as $n \to \infty$, then [109, Lem. 7.1] implies that $q_n \to q$ in $H_0^1(\Omega)$, i.e., the subgradient $g = q$ can be approximated arbitrarily well by the elements $g_n := \mathring{F}'(w)^* q_n = q_n$, $n \in \mathbb{N}$. However, as the exact set D is unknown, the $H^{-1}(D)$-norm of $-\Delta q_n - J'(y_n)$ cannot be evaluated exactly. Therefore, the quasi-open sets $\tilde{D}_n$, $n \in \mathbb{N}$, with $D \subset \tilde{D}_n \subset \Omega$ are introduced. Note that $H_0^1(D) \subset H_0^1(\tilde{D}_n)$ shows

$$\begin{aligned}\|w\|_{H^{-1}(D)} &= \sup_{v \in H_0^1(D), \|v\|_{H_0^1(D)} \leq 1} |\langle w, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}| \\ &\leq \sup_{v \in H_0^1(\tilde{D}_n), \|v\|_{H_0^1(\tilde{D}_n)} \leq 1} |\langle w, v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}| = \|w\|_{H^{-1}(\tilde{D}_n)}\end{aligned}$$

for all $w \in H^{-1}(\tilde{D}_n)$. This results in

$$\|q - q_n\|_{H_0^1(\Omega)} \leq \| -\Delta q_n - J'(y_n)\|_{H^{-1}(\tilde{D}_n)} + \|J'(y_n) - J'(S(\mathring{F}(w)))\|_{H^{-1}(\Omega)}. \tag{5.3.18}$$

Inspecting the proof of [109, Cor. 7.3] shows that if $y_n \to S(\mathring{F}(w))$ in $H_0^1(\Omega)$, $H_0^1(D_n) \to H_0^1(D)$ and $H_0^1(\tilde{D}_n) \to H_0^1(D)$ in the sense of Mosco as $n \to \infty$, then it holds

$$\| -\Delta q_n - J'(y_n)\|_{H^{-1}(\tilde{D}_n)} + \|J'(y_n) - J'(S(\mathring{F}(w)))\|_{H^{-1}(\Omega)} \to 0 \qquad \text{as } n \to \infty,$$

i.e., (5.3.18) is a reliable error estimate. The important issue to compute appropriate sets D_n and $\tilde{D}_n$ is addressed at the end of this section.

While the approximations $g_n = q_n$ defined in (5.3.17) solve the main difficulties in approximating a subgradient, we still cannot compute them exactly. Now we turn to developing a computable approximate subgradient $g^{h_n} \in V^h$ and computable error bounds for $\|g^{h_n} - g\|_{H_0^1(\Omega)}$, where $g \in \partial_C p(w)$ with $w \in H^{-1}(\Omega)$. First, note that the element q_n, as defined in (5.3.17), is the solution to the Dirichlet problem

$$\begin{aligned} -\Delta q_n &= y_n - y_d && \text{in } D_n \\ q_n &= 0 && \text{on } \partial D_n. \end{aligned}$$

We therefore approximate the element q_n by a finite element approximation q^{h_n}. Denote

$$V_D^n := V^{h_n} \cap H_0^1(D_n) \qquad \text{and} \qquad V_{\tilde{D}}^n := V^{h_n} \cap H_0^1(\tilde{D}_n).$$

Choose $y_n := S^{h_n}(w+\mathring{f})$ and define $q^{h_n} \in V_D^n \subset V^{h_n}$ as the solution to the variational equation

$$\text{Find } q^{h_n} \in V_D^n: \quad \langle -\Delta q^{h_n}, v^{h_n}\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = (y_n - y_d, v^{h_n})_{L^2(\Omega)} \quad \forall v^{h_n} \in V_D^n.$$

Note that if w is in V^{h_n}, then $y_n \in V^{h_n}$ can be computed exactly and (5.3.8) yields

$$(y_n - y_d, v^{h_n})_{L^2(\Omega)} = (\mathsf{y_n}^\top \mathsf{M} - \mathsf{y_d}^\top)\mathsf{v^{h_n}} \qquad \text{for all } v^{h_n} \in V^{h_n}. \tag{5.3.19}$$

This shows that $q^{h_n} \in V_D^n$ can be computed exactly. We now develop a computable heuristic for the error $\|q^{h_n} - q\|_{H_0^1(\Omega)}$. Using (5.3.14) and (5.3.18), we find

$$\|q^{h_n} - q\|_{H_0^1(\Omega)} \le \|q^{h_n} - q_n\|_{H_0^1(\Omega)} + \|-\Delta q_n - J'(y_n)\|_{H^{-1}(\tilde{D}_n)} + \|y_n - S(\mathring{F}(w))\|_{H^{-1}(\Omega)}. \tag{5.3.20}$$

In order to approximate the term $\|-\Delta q_n - J'(y_n)\|_{H^{-1}(\tilde{D}_n)}$, we approximately compute the inverse of the inverse Riesz representative of $-\Delta q_n - J'(y_n)$ in the Hilbert space $H_0^1(\tilde{D}_n)$, cf. Section 2.1. Since $q_n \in H_0^1(D_n) \subset H_0^1(\tilde{D}_n)$, the Riesz representative on $H_0^1(\tilde{D}_n)$ of $-\Delta q_n$ is given by q_n. Denoting $r_n := R^{-1}_{H_0^1(\tilde{D}_n)} J'(y_n)$ yields

$$\|-\Delta q_n - J'(y_n)\|_{H^{-1}(\tilde{D}_n)} = \|q_n - R^{-1}_{H_0^1(\tilde{D}_n)} J'(y_n)\|_{H_0^1(\tilde{D}_n)} \le \|q_n - q^{h_n}\|_{H_0^1(\Omega)} + \|q^{h_n} - r_n\|_{H_0^1(\tilde{D}_n)}. \tag{5.3.21}$$

The inverse Riesz representative $r_n = R^{-1}_{H_0^1(\tilde{D}_n)} J'(y_n)$ is characterized via

$$(r_n, v)_{H_0^1(\tilde{D}_n)} = \langle J'(y_n), v\rangle_{H^{-1}(\tilde{D}_n),H_0^1(\tilde{D}_n)} = (y_n - y_d, v)_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\tilde{D}_n).$$

Therefore, r_n is given as the solution to the Dirichlet problem

$$\begin{aligned} -\Delta r_n &= y_n - y_d && \text{in } \tilde{D}_n \\ r_n &= 0 && \text{on } \partial\tilde{D}_n. \end{aligned}$$

Thus, we approximate r_n by the finite element approximation $r^{h_n} \in V_{\tilde{D}}^n$ which is defined as the solution to

$$\text{Find } r^{h_n} \in V_{\tilde{D}}^n: \qquad \langle -\Delta r^{h_n}, v^{h_n}\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = (y_n - y_d, v^{h_n})_{L^2(\Omega)} \qquad \forall v^{h_n} \in V_{\tilde{D}}^n.$$

Note that by (5.3.19), the right hand side can be evaluated exactly and thus $r^{h_n} \in V_{\tilde{D}}^n$ can be computed exactly. Combining (5.3.20) and (5.3.21), we arrive at the estimate

$$\|q^{h_n} - q\|_{H_0^1(\Omega)} \le 2\|q_n - q^{h_n}\|_{H_0^1(\Omega)} + \|q^{h_n} - r^{h_n}\|_{H_0^1(\tilde{D}_n)} + \|r^{h_n} - r_n\|_{H_0^1(\tilde{D}_n)} + \|y_n - S(w+\mathring{f})\|_{H^{-1}(\Omega)}.$$

Since $q^{h_n} \in V_D^n$ and $r^{h_n} \in V_{\tilde{D}}^n$, we can compute $\|q^{h_n} - r^{h_n}\|_{H_0^1(\tilde{D}_n)} = \|q^{h_n} - r^{h_n}\|_{H_0^1(\Omega)}$ exactly, but the other three terms on the right hand side need to be estimated. The errors $\|q_n - q^{h_n}\|_{H_0^1(\Omega)} = \|q_n - q^{h_n}\|_{H_0^1(D_n)}$ and $\|r_n - r^{h_n}\|_{H_0^1(\tilde{D}_n)}$ have to be treated with care. Applying a widely used residual error estimate for a quasi-uniform triangulation yields an estimate of the form $\|q_n - q^{h_n}\|_{H_0^1(\Omega)} \le C h_n^\alpha \|y_n - y_d\|_{L^2(\Omega)}$ with constants $C > 0$ and $\alpha > 0$ independent of h_n. However, the constant C usually depends on the domain D_n and is usually difficult to compute or even estimate. Thus, it is not possible to use a residual error estimate here. Instead, we use the majorant M_D, defined in (5.3.10), to estimate the errors $\|q_n - q^{h_n}\|_{H_0^1(D_n)}$ and $\|r_n - r^{h_n}\|_{H_0^1(\tilde{D}_n)}$ via local gradient averaging, cf. (5.3.11). This yields

$$\|q_n - q^{h_n}\|_{H_0^1(D_n)} \le M_{D_n}(q^{h_n}, G_{h_n} q^{h_n}) \qquad \text{and} \qquad \|r_n - r^{h_n}\|_{H_0^1(\tilde{D}_n)} \le M_{\tilde{D}_n}(r^{h_n}, G_{h_n} r^{h_n})$$

which leads to

$$\|q^{h_n} - q\|_{H_0^1(\Omega)} \le 2M_{D_n}(q^{h_n}, G_{h_n} q^{h_n}) + \|q^{h_n} - r^{h_n}\|_{H_0^1(\tilde{D}_n)} + M_{\tilde{D}_n}(r^{h_n}, G_{h_n} r^{h_n}) + \|y_n - S(\mathring{F}(w))\|_{H^{-1}(\Omega)}.$$

Finally, the term $\|y_n - S(\mathring{F}(w))\|_{H^{-1}(\Omega)}$ needs to be estimated by computable quantities. Recall that the Poincare-Friedrichs constant fulfills $C_{F,\Omega} = \|\iota^*\|_{\mathscr{L}(H_0^1(\Omega),L^2(\Omega))} = \|\iota\|_{\mathscr{L}(H^{-1}(\Omega),L^2(\Omega))}$, cf. Section 2.6. We therefore estimate

$$\|y_n - S(\mathring{F}(w))\|_{H^{-1}(\Omega)} \le C_{F,\Omega}\|y_n - S(\mathring{F}(w))\|_{L^2(\Omega)} \le C_{F,\Omega}^2\|y_n - S(\mathring{F}(w))\|_{H_0^1(\Omega)} \le C_{F,\Omega}^2 e_S(w),$$

where $e_S(w)$ is a computable error estimate of the $H_0^1(\Omega)$-error of the solution operator of the obstacle problem, i.e., $\|y_n - S(\mathring{F}(w))\|_{H_0^1(\Omega)} \le e_S(w)$. In Sections 5.4.1 and 5.5.1, we use a priori and a posteriori techniques to construct such an error estimate e_S. This results in the computable subgradient error estimate

$$\|q^{h_n} - q\|_{H_0^1(\Omega)} \le 2M_{D_n}(q^{h_n}, G_{h_n} q^{h_n}) + \|q^{h_n} - r^{h_n}\|_{H_0^1(\tilde{D}_n)} + M_{\tilde{D}_n}(r^{h_n}, G_{h_n} r^{h_n}) + C_{F,\Omega}^2 e_S(w). \quad (5.3.22)$$

The task to compute sets D_n and $\tilde{D}_n$ which fulfill $D_n \subset D \subset \tilde{D}_n \subset \Omega$, $H_0^1(D_n) \to H_0^1(D)$ and $H_0^1(\tilde{D}_n) \to H_0^1(D)$ in the sense of Mosco is the main difficulty of this approach. Several possibilities are explored in [109]. Each of them relies on the convergence $\|y_n - S(\mathring{F}(w))\|_{L^\infty(\Omega)} \to 0$ as $n \to \infty$ (which can be ensured if sufficient regularity assumptions on the data of the obstacle problem are made) and on further regularity assumptions on the active set $\mathscr{A}$, cf. [109, Thm. 7.31]. In particular, it is assumed that there exists a neighborhood U of $\mathscr{A}$ and a positive number η such that

$$\langle -\Delta\psi - \mathring{F}(w), v\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \ge \eta \int_U v \,\mathrm{d}\lambda \qquad \text{for all } v \in H_0^1(U)_+ \qquad (\mathrm{ND}_\eta)$$

holds. This nondegeneracy condition (ND_η) is vital for the approach of [109] to compute sets D_n and $\tilde{D}_n$ since the parameter η is used to construct $\tilde{D}_n$. However, the condition (ND_η) is not fulfilled in many of our numerical examples. Therefore, we take another approach. For the rest of this section, we assume that $y_n \to S(\mathring{F}(w))$ and $\psi_n := I^{h_n}\psi \to \psi$ in $L^\infty(\Omega)$ as $n \to \infty$. We use $D = \mathscr{I}$, choose a forcing sequence $(\varepsilon_n)_{n\in\mathbb{N}} \subset [0,\infty)$ which fulfills $\varepsilon_n \to 0$ as $n \to \infty$ and

$$\varepsilon_n \ge \|y_n - y\|_{L^\infty(\Omega)} + \|(\psi - \psi_n)_+\|_{L^\infty(\{y_n \le \psi_n + \|y_n - y\|_{L^\infty(\Omega)} + \|(\psi-\psi_n)_+\|_{L^\infty(\Omega)}\})} \qquad \text{for all } n \in \mathbb{N},$$

and define

$$D_n := \{y_n > \psi_n + \varepsilon_n\}. \quad (5.3.23)$$

By [109, Lem. 7.8 and Chap. 7.7], it holds $D_n \subset \mathscr{I}$ for all $n \in \mathbb{N}$ and [109, Thm. 7.10 and Chap. 7.7] implies $H_0^1(D_n) \to H_0^1(\mathscr{I})$ in the sense of Mosco. As stated above, if $y_n \to S(\mathring{F}(w))$ in $H_0^1(\Omega)$, then this yields that $g_n = q_n \to q = g$ in $H_0^1(\Omega)$, i.e., the subgradient g can be approximated arbitrarily well by the elements g_n, $n \in \mathbb{N}$. However, it is very difficult to construct a sequence of sets $\tilde{D}_n \supset \mathscr{I}$ which converge to $\mathscr{I}$ in the sense of Mosco as $n \to \infty$, cf. [109, Chap. 7.6]. We use the choice $\hat{D}_n := \{y_n > \psi_n\} \approx \tilde{D}_n$ instead of a superset $\tilde{D}_n \supset D = \mathscr{I}$ as a heuristic to compute an approximate error bound for the subgradient. In particular, we compute r^{h_n} in the space $V^n_{\hat{D}} := V^{h_n} \cap H_0^1(\hat{D}_n)$ and use

$$e_n^g := 2M_{D_n}(q^{h_n}, G_{h_n} q^{h_n}) + \|q^{h_n} - r^{h_n}\|_{H_0^1(\hat{D}_n)} + M_{\hat{D}_n}(r^{h_n}, G_{h_n} r^{h_n}) + C_{F,\Omega}^2 e_S(w) \tag{5.3.24}$$

as an approximate error bound for the subgradient. Note that we cannot guarantee that $\hat{D}_n \supset D = \mathscr{I}$ and therefore, we cannot guaranteed that $\|q^{h_n} - q\|_{H_0^1(\Omega)} \leq e_n^g$. However, this heuristic for the error of the subgradient approximation is not used in the bundle algorithm. It is only computed to be displayed in the numerical results of Chapter 6.

We now choose

$$O_g^2 : X^{h_n} \times (0,1] \to X^{*h_n}, \qquad O_g^2(y^{h_n}, h_n) := \iota^* q^{h_n} \tag{5.3.25}$$

as a subgradient oracle according to Section 3.5.2. This corresponds to the subdifferential approximation $G := \partial_C p + \bar{B}_{H_0^1(\Omega)}(0, \varepsilon_G)$, where $\varepsilon_G > 0$ is the largest error term e_n^g encountered in the algorithm. Although we cannot guarantee that ε_G is bounded, our numerical experiments suggest this, cf. Figure 6.6(d). In this case, G fulfills Assumption 3.1.1 and thus can be used as a subdifferential. However, since e_n^g is not an upper bound on the error, we cannot compute ε_G. In the numerical results in Chapter 6, we plot e_n^g for all used gradients such that one can get an impression of the size of ε_G.

5.3.5. Computation of a trial iterate

We now apply the theory of Chapter 4 in order to compute a trial iterate. Before we go into detail, we give a short overview of this section. First we define the oracle $O_{F^{-1}}$ needed in Assumption 4.2.7 which gives an approximation of $F^{-1}p'$ for $p' \in P$. Corresponding error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7 are developed in Sections 5.4.2 and 5.5.2 using a priori and a posteriori techniques, respectively. Then Theorem 4.2.8 implies that the output of Algorithm 4.2 can be used as a subproblem oracle O_s and subproblem error bound $\hat{\varepsilon}_s$ which is needed in Algorithm 4.1 to compute a trial iterate. Next we define model value oracles $\underline{O}_\Psi(h)$ and $\overline{O}_\Psi(h)$ which fulfill (4.1.4). Using these oracles and error bounds, Algorithm 4.1 terminates in finitely many steps if the current serious iterate is not G-stationary and Theorem 4.1.1 implies that the computed element can be used for the trial iterate oracle $O_{\tilde{y}}$ in Algorithm 3.4.

We start by defining the oracle $O_{F^{-1}} : P \times (0,1] \to L^2(\Omega)$ where $P := \{p_i',\ i \in I\} \subset L^2(\Omega)$. Here, $p_j' = g_j' + \alpha R_X x_i \in L^2(\Omega)$ is constructed from the subgradients of the current model and the serious iterate, cf. (4.2.2). We first prove that $P \subset V^h$.

LEMMA 5.3.4. *Denote by $\tilde{y}$ the trial iterate and by $\tilde{h}$ the accuracy level computed by Algorithm 4.1 with initial accuracy level $h \in (0,1]$. If the serious iterate $x \in L^2(\Omega)$ is an element of V^h, all subgradients $g_j' \in L^2(\Omega)$ (used in the current model) are elements of V^h and $O_{F^{-1}}(p', \tilde{h}) \in X^{\tilde{h}}$ for all $p' \in V^{\tilde{h}}$, then $\tilde{y} \in V^{\tilde{h}}$.*

Proof. First notice that $\tilde{h} \le h$ and $\tilde{y} = O_s(\hat{\Psi}, \tilde{h})$ for a suitable reduced model $\hat{\Psi}$. Since the oracle O_s is computed via Algorithm 4.2, we find

$$\tilde{y} = O_s(\hat{\Psi}, \tilde{h}) = x - \tilde{d}^* = x - \sum_{j \in I} \tilde{\lambda}_j^* O_{F^{-1}}(p'_j, \tilde{h}) = x - \sum_{j \in I} \tilde{\lambda}_j^* O_{F^{-1}}(g'_j + \alpha R_X x, \tilde{h}).$$

Since $\tilde{h} \le h$ it holds $V^h \subset V^{\tilde{h}}$, cf. (5.3.2). Therefore, $x \in V^{\tilde{h}}$, $g'_j \in V^{\tilde{h}}$, $O_{F^{-1}}(g'_j + \alpha R_X x, \tilde{h}) \in V^{\tilde{h}}$ and $\tilde{y} \in V^{\tilde{h}}$. ◻

LEMMA 5.3.5. *If $x_0 \in V^{h_0}$ and $O_{F^{-1}}(p', h) \in V^h$ for all $p' \in V^h$ and $h \in (0,1]$, then*

$$g_{i+1} \in V^{h_{i+1}}, \quad \tilde{g}_i \in V^{h_{i+1}}, \quad \tilde{y}_i \in V^{\tilde{h}_i}, \quad x_i \in V^{h_i} \qquad \textit{for all } i \in \mathbb{N}.$$

Proof. First notice, that g_{i+1} is always computed at accuracy level $\hat{h}_i = h_{i+1}$. Thus, (5.3.25) implies that $g_{i+1} \in V^{h_{i+1}}$ for all $i \in \mathbb{N}$. Similarly, $\tilde{g}_i$ is always computed at accuracy level $h'_i = h_{i+1}$ which implies $\tilde{g}_i \in V^{h_{i+1}}$ for all $i \in \mathbb{N}$. We show by induction that $\tilde{y}_i \in V^{\tilde{h}_i}$ and $x_i \in V^{h_i}$ for all $i \in \mathbb{N}$. For $i = 0$, the serious iterate x_0 is an element of V^{h_0}. Furthermore, only the exactness gradient $\tilde{g}_0^x \in V^{h_0} \subset V^{\tilde{h}_0}$ is included into the model Ψ_0 and $x_0 \in V^{h_0} \subset V^{\tilde{h}_0}$. Thus Lemma 5.3.4 shows that $\tilde{y}_0 \in V^{\tilde{h}_0}$ which concludes the case $i = 0$. Now assume that $\tilde{y}_i \in V^{\tilde{h}_i}$ and $x_i \in V^{h_i}$ for a given $i \in \mathbb{N}$. In the case of a successful iteration, it holds that $x_{i+1} = \tilde{y}_i \in V^{\tilde{h}_i} \subset V^{h_{i+1}}$ because $h_{i+1} = \hat{h}_i \le \tilde{h}_i$. In the case of a unsuccessful iteration, it holds $x_{i+1} = x_i \in V^{h_i} \subset V^{h_{i+1}}$ because $h_{i+1} \le h_i$. Therefore, $x_{i+1} \in V^{h_{i+1}}$. Now notice that all computed subgradients up to iteration $i+1$ are elements of $V^{h_{i+1}}$ and $x_{i+1} \in V^{h_{i+1}}$. Thus, Lemma 5.3.4 shows that $\tilde{y}_{i+1} \in V^{\tilde{h}_{i+1}}$. ◻

Lemma 5.3.5 implies that $P \subset V^h$. Now recall that the operator $F \in \mathscr{L}(X, X^*)$ is defined by $F = \alpha R_X + \iota^*(Q + \tau R_Y)\iota$, cf. (4.2.2). For the case of no curvature information, i.e., $Q = 0$, it holds (cf. (4.3.4)) for all $p' = \iota^* p^h \in P$ that $F^{-1} p' = F^{-1} \iota^* p^h = R_X^{-1} \iota^* \tilde{D}_\tau^{-1} p^h$ where $\tilde{D}_\tau = \alpha \operatorname{Id}_{Y^*} + \tau R_Y \iota R_X^{-1} \iota^*$. Therefore, we approximate $F^{-1} p'$ via $R_X^{-1} \iota^* \tilde{D}_\tau^{-h} p^h$. By (4.3.7), the coordinates $\mathsf{z} \in \mathbb{R}^{n_V}$ of $z^h := \tilde{D}_\tau^{-h} p^h \in V^h$ can be computed by solving the linear system

$$\mathsf{z}^\top (\alpha \mathsf{K} + \tau \mathsf{M}) \mathsf{w} = \alpha (z^h, w^h)_{H_0^1(\Omega)} + \tau (\iota^* z^h, \iota^* w^h)_{L^2(\Omega)} = (p^h, w^h)_{Y^*} = \mathsf{y}^\top \mathsf{K} \mathsf{w} \quad \forall \mathsf{w} \in \mathbb{R}^{n_V}. \quad (5.3.26)$$

Here, M and K are the mass and stiffness matrices introduced in Section 2.7, respectively. Therefore, $O_{F^{-1}}(p', h) := R_X^{-1} \iota^* \tilde{D}_\tau^{-h} p^h$ is a computable oracle of $F^{-1} p'$ for $Q = 0$. In the case of BFGS curvature, (4.4.13) yields for all $p' = \iota^* p^h \in P$ that $F_{\text{BFGS}}^{-1} p' = R_X^{-1} \iota^* \tilde{F}_{\text{BFGS}}^{-1} p^h$, where $\tilde{F}_{\text{BFGS}}^{-1} p^h$ can be approximated via $\tilde{F}_{\text{BFGS}}^{-h} p^h \in V^h$, cf. Section 4.4.4. We thus use

$$O_{F^{-1}}(p', h) := R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} p^h - R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} U^h (\operatorname{Id}_{\mathbb{R}^{2\tilde{L}}} + V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} U^h)^{-1} V^h \iota R_X^{-1} \iota^* \tilde{D}_{\tau+\mu}^{-h} p^h$$

as a computable oracle of $F^{-1} p'$, cf. (4.4.15). The corresponding error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7 are developed in Sections 5.4.2 and 5.5.2 using a priori and a posteriori techniques, respectively.

In order to compute a subproblem oracle, we run Algorithm 4.2 with the safeguarding parameter $\varepsilon_s^h := \min\{100 \cdot \varepsilon_m \lambda_{\max}(\tilde{\mathsf{H}}_\Delta), h\}$, where ε_m is the machine precision. For $h \approx 1$, this ensures that ε_s^h is reasonably small such that the solution of the approximated problem (4.2.5) is close to the solution of the original problem (4.2.3) while it is sufficiently large to facilitate an efficient solution of (4.2.5). Now, since $p_j^{F^{-1}} = O_{F^{-1}}(p'_j, h) \in X^h$ and $p'_i \in X^{*h}$, the terms $\langle p'_i, p_j^{F^{-1}} \rangle_{X^*,X}$, $i, j \in I$ can be computed exactly. Therefore, Theorem 4.2.8 is applicable which implies that Algorithm 4.2 can be executed. Denote by $\tilde{y}^{*h}$ and e_y^h the output of Algorithm 4.2 for accuracy $h \in (0,1]$. Since $\varepsilon_s^h = h \to 0$ as $h \to 0$, $O_s(\hat{\Psi}, h) := \tilde{y}^{*h}$, $\hat{\varepsilon}_s(\hat{\Psi}, h) := e_y^h$ and $C_{\hat{\Psi}} := \max\{C_F, C_{F^{-1}}, 1\}$ fulfill (4.1.3) and thus can be used as a subproblem oracle with corresponding error estimates.

For the model value oracles $\underline{O}_\Psi(h)$ and $\overline{O}_\Psi(h)$, we choose the oracles defined in Theorems 4.5.2 and 4.5.3. For this, we need a computable error estimator $e_{\tilde{E}}(h) \geq 0$ such that

$$\|(\tilde{E} - \tilde{E}^h)(y^h - x^h)\|_{H_0^1(\Omega)} \leq e_{\tilde{E}}(h) \qquad \text{and} \qquad e_{\tilde{E}}(h) \to 0 \;\text{ as }\; h \to 0. \tag{5.3.27}$$

Note that $z := \tilde{E}(y^h - x^h) = R_Y \iota R_X^{-1} \iota^*(y^h - x^h)$ is the solution of the variational equation $(z, w)_{H_0^1(\Omega)} = (\iota^*(y^h - x^h), \iota^* w)_{L^2(\Omega)}$ for all $w \in H_0^1(\Omega)$. Since this is the weak form of the PDE

$$\begin{aligned} -\Delta z &= \iota^*(y^h - x^h) && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

$\tilde{E}^h(y^h - x^h)$ can be computed exactly via (5.3.9). Therefore, Theorem 5.3.3 implies that for all $w \in H(\Omega, \operatorname{div})$ there holds

$$\|(\tilde{E} - \tilde{E}^h)(y^h - x^h)\|_{H_0^1(\Omega)} \leq M_\Omega(\tilde{E}^h(y^h - x^h), w) = \|\nabla \tilde{E}^h(y^h - x^h) - w\|_{L^2(\Omega)^2} + C_{F,\Omega} \|\operatorname{div} w + f\|_{L^2(\Omega)}.$$

Using $w := w_\beta^h \in H(\Omega, \operatorname{div})$ as in (5.3.12), we define the computable error estimator $e_{\tilde{E}}$ via $e_{\tilde{E}}(h) := M_\Omega(\tilde{E}^h(y^h - x^h), w_\beta^h)$. In order to guarantee $e_{\tilde{E}}(h) \to 0$ as $h \to 0$, we introduce the forcing sequence $v_{\tilde{E}} : (0,1] \to (0, \infty)$ with the property that $v_{\tilde{E}}(h) \to 0$ as $h \to 0$. Now, whenever $e_{\tilde{E}}(h) > v_{\tilde{E}}(h)$, we uniformly refine the mesh and reduce the parameter β. By (5.3.13), repeating this procedure eventually leads to $e_{\tilde{E}}(h) \leq v_{\tilde{E}}(h)$. This guarantees that the conditions (5.3.27) are fulfilled and the model value oracles defined in Theorems 4.5.2 and 4.5.3 fulfill (4.1.4).

In order to define a trial iterate oracle $O_{\tilde{y}} : (0, \infty)^2 \to L^2(\Omega) \times (0, \infty)$ which fulfills (3.5.6), let $\varepsilon > 0$ be a desired accuracy. We run Algorithm 4.1 with the oracles O_s, $\underline{O}_\Psi$ and $\overline{O}_\Psi$ defined above and the error bounds $f^1 := \varepsilon$ and $f^2 := \varepsilon$. Now we distinguish the cases $\Psi(x) = \min_{y \in X} \Psi(y)$ and $\Psi(x) > \min_{y \in X} \Psi(y)$. If the serious iterate minimizes the model Ψ, i.e., if $\Psi(x) = \min_{y \in X} \Psi(y)$, then the serious iterate is already G-stationary (cf. proof of Lemma 3.1.10). However, since we cannot compute the minimizer of Ψ exactly, we cannot detect this case and Algorithm 4.1 refines indefinitely. If $\Psi(x) > \min_{y \in X} \Psi(y)$, then Theorem 4.1.1 shows that Algorithm 4.1 terminates in finitely many steps and the output $(\tilde{y}, \tilde{h})$ fulfills $\Psi(\tilde{y}) \leq \min_{y \in X} \Psi(y) + (C_{\hat{\Psi}} + 1)\varepsilon$ and $\Psi(\tilde{y}) < \Psi(x)$, i.e., $O_{\tilde{y}}(\varepsilon, h) := (\tilde{y}, \tilde{h})$ and $C_{\tilde{y}} := C_{\hat{\Psi}} + 1$ fulfill (3.5.6) and thus can be used as a trial iterate in Algorithm 3.4.

5.4. A priori error estimates

In this section we develop a priori error estimates for quasi-uniform meshes with mesh width h.

5.4.1. A priori error estimates for the solution operator

For convex domains Ω, we use the classical a priori error estimate of [33] for the solution operator of the obstacle problem:

THEOREM 5.4.1 ([33, Thm. 2]). *If Ω is convex, $\psi \in H^2(\Omega)$ and $\mathring{f} \in L^2(\Omega)$, then there exists a constant C_Ω, depending only on Ω but not on h, $\mathring{f}$ and ψ, such that*

$$\|S(u+\mathring{f}) - S^h(u+\mathring{f})\|_{H^1_0(\Omega)} \leq C_\Omega h\big(\|u+\mathring{f}\|_{L^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\big) \qquad \textit{for all } u \in L^2(\Omega).$$

We therefore use $e_S(u) := C_\Omega h\big(\|u\|_{L^2(\Omega)} + \|\mathring{f}\|_{L^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\big)$ as an error estimate for the $H^1_0(\Omega)$-error of the state. For a given domain Ω, we compute a suitable constant C_Ω as described in Section 6.1. The terms $\|\mathring{f}\|_{L^2(\Omega)}$ and $\|\psi\|_{H^2(\Omega)}$ can be computed analytically prior to execution of the algorithm. Therefore, the error estimate e_S can be computed exactly and can be used in Section 5.3.2 and Lemma 5.3.2.

We are not aware of the existence of any a priori error estimates for the solution operator of the obstacle problem for nonconvex domains. Therefore, we use the a posteriori error estimate from Section 5.5.1 for nonconvex domains. Note that we still use only uniform mesh refinements but we estimate the error via the a posteriori error estimate from Section 5.5.1.

5.4.2. A priori error estimates for the trial iterate

In Section 5.3.5, we describe how a trial iterate is computed and define the oracle $O_{F^{-1}}$. The aim of this section is to provide explicit and computable a priori error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ for the oracle $O_{F^{-1}}$ which fulfill Assumption 4.2.7. We start off with the case $Q = 0$ and a convex domain Ω. First we show that problem (4.6.7) is regular in the sense of Definition 4.6.3. Using this regularity result, Corollary 4.6.6 and Theorem 4.6.7 provide error estimates $e^E_{x^h,y^h}$ and $e^{D^{-1}}_{x^h,y^h}$ which fulfill Assumptions 4.3.6 and 4.3.7, respectively. We then use Lemma 4.3.9 to provide error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7.

Throughout this section we work in the setting that $Y^* = H^1_0(\Omega)$, $X^* = L^2(\Omega)$, that $Y^{*h} = V^h$ is a space of finite element functions generated by quasi-uniform meshes with mesh width h and that the initial serious iterate x_0 is an element of V^{h_0}. Thus, Lemma 5.3.5 shows that Assumption 4.3.5 holds true and all relevant quantities are discretized as elements of X^{*h}. In particular, we have $x = R_X^{-1}\iota^* x^h$ and $g'_j = \iota^* g^h_j$ with $x^h, g^h_j \in Y^{*h}$. We set $p^h_j := g^h_j + \alpha x^h \in Y^{*h}$, $j \in I$ and $\tilde{d}^{*h} := -\sum_{j\in I} \tilde{\lambda}^*_j \tilde{D}^{-h}_\tau p^h_j \in Y^{*h}$ where $\tilde{\lambda}^*$ is the solution to the approximated dual problem (4.2.5). The quasi-uniformity parameters of the meshes are denoted by c_1 and c_2.

A priori error estimates for convex, polygonal domains

First note, that (4.3.2) implies that for $x^h \in Y^{*h}$ the adjoint problem (4.6.7),

$$\text{Find } \Phi'_x \in Y^*: \qquad (\tilde{D}_{\tau+\mu}\, w', \Phi'_x)_{Y^*} = (\iota^* x^h, \iota^* w')_{X^*} \qquad \text{for all } w' \in Y^*, \tag{4.6.7}$$

is equivalent to the PDE

$$\alpha(\Phi_x', w')_{H_0^1(\Omega)} + (\tau+\mu)(\Phi_x', w')_{L^2(\Omega)} = (x^h, w')_{L^2(\Omega)} \quad \text{for all } w' \in H_0^1(\Omega). \tag{5.4.1}$$

LEMMA 5.4.2. *If $\Omega \subset \mathbb{R}^2$ is a bounded convex open domain with polygonal boundary, then the variational equation*

$$\textit{Find } z \in H_0^1(\Omega): \qquad (z,w)_{H_0^1(\Omega)} + \beta(z,w)_{L^2(\Omega)} = (y,w)_{L^2(\Omega)} \qquad \textit{for all } w \in H_0^1(\Omega) \tag{5.4.2}$$

with $\beta \geq 0$ and $y \in L^2(\Omega)$ has a unique solution z with regularity $z \in H^2(\Omega) \cap H_0^1(\Omega)$ and it holds

$$\|z\|_{H^2(\Omega)} \leq \Big(1 + C_{F,\Omega}^2 + \frac{2\beta}{C_{F,\Omega}^{-2}+\beta} + \frac{1+\beta^2}{(C_{F,\Omega}^{-2}+\beta)^2}\Big)^{1/2} \|y\|_{L^2(\Omega)} \leq \sqrt{4 + C_{F,\Omega}^2 + C_{F,\Omega}^4}\, \|y\|_{L^2(\Omega)},$$

where $C_{F,\Omega} = \sup_{w \in H_0^1(\Omega)} \|w\|_{L^2(\Omega)} / \|w\|_{H_0^1(\Omega)}$ is the Friedrich constant (which depends only on Ω).

Proof. By the Lax-Milgram theorem, for all $\beta \geq 0$, (5.4.2) has a unique weak solution $z \in H_0^1(\Omega)$. Since Ω is convex, [41, Thm. 3.2.1.2 and Thm. 3.2.1.3] imply that $z \in H^2(\Omega)$. Using integration by parts (cf. [31, Thm. 4.6]), one can see that the solution $z \in H^2(\Omega)$ to (5.4.2) also fulfills

$$\begin{aligned} -\Delta z + \beta z &= y && \text{in } \Omega \\ z &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Since Ω has a polygonal boundary, [41, Eq. (4,3,1,11) and Lem. 4.3.1.3] yield that

$$\|w\|_{H^2(\Omega)}^2 \leq \|\Delta w\|_{L^2(\Omega)}^2 + \|w\|_{H^1(\Omega)}^2 \qquad \text{for all } w \in H^2(\Omega) \cap H_0^1(\Omega).$$

Therefore, we infer

$$\begin{aligned} \|z\|_{H^2(\Omega)}^2 &\leq \|\Delta z\|_{L^2(\Omega)}^2 + \|z\|_{H^1(\Omega)}^2 = \|y - \beta z\|_{L^2(\Omega)}^2 + \|z\|_{H^1(\Omega)}^2 \\ &\leq \|y\|_{L^2(\Omega)}^2 + 2\beta \|y\|_{L^2(\Omega)} \|z\|_{L^2(\Omega)} + \beta^2 \|z\|_{L^2(\Omega)}^2 + \|z\|_{H_0^1(\Omega)}^2 + \|z\|_{L^2(\Omega)}^2. \end{aligned} \tag{5.4.3}$$

By the Poincaré-Friedrich inequality (2.6.1), there exists a constant $C_{F,\Omega} > 0$ which depends only on Ω such that

$$\|w\|_{L^2(\Omega)} \leq C_{F,\Omega} \|\nabla w\|_{L^2(\Omega)^2} = C_{F,\Omega} \|w\|_{H_0^1(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega).$$

Therefore, as $z \in H_0^1(\Omega)$ solves (5.4.2), we find both

$$(C_{F,\Omega}^{-2} + \beta) \|z\|_{L^2(\Omega)}^2 \leq \|z\|_{H_0^1(\Omega)}^2 + \beta \|z\|_{L^2(\Omega)}^2 = (y,z)_{L^2(\Omega)} \leq \|y\|_{L^2(\Omega)} \|z\|_{L^2(\Omega)}$$

and

$$\|z\|_{H_0^1(\Omega)}^2 \leq \|z\|_{H_0^1(\Omega)}^2 + \beta \|z\|_{L^2(\Omega)}^2 = (y,z)_{L^2(\Omega)} \leq C_{F,\Omega} \|y\|_{L^2(\Omega)} \|z\|_{H_0^1(\Omega)}.$$

This gives

$$\|z\|_{L^2(\Omega)} \leq (C_{F,\Omega}^{-2}+\beta)^{-1}\|y\|_{L^2(\Omega)} \quad \text{and} \quad \|z\|_{H_0^1(\Omega)} \leq C_{F,\Omega}\|y\|_{L^2(\Omega)}.$$

Combining this with (5.4.3) shows the first estimate of the lemma via

$$\|z\|_{H^2(\Omega)}^2 \leq (1+\frac{2\beta}{C_{F,\Omega}^{-2}+\beta}+\frac{\beta^2}{(C_{F,\Omega}^{-2}+\beta)^2}+C_{F,\Omega}^2+\frac{1}{(C_{F,\Omega}^{-2}+\beta)^2})\|y\|_{L^2(\Omega)}^2.$$

For $\beta = 0$, the second estimate of the lemma holds true. Now let β be greater than zero. Then $\beta^{-1}C_{F,\Omega}^{-2} > 0$ and

$$\frac{\beta}{C_{F,\Omega}^{-2}+\beta} = \frac{1}{\beta^{-1}C_{F,\Omega}^{-2}+1} < 1 \quad \text{and} \quad \frac{1}{(C_{F,\Omega}^{-2}+\beta)^2} = \frac{1}{C_{F,\Omega}^{-4}+2\beta C_{F,\Omega}^{-2}+\beta^2} < C_{F,\Omega}^4.$$

This shows that

$$\|z\|_{H^2(\Omega)}^2 \leq \Big(1+C_{F,\Omega}^2+\frac{2\beta}{C_{F,\Omega}^{-2}+\beta}+\frac{1+\beta^2}{(C_{F,\Omega}^{-2}+\beta)^2}\Big)\|y\|_{L^2(\Omega)}^2 \leq (4+C_{F,\Omega}^2+C_{F,\Omega}^4)\|y\|_{L^2(\Omega)}^2. \qquad \square$$

Known values of $C_{F,\Omega}$ for several domains Ω can be found in Section 2.6.

LEMMA 5.4.3. *Assume that $\Omega \subset \mathbb{R}^2$ is a convex, bounded and open domain with polygonal boundary. Then the adjoint problem* (4.6.7) *is* 1*-regular and the constant C_{reg} depends only on Ω, the quasi-uniformity parameters c_1, c_2 of the triangulations and α.*

Proof. Let $x^h \in Y^{*h}$ be arbitrary. Dividing both sides of (5.4.1) by $\alpha > 0$, Lemma 5.4.2 shows that the solution Φ_x' of (5.4.1) is an element of $H^2(\Omega) \cap H_0^1(\Omega)$ and there exists a constant $C_\Omega > 0$, only depending on Ω, such that $\|\Phi_x'\|_{H^2(\Omega)} \leq C_\Omega\|\alpha^{-1}\iota^* x^h\|_{L^2(\Omega)} = C_\Omega/\alpha\|\iota^* x^h\|_{L^2(\Omega)}$. Furthermore, by Corollary 2.7.2 there exists a constant $C_{c_1,c_2} > 0$ (depending only on the quasi-uniformity parameters c_1, c_2 of the triangulations) such that

$$\|\Phi_x' - P_{Y^{*h}}\Phi_x'\|_{H_0^1(\Omega)} \leq \|\Phi_x' - I^h\Phi_x'\|_{H_0^1(\Omega)} \leq C_{c_1,c_2}h|\Phi_x'|_{H^2(\Omega)} \leq C_{c_1,c_2}C_\Omega\alpha^{-1}h\|\iota^* x^h\|_{L^2(\Omega)}.$$

As (5.4.1) and (4.6.7) are equivalent, this shows that the adjoint problem (4.6.7) is 1-regular according to Definition 4.6.3. $\square$

THEOREM 5.4.4. *Assume that $\Omega \subset \mathbb{R}^2$ is a convex, bounded and open domain with polygonal boundary. If $Q = 0$, then there exists a constant $C \geq 0$, depending only on Ω, c_1, c_2 and α, such that*

$$O_{F^{-1}}(p_j', h) := R_X^{-1}\iota^*\tilde{D}_\tau^{-h}p_j^h, \qquad C_F := C, \quad C_{F^{-1}} := C,$$

$$\hat{\varepsilon}_{F^{-1}}(p_i', p_j', h) := (1+\tau)h\|p_i^h\|_{L^2(\Omega)}\|p_j^h\|_{H_0^1(\Omega)}, \qquad \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) := \tau h^2\,\|\tilde{d}^{*h}\|_{L^2(\Omega)}^2$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. First note that $\tilde{D}_\tau^{-h}x^h$, $\|x^h\|_{L^2(\Omega)}$ and $\|x^h\|_{H_0^1(\Omega)}$ can be computed exactly for any $x^h \in V^h$, cf. Section 2.7. Thus, $O_{F^{-1}}(p_j', h)$, $\hat{\varepsilon}_{F^{-1}}(p_i', p_j', h)$ and $\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h)$ can be computed. Also, $\hat{\varepsilon}_{F^{-1}}(p_i', p_j', h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) \to 0$ as $h \to 0$. By Lemma 5.4.3, the adjoint problem (4.6.7) for a convex, bounded and open domain $\Omega \subset \mathbb{R}^2$ with polygonal boundary is 1-regular and the constant $C_{\text{reg}} \geq 0$ depends only on Ω, c_1, c_2 and α. Therefore, Corollary 4.6.6 implies that

$$e^{E}_{x^h,y^h} := C^2_{\text{reg}} h^2 \, \|x^h\|_{L^2(\Omega)} \|y^h\|_{L^2(\Omega)} \qquad \text{for all } x^h, y^h \in Y^{*h}$$

fulfills Assumption 4.3.6. Furthermore, Theorem 4.6.7 implies that

$$e^{D^{-1}}_{x^h,y^h} := (1 + \tfrac{\tau}{\alpha} \|\iota^*\|^2_{\mathscr{L}(H^1_0(\Omega),L^2(\Omega))}) C_{\text{reg}} h \|x^h\|_{L^2(\Omega)} \|y^h\|_{H^1_0(\Omega)} \qquad \text{for all } x^h, y^h \in Y^{*h}$$

fulfills Assumption 4.3.7. As Assumptions 4.3.6 and 4.3.7 are fulfilled, Lemma 4.3.9 shows that the approximation $p^{F^{-1}}_j := R_X^{-1} \iota^* \tilde{D}^{-h}_{\tau} p^h_j$, $j \in I$, and the error estimates $e_F := \tau C^2_{\text{reg}} h^2 \, \|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$ and

$$e_{i,j,F^{-1}} := (1 + \tfrac{\tau}{\alpha} \|\iota^*\|^2_{\mathscr{L}(H^1_0(\Omega),L^2(\Omega))}) C_{\text{reg}} h \|p^h_i\|_{L^2(\Omega)} \|p^h_j\|_{H^1_0(\Omega)} \qquad \text{for all } i, j \in I$$

fulfill Assumption 4.2.4. Consequently, $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. ∎

Now we consider the case that the curvature operator is given via the BFGS formula.

LEMMA 5.4.5. *If $\Omega \subset \mathbb{R}^2$ is a domain with Lipschitz boundary, $Y^* = H^1_0(\Omega)$ and $(Y^{*h})_h$ is a family of finite element spaces generated by quasi-uniform meshes with mesh width h, then Assumption 4.6.12 is fulfilled.*

Proof. The space $Y^* = H^1_0(\Omega)$ is a separable and infinite dimensional Hilbert space. By [20, Chap. 1.2], the set $\mathscr{W} := C^\infty_c(\Omega)$ is dense in $H^1_0(\Omega)$. Since Y^{*h} is a finite element space generated by a quasi-uniform mesh with mesh width h, by Corollary 2.7.2, there exists a constant $C > 0$ (independent of h) such that

$$\|v' - I^h v'\|_{H^1_0(\Omega)} \leq Ch|v'|_{H^2(D)} \qquad \text{for all } v' \in H^2(\Omega),$$

where I^h is the Lagrange interpolation operator. Consequently, $\text{dist}(v', Y^{*h}) \leq \|v' - I^h v'\|_{H^1_0(\Omega)} \to 0$ as $h \to 0$ for all $v' \in \mathscr{W} = C^\infty_c(\Omega) \subset H^2(D)$ and Assumption 4.6.12 holds true. ∎

THEOREM 5.4.6. *Assume that $\Omega \subset \mathbb{R}^2$ is a convex, bounded and open domain with polygonal boundary. If Q is given as the BFGS curvature operator (cf. Section 4.4), then there exists constants C, C_{reg} and $\mathring{h} \in (0,1]$ such that for all $h \in (0, \mathring{h}]$, the quantities*

$$O_{F^{-1}}(p'_j, h) := R_X^{-1} \iota^* \tilde{F}^{-h}_{BFGS} p^h_j, \qquad C_F := C^2_{reg}, \quad C_{F^{-1}} := C,$$
$$\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) := h\|p^h_i\|_{L^2(\Omega)} \|p^h_j\|_{H^1_0(\Omega)}, \qquad \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) := (\tau + \mu) h^2 \, \|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. By Lemma 5.4.5, Assumption 4.6.12 is fulfilled and by Lemma 5.4.3, the adjoint problem (4.6.7) is 1-regular. Therefore, Theorem 4.6.17 implies that there exists a $\mathring{h} \in (0,1]$ and a constant C (possibly depending on τ) such that

$$|(\iota^* x^h, \iota^* (\tilde{F}^{-1}_{\text{BFGS}} - \tilde{F}^{-h}_{\text{BFGS}}) y^h)_{X^*}| \leq e^{F^{-1}_{\text{BFGS}}}_{x^h,y^h} := Ch\|\iota^* x^h\|_{X^*} \|y^h\|_{Y^*} \qquad \text{for all } x^h, y^h \in Y^{*h},$$

i.e., $e^{F^{-1}_{\mathrm{BFGS}}}_{x^h,y^h}$ fulfills Assumption 4.4.6 for all $h \in (0,\mathring{h}]$. Furthermore, Corollary 4.6.6 implies that $e^{E}_{x^h,y^h} := C^2_{\mathrm{reg}}h^2\|\iota^* x^h\|_{X^*}\|\iota^* y^h\|_{X^*}$ for $x^h, y^h \in Y^{*h}$ fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.4.6 are fulfilled, Lemma 4.4.7 implies that

$$p_j^{F^{-1}_{\mathrm{BFGS}}} := \tilde{F}^{-h}_{\mathrm{BFGS}} p_j^h, \qquad e_{i,j,F^{-1}_{\mathrm{BFGS}}} := e^{F^{-1}_{\mathrm{BFGS}}}_{p_i^h,p_j^h}, \qquad e_{F_{\mathrm{BFGS}}} := (\tau+\mu)e^{E}_{\tilde{d}^{*h},\tilde{d}^{*h}},$$

fulfill Assumption 4.2.4 for all $h \in (0,\mathring{h}]$. Since $\tilde{F}^{-h}_{\mathrm{BFGS}} x^h$, $\|x^h\|_{L^2(\Omega)}$ and $\|x^h\|_{H^1_0(\Omega)}$ can be computed exactly for any $x^h \in V^h$, cf. Section 2.7, $O_{F^{-1}}(p'_j,h)$, $\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h)$ and $\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h)$ can be computed exactly. From $\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) \to 0$ as $h \to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. ∎

Remark 5.4.7. Note that the constant C in Theorem 5.4.6 might depend on τ. Unfortunately, we were not able to determine in which way C might depend on τ, cf. Remark 4.6.15. In the case that $\tau \to \infty$ during execution of the algorithm, this may cause problems. However, in all our numerical experiments, τ did not tend to infinity.

A priori error estimates for polygonal domains with one reentrant corner

Typically, the solution of the Dirichlet problem (5.4.1) on the nonconvex domain Ω is not an element of $H^2(\Omega)$. Therefore, we need to replace Lemmas 5.4.2 and 5.4.3 with similar statements.

LEMMA 5.4.8. *Suppose that $\Omega \subset \mathbb{R}^2$ is a bounded, open polygonal domain with one reentrant corner of angle $\omega \in (\pi,2\pi)$, let $\beta \geq 0$ and $x \in L^2(\Omega)$ be arbitrary and denote by $z \in H^1_0(\Omega)$ the solution to the variational equation*

$$\text{Find } z \in H^1_0(\Omega): \qquad (z,w)_{H^1_0(\Omega)} + \beta(z,w)_{L^2(\Omega)} = (x,w)_{L^2(\Omega)} \qquad \text{for all } w \in H^1_0(\Omega).$$

Then, for all $p \in (1,2/(2-\frac{\pi}{\omega}))$, the solution $z \in H^1_0(\Omega)$ is an element of $W^{2,p}(\Omega)$ and there exists a constant $C_{\beta,p,\Omega}$ (depending only on β, p and Ω) such that $\|z\|_{W^{2,p}(\Omega)} \leq C_{\beta,p,\Omega}\|x\|_{L^p(\Omega)}$.

Proof. Combine [4, Lem. 1] and [30, Cor. 3.2]. ∎

LEMMA 5.4.9. *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded, open, polygonal domain with one reentrant corner of interior angle $\omega \in (\pi,2\pi)$, $Y^* = H^1_0(\Omega)$. Then the adjoint problem* (4.6.7) *is* $(2-2/p)$*-regular for all $p \in (1,2/(2-\frac{\pi}{\omega}))$.*

Proof. Let $p \in (1,2/(2-\frac{\pi}{\omega}))$ be arbitrary. By Lemma 5.4.8, the solution $\Phi'_x \in H^1_0(\Omega)$ to (4.6.7) is an element of $W^{2,p}(\Omega)$ and there exists a constant $C \geq 0$ (depending only on p, α, $\tau+\mu$ and Ω) such that

$$\|\Phi'_x\|_{W^{2,p}(\Omega)} \leq C\|x^h\|_{L^p(\Omega)} \qquad \text{for all } x^h \in L^2(\Omega).$$

Therefore, by Corollary 2.7.2, there exists a constant $C_{p,c_1,c_2} > 0$ (depending only on p and the quasi-uniformity constants c_1, c_2) such that

$$\|\Phi' - I^h\Phi'\|_{H^1_0(\Omega)} \leq C_{p,c_1,c_2} h^{2-2/p}|\Phi'_x|_{W^{2,p}(\Omega)}.$$

Since Ω is bounded and $1 \le p \le 2$, [1, Thm. 2.14] implies $\|x^h\|_{L^p(\Omega)} \le (\int_\Omega 1 \, \mathrm{d}\lambda)^{1/p-1/2} \|x^h\|_{L^2(\Omega)}$ which yields

$$\|\Phi' - I^h \Phi'\|_{H_0^1(\Omega)} \le C_{p,c_1,c_2} C \left(\int_\Omega 1 \, \mathrm{d}\lambda \right)^{1/p-1/2} h^{2-2/p} \|x^h\|_{L^2(\Omega)}. \qquad \square$$

This regularity result yields a priori error estimates for both the case of no curvature and the case of BFGS curvature.

THEOREM 5.4.10. *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded, open, polygonal domain domain with one reentrant corner of interior angle $\omega \in (\pi, 2\pi)$. If $Q = 0$ and $\varepsilon > 0$ is small, then there exists a constant $C > 0$, (independent of h), such that*

$$O_{F^{-1}}(p'_j, h) := R_X^{-1} \iota^* \tilde{D}_\tau^{-h} p_j^h, \qquad C_F := C, \quad C_{F^{-1}} := C,$$
$$\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) := h^{\pi/\omega - \varepsilon} \|p_i^h\|_{L^2(\Omega)} \|p_j^h\|_{H_0^1(\Omega)}, \qquad \hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) := \tau h^{2\pi/\omega - \varepsilon} \|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. By Lemma 5.4.9, the adjoint problem (4.6.7) for a bounded, open, polygonal domain with one reentrant corner of interior angle $\omega \in (\pi, 2\pi)$ is $2-2/p$-regular for all $p \in (1, 2/(2-\pi/\omega))$. In particular, for small $\varepsilon > 0$, the adjoint problem (4.6.7) is $(\pi/\omega - \varepsilon)$-regular. Therefore, by Theorem 4.6.7 there exists a constant $C > 0$ (independent of h) such that

$$e^{D^{-1}}_{x^h, y^h} := C h^{\pi/\omega - \varepsilon} \|x^h\|_{L^2(\Omega)} \|y^h\|_{H_0^1(\Omega)} \qquad \text{for all } x^h, y^h \in Y^{*h}$$

fulfills Assumption 4.3.7. Furthermore, by Corollary 4.6.6 there exists a constant $\tilde{C} > 0$ (independent of h) such that

$$e^{E}_{x^h, y^h} := \tilde{C} h^{2\pi/\omega - \varepsilon} \|x^h\|_{L^2(\Omega)} \|y^h\|_{L^2(\Omega)} \qquad \text{for all } x^h, y^h \in Y^{*h}$$

fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.3.7 are fulfilled, Lemma 4.3.9 shows that the approximation $p_j^{F^{-1}} := R_X^{-1} \iota^* \tilde{D}_\tau^{-h} p_j^h$, $j \in I$, and the error estimates

$$e_{i,j,F^{-1}} := C h^{\pi/\omega - \varepsilon} \|p_i^h\|_{L^2(\Omega)} \|p_j^h\|_{H_0^1(\Omega)} \;\; i, j \in I \qquad \text{and} \qquad e_F := \tau \tilde{C} h^{2\pi/\omega - \varepsilon} \|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$$

fulfill Assumption 4.2.4. From $\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) \to 0$ as $h \to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. $\square$

Remark 5.4.11. Note that the constant C in Theorem 5.4.10 might depend on τ which might cause problems if $\tau \to \infty$ during execution of the algorithm. Unfortunately, the existence of the constant $C_{\beta,p,\Omega}$ in Lemma 5.4.8 is derived via proof by contradiction and thus we were not able to determine in which way C might depend on τ. However, in all our numerical experiments, τ did not tend to infinity.

Example 5.4.12. Suppose that $\Omega = (-a, a)^2 \setminus [0, a)^2$, $a > 0$, is a L-shaped domain, $Q = 0$ and $\varepsilon > 0$ is small. Then there exists a constant $C > 0$, (independent of h), such that

$$O_{F^{-1}}(p'_j,h) := R_X^{-1}\iota^*\tilde{D}_\tau^{-h}p_j^h, \qquad C_F := C, \quad C_{F^{-1}} := C,$$
$$\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) := h^{2/3-\varepsilon}\|p_i^h\|_{L^2(\Omega)}\|p_j^h\|_{H_0^1(\Omega)}, \qquad \hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) := \tau h^{4/3-\varepsilon}\,\|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

THEOREM 5.4.13. *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded, open, polygonal domain with one reentrant corner of interior angle $\omega \in (\pi, 2\pi)$. If Q is given as the BFGS curvature operator (cf. Section 4.4) and $\varepsilon > 0$ is small, then there exist constants C, C_{reg} and $\mathring{h} \in (0,1]$ (independent of h) such that for all $h \in (0,\mathring{h}]$, the quantities*

$$O_{F^{-1}}(p'_j,h) := R_X^{-1}\iota^*\tilde{F}_{BFGS}^{-h}p_j^h, \qquad C_F := C_{reg}^2, \quad C_{F^{-1}} := C,$$
$$\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) := h^{\pi/\omega-\varepsilon}\|p_i^h\|_{L^2(\Omega)}\|p_j^h\|_{H_0^1(\Omega)}, \qquad \hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) := (\tau+\mu)h^{2\pi/\omega-\varepsilon}\,\|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. By Lemma 5.4.9, the adjoint problem (4.6.7) is $(2-2/p)$-regular for all $p \in (1, 2/(2-\frac{\pi}{\omega}))$. In particular, for small $\varepsilon > 0$, the adjoint problem (4.6.7) is $(\pi/\omega - \varepsilon)$-regular. Furthermore, by Lemma 5.4.5, Assumption 4.6.12 is fulfilled. Therefore, Theorem 4.6.17 implies that there exists a $\mathring{h} \in (0,1]$ and a constant C (possibly depending on τ) such that

$$|(\iota^*x^h, \iota^*(\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h})y^h)_{X^*}| \le Ch^{\pi/\omega-\varepsilon}\|\iota^*x^h\|_{X^*}\|y^h\|_{Y^*} =: e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}} \qquad \text{for all } x^h, y^h \in Y^{*h},$$

i.e., $e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}}$ fulfills Assumption 4.4.6 for all $h \in (0,\mathring{h}]$. Furthermore, Corollary 4.6.6 implies that $e_{x^h,y^h}^E := C_{\mathrm{reg}}^2 h^{2\pi/\omega-\varepsilon}\,\|\iota^*x^h\|_{X^*}\|\iota^*y^h\|_{X^*}$ for $x^h, y^h \in Y^{*h}$ fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.4.6 are fulfilled, Lemma 4.4.7 implies that

$$p_j^{F_{\mathrm{BFGS}}^{-1}} := R_X^{-1}\iota^*\tilde{F}_{\mathrm{BFGS}}^{-h}p_j^h, \qquad e_{i,j,F_{\mathrm{BFGS}}^{-1}} := e_{p_i^h,p_j^h}^{F_{\mathrm{BFGS}}^{-1}}, \qquad e_{F_{\mathrm{BFGS}}} := (\tau+\mu)e^E_{\tilde{d}^{*h},\tilde{d}^{*h}},$$

fulfill Assumption 4.2.4 for all $h \in (0,\mathring{h}]$. From $\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) \to 0$ as $h \to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. ∎

Remark 5.4.14. Similar to Theorems 5.4.6 and 5.4.10, we cannot track the dependence of the constant C in Theorem 5.4.13 on τ. However, in our numerical tests this did not cause any problems, cf. Remarks 5.4.7 and 5.4.11.

Example 5.4.15. Suppose that $\Omega = (-a,a)^2 \setminus [0,a)^2$, $a > 0$ is a L-shaped domain, $\varepsilon > 0$ is small and let Q be given as the BFGS curvature operator. Then there exists constants C, C_{reg} and $\mathring{h} \in (0,1]$ (independent of h) such that for all $h \in (0,\mathring{h}]$, the quantities

$$O_{F^{-1}}(p'_j,h) := R_X^{-1}\iota^*\tilde{F}_{\mathrm{BFGS}}^{-h}p_j^h, \qquad C_F := C_{\mathrm{reg}}^2, \quad C_{F^{-1}} := C,$$
$$\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) := h^{2/3-\varepsilon}\|p_i^h\|_{L^2(\Omega)}\|p_j^h\|_{H_0^1(\Omega)}, \qquad \hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) := (\tau+\mu)h^{4/3-\varepsilon}\,\|\tilde{d}^{*h}\|^2_{L^2(\Omega)}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

5.5. A posteriori error estimates

In this section we develop a posteriori error estimates in order to quantify the function value error and the trial iterate error. The classical adaptive scheme (cf., e.g., [54])

$$\text{solve} \quad \to \quad \text{estimate} \quad \to \quad \text{mark} \quad \to \quad \text{refine}$$

starts by solving the given variational inequality or PDE. From this solution a local error estimate is derived which then is used to mark triangles with large error contribution. Marked triangles are refined in such a way that the resulting family of triangulations is regular, cf. Section 2.7. This process is repeated until a desired accuracy is reached. A posteriori error estimates for the function value error are available in the literature and we use the approach of [13]. For the trial iterates, a posteriori error estimates for the oracle $O_{F^{-1}}$ have to be developed since, to the best of the author's knowledge, error estimates in this form are not available in the literature.

5.5.1. A posteriori error estimates for the solution operator

We use [13] to compute an a posteriori error estimate of the distance from the computed state to the actual state. Alternatively, one could use the error estimators of [54, 35]. The approach of [13] encompasses an error reduction property which leads to linear convergence up to consistency errors. However, it is necessary to refine marked triangles in such a way that the new triangulation has at least one new vertex in the interior of every marked triangle and at least one new vertex in the interior of each side of each marked triangle. To ensure this, the following refinement strategy is used. We divide every marked triangle into four congruent triangles using the midpoints of each side. Then we divide both triangles intersected by the line from the midpoint of the longest side to the vertex opposed to the longest side, cf. [94, Chap. 5.1]. This procedure divides the marked triangle into six new triangles. Possible hanging nodes, i.e., nodes which are on the interior of a side of a triangle, are avoided by bisecting the adjacent triangles, i.e., adding the midpoint of the longest side as a new vertex and dividing the triangle into the two resulting triangles. If the hanging node is not on the longest side of each adjacent triangle, this may introduce new hanging nodes. The procedure is repeated until no hanging nodes exist. This refinement procedure and the evaluation of the local error estimates is implemented in MATLAB. To increase performance, the code is fully vectorized.

5.5.2. A posteriori error estimates for the trial iterate

In Section 5.3.5, we describe how a trial iterate is computed and define the oracle $O_{F^{-1}}$. The goal of this section is to compute a posteriori error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7 in the case of $X^* = L^2(\Omega)$ and $Y^* = H_0^1(\Omega)$, where $\Omega \subset \mathbb{R}^2$ is a general polygonal (not necessarily convex), bounded polygonal domain with Lipschitz boundary. In Section 5.3.5 we already observed that $\tilde{E}x^h$ as well as $\tilde{D}_\tau x^h$ can be characterized as a solutions to a PDE. Thus, classical a posteriori error estimates such as [2, Thm. 2.7], [9, Ex. 3.2] or [130] can be applied to develop estimators $e^E_{x^h,y^h}$ and $e^{D^{-1}}_{x^h,y^h}$ which fulfill Assumptions 4.3.6 and 4.3.7, respectively. However, for such an error estimate, typically convexity of Ω is assumed. In contrast to this, in [133] error estimates for nonconvex domains are developed, but only weighted L^2-estimates are obtained. We therefore generalize the approach of [2, Thm. 2.7] to nonconvex domains. Using Lemma 4.3.9, the obtained estimators $e^E_{x^h,y^h}$ and $e^{D^{-1}}_{x^h,y^h}$ then provide error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7.

Throughout this section we work in the setting that $Y^{*h} = V^h$ is a space of finite element functions generated by the mesh $\mathscr{T}^h$, $0 < h \leq 1$. The family of meshes $(\mathscr{T}^h)_h$, $0 < h \leq 1$, is assumed to be regular with regularity parameter σ, cf. Section 2.7. Furthermore, we assume that the initial serious iterate x_0 is an element of V^{h_0}. Thus, Lemma 5.3.5 shows that Assumption 4.3.5 holds true and all relevant quantities are discretized as elements of X^{*h}. In particular, we have $x = R_X^{-1}\iota^* x^h$ and $g_j' = \iota^* g_j^h$ with $x^h, g_j^h \in Y^{*h}$. We set $p_j^h := g_j^h + \alpha x^h \in Y^{*h}$, $j \in I$ and $\tilde{d}^{*h} := -\sum_{j\in I} \tilde{\lambda}_j^* \tilde{D}_\tau^{-h} p_j^h \in Y^{*h}$ where $\tilde{\lambda}^*$ is the solution to the approximated dual problem (4.2.5).

A posteriori error estimates for an elliptic boundary value problem

We start by Let $V^h \subset H_0^1(\Omega)$ be a finite element subspace corresponding to the triangulation $\mathscr{T}^h$. Suppose $\alpha > 0$, $\beta \geq 0$, $L \in \mathbb{N}$, $\check{y}^h, \hat{y}^h, u_l^h, v_l^h \in V^h$, $1 \leq l \leq L$. We consider the variational equation

$$\text{Find } z \in H_0^1(\Omega): \quad \begin{aligned} &\alpha(z,w)_{H_0^1(\Omega)} + \beta(z,w)_{L^2(\Omega)} + \sum_{l=1}^{L} (u_l^h, w)_{H_0^1(\Omega)} (v_l^h, z)_{L^2(\Omega)} \\ &\quad = (\check{y}^h, w)_{L^2(\Omega)} + (\hat{y}^h, w)_{H_0^1(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega). \end{aligned} \tag{5.5.1}$$

Equation (5.5.1) can be used to evaluate $\tilde{F}_{\mathrm{BFGS}}^{-1}$ and $\tilde{E}$. Indeed, recall that $H_0^1(\Omega)$ is equipped with the inner product $(u,v)_{H_0^1(\Omega)} := \int_\Omega \nabla u^\top \nabla v \,\mathrm{d}\lambda$. Therefore, for arbitrary $\hat{y}^h \in H_0^1(\Omega)$, $z := \tilde{F}_{\mathrm{BFGS}}^{-1}\hat{y}^h$ is characterized by (5.5.1) with $\beta := \tau + \mu$, u_l^h and v_l^h chosen via (4.4.3), $L := 2\tilde{L}$ and $\check{y}^h := 0$. Furthermore, for arbitrary $\check{y}^h \in V^h$, $z := \tilde{E}\check{y}^h$ is characterized by (5.5.1) with $\alpha = 1$, $\beta := 0$, $L := 0$ and $\hat{y}^h := 0$, cf. (4.3.6). We further consider the discretized version of (5.5.1):

$$\text{Find } z^h \in V^h: \quad \begin{aligned} &\alpha(z^h,w^h)_{H_0^1(\Omega)} + \beta(z^h,w^h)_{L^2(\Omega)} + \sum_{l=1}^{L} (u_l^h, w^h)_{H_0^1(\Omega)} (v_l^h, z^h)_{L^2(\Omega)} \\ &\quad = (\check{y}^h, w^h)_{L^2(\Omega)} + (\hat{y}^h, w^h)_{H_0^1(\Omega)} \quad \text{for all } w^h \in V^h. \end{aligned} \tag{5.5.2}$$

Similar to the continuous case, the discrete operators $\tilde{F}_{\mathrm{BFGS}}^{-h}\hat{y}^h$ and $\tilde{E}^h\check{y}^h$ can be characterized as solutions to (5.5.2). We proceed by providing error estimates for the discretization error $e := z - z^h \in H_0^1(\Omega)$. The subsequent analysis adopts the proof of [2, Thm. 2.7] to the given setting. In particular, here the domain is allowed to be nonconvex and the right hand side $(\check{y}^h, \cdot)_{L^2(\Omega)} + (\hat{y}^h, \cdot)_{H_0^1(\Omega)} \in H^{-1}(\Omega)$ is considered. Let the bilinear form $a: H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be defined by

$$a(z,w) := \alpha(z,w)_{H_0^1(\Omega)} + \beta(z,w)_{L^2(\Omega)} + \sum_{l=1}^{L} (u_l^h, w)_{H_0^1(\Omega)} (v_l^h, z)_{L^2(\Omega)} \quad \text{for all } z, w \in H_0^1(\Omega).$$

As $z \in H_0^1(\Omega)$ solves $a(z,w) = (\check{y}^h, w)_{L^2(\Omega)} + (\hat{y}^h, w)_{H_0^1(\Omega)}$ for all $w \in H_0^1(\Omega)$, we find

$$\begin{aligned}
a(e,w) &= (\check{y}^h, w)_{L^2(\Omega)} + (\hat{y}^h, w)_{H_0^1(\Omega)} - a(z^h, w) \\
&= (\hat{y}^h - \alpha z^h, w)_{H_0^1(\Omega)} + (\check{y}^h - \beta z^h, w)_{L^2(\Omega)} + \sum_{l=1}^{L} (u_l^h, w)_{H_0^1(\Omega)} (v_l^h, z^h)_{L^2(\Omega)} \\
&= (\hat{y}^h - \alpha z^h - \sum_{l=1}^{L} (v_l^h, z^h)_{L^2(\Omega)} u_l^h, w)_{H_0^1(\Omega)} + (\check{y}^h - \beta z^h, w)_{L^2(\Omega)} \\
&= \sum_{T \in \mathscr{T}^h} (\nabla s^h, \nabla w)_{L^2(T)^2} + (r^h, w)_{L^2(T)},
\end{aligned}$$

where $s^h, r^h \in V^h$ are defined by

$$s^h := \hat{y}^h - \alpha z^h - \sum_{l=1}^{L} (v_l^h, z^h)_{L^2(\Omega)} u_l^h \quad \text{and} \quad r^h := \check{y}^h - \beta z^h. \tag{5.5.3}$$

In the following, we need to define the edge jump of the function $s^h \in V^h$. Since s^h is affine linear on each triangle $T \in \mathscr{T}^h$, $\operatorname{div}(\nabla s^h|_T) = 0$ and integration by parts (cf. [31, Thm. 4.6]) yields

$$(\nabla s^h, \nabla w)_{L^2(T)^2} = \int_T \nabla s^{h\top} \nabla w \, \mathrm{d}\lambda = \int_{\partial T} \nabla s^{h\top} n_T \operatorname{tr}_T w \, \mathrm{d}S \qquad \text{for all } w \in H_0^1(\Omega),$$

where $n_T : \partial T \to \mathbb{R}^2$ is the outer unit normal to ∂T. As $w \in H_0^1(\Omega)$, it holds $\operatorname{tr}_T w|_\Upsilon = 0$ for each boundary edge $\Upsilon \in \partial T \cap \partial\Omega$, which implies $\int_\Upsilon \frac{\partial s^h}{\partial n_T} \operatorname{tr}_T w \, \mathrm{d}S = 0$. For all other edges, i.e., $\Upsilon \in \partial T \setminus \partial\Omega$, denote by T and T' the triangles adjacent to Υ and define the edge jump along edge Υ by

$$\left[\frac{\partial s^h}{\partial n}\right]_\Upsilon := n_T^\top (\nabla s^h)_T + n_{T'}^\top (\nabla s^h)_{T'} \in \mathbb{R}, \tag{5.5.4}$$

Here, $(\nabla s^h)_T$ denotes the gradient of s^h on the triangle T. Note that $\nabla s^h \in \mathbb{R}^2$ is constant on each triangle as s^h is affine linear on each triangle. Furthermore, as $n_{T'} = -n_T$, (5.5.4) is actually a difference and the value of $[\frac{\partial s^h}{\partial n}]_\Upsilon$ does not depend on the permutation of T and T'. This leads to

DEFINITION 5.5.1 (Boundary residual). *The mapping* $[[\cdot]] : V^h \to L^2(\cup_{T \in \mathscr{T}^h} \partial T)$,

$$[[s^h]](\omega) := \begin{cases} \frac{1}{2}\left[\frac{\partial s^h}{\partial n}\right]_\Upsilon & \text{if } \omega \in \Upsilon \subset \partial T \setminus \partial\Omega \\ 0 & \text{if } \omega \in \Upsilon \subset \partial T \cap \partial\Omega \end{cases}$$

is called the boundary residual.

This enables us to write $(\nabla s^h, \nabla w)_{L^2(T)^2} = \int_{\partial T} [[s^h]] \operatorname{tr}_T w \, \mathrm{d}S$ for all $T \in \mathscr{T}^h$, $w \in H_0^1(\Omega)$ and we get

$$a(e,w) = \sum_{T \in \mathscr{T}^h} \left\{ \int_T r^h w \, \mathrm{d}\lambda + \int_{\partial T} [[s^h]] \operatorname{tr}_T w \, \mathrm{d}S \right\} \qquad \forall w \in H_0^1(\Omega). \tag{5.5.5}$$

This decomposition of the bilinear form into contributions from each triangle leads to an a posteriori error estimate. In order to use the Aubin-Nietsche trick, let $x \in L^2(\Omega)$ be arbitrary, and consider the adjoint problem

$$\text{Find } \Phi'_x \in H^1_0(\Omega): \quad \alpha(w,\Phi'_x)_{H^1_0(\Omega)} + \beta(w,\Phi'_x)_{L^2(\Omega)} + \sum_{l=1}^{L}(u^h_l,\Phi'_x)_{H^1_0(\Omega)}(v^h_l,w)_{L^2(\Omega)} = (x,w)_{L^2(\Omega)} \quad \text{for all } w \in H^1_0(\Omega). \tag{5.5.6}$$

Assumption 5.5.2. There exists constants $p \in (1,2]$ and $\check{C}_{\text{reg}} > 0$ such that for all $x \in L^p(\Omega)$, the corresponding adjoint solution $\Phi'_x \in H^1_0(\Omega)$ of (5.5.6) has the regularity $\Phi'_x \in W^{2,p}(\Omega)$, and

$$\|\Phi'_x\|_{W^{2,p}(\Omega)} \leq \check{C}_{\text{reg}}\|x\|_{L^p(\Omega)}.$$

Combining (5.5.5) with Assumption 5.5.2 leads to the following error estimate.

THEOREM 5.5.3. *Let Ω be an open, bounded polygonal domain with Lipschitz boundary and let $(\mathscr{T}^h)_h$, $0 < h \leq 1$, be a regular family of triangulations of Ω. Let $x \in L^2(\Omega)$ be arbitrary and denote by $z \in H^1_0(\Omega)$ and $z^h \in V^h$ the solutions of* (5.5.1) *and* (5.5.2)*, respectively. If the constants p and $\check{C}_{reg}$ fulfill Assumption 5.5.2, then there exists a constant $C_{p,\sigma,\Omega} > 0$, depending only on p, the regularity parameter σ of the triangulations and the domain Ω, such that*

$$|(x,z-z^h)_{L^2(\Omega)}| \leq \check{C}_{reg}C_{p,\sigma,\Omega}\left(\sum_{T\in\mathscr{T}^h}\eta_T^q\right)^{1/q}\|x\|_{L^2(\Omega)}.$$

Here, $q := p/(p-1)$, the local error indicators η_T are defined via

$$\eta_T := h_T^{3-2/p}\|r^h\|_{L^2(T)} + h_T^{5/2-2/p}\|R\|_{L^2(\partial T)} \qquad \textit{for all } T \in \mathscr{T}^h,$$

and the interior and boundary residuals $r^h \in V^h$ and $R \in L^2(\cup_{T\in\mathscr{T}^h}\partial T)$ are defined via

$$r^h := \check{y}^h - \beta z^h \qquad \textit{and} \qquad R := [[\hat{y}^h - \alpha z^h - \sum_{l=1}^{L}(v^h_l,z^h)_{L^2(\Omega)}u^h_l]]\,.$$

Proof. Denote by $\Phi'_x \in H^1_0(\Omega)$ the solution of the adjoint problem (5.5.6), i.e., Φ'_x fulfills $a(w,\Phi'_x) = (x,w)_{L^2(\Omega)}$ for all $w \in H^1_0(\Omega)$. By Assumption 5.5.2 it holds $\Phi'_x \in W^{2,p}(\Omega)$, $1 < p \leq 2$, and thus the Lagrange interpolation operator $I^h : W^{2,p}(\Omega) \subset C^0(\Omega) \to V^h$ is well-defined. By (5.5.5) and the orthogonality condition for the error in the Galerkin projection, i.e., $a(e,w^h) = 0$ for all $w^h \in V^h$, we find

$$\begin{aligned}(x,e)_{L^2(\Omega)} &= a(e,\Phi'_x) = a(e,\Phi'_x - I^h\Phi'_x)\\ &= \sum_{T\in\mathscr{T}^h}\Big\{\int_T r^h(\Phi'_x - I^h\Phi'_x)\,\mathrm{d}\lambda + \int_{\partial T}R\,\mathrm{tr}_T(\Phi'_x - I^h\Phi'_x)\,\mathrm{d}S\Big\}.\end{aligned} \tag{5.5.7}$$

We now estimate the errors $\Phi'_x - \Phi^h_x$ and $\mathrm{tr}_T(\Phi'_x - \Phi^h_x)$. By Theorem 2.7.1 there exists a constant $C_{p,\sigma}$ (depending only on p and the regularity parameter σ of the triangulation) such that it holds

$$\|v - I^hv\|_{L^2(T)} + h_T|v - I^hv|_{H^1(T)} \leq C_{p,\sigma}h_T^{3-2/p}|v|_{W^{2,p}(T)} \quad \text{for all } v \in W^{2,p}(T), T \in \mathscr{T}^h, 0 < h \leq 1.$$

This gives

$$\left|\int_T r^h(\Phi'_x - \Phi^h_x)\,\mathrm{d}\lambda\right| \le \|r^h\|_{L^2(T)}\|\Phi'_x - \Phi^h_x\|_{L^2(T)} \le C_{p,\sigma}h_T^{3-2/p}\|r^h\|_{L^2(T)}|\Phi'_x|_{W^{2,p}(T)}. \tag{5.5.8}$$

Furthermore, by Lemma 2.7.3 there exists a constant $C_\sigma > 0$ depending only on the regularity parameter σ of the triangulations such that

$$\|\operatorname{tr}_T(\Phi'_x - \Phi^h_x)\|^2_{L^2(T)} \le C_\sigma h_T^{-1}\|\Phi'_x - \Phi^h_x\|^2_{L^2(T)} + C_\sigma h_T|\Phi'_x - \Phi^h_x|^2_{H^1(T)} \le 2C_\sigma C^2_{p,\sigma}h_T^{5-4/p}|\Phi'_x|^2_{W^{2,p}(T)}.$$

Therefore,

$$\begin{aligned}\left|\int_{\partial T} R\operatorname{tr}_T(\Phi'_x - P_{Y^{*h}}\Phi'_x)\,\mathrm{d}S\right| &\le \|R\|_{L^2(\partial T)}\|\operatorname{tr}_T(\Phi'_x - P_{Y^{*h}}\Phi'_x)\|_{L^2(\partial T)} \\ &\le C_{p,\sigma}\sqrt{2C_\sigma}\,h_T^{5/2-2/p}\|R\|_{L^2(\partial T)}|\Phi'_x|_{W^{2,p}(T)}.\end{aligned} \tag{5.5.9}$$

Combining (5.5.7)–(5.5.9) and setting $\tilde{C}_{p,\sigma} := C_{p,\sigma}\max\left\{1, \sqrt{2C_\sigma}\right\}$ results in

$$\begin{aligned}|(x,e)_{L^2(\Omega)}| &\le \sum_{T\in\mathscr{T}^h} C_{p,\sigma}h_T^{3-2/p}\|r^h\|_{L^2(T)}|\Phi'_x|_{W^{2,p}(T)} + C_{p,\sigma}\sqrt{2C_\sigma}\,h_T^{5/2-2/p}\|R\|_{L^2(\partial T)}|\Phi'_x|_{W^{2,p}(T)} \\ &\le \tilde{C}_{p,\sigma}\sum_{T\in\mathscr{T}^h}(h_T^{3-2/p}\|r^h\|_{L^2(T)} + h_T^{5/2-2/p}\|R\|_{L^2(\partial T)})|\Phi'_x|_{W^{2,p}(T)} \\ &= \tilde{C}_{p,\sigma}\sum_{T\in\mathscr{T}^h}\eta_T|\Phi'_x|_{W^{2,p}(T)}.\end{aligned}$$

For $q := p/(p-1)$, Hölder's inequality (cf. [1, Chap. 2.27]) yields

$$|(x,e)_{L^2(\Omega)}| \le \tilde{C}_{p,\sigma}\left(\sum_{T\in\mathscr{T}^h}\eta_T^q\right)^{1/q}\left(\sum_{T\in\mathscr{T}^h}|\Phi'_x|^p_{W^{2,p}(T)}\right)^{1/p} = \tilde{C}_{p,\sigma}\left(\sum_{T\in\mathscr{T}^h}\eta_T^q\right)^{1/q}|\Phi'_x|_{W^{2,p}(\Omega)}.$$

Now, by Assumption 5.5.2 there exists a constant $\check{C}_{\mathrm{reg}} > 0$ (independent of x and h) such that $\|\Phi'_x\|_{W^{2,p}(\Omega)} \le \check{C}_{\mathrm{reg}}\|x\|_{L^p(\Omega)}$. By [1, Thm. 2.14], there exists a constant $C_{p,\Omega}$, depending only on p and the domain Ω such that $\|x\|_{L^p(\Omega)} \le C_{p,\Omega}\|x\|_{L^2(\Omega)}$. This yields the desired estimate

$$|(x, z - z^h)_{L^2(\Omega)}| = |(x,e)_{L^2(\Omega)}| \le \tilde{C}_{p,\sigma}\check{C}_{\mathrm{reg}}C_{p,\Omega}\left(\sum_{T\in\mathscr{T}^h}\eta_T^q\right)^{1/q}\|x\|_{L^2(\Omega)}.$$ □

Remark 5.5.4. Choosing $x = z - z^h$ in Theorem 5.5.3 yields the $L^2(\Omega)$-error estimate

$$\|z - z^h\|_{L^2(\Omega)} \le \check{C}_{\mathrm{reg}}C_{p,\sigma,\Omega}\left(\sum_{T\in\mathscr{T}^h}\eta_T^q\right)^{1/q}.$$

Remark 5.5.5. Theorem 5.5.3 is a generalization of [2, Thm. 2.7]. In fact, since $p \le 2$, it holds $q = p/(p-1) \ge 2$ and $x \mapsto x^q$ is convex. Therefore, Jensen's inequality (cf. [12, Prop. 9.24]) yields

$(\frac{1}{2}(a+b))^q \leq \frac{1}{2}(a^q+b^q)$ for all $a,b \in \mathbb{R}$. This leads to

$$\eta_T^q \leq 2^{q-1}\left(h_T^{(3-2/p)q}\|r^h\|_{L^2(T)}^q + h_T^{(5/2-2/p)q}\|R\|_{L^2(\partial T)}^q\right) \qquad \text{for all } T \in \mathscr{T}^h$$

and thus Theorem 5.5.3 and Remark 5.5.4 yield

$$\|z-z^h\|_{L^2(\Omega)} \leq \check{C}_{\mathrm{reg}} C_{p,\sigma,\Omega} 2^{1-1/q}\left(\sum_{T\in\mathscr{T}^h} h_T^{(3-2/p)q}\|r^h\|_{L^2(T)}^q + h_T^{(5/2-2/p)q}\|R\|_{L^2(\partial T)}^q\right)^{1/q}.$$

In particular, if $\alpha = 1$, $L = 0$, $\hat{y}^h = 0$ and Ω is convex, Lemma 5.4.2 shows that Assumption 5.5.2 is fulfilled with $p = 2$ and $\check{C}_{\mathrm{reg}} > 0$ depends only on Ω. Consequently, the estimate

$$\|z-z^h\|_{L^2(\Omega)} \leq C_{\sigma,\Omega}\left(\sum_{T\in\mathscr{T}^h} h_T^4\|r^h\|_{L^2(T)}^2 + h_T^3\|R\|_{L^2(\partial T)}^2\right)^{1/2}$$

from [2, Thm. 2.7] is recovered.

A posteriori error estimates for polygonal, convex domains

Using Theorem 5.5.3, we develop a posteriori error estimates $\hat{\varepsilon}_F$ and $\hat{\varepsilon}_{F^{-1}}$ which fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

COROLLARY 5.5.6 (A posteriori error estimate for $\tilde{E}$). *Let Ω be a convex, bounded domain with polygonal boundary and let $(\mathscr{T}^h)_h$, $0 < h \leq 1$, be a regular family of triangulations of Ω with regularity parameter σ. For $x^h \in V^h$, define the local error indicator η_T via*

$$\eta_T := h_T^2\|x^h\|_{L^2(T)} + h_T^{3/2}\|\,[[\tilde{E}^h x^h]]\,\|_{L^2(\partial T)} \qquad \textit{for all } T \in \mathscr{T}^h.$$

Then there exists a constant $C_{\Omega,\sigma} > 0$ (depending only on Ω and σ) such that

$$|((\tilde{E}-\tilde{E}^h)x^h, y^h)_{L^2(\Omega)}| \leq C_{\Omega,\sigma}\left(\sum_{T\in\mathscr{T}^h}\eta_T^2\right)^{1/2}\|y^h\|_{L^2(\Omega)} \quad \textit{for all } x^h, y^h \in V^h,$$

i.e., $e^E_{x^h,y^h} := C_{\Omega,\sigma}\left(\sum_{T\in\mathscr{T}^h}\eta_T^2\right)^{1/2}\|y^h\|_{L^2(\Omega)}$ for $x^h, y^h \in Y^{*h}$ fulfills Assumption 4.3.6.

Proof. By setting $\alpha = 1$, $\beta = 0$, $L = 0$, $\check{y}^h = x^h$ and $\hat{y}^h = 0$ in (5.5.1) and (5.5.2), we find $z = \tilde{E}x^h$ and $z^h = \tilde{E}^h x^h$ (cf. (4.3.1) and (4.3.6)). This leads to the adjoint problem

$$\text{Find } \Phi'_y \in H_0^1(\Omega): \qquad (w, \Phi'_y)_{H_0^1(\Omega)} = (y,w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega),$$

where $y \in L^2(\Omega)$ is arbitrary. By Lemma 5.4.2, the solution $\Phi'_y \in H_0^1(\Omega)$ to this adjoint problem is an element of $H^2(\Omega)$ and there exists a constant $C_\Omega > 0$, only depending on Ω, such that $\|\Phi'_y\|_{H^2(\Omega)} \leq C_\Omega\|y\|_{L^2(\Omega)}$. Thus, Assumption 5.5.2 is fulfilled with constants $p = 2$ and $\check{C}_{\mathrm{reg}} := C_\Omega$. Consequently, by Theorem 5.5.3, there exists a constant $C_{\sigma,\Omega} > 0$ (depending only on the regularity parameter σ of the triangulations and the domain Ω) such that

$$|((\tilde{E}-\tilde{E}^h)x^h, y^h)_{L^2(\Omega)}| = |(y^h, z - z^h)_{L^2(\Omega)}| \le C_{\sigma,\Omega} C_\Omega \left(\sum_{T \in \mathscr{T}^h} \eta_T^2 \right)^{1/2} \|y^h\|_{L^2(\Omega)} \quad \text{for all } x^h, y^h \in V^h,$$

where the local error indicator η_T is defined via

$$\eta_T := h_T^2 \|x^h\|_{L^2(T)} + h_T^{3/2} \| [[\tilde{E}^h x^h]] \|_{L^2(\partial T)} \qquad \text{for all } T \in \mathscr{T}^h. \qquad \square$$

THEOREM 5.5.7. *Assume that $\Omega \subset \mathbb{R}^2$ is a convex, bounded and open domain with polygonal boundary. If $Q = 0$, then there exist constants $C_{\Omega,\sigma}, C_{\Omega,\sigma,\alpha} > 0$ such that*

$$\begin{aligned}
&O_{F^{-1}}(p_j', h) := R_X^{-1} \iota^* \tilde{D}_\tau^{-h} p_j^h, \qquad C_F := C_{\Omega,\sigma}, \quad C_{F^{-1}} := C_{\Omega,\sigma,\alpha},\\
&\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) := \tau \left(\sum_{T \in \mathscr{T}^h} \left(h_T^2 \|\tilde{d}^{*h}\|_{L^2(T)} + h_T^{3/2} \| [[\tilde{E}^h \tilde{d}^{*h}]] \|_{L^2(\partial T)} \right)^2 \right)^{1/2} \|\tilde{d}^{*h}\|_{L^2(\Omega)},\\
&\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) := \left(\sum_{T \in \mathscr{T}^h} \left(\tau h_T^2 \|\tilde{D}_\tau^{-h} p_j^h\|_{L^2(T)} + h_T^{3/2} \| [[p_j^h - \alpha \tilde{D}_\tau^{-h} p_j^h]] \|_{L^2(\partial T)} \right)^2 \right)^{1/2} \|p_i^h\|_{L^2(\Omega)}
\end{aligned}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. By setting $\beta = \tau$, $L = 0$, $\check{y}^h = 0$ and $\hat{y}^h = y^h$ in (5.5.1) and (5.5.2), we find $z = \tilde{D}_\tau^{-1} y^h$ and $z^h = \tilde{D}_\tau^{-h} y^h$ (cf. (4.3.5) and (4.3.7)). This leads to the adjoint problem

$$\text{Find } \Phi_x' \in H_0^1(\Omega): \qquad \alpha(w, \Phi_x')_{H_0^1(\Omega)} + \tau(w, \Phi_x')_{L^2(\Omega)} = (x, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega),$$

where $x \in L^2(\Omega)$ is arbitrary. By Lemma 5.4.2, the solution $\Phi_x' \in H_0^1(\Omega)$ to this adjoint problem is an element of $H^2(\Omega)$ and there exists a constant $C_\Omega > 0$, only depending on Ω, such that $\|\Phi_x'\|_{H^2(\Omega)} \le C_\Omega/\alpha \|x\|_{L^2(\Omega)}$. Thus, Assumption 5.5.2 is fulfilled with constants $p = 2$ and $\check{C}_{\text{reg}} := C_\Omega/\alpha$. Consequently, by Theorem 5.5.3, there exists a constant $C_{\sigma,\Omega} > 0$ (depending only on the regularity parameter σ of the triangulations and the domain Ω) such that

$$|(x^h, (\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h}) y^h)_{L^2(\Omega)}| \le \frac{C_{\sigma,\Omega} C_\Omega}{\alpha} \left(\sum_{T \in \mathscr{T}^h} \eta_T^2 \right)^{1/2} \|x^h\|_{L^2(\Omega)} \qquad \text{for all } x^h, y^h \in Y^{*h},$$

where the local error indicator η_T is defined via

$$\eta_T := \tau h_T^2 \|\tilde{D}_\tau^{-h} y^h\|_{L^2(T)} + h_T^{3/2} \| [[y^h - \alpha \tilde{D}_\tau^{-h} y^h]] \|_{L^2(\partial T)} \qquad \text{for all } T \in \mathscr{T}^h.$$

Therefore, the error estimate

$$e_{x^h, y^h}^{D^{-1}} := \frac{C_{\sigma,\Omega} C_\Omega}{\alpha} \left(\sum_{T \in \mathscr{T}^h} \left(\tau h_T^2 \|\tilde{D}_\tau^{-h} y^h\|_{L^2(T)} + h_T^{3/2} \| [[y^h - \alpha \tilde{D}_\tau^{-h} y^h]] \|_{L^2(\partial T)} \right)^2 \right)^{1/2} \|x^h\|_{L^2(\Omega)}$$

fulfills Assumption 4.3.7. Furthermore, by Corollary 5.5.6, there exists a constant $C_{\Omega,\sigma} > 0$ such that for $x^h, y^h \in Y^{*h}$ the error estimate

$$e^E_{x^h,y^h} := C_{\Omega,\sigma}\left(\sum_{T\in\mathscr{T}^h}\left(h_T^2\|x^h\|_{L^2(T)} + h_T^{3/2}\|\,[[\tilde{E}^h x^h]]\,\|_{L^2(\partial T)}\right)^2\right)^{1/2}\|y^h\|_{L^2(\Omega)}$$

fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.3.7 are fulfilled, Lemma 4.3.9 implies that the approximation $p_j^{F^{-1}} := R_X^{-1}\iota^*\tilde{D}_\tau^{-h}p_j^h$ and the error estimates $e_{i,j,F^{-1}} := e^{D^{-1}}_{p_i^h,p_j^h}$ and $e_F := \tau e^E_{\tilde{d}^{*h},\tilde{d}^{*h}}$, $i,j\in I$, fulfill Assumption 4.2.4. From $\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h)\to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h)\to 0$ as $h\to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. ∎

THEOREM 5.5.8. *Assume that $\Omega\subset\mathbb{R}^2$ is a convex, bounded and open domain with polygonal boundary. If Q is given as the BFGS curvature operator (cf. Section 4.4), then there exist constants $C, C_{\sigma,\Omega} > 0$, independent of h, such that*

$$O_{F^{-1}}(p_j',h) := R_X^{-1}\iota^*\tilde{F}_{BFGS}^{-h}p_j^h, \qquad C_F := C_{\sigma,\Omega}, \quad C_{F^{-1}} := C,$$

$$\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) := (\tau+\mu)\left(\sum_{T\in\mathscr{T}^h}\left(h_T^2\|\tilde{d}^{*h}\|_{L^2(T)} + h_T^{3/2}\|\,[[\tilde{E}^h\tilde{d}^{*h}]]\,\|_{L^2(\partial T)}\right)^2\right)^{1/2}\|\tilde{d}^{*h}\|_{L^2(\Omega)},$$

$$\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) := \left(\sum_{T\in\mathscr{T}^h}\left((\tau+\mu)h_T^2\|r^h\|_{L^2(T)} + h_T^{3/2}\|R\|_{L^2(\partial T)}\right)^2\right)^{1/2}\|p_i^h\|_{L^2(\Omega)}$$

$$\textit{with } r^h := \tilde{F}_{BFGS}^{-h}y^h \quad \textit{and} \quad R := [[p_j^h - \alpha\tilde{F}_{BFGS}^{-h}p_j^h - \sum_{l=1}^{2\tilde{L}}(v_l^h,\tilde{F}_{BFGS}^{-h}p_j^h)_{L^2(\Omega)}u_l^h]]$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. In order to apply Lemma 4.4.7, we need an error estimate $e^{F_{\mathrm{BFGS}}^{-1}}_{x^h,y^h}\geq 0$ which fulfills Assumption 4.4.6, i.e.,

$$|(x^h,(\tilde{F}_{\mathrm{BFGS}}^{-1}-\tilde{F}_{\mathrm{BFGS}}^{-h})y^h)_{L^2(\Omega)}| \leq e^{F_{\mathrm{BFGS}}^{-1}}_{x^h,y^h} \qquad \text{for all } x^h,y^h\in V^h.$$

We use Theorem 5.5.3 to construct such an error estimate. Indeed, in (5.5.1) and (5.5.2) set $\beta=\tau+\mu$, $L=2\tilde{L}$, $\check{y}^h=0$, $\hat{y}^h=y^h$ and u_l^h, v_l^h, $1\leq l\leq 2\tilde{L}$, as the BFGS vectors from Section 4.4.1. Then $z=\tilde{F}_{\mathrm{BFGS}}^{-1}y^h$ and $z^h=\tilde{F}_{\mathrm{BFGS}}^{-h}y^h$ (cf. (4.4.11) and (4.4.14)). This leads to the adjoint problem

$$\begin{aligned}\text{Find } \Phi_x'\in H_0^1(\Omega): \quad & \alpha(w,\Phi_x')_{H_0^1(\Omega)} + (\tau+\mu)(w,\Phi_x')_{L^2(\Omega)} + \sum_{l=1}^{2\tilde{L}}(u_l^h,\Phi_x')_{H_0^1(\Omega)}(v_l^h,w)_{L^2(\Omega)}\\ &= (x,w)_{L^2(\Omega)} \qquad \text{for all } w\in H_0^1(\Omega),\end{aligned}$$

where $x\in L^2(\Omega)$ is arbitrary. Using the definition of $\tilde{F}_{\mathrm{BFGS}}$ shows that this is equivalent to

$$\text{Find } \Phi_x'\in H_0^1(\Omega): \quad (\tilde{F}_{\mathrm{BFGS}}w,\Phi_x')_{H_0^1(\Omega)} = (x,w)_{L^2(\Omega)} \qquad \text{for all } w\in H_0^1(\Omega).$$

Choosing $w=\tilde{F}_{\mathrm{BFGS}}^{-1}u_l^h$ yields $(u_l^h,\Phi_x')_{H_0^1(\Omega)} = (x,\tilde{F}_{\mathrm{BFGS}}^{-1}u_l^h)_{L^2(\Omega)}$ for $1\leq l\leq 2\tilde{L}$. Therefore, the adjoint solution Φ_x' also solves

$$\text{Find } \Phi_x' \in H_0^1(\Omega): \quad \alpha(w,\Phi_x')_{H_0^1(\Omega)} + (\tau+\mu)(w,\Phi_x')_{L^2(\Omega)} = (c,w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega),$$

where $c := x - \sum_{l=1}^{2\tilde{L}} (x, \tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h)_{L^2(\Omega)} v_l^h$. Therefore, Lemma 5.4.2 can be applied which shows that Φ_x' has the regularity $\Phi_x' \in H^2(\Omega)$ and there exists a constant $C_{\Omega,\alpha}$ such that

$$\|\Phi_x'\|_{H^2(\Omega)} \leq C_{\Omega,\alpha}\|c\|_{L^2(\Omega)} \leq C_{\Omega,\alpha}(1 + \sum_{l=1}^{2\tilde{L}} \|\tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h\|_{L^2(\Omega)} \|v_l^h\|_{L^2(\Omega)}) \|x\|_{L^2(\Omega)}.$$

Consequently, $p = 2$ and $\check{C}_{\mathrm{reg}} := C_{\Omega,\alpha}(1 + \sum_{l=1}^{2\tilde{L}} \|\tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h\|_{L^2(\Omega)})$ fulfill Assumption 5.5.2. Thus, Theorem 5.5.3 is applicable and there exists a constant $C_{\sigma,\Omega}$ such that for arbitrary $x^h, y^h \in V^h$

$$|(x^h, (\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h}) y^h)_{L^2(\Omega)}| \leq \check{C}_{\mathrm{reg}} C_{\sigma,\Omega} \left(\sum_{T \in \mathscr{T}^h} \eta_T^2 \right)^{1/2} \|x^h\|_{L^2(\Omega)} =: e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}}.$$

Here, for all $T \in \mathscr{T}^h$, the local error indicators η_T are defined via

$$\eta_T := (\tau+\mu) h_T^2 \|\tilde{F}_{\mathrm{BFGS}}^{-h} y^h\|_{L^2(T)} + h_T^{3/2} \| [[y^h - \alpha \tilde{F}_{\mathrm{BFGS}}^{-h} y^h - \sum_{l=1}^{2\tilde{L}} (v_l^h, \tilde{F}_{\mathrm{BFGS}}^{-h} y^h)_{L^2(\Omega)} u_l^h]] \|_{L^2(\partial T)}.$$

In particular, $e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}}$ fulfills Assumption 4.4.6. Furthermore, by Corollary 5.5.6 there exists a constant $C_{\sigma,\Omega} > 0$ such that for $x^h, y^h \in Y^{*h}$ the error estimate

$$e_{x^h,y^h}^E := C_{\sigma,\Omega} \left(\sum_{T \in \mathscr{T}^h} \left(h_T^2 \|x^h\|_{L^2(T)} + h_T^{3/2} \| [[\tilde{E}^h x^h]] \|_{L^2(\partial T)} \right)^2 \right)^{1/2} \|y^h\|_{L^2(\Omega)}$$

fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.4.6 are fulfilled, Lemma 4.4.7 implies that

$$p_j^{F_{\mathrm{BFGS}}^{-1}} := R_X^{-1} \iota^* \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h, \qquad e_{i,j,F_{\mathrm{BFGS}}^{-1}} := e_{p_i^h,p_j^h}^{F_{\mathrm{BFGS}}^{-1}}, \qquad e_{F_{\mathrm{BFGS}}} := (\tau+\mu) e_{\tilde{d}^{*h},\tilde{d}^{*h}}^E,$$

fulfill Assumption 4.2.4. From $\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) \to 0$ as $h \to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. □

A posteriori error estimates for polygonal domains with one reentrant corner

In the last section a posteriori error estimates for convex domains are developed. For nonconvex domains however, the regularity result Lemma 5.4.2 is not valid. Thus, in order to develop a posteriori error estimates for nonconvex domains, we adopt the proofs of the last section by replacing Lemma 5.4.2 with Lemma 5.4.8. We need the following preliminary lemma.

LEMMA 5.5.9 ([1, Chap. 2.27]). *Let $0 < p \le q < \infty$ and $(a_i)_{i\in\mathbb{N}} \subset \mathbb{R}$ with $\sum_{i\in\mathbb{N}} |a_i|^p < \infty$. Then*

$$\left(\sum_{i\in\mathbb{N}} |a_i|^q\right)^{1/q} \le \left(\sum_{i\in\mathbb{N}} |a_i|^p\right)^{1/p}.$$

THEOREM 5.5.10 (A posteriori error estimate for $\tilde{E}$). *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded, open polygonal domain with one reentrant corner of angle $\omega \in (\pi, 2\pi)$, and let $(\mathscr{T}^h)_h$, $0 < h \le 1$, be a regular family of triangulations of Ω with corresponding finite element space V^h. Let $\varepsilon \in (0, \frac{\pi}{\omega})$ and $x^h, y^h \in V^h$ be arbitrary. Define the local error indicator η_T via*

$$\eta_T := h_T^{1+\pi/\omega-\varepsilon} \|x^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \| [[\tilde{E}^h x^h]] \|_{L^2(\partial T)} \qquad \textit{for all } T \in \mathscr{T}^h.$$

Then there exists a constant $C_{\varepsilon,\sigma,\Omega} > 0$ (depending only on ε, σ and Ω) such that

$$|((\tilde{E} - \tilde{E}^h)x^h, y^h)_{L^2(\Omega)}| \le C_{\varepsilon,\sigma,\Omega} \left(\sum_{T\in\mathscr{T}^h} \eta_T^{2\omega/\pi}\right)^{\pi/(2\omega)} \|y^h\|_{L^2(\Omega)} \qquad \textit{for all } x^h, y^h \in V^h,$$

i.e., $e^E_{x^h,y^h} := C_{\varepsilon,\sigma,\Omega} \left(\sum_{T\in\mathscr{T}^h} \eta_T^{2\omega/\pi}\right)^{\pi/(2\omega)} \|y^h\|_{L^2(\Omega)}$ for $x^h, y^h \in Y^{*h} := V^h$ fulfills Assumption 4.3.6.

Proof. By setting $\alpha = 1$, $\beta = 0$, $L = 0$, $\check{y}^h = x^h$ and $\hat{y}^h = 0$ in (5.5.1) and (5.5.2), we find $z = \tilde{E}x^h$ and $z^h = \tilde{E}^h x^h$ (cf. (4.3.1) and (4.3.6)). This leads to the adjoint problem

$$\text{Find } \Phi'_y \in H_0^1(\Omega): \qquad (w, \Phi'_y)_{H_0^1(\Omega)} = (y, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega),$$

where $y \in L^2(\Omega)$ is arbitrary. Let $\varepsilon \in (0, \frac{\pi}{\omega})$ be arbitrary and define $p \in \mathbb{R}$ via $2 - 2/p = \frac{\pi}{\omega} - \varepsilon$. Since $\varepsilon \in (0, \frac{\pi}{\omega})$ and $\omega \in (\pi, 2\pi)$, it holds $1 < p < 2/(2 - \frac{\pi}{\omega}) < 2$. Thus, Lemma 5.4.8 implies that the solution $\Phi'_y \in H_0^1(\Omega)$ is an element of $W^{2,p}(\Omega)$ and there exists a constant $C_{\varepsilon,\Omega}$ (depending only on ε and Ω) such that $\|\Phi'_y\|_{W^{2,p}(\Omega)} \le C_{\varepsilon,\Omega}\|y\|_{L^p(\Omega)}$. Therefore, Assumption 5.5.2 is fulfilled with constants p and $\check{C}_{\text{reg}} := C_{\varepsilon,\Omega}$. Consequently, by Theorem 5.5.3, there exists a constant $C_{\varepsilon,\sigma,\Omega} > 0$ (depending only on ε, the regularity parameter σ of the triangulations and the domain Ω) such that

$$|((\tilde{E} - \tilde{E}^h)x^h, y^h)_{L^2(\Omega)}| = |(y^h, z - z^h)_{L^2(\Omega)}| \le C_{\varepsilon,\sigma,\Omega} C_{\varepsilon,\Omega} \left(\sum_{T\in\mathscr{T}^h} \eta_T^q\right)^{1/q} \|y^h\|_{L^2(\Omega)} \quad \text{for all } x^h, y^h \in V^h,$$

where $q := p/(p-1)$ and the local error indicator η_T is defined via

$$\eta_T := h_T^{1+\pi/\omega-\varepsilon} \|x^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \| [[\tilde{E}^h x^h]] \|_{L^2(\partial T)} \qquad \text{for all } T \in \mathscr{T}^h.$$

Since $p < 2/(2 - \frac{\pi}{\omega})$, it holds $q = 1 + 1/(p-1) > 2\omega/\pi$. Thus, Lemma 5.5.9 yields

$$|((\tilde{E} - \tilde{E}^h)x^h, y^h)_{L^2(\Omega)}| \le C_{\varepsilon,\sigma,\Omega} C_{\varepsilon,\Omega} \left(\sum_{T\in\mathscr{T}^h} \eta_T^{2\omega/\pi}\right)^{\pi/(2\omega)} \|y^h\|_{L^2(\Omega)} \qquad \text{for all } x^h, y^h \in V^h. \qquad \square$$

Remark 5.5.11. Theorem 5.5.10 is most useful for $0 < \varepsilon \ll \frac{\pi}{\omega}$.

THEOREM 5.5.12. *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded, open, polygonal domain with one reentrant corner of angle $\omega \in (\pi, 2\pi)$ and let $(\mathscr{T}^h)_h$, $0 < h \le 1$, be a regular family of triangulations of Ω with corresponding finite element space V^h. Let $\varepsilon \in (0, \frac{\pi}{\omega})$ and $x^h, y^h \in V^h$ be arbitrary. If $Q = 0$, then there exist constants $C_{\varepsilon,\sigma,\Omega}, C_{\varepsilon,\alpha,\tau,\sigma,\Omega} > 0$ such that*

$$
\begin{aligned}
&O_{F^{-1}}(p_j', h) := R_X^{-1}\iota^*\tilde{D}_\tau^{-h} p_j^h, \qquad\qquad C_F := C_{\varepsilon,\sigma,\Omega}, \quad C_{F^{-1}} := C_{\varepsilon,\alpha,\tau,\sigma,\Omega},\\
&\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) := \tau\left(\sum_{T\in\mathscr{T}^h}\left(h_T^{1+\pi/\omega-\varepsilon}\|\tilde{d}^{*h}\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon}\|\,[[\tilde{E}^h\tilde{d}^{*h}]]\,\|_{L^2(\partial T)}\right)^{2\omega/\pi}\right)^{\pi/(2\omega)}\|\tilde{d}^{*h}\|_{L^2(\Omega)},\\
&\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) := \left(\sum_{T\in\mathscr{T}^h}\left(\tau h_T^{1+\pi/\omega-\varepsilon}\|r^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon}\|R\|_{L^2(\partial T)}\right)^{2\omega/\pi}\right)^{\pi/(2\omega)}\|p_i^h\|_{L^2(\Omega)}\\
&\qquad \textit{with } r^h := \tilde{D}_\tau^{-h}p_j^h \textit{ and } R := [[p_j^h - \alpha\tilde{D}_\tau^{-h}p_j^h]]
\end{aligned}
$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. By setting $\beta = \tau$, $L = 0$, $\check{y}^h = 0$ and $\hat{y}^h = y^h$ in (5.5.1) and (5.5.2), we find $z = \tilde{D}_\tau^{-1}y^h$ and $z^h = \tilde{D}_\tau^{-h}y^h$ (cf. (4.3.5) and (4.3.7)). This leads to the adjoint problem

$$
\text{Find } \Phi_x' \in H_0^1(\Omega): \qquad \alpha(w, \Phi_x')_{H_0^1(\Omega)} + \tau(\iota^* w, \iota^*\Phi_x')_{L^2(\Omega)} = (x, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega),
$$

where $x \in L^2(\Omega)$ is arbitrary. Let $\varepsilon \in (0, \frac{\pi}{\omega})$ be arbitrary and define $p \in \mathbb{R}$ via $2 - 2/p = \frac{\pi}{\omega} - \varepsilon$. Since $\varepsilon \in (0, \frac{\pi}{\omega})$ and $\omega \in (\pi, 2\pi)$, it holds $1 < p < 2/(2 - \frac{\pi}{\omega}) < 2$. Thus, Lemma 5.4.8 implies that the solution $\Phi_x' \in H_0^1(\Omega)$ is an element of $W^{2,p}(\Omega)$ and there exists a constant $C_{\varepsilon,\alpha,\tau,\Omega}$ (depending only on ε, α, τ and Ω) such that $\|\Phi_x'\|_{W^{2,p}(\Omega)} \le C_{\varepsilon,\alpha,\tau,\Omega}\|x\|_{L^p(\Omega)}$. Therefore, Assumption 5.5.2 is fulfilled with constants p and $\check{C}_{\mathrm{reg}} := C_{\varepsilon,\alpha,\tau,\Omega}$. Consequently, by Theorem 5.5.3, there exists a constant $C_{\varepsilon,\sigma,\Omega} > 0$ (depending only on ε, the regularity parameter σ of the triangulations and the domain Ω) such that

$$
|(x^h, (\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h})y^h)_{L^2(\Omega)}| \le C_{\varepsilon,\sigma,\Omega}C_{\varepsilon,\alpha,\tau,\Omega}\left(\sum_{T\in\mathscr{T}^h}\eta_T^q\right)^{1/q}\|x^h\|_{L^2(\Omega)} \quad \text{for all } x^h, y^h \in V^h,
$$

where $q := p/(p-1)$ and the local error indicator η_T is defined via

$$
\eta_T := \tau h_T^{1+\pi/\omega-\varepsilon}\|\tilde{D}_\tau^{-h}y^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon}\|\,[[y^h - \alpha\tilde{D}_\tau^{-h}y^h]]\,\|_{L^2(\partial T)} \qquad \text{for all } T \in \mathscr{T}^h.
$$

Since $p < 2/(2 - \frac{\pi}{\omega})$, it holds $q = 1 + 1/(p-1) > 2\omega/\pi$. Thus, for all $x^h, y^h \in V^h$, Lemma 5.5.9 yields

$$
|(x^h, (\tilde{D}_\tau^{-1} - \tilde{D}_\tau^{-h})y^h)_{L^2(\Omega)}| \le C_{\varepsilon,\sigma,\Omega}C_{\varepsilon,\alpha,\tau,\Omega}\left(\sum_{T\in\mathscr{T}^h}\eta_T^{2\omega/\pi}\right)^{\pi/(2\omega)}\|x^h\|_{L^2(\Omega)} =: e_{x^h,y^h}^{D^{-1}},
$$

i.e., $e_{x^h,y^h}^{D^{-1}}$ fulfills Assumption 4.3.7. Furthermore, by Theorem 5.5.10, there exists a constant $C_{\varepsilon,\Omega,\sigma} > 0$

such that for $x^h, y^h \in Y^{*h}$ the error estimate

$$e^E_{x^h,y^h} := C_{\varepsilon,\sigma,\Omega} \left(\sum_{T\in\mathscr{T}^h} \left(h_T^{1+\pi/\omega-\varepsilon} \|x^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \| [[\tilde{E}^h x^h]] \|_{L^2(\partial T)} \right)^{2\omega/\pi} \right)^{\pi/(2\omega)} \|y^h\|_{L^2(\Omega)}$$

fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.3.7 are fulfilled, Lemma 4.3.9 implies that the approximation $p_j^{F^{-1}} := R_X^{-1}\iota^*\tilde{D}_\tau^{-h} p_j^h$ and the error estimates $e_{i,j,F^{-1}} := e^{D^{-1}}_{p_i^h,p_j^h}$ and $e_F := \tau e^E_{\tilde{d}^{*h},\tilde{d}^{*h}}$, $i,j \in I$, fulfill Assumption 4.2.4. From $\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) \to 0$ as $h \to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. □

Example 5.5.13. Suppose that $\Omega = (-a,a)^2 \setminus [0,a)^2$, $a > 0$ is a L-shaped domain and let $(\mathscr{T}^h)_h$, $0 < h \le 1$, be a regular family of triangulations of Ω with corresponding finite element space V^h. Then Ω is a polygonal domain with one reentrant corner of angle $\omega = 3/2\pi$. Let $\varepsilon \in (0,\frac{2}{3})$ and $x^h, y^h \in V^h$ be arbitrary and suppose $Q = 0$. Then there exist constants $C_{\varepsilon,\sigma,\Omega}, C_{\varepsilon,\alpha,\tau,\sigma,\Omega} > 0$ such that

$$\begin{aligned}
O_{F^{-1}}(p'_j,h) &:= R_X^{-1}\iota^*\tilde{D}_\tau^{-h}p_j^h, \qquad C_F := C_{\varepsilon,\sigma,\Omega}, \quad C_{F^{-1}} := C_{\varepsilon,\alpha,\tau,\sigma,\Omega},\\
\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) &:= \tau \left(\sum_{T\in\mathscr{T}^h} \left(h_T^{5/3-\varepsilon} \|\tilde{d}^{*h}\|_{L^2(T)} + h_T^{7/6-\varepsilon} \| [[\tilde{E}^h\tilde{d}^{*h}]] \|_{L^2(\partial T)} \right)^3 \right)^{1/3} \|\tilde{d}^{*h}\|_{L^2(\Omega)},\\
\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) &:= \left(\sum_{T\in\mathscr{T}^h} \left(\tau h_T^{5/3-\varepsilon} \|r^h\|_{L^2(T)} + h_T^{7/6-\varepsilon} \|R\|_{L^2(\partial T)} \right)^3 \right)^{1/3} \|p_i^h\|_{L^2(\Omega)}\\
&\quad \text{with } r^h := \tilde{D}_\tau^{-h}p_j^h \text{ and } R := [[p_j^h - \alpha\tilde{D}_\tau^{-h}p_j^h]]
\end{aligned}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

THEOREM 5.5.14. *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded, open, polygonal domain with one reentrant corner of angle $\omega \in (\pi, 2\pi)$. If Q is given as the BFGS curvature operator (cf. Section 4.4), then there exist constants $C, C_{\varepsilon,\sigma,\Omega} > 0$, independent of h, such that*

$$\begin{aligned}
O_{F^{-1}}(p'_j,h) &:= R_X^{-1}\iota^*\tilde{F}_{BFGS}^{-h}p_j^h, \qquad C_F := C_{\varepsilon,\sigma,\Omega}, \quad C_{F^{-1}} := C,\\
\hat{\varepsilon}_F(\tilde{d}^*,\mathring{d}',h) &:= (\tau+\mu) \left(\sum_{T\in\mathscr{T}^h} \eta_T^{2\omega/\pi} \right)^{\pi/(2\omega)} \|\tilde{d}^{*h}\|_{L^2(\Omega)}\\
&\quad \textit{with } \eta_T := h_T^{1+\pi/\omega-\varepsilon} \|\tilde{d}^{*h}\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \| [[\tilde{E}^h\tilde{d}^{*h}]] \|_{L^2(\partial T)} \qquad \textit{for all } T \in \mathscr{T}^h,\\
\hat{\varepsilon}_{F^{-1}}(p_i,p_j,h) &:= \left(\sum_{T\in\mathscr{T}^h} \left((\tau+\mu) h_T^{1+\pi/\omega-\varepsilon} \|r^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \|R\|_{L^2(\partial T)} \right)^{2\omega/\pi} \right)^{\pi/(2\omega)} \|p_i^h\|_{L^2(\Omega)}\\
&\quad \textit{with } r^h := \tilde{F}_{BFGS}^{-h} y^h \textit{ and } R := [[p_j^h - \alpha\tilde{F}_{BFGS}^{-h}p_j^h - \sum_{l=1}^{2\tilde{L}} (v_l^h, \tilde{F}_{BFGS}^{-h}p_j^h)_{L^2(\Omega)} u_l^h]]
\end{aligned}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

Proof. In order to apply Lemma 4.4.7, we need an error estimate $e^{F^{-1}_{\mathrm{BFGS}}}_{x^h,y^h} \ge 0$ which fulfills Assumption 4.4.6, i.e.,

$$|(x^h, (\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h})y^h)_{L^2(\Omega)}| \leq e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}} \qquad \text{for all } x^h, y^h \in V^h.$$

We use Theorem 5.5.3 to construct such an error estimate. Indeed, in (5.5.1) and (5.5.2) set $\beta = \tau + \mu$, $L = 2\tilde{L}$, $\check{y}^h = 0$, $\hat{y}^h = y^h$ and u_l^h, v_l^h, $1 \leq l \leq 2\tilde{L}$, as the BFGS vectors from Section 4.4.1. Then $z = \tilde{F}_{\mathrm{BFGS}}^{-1} y^h$ and $z^h = \tilde{F}_{\mathrm{BFGS}}^{-h} y^h$ (cf. (4.4.11) and (4.4.14)). This leads to the adjoint problem

$$\begin{aligned} \text{Find } \Phi'_x \in H_0^1(\Omega): \quad & \alpha(w, \Phi'_x)_{H_0^1(\Omega)} + (\tau+\mu)(w, \Phi'_x)_{L^2(\Omega)} + \sum_{l=1}^{2\tilde{L}} (u_l^h, \Phi'_x)_{H_0^1(\Omega)} (v_l^h, w)_{L^2(\Omega)} \\ &= (x, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega) \end{aligned}$$

where $x \in L^2(\Omega)$ is arbitrary. Using the definition of $\tilde{F}_{\mathrm{BFGS}}$, we find that this is equivalent to

$$\text{Find } \Phi'_x \in H_0^1(\Omega): \quad (\tilde{F}_{\mathrm{BFGS}}\, w, \Phi'_x)_{H_0^1(\Omega)} = (x, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega).$$

Choosing $w = \tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h$ yields $(u_l^h, \Phi'_x)_{H_0^1(\Omega)} = (x, \tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h)_{L^2(\Omega)}$ for $1 \leq l \leq 2\tilde{L}$. Therefore, the adjoint solution Φ'_x also solves

$$\text{Find } \Phi'_x \in H_0^1(\Omega): \quad \alpha(w, \Phi'_x)_{H_0^1(\Omega)} + (\tau+\mu)(w, \Phi'_x)_{L^2(\Omega)} = (c, w)_{L^2(\Omega)} \qquad \text{for all } w \in H_0^1(\Omega)$$

where $c := x - \sum_{l=1}^{2\tilde{L}} (x, \tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h)_{L^2(\Omega)} v_l^h$. Let $\varepsilon \in (0, \frac{\pi}{\omega})$ be arbitrary and define $p \in \mathbb{R}$ via $2 - 2/p = \frac{\pi}{\omega} - \varepsilon$. Since $\varepsilon \in (0, \frac{\pi}{\omega})$ and $\omega \in (\pi, 2\pi)$, it holds $1 < p < 2/(2 - \frac{\pi}{\omega}) < 2$. Thus, Lemma 5.4.8 implies that the solution $\Phi'_x \in H_0^1(\Omega)$ is an element of $W^{2,p}(\Omega)$ and there exists a constant $C_{\varepsilon,\alpha,\tau+\mu,\Omega}$ (depending only on ε, α, $\tau+\mu$ and Ω) such that $\|\Phi'_x\|_{W^{2,p}(\Omega)} \leq C_{\varepsilon,\alpha,\tau+\mu,\Omega} \|c\|_{L^p(\Omega)}$. Since $p < 2$ it holds that $q := 1 + 1/(p-1) > 2 > p$ and thus the Sobolev embedding [1, 4.12] implies that $W^{1,2}(\Omega)$ is continuously embedded into $L^q(\Omega)$. Therefore, $\tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h \in H_0^1(\Omega) \subset L^q(\Omega)$ and Hölder's inequality yields $|(x, \tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h)_{L^2(\Omega)}| \leq \|\tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h\|_{L^q(\Omega)} \|x\|_{L^p(\Omega)}$. This results in

$$\|\Phi'_x\|_{W^{2,p}(\Omega)} \leq C_{\varepsilon,\alpha,\tau+\mu,\Omega} \|c\|_{L^p(\Omega)} \leq C_{\varepsilon,\alpha,\tau+\mu,\Omega} (1 + \sum_{l=1}^{2\tilde{L}} \|\tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h\|_{L^q(\Omega)} \|v_l^h\|_{L^p(\Omega)}) \|x\|_{L^p(\Omega)}.$$

Therefore, Assumption 5.5.2 is fulfilled with constants p and

$$\check{C}_{\mathrm{reg}} := C_{\varepsilon,\alpha,\tau+\mu,\Omega} (1 + \sum_{l=1}^{2\tilde{L}} \|\tilde{F}_{\mathrm{BFGS}}^{-1} u_l^h\|_{L^q(\Omega)} \|v_l^h\|_{L^p(\Omega)}).$$

Thus, Theorem 5.5.3 is applicable and there exists a constant $C_{\sigma,\Omega}$ such that, for arbitrary $x^h, y^h \in V^h$,

$$|(x^h, (\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h})y^h)_{L^2(\Omega)}| \leq \check{C}_{\mathrm{reg}} C_{\sigma,\Omega} \left(\sum_{T \in \mathscr{T}^h} \eta_T^q \right)^{1/q} \|x^h\|_{L^2(\Omega)}.$$

Here, for all $T \in \mathscr{T}^h$, the local error indicators η_T are defined via

$$\eta_T := (\tau+\mu) h_T^{1+\pi/\omega-\varepsilon} \|\tilde{F}_{\mathrm{BFGS}}^{-h} y^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \| [[y^h - \alpha \tilde{F}_{\mathrm{BFGS}}^{-h} y^h - \sum_{l=1}^{2\tilde{L}} (v_l^h, \tilde{F}_{\mathrm{BFGS}}^{-h} y^h)_{L^2(\Omega)} u_l^h]] \|_{L^2(\partial T)}.$$

Since $p < 2/(2-\frac{\pi}{\omega})$, it holds $q = 1 + 1/(p-1) > 2\omega/\pi$. Thus, for all $x^h, y^h \in V^h$, Lemma 5.5.9 yields

$$|(x^h, (\tilde{F}_{\mathrm{BFGS}}^{-1} - \tilde{F}_{\mathrm{BFGS}}^{-h}) y^h)_{L^2(\Omega)}| \le \check{C}_{\mathrm{reg}} C_{\sigma,\Omega} \left(\sum_{T \in \mathscr{T}^h} \eta_T^{2\omega/\pi} \right)^{\pi/(2\omega)} \|x^h\|_{L^2(\Omega)} =: e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}},$$

i.e., $e_{x^h,y^h}^{F_{\mathrm{BFGS}}^{-1}}$ fulfills Assumption 4.4.6. Furthermore, by Theorem 5.5.10, there exists a constant $C_{\varepsilon,\Omega,\sigma} > 0$ such that for $x^h, y^h \in Y^{*h}$ the error estimate

$$e_{x^h,y^h}^{E} := C_{\varepsilon,\sigma,\Omega} \left(\sum_{T \in \mathscr{T}^h} \left(h_T^{1+\pi/\omega-\varepsilon} \|x^h\|_{L^2(T)} + h_T^{1/2+\pi/\omega-\varepsilon} \| [[\tilde{E}^h x^h]] \|_{L^2(\partial T)} \right)^{2\omega/\pi} \right)^{\pi/(2\omega)} \|y^h\|_{L^2(\Omega)}$$

fulfills Assumption 4.3.6. As Assumptions 4.3.6 and 4.4.6 are fulfilled, Lemma 4.4.7 implies that

$$p_j^{F_{\mathrm{BFGS}}^{-1}} := \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h, \qquad e_{i,j,F_{\mathrm{BFGS}}^{-1}} := e_{p_i^h,p_j^h}^{F_{\mathrm{BFGS}}^{-1}}, \qquad e_{F_{\mathrm{BFGS}}} := (\tau+\mu) e_{\tilde{d}^{*h},\tilde{d}^{*h}}^{E},$$

fulfill Assumption 4.2.4. From $\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) \to 0$ and $\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) \to 0$ as $h \to 0$, we infer that $O_{F^{-1}}$, C_F, $C_{F^{-1}}$, $\hat{\varepsilon}_{F^{-1}}$ and $\hat{\varepsilon}_F$, as defined above, fulfill Assumption 4.2.7. □

Example 5.5.15. Suppose that $\Omega = (-a,a)^2 \setminus [0,a)^2$, $a > 0$ is a L-shaped domain and let $(\mathscr{T}^h)_h$, $0 < h \le 1$, be a regular family of triangulations of Ω with corresponding finite element space V^h. Then Ω is a polygonal domain with one reentrant corner of angle $\omega = 3/2\pi$. Let $\varepsilon \in (0, \frac{2}{3})$ and $x^h, y^h \in V^h$ be arbitrary and suppose that Q is given as the BFGS curvature operator (cf. Section 4.4). Then there exists a constant $C > 0$, independent of h, such that

$$\begin{aligned}
O_{F^{-1}}(p_j', h) &:= R_X^{-1} \iota^* \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h, \qquad C_F := C, \quad C_{F^{-1}} := C, \\
\hat{\varepsilon}_F(\tilde{d}^*, \mathring{d}', h) &:= (\tau+\mu) \left(\sum_{T \in \mathscr{T}^h} \eta_T^3 \right)^{1/3} \|\tilde{d}^{*h}\|_{L^2(\Omega)} \\
&\quad \text{with } \eta_T := h_T^{5/3-\varepsilon} \|\tilde{d}^{*h}\|_{L^2(T)} + h_T^{7/6-\varepsilon} \| [[\tilde{E}^h \tilde{d}^{*h}]] \|_{L^2(\partial T)} \qquad \text{for all } T \in \mathscr{T}^h \\
\hat{\varepsilon}_{F^{-1}}(p_i, p_j, h) &:= \left(\sum_{T \in \mathscr{T}^h} \left((\tau+\mu) h_T^{5/3-\varepsilon} \|r^h\|_{L^2(T)} + h_T^{7/6-\varepsilon} \|R\|_{L^2(\partial T)} \right)^3 \right)^{1/3} \|p_i^h\|_{L^2(\Omega)} \\
&\quad \text{with } r^h := \tilde{F}_{\mathrm{BFGS}}^{-h} y^h \text{ and } R := [[p_j^h - \alpha \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h - \sum_{l=1}^{2\tilde{L}} (v_l^h, \tilde{F}_{\mathrm{BFGS}}^{-h} p_j^h)_{L^2(\Omega)} u_l^h]]
\end{aligned}$$

fulfill Assumption 4.2.7 and thus can be used in Algorithm 4.2 to compute a trial iterate.

6. Numerical Results

6.1. Implementation

We implement Algorithm 3.4 to search for an approximately stationary point of problem (5.3.1). The programming language MATLAB is used for all implementations.

Solution strategies

We use four distinct strategies to solve the problem (5.3.1).

- **Strategy A: Uniform refinement, no BFGS** This strategy uses a family of triangulations $(\mathscr{T}^h)_h$ obtained by uniformly refining the triangulation in each step. In particular, starting from the initial mesh $\mathscr{T}^{h_0}$, whenever an error estimate needs to be refined, the next mesh is obtained by bisection of every triangle. The oracles defined in Section 5.3 and the a priori error estimates of Section 5.4 are used for the computation. Furthermore, no additional curvature information is used to construct the curvature operator Q, cf. Section 4.3.1.

- **Strategy B: Uniform refinement, with BFGS** This strategy also uses uniform triangulation, but here, the L-BFGS curvature strategy (cf. Section 4.4) is used to construct the curvature operator. In order to include only useful points and subgradients into the BFGS curvature operator, we start with $Q = 0$ and do not add any points and subgradients until the first mesh refinement occurs. From then on, we use the trial iterates with corresponding subgradients for the BFGS curvature operator, cf. (4.4.2). The maximum number of points and subgradients is chosen to be $L = 10$.

- **Strategy C: Adaptive refinement, no BFGS** Strategy C uses the adaptive refinements described in Section 5.5 to construct the next triangulation. The oracles defined in Section 5.3 and the a posteriori error estimates of Section 5.5 are used for the computation.

- **Strategy D: Adaptive refinement, with BFGS** This strategy combines the adaptive mesh refinement of strategy C with the L-BFGS curvature strategy of Strategy B to construct the curvature operator.

Due to constraints on the available memory, the mesh cannot be refined to reach arbitrary small mesh widths. Therefore, a limit of $60,000$ mesh nodes is imposed. Whenever the refined mesh exceeds the maximal number of mesh nodes, no further refinement is allowed. We then set all error estimates used in the computation to zero. This can be seen as solving the discretized version of problem (5.3.1) where the discretization is given via the final mesh. From another point of view, one could argue that the computation has two phases. First, we search for an appropriate discretization of problem (5.3.1) and compute good initial values and starting points. In the second phase the discretized problem is solved.

Parameters of the algorithms

In Algorithm 3.3, we use a maximal number of $n_{it} = 3$ iterations to compute a sufficiently steep subgradient based linearization. In Algorithm 3.4, we used the constants $\gamma = 0.1$, $\tilde{\gamma} = 0.7$, $\Gamma = 0.7$, $\bar{q} = 1000$, $\xi = 1 \cdot 10^{-6}$, $T = 1000$ and the forcing sequences $(\nu_i^{\Psi})_{i\in\mathbb{N}} := (\nu_i^{\text{lin}})_{i\in\mathbb{N}} := 100\tau_i^{-1.1}$, and $(\nu_i^{f})_{i\in\mathbb{N}} := 1000\tau_i^{-0.1}$ which fulfill (3.5.7). The downshift parameter from Section 3.1.4 is set to $c = 10^{-4}$. The start iterate x_0 is chosen as $x_0 = 0$ and the initial proximity parameter τ_0 is set to 1. The initial lift term l_0 differs for each example. Therefore, we report this value for each example.

Additional function value refinement

In the step "Function value refinement" in Algorithm 3.4, one has to choose a new lift term $\hat{l}_i \in (0, \max\{l_i, \nu_i^f \underline{O}_{\|\cdot\|_Y}(\tilde{y}_i - x_i)\}]$. We aim at refining the discretization whenever progress of the computed function values seems to stagnate. Therefore, we reduce the lift term if $O_f(\tilde{y}_i, \hat{h}_i) < O_f(x_i, \hat{h}_i)$ and the relative reduction of the function value $(O_f(x_i, \hat{h}_i) - O_f(\tilde{y}_i, \hat{h}_i))/O_f(\tilde{y}_i, \hat{h}_i)$ is smaller than the threshold $1 \cdot 10^{-4}$. Furthermore, we reduce the lift term if the $L^2(\Omega)$-norm of the gradient of the reduced objective function is below the threshold $1 \cdot 10^{-4}$, i.e., $\|\tilde{g}_i + w'(\tilde{y}_i)\|_{L^2(\Omega)} < 1 \cdot 10^{-4}$. If the lift term is to be reduced, then we use $\hat{l}_i := \max\{l_i/2, \nu_i^f \underline{O}_{\|\cdot\|_Y}(\tilde{y}_i - x_i)\}$, otherwise we reuse $\hat{l}_i := l_i$.

Numerical integration

To evaluate the integrals in (5.3.4) and (5.3.8), we use numerical integration as described below. In our numerical examples, any functions $f \in L^1(\Omega)$ which needs to be integrated is piecewise smooth with possibly jumps between pieces. Given a triangulation $\mathscr{T}^h$ of the domain Ω, we first construct a new triangulation $\mathscr{T}^{h'}$ by refining along any jump of the function $f \in L^1(\Omega)$. We now approximate the integral of f via

$$\int_\Omega f \,\mathrm{d}\lambda = \sum_{T\in\mathscr{T}^{h'}} f \,\mathrm{d}\lambda \approx \sum_{T\in\mathscr{T}^{h'}} \sum_{p\in Q(T)} w_p f(p)$$

where $Q(T) \subset \mathbb{R}^2$ is a discrete set of points with corresponding weights w_p. We use the quadrature formula Q_5 of degree 5 from [29]. There, 7 barycentric coordinates $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}$ with corresponding weights w_i, $1 \le i \le 7$, are given such that $Q_5(T) = \{\alpha_i P_1 + \beta_i P_2 + \gamma_i P_3, 1 \le i \le 7\}$, where P_1, P_2, P_3 are the vertices of triangle T. As $y^h \in V^h$ is linear on each triangle, $y^h(p_i) = y^h(\alpha_i P_1 + \beta_i P_2 + \gamma_i P_3) = \alpha_i y^h(P_1) + \beta_i y^h(P_2) + \gamma_i y^h(P_3)$ can be computed efficiently for any $p_i \in Q_5(T)$. This choice of quadrature points yields the exact value of the integral for any polynomial up to order 5. In our numerical tests, the error introduced by numerical integration is negligible. If this is not the case, other algorithms have to be employed to compute the integral, e.g., the adaptive algorithm presented in [123].

Constants of the error estimates

In order to execute Algorithms 3.4 and 4.2, the oracles O_f and $O_{F^{-1}}$ with corresponding error estimates are needed. The error estimates for $O_{F^{-1}}$ developed in Sections 5.4 and 5.5 involve the constants C_F and $C_{F^{-1}}$. However, due to the convergence theory of Theorem 4.2.8, we do not need to compute a concrete value for these constants. In contrast to this, for the function value oracle O_f an explicit constant for the error estimate has to be determined, both for a priori (cf. Theorem 5.4.1) and a posteriori (cf. Section 5.5.1) error estimates. However, due to the lifting strategy employed in Section 3.5.1, the algorithm is very insensitive to small changes of the constant of the error estimate. Therefore, it suffices to determine an approximate value of this constant. For the a priori error estimate, Theorem 5.4.1, we use

$$C_\Omega := 10 \max \frac{\|S(u+\mathring{f}) - S^h(u+\mathring{f})\|_{H_0^1(\Omega)}}{h\big(\|u+\mathring{f}\|_{L^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\big)},$$

where the maximum is taken over several common values of $u \in L^2(\Omega)$. For the constants in the a posteriori error estimate in Section 5.5.1, we use a similar procedure.

Stopping

In inexact optimization, it is inherently difficult to determine when to stop the algorithm. In order to compare the performance of strategies A-D, we stopped after a prescribed time limit or whenever the algorithm demands a more accurate solution of the bundle subproblem but the refinement of the mesh would lead to a mesh with more than a prescribed number of nodes. To find practical stopping conditions, one can proceed as proposed in [99, Chap. 26.8].

Problem setting

We implemented several examples. All examples have the structure of problem (5.3.1), i.e.,

$$\begin{aligned} &\underset{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)}{\text{minimize}} && \tfrac{1}{2}\|y-y_d\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \\ &\text{subject to} && y \in K, \quad \int_\Omega \nabla y^T \nabla(v-y)\,\mathrm{d}\lambda \geq \int_D (u+\mathring{f})(v-y)\,\mathrm{d}\lambda \quad \forall v \in K. \end{aligned} \tag{5.3.1}$$

Therefore, the data Ω, α, ψ, y_d and $\mathring{f}$ fully determine the problem. All computations were done on an Intel Core i5-7200U laptop with 2 cores and 8 GB of RAM. The color scheme used to plot the figures originates from [113].

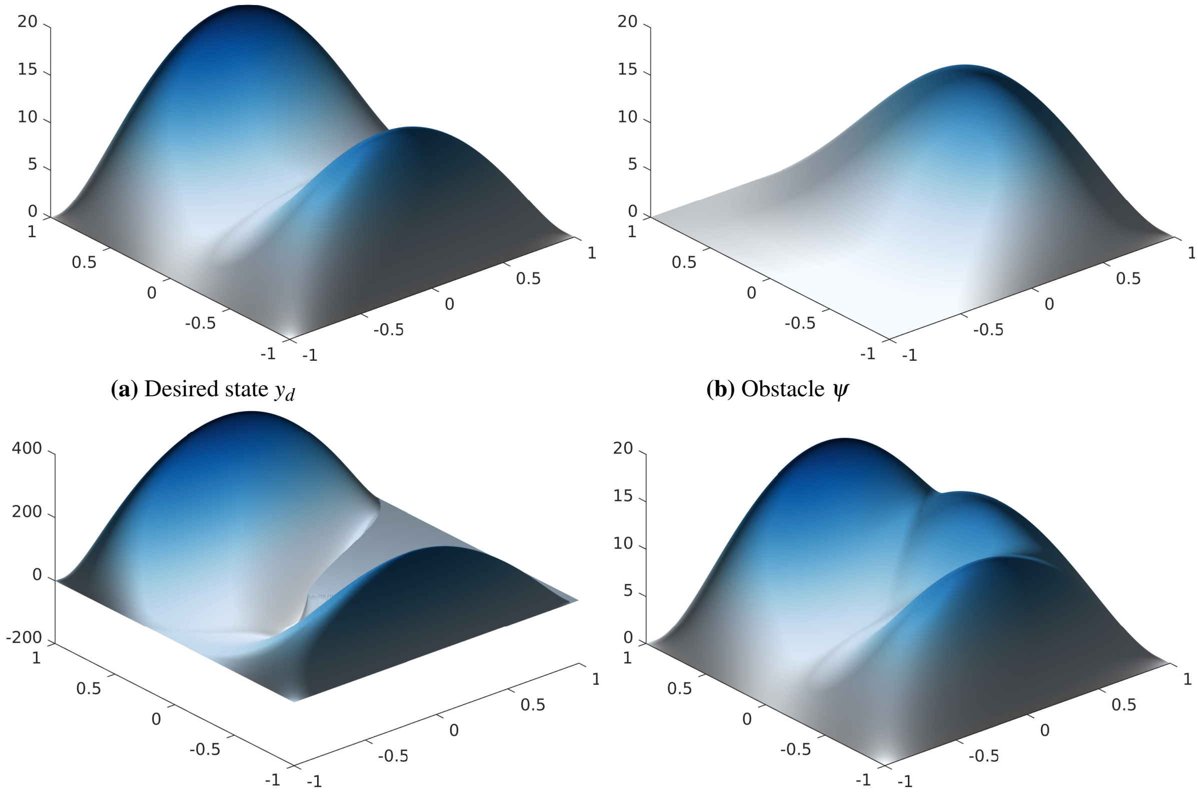
6.2. Example 1

This example was first presented in [50]. The main difference between the solution methods presented here compared to the solution methods of [50] is the lifting strategy for the function value approximations, cf. Section 5.3.2. The domain is given by $\Omega := (-1,1)^2$, the force $\mathring{f}$ is set to zero and $\alpha := 10^{-4}$. The desired state

$$y_d(x_1,x_2) := (1-x_1^2)(x_2+5)(1-11x_2^4+10x_2^2)$$

and the obstacle

$$\psi(x_1,x_2) := 8(x_1-1)(x_1+1)^3(x_2-1)(x_2+1)$$



(a) Desired state y_d

(b) Obstacle ψ

(c) Optimal control u^*

(d) Optimal state y^*

Figure 6.1.: Problem data for example 1.

are depicted in Figures 6.1(a) and 6.1(b), respectively. We start by comparing solution strategies A and B using the subgradient oracle O_g^1. We use an initial triangulation with maximal triangle side length of $h_0 = \sqrt{2}/8$ and 289 nodes. The initial lift term is set to $l_0 := 25$ and the constant for the a priori error estimate for the function value (cf. Section 6.1) is set to $C_\Omega = 0.025$. The final mesh has a maximal triangle side length of $h_4 = \sqrt{2}/128$ and 66049 nodes. We stop the algorithm after a time of 400 seconds has passed.

There is no analytical solution available. In order to be able to asses the performance of strategies A and B, we compute a numerical solution on a high fidelity grid and use this as the optimal solution. In particular, we run the uniform strategy with BFGS for 800 seconds where the maximum number of nodes is restricted such that the final mesh has 263169 nodes with a maximal triangle side length of $h = \sqrt{2}/256$. The resulting optimal control u^* and optimal state y^* are shown in Figures 6.1(c) and 6.1(d), respectively. The active set coincides with the region where the control is zero. On the inactive set, the pointwise distance between the optimal state and the desired state is small.

Figure 6.2 gives an overview over the process of the bundle algorithm. The left column, i.e., figures (a), (c), (e) and (g) correspondent to strategy A and the right column, i.e., figures (b), (d), (f) and (g) correspondent to strategy B. Iterates marked with a red, empty box are serious iterates and the black, filled boxes at the x-axes indicate mesh refinement. Figure 6.2(a) and Figure 6.2(b) depict the error

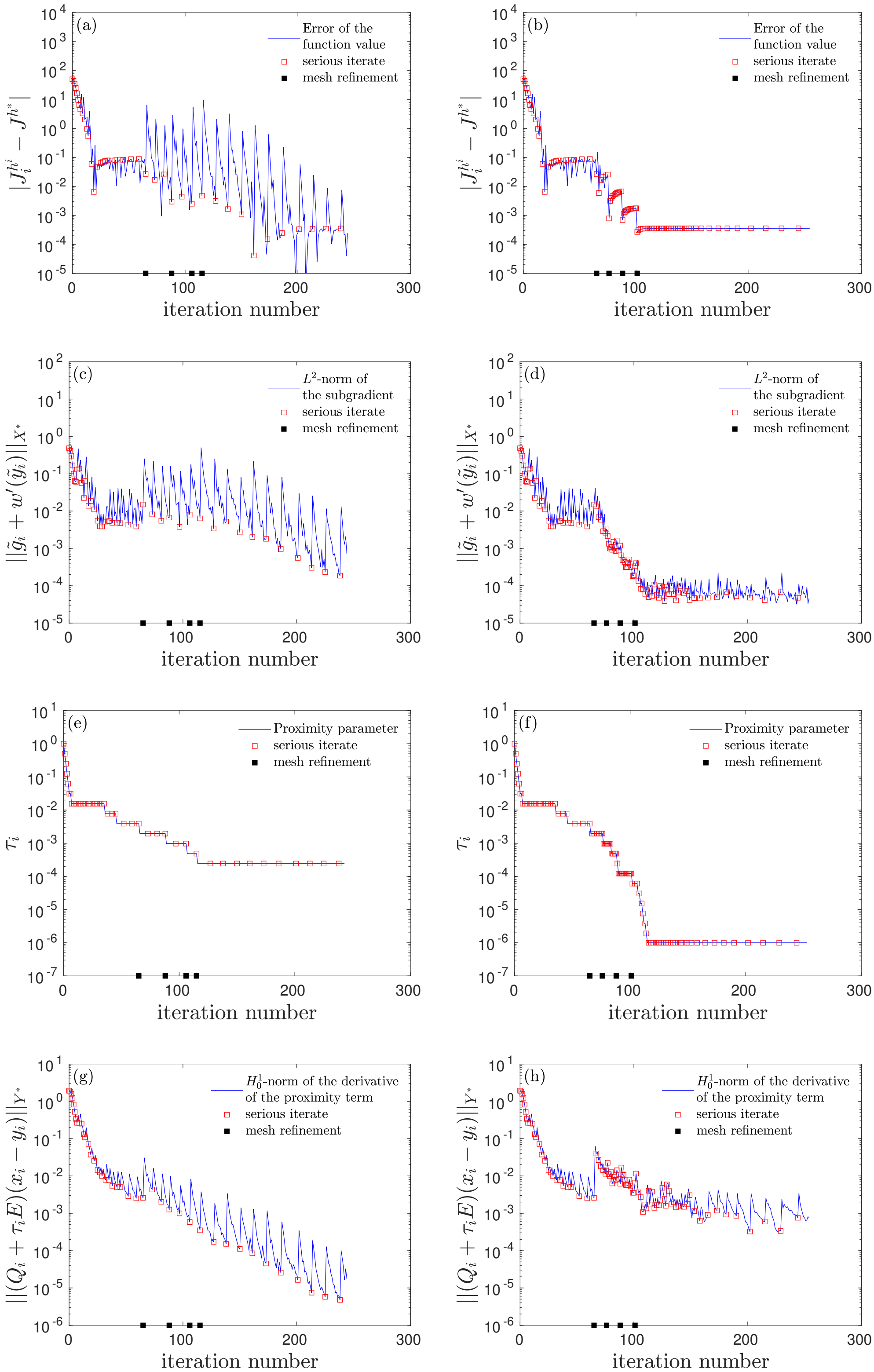


Figure 6.2.: Convergence statistics of Algorithm 3.4 using uniform mesh refinement. No curvature strategy (a),(c),(e),(f). BFGS curvature (b), (d), (f), (h).

between the computed function value $J_i^{h_i}$ at trial iterate u_i and the computed optimal function value J^{h^*}, i.e.,

$$J_i^{h_i} := \tfrac{1}{2}\|S^{\hat{h}_i}(u_i + \mathring{f}) - y_d\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u_i\|^2_{L^2(\Omega)}, \qquad J^{h^*} := \tfrac{1}{2}\|S^{h^*}(u^* + \mathring{f}) - y_d\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u^*\|^2_{L^2(\Omega)}.$$

Note that the computed function values $J_i^{h_i}$ are the function values which were used by the algorithm and thus incorporate the increasing accuracy of the solution operator $S^{\hat{h}_i}$ as the discretization is refined. The computed optimal value J^{h^*} is evaluated by computing the state $y^{h^*} := S^{h^*}(u^* + \mathring{f})$ on the same (high fidelity) mesh as the optimal control u^*. Figure 6.2(c) and Figure 6.2(d) show the $L^2(\Omega)$-norm of the subgradient $\tilde{g}_i + w'(u_i)$ and in Figure 6.2(e) and Figure 6.2(f) the proximity parameter τ_i is depicted. Figure 6.2(g) and Figure 6.2(h) display the $H_0^1(\Omega)$-norm of the derivative of the proximity term $\hat{e}_i = (Q_i + \tau_i R_Y)(x_i - y_i)$. The proximity parameter τ_i and the term $\|\hat{e}_i\|_{H_0^1(\Omega)}$ play an important role in the convergence result Theorem 3.3.3, cf. (3.3.3).

We observe that for the first 66 iterations, both strategies yield the same results. This is due to the fact that strategy B starts to include points and subgradients into the BFGS curvature operator only after the first mesh refinement occurs. However, since the initial computations are carried out on a very coarse grid it takes only 4.0 seconds to execute the first 66 iterations. From this point on, both strategies differ considerably. In particular, the error of the function value for strategy A is jagged. As one can see, after every serious iterate the error of the function value of the trial iterate is large at first and then it is reduced until a new serious iterate is selected. This might be due to the fact that after every serious iterate, Algorithm 3.4 with no curvature strategy discards all cutting planes from the old model and needs to build a completely new model. Compared to strategy A, the graph of the function value error for strategy B is smoother. This shows a stabilizing influence of the BFGS curvature operator, which reuses old subgradient information. From iteration 110 onward, the proximity parameter of the strategy with BFGS curvature has the lowest possible value $\tau_i = \xi = 1 \cdot 10^{-6}$. This shows that the model Ψ_i excellently captures the behavior of the objective function $f + w$. Figure 6.2(c) and Figure 6.2(d) indicate that the $L^2(\Omega)$-norm of the subgradient at the computed solution is of the order of 10^{-4}. Figure 6.2(g) and Figure 6.2(h) indicate that $\|\hat{e}_i\|_{H_0^1(\Omega)}$ goes to $\bar{\varepsilon}$ with $\bar{\varepsilon} < 1 \cdot 10^{-5}$ for the curvature strategy with $Q = 0$ and $\bar{\varepsilon} < 1 \cdot 10^{-3}$ for the BFGS curvature strategy. Thus, Theorem 3.3.3 suggests that the bundle method converges to a $\bar{\varepsilon}$-G-stationary point. However, Figure 6.2(e) and Figure 6.2(f) indicate that the proximity parameter does not go to infinity. This suggests that the hard case, i.e., $\mathscr{E}_{\bar{x}} \neq \emptyset$, does not occur. In this case, Theorem 3.3.3 implies that the limit point is G-stationary.

In order to compare the performance of strategies A and B, we present three distinct benchmarks. Because the objective is to minimize the function value, we aim at comparing the true function value $f(u_i) + w(u_i)$ at the trial iterate u_i to the true optimal function value. Since computing $f(u_i) = \frac{1}{2}\|S(u_i + \mathring{f}) - y_d\|^2_{L^2(\Omega)}$ involves the evaluation of the solution operator S of the variational inequality, we cannot do this. Instead, we compute a high fidelity approximation of the state via

$$y_i^{h^*} := S^{h^*}(u_i + \mathring{f}). \tag{6.2.1}$$

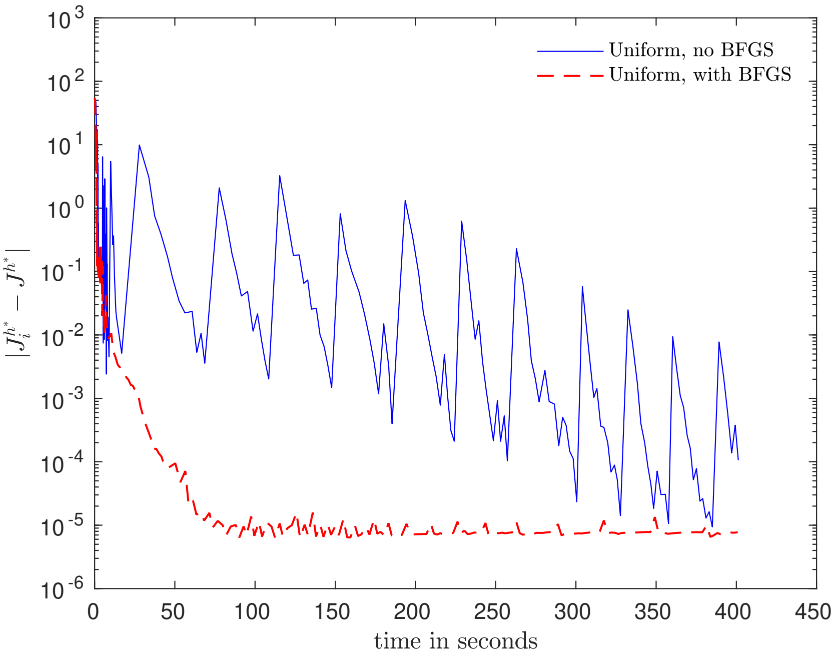


Figure 6.3.: Difference of the function values to the computed optimal value over time.

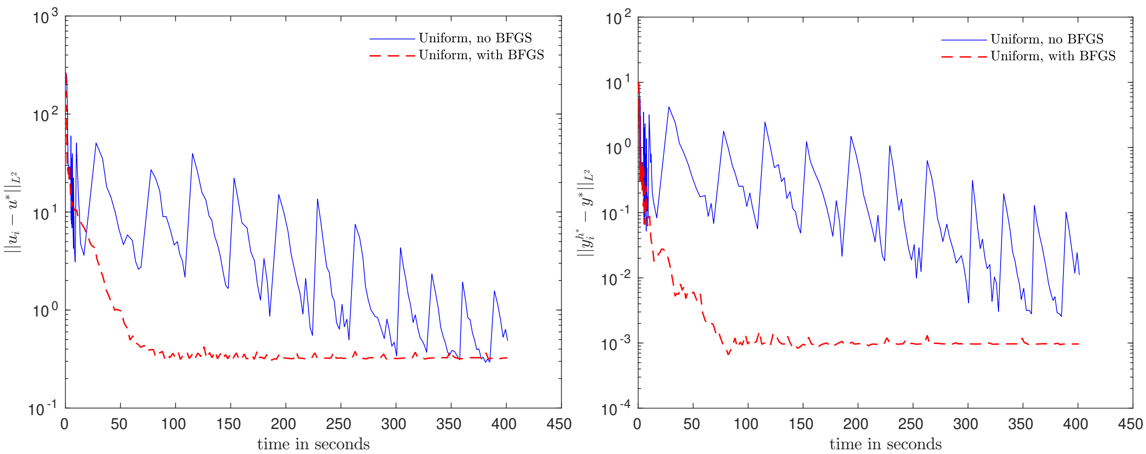


Figure 6.4.: Distances of the control and the state to the computed optimal control and state over time.

Similarly, we use the high fidelity approximation of the function value $f(u_i)+w(u_i)$ for control u_i

$$J_i^{h^*} := \tfrac{1}{2}\|S^{h^*}(u_i+\mathring{f}) - y_d\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|u_i\|^2_{L^2(\Omega)}, \tag{6.2.2}$$

i.e., the state corresponding to each trial iterate is computed on the same mesh as the optimal state. The results are depicted in Figure 6.3. Furthermore, the $L^2(\Omega)$-error of the control, $\|u_i - u^*\|_{L^2(\Omega)}$, and the computed $H_0^1(\Omega)$-error of the state,

$$\|y_i^{h^*} - y^{h^*}\|_{H_0^1(\Omega)} = \|S^{h^*}(u_i+\mathring{f}) - S^{h^*}(u^*+\mathring{f})\|_{H_0^1(\Omega)},$$

are depicted in Figure 6.4. In contrast to the function value and the error of the state, the error of the control can be computed exactly.

We observe that both methods converge to the solution on the final mesh. However, the method with BFGS curvature rapidly converges within 80 seconds whereas the strategy without curvature information only makes slow progress.

The subgradient oracle $\mathbf{O_g^2}$

We now use the subgradient oracle O_g^2 defined in (5.3.25). The advantage of O_g^2 over O_g^1 is that, under reasonable assumptions, it can be guaranteed that for given $x^h \in X^h$ the approximate subgradients $O_g^2(x^h, h_n)$ converge in $H_0^1(\Omega)$ to a true subgradient of the reduced objective function as $n \to \infty$, cf. Section 5.3.4. Recall that the sole difference in the computation of O_g^1 and O_g^2 is the choice of the discrete inactive set. For O_g^1, the discrete inactive set is given via $D^h := \{y^h > I^h \psi\}$ whereas the discrete inactive set for O_g^2 is given via $D_n := \{y_n > \psi_n + \varepsilon_n\}$ where $\varepsilon_n > 0$ is determined via the $L^\infty(\Omega)$-errors of the state and the obstacle discretization, cf. (5.3.23). We now want to determine how this additional term ε_n in the choice of the discrete inactive set influences the performance of the bundle method.

We perform exactly the same numerical test as in the beginning of this chapter with the sole difference that the subgradient oracle O_g^1 is replaced with the subgradient oracle O_g^2. In Figure 6.5, the same data as in Figure 6.2 is plotted. Since the computation of the $L^\infty(\Omega)$-error estimators takes up additional time, only 200 and 100 iterations can be executed within the 400 seconds computation period for strategies A and B, respectively. When comparing the strategy without additional curvature information to the strategy with BFGS curvature, we observe the same behavior as with the first subgradient oracle O_g^1. However, whereas the error of the function value with the first subgradient oracle is of order 10^{-4}, here, strategies A and B produce an errors of order 10^{-2} and 10^{-3}, respectively. Figures 6.6(a) and 6.6(b) depict the control and state error analogous to Figure 6.4. Figure 6.6(c) shows the function value error as in Figure 6.3. Figure 6.6(d) shows the subgradient error oracle e_n^g defined in (5.3.24). Again, the smoothing behavior of the BFGS curvature strategy can be observed. Since the subgradient oracle O_g^1 appears to show better overall performance compared to the oracle O_g^2, all further numerical tests are performed using O_g^1.

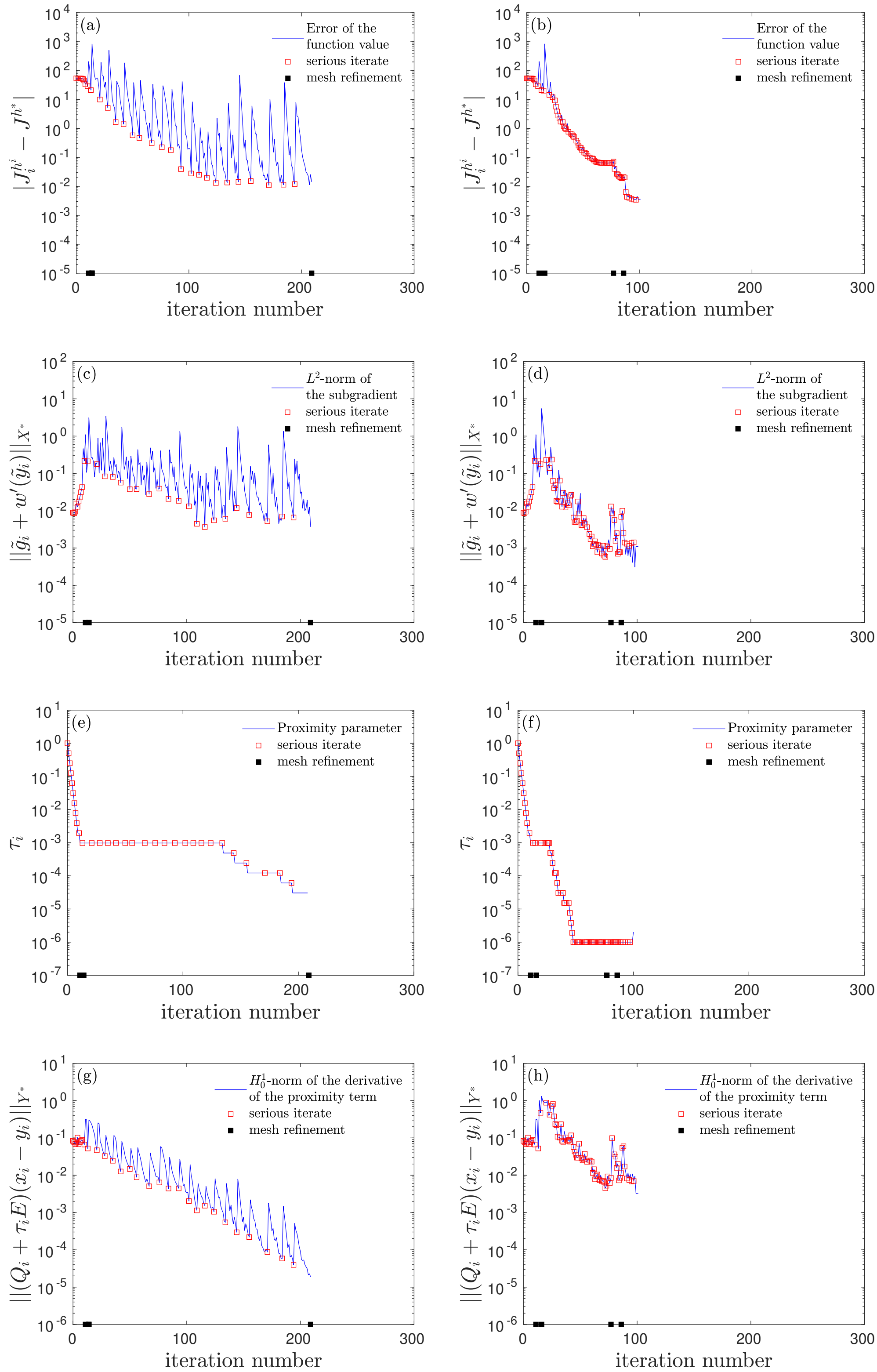
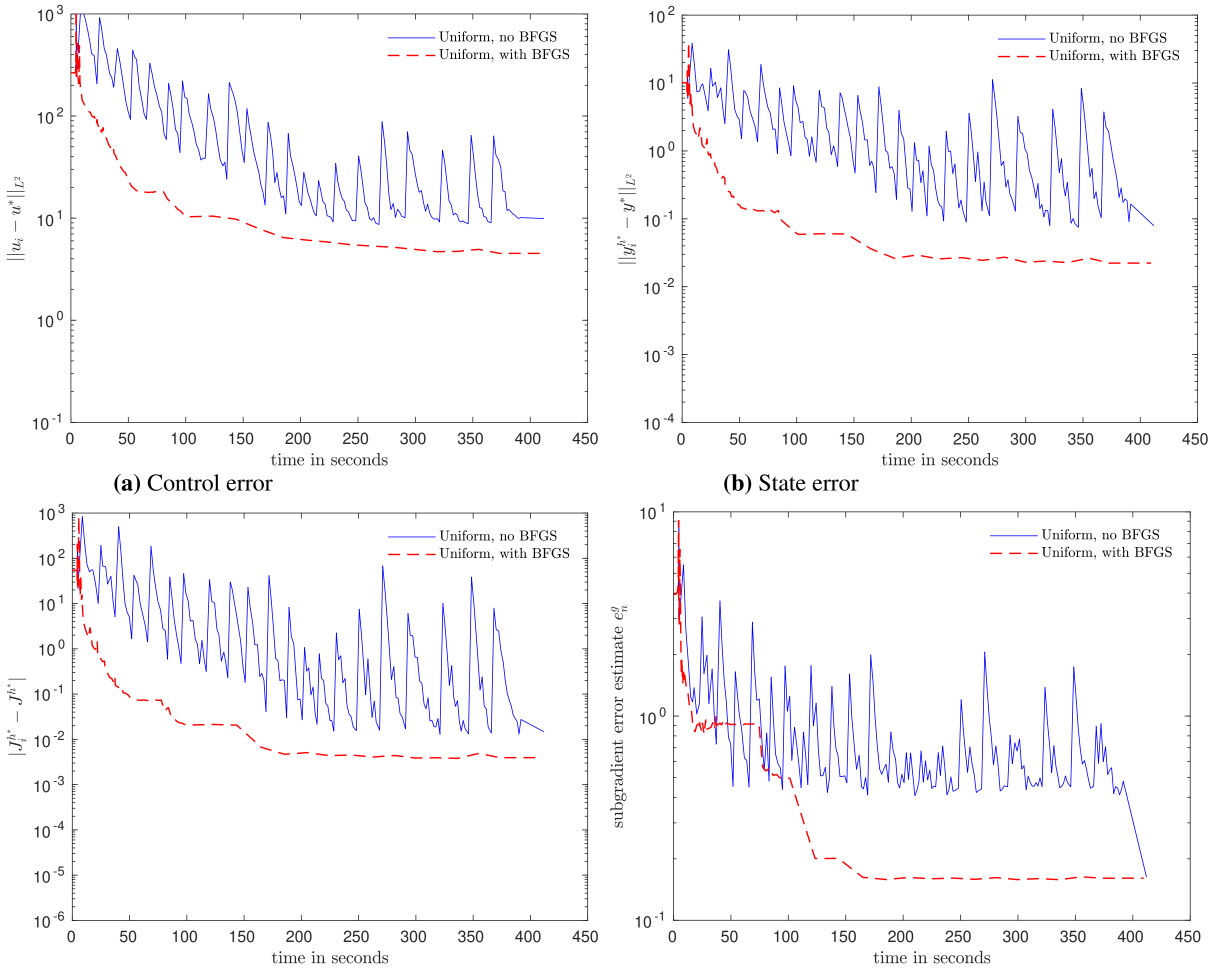


Figure 6.5.: Convergence statistics of Algorithm 3.4 using uniform mesh refinement. No curvature strategy (a),(c),(e),(f). BFGS curvature (b), (d), (f), (h).



(a) Control error

(b) State error

(c) Function value error

(d) Subgradient error estimate

Figure 6.6.: Performance of the bundle method using the subgradient oracle O_g^2 over time.

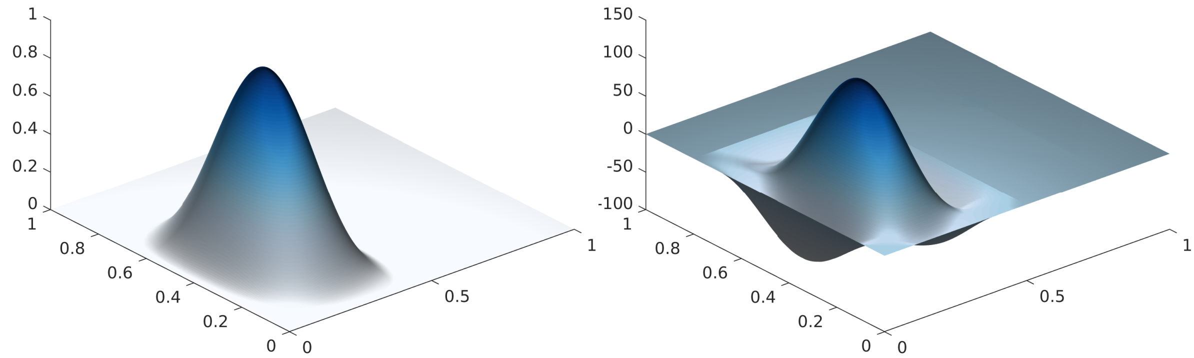
6.3. Example 2

This example was first presented in the paper [55, Ex. 6.1] of M. Hintermüller and I. Kopacka. It is constructed in such a way that the optimal solution is known and that strict complementarity is violated. Here, $\Omega = (0,1)^2$, $\psi \equiv 0$ and $\alpha = 1$. Further data is given via

$$\mathring{f} := -\Delta y^* - u^* - \xi^*,$$
$$y_d := y^* + \xi^* - \alpha u^*,$$

where $u^* := y^*$,

$$y^* : \mathbb{R}^2 \to \mathbb{R}, \qquad y^*(\omega) := z_1(\omega_1) z_2(\omega_2)\, \delta_{(0,0.5)\times(0,0.8)}(\omega),$$
$$\xi^* : \mathbb{R}^2 \to \mathbb{R}, \qquad \xi^*(\omega) := 2\max(0, -|\omega_1 - 0.8| - |(\omega_2 - 0.2)x_1 - 0.3| + 0.35),$$



(a) Optimal control/Optimal state **(b)** Desired state

Figure 6.7.: Problem data of example 2.

and

$$z_1 : \mathbb{R} \to \mathbb{R}, \qquad z_1(\omega_1) := -4096\omega_1^6 + 6144\omega_1^5 - 3072\omega_1^4 + 512\omega_1^3, \tag{6.3.1}$$

$$z_1 : \mathbb{R} \to \mathbb{R}, \qquad z_2(\omega_2) := -244.140625\omega_2^6 + 585.9375\omega_2^5 - 468.75\omega_2^4 + 125\omega_2^3. \tag{6.3.2}$$

The optimal control u^* and the optimal state $y^* = u^*$ are depicted in Figure 6.7(a). The desired state y_d is depicted in Figure 6.7(b). We run the bundle method using strategies A and B and the subgradient oracle O_g^1. We use an initial triangulation with maximal triangle side length of $h_0 = \sqrt{2}/8$ and 81 nodes. The initial lift term is set to $l_0 := 2$ and the constant for the a priori error estimate for the function value (cf. Section 6.1) is set to $C_\Omega = 0.06$. The maximal time is set to 50 seconds and the final mesh has a maximal triangle side length of $h_4 = \sqrt{2}/256$ and 66049 nodes.

Figure 6.8 compares strategies A and B and depicts the same data as in Figure 6.2. Now the course of the bundle method with the no curvature strategy (strategy A) is described. In iteration 0, no refinement of the mesh is necessary. In iterations 1 to 5, the criterion $\|\tilde{g}_1 + w'(u_i[1])\|_{L^2(\Omega)} < 1 \cdot 10^{-4}$ is active and consequently the lift term is halved in each iteration cf. Section 6.1. This leads to a uniform refinement of the mesh in each iteration 1-5. In iteration 5, the mesh cannot be refined anymore and all error estimators return zero error. In iteration 6, Algorithm 4.1 cannot find a new trial iterate with lower model value as the current trial iterate and thus the computation terminates. Strategy B agrees with strategy A for iterations 0 and 1. In iteration 2, the norm of the subgradient is slightly above the threshold of $1 \cdot 10^{-4}$ and therefore the lift term is not reduced and the mesh is not refined. However, this leads to little progress of the algorithm and the distance of the new trial iterate to the serious iterate is small. Therefore, the forcing sequence for the lift term (line 4 in Algorithm 3.4) enforces two mesh refinements in iteration 3. The remaining iterations occur as in strategy A.

Similar to Figures 6.3 and 6.4 we want to compare the performance of strategy A and B for example 2. As before, we cannot exactly compute the state for a given control. Therefore, we compute the states on a fine mesh with 263169 nodes and a mesh width of $h = \sqrt{2}/512$. For this mesh, the $L^2(\Omega)$-distance of the computed optimal state $y^{h^*} := S^{h^*}(u^* + \mathring{f})$ to the true optimal state y^* is $2.52 \cdot 10^{-5}$. The distance of the computed optimal function value $J^{h^*} = \frac{1}{2}\|S^{h^*}(u^* + \mathring{f}) - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u^*\|^2_{L^2(\Omega)}$ to the true optimal function value amounts to $5.16 \cdot 10^{-4}$. In Figure 6.9, the distance of the control u_i to

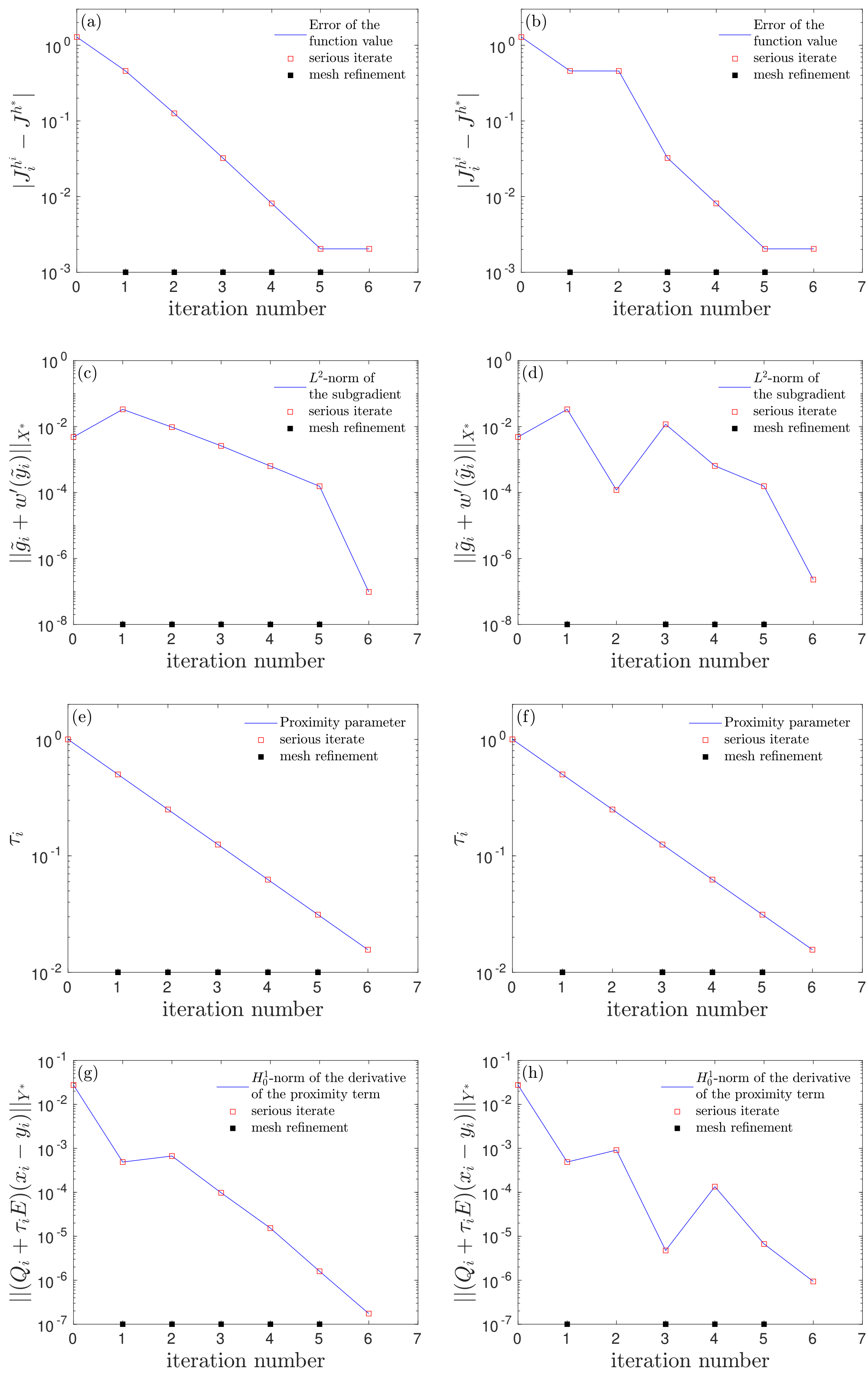


Figure 6.8.: Convergence statistics of Algorithm 3.4 using uniform mesh refinement. No curvature strategy (a),(c),(e),(f). BFGS curvature (b), (d), (f), (h).

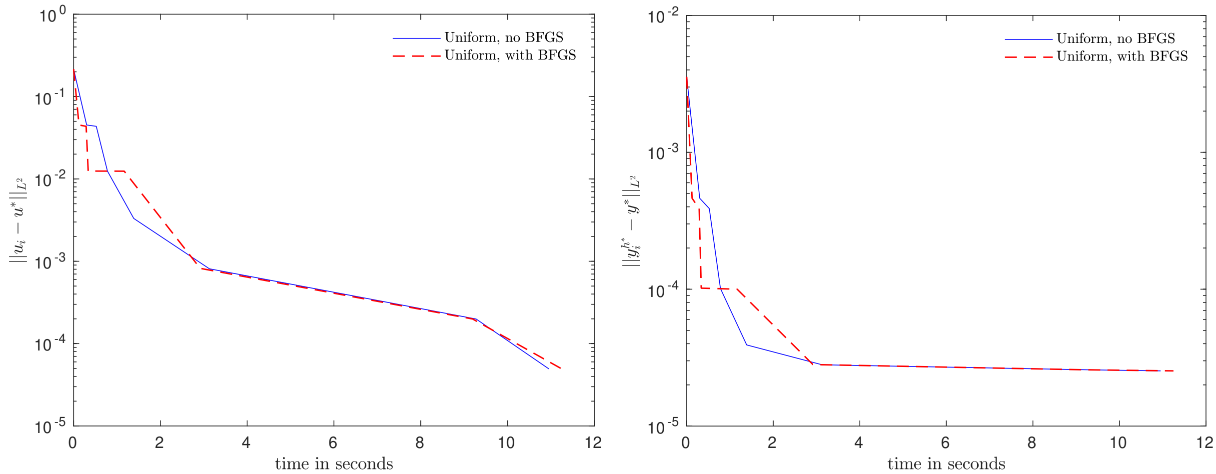


Figure 6.9.: Difference of the controls and states to the optimal values over time.

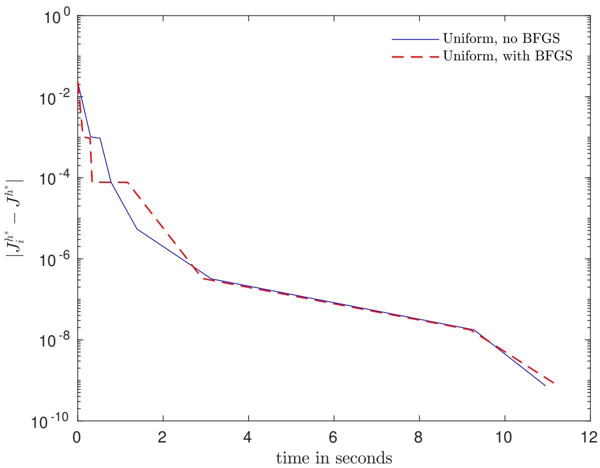


Figure 6.10.: Function value error over time.

the optimal control u^* and the distance of the state $y_i^{h^*}$ to the computed optimal state y^{h^*} is plotted over time. In Figure 6.10, the distance of the function value $J_i^{h^*}$ to the computed optimal function value J^{h^*} is presented. Both strategies A and B find a good approximation of the solution on each mesh within one iteration. Therefore, strategy A and B produce very similar results.

6.4. Example 3

This example is taken from the paper [88, Ex. 7.1] of C. Meyer and O. Thoma and was also published in the OPTPDE problem set [101, Prb. mpccdist1]. There, the problem structure

$$\begin{aligned} &\underset{(y,\bar{u})\in H_0^1(\Omega)\times L^2(\Omega)}{\text{minimize}} && \tfrac{1}{2}\|y-y_d\|_{L^2(\Omega)}^2 + \tfrac{\alpha}{2}\|\bar{u}-u_d\|_{L^2(\Omega)}^2 \\ &\text{subject to} && y \in K_\psi,\ \int_\Omega \nabla y^T \nabla(v-y)\,\mathrm{d}\lambda \geq \int_\Omega \bar{u}(v-y)\,\mathrm{d}\lambda\ \forall v \in K_\psi \end{aligned} \tag{6.4.1}$$

is used. To fit (6.4.1) in the setting of Problem (5.2.1), we use the transformation $u = \bar{u} - u_d$ and thus set the data term to $\mathring{f} := u_d$. Alternatively, one could use $w(u) := \frac{\alpha}{2}\|\bar{u} - u_d\|^2_{L^2(\Omega)}$ in the bundle method. However, if $u_d \in L^2(\Omega)$ is not contained in any of the approximation spaces V^h, the term $\|\bar{u} - u_d\|_{L^2(\Omega)}$ cannot be computed exactly which is required for the setting of the bundle method in Section 4.2. If desired, Section 4.2 can be modified to avoid the transformation $u = \bar{u} - u_d$.

The data of example 3 is given as follows. The domain is $\Omega := (0,1)^2$ and three (pairwise disjoint) subdomains $\Omega_1, \Omega_2, \Omega_3$ are specified via

$$\begin{aligned}
\Omega_1 &:= (0.8, 0.9)^\top + 0.05 \cdot Q\left((-1,1)^2\right), \\
\Omega_2 &:= (0, 0.5) \times (0, 0.8), \\
\Omega_3 &:= (0.5, 1) \times (0, 0.8),
\end{aligned}$$

where $Q \in \mathbb{R}^{2\times 2}$ is the rotation matrix

$$Q := \begin{pmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{pmatrix}.$$

The desired state and the data term are given by

$$y_d(x) := \begin{cases} \Delta p_1(Q^\top x), & x \in \Omega_1 \\ z_1(x_1) z_2(x_2), & x \in \Omega_2 \\ 0 & \text{else} \end{cases}$$

and

$$\mathring{f}(x) = u_d(x) := \begin{cases} p_1(Q^\top x), & x \in \Omega_1 \\ -z_1''(x_1) z_2(x_2) - z_1(x_1) z_2''(x_2), & x \in \Omega_2 \\ -z_1(x_1 - 0.5) z_2(x_2), & x \in \Omega_3 \\ 0 & \text{else} \end{cases},$$

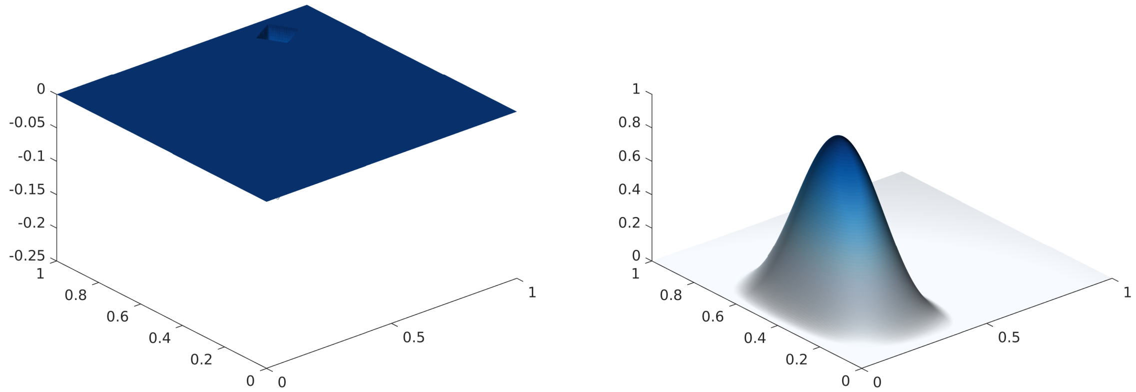
where z_1 and z_2 are defined via (6.3.1) and

$$\begin{aligned}
q_1(y_1) &:= -200(y_1 - 0.8)^2 + 0.5, \\
q_2(y_2) &:= -200(y_2 - 0.9)^2 + 0.5, \\
p_1(y_1, y_2) &:= q_1(y_1) q_2(y_2).
\end{aligned}$$

A local minimum is attained at the control u^* and the state y^* is defined via

$$u^*(x) := \begin{cases} -p_1(Q^\top x) & x \in \Omega_1 \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad y^*(x) := \begin{cases} z_1(x_1) z_2(x_2), & x \in \Omega_2 \\ 0 & \text{else.} \end{cases}$$

The analytic solutions u^* and y^* are plotted in Figure 6.11.



(a) Optimal control **(b)** Optimal state

Figure 6.11.: Analytic solutions of Example 3.

We run the bundle method using strategies A and B with the subgradient oracle O_g^1. The start mesh contained 1089 nodes and has a mesh width of $h = \sqrt{2}/32$. The initial lift term is set to $l_0 := 25$ and the constant for the a priori error estimate for the function value (cf. Section 6.1) is set to $C_\Omega = 0.012$. We stop the algorithm after a time of 400 seconds. The final mesh has a maximal triangle side length of $h_3 = \sqrt{2}/256$ and 66049 nodes.

Figure 6.12 depicts the same data as Figure 6.2. There are three mesh refinements each. The last mesh refinement happens at iteration 77 after 15.4 seconds for strategy A and at iteration 78 after 12.0 seconds for strategy B. The rest of the computation time of 400 seconds is spent on the final mesh with 66049 nodes and mesh width $\sqrt{2}/256$. As one can see, this problem is much more challenging than examples 1 and 2. It can be observed that the process of finding the optimum on the final mesh is tedious. However, at the last iteration, strategy A yields a relative control error of $\|u_i - u^*\|_{L^2(\Omega)}/\|u^*\|_{L^2(\Omega)} = 2.58 \cdot 10^{-2}$ which can be compared to the relative control error of $2.98 \cdot 10^{-2}$ for a uniform mesh of mesh width $\sqrt{2}/250$ in [88, Tab.7.1].

In order to compare the performance of strategy A and B, we compute the high fidelity approximation of the state on a uniform mesh with mesh width $h = \sqrt{2}/512$ and 263169 nodes. The resulting solution operator S^{h^*} yields an error for the computation of the optimal state of $\|y^{h^*} - y^*\|_{L^2(\Omega)} = 2.17 \cdot 10^{-5}$. The error for computation of the function value of the optimal control is $|J^{h^*} - J^*| = 1.03 \cdot 10^{-7}$. The course of control and state error over time is depicted in Figure 6.13. It can be observed that the control error is reduced during the whole computation time. In contrast to this, after approximately 10 seconds, the state error remains constant at the error level of the computation of the optimal state. For this example we do not observe a stabilizing behavior of the BFGS curvature scheme. The computed function value error $|J_i^{h^*} - J^{h^*}|$ is displayed in Figure 6.14. The control u_i and the state y_i enter the objective function via $\frac{1}{2}\|u_i\|^2_{L^2(\Omega)}$ and $\frac{1}{2}\|y_i - y_d\|^2_{L^2(\Omega)}$, respectively. This squared relation explains the behavior of the computed function value error. Since the control error is already of order 10^{-5}, it does not influence the computed function value error, which is of order 10^{-7}. Squaring the order of the control error yields the order of the computed function value error.

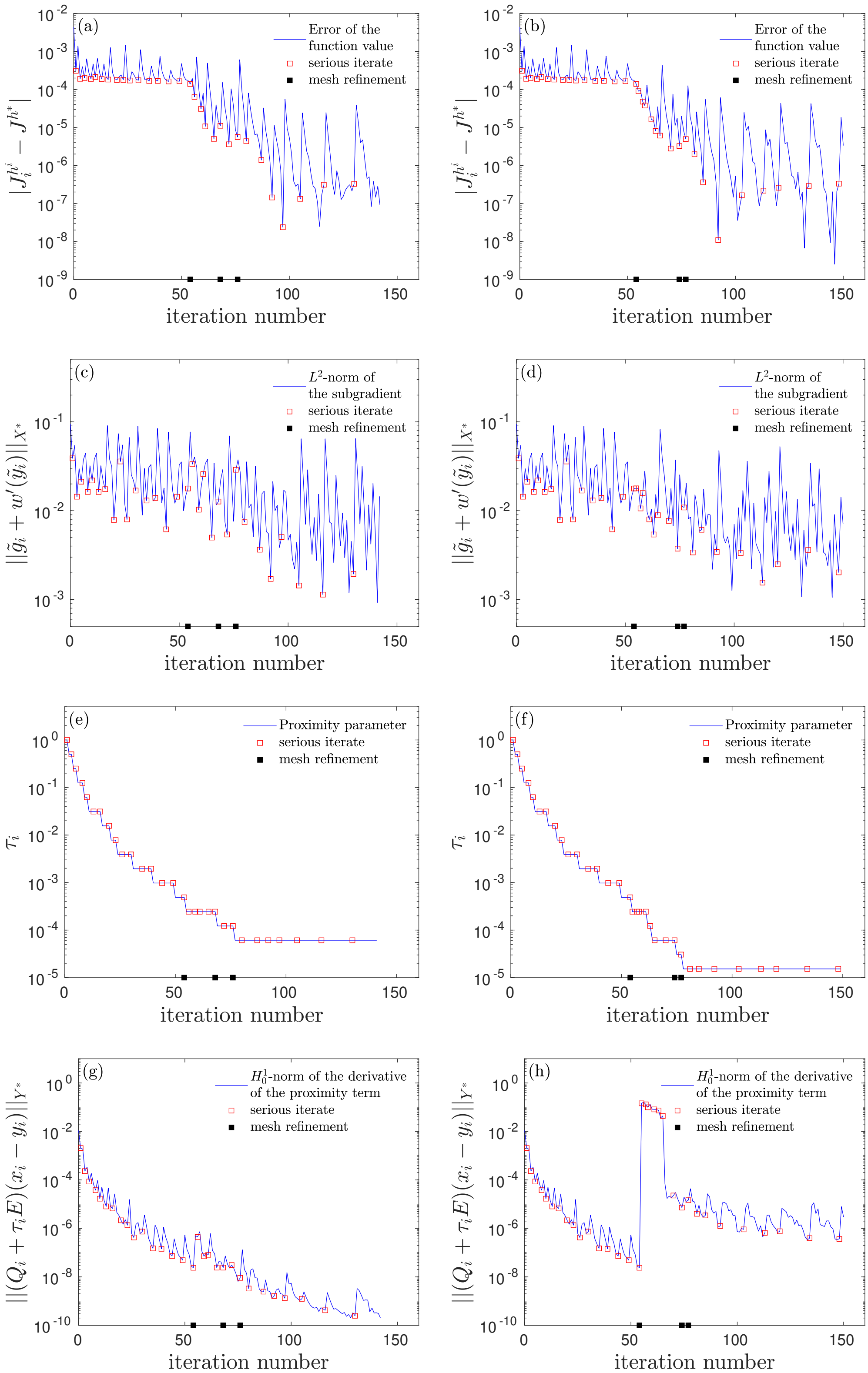


Figure 6.12.: Convergence statistics of Algorithm 3.4 using uniform mesh refinement. No curvature strategy (a),(c),(e),(f). BFGS curvature (b), (d), (f), (h).

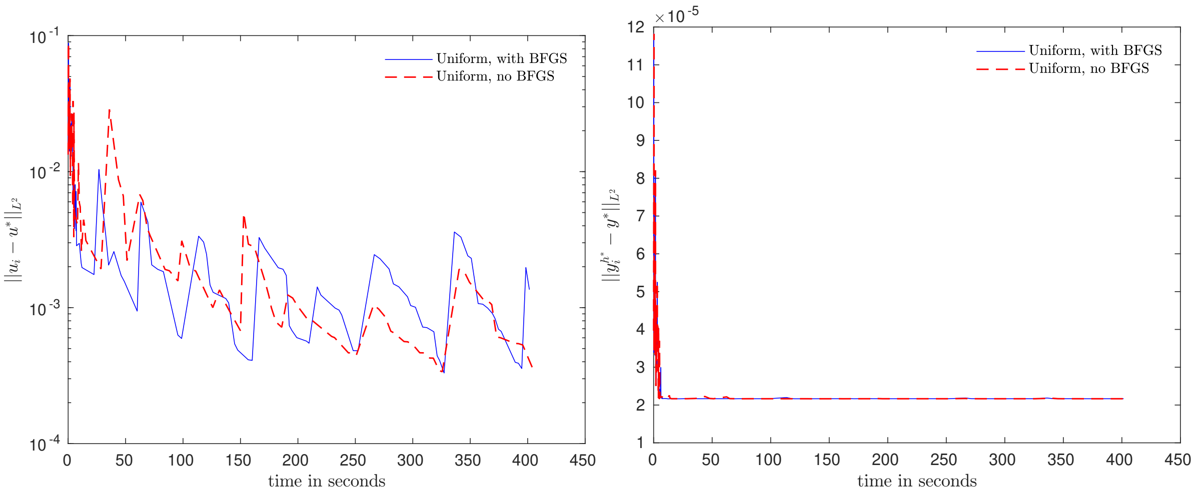


Figure 6.13.: Example 3. Relative difference of the controls and states to the optimal values over time.

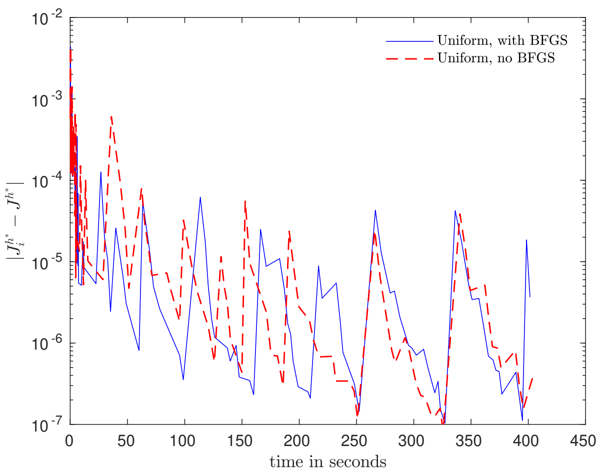


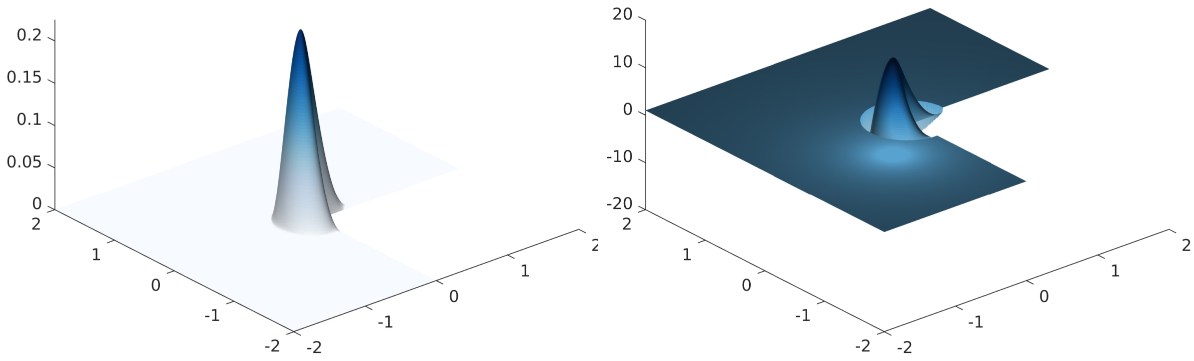
Figure 6.14.: Example 3. Computed difference of the function values to the optimal function value over time.

6.5. Example 4

This example was taken from [35, Ex. 6.1]. We consider the L-shaped domain $\Omega = (-2,2)^2 \setminus ([0,2] \times [-2,0])$ with $\psi \equiv 0$ and $\alpha = 1$. Further data is given via $\mathring{f} := -\Delta y^* - y^* - \xi^*$, $y_d := y^* + \xi^* - \Delta y^*$ and $u^* := y^*$ where, in polar coordinates,

$$y^* : \mathbb{R} \times [0, 2\pi) \to \mathbb{R}, \qquad y^*(r,\varphi) := -(16r^3 - 12r^2 + 1)r^{2/3}\delta_{[0,0.5]}(r)\sin(2/3\varphi),$$
$$\xi^* : \mathbb{R} \times [0, 2\pi) \to \mathbb{R}, \qquad \xi^*(r,\varphi) := \delta_{[0.5,\infty]}(r).$$

It can be shown that the triple (y^*, ξ, u^*) is a strong-stationary point of (5.3.1), cf. [35, Chap. 2.4]. The control u^* and the state y^* are depicted in Figure 6.15(a) and the desired state y_d is depicted in Figure 6.15(b).



(a) Optimal control/optimal state **(b)** Desired state

Figure 6.15.: Data for example 4.

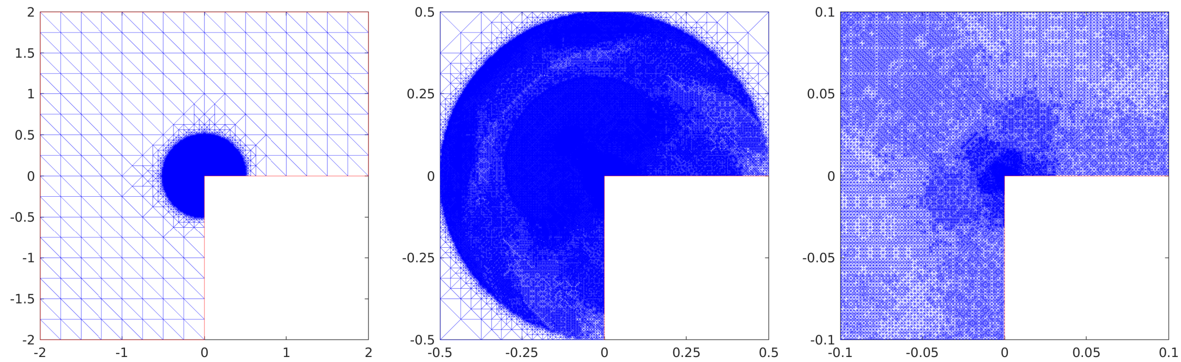


Figure 6.16.: Final mesh and zoom into the vicinity of the origin for example 4 using the adaptive strategy C.

In order to compare adaptive to uniform refinement, we solve problem 4 with strategies A and C with the subgradient oracle O_g^1, i.e., we compare the adaptive refinement strategy C with no curvature operator to the uniform refinement strategy A with no curvature operator. The start mesh has a mesh width of $h = \sqrt{2}/4$ and consists of 225 nodes. We run the algorithm for a maximal time of 100 seconds and allow for a maximum number of 300000 nodes. In both strategy A and C, the initial lift term is set to $l_0 := 25$. The constant for the a priori error estimate for the function value (cf. Section 6.1) is set to $C_\Omega = 0.07$ (strategy A). The constant for the a posteriori error estimate for the function value is set to $C = 0.07$ (strategy C). After 9 iterations, the uniform refinement strategy A encounters the need for a better solution of the bundle subproblem but cannot refine the discretization anymore and therefore stops. The final uniform mesh has a mesh width of $\sqrt{2}/128$ with 197633 nodes. The adaptive refinement strategy C stops after 10 iterations due to the same reason. The adaptively refined mesh after stopping consists of 268882 nodes and is depicted in Figure 6.16. As expected, the mesh is refined in the circle around the origin with radius 0.5 and most refinement occurs at the origin.

In Figure 6.17, the same data as in Figure 6.2 is depicted. Since strategy A and C use the same start mesh, the results agree for both strategies until the first mesh refinement occurs in iteration 3. The mesh refinements occur every or every second iteration. This indicates that on every given mesh, the algorithm finds a close approximation of the solution within one step. Whereas the uniform refinement

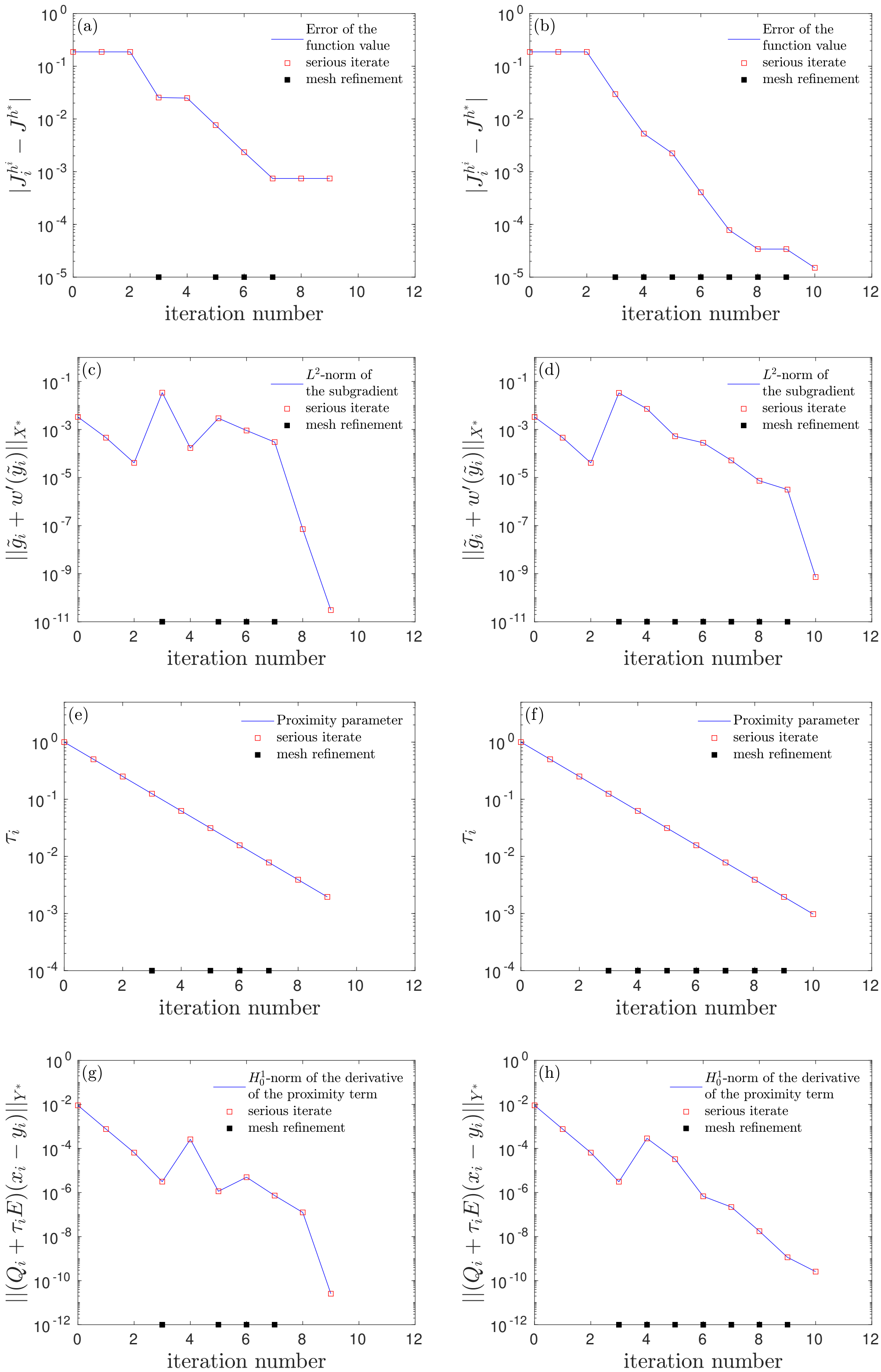


Figure 6.17.: Convergence statistics of Algorithm 3.4 using no curvature strategy. Uniform refinement (a),(c),(e),(f). Adaptive refinement (b), (d), (f), (h).

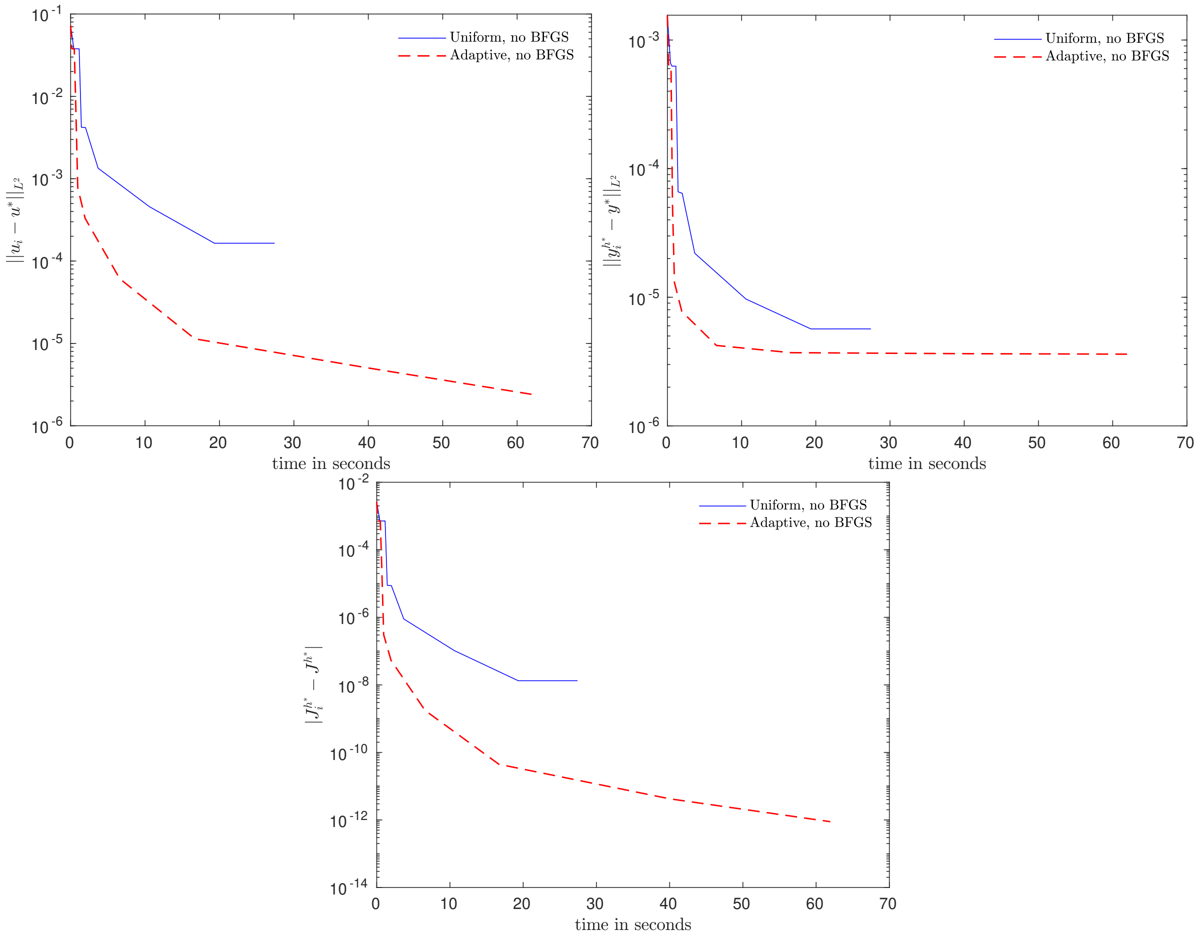


Figure 6.18.: Example 4. Difference of the controls, states and function values to the optimal values over time.

strategy reaches a computed function value error in the order of 10^{-3}, the adaptive strategy reduces the error to 10^{-5}. Both strategies find a subgradient with norm below 10^{-9}. For both strategies, the proximity parameter τ_i does not tend to infinity and the norm of the derivative of the proximity term at the last iteration $\|\hat{e}_{\text{end}}\|_{H_0^1(\Omega)}$ is below 10^{-10} which suggest that the hard case $\mathscr{E}_{\bar{x}} \neq \emptyset$ in Theorem 3.3.3 does not occur. Similar to example 2, our numerical experiments showed no significant differences between the no curvature strategy and the BFGS curvature strategy, i.e., between strategy A and B and strategy C and D. Thus, we do not depict any results for strategies B and D.

In order to compare the true function value $J(u_i)$ at each iterate to the optimal function value J^*, we again compute the state $S(u_i)$ on a fine mesh and approximate $J(u_i)$ and $J(u^*)$ via

$$J_i^{h^*} := \tfrac{1}{2}\|S^{h^*}(u_i + \mathring{f}) - y_d\|_{L^2(\Omega)}^2 + \tfrac{\alpha}{2}\|u_i\|_{L^2(\Omega)}^2, \quad J^{h^*} := \tfrac{1}{2}\|S^{h^*}(u^* + \mathring{f}) - y_d\|_{L^2(\Omega)}^2 + \tfrac{\alpha}{2}\|u^*\|_{L^2(\Omega)}^2,$$

respectively. The fine mesh which induces the high fidelity solution operator S^{h^*} is constructed as follows. We start with a uniform mesh with mesh width $\sqrt{2}/4$ and 225 nodes. This mesh is refined adaptively (such that the optimal state can be computed most efficiently) as described in Section 5.5.1.

This refinement is repeated until a mesh with 268882 nodes is reached. Now we create a uniform mesh with mesh width $h = \sqrt{2}/128$ and 197633 nodes. Finally, we merge both meshes such that the resulting mesh consists of the vertices of both meshes and triangles created by the MATLAB function DELAUNAYTRIANGULATION. The resulting high fidelity solution operator S^{h^*} yields an error for the computation of the optimal state of $\|y^{h^*} - y^*\|_{L^2(\Omega)} = 3.60 \cdot 10^{-6}$. The error for computation of the function value of the optimal control is $|J^{h^*} - J^*| = 1.60 \cdot 10^{-5}$. The resulting control, state and function value errors are depicted in Figure 6.18. We observe that the adaptive mesh refinement results in a control error of order 10^{-6} whereas the control error for uniform refinement is of order 10^{-4}. Since the high fidelity mesh can only resolve a state error of $3.60 \cdot 10^{-6}$, we cannot observe significant differences of the state error. The adaptive mesh refinement strategy performs better with regards to the computed function value error.

7. Optimal control of the stochastic obstacle problem

In this section, we consider a stochastic version of Problem (5.2.1). In real world applications, often it is desirable to know not the optimal solution to Problem (5.2.1) but rather a robust solution, i.e., a solution which performs well also for minor changes in the problem data. This chapter is mainly based on the paper

[49] L. HERTLEIN, A.-T. RAULS, M. ULBRICH, AND S. ULBRICH, *An inexact bundle method and subgradient computations for optimal control of deterministic and stochastic obstacle problems*. Priprint, accepted for publication in SPP1962 Special Issue, Birkhäuser, 2019.

7.1. The stochastic obstacle problem

Let $(\Xi, \mathscr{A}, P)$ be a complete σ-finite measure space and abbreviate $Z := H_0^1(\Omega)$. For $\xi \in \Xi$, we consider a variational inequality of type (5.1.2). In particular, let $A_\xi \in \mathscr{L}(Z, Z^*)$ be an operator, let $b_\xi \in Z^*$ be a force, let $\psi_\xi \in \bar{H} := H^1(\Omega)$ be an obstacle, define the set

$$K_\xi := \{y_\xi \in H_0^1(\Omega) : y_\xi \geq \psi_\xi\}$$

and define the (parametric) obstacle problem

$$\text{Find } y_\xi \in K_\xi, \quad \langle A_\xi y_\xi - b_\xi, v_\xi - y_\xi \rangle_{Z^*,Z} \geq 0 \quad \forall v_\xi \in K_\xi. \tag{VI$_\xi$}$$

The stochastic obstacle problem is given by

$$\text{Find } \mathbf{y} \in \mathbf{K}, \quad \langle \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{v} - \mathbf{y} \rangle_{\mathbf{H}^*,\mathbf{H}} \geq 0 \quad \text{ for all } \mathbf{v} \in \mathbf{K}. \tag{\textbf{VI}}$$

Here, $\mathbf{H} := L^2(\Xi, H_0^1(\Omega))$ is the Bochner space of square integrable functions with values in $H_0^1(\Omega)$ (cf. [62, Def. 1.2.15]), $\mathbf{A} \in \mathscr{L}(\mathbf{H}, \mathbf{H}^*)$, $\mathbf{b} \in \mathbf{H}^*$, $\boldsymbol{\psi} \in \bar{\mathbf{H}} := L^2(\Xi, H^1(\Omega))$ and

$$\mathbf{K} := \{\mathbf{y} \in \mathbf{H} : \mathbf{y}(\xi) \in K_\xi \text{ for } P\text{-a.a. } \xi \in \Xi\}. \tag{7.1.1}$$

We want to relate the solutions to (VI$_\xi$) and (**VI**), see [44, 45] for related results. Using standard techniques, one can show that the projection onto the set $\mathbf{K}$, defined in (7.1.1), agrees pointwise P-a.e. with the projection onto K_ξ:

LEMMA 7.1.1. *If $\boldsymbol{\psi} \in \bar{\mathbf{H}}$ such that $K_\xi \neq \emptyset$ for P-a.a. $\xi \in \Xi$ then $\mathbf{K}$ is a nonempty closed convex subset of $\mathbf{H}$ and $P_{\mathbf{K}}(\mathbf{v})(\xi) = P_{K_\xi}(\mathbf{v}(\xi))$ for P-a.a. $\xi \in \Xi$ and for all $\mathbf{v} \in \bar{\mathbf{H}}$.*

Proof. Let $\mathbf{v} \in \bar{\mathbf{H}}$ be arbitrary and denote by $p_\xi = P_{K_\xi}(\mathbf{v}(\xi))$ the projection of $\mathbf{v}(\xi)$ onto the nonempty closed convex set K_ξ, i.e.,

$$p_\xi \in K_\xi, \qquad (v - p_\xi, p_\xi - w)_Z \geq 0, \qquad \forall w \in K_\xi.$$

By [8, Thm. 8.2.9], the multifunction $\xi \mapsto K_\xi$ is measurable (cf. Section 2.9 for the definition of measurable set valued functions). Thus, [45, Thm. 2.3] implies that $\xi \mapsto p_\xi$ is measurable. As P is a σ-finite measure and $H_0^1(\Omega)$ is separable, Theorem 2.9.1 shows that $\xi \mapsto p_\xi$ is strongly P-measurable. Since the projection P_{K_ξ} is non-expansive, we find $\xi \mapsto P_{K_\xi}(\mathbf{v}(\xi)) \in \mathbf{H}$ for all $\mathbf{v} \in \bar{\mathbf{H}}$. Consequently, $\xi \mapsto P_{K_\xi}(\boldsymbol{\psi}(\xi)) \in \mathbf{H}$ and $\mathbf{K}$ is nonempty. Denote by

$$H_0^1(\Omega)^+ := \{v \in H_0^1(\Omega) : v \geq 0 \text{ a.e. on } \Omega\}$$

the positive cone in $Z = H_0^1(\Omega)$. By [44, Lem. 3.1], the set

$$\mathbf{K} - \boldsymbol{\psi} = \{\mathbf{v} \in \mathbf{H} : \mathbf{v}(\xi) \in H_0^1(\Omega)^+ \text{ for } P\text{-a.a. } \xi \in \Xi\}$$

is closed, and so is $\mathbf{K}$. It is easy to see that $\mathbf{K}$ is convex. Therefore, $\mathbf{K}$ is a nonempty closed and convex subset of the space $\mathbf{H}$. Because Z is a Hilbert space, [18, Thm. 3.1] implies that $\mathbf{H} = L^2(\Xi, Z)$ also is a Hilbert space. Therefore, the projection $P_{\mathbf{K}} : \bar{\mathbf{H}} \to \mathbf{K}$ is well-defined and fulfills

$$\mathbf{p} = P_{\mathbf{K}}(\mathbf{v}) \qquad \Leftrightarrow \qquad \mathbf{p} \in \mathbf{K}, \qquad (\mathbf{v} - \mathbf{p}, \mathbf{p} - \mathbf{w})_{\bar{\mathbf{H}}} \geq 0, \qquad \forall \mathbf{w} \in \mathbf{K}.$$

Since $\xi \mapsto P_{K_\xi}(\mathbf{v}(\xi))$ fulfills this, we find $P_{\mathbf{K}}(\mathbf{v})(\xi) = P_{K_\xi}(\mathbf{v}(\xi))$ for P-a.a. $\xi \in \Xi$ and for all $\mathbf{v} \in \bar{\mathbf{H}}$. □

Using this result, we can show that the solution operator of **(VI)** agrees pointwise P-a.e. with the solution operator of (VI$_\xi$). We need the following definition.

DEFINITION 7.1.2. *A family of operators $(A_\xi)_{\xi \in \Xi} \subset \mathscr{L}(Z, Z^*)$ is called uniformly coercive, if there exists a parameter $C_L > 0$ such that $\langle A_\xi x, x \rangle_{Z^*,Z} \geq C_L \|x\|_Z^2$ for all $x \in Z$ and P-a.a. $\xi \in \Xi$.*

THEOREM 7.1.3. *Assume that $\xi \mapsto A_\xi y$ is strongly P-measurable for every $y \in Z$, that $\xi \mapsto \|A_\xi\|_{\mathscr{L}(Z,Z^*)}$ is in $L^\infty(\Xi)$, $\mathbf{b} : \xi \mapsto b_\xi$ is in $\mathbf{H}^*$, $\boldsymbol{\psi} : \xi \mapsto \psi_\xi$ is in $\bar{\mathbf{H}}$. Then, for every $\mathbf{y} \in \mathbf{H}$, the map $\mathbf{A_y} : \xi \mapsto A_\xi(\mathbf{y}(\xi))$ is in $\mathbf{H}^*$ and $\mathbf{A} : \mathbf{y} \mapsto \mathbf{A_y}$ is in $\mathscr{L}(\mathbf{H}, \mathbf{H}^*)$. Moreover, suppose that $(A_\xi)_{\xi \in \Xi}$ is uniformly coercive and that $K_\xi \neq \emptyset$ for P-a.a. $\xi \in \Xi$. Then, for P-a.a. $\xi \in \Xi$,* (VI$_\xi$) *has a unique solution y_ξ and the solution operator $S_\xi : Z^* \to Z$, $S_\xi(b_\xi) = y_\xi$, is Lipschitz with modulus $1/C_L$. Furthermore,* **(VI)** *has a unique solution, the solution operator $\mathbf{S} : \mathbf{H}^* \to \mathbf{H}$ is Lipschitz with modulus $1/C_L$, and $(\mathbf{S}(\mathbf{b}))(\xi) = S_\xi(b_\xi)$ for P-a.a. $\xi \in \Xi$.*

Proof. Let $\mathbf{y} \in \mathbf{H}$ be arbitrary. Since $\xi \mapsto A_\xi$ and $\xi \mapsto \mathbf{y}(\xi)$ are strongly P-measurable, [62, Prop. 1.1.28] implies that $\mathbf{A_y}$ is strongly P-measurable. Using [62, Prop. 1.2.2], the estimate

$$\|\mathbf{A_y}\|_{\mathbf{H}^*}^2 = \int_\Xi \|A_\xi(\mathbf{y}(\xi))\|_{Z^*}^2 \, dP(\xi) \leq \int_\Xi \|A_\xi\|_{\mathscr{L}(Z,Z^*)}^2 \|\mathbf{y}(\xi)\|_Z^2 \, dP(\xi) < \infty$$

shows that $\mathbf{A_y} \in \mathbf{H}^*$ and that the linear mapping $\mathbf{A} : \mathbf{y} \mapsto \mathbf{A_y}$ is continuous. Now, suppose that $(A_\xi)_{\xi \in \Xi}$

is uniformly coercive with constant C_L and that $K_\xi \neq \emptyset$ for P-a.a. $\xi \in \Xi$. From

$$\langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle_{\mathbf{H}^*,\mathbf{H}} = \int_\Xi \langle A_\xi(\mathbf{y}(\xi)), \mathbf{y}(\xi)\rangle_{Z^*,Z}\, dP(\xi) \geq \int_\Xi C_L \|\mathbf{y}(\xi)\|_Z^2\, dP(\xi) = C_L\|\mathbf{y}\|_{\mathbf{H}}^2 \qquad \forall \mathbf{y} \in \mathbf{H}$$

we deduce that $\mathbf{A}$ is coercive. By Lemma 7.1.1, $\mathbf{K}$ is a nonempty closed and convex subset of a Hilbert space. Thus, the Lions-Stpampacchia theorem, cf. [67, Thm. 2.1], implies that **(VI)** and (VI$_\xi$) are uniquely solvable for P-a.a. $\xi \in \Xi$. Furthermore, by [8, Thm. 8.2.9] the multifunction $\xi \mapsto K_\xi$ is measurable and contains the element $\mathbf{p} \in \mathbf{H}$. Thus, [45, Thm. 2.7] implies $\xi \mapsto S_\xi(b_\xi) \in \mathbf{H}$. Since $\mathbf{y} \in \mathbf{H}$, defined by $\mathbf{y}(\xi) := S_\xi(b_\xi) \in K_\xi$, fulfills **(VI)** and the solution of **(VI)** is unique, we deduce $\mathbf{y} = \mathbf{S}(\mathbf{b})$. □

7.2. Optimal control of the stochastic obstacle problem

We are interested in the following class of optimal control problems governed by the stochastic obstacle problem

$$\min_{u \in U_{\mathrm{ad}}} \mathbf{J}(\mathbf{S}(\mathring{\mathbf{F}}(\iota u)) + \tfrac{\alpha}{2}\|u\|_U^2. \tag{\textbf{P}}$$

Here $\iota \in \mathscr{L}(U, Z^*) = \mathscr{L}(L^2(\Omega), H^{-1}(\Omega))$ is a compact embedding, $\mathring{\mathbf{F}} : Z^* \to \mathbf{H}^*$ is a continuous function which maps the control to the force term, $\mathbf{S} : \mathbf{H}^* \to \mathbf{H}$ is the solution operator of the stochastic variational inequality **(VI)**, $\mathbf{J} : \mathbf{H} \to \mathbb{R}$ is the stochastic objective function and $U_{\mathrm{ad}} \subset U$ is a nonempty, closed and convex set.

LEMMA 7.2.1. *If $\mathbf{J} : \mathbf{H} \to \mathbb{R}$ is lower continuous and bounded below, then the problem* **(P)** *has a solution.*

Proof. We verify the assumptions of Theorem 2.5.1. Since $\mathbf{J}$ and $\frac{\alpha}{2}\|\cdot\|_U^2$ are bounded below, the function $j : \mathbf{H} \times U$, $j(\mathbf{y}, u) := \mathbf{J}(\mathbf{y}) + \frac{\alpha}{2}\|u\|_U^2$ is bounded below. Further, $S : Z^* \to \mathbf{H}$, $S := \mathbf{S} \circ \mathring{\mathbf{F}}$, is continuous. As $\mathbf{J}$ is bounded below, the coercivity of $\frac{\alpha}{2}\|u\|_U^2$ yields that the reduced function $J : U \to \mathbb{R}$, $J(u) := \mathbf{J}(\mathbf{S}(\mathring{\mathbf{F}}(\iota u)) + \frac{\alpha}{2}\|u\|_U^2$, is coercive. As $\mathbf{J}$ is (strongly) lower semicontinuous, $j : \mathbf{H} \times U$ is strong$\times$weak sequentially lower semicontinuous. Therefore, Theorem 2.5.1 is applicable which implies that **(P)** has a solution. □

7.3. Approximate subgradients for the stochastic obstacle problem

If $\mathbf{J}$ and $\mathring{\mathbf{F}}$ are Lipschitz on bounded sets, then **(P)** corresponds to the setting of the bundle method via $X := U = L^2(\Omega)$, $\mathscr{F} := U_{\mathrm{ad}}$, $Y := Z^* = H^{-1}(\Omega)$, $p := \mathbf{J} \circ \mathbf{S} \circ \mathring{\mathbf{F}}$, $w := \frac{\alpha}{2}\|\cdot\|_U^2$. To execute the bundle method, we need to be able to compute a candidate for a trial iterate, a function value approximation and an element of an approximate subdifferential. A trial iterate can be computed using the theory developed in Chapter 4. As $\mathbf{J}$ is a function which maps $L^2(\Xi, H_0^1(\Omega))$ to $\mathbb{R}$, one usually needs to evaluate an integral to obtain a function value (cf. below). In this case, to obtain approximate function values, numerical integration [108, 61, 23] or Monte Carlo methods [10] can be used. In the rest of this section we focus on how to determine an appropriate approximate subdifferential G such that we can compute an element thereof.

In the following, we work under the assumptions of Theorem 7.1.3 such that the solution operators of **(VI)** and (VI_ξ) are connected via $(\mathbf{S}(\mathbf{b}))(\xi) = S_\xi(b_\xi)$ for P-a.a. $\xi \in \Xi$. We concretize the stochastic objective function $\mathbf{J} : \mathbf{H} \to \mathbb{R}$ via $\mathbf{J}(\mathbf{y}) := \mathbb{E}\left[J_\xi(\mathbf{y}(\xi))\right]$. Here, for P-a.a. $\xi \in \Xi$, $J_\xi : Z \to \mathbb{R}$ is an objective function such that $\xi \mapsto J_\xi(\mathbf{y}(\xi))$ is integrable for all $\mathbf{y} \in \mathbf{H}$ and $\mathbb{E}$ denotes the expectation with respect to ξ, i.e., $\mathbb{E}[\mathbf{z}] := \int_\Xi \mathbf{z}(\xi)\, dP(\xi)$. Further, define $\mathring{F}_\xi : Z^* \to Z^*$ via $\mathring{F}_\xi(w) := \mathring{\mathbf{F}}(w)(\xi)$. Then (**P**) takes the form

$$\min_{u \in U_{\mathrm{ad}}} \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(\iota u)))\right] + \tfrac{\alpha}{2}\|u\|_U^2. \qquad \textbf{(P')}$$

We would like to compute an element $g \in Z$ of Clarke's subdifferential of the reduced stochastic objective function $p : Z^* \to \mathbb{R}$, $p := \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(\cdot)))\right]$, i.e.,

$$g \in \partial_C p(w) = \partial_w \mathbf{J}(\mathbf{S}(\mathring{\mathbf{F}}(w))) = \partial_w \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(w)))\right].$$

However, the available calculus rules for the Clarke subdifferential, which often take the form of inclusions, make it difficult to calculate the subdifferential $\partial_w \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(w)))\right]$. As Lemma 5.2.6 shows how to compute a subgradient of the function $J_\xi(S_\xi(\mathring{F}_\xi(\,\cdot\,)))$, we aim at computing elements of

$$\begin{aligned}\mathbb{E}\left[\partial_w J_\xi(S_\xi(\mathring{F}_\xi(w)))\right] := \{\mathbb{E}\left[g(w,\xi)\right] : \xi \mapsto g(w,\xi) \in L^1(\Xi, H^{-1}(\Omega)),\\ g(w,\xi) \in \partial_w J_\xi(S_\xi(\mathring{F}_\xi(w)))\ P\text{-a.e. }\}.\end{aligned} \qquad (7.3.1)$$

Provided that $G(w) := \mathbb{E}\left[\partial_w J_\xi(S_\xi(\mathring{F}_\xi(w)))\right]$ fulfills Assumption 3.1.1, this choice of approximate subdifferential for the bundle method yields convergence to weak stationary points (cf. [136]), i.e., points $\bar{u} \in U$ which fulfill

$$0 \in \iota^* \mathbb{E}\left[\partial_w J_\xi(S_\xi(\mathring{F}_\xi(\iota \bar{u})))\right] + \alpha \bar{u} + N_{U_{\mathrm{ad}}}(\bar{u}) + \iota^* \bar{B}_Z(0, \eta).$$

Under suitable assumptions, [21, Thm. 2.7.2 and Thm. 2.3.10] imply

$$\partial_u \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(\iota u)))\right] \subset \mathbb{E}\left[\partial_u J_\xi(S_\xi(\mathring{F}_\xi(\iota u)))\right] \subset \iota^* \mathbb{E}\left[\partial_w J_\xi(S_\xi(\mathring{F}_\xi(\iota u)))\right]$$

with equality if, for each ξ, the function $J_\xi(S_\xi(\mathring{F}_\xi(\,\cdot\,)))$ is regular in the sense of Clarke. However, this is not necessarily the case for all points of the considered optimal control problem. Under strong assumptions (such as a deterministic obstacle), in [49, Thm. 7.11] a formula for an exact subgradient $g \in \partial_w \mathbb{E}\left[J_\xi(S_\xi(\mathring{F}_\xi(w))))\right]$ is given. Here, we avoid these assumptions but this comes at the cost that we can only guarantee convergence of the bundle method to weak stationary points.

Now we show that the weak subgradients (7.3.1) can be used in the bundle method since they fulfill Assumption 3.1.1. We work in the following setting:

Assumption 7.3.1. Let $\mathscr{F}_B$ be an open subset of a separable reflexive Banach space B. Suppose that for all $\xi \in \Xi$ the functions $p_\xi : \mathscr{F}_B \to \mathbb{R}$ satisfy the following conditions:

1. For all $w \in \mathscr{F}_B$, the map $\xi \mapsto p_\xi(w)$ is measurable.

2. There exists a $w \in \mathscr{F}_B$ such that $\int_\Xi |p_\xi(w)|\,dP(\xi) < \infty$.

3. For all bounded sets $D \subset B$ there exists a function $L_D \in L^1(\Xi)$ such that

$$|p_\xi(w_1) - p_\xi(w_2)| \le L_D(\xi)\|w_1 - w_2\|_B \quad \text{for all } w_1, w_2 \in D \cap \mathscr{F}_B \text{ and for } P\text{-a.a. } \xi \in \Xi.$$

Let $\bar{\mathscr{F}}$ be a closed subset of $\mathscr{F}_B$ and consider the map $G : \bar{\mathscr{F}} \rightrightarrows B^*$ defined by

$$G(w) := \left\{ \int_\Xi g(\xi)\,dP(\xi) : g \in L^1(\Xi, B^*), g(\xi) \in \partial_C p_\xi(w) \ P\text{-a.e.} \right\}. \tag{7.3.2}$$

THEOREM 7.3.2. *Under Assumption 7.3.1, the multifunction $G : \bar{\mathscr{F}} \rightrightarrows B^*$, defined in* (7.3.2)*, fulfills Assumption 3.1.1 and it holds $\partial_C p(w) \subset G(w)$ for all $w \in \bar{\mathscr{F}}$, where $p : \bar{\mathscr{F}} \to \mathbb{R}$ is defined by $p(w) := \int_\Xi p_\xi(w)\,dP(\xi)$.*

Proof. First we show $\partial_C p(w) \subset G(w)$. Let $w \in \bar{\mathscr{F}}$ be arbitrary. By [21, Thm. 2.7.2], p is well-defined, locally Lipschitz and for every $g \in \partial_C p(w)$ there is a corresponding mapping $\xi \mapsto g_\xi$ from Ξ to B^* with $g_\xi \in \partial_C p_\xi(w)$ P-a.e. and such that for every $v \in B$, the function $\xi \mapsto \langle g_\xi, v\rangle_{B^*,B}$ belongs to $L^1(\Xi)$ and one has $\langle g, v\rangle_{B^*,B} = \int_\Xi \langle g_\xi, v\rangle_{B^*,B}\,dP(\xi)$. Consequently, by [62, Cor. 1.1.2], the map $\xi \mapsto g_\xi$ is measurable. Denote by $L_D \in L^1(\Xi)$ the function according to property 3 of Assumption 7.3.1 for $D := \bar{B}_X(w,1)$. From $\int_\Xi \|g_\xi\|_{X^*}\,dP(\xi) \le \int_\Xi L_D(\xi)\,dP(\xi) < \infty$ we deduce that $\xi \mapsto g_\xi$ is in $L^1(\Xi, B^*)$ which shows $\partial_C p(w) \subset G(w)$.
1. For arbitrary $w \in \bar{\mathscr{F}}$, it holds $\emptyset \neq \partial_C p(w) \subset G(w)$. Therefore, $G(w)$ is nonempty. Since $\partial_C p_\xi(w)$ is convex P-a.e. , the set $G(w)$ is convex.
2. Let $D \subset B$ be a bounded set and denote

$$\hat{G} := \{\hat{g} \in L^1(\Xi, B^*) : w \in D \cap \bar{\mathscr{F}}, \hat{g}(\xi) \in \partial_C p_\xi(w) \ P\text{-a.e. }\}. \tag{7.3.3}$$

Choose a neighborhood $\hat{D} \subset B$ of $B \cap \bar{\mathscr{F}}$ and denote by $L_{\hat{D}} \in L^1(\Xi)$ the function which fulfills property 2 of Assumption 7.3.1. By [21, Prop. 2.1.2], there holds $\partial_C p_\xi(w) \subset \bar{B}_{B^*}(0, L_{\hat{D}}(\xi))$ for all $w \in D \cap \bar{\mathscr{F}}$. Consequently, $\hat{G}$ is bounded in $L^1(\Xi, B^*)$ by the constant $\int_\Xi L_{\hat{D}}(\xi)\,dP(\xi) < \infty$ and we find for arbitrary $g \in G(D \cap \bar{\mathscr{F}})$ that there exists a $\hat{g} \in \hat{G}$ such that $g = \int_\Xi \hat{g}(\xi)\,dP(\xi)$ and it holds

$$\|g\|_{B^*} = \| \int_\Xi \hat{g}(\xi)\,dP(\xi)\|_{B^*} \le \int_\Xi \|\hat{g}(\xi)\|_{B^*}\,dP(\xi) \le \int_\Xi L_{\hat{D}}(\xi)\,dP(\xi).$$

3. We verify the assumptions of [105, Thm. 4.2]. Since $\bar{\mathscr{F}}$ is a closed subset of a complete metric space, $(\bar{\mathscr{F}}, \|\cdot\|_B)$ is a complete metric space. By [21, Prop. 2.1.2], the map $(\xi, v) \mapsto \partial_C p_\xi(v)$ is nonempty, closed and convex valued. Using [21, Lem. 2.7.2], [8, Thm. 8.2.11 and Thm. 8.2.9], one sees that the multifunction $\xi \mapsto \partial_C p_\xi(w)$ is measurable for all $w \in \mathscr{F}_B$. By [21, Prop. 2.1.5], for all $\xi \in \Xi$, $\partial_C p_\xi$ has a weakly closed graph. Now let $D \subset B$ be a compact set and denote by $L_{\hat{D}} \in L^1(\Xi)$ a function which fulfills property 3 of Assumption 7.3.1 for a neighborhood $\hat{D}$ of D. Define by $G_D : \Xi \rightrightarrows B^*$ the multifunction $G_D(\xi) := \text{w-cl}\,\text{co}(\partial_C p_\xi(D))$. First note that, since D is bounded, [21, Prop. 2.1.2] implies that $\partial_C p_\xi(D)$ is bounded by $L_{\hat{D}}(\xi)$ P-almost everywhere. This shows that the multifunction G_D is integrably bounded and, for fixed $\xi \in \Xi$, the set $G_D(\xi)$ is bounded. Consequently, by Alaoglu's theorem, $G_D(\xi)$ is weakly compact, and obviously nonempty and convex. As $\partial_C p_\xi(w) \subset G_D(\xi)$ P-a.e. and for all $w \in D$, [105, Thm. 4.2] yields that $w \mapsto \tilde{G}(w)$ is weakly upper semicontinuous, i.e., for every weakly closed set $C \subset Y$ the set $G^-(C) := \{x \in \mathscr{F} : G(r) \cap C \neq \emptyset\}$ is closed in $\mathscr{F}$. By [104, Cor. 3.1], the multifunction $\tilde{G}$ is weakly closed valued. Therefore, [60, Thm. 2.5] implies that $G : \bar{\mathscr{F}} \rightrightarrows B^*$ has a weakly closed graph. □

Example 7.3.3 (Tracking type objective function). For all $\xi \in \Xi$, let $J_\xi : Z \to \mathbb{R}$ be given via $J_\xi(\cdot) := \frac{1}{2}\|O_\xi(\cdot) - y_\xi^d\|_H^2$, where $O_\xi \in \mathscr{L}(Z,H)$ is the stochastic observation operator, $y_\xi^d \in H$ is the stochastic desired state and H is a Hilbert space. Let the assumptions of Theorem 7.1.3 hold and let $\xi \mapsto \|O_\xi\|_{\mathscr{L}(Z,H)}$ be in $L^\infty(\Xi)$ and $\xi \mapsto y_\xi^d$ be in $\mathbf{H}^*$. Further, let $\mathring{\mathbf{f}} \in \mathbf{H}^*$ be a stochastic external force, define $\hat{\boldsymbol{\iota}} \in \mathscr{L}(Z^*, \mathbf{H}^*)$ via $(\hat{\boldsymbol{\iota}} w)(\xi) := w$ and set $\mathring{\mathbf{F}} := \hat{\boldsymbol{\iota}}(\cdot) + \mathring{\mathbf{f}}$.

COROLLARY 7.3.4. *In the situation of Example 7.3.3, the multifunction*

$$G : H^{-1}(\Omega) \rightrightarrows H_0^1(\Omega), \qquad G(w) := \mathbb{E}\left[\partial_w J_\xi(S_\xi(\mathring{F}_\xi(w)))\right] \tag{7.3.4}$$

fulfills Assumption 3.1.1 and thus can be used in the bundle method as a subdifferential.

Proof. Denote by $C_L > 0$ the constant of uniform coercivity of $(A_\xi)_{\xi\in\Xi}$. By Theorem 7.1.3, the solution operators $S_\xi : Z^* \to Z$ are Lipschitz with modulus $1/C_L$ and $\xi \mapsto S_\xi(\mathring{F}_\xi(w))$ is in $\mathbf{H}$ for all $w \in Z^*$. We verify Assumption 7.3.1 with $B := H_0^1(\Omega) = Z$, $\mathscr{F}_B = B$ and $p_\xi(w) := J_\xi(S_\xi(\mathring{F}_\xi(w)))$.

1. Let $w \in \mathscr{F}_B = H_0^1(\Omega)$ be fixed. Since $(\xi, x) \mapsto J_\xi(x)$ is a Cathathéodory mapping and $\xi \mapsto S_\xi(\mathring{F}_\xi(w))$ is measurable, [8, Lem. 8.2.3] implies that $\xi \mapsto J_\xi(S_\xi(\mathring{F}_\xi(w)))$ is measurable.

2. First note that $\mathbf{S}(\mathring{f}) \in \mathbf{H} = L^2(\Xi, Z)$, i.e., $\xi \mapsto \|S_\xi(\mathring{f})\|_Z$ is in $L^2(\Xi)$. Therefore,

$$\begin{aligned}
\sqrt{2\int_\Xi |p_\xi(0)|\,dP(\xi)} &= \sqrt{2\int_\Xi |J_\xi(S_\xi(\mathring{f}))|\,dP(\xi)} \\
&= \|\xi \mapsto \|O_\xi S_\xi(\mathring{f}) - y_\xi^d\|_H\|_{L^2(\Xi)} \\
&\leq \|\xi \mapsto \|O_\xi\|_{\mathscr{L}(Z,H)}\|_{L^\infty(\Xi)} \|\xi \mapsto \|S_\xi(\mathring{f})\|_Z\|_{L^2(\Xi)} + \|\xi \mapsto \|y_\xi^d\|_H\|_{L^2(\Xi)} < \infty.
\end{aligned}$$

3. For each bounded set $D \subset Z^*$ there exists a function $\hat{L}_D \in L^2(\Xi)$ such that for all $w \in D$ and P-a.a. $\xi \in \Xi$ it holds

$$\|S_\xi(\mathring{F}_\xi(w))\|_Z \leq \|S_\xi(w + \mathring{f}) - S_\xi(\mathring{f})\|_Z + \|S_\xi(\mathring{f})\|_Z \leq \tfrac{1}{C_L}\|w\|_{Z^*} + \|S_\xi(\mathring{f})\|_Z \leq \hat{L}_D(\xi).$$

For arbitrary $w_1, w_2 \in D$ and all $\xi \in \Xi$ we get the estimate

$$\begin{aligned}
2|p_\xi(w_1) - p_\xi(w_2)| &\leq \|O_\xi((S_\xi(w_1 + \mathring{f}) + S_\xi(w_2 + \mathring{f})) - 2y_\xi^d\|_H \|O_\xi(S_\xi(w_1 + \mathring{f}) - S_\xi(w_2 + \mathring{f}))\|_H \\
&\leq 2(\|O_\xi\|_{\mathscr{L}(Z,H)} C_D(\xi) + \|y_\xi^d\|_H)\|O_\xi\|_{\mathscr{L}(Z,H)} \|S_\xi(w_1 + \mathring{f}) - S_\xi(w_2 + \mathring{f})\|_Z \\
&\leq L_D(\xi)\|w_1 - w_2\|_{Z^*},
\end{aligned}$$

where $L_D \in L^2(\Xi)$ is defined as $L_D(\xi) := \frac{2}{C_L}(\|O_\xi\|_{\mathscr{L}(Z,H)}\hat{L}_D(\xi) + \|y_\xi^d\|_H)\|O_\xi\|_{\mathscr{L}(Z,H)}$. □

Acknowledgements

The completion of this dissertation would not have been possible without the help of various people.

First and foremost I would like to thank my supervisor Michael Ulbrich for his continuous support, advice and invaluable input. Starting with my Bachelor thesis, he gave me the opportunity to grow as a mathematician not only by assigning me increasingly more complex tasks but also by teaching me mathematical rigor. Many thanks for giving me enough space to create and pursue own ideas during the creation of this dissertation.

I thank the priority program SPP 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" for funding. The program enabled me to meet doctoral students and professors working in the same field and to exchange ideas on various international conferences and workshops.

I am grateful to Anne-Therese Rauls and Stefan Ulbrich for our collaboration. Their characterization of elements of the subdifferential of the reduced objective function fits perfectly into the setting of the bundle method and enriches this dissertation.

Many thanks to all my colleagues in the chair of Michael Ulbrich for pleasant coffee breaks as well as providing a constructive working environment. I am especially thankful to Johannes Haubner, Simon Plazotta, Johannes Milz and Sebastian Garreis for fruitful discussions and leisure activities. In addition, I would like to thank Frauke Bäcker for her organizational work.

I thank all my family for their endorsement and encouragement. My mother's support helped me to focus on my dissertation, especially in the last two years. I am grateful to my father for introducing me into computer programming. Furthermore, I wish to thank Karin Rettenmaier for her patience when I explained all my mathematical insights to her.

A. Complexification of a real Hilbert space

Let H be a real Hilbert space. A linear space $H \times H$ over the field $\mathbb{C}$ with the rule of external multiplication by complex numbers $(\alpha + i\beta)(x,y) := (\alpha x - \beta y, \alpha x + \beta y)$, $\alpha, \beta \in \mathbb{R}$, $(x,y) \in H \times H$ is called *complexification* of the real Hilbert space H and is denoted by $H^{\mathbb{C}}$, cf. [102, 129]. It is convenient to write the elements in $H^{\mathbb{C}}$ as $x + iy$, where $x, y \in H$ and i is the imaginary unit. The vector space $H^{\mathbb{C}}$ is a complex Hilbert space with respect to the scalar product

$$(x_1 + ix_2, y_1 + iy_2)_{H^{\mathbb{C}}} := (x_1, y_1)_H + (x_2, y_2)_H + i(x_2, y_1)_H - i(x_1, y_2)_H$$

and $\|x + iy\|^2_{H^{\mathbb{C}}} = \|x\|^2_H + \|y\|^2_H$, $\|x + i0\|_{H^{\mathbb{C}}} = \|x\|_H$ for all $x, y \in H$. We identify H with the subspace $H \times \{0\}$ of $H^{\mathbb{C}}$. Let H_1, H_2 be two real Hilbert spaces. We define the complexification of an operator $A \in \mathscr{L}(H_1, H_2)$ to be $A^{\mathbb{C}} : H_1^{\mathbb{C}} \to H_2^{\mathbb{C}}$, $A^{\mathbb{C}}(x + iy) := Ax + iAy$. Note that $A^{\mathbb{C}}$ is $\mathbb{C}$-linear and bounded, i.e., $A^{\mathbb{C}} \in \mathscr{L}(H_1^{\mathbb{C}}, H_2^{\mathbb{C}})$.

DEFINITION A.1 (e.g., [129, Def. 4.1.1]). *Let H be a Hilbert space over the field $\mathbb{F} \in \{\mathbb{C}, \mathbb{R}\}$. The spectrum and the point spectrum of an operator $A \in \mathscr{L}(H)$ are defined to be the sets*

$$\begin{aligned} \sigma(A) &:= \{\lambda \in \mathbb{F} : \lambda \operatorname{Id}_H - A \text{ is not invertible in } \mathscr{L}(H)\}, \\ \sigma_p(A) &:= \{\lambda \in \mathbb{F} : \lambda \operatorname{Id}_H - A \text{ is not injective }\}. \end{aligned}$$

LEMMA A.2. *Let H be a real Hilbert space and $A \in \mathscr{L}(H)$ be a (Hilbert space) self-adjoint operator. Then $\sigma_p(A) = \sigma_p(A^{\mathbb{C}})$.*

Proof. Let $\lambda \in \sigma_p(A^{\mathbb{C}})$. As A is self-adjoint, so is $A^{\mathbb{C}}$. [129, Thm. 4.4.2] thus implies $\lambda \in \mathbb{R}$. Further, $\lambda \operatorname{Id}_{H^{\mathbb{C}}} - A^{\mathbb{C}}$ is not injective, i.e., there is a nonzero vector $x \in H^{\mathbb{C}}$ such that $A^{\mathbb{C}} x = \lambda x$ and $x = x_1 + ix_2$, $x_1, x_1 \in H$. Therefore, $Ax_1 + iAx_2 = A^{\mathbb{C}} x = \lambda x_1 + i\lambda x_2$. This shows that $Ax_1 = \lambda x_1$, i.e., $\lambda \in \sigma_p(A)$. Now let $\lambda \in \sigma_p(A)$. Then there is a nonzero vector $x \in H$ such that $Ax = \lambda x$. Consequently, $A^{\mathbb{C}}(x + i0) = Ax = \lambda(x + i0)$ and $\lambda \in \sigma_p(A^{\mathbb{C}})$. □

THEOREM A.3. *Let H be a real Hilbert space. If $Q \in \mathscr{L}(H)$ is given by $Q = \mu \operatorname{Id}_H + UV$ with $\mu \geq 0$, $U \in \mathscr{L}(\mathbb{R}^n, H)$ and $V \in \mathscr{L}(H, \mathbb{R}^n)$, $n \in \mathbb{N}_+$, such that Q is (Hilbert space) self-adjoint, then*

$$(Qv, v)_H \geq \Big(\mu + \min \sigma(VU) \cup \{0\}\Big) \|v\|^2_H \qquad \textit{for all } v \in H$$

and $\|Q\|_{\mathscr{L}(H)} \leq \mu + \max |\sigma(VU)|$.

Proof. Denote by $H^{\mathbb{C}}$ and $Q^{\mathbb{C}}$ the complexifications of H and Q. It holds $Q^{\mathbb{C}}(x_1 + ix_2) = \mu x_1 + i\mu x_2 + UVx_1 + iUVx_2 = (\mu \operatorname{Id}_{H^{\mathbb{C}}} + U^{\mathbb{C}} V^{\mathbb{C}})(x_1 + ix_2)$ for arbitrary $x_1, x_2 \in H$. As $Q \in \mathscr{L}(H)$ is self-adjoint, so is $Q^{\mathbb{C}} \in \mathscr{L}(H^{\mathbb{C}})$. It holds

$$\frac{(Qv,v)_H}{\|v\|_H^2} = \frac{(Q^{\mathbb{C}}v,v)_{H^{\mathbb{C}}}}{\|v\|_{H^{\mathbb{C}}}^2} \geq \inf_{x\neq 0} \frac{(Q^{\mathbb{C}}x,x)_{H^{\mathbb{C}}}}{\|x\|_{H^{\mathbb{C}}}^2} = \inf_{\|x\|_{H^{\mathbb{C}}}=1} (Q^{\mathbb{C}}x,x)_{H^{\mathbb{C}}} \quad \text{for all } v \in H \setminus \{0\}. \tag{A.1}$$

Since the operator $Q^{\mathbb{C}} \in \mathscr{L}(H^{\mathbb{C}})$ is (Hilbert space) self-adjoint, [129, Thm. 4.4.4] and [129, Thm. 4.4.6] can be applied which yield that $\sigma(Q^{\mathbb{C}}) \subset \mathbb{R}$ and

$$\inf_{\|x\|_{H^{\mathbb{C}}}=1} (Q^{\mathbb{C}}x,x)_{H^{\mathbb{C}}} = \min \sigma(Q^{\mathbb{C}}). \tag{A.2}$$

Using [129, Thm. 4.3.1], we calculate

$$\sigma(Q^{\mathbb{C}}) = \sigma(\mu \operatorname{Id}_{H^{\mathbb{C}}} + U^{\mathbb{C}}V^{\mathbb{C}}) = \mu + \sigma(U^{\mathbb{C}}V^{\mathbb{C}}). \tag{A.3}$$

Now, [11, Thm. 3] yields

$$\sigma(U^{\mathbb{C}}V^{\mathbb{C}}) \setminus \{0\} = \sigma(V^{\mathbb{C}}U^{\mathbb{C}}) \setminus \{0\}. \tag{A.4}$$

Consequently,

$$\sigma(U^{\mathbb{C}}V^{\mathbb{C}}) \subset \sigma(U^{\mathbb{C}}V^{\mathbb{C}}) \cup \{0\} = \sigma(V^{\mathbb{C}}U^{\mathbb{C}}) \cup \{0\}. \tag{A.5}$$

Since $V^{\mathbb{C}}U^{\mathbb{C}} = (VU)^{\mathbb{C}} \in \mathscr{L}((\mathbb{R}^n)^{\mathbb{C}})$ and $(\mathbb{R}^n)^{\mathbb{C}}$ is finite dimensional, [129, Rem. 4.1.4 (ii)] and Lemma A.2 yields that

$$\sigma(V^{\mathbb{C}}U^{\mathbb{C}}) = \sigma((VU)^{\mathbb{C}}) = \sigma_p((VU)^{\mathbb{C}}) = \sigma_p(VU) = \sigma(VU). \tag{A.6}$$

Combing (A.1)–(A.3), (A.5), and (A.6) we find for arbitrary $v \in H$ that

$$(Q^{\mathbb{C}}v,v)_{H^{\mathbb{C}}} \geq \Big(\mu + \min \sigma(U^{\mathbb{C}}V^{\mathbb{C}})\Big)\|v\|_{H^{\mathbb{C}}}^2 \geq \Big(\mu + \min \sigma(VU) \cup \{0\}\Big)\|v\|_{H^{\mathbb{C}}}^2.$$

In order to verify the second result, we argue as follows. [129, Thm. 4.4.5] and (A.3) yield

$$\|Q\|_{\mathscr{L}(H)} \leq \|Q^{\mathbb{C}}\|_{\mathscr{L}(H^{\mathbb{C}})} = \max|\sigma(Q^{\mathbb{C}})| \leq \mu + \max|\sigma(U^{\mathbb{C}}V^{\mathbb{C}})|$$

and (A.4) gives

$$\max|\sigma(U^{\mathbb{C}}V^{\mathbb{C}})| \leq \max|(\sigma(U^{\mathbb{C}}V^{\mathbb{C}}) \setminus \{0\}) \cup \{0\}| = \max|(\sigma(V^{\mathbb{C}}U^{\mathbb{C}}) \setminus \{0\}) \cup \{0\}|.$$

By [129, Thm. 4.4.3], $\sigma(V^{\mathbb{C}}U^{\mathbb{C}}) \neq \emptyset$. Together with (A.6), this yields

$$\max|(\sigma(V^{\mathbb{C}}U^{\mathbb{C}}) \setminus \{0\}) \cup \{0\}| = \max|\sigma(V^{\mathbb{C}}U^{\mathbb{C}})| = \max|\sigma(VU)|. \qquad \square$$

Bibliography

[1] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev spaces*, New York, NY, Academic Press, 2nd ed., 2003.

[2] M. AINSWORTH AND J. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, A Wiley-Interscience publication, Wiley, 2000.

[3] H. W. ALT, *Linear functional analysis*, Springer, London, 2016.

[4] T. APEL, A. RÖSCH, AND G. WINKLER, *Optimal control in non-convex domains: a priori discretization error estimates*, Calcolo, 44 (2007), pp. 137–158.

[5] P. APKARIAN, D. NOLL, AND O. PROT, *A proximity control algorithm to minimize nonsmooth and nonconvex semi-infinite maximum eigenvalue functions*, J. Convex Anal., 16 (2009), pp. 641–666.

[6] W. ARENDT, I. CHALENDAR, AND R. EYMARD, *Galerkin approximation of linear problems in Banach and Hilbert spaces*, IMA Journal of Numerical Analysis, (2020).

[7] H. ATTOUCH, G. BUTTAZZO, AND G. MICHAILLE, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, SIAM Ser. Optim., 2014.

[8] J.-P. AUBIN AND H. FRANKOWSKA, *Set-valued analysis*, Birkhäuser, Boston, MA, 2009.

[9] W. BANGERTH AND R. RANNACHER, *Adaptive finite element methods for differential equations*, Birkhäuser, Basel, 2003.

[10] A. BARBU AND S.-C. ZHU, *Monte Carlo methods*, Springer, Singapore, 2020.

[11] B. A. BARNES, *Common operator properties of the linear operators RS and SR*, Proc. Am. Math. Soc., 126 (1998), pp. 1055–1061.

[12] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, New York, NY, 2011.

[13] D. BRAESS, C. CARSTENSEN, AND R. H. W. HOPPE, *Error reduction in adaptive finite element approximations of elliptic obstacle problems*, J. Comput. Math., 27 (2009), pp. 148–169.

[14] S. BRENNER AND R. SCOTT, *The mathematical theory of finite element methods*, vol. 15, Springer Science & Business Media, 2007.

[15] H. BREZIS, *Functional analysis, Sobolev spaces and partial differential equations*, Springer, New York, NY, 2011.

[16] F. BREZZI, W. W. HAGER, AND P. A. RAVIART, *Error estimates for the finite element solution of variational inequalities. Part I. primal theory*, Numer. Math., 28 (1977), pp. 431–443.

[17] F. BREZZI, W. W. HAGER, AND P. A. RAVIART, *Error estimates for the finite element solution of variational inequalities. Part II. Mixed methods*, Numer. Math., 31 (1978), pp. 1–16.

[18] I. CHIŢESCU, R.-C. SFETCU, AND O. COJOCARU, *Köthe-Bochner spaces that are Hilbert spaces*, Carpathian J. Math., 33 (2017), pp. 161–168.

[19] G. CHOQUET, *Topology*, Academic Press, New York, 1966.

[20] P. CIARLET, *The Finite Element Method for Elliptic Problems*, Society for Industrial and Applied Mathematics (SIAM), 2002.

[21] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Montreal, Canada, 1998.

[22] R. CORREA AND C. LEMARÉCHAL, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.

[23] G. DAHLQUIST AND Å. BJÖRCK, *Numerical methods in scientific computing. Vol. 1*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[24] A. DANIILIDIS AND P. GEORGIEV, *Approximate convexity and submonotonicity*, J. Math. Anal. Appl., 291 (2004), pp. 292–301.

[25] W. DE OLIVEIRA, C. SAGASTIZÁBAL, AND C. LEMARÉCHAL, *Convex proximal bundle methods in depth: a unified analysis for inexact oracles*, Math. Program., 148 (2014), pp. 241–277.

[26] W. DE OLIVEIRA AND C. A. SAGASTIZÁBAL, *Bundle methods in the xxist century: A bird's-eye view*, Pesquisa Operacional, 34 (2014), pp. 647–670.

[27] M. C. DELFOUR AND J.-P. ZOLÉSIO, *Shapes and geometries. Metrics, analysis, differential calculus, and optimization*, vol. 22, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2nd ed., 2011.

[28] C. Y. DENG, *A generalization of the Sherman-Morrison-Woodbury formula*, Applied Mathematics Letters, 24 (2011), pp. 1561 – 1564.

[29] D. A. DUNAVANT, *High degree efficient symmetrical Gaussian quadrature rules for the triangle*, Int. J. Numer. Methods Eng., 21 (1985), pp. 1129–1148.

[30] D. E. EDMUNDS, A. KUFNER, AND J. RÁKOSNÍK, *Embeddings of Sobolev spaces with weights of power type*, Z. Anal. Anwend., 4 (1985), pp. 25–34.

[31] L. EVANS, *Measure Theory and Fine Properties of Functions*, CRC Press, 2018.

[32] L. C. EVANS, *Partial differential equations*, American Mathematical Society (AMS), Providence, RI, 2010.

[33] R. FALK, *Error estimates for the approximation of a class of variational inequalities*, Math. Comp., 28 (1974), pp. 963–971.

[34] S. FUNKEN, D. PRAETORIUS, AND P. WISSGOTT, *Efficient implementation of adaptive P1-FEM in Matlab*, Comput. Methods Appl. Math., 11 (2011), pp. 460–490.

[35] A. GAEVSKAYA, M. HINTERMÜLLER, R. H. W. HOPPE, AND C. LÖBHARD, *Adaptive finite elements for optimally controlled elliptic variational inequalities of obstacle type*, in Optimization with PDE constraints. ESF networking program 'OPTPDE', Springer, Cham, 2014, pp. 95–150.

[36] S. GARREIS AND M. ULBRICH, *Constrained optimization with low-rank tensors and applications to parametric problems with PDEs*, SIAM J. Sci. Comput., (2017), pp. A25–A54.

[37] S. A. GERSHGORIN, *Über die Abgrenzung der Eigenwerte einer Matrix*, Bull. Acad. Sci. URSS, 1931 (1931), pp. 749–754.

[38] F. GIANNESSI, A. MAUGERI, AND P. M. PARDALOS, *Equilibrium problems: nonsmooth optimization and variational inequality models*, vol. 58, Springer, New York, NY, 2001.

[39] R. GLOWINSKI, *Numerical methods for nonlinear variational problems. 2nd printing*, Springer, Berlin, 2nd printing ed., 2008.

[40] R. GLOWINSKI, J.-L. LIONS, AND R. TREMOLIERES, *Numerical analysis of variational inequalities. Transl. and rev. ed.* Studies in Mathematics and its Applications, Vol. 8. Amsterdam, New York, Oxford: North-Holland Publishing Company. XXIX, 776 p. $ 109.75; Dfl. 225.00 (1981), 1981.

[41] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Society for Industrial and Applied Mathematics, 2011.

[42] A. GÜNNEL, R. HERZOG, AND E. SACHS, *A note on preconditioners and scalar products in Krylov subspace methods for self-adjoint problems in Hilbert space*, ETNA, Electron. Trans. Numer. Anal., 41 (2014), pp. 13–20.

[43] M. D. GUNZBURGER, H.-C. LEE, AND J. LEE, *Error estimates of stochastic optimal neumann boundary control problems*, SIAM J. Numer. Anal., (2011), pp. 1532–1552.

[44] J. GWINNER, *A note on random variational inequalities and simple random unilateral boundary value problems: Well-posedness and stability results*, in Advances in convex analysis and global optimization. Honoring the memory of C. Carathéodory (1873-1950), Kluwer Academic Publishers, Dordrecht, 2001, pp. 531–543.

[45] J. GWINNER AND F. RACITI, *On a class of random variational inequalities on random sets*, Numer. Funct. Anal. Optim., 27 (2006), pp. 619–636.

[46] N. HAARALA, K. MIETTINEN, AND M. M. MÄKELÄ, *Globally convergent limited memory bundle method for large-scale nonsmooth optimization*, Math. Program., 109 (2007), pp. 181–205.

[47] W. HARE AND C. SAGASTIZÁBAL, *A redistributed proximal bundle method for nonconvex optimization*, SIAM J. Optim., 20 (2010), pp. 2442–2473.

[48] W. HARE, C. SAGASTIZÁBAL, AND M. SOLODOV, *A proximal bundle method for nonsmooth nonconvex functions with inexact information*, Comput. Optim. Appl., 63 (2016), pp. 1–28.

[49] L. HERTLEIN, A.-T. RAULS, M. ULBRICH, AND S. ULBRICH, *An inexact bundle method and subgradient computations for optimal control of deterministic and stochastic obstacle problems*. Priprint, accepted for publication in SPP1962 Special Issue, Birkhäuser, 2019.

[50] L. HERTLEIN AND M. ULBRICH, *An inexact bundle algorithm for nonconvex nonsmooth minimization in Hilbert space*, SIAM J. Control Optim., 57 (2019), pp. 3137–3165.

[51] R. HERZOG AND W. WOLLNER, *A conjugate direction method for linear systems in Banach spaces*, J. Inverse Ill-Posed Probl., 25 (2017), pp. 553–572.

[52] N. J. HIGHAM, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2nd ed., 2002.

[53] M. HINTERMÜLLER, *A proximal bundle method based on approximate subgradients*, Comput. Optim. Appl., 20 (2001), pp. 245–266.

[54] M. HINTERMÜLLER, R. H. W. HOPPE, AND C. LÖBHARD, *Dual-weighted goal-oriented adaptive finite elements for optimal control of elliptic variational inequalities*, ESAIM, Control Optim. Calc. Var., 20 (2014), pp. 524–546.

[55] M. HINTERMÜLLER AND I. KOPACKA, *Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm*, SIAM J. Optim., 20 (2009), pp. 868–902.

[56] M. HINTERMÜLLER AND I. KOPACKA, *A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs*, Comput. Optim. Appl., 50 (2011), pp. 111–145.

[57] M. HINTERMÜLLER, A. LAURAIN, C. LÖBHARD, C. N. RAUTENBERG, AND T. M. SUROWIEC, *Elliptic mathematical programs with equilibrium constraints in function space: optimality conditions and numerical realization*, in Trends in PDE constrained optimization, Birkhäuser/Springer, Cham, 2014, pp. 133–153.

[58] M. HINTERMÜLLER AND T. SUROWIEC, *A bundle-free implicit programming approach for a class of elliptic MPECs in function space*, Math. Program., 160 (2016), pp. 271–305.

[59] M. HINZE, R. PINNAU, M. ULBRICH, AND S. ULBRICH, *Optimization with PDE constraints*, Springer, Dordrecht, 2009.

[60] S. H. HOU, *On property* (Q) *and other semicontinuity properties of multifunctions*, Pacific J. Math., 103 (1982), pp. 39–56.

[61] P. HOUSTON AND T. P. WIHLER, *An adaptive variable order quadrature strategy*, in Spectral and high order methods for partial differential equations, ICOSAHOM 2016. Selected papers from the ICOSAHOM conference, June 27 – July 1, 2016, Rio de Janeiro, Brazil, Springer, Cham, 2017, pp. 533–545.

[62] T. HYTÖNEN, J. VAN NEERVEN, M. VERAAR, AND L. WEIS, *Analysis in Banach Spaces: Volume I: Martingales and Littlewood-Paley Theory*, Springer International Publishing, 2016.

[63] K. ITO AND K. KUNISCH, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.

[64] A. JOFRÉ, D. T. LUC, AND M. THÉRA, *ε-subdifferential and ε-monotonicity*, Nonlinear Anal., Theory Methods Appl., 33 (1998), pp. 71–90.

[65] N. KARMITSA AND M. M. MÄKELÄ, *Limited memory bundle method for large bound constrained nonsmooth optimization: convergence analysis*, Optim. Methods Softw., 25 (2010), pp. 895–916.

[66] N. KIKUCHI AND J. T. ODEN, *Contact problems in elasticity: A study of variational inequalities and finite element methods*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1988.

[67] D. KINDERLEHRER AND G. STAMPACCHIA, *An introduction to variational inequalities and their applications*, Classics Appl. Math., SIAM, Philadelphia, PA, 2000.

[68] K. KIWIEL, *A dual method for certain positive semidefinite quadratic programming problems*, Siam J. Sci. Stat. Comput., 10 (1989), pp. 175 – 186.

[69] K. C. KIWIEL, *A linearization algorithm for nonsmooth minimization*, Math. Oper. Res., 10 (1985), pp. 185–194.

[70] K. C. KIWIEL, *Approximations in proximal bundle methods and decomposition of convex programs*, J. Optim. Theory Appl., 84 (1995), pp. 529–548.

[71] K. C. KIWIEL, *Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization*, SIAM J. Optim., 6 (1996), pp. 227–249.

[72] K. C. KIWIEL, *A proximal bundle method with approximate subgradient linearizations*, SIAM J. Optim., 16 (2006), pp. 1007–1023.

[73] D. KLATTE AND B. KUMMER, *Nonsmooth equations in optimization. Regularity, calculus, methods and applications*, vol. 60, Kluwer Academic Publishers, Dordrecht, 2002.

[74] D. P. KOURI, M. HEINKENSCHLOSS, D. RIDZAL, AND V. B. B. G., *Inexact objective function evaluation in a trust-region algorithm for PDE-constrained optimization under uncertainty*, SIAM J. Sci. Comput., (2014), pp. A3011–A3029.

[75] K. KUNISCH AND D. WACHSMUTH, *Sufficient optimality conditions and semi-smooth newton methods for optimal control of stationary variational inequalities*, ESAIM: Control, Optim. Calc. Var., (2012), pp. 520–547.

[76] N. KUZNETSOV AND A. NAZAROV, *Sharp constants in the Poincaré, Steklov and related inequalities (a survey)*, Mathematika, 61 (2015), pp. 328–344.

[77] C. LEMARÉCHAL, *An extension of Davidon methods to non differentiable problems*. Nondiffer. Optim., Math. Program. Study 3, 95-109, 1975.

[78] C. LEMARÉCHAL, J.-J. STRODIOT, AND A. BIHAIN, *On a bundle algorithm for nonsmooth optimization*, in Nonlinear Programming 4 (Madison, WI, 1980), Academic Press, New York, 1981, pp. 245–282.

[79] A. S. LEWIS AND M. L. OVERTON, *Nonsmooth optimization via quasi-Newton methods*, Math. Program., 141 (2013), pp. 135–163.

[80] G.-H. LIN, X. CHEN, AND M. FUKUSHIMA, *Solving stochastic mathematical programs with equilibrium constraints via approximation and smoothing implicit programming with penalization*, Math. Progam., Ser. B, (2009), pp. 343–368.

[81] C. LÖBHARD, *Optimal Control of Elliptic Variational Inequalities*, PhD thesis, Humboldt-Universität zu Berlin, 2014.

[82] L. LUKŠAN AND J. VLČEK, *A bundle-Newton method for nonsmooth unconstrained minimization*, Math. Programming, 83 (1998), pp. 373–391.

[83] J. LV, L.-P. PANG, AND F.-Y. MENG, *A proximal bundle method for constrained nonsmooth nonconvex optimization with inexact information*, J. Glob. Optim., 70 (2018), pp. 517–549.

[84] M. M. MÄKELÄ, *Survey of bundle methods for nonsmooth optimization*, Optim. Methods Softw., 17 (2002), pp. 1–29.

[85] M. M. MÄKELÄ AND P. NEITTAANMÄKI, *Nonsmooth optimization: Analysis and algorithms with applications to optimal control*, World Scientific Publishing, River Edge, NJ, 1992.

[86] J. MALICK, W. DE OLIVEIRA, AND S. ZAOURAR, *Uncontrolled inexact information within bundle methods*, EURO J. Comput. Optim., 5 (2017), pp. 5–29.

[87] R. E. MEGGINSON, *An introduction to Banach space theory*, vol. 183, Springer, New York, NY, 1998.

[88] C. MEYER AND O. THOMA, *A priori finite element error analysis for optimal control of the obstacle problem*, SIAM J. Numer. Anal., 51 (2013), pp. 605–628.

[89] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[90] R. MIFFLIN, *A modification and extension of Lemaréchal's algorithm for nonsmooth minimization*, Math. Programming Stud., (1982), pp. 77–90. Nondifferential and variational techniques in optimization (Lexington, Ky., 1980).

[91] F. MIGNOT, *Contrôle dans les inéquations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.

[92] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.

[93] W. B. MOORS, *A characterisation of minimal subdifferential mappings of locally Lipschitz functions*, Set-Valued Anal., 3 (1995), pp. 129–141.

[94] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[95] P. NEITTAANMÄKI AND S. REPIN, *Reliable methods for computer simulation. Error control and a posteriori estimates*, Elsevier, Amsterdam, 2004.

[96] J. NOCEDAL, *Updating quasi-Newton matrices with limited storage*, Math. Comput., 35 (1980), pp. 773–782.

[97] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer, New York, USA, second ed., 2006.

[98] R. H. NOCHETTO, K. G. SIEBERT, AND A. VEESER, *Fully located a posteriori error estimators and barrier sets for contact problems*, SIAM J. Numer. Anal., (2005), pp. 2118–2135.

[99] D. NOLL, *Bundle method for non-convex minimization with inexact subgradients and function values*, in Computational and Analytical Mathematics, D. Bailey, H. Bauschke, et al., eds., Springer, 2013, pp. 555–592.

[100] D. NOLL, O. PROT, AND A. RONDEPIERRE, *A proximity control algorithm to minimize nonsmooth and nonconvex functions*, Pac. J. Optim., 4 (2008), pp. 571–604.

[101] *OPTPDE — a collection of problems in PDE-constrained optimization*. `http://www.optpde.net`.

[102] M. N. ORESHINA, *Spectral decomposition of normal operator in real Hilbert space*, Ufim. Mat. Zh., 9 (2017), pp. 87–99.

[103] J. V. OUTRATA, *Mathematical programs with equilibrium constraints: theory and numerical methods*, in Nonsmooth mechanics of solids. Papers based on the presentations at the advanced school, Udine, Italy, October 4–8, 2004, Springer, Wien, 2006, pp. 221–274.

[104] N. S. PAPAGEORGIOU, *On the theory of Banach space valued multifunctions. I. Integration and conditional expectation*, J. Multivariate Anal., 17 (1985), pp. 185–206.

[105] N. S. PAPAGEORGIOU, *On Fatou's lemma and parametric integrals for set-valued functions*, Proc. Indian Acad. Sci. Math. Sci., 103 (1993), pp. 181–195.

[106] D. PAULY AND J. VALDMAN, *Poincaré-Friedrichs type constants for operators involving grad, curl, and div: theory and numerical experiments*, Comput. Math. Appl., 79 (2020), pp. 3027–3067.

[107] J.-P. PENOT, *Favorable classes of mappings and multimappings in nonlinear analysis and optimization*, J. Convex Anal., 3 (1996), pp. 97–116.

[108] W. H. PRESS, S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P. FLANNERY, *Numerical recipes. The art of scientific computing*, Cambridge University Press, Cambridge, 3rd ed., 2007.

[109] A.-T. RAULS, *Generalized Derivatives for Solution Operators of Variational Inequalities of Obstacle Type*, PhD thesis, Universität Darmstadt, 2021.

[110] A.-T. RAULS AND S. ULBRICH, *Computation of a Bouligand generalized derivative for the solution operator of the obstacle problem*, SIAM J. Control Optim., 57 (2019), pp. 3223–3248.

[111] S. REPIN, *A posteriori estimates for partial differential equations*, vol. 4, de Gruyter, Berlin, 2008.

[112] S. REPIN, S. SAUTER, AND A. SMOLIANSKI, *A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions*, Computing, 70 (2003), pp. 205–233.

[113] C. ROBERT, *cbrewer : colorbrewer schemes for Matlab*. https://www.mathworks.com/matlabcentral/fileexchange/34087-cbrewer-colorbrewer-schemes-for-matlab, 2020. MATLAB Central File Exchange. Retrieved September 19, 2020.

[114] R. T. ROCKAFELLAR, *Favorable classes of Lipschitz-continuous functions in subgradient optimization*, in Progress in Nondifferentiable Optimization, IIASA Collab. Proc. Ser. CP-82, vol. 8, Internat. Inst. Appl. Systems Anal., Laxenburg, Austria, 1982, pp. 125–143.

[115] J.-F. RODRIGUES, *Obstacle Problems in Mathematical Physics*, vol. 134, North-Holland, Amsterdam, 1987.

[116] H. SCHEEL AND S. SCHOLTES, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., (2000), pp. 1–22.

[117] A. SCHIELA AND D. WACHSMUTH, *Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints*, ESAIM Math. Model. Numer. Anal., (2013), pp. 771–787.

[118] A. SHAPIRO AND H. XU, *Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation*, Optimization, (2008), pp. 295–418.

[119] J. SHEN, X.-Q. LIU, F.-F. GUO, AND S.-X. WANG, *An approximate redistributed proximal bundle method with inexact data for minimizing nonsmooth nonconvex functions*, Math. Probl. Eng., (2015), p. 9. Id/No 215310.

[120] M. SOFONEA AND A. MATEI, *Variational inequalities with applications. A study of antiplane frictional contact problems*, vol. 18, Springer, New York, NY, 2009.

[121] M. V. SOLODOV, *On approximations with finite precision in bundle methods for nonsmooth optimization*, J. Optim. Theory Appl., 119 (2003), pp. 151–165.

[122] P. SONNEVELD, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52.

[123] R. T. TRÎMBIŢAŞ, *Adaptive cubatures on triangle*, Result. Math., 53 (2009), pp. 453–462.

[124] M. ULBRICH, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, vol. 11, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[125] M. ULBRICH AND S. ULBRICH, *Nichtlineare Optimierung*, Birkhäuser, Basel, 2012.

[126] W. VAN ACKOOIJ, J. Y. BELLO CRUZ, AND W. DE OLIVEIRA, *A strongly convergent proximal bundle method for convex minimization in Hilbert spaces*, Optimization, 65 (2016), pp. 145–167.

[127] H. A. VAN DER VORST, *BI-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 631–644.

[128] H. VAN NGAI, D. T. LUC, AND M. THÉRA, *Approximate convex functions*, J. Nonlinear Convex Anal., 1 (2000), pp. 155–176.

[129] H. L. VASUDEVA, *Elements of Hilbert spaces and operator theory*, Springer, 2017.

[130] R. VERFÜRTH, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comput. Appl. Math., 50 (1994), pp. 67–83.

[131] G. WACHSMUTH, *Strong stationarity for optimal control of the obstacle problem with control constraints*, SIAM J. Optim., 24 (2014), pp. 1914–1932.

[132] G. WACHSMUTH, *Towards M-stationarity for optimal control of the obstacle problem with control constraints*, SIAM J. Control Optim., 54 (2016), pp. 964–986.

[133] T. P. WIHLER, *Weighted L^2-norm a posteriori error estimation of FEM in polygons*, Int. J. Numer. Anal. Model., 4 (2007), pp. 100–115.

[134] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*. Nondiffer. Optim., Math. Program. Study 3, 145-173, 1975.

[135] H. XU AND J. J. YE, *Necessary optimality conditions for two-stage stochastic programs with equilibrium constraints*, SIAM J. Optim., (2010), pp. 1685–1715.

[136] H. XU AND D. ZHANG, *Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications*, Math. Program., 119 (2009), pp. 371–401.

[137] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications - II/ A: Linear Monotone Operators*, Springer Science & Business Media, New York, 1990.