Technische Universität München

Fakultät für Mathematik

Lehrstuhl für Mathematische Optimierung

Interior point methods for optimal control problems with pointwise state constraints

Florian Kruse

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:		Univ.-Prof. Dr. Peter Gritzmann
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Michael Ulbrich
	2.	o. Univ.-Prof. Dr. Karl Kunisch (nur schriftliche Prüfung) Karl-Franzens-Universität, Graz, Österreich
		Univ.-Prof. Dr. Boris Vexler (nur mündliche Prüfung)

Die Dissertation wurde am 19.11.2013 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 23.02.2014 angenommen.

Abstract

This work presents a new approach for the solution of pointwise state constrained optimal control problems that are governed by linear elliptic partial differential equations. The main idea of this approach is to replace the state constraints by a single constraint using a smoothed minimum function. The resulting interior point methods are analyzed in an infinite-dimensional setting using in parts the concept of self-concordance, which is generalized to Banach spaces. Numerical experiments demonstrate the efficiency of the approach.

Zusammenfassung

Diese Arbeit befasst sich mit einem neuen Ansatz zur Lösung von Optimalsteuerungsproblemen mit partiellen Differentialgleichungen und punktweisen Zustandsbeschränkungen. Die Hauptidee ist, die Zustandsbeschränkungen mithilfe einer geglätteten Minimumfunktion durch eine einzige Nebenbedingung zu ersetzen. Für die sich daraus ergebenden Innere-Punkte-Verfahren wird Konvergenztheorie im Unendlichdimensionalen entwickelt. Dabei wird unter anderem Selbstkonkordanztheorie verwendet, die auf Banachräume verallgemeinert wird. Numerische Experimente belegen die Effektivität des Ansatzes.

Contents

1. Introduction

1.1. Optimal control with state constraints

In this thesis we present a new class of interior point methods to tackle pointwise state constrained optimal control problems that are governed by linear elliptic partial differential equations. The problem class that we consider is presented in detail in Section 3. It comprises, in particular, problems of the form

$$\min_{(y,u)\in Y\times U} Q(y,u) + \frac{\hat{\alpha}}{2}\|u\|_U^2 \quad \text{s.t.} \quad y \geq y_a \text{ in } \overline{\Omega}, \quad Ay + Bu = g. \tag{$\mathrm{P}_{\mathrm{full}}$}$$

Here, $\hat{\alpha} > 0$, Y is a Banach space with $Y \hookrightarrow C^{0,\beta}(\overline{\Omega})$ continuously for a given $\beta > 0$ and a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$, Z a Banach space, U a Hilbert space, $Q : Y \times U \to \mathbb{R}$ is quadratic and convex, $y_a \in C^{0,\beta}(\overline{\Omega})$, $A \in \mathcal{L}(Y,Z)$ invertible, $B \in \mathcal{L}(U,Z)$, and $g \in Z$. Also, we assume that $(y^\circ, u^\circ) \in Y \times U$ and $\tau^\circ > 0$ exist with $Ay^\circ + Bu^\circ = g$ as well as $y^\circ - \tau^\circ \geq y_a$ in $\overline{\Omega}$. The equality constraint $Ay + Bu = g$ models a partial differential equation (PDE) in Ω.

It is well-known that the pointwise state constraints $y(x) \geq y_a(x)$ for all $x \in \overline{\Omega}$ complicate the solution of ($\mathrm{P}_{\mathrm{full}}$) since the associated Lagrange multiplier is, in general, only a measure, see, e.g., [Cas86] and [CMV13]. Currently, there exist mainly three types of Newton-based algorithms that can deal with ($\mathrm{P}_{\mathrm{full}}$) and for which an infinite-dimensional analysis is available: Moreau-Yosida regularization, cf., e.g., [IK03, HK06a, HK06b, HK09], Lavrentiev regularization and the closely related virtual control concept, cf., e.g., [MRT06, MPT07, KR09, TY09a, TY09b], and interior point methods, cf., e.g., [PS09, Sch09a, Sch09b, Sch12]. In all these approaches a family of regularized problems is introduced. This family induces a path of solutions that the respective algorithm follows. For a fixed regularization parameter the corresponding regularized problem is solved using (a possibly semismooth) Newton's method. Although all these approaches are quite successful in practice and their convergence analyses are sophisticated, there are still several open questions. For methods based on Moreau-Yosida and Lavrentiev-type regularizations, the convergence of Newton's method for a fixed regularized problem as well as the convergence of the path of solutions to the solution of the original problem have been established. Furthermore, estimates are available regarding Hölder continuity of the path of solutions, cf. [SH11, HSW12] for Moreau-Yosida regularization and [CKR08, KR09] for Lavrentiev-type regularizations. However, there are no complexity estimates available for fixed regularization parameter and there are no results regarding the convergence of the overall path-following algorithms. For interior point methods Hölder continuity of the path of solutions has been established, cf., e.g., [Sch09b], and it has been shown that it is possible to decrease the regularization parameter in such a way that the iterates converge to the optimal

solution of the original problem, cf. [Sch12]. This result concerns a short step method, i.e., each Newton step is accompanied by a decrease of the regularization parameter. Thus, it also provides a complexity estimate for fixed regularization parameter. However, there are no results regarding the rate of convergence of the regularization parameter. Moreover, all these path-following algorithms lack a measure for the proximity of the actual iterate to the path that the respective algorithm follows which can be used in theory *and* practice. In a practical algorithm, therefore, heuristics are used to decide when to stop Newton's method for fixed regularization parameter. However, none of these heuristics guarantee convergence of the overall algorithm on an infinite-dimensional level.

The approach taken in this thesis is based on the idea to replace the pointwise state constraints by the constraint $\min_\varepsilon(y - y_a) \geq 0$. Here, $\min_\varepsilon$ is a smoothed version of the minimum functional $\min_{x \in \overline{\Omega}}$, and $\varepsilon > 0$ denotes the corresponding smoothing parameter. This yields regularized versions of ($\mathrm{P}_{\mathrm{full}}$), parametrized by ε, and induces a path of solutions. We show that this path leads to the optimal solution of ($\mathrm{P}_{\mathrm{full}}$) and is Hölder continuous with order almost $\mathcal{O}(\sqrt{\varepsilon})$. This motivates two schemes: We can fix ε and derive interior point methods that solve the regularized problem associated with this ε. In combination with the estimate from the Hölder continuity of the path of solutions, this approach is viable to solve ($\mathrm{P}_{\mathrm{full}}$) up to a prescribed accuracy corresponding to a given ε. The second idea is to develop interior point methods in which ε is driven to zero.

For all algorithms that we develop we provide a detailed convergence analysis. This includes, in particular:

- A proximity measure through which we can ensure, in theory and practice, that the iterates stay close enough to the path which the respective algorithm follows.
- Linear convergence of the iterates to the solution of the regularized problem together with complexity estimates in the first scheme. Here, all constants are explicit, i.e., we can say in advance how many iterations suffice to produce a δ-optimal solution of the regularized problem.
- Convergence of the iterates to the solution of ($\mathrm{P}_{\mathrm{full}}$) together with a bound on the number of Newton steps required for each regularization parameter and, moreover, a rate of convergence for the regularization parameter in the second scheme.
- In both schemes we have an estimate for the difference in function value between the actual iterate and the solution of ($\mathrm{P}_{\mathrm{full}}$).
- Complexity estimates for phase one, i.e., estimates for the number of Newton steps required to find a suitable starting point if an arbitrary interior point is used as starting point.

Our approach relies in part on the theory of interior point methods for self-concordant barrier functions, which Nesterov and Nemirovski introduced to finite-dimensional optimization, cf., e.g., [NN94]. Since the optimization variables y and u in ($\mathrm{P}_{\mathrm{full}}$), however, belong to infinite-dimensional function spaces, we need to generalize the concept of self-concordance to infinite-dimensional spaces. Therefore, we also develop a rigorous treatment of self-concordance

in Banach spaces in this thesis. To the best of our knowledge self-concordance in an infinite-dimensional setting is only considered in [FM97a] and [FM97b]. However, there the constraints are quadratic, which is rather restrictive and, in fact, not satisfied in our approach for (P_{full}).

This thesis is organized as follows. In Section 2 we generalize the theory of self-concordance to Banach spaces. In Section 3 we present the class of optimal control problems that we consider along with our approach to tackle these problems. In Section 4 we investigate the smoothed problems that result from this approach. In Section 5 we present short step, long step, and phase one methods for the first scheme, i.e., for fixed smoothing parameter ε. In Section 6 we provide the necessary theory to analyze interior point methods for the second scheme, i.e., for $\varepsilon \to 0^+$. In section 7 we examine short step, long step, and phase one methods for the second scheme. In Section 8 we present numerical results for both schemes. In Section 9 we summarize and point out some possible future research directions. Furthermore, this thesis contains an appendix to provide several, sometimes technical results that we consider either to be well-known or that we feel would distract the reader unnecessarily from the presentation of the main ideas. For instance, the appendix contains several results from functional analysis and convex optimization.

1.2. Notation

We employ a rather standard notation. For the reader's convenience, however, we included some of our notation in Section A of the appendix.

2. Self-concordance in Banach spaces

In this section we generalize the theory of self-concordant and self-bounded barrier functions that Nesterov and Nemirovski introduced to finite-dimensional optimization, cf., e.g., [NN94], to Banach spaces. Since this theory makes intensive use of scalar products, one of our assumptions implies that the underlying space is, in fact, a Hilbert space. However, since it is no additional effort to work in Banach spaces and since it allows us to clearly point out the assumption that induces the Hilbert space structure, we use this more general framework. Also, it may be possible to work with a weaker assumption that induces only a pre-Hilbert space structure. However, the investigation of whether or not this is possible is beyond the scope of this thesis.

Many proofs in this section are inspired by their finite-dimensional counterparts, with [NN94] being the main reference. However, there is more literature that influenced the proofs we present in this section; besides the references directly quoted we mention [BV04], [den94], [JS04], [Ren01], and [Ulb10].

To develop the theory of self-concordance in Banach spaces, we employ the following assumption.

Assumption 2.0.1. *Throughout Section 2 let X be a Banach space and $\|\cdot\|_X$ its norm. Moreover, during the entire section we suppose that $K \subset X$ is a nonempty, open, and convex subset of X.*

Our aim in this section is to develop barrier methods to solve the optimization problem

$$\min_{x \in X} j(x) \qquad \text{s.t.} \qquad x \in M, \tag{$\mathrm{P_{SC}}$}$$

where $j : M \to \mathbb{R}$, and M satisfies $K \subset M \subset \overline{K}$. Further details of this problem are fixed later. Before we deal with barrier methods for this problem, we introduce and investigate the class of self-concordant functions, the class of self-concordant barrier functions, and the class of self-bounded functions. These functions are at the heart of the barrier methods that we develop.

2.1. Self-concordant functions

The following definition introduces the class of self-concordant functions. This notion, together with the one of self-concordant *barrier* functions and the concept of self-boundedness (we introduce the first in Section 2.2 and the latter in Section 2.3) are very important since they constitute the framework in which we work throughout this thesis.

Definition 2.1.1 (Cf. [Jar94] and [NN94])**.** Let $f : K \to \mathbb{R}$ be thrice Fréchet differentiable. We call f *self-concordant on* K iff f is convex and satisfies for all $x \in K$ and all $h \in X$

$$f'''(x)[h,h,h] \leq 2\left(f''(x)[h,h]\right)^{\frac{3}{2}}.$$

Example 2.1.2. The standard example for a self-concordant function is $-\ln : K \to \mathbb{R}$ with $K = \mathbb{R}_{>0}$. When we apply the theory of Section 2 to optimal control, we will encounter more complex examples of self-concordant functions.

Remark 2.1.3. The right-hand side of the above inequality is well-defined since $f''(x)$ is positive semidefinite due to the convexity of f, cf. Lemma C.4.9.

Remark 2.1.4. In [NN94], Definition 2.1.1 is stated in more generality, since a-self-concordant functions with $a > 0$ are introduced. Self-concordant functions in the sense of Definition 2.1.1 are 1-self-concordant functions in the sense of [NN94].

Remark 2.1.5. The exponent 3/2 in the defining inequality of self-concordance is the only nonnegative exponent that yields a suitably large class of functions: Assume that the exponent were $p \geq 0$. Inserting th for h we obtain

$$t^{3-2p} f'''(x)[h,h,h] \leq 2\left(f''(x)[h,h]\right)^{p}$$

for every $x \in K$, every $h \in X$, and every $t \in \mathbb{R} \setminus \{0\}$. Considering $t \to 0^+$, $t \to 0^-$, and $t \to \pm\infty$, we deduce that it holds either $f''' \equiv 0$ or $p = 3/2$. Hence, for $p \neq 3/2$ the class of self-concordant functions would only contain quadratic functions. However, to construct interior point methods we also need f to be a barrier function for K, i.e., $f(x^k) \to \infty$ for every sequence $(x^k) \subset K$ that converges to a point x on the topological boundary of K. Consequently, if every self-concordant function were quadratic, we could only provide interior point methods for $K = X$, i.e., for unconstrained optimization problems. Thus, $p = 3/2$ is the only reasonable choice.

Remark 2.1.6. Although the exponent 3/2 is the only reasonable choice in the definition of self-concordance, this is not the case for the factor 2 that appears in this definition. This factor can, in fact, be chosen arbitrarily: If $a > 0$ is chosen instead of 2 in the definition, we see that a function f is self-concordant with respect to the new definition if and only if $\frac{a^2}{4} f$ is self-concordant with respect to the old definition with factor 2. The reason to use exactly 2 is to make the function $-\ln : \mathbb{R}_{>0} \to \mathbb{R}$ self-concordant as it is, i.e., without having to rescale it.

Remark 2.1.7. It follows from Theorem C.2.11 that although f''' is not assumed to be continuous, the trilinear form $f'''(x) : X \times X \times X \to \mathbb{R}$ is symmetric, a property that we employ in several of the following proofs. Moreover, we suspect that for the following theory even thrice Gâteaux differentiability with all differentials being symmetric would suffice, but we do not follow this line of thought since in the optimal control setting we are interested in, differentiability is not an issue. Nonetheless, the assumption of thrice Fréchet differentiability instead of thrice continuous differentiability in Definition 2.1.1 is a small generalization in comparison to the literature.

Definition 2.1.1 is often stated slightly different, namely with the absolute value applied to the left-hand side of the defining inequality. In the next lemma we show that these two definitions coincide and present a third equivalent characterization of self-concordance.

Lemma 2.1.8. *Let $f : K \to \mathbb{R}$ be thrice Fréchet differentiable. Then the following three statements are equivalent:*

1) f is self-concordant on K, i.e., f is convex and satisfies for all $x \in K$ and all $h \in X$

$$f'''(x)[h,h,h] \leq 2\left(f''(x)[h,h]\right)^{\frac{3}{2}}.$$

2) f is convex and satisfies for all $x \in K$ and all $h \in X$

$$\left|f'''(x)[h,h,h]\right| \leq 2\left(f''(x)[h,h]\right)^{\frac{3}{2}}.$$

3) f satisfies for all $x \in K$ and all $h \in X$

$$\left(f'''(x)[h,h,h]\right)^2 \leq 4\left(f''(x)[h,h]\right)^3.$$

Proof. Assume that 1) holds. Fix $x \in K$ and $h \in X$. In the defining inequality of self-concordance we replace h with $-h$ and use $f'''(x)[-h,-h,-h] = -f'''(x)[h,h,h]$ as well as $f''(x)[-h,-h] = f''(x)[h,h]$ to infer $\max\{f'''(x)[h,h,h], -f'''(x)[h,h,h]\} \leq 2\left(f''(x)[h,h]\right)^{\frac{3}{2}}$. This clearly implies 2). Now assume that 2) holds. Fix $x \in K$ and $h \in X$. Squaring both sides of the inequality in 2) yields 3). Suppose that 3) holds. First note that 3) implies $f''(x)[h,h] \geq 0$ for all $x \in K$ and all $h \in X$, which shows the convexity of f. To derive the inequality in 1), take the square root in 3), which is possible due to $f''(x)[h,h] \geq 0$. With $f'''(x)[h,h,h] \leq |f'''(x)[h,h,h]|$ this implies 1). □

We present another characterization of self-concordance. To state this characterization concisely we need the following definition.

Definition 2.1.9. Let $f : K \to \mathbb{R}$ be a function. For $x \in K$ and $h \in X$ we define

$$I_{x,h} := \{t \in \mathbb{R} : x + th \in K\} \qquad \text{and} \qquad f_{x,h} : I_{x,h} \to \mathbb{R}, \quad f_{x,h}(t) := f(x + th).$$

Remark 2.1.10. Since K is open and convex, $I_{x,h}$ is an open interval containing zero.

Using this definition we now present different characterizations of self-concordance which, basically, show that self-concordance is a property "along lines".

Lemma 2.1.11. *Let $f : K \to \mathbb{R}$ be thrice Fréchet differentiable. Then f is self-concordant on K if and only if one of the following four equivalent statements holds:*

1) For every $x \in K$ and every $h \in X$ the function $f_{x,h}$ is convex and satisfies for all $t \in I_{x,h}$

$$f'''_{x,h}(t) \leq 2\left(f''_{x,h}(t)\right)^{\frac{3}{2}}.$$

2) For every $x \in K$ and every $h \in X$ the function $f_{x,h}$ is convex and satisfies for all $t \in I_{x,h}$

$$\left|f'''_{x,h}(t)\right| \leq 2\left(f''_{x,h}(t)\right)^{\frac{3}{2}}.$$

3) For every $x \in K$ and every $h \in X$ the function $f_{x,h}$ satisfies for all $t \in I_{x,h}$

$$\left(f'''_{x,h}(t)\right)^2 \leq 4\left(f''_{x,h}(t)\right)^3.$$

4) For every $x \in K$ and every $h \in X$ the function $f_{x,h}$ is self-concordant on $I_{x,h}$.

Proof. Using $f'''_{x,h}(t) = f'''(y)[h,h,h]$ and $f''_{x,h}(t) = f''(y)[h,h]$ with $y := x+th$, the equivalence of 1), 2), and 3) can be established as in Lemma 2.1.8. We now show that f is self-concordant on K if and only if 3) holds. The inequality in 3) is equivalent to

$$\left(f'''(x+th)[h,h,h]\right)^2 \leq 4\left(f''(x+th)[h,h]\right)^3. \tag{2.1}$$

Since $I_{x,h} \supset \{0\}$ holds, this implies self-concordance of f, as follows by virtue of 3) in Lemma 2.1.8. Conversely, fix $x \in K$, $h \in X$, and $t \in I_{x,h}$. Using 3) from Lemma 2.1.8 at $x + th \in K$ in direction $h \in X$, we obtain (2.1). This concludes the reasoning that 3) is equivalent to f being self-concordant. We now establish that 4) is equivalent to 3). This follows by applying Lemma 2.1.8 3) to $f_{x,h}$ since the direction in which $f''_{x,h}(t)$ and $f'''_{x,h}(t)$ are evaluated belongs to $\mathbb{R}$ and, therefore, cancels out. □

Remark 2.1.12. The characterization in 4) indicates how we can use finite-dimensional self-concordance theory to establish results in the infinite-dimensional case.

Definition 2.1.13. Let $f : K \to \mathbb{R}$ be twice continuously differentiable and convex. For every $x \in K$, $f''(x)$ induces a symmetric, positive semidefinite bilinear form via the definition $(v,w)_{f''(x)} := f''(x)[v,w]$ and a seminorm via $\|v\|_{f''(x)} := \sqrt{(v,v)_{f''(x)}}$. We call this seminorm the *local seminorm of v at x.*

Remark 2.1.14. The fact that a symmetric, positive semidefinite bilinear form induces a seminorm is proven in [MV92, 11.1].

The following lemma presents a geometrical interpretation of self-concordance.

Lemma 2.1.15. *Let $f : K \to \mathbb{R}$ be thrice Fréchet differentiable and convex, and denote $\overline{B}_x := \{h \in X : \|h\|_{f''(x)} \leq 1\}$ for $x \in K$. Then f is self-concordant on K if and only if it satisfies for all $x \in K$*

$$\sup_{h_1,h_2,h_3 \in \overline{B}_x} f'''(x)[h_1,h_2,h_3] \leq 2.$$

Proof. This follows from the definition in combination with Lemma C.1.6. □

Remark 2.1.16. The previous lemma shows that $f'''(x)$ is a bounded trilinear form if the seminorm $\|\cdot\|_{f''(x)}$ is used on X, and that the bound is uniform with respect to x. Note that if $\|\cdot\|_X$ is used instead of $\|\cdot\|_{f''(x)}$, then this uniformity may not hold as $f(x) = -\ln(x)$ shows.

Remark 2.1.17. We will see that self-concordant functions are very well-suited for the application of Newton's method. An informal motivation for this fact is based on the preceding lemma: It is well-known, for instance from the Newton-Kantorovich theorem and the Newton-Mysovskikh theorem, cf. [Deu11, Section 1.2.1], that the convergence properties of Newton's method depend on the Lipschitz constant of f''. Thus, bounding f''', which corresponds to bounding the Lipschitz constant of f'', seems to be a reasonable approach.

The sum of two self-concordant functions is self-concordant, too.

Lemma 2.1.18. *Let $K_1, K_2 \subset X$ be open and convex sets with $K := K_1 \cap K_2 \neq \emptyset$. Let $f_1 : K_1 \to \mathbb{R}$ and $f_2 : K_2 \to \mathbb{R}$ be self-concordant on K_1, respectively, K_2. Then*

$$f : K \to \mathbb{R}, \quad f(x) := f_1(x) + f_2(x)$$

is self-concordant on the nonempty, open, and convex set K.

Proof. Clearly, K is nonempty, open, and convex, and f is thrice Fréchet differentiable. It remains to show the self-concordance of f. The definition of self-concordance implies that the restrictions of f_1 and f_2 to K are self-concordant. Thus, the assertion follows from the inequality $a^{3/2} + b^{3/2} \leq (a+b)^{3/2}$ for $a, b \geq 0$. □

We show that $-C\ln(q)$ is self-concordant for $C \in [1, \infty)$ if q is a quadratic, concave function. Note that if f is self-concordant and $C \in (0, 1)$, then Cf may not be self-concordant.

Lemma 2.1.19. *Let $q : X \to \mathbb{R}$ be quadratic and concave, and define $K := \{x \in X : q(x) > 0\}$. Suppose that K is nonempty and let $C \in [1, \infty)$. Then $f : K \to \mathbb{R}$, $f(x) := -C\ln(q(x))$ is self-concordant on the nonempty, open, and convex set K.*

Proof. Obviously, K is nonempty and open. It is, furthermore, convex since q is concave. To show that f is self-concordant on K, it suffices to consider $C = 1$. The fact that f is thrice Fréchet differentiable can be deduced from Corollary C.2.10 and the product rule. Using $q''' \equiv 0$ and the notation $q := q(x)$, $q' := q'(x)[h]$, $q'' := q''(x)[h,h]$ for $x \in K$ and $h \in X$ we obtain

$$f''(x)[h,h] = \frac{(q')^2 - qq''}{q^2} \geq 0 \quad \text{and} \quad f'''(x)[h,h,h] = \frac{3qq'q'' - 2\,(q')^3}{q^3},$$

where we used $qq'' \leq 0$ due to the concavity of q and the definition of K. Hence, it remains to show $(f'''(x)[h,h])^2 \leq 4(f''(x)[h,h])^3$, which is equivalent to $(3qq'q'' - 2(q')^3)^2 \leq 4((q')^2 - qq'')^3$. For the left-hand side we have $(3qq'q'' - 2(q')^3)^2 \leq 4(q')^6 - 12q(q')^4q'' + 12(qq'q'')^2$. For the right-hand side we use $qq'' \leq 0$ to deduce $4((q')^2 - qq'')^3 \geq 4(q')^6 - 12q(q')^4q'' + 12(qq'q'')^2$. These estimates imply that f is self-concordant on K. □

2.2. Self-concordant barrier functions

In this section we introduce and examine the class of self-concordant barrier functions.

Definition 2.2.1 (Cf. [Jar94])**.** We call a continuous function $f : K \to \mathbb{R}$ a *barrier function for K* iff it has the property

$$(x^k) \subset K, \lim_{k\to\infty} x^k = x \in \partial K \quad \Longrightarrow \quad \lim_{k\to\infty} f(x^k) = +\infty.$$

Here, ∂K denotes the boundary of K.

Example 2.2.2. Obviously, $-\ln : K \to \mathbb{R}$ is a barrier function for $K = \mathbb{R}_{>0}$.

Remark 2.2.3. It is no restriction to assume in this definition that K is open: Suppose that we have a barrier function $f : S \to \mathbb{R}$ for $S \subset X$. From the barrier property together with the continuity of f we infer that $S \cap \partial S$ is empty. Hence, S is open.

Definition 2.2.4. We call $f : K \to \mathbb{R}$ a *self-concordant barrier function for* K iff it is self-concordant on K and a barrier function for K.

Remark 2.2.5. In [NN94], self-concordant barrier functions are called strongly 1-self-concordant functions.

We examine the behaviour of the local seminorm at x with respect to perturbations in x. For self-concordant barrier functions, local seminorms contain feasibility information.

Lemma 2.2.6 (cf. [NN94, Theorem 2.1.1])**.** *Let $f : K \to \mathbb{R}$ be self-concordant on K. Let $x \in K$ and $\tilde{h} \in X$ with $\|\tilde{h}\|_{f''(x)} < 1$ and $x + \tilde{h} \in K$. Then it holds for all $h \in X$*

$$\left(1 - \|\tilde{h}\|_{f''(x)}\right) \|h\|_{f''(x)} \leq \|h\|_{f''(x+\tilde{h})} \leq \frac{1}{1 - \|\tilde{h}\|_{f''(x)}} \|h\|_{f''(x)}.$$

If, in addition, f is a barrier function for K, then $\|\tilde{h}\|_{f''(x)} < 1$ implies $x + \tilde{h} \in K$.

Proof. We start by demonstrating the asserted inequalities. To this end, we define $I := [0, 1]$ and set for every $h \in X$

$$\psi_h : I \to \mathbb{R}, \quad \psi_h(t) := f''(x + t\tilde{h})[h, h].$$

According to Lemma C.1.6, self-concordance of f implies for all $x \in K$ and all $h_1, h_2, h_3 \in X$

$$|f'''(x)[h_1, h_2, h_3]| \leq 2\sqrt{f''(x)[h_1, h_1]} \cdot \sqrt{f''(x)[h_2, h_2]} \cdot \sqrt{f''(x)[h_3, h_3]}.$$

Fixing $h \in X$ this establishes on I the differential inequalities

$$|\psi'_{\tilde{h}}(t)| \leq 2\left(\psi_{\tilde{h}}(t)\right)^{\frac{3}{2}} \qquad \text{and} \qquad |\psi'_h(t)| \leq 2\sqrt{\psi_{\tilde{h}}(t)} \cdot \psi_h(t). \tag{2.2}$$

Applying Gronwall's inequality from Lemma D.0.4 1) with $\alpha(t) = 2\psi_{\tilde{h}}(t)^{\frac{1}{2}}$, we conclude $\psi_{\tilde{h}} \equiv 0$ on I if $\|\tilde{h}\|_{f''(x)} = \psi_{\tilde{h}}(0) = 0$ holds. The application of Lemma D.0.4 1) is possible since (2.2) shows with the continuity of $\psi_{\tilde{h}}$ on the compact interval I that $\psi'_{\tilde{h}}$ is bounded. From Lemma D.0.4 2), applied with $\alpha(t) = -2\psi_{\tilde{h}}(t)^{\frac{1}{2}}$, we deduce that $\|\tilde{h}\|_{f''(x)} > 0$ implies $\psi_{\tilde{h}} > 0$ on I. In the first case, (2.2) shows that ψ_h is constant on I by the mean value theorem so that the asserted inequalities follow from $\|h\|^2_{f''(x+\tilde{h})} = \psi_h(1) = \psi_h(0) = \|h\|^2_{f''(x)}$. It remains to deal with the second case. Thus, we have $\psi_{\tilde{h}} > 0$ on I. The self-concordance of f, therefore, implies $|(\psi_{\tilde{h}}(t)^{-\frac{1}{2}})'| \leq 1$ on I. Integration from 0 to t establishes $(\psi_{\tilde{h}}(t))^{-\frac{1}{2}} \geq (\psi_{\tilde{h}}(0))^{-\frac{1}{2}} - t$ for all $t \in I$. Integration is possible due to Lemma D.0.3 since $(\psi_{\tilde{h}}(t)^{-\frac{1}{2}})'$ is apparently bounded on I. From the last inequality, we infer $\sqrt{\psi_{\tilde{h}}(t)} \leq \sqrt{\psi_{\tilde{h}}(0)}/(1 - t\sqrt{\psi_{\tilde{h}}(0)})$. Note that $1 - t\sqrt{\psi_{\tilde{h}}(0)} \geq 1 - \|\tilde{h}\|_{f''(x)} > 0$ holds on I by assumption. This shows with (2.2) that $|\psi'_h(t)| \leq 2\psi_h(t) \cdot \sqrt{\psi_{\tilde{h}}(0)}/(1 - t\sqrt{\psi_{\tilde{h}}(0)})$ is satisfied for all $t \in I$. Using Gronwall's inequality

again, we either have $\psi_h \equiv 0$ or $\psi_h > 0$ on I depending on whether $\psi_h(0) = 0$ or $\psi_h(0) > 0$. In the first case, ψ_h is constant and as above this implies the asserted inequalities. In the latter case, we divide by ψ_h and integrate from 0 to t to find $|\ln(\psi_h(t)/\psi_h(0))| \leq 2\ln(1/(1-t\sqrt{\psi_{\tilde{h}}(0)}))$ for all $t \in I$. From this, both inequalities of the assertion follow by use of $t = 1$ and application of the exponential function.

For the second part of the proof, let $x \in K$ and $\tilde{h} \in X$ with $\|\tilde{h}\|_{f''(x)} < 1$ be given. We have to show $1 \in I_{x,\tilde{h}}$. To argue by contradiction let us assume $1 \notin I_{x,\tilde{h}}$. The first assertion implies $\sqrt{f''_{x,\tilde{h}}(t)} \leq \sqrt{f''_{x,\tilde{h}}(0)}/(1 - t\sqrt{f''_{x,\tilde{h}}(0)})$ for all $t \in I_{x,\tilde{h}} \cap [0,1]$. Squaring this inequality and applying Corollary C.2.3 twice, we see that this implies boundedness of $f_{x,\tilde{h}}$ on $I_{x,\tilde{h}} \cap [0,1]$. Moreover, we have $f_{x,\tilde{h}}(t) \to +\infty$ for $t \to \sup I_{x,\tilde{h}} \leq 1$ since f is a barrier function for K. However, this contradicts the boundedness of $f_{x,\tilde{h}}$ on $I_{x,\tilde{h}} \cap [0,1]$. □

Remark 2.2.7. The inequalities of the previous lemma are sometimes used to *define* the class of (strongly nondegenerate) self-concordant functions, see, e.g., [Ren01, Section 2.2.1]. This broadens the class of self-concordant functions slightly since then self-concordant functions are allowed to be only twice continuously differentiable instead of thrice Fréchet differentiable. The definition of self-concordance we employ is, however, better suited to check if a given function is self-concordant. And indeed, proving self-concordance of certain barrier functions is an important part of this thesis. Furthermore, it can be shown that in the case of thrice Fréchet differentiable functions with $f''(x)$ being positive definite for all $x \in K$, the definition via the inequalities of Lemma 2.2.6 is equivalent to the definition of a self-concordant barrier function in the sense of Definition 2.2.4, see [Ren01, Section 2.5, in particular Theorem 2.5.3].

To compute Newton steps we require $f''(x) \in \mathcal{L}(X, X^*)$ to be invertible. In a finite-dimensional setting the following property is sufficient to ensure this.

Definition 2.2.8. Let $f : K \to \mathbb{R}$ be twice continuously differentiable. We call f *pd on* K iff $f''(x) \in \mathcal{B}(X \times X, \mathbb{R})$ is positive definite for every $x \in K$, i.e., iff $f''(x)[h,h] > 0$ holds for all $x \in K$ and all $h \in X \setminus \{0\}$. Here, $\mathcal{B}(X \times X, \mathbb{R})$ denotes the space of bounded real-valued bilinear mappings defined on $X \times X$.

Remark 2.2.9. Accordingly, we say that f is *pd self-concordant on* K or *a pd self-concordant barrier function for* K iff f is pd and self-concordant on K or pd and a self-concordant barrier function for K, respectively. We mention that in [NN94], pd self-concordant barrier functions are called nondegenerate strongly 1-self-concordant functions.

We provide sufficient conditions for self-concordant functions to be pd self-concordant.

Lemma 2.2.10. *Let $f : K \to \mathbb{R}$ be self-concordant on K. If there exists $x \in K$ such that $f''(x)$ is positive definite, then f is pd self-concordant on K. Furthermore, if K does not contain a straight line, then every self-concordant barrier function for K is a pd self-concordant barrier function for K.*

Proof. To establish the first assertion we argue by contraposition. To this end, let there be $\tilde{x} \in K$ and $h \in X \setminus \{0\}$ such that $f''(\tilde{x})[h,h] = 0$ holds. We show that this implies

$f''(x)[h,h] = 0$ for all $x \in K$. This suffices. Let $x \in K$ and denote $I := [0,1]$, $\gamma : I \to K$, $\gamma(t) := \tilde{x} + t(x - \tilde{x})$, and $\varphi : I \to \mathbb{R}_{\geq 0}$, $\varphi(t) := f''(\gamma(t))[h,h]$. Then we have

$$|\varphi'(t)| = |f'''(\gamma(t))[h,h,x-\tilde{x}]| \leq 2f''(\gamma(t))[h,h]\sqrt{f''(\gamma(t))[x-\tilde{x},x-\tilde{x}]}$$

for all $t \in I$, where we used the inequality from Lemma C.1.6. Due to the extreme value theorem there exists $M \geq 0$ such that $\varphi'(t) \leq Mf''(\gamma(t))[h,h] = M\varphi(t)$ holds for all $t \in I$. Thus, Gronwall's inequality together with $\varphi(0) = 0$ yields $\varphi \equiv 0$ on I, cf. Lemma D.0.4. In particular, it holds $0 = \varphi(1) = f''(x)[h,h]$.

To establish the second assertion let there again be $\tilde{x} \in K$ and $h \in X \setminus \{0\}$ such that $f''(\tilde{x})[h,h] = 0$ holds. We have shown in the first part that this implies $f''(x)[h,h] = 0$ for all $x \in K$. Now fix $x \in K$ and consider $f_{x,h} : I_{x,h} \to \mathbb{R}$. Since there holds $f''_{x,h} \equiv 0$ on $I_{x,h}$, $f_{x,h}$ is affine linear on $I_{x,h}$. Hence, the barrier property of f yields $I_{x,h} = \mathbb{R}$, i.e., K contains the straight line $x + th$, $t \in \mathbb{R}$. □

If K does not contain a straight line and $f : K \to \mathbb{R}$ is a self-concordant barrier function for K, then f is a pd self-concordant barrier function for K according to Lemma 2.2.10. If X is finite-dimensional, this implies that $f''(x)$ is invertible for all $x \in K$. However, if X is infinite-dimensional, then $f''(x)$ may fail to be invertible, nonetheless. The following example demonstrates this.

Example 2.2.11. Consider the separable Hilbert space $X := \ell_2 := \ell_2(\mathbb{R})$, i.e., the space of square summable real sequences. We endow this space with the usual scalar product and norm, i.e., $(v,w)_X := \sum_{k=1}^{\infty} v_k w_k$ and $\|x\|_X := (\sum_{k=1}^{\infty} x_k^2)^{1/2}$. We consider the concave quadratic function $q(x) := 1 - \frac{1}{2}\|T(x)\|_X^2$, where $T : X \to X$ denotes the bounded linear operator $T(x)_k := \frac{x_k}{k}$ for all $k \in \mathbb{N}$. Note that $T(x) \in X$ as well as the boundedness of T follow from $\|T(x)\|_X \leq \|x\|_X$. Lemma 2.1.19 yields that $f(x) := -\ln(q(x))$ is self-concordant on $K := \{x \in X : q(x) > 0\}$. The continuity of q on X implies $\partial K \subset \{x \in X : q(x) = 0\}$ and, therefore, f is a barrier function for K. For $t \in \mathbb{R}$ and $x, d \in X$ we have $\|T(x+td)\|_X \geq t\|T(d)\|_X - \|T(x)\|_X$ by the reverse triangle inequality, which implies that K does not contain a straight line. Hence, f is pd self-concordant on K. In particular, $f''(\bar{x})$ is positive definite at $\bar{x} := 0 \in K$. We now show that $f''(\bar{x}) \in \mathcal{L}(X, X^*)$ is not invertible. Due to $T(\bar{x}) = 0$ and $q(\bar{x}) = 1$ we have

$$f''(\bar{x})[h_1][h_2] = \frac{(T(h_1),T(h_2))_X}{q(\bar{x})} + \frac{(T(\bar{x}),T(h_1))_X \cdot (T(\bar{x}),T(h_2))_X}{(q(\bar{x}))^2} = (T(h_1),T(h_2))_X$$

for all $h_1, h_2 \in X$. (This computation also confirms that $f''(x)$ is positive definite for $x \in K$.) Hence, $f''(\bar{x})[j^2 e_j] = w_j$, where $e_j \in X$ denotes the sequence whose j-th component is one and whose other components are zero, and $w_j \in X^*$ denotes $w_j(v) := v_j$. Since there hold $\|j^2 e_j\|_X = j^2$ and $\|w_j\|_{X^*} = 1$ for every $j \in \mathbb{N}$, this demonstrates that for $f''(\bar{x})$, the preimage of the unit ball is not bounded. Hence, $f''(\bar{x})$ is not boundedly invertible. With Theorem C.1.1 this shows that $f''(\bar{x})$ is not invertible at all.

We remark that $\|\cdot\|_{f''(\bar{x})}$ is not equivalent to $\|\cdot\|_X$, as is evident from $\|e_j\|_{f''(\bar{x})} = j^{-1}$ and $\|e_j\|_X = 1$ for all $j \in \mathbb{N}$. (However, $\|\cdot\|_X$ is always stronger than $\|\cdot\|_{f''(x)}$, $x \in K$, since $f''(x)$ is a bounded bilinear mapping.)

The following condition ensures that Newton's method can be applied.

Definition 2.2.12. Let $f : K \to \mathbb{R}$ be twice continuously differentiable. We call f *nondegenerate on* K iff $f''(x) \in \mathcal{L}(X, X^*)$ is invertible for every $x \in K$.

Remark 2.2.13. Accordingly, we also speak of *nondegenerate self-concordant functions on* K and *nondegenerate self-concordant barrier functions for* K.

Remark 2.2.14. It follows from the famous Lax-Milgram theorem that $f : K \to \mathbb{R}$ is nondegenerate on K if f is twice continuously differentiable and uniformly convex on K, i.e., if $f''(x)[h,h] \geq \alpha \|h\|_X^2$ holds for all $x \in K$ and all $h \in X$ with a constant $\alpha > 0$, cf. Theorem C.4.15. However, $f(x) = -\ln x$ shows that uniform convexity is not required for f to be nondegenerate. Note also that for finite-dimensional X, pd self-concordance and nondegenerate self-concordance coincide.

Remark 2.2.15. If $f : K \to \mathbb{R}$ is twice continuously differentiable and convex and if there is $x^0 \in K$ such that $f''(x^0) \in \mathcal{L}(X, X^*)$ is invertible, then $f''(x^0)$ is positive definite, $(X, \|\cdot\|_{f''(x^0)})$ is a Hilbert space, and $(X, \|\cdot\|_X)$ is reflexive. In fact, we have $\|h\|_{f''(x^0)} \leq \|f''(x^0)\|^{1/2}_{\mathcal{L}(X,X^*)} \|h\|_X$ and $\|h\|_X \leq \|f''(x^0)^{-1}\|_{\mathcal{L}(X^*,X)} \|f''(x^0)\|^{1/2}_{\mathcal{L}(X,X^*)} \|h\|_{f''(x^0)}$ for every $h \in X$, where we have used the Cauchy-Schwarz inequality. For the application of the Cauchy-Schwarz inequality it suffices that $(\cdot,\cdot)_{f''(x^0)}$ is a positive semidefinite symmetric bilinear form, cf. [MV92, 11.1]. Hence, $f''(x^0)$ is positive definite and $\|\cdot\|_X$ and $\|\cdot\|_{f''(x^0)}$ are equivalent. Therefore, $(X, \|\cdot\|_{f''(x^0)})$ is complete and $(X, \|\cdot\|_X)$ is reflexive, as follows from Theorem C.1.4. This applies, in particular, when we develop barrier methods since then we require $f : K \to \mathbb{R}$ to be a nondegenerate self-concordant barrier function for K. Consequently, it would be interesting to look for weaker assumptions than f being nondegenerate. We suspect that it may be enough to require that Newton's equation is uniquely solvable, i.e., $f'(x) \in \operatorname{ran}(f''(x))$ and $f''(x)$ is injective for $x \in K$, where $\operatorname{ran}(f''(x)) \subset X^*$ denotes the image of $f''(x) \in \mathcal{L}(X, X^*)$. This may be a topic for future research.

A very important quantity in the analysis of interior point methods based on self-concordance is the *Newton decrement*.

Definition 2.2.16. Let $f : K \to \mathbb{R}$ be nondegenerate self-concordant on K. Here and in the following, denote by $n_x \in X$ the Newton step of f at x, i.e., $n_x := -f''(x)^{-1}(f'(x))$. Then $\lambda(x) := \lambda_f(x) := \|n_x\|_{f''(x)}$ is called the *Newton decrement of* f *at* x.

Remark 2.2.17. $\lambda(x)$ is nonnegative with $\lambda(x) = 0$ if and only if x is a stationary point for f, which is equivalent to x being the unique and global minimizer of f on K. The uniqueness follows since $f''(x)$ is positive definite for all $x \in K$ by Remark 2.2.15, which implies strict convexity of f.

The next lemma provides an estimate for the decrease in function value after a suitably damped Newton step.

Lemma 2.2.18. *Let $f : K \to \mathbb{R}$ be a nondegenerate self-concordant barrier function for K. Let $x \in K$ and define $\sigma := \frac{1}{1+\lambda(x)}$. Then there hold $x + \sigma n_x \in K$ and*

$$f(x) - f(x + \sigma n_x) \geq \lambda(x) - \ln(1 + \lambda(x)).$$

Proof. We may assume $n_x \neq 0$ without loss of generality. Since there holds $\|\sigma n_x\|_{f''(x)} = \lambda(x)/(1+\lambda(x)) < 1$, Lemma 2.2.6 implies $x + \sigma n_x \in K$. Defining $h := n_x$ it, hence, suffices to argue that the asserted estimate follows from the finite-dimensional counterpart of Lemma 2.2.18. We have $\sigma < \min\{1, 1/\lambda(x)\}$ and $f''_{x,h}(0) = \|h\|^2_{f''(x)} > 0$. Hence, the Newton step $\tilde{h}$ for $f_{x,h}$ at $t := 0$ is $\tilde{h} = -f'_{x,h}(0)/f''_{x,h}(0)$. Due to $f'_{x,h}(0) = f'(x)[h] = -\|h\|^2_{f''(x)} = -f''_{x,h}(0)$ it holds $\tilde{h} = 1$. Defining $\tilde{\sigma} := 1/(1 + \|\tilde{h}\|_{f''_{x,h}(0)})$ we obtain $\|\tilde{h}\|_{f''_{x,h}(0)} = \|1\|_{f''_{x,h}(0)} = \lambda(x)$, which shows $\tilde{\sigma} = \sigma$ by definition of σ and $\tilde{\sigma}$. In particular, $\tilde{\sigma} < \min\{1, 1/\tilde{\lambda}\}$ is satisfied with $\tilde{\lambda} := \|\tilde{h}\|_{f''_{x,h}(0)}$. The finite-dimensional version of Lemma 2.2.18, which can be found in [NN94, Theorem 2.2.1], yields

$$f_{x,h}(t) - f_{x,h}(t + \tilde{\sigma}\tilde{h}) \geq \tilde{\lambda} - \ln(1 + \tilde{\lambda}). \tag{2.3}$$

We mention that the inequality $\tilde{\sigma} < \min\{1, 1/\tilde{\lambda}\}$ is a prerequisite of the theorem from [NN94]; moreover, the proof of this theorem stays valid if f''' is not continuous since only the inequalities from Lemma 2.2.6 are needed.

Using $t = 0$, $\tilde{h} = 1$, $\tilde{\sigma} = \sigma$, $\tilde{\lambda} = \lambda(x)$, and the definition of $f_{x,h}$ in (2.3) we obtain the assertion. □

We slightly weaken the estimate from the preceding lemma to write it in a form that shows more clearly which decrease we can expect from a damped Newton step. However, in proofs we do not use this weaker estimate but employ the one from Lemma 2.2.18.

Corollary 2.2.19. *In Lemma 2.2.18 we can also estimate as follows:*

$$f(x) - f(x + \sigma n_x) \geq \begin{cases} \frac{\lambda(x)^2}{3} : & \text{if } \lambda(x) \leq \frac{1}{2}, \\ \frac{\lambda(x)^2}{4} + 0.03 : & \text{if } \frac{1}{2} < \lambda(x) < 1, \\ \frac{\lambda(x)-1}{2} + 0.3 : & \text{if } \lambda(x) \geq 1. \end{cases}$$

Proof. Employing the preceding lemma and $1 - \ln 2 > 0.3$ it suffices to show for $t \geq 0$:

$$t - \ln(1+t) \geq \begin{cases} \frac{t^2}{3} : & \text{if } t \leq \frac{1}{2}, \\ \frac{t^2}{4} + 0.032 : & \text{if } \frac{1}{2} < t < 1, \\ \frac{t-1}{2} + 1 - \ln 2 : & \text{if } t \geq 1. \end{cases}$$

We prove these estimate separately.

- Differentiation shows that $t - \ln(1+t) - \frac{t^2}{3}$ is monotone increasing for all t with $0 \leq t \leq \frac{1}{2}$. This implies
$$t - \ln(1+t) - \frac{t^2}{3} \geq 0 - \ln(1+0) - \frac{0^2}{3} = 0$$
for all these t, which proves the first estimate.
- The function $t - \ln(1+t) - \frac{t^2}{4} - 0.032$ is monotone increasing for all t with $0 \leq t \leq 1$. Hence, it holds
$$t - \ln(1+t) - \frac{t^2}{4} - 0.032 \geq \frac{1}{2} - \ln\frac{3}{2} - \frac{1}{16} - 0.032 > 0$$
for all t with $\frac{1}{2} < t < 1$, which establishes the second estimate.

- Again by differentiation we infer that $t-\ln(1+t)-(1-\ln 2+\frac{t-1}{2})$ is monotone increasing for all $t \geq 1$. Thus, we have for these t

$$t-\ln(1+t)-\left(1-\ln 2+\frac{t-1}{2}\right) \geq 1-\ln(1+1)-\left(1-\ln 2+\frac{1-1}{2}\right)=0,$$

which demonstrates that the third estimate is valid. This concludes the proof. □

In the next result we present estimates for the Newton decrement of different points and for $f'(x^+)[n_x]$.

Lemma 2.2.20. *Let $f : K \to \mathbb{R}$ be a nondegenerate self-concordant barrier function for K and let $x \in K$ with $\lambda(x) < 1$ be given. Set $\sigma := \frac{1}{1+\lambda(x)}$ and denote $x^+ := x + n_x$ and $x^\sigma := x + \sigma n_x$. Then there hold $x^+, x^\sigma \in K$ together with the estimates*

$$\lambda(x^+) \leq \left(\frac{\lambda(x)}{1-\lambda(x)}\right)^2, \qquad \lambda(x^\sigma) \leq 2\lambda(x)^2, \quad \textit{and} \quad f'(x^+)[n_x] \leq \frac{\lambda(x)^3}{1-\lambda(x)}.$$

Proof. We number the three asserted estimates as 1), 2), 3). Combining ideas of [JS04, pp. 366-370] and [Nem04, Chapter 2, IX] we demonstrate these three estimates simultaneously. We mention that 1) and 2) are a special case of [NN94, Theorem 2.2.1, (2.2.8)], whose proof is, however, more technical.

We denote by n_x, n_{x^σ} and n_{x^+} the Newton steps at x, x^σ and x^+. Furthermore, we abbreviate $\lambda := \lambda(x)$, $\lambda^\sigma := \lambda(x^\sigma)$ and $\lambda(x^+) := \lambda^+$. Defining $I := [0,1]$ and $\gamma : I \to X$, $\gamma(t) := x + tn_x$ we conclude $\gamma(I) \subset K$ from Lemma 2.2.6. In particular, we have $x^\sigma, x^+ \in K$. Lemma 2.2.6 also implies for all $\tilde{h} \in X$ and all $t \in I$

$$\left|f''(\gamma(t))[\tilde{h},\tilde{h}] - f''(x)[\tilde{h},\tilde{h}]\right| = \left|\|\tilde{h}\|^2_{f''(x+tn_x)} - \|\tilde{h}\|^2_{f''(x)}\right| \leq \left(\frac{1}{(1-t\lambda)^2}-1\right)\|\tilde{h}\|^2_{f''(x)}. \quad (2.4)$$

Here, we used $1-(1-s)^2 \leq \frac{1}{(1-s)^2}-1$ for all $s \in [0,1)$ with $s := t\lambda$. For $h \in X$, set

$$\varphi_h : I \to \mathbb{R}, \quad \varphi_h(t) := (1-t)f'(x)[h] - f'(\gamma(t))[h].$$

Due to (2.4) and the positive definiteness of $f''(x)$ we can apply the generalized Cauchy-Schwarz inequality, see Lemma C.1.5. This yields for all $t \in I$

$$\varphi_h'(t) = f''(x)[n_x,h] - f''(\gamma(t))[n_x,h] \leq r(t)\sqrt{f''(x)[n_x,n_x]}\sqrt{f''(x)[h,h]} = \lambda r(t)\|h\|_{f''(x)}, \quad (2.5)$$

with $r : I \to \mathbb{R}$, $r(t) := \frac{1}{(1-t\lambda)^2}-1$. We now prove 1) and 3). Since we have

$$\varphi_{n_{x^+}}(1) = -f'(x^+)[n_{x^+}] = \left(\lambda^+\right)^2 \qquad \text{and} \qquad \varphi_{n_x}(1) = -f'(x^+)[n_x],$$

it suffices to establish $\varphi_{n_{x^+}}(1) \leq \lambda^+\left(\frac{\lambda}{1-\lambda}\right)^2$ as well as $|\varphi_{n_x}(1)| \leq \frac{\lambda^3}{1-\lambda}$. Since $\varphi'_{n_{x^+}}$ is continuous on I and since $\varphi_{n_{x^+}}(0) = 0$ holds, (2.5) and Lemma 2.2.6 imply 1) via

$$\varphi_{n_{x^+}}(1) = \int_0^1 \varphi'_{n_{x^+}}(t)\,\mathrm{d}t \leq \lambda\|n_{x^+}\|_{f''(x)}\int_0^1 r(t)\,\mathrm{d}t \leq \lambda\,\frac{\lambda^+}{1-\lambda}\,\frac{\lambda}{1-\lambda},$$

and 3) follows from $|\varphi_{n_x}(1)| \leq \lambda^2 \int_0^1 r(t)\,\mathrm{d}t = \frac{\lambda^3}{1-\lambda}$. It remains to prove 2). With the definition of σ we obtain

$$\varphi_{n_{x^\sigma}}(\sigma) = (1-\sigma)f'(x)[n_{x^\sigma}] - f'(x^\sigma)[n_{x^\sigma}] = (\lambda^\sigma)^2 - \frac{\lambda}{1+\lambda} f''(x)[n_x, n_{x^\sigma}].$$

By virtue of the estimate

$$\frac{\lambda}{1+\lambda} f''(x)[n_x, n_{x^\sigma}] \leq \frac{\lambda^2}{1+\lambda} \|n_{x^\sigma}\|_{f''(x)} \leq \frac{\lambda^2}{1+\lambda} \frac{1}{1-\sigma\lambda} \|n_{x^\sigma}\|_{f''(x^\sigma)} = \lambda^2\lambda^\sigma$$

it, thus, suffices to prove $\varphi_{n_{x^\sigma}}(\sigma) \leq \lambda^\sigma\lambda^2$. Analogously as for 1) and 3), this follows from

$$\varphi_{n_{x^\sigma}}(\sigma) \leq \lambda \|n_{x^\sigma}\|_{f''(x)} \int_0^\sigma r(t)\,\mathrm{d}t \leq \frac{\lambda\lambda^\sigma}{1-\sigma\lambda} \left[\frac{1}{\lambda(1-t\lambda)} - t\right]_0^\sigma = \lambda^\sigma\lambda^2. \qquad \square$$

Remark 2.2.21. The second inequality stays valid without the assumption $\lambda(x) < 1$. The proof of this more general version only requires to work with $I = [0, \sigma]$ instead of $I = [0, 1]$ in the proof given above. However, we do not need this generality.

Definition 2.2.22. Let $f : K \to \mathbb{R}$ be a nondegenerate and self-concordant function on K. For $t \geq 0$ we denote $\Lambda(t) := \{x \in K : \lambda(x) \leq t\}$. Moreover, we set $\Lambda := \Lambda(\frac{1}{4})$.

The next result elucidates the convergence behaviour of Newton's method for nondegenerate self-concordant barrier functions. Although the result mimics closely its finite-dimensional counterpart, the proof requires some extra work.

Lemma 2.2.23. *Let $f : K \to \mathbb{R}$ be a nondegenerate self-concordant barrier function for K with $\Lambda \neq \emptyset$. Then there exists a unique and global minimizer $\bar{x} \in K$ of the barrier problem $\min_{x\in K} f(x)$. Moreover, Newton's method with starting point $x^0 \in \Lambda$ generates a sequence $(x^k) \subset \Lambda$ that converges strongly to $\bar{x}$. Also, we have the estimates*

$$\left\|x^0 - \bar{x}\right\|_{f''(x^0)} \leq 10\left(\lambda(x^0)\right)^2 \qquad \text{and} \qquad |f(x^0) - f(\bar{x})| \leq \frac{\left(\lambda(x^0)\right)^2}{1 - \left(\frac{16}{9}\lambda(x^0)\right)^2}.$$

Proof. **Preliminaries:** From Remark 2.2.15 we know that $\|\cdot\|_X$ and $\|\cdot\|_{f''(x^0)}$ are equivalent. Moreover, we deduce from Lemma 2.2.20 that $\lambda(x^k) \leq \frac{1}{4}$ holds true for all k, i.e., $(x^k) \subset \Lambda$. We argue for the case where Newton's method does not terminate finitely; the case of finite termination can be treated analogously but is simpler. Thus, we have $\|n_{x^k}\|_{f''(x^i)} \neq 0$ for all $k, i \in \mathbb{N}_0$. Since f is strictly convex, there exists at most one (necessarily global) minimizer $\bar{x}$.

Part I: We prove existence of $\bar{x}$, $\bar{x} \in K$, $\lim_{k\to\infty} x^k = \bar{x}$, and the first asserted estimate.

Part a: We show by induction for all $k \in \mathbb{N}_0$: $\lambda(x^{k+1}) \leq 4 \cdot (\frac{4}{9})^{k+1} \cdot (\lambda(x^0))^2$. In fact, by virtue of Lemma 2.2.20 and the induction hypothesis we obtain for all $k \in \mathbb{N}$

$$\lambda(x^{k+1}) \leq \left[\frac{\lambda(x^k)}{(1-\lambda(x^k))^2}\right] \cdot \left[4 \cdot \left(\frac{4}{9}\right)^k \cdot \left(\lambda(x^0)\right)^2\right] \leq 4 \cdot \left(\frac{4}{9}\right)^{k+1} \cdot \left(\lambda(x^0)\right)^2.$$

Here, we used $(x^k) \subset \Lambda$. The induction assumption also follows from Lemma 2.2.20.

Part b: We prove by induction for all $k \in \mathbb{N}_0$: $\|n_{x^{k+1}}\|_{f''(x^0)} \leq 4 \cdot \left(\frac{16}{27}\right)^{k+1} \cdot (\lambda(x^0))^2$. In fact, using Lemma 2.2.6 and $\|n_{x^k}\|_{f''(x^i)} \neq 0$ for all $k, i \in \mathbb{N}_0$ we deduce

$$\|n_{x^{k+1}}\|_{f''(x^0)} = \|n_{x^{k+1}}\|_{f''(x^{k+1})} \prod_{i=0}^{k} \frac{\|n_{x^{k+1}}\|_{f''(x^i)}}{\|n_{x^{k+1}}\|_{f''(x^{i+1})}} \leq \lambda(x^{k+1}) \prod_{i=0}^{k} \frac{1}{1 - \|n_{x^i}\|_{f''(x^i)}}.$$

Here, we employed $\|n_{x^i}\|_{f''(x^i)} = \lambda(x^i) \leq \frac{1}{4} < 1$ for all $i \in \mathbb{N}_0$. This also implies

$$\|n_{x^{k+1}}\|_{f''(x^0)} \leq \lambda(x^{k+1}) \prod_{i=0}^{k} \frac{1}{1 - \|n_{x^i}\|_{f''(x^i)}} \leq \lambda(x^{k+1}) \left(\frac{4}{3}\right)^{k+1}.$$

Together with the estimate from part a this concludes part b.

Part c: The estimate in part b shows that (x^k) is a Cauchy sequence with respect to $\|\cdot\|_{f''(x^0)}$ since we have $\|x^n - x^m\|_{f''(x^0)} \leq \sum_{i=m}^{n-1} \|n_{x^i}\|_{f''(x^0)} \leq 4(\lambda(x^0))^2 \sum_{i=m}^{n-1} (\frac{16}{27})^i$ for all $m, n \in \mathbb{N}_0$ with $0 \leq m < n$. Since $\|\cdot\|_{f''(x^0)}$ and $\|\cdot\|_X$ are equivalent, we can define $\bar{x} := \lim_{k\to\infty} x^k$. In particular, (x^k) converges strongly to $\bar{x}$, and for $m = 0$ and $n \to \infty$ we obtain $\|x^0 - \bar{x}\|_{f''(x^0)} \leq 4(\lambda(x^0))^2 \sum_{k=0}^{\infty} (\frac{16}{27})^k \leq 10(\lambda(x^0))^2$, which establishes the first of the two asserted estimates and implies $\bar{x} \in K$ via Lemma 2.2.6 and $x^0 \in \Lambda$.

Part d: By demonstrating $f'(\bar{x}) = 0$ we prove that $\bar{x}$ is a minimizer of f. Due to $|f'(x^k)[h]| = |f''(x^k)[n_{x^k}, h]| \leq \lambda(x^k)\|h\|_{f''(x^k)}$ for all $h \in X$, $k \in \mathbb{N}_0$, and $\lim_{k\to\infty} \lambda(x^k) = 0$, see part a, it suffices to show boundedness of $(\|h\|_{f''(x^k)})$ for a given $h \in X$. Lemma 2.2.6 yields $\|h\|_{f''(x^k)} \leq (\prod_{i=0}^{k-1} \frac{1}{1-\lambda(x^i)})\|h\|_{f''(x^0)}$. Hence, it is sufficient to establish $\sum_{i=0}^{\infty} \ln \frac{1}{1-\lambda(x^i)} < \infty$. From $\lambda(x^i) \leq \frac{1}{4}$ we infer $\frac{1}{1-\lambda(x^i)} \leq 1 + 2\lambda(x^i)$ for all i. Thus, there holds

$$\sum_{i=0}^{\infty} \ln \frac{1}{1 - \lambda(x^i)} \leq \sum_{i=0}^{\infty} \ln(1 + 2\lambda(x^i)) \leq \sum_{i=0}^{\infty} 2\lambda(x^i),$$

where we used $\ln(1 + t) \leq t$. Part a now implies $\sum_{i=0}^{\infty} \ln \frac{1}{1-\lambda(x^i)} < \infty$.

Part II: It remains to establish the second estimate. The convexity of f yields

$$f(x^{k+1}) = f(x^k + n_{x^k}) \geq f(x^k) + f'(x^k)[n_{x^k}] = f(x^k) - \left(\lambda(x^k)\right)^2$$

for all $k \in \mathbb{N}_0$. From Lemma 2.2.20 we infer $\lambda(x^k) \leq (\frac{16}{9}\lambda(x^0))^{2^k-1}\lambda(x^0)$ for all $k \in \mathbb{N}_0$. With $f(x^k) \to f(\bar{x})$ this implies

$$\begin{aligned} f(x^0) - f(\bar{x}) &= \sum_{k=0}^{\infty} \left(f(x^k) - f(x^{k+1})\right) \leq \sum_{k=0}^{\infty} \left(\lambda(x^k)\right)^2 \\ &\leq \left(\lambda(x^0)\right)^2 \sum_{k=0}^{\infty} \left[\left(\frac{16}{9}\lambda(x^0)\right)^{2^k-1}\right]^2 \leq \left(\lambda(x^0)\right)^2 \sum_{k=0}^{\infty} \left[\left(\frac{16}{9}\lambda(x^0)\right)^2\right]^k, \end{aligned}$$

thereby concluding the proof. □

2.3. Self-bounded functions

In addition to the concept of self-concordance we need a property that is called self-boundedness. In this section we introduce and investigate the class of self-bounded functions.

Definition 2.3.1 (Cf. [Jar94])**.** We say that a twice continuously differentiable function $b : K \to \mathbb{R}$ is *self-bounded on* K iff there exists a constant $\vartheta_b \geq 0$ with

$$\left(b'(x)[h]\right)^2 \leq \vartheta_b \cdot b''(x)[h,h]$$

for all $x \in K$ and all $h \in X$. We call ϑ_b the *constant of self-boundedness of* b and also refer to b as a *ϑ_b-self-bounded function.* If b is, in addition, self-concordant on K, we speak of a *ϑ_b-self-concordant function on K.*

Example 2.3.2. The function $-\ln : K \to \mathbb{R}$ is self-bounded on $K = \mathbb{R}_{>0}$ with constant of self-boundedness equal to 1.

Remark 2.3.3. We mention a technicality: We were able to weaken the classical assumption of thrice continuous differentiability for $f : K \to \mathbb{R}$ in the self-concordant case, and required only thrice Fréchet differentiability, instead. In the proof of several results we used that the fundamental theorem of calculus applies to $f''_{x,h}$ on compact intervals since $f'''_{x,h}$ is bounded due to the self-concordance of $f_{x,h}$, cf. Lemma D.0.3. For self-bounded functions we employ the fundamental theorem of calculus for $b'_{x,h}$ in various proofs. Therefore, we require in the definition above that b is twice continuously differentiable. Here, twice Fréchet differentiability would not be sufficient.

As for self-concordant functions we have the following equivalent characterization of self-bounded functions. We recall the definition $I_{x,h} = \{t \in \mathbb{R} : x + th \in K\}$.

Lemma 2.3.4. *A twice continuously differentiable function $b : K \to \mathbb{R}$ is self-bounded on K with constant $\vartheta_b \geq 0$ if and only if for every $x \in K$ and every $h \in X$ the function*

$$b_{x,h} : I_{x,h} \to \mathbb{R}, \quad b_{x,h}(t) := b(x + th)$$

is ϑ_b-self-bounded on $I_{x,h}$, i.e., $\left(b'_{x,h}(t)\right)^2 \leq \vartheta_b \cdot b''_{x,h}(t)$ for all $x \in K$, $h \in X$, and $t \in I_{x,h}$.

Proof. This follows readily from the definition. □

The following lemma presents a geometrical interpretation of self-boundedness.

Lemma 2.3.5. *Let $b : K \to \mathbb{R}$ be twice continuously differentiable and convex, and denote $\overline{B}_x := \{h \in X : \|h\|_{b''(x)} \leq 1\}$ for $x \in K$. Then b is ϑ_b-self-bounded on K if and only if it satisfies for all $x \in K$*

$$\sup_{h \in \overline{B}_x} b'(x)[h] \leq \sqrt{\vartheta_b}.$$

Proof. This follows from the definition of self-boundedness. □

Remark 2.3.6. The previous lemma shows that for $x \in K$ the operator $b'(x)$ is bounded if the seminorm $\|\cdot\|_{b''(x)}$ is used on X, and that the bound is uniform with respect to x. Note that this is not necessarily true if $\|\cdot\|_X$ is used instead of $\|\cdot\|_{b''(x)}$, as we see for $b(x) = -\ln(x)$.

Obviously, every self-bounded function is convex. The converse, however, is not true as follows from the example $b : \mathbb{R} \to \mathbb{R}$, $b(x) := x^2$ for $x \to \infty$. (Yet, note that x^2 is self-bounded on any bounded interval.) The next lemma shows how self-boundedness and convexity are linked.

Lemma 2.3.7. *A twice continuously differentiable function $b : K \to \mathbb{R}$ is self-bounded on K with constant $\vartheta_b > 0$ if and only if the function $\Psi : K \to \mathbb{R}$, $\Psi(x) := e^{-b(x)/\vartheta_b}$ is concave.*

Proof. The function Ψ is concave on K if and only if it holds $\Psi''(x)[h,h] \leq 0$ for all $x \in K$ and all $h \in X$. Computing $\Psi''(x)[h,h]$ we see that this is equivalent to

$$e^{-b(x)/\vartheta_b} \left(\left(\frac{b'(x)[h]}{\vartheta_b} \right)^2 - \frac{b''(x)[h,h]}{\vartheta_b} \right) \leq 0.$$

for all $x \in K$ and all $h \in X$. This implies the assertion. □

Remark 2.3.8. It is easy to see that if Ψ is concave, then so is $-b$. In fact, the preceding lemma states that ϑ_b-self-boundedness of b is equivalent to $-b/\vartheta_b$ being "exponentially concave", i.e., $-b/\vartheta_b$ is not only concave but stays concave even after applying the convex exponential function. The term "exponentially concave" is motivated by the analogue concept of logarithmic convexity, which, for instance, plays a role in characterizing the Gamma function, cf. [Kön04a, Section 17.1 and 17.2].

A large class of self-bounded functions is provided by the following simple construction.

Corollary 2.3.9. *Let $B : X \to \mathbb{R}$ be twice continuously differentiable and concave, and define $K := \{x \in X : B(x) > 0\}$. Suppose that K is nonempty and let $C \geq 0$. Then $b : K \to \mathbb{R}$, $b(x) := -C \ln(B(x))$ is C-self-bounded on the nonempty, open, and convex set K.*

Proof. The function b is well-defined on K, and it is easily seen that K is open and convex. From Lemma 2.3.7 we infer that b is C-self-bounded. □

The next lemma shows how scaling affects self-bounded functions.

Lemma 2.3.10. *Let $b : K \to \mathbb{R}$ be self-bounded on K with constant $\vartheta_b \geq 0$ and let $c \geq 0$ be given. Then $\tilde{b} : K \to \mathbb{R}$, $\tilde{b} := cb$ is self-bounded on K with constant $\vartheta_{\tilde{b}} = c\vartheta_b$.*

Proof. Obviously, there holds

$$\left(\tilde{b}'(x)[h]\right)^2 = \left(cb'(x)[h]\right)^2 \leq c^2 \vartheta_b b''(x)[h,h] = c\vartheta_b \tilde{b}''(x)[h,h]$$

for all $x \in K$ and all $h \in X$. This proves the self-boundedness of $\tilde{b}$ with constant $c\vartheta_b$. □

The following observation on the sum of self-bounded functions is also very basic.

Lemma 2.3.11. *Let $K_1, K_2 \subset X$ be open and convex with $K := K_1 \cap K_2 \neq \emptyset$. Let $b_1 : K_1 \to \mathbb{R}$ and $b_2 : K_2 \to \mathbb{R}$ be self-bounded, with constants ϑ_{b_1} and ϑ_{b_2}, respectively. Then*

$$b : K \to \mathbb{R}, \quad b(x) := b_1(x) + b_2(x)$$

is self-bounded on the nonempty, open, and convex set K, with constant $\vartheta_b = \vartheta_{b_1} + \vartheta_{b_2}$.

Proof. Clearly, K is open and convex, and b is twice continuously differentiable. Moreover, we have for every $x \in K$ and every $h \in X$

$$\left(b'(x)[h]\right)^2 = \left(b_1'(x)[h] + b_2'(x)[h]\right)^2 = \left(b_1'(x)[h]\right)^2 + 2b_1'(x)[h] \cdot b_2'(x)[h] + \left(b_2'(x)[h]\right)^2.$$

From the definition of self-boundedness we infer

$$\left(b'(x)[h]\right)^2 \leq \vartheta_{b_1} b_1''(x)[h,h] + 2\sqrt{\vartheta_{b_1} b_2''(x)[h,h]} \cdot \sqrt{\vartheta_{b_2} b_1''(x)[h,h]} + \vartheta_{b_2} b_2''(x)[h,h].$$

Employing $2\sqrt{a}\sqrt{b} \leq a + b$, which holds for $a, b \geq 0$, we obtain

$$\left(b'(x)[h]\right)^2 \leq \vartheta_{b_1} b_1''(x)[h,h] + \vartheta_{b_1} b_2''(x)[h,h] + \vartheta_{b_2} b_1''(x)[h,h] + \vartheta_{b_2} b_2''(x)[h,h].$$

This implies the assertion. □

Every uniformly convex function with uniformly bounded first derivative is self-bounded.

Lemma 2.3.12. *Let $b : K \to \mathbb{R}$ be twice continuously differentiable with uniformly bounded first derivative, i.e., there exists $C \geq 0$ such that $b'(x)[h] \leq C\|h\|_X$ is satisfied for all $x \in K$ and all $h \in X$. Moreover, let b be uniformly convex with convexity modulus $\alpha > 0$. Then b is self-bounded with constant $\vartheta_b = \frac{C^2}{\alpha}$.*

Proof. Note that the inequality $b'(x)[h] \leq C\|h\|_X$ for all $x \in K$ and all $h \in X$ is equivalent to $|b'(x)[h]| \leq C\|h\|_X$ for all $x \in K$ and all $h \in X$. Fix $x \in K$. There holds $b''(x)[h,h] \geq \alpha \|h\|_X^2$ for all $h \in X$. This implies

$$\left(b'(x)[h]\right)^2 \leq \left(C\|h\|_X\right)^2 \leq \frac{C^2}{\alpha} \cdot b''(x)[h,h]$$

for all $h \in X$, which proves the assertion. □

Remark 2.3.13. For unbounded K the assumption that b has bounded first derivatives on K is very restrictive. However, this is not the case that we are interested in: When we apply the above result, the set K is, in fact, bounded. Thus, the assumption of bounded first derivatives for b on K still allows for great generality in the choice of b.

The next lemma deals with growth rates of self-bounded functions. To state it conveniently we introduce the *Minkowski function*.

Definition 2.3.14. Let $x \in K$. The *Minkowski function of K at x* is defined by

$$\omega_x : K \to [0,1), \quad \omega_x(y) := \inf\left\{t > 0 : \; x + \frac{y-x}{t} \in K\right\}.$$

Remark 2.3.15. Since K is nonempty, open, and convex, ω_x is well-defined and, indeed, satisfies $\omega_x(y) < 1$ for all $y \in K$.

We present the aforementioned result.

Lemma 2.3.16. *Let $b : K \to \mathbb{R}$ be a ϑ_b-self-bounded function. Then the following inequalities hold for all $x, y \in K$:*

$$b'(x)[y-x] \leq \vartheta_b \qquad \text{and} \qquad b(y) - b(x) \leq \vartheta_b \left|\ln\left(1 - \omega_x(y)\right)\right|.$$

Proof. We start by establishing the first estimate. Define $h := y - x$ and consider $b_{x,h}$. We need to show $b'_{x,h}(0) \leq \vartheta_b$ and assume $b'_{x,h}(0) > 0$ without loss of generality. This implies $b'_{x,h}(t) > 0$ for all $t \in I := I_{x,h} \cap [0, \infty)$ via monotonicity of $b'_{x,h}$. Thus, we have $b''_{x,h}(t)/(b'_{x,h}(t))^2 \geq \frac{1}{\vartheta_b}$ for all $t \in I$ by Lemma 2.3.4. Integration yields $b'_{x,h}(0)^{-1} - b'_{x,h}(t)^{-1} \geq \frac{t}{\vartheta_b}$ for all $t \in I$, which establishes $b'_{x,h}(0) \leq \frac{\vartheta_b}{t}$. The assertion follows for $t = 1 \in I$.

We now demonstrate the validity of the second estimate. We define $b_{x,h} : I_{x,h} \to \mathbb{R}$ as above. It follows as in the proof of the first estimate that $b'_{x,h}(t_1) \leq \vartheta_b/(t_2 - t_1)$ holds for all $t_1, t_2 \in I_{x,h}$ with $t_1 < t_2$ and $b'_{x,h}(t_1) > 0$. Obviously, $b'_{x,h}(t_1) \leq \vartheta_b/(t_2 - t_1)$ is also true for all $t_1, t_2 \in I_{x,h}$ with $t_1 < t_2$ and $b'_{x,h}(t_1) \leq 0$. This yields for all $t_2 \in I_{x,h}$ with $t_2 > 1$:

$$b_{x,h}(1) - b_{x,h}(0) = \int_0^1 b'_{x,h}(t)\,\mathrm{d}t \leq \int_0^1 \frac{\vartheta_b}{t_2 - t}\,\mathrm{d}t = \vartheta_b \left|\ln\left(1 - t_2^{-1}\right)\right|.$$

Choosing for t_2 a sequence that converges to the (possibly infinite) supremum of $I_{x,h}$, we obtain $t_2^{-1} \to \omega_x(y) < 1$. By continuity this implies the assertion. □

2.4. Construction of self-concordant, self-bounded barrier functions

In this section we present a sophisticated method to construct self-concordant, self-bounded barrier functions. It works for *appropriate* functions.

Definition 2.4.1 (Cf. [Nem04] and [TN10])**.** A thrice Fréchet differentiable concave function $\mathcal{A} : K \to \mathbb{R}$ is said to be *β-appropriate for $\mathbb{R}_{>0}$ on K* iff there is $\beta \geq 0$ such that

$$\mathcal{A}'''(x)[h,h,h] \leq -3\beta\mathcal{A}''(x)[h,h]$$

is satisfied for all $x \in K$ and all $h \in X$ with $x + h \in K$ and $x - h \in K$.

Remark 2.4.2. The definition from [Nem04] also considers different sets than $\mathbb{R}_{>0}$ and is, therefore, more general. Since we only consider $\mathbb{R}_{>0}$, we may not mention this set and just speak of functions that are *β-appropriate on K*.

Equipped with this terminology we present a result that shows how to create a self-concordant and self-bounded barrier function from an appropriate mapping.

Lemma 2.4.3. *Let $\mathcal{A} : K \to \mathbb{R}$ be β-appropriate for $\mathbb{R}_{>0}$ on K. Let $E := \{x \in K : \mathcal{A}(x) > 0\}$ be nonempty and let $C \in [1, \infty)$ and $\hat{C} \geq \max\{1, \beta^2\}$. Moreover, let $f : K \to \mathbb{R}$ be a self-concordant barrier function for K and a ϑ_f-self-bounded function on E. Then E is nonempty, open, and convex, and*

$$G : E \to \mathbb{R}, \quad G(x) := -C \ln(\mathcal{A}(x)) + \hat{C} f(x)$$

is a ϑ-self-concordant barrier function for E, where ϑ is given by $\vartheta := C + \hat{C}\vartheta_f$.

Remark 2.4.4. Roughly speaking, this result shows that for an appropriate mapping $\mathcal{A}$ it is possible to create a self-concordant and self-bounded barrier function via $x \mapsto -\ln(\mathcal{A}(x))$ if a suitably weighted ϑ_f-self-concordant barrier function f is added. The importance of this result lies in the fact that $x \mapsto -C \ln(\mathcal{A}(x))$ is, in general, *not* self-concordant. (It is, however, self-bounded with constant C, cf. Corollary 2.3.9.)

Remark 2.4.5. The above result is based on [Nem04, Theorem 9.1.1]. We mention that a related, yet in some sense less general version of this result can already be found in [NN94, Proposition 5.1.7], see also the discussion in [TN10, Lemma 2.2].

Remark 2.4.6. The original statement [Nem04, Theorem 9.1.1] is more general than our version: It allows arbitrary self-concordant and self-bounded barrier functions for the composition with $\mathcal{A}$, not only the negative logarithm. However, Lemma 2.4.3 is sufficient for our purposes and its proof is less technical in comparison to the original statement.

Remark 2.4.7. In [Nem04, Theorem 9.1.1] (translated into our terminology) it is required that f is self-bounded on K. This is not necessary in our setting; it suffices to assume that f is self-bounded on E. This may lead to a smaller constant of self-boundedness.

Proof. Clearly, E is open and convex. To show that G is a barrier function for E, let $(x^k) \subset E$ satisfy $\lim_{k\to\infty} x^k = \bar{x} \in \partial E$. We need to establish $\lim_{k\to\infty} G(x^k) = \infty$. Due to $(x^k) \subset K$ there either holds $\bar{x} \in \partial K$ or $\bar{x} \in K$. If $\bar{x} \in K$ holds, we have $\lim_{k\to\infty} \mathcal{A}(x^k) = \mathcal{A}(\bar{x}) = 0^+$ since $\mathcal{A}$ is continuous on K and since $\bar{x} \in \partial E$ is valid. This implies $-C \ln(\mathcal{A}(x^k)) \to \infty$ for $k \to \infty$. Since $\lim_{k\to\infty} \hat{C} f(x^k) = \hat{C} f(\bar{x})$ is true, we obtain $\lim_{k\to\infty} G(x^k) = \infty$. In the case $\bar{x} \in \partial K$ we have $f(x^k) \to \infty$ for $k \to \infty$. Thus, it suffices to show that $(\mathcal{A}(x^k))$ is bounded from above. Since (x^k) is convergent, this follows from $\mathcal{A}(x^k) \leq \mathcal{A}(x^0) + \mathcal{A}'(x^0)[x^k - x^0]$.

It remains to establish that G is ϑ-self-concordant on E. Since $-C \ln(\mathcal{A}(\cdot))$ is self-bounded on E with constant C due to Corollary 2.3.9, and since $\hat{C} f$ is self-bounded with constant $\hat{C}\vartheta_f$ on E, Lemma 2.3.11 yields that their sum G is self-bounded with constant $C + \hat{C}\vartheta_f$, as asserted. Therefore, it only remains to show that G is self-concordant, which we do by following closely the proof of [Nem04, Theorem 9.1.1]. Since f and $\mathcal{A}$ are thrice Fréchet differentiable, it follows by virtue of Corollary C.2.10 that G is thrice Fréchet differentiable. We fix $x \in E$ and $h \in X$ and show $|G'''(x)[h,h,h]| \leq 2(G''(x)[h,h])^{3/2}$. This suffices. Let us denote

$$a := \mathcal{A}(x), \quad a' := \mathcal{A}'(x)[h], \quad a'' := \mathcal{A}''(x)[h,h], \quad \text{and} \quad a''' := \mathcal{A}'''(x)[h,h,h].$$

A simple computation shows

$$G''(x)[h,h] = C\left(\frac{(a')^2}{a^2} - \frac{a''}{a}\right) + \hat{C} f''(x)[h,h]$$

and

$$G'''(x)[h,h,h] = C\left(-2\frac{(a')^3}{a^3} + 3\frac{a'a''}{a^2} - \frac{a'''}{a}\right) + \hat{C}f'''(x)[h,h,h].$$

We now establish

$$3\beta a''\sqrt{f''(x)[h,h]} \leq a''' \leq -3\beta a''\sqrt{f''(x)[h,h]}. \tag{2.6}$$

Let $t \in \mathbb{R}$ with $|t|\sqrt{f''(x)[h,h]} < 1$ be given and define $h_t := th$. Then it holds $f''(x)[h_t,h_t] < 1$. With Lemma 2.2.6 it follows $x+h_t \in K$ and $x-h_t \in K$, where we used that f is a self-concordant barrier function for K. Therefore, the definition of appropriateness implies

$$t^3 a''' = \mathcal{A}'''(x)[h_t,h_t,h_t] \leq -3\beta\mathcal{A}''(x)[h_t,h_t] = -3\beta t^2 a''$$

for all t with $|t|\sqrt{f''(x)[h,h]} < 1$. If $f''(x)[h,h] = 0$, then this implies (2.6) for $t \to \pm\infty$. For $f''(x)[h,h] \neq 0$, (2.6) follows from this via $t \to (1/\sqrt{f''(x)[h,h]})^-$ and $t \to (-1/\sqrt{f''(x)[h,h]})^+$. Since f is self-concordant on E, we have

$$G'''(x)[h,h,h] \leq 2C\left(-\frac{(a')^3}{a^3} + \frac{3}{2}\frac{a'a''}{a^2} - \frac{3}{2}\frac{\beta}{\sqrt{\hat{C}}}\frac{a''}{a}\sqrt{\hat{C}f''(x)[h,h]}\right) + 2\frac{\left(\hat{C}f''(x)[h,h]\right)^{\frac{3}{2}}}{\sqrt{\hat{C}}}.$$

This implies with $\sqrt{\hat{C}} \geq \beta$ and $\sqrt{\hat{C}}, \sqrt{C} \geq 1$ that

$$|G'''(x)[h,h,h]| \leq 2\left(\frac{|\sqrt{C}a'|^3}{a^3} + \frac{3}{2}\frac{|Ca''|}{a}\left(\frac{|\sqrt{C}a'|}{a} + \sqrt{\hat{C}f''(x)[h,h]}\right)\right) + 2\left(\hat{C}f''(x)[h,h]\right)^{\frac{3}{2}}$$

holds. Setting $r_1 := |\sqrt{C}a'|/a \geq 0$, $r_2 := \sqrt{|Ca''|/a} \geq 0$, and $r_3 := \sqrt{\hat{C}f''(x)[h,h]} \geq 0$, we obtain

$$G''(x)[h,h] = r_1^2 + r_2^2 + r_3^2 \quad \text{and} \quad |G'''(x)[h,h,h]| \leq 2\left(r_1^3 + \frac{3}{2}r_2^2(r_1+r_3) + r_3^3\right).$$

To conclude the proof it, thus, suffices to show that

$$r_1^3 + \frac{3}{2}r_2^2(r_1+r_3) + r_3^3 \leq (r_1^2+r_2^2+r_3^2)^{\frac{3}{2}}$$

holds for any given $r_1, r_2, r_3 \geq 0$. For $r_1 = r_2 = r_3 = 0$, this is clear. Defining $c := (r_1^2+r_2^2+r_3^2)^{-1}$ and multiplying the inequality with $c^{\frac{3}{2}}$, we see that it is enough to establish

$$r_1^3 + \frac{3}{2}r_2^2(r_1+r_3) + r_3^3 \leq 1$$

for any given $r_1, r_2, r_3 \geq 0$ with $r_1^2 + r_2^2 + r_3^2 = 1$. We have

$$\begin{aligned} r_1^3 + \frac{3}{2}r_2^2(r_1+r_3) + r_3^3 &= (r_1+r_3)\left(r_1^2 + r_3^2 - r_1r_3 + \frac{3}{2}r_2^2\right) \\ &= (r_1+r_3)\left(\frac{3}{2}\left(r_1^2+r_2^2+r_3^2\right) - \frac{1}{2}(r_1+r_3)^2\right) \\ &= \frac{1}{2}(r_1+r_3)\left(3 - (r_1+r_3)^2\right) \leq 1. \end{aligned}$$

To derive the inequality in this estimate we used $\max_{t\geq 0} t(3-t^2) = 2$. □

2.5. Theoretical background for barrier methods

In this section we provide the theoretical foundation for solving

$$\min_{x\in X} j(x) \quad \text{s.t.} \quad x \in M \tag{$\mathrm{P_{SC}}$}$$

via barrier methods that are based on the theory developed so far. Here, $j : M \to \mathbb{R}$. Further details of this problem are fixed after the following definition.

Definition 2.5.1. The optimal value of ($\mathrm{P_{SC}}$) is denoted by $\bar{j} := \inf_{x\in M} j(x)$.

To tackle ($\mathrm{P_{SC}}$) by barrier methods we require the following assumption.

Assumption 2.5.2.

- *We assume $K \subset M \subset \overline{K} \subset X$, where K is nonempty, open, and convex, and $(X, \|\cdot\|_X)$ is a Banach space.*
- *$j : M \to \mathbb{R}$ is continuous. Its restriction to K is thrice Fréchet differentiable and convex.*
- *There exists $\mu_s > 0$ such that for every $\mu \in I_s := (0, \mu_s]$ the functional*

  $$f_\mu : K \to \mathbb{R}, \quad f_\mu(x) := \frac{j(x)}{\mu} + b(x)$$

  is a nondegenerate self-concordant barrier function for K, where $b : K \to \mathbb{R}$ is ϑ_b-self-bounded with $\vartheta_b \geq 1$. For every $\mu \in I_s$, $x \in K$ we define $\lambda_\mu(x) := \lambda_{f_\mu}(x)$. Moreover, we set $\Lambda_\mu(t) := \{x \in K : \lambda_\mu(x) \leq t\}$ and $\Lambda_\mu := \Lambda_\mu(\frac{1}{4})$.
- *There holds $\Lambda_{\mu_s} \neq \emptyset$.*

Remark 2.5.3. Since some of the results in this section do not require the above assumption, we explicitly mention for each result whether Assumption 2.5.2 is imposed or not.

Remark 2.5.4. Problem ($\mathrm{P_{SC}}$) looks a bit strange in the case that M is not closed. The choice $M = \overline{K}$ is natural. For example, the problem $\min_{x\in\mathbb{R}^n} j(x)$ s.t. $g(x) \leq 0$ with j and g (componentwise) convex has $M = \overline{K}$ with $K = \{x \in \mathbb{R}^n : g(x) < 0\}$ if K is nonempty (Slater's condition), cf. Lemma C.4.2. However, this problem can be reformulated as $\min_{x\in\mathbb{R}^n} -\ln(j(\hat{x}) - j(x))$ s.t. $g(x) \leq 0$, $j(x) < j(\hat{x})$, which gives $M = \{x \in \mathbb{R}^n : g(x) \leq 0,\ j(x) < j(\hat{x})\}$ with $K \subset M \subset \overline{K}$ for $K = \{x \in \mathbb{R}^n : g(x) < 0,\ j(x) < j(\hat{x})\}$. If $g(\hat{x}) \leq 0$ holds and $j(\hat{x})$ is *not* the optimal value of the original problem, then this reformulation possesses the same global minimizers (if any) as the original problem, but $M \neq \overline{K}$, cf. again Lemma C.4.2. When we deal with optimal control, such a reformulation will turn out to be advantageous for the overall convergence rate. To cover this case as well as the standard formulation, we work with the set M rather than with $\overline{K}$.

Remark 2.5.5. Assumption 2.5.2 implies that we explicitly know a self-concordant barrier function and a self-bounded function for K, which may not be true for an arbitrary convex set K.

Remark 2.5.6. We stress that $\frac{j}{\mu}$ does not have to be self-bounded in the above assumption, only b. Also, b does not have to be self-concordant or a barrier function for K, only f_μ, $\mu \in I_s$, does.

Remark 2.5.7. In fact, $\vartheta_b \geq 1$ necessarily holds true if b is a self-bounded, self-concordant barrier function for $K \neq X$. By considering $b_{x,h}$ with $I_{x,h} \neq \mathbb{R}$, this can be proven as in [NN94, Section 2.3.4]. We point out that b is not necessarily self-concordant or a barrier in Assumption 2.5.2. However, the impact of the assumption $\vartheta_b \geq 1$ on the convergence results to come is insignificant, anyway.

Remark 2.5.8. The assumption $\Lambda_{\mu_s} \neq \emptyset$ is, for instance, valid if a minimizer of $f_{\mu_s} : K \to \mathbb{R}$ exists. Since f_{μ_s} is a nondegenerate self-concordant barrier function for K, the existence of a minimizer follows if f_{μ_s} possesses at least one nonempty and bounded lower level set. This can be proven analogously to Corollary C.4.6. Note that the reflexivity of X required in this corollary follows from the nondegenerateness of f_{μ_s}, see Remark 2.2.15. In particular, the existence of a minimizer follows if K is bounded or if f_{μ_s} is uniformly convex. The fact that uniform convexity implies boundedness of lower level sets is proven in Lemma C.4.13.

It is important to understand how $\lambda_\mu(x)$ changes when x is updated to the next Newton iterate x^+ and μ is updated to $\beta\mu < \mu$.

Lemma 2.5.9. *Let Assumption 2.5.2 hold. Fix $\theta \in (0, \frac{1}{4}]$ and $\delta \in [0, \sqrt{\vartheta_b})$, and let $\mu \in I_s$ and $x \in \Lambda_\mu(\theta)$. Choose $\beta \in (0, 1]$ with $\beta \geq 1 - \frac{\delta}{\sqrt{\vartheta_b}}$ and set $x^+ := x + n_x$ and $\mu_+ := \beta\mu$, where n_x denotes the Newton step for f_μ at x. Then there holds*

$$\lambda_{\mu_+}(x^+) \leq \frac{\frac{\theta^2}{(1-\theta)^2} + \delta}{\beta}.$$

In particular, we have $x^+ \in \Lambda_{\mu_+}(\theta)$ for all $\delta \geq 0$ with $\delta \leq \frac{\theta(1-\frac{\theta}{(1-\theta)^2})}{1+\frac{\theta}{\sqrt{\vartheta_b}}}$.

Remark 2.5.10. $\lambda_{\mu_+}(x^+)$ is well-defined since $x \in \Lambda_\mu(\theta)$ implies $x^+ \in K$, cf. Lemma 2.2.6.

Remark 2.5.11. It follows from $\theta \leq \frac{1}{4}$ that there exist $\delta > 0$ that satisfy $\delta \leq \frac{\theta\left(1-\frac{\theta}{(1-\theta)^2}\right)}{1+\frac{\theta}{\sqrt{\vartheta_b}}}$.

Proof. The second estimate follows from the first by a simple computation. We now establish the first estimate. To do so, let us assume without loss of generality $\lambda_{\mu_+}(x^+) > 0$. Fix $\tilde{x} \in K$ and $h \in X$. Let $n_{\tilde{x}}^+$ and $n_{\tilde{x}}$ denote the Newton steps for f_{μ_+} and f_μ at $\tilde{x}$. We have

$$f''_{\mu_+}(\tilde{x})[n_{\tilde{x}}^+, h] = -f'_{\mu_+}(\tilde{x})[h] = -\frac{j'(\tilde{x})[h]}{\mu_+} - b'(\tilde{x})[h].$$

Furthermore, it holds

$$-\frac{\mu}{\mu_+} f''_\mu(\tilde{x})[n_{\tilde{x}}, h] = \frac{j'(\tilde{x})[h]}{\mu_+} + \frac{\mu}{\mu_+} b'(\tilde{x})[h].$$

This shows

$$f''_{\mu_+}(\tilde{x})[n_{\tilde{x}}^+, h] - \frac{\mu}{\mu_+} f''_\mu(\tilde{x})[n_{\tilde{x}}, h] = \left(\frac{\mu}{\mu_+} - 1\right) b'(\tilde{x})[h].$$

Inserting $\tilde{x} = x^+$ and $h = n^+_{x^+}$ into this equality, where $n^+_{x^+}$ denotes the Newton step for f_{μ_+} at x^+, we obtain

$$\lambda_{\mu_+}(x^+)^2 = \frac{\mu}{\mu_+} f''_\mu(x^+)[n_{x^+}, n^+_{x^+}] + \left(\frac{\mu}{\mu_+} - 1\right) b'(x^+)[n^+_{x^+}].$$

The Cauchy-Schwarz inequality and self-boundedness of b yield

$$\lambda_{\mu_+}(x^+)^2 \leq \frac{1}{\beta}\|n_{x^+}\|_{f''_\mu(x^+)}\|n^+_{x^+}\|_{f''_\mu(x^+)} + \frac{1}{\beta}\sqrt{\vartheta_b}\,|1-\beta|\,\sqrt{b''(x^+)[n^+_{x^+}, n^+_{x^+}]}. \tag{2.7}$$

Using the convexity of j and $\mu_+ \leq \mu$ this implies

$$\lambda_{\mu_+}(x^+)^2 \leq \frac{1}{\beta}\lambda_\mu(x^+)\lambda_{\mu_+}(x^+) + \frac{1}{\beta}\sqrt{\vartheta_b}\,|1-\beta|\,\lambda_{\mu_+}(x^+).$$

Dividing by $\lambda_{\mu_+}(x^+) > 0$ the assertion follows from Lemma 2.2.20 with $x \in \Lambda_\mu(\theta)$. □

Corollary 2.5.12. *Let Assumption 2.5.2 hold. Then we have $\Lambda_\mu \neq \emptyset$ for all $\mu \in I_s$.*

Proof. Due to Assumption 2.5.2 there exists $x \in K$ with $\lambda_{\mu_s}(x) \leq \frac{1}{4}$. Hence, the previous lemma applied with $\mu = \mu_s$, $\theta = \frac{1}{4}$, $\delta = \frac{\theta(1-\frac{\theta}{(1-\theta)^2})}{1+\frac{\theta}{\sqrt{\vartheta_b}}} > 0$, and $\beta \in [1-\frac{\delta}{\sqrt{\vartheta_b}}, 1)$ yields $\Lambda_\mu \neq \emptyset$ for all $\mu \in [(1-\frac{\delta}{\sqrt{\vartheta_b}})\mu_s, \mu_s]$. Applying this lemma again we obtain $\Lambda_\mu \neq \emptyset$ for all $\mu \in [(1-\frac{\delta}{\sqrt{\vartheta_b}})^2\mu_s, \mu_s]$. By iteration of this argument we, thus, conclude $\Lambda_\mu \neq \emptyset$ for all $\mu \in (0, \mu_s] = I_s$. □

Corollary 2.5.13. *Let Assumption 2.5.2 hold. Then f_μ possesses exactly one minimizer for every $\mu \in I_s$, denoted by $\bar{x}_\mu \in K$. In addition, this minimizer is global.*

Proof. For every fixed $\mu \in I_s$, Corollary 2.5.12 shows that there exists $x \in K$ with $\lambda_\mu(x) \leq \frac{1}{4}$. Hence, Lemma 2.2.23 implies the assertion. □

Definition 2.5.14. We call $I_s \ni \mu \mapsto \bar{x}_\mu \in K \subset X$ the *central path*.

For later use in phase one methods we need the following version of the preceding lemma.

Lemma 2.5.15. *Let Assumption 2.5.2 hold, but with $I_s = (0, \mu_s]$ replaced by $I_s = [\mu_s, \infty)$. Furthermore, let j in Assumption 2.5.2 be a linear function. Fix $\theta \in (0, \frac{1}{4}]$ and $\delta \in [0, \sqrt{\vartheta_b})$, and let $\mu \in I_s$ and $x \in \Lambda_\mu(\theta)$. Choose $\beta \in [1, \infty)$ with $\beta \leq 1 + \frac{\delta}{\sqrt{\vartheta_b}}$. Set $x^+ := x + n_x$ and $\mu_+ := \beta\mu$, where n_x denotes the Newton step for f_μ at x. Then there holds*

$$\lambda_{\mu_+}(x^+) \leq \frac{\frac{\theta^2}{(1-\theta)^2} + \delta}{\beta}.$$

In particular, we have $x^+ \in \Lambda_{\mu_+}(\theta)$ for all $\delta \geq 0$ with $\delta \leq \frac{\theta(1-\frac{\theta}{(1-\theta)^2})}{1-\frac{\theta}{\sqrt{\vartheta_b}}}$.

Remark 2.5.16. We have $1 - \frac{\theta}{\sqrt{\vartheta_b}} > 0$ due to $\vartheta_b \geq 1$.

Remark 2.5.17. The assumption that the objective j is linear may seem to be a great restriction. However, this is not the case since the objective that is employed in phase one is not the objective of ($\mathrm{P_{SC}}$) but, in fact, a linear function, as we will see when we deal with phase one.

Proof. The second estimate is implied by the first. The first estimate can be proven as in Lemma 2.5.9. The linearity of j is needed in the line where it says "Using the convexity of j [...]" to be able to still conclude $\|x^+\|_{f''_\mu(x^+)} \leq \|x^+\|_{f''_{\mu_+}(x^+)}$. □

The next two lemmas provide estimates for the objective function j. The first lemma deals with the difference between $x \in \Lambda_\mu$ and $\bar{x}_\mu$.

Lemma 2.5.18. *Let Assumption 2.5.2 hold. Fix $\mu \in I_s$ and let $x^0 \in \Lambda_\mu$. Then there holds*

$$\left|j(x^0) - j(\bar{x}_\mu)\right| \leq \frac{\lambda_\mu(x^0)}{1 - \frac{16}{9}\lambda_\mu(x^0)} \cdot \frac{\sqrt{\vartheta_b} + (\lambda_\mu(x^0))^2}{1 - \lambda_\mu(x^0)} \cdot \mu.$$

Proof. We set $x := x^0$ and denote by n_x the Newton step at x. First, we prove

$$\frac{|j(x + n_x) - j(x)|}{\mu} \leq \frac{\lambda_\mu(x)^3 + \sqrt{\vartheta_b} \cdot \lambda_\mu(x)}{1 - \lambda_\mu(x)}. \tag{2.8}$$

Since j is convex on K, we have

$$j'(x)[n_x] \leq j(x + n_x) - j(x) \leq j'(x + n_x)[n_x]. \tag{2.9}$$

From the structure of f_μ we derive $\frac{j'(x)[n_x]}{\mu} = -\lambda_\mu(x)^2 - b'(x)[n_x]$. Self-boundedness of b and convexity of j imply $\frac{j'(x)[n_x]}{\mu} \geq -\lambda_\mu(x)^2 - \sqrt{\vartheta_b \cdot f''_\mu(x)[n_x, n_x]}$. This establishes

$$\frac{j'(x)[n_x]}{\mu} \geq -\lambda_\mu(x)^2 - \sqrt{\vartheta_b} \cdot \lambda_\mu(x). \tag{2.10}$$

Analogously, we obtain

$$\frac{j'(x + n_x)[n_x]}{\mu} \leq f'_\mu(x + n_x)[n_x] + \sqrt{\vartheta_b}\|n_x\|_{f''_\mu(x+n_x)}.$$

Lemma 2.2.20 yields $f'_\mu(x + n_x)[n_x] \leq \frac{\lambda_\mu(x)^3}{1-\lambda_\mu(x)}$, while Lemma 2.2.6 provides the estimate $\|n_x\|_{f''_\mu(x+n_x)} \leq \frac{\lambda_\mu(x)}{1-\lambda_\mu(x)}$. Together, we deduce

$$\frac{j'(x + n_x)[n_x]}{\mu} \leq \frac{\lambda_\mu(x)^3 + \sqrt{\vartheta_b} \cdot \lambda_\mu(x)}{1 - \lambda_\mu(x)}.$$

In combination with (2.10) this estimate implies via (2.9)

$$\frac{|j(x + n_x) - j(x)|}{\mu} \leq \max\left\{\lambda_\mu(x)^2 + \sqrt{\vartheta_b} \cdot \lambda_\mu(x),\ \frac{\lambda_\mu(x)^3 + \sqrt{\vartheta_b} \cdot \lambda_\mu(x)}{1 - \lambda_\mu(x)}\right\}.$$

Using $\vartheta_b \geq 1$ it is readily shown that the maximum in the above expression is given by $\frac{\lambda_\mu(x)^3+\sqrt{\vartheta_b}\cdot\lambda_\mu(x)}{1-\lambda_\mu(x)}$. In conclusion, we proved (2.8). To obtain the assertion we apply Newton's method to f_μ with starting point x^0. This yields a sequence $(x^k) \subset \Lambda_\mu$ with $\lim_{k\to\infty} x^k = \bar{x}_\mu \in K$, see Lemma 2.2.23. We only argue for the case that (x^k) is infinite; the finite case can be treated similarly. Continuity of j implies $\lim_{k\to\infty} j(x^k) = j(\bar{x}_\mu)$. Also, we infer from Lemma 2.2.20 via $t \geq \left(\frac{t}{1-t}\right)^2$ for $t \in [0, \frac{1}{4}]$ that $\lambda_\mu(x^k) \leq \lambda_\mu(x^0)$ holds for all $k \in \mathbb{N}_0$. In combination with (2.8) this implies

$$\frac{|j(x^0) - j(\bar{x}_\mu)|}{\mu} = \frac{1}{\mu} \cdot \left| \sum_{k=0}^{\infty} \left(j(x^k) - j(x^{k+1}) \right) \right| \leq \frac{\lambda_\mu(x^0)^2 + \sqrt{\vartheta_b}}{1 - \lambda_\mu(x^0)} \cdot \sum_{k=0}^{\infty} \lambda_\mu(x^k). \tag{2.11}$$

From Lemma 2.2.20 we infer $\lambda_\mu(x^k) \leq (\frac{16}{9}\lambda_\mu(x^0))^{2^k-1}\lambda_\mu(x^0)$ for all $k \in \mathbb{N}_0$. Therefore, we have

$$\sum_{k=0}^{\infty} \lambda_\mu(x^k) \leq \sum_{k=0}^{\infty} \left[\left(\frac{16}{9} \lambda_\mu(x^0) \right)^{2^k-1} \cdot \lambda_\mu(x^0) \right] \leq \frac{\lambda_\mu(x^0)}{1 - \frac{16}{9}\lambda_\mu(x^0)}. \tag{2.12}$$

Together, (2.11) and (2.12) establish the assertion. □

For points on the central path we have the following estimate.

Lemma 2.5.19. *Let Assumption 2.5.2 hold. Let $\mu \in I_s$. Then we have for all $x \in M$*

$$j(\bar{x}_\mu) - j(x) \leq \mu\vartheta_b.$$

In particular, Problem ($\mathrm{P_{SC}}$) *is bounded from below, i.e. $\bar{j} > -\infty$, and there holds*

$$\left| j(\bar{x}_\mu) - \bar{j} \right| \leq \mu\vartheta_b.$$

Remark 2.5.20. Of course, if j possesses a minimizer $\bar{x}$ on M, then we have $\bar{j} = j(\bar{x})$.

Proof. Obviously, the second estimate as well as boundedness of $\bar{j}$ from below follow from the first. Hence, it only remains to establish the first estimate. Note that it suffices to prove this estimate on K since by continuity of j on M and by $M \subset \overline{K}$ it extends to M. To this end, fix $\mu \in I_s$ and $x \in K$. The assertion follows from the convexity of j, $f'_\mu(\bar{x}_\mu) = 0$, and Lemma 2.3.16:

$$j(\bar{x}_\mu) - j(x) \leq j'(\bar{x}_\mu)[\bar{x}_\mu - x] = \mu b'(\bar{x}_\mu)[x - \bar{x}_\mu] \leq \mu\vartheta_b. \qquad \square$$

For long step based barrier methods we need the following estimate.

Lemma 2.5.21. *Let Assumption 2.5.2 hold. Let $\mu \in I_s$ and $\beta \in (0, 1]$. Set $\mu_+ := \beta\mu$. Let $x \in \Lambda_\mu(\theta)$ with $\theta \in (0, \frac{1}{4}]$. Denote $f_+ := f_{\mu_+}$ and $\bar{x}_+ := \bar{x}_{\mu_+}$. Then we have*

$$f_+(x) - f_+(\bar{x}_+) \leq \frac{10\theta^3}{\beta} + \frac{1-\beta}{\beta}\left(10\theta^2\sqrt{\vartheta_b} + \vartheta_b\right) \leq \frac{\sqrt{\vartheta_b} + \vartheta_b}{\beta}.$$

Proof. Let $\bar{x}$ denote the minimizer of $f := f_\mu$. Furthermore, let n_x and n_x^+ denote the Newton steps for f_μ and f_{μ_+} at x. Set $h := x - \bar{x}$. Then it holds

$$f_+(x) - f_+(\bar{x}) \leq f'_+(x)[h] = -f''_+(x)[n_x^+, h] = -\frac{\mu}{\mu_+} f''(x)[n_x, h] + \left(1 - \frac{\mu}{\mu_+}\right) b'(x)[h]. \quad (2.13)$$

This implies

$$f_+(x) - f_+(\bar{x}) \leq \frac{1}{\beta}\,\theta\,\|h\|_{f''(x)} + \frac{1-\beta}{\beta}\sqrt{\vartheta_b}\,\|h\|_{f''(x)}.$$

Lemma 2.2.23 provides $\|h\|_{f''(x)} \leq 10\lambda(x)^2$. Hence, we have

$$f_+(x) - f_+(\bar{x}) \leq \frac{10\theta^3}{\beta} + \frac{1-\beta}{\beta}\,10\theta^2\sqrt{\vartheta_b} \leq \frac{\sqrt{\vartheta_b}}{\beta}.$$

Due to $j(\bar{x}) - j(\bar{x}_+) \leq j'(\bar{x})[\bar{x} - \bar{x}_+] = \mu b'(\bar{x})[\bar{x}_+ - \bar{x}]$ we deduce from Lemma 2.3.16

$$f_+(\bar{x}) - f_+(\bar{x}_+) = \frac{j(\bar{x}) - j(\bar{x}_+)}{\mu_+} + b(\bar{x}) - b(\bar{x}_+) \leq \left(\frac{\mu}{\mu_+} - 1\right) b'(\bar{x})[\bar{x}_+ - \bar{x}] \leq \frac{1-\beta}{\beta}\,\vartheta_b.$$

Together with (2.13) the assertion follows by use of the triangle inequality. □

For a successful phase one we require that for a fixed $\mu_0 \in I_s$, the function f_{μ_0} from Assumption 2.5.2 is $\vartheta_{f_{\mu_0}}$-self-bounded on K with $\vartheta_{f_{\mu_0}} \geq 1$. We now collect three results that ensure this.

Lemma 2.5.22. *Let Assumption 2.5.2 hold. In addition, let j be self-bounded with constant ϑ_j. Then for every $\mu_0 \in I_s$, f_{μ_0} is self-bounded with constant $\vartheta_{f_{\mu_0}} = \frac{\vartheta_j}{\mu_0} + \vartheta_b$.*

Proof. This follows from Lemma 2.3.10, Lemma 2.3.11, and Assumption 2.5.2. □

Lemma 2.5.23. *Let Assumption 2.5.2 hold. In addition, let $j : K \to \mathbb{R}$ have a uniformly bounded first derivative and be uniformly convex with modulus $\alpha > 0$. Then f_{μ_0} is self-bounded for every $\mu_0 \in I_s$ with constant $\frac{C^2}{\alpha\mu_0} + \vartheta_b$, where $C > 0$ satisfies $\sup_{x\in K}\|j'(x)\|_{X^*} \leq C$.*

Proof. Using Lemma 2.3.12 this is a consequence of Lemma 2.5.22. □

We can also deal with the case where j is neither self-bounded nor uniformly convex. In order to still infer that f_{μ_0} is self-bounded, we require additionally that b is uniformly convex.

Lemma 2.5.24. *Let Assumption 2.5.2 hold. In addition, let $j : K \to \mathbb{R}$ have a uniformly bounded first derivative and let $b : K \to \mathbb{R}$ be uniformly convex with modulus $\alpha > 0$. Set $C := \sup_{x\in K}\|j'(x)\|_{X^*} < \infty$. Then for every $\mu_0 \in I_s$ the function f_{μ_0} is self-bounded with constant $\vartheta_{f_{\mu_0}} = 2\left(\frac{C^2}{\alpha\mu_0^2} + \vartheta_b\right)$.*

Proof. Fix $\mu_0 > 0$. By definition we have $|j'(x)[h]| \leq C \|h\|_X$ for all $x \in K$ and all $h \in X$. Employing Young's inequality, the self-boundedness of b with constant ϑ_b, the uniform convexity of b, and the convexity of j/μ_0, we have for all $x \in K$ and all $h \in X$

$$\begin{aligned}\left(\frac{j'(x)[h]}{\mu_0} + b'(x)[h]\right)^2 &\leq 2\left(\frac{j'(x)[h]}{\mu_0}\right)^2 + 2\left(b'(x)[h]\right)^2 \leq 2\left(\frac{C\|h\|_X}{\mu_0}\right)^2 + 2\vartheta_b b''(x)[h,h] \\ &\leq \frac{2C^2}{\alpha\mu_0^2}\alpha\|h\|_X^2 + 2\vartheta_b b''(x)[h,h] \leq 2\left(\frac{C^2}{\alpha\mu_0^2} + \vartheta_b\right) b''(x)[h,h] \\ &\leq 2\left(\frac{C^2}{\alpha\mu_0^2} + \vartheta_b\right)\left(\frac{j''(x)[h,h]}{\mu_0} + b''(x)[h,h]\right). \qquad \square\end{aligned}$$

To derive complexity estimates for phase one algorithms we introduce the following definition.

Definition 2.5.25 (Cf. [Ren01])**.** Suppose that $K \subset X$ is nonempty, open, convex, and bounded. Let $x \in K$. For $y \in X \setminus \{0\}$ define $l_x(y) > 0$ through

$$l_x(y) := \sup\{t > 0 : x + ty \in K\}.$$

Furthermore, define the *symmetry of K about x in direction $y \in X \setminus \{0\}$* by

$$\operatorname{sym}_x(y) := \min\left\{\frac{l_x(y)}{l_x(-y)}, \frac{l_x(-y)}{l_x(y)}\right\}.$$

Eventually, define the *symmetry of K about x* by

$$\operatorname{sym}(x, K) := \inf_{y \in X\setminus\{0\}} \operatorname{sym}_x(y).$$

Remark 2.5.26. Since K is nonempty, open, and bounded, $l_x(y)$ and $\operatorname{sym}_x(y)$ are well-defined and positive. The well-definition of $\operatorname{sym}(x, K)$ is then obvious.

Remark 2.5.27. The additional assumption that K is bounded, respectively, the above definition is only needed for a particular phase one method. This has nothing to do with the fact that we follow [Ren01]: The boundedness of K is also assumed in [NN94], cf. [NN94, Beginning of Section 3.2.2].

Remark 2.5.28. From a geometrical point of view, $\operatorname{sym}_x(y) \in (0, 1]$ measures the symmetry of K with respect to the point x if one looks only into the directions y and $-y$. Informally speaking, values close to 1 indicate "much symmetry in direction y and $-y$", whereas values close to zero indicate that K is "not very symmetric in direction y and $-y$" . For example, if $K \subset \mathbb{R}^n$ is the image of the open unit ball under an isomorphism, then $\operatorname{sym}_x(y) = 1$ for $x = 0$ and all $y \in X \setminus \{0\}$. Since small values of $\operatorname{sym}_x(y)$ indicate that x is much closer to ∂K in one direction than in the corresponding negative direction, we can interpret $\operatorname{sym}(x, K)$ as a measure for symmetry of K about x.

Remark 2.5.29. If x is chosen to be an *analytical center* of K, i.e., the minimizer of a self-bounded function $b : K \to \mathbb{R}$, then one can show $\operatorname{sym}(x, K) \geq 1/(4\vartheta_b + 1)$, cf. [Ren01, Corollary 2.3.5].

Remark 2.5.30. It is easy to see that $\operatorname{sym}_x(y) = \operatorname{sym}_x(ty)$ holds for all $t > 0$. Thus, an equivalent definition of $\operatorname{sym}(x, K)$ is given by $\operatorname{sym}(x, K) = \inf_{y \in S_1} \operatorname{sym}_x(y)$, where $S_1 \subset X$ is defined by $S_1 := \{y \in X : \|y\|_X = 1\}$. This implies $\operatorname{sym}(x, K) > 0$ since K is open and bounded.

We prove a simple auxiliary result for $\operatorname{sym}_x(y)$.

Lemma 2.5.31. *Suppose that $K \subset X$ is nonempty, open, convex, and bounded. Let $x \in K$ and $y \in K \setminus \{x\}$ be given. Set $s := \operatorname{sym}_x(y - x)$. Then it holds $x - s(y - x) \in K$.*

Remark 2.5.32. Since $s \in (0, 1]$ is valid and since K is convex, we obviously have $x + s(y - x) \in K$. The lemma above states that this is also true if we look from the center x into the other direction, i.e., in direction $x - y$ instead of $y - x$.

Proof. We demonstrate $l_x(x - y) > s$. By definition this implies $x + s(x - y) \in K$ thereby proving the assertion. Clearly, we have $l_x(y - x) > 1$. This yields

$$s = \min\left\{\frac{l_x(y-x)}{l_x(x-y)}, \frac{l_x(x-y)}{l_x(y-x)}\right\} \leq \frac{l_x(x-y)}{l_x(y-x)} < l_x(x-y). \qquad \square$$

Lemma 2.5.33. *Let $K \subset X$ be nonempty, open, convex, and bounded. Let $f : K \to \mathbb{R}$ be a nondegenerate ϑ_f-self-concordant barrier function for K. Eventually, let $x, x^0 \in K$ with $x \neq x^0$ and define $s := -f''(x^0)^{-1} f'(x)$. Then it holds*

$$\|s\|_{f''(x^0)} \leq \left(1 + \frac{1}{\operatorname{sym}_x(x^0 - x)}\right) \vartheta_f.$$

In particular, this implies

$$\|s\|_{f''(x^0)} \leq \left(1 + \frac{1}{\operatorname{sym}(x, K)}\right) \vartheta_f.$$

Remark 2.5.34. We have $\operatorname{sym}(x, K) > 0$, cf. Remark 2.5.30.

Proof. Since it holds $\operatorname{sym}_x(x^0 - x) \geq \operatorname{sym}(x, K)$ by definition, it suffices to prove the first estimate. The following proof is a refined version of [Ren01, Proof of Proposition 2.3.7]. Set $\sigma := \operatorname{sym}_x(x^0 - x) \in (0, 1]$ and $w := x - \sigma(x^0 - x)$. Lemma 2.5.31 shows $w \in K$. For $r > 0$ and $\tilde{x} \in X$ we denote $B_r(\tilde{x}) := \{y \in X : \|y - \tilde{x}\|_{f''(x^0)} < r\}$ during the remainder of the proof. Using $B_1(x^0) \subset K$, cf. Lemma 2.2.6, and the convexity of K we obtain

$$\frac{1}{1+\sigma} w + \frac{\sigma}{1+\sigma} B_1(x^0) \subset K.$$

There holds $x = \frac{w}{1+\sigma} + \frac{\sigma}{1+\sigma} x^0$. With $r := \frac{\sigma}{1+\sigma}$ we deduce therefrom

$$B_r(x) = B_r\left(\frac{w}{1+\sigma} + r x^0\right) = \frac{1}{1+\sigma} w + B_r(r x^0) = \frac{1}{1+\sigma} w + r B_1(x^0).$$

Together, this implies $B_r(x) \subset K$. For $s = 0$, the assertion is trivially fulfilled. For $s \neq 0$ we have

$$\|s\|_{f''(x^0)} \leq \sup_{v \in B_1(x)} f'(x)[v - x] = r^{-1} \sup_{v \in B_r(x)} f'(x)[v - x] \leq r^{-1} \sup_{v \in K} f'(x)[v - x] \leq r^{-1}\vartheta_f,$$

where we used $v = x + ts/\|s\|_{f''(x^0)}$ with $t \to 1^-$ in the first and Lemma 2.3.16 in the last step. The assertion now follows from $r^{-1} = 1 + \frac{1}{\sigma}$. □

2.6. A short step method

In this section we present and analyze a short step method for solving problem ($\mathrm{P_{SC}}$).

For the moment let us suppose that Assumption 2.5.2 holds. Then we can consider the following algorithm for solving ($\mathrm{P_{SC}}$).

Algorithm SSM (short step method)

Input: Parameters $(\theta, \mu_0) \in (0, \frac{1}{4}] \times I_s$, starting point $x^0 \in \Lambda_{\mu_0}(\theta)$.

Set $\delta := \frac{\theta\left(1 - \frac{\theta}{(1-\theta)^2}\right)}{1 + \frac{\theta}{\sqrt{\vartheta_b}}}$ and $\beta := 1 - \frac{\delta}{\sqrt{\vartheta_b}}$.

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k \in X$ by solving $f''_{\mu_k}(x^k)[s^k] = -f'_{\mu_k}(x^k)$ in X^*.

Set $x^{k+1} := x^k + s^k$ and $\mu_{k+1} := \beta\mu_k$.

END

Remark 2.6.1. A starting point x^0 that satisfies $x^0 \in \Lambda_{\mu_0}(\theta)$ can be found by use of a phase one method. Phase one methods only require a starting point $\tilde{x}_0 \in K$. We treat phase one methods in Section 2.9.

Remark 2.6.2. We comment on termination criteria for SSM after the next theorem.

We present one of the main results of Section 2. It states the convergence of SSM with r-linear rate and provides complexity estimates.

Theorem 2.6.3. *Let Assumption 2.5.2 be satisfied. Then Algorithm SSM generates a sequence $(x^k) \subset K$ with $x^k \in \Lambda_{\mu_k}(\theta)$ for all $k \in \mathbb{N}_0$, and for every $k \in \mathbb{N}_0$ we have:*

1) *To reach iteration k (more precisely: to reach the FOR statement in SSM for the $k+1$-th time) Algorithm SSM requires exactly k Newton steps.*

2) *The sequence $(j(x^k))$ converges with r-linear rate β to the optimal value $\bar{j} = \inf_{x \in M} j(x)$ of ($\mathrm{P_{SC}}$). More precisely, there holds*

$$|j(x^k) - \bar{j}| \leq (\vartheta_b + \sqrt{\vartheta_b})\mu_k = (\vartheta_b + \sqrt{\vartheta_b})\beta^k\mu_0. \tag{2.14}$$

3) For every $\hat{\varepsilon} > 0$ we have the complexity estimate

$$k \geq \frac{\sqrt{\vartheta_b}}{\delta} \ln\left(\frac{\mu_0(\vartheta_b + \sqrt{\vartheta_b})}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad |j(x^k) - \bar{j}| \leq \hat{\varepsilon}. \tag{2.15}$$

4) If M is convex and j is uniformly convex on M with respect to a norm $|||\cdot|||$, and if ($\mathrm{P_{SC}}$) *possesses a minimizer $\bar{x} \in M$, then it holds*

$$|||x^k - \bar{x}||| \leq \sqrt{\frac{4}{\alpha}}\sqrt{\vartheta_b + \sqrt{\vartheta_b}}\sqrt{\mu_k}, \tag{2.16}$$

where $\alpha > 0$ denotes the convexity modulus of j with respect to $|||\cdot|||$. In particular, (x^k) then converges r-linearly with rate $\sqrt{\beta}$ and $|||\cdot|||$-strongly to the unique minimizer $\bar{x}$, and we have for every $\hat{\varepsilon} > 0$ the complexity estimate

$$k \geq \frac{2\sqrt{\vartheta_b}}{\delta} \ln\left(\frac{\sqrt{\frac{4}{\alpha}}\sqrt{\mu_0(\vartheta_b + \sqrt{\vartheta_b})}}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad |||x^k - \bar{x}||| \leq \hat{\varepsilon}. \tag{2.17}$$

Proof. Lemma 2.5.9 implies $(x^k) \subset K$ with $x^k \in \Lambda_{\mu_k}(\theta) \subset \Lambda_{\mu_k}$ for all $k \in \mathbb{N}_0$. 1) is evident. Using Lemma 2.5.18 and Lemma 2.5.19 we obtain (2.14):

$$\begin{aligned} j(x^k) - \bar{j} &\leq |j(x^k) - j(\bar{x}_{\mu_k})| + |j(\bar{x}_{\mu_k}) - \bar{j}| \\ &\leq \left[\frac{\lambda_{\mu_k}(x^k)\sqrt{\vartheta_b} + \lambda_{\mu_k}(x^k)^3}{\left(1 - \frac{16}{9}\lambda_{\mu_k}(x^k)\right) \cdot (1 - \lambda_{\mu_k}(x^k))} + \vartheta_b\right] \mu_k \\ &\leq \left[\frac{\frac{1}{4}\sqrt{\vartheta_b} + \frac{1}{64}\sqrt{\vartheta_b}}{(1 - \frac{4}{9}) \cdot (1 - \frac{1}{4})} + \vartheta_b\right] \mu_k \leq \left[\sqrt{\vartheta_b} + \vartheta_b\right] \mu_k = \left[\sqrt{\vartheta_b} + \vartheta_b\right] \beta^k \mu_0. \end{aligned}$$

Together with the inequality $-\frac{1}{\ln \beta} \leq \frac{1}{1-\beta}$, (2.15) follows from (2.14) with a simple computation. Setting $t := 1 - \beta \in (0, 1)$ this inequality is obviously a consequence of the inequality $-\ln(1-t) \geq t$, which holds for all $t \in [0, 1)$. The latter inequality follows from the fact that $g(t) := -\ln(1-t) - t$ satisfies $g(0) = 0$ and $g'(t) \geq 0$ for all $t \in [0, 1)$.

We now establish (2.16). In view of (2.14) it suffices to prove $|||x - \bar{x}||| \leq \sqrt{\frac{4}{\alpha}}\sqrt{j(x) - j(\bar{x})}$ for all $x \in K$. This inequality is a consequence of Lemma C.4.12. Furthermore, it implies (2.17) via (2.15). Since j is, in particular, strictly convex on M, the minimizer $\bar{x}$ of ($\mathrm{P_{SC}}$) is unique. □

Remark 2.6.4. Based on (2.14), (2.15), (2.16), or (2.17), various termination criteria are conceivable for Algorithm SSM.

Remark 2.6.5. Theorem 2.6.3 holds for arbitrary choices $\theta \in (0, \frac{1}{4}]$. It turns out that if θ is chosen suitably, then the complexity estimates in the preceding theorem can be improved, cf. [Gli02, Theorem 2.5 and Corollary 2.1]. However, the improved estimates still contain the leading factor $\sqrt{\vartheta_b}$, which turns out to be the dominant part in the application to optimal control. Therefore, the above version of Theorem 2.6.3 is sufficient for our purposes.

Remark 2.6.6. The existence of a minimizer $\bar{x}$ of ($\mathrm{P_{SC}}$) follows, for instance, if X is reflexive, $M = \overline{K}$ holds, and K is nonempty, bounded, and convex, and $j : M \to \mathbb{R}$ is continuous and convex, cf. Lemma C.4.5. The existence of $\bar{x}$ is also ensured under these conditions if the boundedness of K is replaced by uniform convexity of j on $M = \overline{K}$ together with its Gâteaux differentiability on an open set $D \supset M$, since then j has bounded level sets, cf. Lemma C.4.13.

2.7. A long step method

In this section we present and analyze a long step method for solving problem ($\mathrm{P_{SC}}$).

For the moment let us suppose that Assumption 2.5.2 holds. Then we can consider the following algorithm for solving ($\mathrm{P_{SC}}$).

Algorithm LSM (long step method)

Input: Parameters $(\theta, \mu_0) \in (0, \frac{1}{4}] \times I_s$, $\beta_{\min}, \beta_{\max} \in (0,1)$ with $\beta_{\min} \leq \beta_{\max}$,
starting point $x^0 \in \Lambda_{\mu_0}\big(\tau^{-1}\big)$ with $\tau := 2\sqrt[2^6]{2}$.

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k \in X$ by solving $f''_{\mu_k}(x^k)[s^k] = -f'_{\mu_k}(x^k)$ in X^*.

CALL Algorithm LSMSUB with $(x^k, s^k, \mu_k, \theta)$ and denote its return value by x^{k+1}.

Choose $\beta_k \in [\beta_{\min}, \beta_{\max}]$ and set $\mu_{k+1} := \beta_k \mu_k$.

END

Remark 2.7.1. We point out that x^{k+1}, $k \in \mathbb{N}_0$, is determined in iteration k. We will encounter this index shift in all algorithms that are based on LSM.

Remark 2.7.2. We later present termination criteria for LSM.

We state Algorithm LSMSUB.

Algorithm LSMSUB (subroutine for Algorithm LSM)

Input: $(x, s, \mu, \theta) \in K \times X \times I_s \times (0, \frac{1}{4}]$, where s is the Newton step for f_μ at x.

Output: $\tilde{x} \in \Lambda_\mu(\theta)$.

Set $\tilde{x}^0 := x$, $\tilde{s}^0 := s$, $\tau_1 := \sqrt{\frac{\theta}{2}}$, and $\tau_2 := \min\left\{1 - \frac{1}{\sqrt{2}}, \tau_1\right\}$.

FOR $l = 0, 1, 2, \ldots$:

IF $\lambda_\mu(\tilde{x}^l) \leq \tau_1$, **THEN** $\tilde{x} := \tilde{x}^l + \frac{\tilde{s}^l}{1+\lambda_\mu(\tilde{x}^l)}$.

IF $\lambda_\mu(\tilde{x}^l) \leq \tau_2$ **AND** $f_\mu(\tilde{x}^l + \tilde{s}^l) \leq f_\mu(\tilde{x})$, **THEN** (overwrite $\tilde{x}$ by) $\tilde{x} := \tilde{x}^l + \tilde{s}^l$.

IF $\lambda_\mu(\tilde{x}^l) \leq \tau_1$, **THEN RETURN** $\tilde{x}$.

Set $\hat{x}^{l+1} := \tilde{x}^l + \frac{1}{1+\lambda_\mu(\tilde{x}^l)}\tilde{s}^l$.

IF $\lambda_\mu(\tilde{x}^l) \leq 1 - \frac{1}{\sqrt{2}}$ **AND** $f_\mu(\tilde{x}^l + \tilde{s}^l) \leq f_\mu(\hat{x}^{l+1})$, **THEN** $\tilde{x}^{l+1} := \tilde{x}^l + \tilde{s}^l$, **ELSE** $\tilde{x}^{l+1} := \hat{x}^{l+1}$.

IF $\lambda_\mu(\tilde{x}^l) > 1 - \frac{1}{\sqrt{2}}$ **AND** $\lambda_\mu(\tilde{x}^l) \leq \frac{1}{2 \cdot \sqrt[2^6]{2}}$, **THEN** $\tilde{x}^{l+1} := \hat{x}^{l+1}$.

IF $\lambda_\mu(\tilde{x}^l) > \frac{1}{2 \cdot \sqrt[2^6]{2}}$, **THEN** choose $\tilde{x}^{l+1} \in K$ such that $f_\mu(\tilde{x}^{l+1}) \leq f_\mu(\hat{x}^{l+1})$ holds.

Compute the Newton step $\tilde{s}^{l+1} \in X$ by solving $f''_\mu(\tilde{x}^{l+1})\tilde{s}^{l+1} = -f'_\mu(\tilde{x}^{l+1})$ in X^*.

END

Remark 2.7.3. Making use of Lemma 2.2.6 it is easy to see that all iterates that LSMSUB generates belong to K if Assumption 2.5.2 holds. For instance, $\hat{x}^{l+1} = \tilde{x}^l + \frac{\tilde{s}^l}{1+\lambda_\mu(\tilde{x}^l)} \in K$ is satisfied due to $\|\frac{\tilde{s}^l}{1+\lambda_\mu(\tilde{x}^l)}\|_{f''_\mu(\tilde{x}^l)} = \frac{\lambda_\mu(\tilde{x}^l)}{1+\lambda_\mu(\tilde{x}^l)} < 1$.

Remark 2.7.4. If $\lambda_\mu(\tilde{x}^l) > \frac{1}{2 \cdot \sqrt[2^6]{2}}$ holds, we are free to apply heuristics to find $\tilde{x}^{l+1}$ with $f_\mu(\tilde{x}^{l+1}) \leq f_\mu(\hat{x}^{l+1})$. For instance, we can employ line search strategies, which may (significantly) increase the effectiveness of Algorithm LSMSUB in comparison to the use of $\tilde{x}^{l+1} = \hat{x}^{l+1}$. In practical optimization problems the numerical costs for a line search are often negligible in comparison to the computation of a Newton step. Let us now comment further on how to determine a suitable $\tilde{x}^{l+1}$ in the case $\lambda_\mu(\tilde{x}^l) > \frac{1}{2 \cdot \sqrt[2^6]{2}}$ using line search. Usually, we are interested in taking large steps. Thus, we successively check if $\tilde{x}^l + t_i\tilde{s}^l$, $i = 0, 1, 2, \ldots, N$ for some large N, belongs to K and yields a function value smaller than or equal to $f_\mu(\hat{x}^{l+1})$, where $t_i := 1 - (i/N)(1 - \sigma_l)$ with $\sigma_l := \frac{1}{1+\lambda_\mu(\tilde{x}^l)}$. Of course, for $i = N$ we have $t_i = \sigma_l$, which implies that $\tilde{x}^l + t_i\tilde{s}^l$ is accepted for $i = N$. Hence, this method is well-defined. To check for a given i whether $\tilde{x}^l + t_i\tilde{s}^l$ is accepted, only the evaluation of $f_\mu(\tilde{x}^l + t_i\tilde{s}^l)$ is required. Of course, several refinements of this strategy are conceivable. For instance, we could also incorporate step sizes t_i that are smaller than σ_l. Or we could use the evaluations of $f_\mu(\tilde{x}^l + t_i\tilde{s}^l)$ to construct a model of $t \mapsto f_\mu(\tilde{x}^l + t\tilde{s}^l)$ on some interval containing σ_l, and then choose $\tilde{x}^{l+1} := \tilde{x}^l + \hat{t}\tilde{s}^l$, where $\hat{t}$ is obtained from the model. In this setting, the model is often chosen such that its exact minimizer can be obtained easily and $\hat{t}$ is then chosen to be this minimizer. Also, standard step size rules, e.g., the Armijo rule, could be employed to obtain a candidate for the comparison with $f_\mu(\hat{x}^{l+1})$. Note that $f'_\mu(\tilde{x}^l)$, which appears in the Armijo rule, is already available from the computation of the Newton step $\tilde{s}^l$.

Remark 2.7.5. In contrast to the previous remark on *step sizes*, computing other *search directions*, i.e., search directions different from the Newton direction, is not sensible in general, since the Newton direction is required in each iteration of LSMSUB, anyway, to compute $\lambda_\mu(\tilde{x}^l)$. Hence, in a concrete implementation we only use heuristics on the step size but not on the search direction.

We study how Algorithm LSMSUB affects function value and Newton decrement.

Lemma 2.7.6. *Let Assumption 2.5.2 hold. Let Algorithm LSMSUB be started with $(x, s, \mu, \theta) \in K \times X \times I_s \times (0, \frac{1}{4}]$, where s is the Newton step for f_μ at x. Denote by $\tilde{x}^l$, $l = 0, 1, 2, \ldots$, the (possibly finitely many) iterates that LSMSUB generates. Then there hold for every $L \in \mathbb{N}_0$ for which $\tilde{x}^L$ exists:*

1) $\lambda_\mu(\tilde{x}^L) > \frac{1}{2 \cdot \sqrt[2^6]{2}} \quad \Longrightarrow \quad f_\mu(\tilde{x}^L) - f_\mu(\tilde{x}^{L+1}) \geq \lambda_\mu(\tilde{x}^L) - \ln\Big(1 + \lambda_\mu(\tilde{x}^L)\Big) > 0.0927.$

2) $1 - \frac{1}{\sqrt{2}} < \lambda_\mu(\tilde{x}^L) \leq \frac{1}{2 \cdot \sqrt[2^6]{2}} \quad \Longrightarrow \quad$ *There exists* $i \in \{1, 2, \ldots, 6\}$ *with* $\lambda_\mu(\tilde{x}^{L+i}) \leq 1 - \frac{1}{\sqrt{2}}$.

3) $\lambda_\mu(\tilde{x}^L) \leq 1 - \frac{1}{\sqrt{2}} \quad \Longrightarrow \quad$ *LSMSUB takes maximal* $L + \lceil 1.13 + 1.45 \ln |\ln 2\tau_1| \rceil$ *Newton steps.*

Proof. We start by establishing that 1) holds. The first inequality follows from Lemma 2.2.18 in combination with $f_\mu(\tilde{x}^{L+1}) \leq f_\mu(\hat{x}^{L+1})$. The second inequality is due to monotonicity of $t - \ln(1 + t)$.

Now we demonstrate that 2) is valid. If there holds $\lambda_\mu(\tilde{x}^{L+i}) \leq 1 - \frac{1}{\sqrt{2}}$ for some $i \in \{1, 2, \ldots, 5\}$, then there is nothing to prove. Therefore, we may assume $\lambda_\mu(\tilde{x}^{L+i}) > 1 - \frac{1}{\sqrt{2}}$ for $i = 0, 1, \ldots, 5$. Since $\lambda_\mu(\tilde{x}^L) \leq \frac{1}{2 \cdot \sqrt[2^6]{2}}$ holds, we deduce from Lemma 2.2.20 $\lambda_\mu(\tilde{x}^{L+1}) \leq 2\lambda_\mu(\tilde{x}^L)^2 \leq \frac{1}{2 \cdot \sqrt[2^5]{2}}$, where we used $\tilde{x}^{L+1} = \hat{x}^{L+1}$ due to $\lambda_\mu(\tilde{x}^L) > 1 - \frac{1}{\sqrt{2}}$. Hence, repeated application of Lemma 2.2.20 yields that $\lambda_\mu(\tilde{x}^{L+6}) \leq 2\lambda_\mu(\tilde{x}^{L+5})^2 \leq \frac{1}{4} < 1 - \frac{1}{\sqrt{2}}$ is satisfied.

Eventually, we turn to 3). By $\lambda_\mu(\tilde{x}^L) \leq 1 - \frac{1}{\sqrt{2}}$, Lemma 2.2.20 yields $\lambda_\mu(\tilde{x}^{L+l}) \leq 2\lambda_\mu(\tilde{x}^{L+l-1})^2$ for all $l \in \mathbb{N}$ such that $\tilde{x}^{L+l}$ exists, where we used $1/(1-t)^2 \leq 2$ for $t \leq 1 - \frac{1}{\sqrt{2}}$. From this follows $\lambda_\mu(\tilde{x}^{L+l}) \leq \frac{1}{2} \cdot (\frac{2}{5})^{2^{l-1}}$ for all these l as well as $l = 0$. Hence, it suffices to determine $l \in \mathbb{N}_0$ with $2^{l-1} \geq \ln(2\tau_1)/\ln\frac{2}{5}$, since this implies $\lambda_\mu(\tilde{x}^{L+l}) \leq \tau_1$, i.e., Algorithm LSMSUB terminates after the computation of at most $L + l$ Newton steps. This is ensured if $l \geq 1 - \frac{\ln\ln\frac{5}{2}}{\ln 2} + \frac{\ln|\ln 2\tau_1|}{\ln 2}$ holds. We have $-\frac{\ln\ln\frac{5}{2}}{\ln 2} \leq 0.1262$ and $\frac{1}{\ln 2} \leq 1.4427$, which concludes the proof. □

Remark 2.7.7. The last inequality in 1) may be much too conservative. In fact, for $\lambda_\mu(\tilde{x}^L) \gg 1$ we have $f_\mu(\tilde{x}^L) - f_\mu(\tilde{x}^{L+1}) \geq f_\mu(\tilde{x}^L) - f_\mu(\hat{x}^{L+1}) \approx \frac{\lambda_\mu(\tilde{x}^L)}{2}$, cf. Corollary 2.2.19.

The previous lemma implies an estimate on the maximal number of iterations of LSMSUB.

Corollary 2.7.8. *Let Assumption 2.5.2 hold. When started with $(x, s, \mu, \theta) \in K \times X \times I_s \times (0, \frac{1}{4}]$, where s is the Newton step for f_μ at x, LSMSUB computes*

$$N \leq \lfloor 10.79\,(f_\mu(x) - f_\mu(\bar{x}_\mu))\rfloor + \left\lceil 7.13 + 1.45 \ln \left| \ln \sqrt{2\theta} \right| \right\rceil$$

Newton steps and terminates with $\tilde{x} \in \Lambda_\mu(\theta)$. If f_μ is, in addition, self-bounded on K with constant ϑ_μ, then we also have

$$N \leq \lfloor 10.79\vartheta_\mu \left|\ln(1 - \omega_{\bar{x}_\mu}(x))\right| \rfloor + \left\lceil 7.13 + 1.45 \ln \left| \ln \sqrt{2\theta} \right| \right\rceil,$$

where $\omega_{\bar{x}_\mu} : K \to [0, 1)$ denotes the Minkowski function, cf. Definition 2.3.14.

Proof. By virtue of Lemma 2.3.16 the second estimate is a direct consequence of the first, so it remains to prove the first assertion. If we have $\lambda_\mu(\tilde{x}^0) \leq \frac{1}{2 \cdot \sqrt[2^6]{2}}$, it follows from Lemma 2.7.6 2) and 3) that LSMSUB terminates after at most $6 + \lceil 1.13 + 1.45 \ln|\ln 2\tau_1|\rceil$ Newton steps, which proves the asserted estimate for this case. Let $\lambda_\mu(\tilde{x}^0) > \frac{1}{2 \cdot \sqrt[2^6]{2}}$ and denote by $L \in \mathbb{N}_0$ a number with $\lambda_\mu(\tilde{x}^l) > \frac{1}{2 \cdot \sqrt[2^6]{2}}$ for all $l \leq L$. From $x = \tilde{x}^0$ and Lemma 2.7.6 1) we infer

$$f_\mu(x) - f_\mu(\tilde{x}^{L+1}) = \sum_{l=0}^{L} \left(f_\mu(\tilde{x}^l) - f_\mu(\tilde{x}^{l+1}) \right) > 0.0927(L + 1). \tag{2.18}$$

The function f_μ has a global minimizer $\bar{x}_\mu$, see Corollary 2.5.13. Hence, (2.18) implies $L + 1 \leq \lfloor (f_\mu(x) - f_\mu(\bar{x}_\mu)) \cdot 10.79 \rfloor$. Now, let $L^* \in \mathbb{N}_0$ denote the *maximal* number that satisfies $\lambda_\mu(\tilde{x}^l) > \frac{1}{2 \cdot \sqrt[2^6]{2}}$ for all $l \leq L^*$. Thus, $\lambda_\mu(\tilde{x}^l) \leq \frac{1}{2 \cdot \sqrt[2^6]{2}}$ for $l = L^* + 1$ so that we can apply Lemma 2.7.6 2) and 3). This shows that LSMSUB terminates after at most $L^* + 1 + 6 + \lceil 1.13 + 1.45 \ln|\ln 2\tau_1|\rceil$ Newton steps, which concludes the proof of the asserted estimate. When LSMSUB terminates, there holds $\lambda_\mu(\tilde{x}) \leq 2\lambda_\mu(\tilde{x}^l)^2 \leq 2\tau_1^2 = \theta$, as follows from Lemma 2.2.20. □

Remark 2.7.9. In [KU13] we used a slightly different version of Algorithm LSMSUB. This yields a different estimate for N, namely one where the factor 10.79 is changed to 27.77, while the number 7.13 is replaced by 1.13. The modification in [KU13] allows to use line search also in the case where $\lambda_\mu(\tilde{x}^l) > 1 - \frac{1}{\sqrt{2}}$ and $\lambda_\mu(\tilde{x}^l) \leq \frac{1}{2 \cdot \sqrt[2^6]{2}}$ are satisfied.

We present another main result of Section 2. It states the convergence of LSM with r-linear rate and provides complexity estimates.

Theorem 2.7.10. *Let Assumption 2.5.2 be satisfied. Then Algorithm LSM generates a sequence $(x^k) \subset K$ with $x^{k+1} \in \Lambda_{\mu_k}(\theta)$ for all $k \in \mathbb{N}_0$, and for each $k \in \mathbb{N}_0$ we have:*

1) To reach iteration k (more precisely: to reach the FOR statement in LSM for the $k+1$-th time) Algorithm LSM requires at most

$$k \left(\left\lfloor \frac{10.79}{\beta_{\min}} \left(\vartheta_b + \sqrt{\vartheta_b} \right) \right\rfloor + \lceil 8.13 + 1.45 \ln |\ln 2\tau_1| \rceil \right) \tag{2.19}$$

Newton steps, including the Newton steps from LSMSUB.

2) The sequence $(j(x^k))$ converges with r-linear rate β_k in iteration k to the optimal value $\bar{j} = \inf_{x\in M} j(x)$ of $(\mathrm{P_{SC}})$. More precisely, there holds

$$|j(x^{k+1}) - \bar{j}| \leq \left(\vartheta_b + \sqrt{\vartheta_b}\right) \mu_0 \prod_{i=0}^{k-1} \beta_i = \left(\vartheta_b + \sqrt{\vartheta_b}\right) \mu_k.$$

3) For every $\hat{\varepsilon} > 0$ we have the complexity estimate

$$k \geq \left|\frac{1}{\ln \beta_{\max}}\right| \ln\left(\frac{\mu_0(\vartheta_b + \sqrt{\vartheta_b})}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad |j(x^{k+1}) - \bar{j}| \leq \hat{\varepsilon}.$$

4) If M is convex and j is uniformly convex on M with respect to a norm $|||\cdot|||$, and if $(\mathrm{P_{SC}})$ possesses a minimizer $\bar{x} \in M$, then it holds

$$|||x^{k+1} - \bar{x}||| \leq \sqrt{\frac{4}{\alpha}} \sqrt{\vartheta_b + \sqrt{\vartheta_b}} \sqrt{\mu_k},$$

where $\alpha > 0$ denotes the convexity modulus of j with respect to $|||\cdot|||$. In particular, (x^k) then converges r-linearly with rate $\sqrt{\beta_k}$ in iteration k and $|||\cdot|||$-strongly to the unique minimizer $\bar{x}$, and we have for every $\hat{\varepsilon} > 0$ the complexity estimate

$$k \geq \left|\frac{2}{\ln \beta_{\max}}\right| \ln\left(\frac{\sqrt{\frac{4}{\alpha}}\sqrt{\mu_0(\vartheta_b + \sqrt{\vartheta_b})}}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad |||x^{k+1} - \bar{x}||| \leq \hat{\varepsilon}.$$

Proof. Corollary 2.7.8 implies $(x^k) \subset K$ and $x^{k+1} \in \Lambda_{\mu_k}(\theta)$ for all $k \in \mathbb{N}_0$. Assertions 2), 3) and 4) can be proven as their counterparts in Theorem 2.6.3, so it suffices to establish 1). For $k = 0$ there is nothing to prove. To determine x^1 no more than $1 + \lceil 7.13 + 1.45 \ln|\ln 2\tau_1|\rceil$ Newton steps are required by LSM and LSMSUB together, as follows from Lemma 2.7.6 2) and 3) and the requirement $x^0 \in \Lambda_{\mu_0}(\tau^{-1})$. To determine x^{k+1} for $k \in \mathbb{N}$ we note that during each iteration of LSM exactly one Newton step is computed if we do not count the Newton steps from LSMSUB. If LSMSUB is called to determine x^{k+1}, $k \in \mathbb{N}$, no more than $\lfloor 10.79(f_{\mu_k}(x^k) - f_{\mu_k}(\bar{x}_{\mu_k}))\rfloor + \lceil 7.13 + 1.45 \ln|\ln 2\tau_1|\rceil$ Newton steps are required, cf. Corollary 2.7.8. Hence, 1) is a consequence of $f_{\mu_k}(x^k) - f_{\mu_k}(\bar{x}_{\mu_k}) \leq \frac{\vartheta_b + \sqrt{\vartheta_b}}{\beta_{k-1}}$, which follows from Lemma 2.5.21. To apply this lemma we used $x^k \in \Lambda_{\mu_{k-1}}(\theta)$ for $k \in \mathbb{N}$. □

Remark 2.7.11. Remark 2.6.4 and Remark 2.6.6 also apply here.

Remark 2.7.12. The complexity estimates for Algorithm LSM demonstrate that an $\hat{\varepsilon}$-optimal iterate can be found after $\mathcal{O}(\vartheta_b \ln(\frac{\mu_0(\vartheta_b+\sqrt{\vartheta_b})}{\hat{\varepsilon}}))$ Newton steps, as follows from 1) and 3) in the above theorem. With Algorithm SSM this task requires only $\mathcal{O}(\sqrt{\vartheta_b} \ln(\frac{\mu_0(\vartheta_b+\sqrt{\vartheta_b})}{\hat{\varepsilon}}))$ Newton steps, cf. Theorem 2.6.3 3). Let us, therefore, mention that in practice long step methods are usually superior to short step methods. Or, as Renegar so eloquently puts it, cf. [Ren01, Section 2.4.3]: *"It is one of the ironies of the ipm literature that algorithms which are more efficient in practice often have somewhat worse complexity bounds"*. For linear optimization it is possible to prove $\mathcal{O}(\sqrt{\vartheta_b} \ln(\vartheta_b) \ln(\frac{\mu_0 \vartheta_b}{\hat{\varepsilon}}))$ complexity for long step methods using self-regular

functions rather than self-concordant functions, cf., e.g., [PRT02a, PRT02b]. However, it seems that so far self-regularity has not been generalized to convex optimization problems but only to subclasses, e.g., semidefinite optimization and convex quadratic optimization, cf. [Liu09].

2.8. A predictor-corrector method

Among the fastest interior point methods in practice are predictor-corrector methods, cf., e.g., [Wri97, Chapter 10] and [Ren01, Section 2.4.4], even though their provable complexity is worse than the ones of short step and long step methods. In the following we sketch a predictor-corrector method. Throughout this section, we impose Assumption 2.5.2.

In Corollary 2.5.13 we proved that the barrier function f_μ possesses a unique minimum for all $\mu \in I_s$. Since f_μ is convex, this minimum $\bar{x}_\mu$ is characterized as the unique solution of $f'_\mu(x) = 0$. Moreover, since f_μ depends twice continuously differentiable on $(x, \mu) \in K \times \mathbb{R}_{>0}$ and since $f''_\mu(x) \in \mathcal{L}(X, X^*)$ is invertible for every $(x, \mu) \in K \times I_s$, the central path $\gamma : I_s \to K$, $\gamma(\mu) := \bar{x}_\mu$ defines a continuously differentiable trajectory, as follows from the implicit function theorem. (In fact, by the same argument the central path is infinitely many times continuously differentiable.) This motivates the predictor step: Given $\bar{x}_{\mu_k}$ for $\mu_k \in I_s$, i.e., $\gamma(\mu_k)$, we are interested in finding $\gamma(\mu_{k+1})$ with $\mu_{k+1} < \mu_k$. Taylor expansion $\gamma(\mu) \approx \gamma(\mu_k) + \gamma'(\mu_k)(\mu - \mu_k)$ shows that an approximation for $\gamma(\mu_{k+1})$ can be obtained if $\gamma'(\mu_k)$ is available. Since we have

$$0 = \frac{d}{\mathrm{d}\mu} f'_\mu(\gamma(\mu)) = \frac{j''(\gamma(\mu))}{\mu}\gamma'(\mu) - \frac{j'(\gamma(\mu))}{\mu^2} + b''(\gamma(\mu))\gamma'(\mu),$$

computing $\gamma'(\mu_k)$ means solving the linear system $f''_{\mu_k}(\bar{x}_{\mu_k})s = \frac{j'(\bar{x}_{\mu_k})}{\mu_k^2}$. Of course, in practice we do not have $\bar{x}_{\mu_k}$ at our disposal but use an approximation, e.g., x^k with $\lambda_{\mu_k}(x^k) \leq 0.01$. Since the new point is only an approximation of $\gamma(\mu_{k+1})$, we have to ensure feasibility of this new point, i.e., we have to choose μ_{k+1} suitably. Moreover, each predictor step may have to be accompanied by several corrector steps, i.e., steps that yield an iterate x^{k+1} with $\lambda_{\mu_{k+1}}(x^{k+1}) \leq 0.01$ again. This task can be handled by LSMSUB. If such an x^{k+1} is obtained, we use a predictor step again, and so on. Summarizing, we have sketched a predictor-corrector scheme. However, we leave all theoretical investigations for future research.

2.9. Phase one

Algorithm SSM requires for a given $\mu_0 \in I_s$ a starting point $x^0 \in K$ that satisfies $\lambda_{\mu_0}(x^0) \leq \theta$ with $\theta \in (0, \frac{1}{4}]$. Such a point is also suitable to start Algorithm LSM. The task of finding such a point is called phase one since it has to be applied prior to the actual optimization method. In this section we examine two methods that realize phase one.

2.9.1. Phase one based on a short step method

In this section we describe and investigate an algorithm that can serve as a phase one method and that is based on short steps. This algorithm utilizes the following barrier functions.

Definition 2.9.1. Let Assumption 2.5.2 hold. For $(x^0, \mu_0, \nu) \in K \times I_s \times \mathbb{R}_{>0}$ we abbreviate by f_{ν,μ_0,x^0} the function

$$f_{\nu,\mu_0,x^0} : K \to \mathbb{R}, \quad f_{\nu,\mu_0,x^0}(x) := f_{\mu_0}(x) - \frac{f'_{\mu_0}(x^0)[x]}{\nu}.$$

In the following algorithm $\lambda_{\mu_0}(x)$ denotes the Newton decrement of f_{μ_0}, as always. To state this algorithm we suppose that Assumption 2.5.2 holds and that f_{μ_0} is $\vartheta_{f_{\mu_0}}$-self-bounded on K. We recall that we provided sufficient conditions for f_{μ_0} to be self-bounded in Lemma 2.5.22, 2.5.23, and 2.5.24.

Algorithm POSS (phase one based on short steps)

Input: $(x^0, \mu_0, \theta) \in K \times I_s \times (0, \frac{1}{4}]$.

Output: $\tilde{x} \in \Lambda_{\mu_0}(\theta)$.

Let $\vartheta_{f_{\mu_0}} \geq 1$ be the self-boundedness constant of f_{μ_0} on K. Denote $\tilde{\theta} := \frac{\theta}{2}$, $\nu_0 := 1$, and define

$$\delta := \frac{\tilde{\theta}\left(1 - \frac{\tilde{\theta}}{(1-\tilde{\theta})^2}\right)}{1 - \frac{\tilde{\theta}}{\sqrt{\vartheta_{f_{\mu_0}}}}} \qquad \text{and} \qquad \beta := 1 + \frac{\delta}{\sqrt{\vartheta_{f_{\mu_0}}}}.$$

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k_{\mu_0}$ for f_{μ_0} at x^k by solving $f''_{\mu_0}(x^k)[s^k_{\mu_0}] = -f'_{\mu_0}(x^k)$ in X^*.

IF $\lambda_{\mu_0}(x^k) \leq \theta$, **THEN RETURN** $\tilde{x} := x^k$.

Compute the Newton step s^k for f_{ν_k,μ_0,x^0} at x^k by solving $f''_{\nu_k,\mu_0,x^0}(x^k)[s^k] = -f'_{\nu_k,\mu_0,x^0}(x^k)$ in X^*.

Set $x^{k+1} := x^k + s^k$ and $\nu_{k+1} := \beta\nu_k$.

END

Remark 2.9.2. There holds $f''_{\mu_0}(x^k) = f''_{\nu_k,\mu_0,x^0}(x^k)$, which shows that the Newton steps in POSS are well-defined. Moreover, the linear systems that need to be solved to compute the Newton steps $s^k_{\mu_0}$ and s^k in a practical implementation have the same coefficient matrix. This can be used to decrease the computational costs by either employing the same factorization or the same preconditioner when solving these systems.

Lemma 2.9.3. *Let Assumption 2.5.2 be valid, let K be bounded, and let $\mu_0 \in I_s$. Then for every $x^0 \in K$ and all $\nu > 0$, f_{ν,μ_0,x^0} is a nondegenerate self-concordant barrier function for K and it holds $\Lambda_{\nu,\mu_0,x^0} \neq \emptyset$ for $\nu = 1$. Here, $\Lambda_{\nu,\mu_0,x^0} := \{x \in K : \lambda_{\nu,\mu_0,x^0}(x) \leq \frac{1}{4}\}$ with λ_{ν,μ_0,x^0} denoting the Newton decrement of f_{ν,μ_0,x^0}.*

Proof. Since f_{μ_0} is nondegenerate self-concordant on K due to Assumption 2.5.2, f_{ν,μ_0,x^0} is nondegenerate self-concordant on K. Since K is bounded and since f_{μ_0} is a barrier function for K by Assumption 2.5.2, we infer that f_{ν,μ_0,x^0} is a barrier function for K, too. Moreover, it holds $\lambda_{1,\mu_0,x^0}(x^0) = 0$ due to $f'_{1,\mu_0,x^0}(x^0) = 0$. □

The next lemma indicates that Algorithm POSS is a path-following scheme. In fact, it follows the path $[1,\infty) \ni \nu \mapsto \operatorname{argmin}_{x\in K} f_{\nu,\mu_0,x^0}(x)$ for $\nu \to \infty$, which can be shown to exist under Assumption 2.5.2 and if f_{μ_0} is $\vartheta_{f_{\mu_0}}$-self-bounded with $\vartheta_{f_{\mu_0}} \geq 1$. This can be proven analogously to Corollary 2.5.13. However, the existence of this path is not required for the analysis to come.

Lemma 2.9.4. *Let Assumption 2.5.2 be valid, let K be bounded, and let f_{μ_0} be $\vartheta_{f_{\mu_0}}$-self-bounded on K with $\vartheta_{f_{\mu_0}} \geq 1$ for a fixed $\mu_0 \in I_s$. Then Algorithm POSS generates a sequence $(x^k) \subset K$ with $\lambda_{\nu_k,\mu_0,x^0}(x^k) \leq \tilde{\theta}$ for every $k \in \mathbb{N}_0$.*

Proof. Obviously, there holds $\lambda_{\nu_0,\mu_0,x^0}(x^0) = 0$ due to $f'_{\nu_0,\mu_0,x^0}(x^0) = 0$. Using induction we conclude $\lambda_{\nu_{k+1},\mu_0,x^0}(x^{k+1}) \leq \tilde{\theta}$ from $\lambda_{\nu_k,\mu_0,x^0}(x^k) \leq \tilde{\theta}$ by Lemma 2.5.15 applied to f_{ν,μ_0,x^0} with the role of μ now played by ν. This lemma is applicable due to Lemma 2.9.3. □

Informally speaking, the problems

$$\min_{x\in K} f_{\nu,\mu_0,x^0}(x) \qquad \text{and} \qquad \min_{x\in K} f_{\mu_0}(x)$$

become more and more alike for $\nu \to \infty$. Hence, following the path $\nu \mapsto \operatorname{argmin}_{x\in K} f_{\nu,\mu_0,x^0}(x)$ for $\nu \to \infty$ should lead to the minimizer $\bar{x}_{\mu_0}$ of the problem $\min_{x\in K} f_{\mu_0}(x)$, which satisfies $\lambda_{\mu_0}(\bar{x}_{\mu_0}) = 0$. Therefore, we can expect $\lambda_{\mu_0}(x^k) \leq \theta$ for sufficiently large k, i.e., Algorithm POSS terminates successfully after finitely many iterations. This vague argument is made precise with the following complexity estimate, which is the main result for Algorithm POSS. We recall that $\operatorname{sym}(x^0, K) > 0$, the *symmetry of K about x^0*, is given by Definition 2.5.25.

Theorem 2.9.5. *Let Assumption 2.5.2 be valid, let K be bounded, and let f_{μ_0} be $\vartheta_{f_{\mu_0}}$-self-bounded on K with $\vartheta_{f_{\mu_0}} \geq 1$ for a fixed $\mu_0 \in I_s$. Then Algorithm POSS requires $N \in \mathbb{N}_0$ iterations and terminates with a $\tilde{x} \in \Lambda_{\mu_0}(\theta)$, where N is bounded from above by*

$$N \leq \left\lceil \frac{17}{16} \cdot \frac{\sqrt{\vartheta_{f_{\mu_0}}}}{\delta} \cdot \ln\left(\frac{2\vartheta_{f_{\mu_0}}}{\theta} \left(1 + \frac{1}{\operatorname{sym}(x^0, K)} \right) \right) \right\rceil .$$

During the course of POSS, $2N + 1$ Newton steps have to be computed.

Proof. We need to estimate for which $N \in \mathbb{N}_0$ we have $\lambda_{\mu_0}(x^N) \leq \theta$. It holds

$$\begin{aligned}
\lambda_{\mu_0}(x^N) &= \left\|s^N_{\mu_0}\right\|_{f''_{\mu_0}(x^N)} = \left\|s^N_{\mu_0}\right\|_{f''_{\nu,\mu_0,x^0}(x^N)} = \left\|-f''_{\mu_0}(x^N)^{-1} f'_{\mu_0}(x^N)\right\|_{f''_{\nu,\mu_0,x^0}(x^N)} \\
&= \left\| s^N - \frac{f''_{\nu,\mu_0,x^0}(x^N)^{-1} f'_{\mu_0}(x^0)}{\nu} \right\|_{f''_{\nu,\mu_0,x^0}(x^N)} \\
&\leq \tilde{\theta} + \frac{1}{\nu} \left\| f''_{\nu,\mu_0,x^0}(x^N)^{-1} f'_{\mu_0}(x^0) \right\|_{f''_{\nu,\mu_0,x^0}(x^N)} \\
&= \frac{\theta}{2} + \frac{1}{\nu} \left\| f''_{\mu_0}(x^N)^{-1} f'_{\mu_0}(x^0) \right\|_{f''_{\mu_0}(x^N)} \\
&\leq \frac{\theta}{2} + \frac{1}{\nu} \left(1 + \frac{1}{\operatorname{sym}(x^0, K)}\right) \vartheta_{f_{\mu_0}},
\end{aligned}$$

where we used Lemma 2.9.4 and Lemma 2.5.33. This shows that for N with

$$\nu_N \geq \frac{2}{\theta} \left(1 + \frac{1}{\operatorname{sym}(x^0, K)}\right) \vartheta_{f_{\mu_0}}$$

we have $\lambda_{\mu_0}(x^N) \leq \theta$. Since there holds $\nu_{k+1} = \beta\nu_k$ for all k and $\nu_0 = 1$, a simple computation using $\frac{1}{\ln(1+t)} \leq \frac{17}{16t}$ for $t \in (0, 1/8]$ together with $0 < \delta \leq 1/8$, which follows from $\tilde{\theta} \leq 1/8$, establishes the bound on N. The assertion on the number of Newton steps is obvious. □

2.9.2. Phase one based on a long step method

We can use LSMSUB as a long step based phase one method.

Theorem 2.9.6. *Let Assumption 2.5.2 be valid and let f_{μ_0} be $\vartheta_{f_{\mu_0}}$-self-bounded on K with $\vartheta_{f_{\mu_0}} \geq 1$ for a fixed $\mu_0 \in I_s$. Let Algorithm LSMSUB be started with $(x^0, s^0, \mu_0, \theta) \in K \times X \times I_s \times (0, \frac{1}{4}]$, where s^0 denotes the Newton step for f_{μ_0} at x^0. Then LSMSUB requires $N \in \mathbb{N}_0$ iterations and terminates with a $\tilde{x} \in \Lambda_{\mu_0}(\theta)$, where N is bounded from above by*

$$N \leq \left\lfloor 10.79 \vartheta_{f_{\mu_0}} \left| \ln\left(1 - \omega_{\bar{x}_{\mu_0}}(x^0)\right) \right| \right\rfloor + \left\lceil 7.13 + 1.45 \ln \left| \ln \sqrt{2\theta} \right| \right\rceil .$$

Here, $\omega_{\bar{x}_{\mu_0}} : K \to [0, 1)$ denotes the Minkowski function, see Definition 2.3.14. Also, the number of Newton steps that have to be computed during the course of LSMSUB equals N.

Proof. See Corollary 2.7.8. □

Remark 2.9.7. In contrast to the corresponding result for Algorithm POSS, cf. Theorem 2.9.5, the above complexity is derived without the assumption that K is bounded.

Remark 2.9.8. Similar to the short step and long step method we presented, we observe that the complexity of a long step based phase one is worse with respect to the parameter $\vartheta_{f_{\mu_0}}$ than the complexity of a short step based phase one. Therefore, we mention again that despite their theoretically worse complexity, in practice long steps methods usually perform better than short step methods.

3. Problem class and associated barrier problems

In this section we present the class of optimal control problems we are interested in. We provide results on existence and uniqueness of optimal solutions and establish necessary and sufficient optimality conditions for these problems. Furthermore, given a problem from this class we show how to construct self-concordant and self-bounded barrier functions for a closely related problem. This results in a powerful framework for the convergence analysis of the algorithms that we develop in later sections.

3.1. Problem formulation, reduced problem, general assumptions

The problem under consideration is

$$\min_{(y,u)\in Y\times U} \hat{J}(y,u) \quad \text{s.t.} \quad Ay + Bu = g, \quad y(x) \geq y_a(x) \ \ \forall x \in \overline{\Omega}_a, \tag{$\mathrm{P}_{\mathrm{orig}}$}$$

where Y and Z are Banach spaces, $Y \hookrightarrow C^{0,\beta}(\overline{\Omega}_a)$ continuously with a $\beta > 0$, $\Omega_a \subset \mathbb{R}^d$ open and bounded, U is a Hilbert space, $\hat{J} : Y \times U \to \mathbb{R}$, $A \in \mathcal{L}(Y,Z)$, $B \in \mathcal{L}(U,Z)$, $g \in Z$, and $y_a \in C^{0,\beta}(\overline{\Omega}_a)$. Moreover, we suppose that A is invertible. To impose further assumptions we require some standard definitions.

Definition 3.1.1. For $u \in U$ the unique solution of the state equation $Ay + Bu = g$ is denoted by $y(u) := A^{-1}(g - Bu)$.

Definition 3.1.2. The *reduced objective* is denoted by $\hat{j} : U \to \mathbb{R}$, $\hat{j}(u) := \hat{J}(y(u), u)$.

Definition 3.1.3. The *reduced problem* is given by

$$\min_{u\in U} \hat{j}(u) \quad \text{s.t.} \quad y(u)(x) \geq y_a(x) \ \ \forall x \in \overline{\Omega}_a. \tag{$\mathrm{P}_{\mathrm{red}}$}$$

Definition 3.1.4. We denote the *admissible set* of the reduced problem by U_{ad}, i.e.,

$$U_{\mathrm{ad}} := \left\{ u \in U : y(u)(x) \geq y_a(x) \ \ \forall x \in \overline{\Omega}_a \right\}.$$

Remark 3.1.5. The reduced problem is equivalent to ($\mathrm{P}_{\mathrm{orig}}$) in the sense that $\bar{u}$ is a solution of ($\mathrm{P}_{\mathrm{red}}$) if and only if $(y(\bar{u}), \bar{u})$ is a solution of ($\mathrm{P}_{\mathrm{orig}}$). Hence, we can focus on ($\mathrm{P}_{\mathrm{red}}$).

Remark 3.1.6. The set U_{ad} is closed and convex. This follows since A^{-1} is continuous by the bounded inverse theorem, cf. Theorem C.1.1, and since $Y \hookrightarrow C^{0,\beta}(\overline{\Omega}_a) \hookrightarrow C(\overline{\Omega}_a)$ continuously.

We demonstrate that if $\hat{j}$ is uniformly convex, then $(\mathrm{P}_{\mathrm{red}})$ has a unique optimal solution.

Lemma 3.1.7. *Let the reduced objective $\hat{j} : U \to \mathbb{R}$ be Gâteaux differentiable. Furthermore, assume that it is uniformly convex and lower semi-continuous on U_{ad} and that U_{ad} is nonempty. Then the reduced problem $(\mathrm{P}_{\mathrm{red}})$ possesses a unique optimal solution.*

Proof. Let $\hat{u} \in U_{\mathrm{ad}}$. Then the lower level set $L := \{u \in U_{\mathrm{ad}} : \hat{j}(u) \leq \hat{j}(\hat{u})\}$ is bounded due to Lemma C.4.13. It is, furthermore, nonempty, closed, and convex, where we used that U_{ad} is closed and convex and that $\hat{j}$ is lower semi-continuous. Since U is reflexive and $\hat{j}$ is lower semi-continuous and convex on L, Lemma C.4.5 yields the existence of an optimal solution. Since $\hat{j}$ is, in particular, strictly convex on L, the optimal solution is unique. □

Remark 3.1.8. The uniform convexity is required to deduce the boundedness of the lower level set in the proof above. The existence of an optimal solution is, thus, also ensured if the uniform convexity of $\hat{j}$ is replaced by convexity of $\hat{j}$ and boundedness of U_{ad}. However, we can only expect U_{ad} to be bounded if either bilateral control constraints or bilateral state constraints are present, and we cover neither of these settings in this thesis. We remark that bilateral state constraints are a straightforward extension of the theory we present in this thesis, but are neglected for conciseness of the presentation.

We impose the following assumption on $(\mathrm{P}_{\mathrm{red}})$ throughout this thesis.

Assumption 3.1.9.

1) *Y is a Banach space that allows for the continuous (but not necessarily injective) embedding $Y \hookrightarrow C^{0,\beta}(\overline{\Omega}_a)$ for some fixed $\beta > 0$ and a set $\Omega_a \subset \mathbb{R}^d$ that consists of finitely many nonempty, disjoint and bounded domains $\Omega_{a,i}$, $i = 1, \dots, m$. We suppose that each of the $\Omega_{a,i}$ satisfies the* cone condition*. Since the cone condition is a technical and very weak assumption (it is, e.g., satisfied by Lipschitz domains), we refer to the appendix for details, see Section E. In addition, we assume that U is a Hilbert space and Z is a Banach space.*
2) *There hold $A \in \mathcal{L}(Y, Z)$, $B \in \mathcal{L}(U, Z)$, $g \in Z$, and A is invertible.*
3) *We assume that $(\mathrm{P}_{\mathrm{red}})$ possesses at least one optimal solution.*
4) *We suppose that $\hat{j} : U \to \mathbb{R}$ is either quadratic and uniformly convex with modulus $\hat{\alpha} > 0$ or self-concordant. We refer to the first setting as "case I" and to the second as "case II". In case II, i.e., if $\hat{j}$ is only self-concordant, we additionally require the following:*
   - *It holds $\|\bar{u}\|_U \leq C_{\|\bar{u}\|_U}$ with a known constant $C_{\|\bar{u}\|_U} > 0$ for at least one solution $\bar{u}$ of $(\mathrm{P}_{\mathrm{red}})$;*
   - *$\hat{j} : U \to \mathbb{R}$ has a bounded first derivative on bounded sets.*
5) *It holds $y_a \in C^{0,\beta}(\overline{\Omega}_a)$.*
6) *There exists a point $u^\circ \in U$ such that $y^\circ := y(u^\circ)$ is an interior point of the closed convex set $\{y \in C(\overline{\Omega}_a) : y(x) \geq y_a(x) \ \forall x \in \overline{\Omega}_a\}$. This is, y° has to satisfy*

$$\exists \tau^\circ > 0 : \qquad y^\circ(x) - \tau^\circ \geq y_a(x) \quad \forall x \in \overline{\Omega}_a.$$

Remark 3.1.10. Local and global solutions of ($\mathrm{P}_{\mathrm{red}}$) coincide due to convexity, cf. Lemma C.4.4.

Remark 3.1.11. In case II we suppose that there is an optimal solution $\bar{u}$ such that $\|\bar{u}\|_U$ is bounded by the available constant $C_{\|\bar{u}\|_U}$. This assumption is, for instance, fulfilled if $\hat{j}$ is uniformly convex on U, cf. Lemma C.4.13. It is also satisfied if bilateral control constraints occur in ($\mathrm{P}_{\mathrm{red}}$). In this thesis, however, we do not treat problems with both control and state constraints but instead consider this a topic for future research.

Remark 3.1.12. If $\hat{j}$ is uniformly convex and quadratic, then it is also self-concordant, a constant $C_{\|\bar{u}\|_U}$ that bounds the norm of the optimal solution is available, and $\hat{j}$ has bounded derivatives on bounded sets. For the existence of $C_{\|\bar{u}\|_U}$ see Lemma C.4.13. Hence, case II comprises case I. However, in case I we construct the barrier functionals differently and are, thereby, able to prove better convergence rates. This is why it pays off to not just work with the more general setting.

Remark 3.1.13. In case II it is possible to replace the self-concordance of $\hat{j}$ on U and the boundedness of the first derivative on bounded sets by the weaker assumption that self-concordance and boundedness of the first derivative only hold on $B_r(0)$ for an $r > 0$ with $r > 1 + \frac{1}{2}(\max\{\|u^\circ\|_U, C_{\|\bar{u}\|_U}\})^2$. However, this requires additional technicalities and we refrain from it.

The next lemma introduces an important constant.

Lemma 3.1.14. *There exists a constant $C_{\partial,C(\overline{\Omega}_a)} > 0$ with $\|A^{-1}Bu\|_{C(\overline{\Omega}_a)} \leq C_{\partial,C(\overline{\Omega}_a)} \|u\|_U$ for all $u \in U$ and $\|A^{-1}g\|_{C(\overline{\Omega}_a)} \leq C_{\partial,C(\overline{\Omega}_a)} \|g\|_Z$.*

Proof. This follows from the boundedness of A^{-1} and B in combination with the embeddings $Y \hookrightarrow C^{0,\beta}(\overline{\Omega}_a) \hookrightarrow C(\overline{\Omega}_a)$. □

3.2. A model problem and possible generalizations

We present a model problem that satisfies Assumption 3.1.9. We also discuss which type of problems fall under Assumption 3.1.9 and how the framework that we use in this thesis may be extended.

Example 3.2.1. We consider ($\mathrm{P}_{\mathrm{orig}}$) with $Y := H^2(\Omega) \cap H_0^1(\Omega)$, $U := L^2(\Omega)$, $Z := U$, $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, where Ω is a nonempty and bounded domain that is either convex or of class $\mathcal{C}^2$, $A := -\Delta \in \mathcal{L}(Y, Z)$, $B := -I \in \mathcal{L}(U, Z)$, $g \equiv 0$, and $\hat{J}(y, u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_{L^2(\Omega)}$ with $y_d \in L^2(\Omega)$ and $\hat{\alpha} > 0$. We take $\Omega_a := \Omega$ and use some $y_a \in C^{0,\beta}(\overline{\Omega}_a)$ with $y_a < 0$ on $\partial\Omega_a$, which is consistent with the homogeneous Dirichlet boundary conditions we imposed. Assumption 3.1.9 is fulfilled:

1) From Sobolev embedding theory we obtain that $Y \hookrightarrow C^{0,\beta}(\overline{\Omega}_a)$ holds for every $\beta \in (0, 1]$, $\beta \in (0, 1)$ and $\beta \in (0, 1/2)$, respectively, for $d = 1$, $d = 2$, and $d = 3$, cf., e.g., [Alt06, 8.13] and [Gri11, Theorem 1.4.4.1]. To apply these embedding theorems in the case of a convex Ω we note that convex domains are, in particular, Lipschitz domains, see [Gri11, Corollary 1.2.2.3]. Since we have $\Omega_a = \Omega$, this also implies that Ω_a is a nonempty and

bounded domain that satisfies the cone condition. Of course, U is a Hilbert space and Z is a Banach space.

2) The operator A is well-defined, linear, continuous, and invertible, as is well-known from regularity theory for linear elliptic PDEs of second order, cf. [GT83, Theorem 8.12] or [Eva10, Section 6.3.2, Theorem 4] for the case of a domain with $\mathcal{C}^2$-boundary, and [Gri11, Theorem 3.2.1.2] for the case of a convex domain.

3) Lemma 3.1.7 shows the existence of a unique optimal solution $\bar{u}$ of the reduced problem, provided a feasible point exists. In particular, this is satisfied if u° exists, see 6).

4) The objective $\hat{j}$ is uniformly convex with modulus $\hat{\alpha}$ and quadratic.

5) $y_a \in C^{0,\beta}(\overline{\Omega}_a)$ holds by assumption.

6) We can choose $u^\circ := -\Delta y^\circ$, where $y^\circ \in Y$ denotes a function with $y^\circ > y_a$ on $\overline{\Omega}_a$. The existence of such a function is, for example, ensured if $y_a < 0$ on $\overline{\Omega}_a$. (More generally, if $B \in \mathcal{L}(U, Z)$ is invertible and there is a function $y^\circ \in Y$ with $y^\circ > y_a$ on $\overline{\Omega}_a$, then we can use $u^\circ := B^{-1}(g - Ay^\circ)$.)

We comment on possible generalizations of the model problem that still satisfy Assumption 3.1.9. First of all, it is clear that 4) in this assumption allows for more general objectives, for instance $\hat{J}(y,u) = Q(y,u) + \frac{\hat{\alpha}}{2}\|u - u_0\|_U^2$, where $Q : Y \times U \to \mathbb{R}$ is quadratic and convex, and $u_0 \in U$. Second, we remark that the parts 3), 5), and 6) are standard assumptions in state constrained optimal control.

It remains to deal with 1) and 2). This comes down to the question which PDEs we can allow for the state equation $Ay + Bu = g$. We start with the case $\Omega = \Omega_a$. If we use $-\Delta y = u$ on $\Omega = \Omega_a$ with homogeneous Dirichlet boundary conditions as state equation, then Assumption 3.1.9 does not necessarily require that Ω is convex or $\mathcal{C}^2$. For instance, in the case $\Omega \subset \mathbb{R}^2$ we can work with arbitrary bounded Lipschitz domains: For these domains, $-\Delta y = u$ has a unique solution $y(u) \in H^{3/2-\delta}(\Omega) \cap H_0^1(\Omega)$ for every $u \in L^2(\Omega)$, where $\delta > 0$ is arbitrarily small, and $y(u)$ depends continuously on u, see [JK95, Theorem 0.5 (b)]. Moreover, for these domains we have the embedding $H^{3/2-\delta}(\Omega) \cap H_0^1(\Omega) \hookrightarrow C^{0,\beta}(\overline{\Omega}_a)$ for $\beta \in (0, 1/2 - \delta)$. In the case $\Omega \subset \mathbb{R}^d$ we have $y(u) \in H^2(\Omega) \cap H_0^1(\Omega)$ if the bounded Lipschitz domain Ω satisfies a so-called outer ball condition, cf. [Ado92, Theorem 1.1], which is more general than being $\mathcal{C}^2$ or convex. For $d \leq 3$ this allows for embeddings into Hölder spaces, as explained in the above example. We are also able to work with state equations where the control acts only on the boundary, e.g., $-\Delta y + y = g$, $\frac{\partial y}{\partial \nu} = u$ with $g \in L^2(\Omega)$, $u \in U = L^2(\partial\Omega)$ on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$; the unique solution $y(u)$ of this equation belongs to $H^{3/2-\delta}(\Omega)$, where $\delta > 0$ is arbitrarily small, cf. [Sav98, Theorem 4], and depends continuously on u. Reversing the roles of g and u, [Sav98, Theorem 4] implies that we also have $y(u) \in H^{3/2-\delta}(\Omega)$ with continuous dependence if we consider $-\Delta y + y = u$, $\partial y/\partial \nu = g$ on a bounded Lipschitz domain Ω. For both cases we obtain an embedding into $C^{0,\beta}(\overline{\Omega}_a)$ for $\beta \in (0, 1/2 - \delta)$ if $\Omega \subset \mathbb{R}^2$. We mention that in [Sav98] all results are proven for more general elliptic operators. Also, it is possible to consider in $\mathbb{R}^d$, $2 \leq d \leq 4$, linear elliptic PDEs of second order with *mixed* boundary conditions on bounded Lipschitz domains that satisfy a certain regularity

property, cf. [HDMRS09, Theorem 3.3]. Of course, many more regularity results are available; for a broad account see, e.g., [Ama93, Section 9].

In the case that the PDE is defined on Ω and the state constraints act on a subset $\overline{\Omega}_a \subset \Omega$, only *interior regularity* is required, i.e., it suffices if solutions of the PDE are Hölder continuous on $\overline{\Omega}_a \subset \Omega$ instead of $\overline{\Omega}$. In contrast to the case of *boundary regularity* from above this may be satisfied without assumptions on the regularity of Ω (and, since the Hölder continuity of y is induced by the PDE, without any assumption on the regularity of Ω_a other than the cone condition). For interior regularity of elliptic PDEs, see, e.g., [Eva10, Section 6.3] and [GT83, Section 8.3].

The case where $\Omega_a \subset \Omega$ but $\overline{\Omega}_a$ touches the boundary of Ω may be handled as the case $\Omega = \Omega_a$, but would also allow for more generality than this case, since regularity of the solution must only be valid on the parts of $\partial\Omega$ that intersect with $\overline{\Omega}_a$. In particular, this may be exploited in the case of mixed boundary conditions.

3.3. Two reformulations of the reduced problem

We consider different barrier functionals for ($\mathrm{P}_{\mathrm{red}}$) depending on whether $\hat{j}$ is uniformly convex and quadratic (case I) or only self-concordant (case II), cf. Assumption 3.1.9. The reason for this is that in the special case that $\hat{j}$ is uniformly convex and quadratic it turns out that the overall convergence rate is better if the construction of the barrier functional takes into account this special structure of the objective. This construction can, however, not be employed in the more general case that $\hat{j}$ is self-concordant. If $\hat{j}$ is uniformly convex and quadratic, we reformulate ($\mathrm{P}_{\mathrm{red}}$) as

$$\min_{u \in D_j} j(u) \quad \text{s.t.} \quad y(u)(x) \geq y_a(x) \ \ \forall x \in \overline{\Omega}_a,$$

with

$$D_j := \{u \in U : C_{\hat{j}} - \hat{j}(u) > 0\}, \quad C_{\hat{j}} := 1 + \hat{j}(u^\circ), \quad j(u) := -C_j \ln(C_{\hat{j}} - \hat{j}(u)), \ \text{and} \ C_j > 0.$$

It is obvious that D_j is nonempty, open, and convex, that j is a barrier function for D_j, and that this problem possesses the same unique minimizer as ($\mathrm{P}_{\mathrm{red}}$).

To have a unified notation in case I and case II we consider from now on the problem

$$\min_{u \in D_j} j(u) \quad \text{s.t.} \quad y(u)(x) \geq y_a(x) \ \ \forall x \in \overline{\Omega}_a, \tag{P}$$

where D_j is nonempty, open, and convex, j is a barrier function for D_j, and the set of optimal solutions of ($\mathrm{P}_{\mathrm{red}}$) coincides with the one of (P). If $\hat{j}$ is *not* uniformly convex and quadratic, we can achieve this by use of $j := C_j \hat{j}$ with $C_j := 1$, and $D_j := U$. Thus, (P) comprises both settings we are interested in, but j and D_j differ in these settings. For clarity, let us repeat:

- In case I we employ $D_j := \{u \in U : C_{\hat{j}} - \hat{j}(u) > 0\}$, $C_{\hat{j}} := 1 + \hat{j}(u^\circ)$, $j(u) := -C_j \ln(C_{\hat{j}} - \hat{j}(u))$, and $C_j > 0$ in (P);
- In case II we use $D_j := U$ and $j := C_j \hat{j}$ with $C_j := 1$ in (P).

Our aim from now on is to solve (P).

3.4. KKT conditions

To state the necessary and sufficient optimality conditions of (P) in a convenient form we introduce the following definition.

Definition 3.4.1. We call $\lambda \in C(\overline{\Omega}_a)^*$ *nonpositive* and write $\lambda \leq 0$ iff it satisfies

$$\langle \lambda, y \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)} \leq 0$$

for all $y \in C_{\geq 0}(\overline{\Omega}_a)$. Here, $C_{\geq 0}(\overline{\Omega}_a) := \{y \in C(\overline{\Omega}_a) : y(x) \geq 0 \ \forall x \in \overline{\Omega}_a\} \subset C(\overline{\Omega}_a)$ denotes the cone of nonnegative continuous functions on $\overline{\Omega}_a$.

Remark 3.4.2. The space $C(\overline{\Omega}_a)^*$ can be identified with the space of regular Borel measures on $\overline{\Omega}_a$, cf. [Alt06, 4.22]. A more thorough introduction into this subject can be found in, e.g., [Bau01, §29].

The KKT conditions for a minimizer $\bar{u} \in D_j$ of (P) read as follows.

Lemma 3.4.3. *The point $\bar{u} \in D_j$ is a minimizer of* (P) *if and only if there exists $\bar{\lambda} \in C(\overline{\Omega}_a)^*$ with*

$$\begin{aligned} &j'(\bar{u}) + T^* \bar{\lambda} = 0 \quad \text{in } U^*, \\ &\bar{y} \geq y_a \text{ in } \overline{\Omega}_a, \ \ \bar{\lambda} \leq 0, \ \ \langle \bar{\lambda}, \bar{y} - y_a \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)} = 0. \end{aligned}$$

Here, we used $\bar{y} := y(\bar{u})$ and $T := -A^{-1}B \in \mathcal{L}(U, Y)$.

Remark 3.4.4. The equation in the first line is well-defined due to $Y \hookrightarrow C(\overline{\Omega}_a)$.

Remark 3.4.5. Since (P) is convex, the KKT conditions are sufficient for (local=global) optimality. This assertion can be proven similarly as in the finite-dimensional case, cf., e.g., [GK02, Satz 2.46]; for a proof in the infinite-dimensional setting see, e.g., [Ulb11a, Theorem 3.21]. Also, we mention that we never use the fact that the KKT conditions are sufficent for optimality.

Proof. We argue that the stated conditions are necessary for optimality. We start with case II.

Case II

Here, ($\mathrm{P}_{\mathrm{red}}$) and (P) coincide. We establish that ($\mathrm{P}_{\mathrm{red}}$) satisfies the KKT conditions. To this end, we first argue for the original problem ($\mathrm{P}_{\mathrm{orig}}$). In this problem we consider the pointwise inequality constraints as $(y(u) - y_a) \in C_{\geq 0}(\overline{\Omega}_a)$ and note that $C_{\geq 0}(\overline{\Omega}_a) \subset C(\overline{\Omega}_a)$ is a closed convex cone. By assumption there exists y° with $y^\circ - \tau^\circ \geq y_a$. Obviously, $y^\circ - y_a \in C(\overline{\Omega}_a)$ is an interior point of $C_{\geq 0}(\overline{\Omega}_a)$. Moreover, $(y^\circ, u^\circ) \in Y \times U$ is feasible and A is surjective. Together, these facts imply that Robinson's constraint qualification holds at every feasible point $(y, u) \in Y \times U$ of ($\mathrm{P}_{\mathrm{orig}}$), see [HPUU09, Lemma 1.14, p. 85]. Since $\bar{u}$ is a solution of ($\mathrm{P}_{\mathrm{red}}$), $(\bar{y}, \bar{u})$ with $\bar{y} := y(\bar{u})$ is a solution of ($\mathrm{P}_{\mathrm{orig}}$). Thus, the KKT conditions are fulfilled at $(\bar{y}, \bar{u})$. This implies that there exist $\bar{\lambda} \in C(\overline{\Omega}_a)^*$ and $\bar{p} \in Z^*$ such that there hold

$$\begin{aligned} \hat{J}'_y(\bar{y}, \bar{u}) + \bar{\lambda} + A^* \bar{p} &= 0 \quad \text{in } Y^*, \\ \hat{J}'_u(\bar{y}, \bar{u}) + B^* \bar{p} &= 0 \quad \text{in } U^*, \\ A\bar{y} + B\bar{u} &= g \quad \text{in } Z, \\ (\bar{y} - y_a) \in C_{\geq 0}(\overline{\Omega}_a), \ \ \bar{\lambda} \leq 0, \ \ \langle \bar{\lambda}, \bar{y} - y_a \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)} &= 0, \end{aligned}$$

see [UU12, Section 1.7.3.4]. From the first equation we deduce $\bar{p} = -A^{-*}(\hat{J}'_y(\bar{y},\bar{u}) + \bar{\lambda})$. Inserting this into the second equation we obtain

$$\hat{J}'_u(\bar{y},\bar{u}) - B^*\left(A^{-*}\left(\hat{J}'_y(\bar{y},\bar{u}) + \bar{\lambda}\right)\right) = 0 \quad \text{in } U^*.$$

Since it holds $\hat{j}'(\bar{u}) = \hat{J}'_u(\bar{y},\bar{u}) + T^*\hat{J}'_y(\bar{y},\bar{u})$ with $T := -A^{-1}B$ by the chain rule, we infer that

$$\hat{j}'(\bar{u}) + T^*\bar{\lambda} = \hat{J}'_u(\bar{y},\bar{u}) + T^*\left(\hat{J}'_y(\bar{y},\bar{u}) + \bar{\lambda}\right) = \hat{J}'_u(\bar{y},\bar{u}) - B^*\left(A^{-*}\left(\hat{J}'_y(\bar{y},\bar{u}) + \bar{\lambda}\right)\right) = 0$$

is true in U^*. This establishes the assertion for case II.

Case I
As we have already established, the KKT conditions are satisfied for ($\mathrm{P}_{\mathrm{red}}$) at a minimizer $\bar{u}$, i.e., there exists $\bar{\lambda} \in C(\overline{\Omega}_a)^*$ with

$$\begin{aligned} &\hat{j}'(\bar{u}) + T^*\bar{\lambda} = 0 \quad \text{in } U^*, \\ &\bar{y} \geq y_a \text{ in } \overline{\Omega}_a, \;\; \bar{\lambda} \leq 0, \;\; \langle \bar{\lambda}, \bar{y} - y_a \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)} = 0, \end{aligned}$$

where $\bar{y} = y(\bar{u})$. For the derivative of the functional $j : D_j \to \mathbb{R}$, $j(u) := -C_j \ln(C_{\hat{j}} - \hat{j}(u))$ at $\bar{u}$ we compute

$$j'(\bar{u}) = C_j \cdot \frac{\hat{j}'(\bar{u})}{C_{\hat{j}} - \hat{j}(\bar{u})} = r \cdot \hat{j}'(\bar{u})$$

with $r := \frac{C_j}{C_{\hat{j}} - \hat{j}(\bar{u})} > 0$. Hence, the KKT conditions of ($\mathrm{P}_{\mathrm{red}}$) for $(\bar{u}, \bar{\lambda})$ imply the KKT conditions of (P) in case I for $(\bar{u}, r\bar{\lambda})$. □

3.5. Associated barrier problems

In this section we construct self-concordant barrier functions that can be used to solve (P).

Employing the non-differentiable functional $\min_{x \in \overline{\Omega}_a} : C(\overline{\Omega}_a) \to \mathbb{R}$ we can reformulate (P) equivalently as

$$\min_{u \in D_j} j(u) \qquad \text{s.t.} \qquad \min_{x \in \overline{\Omega}_a} \big(y(u)(x) - y_a(x)\big) \geq 0.$$

In this reformulation we have transferred the infinitely many inequality constraints of (P) to a single nonsmooth inequality constraint. The underlying idea is that we can handle finitely many constraints very well by barrier methods using the theory developed in Section 2, but that for infinitely many constraints this theory is not applicable: For finitely many constraints we can use one barrier per constraint and obtain a barrier for the entire optimization problem by summation. This is common practice in finite-dimensional optimization, but is not sensible for infinitely many constraints since, for instance, the constant of self-boundedness would be infinite. This follows from the fact that for m linear constraints, this constant cannot be smaller than m, cf. [Ren01, Section 2.3.3]. However, it turns out that it is possible to construct self-concordant barriers using the above reformulation with just a single constraint. A first idea may be to consider $-\ln(\min_{x \in \overline{\Omega}_a}(y(u)(x) - y_a(x)))$. Yet, by definition we require a self-concordant barrier to be thrice Fréchet differentiable. Thus, to continue in this direction we smooth the minimum function $\min_{x \in \overline{\Omega}_a}$ and then apply the negative logarithm.

Definition 3.5.1. We call the mapping

$$\min_\varepsilon : C(\overline{\Omega}_a) \to \mathbb{R}, \quad \min_\varepsilon(y) := -\varepsilon \ln\left(\frac{\int_{\Omega_a} e^{-y(x)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}\right)$$

the *smoothed minimum* with *smoothing parameter* $\varepsilon > 0$.

Remark 3.5.2. The finite-dimensional version of this function is well-known in optimization. It has been used in, e.g., [BTT89, CM95, Aus99, CQQT04].

The smoothed minimum is well-defined on $C(\overline{\Omega}_a)$ since $\overline{\Omega}_a$ is compact. Using it we obtain the following family of problems:

$$\min_{u\in D_j} j(u) \qquad \text{s.t.} \qquad \min_\varepsilon\big(y(u) - y_a\big) \geq 0. \tag{3.1}$$

Since we need the mapping $\min_\varepsilon(y - y_a)$ very often, we abbreviate it.

Definition 3.5.3. For every $\varepsilon > 0$ we denote by $B^\varepsilon_{C(\overline{\Omega}_a)}$ the mapping

$$B^\varepsilon_{C(\overline{\Omega}_a)} : C(\overline{\Omega}_a) \to \mathbb{R}, \quad B^\varepsilon_{C(\overline{\Omega}_a)}(y) := \min_\varepsilon(y - y_a).$$

Furthermore, we use the mapping B^ε, which we define by

$$B^\varepsilon : U \to \mathbb{R}, \quad B^\varepsilon(u) := B^\varepsilon_{C(\overline{\Omega}_a)}\big(y(u)\big).$$

We set $D_{b^\varepsilon} := \{u \in U : B^\varepsilon(u) > 0\}$ and create the barrier functional b^ε by

$$b^\varepsilon : D_{b^\varepsilon} \to \mathbb{R}, \quad b^\varepsilon(u) := -\tau(\varepsilon)\ln\big(B^\varepsilon(u)\big),$$

where $\tau(\varepsilon) > 0$.

Later we develop algorithms in which we drive ε to zero, but for the moment it is helpful to assume that ε is fixed and that we would like to solve (3.1) for this particular ε by use of self-concordant barrier functions.

We point out that the idea how to construct suitable barriers is motivated by [TN10] and the references therein, in particular [Nem04]. Note, however, that the infinite-dimensional case is not considered there and that a sum of n barriers occurs, where n denotes the dimension of the optimization variable.

3.5.1. A suitable barrier function for case I

In this section we introduce and investigate the barrier function that we use in case I.

Definition 3.5.4. For $\varepsilon > 0$ we define $U_{\mathrm{ad}}(\varepsilon) := D_j \cap D_{b^\varepsilon}$ in case I.

Definition 3.5.5. In case I we use for $\varepsilon, \mu > 0$ the barrier function

$$f_{\varepsilon,\mu} : U_{\mathrm{ad}}(\varepsilon) \to \mathbb{R}, \quad f_{\varepsilon,\mu}(u) := \frac{j(u)}{\mu} + b^\varepsilon(u).$$

To apply the framework of Section 2 we want to show that for fixed ε and suitable $\mu > 0$, $f_{\varepsilon,\mu}$ satisfies Assumption 2.5.2. This is the goal of this section. The key step is to argue that $f_{\varepsilon,\mu}$ is self-concordant. This can be done using Lemma 2.4.3. To apply Lemma 2.4.3 we have to make sure that B^ε is appropriate. To prove this in a concise way we first establish the following auxiliary result.

Lemma 3.5.6. *Let $\varepsilon > 0$ be given. For fixed $y, h \in Y$ define*

$$p(x) := \frac{e^{-(y(x)-y_a(x))/\varepsilon}}{\int_{\Omega_a} e^{-(y(\tilde{x})-y_a(\tilde{x}))/\varepsilon}\,\mathrm{d}\tilde{x}}, \qquad \mu := \int_{\Omega_a} p(x)\cdot\frac{h(x)}{\varepsilon}\,\mathrm{d}x, \qquad \text{and} \quad s(x) := \frac{h(x)}{\varepsilon} - \mu.$$

Then there hold:

$$\Big(B^\varepsilon_{C(\overline{\Omega}_a)}\Big)''(y)[h,h] = -\varepsilon\int_{\Omega_a} ps^2\,\mathrm{d}x \qquad \text{and} \qquad \Big(B^\varepsilon_{C(\overline{\Omega}_a)}\Big)'''(y)[h,h,h] = \varepsilon\int_{\Omega_a} ps^3\,\mathrm{d}x.$$

Proof. By definition we have $B^\varepsilon_{C(\overline{\Omega}_a)}(y) = -\varepsilon\ln\Big(\frac{\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}\Big)$. Hence, Corollary C.2.18 and Corollary C.2.10 imply in combination with the product rule that $y \mapsto B^\varepsilon_{C(\overline{\Omega}_a)}(y)$ is thrice Fréchet differentiable. Setting $H(y) := \ln(\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}\,\mathrm{d}x)$ it suffices to prove

$$1)\ \ H''(y)[h,h] = \int_{\Omega_a} ps^2\,\mathrm{d}x \qquad \text{and} \qquad 2)\ \ H'''(y)[h,h,h] = -\int_{\Omega_a} ps^3\,\mathrm{d}x.$$

In the following we abbreviate $f := e^{-(y-y_a)/\varepsilon}$.

To 1): From $ps^2 = p\left(\frac{h}{\varepsilon}\right)^2 - 2p\frac{h}{\varepsilon}\int_{\Omega_a} p\frac{h}{\varepsilon}\,\mathrm{d}x + p\left(\int_{\Omega_a} p\frac{h}{\varepsilon}\,\mathrm{d}x\right)^2$ and $\int_{\Omega_a} p\,\mathrm{d}x = 1$ we infer

$$\begin{aligned} H''(y)[h,h] &= \frac{\int_{\Omega_a} f\cdot\left(\frac{h}{\varepsilon}\right)^2\mathrm{d}x}{\int_{\Omega_a} f\,\mathrm{d}x} - \left(\frac{\int_{\Omega_a} f\frac{h}{\varepsilon}\,\mathrm{d}x}{\int_{\Omega_a} f\,\mathrm{d}x}\right)^2 \\ &= \int_{\Omega_a} p\cdot\left(\frac{h}{\varepsilon}\right)^2\mathrm{d}x - \left(\int_{\Omega_a} p\frac{h}{\varepsilon}\,\mathrm{d}x\right)^2 = \int_{\Omega_a} ps^2\,\mathrm{d}x. \end{aligned}$$

To 2): For the third directional derivative we obtain

$$\begin{aligned} &H'''(y)[h,h,h] \\ &= 3\frac{\int_{\Omega_a} f\frac{h}{\varepsilon}\,\mathrm{d}x\cdot\int_{\Omega_a} f\left(\frac{h}{\varepsilon}\right)^2\mathrm{d}x}{\left(\int_{\Omega_a} f\,\mathrm{d}x\right)^2} - 2\left(\frac{\int_{\Omega_a} f\frac{h}{\varepsilon}\,\mathrm{d}x}{\int_{\Omega_a} f\,\mathrm{d}x}\right)^3 - \frac{\int_{\Omega_a} f\left(\frac{h}{\varepsilon}\right)^3\mathrm{d}x}{\int_{\Omega_a} f\,\mathrm{d}x} \\ &= 3\left(\int_{\Omega_a} p(s+\mu)\,\mathrm{d}x\right)\left(\int_{\Omega_a} p(s+\mu)^2\,\mathrm{d}x\right) - 2\left(\int_{\Omega_a} p(s+\mu)\,\mathrm{d}x\right)^3 - \int_{\Omega_a} p(s+\mu)^3\,\mathrm{d}x \\ &= 3\left(\int_{\Omega_a} p\mu\,\mathrm{d}x\right)\left(\int_{\Omega_a} ps^2\,\mathrm{d}x + \int_{\Omega_a} p\mu^2\,\mathrm{d}x\right) - 2\left(\int_{\Omega_a} p\mu\,\mathrm{d}x\right)^3 - \int_{\Omega_a} p(s+\mu)^3\,\mathrm{d}x. \end{aligned}$$

Here, we used $\int_{\Omega_a} ps\,\mathrm{d}x = 0$. From $\int_{\Omega_a} p\mu\,\mathrm{d}x = \mu\int_{\Omega_a} p\,\mathrm{d}x = \mu$ we deduce

$$\begin{aligned} H'''(y)[h,h,h] &= 3\mu\left(\int_{\Omega_a} ps^2\,\mathrm{d}x + \mu^2\right) - 2\mu^3 - \int_{\Omega_a} p(s^3 + 3s^2\mu + 3s\mu^2 + \mu^3)\,\mathrm{d}x \\ &= -\int_{\Omega_a} ps^3\,\mathrm{d}x - 3\mu^2\int_{\Omega_a} ps\,\mathrm{d}x = -\int_{\Omega_a} ps^3\,\mathrm{d}x. \end{aligned}$$

□

In the next lemma we show that $y \mapsto B^{\varepsilon}_{C(\overline{\Omega}_a)}(y)$ is an appropriate mapping.

Lemma 3.5.7. *Fix $\varepsilon > 0$. Let $C, \hat{C} \in \mathbb{R}$ with $\hat{C} \leq C$. Let $y, h \in Y$ with $y \leq C$ and $y \pm h \geq \hat{C}$ in $\overline{\Omega}_a$ be given. Then it holds*

$$\left(B^{\varepsilon}_{C(\overline{\Omega}_a)}\right)'''(y)[h,h,h] \leq -2\frac{C-\hat{C}}{\varepsilon}\left(B^{\varepsilon}_{C(\overline{\Omega}_a)}\right)''(y)[h,h].$$

In particular, $B^{\varepsilon}_{C(\overline{\Omega}_a)} : Y \to \mathbb{R}$ is β-appropriate on $\{y \in Y : \hat{C} \leq y \leq C\}$ and any convex subset thereof if β satisfies $\beta \geq \frac{2(C-\hat{C})}{3\varepsilon}$.

Proof. We start by proving the asserted inequality. Since $|h| \leq y - \hat{C}$ holds in $\overline{\Omega}_a$, we have $\left|\frac{h}{\varepsilon}\right| \leq \frac{y-\hat{C}}{\varepsilon} \leq \frac{C-\hat{C}}{\varepsilon}$. Defining p, μ, and s as in Lemma 3.5.6 we compute

$$|\mu| \leq \int_{\Omega_a} \left|p \cdot \frac{h}{\varepsilon}\right| \mathrm{d}x \leq \int_{\Omega_a} |p| \cdot \left|\frac{C-\hat{C}}{\varepsilon}\right| \mathrm{d}x \leq \frac{C-\hat{C}}{\varepsilon},$$

where we used $\int_{\Omega_a} |p| \,\mathrm{d}x = 1$. Together, this implies

$$|s| = \left|\frac{h}{\varepsilon} - \mu\right| \leq \left|\frac{h}{\varepsilon}\right| + |\mu| \leq 2\frac{C-\hat{C}}{\varepsilon}.$$

With Lemma 3.5.6 we deduce

$$\begin{aligned}\left(B^{\varepsilon}_{C(\overline{\Omega}_a)}\right)'''(y)[h,h,h] &\leq \varepsilon \int_{\Omega_a} ps^2 \,|s| \,\mathrm{d}x \\ &\leq 2(C-\hat{C}) \int_{\Omega_a} ps^2 \,\mathrm{d}x = -2\frac{C-\hat{C}}{\varepsilon}\left(B^{\varepsilon}_{C(\overline{\Omega}_a)}\right)''(y)[h,h].\end{aligned}$$

The second assertion ("In particular, ...") is obvious: In order for $B^{\varepsilon}_{C(\overline{\Omega}_a)}$ to be appropriate on the convex set $\{y \in Y : \hat{C} \leq y \leq C\}$ we only have to consider y with $\hat{C} \leq y \leq C$, and $h \in Y$ with $\hat{C} \leq y \pm h \leq C$, which comprises in particular the inequalities that were needed to prove the first assertion. Moreover, $B^{\varepsilon}_{C(\overline{\Omega}_a)} : Y \to \mathbb{R}$ is concave, see its second directional derivatives in Lemma 3.5.6. In conclusion, $B^{\varepsilon}_{C(\overline{\Omega}_a)}$ is appropriate on the aforementioned set with the asserted constant. Moreover, it follows directly from the definition that a function that is appropriate on a set is also appropriate with the same constant on every convex subset thereof. Finally, it is clear from the definition that any β-appropriate function remains appropriate if β is enlarged. This concludes the proof. □

As main result of this section we show that $f_{\varepsilon,\mu} = \frac{j}{\mu} + b^{\varepsilon} = -\frac{C_j \ln(C_{\hat{j}} - \hat{j})}{\mu} - \tau(\varepsilon)\ln(B^{\varepsilon})$ is a self-concordant and self-bounded barrier function for $U_{\mathrm{ad}}(\varepsilon)$ if C_j is large enough.

Lemma 3.5.8. *Let $\varepsilon > 0$ and $\tau(\varepsilon) \geq 1$. Let $C_\mu > 0$ and let C_j satisfy*

$$C_j \geq \frac{C_\mu}{\varepsilon^2} \max\left\{\varepsilon^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} \left(\frac{2\|\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U + \|g\|_Z\right)^2\right\}$$

with a $\tilde{u} \in U$ that fulfills $\hat{j}(\tilde{u}) \geq C_{\hat{j}} = 1 + \hat{j}(u^\circ)$. Then $f_{\varepsilon,\mu} = \frac{j}{\mu} + b^\varepsilon$ is a nondegenerate $(\frac{C_j}{\mu} + \tau(\varepsilon))$-self-concordant barrier function for $U_{ad}(\varepsilon)$ if $\mu \in (0, C_\mu]$ holds.

Moreover, $f_{\varepsilon,\mu}$ is uniformly convex with modulus $\beta(\varepsilon,\mu) := \frac{\alpha}{\mu}$, where $\alpha > 0$ denotes the convexity modulus of j on $U_{ad}(\varepsilon)$.

Remark 3.5.9. In Lemma C.4.14 we show that j is uniformly convex on $D_j \supset U_{\mathrm{ad}}(\varepsilon)$ and link the convexity modulus α of j to the convexity modulus $\hat{\alpha}$ of $\hat{j}$.

Remark 3.5.10. For the standard example $\hat{j}(u) = \frac{1}{2}\|y(u)-y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_{L^2(\Omega)}$ with $y_d \in L^2(\Omega)$ and state equation $-\Delta y = u$ with homogeneous Dirichlet boundary conditions we readily compute $\|\hat{j}'(\tilde{u})\|_{U^*} = \|q'(\tilde{u})\|_{U^*} + \hat{\alpha}\|\tilde{u}\|_U = \|p(\tilde{u})\|_{L^2(\Omega)} + \hat{\alpha}\|\tilde{u}\|_{L^2(\Omega)}$, where $q(u) := \frac{1}{2}\|y(u) - y_d\|^2_{L^2(\Omega)}$ and $p(\tilde{u}) \in H^2(\Omega) \cap H^1_0(\Omega)$ solves $-\Delta p = y(\tilde{u}) - y_d$. The quantity $\|\hat{j}'(\tilde{u})\|_{U^*}$ can, hence, easily be evaluated numerically.

Proof. We want to apply Lemma 2.4.3. Using Lemma 2.1.19 as well as Corollary 2.3.9 it follows that j/C_j is a 1-self-concordant barrier function for D_j. From Lemma C.4.13 we deduce that $D_j = \{u \in U : \hat{j}(u) < C_{\hat{j}}\}$ is bounded by $2\|\hat{j}'(\tilde{u})\|_{U^*}/\hat{\alpha} + \|\tilde{u}\|_U$, where $\tilde{u}$ satisfies $\hat{j}(\tilde{u}) \geq C_{\hat{j}}$. Due to $\|y(u)\|_{C(\overline{\Omega}_a)} \leq C_{\partial,C(\overline{\Omega}_a)}\,(\|u\|_U + \|g\|_Z)$, see Lemma 3.1.14, we have for all $u \in D_j$

$$\|y(u)\|_{C(\overline{\Omega}_a)} \leq \underbrace{C_{\partial,C(\overline{\Omega}_a)}\left(\frac{2\|\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U + \|g\|_Z\right)}_{=:\gamma}. \tag{3.2}$$

Define $T := -A^{-1}B \in \mathcal{L}(U,Y)$ and let $u \in D_j$ and $h \in U$ with $u \pm h \in D_j$ be given. From (3.2) we infer

$$\|y(u) \pm T(h)\|_{C(\overline{\Omega}_a)} = \|y(u \pm h)\|_{C(\overline{\Omega}_a)} \leq \gamma \qquad \text{and} \qquad \|y(u)\|_{C(\overline{\Omega}_a)} \leq \gamma.$$

Together with Lemma 3.5.7 this implies

$$\begin{aligned}(B^\varepsilon)'''(u)[h,h,h] &= \big(B^\varepsilon_{C(\overline{\Omega}_a)}\big)'''(y(u))[T(h),T(h),T(h)] \\ &\leq \frac{-4\gamma}{\varepsilon}\big(B^\varepsilon_{C(\overline{\Omega}_a)}\big)''(y(u))[T(h),T(h)] = \frac{-4\gamma}{\varepsilon}(B^\varepsilon)''(u)[h,h].\end{aligned}$$

This shows that $B^\varepsilon : D_j \to \mathbb{R}$ is β-appropriate if β satisfies $\beta \geq \frac{4\gamma}{3\varepsilon}$. Hence, all prerequisites of Lemma 2.4.3 are fulfilled (use $\mathcal{A} = B^\varepsilon$, $f = j/C_j$, $K = D_j$, $E = U_{\mathrm{ad}}(\varepsilon) \ni u^\circ$, $C = \tau(\varepsilon) \geq 1$, $\hat{C} = C_j/\mu$ in this lemma). This yields that $f_{\varepsilon,\mu}$ is a $(\frac{C_j}{\mu} + \tau(\varepsilon))$-self-concordant barrier function for $U_{\mathrm{ad}}(\varepsilon)$ if C_j is chosen according to $\frac{C_j}{\mu} \geq \max\{1, \left(\frac{4\gamma}{3\varepsilon}\right)^2\}$. For $\mu \leq C_\mu$ it is, thus, sufficient to choose C_j such that it holds

$$C_j \geq C_\mu \max\left\{1, \left(\frac{4\gamma}{3\varepsilon}\right)^2\right\} = \frac{C_\mu}{\varepsilon^2}\max\left\{\varepsilon^2, \frac{16\gamma^2}{9}\right\}.$$

Inserting the definition of γ we obtain the asserted estimate. The uniform convexity of $f_{\varepsilon,\mu}$ implies that $f_{\varepsilon,\mu}$ is nondegenerate, see Theorem C.4.15. It follows from the uniform convexity of j, which is itself a consequence of Lemma C.4.14. □

Lemma 3.5.11. *For $\varepsilon > 0$, b^ε is $\tau(\varepsilon)$-self-bounded on $U_{ad}(\varepsilon)$.*

Proof. This follows directly from Corollary 2.3.9. □

Corollary 3.5.12. *Under the assumptions of Lemma 3.5.8, $f_{\varepsilon,\mu}$ satisfies Assumption 2.5.2 for $M := D_j \cap \overline{D}_{b^\varepsilon}$, $K := U_{ad}(\varepsilon) = D_j \cap D_{b^\varepsilon}$, $\mu_s := C_\mu$, and $\vartheta_b := \tau(\varepsilon)$.*

Proof. It follows from Lemma C.4.2 that $K \subset M \subset \overline{K}$ is satisfied. Furthermore, $f_{\varepsilon,\mu}$ possesses a global minimizer on $U_{\mathrm{ad}}(\varepsilon)$ for $\mu = \mu_s$. The existence of such a minimizer follows from Corollary C.4.6 since $U_{\mathrm{ad}}(\varepsilon) \subset D_j$ is nonempty, convex, and bounded. The remaining parts of Assumption 2.5.2 follow from Lemma 3.5.8. □

Definition 3.5.13. For $\varepsilon > 0$ we set $\vartheta(\varepsilon) := \tau(\varepsilon)$ in case I.

3.5.2. A suitable barrier function for case II

In this section we construct a self-concordant barrier function $f_{\varepsilon,\mu}$ for case II. In this case the function $\frac{j}{\mu} + b^\varepsilon$ that we used in case I may not be nondegenerate, e.g., if $j = \hat{j}$ is linear. Therefore, we add a uniformly convex barrier term. We also need this term to argue that $f_{\varepsilon,\mu}$ is self-concordant.

Definition 3.5.14. In case II we define

$$\tilde{B} : U \to \mathbb{R}, \quad \tilde{B}(u) := C_{\|\cdot\|} - \frac{1}{2}\|u\|_U^2,$$

where $C_{\|\cdot\|}$ is given by $C_{\|\cdot\|} := 1 + \frac{1}{2}(\max\{\|u^\circ\|_U, C_{\|\bar{u}\|_U}\})^2$. We set $D_{\tilde{b}^\varepsilon} := \{u \in U : \tilde{B}(u) > 0\}$ and define for $\varepsilon > 0$ the barrier functional $\tilde{b}^\varepsilon$ by

$$\tilde{b}^\varepsilon : D_{\tilde{b}^\varepsilon} \to \mathbb{R}, \quad \tilde{b}^\varepsilon(u) := -\tilde{\tau}(\varepsilon)\ln(\tilde{B}(u)),$$

where $\tilde{\tau}(\varepsilon) > 0$.

Definition 3.5.15. For $\varepsilon > 0$ we define $U_{\mathrm{ad}}(\varepsilon) := D_{b^\varepsilon} \cap D_{\tilde{b}^\varepsilon}$ in case II.

Before we introduce the barrier function $f_{\varepsilon,\mu}$, we demonstrate that $b^\varepsilon + \tilde{b}^\varepsilon = -\tau(\varepsilon)\ln(B^\varepsilon) - \tilde{\tau}(\varepsilon)\ln(\tilde{B})$ is a self-concordant and self-bounded barrier function for $U_{\mathrm{ad}}(\varepsilon)$ if $\tilde{\tau}(\varepsilon)$ is large enough.

Lemma 3.5.16. *Let $\varepsilon > 0$ and $\tau(\varepsilon) \geq 1$. Then $b^\varepsilon + \tilde{b}^\varepsilon$ is a $(\tau(\varepsilon) + \tilde{\tau}(\varepsilon))$-self-concordant barrier function for $U_{ad}(\varepsilon)$ if $\tilde{\tau}(\varepsilon)$ satisfies*

$$\tilde{\tau}(\varepsilon) \geq \frac{1}{\varepsilon^2}\max\left\{\varepsilon^2, \frac{16}{9}C_{\partial,C(\overline{\Omega}_a)}^2\left(\sqrt{2C_{\|\cdot\|}} + \|g\|_Z\right)^2\right\}.$$

Proof. To establish the assertion we use Lemma 2.4.3. Let us start by arguing that this lemma is applicable. From Lemma 2.1.19 and Corollary 2.3.9 it follows that $\tilde{b}^\varepsilon/\tilde{\tau}(\varepsilon)$ is a 1-self-concordant barrier function for $D_{\tilde{b}^\varepsilon}$. Furthermore, for $u \in D_{\tilde{b}^\varepsilon} = \{u \in U : \frac{1}{2}\|u\|_U^2 < C_{\|\cdot\|}\}$ we obviously have the bound $\|u\|_U \leq \sqrt{2C_{\|\cdot\|}}$. Together with $\|y(u)\|_{C(\overline{\Omega}_a)} \leq C_{\partial,C(\overline{\Omega}_a)}(\|u\|_U + \|g\|_Z)$, see Lemma 3.1.14, we obtain

$$\|y(u)\|_{C(\overline{\Omega}_a)} \leq \underbrace{C_{\partial,C(\overline{\Omega}_a)}\left(\sqrt{2C_{\|\cdot\|}} + \|g\|_Z\right)}_{=:\gamma} \tag{3.3}$$

for all $u \in D_{\tilde{b}^\varepsilon}$. This yields for all $u \in D_{\tilde{b}^\varepsilon}$ and all $h \in U$ with $u \pm h \in D_{\tilde{b}^\varepsilon}$ the estimate

$$\|y(u) \pm T(h)\|_{C(\overline{\Omega}_a)} = \|y(u \pm h)\|_{C(\overline{\Omega}_a)} \leq \gamma,$$

where $T := -A^{-1}B \in \mathcal{L}(U,Y)$. Using this and (3.3) in combination with Lemma 3.5.7, we deduce for these u, h that

$$\begin{aligned}(B^\varepsilon)'''(u)[h,h,h] &= (B^\varepsilon_{C(\overline{\Omega}_a)})'''(y(u))[T(h),T(h),T(h)] \\ &\leq \frac{-4\gamma}{\varepsilon}(B^\varepsilon_{C(\overline{\Omega}_a)})''(y(u))[T(h),T(h)] = \frac{-4\gamma}{\varepsilon}(B^\varepsilon)''(u)[h,h]\end{aligned}$$

holds. Thus, B^ε is β-appropriate on $D_{\tilde{b}^\varepsilon}$ if β satisfies $\beta \geq \frac{4\gamma}{3\varepsilon}$. Hence, all prerequisites of Lemma 2.4.3 are fulfilled (use $\mathcal{A} = B^\varepsilon$, $f = \tilde{b}^\varepsilon/\tilde{\tau}(\varepsilon)$, $K = D_{\tilde{b}^\varepsilon}$, $E = U_{\mathrm{ad}}(\varepsilon) \ni u^\circ$, $C = \tau(\varepsilon) \geq 1$, and $\hat{C} = \tilde{\tau}(\varepsilon)$ in this lemma). This yields that $b^\varepsilon + \tilde{b}^\varepsilon$ is a $(\tau(\varepsilon) + \tilde{\tau}(\varepsilon))$-self-concordant barrier function for $U_{\mathrm{ad}}(\varepsilon)$ if $\tilde{\tau}(\varepsilon)$ is chosen according to $\tilde{\tau}(\varepsilon) \geq \max\{1, (\frac{4\gamma}{3\varepsilon})^2\}$, which establishes the assertion. □

Definition 3.5.17. In case II we use for $\varepsilon, \mu > 0$ the barrier function

$$f_{\varepsilon,\mu} : U_{\mathrm{ad}}(\varepsilon) \to \mathbb{R}, \quad f_{\varepsilon,\mu}(u) := \frac{j(u)}{\mu} + b^\varepsilon(u) + \tilde{b}^\varepsilon(u).$$

As main result of this section we obtain that $f_{\varepsilon,\mu}$ is a self-concordant and self-bounded barrier function for $U_{\mathrm{ad}}(\varepsilon)$ if $\tilde{\tau}(\varepsilon)$ is large enough.

Lemma 3.5.18. *Let $\varepsilon > 0$, $\tau(\varepsilon) \geq 1$, and $C_\mu \leq 1$. Then for every $\mu \in (0, C_\mu]$ the function $f_{\varepsilon,\mu} = \frac{j}{\mu} + b^\varepsilon + \tilde{b}^\varepsilon$ is a nondegenerate $2\left(\frac{C^2 C_{\|\cdot\|}}{\mu^2 \tilde{\tau}(\varepsilon)} + \tau(\varepsilon) + \tilde{\tau}(\varepsilon)\right)$-self-concordant barrier function for $U_{ad}(\varepsilon)$ if $\tilde{\tau}(\varepsilon)$ satisfies*

$$\tilde{\tau}(\varepsilon) \geq \frac{1}{\varepsilon^2}\max\left\{\varepsilon^2, \frac{16}{9}C^2_{\partial,C(\overline{\Omega}_a)}\left(\sqrt{2C_{\|\cdot\|}} + \|g\|_Z\right)^2\right\}.$$

Here, C denotes the bound on the first derivative of $\hat{j}$ on $U_{ad}(\varepsilon)$.

Moreover, $f_{\varepsilon,\mu}$ is uniformly convex with modulus $\beta(\varepsilon,\mu) := \frac{\alpha}{\mu} + \frac{\tilde{\tau}(\varepsilon)}{C_{\|\cdot\|}}$, where $\alpha \geq 0$ denotes the convexity modulus of j, with $\alpha = 0$ if j is not uniformly convex.

Remark 3.5.19. Since we assume $j = \hat{j}$ to be self-concordant, $\frac{j}{\mu}$ may not be self-concordant for $\mu > 1$. Therefore, we use $C_\mu \leq 1$ in the preceding lemma.

Proof. Let $\mu \in (0, C_\mu]$. Employing Lemma 3.5.16, the self-concordance of $\frac{j}{\mu}$, and its boundedness on the bounded set $U_{\mathrm{ad}}(\varepsilon)$, cf. Corollary C.2.3, we obtain that $f_{\varepsilon,\mu} = \frac{j}{\mu} + b^\varepsilon + \tilde{b}^\varepsilon$ is a self-concordant barrier function for $U_{\mathrm{ad}}(\varepsilon)$. It is nondegenerate since it is uniformly convex, as follows together with the asserted modulus of convexity from Lemma C.4.14. Since $b^\varepsilon + \tilde{b}^\varepsilon$ is uniformly convex with modulus $\frac{\tilde{\tau}(\varepsilon)}{C_{\|\cdot\|}}$ and self-bounded with constant $\tau(\varepsilon) + \tilde{\tau}(\varepsilon)$, Lemma 2.5.24 establishes that $f_{\varepsilon,\mu}$ is self-bounded with constant as asserted. This lemma is applicable since Assumption 2.5.2 is satisfied, as we prove in the succeeding corollary. □

Corollary 3.5.20. *Under the Assumptions of Lemma 3.5.18, $f_{\varepsilon,\mu}$ satisfies Assumption 2.5.2 with $M := \overline{D}_{b^\varepsilon} \cap \overline{D}_{\tilde{b}^\varepsilon}$, $K := U_{ad}(\varepsilon)$, $\mu_s := C_\mu$, and $\vartheta_b := \tau(\varepsilon) + \tilde{\tau}(\varepsilon)$.*

Proof. It follows from Lemma C.4.2 that $K \subset M = \overline{K}$ is satisfied. The last part of Assumption 2.5.2 is valid since the barrier function f_{ε,μ_s} possesses a minimizer on $U_{\mathrm{ad}}(\varepsilon)$ due to Corollary C.4.6. The remaining parts of Assumption 2.5.2 follow from Lemma 3.5.18. □

Definition 3.5.21. For $\varepsilon > 0$ we set $\vartheta(\varepsilon) := \tau(\varepsilon) + \tilde{\tau}(\varepsilon)$ in case II.

3.5.3. Definitions for the barrier functions

Since $f_{\varepsilon,\mu}$ is thrice Fréchet differentiable and uniformly convex for every $(\varepsilon, \mu) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ on $U_{\mathrm{ad}}(\varepsilon)$, the Newton step n_u is well-defined at every $u \in U_{\mathrm{ad}}(\varepsilon)$, cf. Theorem C.4.15.

Definition 3.5.22. For $(\varepsilon, \mu) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ we define the *Newton decrement* of $f_{\varepsilon,\mu}$ at $u \in U_{\mathrm{ad}}(\varepsilon)$ via

$$\lambda_{\varepsilon,\mu}(u) := \sqrt{f''_{\varepsilon,\mu}(u)[n_u, n_u]},$$

where $n_u \in U$ denotes the Newton step for $f_{\varepsilon,\mu}$ at $u \in U_{\mathrm{ad}}(\varepsilon)$.

Definition 3.5.23. For $(\varepsilon, \mu) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and $t \geq 0$ we set

$$\Lambda_{\varepsilon,\mu}(t) := \{u \in U_{\mathrm{ad}}(\varepsilon) : \lambda_{\varepsilon,\mu}(u) \leq t\} \qquad \text{and} \qquad \Lambda_{\varepsilon,\mu} := \Lambda_{\varepsilon,\mu}\left(\frac{1}{4}\right).$$

3.5.4. Associated smoothed problems

So far we have shown how to obtain suitable barrier functions $f_{\varepsilon,\mu}$. It is left to explain how we actually tackle (P) by use of these barriers.

Definition 3.5.24. We call the problem

$$\min_{u \in U} j(u) \qquad \text{s.t.} \qquad u \in M(\varepsilon) \tag{P_ε}$$

the *smoothed problem with smoothing parameter* $\varepsilon > 0$. Here, we define $M(\varepsilon) := \overline{D}_{b^\varepsilon} \cap D_j$ in case I and $M(\varepsilon) := \overline{D}_{b^\varepsilon} \cap \overline{D}_{\tilde{b}^\varepsilon}$ in case II.

Remark 3.5.25. Since closures of convex sets are convex, cf. Lemma C.4.1, $M(\varepsilon)$ is convex.

Remark 3.5.26. In case I, (P_ε) is identical to (3.1) that we used to motivate our approach.

To tackle (P) we use two different strategies. On the one hand, we develop barrier methods for fixed ε that drive μ to zero. This is, for each μ we approximately solve $\min_{u\in U_{\mathrm{ad}}(\varepsilon)} f_{\varepsilon,\mu}(u)$ via (a damped version of) Newton's method and then decrease μ. This yields a solution of (P_ε), as we prove later. This solution can be regarded as an approximation for the solution of (P), in particular since we will provide an error bound. On the other hand, we develop algorithms in which μ and ε are both driven to zero. This is, for each (ε,μ) we approximately solve $\min_{u\in U_{\mathrm{ad}}(\varepsilon)} f_{\varepsilon,\mu}(u)$ via (a damped version of) Newton's method and then decrease ε and μ. This yields a solution of (P), as we establish later.

3.6. Estimates for an important constant

To ensure self-concordance of $f_{\varepsilon,\mu}$ we need to choose C_j and $\tilde{\tau}(\varepsilon)$, respectively, large enough, cf. Lemma 3.5.8 and Lemma 3.5.18. To do so, we need to find a priori estimates for $C_{\partial,C(\overline{\Omega}_a)}$, as the very same lemmas show. This constant depends only on the state equation $Ay + Bu = g$ and Ω_a. For $d \in \{2,3\}$ results from [Plu92] can be employed to derive bounds for $C_{\partial,C(\overline{\Omega}_a)}$. We use this to obtain estimates for one of the two state equations with $d = 2$ that we employ in our numerical experiments. The first estimate, however, deals with the case $d = 1$.

Lemma 3.6.1. *Let $\Omega := (a,b) \subset \mathbb{R}$, $Y := H^2(\Omega) \cap H_0^1(\Omega)$, $U := Z := L^2(\Omega)$, $A := -\Delta$, $B := -I$, and $g \equiv 0$. Then we have*

$$\|y(u)\|_{C(\overline{\Omega}_a)} \leq \frac{(b-a)^{\frac{3}{2}}}{2\sqrt{6}} \|\Delta y\|_{L^2(\Omega)} = \frac{(b-a)^{\frac{3}{2}}}{2\sqrt{6}} \|u\|_{L^2(\Omega)}$$

for all $u \in L^2(\Omega)$ and all $\Omega_a \subset \Omega$, i.e., it holds $C_{\partial,C(\overline{\Omega}_a)} \leq \frac{(b-a)^{\frac{3}{2}}}{2\sqrt{6}}$.

Proof. It suffices to argue for $\Omega_a = \Omega$. Let $a, b \in \mathbb{R}$ with $a < b$ be given. We first show how to reduce the general case of $\Omega = (a,b)$ to the case $\Omega = (0,1)$. To this end, set $\Omega := (a,b)$ and let $u \in L^2(\Omega)$ be given. Then, $y := y(u) \in Y$ satisfies $-y'' = u$ on Ω in the weak sense, and $y(a) = y(b) = 0$. Using the transformation $g : [0,1] \to [a,b]$, $g(t) := a + (b-a)t$ we obtain $\tilde{y} : [0,1] \to \mathbb{R}$, $\tilde{y}(t) := y(g(t))$. The chain rule for Sobolev functions, cf., e.g., [Alt06, 2.25, p. 124], yields $\tilde{y} \in \tilde{Y} := H^2(\tilde{\Omega}) \cap H_0^1(\tilde{\Omega})$, where we used $\tilde{\Omega} := (0,1)$. Defining $\tilde{u}(t) := (b-a)^2 u(g(t))$ we, thus, deduce that $\tilde{y} : [0,1] \to \mathbb{R}$ satisfies

$$-\tilde{y}''(t) = \tilde{u}(t) \text{ for all } t \in \tilde{\Omega} = (0,1), \quad \tilde{y}(0) = \tilde{y}(1) = 0.$$

Using integration by substitution we obtain $\|\tilde{u}\|_{L^2(\tilde{\Omega})} = (b-a)^{3/2}\|u\|_{L^2(\Omega)}$. This yields

$$\|y(u)\|_{C(\overline{\Omega}_a)} = \|y\|_{C(\overline{\Omega}_a)} = \|\tilde{y}\|_{C([0,1])} \leq C\|\tilde{u}\|_{L^2(0,1)} = C(b-a)^{3/2}\|u\|_{L^2(\Omega)},$$

where C denotes $C_{\partial,C(\overline{\Omega}_a)}$ in the special case $\Omega = \Omega_a = (0,1)$. To establish the assertion, it, hence, suffices to prove $\|y(u)\|_{C(\overline{\Omega}_a)} \leq \frac{1}{2\sqrt{6}}\|u\|_{L^2(\Omega)}$ for all $u \in L^2(\Omega)$ with $\Omega = \Omega_a = (0,1)$. To do so, assume without loss of generality $y := y(u) \not\equiv 0$ and let $x_0 \in (0,1)$ with $y(x_0) = \|y\|_{C(\overline{\Omega}_a)}$. Assume, furthermore, that $y(x_0) > 0$ is satisfied (if not, use $-y$ instead of y) and that $x_0 \in [0, \frac{1}{2}]$ holds (if not, use $[x_0, 1]$ instead of $[0, x_0]$). Since $y \in C^1(\overline{\Omega})$ holds due to Sobolev embeddings,

we infer $y'(x_0) = 0$. Using $y(0) = 0$, the fundamental theorem of calculus, integration by parts for Sobolev functions, cf. [Alt06, A6.8 $\langle 2 \rangle$, p. 267], and $y'(x_0) = 0$ we obtain

$$\begin{aligned}\|y\|_{C(\overline{\Omega}_a)} = y(x_0) = y(x_0) - y(0) = \int_0^{x_0} y'(t)\,\mathrm{d}t = [y'(t)t]_0^{x_0} - \int_0^{x_0} y''(t)t\,\mathrm{d}t \\ = -\int_0^{x_0} y''(t)t\,\mathrm{d}t \leq \|y''\|_{L^2(0,x_0)}\|t\|_{L^2(0,x_0)} \leq \|u\|_{L^2(\Omega)}\|t\|_{L^2(0,x_0)},\end{aligned}$$

where we also employed Hölder's inequality. Invoking $x_0 \leq \frac{1}{2}$ we infer $\|t\|_{L^2(0,x_0)} \leq \frac{1}{2\sqrt{6}}$. This implies the assertion. □

Remark 3.6.2. By use of $\Omega = \Omega_a = (a, b)$ and $y(x) = (x - a)(x - b)$ we obtain $\|y\|_{C(\overline{\Omega}_a)} = \frac{(b-a)^{3/2}}{8}\|\Delta y\|_{L^2(\Omega)}$. Since $|\frac{1}{8} - \frac{1}{2\sqrt{6}}| < 0.08$ holds, the bound of Lemma 3.6.1 may be considered sufficiently sharp for practical purposes. Of course, for $\Omega_a \subset \Omega$ with $\Omega_a \neq \Omega$ it may be possible to improve this bound.

Remark 3.6.3. [CGL10, Theorem 1] yields a slightly better estimate than the one we presented.

As state equation in one of our numerical examples we consider $-\Delta y + y = u$ with homogeneous Dirichlet boundary conditions on the unit square.

Lemma 3.6.4. *Let $\Omega := (0, 1) \times (0, 1) \subset \mathbb{R}^2$, $Y := H^2(\Omega) \cap H_0^1(\Omega)$, $U := Z := L^2(\Omega)$, $A := -\Delta + I$, $B := -I$, and $g \equiv 0$. Then we have $\|y(u)\|_{C(\overline{\Omega}_a)} \leq 1.14\|u\|_{L^2(\Omega)}$ for all $u \in L^2(\Omega)$ and all $\Omega_a \subset \Omega$, i.e., it holds $C_{\partial,C(\overline{\Omega}_a)} \leq 1.14$.*

Proof. We only need to argue for the case $\Omega_a = \Omega$. From [Plu92, Theorem 1, (b)], [Plu92, Theorem 4], and [KS84, p. 164, (3.2)] we infer $C_{\partial,C(\overline{\Omega}_a)} \leq 1.14$. To argue more precisely, we use the notation from [Plu92]. Then we have:

- In [Plu92, Theorem 1, (b)] we use $\mathrm{meas}(\Omega) = 1$, $M_\nu(Q, x_0) \leq \sqrt{2^\nu}$, and $\gamma_1 = \sqrt{2}$; this value for γ_1 is contained in the proof. This yields that the choice $\hat{C}_0 = 0$, $\hat{C}_1 = 2$, $\hat{C}_2 = 0.71$ is possible in [Plu92, Theorem 4], i.e., $\|y\|_{C(\overline{\Omega}_a)} \leq 2\|y_x\|_{2,\Omega} + 0.71\|y_{xx}\|_{2,\Omega}$ holds for $y \in Y$.
- In [Plu92, Theorem 4] we take $\hat{\Omega} = \Omega = (0, 1) \times (0, 1)$, $\underline{c} = \overline{c} = 1$, $\hat{C}_0 = 0$, $\hat{C}_1 = 2$, $\hat{C}_2 = 0.71$, and $\tau = 1$ to obtain $K \leq 1.14$; for μ_0 we employ the smallest eigenvalue μ of $-\Delta$ on Ω with respect to Dirichlet boundary conditions. It is well-known that $\mu = 2\pi^2$, see, e.g., [KS84, p. 164, (3.2)].

In conclusion, this yields $K \leq 1.14$, i.e. $\|y(u)\|_{C(\overline{\Omega}_a)} \leq 1.14\|u\|_{L^2(\Omega)}$ for all $u \in U$. □

In another numerical example we employ Poisson's equation on the unit disc.

Lemma 3.6.5. *Let $\Omega := B_1(0) \subset \mathbb{R}^2$, $Y := H^2(\Omega) \cap H_0^1(\Omega)$, $U := Z := L^2(\Omega)$, $A := -\Delta$, $B := -I$, and $g \equiv 0$. Then we have $C_{\partial,C(\overline{\Omega}_a)} = \frac{1}{2\sqrt{2\pi}} \leq 0.2$ for all $\Omega_a \subset \Omega$.*

Proof. As before we establish the assertion for $\Omega_a = \Omega$, which suffices. In this case the estimate is proven in [CGL10, Theorem 2]. □

4. The smoothed problems

This section is devoted to the examination of the problems (P_ε).

4.1. Properties of the smoothed minimum

In this section we study the smoothed minimum $\min_\varepsilon : C(\overline{\Omega}_a) \to \mathbb{R}$.

For constant functions the smoothed minimum yields the exact minimum.

Lemma 4.1.1. *Let $y \in C(\overline{\Omega}_a)$ and $c \in \mathbb{R}$. Then it holds for all $\varepsilon > 0$*

$$\min_\varepsilon(y + c) = \min_\varepsilon(y) + c.$$

In particular, we have $\min_\varepsilon(c) = c$ for all $\varepsilon > 0$.

Proof. For $\varepsilon > 0$ the definition of $\min_\varepsilon$ yields

$$\min_\varepsilon (y + c) = -\varepsilon \ln\left(\frac{\int_{\Omega_a} e^{-(y+c)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}\right) = -\varepsilon \ln\left(e^{-c/\varepsilon} \cdot \frac{\int_{\Omega_a} e^{-y/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}\right) = c + \min_\varepsilon(y). \quad \square$$

In the next lemma we show, in particular, that the smoothed minimum is monotone in the sense that for two functions $y, \tilde{y}$ with $y \le \tilde{y}$ there holds $\min_\varepsilon(y) \le \min_\varepsilon(\tilde{y})$.

Lemma 4.1.2. *Let $y, \tilde{y} \in C(\overline{\Omega}_a)$ with $y \le \tilde{y}$. Then it holds $\min_\varepsilon(y) \le \min_\varepsilon(\tilde{y})$ for all $\varepsilon > 0$. If, in addition, there is $x_0 \in \overline{\Omega}_a$ with $y(x_0) < \tilde{y}(x_0)$, then we have $\min_\varepsilon(y) < \min_\varepsilon(\tilde{y})$ for all $\varepsilon > 0$.*

Proof. We only establish the second assertion; the proof of the first assertion is similar, but simpler. Let $\varepsilon > 0$ and set

$$\Omega_< := \{x \in \Omega_a : y(x) < \tilde{y}(x)\} \qquad \text{and} \qquad \Omega_= := \{x \in \Omega_a : y(x) = \tilde{y}(x)\}.$$

Due to the existence of x_0 and the continuity of y and $\tilde{y}$ there holds $\mathrm{vol}(\Omega_<) > 0$. Therefore,

$$\int_{\Omega_<} \frac{e^{-y(x)/\varepsilon}}{\mathrm{vol}(\Omega_a)}\,\mathrm{d}x > \int_{\Omega_<} \frac{e^{-\tilde{y}(x)/\varepsilon}}{\mathrm{vol}(\Omega_a)}\,\mathrm{d}x \qquad \text{and} \qquad \int_{\Omega_=} \frac{e^{-y(x)/\varepsilon}}{\mathrm{vol}(\Omega_a)}\,\mathrm{d}x = \int_{\Omega_=} \frac{e^{-\tilde{y}(x)/\varepsilon}}{\mathrm{vol}(\Omega_a)}\,\mathrm{d}x$$

are satisfied due to the definition of $\Omega_<$ and $\Omega_=$. Using $\Omega_< \cap \Omega_= = \emptyset$ and $\Omega_a = \Omega_< \cup \Omega_=$ this implies

$$\begin{aligned}\min_\varepsilon(y) &= -\varepsilon \ln\left(\int_{\Omega_<} \frac{e^{-y(x)/\varepsilon}}{\operatorname{vol}(\Omega_a)}\,\mathrm{d}x + \int_{\Omega_=} \frac{e^{-y(x)/\varepsilon}}{\operatorname{vol}(\Omega_a)}\,\mathrm{d}x\right)\\ &< -\varepsilon \ln\left(\int_{\Omega_<} \frac{e^{-\tilde{y}(x)/\varepsilon}}{\operatorname{vol}(\Omega_a)}\,\mathrm{d}x + \int_{\Omega_=} \frac{e^{-\tilde{y}(x)/\varepsilon}}{\operatorname{vol}(\Omega_a)}\,\mathrm{d}x\right) = \min_\varepsilon(\tilde{y}).\end{aligned}$$ □

Corollary 4.1.3. *Let $y \in C(\overline{\Omega}_a)$. Then there holds for all $\varepsilon > 0$*

$$\min(y) \le \min_\varepsilon(y) \le \max(y).$$

Proof. We set $m := \min_{x\in\overline{\Omega}_a} y(x)$ and $M := \max_{x\in\overline{\Omega}_a} y(x)$. The assertion then follows from the application of $\min_\varepsilon$ to

$$m \le y \le M,$$

using the preceding lemma and Lemma 4.1.1. □

The following lemma justifies the name smoothed minimum.

Lemma 4.1.4. *Let $y \in C(\overline{\Omega}_a)$. Then we have*

$$\lim_{\varepsilon\to 0^+} \min_\varepsilon(y) = \min(y).$$

Proof. Set $m := \min(y)$. From Lemma 4.1.1 we know that it holds $\min_\varepsilon(y-m) = \min_\varepsilon(y) - m$ for all $\varepsilon > 0$. Defining $\tilde{y}(x) := y(x) - m$ for $x \in \overline{\Omega}_a$ it, thus, suffices to show $\lim_{\varepsilon\to 0^+} \min_\varepsilon(\tilde{y}) = 0$. Since Corollary 4.1.3 provides $0 = \min(\tilde{y}) \le \liminf_{\varepsilon\to 0^+} \min_\varepsilon(\tilde{y})$, we only need to establish that $\limsup_{\varepsilon\to 0^+} \min_\varepsilon(\tilde{y}) \le 0$ is valid. This is the aim for the remainder of the proof.

Let $\delta > 0$. Since there hold $\tilde{y} \in C(\overline{\Omega}_a)$ and $\{x \in \overline{\Omega}_a : \tilde{y}(x) = 0\} \neq \emptyset$, we know that

$$S := \{x \in \Omega_a : \tilde{y}(x) < \delta\}$$

is nonempty and open, hence satisfies $\operatorname{vol}(S) > 0$. For all $\varepsilon > 0$ we clearly have $e^{-\tilde{y}/\varepsilon} \ge 0$ on Ω_a and $e^{-\tilde{y}/\varepsilon} > e^{-\delta/\varepsilon}$ on S. We deduce that it holds $\int_{\Omega_a} e^{-\tilde{y}/\varepsilon}\,\mathrm{d}x \ge \int_S e^{-\delta/\varepsilon}\,\mathrm{d}x$ for all $\varepsilon > 0$. Hence, we have

$$\begin{aligned}\min_\varepsilon(\tilde{y}) &= -\varepsilon \ln\left(\frac{\int_{\Omega_a} e^{-\tilde{y}/\varepsilon}\,\mathrm{d}x}{\operatorname{vol}(\Omega_a)}\right) \le -\varepsilon \ln\left(\frac{\int_S e^{-\delta/\varepsilon}\,\mathrm{d}x}{\operatorname{vol}(\Omega_a)}\right)\\ &= -\varepsilon\left(-\frac{\delta}{\varepsilon} + \ln\left(\frac{\operatorname{vol}(S)}{\operatorname{vol}(\Omega_a)}\right)\right) = \delta + \varepsilon \ln\left(\frac{\operatorname{vol}(\Omega_a)}{\operatorname{vol}(S)}\right)\end{aligned}$$

for all $\varepsilon > 0$, from which we infer $\limsup_{\varepsilon\to 0^+} \min_\varepsilon(\tilde{y}) \le \delta$. Since $\delta > 0$ was arbitrary, it follows $\limsup_{\varepsilon\to 0^+} \min_\varepsilon(\tilde{y}) \le 0$, which concludes the proof. □

Remark 4.1.5. It is possible to broaden the domain of definition of the smoothed minimum. For example, upper semi-continuous functions are admissible. The integral is, in that case, well-defined since for an upper semi-continuous function y the mapping $x \mapsto e^{-y(x)/\varepsilon}$ is lower semi-continuous, hence integrable. Furthermore, the above lemma holds true for upper semi-continuous functions if "min" is replaced by "inf" (note that upper semi-continuous functions on compact sets do not necessarily attain their infimum). To establish this result one proceeds analogously to the proof given above and uses the fact that for each $\delta \in \mathbb{R}$ the set $\{x \in \Omega_a : y(x) < \delta\}$ is open, which is a consequence of the upper semi-continuity of y. However, in this thesis we only deal with functions that are at least continuous and, thus, choose to work with $C(\overline{\Omega}_a)$ as the domain of definition of the smoothed minimum.

Lemma 4.1.6. *The function* $\min_\varepsilon : C(\overline{\Omega}_a) \to \mathbb{R}$ *is concave for every* $\varepsilon > 0$.

Proof. Let $\varepsilon > 0$ and $y \in C(\overline{\Omega}_a)$. It suffices to show $\min''_\varepsilon(y)[h,h] \leq 0$ for all $h \in C(\overline{\Omega}_a)$. To this end, let $h \in C(\overline{\Omega}_a)$ be given. Then it holds

$$\min''_\varepsilon(y)[h,h] = \frac{1}{\varepsilon}\left(\frac{\left(\int_{\Omega_a} e^{-y/\varepsilon} h \,\mathrm{d}x\right)^2}{\left(\int_{\Omega_a} e^{-y/\varepsilon}\,\mathrm{d}x\right)^2} - \frac{\int_{\Omega_a} e^{-y/\varepsilon} h^2 \,\mathrm{d}x}{\int_{\Omega_a} e^{-y/\varepsilon}\,\mathrm{d}x}\right). \tag{4.1}$$

Moreover, Hölder's inequality implies

$$\left(\int_{\Omega_a} e^{-y/\varepsilon} h \,\mathrm{d}x\right)^2 = \left(\int_{\Omega_a} \sqrt{e^{-y/\varepsilon}}\,\sqrt{e^{-y/\varepsilon}}\, h \,\mathrm{d}x\right)^2 \leq \left\|e^{-y/\varepsilon}\right\|_{L^1(\Omega_a)} \left\|e^{-y/\varepsilon} h^2\right\|_{L^1(\Omega_a)}. \tag{4.2}$$

Inserting (4.2) into (4.1) we obtain the assertion. □

Corollary 4.1.7. *The functions* $B^\varepsilon_{C(\overline{\Omega}_a)} : Y \to \mathbb{R}$ *and* $B^\varepsilon : U \to \mathbb{R}$ *are concave for all* $\varepsilon > 0$.

Proof. The claim on $B^\varepsilon_{C(\overline{\Omega}_a)}$ follows from the previous lemma by the chain rule. To establish the assertion on B^ε let $u \in U$ and $\varepsilon > 0$. We have $(B^\varepsilon)''(u)[h_1,h_2] = (B^\varepsilon_{C(\overline{\Omega}_a)})''(y(u))[Th_1, Th_2]$ for all h_1, $h_2 \in U$, with T denoting the operator $T := -A^{-1}B \in \mathcal{L}(U,Y)$. The first assertion, thus, implies $(B^\varepsilon)''(u)[h,h] \leq 0$ for all $h \in U$. □

The smoothed minimum is monotone with respect to ε.

Lemma 4.1.8. *For every* $y \in C(\overline{\Omega}_a)$ *the mapping* $\varepsilon \mapsto \min_\varepsilon(y)$ *is monotonically increasing on* $\mathbb{R}_{>0}$. *Moreover,* $\varepsilon \mapsto \min_\varepsilon(y)$ *is strictly monotonically increasing if and only if* y *is not constant.*

Proof. We start by deriving an auxiliary estimate. To this end, let $f \in C(\overline{\Omega}_a)$. Using Hölder's inequality with $c > 1$ and $c/(c-1)$ we have

$$\int_{\Omega_a} f(x)\,\mathrm{d}x \leq \|f\|_{L^c(\Omega_a)} \|1\|_{L^{\frac{c}{c-1}}(\Omega_a)} = \|f\|_{L^c(\Omega_a)} \operatorname{vol}(\Omega_a)^{\frac{c-1}{c}}.$$

This yields for all $c \geq 1$ and all $f \in C(\overline{\Omega}_a)$ with $f \geq 0$ the inequality

$$\left(\frac{\int_{\Omega_a} f(x)\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right)^c \leq \frac{\int_{\Omega_a} (f(x))^c\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}. \tag{4.3}$$

We now establish the first assertion. Let $y \in C(\overline{\Omega}_a)$, $\varepsilon > 0$, and $c \geq 1$ be given. Employing (4.3) with $f(x) := e^{-y(x)/\varepsilon} \in C(\overline{\Omega}_a)$ we obtain the assertion via

$$\begin{aligned} \min\nolimits_{\frac{\varepsilon}{c}}(y) &= -\frac{\varepsilon}{c} \ln\left(\frac{\int_{\Omega_a} e^{-cy(x)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right) = -\frac{\varepsilon}{c} \ln\left(\frac{\int_{\Omega_a} \left(e^{-y(x)/\varepsilon}\right)^c\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right) \\ &\overset{(4.3)}{\leq} -\frac{\varepsilon}{c} \ln\left(\left(\frac{\int_{\Omega_a} e^{-y(x)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right)^c \right) = -\varepsilon \ln\left(\frac{\int_{\Omega_a} e^{-y(x)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right) = \min\nolimits_{\varepsilon}(y). \end{aligned}$$

We now demonstrate that the second assertion is valid. To this end, we note that (4.3) holds strictly if and only if f and the function 1 are linearly independent. This follows from the fact that Hölder's inequality becomes an equality if and only if the two integrands are linearly dependent, see, e.g., [Rud87, p. 65]. The linear dependence of f and 1 is, of course, equivalent to f being constant, which in turn is equivalent to y being constant. □

Corollary 4.1.9. *For every $y \in C(\overline{\Omega}_a)$ the mapping $\varepsilon \mapsto B^{\varepsilon}_{C(\overline{\Omega}_a)}(y)$ is monotonically increasing on $\mathbb{R}_{>0}$. Moreover, $\varepsilon \mapsto B^{\varepsilon}_{C(\overline{\Omega}_a)}(y)$ is strictly monotonically increasing if and only if $y - y_a$ is not constant.*

Proof. With the previous lemma this follows directly from $B^{\varepsilon}_{C(\overline{\Omega}_a)}(y) = \min_{\varepsilon}(y - y_a)$. □

Corollary 4.1.10. *There hold $U_{ad}(\varepsilon_1) \subset U_{ad}(\varepsilon_2)$ and $M(\varepsilon_1) \subset M(\varepsilon_2)$ for all $0 < \varepsilon_1 \leq \varepsilon_2$.*

Proof. Using Corollary 4.1.9 this follows from the definition of $U_{\mathrm{ad}}(\varepsilon)$ and $M(\varepsilon)$. □

The next result is an interesting consequence of Lemma 4.1.8.

Corollary 4.1.11. *For all $y \in C(\overline{\Omega}_a)$ with $\int_{\Omega_a} e^{y(x)}\,\mathrm{d}x \geq \mathrm{vol}(\Omega_a)$ it holds $\int_{\Omega_a} e^{y(x)} y(x)\,\mathrm{d}x \geq 0$.*

Proof. Let $y \in C(\overline{\Omega}_a)$ and apply Lemma 4.1.8 with $-y$ instead of y. The monotonicity of $\varepsilon \mapsto \min_{\varepsilon}(-y)$ is equivalent to $\frac{\partial \min_{\varepsilon}(-y)}{\partial \varepsilon} \geq 0$ for all $\varepsilon > 0$. We compute the derivative, which exists due to Lemma C.2.20, and obtain

$$\frac{1}{\varepsilon} \frac{\int_{\Omega_a} e^{y(x)/\varepsilon} y(x)\,\mathrm{d}x}{\int_{\Omega_a} e^{y(x)/\varepsilon}\,\mathrm{d}x} - \ln\left(\frac{\int_{\Omega_a} e^{y(x)/\varepsilon}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right) \geq 0.$$

For $\varepsilon = 1$ this yields the assertion:

$$\frac{\int_{\Omega_a} e^{y(x)} y(x)\,\mathrm{d}x}{\int_{\Omega_a} e^{y(x)}\,\mathrm{d}x} \geq \ln\left(\frac{\int_{\Omega_a} e^{y(x)}\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)} \right) \geq 0,$$

where we used the prerequisite $\int_{\Omega_a} e^{y(x)}\,\mathrm{d}x / \mathrm{vol}(\Omega_a) \geq 1$. □

4.2. Boundedness of the feasible sets

Definition 4.2.1. For every $\varepsilon > 0$ we define

$$Y_{\mathrm{ad}}(\varepsilon) := \{y(u) \in Y : u \in U_{\mathrm{ad}}(\varepsilon)\}.$$

Lemma 4.2.2. *The sets $Y_{ad}(\varepsilon)$, $U_{ad}(\varepsilon)$, and $M(\varepsilon)$ are uniformly bounded for all $\varepsilon > 0$.*

Proof. The uniform boundedness of $U_{\mathrm{ad}}(\varepsilon)$ and $M(\varepsilon)$ follows in case I from $U_{\mathrm{ad}}(\varepsilon), M(\varepsilon) \subset D_j$ and in case II from $U_{\mathrm{ad}}(\varepsilon), M(\varepsilon) \subset \overline{D}_{\tilde{b}^\varepsilon}$. The uniform boundedness of $Y_{\mathrm{ad}}(\varepsilon)$ is then implied by the boundedness of A^{-1} and B. □

Corollary 4.2.3. *The set $Y_{ad}(\varepsilon) \subset C^{0,\beta}(\overline{\Omega}_a) \subset C(\overline{\Omega}_a)$ is uniformly bounded for all $\varepsilon > 0$.*

Proof. By Assumption 3.1.9 we have $Y \hookrightarrow C^{0,\beta}(\overline{\Omega}_a)$. Of course, it holds $C^{0,\beta}(\overline{\Omega}_a) \hookrightarrow C(\overline{\Omega}_a)$. Along with the preceding lemma this implies the claim. □

4.3. Existence of optimal solutions

Lemma 4.3.1. *For every $\varepsilon > 0$ the smoothed problem (P_ε) possesses a global solution, and this solution is unique if j is strictly convex on $M(\varepsilon)$.*

Proof. Let $\varepsilon > 0$. Since $M(\varepsilon)$ is convex as intersection of convex sets, it is clear that a possible solution is unique if the objective j is strictly convex. We now prove existence of solutions. We first deal with case I. In this case we consider j on $L := \{u \in M(\varepsilon) : j(u) \leq j(u^\circ)\}$. We note that we have $u^\circ \in L$. Since $M(\varepsilon) = D_j \cap \overline{D}_{b^\varepsilon}$ and since j is a continuous barrier function for D_j, it is easy to see that L is closed in U. Also, $L \subset M(\varepsilon)$ is bounded due to Lemma 4.2.2. Moreover, L is convex. Since j is continuous and convex on L, Lemma C.4.5 yields the assertion. In case II, $M(\varepsilon) \ni u^\circ$ is nonempty, bounded, closed, and convex, and j is continuous and convex on U, so that Lemma C.4.5 implies the existence of a minimizer. □

The possible non-uniqueness of minimizers of (P_ε) in case II poses no problems, as is emphasized by the next definition.

Definition 4.3.2. For every $\varepsilon > 0$ we denote by $\bar{u}_\varepsilon \in M(\varepsilon)$ an arbitrary minimizer of (P_ε) and by $\bar{y}_\varepsilon \in Y$ the function $\bar{y}_\varepsilon := y(\bar{u}_\varepsilon)$. Moreover, we define the *path of optimal solutions* by $\mathbb{R}_{>0} \ni \varepsilon \mapsto \bar{u}_\varepsilon$.

Remark 4.3.3. To understand why we can work with an arbitrary minimizer $\bar{u}_\varepsilon$ we mention that if j is not uniformly convex, we will not prove results on $\bar{u}_\varepsilon$ but rather on $j(\bar{u}_\varepsilon)$. For example, we later provide an estimate for $|j(\bar{u}_\varepsilon) - j(\bar{u})|$ and deduce from this, under the additional assumption that j is uniformly convex, an estimate for $\|\bar{u}_\varepsilon - \bar{u}\|_U$.

4.4. The path of optimal solutions

In this section we examine the path of optimal solutions.

4.4.1. Maximum constraint violation

To examine the maximal pointwise infeasibility of functions $y \in Y_{\mathrm{ad}}(\varepsilon)$ with respect to the constraint $y \geq y_a$, we need the following auxiliary result.

Lemma 4.4.1. *Let $p > 0$ and define $\gamma_p : \mathbb{R}_{\geq 0} \to \mathbb{R}$, $\gamma_p(s) := \int_0^s e^{-t} t^{p-1} \,\mathrm{d}t$. Then there holds $\gamma_p(s) \geq \frac{e^{-s} s^p}{p}$ for all $s \geq 0$.*

Proof. Integration by parts yields $\gamma_p(s) = \frac{e^{-s} s^p}{p} + \int_0^s \frac{e^{-t} t^p}{p} \,\mathrm{d}t$ for all $s \geq 0$. □

Remark 4.4.2. We point out that $\lim_{s\to\infty} \gamma_p(s) = \Gamma(p)$, where $\Gamma : \mathbb{R}_{>0} \to \mathbb{R}$ denotes the well-known *gamma function* (whose domain of definition can be extended substantially by means of complex analysis). However, in this thesis we do not use this identity.

We recall that d denotes the dimension of Ω_a, i.e., $\Omega_a \subset \mathbb{R}^d$.

Lemma 4.4.3. *Let $S \subset C^{0,\beta}(\overline{\Omega}_a)$ be bounded. Then there exists $C > 0$ such that*

$$0 \leq \min_\varepsilon(y) - \min(y) \leq \varepsilon \left(C + \frac{d}{\beta} \left| \ln \varepsilon \right| \right)$$

holds for all $\varepsilon > 0$ and all $y \in S$.

Proof. Let $y \in S$. The inequality on the left-hand side follows from Corollary 4.1.3. We now establish the inequality on the right-hand side. For $\varepsilon \geq 1$ there holds $\min_\varepsilon(y) - \min(y) \leq \max(y) + \|y\|_{C(\overline{\Omega}_a)} \leq \varepsilon(C + \frac{d}{\beta} |\ln \varepsilon|)$ with $C := 2\|y\|_{C(\overline{\Omega}_a)}$. Here, we employed $\min_\varepsilon(y) \leq \max(y)$, see Corollary 4.1.3. Since S is bounded, this establishes the assertion for $\varepsilon \geq 1$.

Let $\varepsilon \in (0, 1]$. The function y possesses a minimizer $\tilde{x}^*$ on the compact set $\overline{\Omega}_a$. Since $\Omega_a = \cup_{i=1}^m \Omega_{a,i}$, we have $\tilde{x}^* \in \overline{\Omega}_{a,i^*}$ for an $i^* \in \{1, \dots, m\}$. Let $\eta > 0$. Then there exists $x^* \in \Omega_{a,i^*}$ with $y(x^*) \leq y(\tilde{x}^*) + \eta$. We now establish $\min_\varepsilon(y) - \min(y) \leq \varepsilon(C + \frac{d}{\beta} |\ln \varepsilon|) + \eta$ with a constant $C > 0$ that is independent of ε, y, $\tilde{x}^*$, x^*, i^*, and η. This implies the assertion. By assumption there exists a constant $\bar{C} > 0$, that is independent of all these quantities, with $\|y\|_{C^{0,\beta}(\overline{\Omega}_a)} \leq \bar{C}$. Setting $\omega := \min_\varepsilon(y)$ it holds $\int_{\Omega_a} e^{-y/\varepsilon} \,\mathrm{d}x = \mathrm{vol}(\Omega_a) e^{-\omega/\varepsilon}$. We have $-(y(x) - y(x^*)) \geq -\bar{C} \|x - x^*\|^\beta$ for all $x \in \overline{\Omega}_a$. This implies

$$\begin{aligned} \mathrm{vol}(\Omega_a) e^{-\omega/\varepsilon} = \int_{\Omega_a} e^{-y(x)/\varepsilon} \,\mathrm{d}x &\geq e^{-y(x^*)/\varepsilon} \int_{\Omega_a} e^{-\bar{C}\|x-x^*\|^\beta/\varepsilon} \,\mathrm{d}x \\ &\geq e^{-y(x^*)/\varepsilon} \int_{\Omega_{a,i^*}} e^{-\bar{C}\|x-x^*\|^\beta/\varepsilon} \,\mathrm{d}x. \end{aligned} \tag{4.4}$$

Due to the cone condition there are constants $c_i > 0$ and $\delta_i > 0$ such that $\int_{\Omega_{a,i}} e^{-\bar{C}\|x-z_i\|^\beta/\varepsilon} \,\mathrm{d}x \geq c_i \int_{B_{\delta_i}(z_i)} e^{-\bar{C}\|x-z_i\|^\beta/\varepsilon} \,\mathrm{d}x$ is satisfied for all $\varepsilon > 0$ and all $z_i \in \Omega_{a,i}$, $i = 1, \dots, m$. For a rigorous

proof that such constants exist we refer to Lemma E.0.6. We set $c := \min_{i\in\{1,\dots,m\}} c_i$ and $\delta := \min_{i\in\{1,\dots,m\}} \delta_i$, and note that c and δ are independent of ε, y, $\tilde{x}^*$, x^*, i^*, and η. Multiplying (4.4) with $\frac{e^{y(x^*)/\varepsilon}}{\operatorname{vol}(\Omega_a)}$ we deduce

$$e^{(y(x^*)-\omega)/\varepsilon} \geq \frac{c}{\operatorname{vol}(\Omega_a)} \int_{B_\delta(x^*)} e^{-\bar{C}\|x-x^*\|^\beta/\varepsilon}\,\mathrm{d}x \geq \frac{c\tilde{c}}{\operatorname{vol}(\Omega_a)} \int_0^\delta e^{-\bar{C}r^\beta/\varepsilon} r^{d-1}\,\mathrm{d}r$$

with $\tilde{c} := d\operatorname{vol}(B_1(0))$, $B_1(0) \subset \mathbb{R}^d$. Here, we used integration of rotational symmetric functions, cf., e.g., [Kön04b, Satz, p. 311]. Substituting $t := \frac{\bar{C}}{\varepsilon} r^\beta$ we infer

$$e^{(y(x^*)-\omega)/\varepsilon} \geq \frac{c\tilde{c}}{\beta\operatorname{vol}(\Omega_a)} \left(\frac{\varepsilon}{\bar{C}}\right)^{\frac{d}{\beta}} \int_0^{\frac{\bar{C}}{\varepsilon}\delta^\beta} e^{-t} t^{\frac{d-\beta}{\beta}}\,\mathrm{d}t \geq \frac{c\tilde{c}}{\beta\operatorname{vol}(\Omega_a)} \left(\frac{\varepsilon}{\bar{C}}\right)^{\frac{d}{\beta}} \int_0^{\hat{c}} e^{-t} t^{\frac{d-\beta}{\beta}}\,\mathrm{d}t$$

with $\hat{c} := \bar{C}\delta^\beta \leq \frac{\bar{C}}{\varepsilon}\delta^\beta$ due to $\varepsilon \leq 1$. Apparently, $\hat{c}$ is independent of ε, y, $\tilde{x}^*$, x^*, i^*, and η. We remark that the substitution $t := \frac{\bar{C}}{\varepsilon} r^\beta$ we used may not be continuously differentiable in $[0,\delta]$ due to $\beta \in (0,1]$. However, if we start with the integral we derived by substitution and substitute $r := (\frac{\varepsilon}{\bar{C}} t)^{1/\beta}$, which is continuously differentiable, we obtain that the substitution we performed is, in fact, correct.

From Lemma 4.4.1 with $p := \frac{d-\beta}{\beta} + 1 = \frac{d}{\beta} > 0$ we conclude

$$e^{(y(x^*)-\omega)/\varepsilon} \geq \frac{c\tilde{c}}{\beta\operatorname{vol}(\Omega_a)} \left(\frac{\varepsilon}{\bar{C}}\right)^{\frac{d}{\beta}} \gamma_p(\hat{c}) \geq \frac{c\tilde{c}}{\beta\operatorname{vol}(\Omega_a)} \left(\frac{\varepsilon}{\bar{C}}\right)^{\frac{d}{\beta}} \frac{\beta}{d} e^{-\hat{c}} \hat{c}^{\frac{d}{\beta}}.$$

Hence, for

$$\bar{c} := \min\left\{ \frac{c\tilde{c}}{d\operatorname{vol}(\Omega_a)} \left(\frac{\hat{c}}{\bar{C}}\right)^{\frac{d}{\beta}} e^{-\hat{c}},\ \frac{1}{2} \right\}$$

we obtain $e^{(y(x^*)-\omega)/\varepsilon} \geq \bar{c}\varepsilon^{\frac{d}{\beta}}$, where $\bar{c} \leq \frac{1}{2}$ is independent of ε, y, $\tilde{x}^*$, x^*, i^*, and η. From this we infer with the definition of x^*:

$$y(\tilde{x}^*) \geq y(x^*) - \eta \geq \omega + \varepsilon \ln\bigl(\bar{c}\varepsilon^{\frac{d}{\beta}}\bigr) - \eta = \omega - \eta + \varepsilon\left(\ln\bar{c} + \frac{d}{\beta}\ln\varepsilon\right).$$

We set $C := -\ln\bar{c} > 0$ and note that this constant is independent of ε, y, $\tilde{x}^*$, x^*, i^*, and η. Thus, we finish the proof with the conclusion

$$\min_\varepsilon(y) - \min(y) = \omega - y(\tilde{x}^*) \leq \varepsilon\left(C + \frac{d}{\beta}\left|\ln\varepsilon\right|\right) + \eta. \qquad \square$$

Corollary 4.4.4. *There exists $C > 0$ such that for every $\varepsilon > 0$ it holds*

$$\left\|(y(u) - y_a)^-\right\|_{C(\overline{\Omega}_a)} \leq \varepsilon\left(C + \frac{d}{\beta}\left|\ln\varepsilon\right|\right)$$

for all $u \in M(\varepsilon)$. In particular, we have for all $\varepsilon > 0$ the estimate

$$\left\|(\bar{y}_\varepsilon - y_a)^-\right\|_{C(\overline{\Omega}_a)} \leq \varepsilon\left(C + \frac{d}{\beta}\left|\ln\varepsilon\right|\right). \tag{4.5}$$

Proof. By use of $u := \bar{u}_\varepsilon \in M(\varepsilon)$ we infer that it suffices to prove the first assertion. Since the first assertion is trivial for $\min(y(u) - y_a) \geq 0$, we may assume $\min(y(u) - y_a) < 0$ in the following. Lemma 4.2.2 shows that $M := \cup_{\varepsilon>0} M(\varepsilon)$ is bounded. This implies the boundedness of $\{y(u) \in Y : u \in M\} \hookrightarrow C^{0,\beta}(\overline{\Omega}_a)$. Hence, Lemma 4.4.3 yields the existence of $C > 0$ such that for all $u \in M$ and all $\varepsilon > 0$ it holds

$$\left\|(y(u) - y_a)^-\right\|_{C(\overline{\Omega}_a)} = -\min(y(u) - y_a) \leq \varepsilon \left(C + \frac{d}{\beta} \left|\ln \varepsilon\right|\right) - B^\varepsilon(u).$$

For every $\varepsilon > 0$ there holds $M(\varepsilon) \subset \overline{D}_{b^\varepsilon}$, hence $B^\varepsilon(u) \geq 0$ for all $u \in M(\varepsilon)$. □

We show that the estimate of the preceding lemma is sharp with respect to the order of ε if ε is sufficiently small. Note that for sufficiently small ε the right-hand side in this lemma has order $\mathcal{O}(\varepsilon|\ln \varepsilon|)$.

Lemma 4.4.5. *For $\Omega_a = (0, 2) \subset \mathbb{R}$ there exists a bounded family $(y_\varepsilon)_{\varepsilon\in(0,1)} \subset C^{0,1}(\overline{\Omega}_a)$ such that every y_ε satisfies*

$$\min_\varepsilon(y_\varepsilon) \geq 0 \qquad \textit{and} \qquad \left\|y_\varepsilon^-\right\|_{C(\overline{\Omega}_a)} = \varepsilon \left|\ln \varepsilon\right|.$$

Moreover, this is also true if all y_ε are required to satisfy either homogeneous Dirichlet or homogeneous Neumann boundary conditions.

Proof. To establish the assertions it suffices to argue for the case of homogeneous Dirichlet and homogeneous Neumann boundary conditions. In the case of homogeneous Dirichlet boundary conditions consider for $\varepsilon \in (0, 1)$

$$y_\varepsilon(x) := \begin{cases} -x : & \text{if } x \in [0, -\varepsilon \ln \varepsilon], \\ 2\varepsilon \ln \varepsilon + x : & \text{if } x \in [-\varepsilon \ln \varepsilon, 1 - \varepsilon \ln \varepsilon], \\ 2 - x : & \text{if } x \in [1 - \varepsilon \ln \varepsilon, 2]. \end{cases}$$

Note that the last interval in the definition of y_ε is well-defined since $-\varepsilon \ln \varepsilon < 1$ for all $\varepsilon \in (0, 1)$. Obviously, y_ε is Lipschitz with constant 1 and uniformly bounded regardless of ε. This shows the boundedness of $(y_\varepsilon) \subset C^{0,1}(\overline{\Omega}_a)$. Furthermore, there hold $y_\varepsilon(0) = y_\varepsilon(2) = 0$ and $\|y_\varepsilon^-\|_{C(\overline{\Omega}_a)} = \varepsilon|\ln \varepsilon|$. A computation yields $\min_\varepsilon(y_\varepsilon) = -\varepsilon \ln(1 - e^{-1/\varepsilon})$, which is positive.

In the case of homogeneous Neumann boundary conditions consider for $\varepsilon \in (0, 1)$

$$y_\varepsilon(x) := \begin{cases} \varepsilon \ln \varepsilon : & \text{if } x \in [0, \varepsilon], \\ x + \varepsilon(\ln \varepsilon - 1) : & \text{if } x \in [\varepsilon, 2 - \varepsilon], \\ 2 + \varepsilon(\ln \varepsilon - 2) : & \text{if } x \in [2 - \varepsilon, 2]. \end{cases}$$

Obviously, y_ε is Lipschitz with constant 1 and uniformly bounded regardless of ε. This shows the boundedness of $(y_\varepsilon) \subset C^{0,1}(\overline{\Omega}_a)$. Furthermore, it holds $y_\varepsilon'(0) = y_\varepsilon'(2) = 0$, and $\|y_\varepsilon^-\|_{C(\overline{\Omega}_a)} = \varepsilon|\ln \varepsilon|$. A computation yields $\min_\varepsilon(y_\varepsilon) = 0$. This concludes the proof. □

Remark 4.4.6. Since the order of Lemma 4.4.3 is sharp with respect to ε for $d = 1$, it can be expected to be sharp for $d \geq 2$ as well. This is, in particular, true since the key argument in the proof of this lemma is to employ rotational symmetry, which shows that we basically use an argument for $d = 1$, anyway.

Remark 4.4.7. The estimate in Lemma 4.4.3 may be improved in specific situations. For instance, if every $y \in S$ is constant, then the right-hand side is bounded by zero, cf. Lemma 4.1.1. We mention this since it indicates that in specific situations it may be possible to improve the estimate for $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$ from Corollary 4.4.4. This would be desirable since it turns out that the quantity $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$ determines the length of the path of solutions. However, we remark that in one of the numerical examples for $d = 2$ we observe the order $\mathcal{O}(\varepsilon|\ln \varepsilon|)$ for $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$ to be sharp, which confirms the estimate in Corollary 4.4.4 in the general case.

4.4.2. Length

From the previous result we derive an estimate for the length of the path of solutions. Recall that in case I, $j = -C_j \ln(C_{\hat{j}} - \hat{j})$ depends on the choice of $C_j > 0$. Note, however, that the minimizer $\bar{u}_\varepsilon$ of j on $M(\varepsilon)$ does *not* depend on this choice. We also recall that we have $j = C_j \hat{j}$ with $C_j = 1$ in case II.

Theorem 4.4.8. *There exists $C > 0$ such that*

$$|j(\bar{u}_\varepsilon) - j(\bar{u})| \leq C_j \|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \, \varepsilon \left(C + \frac{d}{\beta} \, |\ln \varepsilon| \right)$$

is satisfied for all $\varepsilon > 0$ and, in case I, for all $C_j > 0$. Here, $\bar{\lambda}$ denotes a Lagrange multiplier for $\bar{u}$ in the case $C_j = 1$, cf. Lemma 3.4.3. Moreover, if j/C_j is uniformly convex on $M(\varepsilon)$ with modulus $\alpha > 0$, then it also holds

$$\|\bar{u}_\varepsilon - \bar{u}\|_U \leq \sqrt{\frac{4\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}}{\alpha}} \, \sqrt{\varepsilon} \, \sqrt{C + \frac{d}{\beta} \, |\ln \varepsilon|},$$

with the same C and the same $\bar{\lambda}$ as before.

Proof. We first deal with the estimate for the objective. Dividing this estimate by C_j we see that it suffices to argue for the case $C_j = 1$. To this end, define $T := -A^{-1}B \in \mathcal{L}(U, C(\overline{\Omega}_a))$. We have $\overline{D}_{b^\varepsilon} = \{u \in U : B^\varepsilon(u) \geq 0\}$ due to Lemma C.4.2 and $u^\circ \in D_{b^\varepsilon}$. From $B^\varepsilon(\bar{u}) \geq \min(y(\bar{u}) - y_a) \geq 0$, cf. Lemma 4.4.3, we infer $\bar{u} \in \overline{D}_{b^\varepsilon}$. This implies $\bar{u} \in M(\varepsilon)$. Together with Lemma 3.4.3 this yields

$$\begin{aligned} |j(\bar{u}_\varepsilon) - j(\bar{u})| = j(\bar{u}) - j(\bar{u}_\varepsilon) \leq \; & -j'(\bar{u})[\bar{u}_\varepsilon - \bar{u}] = \langle \overline{\lambda}, T(\bar{u}_\varepsilon - \bar{u}) \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)} \\ = \; & \langle \overline{\lambda}, (T(\bar{u}_\varepsilon) + A^{-1}g) - (T(\bar{u}) + A^{-1}g) \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)} \\ = \; & \langle \overline{\lambda}, \bar{y}_\varepsilon - \bar{y} \rangle_{C(\overline{\Omega}_a)^*, C(\overline{\Omega}_a)}. \end{aligned}$$

By Lemma 3.4.3 we have $\langle \bar{\lambda}, y\rangle_{C(\overline{\Omega}_a)^*,C(\overline{\Omega}_a)} \leq 0$ for all $y \in C(\overline{\Omega}_a)$ with $y \geq 0$. From this we infer $|j(\bar{u}_\varepsilon) - j(\bar{u})| \leq \langle \bar{\lambda}, (\bar{y}_\varepsilon - y_a)^-\rangle_{C(\overline{\Omega}_a)^*,C(\overline{\Omega}_a)} \leq \|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$. Using Corollary 4.4.4 we obtain the assertion.

We now turn to the second estimate. Since $\bar{u}_\varepsilon$ is also the global minimizer of the uniformly convex function j with $C_j = 1$ on $M(\varepsilon)$, it follows from Lemma C.4.12 that we have $\|\bar{u}-\bar{u}_\varepsilon\|_U \leq \sqrt{\frac{4}{\alpha}}\sqrt{j(\bar{u}) - j(\bar{u}_\varepsilon)}$ with $C_j = 1$. Hence, the second inequality follows from the first one. □

Remark 4.4.9. *Regardless of the dimension d of Ω we have order $\mathcal{O}(\sqrt{\varepsilon(1 + |\ln \varepsilon|)})$ in the above estimate for the path of optimal solutions. For interior point methods and Lavrentiev regularization similar estimates have been proven, cf., e.g., [Sch09b, Theorem 6.3] and [KR09, Theorem 3.4]. In Moreau-Yosida regularization for state constrained optimal control problems the length of the path of optimal solutions can also be bounded by some power of the regularization parameter, but this power depends on d, cf. [HSW12, Theorem 2.9] and [SH11, Theorem 2.9].*

We estimate $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ in a simple situation.

Lemma 4.4.10. *We consider a state equation with $Z = U$, $B = -I$, $g = 0$, i.e., a state equation of the form $Ay = u$, with $A \in \mathcal{L}(Y, U)$. Let Y satisfy $Y \hookrightarrow L^2(\Omega)$ as well as $\mathbf{1} \in Y$, where $\mathbf{1}$ denotes the constant function with value 1. Furthermore, let the reduced objective be given by $\hat{j}(u) := \frac{1}{2}\|y(u) - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_U$ with $y_d \in L^2(\Omega)$ and $\hat{\alpha} > 0$. Then for $\bar{\lambda}$ from the previous theorem it holds*

$$\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq \sqrt{2\hat{j}(\bar{u})}\left(\sqrt{\mathrm{vol}(\Omega)} + \|A\mathbf{1}\|_U\right).$$

In particular, for the state equation $-\Delta y + y = u$ on Ω and $\frac{\partial y}{\partial \nu} = 0$ on $\partial\Omega$, this estimate reads $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq 2\sqrt{2\hat{j}(\bar{u})\,\mathrm{vol}(\Omega)}$.

Proof. The optimality condition $\bar{\lambda} \leq 0$ implies $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} = -\langle \bar{\lambda}, \mathbf{1}\rangle_{C(\overline{\Omega}_a)^*,C(\overline{\Omega}_a)}$. Moreover, we have $j'(\bar{u})[h] = \langle \bar{\lambda}, A^{-1}h\rangle_{C(\overline{\Omega}_a)^*,C(\overline{\Omega}_a)}$ for all $h \in U$. There hold $|j'(\bar{u})[h]| \leq |\hat{j}'(\bar{u})[h]|$ and $\hat{j}'(\bar{u})[h] = (y(\bar{u}) - y_d, A^{-1}h)_{L^2(\Omega)} + \hat{\alpha}(\bar{u}, h)_{L^2(\Omega)}$ for all $h \in U$. To derive the first estimate we used $C_j = 1$ and, in case I, the optimality of $\bar{u}$ and $C_{\hat{j}} \geq 1$. Apparently, we have $|\hat{j}'(\bar{u})[A\mathbf{1}]| \leq \|\bar{y}-y_d\|_{L^2(\Omega)}\|\mathbf{1}\|_{L^2(\Omega)}+\hat{\alpha}\|\bar{u}\|_{L^2(\Omega)}\|A\mathbf{1}\|_{L^2(\Omega)}$. Together with $\max\{\|\bar{y}-y_d\|_{L^2(\Omega)}, \hat{\alpha}\|\bar{u}\|_{L^2(\Omega)}\} \leq \sqrt{2\hat{j}(\bar{u})}$ this implies

$$\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq |j'(\bar{u})[A\mathbf{1}]| \leq \left|\hat{j}'(\bar{u})[A\mathbf{1}]\right| \leq \sqrt{2\hat{j}(\bar{u})}\left(\|\mathbf{1}\|_{L^2(\Omega)} + \|A\mathbf{1}\|_{L^2(\Omega)}\right). \qquad \square$$

5. Barrier methods for fixed smoothing parameter

In this section we develop barrier methods to solve (P_ε) for a given ε.

5.1. An estimate for the overall error

Lemma 5.1.1. *Let $\varepsilon > 0$ and $\tau(\varepsilon) \geq 1$. In case I let $C_\mu > 0$ and let C_j be chosen according to Lemma 3.5.8. In case II let $C_\mu \leq 1$ and let $\tilde{\tau}(\varepsilon)$ be chosen according to Lemma 3.5.18. Then for every $\mu \in I_s = (0, C_\mu]$ we have*

$$|j(u) - j(\bar{u}_\varepsilon)| \leq 2\mu\vartheta(\varepsilon)$$

for all $u \in \Lambda_{\varepsilon,\mu}$. If, in addition, j is uniformly convex on $M(\varepsilon)$ with modulus $\alpha > 0$, then it also holds

$$\|u - \bar{u}_\varepsilon\|_U \leq \sqrt{\frac{4}{\alpha}}\sqrt{2\mu\vartheta(\varepsilon)}$$

for these μ and u.

Proof. Considering Corollary 3.5.12 in case I and Corollary 3.5.20 in case II, it follows from Lemma 2.5.18, Lemma 2.5.19, and $\lambda_{\varepsilon,\mu}(u) \leq \frac{1}{4}$ that

$$\begin{aligned}|j(u) - j(\bar{u}_\varepsilon)| &\leq |j(u) - j(\bar{u}_{\varepsilon,\mu})| + |j(\bar{u}_{\varepsilon,\mu}) - j(\bar{u}_\varepsilon)| \\ &\leq \frac{\lambda_{\varepsilon,\mu}(u)}{1 - \frac{16}{9}\lambda_{\varepsilon,\mu}(u)} \cdot \frac{\sqrt{\vartheta(\varepsilon)} + (\lambda_{\varepsilon,\mu}(u))^2}{1 - \lambda_{\varepsilon,\mu}(u)} \cdot \mu + \vartheta(\varepsilon)\mu \leq \left(\frac{3}{5}\sqrt{\vartheta(\varepsilon)} + \frac{3}{80}\right)\mu + \vartheta(\varepsilon)\mu\end{aligned}$$

is satisfied, where $\bar{u}_{\varepsilon,\mu}$ denotes the minimizer of $f_{\varepsilon,\mu}$ on $U_{\mathrm{ad}}(\varepsilon)$. Since we have $\vartheta(\varepsilon) \geq 1$, it holds

$$\frac{3}{5}\sqrt{\vartheta(\varepsilon)} + \frac{3}{80} \leq \sqrt{\vartheta(\varepsilon)} \leq \vartheta(\varepsilon),$$

which yields the first estimate. The second estimate is implied by the first, cf. Lemma C.4.12. □

Remark 5.1.2. The proof shows that the order $\mu\vartheta(\varepsilon)$ in the above estimate does not improve if $u \in \Lambda_{\varepsilon,\mu}(\theta)$ with $\theta < \frac{1}{4}$ is required instead of $u \in \Lambda_{\varepsilon,\mu} = \Lambda_{\varepsilon,\mu}(\frac{1}{4})$.

We present an estimate for the overall error. Clearly, choosing a smaller ε should enable us to decrease this error. However, since a smaller ε may require the choice of a larger C_j in case I, cf. Lemma 3.5.8, the following result also displays the influence of C_j. We recall that in case II we have $C_j = 1$.

Lemma 5.1.3. *Let $\varepsilon > 0$ and $\tau(\varepsilon) \geq 1$. In case I let $C_\mu > 0$ and let C_j be chosen according to Lemma 3.5.8. In case II let $C_\mu \leq 1$ and let $\tilde{\tau}(\varepsilon)$ be chosen according to Lemma 3.5.18. Then for every $\mu \in I_s = (0, C_\mu]$ we have*

$$\frac{|j(u) - j(\bar{u})|}{C_j} \leq \frac{2\mu\vartheta(\varepsilon)}{C_j} + C\varepsilon\,(1 + |\ln\varepsilon|)$$

for all $u \in \Lambda_{\varepsilon,\mu}$, where $C > 0$ is independent of ε, $\tau(\varepsilon)$, $\tilde{\tau}(\varepsilon)$, C_μ, C_j, μ, and u.

If, in addition, j/C_j is uniformly convex on $M(\varepsilon)$ with modulus $\alpha > 0$, then it also holds

$$\|u - \bar{u}\|_U \leq \sqrt{\frac{8\mu\vartheta(\varepsilon)}{C_j\alpha}} + \sqrt{\frac{4C}{\alpha}}\sqrt{\varepsilon\,(1 + |\ln\varepsilon|)}.$$

for these μ, u, and the same C.

Proof. Both estimates follow from Theorem 4.4.8 and Lemma 5.1.1 by use of the triangle inequality. □

Remark 5.1.4. Theorem 4.4.8 provides more information for the right summand in the above estimates.

5.2. A short step method

We consider the following short step method to solve (P_ε).

Algorithm SSM_ε (short step method to solve (P_ε))

Input: Parameters $\varepsilon > 0$, $\mu_0 > 0$ in case I and $\mu_0 \leq 1$ in case II, $\theta \in (0, \frac{1}{4}]$, $\tau(\varepsilon) \geq 1$, starting point $u^0 \in \Lambda_{\varepsilon,\mu_0}(\theta)$.

Set $C_\mu := \mu_0$ and select C_j according to Lemma 3.5.8 in case I and $\tilde{\tau}(\varepsilon)$ according to Lemma 3.5.18 in case II.

Set $\vartheta_b := \vartheta(\varepsilon)$ and define $\delta := \frac{\theta\left(1 - \frac{\theta}{(1-\theta)^2}\right)}{1 + \frac{\theta}{\sqrt{\vartheta_b}}}$ and $\beta := 1 - \frac{\delta}{\sqrt{\vartheta_b}}$. Set $f_\mu := f_{\varepsilon,\mu}$.

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k \in U$ by solving $f''_{\mu_k}(u^k)[s^k] = -f'_{\mu_k}(u^k)$ in U^*.

Set $u^{k+1} := u^k + s^k$ and $\mu_{k+1} := \beta\mu_k$.

END

Remark 5.2.1. The freedom to choose large values for $\tau(\varepsilon)$ may be useful to find a starting point for SSM_ε.

Remark 5.2.2. Termination criteria for SSM_ε can be based on the next theorem.

We have the following theorem on convergence and complexity of Algorithm SSM_ε. It is one of the main results of this thesis.

Theorem 5.2.3. *Algorithm SSM_ε generates a sequence $(u^k) \subset U_{ad}(\varepsilon)$ with $u^k \in \Lambda_{\varepsilon,\mu_k}(\theta)$ for all $k \in \mathbb{N}_0$. Moreover, with $\vartheta_b := \vartheta(\varepsilon)$ we have for every $k \in \mathbb{N}_0$:*

1) To reach iteration k (more precisely: to reach the FOR statement in SSM_ε for the $k+1$-th time) Algorithm SSM_ε requires exactly k Newton steps.

2) The sequence $(j(u^k))$ converges with r-linear rate β to the optimal value of (P_ε). More precisely, there holds

$$\frac{|j(u^k) - j(\bar{u}_\varepsilon)|}{C_j} \leq \frac{\vartheta_b + \sqrt{\vartheta_b}}{C_j}\mu_k = \frac{\vartheta_b + \sqrt{\vartheta_b}}{C_j}\beta^k\mu_0.$$

3) For every $\hat{\varepsilon} > 0$ we have the complexity estimate

$$k \geq \frac{\sqrt{\vartheta_b}}{\delta}\ln\left(\frac{\frac{\mu_0}{C_j}(\vartheta_b + \sqrt{\vartheta_b})}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad \frac{|j(u^k) - j(\bar{u}_\varepsilon)|}{C_j} \leq \hat{\varepsilon}.$$

4) If j/C_j is uniformly convex on $M(\varepsilon)$ with modulus $\alpha > 0$, then it holds

$$\|u^k - \bar{u}_\varepsilon\|_U \leq \sqrt{\frac{4}{C_j\alpha}}\sqrt{\vartheta_b + \sqrt{\vartheta_b}}\sqrt{\mu_k}.$$

In particular, (u^k) converges r-linearly with rate $\sqrt{\beta}$ and $\|\cdot\|_U$-strongly to $\bar{u}_\varepsilon$, and we have for every $\hat{\varepsilon} > 0$ the complexity estimate

$$k \geq \frac{2\sqrt{\vartheta_b}}{\delta}\ln\left(\frac{\sqrt{\frac{4}{C_j\alpha}}\sqrt{\mu_0(\vartheta_b + \sqrt{\vartheta_b})}}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad \|u^k - \bar{u}_\varepsilon\|_U \leq \hat{\varepsilon}.$$

5) There exists a constant $C > 0$ that is independent of ε, $\tau(\varepsilon)$, $\tilde{\tau}(\varepsilon)$, C_μ, C_j, and k such that

$$\frac{|j(u^k) - j(\bar{u})|}{C_j} \leq \frac{2\mu_k\vartheta_b}{C_j} + C\varepsilon\,(1 + |\ln\varepsilon|)$$

is satisfied. If j/C_j is uniformly convex on $M(\varepsilon)$ with modulus $\alpha > 0$, then it also holds

$$\|u^k - \bar{u}\|_U \leq \sqrt{\frac{8\mu_k\vartheta_b}{C_j\alpha}} + \sqrt{\frac{4C}{\alpha}}\sqrt{\varepsilon\,(1 + |\ln\varepsilon|)},$$

with the same C as before.

Proof. 1), 2), 3), and 4) are derived from their counterparts in Theorem 2.6.3, which is applicable due to Corollary 3.5.12 in case I and Corollary 3.5.20 in case II. The estimates in 5) follow from Lemma 5.1.3. □

Remark 5.2.4. All constants that appear in 1), 2), 3), and 4) are known explicitly. Thus, we can determine in advance how many iterations of SSM_ε suffice to ensure a given accuracy for $|j(u^k) - j(\bar{u}_\varepsilon)|$ and, if j is uniformly convex, $\|u^k - \bar{u}_\varepsilon\|_U$.

Remark 5.2.5. In case I, j/C_j is uniformly convex on $D_j \supset M(\varepsilon)$, as follows from the uniform convexity of $\hat{j}$ via Lemma C.4.14. In this lemma we also provide an estimate for the convexity modulus of j/C_j.

Remark 5.2.6. In case II we have $j \equiv \hat{j}$, so the estimates for j are actually estimates for the original objective $\hat{j}$ of $(\mathrm{P}_{\mathrm{red}})$. In case I we use the reformulation $j = -C_j \ln(C_{\hat{j}} - \hat{j})$. However, by use of $\frac{|j(u)-j(v)|}{C_j} = |\ln(1 + \frac{\hat{j}(u)-\hat{j}(v)}{C_{\hat{j}}-\hat{j}(u)})|$, estimates for j may be transferred to $\hat{j}$ also in this case. For instance, if k is large enough, then $\hat{j}(u^k) - \hat{j}(\bar{u}_\varepsilon) \le 2.5(C_{\hat{j}} - \hat{j}(u^k))$ holds. If, in addition, $\hat{j} \ge 0$ is valid, then $\frac{|j(u^k)-j(\bar{u}_\varepsilon)|}{C_j} \ge \frac{|\hat{j}(u^k)-\hat{j}(\bar{u}_\varepsilon)|}{2C_{\hat{j}}}$ is satisfied for all k large enough. In particular, this implies that $(\hat{j}(u^k))$ is r-linear convergent.

Remark 5.2.7. In case I we have $\vartheta_b = \vartheta(\varepsilon) = \tau(\varepsilon)$ and $\tau(\varepsilon)$ only has to satisfy $\tau(\varepsilon) \ge 1$. In case II it holds $\vartheta_b = \vartheta(\varepsilon) = \tau(\varepsilon) + \tilde{\tau}(\varepsilon)$ and $\tilde{\tau}(\varepsilon)$ has to be chosen of order $\mathcal{O}(1/\varepsilon^2)$, cf. Lemma 3.5.18.

Remark 5.2.8. The estimates in 5) indicate that in case I it is not sensible to choose $\vartheta_b = \vartheta(\varepsilon) = \tau(\varepsilon)$ too small, since then the first term on the right-hand side may already be smaller than the second term for $\mu = \mu_0$. This matches the observation from our numerical experiments that phase one may become excessively long for small $\tau(\varepsilon)$ if ε is small. In several experiments for case I choices around $\tau(\varepsilon) = C_j\varepsilon(1 + |\ln \varepsilon|)$ produced good results.

5.3. A long step method

We consider the following long step method for solving (P_ε).

Algorithm LSM_ε (long step method to solve (P_ε))

Input: Parameters $\varepsilon > 0$, $\mu_0 > 0$ in case I and $\mu_0 \le 1$ in case II, $\theta \in (0, \frac{1}{4}]$, $\beta_{\min}, \beta_{\max} \in (0,1)$ with $\beta_{\min} \le \beta_{\max}$, $\tau(\varepsilon) \ge 1$, starting point $u^0 \in \Lambda_{\varepsilon,\mu_0}(\tau^{-1})$ with $\tau := 2\sqrt[6]{2}$.

Set $C_\mu := \mu_0$ and select C_j according to Lemma 3.5.8 in case I and $\tilde{\tau}(\varepsilon)$ according to Lemma 3.5.18 in case II. Set $f_\mu := f_{\varepsilon,\mu}$.

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k \in U$ by solving $f''_{\mu_k}(u^k)[s^k] = -f'_{\mu_k}(u^k)$ in U^*.

CALL Algorithm LSMSUB from Section 2.7 with $(u^k, s^k, \mu_k, \theta)$ (use $X := U$, $K := U_{\mathrm{ad}}(\varepsilon)$, and $I_s = (0, C_\mu]$ in LSMSUB) and denote its return value by u^{k+1}.

Choose $\beta_k \in [\beta_{\min}, \beta_{\max}]$ and set $\mu_{k+1} := \beta_k \mu_k$.

END

Remark 5.3.1. Termination criteria for LSM_ε can be based on the next theorem.

Remark 5.3.2. In LSMSUB line searches may be employed, cf. Remark 2.7.4. We emphasize that line searches based on function values do not require additional solves of the state equation and are, therefore, numerically inexpensive: Given the actual iterate $\tilde{u}^l$ in LSMSUB and the corresponding Newton step $\tilde{s}^l$, line searches evaluate $f_{\mu_k}(\tilde{u}^l + t\tilde{s}^l)$ for some $t \in \mathbb{R}$, which requires $y(\tilde{u}^l + t\tilde{s}^l)$. Assuming that $A^{-1}g$ and $y(\tilde{u}^l)$ are already available the affine linearity of the state equation implies $y(\tilde{u}^l + t\tilde{s}^l) = y(\tilde{u}^l) + ty(\tilde{s}^l) - tA^{-1}g$. This demonstrates that only the state $y(\tilde{s}^l)$ has to be computed for the line search, i.e., one solve of the state equation is required. However, $y(\tilde{s}^l)$ can then be used to compute $y(\tilde{u}^{l+1})$ without solving the state equation. Since the computation of $y(\tilde{u}^{l+1})$ is necessary anyway, this shows that function evaluations for line search do not require additional solves of the state equation.

We have the following theorem on convergence and complexity of Algorithm LSM_ε. It is one of the main results of this thesis. Note that u^{k+1} is determined in iteration $k \in \mathbb{N}_0$ of LSM_ε.

Theorem 5.3.3. *Algorithm LSM_ε generates a sequence $(u^k) \subset U_{ad}(\varepsilon)$ with $u^{k+1} \in \Lambda_{\varepsilon,\mu_k}(\theta)$ for all $k \in \mathbb{N}_0$. Moreover, with $\vartheta_b := \vartheta(\varepsilon)$ we have for all $k \in \mathbb{N}_0$:*

1) *To reach iteration k (more precisely: to reach the FOR statement in LSM_ε for the $k+1$-th time) Algorithm LSM_ε requires at most*

$$k\left(\left\lfloor \frac{10.79}{\beta_{\min}}\left(\vartheta_b + \sqrt{\vartheta_b}\right)\right\rfloor + \left\lceil 8.13 + 1.45\ln\left|\ln\sqrt{2\theta}\right|\right\rceil\right)$$

Newton steps, including the Newton steps from LSMSUB.

2) *The sequence $(j(u^k))$ converges with r-linear rate β_k in iteration k to the optimal value of (P_ε). More precisely, there holds*

$$\frac{|j(u^{k+1}) - j(\bar{u}_\varepsilon)|}{C_j} \le \frac{\vartheta_b + \sqrt{\vartheta_b}}{C_j}\mu_0 \prod_{j=0}^{k-1}\beta_j = \frac{\vartheta_b + \sqrt{\vartheta_b}}{C_j}\mu_k.$$

3) *For every $\hat{\varepsilon} > 0$ we have the complexity estimate*

$$k \ge \left|\frac{1}{\ln\beta_{\max}}\right| \ln\left(\frac{\frac{\mu_0}{C_j}(\vartheta_b + \sqrt{\vartheta_b})}{\hat{\varepsilon}}\right) \quad \Longrightarrow \quad \frac{|j(u^{k+1}) - j(\bar{u}_\varepsilon)|}{C_j} \le \hat{\varepsilon}.$$

4) *Identical to 4) from Theorem 5.2.3, with u^k, β, and (in the complexity estimate) $\frac{2\sqrt{\vartheta_b}}{\delta}$ replaced by u^{k+1}, β_k, and $|\frac{2}{\ln\beta_{\max}}|$.*

5) *Identical to 5) from Theorem 5.2.3, with u^k replaced by u^{k+1}.*

Proof. 1), 2), 3), and 4) follow directly from their counterparts in Theorem 2.7.10, which is applicable due to Corollary 3.5.12 in case I and Corollary 3.5.20 in case II. The estimates in 5) follow from Lemma 5.1.3. □

Remark 5.3.4. The remarks after Theorem 5.2.3 also apply here.

5.4. Phase one

In this section we describe phase one methods for Algorithm SSM_ε and Algorithm LSM_ε. This is, we show how to obtain a starting point for Algorithm SSM_ε and Algorithm LSM_ε if only a point $u^0 \in U_{\mathrm{ad}}(\varepsilon)$ is known.

5.4.1. Phase one based on a short step method

The following Algorithm APOSS determines $\tilde{u} \in U_{\mathrm{ad}}(\varepsilon)$ with $\lambda_{\varepsilon,\mu_0}(\tilde{u}) \leq \theta$. Thus, $\tilde{u}$ is suitable as starting point for Algorithm SSM_ε and LSM_ε.

Algorithm APOSS (phase one based on short steps applied to $f_{\varepsilon,\mu}$)

Input: Parameters $\varepsilon > 0$, $\mu_0 > 0$ in case I and $\mu_0 \leq 1$ in case II, $\theta \in (0, \frac{1}{4}]$, $\tau(\varepsilon) \geq 1$, starting point $u^0 \in U_{\mathrm{ad}}(\varepsilon)$.

Output: $\tilde{u} \in \Lambda_{\varepsilon,\mu_0}(\theta)$.

Set $C_\mu := \mu_0$ and select C_j according to Lemma 3.5.8 in case I and $\tilde{\tau}(\varepsilon)$ according to Lemma 3.5.18 in case II. Define $f_{\mu_0} := f_{\varepsilon,\mu_0}$, $K := U_{\mathrm{ad}}(\varepsilon)$, and set for $\nu > 0$

$$f_{\nu,\mu_0,u^0} : K \to \mathbb{R}, \quad f_{\nu,\mu_0,u^0}(u) := f_{\mu_0}(u) - \frac{f'_{\mu_0}(u^0)[u]}{\nu}.$$

CALL Algorithm POSS from Section 2.9.1 with (u^0, μ_0, θ) (use $X = U$, $I_s = (0, C_\mu]$ in Algorithm POSS) and denote its return value by $\tilde{u}$.
RETURN $\tilde{u}$.

Remark 5.4.1. Algorithm POSS requires the self-boundedness constant $\vartheta_{f_{\varepsilon,\mu_0}}$ of $f_{\mu_0} = f_{\varepsilon,\mu_0}$. From Lemma 3.5.8 and Lemma 3.5.18 we know that this constant is given by

$$\vartheta_{f_{\varepsilon,\mu_0}} = \frac{C_j}{\mu_0} + \tau(\varepsilon) \text{ in case I} \quad \text{and} \quad \vartheta_{f_{\varepsilon,\mu_0}} = 2\left(\frac{C^2 C_{\|\cdot\|}}{\mu_0^2 \tilde{\tau}(\varepsilon)} + \tau(\varepsilon) + \tilde{\tau}(\varepsilon)\right) \text{ in case II,}$$

where C satisfies $\|\hat{j}'(u)\|_{U^*} \leq C$ for all $u \in U_{\mathrm{ad}}(\varepsilon)$.

We have the following complexity result for Algorithm APOSS. We recall that $\mathrm{sym}(u^0, U_{\mathrm{ad}}(\varepsilon))$, the symmetry of $U_{\mathrm{ad}}(\varepsilon)$ about u^0, is introduced in Definition 2.5.25.

Theorem 5.4.2. *Algorithm APOSS returns a $\tilde{u}$ that satisfies $\lambda_{\varepsilon,\mu_0}(\tilde{u}) \leq \theta$ after $N \in \mathbb{N}_0$ iterations of Algorithm POSS, where N is bounded from above by*

$$N \leq \left\lceil \frac{17}{16} \cdot \frac{\sqrt{\vartheta_{f_{\varepsilon,\mu_0}}}}{\delta} \cdot \ln\left(\frac{2\vartheta_{f_{\varepsilon,\mu_0}}}{\theta}\left(1 + \frac{1}{\mathrm{sym}(u^0, U_{ad}(\varepsilon))}\right)\right)\right\rceil$$

with

$$\vartheta_{f_{\varepsilon,\mu_0}} = \frac{C_j}{\mu_0} + \tau(\varepsilon) \text{ in case I}, \quad \vartheta_{f_{\varepsilon,\mu_0}} = 2\left(\frac{C^2 C_{\|\cdot\|}}{\mu_0^2 \tilde{\tau}(\varepsilon)} + \tau(\varepsilon) + \tilde{\tau}(\varepsilon)\right) \text{ in case II},$$

and

$$\delta = \frac{\tilde{\theta}\left(1 - \frac{\tilde{\theta}}{(1-\tilde{\theta})^2}\right)}{1 - \frac{\tilde{\theta}}{\sqrt{\vartheta_{f_{\varepsilon,\mu_0}}}}}, \text{ where } \tilde{\theta} = \frac{\theta}{2}.$$

Here, C satisfies $\|\hat{j}'(u)\|_{U^*} \leq C$ for all $u \in U_{ad}(\varepsilon)$.

During the course of APOSS, $2N+1$ Newton steps have to be computed.

Proof. The boundedness of $U_{\mathrm{ad}}(\varepsilon)$ together with Lemma 3.5.8 and Corollary 3.5.12 in case I, respectively, Lemma 3.5.18 and Corollary 3.5.20 in case II, yield that Theorem 2.9.5 is applicable with constants of self-boundedness as claimed. Theorem 2.9.5 now implies all assertions. □

Remark 5.4.3. $\vartheta_{f_{\varepsilon,\mu_0}}$ has at least order $\mathcal{O}(\frac{1}{\varepsilon^2})$, cf. Lemma 3.5.8 and Lemma 3.5.18.

5.4.2. Phase one based on a long step method

The following Algorithm APOLS determines $\tilde{u} \in U_{\mathrm{ad}}(\varepsilon)$ with $\lambda_{\varepsilon,\mu_0}(\tilde{u}) \leq \theta$. Thus, $\tilde{u}$ is suitable as starting point for Algorithm SSM_ε and LSM_ε.

Algorithm APOLS (phase one based on long steps applied to $f_{\varepsilon,\mu}$)

Input: Parameters $\varepsilon > 0$, $\mu_0 > 0$ in case I and $\mu_0 \leq 1$ in case II, $\theta \in (0, \frac{1}{4}]$, $\tau(\varepsilon) \geq 1$, starting point $u^0 \in U_{\mathrm{ad}}(\varepsilon)$.

Output: $\tilde{u} \in \Lambda_{\varepsilon,\mu_0}(\theta)$.

Set $C_\mu := \mu_0$ and select C_j according to Lemma 3.5.8 in case I and $\tilde{\tau}(\varepsilon)$ according to Lemma 3.5.18 in case II. Define $f_{\mu_0} := f_{\varepsilon,\mu_0}$ and $K := U_{\mathrm{ad}}(\varepsilon)$.

Compute the Newton step $s^0 \in U$ by solving $f''_{\mu_0}(u^0)[s^0] = -f'_{\mu_0}(u^0)$ in U^*.
CALL Algorithm LSMSUB from Section 2.7 with $(u^0, s^0, \mu_0, \theta)$ (use $X = U$, $I_s = (0, C_\mu]$ in LSMSUB) and denote its return value by $\tilde{u}$.
RETURN $\tilde{u}$.

We have the following complexity result for Algorithm APOLS.

Theorem 5.4.4. *Algorithm APOLS returns a $\tilde{u}$ that satisfies $\lambda_{\varepsilon,\mu_0}(\tilde{u}) \leq \theta$ after $N \in \mathbb{N}_0$ iterations of LSMSUB, where N is bounded from above by*

$$N \leq \left\lfloor 10.79 \vartheta_{f_{\varepsilon,\mu_0}} \left| \ln\left(1 - \omega_{\bar{u}_{\varepsilon,\mu_0}}(u^0)\right) \right| \right\rfloor + \left\lceil 7.13 + 1.45 \ln \left| \ln \sqrt{2\theta} \right| \right\rceil.$$

Here, $\bar{u}_{\varepsilon,\mu_0}$ denotes the minimizer of f_{ε,μ_0} on $U_{ad}(\varepsilon)$, $\omega_{\bar{u}_{\varepsilon,\mu_0}} : U_{ad}(\varepsilon) \to [0,1)$ denotes the Minkowski function, see Definition 2.3.14, and $\vartheta_{f_{\varepsilon,\mu_0}}$ denotes the same constant as in Theorem 5.4.2.

During the course of APOLS, $N+1$ Newton steps have to be computed.

Proof. Due to Lemma 3.5.8 and Corollary 3.5.12 in case I, respectively, Lemma 3.5.18 and Corollary 3.5.20 in case II, Theorem 2.9.6 is applicable with constants of self-boundedness as claimed. Theorem 2.9.6 now implies all assertions. □

5.5. Comparison with a grid-based approach

In this section we compare our approach for fixed ε with another approach.

For simplicity let us assume $\Omega = \Omega_a$. We consider a mesh on $\overline{\Omega}$ with nodes $(x_i)_{i=1,\dots,n}$. We can now impose the constraints $y(x_i) \geq y_a(x_i)$ for $i = 1, \dots, n$ and introduce the 1-self-concordant functions $u \mapsto -\ln(y(u)(x_i) - y_a(x_i))$ for $i = 1, \dots, n$. The fact that these functions are 1-self-concordant follows from Lemma 2.1.19 and Corollary 2.3.9. Instead of (P_ε) and the associated barrier functional $f_{\varepsilon,\mu}$ we work with $f_{n,\mu}(u) := \frac{\hat{j}(u)}{\mu} + b_n(u)$ in this approach, where $b_n(u) := -\sum_{i=1}^n \ln(y(u)(x_i) - y_a(x_i))$ is n-self-concordant. Accordingly, we want to solve

$$\min_{u \in U} \hat{j}(u) \qquad \text{s.t.} \qquad y(x_i) \geq y_a(x_i) \text{ for } i = 1, \dots, n,$$

which we call (P_n). We can now apply the theory from Section 2 to obtain, e.g., a short step method for the solution of (P_n) for a fixed n. In fact, with $\vartheta_b := n$ we can directly use Algorithm SSM and the according Theorem 2.6.3 if we write u instead of x and $\hat{j}$ instead of j. If we compare Theorem 2.6.3 2) to Theorem 5.2.3 2) for case II, where we have $j = \hat{j}$, we see that $\vartheta_b = n$ corresponds to $\vartheta_b = \vartheta(\varepsilon)$ in this case, i.e., to obtain the same error bound for the objective $\hat{j}$ when working with $f_{n,\mu}$ instead of $f_{\varepsilon,\mu}$, we have to use a mesh with $n = \lceil \vartheta(\varepsilon) \rceil = \mathcal{O}(\varepsilon^{-2})$ points. For a bounded set $S \subset C^{0,\beta}(\overline{\Omega})$ we have the estimate $|(y(x) - y_a(x))^-| \leq C_S \min_{i \in \{1,\dots,n\}} \|x - x_i\|_2^\beta$ for all $x \in \overline{\Omega}$ and all $y \in S$, where $C_S > 0$ only depends on S. Under a regularity assumption on the mesh we obtain $\min_{i \in \{1,\dots,n\}} \|x - x_i\|_2 \leq Ch$ for all $x \in \overline{\Omega}$, with $h = \frac{1}{n^{1/d}}$ and a constant C that is independent of n. This yields $\|(y - y_a)^-\|_{C(\overline{\Omega})} \leq Cn^{-\beta/d}$. Using $\Omega = (0,1)^d$ it can be argued that this order is sharp, in general. For $n = \tilde{C}\varepsilon^{-2}$ this implies $\|(y - y_a)^-\|_{C(\overline{\Omega})} \leq C\varepsilon^{2\beta/d}$, where the value of C may have changed but is still independent of n, ε, β, and d. Since the maximum pointwise constraint violation directly translates into the length of the path of optimal solutions, cf. Theorem 4.4.8 and its proof, we obtain that for $n = \tilde{C}\varepsilon^{-2}$ points there holds $|\hat{j}(\bar{u}_n) - \hat{j}(\bar{u})| \leq C\varepsilon^{2\beta/d}$, where $\bar{u}_n$ denotes the optimal solution to (P_n). For Algorithm SSM this implies with Theorem 2.6.3 2) that we have

$$|\hat{j}(u^k) - \hat{j}(\bar{u})| \leq 2\mu_k \vartheta_b + C\varepsilon^{2\beta/d}.$$

By comparison of this estimate with the one from Theorem 5.2.3 5) we conjecture that this grid-based approach is favourable if it holds $\varepsilon^{2\beta/d} \leq \varepsilon(1 + |\ln \varepsilon|)$ or, a little less precise, if

$2\beta/d \geq 1$. Assuming $Y \hookrightarrow H^2(\Omega)$ we have $\beta \approx 1$ in dimension $d = 2$ and, thus, conclude that the two approaches may be comparable. However, if H^2-regularity is not available, then our approach seems favourable even for $d = 2$. In particular, we see that our approach is more robust with respect to the parameters β and d. For $d = 3$ our approach seems superior, anyway. Also, let us mention that if we compare this new approach to case I, where we usually work with $\vartheta_b = \mathcal{O}(\frac{1+|\ln \varepsilon|}{\varepsilon})$, we obtain that our approach seems favourable for $d = 2$.

Finally, we note that a comparison of the practical performance of algorithms based on $f_{\varepsilon,\mu}$ with algorithms based on $f_{n,\mu}$ may be interesting.

6. Theoretical background for variable smoothing parameter

In this section we provide results that are required for the convergence analysis of barrier methods that drive ε and μ to zero.

6.1. Standing assumptions

In this section we present additional assumptions that we impose to treat variable ε.

When we dealt with fixed $\varepsilon > 0$, the key to apply self-concordance theory was to make $f_{\varepsilon,\mu}$ fulfill Assumption 2.5.2 for all μ from some interval $(0, \mu_s]$, cf. Corollary 3.5.12 and Corollary 3.5.20. With this Assumption fulfilled a main result was Lemma 5.1.3. When dealing with variable ε we want $f_{\varepsilon,\mu}$ to satisfy Assumption 2.5.2 for all ε from the interval $(0, \varepsilon_s]$ and for all μ from some interval that may depend on ε. We now explain for case I and case II how to choose the parameters that occur in $f_{\varepsilon,\mu}$. In case I we scale $f_{\varepsilon,\mu} = \frac{-C_j \ln(C_j - \hat{j})}{\mu} - \tau(\varepsilon)\ln(B^\varepsilon)$ such that C_j is independent of $\varepsilon \in (0, \varepsilon_s]$. From Lemma 3.5.8 we see that this is possible with $C_j = \max\{\varepsilon_s^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} (\frac{\|2\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U + \|g\|_Z)^2\}$ provided we choose $C_\mu = \varepsilon^2$ in this lemma. This implies that we have to choose $\mu \leq \varepsilon^2$ to make $f_{\varepsilon,\mu}$ fulfill Assumption 2.5.2. If we choose $\mu = \varepsilon^2$, Lemma 5.1.3 tells us to use $\tau(\varepsilon) = C_\tau \frac{1+|\ln \varepsilon|}{\varepsilon}$ with an ε-independent $C_\tau > 0$ to optimally balance the errors with respect to ε (recall that $\vartheta(\varepsilon) = \tau(\varepsilon)$ in case I). In case II we have $\vartheta(\varepsilon) = \tau(\varepsilon) + \tilde{\tau}(\varepsilon)$ and we want to choose $\vartheta(\varepsilon)$ small with respect to the order in ε. Therefore, Lemma 3.5.18 leads to the use of $\tilde{\tau}(\varepsilon) = \frac{1}{\varepsilon^2} \max\{\varepsilon^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} (\sqrt{2C_{\|\cdot\|}} + \|g\|_Z)^2\}$. Considering Lemma 5.1.3 we then work with $\mu = C\varepsilon^3$ to balance the errors suitably, where $C > 0$ needs to be chosen such that $\mu \leq 1$ is satisfied for all $\varepsilon \in (0, \varepsilon_s]$. Here, we neglected the logarithm for simplicity.

Summarizing, we have motivated the following assumption on the parameter choice, that we impose for the rest of this thesis.

Assumption 6.1.1. *From now on we consider a fixed $\varepsilon_s > 0$.*

- *In case I: Choose*

$$C_j := \max\left\{\varepsilon_s^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} \left(\frac{\|2\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U + \|g\|_Z\right)^2\right\}$$

with a $\tilde{u} \in U$ that fulfills $\hat{j}(\tilde{u}) \geq C_{\hat{j}} = 1 + \hat{j}(u^\circ)$. Moreover, let $\mathbb{R}_{>0} \ni \varepsilon \mapsto \tau(\varepsilon) \in \mathbb{R}_{>0}$ be continuously differentiable with $\tau(\varepsilon) \geq 1$ for all $\varepsilon \in (0, \varepsilon_s]$ and $\tau(\varepsilon) := C_\tau \frac{1+|\ln \varepsilon|}{\varepsilon}$ for all $\varepsilon \in (0, \frac{1}{2}]$, where $C_\tau > 0$ is a constant.

- *In case II: Let $j = \hat{j}$ be uniformly convex on U with bounded second derivative on bounded sets. Choose*
$$\tilde{\tau}(\varepsilon) := \frac{1}{\varepsilon^2} \max\left\{\varepsilon_s^2, \frac{16}{9} C_{\partial, C(\overline{\Omega}_a)}^2 \left(\sqrt{2C_{\|\cdot\|}} + \|g\|_Z\right)^2\right\}$$
for all $\varepsilon \in \mathbb{R}_{>0}$. Moreover, let $\mathbb{R}_{>0} \ni \varepsilon \mapsto \tau(\varepsilon) \in \mathbb{R}_{>0}$ be continuously differentiable with $\tau(\varepsilon) \geq 1$ for all $\varepsilon \in (0, \varepsilon_s]$ and $\tau(\varepsilon) = C_\tau \tilde{\tau}(\varepsilon)$ for all $\varepsilon \in (0, \varepsilon_s]$, where $C_\tau > 0$ is a constant.

Remark 6.1.2. We recall that by definition $C_{\partial, C(\overline{\Omega}_a)}$ denotes *any* constant that satisfies the estimates $\|A^{-1}Bu\|_{C(\overline{\Omega}_a)} \leq C_{\partial, C(\overline{\Omega}_a)} \|u\|_U$ for all $u \in U$ and $\|A^{-1}g\|_{C(\overline{\Omega}_a)} \leq C_{\partial, C(\overline{\Omega}_a)} \|g\|_Z$, cf. Lemma 3.1.14. Since the maximum in the definitions of C_j and $\tilde{\tau}(\varepsilon)$ is usually attained for the second term in the max operator, this shows that the choice of C_j and $\tilde{\tau}(\varepsilon)$ is still somewhat flexible.

As explained above we choose $\mu = \varepsilon^2$ in case I and $\mu = C\varepsilon^3$ in case II. To this end, we introduce the following definitions.

Definition 6.1.3. Denote
$$\rho : \mathbb{R}_{>0} \to \mathbb{R}_{>0}, \quad \rho(t) = C_\rho \sqrt[p]{t},$$
with $p = 2$ in case I and $p = 3$ in case II, and $C_\rho > 0$ a constant. In case I let $C_\rho := 1$ and in case II let it be such that $\mu_s := \rho^{-1}(\varepsilon_s) \leq 1$ holds, where ρ^{-1} denotes the inverse function, i.e., $\rho^{-1} \circ \rho \equiv 1$.

Definition 6.1.4. We define $\mu_s := \rho^{-1}(\varepsilon_s)$ and set
$$\mathcal{P}_{\leq} := \{(\varepsilon, \mu) \in (0, \varepsilon_s] \times (0, \mu_s] : \rho(\mu) \leq \varepsilon\}.$$
Furthermore, we define
$$\mathcal{P}_{=} := \{(\varepsilon, \mu) \in (0, \varepsilon_s] \times (0, \mu_s] : \rho(\mu) = \varepsilon\}.$$
For a pair $(\varepsilon, \mu) \in \mathcal{P}_=$ we sometimes write $(\varepsilon(\mu), \mu)$, i.e., we use $\varepsilon(\mu) := \rho(\mu)$.

Remark 6.1.5. For $(\varepsilon, \mu) \in \mathcal{P}_=$ we have $\varepsilon \to 0^+$ if and only if $\mu \to 0^+$. We use this tacitly from now on.

The next lemma presents fundamental properties of $f_{\varepsilon,\mu}$. In particular, it shows that $f_{\varepsilon,\mu}$ allows for the application of self-concordance theory.

Lemma 6.1.6. *We have*

- *In case I: $f_{\varepsilon,\mu}$ is a nondegenerate $(\frac{C_j}{\mu} + \tau(\varepsilon))$-self-concordant barrier for $U_{ad}(\varepsilon)$ for every $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$. For every $\varepsilon \in (0, \varepsilon_s]$, $f_{\varepsilon,\mu}$ satisfies Assumption 2.5.2 with $M := D_j \cap \overline{D}_{b^\varepsilon}$, $K := U_{ad}(\varepsilon) = D_j \cap D_{b^\varepsilon}$, $\mu_s := \rho^{-1}(\varepsilon)$, and $\vartheta_b := \tau(\varepsilon)$.*

- *In case II: $f_{\varepsilon,\mu}$ is a nondegenerate $2\left(\frac{\tilde{C}^2 C_{\|\cdot\|}}{\mu^2\tilde{\tau}(\varepsilon)} + \tau(\varepsilon) + \tilde{\tau}(\varepsilon)\right)$-self-concordant barrier for $U_{ad}(\varepsilon)$ for every $(\varepsilon,\mu) \in \mathcal{P}_{\leq}$. Here, $\tilde{C}$ denotes the bound on the first derivative of $\hat{j}$ on $U_{ad}(\varepsilon)$. For every $\varepsilon \in (0,\varepsilon_s]$, $f_{\varepsilon,\mu}$ satisfies Assumption 2.5.2 with $M := \overline{D}_{b^\varepsilon} \cap \overline{D}_{\tilde{b}^\varepsilon}$, $K := U_{ad}(\varepsilon)$, $\mu_s := \rho^{-1}(\varepsilon)$, and $\vartheta_b := \tau(\varepsilon) + \tilde{\tau}(\varepsilon)$.*
- *In both cases $f_{\varepsilon,\mu}$ possesses exactly one global minimizer $\bar{u}_{\varepsilon,\mu} \in U_{ad}(\varepsilon)$ for every $(\varepsilon,\mu) \in \mathcal{P}_{\leq}$.*
- *In both cases it holds for every $(\varepsilon,\mu) \in \mathcal{P}_{\leq}$*

$$\frac{|j(u) - j(\bar{u})|}{C_j} \leq \frac{2\mu\vartheta(\varepsilon)}{C_j} + C\varepsilon\,(1 + |\ln\varepsilon|)$$

for all $u \in \Lambda_{\varepsilon,\mu}$, where $C > 0$ is independent of ε, $\tau(\varepsilon)$, $\tilde{\tau}(\varepsilon)$, C_j, μ, and u. Moreover, it holds for every $(\varepsilon,\mu) \in \mathcal{P}_{\leq}$

$$\|u - \bar{u}\|_U \leq \sqrt{\frac{8\mu\vartheta(\varepsilon)}{C_j\alpha}} + \sqrt{\frac{4C}{\alpha}}\sqrt{\varepsilon\,(1 + |\ln\varepsilon|)}.$$

for all $u \in \Lambda_{\varepsilon,\mu}$ and the same C, where $\alpha > 0$ denotes the convexity modulus of j/C_j on $\cup_{\varepsilon\in(0,\varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$.
- *In particular, we have in both cases strong convergence $\bar{u}_{\varepsilon,\mu} \to \bar{u}$ for $(\varepsilon,\mu) \in \mathcal{P}_{=}$ and $\varepsilon \to 0^+$.*

Proof. The assertions for case I are special cases of Lemma 3.5.8 and Corollary 3.5.12. The assertions for case II are special cases of Lemma 3.5.18 and Corollary 3.5.20. Since Assumption 2.5.2 is fulfilled for every fixed $\varepsilon \in (0,\varepsilon_s]$, the assertion on $\bar{u}_{\varepsilon,\mu}$ is implied by Corollary 2.5.13. The estimates for both cases follow directly from Lemma 5.1.3, with $\cup_{\varepsilon\in(0,\varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$ being a consequence of Corollary 4.1.10. The convergence of $\bar{u}_{\varepsilon,\mu}$ follows from the estimate for $\|u - \bar{u}\|_U$ by use of $\mu\vartheta(\varepsilon) \to 0^+$ for $(\varepsilon,\mu) \in \mathcal{P}_{=}$ and $\varepsilon \to 0^+$. □

Remark 6.1.7. The convexity modulus α of j/C_j on $M(\varepsilon_s)$ is well-defined. In case I this follows since $M(\varepsilon_s) \subset D_j$ holds and since j is uniformly convex on D_j due to Lemma C.4.14. In case II $j = \hat{j}$ is uniformly convex on U, hence also on $M(\varepsilon_s)$.

We require the existence of a certain test function.

Definition 6.1.8. For a bounded set $\Omega \subset \mathbb{R}^d$ and $\delta > 0$ we denote

$$\Omega_\delta := \{x \in \Omega : \|x - \tilde{x}\|_2 < \delta \text{ for some } \tilde{x} \in \partial\Omega\} \quad \text{and} \quad \Omega_\delta^C := \Omega \setminus \Omega_\delta.$$

Assumption 6.1.9. *Let the equality constraint $Ay + Bu = g$ in ($\mathrm{P}_{\mathrm{orig}}$) be a partial differential equation on $\Omega \supset \Omega_a$.*

1) If we have $\Omega = \Omega_a$ and homogeneous Dirichlet boundary conditions, let $\tilde{\delta} := (\frac{\tau^\circ}{4L})^{1/\beta}$, where $L := \|\bar{y} - y_a\|_{C^{0,\beta}(\overline{\Omega}_a)}$ and τ° denotes the constant from Assumption 3.1.9. We assume that there exist $\hat{u} \in U$, $\hat{\varepsilon} > 0$, and a positive $\hat{\delta} \leq \tilde{\delta}$ with

$$\hat{y}(x) \geq 0 \text{ for all } x \in \Omega \qquad \text{and} \qquad \hat{y}(x) \geq \hat{\varepsilon} \text{ for all } x \in \Omega_{\hat{\delta}}^C,$$

where $\hat{y} \in Y$ is defined via $\hat{y} := -A^{-1}B\hat{u}$.

2) *If we have Neumann or Robin boundary conditions or if it holds* $\overline{\Omega}_a \subset \Omega$*, we assume that there exist* $\hat{u} \in U$ *and* $\hat{\varepsilon} > 0$ *with*

$$\hat{y}(x) \geq \hat{\varepsilon} \text{ for all } x \in \Omega_a,$$

where $\hat{y} \in Y$ *is defined via* $\hat{y} := -A^{-1}B\hat{u}$.

Remark 6.1.10. Notice that we do not require $\hat{u}$ to be feasible but only to belong to U. Also notice that the assumption we make in the case of Neumann or Robin boundary conditions and $\Omega = \Omega_a$ is not sensible in the case of homogeneous Dirichlet boundary conditions.

Remark 6.1.11. In particular, we exclude from now on mixed boundary conditions on $\Omega = \Omega_a$ as well as the case where we have homogeneous Dirichlet boundary conditions on $\partial\Omega$ with $\partial\Omega_a \cap \partial\Omega \neq \emptyset$ but $\Omega_a \neq \Omega$. However, we mention that these scenarios can be included but the previous lemma and assumption become more technical then.

Remark 6.1.12. For $B = -\mathrm{Id}$ the assumption simplifies. In fact, it is then only an assumption on $\hat{y}$ since if $\hat{y} \in Y$ satisfies, e.g., $\hat{y}(x) \geq \hat{\varepsilon}$ for all $x \in \Omega_a$, then we can obtain an according $\hat{u}$ through $\hat{u} := A\hat{y}$. This shows, in particular, that for standard examples such as $-\Delta y = u$ on $\Omega = \Omega_a$ with homogeneous Dirichlet boundary conditions the above assumption is valid (use a *partition of unity* to obtain $\hat{y}$, cf., e.g., [Alt06, Section 2.19]).

6.2. Distance to the boundary I

By definition, $\bar{u}_{\varepsilon,\mu} \in U_{\mathrm{ad}}(\varepsilon)$ is a minimizer of $f_{\varepsilon,\mu}$, and this functional is a barrier for $U_{\mathrm{ad}}(\varepsilon)$. The aim of this section is to show in some sense that the distance between $\bar{u}_{\varepsilon,\mu}$ and $\partial U_{\mathrm{ad}}(\varepsilon)$ is uniformly bounded away from zero if ε and μ are uniformly bounded away from zero. The following lemma states the precise result. For a convenient notation we introduce a definition first. We recall that in case II we have $\tilde{B}(u) := C_{\|\cdot\|} - \frac{1}{2}\|u\|_U^2$, whereas in case I we have not defined $\tilde{B}$ so far.

Definition 6.2.1. In case I we define

$$\tilde{B} : U \to \mathbb{R}, \quad \tilde{B}(u) := C_{\hat{j}} - \hat{j}(u).$$

Lemma 6.2.2. *Let* $\tilde{\varepsilon}, \tilde{\mu} > 0$*. Then there exists* $c > 0$ *such that for every* $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$ *with* $\varepsilon \geq \tilde{\varepsilon}$ *and* $\mu \geq \tilde{\mu}$ *we have* $B^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) \geq c$ *and* $\tilde{B}(\bar{u}_{\varepsilon,\mu}) \geq c$.

Proof. It suffices to show that $(\varepsilon, \mu) \mapsto \bar{u}_{\varepsilon,\mu}$ is continuous on $\mathcal{P}_{\leq}$ since then $(\varepsilon, \mu) \mapsto B^{\varepsilon}(\bar{u}_{\varepsilon,\mu})$ and $(\varepsilon, \mu) \mapsto \tilde{B}(\bar{u}_{\varepsilon,\mu})$ are continuous and positive on $\mathcal{P}_{\leq}$, which implies the assertion by compactness.

For every $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$, the point $\bar{u}_{\varepsilon,\mu}$ is the unique root of $f'_{\varepsilon,\mu} : U_{\mathrm{ad}}(\varepsilon) \to U^*$. Moreover, since $f_{\varepsilon,\mu}$ is uniformly convex, it follows from Lemma C.1.4 that $f''_{\varepsilon,\mu}(\bar{u}_{\varepsilon,\mu}) \in \mathcal{L}(U, U^*)$ is invertible for each $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$. Application of the implicit function theorem, cf. Theorem C.2.25, yields, in particular, that $(\varepsilon, \mu) \mapsto \bar{u}_{\varepsilon,\mu}$ is continuous on $\mathcal{P}_{\leq}$. □

Remark 6.2.3. We could prove the preceding lemma by directly estimating $B^{\varepsilon}(\bar{u}_{\varepsilon,\mu})$ and $\tilde{B}(\bar{u}_{\varepsilon,\mu})$, in which case we would argue without using the implicit function theorem. However, the proof given above is much shorter.

6.3. Distance to the boundary II

In this section we derive a positive lower bound for the term $\tilde{B}(u)$, where u belongs to any of the neighborhoods $\Lambda_{\varepsilon(\mu),\mu}$ of $\bar{u}_{\varepsilon(\mu),\mu}$. This is, the neighborhoods $\Lambda_{\varepsilon(\mu),\mu}$ provide a certain distance to the boundary of $U_{\mathrm{ad}}(\varepsilon)$ with respect to $\tilde{B}$. From this result we infer that the mapping $u \mapsto \tilde{b}^\varepsilon(u)$ is Lipschitz on $\Lambda_{\varepsilon(\mu),\mu}$. Furthermore, we provide a bound for $(\tilde{b}^\varepsilon)'$ on these neighborhoods.

Lemma 6.3.1. *There exists $c > 0$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds $\tilde{B}(u) \geq c$ for all $u \in \Lambda_{\varepsilon,\mu}$.*

Proof. We have $\tilde{B}(\bar{u}) \geq 1$ in case I and case II. By continuity of $u \mapsto \tilde{B}(u)$ at $u = \bar{u}$ we infer that there exists a $\delta > 0$ such that for all u with $\|u - \bar{u}\|_U \leq \delta$ it holds $\tilde{B}(u) \geq 1/2$. Lemma 6.1.6 implies that there is $\tilde{\varepsilon} > 0$ such that $\|u - \bar{u}\|_U \leq \delta$ is satisfied for all $(\varepsilon, \mu) \in \mathcal{P}_\leq$ with $\varepsilon \leq \tilde{\varepsilon}$ and all $u \in \Lambda_{\varepsilon,\mu}$. Thus, we have

$$\tilde{B}(u) \geq \frac{1}{2} \text{ for all } (\varepsilon, \mu) \in \mathcal{P}_\leq \text{ with } \varepsilon \leq \tilde{\varepsilon} \text{ and all } u \in \Lambda_{\varepsilon,\mu}. \tag{6.1}$$

Now we consider $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\varepsilon \in [\tilde{\varepsilon}, \varepsilon_s]$. From Lemma 2.2.23 we know that $f_{\varepsilon,\mu}(u) - f_{\varepsilon,\mu}(\bar{u}_{\varepsilon,\mu}) \leq C$ is satisfied for $u \in \Lambda_{\varepsilon,\mu}$ with a constant $C > 0$ that is independent of ε, μ, and u. Using the definition of $f_{\varepsilon,\mu}$ this yields

$$-C_j \ln\Big(\tilde{B}(u)\Big) \leq C\mu + \mu \max\{0, b^\varepsilon(\bar{u}_{\varepsilon,\mu}) - b^\varepsilon(u)\} - C_j \ln\Big(\tilde{B}(\bar{u}_{\varepsilon,\mu})\Big)$$

in case I and

$$-\ln\Big(\tilde{B}(u)\Big) \leq C + \frac{|j(\bar{u}_{\varepsilon,\mu}) - j(u)|}{+} \max\{0, b^\varepsilon(\bar{u}_{\varepsilon,\mu}) - b^\varepsilon(u)\} - \ln\Big(\tilde{B}(\bar{u}_{\varepsilon,\mu})\Big)$$

in case II, where we used $\tilde{\tau}(\varepsilon) \geq 1$.

We begin by arguing for the first case:

- Clearly, there holds $C\mu \leq C\mu_s$.
- The term $\mu b^\varepsilon(\bar{u}_{\varepsilon,\mu}) = -\mu\tau(\varepsilon)\ln(B^\varepsilon(\bar{u}_{\varepsilon,\mu}))$ is bounded from above on $\mathcal{P}_=$ with $\varepsilon \geq \tilde{\varepsilon}$. This follows from the fact that $\mu\tau(\varepsilon) = \rho^{-1}(\varepsilon)\tau(\varepsilon)$ is bounded on $[\tilde{\varepsilon}, \varepsilon_s]$ due to continuity, together with Lemma 6.2.2.
- For the term $-\mu b^\varepsilon(u) = \mu\tau(\varepsilon)\ln(B^\varepsilon(u))$ we have $\ln(B^\varepsilon(u)) \leq \ln(\max(y(u) - y_a))$, see Corollary 4.1.3, which is bounded from above independently of u, μ, and ε, cf. Corollary 4.2.2.
- The term $-C_j \ln(\tilde{B}(\bar{u}_{\varepsilon,\mu}))$, too, is bounded from above on $\mathcal{P}_=$with $\varepsilon \geq \tilde{\varepsilon}$, see Lemma 6.2.2.

From these arguments we see in case I that $-\ln(\tilde{B}(u))$ is bounded from above for all $u \in \Lambda_{\varepsilon,\mu}$, where $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\varepsilon \geq \tilde{\varepsilon}$ holds. Thus, $\tilde{B}(u)$ is bounded away from zero for all these (ε, μ) and u. Together with (6.1) the assertion follows in case I. In case II we can reason analogously, the only additional argument being that that $|j(\bar{u}_{\varepsilon,\mu}) - j(u)|$ is bounded since $j = \hat{j}$ is bounded on bounded sets. □

Corollary 6.3.2. *There exists $L > 0$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

- $|j(u_1) - j(u_2)| \leq L\|u_1 - u_2\|_U$ *in case I,*
- $|\tilde{b}^\varepsilon(u_1) - \tilde{b}^\varepsilon(u_2)| \leq L\tilde{\tau}(\varepsilon)\|u_1 - u_2\|_U$ *in case II*

for all $u_1, u_2 \in \Lambda_{\varepsilon,\mu}$.

Proof. By definition we have $j(u) = -C_j \ln(\tilde{B}(u))$ in case I and $\tilde{b}^\varepsilon(u) = -\tilde{\tau}(\varepsilon)\ln(\tilde{B}(u))$ in case II. From Lemma 4.2.2 and the fact that $\tilde{B}$ is quadratic we infer that $\tilde{B}$ is Lipschitz continuous on the bounded set $\cup_{\varepsilon>0} U_{\mathrm{ad}}(\varepsilon) \supset \cup_{(\varepsilon,\mu)\in\mathcal{P}_=} \Lambda_{\varepsilon,\mu}$. Since the natural logarithm is Lipschitz continuous on every interval $[c, \infty)$, $c > 0$, the assertion follows using the preceding lemma. □

Corollary 6.3.3. *There exists $L > 0$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

- $\|j'(u)\|_{U^*} \leq L$ *in case I,*
- $\|(\tilde{b}^\varepsilon)'(u)\|_{U^*} \leq L\tilde{\tau}(\varepsilon)$ *in case II*

for all $u \in \Lambda_{\varepsilon,\mu}$.

Proof. There hold

$$\|j'(u)\|_{U^*} = \frac{C_j \left\|\tilde{B}'(u)\right\|_{U^*}}{\tilde{B}(u)} \quad \text{and} \quad \left\|(\tilde{b}^\varepsilon)'(u)\right\|_{U^*} = \frac{\tilde{\tau}(\varepsilon)\left\|\tilde{B}'(u)\right\|_{U^*}}{\tilde{B}(u)}$$

in case I, respectively, case II. By virtue of Lemma 6.3.1 it suffices to show that $\tilde{B}'(u)$ is uniformly bounded for all $u \in \cup_{(\varepsilon,\mu)\in\mathcal{P}_=} \Lambda_{\varepsilon,\mu}$. Since $\tilde{B}$ is quadratic, it has bounded derivatives on bounded sets. Using $\cup_{(\varepsilon,\mu)\in\mathcal{P}_=} \Lambda_{\varepsilon,\mu} \subset \cup_{\varepsilon>0} U_{\mathrm{ad}}(\varepsilon)$ we obtain the assertion. □

Remark 6.3.4. To infer Corollary 6.3.2 from Corollary 6.3.3 we would have to show that $\Lambda_{\varepsilon,\mu}$ is convex. Therefore, we prefer a direct proof of Corollary 6.3.2.

6.4. Distance to the boundary III

We show in two steps that $B^\varepsilon(u)$ is bounded away from zero for $(\varepsilon, \mu) \in \mathcal{P}_=$ and $u \in \Lambda_{\varepsilon,\mu}$, where the bound depends on ε and μ and tends to zero for $\varepsilon, \mu \to 0^+$. This is, the neighborhood $\Lambda_{\varepsilon(\mu),\mu}$ provides a certain distance to the boundary of $U_{\mathrm{ad}}(\varepsilon)$ with respect to the smoothed minimum.

6.4.1. Step I: An estimate for $b^{\varepsilon(\mu)}(\bar{u}_{\varepsilon(\mu),\mu})$

We deduce an upper bound for the term $b^{\varepsilon(\mu)}(\bar{u}_{\varepsilon(\mu),\mu})$ from the following result.

Lemma 6.4.1. *There exists $c > 0$ such that for all $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

$$B^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) \geq c\mu\vartheta(\varepsilon).$$

Remark 6.4.2. This lemma indicates how fast $B^{\varepsilon}(\bar{u}_{\varepsilon,\mu})$ converges to zero for $(\varepsilon, \mu) = (\varepsilon(\mu), \mu)$ and $\mu \to 0^+$. Note that in the case $\min(\bar{y} - y_a) = 0$ the condition $B^{\varepsilon}(\bar{u}_{\varepsilon(\mu),\mu}) \to 0^+$ for $\mu \to 0^+$ is necessary to achieve convergence $\bar{u}_{\varepsilon(\mu),\mu} \xrightarrow{\|\cdot\|_U} \bar{u}$ for $\mu \to 0^+$. This can be shown by virtue of Lemma 4.1.1, Lemma 4.1.2, and Lemma 4.1.4, using that strong convergence $\bar{u}_{\varepsilon(\mu),\mu} \xrightarrow{\|\cdot\|_U} \bar{u}$ implies uniform convergence $\bar{y}_{\varepsilon(\mu),\mu} \xrightarrow{\|\cdot\|_{C(\overline{\Omega}_a)}} \bar{y}$, where $\bar{y}_{\varepsilon(\mu),\mu} := y(\bar{u}_{\varepsilon(\mu),\mu})$.

Proof. Before we start the actual proof, we recall the definition $B^{\varepsilon}(u) = B^{\varepsilon}_{C(\overline{\Omega}_a)}(y(u))$. In particular, we have $B^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) = B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu})$ and $B^{\varepsilon}(\bar{u}) = B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y})$, where $\bar{y}_{\varepsilon(\mu),\mu} := y(\bar{u}_{\varepsilon(\mu),\mu})$. We establish the assertions with $B^{\varepsilon}(\bar{u}_{\varepsilon,\mu})$ replaced by $B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu})$.

If we have $\bar{y} > y_a$ in $\overline{\Omega}_a$, we infer by continuity the existence of a constant $\eta > 0$ such that $\bar{y} - y_a \geq \eta$ is satisfied. Hence, we have

$$B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y}) = \min_{\varepsilon}(\bar{y} - y_a) \geq \min_{x \in \overline{\Omega}_a} (\bar{y}(x) - y_a(x)) \geq \eta$$

for all $\varepsilon \in (0, \infty)$ by Corollary 4.1.3. Using $\bar{y}_{\varepsilon,\mu} \to \bar{y}$ for $\mu \to 0^+$ with respect to $\|\cdot\|_{C(\overline{\Omega}_a)}$ we deduce that there is $\hat{\mu} > 0$ such that it holds $\bar{y}_{\varepsilon,\mu} - y_a \geq \eta/2$ in $\overline{\Omega}_a$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$. Therefore, we have $B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu}) \geq \eta/2$ for all these (ε, μ). The application of Lemma 6.2.2 yields a constant $\tilde{\eta} > 0$ such that $B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu}) \geq \tilde{\eta}$ holds for all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \geq \hat{\mu}$. Setting $c := \min\{\eta/2, \tilde{\eta}\}$ we obtain $B^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) = B^{\varepsilon}_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu}) \geq c$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$, which establishes the assertion in the case $\bar{y} > y_a$ in $\overline{\Omega}_a$. Hence, we may assume in the following that there exists at least one $x^* \in \overline{\Omega}_a$ such that $\bar{y}(x^*) = y_a(x^*)$ is satisfied.

Let $(\varepsilon, \mu) \in \mathcal{P}_=$. We have $f'_{\varepsilon,\mu}(\bar{u}_{\varepsilon,\mu}) = 0$ in U^*, that is

$$\frac{j'(\bar{u}_{\varepsilon,\mu})[h]}{\mu} + (b^{\varepsilon})'(\bar{u}_{\varepsilon,\mu})[h] = 0$$

for all $h \in U$ in case I and

$$\frac{j'(\bar{u}_{\varepsilon,\mu})[h]}{\mu} + (b^{\varepsilon})'(\bar{u}_{\varepsilon,\mu})[h] + (\tilde{b}^{\varepsilon})'(\bar{u}_{\varepsilon,\mu})[h] = 0$$

for all $h \in U$ in case II. We treat case II first.

Case II

We use $h = \hat{u}$ with $\hat{u}$ from Assumption 6.1.9. Employing the boundedness of j' on bounded sets, see Assumption 3.1.9, we obtain

$$-\frac{(b^{\varepsilon})'(\bar{u}_{\varepsilon,\mu})[\hat{u}]}{\tau(\varepsilon)} \leq \frac{L\,\|\hat{u}\|_U}{\tau(\varepsilon)\mu} + \frac{\left\|(\tilde{b}^{\varepsilon})'(\bar{u}_{\varepsilon,\mu})\right\|_{U^*} \cdot \|\hat{u}\|_U}{\tau(\varepsilon)},$$

where $L > 0$ is independent of μ and ε. We use Corollary 6.3.3, $\tau(\varepsilon) = \frac{\tilde{C}}{\varepsilon^2}$ with a constant $\tilde{C} > 0$, and $\frac{\tilde{\tau}(\varepsilon)}{\tau(\varepsilon)} = \frac{1}{C_\tau}$ to infer that the right-hand side can be bounded by $C(\frac{\varepsilon^2}{\mu} + 1)$ with a μ- and ε-independent constant $C > 0$. With the chain rule and $\hat{y} = -A^{-1}B\hat{u}$ this yields

$$\frac{(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}]}{B^\varepsilon_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu})} = -\frac{(b^\varepsilon)'(\bar{u}_{\varepsilon,\mu})[\hat{u}]}{\tau(\varepsilon)} \leq C\left(1 + \frac{\varepsilon^2}{\mu}\right).$$

Multiplying this inequality with $\frac{B^\varepsilon_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu})}{C(1+\frac{\varepsilon^2}{\mu})}$ we see, thus, that we only need to derive a μ- and ε-independent positive lower bound for $(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}]$ to establish the assertion. (Note that $\frac{\varepsilon^2}{\mu} = \frac{C}{\varepsilon}$ with a suitable constant $C > 0$ and that $\vartheta(\varepsilon) = \frac{C}{\varepsilon^2}$ with another constant.) Before we do this, we show that in case I we have to prove the very same.

Case I
We use $h = \hat{u}$ with $\hat{u}$ from Assumption 6.1.9. Corollary 6.3.3 yields

$$-\frac{(b^\varepsilon)'(\bar{u}_{\varepsilon,\mu})[\hat{u}]}{\tau(\varepsilon)} \leq \frac{L\,\|\hat{u}\|_U}{\tau(\varepsilon)\mu},$$

where $L > 0$ is independent of μ and ε. Hence, we obtain

$$\frac{(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}]}{B^\varepsilon_{C(\overline{\Omega}_a)}(\bar{y}_{\varepsilon,\mu})} = -\frac{(b^\varepsilon)'(\bar{u}_{\varepsilon,\mu})[\hat{u}]}{\tau(\varepsilon)} \leq \frac{C}{\vartheta(\varepsilon)\mu}.$$

Thus, we only need to derive a μ- and ε-independent positive lower bound for $(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}]$ to establish the assertion.

We now demonstrate that $(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}]$ is bounded from below by a positive constant that is independent of μ and ε. To do so, we have to differ between Assumption 6.1.9 1) and 2).

Assumption 6.1.9 2) is satisfied
We use $\hat{y} \geq \hat{\varepsilon} > 0$ on Ω_a to infer

$$(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}] = \frac{\int_{\Omega_a} e^{-(\bar{y}_{\varepsilon,\mu}-y_a)/\varepsilon}\cdot\hat{y}\,\mathrm{d}x}{\int_{\Omega_a} e^{-(\bar{y}_{\varepsilon,\mu}-y_a)/\varepsilon}\,\mathrm{d}x} \geq \hat{\varepsilon}.$$

This concludes the proof in the case where Assumption 6.1.9 2) is satisfied.

Assumption 6.1.9 1) is satisfied
By assumption there exists at least one $x^* \in \overline{\Omega}_a$ such that $\bar{y}(x^*) = y_a(x^*)$ is satisfied. From the existence of the interior point y° and the fact that $y^\circ \equiv \bar{y} \equiv 0$ holds true on $\partial\Omega = \partial\Omega_a$, we infer that

$$\bar{y}(x) - y_a(x) = y^\circ(x) - y_a(x) \geq \tau^\circ > 0$$

is satisfied for all $x \in \partial\Omega$, see Assumption 3.1.9 6). From the Hölder continuity of $\bar{y} - y_a$ on the compact set $\overline{\Omega} = \overline{\Omega}_a$ we infer that with $\hat{\delta}$ from Assumption 6.1.9 1) we have

$$\bar{y}(x) - y_a(x) \geq \frac{3}{4}\,\tau^\circ \tag{6.2}$$

for all $x \in \Omega_{\hat{\delta}}$. Of course, $\hat{\delta}$ is independent of ε and μ. Using $\bar{y}_{\varepsilon,\mu} \to \bar{y}$ for $\mu \to 0^+$ with respect to $\|\cdot\|_{C(\overline{\Omega}_a)}$ we infer that we can choose $\hat{\mu} > 0$ such that it holds $\|\bar{y}_{\varepsilon,\mu} - \bar{y}\|_{C(\overline{\Omega}_a)} \leq \tau^\circ/4$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$. For all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \geq \hat{\mu}$ the assertion follows from Lemma 6.2.2. Thus, it only remains to establish the assertion for $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$. To this end, it suffices to prove that $(B^\varepsilon_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}]$ is bounded away from zero for all these ε and μ.

From $\|\bar{y}_{\varepsilon,\mu} - \bar{y}\|_{C(\overline{\Omega}_a)} \leq \tau^\circ/4$ for $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$ we deduce by virtue of (6.2) that

$$\bar{y}_{\varepsilon,\mu}(x) - y_a(x) \geq \frac{\tau^\circ}{2} \tag{6.3}$$

is valid for all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$ and all $x \in \Omega_{\hat{\delta}}$. Choose a $\delta > 0$ such that $B_\delta(x^*) \subset \Omega^C_{\hat{\delta}}$ holds true and such that

$$\bar{y}(x) - y_a(x) \leq \frac{\tau^\circ}{4} \tag{6.4}$$

is satisfied for all $x \in B_\delta(x^*)$. Such a choice is possible because $\bar{y} - y_a$ is continuous, $\bar{y}(x^*) = y_a(x^*)$ is satisfied, and (6.2) holds true, which says that x^* is an interior point of $\Omega^C_{\hat{\delta}}$. We stress that δ is independent of μ and ε. Again using $\|\bar{y}_{\varepsilon,\mu} - \bar{y}\|_{C(\overline{\Omega}_a)} \leq \tau^\circ/4$ for $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$ we infer from (6.4) that for all these (ε, μ) it holds

$$\bar{y}_{\varepsilon,\mu}(x) - y_a(x) \leq \frac{\tau^\circ}{2}$$

for all $x \in B_\delta(x^*)$. This yields for all these (ε, μ)

$$\int_{B_\delta(x^*)} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x \geq e^{-\tau^\circ/(2\varepsilon)} \cdot \mathrm{vol}(B_\delta(x^*)).$$

From (6.3) it follows for all these (ε, μ) that

$$\int_{\Omega_{\hat{\delta}}} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x \leq e^{-\tau^\circ/(2\varepsilon)} \cdot \mathrm{vol}(\Omega_{\hat{\delta}})$$

holds true. We set $\sigma := \frac{\mathrm{vol}(B_\delta(x^*))}{\mathrm{vol}(\Omega_{\hat{\delta}})}$ and note that $\sigma > 0$ is independent of μ and ε since this is true for δ and $\hat{\delta}$. So far we have proven that for all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$ we have

$$\int_{B_\delta(x^*)} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x \geq \sigma \cdot \int_{\Omega_{\hat{\delta}}} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x,$$

where $\sigma > 0$, $\delta > 0$, and $\hat{\delta} > 0$ are independent of μ and ε. Using $B_\delta(x^*) \subset \Omega^C_{\hat{\delta}}$ we are able to infer therefrom for all these (ε, μ)

$$\int_{\Omega^C_{\hat{\delta}}} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x \geq \sigma \cdot \int_{\Omega_{\hat{\delta}}} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x.$$

Adding $\sigma \int_{\Omega^C_{\hat{\delta}}} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x$ to both sides and dividing by $1 + \sigma$ we obtain

$$\int_{\Omega^C_{\hat{\delta}}} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x \geq \frac{\sigma}{1+\sigma} \cdot \int_{\Omega_a} e^{-(\bar{y}_{\varepsilon,\mu} - y_a)/\varepsilon} \,\mathrm{d}x$$

for all these (ε, μ). We use $\hat{y} \geq \hat{\varepsilon}$ on $\Omega_{\hat{\delta}}^C$ and $\hat{y} \geq 0$ on $\Omega = \Omega_a$ to, eventually, infer the desired positive lower bound for all $(\varepsilon, \mu) \in \mathcal{P}_=$ with $\mu \leq \hat{\mu}$:

$$(B^{\varepsilon}_{C(\overline{\Omega}_a)})'(\bar{y}_{\varepsilon,\mu})[\hat{y}] = \frac{\int_{\Omega_a} e^{-(\bar{y}_{\varepsilon,\mu}-y_a)/\varepsilon} \cdot \hat{y}\,\mathrm{d}x}{\int_{\Omega_a} e^{-(\bar{y}_{\varepsilon,\mu}-y_a)/\varepsilon}\,\mathrm{d}x} \geq \hat{\varepsilon} \cdot \frac{\int_{\Omega_{\hat{\delta}}^C} e^{-(\bar{y}_{\varepsilon,\mu}-y_a)/\varepsilon}\,\mathrm{d}x}{\int_{\Omega_a} e^{-(\bar{y}_{\varepsilon,\mu}-y_a)/\varepsilon}\,\mathrm{d}x} \geq \hat{\varepsilon} \cdot \frac{\sigma}{1+\sigma},$$

where $\sigma > 0$ and $\hat{\varepsilon} > 0$ are independent of μ and ε. This concludes the proof in the case where Assumption 6.1.9 1) is satisfied and, thus, also finishes the proof in total. □

Corollary 6.4.3. *There exists $C \in \mathbb{R}$ such that for all $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

$$b^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) \leq \tau(\varepsilon)\left(C - \ln\big(\mu\vartheta(\varepsilon)\big)\right).$$

Proof. Using the definition $b^{\varepsilon}(u) = -\tau(\varepsilon)\ln(B^{\varepsilon}(u))$ the assertion follows readily from the previous lemma. □

6.4.2. Step II: An estimate for $b^{\varepsilon(\mu)}$ on $\Lambda_{\varepsilon(\mu),\mu}$

The next lemma contains an estimate for $b^{\varepsilon(\mu)}(u)$, where u lies in the neighborhood $\Lambda_{\varepsilon(\mu),\mu}$ of $\bar{u}_{\varepsilon(\mu),\mu}$. It is derived from the above result for $b^{\varepsilon(\mu)}(\bar{u}_{\varepsilon(\mu),\mu})$.

Lemma 6.4.4. *There exists $C \in \mathbb{R}$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

$$b^{\varepsilon}(u) \leq \tau(\varepsilon)\left(C - \ln\big(\mu\vartheta(\varepsilon)\big)\right)$$

for all $u \in \Lambda_{\varepsilon,\mu}$.

Proof. Let $(\varepsilon, \mu) \in \mathcal{P}_=$ and let $u \in \Lambda_{\varepsilon,\mu}$. Then it follows from Lemma 2.2.23 that it holds

$$f_{\varepsilon,\mu}(u) - f_{\varepsilon,\mu}(\bar{u}_{\varepsilon,\mu}) \leq \frac{(\lambda_{\varepsilon,\mu}(u))^2}{1 - \left(\frac{16}{9}\lambda_{\varepsilon,\mu}(u)\right)^2} \leq \frac{1}{10}. \tag{6.5}$$

By definition we have $f_{\varepsilon,\mu}(u) = j(u)/\mu + b^{\varepsilon}(u)$ in case I and $f_{\varepsilon,\mu}(u) = j(u)/\mu + b^{\varepsilon}(u) + \tilde{b}^{\varepsilon}(u)$ in case II. We deal with case II first.

Case II:
We infer from (6.5) that it holds

$$\frac{b^{\varepsilon}(u)}{\tau(\varepsilon)} \leq \frac{1}{10} + \frac{|j(u) - j(\bar{u}_{\varepsilon,\mu})|}{\tau(\varepsilon)\mu} + \frac{b^{\varepsilon}(\bar{u}_{\varepsilon,\mu})}{\tau(\varepsilon)} + \frac{\tilde{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) - \tilde{b}^{\varepsilon}(u)}{\tau(\varepsilon)}, \tag{6.6}$$

where we used $\tau(\varepsilon) \geq 1$. We estimate the three unknown summands on the right-hand side.

- **An upper bound for $\frac{|j(u)-j(\bar{u}_{\varepsilon,\mu})|}{\tau(\varepsilon)\mu}$:**
  By Lemma 2.5.18 we have

$$\frac{|j(u)-j(\bar{u}_{\varepsilon,\mu})|}{\mu} \leq \frac{\lambda_{\varepsilon,\mu}(u)\sqrt{\vartheta(\varepsilon)}+\lambda_{\varepsilon,\mu}(u)^3}{\left(1-\frac{16}{9}\lambda_{\varepsilon,\mu}(u)\right)\cdot(1-\lambda_{\varepsilon,\mu}(u))} \leq C\sqrt{\frac{1}{\varepsilon^2}} = \frac{C}{\varepsilon},$$

  where $C>0$ is independent of ε, μ, and u. Here, we have used $\lambda_{\varepsilon,\mu}(u) \leq \frac{1}{4}$ and $\vartheta(\varepsilon) = \tau(\varepsilon)+\tilde{\tau}(\varepsilon) = (1+C_\tau)\tilde{\tau}(\varepsilon) = C/\varepsilon^2$ with a suitable C that is independent of ε, μ, and u. We infer

$$\frac{|j(u)-j(\bar{u}_{\varepsilon,\mu})|}{\tau(\varepsilon)\mu} \leq C\varepsilon \leq C\varepsilon_s,$$

  where $C>0$ is independent of ε, μ, and u.

- **An upper bound for $b^\varepsilon(\bar{u}_{\varepsilon,\mu})$:**
  We know from Corollary 6.4.3 that $b^\varepsilon(\bar{u}_{\varepsilon,\mu})/\tau(\varepsilon)$ is bounded from above by $C-\ln(\mu\vartheta(\varepsilon))$ with a constant $C\in\mathbb{R}$ that is independent of ε, μ, and u.

- **An upper bound for $\frac{\tilde{b}^\varepsilon(\bar{u}_{\varepsilon,\mu})-\tilde{b}^\varepsilon(u)}{\tau(\varepsilon)}$:**
  By Corollary 6.3.2 there exists an ε-, μ-, and u-independent constant $L>0$ with

$$\frac{\left|\tilde{b}^\varepsilon(\bar{u}_{\varepsilon,\mu})-\tilde{b}^\varepsilon(u)\right|}{\tau(\varepsilon)} \leq \frac{L\tilde{\tau}(\varepsilon)\,\|\bar{u}_{\varepsilon,\mu}-u\|_U}{\tau(\varepsilon)}.$$

  Since $\tilde{\tau}(\varepsilon)/\tau(\varepsilon) = C_\tau$ and since $U_{\mathrm{ad}}(\varepsilon)$ is bounded independently of ε, we obtain that $\frac{|\tilde{b}^\varepsilon(\bar{u}_{\varepsilon,\mu})-\tilde{b}^\varepsilon(u)|}{\tau(\varepsilon)}$ is bounded from above independently of ε, μ, and u.

With (6.6) these three estimates yield the existence of a constant $\tilde{C}>0$ that is independent of ε, μ, and u such that

$$b^\varepsilon(u) \leq \tau(\varepsilon)\left(\tilde{C}-\ln(\mu\vartheta(\varepsilon))\right)$$

is satisfied. This establishes the proof in case II.

Case I:
We deduce from (6.5) that it holds

$$\frac{b^\varepsilon(u)}{\tau(\varepsilon)} \leq \frac{1}{10} + \frac{|j(u)-j(\bar{u}_{\varepsilon,\mu})|}{\tau(\varepsilon)\mu} + \frac{b^\varepsilon(\bar{u}_{\varepsilon,\mu})}{\tau(\varepsilon)}.$$

We infer from Corollary 6.4.3 that $b^\varepsilon(\bar{u}_{\varepsilon,\mu})/\tau(\varepsilon)$ is bounded from above by $C-\ln(\mu\vartheta(\varepsilon))$ with a constant $C\in\mathbb{R}$ that is independent of ε, μ, and u. It remains to estimate $|j(u)-j(\bar{u}_{\varepsilon,\mu})|/(\tau(\varepsilon)\mu)$. We deal with this term analogously to case II and obtain

$$\frac{|j(u)-j(\bar{u}_{\varepsilon,\mu})|}{\mu} \leq C\sqrt{\tau(\varepsilon)},$$

which shows that $\frac{|j(u)-j(\bar{u}_{\varepsilon,\mu})|}{\tau(\varepsilon)\mu}$ is bounded from above by a constant that is independent of ε, μ, and u. This establishes the assertion in case I, thereby finishing the proof. □

The following result is crucial for large parts of the theory to come.

Corollary 6.4.5. *There exists $c > 0$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

$$B^\varepsilon(u) \geq c\mu\vartheta(\varepsilon)$$

for all $u \in \Lambda_{\varepsilon,\mu}$.

Proof. We use $B^\varepsilon(u) = e^{-\frac{b^\varepsilon(u)}{\tau(\varepsilon)}}$ to infer from Lemma 6.4.4 that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ we have

$$B^\varepsilon(u) \geq e^{-C}\mu\vartheta(\varepsilon)$$

for all $u \in \Lambda_{\varepsilon,\mu}$, which proves the assertion. □

As a consequence we obtain the following bound for $\|(b^\varepsilon)'(u)\|_{U^*}$.

Lemma 6.4.6. *There exists $C > 0$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_=$ it holds*

$$\left\|(b^\varepsilon)'(u)\right\|_{U^*} \leq \frac{C\tau(\varepsilon)}{\mu\vartheta(\varepsilon)}$$

for all $u \in \Lambda_{\varepsilon,\mu}$.

Proof. Due to Corollary 6.4.5 we have for every $(\varepsilon, \mu) \in \mathcal{P}_=$

$$\left\|(b^\varepsilon)'(u)\right\|_{U^*} = \frac{\tau(\varepsilon)\,\|(B^\varepsilon)'(u)\|_{U^*}}{B^\varepsilon(u)} \leq \frac{\tau(\varepsilon)\,\|(B^\varepsilon)'(u)\|_{U^*}}{c\mu\vartheta(\varepsilon)}$$

for all $u \in \Lambda_{\varepsilon,\mu}$. Thus, it remains to show that $\|(B^\varepsilon)'(u)\|_{U^*} \leq C$ is satisfied for all these ε, μ, and u. The chain rule implies for all $u \in U$ and all $h \in U$ that it holds

$$\begin{aligned}
|(B^\varepsilon)'(u)[h]| &= \left|(B^\varepsilon_{C(\overline{\Omega}_a)})'(y(u))[Th]\right| \\
&= \left|\frac{\int_{\Omega_a} e^{-(y(u)-y_a)/\varepsilon} \cdot Th\,\mathrm{d}x}{\int_{\Omega_a} e^{-(y(u)-y_a)/\varepsilon}\,\mathrm{d}x}\right| \leq \|Th\|_{C(\overline{\Omega}_a)} \leq C_{Y,C(\overline{\Omega}_a)}\,\|T\|_{\mathcal{L}(U,Y)}\,\|h\|_U\,,
\end{aligned}$$

where we used $T := -A^{-1}B \in \mathcal{L}(U,Y)$ and the embedding $Y \hookrightarrow C(\overline{\Omega}_a)$ with constant $C_{Y,C(\overline{\Omega}_a)}$. This concludes the proof. □

6.5. An estimate for a derivative of the smoothed minimum

Definition 6.5.1. Subsequently, we always denote by a the function

$$a : \mathbb{R}_{>0} \to \mathbb{R}_{>0}, \quad a(\varepsilon) := \varepsilon\,(1 + |\ln\varepsilon|)\,.$$

Lemma 6.5.2. *There exists $C > 0$ such that for every $\tilde{\varepsilon}, \bar{\varepsilon} \in (0, \varepsilon_s]$ with $\tilde{\varepsilon} \leq \bar{\varepsilon}$ it holds*

$$\frac{\partial B^{\tilde{\varepsilon}}(u)}{\partial\varepsilon} \leq \frac{1}{\tilde{\varepsilon}}\left(B^{\tilde{\varepsilon}}(u) + Ca(\bar{\varepsilon})\right)$$

for all $u \in U_{ad}(\tilde{\varepsilon})$.

Proof. From the definition it follows

$$\frac{\partial B^{\tilde{\varepsilon}}(u)}{\partial \varepsilon} = \frac{B^{\tilde{\varepsilon}}(u)}{\tilde{\varepsilon}} + \frac{1}{\tilde{\varepsilon}} \cdot \frac{\int_{\Omega_a} e^{-(y(u)-y_a)/\tilde{\varepsilon}} \cdot (y_a - y(u))\,\mathrm{d}x}{\int_{\Omega_a} e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}.$$

This implies

$$\begin{aligned}\frac{\partial B^{\tilde{\varepsilon}}(u)}{\partial \varepsilon} &\leq \frac{B^{\tilde{\varepsilon}}(u)}{\tilde{\varepsilon}} + \frac{1}{\tilde{\varepsilon}} \cdot \frac{\int_{\{y_a - y(u) \geq 0\}} e^{-(y(u)-y_a)/\tilde{\varepsilon}} \cdot (y_a - y(u))\,\mathrm{d}x}{\int_{\Omega_a} e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x} \\ &\leq \frac{B^{\tilde{\varepsilon}}(u)}{\tilde{\varepsilon}} + \frac{1}{\tilde{\varepsilon}} \cdot \left\|(y(u) - y_a)^-\right\|_{C(\overline{\Omega}_a)} \cdot \frac{\int_{\Omega_a} e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}{\int_{\Omega_a} e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}.\end{aligned}$$

The assertion follows from Corollary 4.4.4 since it holds $u \in U_{\mathrm{ad}}(\tilde{\varepsilon}) \subset U_{\mathrm{ad}}(\bar{\varepsilon})$. □

6.6. Lipschitz continuity of the first derivative of the barrier function

In this section we prove that for suitable μ and u the mapping $\varepsilon \mapsto f'_{\varepsilon,\mu}(u)$ is Lipschitz continuous on a certain interval. Note that the feasible set $U_{\mathrm{ad}}(\varepsilon)$ changes when ε changes. Since we want to prove Lipschitz continuity with respect to ε, we need to make sure that u stays away from the boundary of $U_{\mathrm{ad}}(\varepsilon)$ for all ε that we consider. We start with two definitions.

Definition 6.6.1. For $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$ and $\theta \in (0, \frac{1}{4}]$ we denote by $\Lambda^2_{\varepsilon,\mu}(\theta) \subset U$ the set

$$\Lambda^2_{\varepsilon,\mu}(\theta) := \left\{ u \in U_{\mathrm{ad}}(\varepsilon) : \lambda_{\varepsilon,\mu}(u) \leq \left(\frac{\theta}{1-\theta}\right)^2 \right\}.$$

Furthermore, we define $\Lambda^2_{\varepsilon,\mu} := \Lambda^2_{\varepsilon,\mu}(\frac{1}{4})$.

Remark 6.6.2. We have $\Lambda^2_{\varepsilon,\mu}(\theta) \subset \Lambda_{\varepsilon,\mu}(\theta) \subset U_{\mathrm{ad}}(\varepsilon)$ for all $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$ and all $\theta \in (0, \frac{1}{4}]$.

Remark 6.6.3. Note that $\Lambda^2_{\varepsilon_k,\mu_k}(\theta)$ is the set to which $u^{k+1} = u^k + n_{u^k}$ belongs if $u^k \in \Lambda_{\varepsilon_k,\mu_k}(\theta)$ and n_{u^k} is the Newton step for f_{ε_k,μ_k} at u^k, where $(\varepsilon_k, \mu_k) \in \mathcal{P}_{\leq}$. This indicates why the neighborhoods $\Lambda^2_{\varepsilon_k,\mu_k}(\theta)$ are important for short step methods.

Definition 6.6.4. For every $(\varepsilon, \mu) \in \mathcal{P}_{=}$ let a set $M_{\varepsilon,\mu} \subset U$ be given. Then we denote

$$T(M_{\varepsilon,\mu}) := \{(\varepsilon, \mu, u) \in \mathcal{P}_{=} \times U : u \in M_{\varepsilon,\mu}\}.$$

Via the following assumption we ensure for suitable u that the distance from u to the boundary of $U_{\mathrm{ad}}(\varepsilon)$ is sufficiently large for all ε from some interval.

Assumption 6.6.5. *Let a family $(M_{\varepsilon,\mu})_{(\varepsilon,\mu)\in\mathcal{P}_{=}}$ with $M_{\varepsilon,\mu} \subset U$ for all $(\varepsilon, \mu) \in \mathcal{P}_{=}$ be given. Furthermore, let $(I(\varepsilon, \mu, u))_{(\varepsilon,\mu,u)\in T(M_{\varepsilon,\mu})}$ be a family of compact intervals. Suppose that there exists a constant $c > 0$ with the following property:*

For every $(\varepsilon,\mu,u)\in T\left(M_{\varepsilon,\mu}\right)$ *the interval* $I\left(\varepsilon,\mu,u\right)$ *satisfies* $\{\varepsilon\}\subsetneq I\left(\varepsilon,\mu,u\right)\subset[\frac{\varepsilon}{2},\varepsilon]$*, and for every* $(\varepsilon,\mu,u)\in T\left(M_{\varepsilon,\mu}\right)$ *we have*

$$B^{\tilde{\varepsilon}}(u)\geq c\mu\vartheta(\varepsilon)\qquad\text{and}\qquad\tilde{B}(u)\geq c$$

for all $\tilde{\varepsilon}\in I\left(\varepsilon,\mu,u\right)$*.*

We now want to prove that the family $(\Lambda^2_{\varepsilon,\mu}(\theta))_{(\varepsilon,\mu)\in\mathcal{P}_=}$ and a suitably chosen family of intervals fulfill Assumption 6.6.5. To this end, we first establish Assumption 6.6.5 for $(\Lambda_{\varepsilon,\mu})$ and suitable intervals.

Lemma 6.6.6. *There exists a constant* $c>0$ *with the following property:*
For every $(\varepsilon,\mu,u)\in T(\Lambda_{\varepsilon,\mu})$ *the interval* $I\left(\varepsilon,\mu,u\right):=[\varepsilon(1-\frac{c\mu\vartheta(\varepsilon)}{a(\varepsilon)}),\varepsilon]$ *satisfies* $\{\varepsilon\}\subsetneq I\left(\varepsilon,\mu,u\right)\subset[\frac{\varepsilon}{2},\varepsilon]$ *and there hold*

$$B^{\tilde{\varepsilon}}(u)\geq c\mu\vartheta(\varepsilon)\qquad\text{and}\qquad\tilde{B}(u)\geq c$$

for all $\tilde{\varepsilon}\in I\left(\varepsilon,\mu,u\right)$*. This is, with* $M_{\varepsilon,\mu}:=\Lambda_{\varepsilon,\mu}$ *for* $(\varepsilon,\mu)\in\mathcal{P}_=$ *and the compact intervals* $I\left(\varepsilon,\mu,u\right)$ *defined above for* $(\varepsilon,\mu,u)\in T(\Lambda_{\varepsilon,\mu})$*, Assumption 6.6.5 is fulfilled. Moreover, all these assertions stay true if* c *is replaced by a positive constant smaller than* c*.*

Proof. It suffices to prove the existence of c as described in the lemma. To this end, fix $(\varepsilon,\mu,u)\in T(\Lambda_{\varepsilon,\mu})$. From Corollary 6.4.5 and Lemma 6.3.1 we deduce the existence of a constant $\tilde{c}>0$ that is independent of u, ε, and μ such that the inequalities

$$B^{\varepsilon}(u)\geq\tilde{c}\mu\vartheta(\varepsilon)\quad\text{ and }\quad\tilde{B}(u)\geq\tilde{c}\tag{6.7}$$

are satisfied. Obviously, $\tilde{B}(u)$ is constant on $I\left(\varepsilon,\mu,u\right)$. Summarizing, it is sufficient to prove the existence of c with $B^{\varepsilon}(u)\geq c\mu\vartheta(\varepsilon)$ for all $\tilde{\varepsilon}\in[\varepsilon(1-\frac{c\mu\vartheta(\varepsilon)}{a(\varepsilon)}),\varepsilon]\subset[\frac{\varepsilon}{2},\varepsilon]$, where c is independent of u, $\tilde{\varepsilon}$, ε, and μ. In the remainder of the proof we establish this.

By definition of $\mathcal{P}_=$, $\vartheta(\varepsilon)$, and $a(\varepsilon)$ we have $\frac{\mu\vartheta(\varepsilon)}{a(\varepsilon)}\leq\tilde{C}$ for all $(\varepsilon,\mu)\in\mathcal{P}_=$ and a suitable $\tilde{C}>0$ that is independent of u, $\tilde{\varepsilon}$, ε, and μ. Moreover, it holds

$$\left\|(y(u)-y_a)^-\right\|_{C(\overline{\Omega}_a)}\leq\hat{C}a(\varepsilon),$$

see Corollary 4.4.4. To apply this corollary note that $u\in U_{\mathrm{ad}}(\varepsilon)\subset M(\varepsilon)$ is valid due to (6.7). Here, $\hat{C}$ is independent of u, $\tilde{\varepsilon}$, ε, and μ. Thus, we obtain

$$\frac{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\cdot(y_a-y(u))\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}\leq\left\|(y(u)-y_a)^-\right\|_{C(\overline{\Omega}_a)}\cdot\frac{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}\leq\hat{C}a(\varepsilon)\tag{6.8}$$

for all $\tilde{\varepsilon}\in(0,\varepsilon_s]$. Here and in the following, the domain of integration is always Ω_a. Thus, for all $\tilde{\varepsilon}\in(0,\varepsilon_s]$ we have

$$\tilde{c}\mu\vartheta(\varepsilon)+\frac{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\cdot(y_a-y(u))\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x}\leq\tilde{C}\tilde{c}a(\varepsilon)+\hat{C}a(\varepsilon)\leq Ca(\varepsilon),$$

where we set $C := \tilde{C}\tilde{c} + \hat{C}$. Defining $c := \min\{\tilde{c}/2, \tilde{c}/(4C), 1/(2\tilde{C})\}$ we obtain that c is independent of u, $\tilde{\varepsilon}$, ε, and μ since this is true for $\tilde{C}$, C, and $\tilde{c}$. This shows

$$\tilde{c}\mu\vartheta(\varepsilon) + \frac{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}} \cdot (y_a - y(u))\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\tilde{\varepsilon}}\,\mathrm{d}x} \leq Ca(\varepsilon) \tag{6.9}$$

for all $\tilde{\varepsilon} \in (0, \varepsilon_s]$. Let us define $I(\varepsilon, \mu, u) := [\varepsilon(1 - \frac{c\mu\vartheta(\varepsilon)}{a(\varepsilon)}), \varepsilon]$. Note that the choice of c implies, in particular, $I(\varepsilon, \mu, u) \subset [\varepsilon/2, \varepsilon]$. If there holds $B^{\tilde{\varepsilon}}(u) > \tilde{c}\mu\vartheta(\varepsilon)/2$ for all $\tilde{\varepsilon} \in I(\varepsilon, \mu, u)$, then the assertion is true with $\tilde{c}/2$, and thus with the constant c as defined above. Using the continuity of $\varepsilon \mapsto B^{\varepsilon}(u)$ as well as $B^{\varepsilon}(u) \geq \tilde{c}\mu\vartheta(\varepsilon)$, see (6.7), we may, therefore, assume that there exists $\bar{\varepsilon} \in I(\varepsilon, \mu, u)$ with $\bar{\varepsilon} < \varepsilon$ and

$$B^{\bar{\varepsilon}}(u) = \frac{\tilde{c}\mu\vartheta(\varepsilon)}{2} \qquad \text{and} \qquad B^{\tilde{\varepsilon}}(u) \geq \frac{\tilde{c}\mu\vartheta(\varepsilon)}{2} \quad \text{for all } \tilde{\varepsilon} \in [\bar{\varepsilon}, \varepsilon]. \tag{6.10}$$

Let $\varepsilon^* \in (\bar{\varepsilon}, \varepsilon]$ denote the smallest number greater than $\bar{\varepsilon}$ with $B^{\varepsilon^*}(u) = \tilde{c}\mu\vartheta(\varepsilon)$. We now prove that $\varepsilon^* - \bar{\varepsilon} \geq c\varepsilon\mu\vartheta(\varepsilon)/a(\varepsilon)$ is satisfied. Using the mean value theorem we obtain an $\varepsilon^\dagger \in [\bar{\varepsilon}, \varepsilon^*]$ such that it holds

$$\frac{\tilde{c}\mu\vartheta(\varepsilon)}{2} = B^{\varepsilon^*}(u) - B^{\bar{\varepsilon}}(u) = \frac{\partial B^{\varepsilon^\dagger}(u)}{\partial\varepsilon} \cdot (\varepsilon^* - \bar{\varepsilon}).$$

Employing the definition of $B^{\varepsilon}(u)$ we compute

$$\frac{\partial B^{\varepsilon^\dagger}(u)}{\partial\varepsilon} = \frac{1}{\varepsilon^\dagger}\left(B^{\varepsilon^\dagger}(u) - \frac{\int e^{-(y(u)-y_a)/\varepsilon^\dagger} \cdot (y(u) - y_a)\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\varepsilon^\dagger}\,\mathrm{d}x}\right).$$

This yields

$$\begin{aligned}
\frac{\tilde{c}\mu\vartheta(\varepsilon)}{2} = \frac{\partial B^{\varepsilon^\dagger}(u)}{\partial\varepsilon} \cdot (\varepsilon^* - \bar{\varepsilon}) &= \frac{1}{\varepsilon^\dagger}\left(B^{\varepsilon^\dagger}(u) - \frac{\int e^{-(y(u)-y_a)/\varepsilon^\dagger} \cdot (y(u) - y_a)\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\varepsilon^\dagger}\,\mathrm{d}x}\right) \cdot (\varepsilon^* - \bar{\varepsilon}) \\
&\leq \frac{1}{\varepsilon^\dagger}\left(\tilde{c}\mu\vartheta(\varepsilon) + \frac{\int e^{-(y(u)-y_a)/\varepsilon^\dagger} \cdot (y_a - y(u))\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\varepsilon^\dagger}\,\mathrm{d}x}\right) \cdot (\varepsilon^* - \bar{\varepsilon}) \\
&\leq \frac{2}{\varepsilon}\left(\tilde{c}\mu\vartheta(\varepsilon) + \frac{\int e^{-(y(u)-y_a)/\varepsilon^\dagger} \cdot (y_a - y(u))\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\varepsilon^\dagger}\,\mathrm{d}x}\right) \cdot (\varepsilon^* - \bar{\varepsilon}),
\end{aligned}$$

where we used $\varepsilon^\dagger \geq \varepsilon/2$ due to $I(\varepsilon, \mu, u) \subset [\varepsilon/2, \varepsilon]$ (note that the left-hand side is positive and, therefore, this is true for the term in large brackets, too). Using (6.9) we obtain

$$\frac{\tilde{c}\mu\vartheta(\varepsilon)}{2} \leq \frac{2}{\varepsilon}\left(\tilde{c}\mu\vartheta(\varepsilon) + \frac{\int e^{-(y(u)-y_a)/\varepsilon^\dagger} \cdot (y_a - y(u))\,\mathrm{d}x}{\int e^{-(y(u)-y_a)/\varepsilon^\dagger}\,\mathrm{d}x}\right) \cdot (\varepsilon^* - \bar{\varepsilon}) \leq \frac{2Ca(\varepsilon)}{\varepsilon} \cdot (\varepsilon^* - \bar{\varepsilon}).$$

From this we deduce

$$\varepsilon^* - \bar{\varepsilon} \geq \frac{\tilde{c}\varepsilon\mu\vartheta(\varepsilon)}{4Ca(\varepsilon)}.$$

Therefore, it holds $c\varepsilon\mu\vartheta(\varepsilon)/a(\varepsilon) \leq \varepsilon^* - \bar{\varepsilon} \leq \varepsilon - \bar{\varepsilon}$, which yields $\bar{\varepsilon} \leq \varepsilon(1 - c\mu\vartheta(\varepsilon)/a(\varepsilon))$. Using $\bar{\varepsilon} \in I(\varepsilon, \mu, u)$ we infer $\bar{\varepsilon} = \varepsilon(1 - c\mu\vartheta(\varepsilon)/a(\varepsilon))$. The assertion then follows from (6.10). □

Corollary 6.6.7. *Let $\theta \in (0, \frac{1}{4}]$. Then Assumption 6.6.5 is fulfilled for the families $(\Lambda^2_{\varepsilon,\mu}(\theta))$ and $(I(\varepsilon,\mu,u))$, the latter being defined as in the previous lemma.*

Proof. The assertion follows from $\Lambda^2_{\varepsilon,\mu}(\theta) \subset \Lambda_{\varepsilon,\mu}(\theta) \subset \Lambda_{\varepsilon,\mu}$ for all $(\varepsilon,\mu) \in \mathcal{P}_=$. □

The succeeding lemma gives precise information about the Lipschitz continuity of $\varepsilon \mapsto f'_{\varepsilon,\mu}(u)$. The proof demonstrates what the use of Assumption 6.6.5 is.

Lemma 6.6.8. *Let Assumption 6.6.5 be fulfilled for the families $(M_{\varepsilon,\mu})$ and $(I(\varepsilon,\mu,u))$. Then there exists $C > 0$ such that for every $(\varepsilon,\mu,u) \in T(M_{\varepsilon,\mu})$ it holds*

$$\left\| f'_{\varepsilon_1,\mu}(u) - f'_{\varepsilon_2,\mu}(u) \right\|_{U^*} \leq \frac{C}{\varepsilon^{p+2}} \cdot |\varepsilon_1 - \varepsilon_2|$$

for all $\varepsilon_1, \varepsilon_2 \in I(\varepsilon,\mu,u)$, with $p = 2$ in case I and $p = 3$ in case II.

Proof. In the following we only argue for case II. Case I can be treated analogously. Fix $(\varepsilon,\mu,u) \in T(M_{\varepsilon,\mu})$. We show that

$$\varphi : I(\varepsilon,\mu,u) \to U^*, \quad \varphi(t) := f'_{t,\mu}(u)$$

is Lipschitz continuous and that $\frac{C}{\varepsilon^5}$ is an upper bound for the Lipschitz constant, where the constant $C > 0$ is independent of μ, ε, and u. This proves the assertion.

Note first that φ is well-defined: We have $B^{\tilde{\varepsilon}}(u) \geq c\mu\vartheta(\varepsilon) > 0$ and $\tilde{B}(u) \geq c > 0$ for all $\tilde{\varepsilon} \in I(\varepsilon,\mu,u)$, hence there holds $u \in U_{\mathrm{ad}}(\tilde{\varepsilon})$ for all $\tilde{\varepsilon}$ from an open neighborhood of $I(\varepsilon,\mu,u)$. With Lemma C.3.1 we infer that $f'_{t,\mu}(u)$ exists for all t from this neighborhood, which shows the well-definition of φ on this neighborhood.

It suffices to demonstrate that φ is Fréchet differentiable in an open neighborhood of $I(\varepsilon,\mu,u)$ and that there exists $C > 0$ that is independent of μ, ε, and u, such that

$$\left\| \varphi'(t)[1] \right\|_{U^*} \leq \frac{C}{t^5} \tag{6.11}$$

is satisfied for all $t \in I(\varepsilon,\mu,u) \subset [\varepsilon/2, \varepsilon]$, cf. Lemma C.2.1. In the following, we use C, C_1, C_2, and C_3 for positive constants that are independent of μ, ε, and u. Also, we identify $\varphi'(t)[1]$ with $\varphi'(t)$, cf. also [Zei93, Corollary 4.12]. Fréchet differentiability of φ is not hard to see. It mainly follows from the product rule, cf. Lemma C.2.7, the chain rule, cf. Lemma C.2.9, Lemma C.2.20, and Corollary C.2.22. We spare the details and focus on estimating the Lipschitz constant. Lemma C.3.1 yields the identity

$$\varphi'(t) = -\frac{\partial}{\partial t}\left(\frac{\tau(t)\left(B^t\right)'(u)}{B^t(u)}\right) - \tilde{\tau}'(t)\frac{\tilde{B}'(u)}{\tilde{B}(u)} = \frac{C_1\left(B^t\right)'(u)}{t^3 B^t(u)} + \frac{C_2}{t^2}\frac{\partial}{\partial t}\left(-\frac{\left(B^t\right)'(u)}{B^t(u)}\right) + \frac{C_3\tilde{B}'(u)}{t^3\tilde{B}(u)}. \tag{6.12}$$

For the right summand we have $\frac{C_3}{t^3} \cdot \frac{\|\tilde{B}'(u)\|_{U^*}}{\tilde{B}(u)} \leq \frac{C\|\tilde{B}'(u)\|_{U^*}}{ct^3}$ for all $t \in I(\varepsilon,\mu,u)$. This is bounded from above by C/t^3 since $\tilde{B}'$ is bounded on bounded sets. The summand on the left is bounded in U^* by $C/(\mu\vartheta(\varepsilon)t^3) \leq C/t^4$, since $\|\left(B^t\right)'(u)\|_{U^*} \leq C$, cf. the proof of Lemma 6.4.6,

and since $B^t(u) \geq c\mu\vartheta(\varepsilon)$. It remains to estimate the summand in the middle. For this term we have

$$\frac{\partial}{\partial t}\left(-\frac{\left(B^t\right)'(u)}{B^t(u)}\right) = \frac{\left(B^t\right)'(u)}{\left(B^t(u)\right)^2} \cdot \frac{\partial B^t(u)}{\partial t} + \frac{1}{B^t(u)} \cdot \frac{-\partial\left(\left(B^t\right)'(u)\right)}{\partial t}. \tag{6.13}$$

We have $1/B^t(u) \leq 1/(c\mu\vartheta(\varepsilon))$ and $\|\left(B^t\right)'(u)\|_{U^*} \leq C$. Thus, to obtain an estimate for the right-hand side of (6.13) it remains to deduce upper bounds for the two terms $\frac{\partial B^t(u)}{\partial t}$ and $\|\frac{\partial((B^t)'(u))}{\partial t}\|_{U^*}$.

1) Due to $\frac{\partial B^t(u)}{\partial t} \geq 0$, cf. Corollary 4.1.9, we have

$$\left|\frac{\partial B^t(u)}{\partial t}\right| \cdot \frac{1}{B^t(u)} \leq \frac{1}{t} + \frac{Ca(t)}{\mu\vartheta(\varepsilon)t} \leq \frac{Ca(t)}{\mu\vartheta(\varepsilon)t},$$

where we used Lemma 6.5.2 and two different C.

2) We compute for all $h \in U$

$$\begin{aligned}&\frac{-\partial\left(\left(B^t\right)'(u)\right)}{\partial t}(h)\\ &= \frac{1}{t^2}\left(\frac{\int q(t,u)(y_a - y(u))Th\,\mathrm{d}x}{\int q(t,u)\,\mathrm{d}x} - \frac{\int q(t,u)Th\,\mathrm{d}x \int q(t,u)(y_a - y(u))\,\mathrm{d}x}{\left(\int q(t,u)\,\mathrm{d}x\right)^2}\right)\end{aligned}$$

where $T := -A^{-1}B$ and $q(t,u)(x) := e^{-(y(u)(x)-y_a(x))/t}$. Since $\|y(u) - y_a\|_{C(\overline{\Omega}_a)} \leq C$ and $\|Th\|_{C(\overline{\Omega}_a)} \leq C\|h\|_U$ are satisfied, this implies

$$\left\|\frac{\partial\left(\left(B^t\right)'(u)\right)}{\partial t}\right\|_{U^*} \leq \frac{C}{t^2}.$$

Summarizing, there exists $C > 0$ with

$$\begin{aligned}\left\|\frac{\partial}{\partial t}\left(-\frac{\left(B^t\right)'(u)}{B^t(u)}\right)\right\|_{U^*} &\leq \frac{\left\|\left(B^t\right)'(u)\right\|_{U^*}}{\left(B^t(u)\right)^2}\left|\frac{\partial B^t(u)}{\partial t}\right| + \frac{1}{B^t(u)}\left\|\frac{\partial\left(\left(B^t\right)'(u)\right)}{\partial t}\right\|_{U^*}\\ &\leq \frac{Ca(t)}{\mu^2\vartheta(\varepsilon)^2 t} + \frac{C}{\mu\vartheta(\varepsilon)t^2}.\end{aligned}$$

To obtain the first inequality we used (6.13). In conclusion, the summand in the middle of (6.12) can be bounded by

$$\frac{Ca(t)}{\mu^2\vartheta(\varepsilon)^2 t^3} + \frac{C}{\mu\vartheta(\varepsilon)t^4} \leq \frac{C(1+|\ln t|)}{t^4} + \frac{C}{t^5} \leq \frac{C}{t^5},$$

which concludes the proof. □

As a direct consequence of the previous lemma we note:

Corollary 6.6.9. *Define $(I(\varepsilon,\mu,u))$ as in Lemma 6.6.6. Then there exists $C>0$ such that for every $(\varepsilon,\mu,u)\in T(\Lambda^2_{\varepsilon,\mu})$ it holds*

$$\left\|f'_{\varepsilon_1,\mu}(u)-f'_{\varepsilon_2,\mu}(u)\right\|_{U^*}\leq\frac{C}{\varepsilon^{p+2}}\cdot|\varepsilon_1-\varepsilon_2|$$

for all $\varepsilon_1,\varepsilon_2\in I(\varepsilon,\mu,u)$, with $p=2$ in case I and $p=3$ in case II.

Proof. We apply Lemma 6.6.8 to $(\Lambda^2_{\varepsilon,\mu})$ and $(I(\varepsilon,\mu,u))$. This is possible, see Corollary 6.6.7. □

6.7. An estimate for $(\hat{b}^\varepsilon)''$

Definition 6.7.1. For $\varepsilon>0$ we define

$$\hat{b}^\varepsilon:U_{\mathrm{ad}}(\varepsilon)\to\mathbb{R},\quad \hat{b}^\varepsilon:=\begin{cases}b^\varepsilon: & \text{in case I,}\\ b^\varepsilon+\tilde{b}^\varepsilon: & \text{in case II.}\end{cases}$$

We estimate $\|(\hat{b}^{\varepsilon_+})''(u)\|_{\mathcal{L}(U,U^*)}$ uniformly with respect to $\varepsilon_+\in I(\varepsilon,\mu,u)$.

Lemma 6.7.2. *Let Assumption 6.6.5 be fulfilled for the families $(M_{\varepsilon,\mu})$ and $(I(\varepsilon,\mu,u))$. There exists $C>0$ such that for every $(\varepsilon,\mu,u)\in T(M_{\varepsilon,\mu})$ it holds*

$$\left\|(\hat{b}^{\varepsilon_+})''(u)\right\|_{\mathcal{L}(U,U^*)}\leq\frac{C}{\varepsilon^{p+1}}$$

for all $\varepsilon_+\in I(\varepsilon,\mu,u)$, with $p=2$ in case I and $p=3$ in case II.

Proof. We argue for case II since case I is simpler. Hence, we have

$$(\hat{b}^{\varepsilon_+})''(u)=(b^{\varepsilon_+})''(u)+(\tilde{b}^{\varepsilon_+})''(u).$$

We estimate the two summands on the right-hand side separately.

- For the first summand there holds for all $h_1,h_2\in U$

$$\frac{(b^{\varepsilon_+})''(u)[h_1,h_2]}{\tau(\varepsilon_+)}=-\frac{(B^{\varepsilon_+})''(u)[h_1,h_2]}{B^{\varepsilon_+}(u)}+\frac{(B^{\varepsilon_+})'(u)[h_1]\cdot(B^{\varepsilon_+})'(u)[h_2]}{\left(B^{\varepsilon_+}(u)\right)^2}.$$

  We have

$$\|(B^{\varepsilon_+})'(u)\|_{U^*}\leq C_{\partial,C(\overline{\Omega}_a)}.$$

  Moreover, it is easy to see that it holds

$$\|(B^{\varepsilon_+})''(u)\|_{\mathcal{L}(U,U^*)}\leq\frac{C^2_{\partial,C(\overline{\Omega}_a)}}{\varepsilon_+}\leq\frac{2C^2_{\partial,C(\overline{\Omega}_a)}}{\varepsilon}.$$

Also, we have $B^{\varepsilon_+}(u) \geq c\mu\vartheta(\varepsilon)$ due to Assumption 6.6.5. Using $\mu\vartheta(\varepsilon) = C\varepsilon$ and $\tau(\varepsilon_+) = C/(\varepsilon_+)^2 \leq 4C/\varepsilon^2$ these considerations yield for the first summand

$$\|(b^{\varepsilon_+})''(u)\|_{\mathcal{L}(U,U^*)} \leq C\left(\frac{1}{\varepsilon^3\mu\vartheta(\varepsilon)} + \frac{1}{(\varepsilon\mu\vartheta(\varepsilon))^2}\right) \leq \frac{C}{\varepsilon^4},$$

where we used $C > 0$ for different constants that are all independent of ε, ε_+, μ, and u.

- For the second summand we have for all $h_1, h_2 \in U$

$$\frac{(\tilde{b}^{\varepsilon_+})''(u)[h_1,h_2]}{\tilde{\tau}(\varepsilon_+)} = -\frac{\tilde{B}''(u)[h_1,h_2]}{\tilde{B}(u)} + \frac{\tilde{B}'(u)[h_1]\cdot\tilde{B}'(u)[h_2]}{\left(\tilde{B}(u)\right)^2}.$$

  Due to Assumption 6.6.5 we have $\tilde{B}(u) \geq c$. From this we easily infer

$$\left\|(\tilde{b}^{\varepsilon_+})''(u)\right\|_{\mathcal{L}(U,U^*)} \leq C\tilde{\tau}(\varepsilon_+) \leq \frac{C}{\varepsilon^2},$$

  where $C > 0$ is independent of ε, ε_+, μ, and u.

Together, these two estimates imply the assertion. □

6.8. Uniform Lipschitz continuity of the Newton decrement

Lemma 6.8.1. *Let $\varepsilon_1, \varepsilon_2 \in (0, \varepsilon_s]$ with $\varepsilon_1 < \varepsilon_2$ and define $N := [\varepsilon_1, \varepsilon_2]$. For a given $c > 0$ set $A(N,c) := \{u \in U : B^\varepsilon(u) \geq c$ for all $\varepsilon \in N$ and $\tilde{B}(u) \geq c\}$. Then for every $u \in A(N,c)$ the mapping*

$$F_u : N \to \mathbb{R}, \quad F_u(\varepsilon) := \left(\lambda_{\varepsilon,\rho^{-1}(\varepsilon)}(u)\right)^2$$

is well-defined and Lipschitz, and the Lipschitz constant is uniformly bounded for all $u \in A(N,c)$. In particular, for $\bar{\varepsilon} \in (0, \varepsilon_s]$ and $u \in U_{ad}(\bar{\varepsilon})$ there exists a neighborhood N of $\bar{\varepsilon}$ such that F_u is Lipschitz in N.

Proof. The second assertion follows from the first: Since $u \in U_{\mathrm{ad}}(\bar{\varepsilon})$ implies $B^{\bar{\varepsilon}}(u) > 0$ and $\tilde{B}(u) > 0$, there exists an interval $N := [\varepsilon_1, \varepsilon_2]$ with $\bar{\varepsilon} \in (\varepsilon_1, \varepsilon_2)$ such that $B^\varepsilon(u) \geq B^{\bar{\varepsilon}}(u)/2$ and $\tilde{B}(u) > 0$ are satisfied due to continuity. With $c := \min\{B^{\bar{\varepsilon}}(u)/2, \tilde{B}(u)\}$ we obtain $u \in A(N,c)$ and, thus, Lipschitz continuity of F_u on N follows from the first assertion. It remains to prove the first assertion.

Let us begin by explaining that for $u \in A(N,c)$ the mapping F_u is well-defined on N. In fact, $u \in A(N,c)$ implies $u \in U_{\mathrm{ad}}(\varepsilon)$ for all $\varepsilon \in N$. Since $f_{\varepsilon,\mu}$ is nondegenerate on $U_{\mathrm{ad}}(\varepsilon)$ for every $(\varepsilon,\mu) \in \mathcal{P}_{\leq}$, cf. Lemma 6.1.6, the Newton step and, hence, the Newton decrement for $f_{\varepsilon,\mu}$ at u is well-defined for every $(\varepsilon,\mu) \in \mathcal{P}_=$ with $\varepsilon \in N$. This shows the well-definition of F_u on N since $(\varepsilon, \rho^{-1}(\varepsilon)) \in \mathcal{P}_=$.

We now argue briefly that $\varepsilon \mapsto f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u) \in U^*$ and $\varepsilon \mapsto f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u) \in \mathcal{L}(U,U^*)$ are continuously differentiable in an open neighborhood of N. We demonstrate this for the first mapping in case I. The reasoning for case II and the second mapping are similar. We have

$$f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u) = \frac{C_j}{\varepsilon^2}\frac{\hat{j}'(u)}{\tilde{B}(u)} - \tau(\varepsilon)\frac{(B^\varepsilon)'(u)}{B^\varepsilon(u)}, \tag{6.14}$$

see Lemma C.3.1, where we used that in case I we have $\rho^{-1}(\varepsilon) = \varepsilon^2$. Since $\varepsilon \mapsto B^\varepsilon(u)$ and $\varepsilon \mapsto (B^\varepsilon)'(u)$ are continuously differentiable in an open neighborhood of N, as follows with the use of Lemma C.2.20 and Corollary C.2.22, we deduce that $\varepsilon \mapsto f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u)$ is continuously differentiable in an open neighborhood of N. We compute this derivative and obtain for all $h \in U$

$$\begin{aligned}\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u)\right)(h) = &-\frac{C_j}{\varepsilon^3}\frac{\hat{j}'(u)[h]}{\tilde{B}(u)} - \tau'(\varepsilon)\frac{(B^\varepsilon)'(u)[h]}{B^\varepsilon(u)} \\ &- \tau(\varepsilon)\frac{(B^\varepsilon)'(u)[h]\cdot\left(\ln\left(\frac{\int q(\varepsilon,u)\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}\right) + \frac{\int q(\varepsilon,u)(y(u)-y_a)\,\mathrm{d}x}{\varepsilon\int q(\varepsilon,u)\,\mathrm{d}x}\right)}{\left(B^\varepsilon(u)\right)^2} \\ &- \tau(\varepsilon)\frac{\frac{\int q(\varepsilon,u)(y(u)-y_a)Th\,\mathrm{d}x}{\int q(\varepsilon,u)\,\mathrm{d}x} - \frac{\int q(\varepsilon,u)Th\,\mathrm{d}x\int q(\varepsilon,u)(y(u)-y_a)\,\mathrm{d}x}{\left(\int q(\varepsilon,u)\,\mathrm{d}x\right)^2}}{\varepsilon^2 B^\varepsilon(u)},\end{aligned} \tag{6.15}$$

where $T := -A^{-1}B$, $q(\varepsilon,u)(x) := e^{-(y(u)(x)-y_a(x))/\varepsilon}$, and the domain of integration is always Ω_a. We now show that

$$\left|\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u)\right)(h)\right| \leq L\,\|h\|_U$$

is satisfied for all $\varepsilon \in N$, all $u \in A(N,c)$, and all $h \in U$, where L is independent of ε, u, and h. This implies via Lemma C.2.1 that for every $u \in A(N,c)$ the mapping $\varepsilon \mapsto f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u)$ is Lipschitz in N, where the Lipschitz constant is independent of u.

- For the first summand in (6.15) this is clear since $1/\varepsilon^3 \leq 1/\varepsilon_1^3$ on N, since $\tilde{B}(u) \geq c$ for $u \in A(N,c)$, and since $\hat{j}'(u)$ is bounded on bounded sets.
- For the second summand we note that τ' is bounded on N since it is continuous, that $B^\varepsilon(u) \geq c$ holds for all $u \in A(N,c)$ and all $\varepsilon \in N$, and that $|(B^\varepsilon)'(u)[h]| \leq C\|h\|_U$ with a C that is independent of ε, u, and h, as we have already shown in the proof of Lemma 6.4.6.
- For most parts of the third summand we can use the same arguments as before. It remains to argue that

  $$\left|\ln\left(\frac{\int q(\varepsilon,u)\,\mathrm{d}x}{\mathrm{vol}(\Omega_a)}\right) + \frac{\int q(\varepsilon,u)(y(u)-y_a)\,\mathrm{d}x}{\varepsilon\int q(\varepsilon,u)\,\mathrm{d}x}\right| \leq \frac{1}{\varepsilon}\left(|B^\varepsilon(u)| + \left|\frac{\int q(\varepsilon,u)(y(u)-y_a)\,\mathrm{d}x}{\int q(\varepsilon,u)\,\mathrm{d}x}\right|\right)$$

  is bounded from above independently of ε, u, and h. Using $\frac{1}{\varepsilon} \leq \frac{1}{\varepsilon_1}$ as well as $B^\varepsilon(u) = B^\varepsilon_{C(\overline{\Omega}_a)}(y(u)) \leq \max(y(u)-y_a) \leq \|y(u)\|_{C(\overline{\Omega}_a)} + \|y_a\|_{C(\overline{\Omega}_a)}$ together with the uniform boundedness of $Y_{\mathrm{ad}}(\varepsilon)$, cf. Corollary 4.2.3, this follows with $\frac{\int q(\varepsilon,u)(y(u)-y_a)\,\mathrm{d}x}{\int q(\varepsilon,u)\,\mathrm{d}x} \leq \|y(u)\|_{C(\overline{\Omega}_a)} + \|y_a\|_{C(\overline{\Omega}_a)}$.

- For the last summand no new arguments are required.

In conclusion, we have established in case I that for every $u \in A(N,c)$ the mapping $\varepsilon \mapsto f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u)$ is Lipschitz in N, where the Lipschitz constant is independent of u. Similar arguments can be used to show the same for $\varepsilon \mapsto f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u)$ and for both these mappings in case II.

Furthermore, for every $\varepsilon \in N$, $f_{\varepsilon,\rho^{-1}(\varepsilon)}$ is uniformly convex on $U_{\mathrm{ad}}(\varepsilon)$ with convexity modulus uniformly bounded away from zero, cf. Lemma 3.5.8 and Lemma 3.5.18. With Corollary C.2.29 this implies that the concatenation $\varepsilon \mapsto (f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u))^{-1}$ is Lipschitz in N for every $u \in A(N,c)$ with a Lipschitz constant that is independent of u. Moreover, Theorem C.1.4 shows that $\|(f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u))^{-1}\|_{\mathcal{L}(U^*,U)}$ is bounded from above independently of $\varepsilon \in N$ and $u \in A(N,c)$. With (6.14) it is easy to argue that this holds true for $\|f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u)\|_{U^*}$ as well. Together, it follows that for every $u \in A(N,c)$ the mapping $n_u(\varepsilon) := -(f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u))^{-1}(f'_{\varepsilon,\rho^{-1}(\varepsilon)}(u))$ is bounded from above independently of u and ε, and Lipschitz in N with a Lipschitz constant independent of u. Furthermore, with arguments as before it can be established that $\|f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u)\|_{\mathcal{L}(U,U^*)}$ is bounded from above independently of $\varepsilon \in N$ and $u \in A(N,c)$, cf. also Lemma C.3.1.

In conclusion, for every $u \in A(N,c)$ the mapping

$$\varepsilon \mapsto f''_{\varepsilon,\rho^{-1}(\varepsilon)}(u)(n_u(\varepsilon), n_u(\varepsilon)) = F_u(\varepsilon)$$

is Lipschitz in N with a Lipschitz constant independent of u. □

6.9. An estimate for the Newton decrement after an update of the barrier parameter

Definition 6.9.1. We use the notation $f_+(u) := f_{\varepsilon_+,\mu_+}(u)$ for $(\varepsilon_+,\mu_+) \in \mathcal{P}_{\leq}$ and $u \in U_{\mathrm{ad}}(\varepsilon_+)$.

Lemma 6.9.2. *Let $(\varepsilon,\mu),(\varepsilon_+,\mu_+) \in \mathcal{P}_{\leq}$ with $\varepsilon_+ \leq \varepsilon$ and $u \in U_{ad}(\varepsilon_+)$ be given. Let n_u and n_u^+ denote the Newton steps for $f_{\varepsilon,\mu}$ at u and f_+ at u, respectively. Then for all $h \in U$ it holds*

$$f''_+(u)[n_u^+,h] = \frac{\mu}{\mu_+}\cdot f''_{\varepsilon,\mu}(u)[n_u,h] + \frac{\mu}{\mu_+}\cdot\Big((\hat{b}^{\varepsilon})'(u)[h] - (\hat{b}^{\varepsilon_+})'(u)[h]\Big) + \left(\frac{\mu}{\mu_+}-1\right)\cdot(\hat{b}^{\varepsilon_+})'(u)[h].$$

Remark 6.9.3. Since $u \in U_{\mathrm{ad}}(\varepsilon_+)$ holds, $f_+(u)$, $\hat{b}^{\varepsilon_+}(u)$ and all derivatives thereof are well-defined. Moreover, $u \in U_{\mathrm{ad}}(\varepsilon_+)$ implies $u \in U_{\mathrm{ad}}(\varepsilon)$, cf. Corollary 4.1.10. This shows the well-definition of $f_{\varepsilon,\mu}(u)$ and all related quantities.

Proof. We use $f'_{\varepsilon,\mu}(u) = \frac{j'(u)}{\mu} + (\hat{b}^{\varepsilon})'(u)$ and Newton's equation for n_u to derive for all $h \in U$

$$\begin{aligned}\frac{\mu}{\mu_+}\cdot f''_{\varepsilon,\mu}(u)[n_u,h] = -\frac{\mu}{\mu_+}\cdot f'_{\varepsilon,\mu}(u)[h] &= -\frac{j'(u)[h]}{\mu_+} - \frac{\mu}{\mu_+}\cdot(\hat{b}^{\varepsilon})'(u)[h]\\ &= -f'_+(u)[h] + (\hat{b}^{\varepsilon_+})'(u)[h] - \frac{\mu}{\mu_+}\cdot(\hat{b}^{\varepsilon})'(u)[h].\end{aligned}$$

Using Newton's equation for n_u^+ the assertion follows. □

In the proof of the theorem on the behaviour of the Newton decrement with respect to changes in ε and μ, we switch from the norm $\|\cdot\|_U$ to the local norm, and vice versa. Therefore, the following two lemmas are helpful.

Lemma 6.9.4. *There exists $C > 0$ such that for all $(\varepsilon, \mu, u) \in T(\Lambda_{\varepsilon,\mu})$ and all $h \in U$ it holds*

$$\|h\|_{f''_{\varepsilon,\mu}(u)} \leq \frac{C}{\sqrt{\varepsilon\mu}} \|h\|_U .$$

Proof. We argue for case II since the other case is simpler. Using the formula for the second derivative from Lemma C.3.1 we can estimate

$$\begin{aligned} f''_{\varepsilon,\mu}(u)[h,h] &= \frac{j''(u)[h,h]}{\mu} - \tau(\varepsilon) \left(\frac{(B^\varepsilon)''(u)[h,h]}{B^\varepsilon(u)} - \left(\frac{(B^\varepsilon)'(u)[h]}{B^\varepsilon(u)} \right)^2 \right) \\ &\quad - \tilde{\tau}(\varepsilon) \cdot \left(\frac{\tilde{B}''(u)[h,h]}{\tilde{B}(u)} - \left(\frac{\tilde{B}'(u)[h]}{\tilde{B}(u)} \right)^2 \right) \\ &\leq C \|h\|_U^2 \left(\frac{1}{\mu} + \frac{1}{\varepsilon\mu} + \frac{1}{\mu^2\vartheta(\varepsilon)} + \frac{1}{\varepsilon^2} \right), \end{aligned}$$

where we employed $\tau(\varepsilon), \tilde{\tau}(\varepsilon) \leq \vartheta(\varepsilon)$, $B^\varepsilon(u) \geq c\mu\vartheta(\varepsilon)$ and $\tilde{B}(u) \geq c$ as well as estimates for the first and second derivatives of $B^\varepsilon(u)$ and $\tilde{B}(u)$ that we have used in the proof of Lemma 6.7.2, and the fact that $j = \hat{j}$ has uniformly bounded second derivatives on the bounded set $T(\Lambda_{\varepsilon,\mu})$. The constant $C > 0$ is independent of ε, μ, u, and h. Using $\mu\vartheta(\varepsilon) \geq C\varepsilon$ the above estimate implies

$$\|h\|^2_{f''_{\varepsilon,\mu}(u)} = f''_{\varepsilon,\mu}(u)[h,h] \leq \frac{C \|h\|_U^2}{\varepsilon\mu},$$

where $C > 0$ is still independent of ε, μ, u, and h. This establishes the assertion. □

Lemma 6.9.5. *There exists $C > 0$ such that for every $(\varepsilon, \mu) \in \mathcal{P}_{\leq}$ it holds*

$$\|h\|_U \leq C\sqrt{\mu} \|h\|_{f''_{\varepsilon,\mu}(u)}$$

for all $u \in U_{ad}(\varepsilon)$ and all $h \in U$.

Proof. Using the uniform convexity of $f_{\varepsilon,\mu}$ on $U_{\mathrm{ad}}(\varepsilon)$ with modulus $\beta(\varepsilon, \mu) \geq \alpha/\mu$, where α denotes the convexity modulus of j on $\cup_{\varepsilon\in(0,\varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$ we directly obtain

$$\|h\|^2_{f''_{\varepsilon,\mu}(u)} \geq \frac{\alpha \|h\|_U^2}{\mu}$$

for all $u \in U_{\mathrm{ad}}(\varepsilon)$ and all $h \in U$, where α is independent of ε, μ, u, and h. □

The next statement is the aforementioned theorem that allows us to estimate the Newton decrement after an update of ε and μ to ε_+ and μ_+. It is a key result to derive convergence rates for short step methods.

Theorem 6.9.6. *There exists $C > 0$ such that for every $(\varepsilon, \mu, u) \in T(\Lambda_{\varepsilon,\mu})$ it holds*

$$\lambda_{\varepsilon_+,\mu_+}(u) \leq C \left(\lambda_{\varepsilon,\mu}(u) \cdot \sqrt{\frac{1}{\varepsilon}} + \frac{1}{\varepsilon^2 \sqrt{\mu}} \cdot (\varepsilon - \varepsilon_+) \right)$$

for all $(\varepsilon_+, \mu_+) \in \mathcal{P}_=$ with $\varepsilon_+ \in I(\varepsilon, \mu, u)$, where $I(\varepsilon, \mu, u)$ is defined as in Lemma 6.6.6.

Proof. Inserting $h = n_u^+$ in Lemma 6.9.2 we obtain

$$\begin{aligned} \left\| n_u^+ \right\|_{f_+''(u)}^2 = {} & \frac{\mu}{\mu_+} \cdot f_{\varepsilon,\mu}''(u)[n_u, n_u^+] \\ & + \frac{\mu}{\mu_+} \cdot \left((\hat{b}^{\varepsilon})'(u)[n_u^+] - (\hat{b}^{\varepsilon_+})'(u)[n_u^+] \right) + \left(\frac{\mu}{\mu_+} - 1 \right) \cdot (\hat{b}^{\varepsilon_+})'(u)[n_u^+]. \end{aligned} \tag{6.16}$$

We estimate the three summands on the right-hand side of this equation separately.

- The first summand can be estimated via the Cauchy-Schwarz inequality for the local norm at u and the inequality from Lemma 6.9.4. This yields

  $$\frac{\mu}{\mu_+} \cdot f_{\varepsilon,\mu}''(u)[n_u, n_u^+] \leq \frac{\mu}{\mu_+} \cdot \|n_u\|_{f_{\varepsilon,\mu}''(u)} \cdot \left\| n_u^+ \right\|_{f_{\varepsilon,\mu}''(u)} \leq \frac{\mu}{\mu_+} \cdot \lambda_{\varepsilon,\mu}(u) \cdot \frac{C}{\sqrt{\varepsilon\mu}} \cdot \left\| n_u^+ \right\|_U,$$

  where here and in the following $C > 0$ denotes a constant that is independent of ε, ε_+, μ, μ_+, and u. Employing Lemma 6.9.5, which is possible due to $u \in U_{\mathrm{ad}}(\varepsilon_+)$, we obtain

  $$\frac{\mu}{\mu_+} \cdot f_{\varepsilon,\mu}''(u)[n_u, n_u^+] \leq C \cdot \lambda_{\varepsilon,\mu}(u) \cdot \sqrt{\frac{\mu}{\mu_+}} \cdot \sqrt{\frac{1}{\varepsilon}} \cdot \left\| n_u^+ \right\|_{f_+''(u)}.$$

  It follows from $\mu/\mu_+ = (\varepsilon/\varepsilon_+)^p$ (with $p = 2$ in case I and $p = 3$ in case II) and $\varepsilon_+ \geq \varepsilon/2$ that we have $\mu/\mu_+ \leq 2^p$. Hence, there exists $C > 0$ that is independent of ε, ε_+, μ, μ_+, and u with

  $$\frac{\mu}{\mu_+} \cdot f_{\varepsilon,\mu}''(u)[n_u, n_u^+] \leq C \cdot \lambda_{\varepsilon,\mu}(u) \cdot \sqrt{\frac{1}{\varepsilon}} \cdot \left\| n_u^+ \right\|_{f_+''(u)}.$$

- For the second summand we obtain $C > 0$ such that it holds

  $$\frac{\mu}{\mu_+} \cdot \left((\hat{b}^{\varepsilon})'(u)[n_u^+] - (\hat{b}^{\varepsilon_+})'(u)[n_u^+] \right) \leq \frac{C}{\varepsilon^{p+2}} \cdot (\varepsilon - \varepsilon_+) \cdot \left\| n_u^+ \right\|_U.$$

  Here, we used Lemma 6.6.8. Applying Lemma 6.9.5 and $\mu = C\varepsilon^p$ we deduce

  $$\frac{\mu}{\mu_+} \cdot \left((\hat{b}^{\varepsilon})'(u)[n_u^+] - (\hat{b}^{\varepsilon_+})'(u)[n_u^+] \right) \leq \frac{C}{\varepsilon^2 \sqrt{\mu}} \cdot (\varepsilon - \varepsilon_+) \cdot \left\| n_u^+ \right\|_{f_+''(u)}.$$

- For the last summand we use the self-boundedness of $\hat{b}^{\varepsilon_+}$ and $\varepsilon/\varepsilon_+ \leq 2$ to infer

  $$\begin{aligned} \left(\frac{\mu}{\mu_+} - 1 \right) \cdot (\hat{b}^{\varepsilon_+})'(u)[n_u^+] &\leq \left(\frac{\mu}{\mu_+} - 1 \right) \cdot \sqrt{\vartheta(\varepsilon_+)} \cdot \sqrt{(\hat{b}^{\varepsilon_+})''(u)[n_u^+, n_u^+]} \\ &\leq \left(\frac{\mu}{\mu_+} - 1 \right) \cdot \sqrt{\vartheta(\varepsilon_+)} \cdot \sqrt{f_+''(u)[n_u^+, n_u^+]} \\ &\leq \left(\frac{\mu}{\mu_+} - 1 \right) \cdot C \cdot \sqrt{\vartheta(\varepsilon)} \cdot \left\| n_u^+ \right\|_{f_+''(u)}. \end{aligned}$$

Together, these considerations imply via (6.16) the estimate

$$\left\|n_u^+\right\|_{f_+''(u)}^2 \leq C \left\|n_u^+\right\|_{f_+''(u)} \left(\lambda_{\varepsilon,\mu}(u) \cdot \sqrt{\frac{1}{\varepsilon}} + \frac{1}{\varepsilon^2\sqrt{\mu}} \cdot (\varepsilon - \varepsilon_+) + \left(\frac{\mu}{\mu_+} - 1\right) \cdot \sqrt{\vartheta(\varepsilon)}\right).$$

Since we have $\frac{\mu}{\mu_+} - 1 = |1 - \frac{\varepsilon^p}{\varepsilon_+^p}| = 2^p \frac{\varepsilon^p - \varepsilon_+^p}{\varepsilon^p}$, it is elementary to see that $\frac{\mu}{\mu_+} - 1 \leq C\frac{\varepsilon - \varepsilon_+}{\varepsilon}$ is satisfied. This implies

$$\left\|n_u^+\right\|_{f_+''(u)}^2 \leq C \left\|n_u^+\right\|_{f_+''(u)} \left(\lambda_{\varepsilon,\mu}(u) \cdot \sqrt{\frac{1}{\varepsilon}} + \frac{1}{\varepsilon^2\sqrt{\mu}} \cdot (\varepsilon - \varepsilon_+)\right).$$

Since by definition it holds $\|n_u^+\|_{f_+''(u)} = \lambda_{\varepsilon_+,\mu_+}(u)$, the assertion follows. □

6.10. Estimates on function values

Lemma 6.10.1. *There exists $C > 0$ such that for every $(\varepsilon, \mu, u) \in T(\Lambda_{\varepsilon,\mu})$ it holds*

$$f_+(u) - f_+(\bar{u}_{\varepsilon,\mu}) \leq C\lambda_{\varepsilon,\mu}(u)^2 \left(\lambda_{\varepsilon,\mu}(u) + \sqrt{\frac{\vartheta(\varepsilon)}{\varepsilon}}\right)$$

for all $(\varepsilon_+, \mu_+) \in \mathcal{P}_=$ with $\varepsilon_+ \in I(\varepsilon, \mu, u)$, where $I(\varepsilon, \mu, u)$ is defined as in Lemma 6.6.6.

Remark 6.10.2. $f_+(\bar{u}_{\varepsilon,\mu})$ is well-defined since it holds $\bar{u}_{\varepsilon,\mu} \in U_{\mathrm{ad}}(\varepsilon_+)$ due to $\varepsilon_+ \in I(\varepsilon, \mu, u) = I(\varepsilon, \mu, \bar{u}_{\varepsilon,\mu})$, cf. Lemma 6.6.6.

Proof. We denote by n_u and n_u^+ the Newton steps for $f_{\varepsilon,\mu}$ at u and f_+ at u, respectively. Moreover, we set $h := u - \bar{u}_{\varepsilon,\mu}$. Since f_+ is convex, we have

$$\begin{aligned} f_+(u) - f_+(\bar{u}_{\varepsilon,\mu}) &\leq f_+'(u)[h] = -f_+''(u)[n_u^+, h] \\ &= -\frac{\mu}{\mu_+} \cdot f_{\varepsilon,\mu}''(u)[n_u, h] - \frac{\mu}{\mu_+} \cdot (\hat{b}^\varepsilon)'(u)[h] + (\hat{b}^{\varepsilon_+})'(u)[h]. \end{aligned} \tag{6.17}$$

For the second equality we employed Lemma 6.9.2. Using the Cauchy-Schwarz inequality for $\|\cdot\|_{f_{\varepsilon,\mu}''(u)}$ and the self-boundedness of b^ε and b^{ε_+} we infer from (6.17) that it holds

$$\begin{aligned} f_+(u) - f_+(\bar{u}_{\varepsilon,\mu}) \leq{}& \frac{\mu}{\mu_+} \cdot \|n_u\|_{f_{\varepsilon,\mu}''(u)} \|h\|_{f_{\varepsilon,\mu}''(u)} + \frac{\mu}{\mu_+} \cdot \sqrt{\vartheta(\varepsilon)} \cdot \sqrt{(\hat{b}^\varepsilon)''(u)[h,h]} \\ &+ \sqrt{\vartheta(\varepsilon_+)} \cdot \sqrt{(\hat{b}^{\varepsilon_+})''(u)[h,h]}. \end{aligned} \tag{6.18}$$

We estimate the three summands on the right-hand side of (6.18) separately.

- For the first summand we have by definition $\|n_u\|_{f''_{\varepsilon,\mu}(u)} = \lambda_{\varepsilon,\mu}(u)$. From Lemma 2.2.23 we deduce $\|h\|_{f''_{\varepsilon,\mu}(u)} = \|u - \bar{u}_{\varepsilon,\mu}\|_{f''_{\varepsilon,\mu}(u)} \leq 10\,(\lambda_{\varepsilon,\mu}(u))^2$. We have noted in other proofs that $\mu/\mu_+ \leq C$, where here and in the following $C > 0$ denotes constants that are independent of ε, ε_+, μ, μ_+, h, and u. Together, we have

$$\frac{\mu}{\mu_+} \cdot \|n_u\|_{f''_{\varepsilon,\mu}(u)} \|h\|_{f''_{\varepsilon,\mu}(u)} \leq C\lambda_{\varepsilon,\mu}(u)^3.$$

- For the second summand we can estimate

$$\frac{\mu}{\mu_+} \cdot \sqrt{\vartheta(\varepsilon)} \cdot \sqrt{(\hat{b}^{\varepsilon})''(u)[h,h]} \leq C \cdot \sqrt{\vartheta(\varepsilon)} \cdot \sqrt{f''_{\varepsilon,\mu}(u)[h,h]} \leq 10C \cdot \sqrt{\vartheta(\varepsilon)} \cdot \lambda_{\varepsilon,\mu}(u)^2.$$

- For the third summand we apply Lemma 6.7.2 and Lemma 6.9.5. Since we have $\mu = C\varepsilon^p$ with $p = 2$ in case I and $p = 3$ in case II, this yields

$$\sqrt{(\hat{b}^{\varepsilon_+})''(u)[h,h]} \leq \sqrt{\frac{C}{\varepsilon^{p+1}}} \cdot \|h\|_U \leq \frac{C}{\sqrt{\varepsilon^{p+1}}} \cdot \sqrt{\mu} \cdot (\lambda_{\varepsilon,\mu}(u))^2 \leq \frac{C\lambda_{\varepsilon,\mu}(u)^2}{\sqrt{\varepsilon}}.$$

  Moreover, we have $\sqrt{\vartheta(\varepsilon_+)} \leq C\sqrt{\vartheta(\varepsilon)}$. Putting these inequalities together we obtain for the third summand the upper bound

$$\sqrt{\vartheta(\varepsilon_+)} \cdot \sqrt{(\hat{b}^{\varepsilon_+})''(u)[h,h]} \leq \frac{C\sqrt{\vartheta(\varepsilon)}\lambda_{\varepsilon,\mu}(u)^2}{\sqrt{\varepsilon}}.$$

Together with (6.18) the estimates for the three summands imply

$$f_+(u) - f_+(\bar{u}_{\varepsilon,\mu}) \leq C\lambda_{\varepsilon,\mu}(u)^2 \left(\lambda_{\varepsilon,\mu}(u) + \sqrt{\frac{\vartheta(\varepsilon)}{\varepsilon}}\right). \qquad \square$$

Lemma 6.10.3. *There exists $C > 0$ such that for every $(\varepsilon,\mu),(\varepsilon_+,\mu_+) \in \mathcal{P}_=$ with $\varepsilon_+ \in I\,(\varepsilon,\mu,\bar{u}_{\varepsilon,\mu})$ it holds*

$$f_+(\bar{u}_{\varepsilon,\mu}) - f_+(\bar{u}_{\varepsilon_+,\mu_+}) \leq \frac{Ca(\varepsilon)\,|\ln(c\mu\vartheta(\varepsilon))|}{\mu\varepsilon}\,|\varepsilon - \varepsilon_+|\,,$$

where $I\,(\varepsilon,\mu,\bar{u}_{\varepsilon,\mu})$ and c are defined as in Lemma 6.6.6.

Proof. Using the definition of f_+ we have

$$f_+(\bar{u}_{\varepsilon,\mu}) - f_+(\bar{u}_{\varepsilon_+,\mu_+}) = \frac{j(\bar{u}_{\varepsilon,\mu}) - j(\bar{u}_{\varepsilon_+,\mu_+})}{\mu_+} + \hat{b}^{\varepsilon_+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon_+}(\bar{u}_{\varepsilon_+,\mu_+}). \tag{6.19}$$

We estimate the first summand by using the convexity of j, the optimality of $\bar{u}_{\varepsilon,\mu}$, i.e., $f'_{\varepsilon,\mu}(\bar{u}_{\varepsilon,\mu}) = 0$, and the convexity of $\hat{b}^{\varepsilon}$. This yields

$$\begin{aligned}\frac{j(\bar{u}_{\varepsilon,\mu}) - j(\bar{u}_{\varepsilon_+,\mu_+})}{\mu_+} &= \frac{\mu}{\mu_+} \cdot \frac{j(\bar{u}_{\varepsilon,\mu}) - j(\bar{u}_{\varepsilon_+,\mu_+})}{\mu} \leq \frac{\mu}{\mu_+} \cdot \frac{j'(\bar{u}_{\varepsilon,\mu})[\bar{u}_{\varepsilon,\mu} - \bar{u}_{\varepsilon_+,\mu_+}]}{\mu} \\ &= \frac{\mu}{\mu_+} \cdot (\hat{b}^{\varepsilon})'(\bar{u}_{\varepsilon,\mu})[\bar{u}_{\varepsilon_+,\mu_+} - \bar{u}_{\varepsilon,\mu}] \leq \frac{\mu}{\mu_+} \cdot \left(\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu})\right).\end{aligned}$$

From this inequality we infer with (6.19) that it holds

$$\begin{aligned} f_+(\bar{u}_{\varepsilon,\mu}) - f_+(\bar{u}_{\varepsilon_+,\mu_+}) &\leq \hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon_+,\mu_+}) + \frac{\mu}{\mu_+}\left(\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu})\right) \\ &= \left(\hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu})\right) + \left(\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+}) - \hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon_+,\mu_+})\right) \\ &\quad + \left(1 - \frac{\mu}{\mu_+}\right) \cdot \left(\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+})\right). \end{aligned} \tag{6.20}$$

We derive an upper bound for the sum on the right-hand side by estimating its three summands separately. To this end, we set $I := [\varepsilon_+, \varepsilon]$. To make clear that derivatives with respect to ε are not applied to $\bar{u}_{\varepsilon,\mu}$ and $\bar{u}_{\varepsilon_+,\mu_+}$, respectively, we set $u := \bar{u}_{\varepsilon,\mu}$ and $u_+ := \bar{u}_{\varepsilon_+,\mu_+}$.

- We start with the term $\hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu})$. Using the mean value theorem we obtain $\tilde{\varepsilon} \in I$ with

  $$\hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) = \hat{b}^{\varepsilon+}(u) - \hat{b}^{\varepsilon}(u) = \frac{\partial \hat{b}^{\tilde{\varepsilon}}(u)}{\partial \varepsilon} \cdot (\varepsilon_+ - \varepsilon).$$

  We first deal with case I, i.e., $\hat{b}^{\tilde{\varepsilon}} = b^{\tilde{\varepsilon}} = -\tau(\tilde{\varepsilon}) \ln(B^{\tilde{\varepsilon}})$. Due to Lemma 6.6.6 we have $B^{\tilde{\varepsilon}}(u) \geq c\mu\vartheta(\varepsilon)$, where $c > 0$ is independent of ε, $\tilde{\varepsilon}$, ε_+, μ, μ_+, $\bar{u}_{\varepsilon,\mu}$, and $\bar{u}_{\varepsilon_+,\mu_+}$. Moreover, from Lemma 6.5.2 we deduce

  $$\begin{aligned} \frac{\partial \hat{b}^{\tilde{\varepsilon}}(u)}{\partial \varepsilon} \cdot (\varepsilon_+ - \varepsilon) &= \left(\tau'(\tilde{\varepsilon}) \ln\big(B^{\tilde{\varepsilon}}(u)\big) + \frac{\tau(\tilde{\varepsilon})}{B^{\tilde{\varepsilon}}(u)} \cdot \frac{\partial B^{\tilde{\varepsilon}}(u)}{\partial \varepsilon}\right) \cdot (\varepsilon - \varepsilon_+) \\ &\leq \left(C\left|\tau'(\tilde{\varepsilon})\right| \left|\ln\big(c\mu\vartheta(\varepsilon)\big)\right| + \frac{C\tau(\varepsilon)}{\varepsilon} + \frac{C\tau(\varepsilon)a(\varepsilon)}{c\mu\vartheta(\varepsilon)\varepsilon}\right) \cdot (\varepsilon - \varepsilon_+), \end{aligned}$$

  where we used that $\tau(\tilde{\varepsilon}) \leq C\tau(\varepsilon)$ and $\tilde{\varepsilon} \geq \varepsilon/2$ hold as well as $B^{\tilde{\varepsilon}}(u) \leq \max(y(u) - y_a) \leq C$ and $c\mu\vartheta(\varepsilon) \leq \frac{1}{2}$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$ without loss of generality. Here and in the following, $C > 0$ denotes different constants that are all independent of ε, $\tilde{\varepsilon}$, ε_+, μ, μ_+, $\bar{u}_{\varepsilon,\mu}$, and $\bar{u}_{\varepsilon_+,\mu_+}$. It follows from elementary calculus that $|\tau'(\tilde{\varepsilon})| \leq |\frac{C\tau(\tilde{\varepsilon})}{\tilde{\varepsilon}}| \leq \frac{C\tau(\varepsilon)}{\varepsilon}$. Employing $\tau(\varepsilon) \leq \frac{Ca(\varepsilon)}{\mu}$ and $\frac{\tau(\varepsilon)}{\vartheta(\varepsilon)} = 1$ we have

  $$\hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) \leq \frac{Ca(\varepsilon)\left|\ln\big(c\mu\vartheta(\varepsilon)\big)\right|}{\mu\varepsilon} \left|\varepsilon - \varepsilon_+\right|.$$

  In case II we have $\hat{b}^{\varepsilon} = b^{\varepsilon} + \tilde{b}^{\varepsilon}$. For b^{ε} we can argue exactly as in case I and obtain the same bound for the derivative with respect to ε. For $\tilde{b}^{\varepsilon}$ it holds $\tilde{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) = -\tilde{\tau}(\varepsilon) \ln(\tilde{B}(\bar{u}_{\varepsilon,\mu}))$ and $\tilde{B}(\bar{u}_{\varepsilon,\mu}) \geq c$. With $|\tilde{\tau}'(\varepsilon)| = C/\varepsilon^3 = C/\mu$ this yields a smaller bound for $\frac{\partial \tilde{b}^{\tilde{\varepsilon}}}{\partial \varepsilon}$ than for $\frac{\partial b^{\tilde{\varepsilon}}}{\partial \varepsilon}$. This shows that in case II we also have

  $$\hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon,\mu}) - \hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu}) \leq \frac{Ca(\varepsilon)\left|\ln\big(c\mu\vartheta(\varepsilon)\big)\right|}{\mu\varepsilon} \left|\varepsilon - \varepsilon_+\right|.$$

- For the second summand we start with case I. The monotonicity of $\varepsilon \mapsto B^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+})$, cf. Corollary 4.1.9, implies

  $$\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+}) - \hat{b}^{\varepsilon+}(\bar{u}_{\varepsilon_+,\mu_+}) = b^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+}) - b^{\varepsilon+}(\bar{u}_{\varepsilon_+,\mu_+}) \leq (\tau(\varepsilon_+) - \tau(\varepsilon)) \ln\big(B^{\varepsilon+}(\bar{u}_{\varepsilon_+,\mu_+})\big).$$

This can be estimated by $|\tau'(\tilde{\varepsilon})||\varepsilon-\varepsilon_+||\ln(c\mu_+\vartheta(\varepsilon_+))|$ with $\tilde{\varepsilon}\in I$, cf. Lemma 6.4.1, which subsequently yields the same estimate as for the first summand using $|\ln(c\mu_+\vartheta(\varepsilon_+))|\leq C|\ln(c\mu\vartheta(\varepsilon))|$ (we recall that c is chosen so small that we have $c\mu\vartheta(\varepsilon)\leq\frac{1}{2}$ for all $(\mu,\varepsilon)\in\mathcal{P}_=$). In case II we can use the same argument but have an additional term $\tilde{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+})-\tilde{b}^{\varepsilon_+}(\bar{u}_{\varepsilon_+,\mu_+})$. For this term we can estimate

$$\tilde{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+})-\tilde{b}^{\varepsilon_+}(\bar{u}_{\varepsilon_+,\mu_+})\leq(\tilde{\tau}(\varepsilon_+)-\tilde{\tau}(\varepsilon))\ln(\tilde{B}(\bar{u}_{\varepsilon_+,\mu_+}))\leq C\left|\tilde{\tau}'(\tilde{\varepsilon})\right|\left|\varepsilon-\varepsilon_+\right|,$$

where we used $\tilde{\tau}(\varepsilon_+)-\tilde{\tau}(\varepsilon)\geq 0$ and $\ln(\tilde{B}(\bar{u}_{\varepsilon_+,\mu_+}))\leq C$. This shows that we obtain the same bound as in case I.

- For the last summand we use arguments as for the other summands and obtain that in case I it holds

$$\begin{aligned}\left(1-\frac{\mu}{\mu_+}\right)\cdot\left(\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu})-\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+})\right)&\leq\left|1-\frac{\mu}{\mu_+}\right|\tau(\varepsilon)\left(\left|\ln\big(B^{\varepsilon}(\bar{u}_{\varepsilon,\mu})\big)\right|-\ln\big(B^{\varepsilon_+}(\bar{u}_{\varepsilon_+,\mu_+})\big)\right)\\&\leq\left|1-\frac{\varepsilon^p}{\varepsilon_+^p}\right|\frac{Ca(\varepsilon)}{\mu}\left(C\left|\ln(c\mu\vartheta(\varepsilon))\right|\right),\end{aligned}\tag{6.21}$$

with $p=2$. Since we have $|1-\frac{\varepsilon^p}{\varepsilon_+^p}|=2^p\frac{\varepsilon^p-\varepsilon_+^p}{\varepsilon^p}$, it is elementary to see that the inequality $|1-\frac{\varepsilon^p}{\varepsilon_+^p}|\leq C\frac{|\varepsilon-\varepsilon_+|}{\varepsilon}$ is satisfied. In case I we, therefore, obtain

$$\left(1-\frac{\mu}{\mu_+}\right)\cdot\left(\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon,\mu})-\hat{b}^{\varepsilon}(\bar{u}_{\varepsilon_+,\mu_+})\right)\leq\frac{Ca(\varepsilon)\left|\ln\big(c\mu\vartheta(\varepsilon)\big)\right|}{\mu\varepsilon}\left|\varepsilon-\varepsilon_+\right|.$$

In case II we can estimate similar to (6.21) with $p=3$ and use $C\geq\tilde{B}(\bar{u}_{\varepsilon,\mu}),\tilde{B}(\bar{u}_{\varepsilon_+,\mu_+})\geq c$, cf. Lemma 6.3.1, to derive the same upper bound as in case I.

In conclusion, all three summands on the right-hand side of (6.20) allow the upper bound $\frac{Ca(\varepsilon)|\ln(c\mu\vartheta(\varepsilon))|}{\mu\varepsilon}|\varepsilon-\varepsilon_+|$. □

6.11. A result for the derivation of complexity estimates in the case of sublinear convergence

In this section we provide a result that allows to establish complexity estimates in the presence of sublinear convergence. From this result we deduce an upper bound for the number of iterations that is necessary to reach a prescribed value ε_* from a starting point ε_0 if the underlying sequence (ε_k) converges to zero at a sublinear rate.

Lemma 6.11.1. *Let a sequence $(\varepsilon_k)\subset\mathbb{R}_{>0}$ be given with $\varepsilon_{k+1}\leq(1-c\varepsilon_k^r)\,\varepsilon_k$ for all $k\in\mathbb{N}_0$, where $c,r,\varepsilon_0>0$ are real numbers. Then the function*

$$\tilde{y}:[0,\infty)\to\mathbb{R},\quad\tilde{y}(t):=\left(crt+\frac{1}{\varepsilon_0^r}\right)^{-\frac{1}{r}}$$

satisfies $\varepsilon_k\leq\tilde{y}(k)$ for all $k\in\mathbb{N}_0$.

Proof. A simple computation shows that $\tilde{y}$ solves the initial value problem $y' = -cy^{r+1}$ with $y(0) = \varepsilon_0$. We now prove by induction that $\varepsilon_k \leq \tilde{y}(k)$ holds for all $k \in \mathbb{N}_0$. For $k = 0$ this is clear. Hence, we can assume $k \geq 1$ and $\varepsilon_{k-1} \leq \tilde{y}(k-1)$. Defining

$$y^* : [k-1, \infty) \to \mathbb{R}, \quad y^*(t) := \left(cr(t-k+1) + \frac{1}{\varepsilon_{k-1}^r} \right)^{-\frac{1}{r}}$$

we see that y^* solves $y' = -cy^{r+1}$ and satisfies $y^*(k-1) = \varepsilon_{k-1}$. From $y^*(k-1) = \varepsilon_{k-1} \leq \tilde{y}(k-1)$ we infer with Lemma D.0.6 that it holds $y^* \leq \tilde{y}$ on $[k-1, \infty)$. In particular, $y^*(k) \leq \tilde{y}(k)$ is satisfied. Thus, it suffices to establish $\varepsilon_k \leq y^*(k)$. Since $t \mapsto (y^*(t))^{r+1}$ decreases monotonically, we infer that $(y^*)' = -c(y^*)^{r+1}$ increases monotonically. This implies $(y^*)'(k-1) \leq (y^*)'(t)$ for all $t \in [k-1, k]$. Using the mean value theorem we deduce that there exists a $\xi \in (k-1, k)$ such that it holds

$$y^*(k) = y^*(k-1) + (y^*)'(\xi) \geq y^*(k-1) + (y^*)'(k-1) = \varepsilon_{k-1} - c\varepsilon_{k-1}^{r+1} \geq \varepsilon_k,$$

which concludes the proof. □

Corollary 6.11.2. *Let a sequence $(\varepsilon_k) \subset \mathbb{R}_{>0}$ be given with $\varepsilon_{k+1} \leq (1 - c\varepsilon_k^r)\,\varepsilon_k$ for all $k \in \mathbb{N}_0$, where $c, r, \varepsilon_0 > 0$ are real numbers. Let $0 < \varepsilon_* \leq \varepsilon_0$ and define*

$$K := \left\lceil \frac{1 - \left(\frac{\varepsilon_*}{\varepsilon_0}\right)^r}{cr\varepsilon_*^r} \right\rceil.$$

Then it holds $\varepsilon_K \leq \varepsilon_*$.

Proof. Denoting by $\tilde{y}$ the same function as in the preceding lemma we infer from this lemma that it suffices to establish $\tilde{y}(K) \leq \varepsilon_*$. By requiring $\tilde{y}(T) = \varepsilon_*$ the definition of $\tilde{y}$ implies

$$T = \frac{1 - \left(\frac{\varepsilon_*}{\varepsilon_0}\right)^r}{cr\varepsilon_*^r}.$$

Since $\tilde{y}$ is monotonically decreasing, this yields $\tilde{y}(\lceil T \rceil) \leq \varepsilon_*$, which proves $\tilde{y}(K) \leq \varepsilon_*$. □

Remark 6.11.3. In the previous lemma and corollary it is possible to include the case $r = 0$, i.e., the case of linear convergence. The proof runs completely parallel but the structure of $\tilde{y}$ and y^* changes, hence extra notational effort is needed. Since we are not interested in the case $r = 0$ and, also, in this case an estimate for K can be computed directly, we omit including a proof for this case. However, we mention that our approach yields that $K = \lceil \frac{1}{c} \ln(\frac{\varepsilon_0}{\varepsilon_*}) \rceil$ iterations suffice to obtain $\varepsilon_K \leq \varepsilon_*$ in this case. The bound obtained by direct calculation reads $K = \lceil -\frac{1}{\ln(1-c)} \ln(\frac{\varepsilon_0}{\varepsilon_*}) \rceil$. Since $-\frac{1}{\ln(1-c)}$ can be approximated (via its Laurent series at $c = 0$) by $1/c - 1/2 - c/12 - \mathcal{O}(c^2)$, which is close to $1/c$ if c is close to zero, we suspect that our bound is reasonably sharp in the sublinear case as well. This is further backed up by numerical tests. For instance, using $\varepsilon_{k+1} := (1 - c\varepsilon_k^r)\varepsilon_k$ for $k \in \mathbb{N}_0$ with $c = r = 1$ and $\varepsilon_0 = 0.5$ we obtain $\varepsilon_{1000} \approx 9.914 \cdot 10^{-4}$, while we have $K = 1007$ for $\varepsilon_* = \varepsilon_{1000}$. In the same setting we can, moreover, conclude from numerical tests that the asymptotical behaviour of (ε_k) is captured very well. For instance, using $\varepsilon_* = \varepsilon_{10^6}$ we have $K = 10^6 + 14$, using $\varepsilon_* = \varepsilon_{10^9}$ we obtain $K = 10^9 + 21$.

7. Barrier methods for variable smoothing parameter

7.1. The short step method SSM$_{(\mathrm{P})}$

In this section we introduce and examine the short step method SSM$_{(\mathrm{P})}$ that is able to solve the state constrained optimal control problem (P).

We state the algorithm of SSM$_{(\mathrm{P})}$.

Algorithm SSM$_{(\mathrm{P})}$ (short step method to solve (P))

Input: Parameters $(\varepsilon_0, \mu_0) \in \mathcal{P}_=$, $\theta \in (0, \frac{1}{4}]$, $\delta \in (0,1)$, starting point $u^0 \in \Lambda_{\varepsilon_0,\mu_0}(\phi(\varepsilon_0))$, where $\phi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is given by $\phi(\varepsilon) := \min\{\theta, \varepsilon^{\frac{1+\delta}{2}}\}$.

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k \in U$ by solving $f''_{\varepsilon_k,\mu_k}(u^k)[s^k] = -f'_{\varepsilon_k,\mu_k}(u^k)$ in U^*.

Set $u^{k+1} := u^k + s^k$.

Choose $\beta_k \in [\frac{1}{2}, 1)$ via backtracking such that $\lambda_{\beta_k\varepsilon_k,\rho^{-1}(\beta_k\varepsilon_k)}(u^{k+1}) \leq \phi(\beta_k\varepsilon_k)$ is satisfied.

Set $\varepsilon_{k+1} := \beta_k\varepsilon_k$ and $\mu_{k+1} := \rho^{-1}(\varepsilon_{k+1})$.

END

Remark 7.1.1. In the following we consider SSM$_{(\mathrm{P})}$ only with the two following backtracking strategies: In *strategy A* we successively test if $\tilde{\beta}_i = 1 - (\frac{1}{2})^i$, $i = 1, 2, \ldots$, is a possible value for β_k and set $\beta_k := \tilde{\beta}_i$ for the first i that passes the test. Of course, other factors than $\frac{1}{2}$ can be used. However, to simplify notation in the following we use precisely this backtracking strategy as strategy A. In *strategy B* we successively test if $\tilde{\beta}_i = 1 - \tilde{c}\varepsilon_k^{\frac{3+p}{2}+\delta}(\frac{1}{2})^i$, $i = 1, 2, \ldots$, is a possible value for β_k and set $\beta_k := \tilde{\beta}_i$ for the first i that passes the test. Here, $p = 2$ in case I and $p = 3$ in case II, and, moreover, $\tilde{c}$ is chosen such that $\tilde{c}\varepsilon_k^{\frac{3+p}{2}+\delta} \leq 1$ for all $k \in \mathbb{N}_0$. This can, for instance, be ensured by taking $\tilde{c} \leq 1/\varepsilon_0^{\frac{3+p}{2}+\delta}$.

Remark 7.1.2. Note that the i-th backtracking step requires to check for $\tilde{\varepsilon}_i := \tilde{\beta}_i\varepsilon_k$ if $\lambda_{\tilde{\varepsilon}_i,\rho^{-1}(\tilde{\varepsilon}_i)}(u^{k+1}) \leq \phi(\tilde{\varepsilon}_i)$ is fulfilled. This is expensive since the evaluation of $\lambda_{\tilde{\varepsilon}_i,\rho^{-1}(\tilde{\varepsilon}_i)}(u^{k+1})$ requires the computation of a Newton step, and hence it can be necessary to compute several Newton steps in each iteration of SSM$_{(\mathrm{P})}$. As a speedup one could, for example, combine the backtracking with the following interpolation strategy: Suppose that $(\tilde{\varepsilon}_i, \lambda_{\tilde{\varepsilon}_i,\rho^{-1}(\tilde{\varepsilon}_i)}(u^{k+1}))$ have

been computed for $i = 1, \dots, K$, i.e., these backtracking steps were unsuccessful. Interpolate these points (or a subset thereof) by a polynomial $p : \mathbb{R} \to \mathbb{R}$. Determine the next candidate $\tilde{\varepsilon}_{K+1}$ for ε_{k+1} as (approximate) solution of the equation $p(\varepsilon) = \phi(\varepsilon)$, or, to give another example, of the equation $p(\varepsilon) = \frac{3}{4}\phi(\varepsilon)$, where a safeguard is included. If this equation has several solutions, choose the largest among those which are smaller than ε_k. If there is no solution that is smaller than ε_k, use $\tilde{\varepsilon}_{K+1}$ from the backtracking instead. Check if $\tilde{\varepsilon}_{K+1}$ can be used as ε_{k+1}, and if this is not the case, set $K + 1 \leftarrow K$ and repeat the whole procedure.

Remark 7.1.3. Below we show that all β sufficiently close to 1 are valid choices for β_k. Thus, the backtracking technique and the overall algorithm are well-defined.

Remark 7.1.4. To obtain a starting point $u^0 \in \Lambda_{\varepsilon_0,\mu_0}(\phi(\varepsilon_0))$ a phase one may be required. We treat phase one methods in Section 7.3.

Remark 7.1.5. Termination criteria for an implementation of $\text{SSM}_{(\text{P})}$ can be based, e.g., on convergence of the objective value or other quantities of interest. We present a termination criterion that can also be used in $\text{SSM}_{(\text{P})}$ when we conduct numerical experiments for variable smoothing parameter in Section 8.3.

Lemma 7.1.6. *In every iteration of Algorithm $\text{SSM}_{(\text{P})}$ the backtracking to obtain β_k terminates successfully after finitely many steps, i.e., $\text{SSM}_{(\text{P})}$ is well-defined.*

Proof. For $k = 0$ we have $u^k \in \Lambda_{\varepsilon_k,\mu_k}(\phi(\varepsilon_k))$. Hence, u^1 satisfies $\lambda_{\varepsilon_k,\mu_k}(u^1) \leq \frac{\phi(\varepsilon_k)^2}{(1-\phi(\varepsilon_k))^2} \leq \frac{\phi(\varepsilon_k)}{2}$ due to Lemma 2.2.20. With Lemma 6.8.1 this implies that there exists a neighborhood N of ε_k such that $\lambda_{\varepsilon,\rho^{-1}(\varepsilon)}(u^1) \leq \phi(\varepsilon)$ for all $\varepsilon \in N$. Therefore, the backtracking terminates successfully after finitely many steps for $k = 0$. Using the same argument for $k > 0$ the assertion follows by induction. □

7.1.1. Convergence of $\text{SSM}_{(\text{P})}$

In this section we show that Algorithm $\text{SSM}_{(\text{P})}$ is convergent, i.e., it generates a sequence $(u^k) \subset U$ that converges strongly to the unique solution of (P).

The following lemma is the base of the convergence proof.

Lemma 7.1.7. *Algorithm $\text{SSM}_{(\text{P})}$ generates sequences $(u^k) \subset U$, $(\varepsilon_k) \subset \mathbb{R}_{>0}$, and $(\mu_k) \subset \mathbb{R}_{>0}$ with*

$$\lambda_{\varepsilon_k,\mu_k}(u^k) \leq \phi(\varepsilon_k) \quad \textit{for all } k \in \mathbb{N}_0 \qquad \textit{and} \qquad (\varepsilon_k, \mu_k) \to (0^+, 0^+) \quad \textit{for } k \to \infty.$$

Proof. Obviously, for all $k \in \mathbb{N}_0$ we have $\lambda_{\varepsilon_k,\mu_k}(u^k) \leq \phi(\varepsilon_k)$. Hence it remains to show that it is possible to choose (β_k) such that it holds $(\varepsilon_k, \mu_k) \to (0^+, 0^+)$ for $k \to \infty$. Since we have $\varepsilon_k \to 0^+$ if and only if $\mu_k \to 0^+$ for $k \to \infty$, it suffices to establish $\varepsilon_k \to 0^+$ for $k \to \infty$. Hence, we need to take a look at the backtracking in $\text{SSM}_{(\text{P})}$. We argue for strategy A, which works as follows: Successively choose $\tilde{\beta}_i := 1 - \left(\frac{1}{2}\right)^i$, $i = 1, 2, 3, \dots$ and check if $\tilde{\beta}_i$ is admissible for β_k. If so, set $\beta_k := \tilde{\beta}_i$ and denote this value of i by $i[k]$. We now prove that for this choice of (β_k) the resulting sequence (ε_k) converges to zero. Obviously, the sequence $(\varepsilon_k) \subset \mathbb{R}_{>0}$ is strictly

monotone decreasing and, thus, converges to $\varepsilon_* \in \mathbb{R}_{\geq 0}$. Assume that $\varepsilon_* > 0$ is true. We demonstrate that there is a constant $c > 0$, a neighborhood N of ε_*, and an index $K \in \mathbb{N}_0$ such that there hold for all $k \geq K$ the inequalities $B^\varepsilon(u^k) \geq c$ for all $\varepsilon \in N$ and $\tilde{B}(u^k) \geq c$. Since we have $\tilde{B}(u^k) \geq c$ for all k due to $u^k \in \Lambda_{\varepsilon_k,\mu_k}$, cf. Lemma 6.3.1, we only need to show that there holds $B^\varepsilon(u^k) \geq c$ for all $\varepsilon \in N$ and all $k \geq K$ with suitable K, c, and N. There exist $K \in \mathbb{N}_0$ and $0 < \eta < \varepsilon_*/2$ such that $(1 - c\mu_k\vartheta(\varepsilon_k)/a(\varepsilon_k))\varepsilon_k \leq \varepsilon_* - \eta$ is satisfied for all $k \geq K$, where c denotes the constant from Lemma 6.6.6. This is true since (ε_k) converges monotonically decreasing to $\varepsilon_* > 0$ and since $c\rho^{-1}(\varepsilon)\vartheta(\varepsilon)/a(\varepsilon)$ is positive and continuous on $[\varepsilon_*, \varepsilon_0]$. We define $N := [\varepsilon_* - \eta, \varepsilon_K]$. Due to Lemma 6.6.6 there holds for all $k \geq K$: $B^\varepsilon(u^k) \geq c\mu_k\vartheta(\varepsilon_k)$ for all $\varepsilon \in [\varepsilon_* - \eta, \varepsilon_k]$. Hence, Corollary 4.1.9 shows that we have for all $k \geq K$: $B^\varepsilon(u^k) \geq c\mu_k\vartheta(\varepsilon_k)$ for all $\varepsilon \in [\varepsilon_* - \eta, \infty) \supset N$. This implies $B^\varepsilon(u^k) \geq c \min_{\varepsilon\in N}\{\rho^{-1}(\varepsilon)\vartheta(\varepsilon)\}$ for all $k \geq K$ and all $\varepsilon \in N$. In conclusion, we have established that there are $c > 0$, N, and $K \in \mathbb{N}_0$ such that there hold for all $k \geq K$ the inequalities $B^\varepsilon(u^k) \geq c$ for all $\varepsilon \in N$ and $\tilde{B}(u^k) \geq c$. Therefore, Lemma 6.8.1 implies that

$$G_{u^k} : N \to \mathbb{R}, \quad G_{u^k}(\varepsilon) := \left(\lambda_{\varepsilon,\rho^{-1}(\varepsilon)}(u^k)\right)^2 - (\phi(\varepsilon))^2$$

is Lipschitz in N for all $k \geq K$ and the Lipschitz constant is independent of k. This is, there exists $L > 0$ such that for all $k \geq K$ it holds

$$|G_{u^k}(\varepsilon) - G_{u^k}(\tilde{\varepsilon})| \leq L\,|\varepsilon - \tilde{\varepsilon}|$$

for all $\varepsilon, \tilde{\varepsilon} \in N$.

We may assume that $i[k] \geq 2$ is satisfied for all sufficiently large k with $k \geq K$. This is possible since otherwise we would have $\varepsilon_k \to 0^+$ for $k \to \infty$. Using the definition of $\phi(\varepsilon)$ and the fact that N does not contain zero and is compact, we have

$$\left(\frac{\phi(\varepsilon)}{1 - \phi(\varepsilon)}\right)^4 - \phi(\varepsilon)^2 \leq -\tilde{c}$$

for all $\varepsilon \in N$ and a suitable $\tilde{c} > 0$ (the function $(t/(1-t))^4 - t^2$ is strictly monotone decreasing on $[0, 1/4]$ with function value 0 at $t = 0$). We recall the definition $\tilde{\beta}_i = 1 - (\frac{1}{2})^i$. Using $G_{u^{k+1}}(\varepsilon_k\tilde{\beta}_{i[k]-1}) > 0$ it follows for all sufficiently large k

$$\begin{aligned} \tilde{c} &\leq \phi(\varepsilon_k)^2 - \left(\frac{\phi(\varepsilon_k)}{1 - \phi(\varepsilon_k)}\right)^4 \leq -G_{u^{k+1}}(\varepsilon_k) \\ &\leq G_{u^{k+1}}(\varepsilon_k\tilde{\beta}_{i[k]-1}) - G_{u^{k+1}}(\varepsilon_k) \leq L(\varepsilon_k - \varepsilon_k\tilde{\beta}_{i[k]-1}) = L\varepsilon_k\left(\frac{1}{2}\right)^{i[k]-1}. \end{aligned} \tag{7.1}$$

Here, we used $\lambda_{\varepsilon_k,\rho^{-1}(\varepsilon_k)}(u^{k+1}) \leq \frac{\phi(\varepsilon_k)^2}{(1-\phi(\varepsilon_k))^2}$, which follows from $\lambda_{\varepsilon_k,\rho^{-1}(\varepsilon_k)}(u^k) \leq \phi(\varepsilon_k)$, and that for all sufficiently large k we have $\varepsilon_k\tilde{\beta}_{i[k]-1} \in N$: It holds $\varepsilon_k(1 - (\frac{1}{2})^{i[k]}) = \varepsilon_k\beta_k \to \varepsilon_*^+$ for $k \to \infty$. This implies

$$\varepsilon_k\tilde{\beta}_{i[k]-1} = \varepsilon_k\left(1 - \left(\frac{1}{2}\right)^{i[k]-1}\right) = 2\varepsilon_k\left(1 - \left(\frac{1}{2}\right)^{i[k]}\right) - \varepsilon_k \to \varepsilon_*$$

for $k \to \infty$.

Using $i[k] \to \infty$ for $k \to \infty$ (otherwise, $\varepsilon_k \to 0^+$) we can take $k \to \infty$ in (7.1) and arrive at the contradiction $\tilde{c} \leq 0$. This establishes $\varepsilon_k \to 0^+$ for $k \to \infty$ and, thus, finishes the proof if strategy A is used in the backtracking. The proof for strategy B is quasi identical and is, therefore, omitted. □

The next theorem states convergence of $\mathrm{SSM}_{(\mathrm{P})}$.

Theorem 7.1.8. *Algorithm $SSM_{(\mathrm{P})}$ generates a sequence $((u^k, \varepsilon_k, \mu_k)) \subset U \times \mathcal{P}_=$ that satisfies $(\varepsilon_k, \mu_k) \to (0^+, 0^+)$ as well as the estimates*

$$\frac{|j(u^k) - j(\bar{u})|}{C_j} \leq \frac{2\mu_k \vartheta(\varepsilon_k)}{C_j} + C\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)$$

and

$$\|u^k - \bar{u}\|_U \leq \sqrt{\frac{8\mu_k \vartheta(\varepsilon_k)}{C_j \alpha}} + \sqrt{\frac{4C}{\alpha}} \sqrt{\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)}$$

for all $k \in \mathbb{N}_0$, where $C > 0$ is independent of k and α denotes the convexity modulus of j/C_j on $\cup_{\varepsilon \in (0, \varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$. In particular, u^k converges strongly to $\bar{u}$.

Proof. The estimates are a consequence of Lemma 6.1.6 and $u^k \in \Lambda_{\varepsilon_k, \mu_k}$, which follows from $\lambda_{\varepsilon_k, \mu_k}(u^k) \leq \phi(\varepsilon_k) \leq \theta \leq \frac{1}{4}$. The assertion $(\varepsilon_k, \mu_k) \to (0^+, 0^+)$ was proven in the previous lemma. □

Remark 7.1.9. Due to $((\varepsilon_k, \mu_k)) \subset \mathcal{P}_=$ we have $\mu_k \vartheta(\varepsilon_k) = \mathcal{O}(\varepsilon_k(1 + |\ln \varepsilon_k|))$ so that the above estimates imply $|j(u^k) - j(\bar{u})| = \mathcal{O}(\varepsilon_k(1 + |\ln \varepsilon_k|))$ and $\|u^k - \bar{u}\|_U = \mathcal{O}(\sqrt{\varepsilon_k(1 + |\ln \varepsilon_k|)})$.

Remark 7.1.10. It is easy to see that the above theorem is still valid if only $\lambda_{\varepsilon_k, \mu_k}(u^k) \leq \theta$ with $\theta \in (0, \frac{1}{4}]$ is ensured in each iteration. However, to prove a rate of convergence we need the stronger requirement $\lambda_{\varepsilon_k, \mu_k}(u^k) \leq \phi(\varepsilon_k)$.

7.1.2. Rate of convergence and complexity of $\mathrm{SSM}_{(\mathrm{P})}$

In this section we prove that the sequences generated by Algorithm $\mathrm{SSM}_{(\mathrm{P})}$ converge with certain rates.

The succeeding lemma establishes a convergence rate for (ε_k).

Lemma 7.1.11. *Let the sequence $(\beta_k) \subset [\frac{1}{2}, 1)$ be generated by Algorithm $SSM_{(\mathrm{P})}$. Then there exists $c > 0$ such that*

$$\beta_k \leq 1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}$$

is satisfied for all $k \in \mathbb{N}_0$, with $p = 2$ in case I and $p = 3$ in case II.

Moreover, if strategy B is used in the backtracking, then there exists a constant $C \in \mathbb{N}$ such that in every iteration the required number of backtracking steps is bounded by C.

Proof. The second assertion is implied by the first: In iteration $k \in \mathbb{N}_0$ of $\mathrm{SSM}_{(\mathrm{P})}$ the first assertion yields that strategy B terminates for an $i \in \mathbb{N}$ that satisfies $\tilde{c}\,(\frac{1}{2})^i \geq c$, since $\tilde{c}\,(\frac{1}{2})^i < c$ would imply $\beta_k = 1 - \tilde{c}\varepsilon_k^{\frac{3+p}{2}+\delta}(\frac{1}{2})^i > 1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}$. Apparently, this implies that i is bounded from above by a constant that is independent of k, which proves the second assertion.

Thus, it remains to establish the first assertion. Let c denote the constant from Lemma 6.6.6. We show that for every k sufficiently large every $\tilde{\beta}$ that satisfies

$$\tilde{\beta} \in \left[1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}, 1\right] \tag{7.2}$$

is admissible as β_k. Due to the backtracking in $\mathrm{SSM}_{(\mathrm{P})}$ this implies that

$$\beta_k \in \left[1 - \hat{c}\varepsilon_k^{\frac{3+p}{2}+\delta}, 1 - \frac{\hat{c}}{2}\varepsilon_k^{\frac{3+p}{2}+\delta}\right)$$

or an even smaller β_k is selected for all k sufficiently large, which yields the first assertion. Here, we use $\hat{c} := c$ in strategy A and $\hat{c} := \min\{c, \tilde{c}/2\}$ in strategy B.

Since $\lambda_{\varepsilon_k,\mu_k}(u^k) \leq \phi(\varepsilon_k) \leq \theta \leq \frac{1}{4}$, we have $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \frac{\theta^2}{(1-\theta)^2} \leq \frac{1}{4}$, which shows $u^{k+1} \in \Lambda_{\varepsilon_k,\mu_k}$ for all $k \in \mathbb{N}_0$. Hence, with I from Lemma 6.6.6 it holds

$$\begin{aligned} I(\varepsilon_k, \mu_k, u^{k+1}) = \left[\varepsilon_k\left(1 - \frac{c\mu_k\vartheta(\varepsilon_k)}{a(\varepsilon_k)}\right), \varepsilon_k\right] &\supset \left[\varepsilon_k\left(1 - \frac{c\varepsilon_k}{a(\varepsilon_k)}\right), \varepsilon_k\right] \\ &= \left[\varepsilon_k\left(1 - \frac{c}{1 + |\ln \varepsilon_k|}\right), \varepsilon_k\right], \end{aligned}$$

where c may have changed but is still independent of k. Thus, we may assume that for all k sufficiently large $[\varepsilon_k(1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}), \varepsilon_k] \subset I(\varepsilon_k, \mu_k, u^{k+1})$ is satisfied. Therefore, Theorem 6.9.6 yields for every k sufficiently large

$$\lambda_{\varepsilon_+,\mu_+}(u^{k+1}) \leq C\left(\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \cdot \sqrt{\frac{1}{\varepsilon_k}} + \frac{1}{\varepsilon_k^2\sqrt{\mu_k}} \cdot (\varepsilon_k - \varepsilon_+)\right)$$

for all $(\varepsilon_+, \mu_+) \in \mathcal{P}_=$ with $\varepsilon_+ \in [\varepsilon_k(1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}), \varepsilon_k]$. Here and in the following, $C > 0$ denotes different constants that are all independent of k.

Since $\lambda_{\varepsilon_k,\mu_k}(u^k) \leq \phi(\varepsilon_k)$, we have $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \frac{\phi(\varepsilon_k)^2}{(1-\phi(\varepsilon_k))^2} \leq 2\phi(\varepsilon_k)^2 \leq 2\varepsilon_k^{1+\delta}$ for sufficiently large k due to $\varepsilon_k \to 0^+$, see Lemma 7.1.7. Moreover, $\varepsilon_+ \in [\varepsilon_k(1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}), \varepsilon_k]$ implies $\varepsilon_k - \varepsilon_+ \leq c\varepsilon_k\varepsilon_k^{\frac{3+p}{2}+\delta}$. Also, we have $\sqrt{\mu_k} \geq \varepsilon_k^{p/2}/C$. Together, this shows for every k sufficiently large

$$\lambda_{\varepsilon_+,\mu_+}(u^{k+1}) \leq C\left(\varepsilon_k^{1+\delta} \cdot \sqrt{\frac{1}{\varepsilon_k}} + \frac{1}{\varepsilon_k^{2+\frac{p}{2}}} \cdot \varepsilon_k\varepsilon_k^{\frac{3+p}{2}+\delta}\right) = C\varepsilon_k^{\frac{\delta}{2}}\varepsilon_k^{\frac{1+\delta}{2}} \leq 2^{\frac{1+\delta}{2}}C\varepsilon_k^{\frac{\delta}{2}}\phi(\varepsilon_+)$$

for all $(\varepsilon_+, \mu_+) \in \mathcal{P}_=$ with $\varepsilon_+ \in [\varepsilon_k(1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}), \varepsilon_k]$, where we used $\varepsilon_+ \geq \varepsilon_k/2$. In conclusion, we have $\lambda_{\varepsilon_+,\mu_+}(u^{k+1}) \leq \phi(\varepsilon_+)$ for every k sufficiently large and all $(\varepsilon_+, \mu_+) \in \mathcal{P}_=$ with $\varepsilon_+ \in [\varepsilon_k(1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}), \varepsilon_k]$. This establishes the assertion. □

Corollary 7.1.12. *Let (ε_k) and (μ_k) be sequences generated by Algorithm SSM$_{(\mathrm{P})}$ and denote by $\sigma_k \in (0,1)$ the ratio $\sigma_k := \frac{\mu_{k+1}}{\mu_k}$. Then for all $k \in \mathbb{N}_0$ it holds*

$$\sigma_k \leq 1 - c\sqrt[p]{\mu_k^{\frac{3+p}{2}+\delta}},$$

where $c > 0$ is independent of k, with $p = 2$ in case I and $p = 3$ in case II.

Proof. From Lemma 7.1.11 we know that

$$\beta_k \leq 1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}$$

is satisfied for all $k \in \mathbb{N}_0$. This implies

$$\sigma_k = \frac{\mu_{k+1}}{\mu_k} = \frac{\rho^{-1}(\varepsilon_{k+1})}{\rho^{-1}(\varepsilon_k)} = \left(\frac{\beta_k \varepsilon_k}{\varepsilon_k}\right)^p \leq \left(1 - c\varepsilon_k^{\frac{3+p}{2}+\delta}\right)^p.$$

Hence, we obtain

$$\sigma_k \leq \left(1 - c\sqrt[p]{\mu_k^{\frac{3+p}{2}+\delta}}\right)^p$$

for all $k \in \mathbb{N}_0$, where we used $\varepsilon_k = C\sqrt[p]{\mu_k}$. Application of the reverse version of Bernoulli's inequality, cf. Lemma D.0.2, implies the assertion. □

The next lemma contains a bound for the number of iterations K of SSM$_{(\mathrm{P})}$ that is necessary to obtain u^K with $|j(u^K) - j(\bar{u})| \leq \tau$ or $\|u^K - \bar{u}\|_U \leq \tau$, where $\tau > 0$ represents a prescribed tolerance.

Lemma 7.1.13. *Let (ε_k), (μ_k), and (u^k) be generated by Algorithm SSM$_{(\mathrm{P})}$. Then there exists $C > 0$ such that for all $\tau > 0$ it holds*

$$K \geq \left\lceil \frac{\left(\frac{C}{\tau}\right)^{\frac{r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{cr} \right\rceil \qquad \Longrightarrow \qquad \left|j(u^K) - j(\bar{u})\right| \leq \tau$$

and

$$K \geq \left\lceil \frac{\left(\frac{C}{\tau}\right)^{\frac{2r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{cr} \right\rceil \qquad \Longrightarrow \qquad \left\|u^K - \bar{u}\right\|_U \leq \tau.$$

Here, c is the constant from the rate of convergence of (ε_k), cf. Lemma 7.1.11, and $r = \frac{3+p}{2} + \delta$, with $p = 2$ in case I and $p = 3$ in case II.

Remark 7.1.14. The preceding lemma implies, for instance, that in case I we have $|j(u^K) - j(\bar{u})| \leq \tau$ for small values of τ after approximately $\mathcal{O}(\tau^{-\frac{5}{2}})$ iterations, and $\|u^K - \bar{u}\|_U \leq \tau$ after roughly $\mathcal{O}(\tau^{-5})$ iterations. Of course, these are worst-case bounds.

Proof. By Theorem 7.1.8 we have for all $k \in \mathbb{N}_0$ the estimates

$$|j(u^k) - j(\bar{u})| \leq C\varepsilon_k \left(1 + |\ln \varepsilon_k|\right) \qquad \text{and} \qquad \|u^k - \bar{u}\|_U \leq C\sqrt{\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)}.$$

Enlarging C if necessary this implies for all $k \in \mathbb{N}_0$ the estimates

$$|j(u^k) - j(\bar{u})| \leq C\varepsilon_k^{1-\delta} \qquad \text{and} \qquad \|u^k - \bar{u}\|_U \leq C\varepsilon_k^{\frac{1-\delta}{2}}.$$

We now show that $|j(u^K) - j(\bar{u})| \leq \tau$ holds by establishing $C\varepsilon_K^{1-\delta} \leq \tau$ provided $K \geq \left\lceil \frac{(\frac{C}{\tau})^{\frac{r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{cr} \right\rceil$. Since $C\varepsilon_K^{\frac{1-\delta}{2}} \leq \tau$ is equivalent to $C^2\varepsilon_K^{1-\delta} \leq \tau^2$, this also implies the complexity estimate for $\|u^K - \bar{u}\|_U \leq \tau$.

Applying Lemma 6.11.1 we obtain $\varepsilon_k \leq \tilde{y}(k)$ for all $k \in \mathbb{N}_0$ for the function

$$\tilde{y} : [0, \infty), \quad \tilde{y}(t) := \left(crt + \frac{1}{\varepsilon_0^r}\right)^{-\frac{1}{r}}.$$

Here, we used that (ε_k) satisfies $\varepsilon_{k+1} \leq (1 - c\varepsilon_k^r)\varepsilon_k$ for all $k \in \mathbb{N}_0$ with $r = \frac{3+p}{2} + \delta$, cf. Lemma 7.1.11. To establish the assertion it, thus, suffices to demonstrate that $C\tilde{y}(K)^{1-\delta} \leq \tau$ is valid. Solving the equation $C\tilde{y}(T)^{1-\delta} = \tau$ yields

$$T = \frac{\left(\frac{C}{\tau_1}\right)^{\frac{r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{cr},$$

which finishes the proof. □

The next theorem is one of the main results of this thesis.

Theorem 7.1.15. *Algorithm $SSM_{(\mathrm{P})}$ generates a sequence $((u^k, \varepsilon_k, \mu_k)) \subset U \times \mathcal{P}_=$ with $(u^k, \varepsilon_k, \mu_k) \to (\bar{u}, 0^+, 0^+)$ for $k \to \infty$, and such that*

$$\varepsilon_{k+1} = \beta_k \varepsilon_k \quad \text{and} \quad \mu_{k+1} = \sigma_k \mu_k$$

are satisfied for all $k \in \mathbb{N}_0$. Here, $\beta_k \in (0, 1)$ and $\sigma_k \in (0, 1)$ satisfy for all k the inequalities

$$\beta_k \leq 1 - c\varepsilon_k^{\frac{3+p}{2}+\delta} \qquad \text{and} \qquad \sigma_k \leq 1 - c\sqrt[p]{\varepsilon_k^{\frac{3+p}{2}+\delta}}$$

with a constant $c > 0$ that is independent of k, where $p = 2$ in case I and $p = 3$ in case II. If strategy B is used in the backtracking, then the number of backtracking steps required to obtain β_k is bounded by a constant that is independent of k.

Moreover, there exists $C > 0$ such that for all $k \in \mathbb{N}_0$ there hold

$$\frac{|j(u^k) - j(\bar{u})|}{C_j} \leq \frac{2\mu_k \vartheta(\varepsilon_k)}{C_j} + C\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)$$

and

$$\|u^k - \bar{u}\|_U \leq \sqrt{\frac{8\mu_k \vartheta(\varepsilon_k)}{C_j \alpha}} + \sqrt{\frac{4C}{\alpha}} \sqrt{\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)},$$

where α denotes the convexity modulus of j/C_j on $\cup_{\varepsilon\in(0,\varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$.

Also, there exists $\tilde{C} > 0$ such that for all $\tau > 0$ we have the complexity estimates

$$K \geq \left\lceil \frac{\left(\frac{\tilde{C}}{\tau}\right)^{\frac{r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{cr} \right\rceil \qquad \Longrightarrow \qquad \left| j(u^K) - j(\bar{u}) \right| \leq \tau$$

and

$$K \geq \left\lceil \frac{\left(\frac{\tilde{C}}{\tau}\right)^{\frac{2r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{cr} \right\rceil \qquad \Longrightarrow \qquad \left\| u^K - \bar{u} \right\|_U \leq \tau.$$

Here, c denotes the same constant as above and $r = \frac{3+p}{2} + \delta$.

Proof. Combine Theorem 7.1.8, Lemma 7.1.11, Corollary 7.1.12, and Lemma 7.1.13. □

Remark 7.1.16. Of course, the rates of convergence of (ε_k) and (μ_k) and also the complexity estimates are worst-case bounds. This is, in practice we can hope for better rates and better complexity.

Remark 7.1.17. We remark again that due to $((\varepsilon_k, \mu_k)) \subset \mathcal{P}_=$ the above estimates imply $|j(u^k) - j(\bar{u})| = \mathcal{O}(\varepsilon_k(1 + |\ln \varepsilon_k|))$ and $\|u^k - \bar{u}\|_U = \mathcal{O}(\sqrt{\varepsilon_k(1 + |\ln \varepsilon_k|)})$.

Remark 7.1.18. Also, let us mention again that in case II we have $j = \hat{j}$, i.e., estimates for j are estimates for the objective of the original reduced problem ($\mathrm{P}_{\mathrm{red}}$). In case I estimates for j can be transferred to $\hat{j}$ as described in Remark 5.2.6.

Remark 7.1.19. We point out that the number of backtracking steps and, thus, the number of Newton steps required for each iteration of $\mathrm{SSM}_{(\mathrm{P})}$ may tend to infinity as ε tends to zero if strategy A is used for backtracking. In strategy B this cannot happen; the number of Newton steps required in each iteration of $\mathrm{SSM}_{(\mathrm{P})}$ is bounded by a constant. This shows that from a theoretical point of view the convergence result for $\mathrm{SSM}_{(\mathrm{P})}$ with strategy B is stronger. However, since this strategy limits the practical speed of convergence severely, we also provided a convergence result for the more practically oriented strategy A.

7.2. The long step method $\mathrm{LSM}_{(\mathrm{P})}$

In this section we introduce and examine the long step method $\mathrm{LSM}_{(\mathrm{P})}$ that is able to solve the state constrained optimal control problem (P).

We state the algorithm of the long step method $\mathrm{LSM}_{(\mathrm{P})}$. In fact, we state versions of $\mathrm{LSM}_{(\mathrm{P})}$, Version A and Version B. We indicate the different versions by writing "A: Command" or "B: Command", respectively, if "Command" is to be executed in Version A or Version B only.

Algorithm $LSM_{(P)}$ (long step method to solve (P))

Input: Parameters $(\varepsilon_0, \mu_0) \in \mathcal{P}_=$, $\theta \in (0, \frac{1}{4}]$, $\beta_{\min}, \beta_{\max} \in (0,1)$ with $\beta_{\min} \leq \beta_{\max}$, starting point $u^0 \in U_{\mathrm{ad}}(\varepsilon_0)$.

Denote by $\phi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ a function with $\phi(\varepsilon) \leq \theta$ for all $\varepsilon > 0$.

FOR $k = 0, 1, 2, \ldots$:

Compute the Newton step $s^k \in U$ by solving $f''_{\varepsilon_k,\mu_k}(u^k)[s^k] = -f'_{\varepsilon_k,\mu_k}(u^k)$ in U^*.

CALL Algorithm LSMSUB from Section 2.7 with $(u^k, s^k, \mu_k, \phi(\varepsilon_k))$ (use $X := U$, $K := U_{\mathrm{ad}}(\varepsilon)$, $I_s := (0, \mu_k]$, and $f_{\mu_k} := f_{\varepsilon_k,\mu_k}$ in LSMSUB) and denote its return value by u^{k+1}.

A: Choose $\beta_k \in [\beta_{\min}, \beta_{\max}]$.

B: Choose $\beta_k \in [\beta_{\min}, 1)$ via backtracking such that $u^{k+1} \in U_{\mathrm{ad}}(\beta_k \varepsilon_k)$ is satisfied.

Set $\varepsilon_{k+1} := \beta_k \varepsilon_k$ and $\mu_{k+1} := \rho^{-1}(\varepsilon_{k+1})$.

A: **IF** $u^{k+1} \notin U_{\mathrm{ad}}(\varepsilon_{k+1})$, **THEN** choose $\lambda_{k+1} \in [0,1)$ via backtracking such that $\tilde{u}^{k+1} := \lambda_{k+1} u^{k+1} + (1 - \lambda_{k+1}) u^\circ \in U_{\mathrm{ad}}(\varepsilon_{k+1})$ holds and redefine $u^{k+1} := \tilde{u}^{k+1}$.

END

Remark 7.2.1. Obviously, the difference between Version A and Version B of $LSM_{(P)}$ is the following:

- In Version A the reduction of ε_k happens linearly. Feasibility of the new iterate u^{k+1}, i.e. $u^{k+1} \in U_{\mathrm{ad}}(\varepsilon_{k+1})$, is ensured via a shift towards the interior if necessary.
- In Version B, however, the reduction of ε_k is allowed to be sublinear. Feasibility of u^{k+1} is ensured via the choice of β_k.

In conclusion, Version A of $LSM_{(P)}$ aims at a large reduction of ε in each iteration. The drawback of this is that after reducing ε the actual iterate may be infeasible, which is why we employ a shift towards u°. In Version B this shift is not necessary since the update for ε is chosen such that the actual iterate stays feasible. The disadvantage of this is that in Version B the reduction for ε is generally smaller than in Version A. However, if the shift towards u° is large, it is reasonable to expect that this affects the number of Newton steps required by LSMSUB negatively.

Remark 7.2.2. We later discuss different choices for $\phi(\varepsilon)$.

Remark 7.2.3. We assume that the backtracking for λ_{k+1} in Version A has the following structure: It generates a sequence $(\tilde{\lambda}_i)_{i \in \mathbb{N}_0} \subset (0,1)$ with $\tilde{\lambda}_i \to 0^+$ for $i \to \infty$ and uses $\lambda_{k+1} := \tilde{\lambda}_i$ for the smallest i that satisfies $\tilde{\lambda}_i u^{k+1} + (1 - \tilde{\lambda}_i) u^\circ \in U_{\mathrm{ad}}(\varepsilon_{k+1})$. For the backtracking of β_k in Version B we impose a similar structure: It generates a sequence $(\tilde{\beta}_i)_{i \in \mathbb{N}_0} \subset [\beta_{\min}, 1)$ with $\tilde{\beta}_i \to 1^-$ for $i \to \infty$ and uses $\beta_k := \tilde{\beta}_i$ for the smallest i that satisfies $u^{k+1} \in U_{\mathrm{ad}}(\tilde{\beta}_i \varepsilon_k)$. In particular, we allow in Version A and Version B that the backtracking changes during the course of the algorithm in the sense that different sequences $(\tilde{\lambda}_i)$ and $(\tilde{\beta}_i)$, respectively, can be employed in different iterations of Version A and Version B.

Remark 7.2.4. We emphasize that the shift in Version A of $LSM_{(P)}$ does not require additional solves of the state equation apart from computing $y^\circ = y(u^\circ)$ during the initialization of $LSM_{(P)}$. This can be argued similar as in Remark 5.3.2.

Remark 7.2.5. Termination criteria for an implementation of $\mathrm{LSM}_{(\mathrm{P})}$ can be based, e.g., on convergence of the objective value or other quantities of interest. We develop a termination criterion for $\mathrm{LSM}_{(\mathrm{P})}$ when we conduct numerical experiments for variable smoothing parameter in Section 8.3.

We have the following bound on the maximal number of iterations required by LSMSUB.

Lemma 7.2.6. *When called from Algorithm $\mathrm{LSM}_{(\mathrm{P})}$ with $(u^k, s^k, \mu_k, \phi(\varepsilon_k))$, LSMSUB terminates after finitely many iterations with a point $\tilde{u} \in U_{ad}(\varepsilon_k)$ that satisfies $\lambda_{\varepsilon_k,\mu_k}(\tilde{u}) \leq \phi(\varepsilon_k)$. More precisely, the number of iterations of LSMSUB is bounded from above by*

$$\left\lfloor 10.79 \left(f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \right) \right\rfloor + \left\lceil 7.13 + 1.45 \ln \left| \ln \sqrt{2\phi(\varepsilon_k)} \right| \right\rceil .$$

Proof. We recall the fact that Assumption 2.5.2 is satisfied, cf. Lemma 6.1.6. Hence, the assertion follows from Corollary 2.7.8. □

Remark 7.2.7. The maximal number of iterations of LSMSUB depends only very weakly on $\phi(\varepsilon_k)$. This is to say that this number increases only slightly if we require, say, $\lambda_{\varepsilon_k,\mu_k}(\tilde{u}) \leq \varepsilon_k$ instead of $\lambda_{\varepsilon_k,\mu_k}(\tilde{u}) \leq \frac{1}{4}$. This is due to the quadratic convergence that Newton's method achieves in $\Lambda_{\varepsilon_k,\mu_k}$. However, the above bound also tells us that the number of iterations of LSMSUB may become arbitrarily large for $k \to \infty$ if we have $\phi(\varepsilon_k) \to 0^+$, which yields that the second summand tends to infinity. For practical purposes, however, we note that for the choice $\phi(\varepsilon) = \min\{\varepsilon, \theta\}$, and $\varepsilon_k = 10^{-16}$, the second summand equals 13. We will see in the numerical experiments in Section 8 that $\varepsilon_k = 10^{-16}$ is far smaller than a practically reasonable choice for the final value of ε_k. Hence, in practical applications the second summand can be estimated by 13 or a smaller value, i.e., the fact that it tends to infinity can be neglected.

Remark 7.2.8. The first summand shows how the number of iterations of LSMSUB depends on the choice of u^k and ε_k. We see that this number depends on the function value $f_{\varepsilon_k,\mu_k}(u^k)$ and the optimal value $f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})$. In a practical implementation of $\mathrm{LSM}_{(\mathrm{P})}$, this can be used to develop heuristics for the choice of ε_k and u^k. As a very simple example we consider the shifting step in Version A of $\mathrm{LSM}_{(\mathrm{P})}$. In this step we have the freedom to further decrease λ_{k+1} once a feasible λ_{k+1} is found. The preceding result (rather unsurprisingly) indicates that a further decrease of λ_{k+1} may be beneficial if it lowers the function value. As a more complex example we mention that if $\phi(\varepsilon_k)$ is close to zero for small ε_k, we can regard $f_{\varepsilon_k,\mu_k}(u^k)$ as good approximation of $f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})$ due to $\lambda_{\varepsilon_k,\mu_k}(u^k) \leq \phi(\varepsilon_k)$, cf. Lemma 2.2.23. Thus, along with the iterates we obtain an approximation of the sequence $((\varepsilon_k, f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})))$. This can be used for extrapolation, i.e., for estimating $f_{\varepsilon_{k+1},\mu_{k+1}}(\bar{u}_{\varepsilon_{k+1},\mu_{k+1}})$. The choice of ε_{k+1} could then, for instance, be based on a comparison of $f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})$ and $f_{\varepsilon_{k+1},\mu_{k+1}}(u^{k+1}) - f_{\varepsilon_{k+1},\mu_{k+1}}(\bar{u}_{\varepsilon_{k+1},\mu_{k+1}})$.

We take a look at the shifting in Version A of $\mathrm{LSM}_{(\mathrm{P})}$.

Lemma 7.2.9. *We consider iteration $k \in \mathbb{N}_0$ of Version A of Algorithm $\mathrm{LSM}_{(\mathrm{P})}$. If it holds $u^{k+1} \notin U_{ad}(\varepsilon_{k+1})$, then there exists $\delta_{k+1} > 0$ such that $(\lambda_{k+1}u^{k+1} + (1 - \lambda_{k+1})u^\circ) \in U_{ad}(\varepsilon_{k+1})$ is satisfied for all $\lambda_{k+1} \in [0, \delta_{k+1}]$.*

Remark 7.2.10. Of course, the assertion is also true in the case $u^{k+1} \in U_{\mathrm{ad}}(\varepsilon_{k+1})$, with $\delta_{k+1} = 1$.

Proof. We have $B^{\varepsilon}(u^{\circ}) = \min_{\varepsilon}(y^{\circ} - y_a) \geq \min(y^{\circ} - y_a) \geq \tau^{\circ}$ for every $\varepsilon > 0$, i.e. $u^{\circ} \in D_{b^{\varepsilon}}$ for all $\varepsilon > 0$. Hence, from the definition of $C_{\hat{j}}$ in case I and $C_{\|\cdot\|}$ in case II we infer $u^{\circ} \in U_{\mathrm{ad}}(\varepsilon_{k+1})$. Since $U_{\mathrm{ad}}(\varepsilon_{k+1})$ is open, continuity implies that $(\lambda_{k+1}u^{k+1} + (1 - \lambda_{k+1})u^{\circ}) \in U_{\mathrm{ad}}(\varepsilon_{k+1})$ holds true for all λ_{k+1} sufficiently close to zero. □

The following lemma contains the well-definition of Version A of Algorithm $\mathrm{LSM}_{(P)}$. It also constitutes the base for the convergence proof of this version.

Lemma 7.2.11. *Version A of Algorithm $LSM_{(P)}$ generates a sequence $\big((u^k, \varepsilon_k, \mu_k)\big) \subset U \times \mathcal{P}_{=}$ with*

$$\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \phi(\varepsilon_k) \quad \textit{for all } k \in \mathbb{N}_0 \qquad \textit{and} \qquad (\varepsilon_k, \mu_k) \to (0^+, 0^+) \quad \textit{for } k \to \infty.$$

Moreover, the convergence of (ε_k) and (μ_k) is q-linear.

Proof. Newton steps exist since f_{ε_k,μ_k} is nondegenerate, cf. Lemma 6.1.6. Lemma 7.2.6 implies that in every iteration of $\mathrm{LSM}_{(P)}$, LSMSUB terminates finitely, while Lemma 7.2.9 shows that this is also true for the backtracking to obtain λ_{k+1}. Together, it follows that Version A of $\mathrm{LSM}_{(P)}$ generates a sequence $\big((u^k, \varepsilon_k, \mu_k)\big) \subset U \times \mathcal{P}_{=}$ with $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \phi(\varepsilon_k)$ for all $k \in \mathbb{N}_0$. Moreover, we obviously have $\varepsilon_k \leq \beta_{\max}^k \varepsilon_0$ for all $k \in \mathbb{N}_0$, and hence $\varepsilon_k \to 0^+$ for $k \to \infty$ as well as q-linear convergence. From

$$\frac{\mu_{k+1}}{\mu_k} = \frac{\rho^{-1}(\varepsilon_{k+1})}{\rho^{-1}(\varepsilon_k)} = \left(\frac{\beta_k \varepsilon_k}{\varepsilon_k}\right)^p = \beta_k^p \leq \beta_{\max}^p$$

with $p = 2$ in case I and $p = 3$ in case II we deduce that (μ_k) also converges q-linearly to zero. □

For Version B of $\mathrm{LSM}_{(P)}$ we have the following result that contains, in particular, its well-definition.

Lemma 7.2.12. *Version B of Algorithm $LSM_{(P)}$ generates a sequence $\big((u^k, \varepsilon_k, \mu_k)\big) \subset U \times \mathcal{P}_{=}$ with $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \phi(\varepsilon_k)$ for all $k \in \mathbb{N}_0$.*

Proof. Newton steps can be computed since f_{ε_k,μ_k} is nondegenerate, cf. Lemma 6.1.6. Taking Lemma 7.2.6 into account it only remains to argue that the backtracking strategy in iteration $k \in \mathbb{N}_0$ yields $\beta_k \in [\beta_{\min}, 1)$ with $u^{k+1} \in U_{\mathrm{ad}}(\beta_k \varepsilon_k)$ after finitely many steps. Lemma 7.2.6 shows that $u^{k+1} \in U_{\mathrm{ad}}(\varepsilon_k)$ holds true. Hence, we have $B^{\varepsilon_k}(u^{k+1}) > 0$ and $\tilde{B}(u^{k+1}) > 0$. We now use $\varepsilon_k > 0$ and the continuity of $\varepsilon \mapsto B^{\varepsilon}(u^{k+1})$ on $(0, \infty)$. Therefrom we infer that for all $\tilde{\varepsilon}$ sufficiently close to ε_k there hold $B^{\tilde{\varepsilon}}(u^{k+1}) > 0$ and $\tilde{B}(u^{k+1}) > 0$, i.e. $u^{k+1} \in U_{\mathrm{ad}}(\tilde{\varepsilon})$. This shows that for β_k sufficiently close to 1 there holds $u^{k+1} \in U_{\mathrm{ad}}(\beta_k \varepsilon_k)$. Since the sequence $(\tilde{\beta}_i)$ generated by backtracking is supposed to satisfy $\tilde{\beta}_i \to 1^-$ for $i \to \infty$, we obtain after finitely many steps a $\tilde{\beta}_i$ that is accepted as β_k. □

7.2.1. Convergence of Version A of LSM$_{(\mathrm{P})}$

In this section we only deal with Version A of Algorithm LSM$_{(\mathrm{P})}$. We show that the shift towards u° becomes arbitrarily small and that this version of LSM$_{(\mathrm{P})}$ converges.

The next lemma demonstrates that the necessary shift tends to zero as ε_k tends to zero. We recall that the constant τ° stems from the interior point y°, cf. Assumption 3.1.9 6).

Lemma 7.2.13. *There exists $C > 0$ that is independent of k such that in iteration k of Version A of Algorithm LSM$_{(\mathrm{P})}$ it holds $(\lambda_{k+1}u^{k+1} + (1-\lambda_{k+1})u^\circ) \in U_{ad}(\varepsilon_{k+1})$ for all $\lambda_{k+1} \in [0,1)$ with*

$$\lambda_{k+1} < \frac{\tau^\circ}{\tau^\circ + C\varepsilon_k\,(1 + \ln|\varepsilon_k|)}.$$

In particular, λ_{k+1} can be chosen close to 1 if k is large.

Remark 7.2.14. The fact that the necessary shift tends to zero when ε tends to zero should be incorporated in the backtracking strategy for λ_{k+1} since large shifts can be expected to result in large iteration numbers of LSMSUB. As a very basic example a backtracking of the form $\tilde{\lambda}_i = (\frac{k+1}{k+2})^i$ for $i = 0, 1, 2$ may be used in iteration $k \in \mathbb{N}_0$ of LSM$_{(\mathrm{P})}$. We mention also that the backtracking strategy may be improved in practice by taking into account not only the feasibility of $\lambda_{k+1}u^{k+1} + (1-\lambda_{k+1})u^\circ$ but also its function value $f_{\varepsilon_{k+1},\mu_{k+1}}(\lambda_{k+1}u^{k+1} + (1-\lambda_{k+1})u^\circ)$ since a lower function value may result in less iterations of LSMSUB, cf. the estimate in Lemma 7.2.6 and the discussion in Remark 7.2.8.

Proof. For $k \to \infty$ we have $\varepsilon_k \to 0^+$, see Lemma 7.2.11. Hence, there holds

$$\frac{\tau^\circ}{\tau^\circ + C\varepsilon_{k+1}\,(1 + \ln|\varepsilon_{k+1}|)} \to 1^- \quad \text{for } k \to \infty,$$

which shows that λ_{k+1} can be chosen close to 1 for large values of k provided we can prove the estimate for λ_{k+1}. Hence, it remains to demonstrate the existence of $C > 0$ such that $\lambda_{k+1} < \frac{\tau^\circ}{\tau^\circ + C\varepsilon_k(1+\ln|\varepsilon_k|)}$ implies $(\lambda_{k+1}u^{k+1} + (1-\lambda_{k+1})u^\circ) \in U_{\mathrm{ad}}(\varepsilon_{k+1})$, i.e., $B^{\varepsilon_{k+1}}(\lambda_{k+1}u^{k+1} + (1-\lambda_{k+1})u^\circ) > 0$ and $\tilde{B}(\lambda_{k+1}u^{k+1} + (1-\lambda_{k+1})u^\circ) > 0$. Using $u^{k+1} \in U_{\mathrm{ad}}(\varepsilon_k)$ we obtain from Corollary 4.4.4 a constant $C > 0$ independent of u^{k+1} and ε_k and such that the pointwise estimate

$$y(u^{k+1}) - y_a \geq -C\varepsilon_k\,(1 + |\ln\varepsilon_k|)$$

is satisfied. Moreover, there holds $y^\circ - y_a \geq \tau^\circ$, see Assumption 3.1.9. Together, we infer that

$$\begin{aligned}\Big(\lambda y(u^{k+1}) + (1-\lambda)y^\circ\Big) - y_a &= \lambda\Big(y(u^{k+1}) - y_a\Big) + (1-\lambda)\,(y^\circ - y_a)\\ &\geq -C\lambda\varepsilon_k\,(1 + |\ln\varepsilon_k|) + (1-\lambda)\tau^\circ\end{aligned}$$

holds true for all $\lambda \in [0,1]$. The expression on the right-hand side is positive for all $\lambda \in [0,1)$ that satisfy

$$\lambda < \frac{\tau^\circ}{\tau^\circ + C\varepsilon_k\,(1 + \ln|\varepsilon_k|)}.$$

Using the affine linearity of $u \mapsto y(u)$ and Corollary 4.1.3 we, hence, obtain for all these λ

$$\begin{aligned} B^{\varepsilon_{k+1}}\Big(\lambda u^{k+1} + (1-\lambda)u^\circ\Big) &= B^{\varepsilon_{k+1}}_{C(\overline{\Omega}_a)}\Big(y(\lambda u^{k+1} + (1-\lambda)u^\circ)\Big) \\ &= B^{\varepsilon_{k+1}}_{C(\overline{\Omega}_a)}\Big(\lambda y(u^{k+1}) + (1-\lambda)y^\circ\Big) \\ &= \min\nolimits_{\varepsilon_{k+1}}\Big((\lambda y(u^{k+1}) + (1-\lambda)y^\circ) - y_a\Big) \\ &\geq \min\Big((\lambda y(u^{k+1}) + (1-\lambda)y^\circ) - y_a\Big) > 0. \end{aligned}$$

To finish the proof it remains to establish $\tilde{B}(\lambda u^{k+1} + (1-\lambda)u^\circ) > 0$ for all $\lambda \in [0,1)$ that satisfy $\lambda < \frac{\tau^\circ}{\tau^\circ + C\varepsilon_k(1+\ln|\varepsilon_k|)}$. Since we have $u^{k+1} \in U_{\mathrm{ad}}(\varepsilon_k)$, there holds $\tilde{B}(u^{k+1}) > 0$. Furthermore, $\tilde{B}(u^\circ) > 0$ is satisfied. Since $u \mapsto \tilde{B}(u)$ is concave, we infer $\tilde{B}(\lambda u^{k+1} + (1-\lambda)u^\circ) > 0$. □

The next theorem is one of the main results of this thesis.

Theorem 7.2.15. *Version A of Algorithm $LSM_{(P)}$ generates a sequence $((u^k, \varepsilon_k, \mu_k)) \subset U \times \mathcal{P}_=$ that satisfies $(u^k, \varepsilon_k, \mu_k) \to (\bar{u}, 0^+, 0^+)$, where the convergence of (ε_k) and (μ_k) is q-linear. Also, we have the estimates*

$$\frac{|j(u^{k+1}) - j(\bar{u})|}{C_j} \leq \frac{2\mu_k \vartheta(\varepsilon_k)}{C_j} + C\varepsilon_k\left(1 + |\ln \varepsilon_k|\right)$$

and

$$\left\|u^{k+1} - \bar{u}\right\|_U \leq \sqrt{\frac{8\mu_k\vartheta(\varepsilon_k)}{C_j\alpha}} + \sqrt{\frac{4C}{\alpha}}\sqrt{\varepsilon_k\left(1 + |\ln\varepsilon_k|\right)}$$

for all $k \in \mathbb{N}_0$ and a constant $C > 0$. Here, α denotes the convexity modulus of j/C_j on $\cup_{\varepsilon\in(0,\varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$.

Moreover, the number of iterations required by LSMSUB in iteration $k \in \mathbb{N}_0$ is bounded from above by

$$\left\lfloor 10.79\left(f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})\right)\right\rfloor + \left\lceil 7.13 + 1.45\ln\left|\ln\sqrt{2\phi(\varepsilon_k)}\right|\right\rceil.$$

Proof. The estimates for $|j(u^{k+1}) - j(\bar{u})|$ and $\|u^{k+1} - \bar{u}\|_U$ follow from Lemma 6.1.6, where we used $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \phi(\varepsilon_k) \leq \frac{1}{4}$, cf. Lemma 7.2.11. Lemma 7.2.11 also shows the q-linear convergence of (ε_k) and (μ_k) to zero. The estimate for $\|u^{k+1} - \bar{u}\|_U$ now implies $u^k \to \bar{u}$ for $k \to \infty$. The bound on the number of iterations required by LSMSUB is proven in Lemma 7.2.6. □

Remark 7.2.16. The above theorem establishes convergence of $((u^k, \varepsilon_k, \mu_k))$ and also provides convergence rates. Furthermore, it features an estimate for the number of iterations required by LSMSUB in each iteration of $LSM_{(P)}$. However, we note that this estimate is very basic since it still allows, for instance, that the iteration numbers of LSMSUB increase arbitrarily fast as ε converges to zero. Note that this number equals the number of Newton steps required by LSMSUB. For Version B of $LSM_{(P)}$ we will provide a convergence result that contains a more concrete bound for the iteration numbers of LSMSUB.

Remark 7.2.17. We remark that the above estimates imply $|j(u^{k+1}) - j(\bar{u})| = \mathcal{O}(\varepsilon_k(1 + |\ln \varepsilon_k|))$ and $\|u^{k+1} - \bar{u}\|_U = \mathcal{O}(\sqrt{\varepsilon_k(1 + |\ln \varepsilon_k|)})$.

Remark 7.2.18. Also, let us mention again that in case II we have $j = \hat{j}$, i.e., estimates for j are estimates for the objective of the original reduced problem ($\mathrm{P}_{\mathrm{red}}$). In case I estimates for j can be transferred to $\hat{j}$ as described in Remark 5.2.6.

7.2.2. Convergence of Version B of LSM$_{(\mathrm{P})}$

In this section we prove that Version B of Algorithm LSM$_{(\mathrm{P})}$ converges for two different backtracking strategies to update ε_k.

Lemma 7.2.12 implies that we have the error estimates from Lemma 6.1.6. Thus, we see that for a convergence proof we only have to ensure $\varepsilon_k \to 0^+$ for $k \to \infty$. Considering the update formula $\varepsilon_{k+1} = \beta_k \varepsilon_k$ we, hence, need to show that $\beta_k \in [\beta_{\min}, 1)$ can be chosen sufficiently small.

Lemma 7.2.19. *There exists $c > 0$ such that for every $k \in \mathbb{N}_0$ it holds $u^{k+1} \in U_{ad}(\beta_k \varepsilon_k)$ in iteration k of Version B of LSM$_{(\mathrm{P})}$ for all $\beta_k \in [1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)}, 1]$.*

Proof. Applying Lemma 6.6.6 we readily obtain the assertion with c from that lemma. Lemma 6.6.6 is applicable since $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \phi(\varepsilon_k) \leq \frac{1}{4}$, cf. Lemma 7.2.12. □

In the next two corollaries we present two different backtracking strategies to obtain β_k in iteration k of Version B of LSM$_{(\mathrm{P})}$.

Corollary 7.2.20. *We consider Version B of LSM$_{(\mathrm{P})}$. For the backtracking in iteration k we choose $\tilde{\beta}_0 \in [\beta_{\min}, \beta_{\max}]$ and define $\tilde{\beta}_i := 1 - (\frac{1}{2})^i(1 - \tilde{\beta}_0)$ for $i \in \mathbb{N}$. We select the smallest $i \in \mathbb{N}_0$ such that $u^{k+1} \in U_{ad}(\tilde{\beta}_i \varepsilon_k)$ is satisfied and set $\beta_k := \tilde{\beta}_i$. Then there exists $c > 0$ such that for every $k \in \mathbb{N}_0$ it holds $\beta_k \leq 1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)}$ in iteration k of Version B of LSM$_{(\mathrm{P})}$.*

Proof. We have $\frac{\mu\vartheta(\varepsilon)}{a(\varepsilon)} \leq \tilde{c}$ for a suitable $\tilde{c} > 0$ and all $(\varepsilon, \mu) \in \mathcal{P}_=$. Hence, we can choose $c > 0$ so small that the assertion of the previous lemma holds true and $1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)} \geq \beta_{\max}$ is valid for all $k \in \mathbb{N}_0$. The latter implies $\tilde{\beta}_0 \leq 1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)}$ for every $k \in \mathbb{N}_0$, where $\tilde{\beta}_0$ stems from the backtracking strategy. Hence, for a suitable $i = i(k) \in \mathbb{N}_0$ we have $\tilde{\beta}_i \in [1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)}, 1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{2a(\varepsilon_k)})$. In combination with the previous lemma this yields that in iteration k a $\beta_k \in [1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)}, 1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{2a(\varepsilon_k)})$ (or an even smaller β_k) is selected, which establishes the assertion. □

Remark 7.2.21. In case I we have $\frac{\mu\vartheta(\varepsilon)}{a(\varepsilon)} = \tilde{c}$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$ and a suitable $\tilde{c} > 0$. This demonstrates that (ε_k) converges q-linearly to zero in this case if the backtracking from the preceding corollary is used, albeit the rate may be close to 1. In case II we have $\frac{\mu\vartheta(\varepsilon)}{a(\varepsilon)} = \frac{\tilde{c}}{1+|\ln \varepsilon|}$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$ and a suitable $\tilde{c} > 0$, which displays that (ε_k) converges to zero at a rate that converges very slowly to 1.

Corollary 7.2.22. *We consider Version B of $LSM_{(P)}$. For the backtracking in iteration k we define $\tilde{\beta}_i := 1 - (\frac{1}{2})^i \frac{\tilde{c}\mu_k}{a(\varepsilon_k)(1+|\ln \varepsilon_k|)}$ for $i \in \mathbb{N}_0$, where $\tilde{c}$ is so small that $1 - \frac{\tilde{c}\mu}{a(\varepsilon)(1+|\ln \varepsilon|)} \geq \beta_{\min}$ holds for all $(\varepsilon, \mu) \in \mathcal{P}_=$ (we use this $\tilde{c}$ for all k). We select the smallest $i \in \mathbb{N}_0$ such that $u^{k+1} \in U_{ad}(\tilde{\beta}_i \varepsilon_k)$ is satisfied and set $\beta_k := \tilde{\beta}_i$. Then there exists $c > 0$ such that for every $k \in \mathbb{N}_0$ it holds $\beta_k \leq 1 - \frac{c\mu_k}{a(\varepsilon_k)(1+|\ln \varepsilon_k|)}$ in iteration k of Version B of $LSM_{(P)}$.*

Remark 7.2.23. The constant $\tilde{c}$ exists due to $\frac{\mu}{a(\varepsilon)} \leq C\varepsilon$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$, with a suitable constant C. Moreover, it is easy to obtain an explicit value for $\tilde{c}$. For example, in case I we have $\frac{\mu}{a(\varepsilon)(1+|\ln \varepsilon|)} \leq \varepsilon \leq \varepsilon_s$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$, hence $\tilde{c} := \frac{1-\beta_{\min}}{\varepsilon_s}$ can be used. This shows that it is possible to actually implement the backtracking strategy of the preceding corollary.

Proof. Due to $\vartheta(\varepsilon) \geq \frac{C}{\varepsilon}$ for all $\varepsilon \in (0, \varepsilon_s]$ and a suitable $C > 0$, we have $1 - \frac{c\mu_k \vartheta(\varepsilon_k)}{a(\varepsilon_k)} \leq 1 - \frac{\hat{c}\mu_k}{a(\varepsilon_k)(1+|\ln \varepsilon_k|)}$ for all $k \in \mathbb{N}_0$ and a suitable k-independent $\hat{c} > 0$ with $\hat{c} \leq \tilde{c}$, where $c > 0$ denotes the same constant as in Lemma 7.2.19. With Lemma 7.2.19 the backtracking now yields that in iteration k a $\beta_k \in [1 - \frac{\hat{c}\mu_k}{a(\varepsilon_k)(1+|\ln \varepsilon_k|)}, 1 - \frac{\hat{c}\mu_k}{2a(\varepsilon_k)(1+|\ln \varepsilon_k|)})$ (or an even smaller β_k) is selected, which establishes the assertion. □

Corollary 7.2.24. *For the sequence $((\varepsilon_k, \mu_k))$ generated by Version B of $LSM_{(P)}$ it holds $(\varepsilon_k, \mu_k) \to (0^+, 0^+)$ for $k \to \infty$ provided that either the backtracking strategy from Corollary 7.2.20 or the one from Corollary 7.2.22 is employed.*

Proof. Since (ε_k) decreases monotonically, it converges to $\varepsilon_* \geq 0$. Assuming $\varepsilon_* > 0$ we obtain from Corollary 7.2.20, respectively, Corollary 7.2.22 that (β_k) is uniformly bounded away from 1 for all $k \in \mathbb{N}_0$, which yields the contradiction $\varepsilon_k \to 0^+$ for $k \to \infty$. Hence, there holds $\varepsilon_* = 0$, i.e., $\varepsilon_k \to 0^+$. Due to $(\varepsilon_k, \mu_k) \in \mathcal{P}_=$ for all $k \in \mathbb{N}_0$ this implies $\mu_k \to 0^+$ for $k \to \infty$. □

To estimate the rate of convergence of (μ_k) we use the following result.

Lemma 7.2.25. *Let (ε_k) and (μ_k) be generated by Version B of Algorithm $LSM_{(P)}$ and denote by $\sigma_k \in (0, 1)$ the ratio $\sigma_k := \frac{\mu_{k+1}}{\mu_k}$. Then for all $k \in \mathbb{N}_0$ it holds*

$$\sigma_k = \beta_k^p,$$

where $p = 2$ in case I and $p = 3$ in case II.

Proof. We have

$$\sigma_k = \frac{\mu_{k+1}}{\mu_k} = \frac{\rho^{-1}(\varepsilon_{k+1})}{\rho^{-1}(\varepsilon_k)} = \left(\frac{\beta_k \varepsilon_k}{\varepsilon_k}\right)^p. \qquad \square$$

We now establish convergence of Version B of Algorithm $LSM_{(P)}$. This is one of the main results of this thesis.

Theorem 7.2.26. *We consider Version B of Algorithm LSM$_{(\mathrm{P})}$ with either the backtracking strategy from Corollary 7.2.20 or the one from Corollary 7.2.22. Then it holds:*
Version B of Algorithm LSM$_{(\mathrm{P})}$ generates a sequence $((u^k, \varepsilon_k, \mu_k)) \subset U \times \mathcal{P}_=$ that satisfies $(u^k, \varepsilon_k, \mu_k) \to (\bar{u}, 0^+, 0^+)$ as well as the estimates

$$\frac{|j(u^{k+1}) - j(\bar{u})|}{C_j} \leq \frac{2\mu_k \vartheta(\varepsilon_k)}{C_j} + C\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)$$

and

$$\|u^{k+1} - \bar{u}\|_U \leq \sqrt{\frac{8\mu_k \vartheta(\varepsilon_k)}{C_j \alpha}} + \sqrt{\frac{4C}{\alpha}} \sqrt{\varepsilon_k \left(1 + |\ln \varepsilon_k|\right)}$$

for all $k \in \mathbb{N}_0$ and a constant $C > 0$. Here, α denotes the convexity modulus of j/C_j on $\cup_{\varepsilon \in (0, \varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$.

Moreover, we have

$$\varepsilon_{k+1} = \beta_k \varepsilon_k \quad \text{and} \quad \mu_{k+1} = \sigma_k \mu_k$$

for all $k \in \mathbb{N}_0$, where $\beta_k \in (0, 1)$ satisfies for all k the estimate from Corollary 7.2.20 or Corollary 7.2.22, respectively, and $\sigma_k \in (0, 1)$ satisfies $\sigma_k = \beta_k^p$ for all k, with $p = 2$ in case I and $p = 3$ in case II.

Also, the number of iterations required by LSMSUB in iteration $k \in \mathbb{N}_0$ is bounded from above by

$$\left\lfloor 10.79 \left(f_{\varepsilon_k, \mu_k}(u^k) - f_{\varepsilon_k, \mu_k}(\bar{u}_{\varepsilon_k, \mu_k}) \right) \right\rfloor + \left\lceil 7.13 + 1.45 \ln \left| \ln \sqrt{2\phi(\varepsilon_k)} \right| \right\rceil .$$

Proof. The estimates follow from Lemma 6.1.6, where we used $\lambda_{\varepsilon_k, \mu_k}(u^{k+1}) \leq \phi(\varepsilon_k) \leq \frac{1}{4}$, cf. Lemma 7.2.12. The remaining assertions follow from Lemma 7.2.6, Corollary 7.2.20, Corollary 7.2.22, Corollary 7.2.24, and Lemma 7.2.25. □

Remark 7.2.27. In the above theorem the number of iterations required by LSMSUB in each iteration of LSM$_{(\mathrm{P})}$ can still tend to infinity at an arbitrarily fast rate. Note that this number equals the number of Newton steps required by LSMSUB. Therefore, we consider it a convergence result rather than a result on the rate of convergence.

Remark 7.2.28. As for Version A we remark that the above estimates imply $|j(u^{k+1}) - j(\bar{u})| = \mathcal{O}(\varepsilon_k(1 + |\ln \varepsilon_k|))$ and $\|u^{k+1} - \bar{u}\|_U = \mathcal{O}(\sqrt{\varepsilon_k(1 + |\ln \varepsilon_k|)})$ and that in case II we have $j = \hat{j}$, i.e., estimates for j are estimates for the objective of the original reduced problem ($\mathrm{P}_{\mathrm{red}}$). In case I estimates for j can be transferred to $\hat{j}$ as described in Remark 5.2.6.

7.2.3. Rate of convergence and complexity of Version B of LSM$_{(\mathrm{P})}$

In this section we establish a rate of convergence and complexity estimates for Version B of LSM$_{(\mathrm{P})}$ with the backtracking from Corollary 7.2.22. Using similar arguments it would also be possible to derive a rate of convergence and complexity estimates for Version B of LSM$_{(\mathrm{P})}$ with the backtracking from Corollary 7.2.20.

Version B of $\mathrm{LSM}_{(\mathrm{P})}$ consists of two parts: An inner iteration that is given by LSMSUB and an outer iteration in which ε and μ are decreased. To obtain a meaningful rate of convergence we need to derive estimates for both parts of the algorithm. In view of Theorem 7.2.26 it, therefore, only remains to examine Algorithm LSMSUB, more precisely we need to estimate the quantity $f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})$. This is done in the subsequent lemma.

Lemma 7.2.29. *We consider Version B of Algorithm $LSM_{(P)}$ with the backtracking strategy from Corollary 7.2.22. Moreover, we choose $\phi(\varepsilon) = \min\{\varepsilon^{\frac{1}{2}}/(1+|\ln\varepsilon|)^{\frac{1}{4}}, \theta\}$ in case I and $\phi(\varepsilon) = \min\{\varepsilon^{\frac{3}{4}}, \theta\}$ in case II. Then there exists $C > 0$ such that for all $k \in \mathbb{N}_0$ it holds*

$$f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \leq C.$$

Remark 7.2.30. By use of a slightly more restrictive $\phi(\varepsilon)$ and a slightly more restrictive backtracking technique we could prove that $f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})$ converges to zero for $k \to \infty$. However, the estimate from Lemma 7.2.6 shows that the number of iterations required by LSMSUB tends to infinity, anyway, if $\phi(\varepsilon)$ converges to zero for $\varepsilon \to 0^+$. Therefore, a further improvement of the bound for $f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k})$ does not yield a better order for the number of iterations required by LSMSUB.

Proof. It suffices to establish the assertion for all $k \in \mathbb{N}_0$ sufficiently large. Let c denote the constant from Lemma 6.6.6. We show that for every k sufficiently large $\tilde{\beta}_0$ from the backtracking satisfies $\tilde{\beta}_0\varepsilon_k \in I(\varepsilon_k, \mu_k, u^{k+1})$. This implies $\tilde{\beta}_i\varepsilon_k \in I(\varepsilon_k, \mu_k, u^{k+1})$ for all $i \in \mathbb{N}_0$ and all these k, which yields $\varepsilon_{k+1} \in I(\varepsilon_k, \mu_k, u^{k+1})$ for all k sufficiently large. Note that $I(\varepsilon_k, \mu_k, u^{k+1})$ is well-defined since $\lambda_{\varepsilon_k,\mu_k}(u^{k+1}) \leq \phi(\varepsilon_k) \leq \theta \leq \frac{1}{4}$ implies $u^{k+1} \in \Lambda_{\varepsilon_k,\mu_k}$ for all $k \in \mathbb{N}_0$. To deduce $\tilde{\beta}_0\varepsilon_k \in I(\varepsilon_k, \mu_k, u^{k+1}) = [\varepsilon_k(1 - \frac{c\mu_k\vartheta(\varepsilon_k)}{a(\varepsilon_k)}), \varepsilon_k]$ we have to prove $1 - \frac{\tilde{c}\mu_k}{a(\varepsilon_k)(1+|\ln\varepsilon_k|)} = \tilde{\beta}_0 \geq 1 - \frac{c\mu_k\vartheta(\varepsilon_k)}{a(\varepsilon_k)}$ for all k sufficiently large. This inequality follows from $\vartheta(\varepsilon) \geq \frac{\tilde{C}}{\varepsilon}$ for all $\varepsilon \in (0, \varepsilon_s]$ and a suitable $\tilde{C} > 0$ since we know from Corollary 7.2.24 that $\varepsilon_k, \mu_k \to 0^+$ for $k \to \infty$. In conclusion, we have established $\varepsilon_{k+1} \in I(\varepsilon_k, \mu_k, u^{k+1})$ for all k sufficiently large. We now show $f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \leq C$ for all k sufficiently large, which finishes the proof. We do this in two steps.

1) Without loss of generality we assume $k \geq 1$. Inserting $\varepsilon := \varepsilon_{k-1}$, $\mu := \mu_{k-1}$, $\varepsilon_+ := \varepsilon_k$, and $\mu_+ := \mu_k$ into Lemma 6.10.1 we obtain the existence of $C > 0$ such that it holds for all sufficiently large k

$$f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_{k-1},\mu_{k-1}}) \leq C\left(\lambda_{\varepsilon_{k-1},\mu_{k-1}}(u^k)\right)^2\left(\lambda_{\varepsilon_{k-1},\mu_{k-1}}(u^k) + \sqrt{\frac{\vartheta(\varepsilon_{k-1})}{\varepsilon_{k-1}}}\right),$$

where $C > 0$ is independent of k. Lemma 6.10.1 can be applied since u^k satisfies $\lambda_{\varepsilon_{k-1},\mu_{k-1}}(u^k) \leq \phi(\varepsilon_{k-1}) \leq \frac{1}{4}$, i.e., $(\varepsilon_{k-1}, \mu_{k-1}, u^k) \in T(\Lambda_{\varepsilon_{k-1},\mu_{k-1}})$, cf. Lemma 7.2.12, and since $\varepsilon_k \in I(\varepsilon_{k-1}, \mu_{k-1}, u^k)$ for sufficiently large k, as argued above. Using $\frac{\vartheta(\varepsilon)}{\varepsilon} \leq \frac{C(1+|\ln\varepsilon|)}{\varepsilon^2}$ in case I and $\frac{\vartheta(\varepsilon)}{\varepsilon} = \frac{C}{\varepsilon^3}$ in case II, where C is independent of $\varepsilon \in (0, \varepsilon_s]$, we obtain

$$f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_{k-1},\mu_{k-1}}) \leq C \tag{7.3}$$

for all sufficiently large k from $\lambda_{\varepsilon_{k-1},\mu_{k-1}}(u^k) \leq \phi(\varepsilon_{k-1})$ and the definition of ϕ using $\varepsilon_k \to 0^+$ for $k \to \infty$.

2) Since we have $I(\varepsilon_{k-1}, \mu_{k-1}, u^k) = I(\varepsilon_{k-1}, \mu_{k-1}, \bar{u}_{\varepsilon_{k-1},\mu_{k-1}})$, cf. Lemma 6.6.6, it follows that $\varepsilon_k \in I(\varepsilon_{k-1}, \mu_{k-1}, \bar{u}_{\varepsilon_{k-1},\mu_{k-1}})$ holds for all k sufficiently large. Therefore, Lemma 6.10.3 is applicable for all these k. This yields a k-independent constant $C > 0$ such that

$$f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_{k-1},\mu_{k-1}}) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \leq \frac{Ca(\varepsilon_{k-1})\,|\ln(c\mu_{k-1}\vartheta(\varepsilon_{k-1}))|}{\mu_{k-1}\varepsilon_{k-1}}\,|\varepsilon_{k-1} - \varepsilon_k|$$

is satisfied for all k large enough. We have $\varepsilon_{k-1} - \varepsilon_k = (1-\beta_{k-1})\varepsilon_{k-1}$ and $\beta_{k-1} \geq 1 - \frac{\tilde{c}\mu_{k-1}}{a(\varepsilon_{k-1})(1+|\ln\varepsilon_{k-1}|)}$ due to the backtracking we employ. Hence, we obtain $|\varepsilon_{k-1} - \varepsilon_k| \leq \frac{\tilde{c}\varepsilon_{k-1}\mu_{k-1}}{a(\varepsilon_{k-1})(1+|\ln\varepsilon_{k-1}|)}$. This implies

$$f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_{k-1},\mu_{k-1}}) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \leq \frac{C\tilde{c}\,|\ln(c\mu_{k-1}\vartheta(\varepsilon_{k-1}))|}{1+|\ln\varepsilon_{k-1}|}$$

for all k sufficiently large. To establish

$$f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_{k-1},\mu_{k-1}}) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \leq C \tag{7.4}$$

for all k sufficiently large and a suitable $C > 0$, it, thus, suffices to argue that the term $|\ln(c\mu_{k-1}\vartheta(\varepsilon_{k-1}))|/(1+|\ln\varepsilon_{k-1}|)$ is bounded from above independently of k. Due to $|\ln(c\mu_{k-1}\vartheta(\varepsilon_{k-1}))| \leq |\ln c| + |\ln(\mu_{k-1}\vartheta(\varepsilon_{k-1}))|$ we only need to establish this for the term $|\ln(\mu_{k-1}\vartheta(\varepsilon_{k-1}))|/(1+|\ln\varepsilon_{k-1}|)$. Since it holds $\mu_{k-1}\vartheta(\varepsilon_{k-1}) \to 0^+$ for $k \to \infty$ and $\mu_{k-1}\vartheta(\varepsilon_{k-1}) \geq c\varepsilon_{k-1}$ with a k-independent constant $c > 0$, this is elementary to see.

Summarizing, (7.3) and (7.4) show $f_{\varepsilon_k,\mu_k}(u^k) - f_{\varepsilon_k,\mu_k}(\bar{u}_{\varepsilon_k,\mu_k}) \leq C$ for all k sufficiently large. □

We conclude this section with the following theorem that presents a detailed description of the convergence behaviour of Version B of Algorithm $\mathrm{LSM}_{(\mathrm{P})}$ with the backtracking from Corollary 7.2.22.

Theorem 7.2.31. *Version B of Algorithm $\mathrm{LSM}_{(\mathrm{P})}$ with the backtracking strategy from Corollary 7.2.22 and the choice $\phi(\varepsilon) = \min\{\varepsilon^{\frac{1}{2}}/(1+|\ln\varepsilon|)^{\frac{1}{4}}, \theta\}$ in case I and $\phi(\varepsilon) = \min\{\varepsilon^{\frac{3}{4}}, \theta\}$ in case II generates a sequence $((u^k, \varepsilon_k, \mu_k)) \subset U \times \mathcal{P}_=$ that satisfies $(u^k, \varepsilon_k, \mu_k) \to (\bar{u}, 0^+, 0^+)$ as well as the estimates*

$$\frac{|j(u^{k+1}) - j(\bar{u})|}{C_j} \leq \frac{2\mu_k\vartheta(\varepsilon_k)}{C_j} + C\varepsilon_k\,(1+|\ln\varepsilon_k|)$$

and

$$\|u^{k+1} - \bar{u}\|_U \leq \sqrt{\frac{8\mu_k\vartheta(\varepsilon_k)}{C_j\alpha}} + \sqrt{\frac{4C}{\alpha}}\,\sqrt{\varepsilon_k\,(1+|\ln\varepsilon_k|)}$$

for all $k \in \mathbb{N}_0$ and a constant $C > 0$. Here, α denotes the convexity modulus of j/C_j on $\cup_{\varepsilon\in(0,\varepsilon_s]} M(\varepsilon) = M(\varepsilon_s)$.

Moreover, we have

$$\varepsilon_{k+1} = \beta_k\varepsilon_k \quad \text{and} \quad \mu_{k+1} = \sigma_k\mu_k$$

for all $k \in \mathbb{N}_0$, where $\beta_k \in (0,1)$ satisfies for all k the estimate

$$\beta_k \leq 1 - \frac{c\varepsilon_k^{p-1}}{(1+|\ln \varepsilon_k|)^2}$$

with a k-independent $c > 0$ and $\sigma_k \in (0,1)$ satisfies $\sigma_k = \beta_k^p$ for all k. Here, it holds $p = 2$ in case I and $p = 3$ in case II.

Also, the number of iterations required by LSMSUB in iteration $k \in \mathbb{N}_0$ is bounded from above by

$$C + \left\lceil 1.45 \ln \left| \ln \sqrt{2\phi(\varepsilon_k)} \right| \right\rceil,$$

where $C > 0$ is independent of k. This is, $\mathrm{LSM}_{(\mathrm{P})}$ requires at most $C+1+\lceil 1.45 \ln|\ln \sqrt{2\phi(\varepsilon_k)}|\rceil$ Newton steps in iteration k.

Finally, for every $\delta \in (0,1)$ there exist $\tilde{C}, \tilde{c} > 0$ such that for all $\tau > 0$ we have the complexity estimates

$$K \geq \left\lceil \frac{\left(\frac{\tilde{C}}{\tau}\right)^{\frac{r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{\tilde{c}r} \right\rceil \qquad \Longrightarrow \qquad \left| j(u^{K+1}) - j(\bar{u}) \right| \leq \tau$$

and

$$K \geq \left\lceil \frac{\left(\frac{\tilde{C}}{\tau}\right)^{\frac{2r}{1-\delta}} - \frac{1}{\varepsilon_0^r}}{\tilde{c}r} \right\rceil \qquad \Longrightarrow \qquad \left\| u^{K+1} - \bar{u} \right\|_U \leq \tau.$$

Here, it holds $r = p - 1 + \delta$ with $p = 2$ in case I and $p = 3$ in case II.

Remark 7.2.32. Note that for $k \to \infty$ the term $\ln|\ln \sqrt{2\phi(\varepsilon_k)}|$ goes to infinity. However, since its order is $\mathcal{O}(\ln|\ln \sqrt{2\varepsilon_k}|)$ for $k \to \infty$, it grows very slowly. In practice, we can assume that the number of iterations of LSMSUB is bounded finitely, cf. also Remark 7.2.7.

Remark 7.2.33. The complexity estimates show, for instance, that in case I we have $|j(u^{K+1}) - j(\bar{u})| \leq \tau$ for small values of τ after approximately $\mathcal{O}(\tau^{-1})$ iterations and $\|u^{K+1} - \bar{u}\|_U \leq \tau$ after roughly $\mathcal{O}(\tau^{-2})$ iterations. Of course, these are worst-case bounds.

Proof. All assertions except the complexity estimates follow from Theorem 7.2.26 in combination with Lemma 7.2.29. To establish the complexity estimates we choose $\tilde{c} > 0$ so small that it holds $\beta_k \leq 1 - \tilde{c}\varepsilon_k^{p-1+\delta}$ for all $k \in \mathbb{N}_0$. This is possible due to the estimate we have for β_k in the above theorem and since $\varepsilon_k^\delta \leq 1/(1+|\ln \varepsilon_k|)^2$ is satisfied for all k sufficiently large, as is elementary to see. The complexity estimates can now be proven exactly as in Lemma 7.1.13 the only difference being the different value of r. □

7.3. Phase one

In this section we describe phase one methods for Algorithm $\mathrm{SSM}_{(\mathrm{P})}$ and Algorithm $\mathrm{LSM}_{(\mathrm{P})}$. This is, we show how to obtain a starting point for Algorithm $\mathrm{SSM}_{(\mathrm{P})}$ and Algorithm $\mathrm{LSM}_{(\mathrm{P})}$ if

only a point $u^0 \in U_{\mathrm{ad}}(\varepsilon_0)$ is known and provide complexity estimates for this process. Note that for $\mathrm{LSM}_{(\mathrm{P})}$ we only require a starting point $u^0 \in U_{\mathrm{ad}}(\varepsilon_0)$ since $\mathrm{LSM}_{(\mathrm{P})}$ calls Algorithm LSMSUB, which is used as a phase one method if it is the first iteration of $\mathrm{LSM}_{(\mathrm{P})}$. However, in this section we obtain a complexity result that is more precise than the one from Theorem 7.2.31.

7.3.1. Phase one based on a short step method

The following Algorithm APOSS determines $\tilde{u} \in U_{\mathrm{ad}}(\varepsilon_0)$ with $\lambda_{\varepsilon_0,\mu_0}(\tilde{u}) \leq \theta$. We use the name APOSS since this algorithm is the same as the one from Section 5.4.1.

Algorithm APOSS (phase one based on short steps applied to $f_{\varepsilon,\mu}$)

Input: Parameters $(\varepsilon_0, \mu_0) \in \mathcal{P}_=$, $\theta \in (0, \frac{1}{4}]$, starting point $u^0 \in U_{\mathrm{ad}}(\varepsilon_0)$.

Output: $\tilde{u} \in \Lambda_{\varepsilon_0,\mu_0}(\theta)$.

Define $f_{\mu_0} := f_{\varepsilon_0,\mu_0}$, $K := U_{\mathrm{ad}}(\varepsilon_0)$, and set for $\nu > 0$

$$f_{\nu,\mu_0,u^0} : K \to \mathbb{R}, \quad f_{\nu,\mu_0,u^0}(u) := f_{\mu_0}(u) - \frac{f'_{\mu_0}(u^0)[u]}{\nu}.$$

CALL Algorithm POSS from Section 2.9.1 with (u^0, μ_0, θ) (use $X = U$, $I_s = (0, \mu_0]$ in Algorithm POSS) and denote its return value by $\tilde{u}$.
RETURN $\tilde{u}$.

Remark 7.3.1. Algorithm POSS requires the self-boundedness constant $\vartheta_{f_{\varepsilon_0,\mu_0}}$ of $f_{\mu_0} = f_{\varepsilon_0,\mu_0}$. From Lemma 6.1.6 we know that this constant is given by

$$\vartheta_{f_{\varepsilon_0,\mu_0}} = \frac{C_j}{\mu_0} + \tau(\varepsilon_0) \text{ in case I} \quad \text{and} \quad \vartheta_{f_{\varepsilon_0,\mu_0}} = 2\left(\frac{C^2 C_{\|\cdot\|}}{\mu_0^2 \tilde{\tau}(\varepsilon_0)} + \tau(\varepsilon_0) + \tilde{\tau}(\varepsilon_0)\right) \text{ in case II,}$$

where C satisfies $\|\hat{j}'(u)\|_{U^*} \leq C$ for all $u \in U_{\mathrm{ad}}(\varepsilon_0)$.

We have the following complexity result for Algorithm APOSS. We recall that $\mathrm{sym}(u^0, U_{\mathrm{ad}}(\varepsilon_0))$, the symmetry of $U_{\mathrm{ad}}(\varepsilon_0)$ about u^0, is introduced in Definition 2.5.25.

Theorem 7.3.2. *Algorithm APOSS returns a $\tilde{u}$ that satisfies $\lambda_{\varepsilon_0,\mu_0}(\tilde{u}) \leq \theta$ after $N \in \mathbb{N}_0$ iterations of Algorithm POSS, where N is bounded from above by*

$$N \leq \left\lceil \frac{17}{16} \cdot \frac{\sqrt{\vartheta_{f_{\varepsilon_0,\mu_0}}}}{\delta} \cdot \ln\left(\frac{2\vartheta_{f_{\varepsilon_0,\mu_0}}}{\theta}\left(1 + \frac{1}{\mathrm{sym}(u^0, U_{ad}(\varepsilon_0))}\right)\right)\right\rceil$$

with

$$\vartheta_{f_{\varepsilon_0,\mu_0}} = \frac{C_j}{\mu_0} + \tau(\varepsilon_0) \text{ in case I}, \quad \vartheta_{f_{\varepsilon_0,\mu_0}} = 2\left(\frac{C^2 C_{\|\cdot\|}}{\mu_0^2 \tilde{\tau}(\varepsilon_0)} + \tau(\varepsilon_0) + \tilde{\tau}(\varepsilon_0)\right) \text{ in case II,}$$

and

$$\delta = \frac{\tilde{\theta}\left(1 - \frac{\tilde{\theta}}{(1-\tilde{\theta})^2}\right)}{1 - \frac{\tilde{\theta}}{\sqrt{\vartheta_{f_{\varepsilon_0,\mu_0}}}}}, \text{ where } \tilde{\theta} = \frac{\theta}{2}.$$

Here, C satisfies $\|\hat{j}'(u)\|_{U^*} \leq C$ for all $u \in U_{ad}(\varepsilon_0)$.

During the course of APOSS, $2N+1$ Newton steps have to be computed.

Proof. The boundedness of $U_{\mathrm{ad}}(\varepsilon_0)$ together with Lemma 6.1.6 yields that Theorem 2.9.5 is applicable with constants of self-boundedness as asserted. Theorem 2.9.5 now implies all assertions. □

Remark 7.3.3. We stress that in $\mathrm{SSM}_{(\mathrm{P})}$ and $\mathrm{LSM}_{(\mathrm{P})}$ we usually start with a relatively large ε_0 in comparison to SSM_ε and LSM_ε, where ε is chosen small from the beginning. This implies that the complexity of phase one may be much worse for SSM_ε and LSM_ε than for $\mathrm{SSM}_{(\mathrm{P})}$ and $\mathrm{LSM}_{(\mathrm{P})}$.

7.3.2. Phase one based on a long step method

The following Algorithm APOLS determines $\tilde{u} \in U_{\mathrm{ad}}(\varepsilon_0)$ with $\lambda_{\varepsilon_0,\mu_0}(\tilde{u}) \leq \theta$. We use the name APOLS since this algorithm is the same as the one from Section 5.4.2.

Algorithm APOLS (phase one based on long steps applied to $f_{\varepsilon,\mu}$)

Input: Parameters $(\varepsilon_0, \mu_0) \in \mathcal{P}_=$, $\theta \in (0, \frac{1}{4}]$, starting point $u^0 \in U_{\mathrm{ad}}(\varepsilon_0)$.

Output: $\tilde{u} \in \Lambda_{\varepsilon_0,\mu_0}(\theta)$.

Define $f_{\mu_0} := f_{\varepsilon_0,\mu_0}$ and $K := U_{\mathrm{ad}}(\varepsilon_0)$.

Compute the Newton step $s^0 \in U$ by solving $f''_{\mu_0}(u^0)[s^0] = -f'_{\mu_0}(u^0)$ in U^*.
CALL Algorithm LSMSUB from Section 2.7 with $(u^0, s^0, \mu_0, \theta)$ (use $X = U$, $I_s = (0, \mu_0]$ in LSMSUB), and denote its return value by $\tilde{u}$.
RETURN $\tilde{u}$.

We have the following complexity result for Algorithm APOLS.

Theorem 7.3.4. *Algorithm APOLS returns a $\tilde{u}$ that satisfies $\lambda_{\varepsilon_0,\mu_0}(\tilde{u}) \leq \theta$ after $N \in \mathbb{N}_0$ iterations of LSMSUB, where N is bounded from above by*

$$N \leq \left\lfloor 10.79\vartheta_{f_{\varepsilon_0,\mu_0}} \left|\ln\left(1 - \omega_{\bar{u}_{\varepsilon_0,\mu_0}}(u^0)\right)\right| \right\rfloor + \left\lceil 7.13 + 1.45 \ln\left|\ln\sqrt{2\theta}\right| \right\rceil.$$

Here, $\omega_{\bar{u}_{\varepsilon_0,\mu_0}} : U_{ad}(\varepsilon_0) \to [0,1)$ denotes the Minkowski function, see Definition 2.3.14, and $\vartheta_{f_{\varepsilon_0,\mu_0}}$ denotes the same constant as in Theorem 7.3.2.

During the course of APOLS, $N+1$ Newton steps have to be computed.

Proof. Due to Lemma 6.1.6, Theorem 2.9.6 is applicable with constants of self-boundedness as asserted. Theorem 2.9.6 now implies all assertions. □

8. Numerics

In this section we deal with numerical aspects of the developed algorithms and present numerical results. It is well-known that in practice short step methods are usually inferior to long step methods. Therefore, we deal with Algorithm LSM_ε and $\mathrm{LSM}_{(\mathrm{P})}$ in this section and leave SSM_ε and $\mathrm{SSM}_{(\mathrm{P})}$ aside. All examples that we consider are treated within the framework of case I.

8.1. Discretization

We used MATLAB to implement Algorithm LSM_ε and $\mathrm{LSM}_{(\mathrm{P})}$. Before we explain the discretization strategy in more detail, let us outline the most important parts. For the discretization of the Laplace operator in two dimensions we use a standard regular five-point finite difference stencil with mesh size h. Most of our numerical examples use the unit square $\Omega = (0,1) \times (0,1)$ as domain for the state equation, where this discretization strategy is equivalent to the use of linear finite elements on a uniformly triangulated grid, cf. [GR05, Beispiel 3.4, pp. 104]. Furthermore, the state constraints act on all of Ω, i.e. $\Omega = \Omega_a$. For integration on the unit square we use the weights h^2 for all nodes in Ω. To develop fast code several techniques are helpful. For instance, by suitably rewriting the Newton system and using structural properties of the corresponding Hessians, the main costs for the computation of a Newton step in our test problems amount to two solves of a linear system of the form $(A^*A + \gamma \cdot I)y = f$, with $\gamma \in C^{0,\beta}(\overline{\Omega})$. If, for instance, $A = -\Delta$, then this system corresponds to a biharmonic equation. Also, only the right-hand side f varies, which further diminishes the numerical costs.

8.1.1. Discretization strategy

We use the same discretization strategy for all numerical examples. To explain this strategy we consider the model problem

$$\min_{(y,u)\in Y\times U} \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_{L^2(\Omega)} \quad \text{s.t.} \quad y \geq y_a \text{ in } \overline{\Omega}_a, \quad \begin{cases} -\Delta y = u & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases} \tag{MP}$$

with $Y = H^2(\Omega) \cap H^1_0(\Omega)$, $U = L^2(\Omega)$, $\Omega = \Omega_a = (0,1)^2 \subset \mathbb{R}^2$, $y_d \in L^2(\Omega)$, $y_a \in C^{0,\beta}(\overline{\Omega}_a)$, and $\hat{\alpha} > 0$.

To discretize this problem we use a uniform grid on $\overline{\Omega}$ with $(N+1)^2$ points to cover $\overline{\Omega}$. This grid has mesh width $h = 1/N$. We number the points from the origin to the right, i.e., $x_0 = (0,0)$,

$x_1 = (h, 0)$, …, $x_N = (1, 0)$, $x_{N+1} = (0, h)$, …, $x_{2N+1} = (1, h)$, …, $x_{(N+1)^2-1} = (1, 1)$. The points that belong to Ω are exactly the x_i with $i \in I$, where I denotes the index set

$$I := \left\{0, 1, 2, \ldots, (N+1)^2 - 1\right\} \setminus \tilde{I}$$

with

$$\begin{aligned}\tilde{I} := &\{0, 1, 2, \ldots, N\} \\ &\cup \{N+1, 2(N+1)-1, 2(N+1), 3(N+1)-1, 3(N+1), \ldots, N(N+1)-1\} \\ &\cup \left\{N(N+1), N(N+1)+1, N(N+1)+2, \ldots, (N+1)^2 - 1\right\}.\end{aligned}$$

We replace the functions in (MP) by vectors from $\mathbb{R}^{(N-1)^2} = \mathbb{R}^{|I|}$, whose values represent the function values at the nodes in Ω. The replacements are denoted by an additional index h; for example, $y \in Y$ becomes $y_h \in Y_h := \mathbb{R}^{(N-1)^2}$ and $u \in U$ becomes $u_h \in U_h := \mathbb{R}^{(N-1)^2}$. We employ a superscript i to denote the i-th component of a vector. We consider the elements in I to be ordered by size from the smallest to the largest and use $I(i)$ to denote the i-th element with respect to this order. Then y_h^i represents $y(x_{I(i)})$ for $i \in \{1, 2, \ldots, (N-1)^2\}$.

We use the classical five-point stencil to discretize Δ, cf., e.g., [Bra13, Chapter 1]. This is, the PDE in (MP) is replaced by $A_h y_h = u_h$ with

$$A_h := \frac{1}{h^2} \begin{pmatrix} T & -I & & & 0 \\ -I & T & -I & & \\ & -I & \ddots & \ddots & -I \\ & & \ddots & \ddots & \\ 0 & & & -I & T \end{pmatrix} \in \mathbb{R}^{(N-1)^2 \times (N-1)^2},$$

where $I \in \mathbb{R}^{(N-1)\times(N-1)}$ denotes the identity matrix and

$$T := \begin{pmatrix} 4 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}.$$

Note that linear finite elements on a uniformly triangulated grid would result in the same discretization of Δ apart from the scaling $1/h^2$, cf. [GR05, Beispiel 3.4, pp. 104].

To evaluate the integrals in (MP) we use $\int_\Omega f(x)\,\mathrm{d}x \approx h^2 \sum_{i=1}^{(N-1)^2} f(x_{I(i)})$. This formula can, for instance, be obtained via Fubini's theorem $\int_\Omega f(x)\,\mathrm{d}x = \int_0^1 \int_0^1 f(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2$ and application of the trapezoidal rule to the iterated integrals assuming that f is zero on $\partial\Omega$. Another motivation for this formula is given in [Trö05, Section 2.12.3].

Summarizing, we replace (MP) by the discrete problem

$$\min_{(y_h, u_h) \in Y_h \times U_h} \frac{1}{2} \sum_{i=1}^{(N-1)^2} \left((y_h^i - y_{d,h}^i)^2 + \hat{\alpha}(u_h^i)^2\right) \quad \text{s.t.} \quad y_h^i \geq y_{a,h}^i,\ 1 \leq i \leq (N-1)^2,\ A_h y_h = u_h, \tag{MP$_h$}$$

where we divided the objective by h^2.

To apply an algorithm, e.g., LSM_ε, to (MP_h) we need a discretization of $f_{\varepsilon,\mu}$. This discretization can be derived in the same fashion as above. A particularity is that $f_{\varepsilon,\mu}$ contains the integral $\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}\,\mathrm{d}x$. More generally, at different points in the algorithms that we use we have to evaluate integrals of the form $\int_{\Omega_a} e^{-(y-y_a)/\varepsilon} f\,\mathrm{d}x$ with a function f. This can be a cause of instability if $y - y_a$ is negative, respectively, if $y_h^i - y_{a,h}^i$ is negative for some i. To compute these expressions in a stable manner we employ the identity $\int_{\Omega_a} e^{-(y-y_a)/\varepsilon} f\,\mathrm{d}x = e^{-\min(y-y_a)/\varepsilon} \int_{\Omega_a} e^{-(y-y_a-\min(y-y_a))/\varepsilon} f\,\mathrm{d}x$, where min denotes the pointwise minimum in $\overline{\Omega}_a$. On the discrete level the corresponding identity reads $h^2 \sum_{i=1}^{(N-1)^2} e^{-(y_h^i - y_{a,h}^i)/\varepsilon} f_h^i = e^{-\min(y_h - y_{a,h})/\varepsilon} h^2 \sum_{i=1}^{(N-1)^2} e^{-(y_h^i - y_{a,h}^i - \min(y_h - y_{a,h}))/\varepsilon} f_h^i$, where $\min(y_h - y_{a,h})$ denotes the minimal component of the vector $y_h - y_{a,h}$.

So far we have described the discretization strategy that we use. Let us now cover an important aspect of the implementation, the computation of Newton steps. In doing so, we will also derive the discrete Newton system that we solve.

8.1.2. Efficient computation of Newton steps in the case of a linear state equation

In the following we demonstrate for case I how Newton steps can be computed efficiently since all our numerical examples use case I. We mention that case II can be handled similarly. Also, we allow $\Omega_a \subsetneqq \Omega$ in this section.

The state equations that we consider in the numerical examples are of the form $Ay = u$ with $A \in \mathcal{L}(Y, U)$, where $Y = H^2(\Omega) \cap H_0^1(\Omega)$ and $U = L^2(\Omega)$. Since A is bijective, we can obtain a state reduced version $f_{\varepsilon,\mu}(y)$ instead of the control reduced version that we worked with in the preceding sections. However, since $Ay = u$ couples y and u linearly and bijectively, it can be argued that the sequence of iterates (y^k) obtained by use of the state reduced $f_{\varepsilon,\mu}$, is exactly $(A^{-1}u^k)$, where (u^k) denotes the sequence of iterates generated by use of the control reduced $f_{\varepsilon,\mu}$. This shows that the theory developed in this thesis still applies if we use a state reduced $f_{\varepsilon,\mu}$. In fact, in a former version of this thesis we worked with reduction to the state since that seemed to simplify the theory a bit. While it is a standard requirement that there is a unique state for every control, this is not the case the other way round, so we changed the presentation and now use reduction to the control in this thesis to be more general.

We start by considering the Newton system for $f_{\varepsilon,\mu}$ in function space. From this system we derive a discretized Newton system, which is the one that is actually solved to compute the Newton step in our implementation. We remark that the Newton step that results from this process is the same as the one we obtain by first discretizing $f_{\varepsilon,\mu}$ and then computing the Newton step for the discretized $f_{\varepsilon,\mu}$.

The Newton system reads $f''_{\varepsilon,\mu}(y)[n_y] = -f'_{\varepsilon,\mu}(y)$ for a given $y \in Y_{\mathrm{ad}}(\varepsilon)$. Since we have

$f_{\varepsilon,\mu}(y) = -C_j \frac{\ln(C_{\hat{j}} - \hat{j}(y))}{\mu} - \tau(\varepsilon)\ln(B^\varepsilon_{C(\overline{\Omega}_a)}(y))$, we obtain for the derivatives

$$f'_{\varepsilon,\mu}(y)[h_1] = \frac{C_j}{\mu}\frac{\hat{j}'(y)[h_1]}{\tilde{B}(y)} - \tau(\varepsilon)\frac{(B^\varepsilon_{C(\overline{\Omega}_a)})'(y)[h_1]}{B^\varepsilon_{C(\overline{\Omega}_a)}(y)}$$

and

$$\begin{aligned} f''_{\varepsilon,\mu}(y)[h_1,h_2] &= \frac{C_j}{\mu}\left(\frac{\hat{j}''(y)[h_1,h_2]}{\tilde{B}(y)} + \frac{\hat{j}'(y)[h_1]\cdot\hat{j}'(y)[h_2]}{\left(\tilde{B}(y)\right)^2}\right) \\ &\quad - \tau(\varepsilon)\left(\frac{(B^\varepsilon_{C(\overline{\Omega}_a)})''(y)[h_1,h_2]}{B^\varepsilon_{C(\overline{\Omega}_a)}(y)} - \frac{(B^\varepsilon_{C(\overline{\Omega}_a)})'(y)[h_1]\cdot(B^\varepsilon_{C(\overline{\Omega}_a)})'(y)[h_2]}{\left(B^\varepsilon_{C(\overline{\Omega}_a)}(y)\right)^2}\right), \end{aligned}$$

where $\hat{j}(y)$ denotes the state reduced objective and $\tilde{B}(y) := C_{\hat{j}} - \hat{j}(y)$. In the numerical examples we always use $\hat{J}(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_{L^2(\Omega)}$ as objective function, i.e., $\hat{j}(y) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|Ay\|^2_{L^2(\Omega)}$. Thus, we have

$$\hat{j}'(y)[h_1] = (y - y_d, h_1)_{L^2(\Omega)} + \hat{\alpha}(Ay, Ah_1)_{L^2(\Omega)},$$

$$\hat{j}''(y)[h_1,h_2] = (h_1,h_2)_{L^2(\Omega)} + \hat{\alpha}(Ah_1, Ah_2)_{L^2(\Omega)},$$

$$\left(B^\varepsilon_{C(\overline{\Omega}_a)}\right)'(y)[h_1] = \frac{\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}h_1\,\mathrm{d}x}{\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}\,\mathrm{d}x},$$

and

$$\left(B^\varepsilon_{C(\overline{\Omega}_a)}\right)''(y)[h_1,h_2] = \frac{1}{\varepsilon}\left(\frac{\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}h_1\,\mathrm{d}x\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}h_2\,\mathrm{d}x}{\left(\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}\,\mathrm{d}x\right)^2} - \frac{\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}h_1h_2\,\mathrm{d}x}{\int_{\Omega_a} e^{-(y-y_a)/\varepsilon}\,\mathrm{d}x}\right).$$

We note that here and in the following all derivatives are Fréchet derivatives, as can be established by arguments similar to those used in the control reduced case or, alternatively, the chain rule. Moreover, we point out that the state reduced $f_{\varepsilon,\mu}$ is uniformly convex since this is true for $\hat{j}(y)$. To infer that this suffices, see Lemma C.4.14. To infer uniform convexity of $\hat{j}$ note that the bijectivity of $A : Y \to U$ implies $\|Ah\|_U \geq \frac{\|h\|_Y}{\|A^{-1}\|_{\mathcal{L}(U,Y)}}$. The uniform convexity of $f_{\varepsilon,\mu}$ implies that Newton's equation is uniquely solvable, cf. Theorem C.4.15. Alternatively, this also follows from the fact that the control reduced version of Newton's equation is uniquely solvable in combination with the bijectivity of A.

Setting $q := e^{-(y-y_a)/\varepsilon}$ on Ω_a and $q \equiv 0$ on $\Omega \setminus \Omega_a$, we can write Newton's equation as operator equation $f''_{\varepsilon,\mu}(y)[n_y] = -f'_{\varepsilon,\mu}(y)$ in Y^* using the $L^2(\Omega)$ scalar product as well as

$$f'_{\varepsilon,\mu}(y) = \frac{C_j}{\mu\tilde{B}(y)}(y - y_d + \hat{\alpha}A^*Ay) - \frac{\tau(\varepsilon)}{B^\varepsilon_{C(\overline{\Omega}_a)}(y)}\frac{q}{\int_{\Omega_a} q\,\mathrm{d}x} \tag{8.1}$$

and

$$f''_{\varepsilon,\mu}(y)[h_1] = \frac{C_j}{\mu\tilde{B}(y)}\left(h_1 + \hat{\alpha}A^*Ah_1 + \frac{(y - y_d + \hat{\alpha}A^*Ay)\cdot\int_\Omega (y - y_d + \hat{\alpha}A^*Ay)h_1\,\mathrm{d}x}{\tilde{B}(y)}\right) - \frac{\tau(\varepsilon)}{\varepsilon B^\varepsilon_{C(\overline{\Omega}_a)}(y)\int_{\Omega_a} q\,\mathrm{d}x}\left(\frac{q\int_{\Omega_a} qh_1\,\mathrm{d}x}{\int_{\Omega_a} q\,\mathrm{d}x}\cdot\left(1 - \frac{\varepsilon}{B^\varepsilon_{C(\overline{\Omega}_a)}(y)}\right) - qh_1\right).$$

We now replace the quantities in the Newton system $f''_{\varepsilon,\mu}(y)[n_y] = -f'_{\varepsilon,\mu}(y)$ by their discrete counterparts, e.g., functions are replaced by their node vectors as described in the previous section. For conciseness we perform these replacements without change of notation. This is, from now on y, y_d, etc. denote vectors, A and $A^* = A^T$ denote matrices, and integrals are to be read in the sense of the quadrature formula described in the previous section. We emphasize again that the discrete Newton system that results from this process is the same as the one we obtain if we compute the Newton system for the discretized $f_{\varepsilon,\mu}$.

We assume that the uniform mesh used for discretization contains M nodes in Ω. In the notation of the previous section we would have, e.g., $M = (N-1)^2$ for the unit square. Introducing the abbreviations

$$c_1 := \frac{C_j}{\mu\tilde{B}(y)} \qquad \text{and} \qquad c_2 := \frac{\tau(\varepsilon)}{\varepsilon B^\varepsilon_{C(\overline{\Omega}_a)}(y)\int_{\Omega_a} q\,\mathrm{d}x},$$

$f''_{\varepsilon,\mu}(y)[n_y]$ takes the form $(E + v_1w_1^T + v_2w_2^T)n_y$ with

$$E := c_1\left(I + \hat{\alpha}A^TA\right) + c_2\mathrm{diag}(q), \tag{8.2}$$

where $\mathrm{diag}(q) \in \mathbb{R}^{M\times M}$ denotes the square matrix with zero entries and diagonal q,

$$v_1 := c_1\frac{y - y_d + \hat{\alpha}A^TAy}{\tilde{B}(y)}, \quad w_1 := Q^T(y - y_d + \hat{\alpha}A^TAy),$$

and

$$v_2 := c_2\left(\frac{\varepsilon}{B^\varepsilon_{C(\overline{\Omega}_a)}(y)} - 1\right)q, \quad w_2 := \frac{Q^Tq}{\int_{\Omega_a} q\,\mathrm{d}x}.$$

Here, $Q \in \mathbb{R}^{M\times M}$ denotes the matrix whose entries are the weights for integration, i.e., $f_h^TQg_h \approx \int_\Omega fg\,\mathrm{d}x$ if the vectors f_h and g_h represent the nodal values of functions f and g in Ω. Note that Q does not have to be assembled, as we explain below. In effect, we have to solve a linear system of the form $(E + v_1w_1^T + v_2w_2^T)n_y = r$, with r the discretization of $-f'_{\varepsilon,\mu}(y)$ from (8.1). The rank-1 updates $v_1w_1^T$ and $v_2w_2^T$ can be treated with the Sherman-Morrison formula, as we explain in a moment. This leaves as crucial step the solution of a linear system of the form $Es = b$. The coefficient matrix E contains A^TA. However, since A^TA has the squared condition of A, we want to avoid working with A^TA. To this end, we remark that it is possible to solve instead of $Es = b$ the linear system

$$\begin{pmatrix} \frac{1}{\hat{\alpha}}\mathrm{diag}(1 + \frac{c_2}{c_1}q) & A^T \\ A & -I \end{pmatrix}\begin{pmatrix} s \\ \tilde{s} \end{pmatrix} = \begin{pmatrix} \frac{b}{c_1\hat{\alpha}} \\ 0 \end{pmatrix}.$$

This is a 2×2 elliptic system. Systems like this one frequently occur in optimal control, cf., e.g., [HPUU09, Section 2.8.2]. They can be solved efficiently by multigrid methods, cf. [Bra13, Chapter 5]. An alternative to multigrid methods is the preconditioned conjugate gradients method, cf. [GVL07, Section 10.3] and [BBC+93, Section 2 and 3]. This method, as well as others, are MATLAB built-in functions and we tested several of them. However, we observed that MATLAB's sparse backslash operator is more efficient and, therefore, in the implementation for this thesis we use backslash to solve the occuring sparse linear systems. For more on how MATLAB deals with sparsity, cf. [GMS92]. We note that the relative residuals we observed by use of backslash are usually below 10^{-10}, measured in the (discrete) L^2-norm. As an interesting alternative, for instance on very fine meshes or in 3D where backslash may be to expensive, we mention that quite recently preconditioners that are tailored for optimal control with PDEs in the presence of state constraints have been proposed in [HS10] and [SU12].

We handle the two rank-1 updates by the Sherman-Morrison formula. Applying this formula twice we see that we can solve $(E + v_1 w_1^T + v_2 w_2^T) n_y = r$ as follows:

1) Solve $Es_1 = r$, $Es_2 = v_1$, and $Es_3 = v_2$.

2) Define $p_1 := s_1 - \frac{s_2 w_1^T s_1}{1 + w_1^T s_2}$ and $p_2 := s_3 - \frac{s_2 w_1^T s_3}{1 + w_1^T s_2}$. Set $n_y := p_1 - \frac{p_2 w_2^T p_1}{1 + w_2^T p_2}$.

Alternatively, we can use the Sherman-Morrison-Woodbury formula and develop the same routine from inverting $E + VW^T$ with $V = (v_1, v_2) \in \mathbb{R}^{M \times 2}$ and $W = (w_1, w_2) \in \mathbb{R}^{M \times 2}$. We point out that r is a linear combination of v_1 and v_2, so in 1) only the systems $Es_2 = v_1$ and $Es_3 = v_2$ have to be solved. Furthermore, we note that the vectors w_1 and w_2 only occur in 2) and only in scalar products, i.e., in the form $w_1^T s$ and $w_2^T s$. This means that we do not have to assemble the matrix Q with the integration weights; it suffices to evaluate integrals of the form $\int_\Omega (y - y_d + \hat{\alpha} A^T A y) s \, \mathrm{d}x$ and $\int_\Omega qs \, \mathrm{d}x$.

8.1.3. Efficient computation of Newton steps in the case of a semilinear state equation

In one of our numerical examples we consider $\hat{J}(y, u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_{L^2(\Omega)}$ together with the semilinear state equation $-\Delta y + y + y^3 = u$ in Ω, $y = 0$ on $\partial\Omega$. Of course, the nonlinearity y^3 introduces nonconvexity to j and $f_{\varepsilon,\mu}$, and, therefore, the theory developed in this thesis is not applicable. Nonetheless, it is interesting to see how our algorithms perform in this case.

To derive the state reduced Newton system we insert $F : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$, $F(y) := -\Delta y + y + y^3$ for u in $\hat{J}$. This approach is sensible since for every $u \in L^2(\Omega)$ there is a unique solution $y = y(u) \in H^2(\Omega) \cap H_0^1(\Omega)$ of $F(y) = u$, as we argue now. First, we note that $-\Delta y + y + y^3 = u$ possesses a unique weak solution $y = y(u)$ in $H_0^1(\Omega)$ for $\Omega \subset \mathbb{R}^d$, $d \leq 3$, a bounded Lipschitz domain. This can be proven exactly as in [Trö05, Section 7.1.1] using the theory of monotone operators. Alternatively, we can prove this by establishing that for given $u \in L^2(\Omega)$ the Fréchet differentiable functional $G : H_0^1(\Omega) \to \mathbb{R}$, $G(y) := \frac{1}{2}\|\nabla y\|^2_{L^2(\Omega)^d} + \frac{1}{2}\|y\|^2_{L^2(\Omega)} + \frac{1}{4}\|y\|^4_{L^4(\Omega)} - (y, u)_{L^2(\Omega)}$ possesses a unique minimizer $\bar{y}$. This minimizer, of course, satisfies $G'(\bar{y})[h] = 0$ for all $h \in H_0^1(\Omega)$, which states that $\bar{y}$ is the weak solution of the semilinear PDE we are interested in. To demonstrate H^2-regularity

let $\bar{y} \in H_0^1(\Omega)$ denote the unique weak solution of $-\Delta y + y + y^3 = u$. For $d \leq 3$ it holds $H_0^1(\Omega) \hookrightarrow L^6(\Omega)$ and, thus, $\bar{y}^3 \in L^2(\Omega)$. This shows $w := u - \bar{y}^3 \in L^2(\Omega)$ and, hence, the solution of $-\Delta y + y = w$, which is $\bar{y}$, satisfies $\bar{y} \in H^2(\Omega)$ if Ω is a C^2 domain or convex. For references concerning these and other regularity results, see Section 3.2. We remark that in our numerical example, $\Omega = (0,1)^2$ is, indeed, convex.

For $F(y) = -\Delta y + y + y^3$ we have $F'(y)[h_1] = -\Delta h_1 + h_1 + 3y^2 h_1$ and $F''(y)[h_1, h_2] = 6y h_1 h_2$, which are Fréchet derivatives as one can argue. Thus, we obtain

$$f'_{\varepsilon,\mu}(y) = c_1 \left(y - y_d + \hat{\alpha} F'(y)^* F(y)\right) - \frac{\tau(\varepsilon)}{B^{\varepsilon}_{C(\overline{\Omega}_a)}(y)} \frac{q}{\int_{\Omega_a} q \,\mathrm{d}x}$$

and

$$\begin{aligned} f''_{\varepsilon,\mu}(y)[h_1] \;=\; & c_1 \Bigg(h_1 + \hat{\alpha} \left(F'(y)^* F'(y)[h_1] + F(y) \cdot 6y h_1\right) \\ & \quad + \frac{\left(y - y_d + \hat{\alpha} F'(y)^* F(y)\right) \cdot \int_\Omega (y - y_d + \hat{\alpha} F'(y)^* F(y)) h_1 \,\mathrm{d}x}{\tilde{B}(y)} \Bigg) \\ & - c_2 \left(\frac{q \int_{\Omega_a} q h_1 \,\mathrm{d}x}{\int_{\Omega_a} q \,\mathrm{d}x} \cdot \left(1 - \frac{\varepsilon}{B^{\varepsilon}_{C(\overline{\Omega}_a)}(y)}\right) - q h_1 \right), \end{aligned}$$

where we used the same notation as in the previous section. The discretization can now be carried out analogously to the case with a linear state equation.

The nonconvexity of $f_{\varepsilon,\mu}$ may prevent the unique solvability of Newton's equation. This can be overcome by using $f''_{\varepsilon,\mu} + \beta I$ with a sufficiently large $\beta > 0$ instead of $f''_{\varepsilon,\mu}$. Here, β may be chosen different in each iteration. This technique is known as Levenberg-Marquardt regularization, cf., e.g., [GK99, Aufgabe 9.11] and [Ber99, Section 1.5.1]. However, in the numerical experiments that we conduct such a regularization is not necessary.

8.1.4. Line search

We recall that if iterates with large Newton decrement occur in LSMSUB, then line search can be used to find a better step size than the one induced by the Newton decrement, cf. Remark 2.7.4. If not mentioned otherwise, we employ as line search strategy a (numerical) minimum rule, where we first determine the maximal interval in $[0,1]$ such that $y + t n_y$ is feasible using MATLAB's built-in *fzero*, and then employ MATLAB's *fminbnd* to determine the unique (local=global) minimum of $\varphi(t) := f_{\varepsilon,\mu}(y + t n_y)$ in this interval. As tolerance in *fzero* and *fminbnd* we use machine precision. We note that *fzero* requires only function evaluations of $B^{\varepsilon}_{C(\overline{\Omega}_a)}$ and $\tilde{B}$, and *fminbnd* requires only function evaluations of $f_{\varepsilon,\mu}$. Since we work with reduction to the state, these evaluations are numerically inexpensive because no additional PDEs have to be solved. However, this would still be true if we used reduction to the control since the state equation is affine linear; see also the discussion in Remark 5.3.2. Moreover, we point out that φ is uniformly convex, which is certainly a good structural property for minimization. Summarizing, we expect the line search strategy to be computationally

inexpensive. In fact, in the numerical experiments whose results we present in the following sections, we usually observed on the finest mesh that all line searches together require less than two percent of the total running time, whereas the computation of Newton steps consumes almost all remaining time. Of course, a larger tolerance in *fzero* and *fminbnd* or other line search strategies are conceivable. For more on the line search strategy in LSMSUB, see the discussion in Remark 2.7.4.

8.2. Numerical results for fixed smoothing parameter

In this section we present numerical experiments with Algorithm LSM_ε as stated in Section 5.3. Some of these experiments are similar to the ones conducted in [KU13, Section 4], while others are completely new. In comparison to [KU13] we use a different termination criterion and a slightly different version of Algorithm LSMSUB, which is at the heart of LSM_ε. We recall that the main result for LSM_ε is Theorem 5.3.3.

8.2.1. Test Problem I

We consider the problem

$$\min_{(y,u)\in Y\times U} \frac{1}{2}\|y-y_d\|^2_{L^2(\Omega)} + \frac{\hat\alpha}{2}\|u\|^2_{L^2(\Omega)} \qquad \text{s.t.} \qquad y \geq y_a \text{ in } \overline{\Omega}_a, \quad \begin{cases} -\Delta y = u & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases}$$

with $Y := H^2(\Omega)\cap H^1_0(\Omega)$, $U := L^2(\Omega)$, $\Omega := \Omega_a := B_1(0) \subset \mathbb{R}^2$, and $\hat\alpha := 10^{-2}$. The functions y_d and y_a are given in polar coordinates: $y_d(r) := \frac{1}{8\pi\hat\alpha}(1 + r^2(\ln(r) - 1))$ and $y_a(r) := \frac{1}{8\pi\hat\alpha}(1-2r)$. Using $Z := U$, $A := -\Delta \in \mathcal{L}(Y,Z)$, $B := -I \in \mathcal{L}(U,Z)$, and $g \equiv 0$, this problem is a special instance of Example 3.2.1 and, hence, satisfies Assumption 3.1.9.

This problem is also considered in [CKR08, Section 5.1] and [BV09, Section 5.3], and in slightly different form in [MPT07, Section 6.2] and [HPUU09, Section 3.3.1.4]. Using the optimality conditions from Lemma 3.4.3 for case II it is easy to check that $\bar y \equiv y_d$ and $\bar u(r) = -\frac{\ln(r)}{2\pi\hat\alpha}$ are optimal and that the corresponding multiplier $\bar\lambda$ is a Dirac measure concentrated in the origin. The latter is not surprising since the origin is the only active point of the constraints $y \geq y_a$, cf. also [Cas86, Corollary 1]. Recalling that $\hat j$ is the objective of the reduced original problem, this yields $\hat j(\bar u) = 6.25/\pi$.

Throughout the experiments for this test problem we use $\theta = 0.25$, $C_j = \frac{2\cdot 10^3}{\varepsilon^2}$, $\mu_0 = 1$, and $C_{\hat j} = 1 + \hat j(u^0)$ with $u^0 \equiv \frac{1}{2\pi\hat\alpha}$. Unless stated otherwise, we employ $\tau(\varepsilon) = 2\cdot 10^5 \cdot \frac{1+|\ln\varepsilon|}{\varepsilon}$. Lemma 3.6.5 provides $C_{\partial,C(\overline{\Omega}_a)} \leq 0.2$. Thus, by use of $\tilde u \equiv 1.1u^0$ Lemma 3.5.8 yields self-concordance of $f_{\varepsilon,\mu}$ for $\mu \in (0,\mu_0]$ and $\varepsilon \leq 1$, since it holds

$$\frac{2\cdot 10^3}{\varepsilon^2} = C_j \geq \frac{1}{\varepsilon^2}\max\left\{\varepsilon^2, \frac{16}{9}C^2_{\partial,C(\overline{\Omega}_a)}\left(\frac{2\|\hat j'(\tilde u)\|_{U^*}}{\hat\alpha} + \|\tilde u\|_U\right)^2\right\}.$$

Here, we computed the quantity $\|\hat j'(\tilde u)\|_{U^*}$ numerically on a uniform mesh with width $h = 2^{-9}$, cf. Remark 3.5.10. This shows that every $C_j \geq \frac{1.31\cdot 10^3}{\varepsilon^2}$ ensures self-concordance. The actual

choice for C_j includes a safeguard to take into account the numerical computation involved in this process.

The main result regarding Algorithm LSM_ε is Theorem 5.3.3. To derive the estimates in part 5) of this theorem, Corollary 4.4.4 is crucial. We start with a numerical validation of this theorem and this corollary. Later, we also examine the efficiency of LSM_ε.

First, we investigate the convergence of $(j(u^k))$ and $(\hat{j}(u^k))$ to $j(\bar{u}_\varepsilon)$ and $\hat{j}(\bar{u}_\varepsilon)$, respectively, cf. Theorem 5.3.3 2). We use mesh width $h = 2^{-7}$. In Figure 8.1 we display $\frac{1}{C_j}(j(u^{k+1}) - j(\bar{u}_\varepsilon))$ and $\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)$ for $k \in \mathbb{N}_0$, where we use for $\bar{u}_\varepsilon$ the last iterate of LSM_ε with $h = 2^{-7}$ and $\mu_{\text{final}} = 10^{-10}$. We have $1 \leq C_{\hat{j}} - \hat{j}(u^{k+1}) = 1 + \hat{j}(u^0) - \hat{j}(u^{k+1}) \leq 1 + \hat{j}(u^0) \leq \frac{11}{2}$ for sufficiently large k in this experiment. Hence,

$$j(u^{k+1}) - j(\bar{u}_\varepsilon) = C_j \ln\left(1 + \frac{\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)}{C_{\hat{j}} - \hat{j}(u^{k+1})}\right) \approx C_j \frac{\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)}{C_{\hat{j}} - \hat{j}(u^{k+1})}$$

shows that we can expect $\frac{1}{C_j}|j(u^{k+1}) - j(\bar{u}_\varepsilon)|$ to range roughly from $\frac{2}{11}|\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)|$ to $|\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)|$ for large k, and to behave similar to $|\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)|$. Figure 8.1 confirms this. It also shows that the convergence rate of $(j(u^k))$ is linear, in accordance with the developed theory, cf. Theorem 5.3.3 2). In addition, this theorem provides a fairly good estimate for $|j(u^{k+1}) - j(\bar{u}_\varepsilon)|$ if μ_k is small, as is displayed in Figure 8.2.

To assess the discretization error we show in Figure 8.3 $\frac{1}{C_j}|j(u^{k+1}) - j(\bar{u}_\varepsilon)|$ and $|\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)|$ for $k \in \mathbb{N}_0$, where $\bar{u}_\varepsilon$ is now computed with $h = 2^{-8}$ while the iterates are still computed with $h = 2^{-7}$. In [BV09] the discretization error for the objective value is estimated using linear finite elements on uniform triangulations and the trapezoidal rule for integration. For a mesh with roughly $8 \cdot 10^4$ nodes the discretization error is around $8 \cdot 10^{-3}$. We use mesh width $h = 2^{-7}$ with roughly $5 \cdot 10^4$ nodes and observe in Figure 8.3 that $|\hat{j}(u^{k+1}) - \hat{j}(\bar{u}_\varepsilon)|$ lies in $[3 \cdot 10^{-3}, 7 \cdot 10^{-3}]$ for large k. Note that since LSM_ε is an infeasible method, the values for $j(\bar{u}_\varepsilon)$ and $\hat{j}(\bar{u}_\varepsilon)$ using $h = 2^{-7}$ may be smaller than the ones obtained with $h = 2^{-8}$. This is what causes the regions of non-monotonic convergence in Figure 8.3. For an assessment of the discretization error in the case that the iterates are computed with $h = 2^{-8}$ and $\bar{u}_\varepsilon$ is computed with $h = 2^{-9}$, we refer to [KU13, Section 4].

We now focus on the infeasibility of $(y(u^k))$ with respect to the pointwise state constraints. An estimate for this infeasibility can be found in Corollary 4.4.4. We use mesh widths $h = 2^{-7}$ and $h = 2^{-9}$. The results are depicted in Figure 8.4. The iterates are feasible at the beginning of the algorithm and become infeasible during the course of the algorithm. This happens at $\mu_k \approx 5.5 \cdot 10^{-3}$ and $\mu_k \approx 3 \cdot 10^{-3}$, respectively. We observe that the last iterates, which approximate $y(\bar{u}_\varepsilon)$, violate feasibility the most. This is not surprising since the optimal solution $\bar{u}_\varepsilon$ of (P_ε) can be expected to lie on the boundary of D_{b^ε}. Moreover, we observe that the choice of $h = 2^{-7}$ or $h = 2^{-9}$ does not make a significant difference. This may indicate that the discretization error for (y^k) is already rather small for $h = 2^{-7}$. Taking a look at the infeasibility of the final iterates in Figure 8.4 we suspect that $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$ is of order $\mathcal{O}(\varepsilon)$, which is slightly better than the estimate provided in Corollary 4.4.4.

Next, we examine the error $\|u^{k+1} - \bar{u}\|_{L^2(\Omega)}$ for $k \in \mathbb{N}_0$, cf. Theorem 5.3.3 5). Recall that $\bar{u}$ is known exactly. We employ mesh width $h = 2^{-10}$. The results are depicted in Figure 8.5. Since

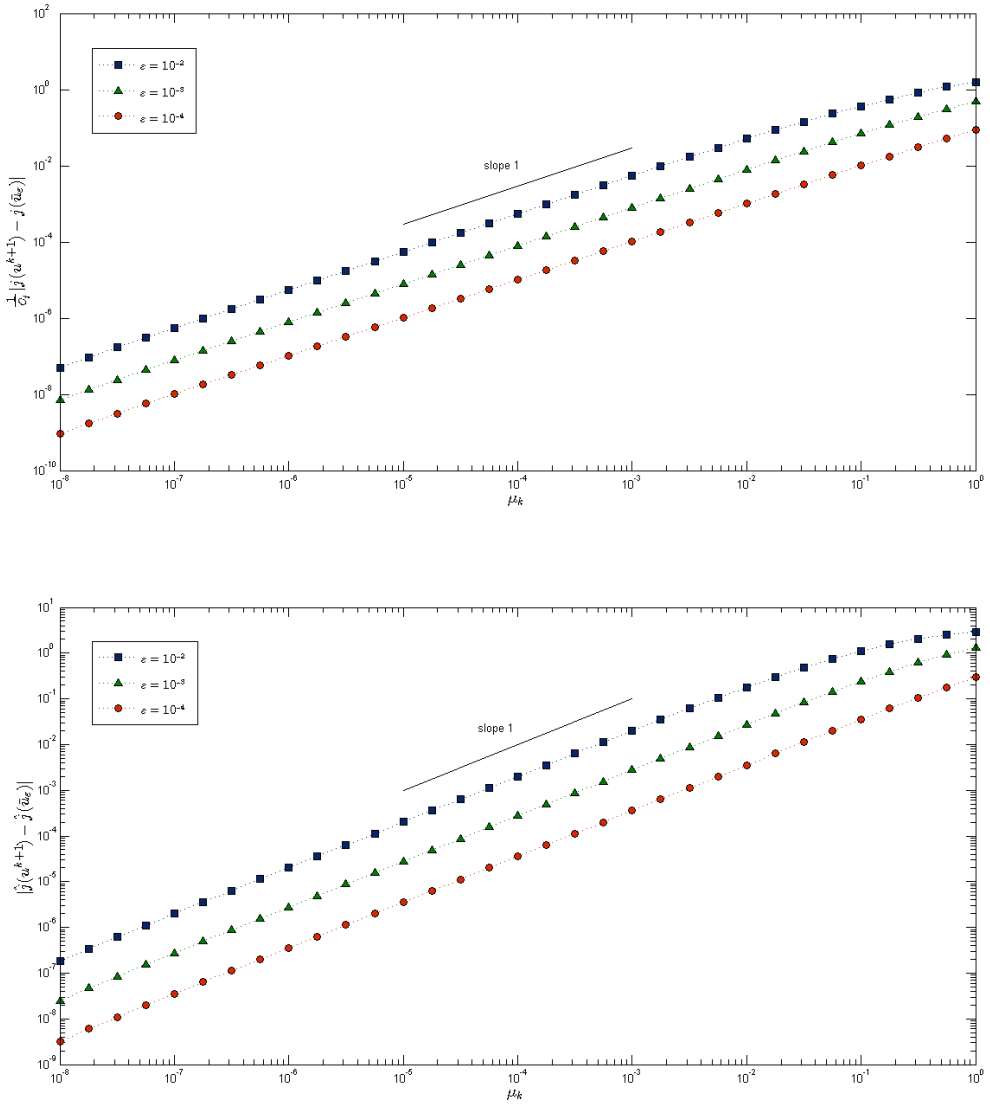

Figure 8.1. Test problem I: Error in j and $\hat{j}$ with respect to $\bar{u}_\varepsilon$ computed on the same grid as the iterates

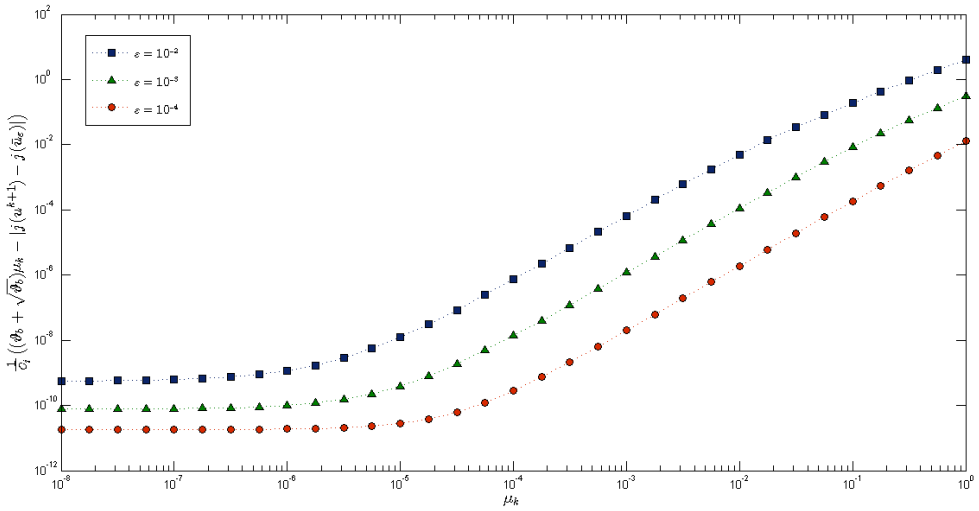

Figure 8.2. Test problem I: Error in j compared to its prediction from Theorem 5.3.3 2)

(u^k) converges to $\bar{u}_\varepsilon$, the last iterates display the error $\|\bar{u}_\varepsilon - \bar{u}\|_{L^2(\Omega)}$. The choice $h = 2^{-10}$ is motivated by the fact that we want to take a look at the asymptotic for $\|\bar{u}_\varepsilon - \bar{u}\|_{L^2(\Omega)}$ as ε decreases. Therefore, we choose h so small that $\|\bar{u}_\varepsilon - \bar{u}\|_{L^2(\Omega)}$ (almost) does not change if h is further diminished. This makes it most likely that this error does not stem from discretization. We conjecture that for this problem $\|\bar{u}_\varepsilon - \bar{u}\|_{L^2(\Omega)}$ has order $\mathcal{O}(\sqrt{\varepsilon(1 + |\ln \varepsilon|)})$, as predicted by the estimate from Theorem 5.3.3 5).

In Figure 8.6 we display $-\bar{u}_\varepsilon$ and $-\bar{y}_\varepsilon$ for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$, together with $-\bar{u}$ and $-\bar{y}$. We use negative functions to generate good-looking plots. Also, we recall that we have $y_d \equiv \bar{y}$. Apparently, the structure of the optimal control $\bar{u}$ and the optimal state $\bar{y}$ are well replicated by $\bar{u}_\varepsilon$ and $\bar{y}_\varepsilon$, respectively.

We now examine the efficiency of LSM$_\varepsilon$. To this end, we use mesh sizes up to $h = 2^{-9}$ and choose $\varepsilon \geq 10^{-5}$. We use the same parameters as before except for $\tau(\varepsilon)$, which we change to $2 \cdot 10^2 \cdot \frac{1+|\ln \varepsilon|}{\varepsilon}$. The choice $\tau(\varepsilon) = C_j \varepsilon(1 + |\ln \varepsilon|)/10$ also produced good efficiency results for other test problems and is based on the error estimate from Theorem 5.3.3 5). The value for $\tau(\varepsilon)$ that we employed before, however, is better suited to generate nice plots; for the new choice of $\tau(\varepsilon)$ the overall errors and infeasibilities are already for $\mu_0 = 1$ rather close to their final values, which explains why we used a larger value for $\tau(\varepsilon)$ before. Note that regardless of the value for $\tau(\varepsilon)$, LSM$_\varepsilon$ computes $\bar{u}_\varepsilon$. This is, $\tau(\varepsilon)$ influences the path $\mu \mapsto \bar{u}_{\varepsilon,\mu}$, which the iterates generated by LSM$_\varepsilon$ follow, but not the endpoint $\bar{u}_\varepsilon$ of this path, to which the iterates converge. This knowledge can be used to determine a suitable value for $\tau(\varepsilon)$ in the following way: We apply LSM$_\varepsilon$ on a coarse grid for different choices of $\tau(\varepsilon)$ and pick for $\tau(\varepsilon)$ the value for which the number of Newton steps is minimal (using a termination criterion based on convergence of the objective since the use of a fixed μ_{final} would not yield comparable final iterates). Here, we focus on the number of Newton steps since the computation of Newton

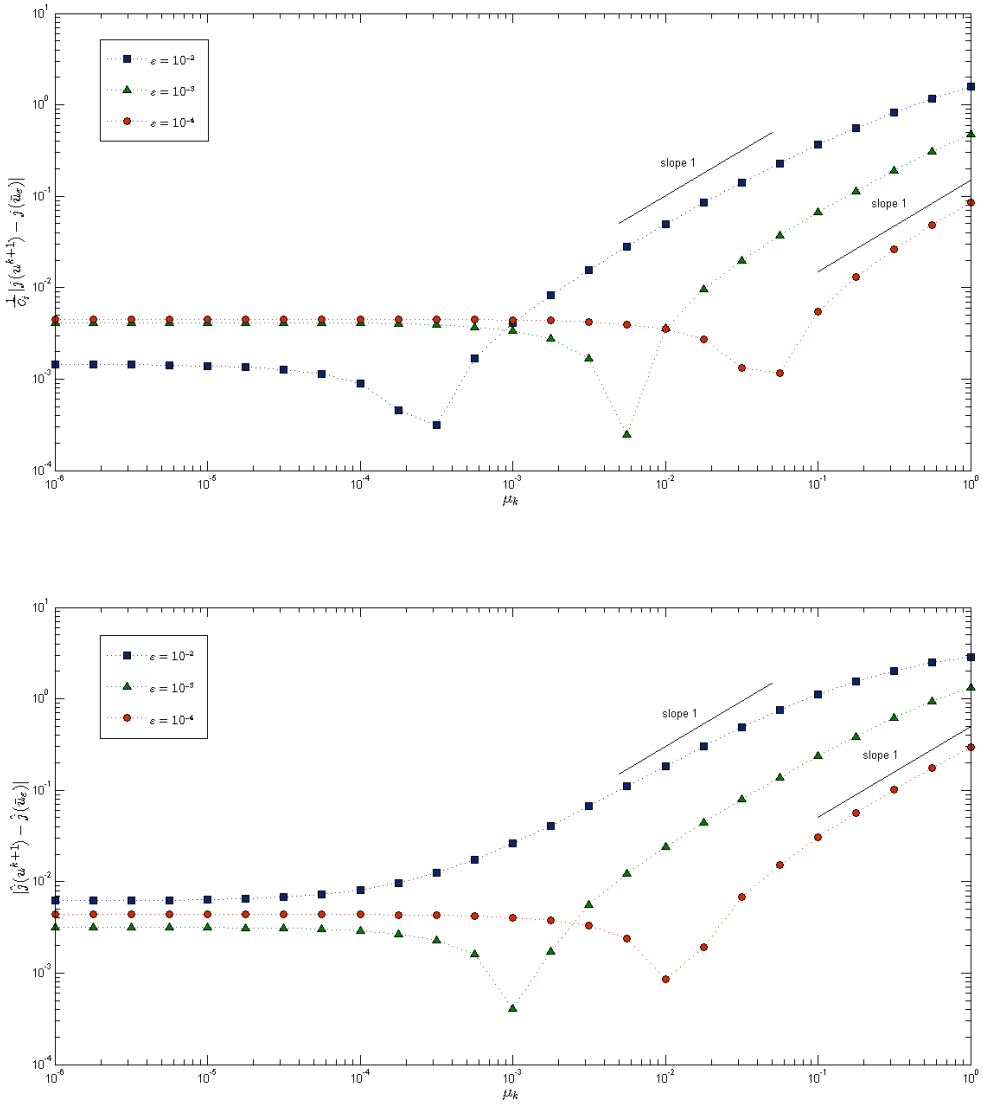

Figure 8.3. Test problem I: Error in j and $\hat{j}$ with respect to $\bar{u}_\varepsilon$ computed on a finer grid than the iterates

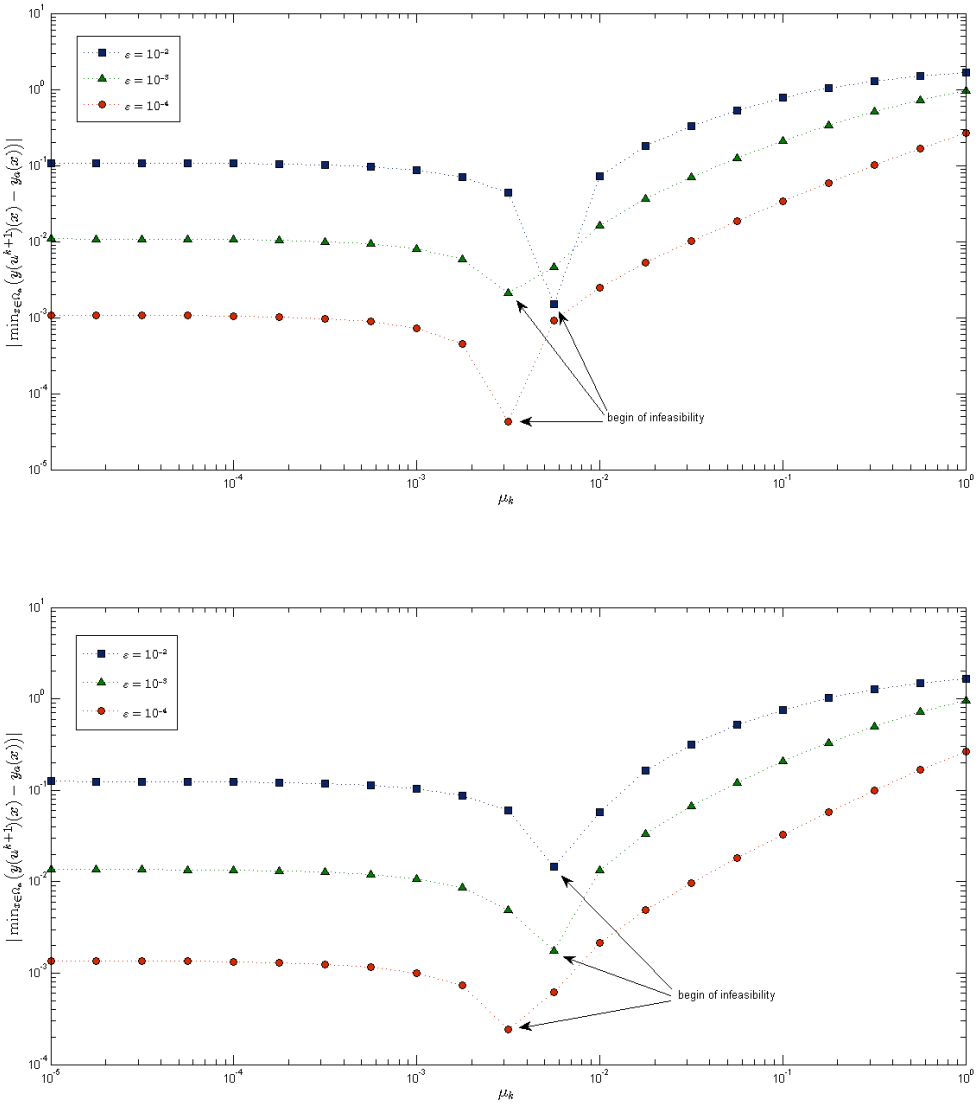

Figure 8.4. Test problem I: Infeasibility of $y(u^{k+1})$ for $h = 2^{-7}$ (top) and $h = 2^{-9}$ (bottom)

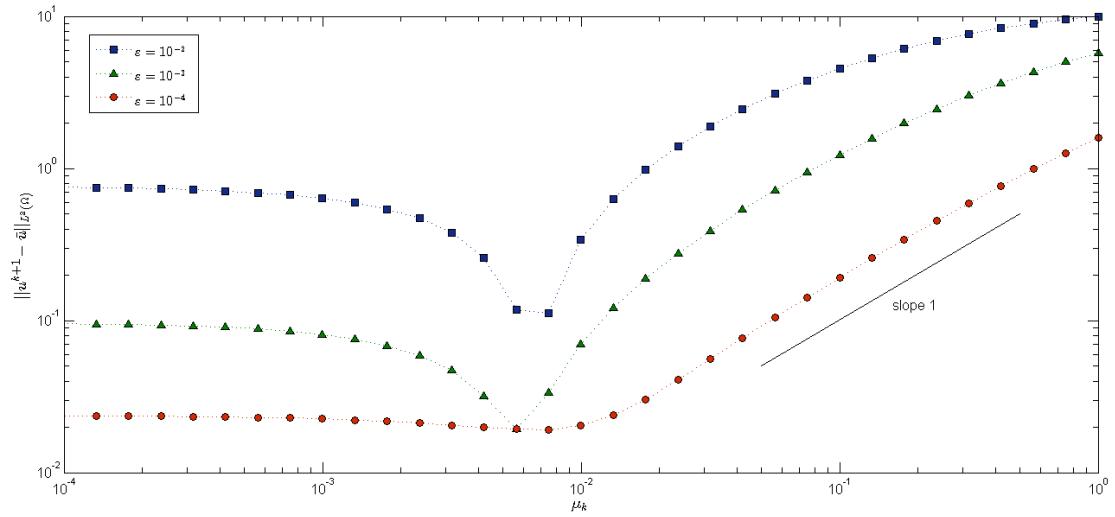

Figure 8.5. Test problem I: Overall error of the control

steps dominates the amount of time that is required by the algorithm. We demonstrate later that LSM_ε is mesh independent. This shows that the use of a coarse grid is possible in this strategy and, therefore, this strategy is practicable. To improve this strategy an estimate for $\tau(\varepsilon)$ can be derived from Theorem 5.3.3 5) and Lemma 4.4.10, as we explain in more detail when we conduct experiments for $\varepsilon \to 0^+$. We emphasize that this estimate usually yields a good final value for $\tau(\varepsilon)$, already.

Furthermore, we choose for LSM_ε the parameters $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 0.5$, and $\beta_0 = 0.1$.

To adaptively determine the update β_k for μ_k at the end of iteration $k \in \mathbb{N}$, we use the following simple rule: We prescribe a natural number $m \geq 2$. If LSMSUB is called in iteration k and takes less than $m-1$ iterations, then we choose $\beta_k \in (\beta_{\min}, \beta_{\max})$ with $\beta_k < \beta_{k-1}$. If LSMSUB takes more than $m+1$ iterations, then we choose $\beta_k \in (\beta_{\min}, \beta_{\max})$ with $\beta_k > \beta_{k-1}$. The concrete size of β_k is determined depending on the deviation from m, e.g., if LSMSUB takes $m+4$ iterations, $\beta_k > \beta_{k-1}$ is chosen larger than if LSMSUB takes $m+3$ iterations. We employ $m = 4$. We do not apply this strategy after phase one, i.e., we always have $\mu_1 = \beta_0 \mu_0$ with the initial β_0. We point out that this strategy for the determination of (β_k) is completely within the theoretical framework developed in this thesis.

Motivated by Theorem 5.3.3 5) we use $\max\{\frac{|j(u^{k+1-i}) - j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i}) - \hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq \frac{\varepsilon(1+|\ln \varepsilon|)}{10}$ for $i = 1, 2$ as termination criterion, which seems rather strict. As an alternative to this termination criterion it is also possible to directly prescribe μ_{final}, cf. [KU13, Section 4].

We first display the mesh independence of LSM_ε. To this end, we apply LSM_ε for different values of ε and different mesh sizes, where the finest mesh has width $h = 2^{-9}$. The total numbers of Newton steps that have to be computed during the course of LSM_ε can be found in Table 8.1 and clearly indicate that LSM_ε is mesh independent.

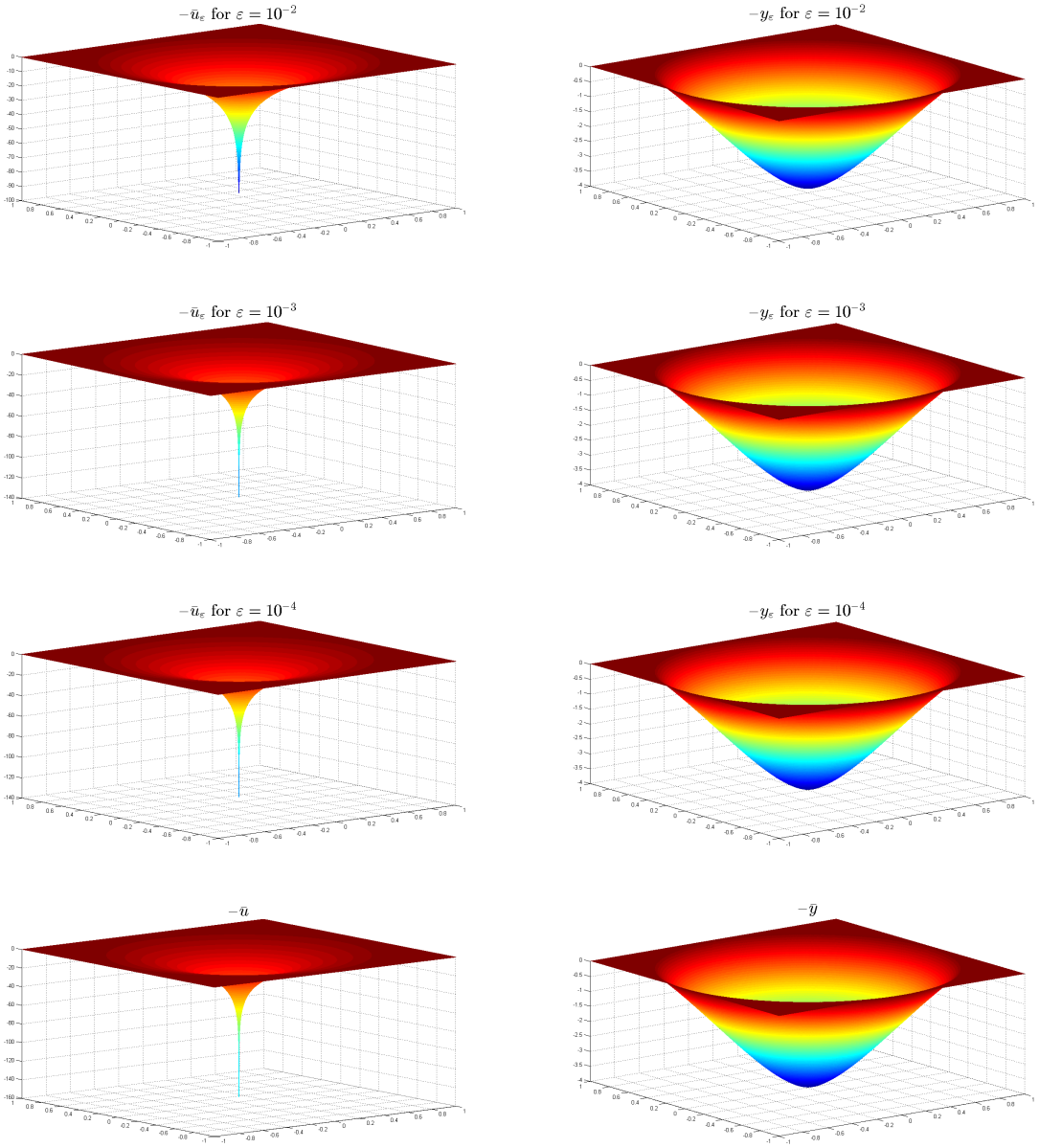

Figure 8.6. Test problem I: $-\bar{u}_\varepsilon$ and $-\bar{y}_\varepsilon$ for $\varepsilon = 10^{-\{2,3,4\}}$ together with $-\bar{u}$ and $-\bar{y}$

Algorithm LSM$_\varepsilon$ requires a starting point $u^0 \in \Lambda_{\mu_0}(\tau)$. Since the u^0 that we employ does not satisfy $u^0 \in \Lambda_{\mu_0}(\tau)$, the overall iteration numbers displayed in Table 8.1 contain a phase one. To be in accordance with Algorithm APOSS and the corresponding Theorem 5.4.4 we count all Newton steps as phase one until an iterate $\tilde{u}$ is obtained for which $\Lambda_{\mu_0}(\tilde{u}) \leq \theta$ is ensured. This is, we count all Newton steps until μ_0 is decreased to μ_1, which is a little stricter than just counting the Newton steps until a $\tilde{u}$ with $\Lambda_{\mu_0}(\tilde{u}) \leq \tau$ is found, since $\tau \approx 0.5$ while $\theta = 0.25$. However, in practice we observed that this often does not change the number of Newton steps that we count, anyway. If it changes the number, then usually only by one step. The Newton steps required by phase one are displayed in Table 8.1 in brackets for the finest mesh; for coarser mesh sizes this number is basically the same. From now on, such a number in brackets always denotes iterations required during phase one in the sense just explained.

Table 8.1 also contains the infeasibility $\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$ along with the errors $\frac{|j(u^K)-j(\bar{u})|}{C_j}$ and $|\hat{j}(u^K) - \hat{j}(\bar{u})|$, and the final value of μ_k. Here and in all experiments that follow, (y^K, u^K) denotes the final iterate of LSM$_\varepsilon$. Since LSM$_\varepsilon$ generates a sequence (u^k) that converges to $\bar{u}_\varepsilon$, we interpret (y^K, u^K) as $(\bar{y}_\varepsilon, \bar{u}_\varepsilon)$ and, consequently, understand the infeasibility and errors in Table 8.1 as values for $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$, $\frac{|j(\bar{u}_\varepsilon)-j(\bar{u})|}{C_j}$, and $|\hat{j}(\bar{u}_\varepsilon) - \hat{j}(\bar{u})|$. We remark that in separate experiments we confirmed that these values do not change when a stricter termination criterion is employed.

When we displayed $\|\bar{u}_\varepsilon - \bar{u}\|_{L^2(\Omega)}$, we used $h = 2^{-10}$ since for larger h the numerically computed value for $\|\bar{u}_\varepsilon - \bar{u}\|_{L^2(\Omega)}$ decreased for $\varepsilon = 10^{-4}$ when passing from h to $h/2$. Therefore, we expect that the numerically computed value for $\frac{|j(\bar{u}_\varepsilon)-j(\bar{u})|}{C_j}$ with $h = 2^{-9}$ may contain a notable discretization error for $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-5}$. This matches the fact that in Table 8.1 the error for $\frac{|j(\bar{u}_\varepsilon)-j(\bar{u})|}{C_j}$ seems to be too large in comparison to the predicted $\mathcal{O}((1 + |\ln \varepsilon|)\varepsilon)$ for $\varepsilon = 10^{-4}$ and almost stays the same when passing from $\varepsilon = 10^{-4}$ to $\varepsilon = 10^{-5}$. The infeasibility $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$, however, seems to be well-resolved for $h = 2^{-9}$ and all displayed values of ε; we suspect convergence with order $\mathcal{O}(\varepsilon)$, which is slightly better than the result from Corollary 4.4.4. This convergence order is in accordance with Figure 8.4.

ε	Mesh size $h = 2^{-i},\ i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	12	12	12	12	12	12(6)	1.25×10^{-1}	1.22×10^{-1}	3.46×10^{-2}	2.43×10^{-6}
10^{-3}	16	15	16	16	17	17(11)	1.36×10^{-2}	1.62×10^{-2}	4.68×10^{-3}	2.43×10^{-6}
10^{-4}	18	17	16	20	16	17(11)	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}	2.43×10^{-6}
10^{-5}	17	16	21	20	17	21(10)	1.36×10^{-4}	2.81×10^{-3}	8.12×10^{-4}	7.65×10^{-6}

Table 8.1. Test problem I: Total number of Newton steps required by LSM$_\varepsilon$; the Newton steps from LSMSUB and phase one are included; displayed in brackets is the number of Newton steps required by phase one; (y^K, u^K) denotes the final iterate

In Table 8.2 we display for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$, in more detail how μ_k is decreased to μ_{final} for $h = 2^{-9}$. We observe that the overall errors with respect to j and $\hat{j}$ increase for $\mu_k \leq 10^{-1}$. This can be attributed to the fact that (u^k) converges to $\bar{u}_\varepsilon$, not to $\bar{u}$. To compute an approximation

of $j(\bar{u})$ in this test problem, it would be sufficient to use $\mu_{\text{final}} \approx 1$ in LSM$_\varepsilon$. Of course, this value for μ_{final} is based on the fact that we know $\bar{u}$ and $\hat{j}(\bar{u})$ in this example. In a practical optimization problem, however, such optimal values for μ_{final} are unknown. In this case we suggest to determine μ_{final} within the algorithm in the way presented here, i.e., by use of a termination criterion that is based on convergence of the quantity of interest, e.g., the objective value.

Figure 8.7 depicts the development of the Newton decrement during the course of LSM$_\varepsilon$ for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$, and $h = 2^{-9}$. We can clearly recognize phase one as well as the points at which μ_k is decreased since these are exactly the points at which the Newton decrement increases. We note that LSMSUB is capable of substantially reducing the Newton decrement with only a single damped Newton step.

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	6	1.04×10^{-1}	1.02×10^{-1}	2.91×10^{-2}
10^{-1}	2	1.23×10^{-1}	1.20×10^{-1}	3.41×10^{-2}
1.26×10^{-3}	2	1.25×10^{-1}	1.22×10^{-1}	3.46×10^{-2}
2.43×10^{-6}	2	1.25×10^{-1}	1.22×10^{-1}	3.46×10^{-2}

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	11	1.09×10^{-2}	1.35×10^{-2}	3.89×10^{-3}
10^{-1}	2	1.33×10^{-2}	1.59×10^{-2}	4.60×10^{-3}
1.26×10^{-3}	2	1.36×10^{-2}	1.62×10^{-2}	4.68×10^{-3}
2.43×10^{-6}	2	1.36×10^{-2}	1.62×10^{-2}	4.68×10^{-3}

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	11	1.01×10^{-3}	3.68×10^{-3}	1.06×10^{-3}
10^{-1}	2	1.33×10^{-3}	4.00×10^{-3}	1.16×10^{-3}
1.26×10^{-3}	2	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}
2.43×10^{-6}	2	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}

Table 8.2. Test problem I: Course of LSM$_\varepsilon$ with $\varepsilon = 10^{-2}$ (top), $\varepsilon = 10^{-3}$ (middle), and $\varepsilon = 10^{-4}$ (bottom)

Table 8.2 may be read in the sense that the termination criterion is too strict. Thus, we now change the termination criterion to $\max\{\frac{|j(u^{k+1-i}) - j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i}) - \hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq 10\varepsilon(1 + |\ln \varepsilon|)$ for $i = 1, 2$, which is less strict than the criterion we used before. It yields the iteration numbers and errors displayed in Table 8.3. Comparing these overall errors to the ones from Table 8.1 we conclude that in this test problem the less strict termination criterion is more efficient in approximating $j(\bar{u})$ since it requires less iterations and produces equally good solutions. In the following experiments for Test Problem I, we, hence, work with this less strict termination criterion.

To increase the practical efficiency of LSM$_\varepsilon$ for Test Problem I we employ a nested grid strategy. More precisely, we use a hierarchy of six grids that have mesh widths $h = 2^{-i}$, $i = 4, 5, \ldots, 9$.

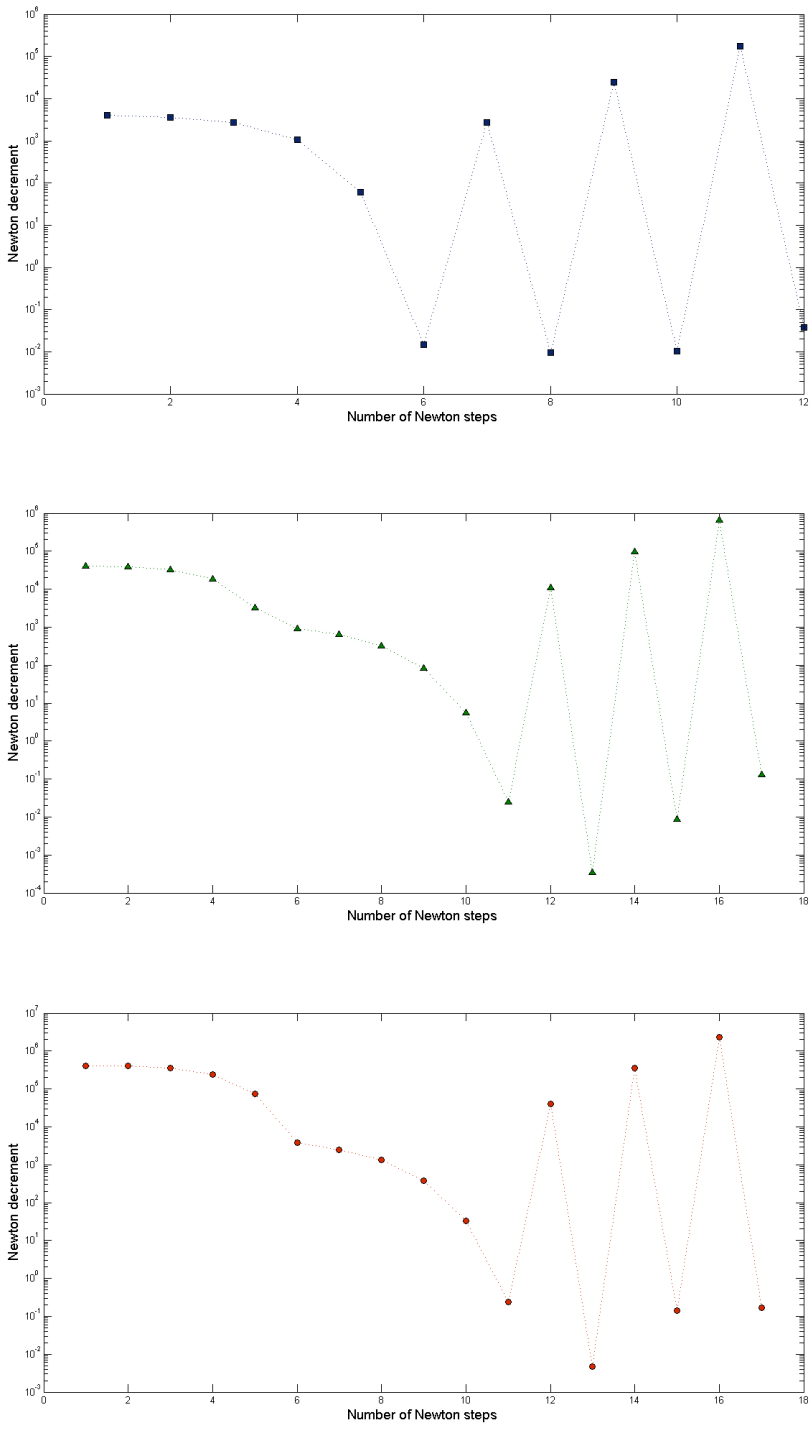

Figure 8.7. Test problem I: Newton decrements of LSM_ε for $\varepsilon = 10^{-2}$ (top), $\varepsilon = 10^{-3}$ (middle), and $\varepsilon = 10^{-4}$ (bottom)

ε	Mesh size $h = 2^{-i}$, $i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	10	10	10	10	10	10	1.25×10^{-1}	1.22×10^{-1}	3.46×10^{-2}	1.26×10^{-3}
10^{-3}	13	13	14	14	15	15	1.36×10^{-2}	1.62×10^{-2}	4.68×10^{-3}	1.26×10^{-3}
10^{-4}	13	13	13	14	14	15	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}	1.26×10^{-3}
10^{-5}	13	14	15	15	14	15	1.36×10^{-4}	2.81×10^{-3}	8.12×10^{-4}	1.26×10^{-3}

Table 8.3. Test problem I: Results of LSM_ε with less strict termination criterion

Applying LSM_ε on the coarsest grid of the hierarchy we obtain as final iterate an approximation of $\bar{u}_\varepsilon$ on this grid for some value μ_{final}. We denote this iterate by $\bar{u}_\varepsilon^h$ to take into account the grid on which this iterate lives. We prolongate $\bar{u}_\varepsilon^h$ onto the next finer grid to obtain an approximation of $\bar{u}_\varepsilon$ on this finer grid with mesh width $h/2$. We use this approximation as starting point for LSM_ε on the grid with mesh width $h/2$. On this grid we carry out LSM_ε only for $\mu = \mu_{\text{final}}$, where μ_{final} stems from LSM_ε on the coarsest grid. This is, we assume that the value for μ_{final} obtained on the coarsest grid is somewhat close to the value for μ_{final} that we would obtain on finer grids, a hypothesis that is confirmed by a comparison of the values μ_{final} from experiments without nested grids, see Table 8.3, with the ones we obtain here. To further safeguard this value of μ_{final} we could, for instance, work with a stricter termination criterion on the coarsest grid. However, we will see that even without such a safeguard our strategy works well. The final iterate generated by LSM_ε on the grid with mesh width $h/2$ is denoted by $\bar{u}_\varepsilon^{h/2}$. We use the prolongation of $\bar{u}_\varepsilon^{h/2}$ onto the next finer grid as starting point for LSM_ε with $\mu = \mu_{\text{final}}$ on this finer grid. We repeat this procedure until we obtain the final iterate of LSM_ε on the finest grid.

It remains to describe how we prolongate a given $\bar{u}_\varepsilon^h$ onto a grid with mesh width $h/2$ to generate a starting point for LSM_ε on this finer mesh. As a first step we use linear interpolation to create from $\bar{u}_\varepsilon^h$ the interpolant $\hat{u}_\varepsilon$ that lives on the grid with mesh width $h/2$. Unfortunately, $\hat{u}_\varepsilon$ may be infeasible, i.e. $B^\varepsilon(\hat{u}_\varepsilon) \leq 0$ or $\tilde{B}(\hat{u}_\varepsilon) \leq 0$, even though $\bar{u}_\varepsilon^h$ is feasible on the grid with mesh width h. A simple idea to overcome this difficulty is to shift $\hat{u}_\varepsilon$ towards a feasible point, e.g., the starting point $u^0 \equiv \frac{1}{2\pi\hat{\alpha}}$ from before or, more precisely, its discretization on the mesh with width $h/2$. Another idea is based on the observation that the infeasibility increases as μ decreases: We could use all grids except for the finest to compute $\bar{u}_{\varepsilon,\mu}^h$ for a rather large μ instead of aiming for $\bar{u}_\varepsilon^h$. The decrease of μ to μ_{final} would then be carried out on the finest mesh resulting in $\bar{u}_\varepsilon^h$ on the finest mesh. If $\mu = \mu_0$ is used, then this basically means that we try to reduce the iterations required by phase one on the finest grid. We follow the first idea since it seems more promising. We now explain how we determine the size of the shift towards u^0. More precisely, we want to obtain $t^* \in [0,1]$ such that $\tilde{u}_\varepsilon := t^*\hat{u}_\varepsilon + (1-t^*)u^0$ is feasible, i.e. $B^\varepsilon(\tilde{u}_\varepsilon) > 0$ and $\tilde{B}(\tilde{u}_\varepsilon) > 0$. First, we check if $B^\varepsilon(\hat{u}_\varepsilon) > 0$. If so, then we set $t_1 := 1$. If not, then we use MATLAB's built-in function *fzero* to determine a root $t_1 \in (0,1]$ of $\varphi_1 : [0,1] \to \mathbb{R}$, $\varphi_1(t) := B^\varepsilon(t\hat{u}_\varepsilon + (1-t)u^0)$. The concavity of B^ε, cf. Corollary 4.1.7, implies concavity of φ_1. Together with $\varphi_1(0) > 0$ and $\varphi_1(1) \leq 0$ this can be used to argue that φ_1 has a unique root $t_1 \in (0,1]$ and that $\varphi_1(t) > 0$ is satisfied for all $0 \leq t < t_1$. The same argument shows that either $\tilde{B}(\hat{u}_\varepsilon) > 0$, in which case we set $t_2 := 1$, or $\varphi_2 : [0,1] \to \mathbb{R}$, $\varphi_2(t) := \tilde{B}(t\hat{u}_\varepsilon + (1-t)u^0)$ possesses a unique root $t_2 \in (0,1]$ and it holds $\varphi_2(t) > 0$ for all

$0 \leq t < t_2$. Thus, $t\hat{u}_\varepsilon + (1-t)u^0$ is feasible for every $0 \leq t < \min\{t_1, t_2\}$. Finally, we obtain t^* by setting $t^* := \min\{t_1, t_2\} \cdot \kappa^{\frac{|B^\varepsilon(\hat{u}_\varepsilon)| + B^\varepsilon(u^0)}{B^\varepsilon(u^0)}}$ for a $\kappa \in (0,1)$. In this test problem we employ $\kappa := 0.9$. The idea behind the definition of t^* is to ensure a certain distance to the boundary of the feasible set. This resembles the fraction-to-the-boundary rule, which is, for instance, used in the very successful IPOPT, cf. [WB06, Section 2.2]. Moreover, the definition of t^* is based on the observation that in all numerical experiments feasibility is only violated with respect to B^ε, i.e., $t_2 = 1$ and $\min\{t_1, t_2\} = t_1$. If $B^\varepsilon(u^0)$ is significantly larger than $|B^\varepsilon(\hat{u}_\varepsilon)|$, we want $t^* < t_1$ to be rather close to $\kappa t_1 = \kappa \min\{t_1, t_2\}$ since we expect that this already ensures enough feasibility for the starting point. If, however, $B^\varepsilon(u^0)$ is significantly smaller than $|B^\varepsilon(\hat{u}_\varepsilon)|$, we want t^* to deviate stronger from t_1 since we expect that values close to t_1 may not yield enough feasibility for the starting point. Eventually, we define $\tilde{u}_\varepsilon := t^*\hat{u}_\varepsilon + (1-t^*)u^0$.

We remark that all states that occur during the determination process for t^* are of the form $y(t\hat{u}_\varepsilon + (1-t)u^0) = ty(\hat{u}_\varepsilon) + (1-t)y(u^0)$, where we used the affine linearity of the state equation. Since interpolation and the use of *fzero* are numerically cheap, the prolongation onto a finer grid essentially only requires two additional solves of the state equation on this finer grid. To judge these costs we recall that a Newton step on this grid is more expensive, cf. Section 8.1.2.

In Table 8.4 we show the Newton steps required on each of the nested grids together with errors on the finest grid. We use an additional (s) to mark that prolongation from a grid onto the next finer one requires a shift. Except for $\varepsilon = 10^{-2}$ we observe that the number of Newton steps required on finer grid seems to be rather large for a nested grid strategy; we attribute this to the shift towards u^0 that ensures feasibility. We believe that it is possible to further reduce this number by use of more sophisticated techniques to ensure feasibility. For instance, it may be possible to determine a better t^* by involving the objective j or the barrier $f_{\varepsilon,\mu}$. We leave this as a topic for future research. Nonetheless, a comparison with Table 8.3 shows that the nested grid strategy increases the practical efficiency of LSM_ε.

In Figure 8.8 and 8.9 we depict for $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-5}$ on each grid how the Newton decrement is reduced. This is, we show the complete development of the Newton decrement during the iterations of the nested grid strategy.

ε	Mesh size $h = 2^{-i},\ i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	10	5(s)	4(s)	3	3	3	1.25×10^{-1}	1.22×10^{-1}	3.46×10^{-2}	1.26×10^{-3}
10^{-3}	13(s)	6(s)	6(s)	6(s)	5(s)	5	1.36×10^{-2}	1.62×10^{-2}	4.68×10^{-3}	1.26×10^{-3}
10^{-4}	13(s)	7(s)	6(s)	6(s)	6(s)	6	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}	1.26×10^{-3}
10^{-5}	13(s)	8(s)	7(s)	9(s)	8(s)	7	1.36×10^{-4}	2.81×10^{-3}	8.12×10^{-4}	1.26×10^{-3}

Table 8.4. Test problem I: Results of LSM_ε with a nested grid strategy; (s) indicates that the prolongation from this mesh onto the next finer one involves a shift

We recall that the barrier $f_{\varepsilon,\mu}(u) = -\frac{C_j(\ln(C_j - \hat{j}(u))}{\mu} - \tau(\varepsilon)\ln(B^\varepsilon(u))$ contains the weights C_j and $\tau(\varepsilon)$. The choice for C_j is induced by the necessity to make $f_{\varepsilon,\mu}$ self-concordant, cf. Lemma 3.5.8. The choice for $\tau(\varepsilon)$ is motivated by the error estimate from Theorem 5.3.3 5).

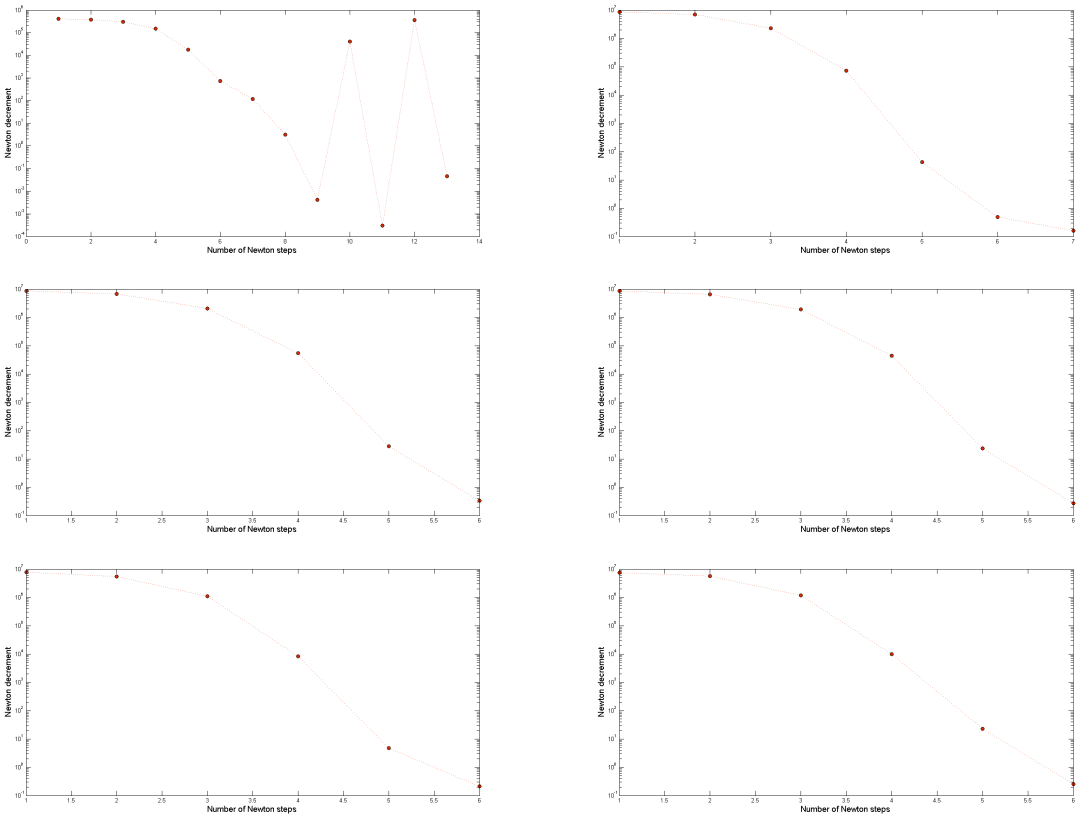

Figure 8.8. Test problem I: Newton decrements of LSM_ε on the different grids in a nested grid strategy for $\varepsilon = 10^{-4}$, from $h = 2^{-4}$ (top left) over $h = 2^{-5}$ (top right) to $h = 2^{-9}$ (bottom right)

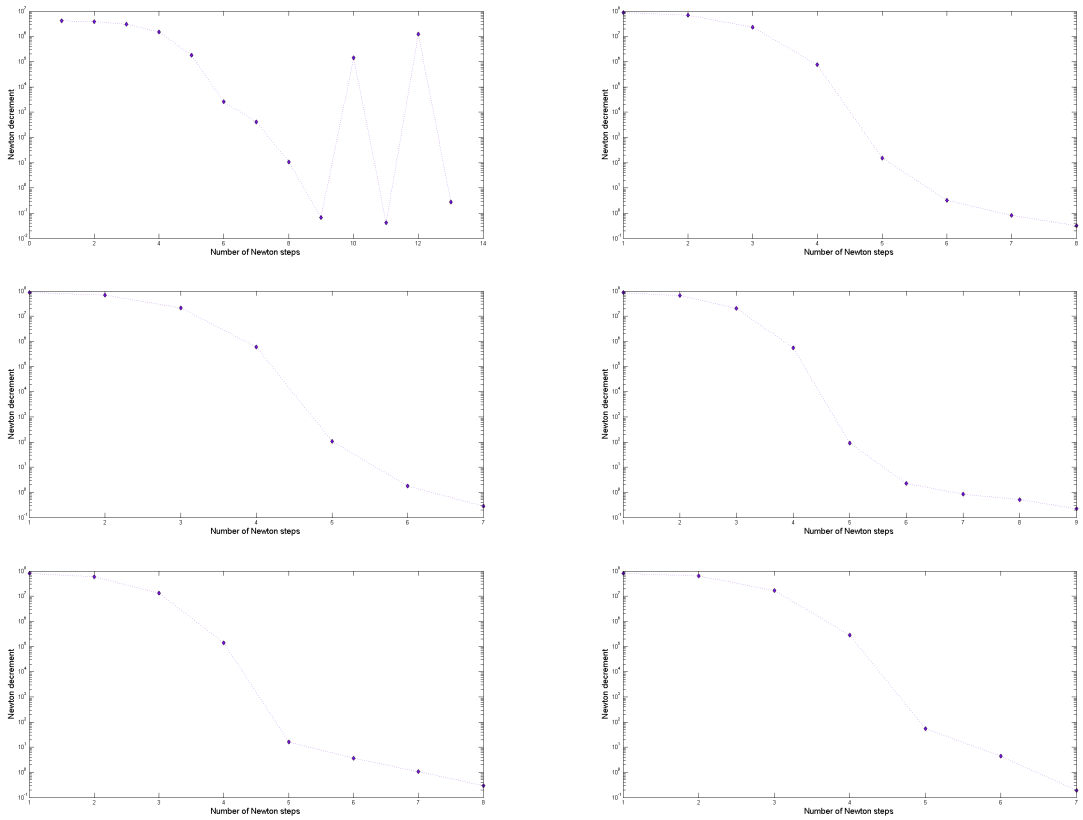

Figure 8.9. Test problem I: Newton decrements of LSM_ε on the different grids in a nested grid strategy for $\varepsilon = 10^{-5}$, from $h = 2^{-4}$ (top left) over $h = 2^{-5}$ (top right) to $h = 2^{-9}$ (bottom right)

In the experiments we conducted for efficiency so far, we worked with $C_j = \frac{2 \cdot 10^3}{\varepsilon^2}$ and $\tau(\varepsilon) = 2 \cdot 10^2 \cdot \frac{1+|\ln \varepsilon|}{\varepsilon}$. This shows that these weights become rather large, in particular for small values of ε. It is interesting to see what happens if smaller values are used, although self-concordance is no longer guaranteed and, therefore, our theory is no longer applicable. To this end, we rescale these weights such that $\tau(\varepsilon) = 1$ holds. More precisely, we use $C_j = \frac{10}{\varepsilon(1+|\ln \varepsilon|)}$ and $\tau(\varepsilon) = 1$. This yields the values and errors displayed in Table 8.5 and shows, in particular, that the overall errors are the same as before, cf. Table 8.3. This may indicate that it is possible to use these rescaled weights in practice. This is, it may not be necessary to know an exact lower bound for C_j, which is somewhat of a relief since such a bound seems to be hard to obtain, in general, because it involves the constant $C_{\partial, C(\overline{\Omega}_a)}$, cf. Lemma 3.5.8. In Table 8.6 we show the course of LSM_ε with rescaled weights for $\varepsilon = 10^{-4}$ and $h = 2^{-9}$ in detail.

ε	Mesh size $h = 2^{-i}$, $i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	9	9	9	9	9	9(5)	1.25×10^{-1}	1.22×10^{-1}	3.46×10^{-2}	1.26×10^{-3}
10^{-3}	10	10	11	11	12	12(8)	1.36×10^{-2}	1.62×10^{-2}	4.68×10^{-3}	1.26×10^{-3}
10^{-4}	10	10	11	11	11	12(8)	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}	1.26×10^{-3}
10^{-5}	10	10	11	11	11	11(7)	1.36×10^{-4}	2.81×10^{-3}	8.12×10^{-4}	1.26×10^{-3}

Table 8.5. Test problem I: Results of LSM_ε with modified weights C_j and $\tau(\varepsilon)$

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	8	1.04×10^{-3}	3.71×10^{-3}	1.07×10^{-3}
10^{-1}	2	1.33×10^{-3}	4.00×10^{-3}	1.16×10^{-3}
1.26×10^{-3}	2	1.36×10^{-3}	4.03×10^{-3}	1.17×10^{-3}

Table 8.6. Test problem I: Course of LSM_ε with modified weights C_j and $\tau(\varepsilon)$ for $\varepsilon = 10^{-4}$

8.2.2. Test Problem II

We further illustrate the efficiency of LSM_ε. To this end, we consider

$$\min_{(y,u) \in Y \times U} \frac{1}{2} \|y - y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2} \|u\|^2_{L^2(\Omega)} \qquad \text{s.t.} \qquad y \geq y_a \text{ in } \overline{\Omega}_a, \quad \begin{cases} -\Delta y + y = u & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases}$$

with $Y := H^2(\Omega) \cap H^1_0(\Omega)$, $U := L^2(\Omega)$, $\Omega := \Omega_a := (0,1) \times (0,1) \subset \mathbb{R}^2$, $\hat{\alpha} := 10^{-2}$, $y_d(x) := -\frac{1}{2}(\sin(2\pi x_1) + x_2)$, and $y_a \equiv -0.01$ on $\overline{\Omega}_a$. Using $Z := U$, $A := -\Delta + I \in \mathcal{L}(Y, Z)$, $B := -I \in \mathcal{L}(U, Z)$, and $g \equiv 0$, this problem is a small variation of Example 3.2.1 and it can be argued as for Example 3.2.1 that it satisfies Assumption 3.1.9 with, e.g., $u^\circ \equiv 0$.

Inspection of the numerical solution, for which we explain later how we obtained it, reveals that the optimal state $\bar{y}$ touches the bound y_a on a rather large, connected set. This stands in

contrast to Test Problem I, where the optimal state is active at a single point only. Figure 8.10 shows $\bar{y}_\varepsilon$ and $\bar{u}_\varepsilon$ for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$, together with $\bar{y}$, $\bar{u}$, and y_d. It is, for instance, visible how the feasibility violation of $\bar{y}_\varepsilon$ decreases as ε decreases and how $\bar{y}_\varepsilon$ approximates $\bar{y}$ with increasing accuracy (note that $\bar{u}_\varepsilon \to \bar{u}$ in U for $\varepsilon \to 0^+$ implies $\bar{y}_\varepsilon \to \bar{y}$ in Y and, due to $Y \hookrightarrow C(\overline{\Omega}_a)$, uniform convergence $\bar{y}_\varepsilon \to \bar{y}$ for $\varepsilon \to 0^+$).

We use mesh sizes up to $h = 2^{-10}$ and choose $\varepsilon \geq 10^{-5}$. We employ $\theta = 0.25$, $C_j = \frac{8 \cdot 10^3}{\varepsilon^2}$, $\tau(\varepsilon) = 8 \cdot 10^2 \cdot \frac{1+|\ln \varepsilon|}{\varepsilon}$, $C_{\hat{j}} = 1 + \hat{j}(u^0)$ with $u^0 \equiv 0$, $\mu_0 = 1$, $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 0.5$, and $\beta_0 = 0.1$. Lemma 3.6.4 provides $C_{\partial, C(\overline{\Omega}_a)} \leq 1.14$. Lemma 3.5.8 implies via $\tilde{u} \equiv 15$ that $f_{\varepsilon,\mu}$ is self-concordant for all $\mu \in (0, \mu_0]$ and all $\varepsilon \leq 1$, since it holds

$$\frac{8 \cdot 10^3}{\varepsilon^2} = C_j \geq \frac{1}{\varepsilon^2} \max \left\{ \varepsilon^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} \left(\frac{2 \|\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U \right)^2 \right\}.$$

As for Test Problem I we computed the quantity $\|\hat{j}'(\tilde{u})\|_{U^*}$ numerically on a uniform mesh with width $h = 2^{-9}$, cf. Remark 3.5.10, and incorporated a safeguard for the final choice of C_j.

The update β for μ is adaptively determined in the same way as for Test Problem I. Also, we use the same termination criterion, namely $\max\{\frac{|j(u^{k+1-i}) - j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i}) - \hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq \frac{\varepsilon(1+|\ln \varepsilon|)}{10}$ for $i = 1, 2$. Alternatively, we could prescribe μ_{final}, see [KU13, Section 4]. As optimal values $j(\bar{u})$ and $\hat{j}(\bar{u})$ we take the final values for $j(u^{k+1})$ and $\hat{j}(u^{k+1})$ obtained by LSM_ε with $\varepsilon = 10^{-7}$, $h = 2^{-10}$, and the termination criterion $\max\{\frac{|j(u^{k+1-i}) - j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i}) - \hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq \frac{\varepsilon(1+|\ln \varepsilon|)}{100}$.

We apply LSM_ε for different values of ε and different mesh sizes. The total numbers of Newton steps that have to be computed during the course of LSM_ε can be found in Table 8.7 and clearly indicate that LSM_ε is mesh independent. Moreover, this table contains the infeasibility $\|(\bar{y}_\varepsilon - y_a)^-\|_{C(\overline{\Omega}_a)}$ and the errors $\frac{|j(\bar{u}_\varepsilon) - j(\bar{u})|}{C_j}$ and $|\hat{j}(\bar{u}_\varepsilon) - \hat{j}(\bar{u})|$. We see that the infeasibility of $\bar{y}_\varepsilon$, indeed, behaves like $\mathcal{O}((1 + |\ln \varepsilon|)\varepsilon)$, cf. Corollary 4.4.4. Moreover, as the theory suggests, cf. Theorem 4.4.8, the error $\frac{|j(\bar{u}_\varepsilon) - j(\bar{u})|}{C_j}$ is of order $\mathcal{O}((1 + |\ln \varepsilon|)\varepsilon)$, too. This also seems to hold for the error $|\hat{j}(\bar{u}_\varepsilon) - \hat{j}(\bar{u})|$, which is not surprising since we expect this error to behave similar to $\frac{|j(\bar{u}_\varepsilon) - j(\bar{u})|}{C_j}$.

In Table 8.8 we display for $\varepsilon = 10^{-2}$, $i = 2, 3, 4$, and $h = 2^{-10}$ in detail how μ_k is decreased to μ_{final}. In particular, we observe that the errors with respect to j and $\hat{j}$ increase for $\mu_k \leq 10^{-2}$. This can be attributed to the fact that (u^k) converges to $\bar{u}_\varepsilon$, not to $\bar{u}$. In addition, Figure 8.11 shows the development of the Newton decrement during the course of LSM_ε for these values of ε and h. We can clearly recognize phase one as well as the points at which μ is decreased since these are exactly the points at which the Newton decrement increases.

If we employ $\max\{\frac{|j(u^{k+1-i}) - j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i}) - \hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq 10\varepsilon(1 + |\ln \varepsilon|)$ for $i = 1, 2$ as termination criterion, we obtain the iteration numbers and errors displayed in Table 8.9. Again, we clearly observe mesh independence. Furthermore, a comparison of the errors from this table with the ones from Table 8.7 shows that the less strict termination criterion is superior. In the remaining experiments for this test problem we, therefore, always employ this criterion.

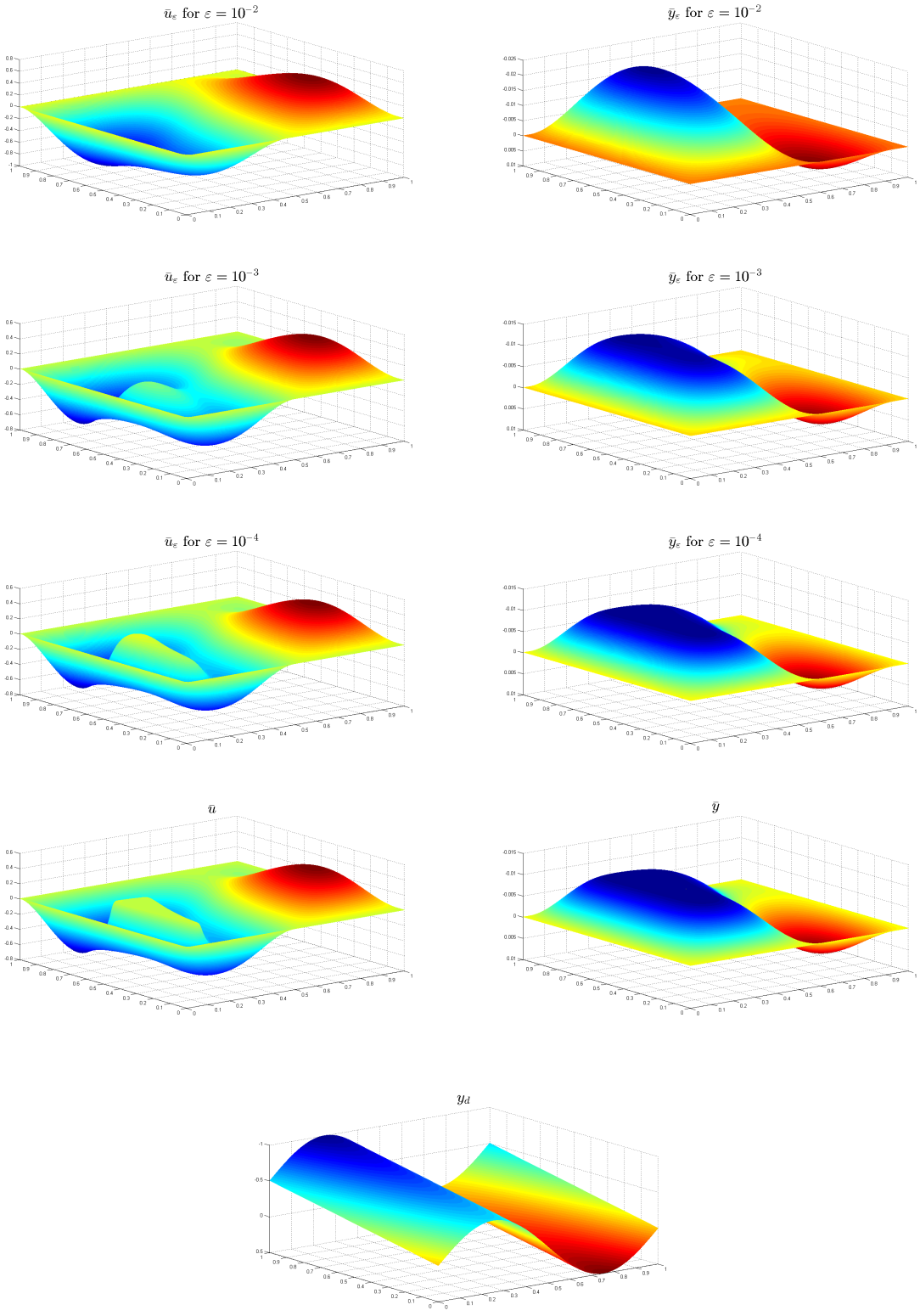

Figure 8.10. Test problem II: $\bar{u}_\varepsilon$ and $\bar{y}_\varepsilon$ (with inverted z-axis) for $\varepsilon = 10^{-\{2,3,4\}}$ together with $\bar{u}$ and $\bar{y}$ (with inverted z-axis) as well as y_d (with inverted z-axis)

ε	Mesh size $h = 2^{-i},\ i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	16	16	16	16	16	16(5)	1.46×10^{-2}	1.33×10^{-3}	1.33×10^{-3}	1×10^{-3}
10^{-3}	23	19	19	19	19	19(5)	2.44×10^{-3}	2.68×10^{-4}	2.68×10^{-4}	1×10^{-3}
10^{-4}	28	21	21	21	21	21(8)	3.11×10^{-4}	3.44×10^{-5}	3.43×10^{-5}	1×10^{-3}
10^{-5}	38	33	31	31	31	28(14)	4.13×10^{-5}	4.00×10^{-6}	3.99×10^{-6}	1×10^{-3}

Table 8.7. Test problem II: Total number of Newton steps required by LSM_ε; the Newton steps from LSMSUB and phase one are included; displayed in brackets is the number of Newton steps required by phase one; (y^K, u^K) denotes the final iterate

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	6	0	2.18×10^{-3}	2.18×10^{-3}
10^{-1}	4	1.01×10^{-2}	8.21×10^{-4}	8.19×10^{-4}
10^{-2}	4	1.41×10^{-2}	1.28×10^{-3}	1.28×10^{-3}
10^{-3}	3	1.46×10^{-2}	1.33×10^{-3}	1.33×10^{-3}

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	0	4.51×10^{-4}	4.50×10^{-4}
10^{-1}	5	1.85×10^{-3}	1.91×10^{-4}	1.90×10^{-4}
10^{-2}	5	2.39×10^{-3}	2.61×10^{-4}	2.61×10^{-4}
10^{-3}	4	2.44×10^{-3}	2.68×10^{-4}	2.68×10^{-4}

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	8	0	6.67×10^{-4}	6.66×10^{-5}
10^{-1}	5	2.37×10^{-4}	2.42×10^{-5}	2.42×10^{-5}
10^{-2}	4	3.04×10^{-4}	3.34×10^{-5}	3.34×10^{-5}
10^{-3}	4	3.11×10^{-4}	3.44×10^{-5}	3.43×10^{-5}

Table 8.8. Test problem II: Course of LSM_ε for $\varepsilon = 10^{-2}$ (top), $\varepsilon = 10^{-3}$ (middle), and $\varepsilon = 10^{-4}$ (bottom)

ε	Mesh size $h = 2^{-i},\ i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	13	13	13	13	13	13	1.41×10^{-2}	1.28×10^{-3}	1.28×10^{-3}	1×10^{-2}
10^{-3}	15	15	15	15	15	15	2.39×10^{-3}	2.61×10^{-4}	2.61×10^{-4}	1×10^{-2}
10^{-4}	18	17	17	17	17	17	3.04×10^{-4}	3.34×10^{-5}	3.34×10^{-5}	1×10^{-2}
10^{-5}	26	25	25	24	25	23	4.05×10^{-5}	3.89×10^{-6}	3.88×10^{-6}	1×10^{-2}

Table 8.9. Test problem II: Results of LSM_ε with less strict termination criterion

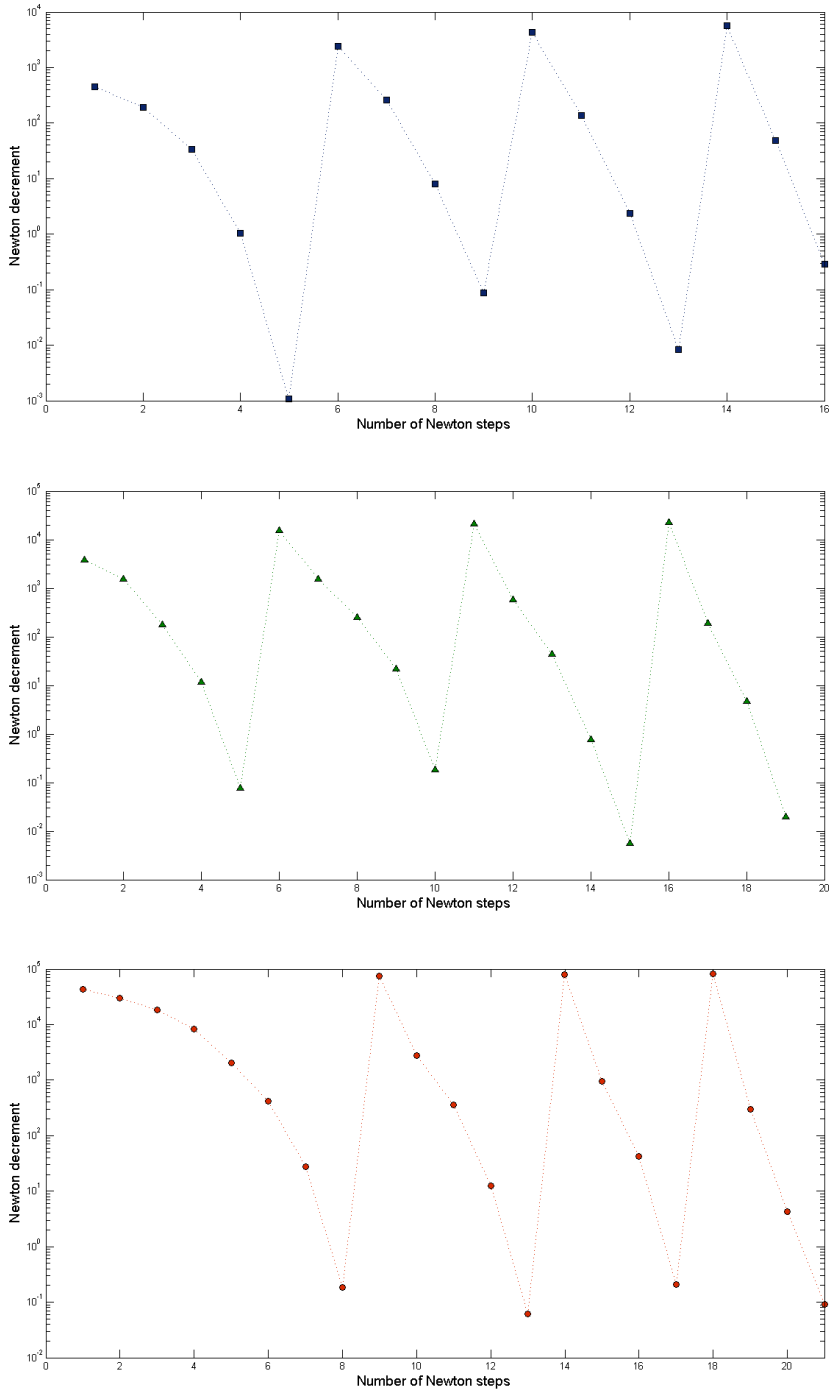

Figure 8.11. Test problem II: Newton decrements of LSM_ε for $\varepsilon = 10^{-2}$ (top), $\varepsilon = 10^{-3}$ (middle), and $\varepsilon = 10^{-4}$ (bottom)

We now add a nested grid strategy to LSM$_\varepsilon$. We use the same strategy as for Test Problem I with a grid hierarchy ranging from $h = 2^{-5}$ to $h = 2^{-10}$. As for Test Problem I we have to ensure feasibility when prolongating onto a finer grid, which may require a shift towards a feasible point. In contrast to Test Problem I, where a shift is often necessary, we observe in this test problem that a shift is only required for $\varepsilon = 10^{-5}$ and only on the coarsest grid. In this test problem we shift towards $u^0 \equiv 0$. The size of the shift is computed as for Test Problem I, only that we change the value of κ to $\kappa := 0.999$.

In Table 8.10 we show the results of LSM$_\varepsilon$ with this nested grid strategy. An additional (s) indicates that a shift is necessary. In Figure 8.12 and 8.13 we display for $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-5}$ how the Newton decrement develops during the course of the algorithm.

For Test Problem I we pointed out that the numbers of Newton steps required on the finer grids seems to be rather large for a nested grid strategy, cf. Table 8.4. We attributed this to the shift that is necessary to ensure feasibility. This view is further encouraged by the fact that in this example, where a shift is not required on finer meshes, the number of Newton steps on finer grids is substantially lower, in particular in comparison to the overall iteration number without a nested strategy. For instance, for $\varepsilon = 10^{-4}$ only 2 Newton steps are required on the finest mesh, which is dramatically lower than the 17 Newton steps needed without nesting, cf. Table 8.9. Moreover, a comparison of the values for μ_{final} from Table 8.10 with the ones from Table 8.9 shows that the the determination of μ_{final} on the coarsest mesh works very well in this test problem. Hence, the overall errors are exactly the same as for LSM$_\varepsilon$ without nesting, as is confirmed by the same tables. We conclude that nesting improves the efficiency of LSM$_\varepsilon$ greatly for this test problem.

ε	Mesh size $h = 2^{-i}$, $i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	13	3	3	3	2	2	1.41×10^{-2}	1.28×10^{-3}	1.28×10^{-3}	1×10^{-2}
10^{-3}	15	4	3	3	2	2	2.39×10^{-3}	2.61×10^{-4}	2.61×10^{-4}	1×10^{-2}
10^{-4}	18	4	4	3	3	2	3.04×10^{-4}	3.34×10^{-5}	3.34×10^{-5}	1×10^{-2}
10^{-5}	26(s)	7	6	5	6	4	4.05×10^{-5}	3.89×10^{-6}	3.88×10^{-6}	1×10^{-2}

Table 8.10. Test problem II: Results of LSM$_\varepsilon$ with a nested grid strategy

In the last experiment for Test Problem II we rescale the weights C_j and $\tau(\varepsilon)$ to $C_j = \frac{10}{\varepsilon(1+|\ln \varepsilon|)}$ and $\tau(\varepsilon) = 1$. The results are displayed in Table 8.11. The table shows, in particular, that the overall errors are basically the same as for the original weights but with the more strict termination criterion, cf. Table 8.7. This is, the rescaled weights yield the same accuracy with a less strict termination criterion than the original weights with a more strict termination criterion. This encourages the view that it is possible to use these rescaled weights in practice. In Table 8.12 we show the course of LSM$_\varepsilon$ with these rescaled weights for $\varepsilon = 10^{-4}$ and $h = 2^{-10}$ in detail.

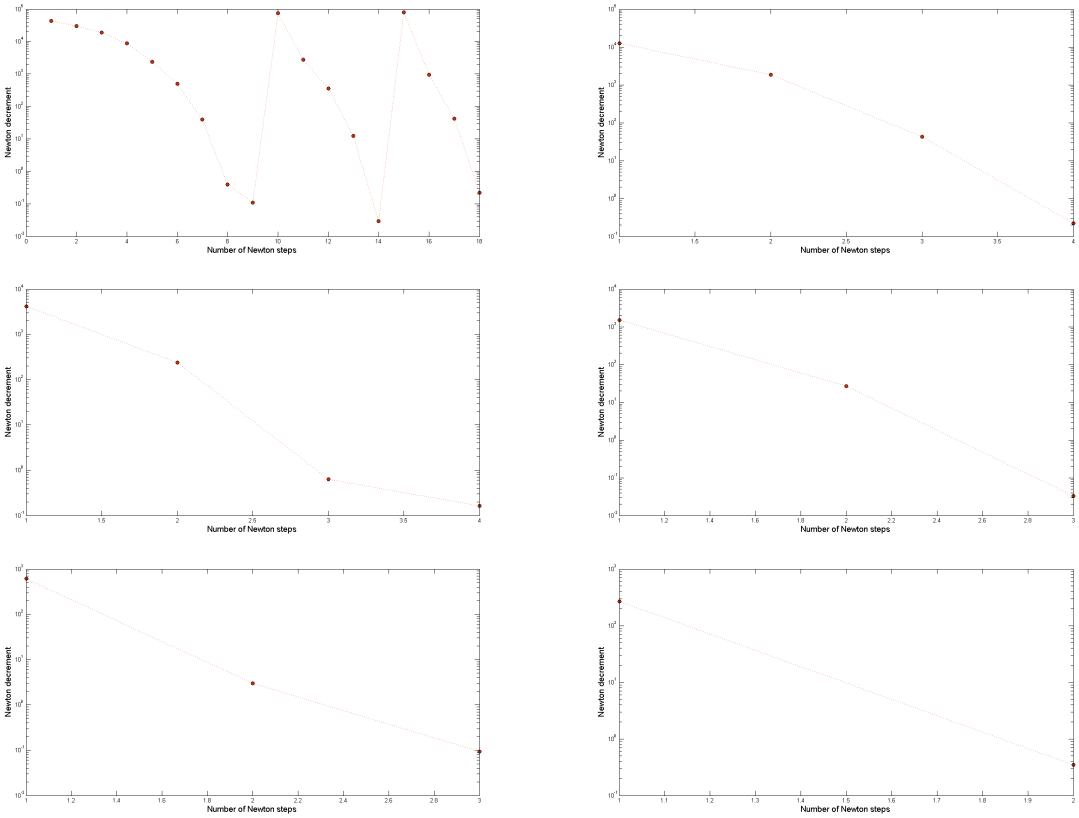

Figure 8.12. Test problem II: Newton decrements of LSM_ε on the different grids in a nested grid strategy for $\varepsilon = 10^{-4}$, from $h = 2^{-5}$ (top left) over $h = 2^{-6}$ (top right) to $h = 2^{-10}$ (bottom right)

ε	Mesh size $h = 2^{-i}$, $i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	8	7	7	7	7	7(2)	1.44×10^{-2}	1.31×10^{-3}	1.31×10^{-3}	4.59×10^{-3}
10^{-3}	9	9	9	9	9	9(3)	2.42×10^{-3}	2.66×10^{-4}	2.65×10^{-4}	4.59×10^{-3}
10^{-4}	11	11	11	11	11	11(5)	3.11×10^{-4}	3.43×10^{-5}	3.43×10^{-5}	1.26×10^{-3}
10^{-5}	14	14	14	14	14	14(10)	4.12×10^{-5}	4.00×10^{-6}	3.99×10^{-6}	1.26×10^{-3}

Table 8.11. Test problem II: Results of LSM_ε with modified weights C_j and $\tau(\varepsilon)$

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	0	6.36×10^{-5}	6.35×10^{-5}
10^{-1}	2	2.35×10^{-4}	2.44×10^{-5}	2.44×10^{-5}
1.26×10^{-3}	4	3.11×10^{-4}	3.43×10^{-5}	3.43×10^{-5}

Table 8.12. Test problem II: Course of LSM_ε with modified weights C_j and $\tau(\varepsilon)$ for $\varepsilon = 10^{-4}$

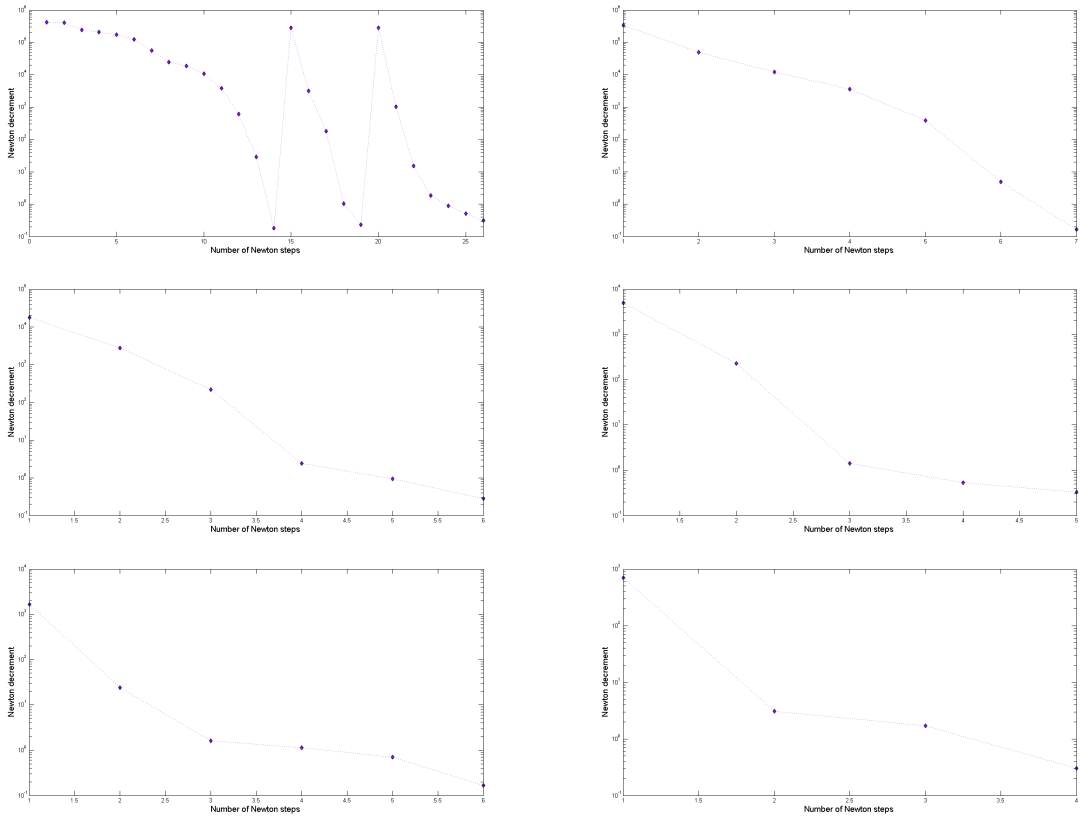

Figure 8.13. Test problem II: Newton decrements of LSM_ε on the different grids in a nested grid strategy for $\varepsilon = 10^{-5}$, from $h = 2^{-5}$ (top left) over $h = 2^{-6}$ (top right) to $h = 2^{-10}$ (bottom right)

8.2.3. Test Problem III

We present an example with a semilinear state equation to further examine LSM_ε. In particular, we want to demonstrate that LSM_ε can also be used to solve problems that are not covered by the theory developed in this thesis. We consider the problem

$$\min_{(y,u)\in Y\times U} \frac{1}{2}\|y-y_d\|^2_{L^2(\Omega)} + \frac{\hat{\alpha}}{2}\|u\|^2_{L^2(\Omega)} \quad \text{s.t.} \quad y \geq y_a \text{ in } \overline{\Omega}_a, \quad \begin{cases} -\Delta y + y + y^3 = u & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases}$$

with $Y := H^2(\Omega)\cap H^1_0(\Omega)$, $U := L^2(\Omega)$, $\Omega_a := \Omega := (0,1)^2$, $y_d(x) := 3\sin(4\pi x_1)\cos(8\pi x_2)e^{2x_1}$, $\hat{\alpha} := 10^{-5}$, and $y_a \equiv -2$ in $\overline{\Omega}_a$. The choice of y_d in this problem is inspired by [Sta09, Example 3]. It can be proven by standard arguments, see, e.g., [HPUU09, Section 1.5.2], that this problem possesses at least one optimal solution $(\bar{y},\bar{u}) \in Y \times U$. We explained in Section 8.1.3 how to compute Newton steps for this semilinear state equation. We remark again that this nonlinearity introduces nonconvexity into the reduced problem and, therefore, the theory developed in this thesis is not applicable.

As optimal state $\bar{y}$ and optimal control $\bar{u}$ we employ the final iterate of LSM_ε with $\varepsilon = 10^{-7}$ and $h = 2^{-10}$. Note that the termination criterion of LSM_ε is based on convergence of the objective function. Therefore, the fact that LSM_ε terminates suggests, at least, convergence of $(j(u^k))$ and $(\hat{j}(u^k))$ for the generated sequence (u^k). Inspection of this numerical solution reveals that the optimal state $\bar{y}$ touches the bound y_a at three different small sets, which are approximately located at $(0.9, 0.25)$, $(0.9, 0.5)$, and $(0.9, 0.75)$. We depict $\bar{y}$, $\bar{u}$, and y_d in Figure 8.14, where we also display the final iterate of LSM_ε for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$. As before we call the final iterate $(\bar{u}_\varepsilon, \bar{y}_\varepsilon)$ in this figure.

In Test Problem I and II we observed that LSM_ε performs very well with the rescaled weights $C_j = \frac{10}{\varepsilon(1+|\ln\varepsilon|)}$ and $\tau(\varepsilon) = 1$. Therefore, we use these weights in all experiments for Test Problem III. Furthermore, we employ $C_{\hat{j}} = 1 + \hat{j}(u^0)$ with $u^0 \equiv 0$. Apart from this all other parameters remain unchanged in comparison to Test Problem I and Test Problem II.

The update β_k for μ_k in iteration $k \in \mathbb{N}$ is adaptively determined in the same way as for Test Problem I and II. Also, we use the same (strict) termination criterion, i.e., we require $\max\{\frac{|j(u^{k+1-i})-j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i})-\hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq \frac{\varepsilon(1+|\ln\varepsilon|)}{10}$ for $i = 1, 2$. To obtain the optimal values $j(\bar{u})$ and $\hat{j}(\bar{u})$ we even used $\max\{\frac{|j(u^{k+1-i})-j(u^{k+1})|}{C_j}, |\frac{\hat{j}(u^{k+1-i})-\hat{j}(u^{k+1})}{\hat{j}(u^{k+1})}|\} \leq \frac{\varepsilon(1+|\ln\varepsilon|)}{100}$ for $i = 1, 2$.

The barrier $f_{\varepsilon,\mu}$ cannot be self-concordant in this example since the very definition of self-concordance requires convexity. Hence, we modify the choice of the step size and the termination criterion in LSMSUB. The reason is that this choice and this criterion rely heavily on self-concordance. In the absence of self-concordance it could, for instance, happen that infeasible iterates occur in LSMSUB or that the point $\tilde{y}$ returned by LSMSUB does not satisfy $\lambda_{\varepsilon,\mu_k}(\tilde{y}) \leq \theta$ when LSMSUB is called in iteration k of LSM_ε (we write $\tilde{y}$ rather than $\tilde{u}$ since in the implementation we use reduction to the state, cf. also Section 8.1.2). The latter problem is solved by changing the termination criterion: We terminate LSMSUB if it produces an iterate $\tilde{y}^l$ with $\lambda_{\varepsilon,\mu_k}(\tilde{y}^l) \leq \theta$. To detect $\lambda_{\varepsilon,\mu_k}(\tilde{y}^l) \leq \theta$ we have to compute the corresponding Newton step $n_{\tilde{y}^l}$. Since this Newton step is available then, we can cheaply compute a new iterate $\tilde{y}^{l+1}$

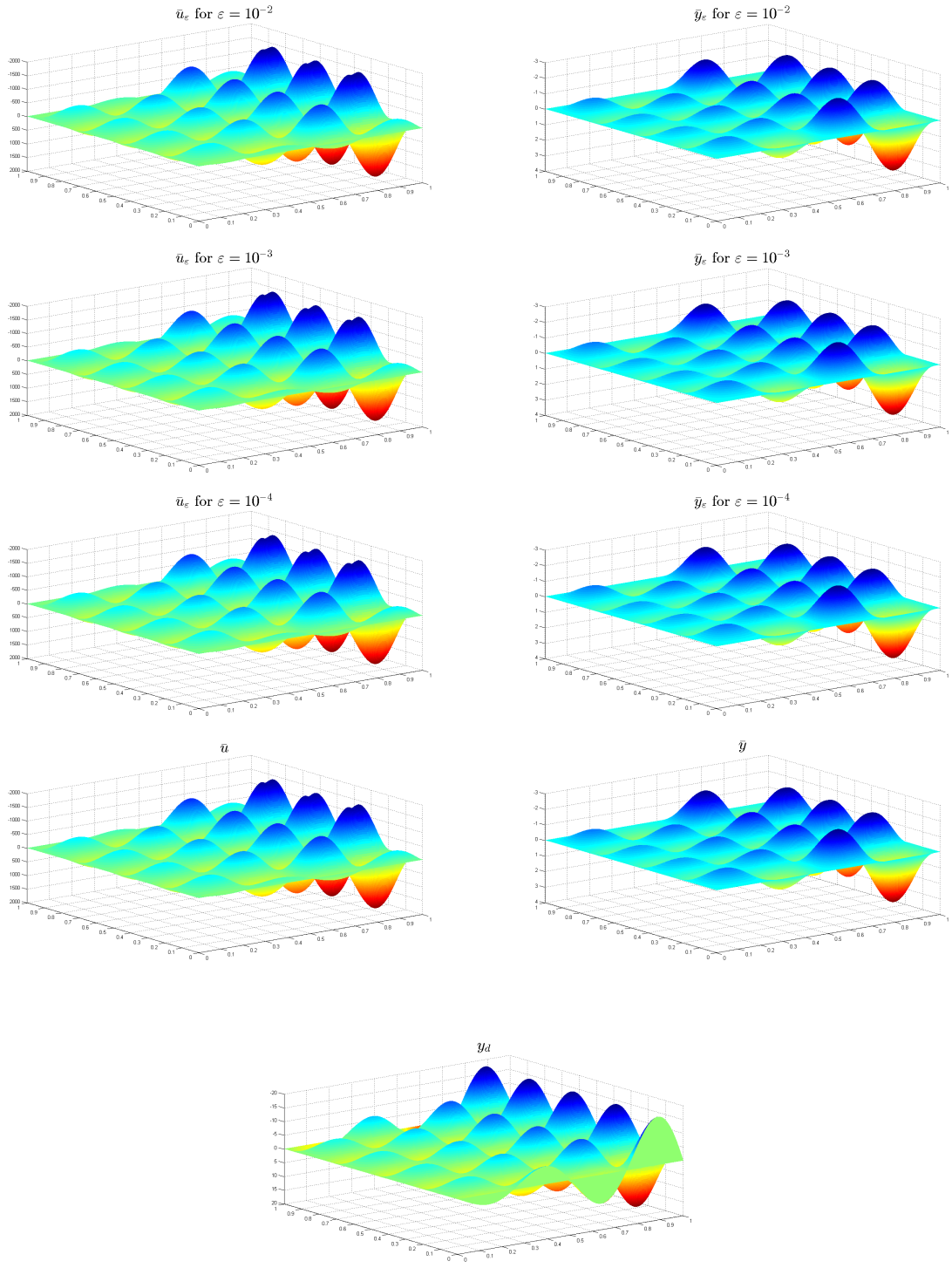

Figure 8.14. Test problem III: $\bar{u}_\varepsilon$ and $\bar{y}_\varepsilon$ for $\varepsilon = 10^{-\{2,3,4\}}$ together with $\bar{u}$, $\bar{y}$, and y_d (all with inverted z-axis)

by application of a line search method. Since this implies that $\tilde{y}^{l+1}$ has smaller function value than $\tilde{y}^l$, LSMSUB returns $\tilde{y} := \tilde{y}^{l+1}$. We remark that $\lambda_{\varepsilon,\mu_k}(\tilde{y}^l)$ may not even be a sensible expression if f_{ε,μ_k} is not convex. However, during our numerical experiments $\lambda_{\varepsilon,\mu_k}(\tilde{y}^l)$ could always be computed.

It remains to explain how we modify the choice of the step size. We mention in advance that the strategy described in the following can be carried out without solving the state equation, since we work with reduction to the state. Let LSMSUB be called in iteration $k \in \mathbb{N}_0$ of LSM$_\varepsilon$ and let $\tilde{y}^l$, $l \in \mathbb{N}_0$, denote a state during the course of LSMSUB. Moreover, denote the corresponding Newton step by $n_{\tilde{y}^l}$. We want to find a suitable step size $t^* \in (0,1]$ and use $\tilde{y}^{l+1} := \tilde{y}^l + t^* n_{\tilde{y}^l}$ as new iterate. In particular, this means that $\tilde{y}^l + t^* n_{\tilde{y}^l}$ should be feasible. To attain this, we first check if $\tilde{y}^l + n_{\tilde{y}^l}$ is feasible and if $f_{\varepsilon,\mu_k}(\tilde{y}^l + n_{\tilde{y}^l}) < f_{\varepsilon,\mu_k}(\tilde{y}^l)$. If so, we use $t^* = 1$. In particular, we hope that this choice ensures the locally quadratic convergence of Newton's method. If $\tilde{y}^l + n_{\tilde{y}^l}$ is not feasible or does not decrease the value of f_{ε,μ_k}, then we use *fzero* to detect a $\hat{t} \in (0,1]$ such that $\tilde{y}^l + \hat{t} n_{\tilde{y}^l}$ lies on the boundary of the feasible set. Since we observed in the numerical experiments that feasibility seems to be a problem only with respect to $B^\varepsilon_{C(\overline{\Omega}_a)}$, we can expect that $\hat{t}$ is unique and that all smaller t are feasible due to the convexity of $\{y \in Y : B^\varepsilon_{C(\overline{\Omega}_a)}(y) \geq 0\}$; we presented a similar argument in more detail in Section 8.2.1 when we discussed how to determine the size of the shift in the nested grid strategy. We now employ *fminbnd* to determine a step size $t^* \in [0, \iota\hat{t}]$ that (locally) minimizes $f_{\varepsilon,\mu_k}(y^k + t n_{y^k})$ on $[0, \iota\hat{t}]$, where $\iota \in [0.5, 1)$ is a factor that safeguards the iterates against coming to close to the boundary of the feasible set. This resembles the fraction-to-the-boundary rule, cf. [WB06, Section 2.2]. We choose ι depending on the size of $\hat{t}$, with ι close to 1 for $\hat{t}$ close to 1. We note that t^* may be very small or even zero, since $\hat{t}$ may be small and since n_{y^k} may not be a (good) descent direction. This issue could, for instance, be addressed by incorporation of negative gradient steps if Y is a Hilbert space, or a Levenberg-Marquardt-type regularization. The use of negative gradient steps in case that some other search direction, e.g., the Newton step, fails to be a descent direction is a common technique to ensure global convergence of optimization algorithms in Hilbert spaces. More generally speaking it is certainly possible to further adapt LSM$_\varepsilon$ to nonconvex problems. However, with the modifications described above we can already demonstrate that LSM$_\varepsilon$ is capable of solving problems that are not covered theoretically. In particular, in all experiments that we conduct for this test problem LSM$_\varepsilon$ converges successfully.

We apply the modified version of LSM$_\varepsilon$ to Test Problem III for different values of ε. In Table 8.13 we display the total number of Newton steps that is required during the course of LSM$_\varepsilon$, along with $\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$, $\frac{|j(u^K) - j(\bar{u})|}{C_j}$, $|\hat{j}(u^K) - \hat{j}(\bar{u})|$, and μ_{final} on the finest mesh. The results indicate that LSM$_\varepsilon$ is mesh independent and can be successfully applied even in cases that are not covered by the theory developed in this thesis.

We note that the numbers of Newton steps in Table 8.13 seem to be rather large in comparison to the computations with the same modified weights for Test Problem I and II, cf. Table 8.5 and 8.11 (although in the computations for Table 8.5 and 8.11 we used a less strict termination criterion, but this does not make much of a difference). We conjecture that this is due to the rather small value of $\hat{\alpha}$. Indeed, if we choose $\hat{\alpha} \geq 10^{-4}$, the required numbers of Newton steps are similar to the ones in Table 8.5 and 8.11. Moreover, we observe that basically only the

number of Newton steps required by phase one increases as ε decreases. This indicates that the iterate obtained by phase one is already close to the final iterate.

In Table 8.14 we display the course of LSM_ε for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$, and $h = 2^{-10}$ in detail.

ε	Mesh size $h = 2^{-i},\ i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	11	10	10	10	10	10(5)	7.92×10^{-2}	2.06×10^{-3}	7.07×10^{-4}	1.26×10^{-3}
10^{-3}	15	16	16	14	14	14(8)	9.87×10^{-3}	2.82×10^{-4}	9.71×10^{-5}	1.26×10^{-3}
10^{-4}	18	25	27	23	18	17(11)	1.10×10^{-3}	3.18×10^{-5}	1.09×10^{-5}	4.59×10^{-3}
10^{-5}	37	37	34	34	36	34(28)	1.17×10^{-4}	3.36×10^{-6}	1.16×10^{-6}	1.26×10^{-3}

Table 8.13. Test problem III: Results of LSM_ε

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	0	8.91×10^{-3}	3.07×10^{-3}
10^{-1}	2	2.18×10^{-2}	5.98×10^{-4}	2.06×10^{-4}
1.26×10^{-3}	3	7.92×10^{-2}	2.06×10^{-3}	7.07×10^{-4}

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	8	0	1.67×10^{-3}	5.76×10^{-4}
10^{-1}	2	2.10×10^{-3}	5.76×10^{-5}	1.98×10^{-5}
1.26×10^{-3}	4	9.87×10^{-3}	2.82×10^{-4}	9.71×10^{-5}

μ_k	#steps	$\|(y^{k+1} - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	11	0	2.47×10^{-4}	8.50×10^{-5}
10^{-1}	3	1.39×10^{-4}	3.54×10^{-6}	1.22×10^{-6}
4.59×10^{-3}	3	1.10×10^{-3}	3.18×10^{-5}	1.09×10^{-5}

Table 8.14. Test problem III: Course of LSM_ε for $\varepsilon = 10^{-2}$ (top), $\varepsilon = 10^{-3}$ (middle), and $\varepsilon = 10^{-4}$ (bottom)

Lastly, we employ a nested grid strategy. We use the same strategy as for Test Problem I and II with a grid hierarchy ranging from $h = 2^{-5}$ to $h = 2^{-10}$. As for the other test problems we have to ensure feasibility when prolongating onto a finer grid. Before, we used interpolation and shifting for the control to realize this. This requires two additional solves of the state equation and is, therefore, not very costly. However, in this test problem the state equation is nonlinear and, therefore, more costly to solve. We circumvent this by using the state rather than the control for interpolation and shifting. The shifting is done in the same manner as before for the control. Since the state is smoother than the control, we use spline interpolation instead of linear interpolation. This is, in particular, sensible since linear interpolation can be expected to cause problems when applying (a discretized version of) Δ. In this test problem we shift towards $y^0 \equiv 0$ and use $\kappa := 0.999$.

In Table 8.15 we show the results of LSM_ε with nested grid strategy. As in Test Problem I we

observe that when a shift is required the number of Newton steps seems to be rather large for a nested grid strategy. In particular, on the mesh with width $h = 2^{-6}$ these numbers are even larger than without nesting, cf. Table 8.13. However, on the finer meshes the nesting technique still reduces the required number of Newton steps significantly. We suspect that the nesting strategy can be further improved.

ε	Mesh size $h = 2^{-i},\ i =$						$\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	μ_{final}
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-2}	11(s)	9	4	2	1	1	7.69×10^{-2}	2.00×10^{-3}	6.89×10^{-4}	4.59×10^{-3}
10^{-3}	15(s)	17(s)	10	3	3	1	9.60×10^{-3}	2.74×10^{-4}	9.45×10^{-5}	4.59×10^{-3}
10^{-4}	18(s)	27(s)	17(s)	13(s)	8(s)	6	1.14×10^{-3}	3.28×10^{-5}	1.13×10^{-5}	1.26×10^{-3}
10^{-5}	37(s)	39(s)	22(s)	19(s)	17(s)	12	1.17×10^{-4}	3.36×10^{-6}	1.16×10^{-6}	1.26×10^{-3}

Table 8.15. Test problem III: Results of LSM_ε with a nested grid strategy

8.3. Numerical results for variable smoothing parameter

In this section we present numerical experiments for Version B of Algorithm $\text{LSM}_{(\text{P})}$. Since the backtracking from Corollary 7.2.22 only allows for small updates of ε and μ, we use the one from Corollary 7.2.20. The main result for this algorithm is Theorem 7.2.26.

8.3.1. Test Problem I

The first problem under consideration is identical to Test Problem I for fixed ε, see Section 8.2.1. We choose $\varepsilon_s = 1$, $C_j = 2 \cdot 10^3$, $\tau(\varepsilon) = C_\tau \frac{1+|\ln \varepsilon|}{\varepsilon}$ with $C_\tau = 2 \cdot 10^3$ for all $\varepsilon \in (0, \varepsilon_s]$, and $C_{\hat{j}} = 1 + \hat{j}(u^0)$ with $u^0 \equiv \frac{1}{2\pi\hat{\alpha}}$. We have to show that these parameters satisfy Assumption 6.1.1. To this end, we note that $\tau(\varepsilon)$ can be extended to a continuously differentiable, positive function in $\mathbb{R}_{>0}$ and that we obviously have $\tau(\varepsilon) \geq 1$ in $(0, \varepsilon_s]$. Moreover, we reasoned in Section 8.2.1 that $C_j = \frac{2 \cdot 10^3}{\varepsilon^2}$ ensures self-concordance of $f_{\varepsilon,\mu}$ for all $\mu \in (0, 1]$ since it holds

$$\frac{2 \cdot 10^3}{\varepsilon^2} \geq \frac{1}{\varepsilon^2} \max \left\{ \varepsilon^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} \left(\frac{\|2\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U \right)^2 \right\}$$

for all $\varepsilon \leq 1$ and $\tilde{u} \equiv 1.1 u^0$. Enlarging $C_{\partial, C(\overline{\Omega}_a)}$ if necessary this implies

$$2 \cdot 10^3 = \max \left\{ \varepsilon_s^2, \frac{16}{9} C^2_{\partial, C(\overline{\Omega}_a)} \left(\frac{\|2\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U \right)^2 \right\}.$$

Summarizing, this shows that Assumption 6.1.1 is fulfilled for this test problem. Moreover, Assumption 6.1.9 is valid as follows by use of $\hat{y}(x) := 1 - \|x\|_2^2 \in Y = H^2(\Omega) \cap H_0^1(\Omega)$ and $\hat{u} := -\Delta \hat{y}$. Also, we argued in Section 8.2.1 that this test problem satisfies Assumption 3.1.9. Together, we have established all assumptions that are required to apply the theory developed in this thesis.

The choice of C_τ is based on the overall error estimate

$$\frac{|j(u^{k+1}) - j(\bar{u})|}{C_j} \leq \frac{2C_\tau \varepsilon_k (1 + |\ln \varepsilon_k|)}{C_j} + \|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \varepsilon_k \left(C + \frac{d}{\beta} |\ln \varepsilon_k| \right), \tag{8.3}$$

where $\bar{\lambda}$ denotes a Lagrange multiplier associated to $\bar{u}$ for $C_j = 1$, cf. Lemma 3.4.3. This estimate is a more detailed version of the estimate from Theorem 7.2.26, cf. also Theorem 4.4.8. The aim is to choose C_τ such that the summands on the right-hand side are well-balanced for $\varepsilon_k \to 0^+$, i.e., have similar order of magnitude. To explain how we make this choice let us assume for the moment that we have an estimate for $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$. For $\varepsilon_k \to 0^+$ the dominant parts on the right-hand side in (8.3) are $\frac{2C_\tau}{C_j} \varepsilon_k |\ln \varepsilon_k|$ and $\frac{d}{\beta} \|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \varepsilon_k |\ln \varepsilon_k|$. Due to Sobolev embeddings every $\beta < 1$ is admissible in this problem, which yields $\frac{d}{\beta} \approx 2$. This leads to the choice $C_\tau \approx C_j \|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$.

It remains to explain how to estimate $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$. To do this, we use the result

$$\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq \sqrt{2\hat{j}(\bar{u})} \left(\sqrt{\mathrm{vol}(\Omega)} + \|A\mathbf{1}\|_U \right) \tag{8.4}$$

from Lemma 4.4.10. We note that the prerequisite $\mathbf{1} \in Y$ is not satisfied here since the state equation contains homogeneous Dirichlet boundary conditions. However, we suspect that the order of magnitude that Lemma 4.4.10 provides for $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ is correct, anyway. From (8.4) we derive $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq \sqrt{2\hat{j}(\bar{u})\pi} \leq \sqrt{5\pi} \approx 4$, where we used $\hat{j}(\bar{u}) \leq 2.5$. Of course, in this example we even know $\hat{j}(\bar{u}) = 6.25/\pi \leq 2$. Since, in general, we do not know $\hat{j}(\bar{u})$, we remark that an upper bound for $\hat{j}(\bar{u})$ is, for instance, provided by any u that is feasible for the reduced problem. Also, $\mathrm{LSM}_{(\mathrm{P})}$ may be used to refine an estimate for $\hat{j}(\bar{u})$: We can choose C_τ based on $\hat{j}(u)$ with a feasible u and then use $\mathrm{LSM}_{(\mathrm{P})}$ with this C_τ to compute a better approximation for $\hat{j}(\bar{u})$; we can use a coarse grid for this since we are only interested in a rough approximation. Since we estimated $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq 4$ and since the real value of $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ is somewhere around $1/2$ (in this example we know that the multiplier associated to $\hat{j}$ is a Dirac at the origin, which implies that $\bar{\lambda}$ is a scaled version of this Dirac; note, furthermore, that the estimate for $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ has the right order of magnitude, indeed), we include as safeguard a factor of $\frac{1}{4}$ to be closer to the real value of $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$. We mention that we employ the same safeguard also in examples where $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ is unknown. Finally, we arrive at the choice $C_\tau = C_j = 2 \cdot 10^3$. We later also investigate how $\mathrm{LSM}_{(\mathrm{P})}$ performs for other choices of C_τ.

We choose $\theta = 0.25$, $\varepsilon_0 = 1$, $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 0.5$, and $\tilde{\beta}_0 = 0.1$ for $k = 0$ in the backtracking strategy. Furthermore, we incorporate a strategy to adaptively choose $\tilde{\beta}_0$ in the backtracking strategy. We recall that the convergence theory we developed covers this case, provided $\tilde{\beta}_0 \in [\beta_{\min}, \beta_{\max}]$ holds throughout $\mathrm{LSM}_{(\mathrm{P})}$, cf. Corollary 7.2.20. The strategy we use works as follows: We prescribe a natural number $m \geq 2$. If LSMSUB is called from $\mathrm{LSM}_{(\mathrm{P})}$ in iteration $k \in \mathbb{N}$ and requires more than $m + 1$ Newton steps, then we choose in iteration $k + 1$ a $\tilde{\beta}_0 \in (\beta_{\min}, \beta_{\max})$ that is larger than the $\tilde{\beta}_0$ that was used in iteration k. In addition, we choose the new $\tilde{\beta}_0$ the closer to $\beta_{\max}$, the larger m is. Analogously, if LSMSUB requires less than $m - 1$ Newton steps, then the new $\tilde{\beta}_0 \in (\beta_{\min}, \beta_{\max})$ is smaller than the previous $\tilde{\beta}_0$, with

the new $\tilde{\beta}_0$ the closer to $\beta_{\min}$, the smaller m is. We use $m = 6$ in all experiments, also for other test problems.

In the first couple of experiments and unless stated otherwise we prescribe a value $\varepsilon_{\text{final}}$ and terminate $\text{LSM}_{(\text{P})}$ when $\varepsilon_k = \varepsilon_{\text{final}}$ is satisfied. To ensure that $\varepsilon_k = \varepsilon_{\text{final}}$ actually occurs we modify the update strategy for ε such that if $\varepsilon_k \beta_k < \varepsilon_{\text{final}}$ holds, then $\varepsilon_{k+1} = \varepsilon_{\text{final}}$ is used instead of $\varepsilon_{k+1} = \varepsilon_k \beta_k$.

In the first experiment we apply $\text{LSM}_{(\text{P})}$ for different values of $\varepsilon_{\text{final}}$ on uniform meshes with different mesh sizes h. We choose $\varepsilon_{\text{final}} \geq 10^{-5}$ and use $h \geq 2^{-9}$. The total numbers of Newton steps that have to be computed during the course of $\text{LSM}_{(\text{P})}$ are listed in Table 8.16 and clearly indicate that $\text{LSM}_{(\text{P})}$ is mesh independent.

The numbers of Newton steps displayed in Table 8.16 contain a phase one. The Newton steps required by phase one are displayed in brackets for $\varepsilon_{\text{final}} = 10^{-2}$ on the finest mesh. For other mesh sizes phase one requires exactly the same number of Newton steps. Moreover, for other values of $\varepsilon_{\text{final}}$ these numbers are the same since phase one is carried out for $\varepsilon_0 = 1$ and is, therefore, not affected by a change of $\varepsilon_{\text{final}}$. The fact that phase one is carried out for $\varepsilon_0 = \mu_0 = 1$ implies that the number of Newton steps required by phase one in $\text{LSM}_{(\text{P})}$ can be expected to be smaller than in LSM_ε with $\mu_0 = 1$ and a small value of ε. A comparison with, e.g., Table 8.1 confirms this. We remark that, as in our tests with fixed ε, we count as phase one all Newton steps until a $\tilde{u}$ is found for which $\tilde{u} \in \lambda_{\varepsilon_0,\mu_0}(\tilde{u}) \leq \theta$ is ensured. This equals the number of Newton steps that are taken until ε_0 is decreased to ε_1. In all further tables numbers in brackets denote the Newton steps required by phase one in the sense just explained. Also, we always denote by (y^K, u^K) the final iterate.

Table 8.16 shows that the final state is feasible with respect to $\min(y - y_a) \geq 0$ for all displayed values of $\varepsilon_{\text{final}}$. Of course, this is not true in general for $\text{LSM}_{(\text{P})}$. In fact, it can be expected that the size of the feasibility, respectively, infeasibility with respect to the smoothed minimum depends on the choice of C_τ or, more precisely, on the ratio C_j/C_τ: In $f_{\varepsilon,\mu} = -\frac{C_j \ln(C_{\hat{j}} - \hat{j})}{\mu} - \frac{C_\tau (1+|\ln \varepsilon|) \ln(B^\varepsilon)}{\varepsilon}$ the constant C_j is a weight for the reformulated objective $-\ln(C_{\hat{j}} - \hat{j})$ and C_τ is a weight for the barrier part $-\ln(B^\varepsilon)$. Thus, we suspect that for a given ε_k, a larger value of C_τ keeps the iterate u^{k+1} farther away from the boundary $\{u \in U : B^{\varepsilon_k}(u) = 0\}$, which, in turn, may result in $y(u^{k+1})$ being more feasible, respectively, less infeasible with respect to $\min(y - y_a) \geq 0$. We investigate this hypothesis when we examine $\text{LSM}_{(\text{P})}$ for different choices of C_τ.

At first glance it may seem odd that the objective value of the last iterate increases when $\varepsilon_{\text{final}}$ changes from 10^{-4} to 10^{-5}. We attribute this to discretization effects and can, in fact, see this in Table 8.16: From Theorem 7.2.26 we know that $\frac{1}{C_j}|j(u^K) - j(\bar{u})|$ should converge like $\mathcal{O}(\varepsilon_{\text{final}}(1 + |\ln \varepsilon_{\text{final}}|))$, and for $\varepsilon_{\text{final}} \geq 10^{-4}$ we actually observe this. For $\varepsilon_{\text{final}} = 10^{-5}$, however, this does not seem to hold any more, which indicates a discretization effect. Moreover, the following observation directly confirms such an effect: For $\varepsilon_{\text{final}} = 10^{-5}$ the last iterate u^K has objective value $\hat{j}(u^K) < 1.9871$ and satisfies $y(u^K) \geq y_a$. However, since the minimal objective value is $\hat{j}(\bar{u}) = 6.25/\pi > 1.989$, the existence of such a u^K is not possible in U; it can only be explained by discretization effects.

By a comparison of Table 8.16 and Table 8.1 we conclude that Algorithm LSM_ε and $\mathrm{LSM}_{(\mathrm{P})}$ require similar iteration numbers for this test problem. However, the final iterates generated by $\mathrm{LSM}_{(\mathrm{P})}$ are closer to the optimal solution $\bar{u}$ in objective value, both with respect to j and $\hat{j}$. Also, this stays true if we compare to LSM_ε with less strict termination criterion, see Table 8.3. Furthermore, the final iterates of $\mathrm{LSM}_{(\mathrm{P})}$ bear the advantage that they are feasible for the original problem. This may be beneficial in practical applications if solutions are strictly forbidden to contain infeasibility.

$\varepsilon_{\text{final}}$	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)
10^{-2}	13	13	13	13	13	13(6)	5.74×10^{-2}	5.60×10^{-2}	1.63×10^{-2}
10^{-3}	15	15	15	15	16	16	1.34×10^{-2}	1.08×10^{-2}	3.11×10^{-3}
10^{-4}	17	17	17	18	18	18	2.16×10^{-3}	5.08×10^{-4}	1.47×10^{-4}
10^{-5}	17	17	17	18	18	18	2.97×10^{-4}	2.38×10^{-3}	6.87×10^{-4}

Table 8.16. Test problem I: Total number of Newton steps required by $\mathrm{LSM}_{(\mathrm{P})}$ with $C_\tau = 2 \cdot 10^3$; the Newton steps from LSMSUB and phase one are included; displayed in brackets is the number of Newton steps required by phase one; (y^K, u^K) denotes the final iterate

In Table 8.17 we show in detail the course of $\mathrm{LSM}_{(\mathrm{P})}$ for $\varepsilon_{\text{final}} = 10^{-5}$ and $h = 2^{-9}$. We observe that the first iterate is infeasible with respect to $\min(y - y_a) \geq 0$, while all others are feasible.

A crucial ingredient to establish several main results for $\mathrm{SSM}_{(\mathrm{P})}$ and $\mathrm{LSM}_{(\mathrm{P})}$ is the estimate from Corollary 6.4.5. This estimate states that

$$B^{\varepsilon_k}(u^{k+1}) \geq c\mu_k \vartheta(\varepsilon_k) = cC_\tau \varepsilon_k (1 + |\ln \varepsilon_k|)$$

is satisfied for all $k \in \mathbb{N}_0$, where $c > 0$ is independent of k. If the order on the right-hand side of this estimate could be improved, then we would obtain stronger results, e.g., better rates of convergence. Therefore, we check numerically if $B_k := B^{\varepsilon_k}(u^{k+1})/\mu_k \vartheta(\varepsilon_k)$ is approximately constant during the course of $\mathrm{LSM}_{(\mathrm{P})}$. We observe in Table 8.17 that B_k is growing at a very slow rate at first, but seems to converge for smaller values of ε_k. We conclude that the order of $B^{\varepsilon_k}(u^{k+1})$ is well captured by our theoretical estimate.

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	6	1.21×10^{-3}	-2.65×10^{-1}	1.50×10^{-1}	4.26×10^{-2}
10^{-1}	4	1.44×10^{-3}	1.63×10^{-1}	1.72×10^{-1}	5.12×10^{-2}
6.00×10^{-3}	3	1.66×10^{-3}	4.21×10^{-2}	4.01×10^{-2}	1.17×10^{-2}
1.45×10^{-4}	3	1.73×10^{-3}	2.94×10^{-3}	2.69×10^{-4}	7.79×10^{-5}
10^{-5}	2	1.73×10^{-3}	2.97×10^{-4}	2.38×10^{-3}	6.87×10^{-4}

Table 8.17. Test problem I: Course of $\mathrm{LSM}_{(\mathrm{P})}$ with $C_\tau = 2 \cdot 10^3$ and $\varepsilon_{\text{final}} = 10^{-5}$

Figure 8.15 depicts for $\varepsilon_{\text{final}} = 10^{-5}$ and $h = 2^{-9}$ the convergence rates (β_k) of (ε_k) that are achieved during the course of $\mathrm{LSM}_{(\mathrm{P})}$. In addition, this figure displays the Newton decrements

that occur. Theorem 7.2.26 together with Remark 7.2.21 predicts that (ε_k) converges q-linearly to zero. This is, indeed, visible in Figure 8.15 since we have $\beta_k \leq 0.1$ for all k. Moreover, this figure may even indicate q-superlinear convergence. Note that the fast rate of convergence of (ε_k) is achieved with a fairly small amount of Newton steps for each ε_k. We emphasize this since the bound for the number of Newton steps in Theorem 7.2.26 could go to infinity at an arbitrarily fast rate for $\varepsilon_k \to 0^+$, which would render the q-linear convergence of (ε_k) meaningless. For the convergence rates we do not display the rate in the last iteration since this rate is modified such that $\varepsilon_{\text{final}}$ is obtained, as we described above.

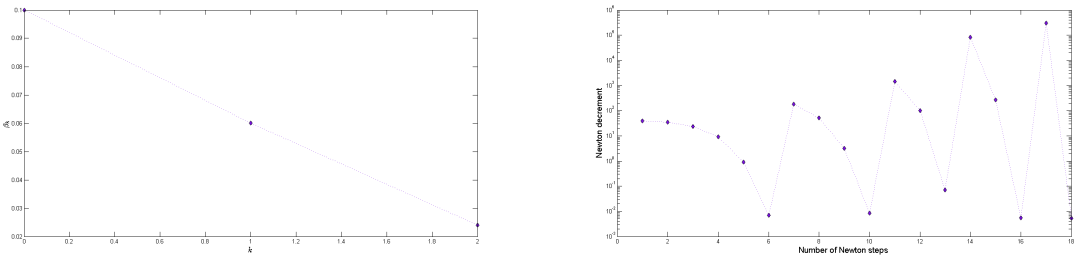

Figure 8.15. Test problem I: Convergence rates and Newton decrements of $\text{LSM}_{(\text{P})}$ with $\varepsilon_{\text{final}} = 10^{-5}$

Figure 8.16 displays the final control u^K and the final state y^K obtained by $\text{LSM}_{(\text{P})}$ for $\varepsilon_{\text{final}} = 10^{-i}$, $i = 2, 3, 4$. We recall that optimal control and optimal state are shown in Figure 8.6 and that this figure also contains $\bar{u}_\varepsilon$ and $\bar{y}_\varepsilon$ for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$. Since the final control of $\text{LSM}_{(\text{P})}$ approximates $\bar{u}_{\varepsilon_{\text{final}},(\varepsilon_{\text{final}})^2}$, which itself can be regarded as an approximation of $\bar{u}_{\varepsilon_{\text{final}}}$, we expect u^K for $\varepsilon_{\text{final}} = 10^{-i}$ to look similar to $\bar{u}_\varepsilon$ for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$. Of course, we have the same expectation for y^K and $\bar{y}_\varepsilon$. A comparison of Figure 8.6 and 8.16 confirms this expectation and, moreover, shows that u^K and y^K replicate quite well the structure of $\bar{u}$ and $\bar{y}$.

So far we used the weight $C_\tau = 2 \cdot 10^3$ for all computations in this section. We now examine the effect of different weights C_τ on $\text{LSM}_{(\text{P})}$. In Table 8.18 we display results for $C_\tau = 2 \cdot 10^2$, $C_\tau = 2 \cdot 10^4$, and $C_\tau = 2 \cdot 10^5$. In Table 8.19 we show in detail the course of $\text{LSM}_{(\text{P})}$ for these weights and $\varepsilon_{\text{final}} = 10^{-5}$, computed with $h = 2^{-9}$. We mention that all these choices for C_τ are covered by the developed theory.

For $C_\tau = 2 \cdot 10^i$ with $i = 3, 4, 5$, we observe that $\text{LSM}_{(\text{P})}$ is clearly mesh independent. For $C_\tau = 2 \cdot 10^2$ the required number of Newton steps varies a little with the mesh. Below we give a reason for this effect.

Table 8.18 shows for $C_\tau \geq 2 \cdot 10^4$ that $\frac{1}{C_j}|j(u^K) - j(\bar{u})|$ seems to have the predicted order $\mathcal{O}(\varepsilon_{\text{final}}(1 + |\ln \varepsilon_{\text{final}}|))$. We can also observe this order of convergence for $C_\tau = 2 \cdot 10^2$ for $\varepsilon_{\text{final}} \geq 10^{-3}$; for $\varepsilon_{\text{final}} \leq 10^{-4}$ the level of the discretization error seems to be reached.

We observe that for a given ε, choosing C_τ larger than $2 \cdot 10^3$ increases feasibility with respect to both the original constraint $\min(y(u) - y_a) \geq 0$, cf. Table 8.16 and 8.18, and the constraint $B^\varepsilon(u) \geq 0$ we replace the original constraint with, cf. the values of B_k in Table 8.17

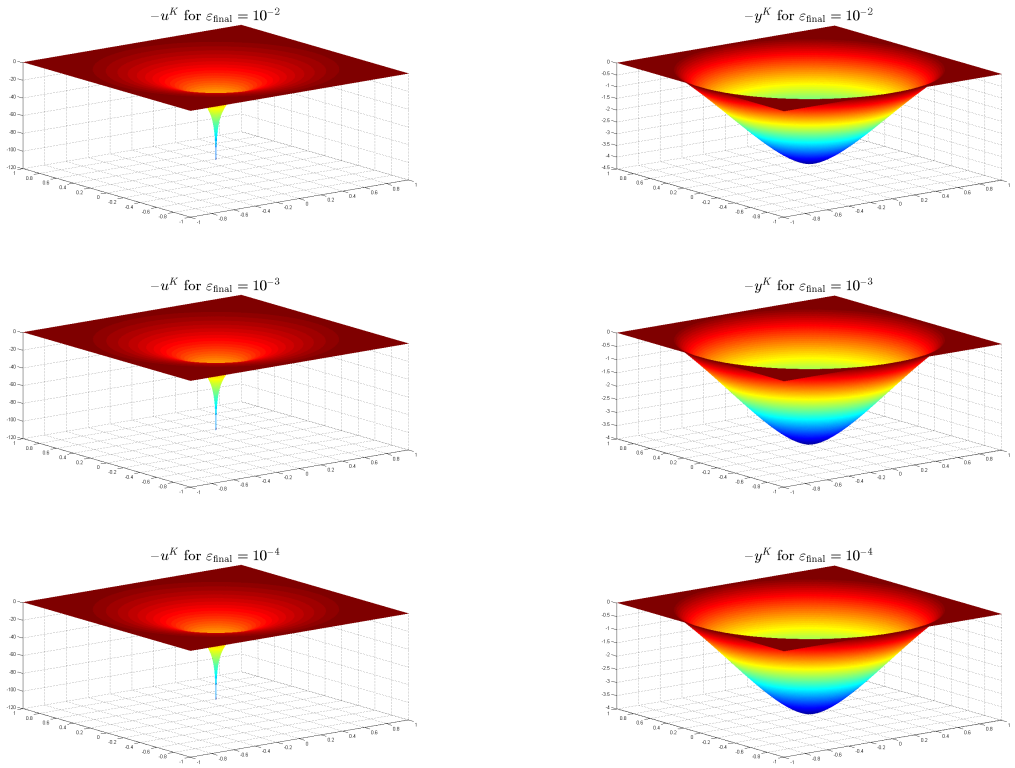

Figure 8.16. Test problem I: Final iterates $-u^K$ and $-y^K$ of $\text{LSM}_{(\text{P})}$ for $\varepsilon_{\text{final}} = 10^{-\{2,3,4\}}$

and 8.19 and recall that $B_k = B^{\varepsilon_k}(u^{k+1})/(C_\tau \varepsilon_k(1 + |\ln \varepsilon_k|))$ depends on C_τ. Similarly, a smaller C_τ decreases feasibility, respectively, increases infeasibility. This observation can be explained by the fact that C_τ is a weight for the barrier part $-\ln(B^\varepsilon)$ in $f_{\varepsilon,\mu}$, since, therefore, increasing C_τ is likely to push the path $\varepsilon \mapsto \bar{u}_{\varepsilon,\varepsilon^2}$ for each ε farther away from the boundary $\{u \in U : B^\varepsilon(u) = 0\}$. This weight property of C_τ also explains why for a given ε the overall errors become larger, in general, when $C_\tau = 2 \cdot 10^3$ is increased, see Table 8.16 and 8.18. It is interesting to note that they are, however, not smaller when C_τ is decreased to $2 \cdot 10^2$. We attribute this to the fact that for $C_\tau = 2 \cdot 10^2$ infeasibility occurs; in fact, if C_τ is further decreased, we expect the objective error to become larger since infeasibility is increased, which means that the path $\varepsilon \mapsto \bar{u}_{\varepsilon,\varepsilon^2}$ is for each ε closer to $\bar{u}_\varepsilon$, cf. also the objective errors of $\bar{u}_\varepsilon$ in Table 8.1.

Figure 8.17 provides the according convergence rates (β_k) for (ε_k) and the development of the Newton decrement, computed with $h = 2^{-9}$. This figure confirms that the convergence rate of (ε_k) is q-linear, even for $C_\tau = 2 \cdot 10^2$, where it is closer to 1 than for the other choices. Moreover, the number of Newton steps required by LSMSUB does not seem to grow as ε_k decreases, cf. Table 8.17 and 8.19.

Since we chose $\beta_{\max} = 0.5$, we have $\tilde{\beta}_0 < 0.5$ in every iteration. Together with the backtracking strategy that we use this implies that $\beta_k > 0.5$ holds if and only if backtracking is actually

$\varepsilon_{\text{final}}$	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)
10^{-2}	70	82	91	89	88	88(6)	-1.04×10^{-1}	1.02×10^{-1}	2.91×10^{-2}
10^{-3}	112	133	164	171	174	184	-1.09×10^{-2}	1.35×10^{-2}	3.89×10^{-3}
10^{-4}	144	177	217	237	251	271	-1.01×10^{-3}	3.68×10^{-3}	1.06×10^{-3}
10^{-5}	165	209	257	284	310	332	-9.29×10^{-5}	2.77×10^{-3}	8.00×10^{-4}

$\varepsilon_{\text{final}}$	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)
10^{-2}	14	14	15	15	15	15(5)	7.56×10^{-1}	9.73×10^{-1}	3.30×10^{-1}
10^{-3}	17	17	17	18	18	18	2.07×10^{-1}	2.21×10^{-1}	6.61×10^{-2}
10^{-4}	19	19	19	20	20	20	3.28×10^{-2}	3.05×10^{-2}	8.86×10^{-3}
10^{-5}	19	19	19	22	22	22	4.18×10^{-3}	1.51×10^{-3}	4.36×10^{-4}

$\varepsilon_{\text{final}}$	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)
10^{-2}	28	28	28	28	28	28(9)	1.67×10^{0}	2.73×10^{0}	1.56×10^{0}
10^{-3}	34	34	34	34	34	34	9.62×10^{-1}	1.31×10^{0}	4.78×10^{-1}
10^{-4}	35	35	35	35	35	36	2.67×10^{-1}	2.92×10^{-1}	8.81×10^{-2}
10^{-5}	37	37	37	37	37	38	4.13×10^{-2}	3.93×10^{-2}	1.14×10^{-2}

Table 8.18. Test problem I: Results of $\text{LSM}_{(\text{P})}$ with $C_\tau = 2 \cdot 10^2$ (top), $C_\tau = 2 \cdot 10^4$ (middle), and $C_\tau = 2 \cdot 10^5$ (bottom)

necessary, i.e., if and only if $u^{k+1} \notin U_{\text{ad}}(\tilde{\beta}_0 \varepsilon_k)$. We observe in Figure 8.17 that for $C_\tau = 2 \cdot 10^i$ with $i = 3, 4, 5$, no backtracking is necessary. This is, the value for $\tilde{\beta}_0$ that is picked by our adaptive strategy is accepted throughout the course of $\text{LSM}_{(\text{P})}$. For $C_\tau = 2 \cdot 10^2$ the situation changes; backtracking is necessary. Of course, this slows down the rate of convergence, which is also obvious in Figure 8.17 and illustrated by the larger numbers of Newton steps required for $C_\tau = 2 \cdot 10^2$. The necessity of backtracking for smaller values of C_τ can be explained as follows: The only situation that requires backtracking is if $B^{\tilde{\beta}_0 \varepsilon_k}(u^{k+1}) \leq 0$ holds. We can expect that it is possible to decrease ε_k the stronger, the larger $B^{\varepsilon_k}(u^{k+1}) > 0$ is. We already argued and observed in practice that, generally, $B^{\varepsilon_k}(u^{k+1})$ is the larger, the larger C_τ is. If u^{k+1} satisfies $\min(y(u^{k+1}) - y_a) > 0$, an additional argument is provided by the inequality from Corollary 4.1.3, which implies $B^{\varepsilon}(u^{k+1}) > 0$ for all $\varepsilon > 0$ and, therefore, backtracking cannot occur (at least if we ignore discretization effects). This explains why we see backtracking only for smaller values of C_τ. Moreover, this leads us to the believe that we will see even more backtracking, respectively, even slower convergence rates if we choose C_τ smaller than $2 \cdot 10^2$. To examine this claim and to prove that despite an even smaller value for C_τ the convergence of (ε_k) stays q-linear, we apply $\text{LSM}_{(\text{P})}$ with $C_\tau = 7 \cdot 10^1$ and $C_\tau = 2 \cdot 10^1$. The resulting convergence rates are shown in Figure 8.18. They display q-linear convergence and, in addition, encourage the view that the smaller distance to the boundary of the feasible set requires more backtracking, thereby slowing down the speed of convergence.

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	6	9.16×10^{-3}	-1.13×10^{0}	6.40×10^{-1}	1.70×10^{-1}
5.50×10^{-1}	2	7.65×10^{-3}	-1.02×10^{0}	6.22×10^{-1}	1.65×10^{-1}
2.81×10^{-1}	3	5.85×10^{-3}	-8.82×10^{-1}	5.85×10^{-1}	1.56×10^{-1}
1.41×10^{-1}	4	4.08×10^{-3}	-6.83×10^{-1}	5.06×10^{-1}	1.36×10^{-1}
1.06×10^{-1}	4	3.53×10^{-3}	-5.89×10^{-1}	4.58×10^{-1}	1.24×10^{-1}
7.97×10^{-2}	4	3.10×10^{-3}	-4.97×10^{-1}	4.05×10^{-1}	1.11×10^{-1}
5.98×10^{-2}	5	2.78×10^{-3}	-4.12×10^{-1}	3.50×10^{-1}	9.63×10^{-2}
4.49×10^{-2}	4	2.53×10^{-3}	-3.38×10^{-1}	2.96×10^{-1}	8.23×10^{-2}
3.37×10^{-2}	8	2.34×10^{-3}	-2.74×10^{-1}	2.47×10^{-1}	6.91×10^{-2}
2.95×10^{-2}	4	2.27×10^{-3}	-2.48×10^{-1}	2.27×10^{-1}	6.35×10^{-2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
4.25×10^{-5}	6	1.73×10^{-3}	-4.16×10^{-4}	3.09×10^{-3}	8.93×10^{-4}
3.38×10^{-5}	6	1.73×10^{-3}	-3.28×10^{-4}	3.00×10^{-3}	8.68×10^{-4}
2.69×10^{-5}	6	1.73×10^{-3}	-2.59×10^{-4}	2.93×10^{-3}	8.48×10^{-4}
2.14×10^{-5}	6	1.73×10^{-3}	-2.04×10^{-4}	2.88×10^{-3}	8.31×10^{-4}
1.70×10^{-5}	6	1.73×10^{-3}	-1.61×10^{-4}	2.83×10^{-3}	8.19×10^{-4}
1.35×10^{-5}	6	1.73×10^{-3}	-1.27×10^{-4}	2.80×10^{-3}	8.09×10^{-4}
1.08×10^{-5}	6	1.73×10^{-3}	-1.00×10^{-4}	2.77×10^{-3}	8.02×10^{-4}
10^{-5}	4	1.73×10^{-3}	-9.29×10^{-5}	2.77×10^{-3}	8.00×10^{-4}

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	1.63×10^{-4}	1.02×10^{0}	2.10×10^{0}	9.34×10^{-1}
10^{-1}	5	3.25×10^{-4}	1.37×10^{0}	2.10×10^{0}	9.34×10^{-1}
8.00×10^{-3}	5	8.48×10^{-4}	6.88×10^{-1}	8.67×10^{-1}	2.89×10^{-1}
5.12×10^{-4}	3	1.52×10^{-3}	1.26×10^{-1}	1.30×10^{-1}	3.83×10^{-2}
1.32×10^{-5}	2	1.72×10^{-3}	5.36×10^{-3}	2.70×10^{-3}	7.81×10^{-4}
10^{-5}	2	1.72×10^{-3}	4.18×10^{-3}	1.51×10^{-3}	4.36×10^{-4}

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	9	1.75×10^{-5}	1.43×10^{0}	3.27×10^{0}	2.91×10^{0}
10^{-1}	9	4.01×10^{-5}	1.88×10^{0}	3.26×10^{0}	2.84×10^{0}
1.50×10^{-2}	6	1.22×10^{-4}	1.73×10^{0}	2.89×10^{0}	1.80×10^{0}
2.25×10^{-3}	6	4.04×10^{-4}	1.26×10^{0}	1.87×10^{0}	7.76×10^{-1}
3.38×10^{-4}	4	9.56×10^{-4}	5.76×10^{-1}	7.00×10^{-1}	2.26×10^{-1}
3.04×10^{-5}	2	1.55×10^{-3}	1.07×10^{-1}	1.09×10^{-1}	3.19×10^{-2}
10^{-5}	2	1.65×10^{-3}	4.13×10^{-2}	3.93×10^{-2}	1.14×10^{-2}

Table 8.19. Test problem I: Course of $\mathrm{LSM}_{(\mathrm{P})}$ with $\varepsilon_{\mathrm{final}} = 10^{-5}$ for $C_\tau = 2 \cdot 10^2$ (top), $C_\tau = 2 \cdot 10^4$ (middle), and $C_\tau = 2 \cdot 10^5$ (bottom)

The values of B_k in Table 8.19 for smaller ε_k lead to the conjecture that $B^{\varepsilon_k}(u^{k+1})$ may have exactly the order $\varepsilon_k(1 + |\ln \varepsilon_k|)$ regardless of C_τ, with (B_k) converging to approximately $1.73 \cdot 10^{-3}$. We further investigated this hypothesis by use of $\varepsilon_{\text{final}} = 10^{-9}$ and found it to be true. This underlines the accuracy of the estimate in Corollary 6.4.5. Moreover, it indicates that the constant c in this corollary may be independent of C_τ, which is neither part of the assertion in Corollary 6.4.5, nor can it be inferred from its proof.

We noted above that for $C_\tau = 2 \cdot 10^2$ the overall number of Newton steps varies stronger with the mesh size. We attribute this to the fact that backtracking is required, since feasibility, respectively, infeasibility with respect to the nonlinear term B^ε can be expected to vary with changing mesh size. We encountered a similar dependence during the experiments for Test Problem I for fixed ε in Section 8.2.1, where we observed in a nesting strategy that a function may be feasible with respect to B^ε on a certain grid, but its interpolant on a finer grid may be infeasible, nevertheless.

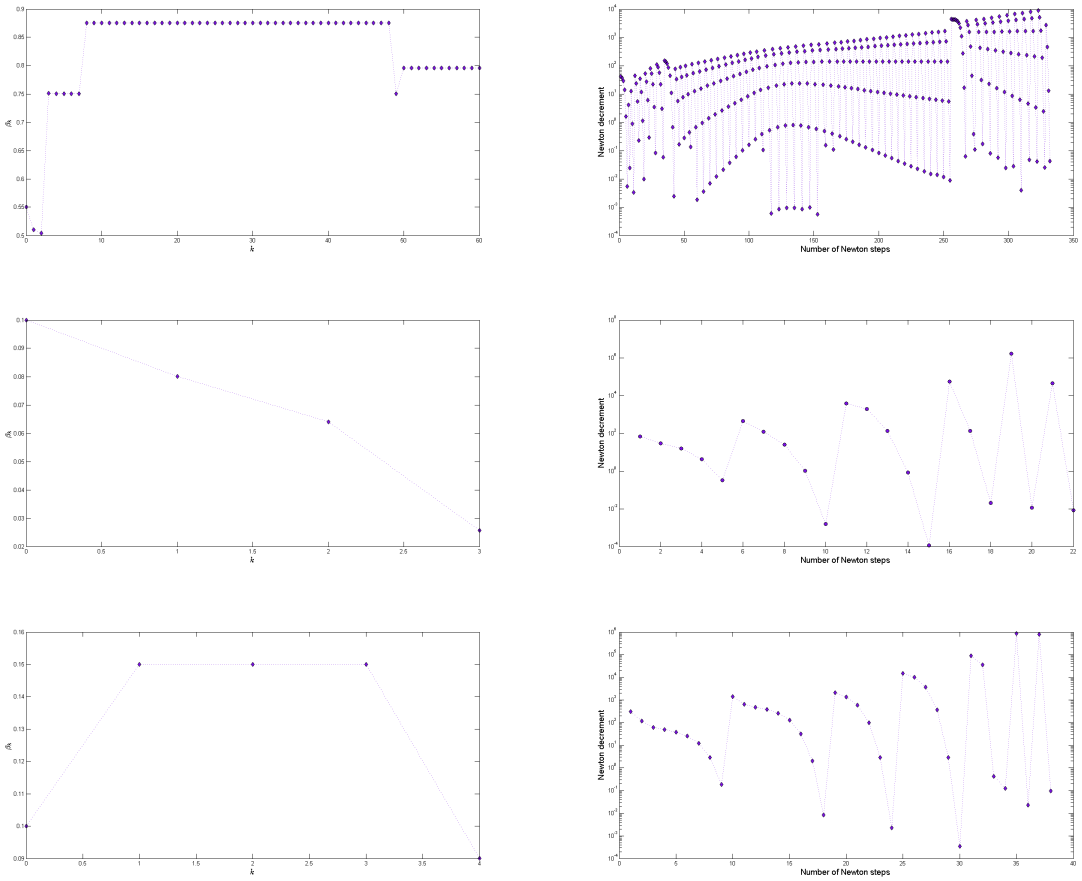

Figure 8.17. Test problem I: Convergence rates and Newton decrements of $\text{LSM}_{(\text{P})}$ with $\varepsilon_{\text{final}} = 10^{-5}$ for $C_\tau = 2 \cdot 10^2$ (top), $C_\tau = 2 \cdot 10^4$ (middle), and $C_\tau = 2 \cdot 10^5$ (bottom)

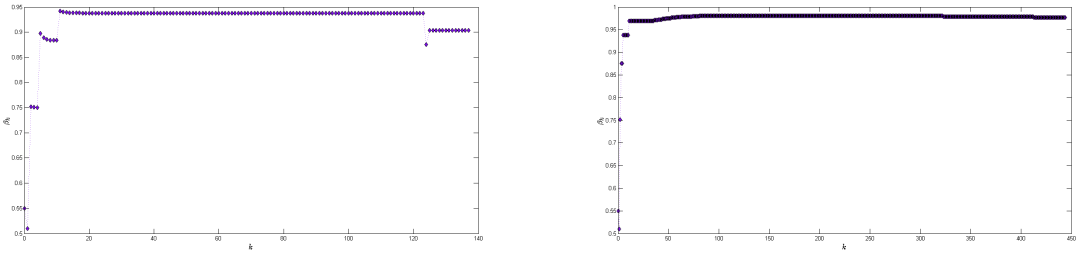

Figure 8.18. Test problem I: Convergence rates of $\mathrm{LSM}_{(\mathrm{P})}$ with $\varepsilon_{\mathrm{final}} = 10^{-5}$ for $C_\tau = 7 \cdot 10^1$ (left) and $C_\tau = 2 \cdot 10^1$ (right)

We now conduct experiments in which $\varepsilon_{\mathrm{final}}$ is determined during the course of the algorithm. We want to terminate $\mathrm{LSM}_{(\mathrm{P})}$ in iteration k if $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\} \leq \mathrm{TOL}$ with $\mathrm{TOL} = 10^{-2}$ is satisfied. To explain the shift in the index we recall that (y^{k+1}, u^{k+1}) is determined in iteration k and satisfies $u^{k+1} \in \Lambda_{\varepsilon_k, \mu_k}(\theta)$. This termination criterion ensures a certain accuracy for the objective while also requiring that the infeasibility is not too large. Of course, many different termination criteria are conceivable. In particular, the error tolerance TOL could be based on an (a posteriori) estimate of the discretization error with respect to the quantities of interest, e.g., $\hat{j}$ and the infeasibility. It would be a desirable feature to incorporate estimates for the discretization error into the implementation of $\mathrm{LSM}_{(\mathrm{P})}$, but this is beyond the scope of this work. We mention that, nevertheless, the choice $\mathrm{TOL} = 10^{-2}$ takes into account that we expect the discretization error for $\hat{j}$ to be smaller than 10^{-2} on meshes with $h \leq 2^{-7}$. This estimate stems from Section 8.2.1.

To check if $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\} \leq 10^{-2}$ is satisfied, we have to estimate $\hat{j}(\bar{u})$ since, in general, we do not know this quantity. To this end, we interpolate linearly between $(\varepsilon_{k-2}, \hat{j}(u^{k-1}))$ and $(\varepsilon_{k-1}, \hat{j}(u^k))$ and also between $(\varepsilon_{k-1}, \hat{j}(u^k))$ and $(\varepsilon_k, \hat{j}(u^{k+1}))$. Evaluating the two interpolants at $\varepsilon = 0$ yields two estimates for $\hat{j}(\bar{u})$. We take the mean value of these estimates as final estimate for $\hat{j}(\bar{u})$. We note that linear interpolation is sensible since $(|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|)_k$ can be expected to behave like $(\frac{|j(u^{k+1}) - j(\bar{u})|}{C_j})_k$ as we argue, e.g., in Remark 5.2.6, and since we have $\frac{|j(u^{k+1}) - j(\bar{u})|}{C_j} = \mathcal{O}(\varepsilon_k(1 + |\ln \varepsilon_k|)) \approx \mathcal{O}(\varepsilon_k)$ by Theorem 7.2.26. We increase $\beta_{\min}$ to $\beta_{\min} = 10^{-2}$ since we believe that very small values of β_k to update ε_k to ε_{k+1} may affect the accuracy of the interpolation strategy negatively.

In Table 8.20 we display results obtained by $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$. Table 8.21 shows partly the corresponding detailed development of $\mathrm{LSM}_{(\mathrm{P})}$, computed with $h = 2^{-9}$; in this table we focus on the last four iterations to demonstrate the accuracy of the termination criterion. Also, the first iterations with automatic determination of $\varepsilon_{\mathrm{final}}$ are very similar to the first iterations with prescribed $\varepsilon_{\mathrm{final}}$. Therefore, the beginning of a detailed development of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$ looks similar to Table 8.17 and 8.19, respectively.

We observe in Table 8.20 that the initially proposed value $C_\tau = 2 \cdot 10^3$ requires the lowest number of Newton steps. Furthermore, Table 8.20 shows that $\varepsilon_{\text{final}}$ is the smaller, the larger C_τ is. This goes together well with our statements from before that C_τ is a weight for the barrier part of $f_{\varepsilon,\mu}$, since weighing the barrier more means weighing the objective less.

In Table 8.21 we see that our termination criterion works relatively well: $\text{LSM}_{(\text{P})}$ is terminated at most one iteration after or one iteration before $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\} \leq 10^{-2}$ is satisfied for the first time. We observe for $C_\tau = 2 \cdot 10^2$ that $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\}$ is slightly larger than 10^{-2}. To avoid this we could incorporate a safeguard into the termination criterion. However, we also see that the interpolation strategy works well for $C_\tau = 2 \cdot 10^2$ since $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\}$ is very close to 10^{-2}. Therefore, we leave the termination criterion unchanged.

Table 8.20 shows that the obtained accuracy is similar for all choices of C_τ and that the iteration numbers are rather mesh independent. Thus, we propose the following strategy to determine a suitable value for C_τ: First, we use theoretical results to estimate C_τ, as shown in detail at the beginning of this section. Then we apply $\text{LSM}_{(\text{P})}$ on a coarse mesh for different values of C_τ that are close to the estimated value. As actual value for C_τ we take the one for which $\text{LSM}_{(\text{P})}$ requires the lowest number of Newton steps. We note, however, that for this test problem the use of C_τ as estimated at the beginning of this section seems to be a good choice, anyway.

C_τ	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	$\varepsilon_{\text{final}}$
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
2×10^2	104	124	163	177	179	193(6)	-8.13×10^{-3}	1.08×10^{-2}	3.11×10^{-3}	7.56×10^{-4}
2×10^3	15	15	15	16	16	16(6)	3.87×10^{-3}	1.20×10^{-3}	3.48×10^{-4}	2.02×10^{-4}
2×10^4	24	22	22	24	25	23(5)	1.48×10^{-4}	2.53×10^{-3}	7.31×10^{-4}	2.68×10^{-7}
2×10^5	43	45	45	45	46	44(9)	7.57×10^{-5}	2.60×10^{-3}	7.51×10^{-4}	1.14×10^{-8}

Table 8.20. Test problem I: Total number of Newton steps required by $\text{LSM}_{(\text{P})}$ with different values of C_τ and automatic determination of $\varepsilon_{\text{final}}$

To increase the practical efficiency of $\text{LSM}_{(\text{P})}$ we add a nested grid strategy. As in the experiments for fixed ε for Test Problem I we use a hierarchy of six grids ranging from $h = 2^{-4}$ to $h = 2^{-9}$. Since we observed during our experiments that the automatic determination of $\varepsilon_{\text{final}}$ yields on coarse grids roughly the same value for $\varepsilon_{\text{final}}$ as on finer grids, we start $\text{LSM}_{(\text{P})}$ on the coarsest grid and use the $\varepsilon_{\text{final}}$ from the coarsest grids for all finer grids, i.e., we only carry out $\text{LSM}_{(\text{P})}$ on all finer grids with $\varepsilon = \varepsilon_{\text{final}}$. More sophisticated techniques to determine $\varepsilon_{\text{final}}$ may be a topic of future research. However, we will see that this strategy works relatively well. When prolongating the final iterate of $\text{LSM}_{(\text{P})}$ onto a finer grid, a shift to restore feasibility may be required. We discussed this in detail in Section 8.2.1, where we also presented a strategy to determine the size of the shift. We use exactly the same strategy here.

The results of $\text{LSM}_{(\text{P})}$ with nested grid strategy are shown for different choices of C_τ in Table 8.22, where an additional (s) indicates that a shift is necessary for the final iterate on the current mesh. We see that $C_\tau = 2 \cdot 10^3$ performs well in comparison to the other choices of

ε_k	#steps	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	6	-1.13×10^{0}	6.40×10^{-1}	1.70×10^{-1}
⋮	⋮	⋮	⋮	⋮
1.12×10^{-3}	5	-1.22×10^{-2}	1.48×10^{-2}	4.28×10^{-3}
9.84×10^{-4}	5	-1.07×10^{-2}	1.33×10^{-2}	3.84×10^{-3}
8.62×10^{-4}	5	-9.32×10^{-3}	1.20×10^{-2}	3.45×10^{-3}
7.56×10^{-4}	5	-8.13×10^{-3}	1.08×10^{-2}	3.11×10^{-3}

ε_k	#steps	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	6	-2.65×10^{-1}	1.50×10^{-1}	4.26×10^{-2}
10^{-1}	4	1.63×10^{-1}	1.73×10^{-1}	5.12×10^{-2}
6.40×10^{-3}	3	4.38×10^{-2}	4.18×10^{-2}	1.22×10^{-2}
2.02×10^{-4}	3	3.87×10^{-3}	1.20×10^{-3}	3.48×10^{-4}

ε_k	#steps	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	1.03×10^{0}	2.10×10^{0}	9.34×10^{-1}
10^{-1}	5	1.37×10^{0}	2.10×10^{0}	9.34×10^{-1}
8.20×10^{-3}	5	6.95×10^{-1}	8.78×10^{-1}	2.93×10^{-1}
5.54×10^{-4}	3	1.34×10^{-1}	1.38×10^{-1}	4.09×10^{-2}
1.83×10^{-5}	2	7.24×10^{-3}	4.59×10^{-3}	1.33×10^{-3}
2.68×10^{-7}	3	1.48×10^{-4}	2.53×10^{-3}	7.31×10^{-4}

ε_k	#steps	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	9	1.43×10^{0}	3.27×10^{0}	2.92×10^{0}
⋮	⋮	⋮	⋮	⋮
3.38×10^{-4}	4	5.76×10^{-1}	7.00×10^{-1}	2.26×10^{-1}
3.17×10^{-5}	2	1.11×10^{-1}	1.13×10^{-1}	3.31×10^{-2}
8.50×10^{-7}	2	4.38×10^{-3}	1.71×10^{-3}	4.94×10^{-4}
1.14×10^{-8}	6	7.57×10^{-5}	2.60×10^{-3}	7.51×10^{-4}

Table 8.21. Test problem I: Course of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\text{final}}$ for $C_\tau = 2 \cdot 10^2$ (top), $C_\tau = 2 \cdot 10^3$ (below top), $C_\tau = 2 \cdot 10^4$ (above bottom), and $C_\tau = 2 \cdot 10^5$ (bottom)

C_τ that we tested. Furthermore, we observe that nesting does not work well for larger choices of C_τ in this test problem. We attribute this to the fact that a (rather large) shift is required even on finer meshes. Since our initial choice $C_\tau = 2 \cdot 10^3$ yields acceptable results, we do not investigate further how to increase the performance of $\text{LSM}_{(\text{P})}$ with nesting for larger weights. Moreover, for $C_\tau = 2 \cdot 10^2$ we see that the termination criterion is slightly violated on the finest mesh. In fact, even the infeasibility is slightly too large, which can be attributed to the fact that $\varepsilon_{\text{final}}$ is determined on the coarsest mesh. However, the violation of the termination criterion is fairly small and could certainly be addressed by incorporation of a safeguard for $\varepsilon_{\text{final}}$.

C_τ	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_{\hat{j}}}\lvert j(u^K) - j(\bar{u})\rvert$	$\varepsilon_{\text{final}}$
	4	5	6	7	8	9	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
2×10^2	104(s)	10(s)	5(s)	4	6	4	-1.69×10^{-2}	1.95×10^{-2}	5.62×10^{-3}	1.54×10^{-3}
2×10^3	15(s)	5(s)	4(s)	6	6	5	3.87×10^{-3}	1.20×10^{-3}	3.48×10^{-4}	2.02×10^{-4}
2×10^4	24(s)	9(s)	10(s)	13(s)	13(s)	14	8.30×10^{-5}	2.59×10^{-3}	7.49×10^{-4}	1.47×10^{-7}
2×10^5	43(s)	12(s)	13(s)	13(s)	16(s)	18	4.68×10^{-5}	2.63×10^{-3}	7.59×10^{-4}	6.84×10^{-9}

Table 8.22. Test problem I: Results of $\text{LSM}_{(\text{P})}$ with automatic determination of $\varepsilon_{\text{final}}$ and a nested grid strategy for different values of C_τ

8.3.2. Test Problem II

The second problem under consideration is identical to Test Problem II for fixed ε, see Section 8.2.2. We choose $\varepsilon_s = 1$, $C_j = 8 \cdot 10^3$, $\tau(\varepsilon) = C_\tau \frac{1+|\ln \varepsilon|}{\varepsilon}$ with $C_\tau = 2 \cdot 10^3$ for all $\varepsilon \in (0, \varepsilon_s]$, and $C_{\hat{j}} = 1 + \hat{j}(u^0)$ with $u^0 \equiv 0$. We have to show that these parameters satisfy Assumption 6.1.1. We reasoned in Section 8.2.2 that $C_j = \frac{8 \cdot 10^3}{\varepsilon^2}$ ensures self-concordance of $f_{\varepsilon,\mu}$ for all $\mu \in (0, 1]$ and all $\varepsilon \in (0, 1]$ since

$$\frac{8 \cdot 10^3}{\varepsilon^2} \geq \frac{1}{\varepsilon^2} \max \left\{ \varepsilon^2, \frac{16}{9} C_{\partial, C(\overline{\Omega}_a)}^2 \left(\frac{\|2\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U \right)^2 \right\}$$

is satisfied for $\tilde{u} \equiv 15$. Enlarging $C_{\partial, C(\overline{\Omega}_a)}$ if necessary this implies

$$8 \cdot 10^3 = \max \left\{ \varepsilon_s^2, \frac{16}{9} C_{\partial, C(\overline{\Omega}_a)}^2 \left(\frac{\|2\hat{j}'(\tilde{u})\|_{U^*}}{\hat{\alpha}} + \|\tilde{u}\|_U \right)^2 \right\}.$$

This shows that Assumption 6.1.1 is fulfilled for this test problem. Moreover, Assumption 6.1.9 is valid as follows by use of $\hat{y}(x_1, x_2) := x_1 x_2 (1 - x_1)(1 - x_2) \in Y = H^2(\Omega) \cap H_0^1(\Omega)$ and $\hat{u} := -\Delta \hat{y} + \hat{y}$. Also, we argued in Section 8.2.2 that this test problem satisfies Assumption 3.1.9. Together, we have established all assumptions that are required to apply the theory developed in this thesis.

At the beginning of Section 8.3.1 we argued that the choice $C_\tau \approx C_j \|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ is sensible. Applying the same reasoning as in that section we, furthermore, obtain for $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*}$ the

estimate $\|\bar{\lambda}\|_{C(\overline{\Omega}_a)^*} \leq 2\sqrt{2\hat{j}(\bar{u})} \leq 2\sqrt{0.25} = 1$, where we used the estimate $\hat{j}(\bar{u}) \leq 0.125$. We include as safeguard a factor of $\frac{1}{4}$, which yields $C_\tau = \frac{C_j}{4} = 2 \cdot 10^3$. We later also investigate how $\text{LSM}_{(\text{P})}$ performs for other choices of C_τ.

We choose $\theta = 0.25$, $\varepsilon_0 = 1$, $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 0.5$, and $\tilde{\beta}_0 = 0.1$ for $k = 0$ in the backtracking strategy. Furthermore, we apply the same strategy as for Test Problem I to adaptively choose $\tilde{\beta}_0$ in iteration $k \in \mathbb{N}$.

For the unknown values $\hat{j}(\bar{u})$ and $j(\bar{u})$ we employ $\hat{j}(\bar{u}_\varepsilon)$ and $j(\bar{u}_\varepsilon)$ for $\varepsilon = 10^{-7}$ computed with LSM_ε for $h = 2^{-10}$ in Section 8.2.2.

In the first experiment we prescribe $\varepsilon_{\text{final}}$ and terminate $\text{LSM}_{(\text{P})}$ when $\varepsilon_k = \varepsilon_{\text{final}}$ is satisfied. We apply $\text{LSM}_{(\text{P})}$ for several values of $\varepsilon_{\text{final}}$ on uniform meshes with different widths h; we choose $\varepsilon_{\text{final}} \geq 10^{-5}$ and $h \geq 2^{-10}$. The total numbers of Newton steps that have to be computed during the course of $\text{LSM}_{(\text{P})}$ are listed in Table 8.23 and clearly indicate that $\text{LSM}_{(\text{P})}$ is mesh independent. By a comparison of Table 8.23 and Table 8.7 we conclude that Algorithm LSM_ε and $\text{LSM}_{(\text{P})}$ require similar iteration numbers for this test problem but that the final iterates generated by LSM_ε are closer to the optimal solution $\bar{u}$ in objective value, both with respect to j and $\hat{j}$. Moreover, the final iterates generated by $\text{LSM}_{(\text{P})}$ are feasible with respect to the constraint $\min(y - y_a) \geq 0$. From Theorem 7.2.26 we know that $\frac{1}{C_j}|j(u^K) - j(\bar{u})|$ has the order $\mathcal{O}(\varepsilon_{\text{final}}(1 + |\ln \varepsilon_{\text{final}}|))$. Table 8.23 confirms this. Moreover, for $\varepsilon_{\text{final}} \leq 10^{-3}$ it seems that this order cannot be improved.

$\varepsilon_{\text{final}}$	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)
10^{-2}	15	15	15	15	15	15(7)	1.00×10^{-2}	5.20×10^{-3}	5.20×10^{-3}
10^{-3}	20	20	20	20	20	20	7.31×10^{-3}	1.22×10^{-3}	1.21×10^{-3}
10^{-4}	26	26	26	26	26	26	1.47×10^{-3}	2.14×10^{-4}	2.14×10^{-4}
10^{-5}	32	32	32	32	32	32	1.86×10^{-4}	2.72×10^{-5}	2.72×10^{-5}

Table 8.23. Test problem II: Results of $\text{LSM}_{(\text{P})}$ with $C_\tau = 2 \cdot 10^3$

In Table 8.24 we show in detail the course of $\text{LSM}_{(\text{P})}$ for $\varepsilon_{\text{final}} = 10^{-5}$ on the finest mesh that we employed. As for Test Problem I in Section 8.3.1 we check numerically if the quantity $B_k := B^{\varepsilon_k}(u^{k+1})/\mu_k\vartheta(\varepsilon_k)$ is approximately constant during the course of $\text{LSM}_{(\text{P})}$. We observe that B_k grows at a slow rate, which indicates that Corollary 6.4.5 estimates B_k well.

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	7	7.43×10^{-5}	1.00×10^{-2}	1.17×10^{-1}	1.24×10^{-1}
10^{-1}	3	1.06×10^{-4}	1.00×10^{-2}	3.75×10^{-2}	3.81×10^{-2}
4.00×10^{-3}	5	2.65×10^{-4}	1.00×10^{-2}	2.46×10^{-3}	2.46×10^{-3}
1.28×10^{-4}	8	8.54×10^{-4}	1.82×10^{-3}	2.66×10^{-4}	2.65×10^{-4}
10^{-5}	9	9.10×10^{-4}	1.86×10^{-4}	2.72×10^{-5}	2.72×10^{-5}

Table 8.24. Test problem II: Course of $\text{LSM}_{(\text{P})}$ with $C_\tau = 2 \cdot 10^3$ and $\varepsilon_{\text{final}} = 10^{-5}$

Figure 8.19 displays the final controls u^K and the final states y^K obtained by $\mathrm{LSM}_{(\mathrm{P})}$ for $\varepsilon_{\mathrm{final}} = 10^{-i}$, $i = 2, 3, 4$. We recall that optimal control and optimal state are shown in Figure 8.10. This figure also contains $\bar{u}_\varepsilon$ and $\bar{y}_\varepsilon$ for $\varepsilon = 10^{-i}$, $i = 2, 3, 4$. For $\varepsilon_{\mathrm{final}} = 10^{-i}$, $i = 3, 4$, it is clearly visible that u^K and y^K are somewhat similar to $\bar{u}_\varepsilon$ and $\bar{y}_\varepsilon$ and that they replicate well the structure of $\bar{u}$ and $\bar{y}$.

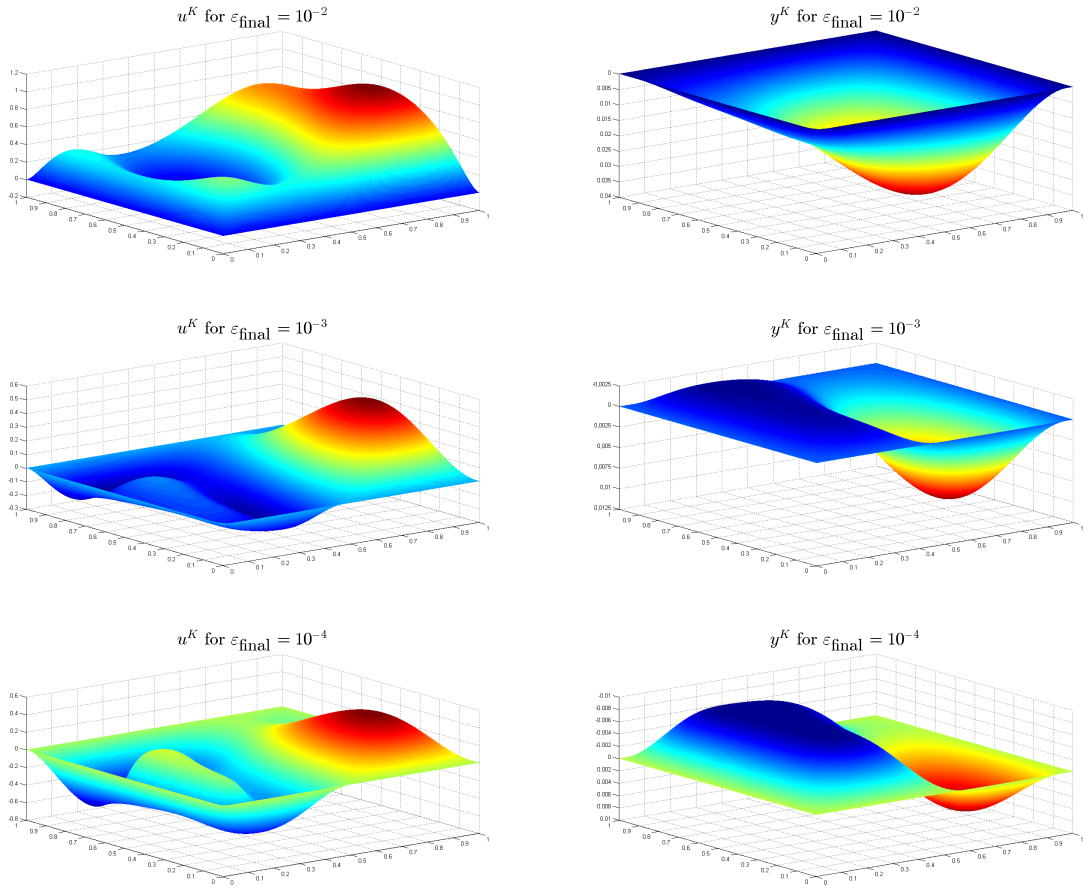

Figure 8.19. Test problem II: Final iterates u^K and y^K (with inverted z-axis) of $\mathrm{LSM}_{(\mathrm{P})}$ for $\varepsilon_{\mathrm{final}} = 10^{-\{2,3,4\}}$

We now incorporate into $\mathrm{LSM}_{(\mathrm{P})}$ the strategy proposed in Section 8.3.1 to determine $\varepsilon_{\mathrm{final}}$ automatically. In the termination criterion we use $\mathrm{TOL} = 10^{-4}$. In Table 8.25 we display the results for different choices of C_τ, while Table 8.26 shows the course of the algorithm in more detail for these choices of C_τ, computed with $h = 2^{-10}$. Figure 8.20 provides the according convergence rates (β_k) for (ε_k) and the development of the Newton decrement.

We see in Table 8.25 that the proposed value $C_\tau = 2 \cdot 10^3$ requires the lowest number of Newton steps. Furthermore, we can observe similar connections as in Test Problem I between the quantities displayed in Table 8.26 and the weight C_τ. For instance, the larger C_τ becomes, the larger is the feasibility of the iterates for a given ε. Moreover, Table 8.26 shows that our termination criterion works relatively well: $\mathrm{LSM}_{(\mathrm{P})}$ terminates either when

$\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\} \leq 10^{-4}$ is satisfied for the first time or one iteration before. If $\mathrm{LSM}_{(\mathrm{P})}$ terminates before this criterion is fulfilled, which is only the case for $C_\tau = 2 \cdot 10^4$, then $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\}$ is only slightly larger than 10^{-4}. Furthermore, Table 8.26 leads to the conjecture that (B_k) converges to a value somewhere around $9.13 \cdot 10^{-4}$. We further investigated this hypothesis by use of $\varepsilon_{\mathrm{final}} = 10^{-10}$ and found it to be true, indeed; (B_k) converges to approximately $9.16 \cdot 10^{-4}$ regardless of the choice of C_τ. This underlines the accuracy of the estimate in Corollary 6.4.5 and, again, seems to indicate that the constant c in this corollary is independent of C_τ. We mention that if this were true, then we could prove by use of Corollary 6.4.5 and an argument as in the proof of Corollary 4.4.4 that for C_τ sufficiently large, the neighborhoods $\Lambda_{\varepsilon,\mu}$ are feasible with respect to $\min(y - y_a) \geq 0$ for all $(\varepsilon, \mu) \in \mathcal{P}_=$. In particular, this would imply that all iterates are feasible if C_τ is chosen large enough.

Figure 8.20 confirms that the convergence rate of (ε_k) is q-linear, even for $C_\tau = 2 \cdot 10^2$, where backtracking is required for $k \geq 2$. In Figure 8.21 we display convergence rates for smaller choices than $C_\tau = 2 \cdot 10^2$. We observe q-linear convergence also for these smaller values of C_τ, with convergence rates closer to 1. This fits nicely with what we have said in Section 8.3.1 about the size of C_τ and its influence on backtracking and the convergence rate. We point out that for $C_\tau = 10^i$, $i = 3, 4$, the number of Newton steps required by LSMSUB seems to increase as ε_k decreases, cf. Table 8.26.

C_τ	Mesh size $h = 2^{-i},\ i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	$\varepsilon_{\mathrm{final}}$
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
2×10^2	40	46	46	46	46	46(5)	-6.02×10^{-5}	3.77×10^{-6}	3.76×10^{-6}	4.17×10^{-5}
2×10^3	36	34	34	34	34	34(7)	1.33×10^{-4}	1.95×10^{-5}	1.94×10^{-5}	6.93×10^{-6}
2×10^4	61	60	47	47	47	46(7)	9.18×10^{-4}	1.28×10^{-4}	1.28×10^{-4}	3.91×10^{-6}

Table 8.25. Test problem II: Results of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$

In the last experiment for Test Problem II we add to $\mathrm{LSM}_{(\mathrm{P})}$ a nested grid strategy. We employ a hierarchy of six grids ranging from $h = 2^{-5}$ to $h = 2^{-10}$ and use automatic determination of $\varepsilon_{\mathrm{final}}$ on the coarsest grid and $\varepsilon = \varepsilon_{\mathrm{final}}$ on all finer grids. The shift to restore feasibility as part of the prolongation onto a finer grid can be computed in exactly the same manner as when we applied LSM_ε with nesting to Test Problem II, cf. Section 8.2.2. However, it turns out that a shift is never required. The results of $\mathrm{LSM}_{(\mathrm{P})}$ with nested grid strategy are shown for different choices of C_τ in Table 8.27. Apparently, nesting increases the efficiency of $\mathrm{LSM}_{(\mathrm{P})}$ significantly for this test problem. We note that the determination of $\varepsilon_{\mathrm{final}}$ on the coarsest grid works well, although for $C_\tau = 2 \cdot 10^2$ the termination criterion is not satisfied on the finest mesh since $\|(y^K - y_a)^-\|_{C(\overline{\Omega}_a)}$ is slightly larger than 10^{-4}; however, the violation is small.

8.3.3. Test Problem III

The last problem that we consider is identical to Test Problem III for fixed ε, see Section 8.2.3. We recall that the state equation in this problem is semilinear and, therefore, the theory

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	2.27×10^{-4}	1.00×10^{-2}	1.37×10^{-2}	1.38×10^{-2}
10^{-1}	3	3.33×10^{-4}	4.60×10^{-3}	3.62×10^{-3}	3.62×10^{-3}
4.60×10^{-3}	5	7.12×10^{-4}	-3.42×10^{-3}	2.10×10^{-4}	2.10×10^{-4}
2.39×10^{-3}	4	8.10×10^{-4}	-2.08×10^{-1}	1.49×10^{-4}	1.48×10^{-4}
1.23×10^{-3}	4	8.63×10^{-4}	-1.18×10^{-3}	8.97×10^{-5}	8.95×10^{-5}
6.26×10^{-4}	5	8.87×10^{-4}	-6.45×10^{-4}	5.05×10^{-5}	5.04×10^{-5}
3.19×10^{-4}	5	9.00×10^{-4}	-3.49×10^{-4}	2.73×10^{-5}	2.72×10^{-5}
1.62×10^{-4}	5	9.07×10^{-4}	-1.90×10^{-4}	1.43×10^{-5}	1.43×10^{-5}
8.23×10^{-5}	5	9.11×10^{-4}	-1.05×10^{-4}	7.40×10^{-6}	7.38×10^{-6}
4.17×10^{-5}	5	9.13×10^{-4}	-6.02×10^{-5}	3.77×10^{-6}	3.76×10^{-6}

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	7	7.43×10^{-5}	1.00×10^{-2}	1.17×10^{-1}	1.24×10^{-1}
10^{-1}	3	1.06×10^{-4}	1.00×10^{-2}	3.75×10^{-2}	3.81×10^{-2}
4.60×10^{-3}	5	2.48×10^{-4}	1.00×10^{-2}	2.72×10^{-3}	2.72×10^{-3}
1.78×10^{-4}	8	8.33×10^{-4}	2.38×10^{-3}	3.52×10^{-4}	3.51×10^{-4}
6.93×10^{-6}	11	9.11×10^{-4}	1.33×10^{-4}	1.95×10^{-5}	1.94×10^{-5}

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	7	1.62×10^{-5}	1.00×10^{-2}	5.48×10^{-1}	7.92×10^{-1}
10^{-1}	4	2.14×10^{-5}	1.00×10^{-2}	2.15×10^{-1}	2.42×10^{-1}
6.40×10^{-3}	5	3.29×10^{-5}	1.00×10^{-2}	1.71×10^{-2}	1.73×10^{-2}
3.40×10^{-4}	8	1.77×10^{-4}	1.00×10^{-2}	1.93×10^{-3}	1.93×10^{-3}
1.81×10^{-5}	12	8.16×10^{-4}	3.46×10^{-3}	5.03×10^{-4}	5.02×10^{-4}
3.91×10^{-6}	10	8.90×10^{-4}	9.18×10^{-4}	1.28×10^{-4}	1.28×10^{-4}

Table 8.26. Test problem II: Course of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\text{final}}$ for $C_\tau = 2 \cdot 10^2$ (top), $C_\tau = 2 \cdot 10^3$ (middle), and $C_\tau = 2 \cdot 10^4$ (bottom)

C_τ	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	$\varepsilon_{\text{final}}$
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
2×10^2	40	3	3	2	2	2	-1.05×10^{-4}	7.36×10^{-6}	7.35×10^{-6}	8.19×10^{-5}
2×10^3	36	5	4	3	3	2	1.33×10^{-4}	1.95×10^{-5}	1.94×10^{-5}	6.93×10^{-6}
2×10^4	61	9	5	5	5	4	3.59×10^{-4}	4.94×10^{-5}	4.93×10^{-5}	1.39×10^{-6}

Table 8.27. Test problem II: Results of $\mathrm{LSM}_{(\mathrm{P})}$ with a nested grid strategy

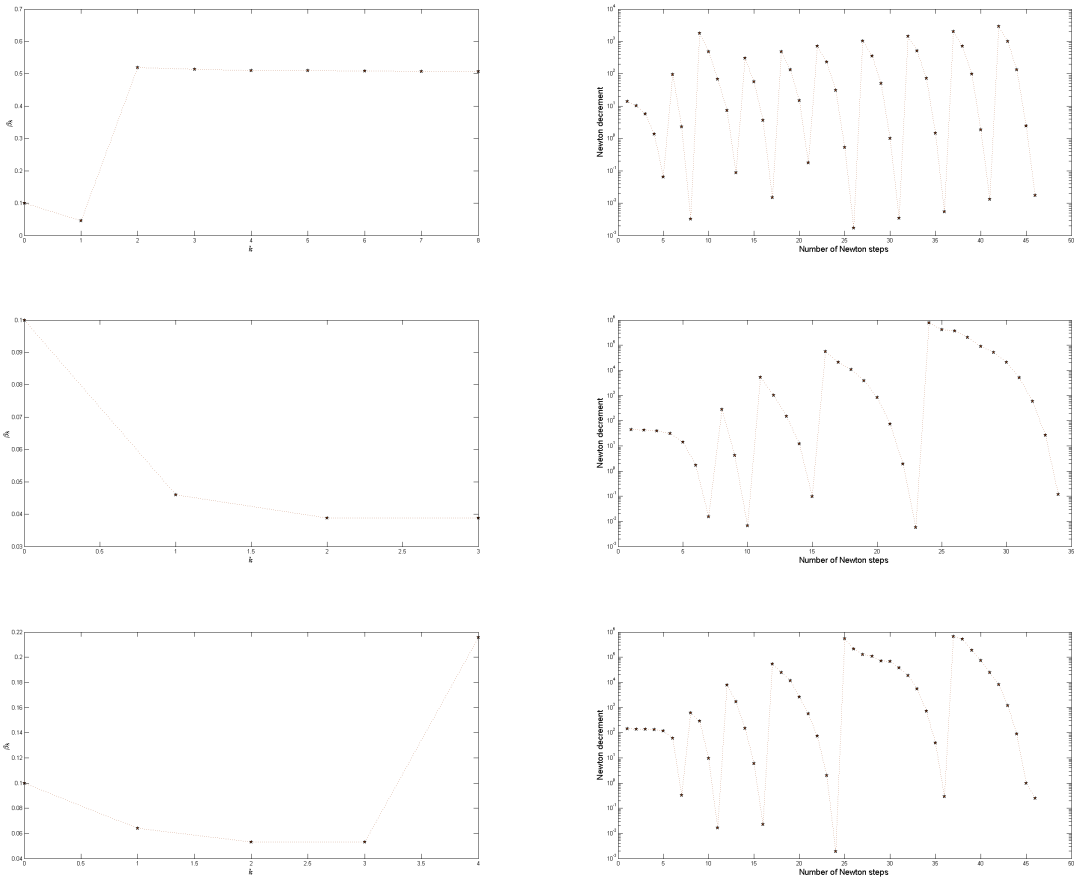

Figure 8.20. Test problem II: Convergence rates and Newton decrements of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$ for $C_\tau = 2 \cdot 10^2$ (top), $C_\tau = 2 \cdot 10^3$ (middle), and $C_\tau = 2 \cdot 10^4$ (bottom)

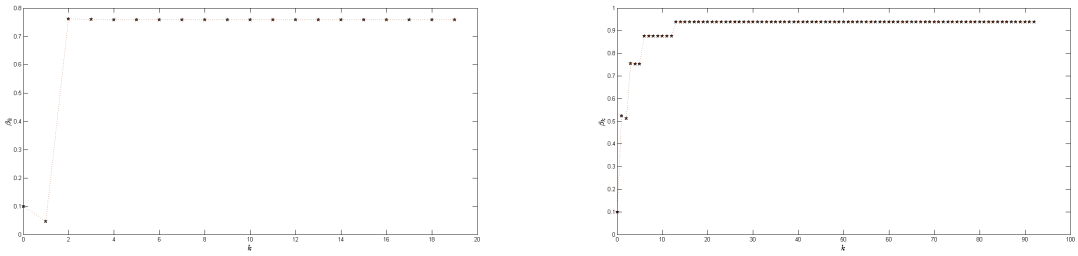

Figure 8.21. Test problem II: Convergence rates of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$ for $C_\tau = 7 \cdot 10^1$ (left) and $C_\tau = 2 \cdot 10^1$ (right)

developed in this thesis is not applicable. Nonetheless, it is interesting to see how $\mathrm{LSM}_{(\mathrm{P})}$ can handle this problem. In what follows we use as optimal solution $(\bar{y}, \bar{u})$ the final iterate of LSM_ε with $\varepsilon = 10^{-7}$ and $h = 2^{-10}$, see Section 8.2.3. Consequently, we employ for $j(\bar{u})$ and $\hat{j}(\bar{u})$ the final values for $j(u^k)$ and $\hat{j}(u^k)$ obtained by LSM_ε with $\varepsilon = 10^{-7}$ and $h = 2^{-10}$. We recall that optimal state, optimal control, and desired state y_d are displayed in Figure 8.14.

We choose $\varepsilon_s = 1$, $C_j = 4 \cdot 10^1$, $\tau(\varepsilon) = C_\tau \frac{1+|\ln \varepsilon|}{\varepsilon}$ with $C_\tau = 1$ for all $\varepsilon \in (0, \varepsilon_s]$, and $C_{\hat{j}} = 1 + \hat{j}(u^0)$ with $u^0 \equiv 0$. Similar to Test Problem III for fixed ε we scaled the weights such that $C_\tau = 1$. To find a good ratio for C_j/C_τ we applied $\mathrm{LSM}_{(\mathrm{P})}$ with automatic detection of $\varepsilon_{\text{final}}$ on coarse meshes for different ratios and observed how many Newton steps were necessary. We chose a ratio that requires a small number of Newton steps. We mention that the automatic detection of $\varepsilon_{\text{final}}$ is exactly the same as for Test Problem I and II. Other parameter values that we employ are $\theta = 0.25$, $\beta_{\min} = 10^{-4}$, and $\beta_{\max} = 0.5$. In the backtracking for $k = 0$ we use $\tilde{\beta}_0 = 0.1$, while in all following iterations $\tilde{\beta}_0$ is determined adaptively in the same way as for Test Problem I and II. Moreover, we modify the choice of the step size and the termination criterion in LSMSUB in the same way as for Test Problem III with fixed ε, cf. Section 8.2.3; in that section we also discuss why this is necessary.

We apply the modified version of $\mathrm{LSM}_{(\mathrm{P})}$ to Test Problem III for different values of $\varepsilon_{\text{final}}$ and on different uniform meshes. In Table 8.28 we display the results. They indicate that $\mathrm{LSM}_{(\mathrm{P})}$ is mesh independent and can be successfully applied to this test problem, although it is not covered by the developed theory. Moreover, $\mathrm{LSM}_{(\mathrm{P})}$ performs similar to LSM_ε for this test problem, cf. Table 8.13. We mention that a modification of the weight C_τ has very similar effects as in Test Problem I and II and, therefore, we do not display results for other choices of C_τ.

Figure 8.22 provides for $\varepsilon_{\text{final}} = 10^{-5}$ the convergence rates (β_k) of (ε_k) and the development of the Newton decrement, computed with $h = 2^{-10}$. As in Test Problem I and II we observe q-linear convergence.

$\varepsilon_{\text{final}}$	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)
10^{-2}	15	12	11	11	11	11(5)	3.65×10^{-2}	1.22×10^{-3}	4.21×10^{-4}
10^{-3}	20	17	21	19	19	19	9.09×10^{-3}	2.43×10^{-4}	9.32×10^{-5}
10^{-4}	31	21	30	26	23	22	1.41×10^{-3}	1.08×10^{-5}	1.34×10^{-5}
10^{-5}	40	33	41	31	32	24	1.17×10^{-4}	3.36×10^{-6}	1.16×10^{-6}

Table 8.28. Test problem III: Results of $\mathrm{LSM}_{(\mathrm{P})}$

We now incorporate into $\mathrm{LSM}_{(\mathrm{P})}$ the strategy proposed in Section 8.3.1 to determine $\varepsilon_{\text{final}}$ automatically. In the termination criterion we use $\mathrm{TOL} = 10^{-3}$ and $\mathrm{TOL} = 10^{-4}$. We note that we have $\hat{j}(\bar{u}) \approx 12.8$ so that requiring $\max\{|\hat{j}(u^{k+1}) - \hat{j}(\bar{u})|, \|(y(u^{k+1}) - y_a)^-\|_{C(\overline{\Omega}_a)}\} \leq 10^{-3}$ is already relatively strict. In Table 8.29 we display the results, while Table 8.30 shows the course of the algorithm on the finest mesh in more detail. We observe in Table 8.30 that our termination criterion works well: $\mathrm{LSM}_{(\mathrm{P})}$ is terminated exactly when the termination criterion is satisfied for the first time. Also, we suspect that (B_k) converges to a value around 2.46. We inspected this by use of $\varepsilon_{\text{final}} = 10^{-8}$ and observed convergence to approximately 2.47.

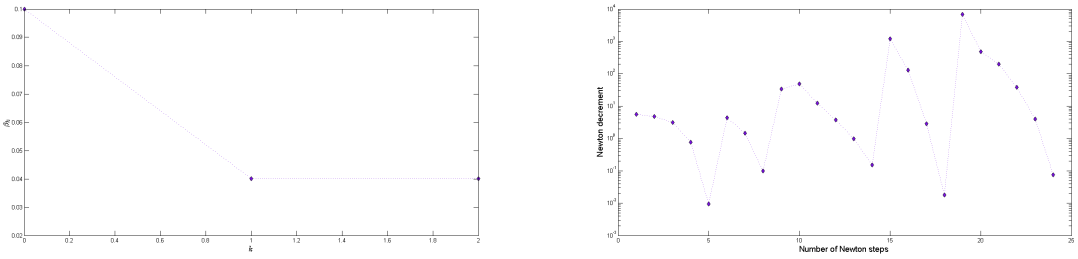

Figure 8.22. Test problem III: Convergence rates and Newton decrements of $\mathrm{LSM}_{(\mathrm{P})}$ with $\varepsilon_{\mathrm{final}} = 10^{-5}$

TOL	Mesh size $h = 2^{-i}$, $i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	$\varepsilon_{\mathrm{final}}$
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-3}	15	12	11	14	14	14(5)	2.51×10^{-2}	7.66×10^{-4}	2.73×10^{-4}	4.60×10^{-3}
10^{-4}	23	21	20	22	19	18	2.64×10^{-3}	4.76×10^{-5}	2.61×10^{-5}	2.12×10^{-4}

Table 8.29. Test problem III: Results of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	1.77×10^{0}	-2.93×10^{-1}	4.46×10^{-3}	1.52×10^{-3}
10^{-1}	3	1.72×10^{0}	1.61×10^{-2}	1.19×10^{-3}	4.19×10^{-4}
4.60×10^{-3}	6	2.19×10^{0}	2.51×10^{-2}	7.67×10^{-4}	2.73×10^{-4}

ε_k	#steps	B_k	$\min(y^{k+1} - y_a)$	$\lvert\hat{j}(u^{k+1}) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^{k+1}) - j(\bar{u})\rvert$
1	5	1.77×10^{0}	-2.93×10^{-1}	4.46×10^{-3}	1.52×10^{-3}
10^{-1}	3	1.72×10^{0}	1.61×10^{-2}	1.19×10^{-3}	4.19×10^{-4}
4.60×10^{-3}	6	2.19×10^{0}	2.51×10^{-2}	7.67×10^{-4}	2.73×10^{-4}
2.12×10^{-4}	4	2.46×10^{0}	2.64×10^{-3}	4.76×10^{-5}	2.61×10^{-5}

Table 8.30. Test problem III: Course of $\mathrm{LSM}_{(\mathrm{P})}$ with automatic determination of $\varepsilon_{\mathrm{final}}$ for $\mathrm{TOL} = 10^{-3}$ (top) and $\mathrm{TOL} = 10^{-4}$ (bottom)

Finally, we add to $\mathrm{LSM}_{(\mathrm{P})}$ a nested grid strategy. As before we use a hierarchy of six grids ranging from $h = 2^{-5}$ to $h = 2^{-10}$. We employ automatic determination of $\varepsilon_{\text{final}}$ on the coarsest grid and $\varepsilon = \varepsilon_{\text{final}}$ on all finer grids. The prolongation onto a finer grid is carried out in exactly the same way as when we applied $\mathrm{LSM}_{\varepsilon}$ with nesting to Test Problem III, cf. Section 8.2.3. If the prolongation involves a shift, we indicate this by writing (s). In Table 8.31 we show the results. We observe that nesting improves the practical performance of $\mathrm{LSM}_{(\mathrm{P})}$ dramatically for this problem. As in Test Problem II we see that the determination of $\varepsilon_{\text{final}}$ on the coarsest mesh has the effect that the termination criterion is violated on the finest mesh. Yet, the violation is very small.

TOL	Mesh size $h = 2^{-i},\ i =$						$\min(y^K - y_a)$	$\lvert\hat{j}(u^K) - \hat{j}(\bar{u})\rvert$	$\frac{1}{C_j}\lvert j(u^K) - j(\bar{u})\rvert$	$\varepsilon_{\text{final}}$
	5	6	7	8	9	10	(on finest mesh)	(o. f. m.)	(o. f. m.)	(o. f. m.)
10^{-3}	15	4	3	2	2	1	3.35×10^{-2}	1.11×10^{-3}	3.81×10^{-4}	8.20×10^{-3}
10^{-4}	20(s)	9	6	6	3	3	5.67×10^{-3}	1.70×10^{-4}	5.86×10^{-5}	5.54×10^{-4}

Table 8.31. Test problem III: Results of $\mathrm{LSM}_{(\mathrm{P})}$ with a nested grid strategy

9. Conclusions and outlook

In this thesis we have presented a new approach to tackle optimal control problems with pointwise state constraints governed by linear elliptic PDEs. The main idea of this approach is to replace the state constraints by a single constraint using a smoothed minimum function. The smoothing parameter induces a family of optimal control problems whose solutions form a path that converges to the optimal solution of the original problem. We call this path the path of solutions. We pursued two ideas:

1) We developed interior point methods for fixed smoothing parameter. These methods converge to a point on the path of solutions. For the remaining length of this path we provided an estimate.

2) We developed interior point methods that drive the smoothing parameter to zero. These methods converge to the optimal solution of the original problem, i.e., the endpoint of the path of solutions.

The methods in 1) admit a very complete convergence analysis in an infinite-dimensional setting; the obtained results are similar to results from finite-dimensional interior point theory. The drawback is, of course, that these methods do not aim for the optimal solution of the original problem. However, by choosing the smoothing parameter small, a solution generated by these methods is close to the solution of the original problem.

For the methods in 2) we provided a detailed convergence analysis in an infinite-dimensional setting, too. In particular, we proved convergence of the iterates to the optimal solution together with an estimate for the error in each iteration. Furthermore, we provided convergence rates for the smoothing parameter and a bound for the required number of Newton steps. We point out that for the problem class under consideration comparable results on convergence rates have not been established before.

A special feature that all methods presented in this thesis share is that there is a quantity, namely the Newton decrement, that provides a theoretically rigorous termination criterion for the inner iteration of the interior point method *and* that can be computed during the course of the algorithm.

We investigated methods from both 1) and 2) in numerical experiments and found that they can be successfully applied in practice, probably even in cases that are not covered by the theory in this thesis. Moreover, we tested some key estimates from theory numerically and found all of them to be sharp.

Yet, there are several open questions. For instance, we only discussed problems with pointwise state constraints; it would be desirable to add control constraints to this setting. Although

control constraints are not considered as challenging as state constraints, it is not straightforward to incorporate these constraints into the self-concordance based framework that we used. For instance, in a standard setting the controls would be elements of $L^2(\Omega)$, but the self-concordant barrier functions that we employed are not well-defined on all of $L^2(\Omega)$.

Another generalization would be to allow for more general pointwise state constraints. We suspect that this is possible as long as the resulting barrier functions are still self-concordant. Moreover, it would be interesting to further investigate the predictor-corrector approach from Section 2.8 in the context of optimal control.

On a more practical note it would be desirable that the developed algorithms take discretization errors into account. Also, the practical efficiency would certainly benefit if goal-oriented mesh refinement was used. Last but not least, it would be interesting to conduct further numerical experiments, for instance for problems where the control acts only on the boundary, where Ω_a is a strict subset of Ω, or where Ω belongs to $\mathbb{R}^3$.

Acknowledgements

At first I would like to express my gratitude to my supervisor Michael Ulbrich for suggesting this interesting subject and letting me explore it with great freedom. Furthermore, I would like to thank him for giving me the opportunity to gain valuable teaching experience; I have always tremendously enjoyed to teach students.

My special thanks are addressed to Boris von Loesch and Dominik Meidner. With Boris I shared an office for over four years; the numerous discussions with him enriched my mathematical work and my personal life. Dominik's help as a colleague and his friendship are invaluable. I am deeply grateful that I have met both of them.

I would also like to thank all my actual and former colleagues at university, particularly Alana, Andre, Christian, Dennis, Florian, Martin, Moritz, Moritz, Sebastian, and Sonja, for the wonderful time we have spent together.

I like to thank Anton, Chen, Dana, Fynn, and Hannah, who studied mathematics with me in Hamburg and who have always encouraged me.

Finally, I thank my family for their support in all of its forms.

Appendices

A. Notation

All vector spaces in this thesis are real vector spaces.

The natural numbers 1, 2, 3, ... are denoted by $\mathbb{N}$. If we want to include zero, we write $\mathbb{N}_0$, i.e., we set $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

We define $\mathbb{R}_{>0}$ and $\mathbb{R}_{\geq 0}$ via $\mathbb{R}_{>0} := \{x \in \mathbb{R} : x > 0\}$ and $\mathbb{R}_{\geq 0} := \{x \in \mathbb{R} : x \geq 0\}$. The sets $\mathbb{R}_{<0}$ and $\mathbb{R}_{\leq 0}$ are defined analogously.

We use $\lceil x \rceil \in \mathbb{Z}$ to denote the smallest integer that is equal to or greater than $x \in \mathbb{R}$, and $\lfloor x \rfloor \in \mathbb{Z}$ for the largest integer that is equal to or smaller than $x \in \mathbb{R}$. This is, $\lceil \cdot \rceil$ rounds up and $\lfloor \cdot \rfloor$ rounds down.

To avoid confusion when dealing with norms we consistently use an index to make clear which norm is meant. For instance, we use $\|\cdot\|_p$ to denote the usual p-norm in $\mathbb{R}^n$, $p \in [1, \infty]$. In particular, $\|x\|_2$ is the Euclidean norm of the vector x. Since this work often deals with several norms on a fixed vector space, it is our opinion that this notation increases accessibility.

Let M and N be sets and $f, g : M \to N$ be functions. Then $f \equiv g$ means $f(x) = g(x)$ for all $x \in M$. In particular, we write $f \equiv 0$ to indicate that f vanishes everywhere on its domain of definition.

For a function $y : M \to \mathbb{R}$ on a set M we define $y^- : M \to \mathbb{R}_{\leq 0}$ via $y^-(x) := \min\{0, y(x)\}$.

Let $D \subset X$ be an open subset of the normed vector space X. We call a three times Gâteaux differentiable function $q : D \to \mathbb{R}$ quadratic on D iff it satisfies $q''' \equiv 0$ on D. Note that q is then infinitely many times continuously differentiable.

A set $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is said to be a domain iff it is open and connected.

For a continuous function $y \in C(\overline{\Omega}_a)$, respectively a function that belongs to a vector space Y with $Y \hookrightarrow C(\overline{\Omega}_a)$, we denote by $\min(y)$ its minimum on $\overline{\Omega}_a$. If the minimum is to be taken on a subset $A \subset \overline{\Omega}_a$ only, we write $\min_{x \in A}(y(x))$.

For two continuous functions $y \in C(\overline{\Omega}_a)$ and $\tilde{y} \in C(\overline{\Omega}_a)$ we often abbreviate pointwise inequalities that are satisfied on all of $\overline{\Omega}_a$ by suppressing the point x. For instance, instead of $y(x) > \tilde{y}(x)\ \forall x \in \overline{\Omega}_a$ we just write $y > \tilde{y}$. Another example of this notation is given by $y \geq 0$.

Newton steps at x are often denoted by n_x. More precisely, for a twice Gâteaux differentiable function $f : D \to \mathbb{R}$ with $D \subset X$ open, X a Banach space, and $f''(x) \in \mathcal{L}(X, X^*)$ invertible, we define $n_x := -f''(x)^{-1} f'(x)$.

To indicate for differentials whether we are dealing with a point x or a direction h, we use different brackets. For example, by $f'(x)[h] \in Z$ we denote the directional derivative of

$f : D \to Z$, $D \subset X$ open, at $x \in D$ in direction $h \in X$, where X and Z are normed vector spaces. If f is at least Gâteaux differentiable at x, $f'(x) \in \mathcal{L}(X, Z)$ denotes the differential of f at x, and $f'(x)[h]$ can then also be interpreted as the evaluation of $f'(x)$ in direction h. In particular, if $Z = \mathbb{R}$ and f is Gâteaux differentiable at x, that is, $f'(x) : X \to \mathbb{R}$ is an element of X^*, the notation $f'(x)[h]$ coincides with the *dual pairing* $\langle f'(x), h\rangle_{X^*,X}$, which we, however, avoid for derivatives.

There are two common ways of looking at differentials of second derivatives. The first one stems from the definition: The twice Gâteaux differentiable function $f : D \to \mathbb{R}$, $D \subset X$ open, has the differential $f''(x) \in \mathcal{L}(X, X^*)$ at $x \in D$. Using the notation from above we write $f''(x)[h_1][h_2]$ for $h_1, h_2 \in X$, when we evaluate the differential $f''(x)[h_1] \in X^*$ in the direction $h_2 \in X$. The second way of regarding the differential $f''(x)$ is to identify $\mathcal{L}(X, X^*)$ with the space $\mathcal{B}(X \times X, \mathbb{R})$, that is the space of bounded bilinear forms. It is easy to see that $\mathcal{L}(X, X^*)$ and $\mathcal{B}(X \times X, \mathbb{R})$ are, indeed, isometrically isomorphic and can, hence, be identified. However, this changes the notation a bit: Taking this point of view we rather write $f''(x)[h_1, h_2]$ for $h_1, h_2 \in X$. Besides notational changes this has an important effect on the existence of inverse operators: While clearly there is no inverse to the bilinear form $f''(x) \in \mathcal{B}(X \times X, \mathbb{R})$, the inverse of the mapping $f''(x) \in \mathcal{L}(X, X^*)$ does exist under certain conditions. In this work we prefer to write $f''(x)[h_1, h_2]$, mainly since it is easier to read. However, if we want to emphasize invertibility, we may use $f''(x)[h_1][h_2]$. As a generalization of these considerations, the differential $f^{(n)}(x)$ of a functional $f : D \to \mathbb{R}$ that is n times Gâteaux differentiable may be considered a bounded multilinear mapping from X^n to $\mathbb{R}$. For more on this topic, see [Zei93, Section 4.5].

B. Sublinear rates of convergence

With the following definitions we give precise meaning to the term sublinear rate of convergence.

Definition B.0.1. Let X be a normed vector space. Let $(x^k) \subset X$ be a convergent sequence with limit point $\bar{x}$ and denote by U the set $U := \{x^k : k \in \mathbb{N}\}$. We say that the sequence (x^k) converges *q-sublinearly* or *at a q-sublinear rate* to $\bar{x} \in X$ iff there exists a function $f : U \to [0, 1)$ and an index $K \in \mathbb{N}$ such that it holds

$$\|x^{k+1} - \bar{x}\|_X \leq f(x^k)\|x^k - \bar{x}\|_X \quad \forall k \geq K.$$

We call $f(x^k)$ the *rate of convergence in iteration k*.

Obviously, every q-linearly convergent sequence is q-sublinearly convergent, too. The converse is not true, as the following example shows.

Example B.0.2. We choose $X = \mathbb{R}$ and consider the function $f : X \to \mathbb{R}$, $f(t) := 1 - t$. We define the recursive sequence (τ_k) via $\tau_1 := \frac{1}{2}$, $\tau_{k+1} := f(\tau_k) \cdot \tau_k$, $k \in \mathbb{N}$. It is positive and monotone decreasing, hence convergent with limit point $\bar{\tau} \geq 0$. From the definition of (τ_k) it follows that $\bar{\tau}$ satisfies $\bar{\tau} = f(\bar{\tau}) \cdot \bar{\tau}$, which yields $\bar{\tau} = 0$. With this it follows from the definition of (τ_k) that this sequence is q-sublinearly convergent to zero with exact rate $f(\tau_k) = 1 - \tau_k$ in iteration k. However, since by definition we have

$$\lim_{k\to\infty} \frac{\tau_{k+1}}{\tau_k} = \lim_{k\to\infty} f(\tau_k) = 1$$

due to $\tau_k \to 0$ for $k \to \infty$, the sequence (τ_k) is not q-linearly convergent.

Definition B.0.3. Let X be a normed vector space. We say that a sequence $(x^k) \subset X$ converges *r-sublinearly* or *at an r-sublinear rate* to $\bar{x} \in X$ iff there exists a sequence $(\tau_k) \subset \mathbb{R}_{>0}$ that converges q-sublinearly to zero and an index $K \in \mathbb{N}$ such that it holds

$$\|x^k - \bar{x}\|_X \leq \tau_k \quad \forall k \geq K.$$

As *rate of convergence in iteration k* we define the corresponding rate of the sequence (τ_k).

Remark B.0.4. Note that q-sublinearly convergent sequences as well as r-sublinearly convergent sequences are, in particular, convergent: For q-sublinearly convergent sequences this is required by definition. For r-sublinearly convergent sequences this follows from the fact that the sequence (τ_k) in the definition of r-sublinear convergence converges to zero.

Remark B.0.5. The notions of q-sublinear and r-sublinear convergence alone are not very strong. For instance, q-sublinear convergence of a sequence (x^k) to $\bar{x}$ is equivalent to the strictly monotone convergence of $(\|x^k - \bar{x}\|_X)_{k\geq K}$ to zero. Of course, these concepts become more meaningful when a rate of convergence is provided.

C. Analysis in normed vector spaces

C.1. (Multi-)Linear operators

We start with the *bounded inverse theorem*, a fundamental result from functional analysis.

Theorem C.1.1. *Let X and Y be Banach spaces and let $A \in \mathcal{L}(X, Y)$ be invertible. Then there holds $A^{-1} \in \mathcal{L}(Y, X)$.*

Proof. See [Yos94, Section 5, Corollary on p. 77], where the spaces X and Y are even allowed to be Fréchet spaces. □

When working with n-th derivatives we want the corresponding differentials to be *symmetric*.

Definition C.1.2. Let X and Y be normed vector spaces and $n \in \mathbb{N}$. Let X^n denote the direct product of n copies of X. We call a multilinear mapping $A : X^n \to Y$ *symmetric* iff it holds $A(h_1, h_2, \ldots, h_n) = A(h_{\sigma(1)}, h_{\sigma(2)}, \ldots, h_{\sigma(n)})$ for all $h_1, h_2, \ldots, h_n \in X$ and every permutation $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. Here, as usual, permutation means bijection.

Remark C.1.3. In the special case $n = 2$ and $Y = \mathbb{R}$ we sometimes consider the bilinear operator $A : X \times X \to \mathbb{R}$ in the preceding definition as $A : X \to X^*$. Then A is symmetric iff it holds $A(x)(y) = A(y)(x)$ for all $x, y \in X$.

The following result is used to demonstrate that Newton's equation is uniquely solvable. It resembles the famous *Lax-Milgram theorem*. Its proof, however, only requires the *Riesz representation theorem* since we work with a symmetric operator.

Theorem C.1.4. *Let $(X, \|\cdot\|_X)$ be a Banach space. Let $c > 0$ and $A \in \mathcal{L}(X, X^*)$ be a symmetric operator that satisfies the inequality*

$$A(h)(h) \geq c \|h\|_X^2 \tag{C.1}$$

for all $h \in X$. Then $(X, \|\cdot\|_X)$ is reflexive and for every $f \in X^*$ the equation

$$A(s) = f$$

has a unique solution $s \in X$. Moreover, A has a continuous inverse $A^{-1} \in \mathcal{L}(X^*, X)$ and there holds

$$\left\| A^{-1} \right\|_{\mathcal{L}(X^*, X)} \leq \frac{1}{c}.$$

In addition, the scalar product induced by A, i.e., $(x, y)_A := A(x)(y)$, yields a norm $\|\cdot\|_A$ on X that is equivalent to $\|\cdot\|_X$. In particular, X is a Hilbert space with respect to $\|\cdot\|_A$.

Proof. Using the properties of A we see that $(x,y)_A$ defines a scalar product, indeed. Furthermore, we infer from (C.1) and the continuity of A that the norm $\|\cdot\|_A$ induced by this scalar product is equivalent to $\|\cdot\|_X$. This implies, firstly, that $(X, \|\cdot\|_A)$ is a Hilbert space, and, secondly, that the dual spaces of $(X, \|\cdot\|_A)$ and $(X, \|\cdot\|_X)$ coincide as sets. Furthermore, the bidual spaces of $(X, \|\cdot\|_A)$ and $(X, \|\cdot\|_X)$ also coincide as sets. It follows that $(X, \|\cdot\|_X)$ is reflexive since this is true for the Hilbert space $(X, \|\cdot\|_A)$. Applying the Riesz representation theorem, see, e.g., [Yos94, Section 6, pp. 90], we deduce that $A(s) = f$ is uniquely solvable for every $f \in X^*$. Thus, A is invertible. Due to (C.1) we have $\|A(h)\|_{X^*} \geq c\|h\|_X$ for all $h \in X$. Using the bijectivity of A this implies that A^{-1} is bounded with $\|A^{-1}\|_{\mathcal{L}(X^*,X)} \leq \frac{1}{c}$, as asserted. Of course, this also shows the continuity of A^{-1}. □

The next lemma provides a generalized version of the well-known *Cauchy-Schwarz inequality.*

Lemma C.1.5. *Let X be a normed vector space. Let $A, B : X \times X \to \mathbb{R}$ be symmetric bilinear forms with*

$$|A(h,h)| \leq B(h,h)$$

for all $h \in X$. Furthermore, let B be positive definite, i.e., for all $h \in X \setminus \{0\}$ it holds $B(h,h) > 0$. Then we have for all $h_1, h_2 \in X$

$$|A(h_1,h_2)| \leq \sqrt{B(h_1,h_1)}\,\sqrt{B(h_2,h_2)}.$$

Proof. We start with an auxiliary consideration: For arbitrary $\tilde{h}_1, \tilde{h}_2 \in X$ we have

$$A(\tilde{h}_1,\tilde{h}_2) = \frac{1}{4}\left(A(\tilde{h}_1+\tilde{h}_2,\tilde{h}_1+\tilde{h}_2) - A(\tilde{h}_1-\tilde{h}_2,\tilde{h}_1-\tilde{h}_2)\right).$$

Invoking the prerequisite we obtain

$$A(\tilde{h}_1,\tilde{h}_2) \leq \frac{1}{4}\left(B(\tilde{h}_1+\tilde{h}_2,\tilde{h}_1+\tilde{h}_2) + B(\tilde{h}_1-\tilde{h}_2,\tilde{h}_1-\tilde{h}_2)\right) = \frac{1}{2}\left(B(\tilde{h}_1,\tilde{h}_1) + B(\tilde{h}_2,\tilde{h}_2)\right). \tag{C.2}$$

We now prove the assertion for $h_1 \in X$ and $h_2 \in X$. Apparently, it suffices to establish $A(h_1,h_2) \leq \sqrt{B(h_1,h_1)}\,\sqrt{B(h_2,h_2)}$. Without loss of generality we assume $h_1, h_2 \neq 0$. We define

$$\mu := \sqrt[4]{B(h_1,h_1)/B(h_2,h_2)} > 0.$$

Application of (C.2) to $\tilde{h}_1 := h_1/\mu$ and $\tilde{h}_2 := \mu h_2$ yields together with the definition of μ

$$A(h_1,h_2) = A(\tilde{h}_1,\tilde{h}_2) \leq \frac{1}{2}\left(\frac{1}{\mu^2}B(h_1,h_1) + \mu^2 B(h_2,h_2)\right) = \sqrt{B(h_1,h_1)}\,\sqrt{B(h_2,h_2)}. \qquad \square$$

The next result is an inequality between symmetric bilinear and trilinear forms.

Lemma C.1.6. *Let X be a normed vector space. Let $A : X \times X \times X \to \mathbb{R}$ be a symmetric trilinear form on X and $B : X \times X \to \mathbb{R}$ be a symmetric bilinear form on X such that*

$$A(h,h,h)^2 \leq \alpha B(h,h)^3$$

is satisfied for all $h \in X$ and a positive constant α. Then it holds for all $h_1, h_2, h_3 \in X$

$$A(h_1,h_2,h_3)^2 \leq \alpha B(h_1,h_1)B(h_2,h_2)B(h_3,h_3).$$

Proof. Let $(h_1, h_2, h_3) \in X \times X \times X$ be given. Defining $V := \mathrm{Span}\{h_1, h_2, h_3\}$ it obviously suffices to prove that

$$A(h,h,h)^2 \leq \alpha B(h,h)^3 \quad \text{for all } h \in V$$

implies

$$A(h_1,h_2,h_3)^2 \leq \alpha B(h_1,h_1)B(h_2,h_2)B(h_3,h_3).$$

We now show that this implication follows from the finite-dimensional version of the assertion. To this end, we define for all $v \in \mathbb{R}^3$ the multiplication $v \cdot (h_1,h_2,h_3) := v_1h_1 + v_2h_2 + v_3h_3$ and set

$$\tilde{A} : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}, \quad \tilde{A}(v_1,v_2,v_3) := A(v_1 \cdot (h_1,h_2,h_3), v_2 \cdot (h_1,h_2,h_3), v_3 \cdot (h_1,h_2,h_3))$$

and

$$\tilde{B} : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}, \quad \tilde{B}(v_1,v_2) := B(v_1 \cdot (h_1,h_2,h_3), v_2 \cdot (h_1,h_2,h_3)).$$

With these definitions it is sufficient to prove that

$$\tilde{A}(v,v,v)^2 \leq \alpha \tilde{B}(v,v)^3 \quad \text{for all } v \in \mathbb{R}^3$$

implies

$$\tilde{A}(e_1,e_2,e_3)^2 \leq \alpha \tilde{B}(e_1,e_1)\tilde{B}(e_2,e_2)\tilde{B}(e_3,e_3),$$

where $e_i \in \mathbb{R}^3$, $i \in \{1,2,3\}$, denotes the canonical unit vector. It is easy to see that $\tilde{A}$ is a symmetric trilinear form and that $\tilde{B}$ is a symmetric bilinear form. In fact, this is directly implied by the corresponding properties of A and B. Thus, it suffices to prove the assertion for the finite-dimensional case of $X = \mathbb{R}^3$. A proof for this case can be found in [Jar92, Appendix, A.1]. (Note that in [Jar92] the form A is additionally assumed to be homogeneous. However, by saying that A is trilinear, we consider it to be homogeneous, too. Thus, we can apply the result from [Jar92].) □

C.2. Differential calculus

The following lemma provides a large class of Lipschitz continuous mappings.

Lemma C.2.1. *Let X and Y be normed vector spaces. Let $D \subset X$ be open and let $K \subset D$ be convex. Let $F : D \to Y$ be Gâteaux differentiable. Then it holds*

$$\|F(x) - F(y)\|_Y \leq \sup_{z \in K} \|F'(z)\|_{\mathcal{L}(X,Y)} \|x - y\|_X$$

for all $x, y \in K$. In particular, F is Lipschitz in K if $\sup_{z\in K} \|F'(z)\|_{\mathcal{L}(X,Y)} < \infty$ is valid.

Proof. Since it is hard to find a proof in the literature if only Gâteaux differentiability is assumed, we provide one here. It is based on [Kön04b, Section 3.2]. If $\sup_{z\in K}\|F'(z)\|_{\mathcal{L}(X,Y)} = \infty$, then there is nothing to prove. Hence, we may assume $L := \sup_{z\in K}\|F'(z)\|_{\mathcal{L}(X,Y)} < \infty$. Fix $x, y \in K$. Without loss of generality we may assume $x \neq y$. The Gâteaux differentiability of F implies that $t \mapsto F(y + t(x - y))$ is continuous in $[0,1]$. For $\varepsilon > 0$ we consider

$$G_\varepsilon : [0,1] \to \mathbb{R}, \quad G_\varepsilon(t) := \|F(y + t(x-y)) - F(y)\|_Y - t\,(L + \varepsilon)\,\|x - y\|_X\,.$$

We establish $G_\varepsilon(1) \leq 0$ for all $\varepsilon > 0$, which implies the assertion. To this end, assume that there is $\varepsilon > 0$ with $G_\varepsilon(1) > 0$. By virtue of $G_\varepsilon(0) = 0$ and the intermediate value theorem (G_ε is continuous) this shows that there exists $t_0 \in (0,1)$ with $G_\varepsilon(t_0) = G_\varepsilon(1)/2$ and $G_\varepsilon(t) > G_\varepsilon(1)/2$ for all $t \in (t_0, 1]$. Hence, it follows

$$\varphi(t) := \frac{G_\varepsilon(t) - G_\varepsilon(t_0)}{t - t_0} > 0 \quad \text{for all } t \in (t_0, 1].$$

Using the reverse triangle inequality the definition of φ implies

$$\varphi(t) \leq \left\| \frac{F(y + t(x-y)) - F(y + t_0(x-y))}{t - t_0} \right\|_Y - (L + \varepsilon) \|x - y\|_X \quad \text{for all } t \in (t_0, 1].$$

Together, we have

$$0 < \left\| \frac{F(y + t(x-y)) - F(y + t_0(x-y))}{t - t_0} \right\|_Y - (L + \varepsilon) \|x - y\|_X \quad \text{for all } t \in (t_0, 1]. \quad \text{(C.3)}$$

With the Gâteaux differentiability of F and the convexity of K we infer

$$\lim_{t \to t_0} \left\| \frac{F(y + t(x-y)) - F(y + t_0(x-y))}{t - t_0} \right\|_Y = \|F'(y + t_0(x-y))[x-y]\|_Y \leq L \|x - y\|_X .$$

Thus, taking the limit $t \to t_0^+$ in (C.3) yields the contradiction

$$0 \leq \|F'(y + t_0(x-y))[x-y]\|_Y - (L + \varepsilon) \|x - y\|_X \leq -\varepsilon \|x - y\|_X < 0. \qquad \square$$

Lemma C.2.2. *Let X and Y be normed vector spaces. Let $K \subset X$ be bounded. If $F : K \to Y$ is Lipschitz continuous, then F is bounded in K.*

Proof. If K is empty, the assertion is trivial. If K is nonempty, choose $y \in K$ arbitrarily. By the triangle inequality and the Lipschitz continuity with constant L we have

$$\|F(x)\|_Y \leq \|F(y)\|_Y + \|F(x) - F(y)\|_Y \leq \|F(y)\|_Y + L \|x - y\|_X$$

for all $x \in K$. Since K is bounded, this inequality proves the assertion. $\square$

Corollary C.2.3. *Let X and Y be normed vector spaces. Let $D \subset X$ be open and let $K \subset D$ be bounded and convex. Let $F : D \to Y$ be Gâteaux differentiable with $\sup_{z \in K} \|F'(z)\|_{\mathcal{L}(X,Y)} < \infty$. Then F is bounded in K.*

Proof. Lemma C.2.1 yields the Lipschitz continuity of F in K, and from Lemma C.2.2 the assertion follows. $\square$

Well-known from finite-dimensional analysis and often very useful is the technique to argue total differentiability through showing continuous partial differentiability. In infinite-dimensional vector spaces this is possible, too, if one replaces the terms "total differentiability" and "partial differentiability" by "Fréchet differentiability" and "Gâteaux differentiability".

Lemma C.2.4. *Let X and Y be normed vector spaces and $D \subset X$ be open. Let $F : D \to Y$ be Gâteaux differentiable in D and let $x \mapsto F'(x)$ be continuous at $x_0 \in D$. Then F is Fréchet differentiable at x_0.*

Proof. There exists $\delta > 0$ with $x_0 + B_\delta(0) \subset D$. Let $(h^k) \subset B_\delta(0) \setminus \{0\}$ with $\lim_{k\to\infty} h^k = 0$ be given. It suffices to establish

$$\lim_{k\to\infty} \frac{\left\|F(x_0 + h^k) - F(x_0) - F'(x_0)[h^k]\right\|_Y}{\|h^k\|_X} = 0. \tag{C.4}$$

Consider for fixed $k \in \mathbb{N}$ the function

$$G : I \to Y, \quad G(t) := F(x_0 + th^k) - tF'(x_0)[h^k],$$

where the open interval I is given by $I := \{t \in \mathbb{R} : th^k \in B_\delta(0)\}$. In particular, we have $[0, 1] \subset I$. The definition of Gâteaux differentiability implies that G is Gâteaux differentiable in I with $G'(t)[s] = s(F'(x_0 + th^k)[h^k] - F'(x_0)[h^k])$. Hence, from Lemma C.2.1 we deduce

$$\begin{aligned}\left\|F(x_0 + h^k) - F(x_0) - F'(x_0)[h^k]\right\|_Y = \|G(1) - G(0)\|_Y &\le \sup_{t\in[0,1]} \|G'(t)\|_{\mathcal{L}(\mathbb{R},Y)} \\ &\le \sup_{t\in[0,1]} \left\|F'(x_0 + th^k) - F'(x_0)\right\|_{\mathcal{L}(X,Y)} \left\|h^k\right\|_X.\end{aligned}$$

This estimate holds for all $k \in \mathbb{N}$. To demonstrate that (C.4) is valid it, thus, remains to argue $\lim_{k\to\infty} \sup_{t\in[0,1]} \|F'(x_0 + th^k) - F'(x_0)\|_{\mathcal{L}(X,Y)} = 0$. However, it is elementary to see that this follows from the continuity of F' at x_0. □

Remark C.2.5. Of course, in the preceding lemma Gâteaux and Fréchet differential coincide.

The next result implies that we do not need to differ between continuous Fréchet differentiability and continuous Gâteaux differentiability; we can just speak of continuous differentiability.

Corollary C.2.6. *Let X and Y be normed vector spaces and $D \subset X$ be open. Let $F : D \to Y$ be continuously Gâteaux differentiable in D. Then F is continuously Fréchet differentiable in D.*

Proof. Using the coincidence of Gâteaux and Fréchet derivative in the case at hand, this follows from the preceding lemma. □

Two basic tools for differentiation in normed vector spaces are product rule and chain rule.

Lemma C.2.7. *Let X, Y_1, Y_2, and Z be normed vector spaces. Let $D \subset X$ be an open set and let $F : D \to Y_1$ and $G : D \to Y_2$ be (continuously) Fréchet differentiable. Moreover, let $a : Y_1 \times Y_2 \to Z$ be a bounded bilinear form. Then*

$$H : D \to Z, \quad H(x) := a\big(F(x), G(x)\big)$$

is (continuously) Fréchet differentiable. Its derivative in direction $h \in X$ is given by

$$H'(x)[h] = a\big(F'(x)[h], G(x)\big) + a\big(F(x), G'(x)[h]\big).$$

Proof. For the proof in the Fréchet differentiable case, see [Zei93, Proposition 4.11]. The prerequisite in [Zei93] that all vector spaces are Banach spaces is not required for the proof given there.

It remains to prove continuity of H' in the case that F' and G' are continuous. Given $x \in D$ and $y \in X$ with $x + y \in D$ we have for all $h \in X$ the estimate

$$\begin{aligned}\|H'(x+y)[h] - H'(x)[h]\|_Z \leq{}& \|a(F'(x+y)[h], G(x+y)) - a(F'(x+y)[h], G(x))\|_Z \\ &+ \|a(F'(x+y)[h], G(x)) - a(F'(x)[h], G(x))\|_Z \\ &+ \|a(F(x+y), G'(x+y)[h]) - a(F(x), G'(x+y)[h])\|_Z \\ &+ \|a(F(x), G'(x+y)[h]) - a(F(x), G'(x)[h])\|_Z .\end{aligned}$$

It is easy to see that the continuity of H' at x follows from this estimate in combination with the boundedness of a and the continuity of F, F', G, and G'. □

Remark C.2.8. An important special case of the preceding lemma is the one where $Z = \mathbb{R}$, $Y_1 = Y_2$, and the bilinear form is a scalar product. If, in addition, $D = X = Y_1 = Y_2$ and $F \equiv G \equiv \mathrm{Id}$ are used, then we obtain the well-known formula for the derivative of $x \mapsto \|x\|_X^2$.

Lemma C.2.9. *Let X, Y, and Z be normed vector spaces. Let $D \subset X$ and $E \subset Y$ be open. Let $F : D \to Y$ and $G : E \to Z$ be (continuously) Fréchet differentiable with $F(D) \subset E$. Then*

$$H : D \to Z, \quad H(x) := G(F(x))$$

is (continuously) Fréchet differentiable. Its derivative in direction $h \in X$ is given by

$$H'(x)[h] = G'(F(x))\,[F'(x)[h]].$$

Proof. In the case of Fréchet differentiability the proof is the same as in finite dimensions, see, e.g., [Kön04b, Section 3.1, II]. In the case of continuous differentiability the continuity of $x \mapsto G'(F(x))F'(x) \in \mathcal{L}(X, Z)$ follows by an estimate similar to the one in the proof of Lemma C.2.7. □

In an important special case the chain rule reads as follows.

Corollary C.2.10. *Let X be a normed vector space. Let $D \subset X$ and $E \subset \mathbb{R}$ be open. Let $F : D \to \mathbb{R}$ be (continuously) Fréchet differentiable with $F(D) \subset E$. Moreover, let $\varphi : E \to \mathbb{R}$ be a (continuously) differentiable function. Then*

$$H : D \to \mathbb{R}, \quad H(x) := \varphi(F(x))$$

is (continuously) Fréchet differentiable. Its derivative in direction $h \in X$ is given by

$$H'(x)[h] = \varphi'(F(x)) \cdot F'(x)[h].$$

Proof. The assertion follows from the preceding lemma by use of $Y = \mathbb{R}$, $Z = \mathbb{R}$, and $G = \varphi$. Note that (continuous) Fréchet differentiability of G translates into (continuous) differentiability of φ, and that there holds $\varphi'(t)[s] = \varphi'(t) \cdot s$ for all directions $s \in \mathbb{R}$. □

A well-known result is *Schwarz's Theorem*, which is also called *Clairaut's Theorem*. We present a version for functionals, i.e., mappings whose image space is $\mathbb{R}$. This is sufficient for our purposes and allows for a particularly simple proof. Moreover, it allows us to slightly weaken the assumption from the frequently required continuous differentiability to Fréchet differentiability only.

Theorem C.2.11. *Let X be a normed vector space and $D \subset X$ be an open set. Let $f : D \to \mathbb{R}$ be n times Fréchet differentiable. Then $f^{(n)}(x) : X^n \to \mathbb{R}$ is symmetric for every $x \in D$.*

Proof. For $n = 1$ the assertion is trivially fulfilled. Hence, let $n \geq 2$. Since every permutation can be written as composition of commutations, cf. [Fis08, Section 3.2.2], it suffices to establish the assertion in the case $n = 2$. To this end, fix $x \in D$ and $h_1, h_2 \in X$. Define

$$\varphi : B_\delta(0) \subset \mathbb{R}^2 \to \mathbb{R}, \quad \varphi(t, s) := f(x + th_1 + sh_2),$$

where $\delta > 0$ is chosen such that $x + th_1 + sh_2 \in D$ holds for all $(t, s) \in B_\delta(0)$. It is elementary to see that φ is twice Fréchet differentiable with

$$\varphi''(t, s)[v][w] = f''(x + th_1 + sh_2)[v_1h_1 + v_2h_2][w_1h_1 + w_2h_2]$$

for all $v, w \in \mathbb{R}^2$. We apply the finite-dimensional version of Schwarz's Theorem, cf. [BF96, Satz, Seite 125ff], to φ at $t = s = 0$. This establishes the assertion:

$$f''(x)[h_1][h_2] = \varphi''(0, 0)[e_1][e_2] = \varphi''(0, 0)[e_2][e_1] = f''(x)[h_2][h_1]. \qquad \square$$

Remark C.2.12. As an alternative to the proof given above we mention that the proof from [BF96] also applies to $f : D \to \mathbb{R}$.

The next lemma shows how to differentiate in a continuously embedded subspace $W \hookrightarrow X$ when a derivative in X is known.

Lemma C.2.13. *Let X and Y be normed vector spaces. Let $D \subset X$ be an open set and let $F : D \to Y$ be (continuously) Fréchet differentiable in D. Moreover, let $W \hookrightarrow X$ be continuously embedded and denote the embedding by $T \in \mathcal{L}(W, X)$. Then the "restriction $\tilde{F}$ of F to W", i.e., $\tilde{F} := F \circ T : T^{-1}(D) \to Y$, is (continuously) Fréchet differentiable in $T^{-1}(D)$, and it holds $\tilde{F}'(w)[h] = F'(x)[T(h)]$ for all $w \in T^{-1}(D)$ and all $h \in W$, where w satisfies $T(w) = x$.*

Remark C.2.14. The set $T^{-1}(D)$ is W-open since T is continuous.

Proof. All assertions follow from the chain rule, cf. Lemma C.2.9. $\square$

Lemma C.2.15. *Let $f : \mathbb{R} \to \mathbb{R}$ be k times continuously differentiable and let Ω be a bounded domain. Then*

$$F : C(\overline{\Omega}) \to \mathbb{R}, \quad F(y) := \int_\Omega f(y(x)) \, \mathrm{d}x$$

is k times continuously differentiable and its r-th differential $F^{(r)}(y) \in (C(\overline{\Omega})^r)^*$ at $y \in C(\overline{\Omega})$ is given by

$$F^{(r)}(y)[h_1, \ldots, h_r] = \int_\Omega f^{(r)}(y(x))h_1(x) \cdots h_r(x) \, \mathrm{d}x.$$

Remark C.2.16. Due to continuity all integrals in the preceding lemma are well-defined.

Remark C.2.17. Apparently, the statement holds true if $\Omega = \cup_{i=1}^{m} \Omega_i$ with disjoint and bounded domains Ω_i, $i = 1, \dots, m$.

Proof. Using the linearity and continuity of the operator $\int_\Omega : C(\overline{\Omega}) \to \mathbb{R}$, the assertion follows from [Ulb11b, Proposition A.13] and the chain rule. □

Corollary C.2.18. *Let Y be a function space with $Y \hookrightarrow C(\overline{\Omega})$. Then the assertions of Lemma C.2.15 hold true with $C(\overline{\Omega})$ replaced by Y and $y(x), h(x)$ replaced by $T(y)(x), T(h)(x)$, where $T \in \mathcal{L}(Y, C(\overline{\Omega}))$ denotes the embedding.*

Proof. Apply Lemma C.2.13 to Lemma C.2.15. □

Remark C.2.19. It is customary not to write embeddings explicitly. Then the preceding result states that we can directly replace $C(\overline{\Omega})$ by Y in Lemma C.2.15.

The next lemma shows how to differentiate integrals that depend on a real parameter. It basically states that we are allowed to "differentiate under the integral". We choose not to give a more general version since this simple version is sufficient for our purposes.

Lemma C.2.20. *Let $I \subset \mathbb{R}$ be an open interval and $\Omega \subset \mathbb{R}^n$ a bounded open set. Furthermore, let $f : I \times \overline{\Omega} \to \mathbb{R}$, $(t, x) \mapsto f(t, x)$ be a function with the following properties:*

- *For every fixed $t \in I$ the function $x \mapsto f(t, x)$ is continuous in $\overline{\Omega}$.*
- *For every fixed $x \in \overline{\Omega}$ the function $t \mapsto f(t, x)$ is differentiable in I.*
- *The function $(t, x) \mapsto \frac{\partial f(t,x)}{\partial t}$ is continuous in $I \times \overline{\Omega}$.*

Then

$$F : I \to \mathbb{R}, \quad F(t) := \int_\Omega f(t, x) \, \mathrm{d}x$$

is continuously differentiable. Its derivative is given by

$$F' : I \to \mathbb{R}, \quad F'(t) := \int_\Omega \frac{\partial f(t, x)}{\partial t} \, \mathrm{d}x.$$

Remark C.2.21. All integrals in the preceding lemma are well-defined due to continuity.

Proof. This result follows from [Kön04b, Section 8.4, Differentiationssatz]. □

The following two corollaries show how to differentiate certain mappings.

Corollary C.2.22. *Let $I \subset \mathbb{R}$ be an open interval and $\Omega \subset \mathbb{R}^n$ a bounded open set. Let $f : I \times \overline{\Omega} \to \mathbb{R}$, $(t, x) \mapsto f(t, x)$ be a function with the same properties as in Lemma C.2.20. Moreover, let Y be a normed vector space with $Y \hookrightarrow C(\overline{\Omega})$. Then*

$$F : I \to Y^*, \quad F(t)(h) := \int_\Omega f(t, x)h(x)\,\mathrm{d}x$$

is continuously differentiable. Its derivative at $t \in I$ in direction $s \in \mathbb{R}$ is given by

$$F'(t)[s] \in Y^*, \quad F'(t)[s](h) = s \cdot \int_\Omega \frac{\partial f(t, x)}{\partial t} \cdot h(x)\,\mathrm{d}x.$$

In the integrals we write h rather than $T(h)$, where T denotes the embedding $Y \hookrightarrow C(\overline{\Omega})$.

Remark C.2.23. $F'(t)$ is well-defined since the occurring integrand is continuous with respect to x.

Proof. We abbreviate by $f'_t(t, x)$ the continuous function $(t, x) \mapsto \frac{\partial f(t,x)}{\partial t}$. Furthermore, we denote by C the constant of the embedding $Y \hookrightarrow C(\overline{\Omega})$. At every point $t_0 \in I$ the asserted differential $F'(t_0)[s] \in Y^*$ is obviously linear in s. It is, furthermore, continuous:

$$\begin{aligned}\sup_{|s|=1} \left\|F'(t_0)[s]\right\|_{Y^*} &= \sup_{|s|=1} |s| \cdot \left\|F'(t_0)[1]\right\|_{Y^*} = \sup_{\|h\|_Y=1} \left|\int_\Omega f'_t(t_0, x)h(x)\,\mathrm{d}x\right| \\ &\leq \sup_{\|h\|_Y=1} \left\|f'_t(t_0, \cdot)\right\|_{L^1(\Omega)} \|h\|_{C(\overline{\Omega})} \leq C \left\|f'_t(t_0, \cdot)\right\|_{L^1(\Omega)}.\end{aligned}$$

Here, we used that $x \mapsto f'_t(t_0, x)$ is continuous in $\overline{\Omega}$, which implies that $\|f'_t(t_0, \cdot)\|_{L^1(\Omega)}$ is finite.

Lemma C.2.20 now yields the continuous differentiability of F in I, as is easy to see. □

Corollary C.2.24. *Let $I \subset \mathbb{R}$ be an open interval and $\Omega \subset \mathbb{R}^n$ a bounded open set. Let $f : I \times \overline{\Omega} \to \mathbb{R}$, $(t, x) \mapsto f(t, x)$ be a function with the same properties as in Lemma C.2.20. Moreover, let Y be a normed vector space with $Y \hookrightarrow C(\overline{\Omega})$. Then*

$$F : I \to \mathcal{L}(Y, Y^*), \quad F(t)(h_1)(h_2) := \int_\Omega f(t, x)h_1(x)h_2(x)\,\mathrm{d}x$$

is continuously differentiable. Its derivative at $t \in I$ in direction $s \in \mathbb{R}$ is given by

$$F'(t)[s] \in \mathcal{L}(Y, Y^*), \quad F'(t)[s](h_1)(h_2) = s \cdot \int_\Omega \frac{\partial f(t, x)}{\partial t} \cdot h_1(x)h_2(x)\,\mathrm{d}x.$$

Proof. Analogue to the proof of the preceding corollary. □

The following result is the well-known *implicit function theorem.*

Theorem C.2.25. *Let X, Y, and Z be Banach spaces and let $D \subset X \times Y$ be open. Let $F : D \to Z$ be (m times) continuously differentiable and let $(x_0, y_0) \in D$ with $F(x_0, y_0) = 0$ be given. Moreover, let $F_y(x_0, y_0) \in \mathcal{L}(Y, Z)$ be invertible. Then there exist $\delta_1, \delta_2 > 0$ such that for each $x \in B_{\delta_1}(x_0)$ there is exactly one $y = y(x) \in B_{\delta_2}(y_0)$ with $F(x, y) = 0$. Furthermore, $x \mapsto y(x)$ is (m times) continuously differentiable.*

Remark C.2.26. $F_y(x_0, y_0)$ denotes the derivative of $y \mapsto F(x_0, y)$ evaluated at $y = y_0$.

Proof. See [Zei93, Theorem 4.B]. □

The next lemma shows that taking the inverse is a Fréchet differentiable mapping.

Lemma C.2.27. *Let X and Y be Banach spaces. Denote by $\mathcal{M} \subset \mathcal{L}(X, Y)$ the set of bounded linear operators that are invertible. Then $\mathcal{M}$ is open and*

$$\mathrm{Inv} : \mathcal{M} \to \mathcal{L}(Y, X), \quad A \mapsto \mathrm{Inv}(A) := A^{-1}$$

is continuously differentiable. Its derivative at $A \in \mathcal{M}$ in direction $H \in \mathcal{L}(X, Y)$ is given by

$$\mathrm{Inv}'(A)[H] = -\mathrm{Inv}(A) \circ H \circ \mathrm{Inv}(A) \quad \in \mathcal{L}(Y, X).$$

Proof. The fact that $\mathcal{M}$ is open is well-known, cf., e.g., [Alt06, Section 3.8]. To infer the continuous differentiability apply the implicit function theorem to the continuously differentiable mapping

$$F : \mathcal{M} \times \mathcal{L}(Y, X) \to \mathcal{L}(Y, Y), \quad F(A, B) := A \circ B - \mathrm{Id}.$$

This is possible since $\mathcal{L}(X, Y)$, $\mathcal{L}(Y, X)$, and $\mathcal{L}(Y, Y)$ are Banach spaces. This yields that the mapping $A \mapsto \mathrm{Inv}(A)$ is continuously differentiable. The formula for the derivative can be deduced from differentiating $A \mapsto F(A, \mathrm{Inv}(A))$ using the chain rule. □

Corollary C.2.28. *Let the mapping $\mathrm{Inv} : \mathcal{M} \to \mathcal{L}(Y, X)$ be defined as in Lemma C.2.27 and let $K \subset \mathcal{M}$ be a convex set in which the inequality $\sup_{A \in K} \|\mathrm{Inv}(A)\|_{\mathcal{L}(Y,X)} < \infty$ is satisfied. Then Inv is Lipschitz continuous in K and the Lipschitz constant L_{Inv} is bounded by*

$$L_{\mathrm{Inv}} \leq \sup_{A \in K} \|\mathrm{Inv}(A)\|^2_{\mathcal{L}(Y,X)} < \infty.$$

Proof. Lipschitz continuity and the first inequality are a direct consequence of Lemma C.2.27 and Lemma C.2.1. The second inequality is obvious. □

At some point we investigate how changes of an invertible operator affect the inverse of this operator. The following corollary provides a suitable estimate.

Corollary C.2.29. *Let X be a Banach space. Let $c > 0$ and assume that the set*

$$P_c := \left\{ A \in \mathcal{L}(X, X^*) : A \text{ symmetric with } A(h)(h) \geq c \|h\|^2_X \text{ for all } h \in X \right\}$$

is nonempty. Then it holds: X is reflexive, all elements of P_c are invertible, and taking the inverse is a Lipschitz continuous operation. More precisely, for all $A, B \in P_c$ we have

$$\left\| A^{-1} - B^{-1} \right\|_{\mathcal{L}(X^*,X)} \leq \frac{1}{c^2} \|A - B\|_{\mathcal{L}(X,X^*)}.$$

Proof. Theorem C.1.4 implies that X is reflexive and that every $A \in P_c$ is invertible with continuous inverse. It remains to estimate the Lipschitz constant. To this end, we define

$$\mathrm{Inv} : \mathcal{M} \to \mathcal{L}(X^*, X), \quad \mathrm{Inv}(A) := A^{-1},$$

with $\mathcal{M} \subset \mathcal{L}(X, X^*)$ as in the two preceding results (note that X^* is a Banach space). Obviously, P_c is a convex subset of $\mathcal{M}$. Furthermore, from Theorem C.1.4 we know that it holds $\|\mathrm{Inv}(A)\|_{\mathcal{L}(X^*,X)} \leq \frac{1}{c}$ for every $A \in P_c$. Therefore, it follows from Corollary C.2.28 that the mapping Inv is Lipschitz continuous on P_c with Lipschitz constant smaller than or equal to $\frac{1}{c^2}$, which shows the assertion. □

C.3. Derivatives of the barrier functions

We compute the derivatives of the barrier functions $f_{\varepsilon,\mu}$.

Lemma C.3.1. *Let $(\varepsilon, \mu) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$. Then it holds:*

The mapping $U_{ad}(\varepsilon) \ni u \mapsto f_{\varepsilon,\mu}(u)$ is thrice Fréchet differentiable. Its first derivative in direction $h \in U$ is given by

$$f'_{\varepsilon,\mu}(u)[h] = \frac{C_j}{\mu} \frac{\hat{j}'(u)[h]}{C_{\hat{j}} - \hat{j}(u)} - \tau(\varepsilon) \frac{(B^\varepsilon)'(u)[h]}{B^\varepsilon(u)}$$

in case I. In case II we have

$$f'_{\varepsilon,\mu}(u)[h] = \frac{C_j}{\mu} \hat{j}'(u)[h] - \tau(\varepsilon) \frac{(B^\varepsilon)'(u)[h]}{B^\varepsilon(u)} - \tilde{\tau}(\varepsilon) \frac{\tilde{B}(u)'[h]}{\tilde{B}(u)}.$$

Its second derivative in directions $(h_1, h_2) \in U \times U$ is given by

$$\begin{aligned} f''_{\varepsilon,\mu}(u)[h_1, h_2] \;=\; & \frac{C_j}{\mu} \left(\frac{\hat{j}''(u)[h_1, h_2]}{C_{\hat{j}} - \hat{j}(u)} - \frac{\hat{j}'(u)[h_1] \cdot \hat{j}'(u)[h_2]}{\left(C_{\hat{j}} - \hat{j}(u)\right)^2} \right) \\ & - \tau(\varepsilon) \left(\frac{(B^\varepsilon)''(u)[h_1, h_2]}{B^\varepsilon(u)} - \frac{(B^\varepsilon)'(u)[h_1] \cdot (B^\varepsilon)'(u)[h_2]}{(B^\varepsilon(u))^2} \right) \end{aligned}$$

in case I. In case II we have

$$\begin{aligned} f''_{\varepsilon,\mu}(u)[h_1, h_2] \;=\; & \frac{C_j}{\mu} \hat{j}''(u)[h_1, h_2] \\ & - \tau(\varepsilon) \left(\frac{(B^\varepsilon)''(u)[h_1, h_2]}{B^\varepsilon(u)} - \frac{(B^\varepsilon)'(u)[h_1] \cdot (B^\varepsilon)'(u)[h_2]}{(B^\varepsilon(u))^2} \right) \\ & - \tilde{\tau}(\varepsilon) \left(\frac{\tilde{B}''(u)[h_1, h_2]}{\tilde{B}(u)} - \frac{\tilde{B}'(u)[h_1] \cdot \tilde{B}'(u)[h_2]}{\left(\tilde{B}(u)\right)^2} \right). \end{aligned}$$

Proof. By definition, $\hat{j}$ is thrice Fréchet differentiable in U. Moreover, $\tilde{B}$ is a quadratic function and, hence, thrice Fréchet differentiable in U. The smoothed minimum $\min_\varepsilon$ is thrice Fréchet differentiable in $C(\overline{\Omega}_a)$, as can be argued by use of Corollary C.2.10, Lemma C.2.15, and the product rule. The chain rule then implies that B^ε is thrice Fréchet differentiable.

The fact that $f_{\varepsilon,\mu}$ is thrice Fréchet differentiable as well as the asserted formulas for the first and second derivative can now be established by use of Corollary C.2.10 and the product rule. □

C.4. Convex analysis

In this section we suppose that $(X, \|\cdot\|_X)$ is a normed vector space if not stated otherwise.

C.4.1. Convex sets

The following lemma shows that convexity of K implies convexity of $\overline{K}$.

Lemma C.4.1. *Let $K \subset X$ be convex. Then $\overline{K}$ is convex.*

Proof. If K is empty, then the assertion is true. Hence, we may suppose that K is nonempty. Let $x, y \in \overline{K}$. There exist sequences $(x^k), (y^k) \subset K$ with $x^k \to x$ and $y^k \to y$ for $k \to \infty$. Hence, for $\lambda \in [0,1]$ and all $k \in \mathbb{N}$ we have $\lambda x^k + (1-\lambda)y^k \in K$. For $k \to \infty$ this shows $\lambda x + (1-\lambda)y \in \overline{K}$. □

The next result characterizes the interior of feasible sets for convex optimization problems.

Lemma C.4.2. *Let $g : X \to \mathbb{R}^m$ be continuous with convex components g_i, $i = 1, \dots, m$. Define $K \subset X$ by $K := \{x \in X : g(x) \leq 0\}$. Then K is convex. Furthermore, if Slater's condition holds, i.e., if there exists $x^\circ \in K$ with $g(x^\circ) < 0$, then the interior of K is given by*

$$\operatorname{int}(K) = \{x \in X : g(x) < 0\},$$

and the closure of $\{x \in X : g(x) < 0\}$ is K. Here, all inequalities are meant componentwise.

Remark C.4.3. Of course, if g is continuous there always holds $\{x \in X : g(x) < 0\} \subset \operatorname{int}(K)$. The result above says that the converse is true if the components of g are convex and $\{x \in X : g(x) < 0\}$ is nonempty. As simple examples describing why these two prerequisites are necessary, consider $g : \mathbb{R} \to \mathbb{R}$, $g(x) = -x^2$ (g is not convex) and $g \equiv 0$ (the set $\{x \in X : g(x) < 0\}$ is empty). Furthermore, it is clear that the closure of $\{x \in X : g(x) < 0\}$ is contained in K. However, these two sets are, in general, not equal. The result above states that the convexity of the components of g and $\{x \in X : g(x) < 0\} \neq \emptyset$ are sufficient conditions for this equality to hold. As examples for the necessity of these two prerequisites use $g : \mathbb{R} \to \mathbb{R}$, $g(x) = \frac{x^3}{3} - x + \frac{2}{3}$ (this function is not convex, and it holds $\tilde{x} := 1 \in K$ but $\tilde{x} \notin \overline{\{x \in X : g(x) < 0\}}$ since $\tilde{x}$ is a local minimizer of g) and, again, $g \equiv 0$.

Proof. It is elementary to show the convexity of K. To prove $\operatorname{int}(K) = \{x \in X : g(x) < 0\}$ it suffices to show $\operatorname{int}(K) \subset \{x \in X : g(x) < 0\}$, as pointed out in the remark. We argue by contradiction. Thus, assume that there exists $\tilde{x} \in \operatorname{int}(K)$ with $g(\tilde{x}) = 0$. Then there is $\varepsilon > 0$ such that all $x \in X$ with $\|x - \tilde{x}\|_X < \varepsilon$ satisfy $g(x) \leq 0$. This implies that there is $\tau > 1$ such that $y := x^\circ + \tau(\tilde{x} - x^\circ)$ satisfies $g(y) \leq 0$. Due to convexity of g_i, $i = 1, \ldots, m$, and $g(x^\circ) < 0$ we infer that it holds for all $\lambda \in [0, 1)$

$$g(\lambda y + (1 - \lambda)x^\circ) < 0.$$

For $\lambda = \frac{1}{\tau} \in (0, 1)$ we obtain $\lambda y + (1 - \lambda)x^\circ = \tilde{x}$ and hence $g(\tilde{x}) < 0$, which is a contradiction.

Using $\overline{\{x \in X : g(x) < 0\}} \subset K$ the last assertion can be established by showing that it holds $K \subset \overline{\{x \in X : g(x) < 0\}}$. To this end, let $x \in K$ be given, i.e., we have $g(x) \leq 0$. Considering $y^k := \lambda_k x + (1 - \lambda_k)x^\circ$ for a sequence $(\lambda_k) \subset [0, 1)$ with $\lambda_k \to 1^-$ for $k \to \infty$, we obtain $y^k \to x$ for $k \to \infty$, and $(y^k) \subset \{x \in X : g(x) < 0\}$ from the convexity of the g_i, $i = 1, \ldots, m$. This implies $x \in \overline{\{x \in X : g(x) < 0\}}$, which concludes the proof. □

C.4.2. Minimizers of convex optimization problems

For convex problems every local minimizer is a global minimizer.

Lemma C.4.4. *Let $f : K \to \mathbb{R}$ be convex on the convex set $K \subset X$. Then every local minimizer of f is a global minimizer. If f is strictly convex, then there is at most one minimizer.*

Proof. Let $\bar{x} \in K$ be a local minimizer of f, i.e., there exists $\delta > 0$ such that $\bar{x}$ is a global minimizer of f on $B_\delta(\bar{x}) \cap K$. Assume that $\bar{x}$ is not a global minimizer of f on K, i.e., there exists $x^* \in K$ with $f(x^*) < f(\bar{x})$. Then for sufficiently small $\lambda \in (0, 1)$ there holds $x := \lambda x^* + (1 - \lambda)\bar{x} \in (B_\delta(\bar{x}) \cap K)$ since K is convex. This implies $f(x) \geq f(\bar{x})$. However, the convexity of f yields the contradiction $f(x) \leq \lambda f(x^*) + (1 - \lambda)f(\bar{x}) < f(\bar{x})$.

If f is strictly convex and $\bar{x}, x^* \in K$ are two (local=global) minimizers with $\bar{x} \neq x^*$, then $f(\frac{1}{2}\bar{x} + \frac{1}{2}x^*) < \frac{1}{2}f(\bar{x}) + \frac{1}{2}f(x^*) = f(\bar{x})$ yields a contradiction. □

The next lemma presents *the* fundamental result for existence of minimizers in reflexive Banach spaces.

Lemma C.4.5. *Let X be a reflexive Banach space. Let $K \subset X$ be nonempty, bounded, closed, and convex. Let $j : K \to \mathbb{R}$ be lower semi-continuous and convex on K. Then there exists a global minimizer of j on K.*

Proof. The proof can, e.g., be found in the classical book of Ekeland and Temam, see [ET99, Chapter 2, Proposition 1.2]. □

The following corollary is very useful in the context of barrier methods.

Corollary C.4.6. *Let X be a reflexive Banach space. Let $K \subset X$ be nonempty, bounded, and convex. Let $f : K \to \mathbb{R}$ be a continuous and convex barrier function for K. Then f possesses a global minimizer on K.*

Proof. Since f is convex on K, any minimizer is global. Fix $\tilde{x} \in K$ and define

$$N := \{x \in K : f(x) \leq f(\tilde{x})\}.$$

We have $\tilde{x} \in N$. Moreover, $N \subset K$ is obviously convex and bounded. Due to the barrier property of f, the set N is also closed in X (not only in K). Hence, by Lemma C.4.5 the function f possesses a minimizer on N. This, of course, is also a minimizer of f on K. □

C.4.3. Convex functions I: Characterizations via derivatives

The following lemmas are often helpful for proving convexity of differentiable functions.

Lemma C.4.7. *Let $D \subset X$ be open, $K \subset D$ convex, and $f : D \to \mathbb{R}$ Gâteaux differentiable. Then f is*

1) convex on K if and only if

$$f(x) + f'(x)[y - x] \leq f(y)$$

2) strictly convex on K if and only if

$$f(x) + f'(x)[y - x] < f(y)$$

3) uniformly convex on K with modulus $\alpha > 0$ if and only if

$$f(x) + f'(x)[y - x] + \frac{\alpha}{2}\|x - y\|_X^2 \leq f(y)$$

holds for all $x, y \in K$ with $x \neq y$.

Proof. This can be proven in the same way as for finite-dimensional X. A proof of the finite-dimensional version can be found in many introductory textbooks on optimization that cover convexity, e.g., [GK99, Satz 3.5] and [UU12, Satz 6.3]. (In these books it is assumed that f is continuously differentiable, however, the proofs do not make use of this since only directional derivatives are considered.) □

Corollary C.4.8. *Let $D \subset X$ be open, $K \subset D$ convex, and $f : D \to \mathbb{R}$ Gâteaux differentiable. Moreover, let f be convex on K and let $\bar{x} \in K$ with $f'(\bar{x}) = 0 \in X^*$ be given. Then $\bar{x}$ is a global minimizer of f on K.*

Proof. This is an immediate consequence of 1) from the preceding lemma. □

Lemma C.4.9. *Let $K \subset X$ be open and convex. Furthermore, let $f : K \to \mathbb{R}$ be twice continuously differentiable. Then f is convex on K (respectively, uniformly convex on K with modulus $\alpha > 0$) if and only if*

$$f''(x)[h,h] \geq 0 \quad \textit{(respectively, } f''(x)[h,h] \geq \alpha\|h\|^2\textit{)}$$

is satisfied for all $x \in K$, $h \in X$. f is strictly convex on K if for all $x \in K$, $h \in X$ we have

$$f''(x)[h,h] > 0.$$

Proof. This can be proven in the same way as for finite-dimensional X, e.g., as in [UU12, Satz 6.4]. □

Remark C.4.10. In the preceding lemma the assumption of twice continuous differentiability can be replaced by twice Gâteaux differentiability. This can be proven as in [BC11, Proposition 17.10]. The standard proofs, however, use Taylor expansion and are, therefore, not valid in this more general setting.

C.4.4. Convex functions II: Uniform convexity

First we state the definition of a uniformly convex function. We do this in particular because we want to make clear what we mean by the *modulus of a uniformly convex function.*

Definition C.4.11. Let $K \subset X$ be a convex set. We call a function $j : K \to \mathbb{R}$ *uniformly convex with (convexity) modulus* $\alpha > 0$ iff it holds

$$j\left(\lambda x + (1-\lambda)\tilde{x}\right) + \frac{\alpha}{2}\lambda(1-\lambda)\left\|x - \tilde{x}\right\|_X^2 \leq \lambda j(x) + (1-\lambda)j(\tilde{x})$$

for all $x, \tilde{x} \in K$ and all $\lambda \in [0,1]$.

Uniform convexity plays an important role in our considerations. This is due to two properties that uniformly convex functions possess: Firstly, uniform convexity implies certain growth properties. For instance, for uniformly convex functions the distance of two points can be measured by the distance of their function values. This allows us to pass from convergence in function value to convergence of iterates. Secondly, uniform convexity implies that Newton's equation has a unique solution. We start with the growth properties.

Lemma C.4.12. *Let $j : K \to \mathbb{R}$ be uniformly convex with modulus $\alpha > 0$ on the convex set $K \subset X$. Assume that j possesses a global minimizer $\overline{x}$ on K. Then it holds for all $x \in K$*

$$\|x - \overline{x}\|_X \leq \frac{2}{\sqrt{\alpha}} \cdot \sqrt{j(x) - j(\overline{x})}.$$

Proof. Let $x \in K$ be arbitrary. Due to the uniform convexity of j with convexity modulus $\alpha > 0$ on K, there holds

$$\frac{\alpha}{4}\|x - \overline{x}\|^2 \leq j(x) + j(\overline{x}) - 2j(\frac{1}{2}x + \frac{1}{2}\overline{x}).$$

we have $-2j(\frac{1}{2}x + \frac{1}{2}\overline{x}) \leq -2j(\overline{x})$ since $\overline{x}$ is the global minimum of j on K. From this we deduce

$$j(x) + j(\overline{x}) - 2j(\frac{1}{2}x + \frac{1}{2}\overline{x}) \leq j(x) - j(\overline{x})$$

and, hence, the assertion follows. □

Lemma C.4.13. *Let $D \subset X$ be open, $K \subset D$ convex, and $f : D \to \mathbb{R}$ Gâteaux differentiable. Moreover, let f be uniformly convex with modulus $\alpha > 0$ on K. Furthermore, let $\tilde{x}, x \in K$. Then it holds*

$$f(x) \leq f(\tilde{x}) \quad \Longrightarrow \quad \|x\|_X \leq \frac{2\|f'(\tilde{x})\|_{X^*}}{\alpha} + \|\tilde{x}\|_X.$$

Moreover, for every $\gamma \in \mathbb{R}$ there exists a $C \geq 0$ such that every $x \in K$ with $\|x\|_X \geq C$ satisfies $f(x) \geq \gamma$. In particular, $L_f(\gamma) := \{x \in K : f(x) < \gamma\}$ is bounded.

Proof. The first assertion is clear for $x = \tilde{x}$. Hence, let $x \neq \tilde{x}$ be satisfied. Due to the uniform convexity of f with modulus $\alpha > 0$ and $f(x) - f(\tilde{x}) \leq 0$ we have

$$\begin{aligned} \frac{\alpha}{8}\|x - \tilde{x}\|_X^2 &\leq \frac{1}{2}\left(f(x) - f(\tilde{x})\right) - \left(f\left(\frac{x}{2} + \frac{\tilde{x}}{2}\right) - f\left(\tilde{x}\right)\right) \\ &\leq -\frac{1}{2}f'(\tilde{x})[x - \tilde{x}] - \frac{\alpha}{8}\|x - \tilde{x}\|_X^2 \leq \frac{1}{2}\|f'(\tilde{x})\|_{X^*}\|x - \tilde{x}\|_X - \frac{\alpha}{8}\|x - \tilde{x}\|_X^2. \end{aligned} \tag{C.5}$$

To derive the second inequality we used Lemma C.4.7. Adding $\frac{\alpha}{8}\|x - \tilde{x}\|_X^2$ to both sides of (C.5), multiplying it with $4/\alpha$, dividing by $\|x - \tilde{x}\|_X$, and applying the reverse triangle inequality $\|x - \tilde{x}\|_X \geq \|x\|_X - \|\tilde{x}\|_X$, we obtain the first assertion.

To deduce the validity of the second assertion fix $\tilde{x} \in K$ and choose $C \geq 0$ so large that

$$f(\tilde{x}) - \|f'(\tilde{x})\|_{X^*}\|x - \tilde{x}\|_X + \frac{\alpha}{4}\|x - \tilde{x}\|_X^2 \geq \gamma \tag{C.6}$$

holds true for all $x \in K$ with $\|x\|_X \geq C$. It is easy to see that such a C exists. Due to the uniform convexity of f we deduce for all these x

$$\frac{\alpha}{8}\|x - \tilde{x}\|_X^2 \leq \frac{1}{2}\left(f(x) - f(\tilde{x})\right) - \left(f\left(\frac{x}{2} + \frac{\tilde{x}}{2}\right) - f\left(\tilde{x}\right)\right) \leq \frac{1}{2}\left(f(x) - f(\tilde{x})\right) - \frac{1}{2}f'(\tilde{x})[x - \tilde{x}],$$

where we used Lemma C.4.7. This yields

$$\frac{\alpha}{4}\|x - \tilde{x}\|_X^2 + f'(\tilde{x})[x - \tilde{x}] + f(\tilde{x}) \leq f(x)$$

for all these x. Together with (C.6) this implies $f(x) \geq \gamma$ for all $x \in K$ with $\|x\|_X \geq C$. □

The next lemma shows that concatenation with the natural logarithm preserves uniform concavity.

Lemma C.4.14. *Let $f : X \to \mathbb{R}$ be twice continuously differentiable and uniformly concave with modulus $\alpha > 0$, i.e., there holds $f''(x)[h, h] \leq -\alpha\|h\|_X^2$ for all $x \in X$ and all $h \in X$. Define $K := \{x \in X : f(x) > 0\}$. Then K is bounded. Let $C_f > 0$ and denote $\bar{f} := \sup_{x \in K} f(x)$. If $\bar{f}$ is finite, then $\tilde{f} : K \to \mathbb{R}$, $\tilde{f}(x) := -C_f \ln(f(x))$ is uniformly convex with modulus $\frac{C_f \alpha}{\bar{f}}$. If K is nonempty and X a Banach space, then $\bar{f}$ is guaranteed to be finite.*

Proof. The boundedness of K follows from Lemma C.4.13. Supposing that $\bar{f}$ is finite the assertion on the uniform convexity of $\tilde{f}$ follows from a simple computation. In fact, we have for all $x \in K$ and $h \in X$

$$\begin{aligned} \tilde{f}''(x)[h,h] &= C_f \left(\left(\frac{f'(x)[h]}{f(x)} \right)^2 - \frac{f''(x)[h,h]}{f(x)} \right) \\ &\geq C_f \left(\frac{-f''(x)[h,h]}{f(x)} \right) \geq C_f \alpha \frac{\|h\|_X^2}{f(x)} \geq \frac{C_f \alpha}{\bar{f}} \|h\|_X^2. \end{aligned} \tag{C.7}$$

We show that $\bar{f}$ is finite under the additional assumptions by establishing $\inf_{x \in K} \tilde{f}(x) > -\infty$. K is nonempty, convex, and bounded. Moreover, $\tilde{f}$ is a continuous barrier function for K since $\partial K \subset \{x \in X : f(x) = 0\}$, as follows from the continuity of f. Without assuming finiteness of $\bar{f}$, (C.7) still shows that $\tilde{f}$ is convex on K. Therefore, Corollary C.4.6 implies $\inf_{x \in K} \tilde{f}(x) > -\infty$. To apply this corollary we note that X is reflexive due to Theorem C.1.4. □

Finally, we establish that Newton's equation has a unique solution for uniformly convex functions.

Theorem C.4.15. *Let $(X, \|\cdot\|_X)$ be a Banach space. Let $K \subset X$ be nonempty, open, and convex. Let $f : K \to \mathbb{R}$ be twice continuously differentiable and uniformly convex on K with modulus $\alpha > 0$. Then X is reflexive and Newton's equation at $x \in K$, i.e.,*

$$f''(x)[s] = -f'(x),$$

posed in X^*, has a unique solution $s \in X$. Moreover, $f''(x) \in \mathcal{L}(X, X^*)$ has a continuous inverse $f''(x)^{-1} \in \mathcal{L}(X^*, X)$ and there holds

$$\left\| f''(x)^{-1} \right\|_{\mathcal{L}(X^*, X)} \leq \frac{1}{\alpha}.$$

In addition, the scalar product induced by $f''(x)$ yields a norm $\|\cdot\|_{f''(x)}$ on X that is equivalent to $\|\cdot\|_X$. In particular, X is a Hilbert space with respect to $\|\cdot\|_{f''(x)}$.

Proof. Using the uniform convexity of f with modulus α and the differentiability properties, we have for all $h \in X$

$$f''(x)[h,h] \geq \alpha \|h\|_X^2 .$$

Since there also holds $f''(x) \in \mathcal{L}(X, X^*)$ and since this operator is symmetric due to Schwarz's theorem, see Theorem C.2.11, we can apply Theorem C.1.4. This theorem yields all assertions. □

D. Inequalities

The first inequality we present is a generalization of *Bernoulli's inequality*.

Lemma D.0.1. *Let $p \in \mathbb{R}$ with $p \geq 1$ and $t \in \mathbb{R}$ with $t \geq -1$. Then it holds $(1+t)^p \geq 1 + pt$.*

Proof. The function $f : \overline{J} \to \mathbb{R}$, $f(t) := (1+t)^p - (1+pt)$ satisfies $f''(t) \geq 0$ on $J := (-1, \infty)$, as a simple computation shows. Hence, f is convex on J, which implies $f(0) + f'(0)(t-0) \leq f(t)$ for all $t \in J$. Since we have $f(0) = 0$ and $f'(0) = 0$, this yields $f(t) \geq 0$ on $J = (-1, \infty)$. By continuity of f we, thus, obtain $f(t) \geq 0$ on $\overline{J} = [-1, \infty)$, which proves the assertion. □

The second inequality can be considered as a *reverse version of Bernoulli's inequality* on a restricted interval.

Lemma D.0.2. *Let $p \geq 1$, $t \in [-\frac{1}{2}, 0]$, and $c := (\frac{1}{2})^{p-1}$. Then it holds $(1+t)^p \leq 1 + cpt$.*

Proof. Let $J := [-\frac{1}{2}, 0]$ and define $f : J \to \mathbb{R}$, $f(t) := 1 + cpt - (1+t)^p$. There holds $f'(t) \leq 0$ for all $t \in J$, as a simple computation shows. Hence, f is monotonically decreasing on J. This implies $f(t) \geq f(0) = 0$ on J, thus establishing the assertion. □

The next result is a very general version of the *fundamental theorem of calculus*.

Lemma D.0.3. *Let $f : [a, b] \to \mathbb{R}$ be differentiable with bounded derivative. Then f' is Lebesgue integrable on $[a, b]$ and it holds $f(b) - f(a) = \int_a^b f'(t) \, \mathrm{d}t$.*

Proof. See [Kön04b, Satz, p. 280]. □

We present a differential version of *Gronwall's inequality*. We provide a small extension in comparison to versions that are usually found in the literature, cf., e.g., [Eva10, Appendix B j].

Lemma D.0.4.

1) Let $f : I \to \mathbb{R}_{\geq 0}$ be nonnegative and differentiable with bounded derivative on $I := [t_0, T]$, where $t_0 < T$. Also, let f satisfy $f'(t) \leq \alpha(t) f(t) + \beta(t)$ for all $t \in I$, where $\alpha, \beta : I \to \mathbb{R}_{\geq 0}$ are nonnegative and continuous. Then it holds for all $t \in I$

$$f(t) \leq e^{\int_{t_0}^t \alpha(s) \, \mathrm{d}s} \left(f(t_0) + \int_{t_0}^t \beta(s) \, \mathrm{d}s \right).$$

2) Let $f : I \to \mathbb{R}_{\geq 0}$ be nonnegative and differentiable with bounded derivative on $I := [t_0, T]$, where $t_0 < T$. Also, let f satisfy $f'(t) \geq \alpha(t)f(t) + \beta(t)$ for all $t \in I$, where $\alpha : I \to \mathbb{R}_{\leq 0}$ is nonpositive, $\beta : I \to \mathbb{R}_{\geq 0}$ is nonnegative, and both functions are continuous. Then it holds for all $t \in I$

$$f(t) \geq e^{\int_{t_0}^{t} \alpha(s)\,\mathrm{d}s} \left(f(t_0) + \int_{t_0}^{t} \beta(s)\,\mathrm{d}s \right).$$

Remark D.0.5. Obviously, the boundedness of f' from above is implied for 1) by the assumed inequality. In 2) this is true for the boundedness of f' from below.

Proof. In the setting of 1) we have for all $s \in I$

$$\frac{\mathrm{d}}{\mathrm{d}s} \left(f(s) e^{-\int_{t_0}^{s} \alpha(r)\,\mathrm{d}r} \right) = e^{-\int_{t_0}^{s} \alpha(r)\,\mathrm{d}r} \left(f'(s) - \alpha(s) f(s) \right) \leq e^{-\int_{t_0}^{s} \alpha(r)\,\mathrm{d}r} \beta(s).$$

The left-hand side is uniformly bounded on I due to boundedness of f' and continuity of α and f on the compact set I. This shows that the fundamental theorem of calculus applies, cf. Lemma D.0.3. Integrating from t_0 to $t \in I$ we obtain

$$f(t) e^{-\int_{t_0}^{t} \alpha(r)\,\mathrm{d}r} \leq f(t_0) + \int_{t_0}^{t} e^{-\int_{t_0}^{s} \alpha(r)\,\mathrm{d}r} \beta(s)\,\mathrm{d}s \leq f(t_0) + \int_{t_0}^{t} \beta(s)\,\mathrm{d}s.$$

From this the assertion of 1) is evident.

To establish 2) we can use the same arguments as for 1), but with $\leq$ replaced by $\geq$. □

Another result on differential inequalities is the following. It is inspired by [Wal00, §9, IX].

Lemma D.0.6. *Let $I := (a, b)$ and (c, d) be intervals, where each of the choices $b = \infty$, $c = -\infty$, and $d = \infty$ is allowed. Let $f : [a, b) \times (c, d) \to \mathbb{R}$ be locally Lipschitz with respect to the second argument, i.e., for every $(t_0, y_0) \in [a, b) \times (c, d)$ there exists a neighborhood $U = U(t_0, y_0) \subset [a, b) \times (c, d)$ of (t_0, y_0) and an $L = L(t_0, y_0) \geq 0$ with $|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$ provided $(t, y_1), (t, y_2) \in U$. Let $y_a \in (c, d)$ be given and assume that there exists a continuous function $\bar{y} : [a, b) \to (c, d)$ that is differentiable in I and solves the initial value problem*

$$y'(t) = f(t, y(t)) \text{ in } I, \quad y(a) = y_a.$$

Furthermore, assume that there exists a continuous function $\hat{y} : [a, b) \to (c, d)$ that is differentiable in I and satisfies

$$\hat{y}'(t) \leq f(t, \hat{y}(t)) \text{ in } I, \quad \hat{y}(a) \leq y_a.$$

Then it holds $\hat{y}(t) \leq \bar{y}(t)$ for all $t \in [a, b)$.

Proof. Suppose the assertion is wrong. Then there exists a nontrivial interval $[\alpha, \beta] \subset [a, b)$ with $\hat{y}(\alpha) = \bar{y}(\alpha)$ and $\hat{y}(t) > \bar{y}(t)$ for all $t \in (\alpha, \beta]$. Let $L \geq 0$ denote the Lipschitz constant of f in a neighborhood $U \subset [a, b) \times (c, d)$ of $(\alpha, \hat{y}(\alpha)) = (\alpha, \bar{y}(\alpha))$. Due to the continuity of $\hat{y}$ and $\bar{y}$ on $[a, b)$ we can shrink β until it so small that $(t, \hat{y}(t)) \in U$ and $(t, \bar{y}(t)) \in U$ are satisfied for all $t \in [\alpha, \beta]$. Of course, we still have $\hat{y}(t) > \bar{y}(t)$ for all $t \in (\alpha, \beta]$.

Denote by $y : [a, b) \to \mathbb{R}$ the continuous function $y(t) := \hat{y}(t) - \bar{y}(t)$. It holds

$$y(\alpha) = 0, \qquad y(t) > 0 \text{ for all } t \in (\alpha, \beta].$$

In contradiction to this we now show that $y(t) \leq 0$ is valid for all $t \in [\alpha, \beta]$. This establishes the assertion. Since y is differentiable in $I \supset (\alpha, \beta]$, we have for all $t \in (\alpha, \beta]$:

$$y'(t) = \underbrace{\hat{y}'(t) - f(t, \hat{y}(t))}_{\leq 0} - \underbrace{(\bar{y}'(t) - f(t, \bar{y}(t)))}_{=0} + \underbrace{f(t, \hat{y}(t)) - f(t, \bar{y}(t))}_{\leq L(\hat{y}(t) - \bar{y}(t))} \leq Ly(t).$$

We set $g : [\alpha, \beta] \to \mathbb{R}$, $g(t) := y(t)e^{-Lt}$. Then we have

$$g'(t) = \left(y(t)e^{-Lt}\right)' = \left(y'(t) - Ly(t)\right) e^{-Lt} \leq 0$$

for all $t \in (\alpha, \beta]$. With $g(\alpha) = 0$ it follows $g(t) \leq 0$ on $[\alpha, \beta]$ since the existence of $\hat{t} \in (\alpha, \beta]$ with $g(\hat{t}) > 0$ yields $\tilde{t} \in (\alpha, \hat{t})$ with $g'(\tilde{t}) > 0$ via the mean value theorem, cf. [Kön04a, Section 9.3]. From $g(t) \leq 0$ on $[\alpha, \beta]$ we infer $y(t) \leq 0$ for all $t \in [\alpha, \beta]$ thus finishing the proof. □

E. Cone condition

In this section we offer a rigorous treatment of the *cone condition*, cf. [Ada75, p. 66].

Definition E.0.1. Let $v \in \mathbb{R}^d \setminus \{0\}$, $0 < \kappa \leq \pi$, and $\rho > 0$. Then the set

$$C := \left\{ x \in \mathbb{R}^d : \|x\|_2 \leq \rho \quad \text{and} \quad \angle(x, v) \leq \frac{\kappa}{2} \right\}$$

is called a *finite cone* of height ρ, axis direction v, and aperture angle κ, with vertex at the origin. Here, $\angle(x, v) \in [0, \pi]$ denotes the unoriented angle between x and v.

Definition E.0.2. Let $A, B \subset \mathbb{R}^d$. We say that A *is congruent to* B iff there exists a vector $w \in \mathbb{R}^d$ and an orthogonal linear mapping $T : \mathbb{R}^d \to \mathbb{R}^d$, i.e., $\|Tv\|_2 = \|v\|_2$ for all $v \in \mathbb{R}^d$, such that $B = w + T(A)$ holds.

Definition E.0.3. We say that the domain $\Omega \subset \mathbb{R}^d$ satisfies the *cone condition* iff there exists a finite cone $C \subset \mathbb{R}^d$ with vertex at the origin such that each point $x \in \Omega$ is the vertex of a finite cone C_x contained in Ω and congruent to C. More precisely, for each $x \in \Omega$ there is an orthogonal linear mapping T_x with $C_x := x + T_x(C) \subset \Omega$.

Remark E.0.4. Every Lipschitz domain satisfies the cone condition, see [Ada75, p. 66f].

We make use of the cone condition during the integration of *rotational symmetric* functions.

Definition E.0.5. We say $f : \mathbb{R}^d \to \mathbb{R}$ is *rotational symmetric with respect to* $x_0 \in \mathbb{R}^d$ iff there exists a function $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that $f(x) = \varphi(\|x - x_0\|_2)$ holds for all $x \in \mathbb{R}^d$.

The cone condition comes into play since we want to change the domain of integration of certain nonnegative functions, defined on $\mathbb{R}^d$, from Ω to $B_\delta(x_0)$, $x_0 \in \Omega$. Since it may happen that $B_\delta(x_0)$ is not completely contained in Ω, it is not at all clear how the domain change affects the value of the integral. However, if Ω satisfies the cone condition and the integrand is, in addition, rotational symmetric, we have the following estimate.

Lemma E.0.6. *Let $\Omega \subset \mathbb{R}^d$ satisfy the cone condition. Let $x_0 \in \Omega$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be measurable, nonnegative, and rotational symmetric with respect to x_0. Then there exist constants $c, \delta > 0$ such that it holds*

$$\int_\Omega f(x)\,\mathrm{d}x \geq c \int_{B_\delta(x_0)} f(x)\,\mathrm{d}x.$$

The constants $c, \delta > 0$ depend only on Ω, but neither on f nor x_0.

Remark E.0.7. The integrals in the above inequality are well-defined (with possible value $+\infty$) since f is measurable and since the sets Ω and $B_\delta(x_0)$ are open, hence measurable.

Proof. In what follows, all integrals are well-defined since all integration domains are measurable, as is readily seen. Also, it poses no difficulties to see that the proof stays valid if one or both of the integrals in the assertion take the value $+\infty$. Therefore, we do not mention these properties during the remainder of the proof.

Let $C_{x_0} = x_0 + T_{x_0}(C) \subset \Omega$ denote the cone at x_0 given by the cone condition. Obviously, we have $\int_\Omega f(x)\,\mathrm{d}x \geq \int_{C_{x_0}} f(x)\,\mathrm{d}x$. To establish the assertion it, thus, suffices to argue the existence of constants $c, \delta > 0$ that are independent of f and x_0 and that satisfy $\int_{C_{x_0}} f(x)\,\mathrm{d}x \geq c \int_{B_\delta(x_0)} f(x)\,\mathrm{d}x$. To this end, set $\delta := \rho/2 > 0$, where ρ is the height of C. Apparently, δ only depends on Ω. We now prove rigorously that it is possible to cover $\overline{B_\delta(x_0)}$ with finitely many congruent copies of C_{x_0}, all originating at x_0. To do so, note first that an open cover for $\overline{B_\delta(x_0)}$ is given by

$$\left(x_0 + \bigcup_{\|v\|=\rho} \left\{ x \in \mathbb{R}^d : \, 0 < \|x\|_2 < \rho \quad \text{and} \quad \angle(x, v) < \frac{\kappa}{2} \right\} \right) \cup B_{\delta/2}(x_0),$$

with κ denoting the aperture angle of C. The compactness of $\overline{B_\delta(x_0)}$ implies that this cover contains a finite subcover. Let m denote the number of elements of this subcover. Clearly, m is independent of f. But m is also independent of x_0 since for a different point $\tilde{x}_0 \in \Omega$ we can use the same open cover and finite subcover with translation $\tilde{x}_0$ instead of x_0 to cover $\overline{B_\delta(\tilde{x}_0)}$. Since x_0 needs to be covered, $B_{\delta/2}(x_0)$ is one of the m elements of the finite subcover. However, the other $m - 1 \geq 1$ elements still cover $\overline{B_\delta(x_0)} \setminus \{x_0\}$, as we argue now. In fact, let $x \in \overline{B_\delta(x_0)} \setminus \{x_0\}$. Then $y := x_0 + \delta \frac{x - x_0}{\|x - x_0\|_2} \in \overline{B_\delta(x_0)} \setminus B_{\delta/2}(x_0)$ is covered by an element of the finite subcover, i.e., there is $v \in \mathbb{R}^d$ with $\|v\|_2 = \rho$ such that $y \in x_0 + \hat{C}$ is satisfied, where $\hat{C} := \{x \in \mathbb{R}^d : \, 0 < \|x\|_2 < \rho \quad \text{and} \quad \angle(x, v) < \frac{\kappa}{2}\}$. Since $w \in \hat{C}$ implies $tw \in \hat{C}$ for all $t \in (0, 1]$, we obtain $t\delta \frac{x - x_0}{\|x - x_0\|_2} \in \hat{C}$ for all $t \in (0, 1]$. With $t = \|x - x_0\|_2/\delta \in (0, 1/2)$ this shows that it holds $x \in x_0 + \hat{C}$, i.e., x is covered by an element of the finite subcover different from $B_{\delta/2}(x_0)$. In conclusion, there are v_i with $\|v_i\| = \rho$, $i = 1, \ldots, m - 1$, such that

$$\overline{B_\delta(x_0)} \subset x_0 + \bigcup_{i=1}^{m-1} C_i = \bigcup_{i=1}^{m-1} (x_0 + C_i) \tag{E.1}$$

holds, where the C_i are given by

$$C_i := \left\{ x \in \mathbb{R}^d : \, \|x\|_2 \leq \rho \quad \text{and} \quad \angle(x, v_i) \leq \frac{\kappa}{2} \right\}, \quad i = 1, \ldots, m - 1.$$

Note that $C_{x_0} - x_0$ also has this form, i.e., there exists v_0 with $\|v_0\| = \rho$ such that it holds

$$C_0 := C_{x_0} - x_0 = \left\{ x \in \mathbb{R}^d : \, \|x\|_2 \leq \rho \quad \text{and} \quad \angle(x, v_0) \leq \frac{\kappa}{2} \right\}.$$

It is easy to see that each C_i, $i = 0, \ldots, m - 1$, is the image of C under an orthogonal transformation $T_i : \mathbb{R}^d \to \mathbb{R}^d$. In fact, T_i can be any orthogonal linear mapping that maps $v/\|v\|$ to $v_i/\|v_i\|$, where v denotes the axis direction of C. It follows from elementary linear algebra that such a mapping exists for each i, regardless of the dimension d. More precisely,

take for fixed i an orthonormal basis $h_1, h_2, \ldots, h_d \in \mathbb{R}^d$ with $h_1 = v/\|v\|_2$ and an orthonormal basis $w_1, w_2, \ldots, w_d \in \mathbb{R}^d$ with $w_1 = v_i/\|v_i\|_2$. Let $A \in \mathbb{R}^{d\times d}$ denote the matrix that maps $(h_1, \ldots, h_d)$ to the canonical base, i.e., $(h_1, \ldots, h_d)^{-1}$, and let $B \in \mathbb{R}^{d\times d}$ denote the matrix $(w_1, \ldots, w_d)$, that maps the canonical base to $(w_1, \ldots, w_d)$. Then A and B are orthogonal and, hence, the mapping $T_i \in \mathbb{R}^{d\times d}$ given by $T_i := BA$ is orthogonal, too. Also, $T_i(h_j) = w_j$ for $j = 1, \ldots, d$, and in particular $T_i(v/\|v\|_2) = v_i/\|v_i\|_2$, as required.

We define $K_i := x_0 + C_i$ for $i = 1, \ldots, m-1$, and use the rotational symmetry of f to infer

$$\int_{C_{x_0}} f(x)\,\mathrm{d}x = \int_{C_0} \varphi(\|x\|)\,\mathrm{d}x \quad \text{and} \quad \int_{K_i} f(x)\,\mathrm{d}x = \int_{C_i} \varphi(\|x\|)\,\mathrm{d}x,\ \ i = 1, \ldots, m-1. \tag{E.2}$$

The orthogonality implies for $i = 0, \ldots, m-1$ that $|\det(T_i)| = 1$ holds and that for all $x \in \mathbb{R}^d$ we have $\|T_i(x)\| = \|x\|$. Using the change of variables formula we deduce from this

$$\int_{C_i} \varphi(\|x\|)\,\mathrm{d}x = \int_{T_i(C)} \varphi(\|x\|)\,\mathrm{d}x = \int_C \varphi(\|T_i(x)\|)\,|\det(T_i)|\,\mathrm{d}x = \int_C \varphi(\|x\|)\,\mathrm{d}x$$

for $i = 0, \ldots, m-1$. With (E.2) this yields $\int_{C_{x_0}} f(x)\,\mathrm{d}x = \int_{K_1} f(x)\,\mathrm{d}x = \ldots = \int_{K_{m-1}} f(x)\,\mathrm{d}x$. Employing $K := \cup_{i=1}^{m-1} K_i \supset B_\delta(x_0)$, see (E.1), in combination with the nonnegativity of f, we infer

$$(m-1)\int_{C_{x_0}} f(x)\,\mathrm{d}x = \sum_{i=1}^{m-1} \left(\int_{K_i} f(x)\,\mathrm{d}x \right) \geq \int_K f(x)\,\mathrm{d}x \geq \int_{B_\delta(x_0)} f(x)\,\mathrm{d}x.$$

As pointed out before, m only depends on Ω. Thus, the assertion follows with $c := 1/(m-1)$. $\square$

Bibliography

[Ada75] R. A. Adams. *Sobolev spaces.* Academic Press, 1975.

[Ado92] V. Adolfsson. L^2-integrability of second order derivatives for Poisson's equation in nonsmooth domains. *Math. Scand.*, 70(1):146–160, 1992.

[Alt06] H. W. Alt. *Lineare Funktionalanalysis. 5. Auflage.* Springer. Berlin, 2006.

[Ama93] H. Amann. Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems. In *Function Spaces, Differential Operators and Nonlinear Analysis.* B. G. Teubner Verlagsgesellschaft. Stuttgart, 1993.

[Aus99] A. Auslender. Penalty and barrier methods: A unified framework. *SIAM J. Optim.*, 10(1):211–230, 1999.

[Bau01] H. Bauer. *Measure and integration theory. Transl. from the German by Robert B. Burckel.* de Gruyter. Berlin, 2001.

[BBC+93] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst. *Templates for the solution of linear systems: building blocks for iterative methods.* SIAM. Philadelphia, PA, 1993.

[BC11] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* Springer. Berlin, 2011.

[Ber99] D. P. Bertsekas. *Nonlinear programming. 2nd ed.* Athena Scientific. Belmont, MA, 1999.

[BF96] M. Barner and F. Flohr. *Analysis II. 3. Auflage.* De Gruyter Lehrbuch. Berlin, 1996.

[Bra13] D. Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie. 4., überarbeitete und erweiterte Auflage.* Springer Spektrum. Berlin, 2013.

[BTT89] A. Ben-Tal and M. Teboulle. A smoothing technique for nondifferentiable optimization problems. In Szymon Dolecki, editor, *Optimization*, volume 1405 of *Lecture Notes in Mathematics*, pages 1–11. Springer. Berlin, 1989.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[BV09] O. Benedix and B. Vexler. A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints. *Comput. Optim. Appl.*, 44(1):3–25, 2009.

[Cas86] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24:1309–1318, 1986.

[CGL10] A. Caboussat, R. Glowinski, and A. Leonard. Looking for the best constant in a Sobolev inequality: a numerical approach. *Calcolo*, 47(4):211–238, 2010.

[CKR08] S. Cherednichenko, K. Krumbiegel, and A. Rösch. Error estimates for the Lavrentiev regularization of elliptic optimal control problems. *Inverse Problems*, 24(5):21 p., 2008.

[CM95] C. Chen and O. L. Mangasarian. Smoothing methods for convex inequalities and linear complementarity problems. *Math. Program.*, 71(1(A)):51–69, 1995.

[CMV13] E. Casas, M. Mateos, and B. Vexler. New regularity results and improved error estimates for optimal control problems with state constraints. *ESAIM Control Optim. Calc. Var.*, submitted, 2013.

[CQQT04] X. Chen, H. Qi, L. Qi, and K.-L. Teo. Smooth convex approximation to the maximum eigenvalue function. *J. Global Optim.*, 30(2):253–270, 2004.

[den94] D. den Hertog. *Interior point approach to linear, quadratic and convex programming: algorithms and complexity.* Kluwer Academic Publishers. Dordrecht, 1994.

[Deu11] P. Deuflhard. *Newton methods for nonlinear problems. Affine invariance and adaptive algorithms. Softcover reprint of the 2006 2nd corrected printing.* Springer Series in Computational Mathematics 35. Berlin, 2011.

[ET99] I. Ekeland and R. Témam. *Convex analysis and variational problems. Unabridged, corrected republication of the 1976 English original.* SIAM. Philadelphia, PA, 1999.

[Eva10] L. C. Evans. *Partial differential equations. 2nd ed.* Graduate Studies in Mathematics 19. AMS, Providence, RI, 2010.

[Fis08] G. Fischer. *Lineare Algebra. Eine Einführung für Studienanfänger. 16. Auflage.* Vieweg. Wiesbaden, 2008.

[FM97a] L. Faybusovich and J. B. Moore. Infinite-dimensional quadratic optimization: Interior-point methods and control applications. *Appl. Math. Optim.*, 36(1):43–66, 1997.

[FM97b] L. Faybusovich and J. B. Moore. Long-step path-following algorithm for convex quadratic programming problems in a Hilbert space. *J. Optim. Theory Appl.*, 95(3):615–635, 1997.

[GK99] C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben.* Springer. Berlin, 1999.

[GK02] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben.* Springer. Berlin, 2002.

[Gli02] F. Glineur. Improving complexity of structured convex optimization problems using self-concordant barriers. *European J. Oper. Res.*, 143(2):291–310, 2002.

[GMS92] J. R. Gilbert, C. Moler, and R. Schreiber. Sparse matrices in MATLAB: Design and implementation. *SIAM J. Matrix Anal. Appl.*, 13(1):333–356, 1992.

[GR05] C. Großmann and H.-G. Roos. *Numerische Behandlung partieller Differentialgleichungen. 3. Auflage.* Teubner. Wiesbaden, 2005.

[Gri11] P. Grisvard. *Elliptic problems in nonsmooth domains. Reprint of the 1985 hardback ed.* Classics in Applied Mathematics 69. SIAM. Philadelphia, PA, 2011.

[GT83] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order. 2nd ed.* Grundlehren der Mathematischen Wissenschaften, 224. Springer. Berlin, 1983.

[GVL07] G. Golub and C. F. Van Loan. *Matrix computations. 3rd ed.* Hindustan Book Agency. New Delhi, 2007.

[HDMRS09] R. Haller-Dintelmann, C. Meyer, J. Rehberg, and A. Schiela. Hölder continuity and optimal control for nonsmooth elliptic problems. *Appl. Math. Optim.*, 60(3):397–428, 2009.

[HK06a] M. Hintermüller and K. Kunisch. Feasible and noninterior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.

[HK06b] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.*, 17(1):159–187, 2006.

[HK09] M. Hintermüller and K. Kunisch. PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative. *SIAM J. Optim.*, 20(3):1133–1156, 2009.

[HPUU09] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints.* Mathematical Modelling: Theory and Applications 23. Springer. Berlin, 2009.

[HS10] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31(5):2291–2317, 2010.

[HSW12] M. Hintermüller, A. Schiela, and W. Wollner. The length of the primal-dual path in Moreau-Yosida-based path-following for state constrained optimal control. *Hamburger Beiträge zur Angewandten Mathematik Nr. 2012-03*, 2012.

[IK03] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems Control Lett.*, 50(3):221–228, 2003.

[Jar92] F. Jarre. Interior-point methods for convex programming. *Appl. Math. Optim.*, 26(3):287–311, 1992.

[Jar94] F. Jarre. *Interior-point methods via self-concordance or relative Lipschitz condition.* Habilitationsschrift, 1994.

[JK95] D. Jerison and C. E. Kenig. The inhomogeneous Dirichlet problem in Lipschitz domains. *J. Funct. Anal.*, 130(1):161–219, 1995.

[JS04] F. Jarre and J. Stoer. *Optimierung.* Springer. Berlin, 2004.

[Kön04a] K. Königsberger. *Analysis 1. 6., durchgesehene Auflage.* Springer. Berlin, 2004.

[Kön04b] K. Königsberger. *Analysis 2. 5., korrigierte Auflage.* Springer. Berlin, 2004.

[KR09] K. Krumbiegel and A. Rösch. A virtual control concept for state constrained optimal control problems. *Comput. Optim. Appl.*, 43(2):213–233, 2009.

[KS84] J. R. Kuttler and V. G. Sigillito. Eigenvalues of the Laplacian in two dimensions. *SIAM Rev.*, 26:163–193, 1984.

[KU13] F. Kruse and M. Ulbrich. A self-concordant interior point approach for optimal control with state constraints. *SIAM J. Optim.*, submitted, 2013.

[Liu09] Z. Liu. Polynomial complexity of primal-dual interior-point methods for convex quadratic programming with self-regular proximity. *Math. Appl.*, 22(2):326–334, 2009.

[MPT07] C. Meyer, U. Prüfert, and F. Tröltzsch. On two numerical methods for state-constrained elliptic control problems. *Optim. Methods Softw.*, 22(6):871–899, 2007.

[MRT06] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Comput. Optim. Appl.*, 33(2-3):209–228, 2006.

[MV92] R. Meise and D. Vogt. *Einführung in die Funktionalanalysis.* Vieweg Studium. 62, Aufbaukurs Mathematik. Vieweg. Braunschweig, 1992.

[Nem04] A. Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes at Georgia Institute of Technology School of Industrial and Systems Engineering*, 2004.

[NN94] Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming.* SIAM Studies in Applied Mathematics, vol. 13. Philadelphia, PA, 1994.

[Plu92] M. Plum. Explicit H_2-estimates and pointwise bounds for solutions of second-order elliptic boundary value problems. *J. Math. Anal. Appl.*, 165(1):36–61, 1992.

[PRT02a] J. Peng, C. Roos, and T. Terlaky. Self-regular functions and new search directions for linear and semidefinite optimization. *Math. Program.*, 93(1(A)):129–171, 2002.

[PRT02b] J. Peng, C. Roos, and T. Terlaky. *Self-regularity: a new paradigm for primal-dual interior-point algorithms.* Princeton University Press, Princeton, NJ, 2002.

[PS09] U. Prüfert and A. Schiela. The minimization of a maximum-norm functional subject to an elliptic PDE and state constraints. *ZAMM, Z. Angew. Math. Mech.*, 89(7):536–551, 2009.

[Ren01] J. Renegar. *A mathematical view of interior-point methods in convex optimization.* MPS/SIAM Series on Optimization. Philadelphia, PA, 2001.

[Rud87] W. Rudin. *Real and complex analysis. 3rd ed.* McGraw-Hill, New York, NY, 1987.

[Sav98] G. Savaré. Regularity results for elliptic equations in Lipschitz domains. *J. Funct. Anal.*, 152(1):176–201, 1998.

[Sch09a] A. Schiela. An extended mathematical framework for barrier methods in function space. In *Domain decomposition methods in science and engineering XVIII. Selected papers based on the presentations at the 18th international conference of domain decomposition methods, Jerusalem, Israel, January 12–17, 2008.* Springer. Berlin, 2009.

[Sch09b] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. Optim.*, 20(2):1002–1031, 2009.

[Sch12] A. Schiela. An interior point method in function space for the efficient solution of state constrained optimal control problems. *Math. Program., Ser. A*, 2012.

[SH11] A. Schiela and M. Hintermüller. On the length of the primal-dual path in Moreau-Yosida-based path-following for state constrained optimal control: Analysis and numerics. *ZIB-Report 11-37*, 2011.

[Sta09] G. Stadler. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44(2):159–181, 2009.

[SU12] A. Schiela and S. Ulbrich. Operator preconditioning for a class of constrained optimal control problems. *Preprint series of the Institute of Mathematics, Technische Universität Berlin*, Preprint 18-2012, 2012.

[TN10] L. Tunçel and A. Nemirovski. Self-concordant barriers for convex approximations of structured convex sets. *Found. Comput. Math.*, 10(5):485–525, 2010.

[Trö05] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen.* Vieweg. Wiesbaden, 2005.

[TY09a] F. Tröltzsch and I. Yousept. A regularization method for the numerical solution of elliptic boundary control problems with pointwise state constraints. *Comput. Optim. Appl.*, 42(1):43–66, 2009.

[TY09b] F. Tröltzsch and I. Yousept. Source representation strategy for optimal boundary control problems with state constraints. *Z. Anal. Anwend.*, 28(2):189–203, 2009.

[Ulb10] S. Ulbrich. Innere-Punkte-Verfahren der konvexen Optimierung. *Lecture Notes. Technische Universität Darmstadt*, 2010.

[Ulb11a] M. Ulbrich. Optimization in Banach spaces. *Lecture Notes. Technische Universität München*, 2011.

[Ulb11b] M. Ulbrich. *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces.* MOS-SIAM Series on Optimization. SIAM. Philadelphia, PA, 2011.

[UU12] M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung.* Birkhäuser. Basel, 2012.

[Wal00] W. Walter. *Gewöhnliche Differentialgleichungen. Eine Einführung. 7., neu bearb. und erweiterte Auflage.* Springer. Berlin, 2000.

[WB06] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106(1(A)):25–57, 2006.

[Wri97] S. J. Wright. *Primal-dual interior-point methods.* SIAM. Philadelphia, PA, 1997.

[Yos94] K. Yosida. *Functional analysis. Repr. of the 6th ed.* Springer. Berlin, 1994.

[Zei93] E. Zeidler. *Nonlinear functional analysis and its applications. Volume I: Fixed-point theorems. Translated from the German by Peter R. Wadsack. 2. corr. printing.* Springer. New York, NY, 1993.