Optimal Control of Quasilinear Parabolic PDEs: Theory and Numerics

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Fabian Hoppe

aus

Bamberg

Bonn 2022

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachterin: Prof. Dr. Ira Neitzel
Gutachter: Prof. Dr. Michael Hinze

Tag der Promotion: 09.12.2022
Erscheinungsjahr: 2022

Abstract

This thesis is concerned with theory and numerics of optimal control of quasilinear parabolic PDEs. The underlying parabolic PDE models, e.g., heat conduction with temperature-dependent thermal conductivity and is highly nonlinear with a nonmonotone nonlinearity in the elliptic operator. This makes its analysis and the analysis of the entire control problem as interesting as challenging. In our work we address optimal control problems with additional pointwise state-constraints and problems with sparse solutions, and analyze convergence of the SQP method. Moreover, we consider model order reduction by proper orthogonal decomposition both for the state equation and the control problem. On the one hand, our contributions can be regarded as extension of results on control-constrained optimal control of quasilinear parabolic PDEs towards the abovementioned additional aspects. In particular, difficulties associated with the nonlinear structure of our state equation are a recurrent issue in our analysis. On the other hand, we also contribute to the fields of state-constrained or sparse optimal control, the analysis of optimization algorithms, and model order reduction by extending them towards quasilinear parabolic PDEs. Consequently, we also encounter the typical challenges due to these respective areas, especially such challenges related to optimality conditions in infinite dimensions.

In Chapter 2 we start our analysis by considering problems with additional pointwise state-constraints. Under appropriate regularity assumptions on the domain, the boundary conditions, and the coefficient functions of the equation we prove first-order necessary and second-order sufficient optimality conditions for pointwise in space and time and pointwise in space and averaged in time state-constraints. Besides typical difficulties associated with the second-order analysis of optimization problems in infinite dimensions, the high regularity requirements coming along with state-constrained problems pose a particular difficulty. As a consequence, we need to perform a detailed regularity analysis of our highly nonlinear state equation and its linearization, the latter requiring careful estimation of the derivatives of the nonlinearity of the state equation. Hereby, the presence of differential operators in the derivatives of the nonlinearity is a particular difference compared to the case of semilinear equations and poses a new difficulty that is specific for quasilinear problems.

Subsequently, we come back to a control-constrained problem in Chapter 3 but now with additional cost terms in the objective functional that enforce so-called

sparse solutions. Based on available results from the literature concerned with semilinear parabolic problems we prove first- and second-order optimality conditions for this problem type. Besides the already mentioned issues specifically related to the quasilinear state equation and the general intricacy of second-order optimality conditions in infinite-dimensional spaces, an additional difficulty now arises from the fact that the sparsity-enforcing cost terms are convex, but nonsmooth. Special emphasis is payed to the practically relevant case of purely time-dependent controls for which we analyze the sparsity patterns resulting from seven different cost terms.

In Chapter 4 we analyze convergence in function space of the SQP method for a control-constrained model problem. Here, we can build again on results from the literature on semilinear parabolic problems. However, since the properties of quasilinear and semilinear equations are quite different, it is not clear a priori that this transfer of techniques works in the end. In particular, we have to restrict ourselves to the so-called purely time-dependent control setting and need, again, to carry out a careful regularity analysis of the appearing equations. Moreover, we put special emphasis on the interplay of second-order sufficient optimality conditions and the resulting restriction of the SQP subproblems onto certain neighbourhoods of the optimal control. Some progress has been made in the theory of such second-order optimality conditions after the existing results on convergence of the SQP method for semilinear parabolic problems have been published. Therefore, revisiting the converegence theory of the SQP method, and in particular the localization of the subproblems, in the light of new techniques and results on second-order conditions is of particular interest to us.

The topic of the final Chapter 5 is model order reduction by proper orthogonal decomposition (POD). First, we apply this technique to the state equation and observe that this allows for a heavy reduction in the computational efforts related to the numerical solution of the state equation. Moreover, we prove a-posteriori estimates for the error due to this POD-reduction. As observed already in the literature on model order reduction for other nonlinear equations, the main difficulty is to combine the projection-based, and hence linear, POD-model order reduction with a highly nonlinear equation. A particular difference compared to previous results on a-posteriori POD error estimates for, e.g., semilinear or quasilinear equations is the presence of a nonmonotone nonlinearity that makes the analysis more difficult. Finally, we demostrate by numerical examples that application of POD-reduction also allows to speed up the numerical solution of the entire control problem significantly.

Acknowledgements

First, I want to express my gratitude to my PhD advisor, Prof. Dr. Ira Neitzel, for giving me the opportunity to work on this interesting and challenging topic in a friendly and relaxed working atmosphere and for support in mathematical and formal questions. Her lecture "Numerical Simulation" in the summer term 2016 also convinced me to turn my scientific interest towards the field of numerical mathematics and optimal control. Of course, I also want to thank Prof. Dr. Michael Hinze (University of Koblenz-Landau) for being the co-referee of this thesis.

For the ongoing colaboration on two further questions related to optimal control of quasilinear parabolic PDEs I want to thank Lucas Bonifacius und Prof. Dr. Hannes Meinlschmidt (University of Erlangen). My collegues at the Institute of Numerical Simulation of the University of Bonn I want thank for a great time at the institute with many interesting and enjoyable, mathematical and non-mathematical conversations.

I gratefully acknowledge the financial support by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) that partially funded my position and several business trips within the Collaborative Research Center 1060 "The Mathematics of Emergent Effecs" (project no. 211504053, from 2021 within the subproject C10 "Sparse controls in optimization of quasilinear partial differential equations").

Last, but not least, I want to say thankyou to my parents, to my aunt, and to my friends: you do not know much about (applied) mathematics, but yet this work would not have been possible without your support!

Bonn, June 2022,

F.H.

Contents

Abstract iii

Acknowledgements v

Introduction 1

Chapter 0. Notation and definitions 13

Part I. Analysis of the control problems 19

Chapter 1. Preliminaries: the control-constrained problem 21

Chapter 2. Pointwise constraints on the state 43

Chapter 3. Sparse purely time-dependent optimal control 83

Part II. Algorithms and model order reduction 117

Chapter 4. Convergence of the SQP method 119

Chapter 5. Model order reduction by proper orthogonal decomposition 157

Conclusion and outlook 195

Bibliography 199

Introduction

> Eine einzelwissenschaftliche, etwa eine physikalische Untersuchung kann ohne weitere Umschweife mit der Bearbeitung ihres Problems beginnen. Sie kann, sozusagen, mit der Tür ins Haus fallen; es ist ja ein "Haus" da: ein wissenschaftliches Lehrgebäude, eine allgemein anerkannte Problemsituation. Der Forscher kann es deshalb auch dem Leser überlassen, die Arbeit in den Zusammenhang der Wissenschaft einzuordnen.
>
> K. Popper, *Logik der Forschung*[1]

The overall topic of this thesis is optimal control of quasilinear parabolic partial differential equations (PDEs). It seems to be appropriate to begin with a short explanation of these termini technici, before we give more details. The generic formulation of an *optimal control problem* reads as follows; see, e.g., [**201, 156, 270**]: given Banach spaces Y, U, Z, subsets $U_{\mathrm{ad}} \subset U$, $Y_{\mathrm{ad}} \subset Y$ and maps $J\colon Y \times U \to \mathbb{R}$, $E\colon Y \times U \to Z$, solve

$$\text{(OCP)} \qquad \begin{cases} \min\limits_{(y,u)\in Y\times U} J(y,u) \\ \quad \text{subject to} \quad u \in U_{\mathrm{ad}}, \\ \qquad\qquad\quad\; y \in Y_{\mathrm{ad}}, \\ \quad \text{and} \quad E(y,u) = 0. \end{cases}$$

We call Y and U state- and control-space, $Y_{\mathrm{ad}} \subset Y$ and $U_{\mathrm{ad}} \subset U$ the sets of admissible state and controls, respectively, J the objective functional, and "$E(y,u) = 0$" the state equation. We will refer to elements $y \in Y$ and $u \in U$ as states and controls, respectively. To make (OCP) a control problem —so far we have only described a constrained optimization problem— we need that each control $u \in U$ uniquely determines a state $y = y(u) \in Y$ (which we will call the state associated with u) such that $E(y(u), u) = 0$. In other words: there is a map $S\colon U \to Y$, the so-called control-to-state map, such that

$$E(S(u), u) = 0, \qquad \text{and} \quad z \in Y,\ E(z,u) = 0 \quad \Leftrightarrow \quad z = S(u).$$

[1] Preface to the first german edition 1934, quoted from: K. R. Popper, Gesammelte Werke in deutscher Sprache, Band 3, Mohr Siebeck, Tübingen, 2005.

The property that the state is uniquely determined by the control, distinguishes the class of optimal control problems from the class of constrained optimization problems. In particular, by eliminating the state-variable we can reformulate (OCP) equivalently as follows:

$$(\mathrm{OCP}_{\mathrm{red}}) \qquad \begin{cases} \min\limits_{u \in U} \; j(u) := J(S(u), u) \\ \qquad\qquad \text{subject to} \quad u \in U_{\mathrm{ad}}, \\ \qquad\qquad\qquad\quad S(u) \in Y_{\mathrm{ad}}. \end{cases}$$

We call $(\mathrm{OCP}_{\mathrm{red}})$ the reduced form of problem (OCP) and refer to j as the reduced functional. Now, a *quasilinear parabolic optimal control problem* simply is an optimal control problem in which the state equation "$E(y, u) = 0$" involves a quasilinear parabolic PDE. Of course, Y and Z have to be appropriate function spaces in that case. In general, if the state equation of an optimal control problem consists of a PDE, we may also speak of "PDE-constrained optimization" which will be used as synonym for "optimal control of PDEs" in the following. Of course, different types of optimal control problems are also known from the literature. We mention, e.g., optimal control of ordinary differential equations (ODEs) or differential-algebraic equations [**112**] or recent research on optimal control of variational inequalities for which we cite exemplarily [**85**]. Nevertheless, for the reason of shortness, the notion "optimal control (problem)" always refers to optimal control problems with PDEs throughout this thesis. Sometimes, we just write "control (problem)" and still refer to optimal control since control problems in the control- or system-theoretic sense, cf., e.g., [**259**], will not be addressed in the present work.

It remains to make precise the notion of a *quasilinear parabolic PDE*. In order to avoid to get bogged down in lengthy definitions of excessively technical nature, we do not explain the term "quasilinear" in its full generality. Instead, we just mention that a quasilinear PDE is roughly characterized by the fact that its nonlinearity appears in the highest order term —in our case in the elliptic operator of the parabolic PDE. Note that this distinguishes this class of PDEs from, e.g., semilinear PDEs, in which nonlinearities are only allowed in the lower order terms. One may therefore imagine a scale of PDE-classes in ascending order of distance to the linear class —and hence, in general, also ascending difficulty— as follows:

$$\text{linear} \quad \longrightarrow \quad \text{semilinear} \quad \longrightarrow \quad \text{quasilinear} \quad \longrightarrow \quad \text{fully nonlinear}.$$

For the precise form of the state equation under consideration in this thesis, however, we refer the reader to Chapter 1. Let us just mention that the nonlinearity considered in the present work appears, e.g., in the modelling of heat conduction in an object with temperature-dependent thermal conductivity. In practice, this is of interest in, e.g., the induction heating of steel [**214, 215**] or the modelling of semiconductor devices [**253**, Chapter 4.3]. In general, the analysis of highly nonlinear equations and related optimal control problems as in this thesis can be motivated by the following heuristic rule of thumb: the more realistic and accurate the underlying physical model, the more nonlinear, and the more difficult the resulting PDE.

We are now able to summarize the goals of this thesis with a bit more details as follows: in [**35, 45**], which will serve as a starting point of our considerations,

a quasilinear parabolic optimal control problem with pure control-constraints, i.e. $Y_{\mathrm{ad}} = Y$, has been discussed. In this thesis we present and discuss extensions hereof from [**167, 166, 168, 169**] recently obtained by the author in joint work with I. Neitzel. More precisely, in Part I we modify the two componenents of the optimal control problem, i.e. additional constraints and the objective functional: first, we add certain pointwise constraints on the state and analyze the resulting problems. Second, we add nonsmooth penalization terms to the objective functional that ensure so-called sparsity of the optimal controls. Afterwards, in Part II, we come back to the original problem from [**35**] and focus on the numerical solution hereof. We prove convergence of the SQP method in function spaces, and we investigate how model order reduction can help to reduce the computational effort related to solving the state equation or the control problem. Consequently, our contributions can always be seen from two perspectives. On the one hand, we extend the results on control-constrained optimal control of quasilinear parabolic PDEs from [**35, 45**] towards the abovementioned additional aspects. This is the read thread of this thesis because —besides the challenges specific to the abovementioned aspects— the ubiquituous difficulty in our work is posed by the highly nonlinear structure of the underlying quasilinear parabolic PDE. On the other hand, of course, we also build on results and techniques from the fields of state-constrained or sparse optimal control, as well as the analysis of optimization algorithms and model order reduction within PDE-constrained optimization. Consequently, our results have to be regarded as contributions to these areas of research, too.

Since we have provided the reader at least a first and rough impression of our topic for the moment, let us start with concise literature overview in order to put our work into its context. A more detailed literature overview on the respective most related aspects will be given at the beginning of each chapter of this thesis. Optimal control of PDEs has become a flourishing and broad area of research in the last decades; see, e.g., the early monography [**201**] or the more recent textbooks [**156, 270, 90**]. Among the almost countless number of real-world applications in different areas we cite exemplarily induction heating of metals [**214, 215**], optimal cooling of glass [**156**, Chapter 4.1] or steel [**278**], multiphase steel production [**161**], optimization of semiconductor devices [**156**, Chapter 4.2] or crystal growth [**218**], optimal planning of hypothermia treatment in cancer therapies [**95**], inverse problems in fincance [**252**] and geophysics [**33**], and last, but not least, optimization in oenology [**217**]. Also on the mathematical side, the contributions are very diverse. They extend from rather theoretical considerations with strong functional analytic flavour up to highly application-oriented results with computational focus. Typical problems to be considered are, e.g., well-posedness of the problem, i.e. existence of optimal controls, the derivation of first- and second-order optimality conditions, discretization of problems and corresponding a-priori discretization error estimates, algorithms and their convergence analysis both on the function space and discrete level, and, finally, analysis and efficient implementation of computational procedures, such as, e.g., regularization of additional constraints, adaptive discretization and optimization strategies, preconditioning, parallel computing, or model order reduction. After this short glimpse on the area in its entirety, we will focus in the following on some particular aspects that are closely related to the present work:

quasilinear (or more general: nonlinear) state equations, additional constraints on the state-variable, sparsity-enforcing functionals, SQP methods, and model order reduction.

We begin our overview with literature on quasilinear problems. In the recent years substantial progress has been made in this topic. We mention in particular the papers [**216, 214, 215, 35, 45, 167, 166, 168, 169**] and [**68, 50, 69, 70, 88, 87, 89, 164**] concerned with parabolic and elliptic problems, respectively. For earlier results we refer the reader to the introduction of [**35**] for instance. In [**216**] the quasilinear parabolic equation

$$\begin{cases} \partial_t y - \nabla \cdot \xi(y)\mu\nabla y + y = \mathcal{F}(t,y), & \text{in } L^s(I, W_{\Gamma_D}^{-1,p}), \\ y = 0, & \text{on } (0,T)\times\Gamma_D, \\ y(0) = y_0, & \text{on } \Omega, \end{cases} \tag{0.1}$$

has been analyzed in a $W_{\Gamma_D}^{-1,p}$-$W_{\Gamma_D}^{1,p}$-setting with a bounded, zero-order semilinear term $\mathcal{F}$. Moreover, existence of solutions to a respective optimal control problem has been proven. For an optimal control problem governed by a slightly different equation, without id-term and semilinear term, first- and second-order optimality conditions were obtained in [**35**]. Further, an improved regularity analysis of the underlying equation on certain Bessel potential spaces $H_D^{-\zeta,p}$ has been provided. Existence of optimal controls and first-order optimality conditions for optimal control of the so-called thermistor problem, a coupled system of a quasilinear parabolic and a quasilinear elliptic PDE, have been derived in [**214, 215**]. Optimal control of a quasilinear parabolic system with a different structure, the so-called chemotaxis system, is considered in [**108**]. The papers [**216, 35, 214, 215**] deal with assumptions on the underlying problem data that are often referred to as "rough" setting: they allow, e.g., for nonsmooth domains and coefficient functions and mixed boundary conditions, as they often arise in real-world constellations; see, e.g., [**97, 98**]. We also mention that first- and second-order optimality conditions for a problem similar to the one from [**35**] have been obtained in [**45**] for a slightly different setting that is more regular w.r.t. domain and boundary conditions, but allows, e.g., for unbounded coefficients and a mononote semilinear term of order zero. Finite element discretization error estimates for the corresponding state equation have been established in [**46**]. In joint work with I. Neitzel the author has addressed quasilinear parabolic problems with additional state-constraints [**168**], sparse purely time-dependent optimal control [**169**] and the convergence analysis of the SQP method [**167**]. The respective results form the content of Chapters 2 to 4 of this thesis. Hereby, it is both one of the goals and one of the challenges to perform as much of the analysis as possible within the rough regularity setting of [**216, 35**], i.e. under minimal assumptions; only if the specific problem type under discussion cannot be handled in a satisfactory manor within this setting, we finally change to a smoother setup based on [**45**].

Optimal control problems with quasilinear parabolic PDEs share some typical difficulties with optimal control problems governed by other nonlinear PDEs. In order to derive first-order necessary optimality conditions, differentiability of the control-to-state map needs to be addressed. The latter requires differentiability of the nonlinearity in the state equation, which itself forces us to consider states in

sufficiently regular function spaces on which the respective superposition operators are differentiable. Usually, obtaining L^∞-regularity of solutions of the state equation is necessary. The latter is certainly nontrivial, in particular for time-dependent problems and in a rough setting, and requires a careful regularity analysis. In [**35**] this analysis relies on [**216**] and the functional analytic concept of maximal parabolic regularity for nonautonomous operators; see [**15**]. Since first-order necessary optimality conditions are, in general, not sufficient for optimality in the case of nonconvex problems, the derivation of second-order sufficient conditions plays an important role in optimal control of nonlinear PDEs. Moreover, such conditions often serve as starting point for the investigation of discretization error estimates or the convergence analysis of numerical algorithms; for the latter see, e.g., Chapter 4 of this thesis. For an overwiew covering different aspects of the topic we refer the reader to, e.g., the recent survey [**73**] on second-order optimality conditions in PDE-constrained optimization. Let us just mention the following prominent example from [**71**, Example 1.2] that highlights the difficulties associated with optimality conditions in infinite dimensions. Consider the following optimization problem without additional constraints:

$$\min_{u\in L^2(0,1)} \; j(u) := \frac{1}{2}\int_0^1 \sin(u(x))\mathrm{d}x. \tag{0.2}$$

Clearly, $\bar u \equiv -\frac{\pi}{2}$ is a global solution of (0.2) and purely formal computations yield

$$j'(\bar u)v = 0, \qquad j''(\bar u) = \|v\|^2_{L^2(0,1)}, \qquad \forall v \in L^2(0,1).$$

One might be tempted to conclude from $j'(\bar u) = 0$ and coercivity of $j''(\bar u)$ that $\bar u$ is a strict local solution of (0.2), i.e. that it holds $j(u) > j(\bar u)$ for all $u \neq \bar u$ sufficiently close to $\bar u$. Such a conclusion would be correct in finite dimensions, but for the present infinite-dimensional example it is false, at least if "local" is meant w.r.t. the $L^2(0,1)$-topology: for any $\epsilon > 0$ we can define $u_\epsilon = -\frac{\pi}{2}\mathbf{1}_{(0,1-\epsilon)} + \frac{3\pi}{2}\mathbf{1}_{(1-\epsilon,1)}$ and observe

$$\|u_\epsilon - \bar u\|_{L^2(0,1)} = 2\pi\sqrt{\epsilon} \qquad \text{and} \qquad j(u_\epsilon) = j(\bar u).$$

The reason for this unexpected behaviour is that the purely formal computations that lead to $j'(\bar u)$ and $j''(\bar u)$ were not correct in a precise mathematical sense. In fact, the functional j is not differentiable as map $L^2(0,1) \to \mathbb{R}$. Regarding j as a map $L^\infty(0,1) \to \mathbb{R}$, however, we can prove twice Fréchet differentiability with the above given formulas for the derivatives; but then $j''(\bar u)$ is no longer coercive w.r.t. the underlying norm, i.e. the $L^\infty(0,1)$-norm. This so-called two-norm discrepancy —differentiability of the functional and coercivity of its second derivative only hold w.r.t. to different, nonequivalent norms— goes back to [**174**] and is typical for optimization problems in inifinite-dimensional function spaces, in particular for PDE-constrained optimization [**71**]. In the case of (0.2) we are only able to draw the following conclusion, a so-called quadratic growth condition with two-norm gap: there are $\delta, \rho > 0$ such that

$$j(u) \geq j(\bar u) + \frac{\rho}{2}\|u - \bar u\|^2_{L^2(0,1)} \qquad \forall u \in L^2(0,1) \text{ s.t. } \|u - \bar u\|_{L^\infty(0,1)} < \delta;$$

cf. [**71**, Theorem 1.3]. Only in certain situations, see, e.g., [**71**], such a two-norm gap can be avoided when addressing second-order optimality conditions in infinite dimensions in the presence of a two-norm discrepancy. This is the case, e.g.,

in both [**35**] and [**45**] where second-order necessary (SNCs) and sufficient (SSCs) optimality conditions for quasilinear parabolic optimal control problems without norm gap have been achieved despite the presence of the two-norm discrepancy. In any case, when proving second-order optimality conditions for optimal control problems with nonlinear state equation the choice of the appearing function spaces and a careful regularity analysis of the state equation and its linearizations become highly important.

A topic that gained some attention in the last years is optimal control of PDEs with additional constraints on the state. In an application, e.g., in which one has to control the temperature of some object, it may be necessary to keep the temperature always below the melting point in order to avoid damage. Mathematically, this or similar situations can be formulated as pointwise inequality constraints on the state ("state-constraints"). Among the applications of PDE-constrained optimization mentioned before, e.g., the problems related to induction heating of metal [**214, 215**], crystal growth [**218**], and cancer therapy [**95**] make use of this concept. Problems with such additional state-constraints are particularly interesting, but their analysis is challenging. We only mention, e.g., the early contributions [**40, 41**] dealing with linear and semilinear elliptic equations and state-constraints, and refer the reader to the introduction of Chapter 2 for a more detailed literature overview. In our paper [**168**] we have addressed this challenging problem type in the case of quasilinear parabolic PDEs; the results hereof are presented in Chapter 2 of this thesis. To illustrate one of the main difficulties associated with state-constraints, let us briefly recall the following classical result on first-order optimality conditions in infinite dimensions.

Theorem 0.1 (KKT conditions in Banach spaces, [**41**], Theorem 5.2)**.** *Let U, Z be Banach spaces, and $K \subset U$, $C \subset Z$ convex sets such that C has nonempty interior. Let $\bar{u} \in K$ be a solution of the problem*

$$\min_{u \in K} \; j(u) \qquad s.t. \quad g(u) \in C,$$

where $j\colon U \to \mathbb{R}$ and $g\colon U \to Z$ are Gâteaux differentiable at $\bar{u}$. Moreover, assume that the following linearized Slater condition is satisfied at $\bar{u}$: there is $u_{\mathrm{Sl}} \in K$ such that

$$g(\bar{u}) + g'(\bar{u})(u_{\mathrm{Sl}} - \bar{u}) \in \overset{\circ}{C}.$$

Then, there exists a Lagrange multiplier $\bar{\mu} \in Z^*$ such that the following optimality system holds true:

$$\langle \bar{\mu}, z - g(\bar{u}) \rangle_{Z^*,Z} \leq 0 \qquad \forall z \in C, \tag{0.3}$$

$$\langle \bar{\lambda} j'(\bar{u}) + g'(\bar{u})^* \bar{\mu}, u - \bar{u} \rangle_{U^*,U} \geq 0 \qquad \forall u \in K. \tag{0.4}$$

The vast majority of results on first-order conditions in PDE-constrained optimization with state-constraints is based on this theorem or similar results. Relying on a typical Slater-type constraint qualification, however, is known to infer substantial difficulties: in order to prove first-order conditions for a problem with pointwise state-constraints in Chapter 2, we will apply Theorem 0.1 with $K = U_{\mathrm{ad}}$, $C = Y_{\mathrm{ad}}$, j being the reduced functional of our problem, and g being the control-to-state map S. The difficulty arises from the fact that one requires Y_{ad} to have nonempty interior in $Z = Y$ in order that the Slater-type constraint qualification

can be fulfilled. If Y_{ad} is given by pointwise inequality constraints, this excludes to choose an L^p-space with $p < \infty$ for the state space Y. Instead, Y is typically chosen to be a space of continuous functions, which in return usually results in the presence of regular Borel measures, i.e. the corresponding dual objects, as Lagrange multiplier $\bar{\mu} \in Z^* = Y^*$ in the KKT system. In particular, $g'(\bar{u})^* \bar{\mu}$ will be related to the solution of a certain linear PDE with a right-hand side containing the measure $\bar{\mu}$, in general. Having in mind the problems coming along even with the "simple" problem (0.2) without additional constraints, it is not surprising that second-order optimality conditions for state-constrained optimal control problems are a challenging topic; see, e.g., [**73**, section 7.2] for a concise overview. In essence, a deeper regularity analysis of the underlying PDE and its linearizations than in the presence of control-constraints only is needed. Of course, this is particulary challenging in the present case of a quasilinear, i.e. highly nonlinear, state equation and rough regularity assumptions. In fact, the regularity results required to address second-order conditions for pointwise in space and time state-constraints cannot be achieved in the rough setting. Hence, we will either have to enforce our regularity assumptions on the state equation in the flavour of [**45**] in order to meet the regularity requirements, or we have to modify the type of state-constraints towards averaged-type state-constraints in order to relax the regularity requirements towards such that can be achieved in the rough setting.

Another well-known field in PDE-constrained optimization is so-called sparse optimal control. Let us first recall the following motivating example related to the design of piezoelectric plates mentioned in the introduction of [**261**]: in order to achieve a prescribed desired displacement of the plate, electrodes need to be positioned suitably on this plate. Hereby, it is crucial to find —in some sense— the optimal location of these electrodes. It is clear that also in other applications it may be necessary, or at least desirable, to determine controls that are nonzero only on a small ("sparse") part of the (space-time-)domain under consideration. For classical L^2-tracking-type functionals with L^2-Tikhonov regularization for the control-variable we cannot expect such a property to hold for the optimal control, in general, due to the structure of the optimality system. Therefore, sparsity of the optimal control has to be enforced by modifying the control problem. In our paper [**169**] we follow a popular approach to do so: a multiple of the L^1-norm, or a mixed L^1-L^2- or L^2-L^1-norm of the control is added to the objective functional as additional cost/penalization term. We will present the corresponding results in Chapter 3. The usage of L^1-penalization goes back to [**261**] which is the first contribution on sparsity in the context of PDE-constrained optimization. Penalization with mixed norms has been introduced in [**138**] and results in so-called directional sparsity, i.e. different sparsity patterns for different spatial/temporal directions. We refer the reader to the recent survey [**44**] or the introduction of Chapter 3 for a detailed literature overview. In Chapter 3 we will apply directional sparsity to the so-called purely time-dependent control setting originally introduced in [**91**]: there are $m \in \mathbb{N}$ fixed actuators and the control is a function $(0,T) \to \mathbb{R}^m$ that encodes the intensity of each actuator depending on time. For instance, when considering optimal cooling of a steel profile by water, cf. [**91**], the actuators are given by a finite number of nozzles spraying water on the profile and the controls are their time-dependent intensities. The combination of directional sparsity and purely

time-dependent controls allows to enforce seven different, interesting sparsity patterns for the optimal controls. For instance, one can select from a given, large set of possible actuators only a small number of actuators that are allowed to become active at every time point while all other actuators are not used at all, or one can select a sparse subset of time points at which all given actuators are allowed to become active while they have to be inactive at all other time points. The typical difficulty related to sparse optimal control arises from the fact that any method to enforce sparsity infers some kind of nonsmoothness. The L^1-norm or mixed norms used for penalization are convex and Lipschitz continuous, but nonsmooth functionals. Of course, this infers additional difficulties in the derivation of optimality conditions as well as in the numerical solution of such problems. The above outlined challenges related to second-order optimality conditions in infinite dimensions, e.g., are now enriched by the question how to deal with the nonsmooth part of the functional that does not even have a second derivative at all. For problems with additional L^2-Tikhonov regularization we are able to handle these problems in a similar way as done in [**54**] for semilinear parabolic problems. The fact that without L^2-Tikhonov regularization we are not able to prove second-order conditions, illustrates that the transfer of techniques successfully applied to semilinear problems to quasilinear ones is not straightforward. Regarding the numerical solution of sparse optimal control problems, we finally note that so-called proximal algorithms [**250**] are easy to implement, relatively fast, state of the art solvers for problems with L^1-penalization. The application to L^1-L^2-penalized problems without control-constraints is straightforward, but —to our best knowledge— a generalization to L^2-L^1-penalization has not been done so far. Therefore, in addition to those cost terms discussed in [**54**], we analyze a slightly modified alternative to L^2-L^1-penalization that allows the application of proximal methods and that has already been applied successfully in a finite dimensional setting in the context of data science.

The issues addressed so far are of more or less theoretical nature. In the following we continue with some more computational aspects. Although this will not be addressed explicitely in the present thesis, we start with the important topic of discretization of optimal control problems. Herein we will restrict ourselves completely to literature on discretization by the finite element method (FEM). In general, discretization techniques can be divided into approaches where an infinite dimensional optimality system is discretized ("optimize then discretize") and approaches where the whole optimization problem is discretized ("discretize then optimize"); see, e.g., [**156**, Chapter 3]. We just note that under certain conditions both approaches yield the same result and focus our brief literature overview on the second approach. Hence, one has to consider discretization of the state-variable and the state equation, and discretization of the control-variable. Unlike in the "optimize then discretize"-approach, the discrete counterpart of the adjoint equation is now determined by the optimality system of the discrete optimal control problem. For a survey on the main techniques in the case of FEM discretization error estimates for elliptic problems we refer the reader to, e.g., [**159**]. For parabolic problems we exemplarily mention discretization of the state-variable and equation by discontinuous Galerkin in time and continuous Galerkin FEM discretization in space for

linear [**212, 213**] and semilinear problems [**224**], or the Crank-Nicolson scheme for time discretization [**284**]. For the discretization of the control-variable there are several well-established concepts, e.g., variational discretization [**149, 284**], piecewise linear [**237, 238**], and piecewise constant [**17**] control discretization. The last one can be combined with a so-called post-processing step [**219**] to improve the order of convergence. For respective a-priori FEM discretization error-estimates we refer the reader to, e.g., [**212, 213**] for linear quadratic parabolic optimal control, or to [**224**] for a semilinear parabolic state equation. As far as we know, there are yet no results on discretization error estimates for quasilinear parabolic optimal control problems. A-priori FEM error estimates for the quasilinear parabolic state equation from [**45**], however, have recently been derived in [**46**]. Time discretization of an optimal control problem governed by a quasilinear parabolic PDE with a different, monotone nonlinearity is addressed in [**32**]. Moreover, we exemplarily mention the earlier results [**101, 99**] on discretization error estimates for quasilinear parabolic PDEs in different settings. A-priori error estimates for the FEM discretization of smooth quasilinear elliptic control problems have been considered in [**69, 70**], for instance, and error estimates for a problem involving a nonsmooth quasilinear elliptic PDE have been obtained recently in [**89**].

Let us now comment on the two computational aspects of PDE-constrained optimization that will be addressed in this thesis: the analysis of SQP methods and the application of model order reduction. Sequential quadratic programming (SQP) methods form a prominent class of state of the art algorithms for the efficient numerical solution of nonlinear optimal control problems; see, e.g., [**156**, Chapter 2.6] for an overview. In essence, the nonlinear optimization problem is approximated by a sequence of linear quadratic subproblems that can be solved, e.g., by application of the well-understood primal dual active set strategy. More precisely, given current approximations y_k, u_k, p_k for the state, control, and adjoint state, the next iterates y_{k+1}, u_{k+1}, p_{k+1} are obtained as state, control, and adjoint state of the problem

$$
\left\{
\begin{aligned}
\min_{y_{k+1},u_{k+1}} \quad & \frac{1}{2}\mathcal{L}''(y_k,u_k,p_k)[(y_{k+1}-y_k,u_{k+1}-u_k)]^2 \\
& \qquad + J'(y_k,u_k)(y_{k+1}-y_k,u_{k+1}-u_k) \\
\text{s.t.} \quad & 0 = E(y_k,u_k) + E_y(y_k,u_k)(y_{k+1}-y_k) \\
& \qquad + E_u(y_k,u_k)(u_{k+1}-u_k), \\
& u_{k+1} \in U_{\mathrm{ad}},
\end{aligned}
\right. \tag{0.5}
$$

where $\mathcal{L}''$ denotes the second derivative of the so-called Lagrangian $\mathcal{L}(y,u,p) = J(y,u) - \langle p, E(y,u)\rangle_{Z^*,Z}$ associated with (OCP). The first convergence analysis of this algorithm in function space for optimal control of a semilinear parabolic equation has been carried out in [**114, 268**]; see the introduction of Chapter 4 for further literature. In our paper [**167**] we have extended this towards the case of a quasilinear parabolic state equation. Our respective results form the content of Chapter 4 of this thesis. Revisiting the convergence analysis for our quasilinear parabolic problem is certainly already of interest, just because quasilinear equations are quite different from semilinear ones in several aspects and hence the adaptation of results known for semilinear PDEs to quasilinear ones is by no means straightforward. Let us briefly explain a second motivation. Besides the correct choice of

function spaces and the analysis of the underlying equations, a typical issue in the analysis of SQP methods in infinite dimensions is that a localization of the linear quadratic subproblems is necessary in order to guarantee their well-definedness. Of course, one will be tempted to keep this artificial restriction as low as possible. It is well-known that convexity of (0.5) for y_k, u_k, p_k in a neighbourhood the solution of the original control problem is closely related to coercivity of the second derivative of the reduced functional at the optimal control; this closes the circle to the topic of second-order optimality conditions mentioned before. In the abovementioned references concerned with the convergence of the SQP method second-order conditions with two-norm gap have been used as a starting point. Since for our quasilinear parabolic model problem second-order conditions without two-norm gap are available, the question naturally arises how this influences the localization of the SQP subproblems. Hence, we pay particular attention to this technical, but interesting problem during our analysis in Chapter 4.

We conclude our overview with the aspect model order reduction. The numerical solution of optimal control problems governed by time-dependent PDEs is expensive because their discretization leads to large-scale optimization problems. A possibility to reduce the computational effort is model order reduction (MOR). The aim of MOR is to replace the high-dimensional original model by a suitable model with less degrees of freedom, the so-called reduced-order model. Typical approaches are, e.g., proper orthogonal decomposition (POD) [**121, 131**] or reduced basis (RB) methods [**139**], the first being a particular instance of the second technique. Both techniques are so-called data-driven, i.e. the approximation properties of the reduced-order model depend on the data, the so-called snapshots, that have been used for its generation. Consequently, these snapshots have to be chosen carefully to ensure that the resulting reduced-order model is able to capture the essential properties of the underlying problem. This means that, in general, there are no a-priori POD/RB error estimates of practical relevance. Instead, a-posteriori error estimates are needed in order to assess the quality of a POD/RB-reduced model. In the context of quasilinear parabolic optimal control problems, POD/RB-MOR poses two main difficulties that we will be confronted with in Chapter 5. First, the POD/RB approach is projection-based, and hence of linear nature which makes the treatment of nonlinear problems particularly challenging. In order to incorporate nonlinearities efficiently into the reduced-order model so-called hyperreduction techniques have to be utilized and also in the derivation of a-posteriori error estimates for the POD/RB error the presence of nonlinearities infers substantial difficulties. In our case, the nonmonotone character of our nonlinearity is a particular challenge when deriving a-posteriori POD/RB error estimates. The respective results have been published in our paper [**166**]; we will present them in Section 5.2.2 of this thesis. Second, the fact that POD/RB are data-driven techniques implies that in applications related to PDE-constrained optimization it is important to couple MOR and numerical optimization accordingly. This is a highly active area of research; cf., e.g., [**272, 38, 232, 236, 21, 183**]. Typically, some kind of a-posteriori estimates for the POD/RB error of the whole control problem are required. The latter, however, may not be available for highly nonlinear control problems. In our case, e.g., we have at hand error estimates for

the state equation, but not for the adjoint equation. Therefore, we present numerical experiments with a heuristic, alternative coupling based on an SQP-type method that avoids the usage of such estimates. Nevertheless, we still achieve a considerable amount of reduction of the computing time.

Let us now give a short overview over the structure of this thesis. In Chapter 0 we introduce our notation and conventions and summarize some basic definitions. In Chapter 1 we state some assumptions that we will rely on in the following and discuss results on control-constrained optimal control of quasilinear parabolic PDEs from [**35, 216, 45**]. They will serve as a fundament of our further analysis. For later reference we also include a summary of important results on maximal parabolic regularity. Chapters 2 to 5 contain the main contributions of this thesis. Most of the respective results have already been published in the authors publications [**167, 166, 168, 169**] (joint work with I. Neitzel).

- **Chapter 2: Pointwise constraints on the state** (based on [**168**]) — We discuss state-constrained optimal control of a quasilinear parabolic PDE. Existence of optimal controls and first-order necessary optimality conditions are derived for a rather general, rough setting including pointwise in space and time constraints on the state. Second-order sufficient optimality conditions are obtained for averaged in time and pointwise in space state-constraints under general regularity assumptions for the equation, and for pointwise in space and time state-constraints when restricting in return to a more regular setting for the state equation.
- **Chapter 3: Sparse purely time-dependent optimal control** (based on [**169**]) — We prove first- and second-order optimality conditions for sparse purely time-dependent optimal control problems governed by a quasilinear parabolic PDE. In particular, we analyze sparsity patterns of the optimal controls induced by different sparsity-enforcing functionals in the purely time-dependent control case and illustrate them by numerical examples. Moreover, we obtain second-order necessary and sufficient optimality conditions. Our findings are based on results obtained by abstraction of well-known techniques from the literature.
- **Chapter 4: Convergence of the SQP method** (based on [**167**]) — We discuss the SQP method for the numerical solution of a quasilinear parabolic optimal control problem with purely time-dependent controls and control-constraints. Following well-known techniques, convergence of the method in appropriate function spaces is proven under some common technical restrictions. Particular attention is payed to how the second-order sufficient conditions for the optimal control problem and the resulting L^2-local quadratic growth condition influence the notion of locality in the SQP method. Numerical examples illustrate the theoretical results.
- **Chapter 5: Model order reduction by proper orthogonal decomposition** — This chapter addresses two aspects. First, we discuss POD/RB-MOR for our quasilinear parabolic state equation and prove corresponding a-posteriori error estimates. We consider the solution of a semidiscrete counterpart of the state equation as reference, and therefore incorporate POD/RB, empirical interpolation, and time discretization errors in our

consideration. This is joint work with I. Neitzel and has been published in [**166**]. Second, we demonstrate the ability of POD-MOR to reduce the computational efforts in the numerical solution of our model problem on behalf of a simple, heuristic, and a-posteriori error estimate-free coupling of POD-MOR and a slightly modified SQP-type algorithm.

We conclude the thesis by an outlook to ongoing work.

As can be seen from this overview, there are two recurrent topics that can be seen as some kind of read thread of this thesis. First, and most important, we will have to deal with difficulties associated with the particular, highly nonlinear structure of our quasilinear parabolic state equation. This is an issue that appears —more or less explicitely— in every chapter of our work and that we will put special emphasis on. Second, as explained before, the two-norm discrepancy and/or second-order optimality conditions with or without two-norm gap are specific for and important in PDE-constrained optimization. This topic also appears multiple times in this thesis and plays a central role particularly in Chapters 2 to 4.

Chapter 0

Notation and definitions

In this chapter we briefly summarize conventions and abbreviations. Moreover, we fix the general notation and recall some basic definitions. In particular, we precisely introduce the required function spaces on rough domains.

Abbreviations and conventions

Troughout this thesis we will make use of the following, common abbreviations:

a.a./a.e.	almost all/almost everywhere (with respect to a measure),
DoF	degrees of freedom,
EIM	empirical interpolation method,
FEM	finite element method,
FON	first-order necessary optimality condition,
MOR	model order reduction,
ODE	ordinary differential equation,
PDE	partial differential equation,
POD	proper orthogonal decomposition,
SNC	second-order necessary optimality condition,
SQP	sequential quadratic programming,
SSC	second-order sufficient optimality condition,
SSN	semismooth Newton method,
s.t.	such that, or —depending on the context— subject to,
w.l.o.g.	without loss of generality,
w.r.t.	with respect to.

Sometimes, we will not distinguish properly between a "minimum" of a minimization problem (i.e. the minimal attained value of the objective functional) and the "minimizer" (i.e. the argument for which the minimal value is attained). The true meaning, however, will always become clear from the context. Moreover, for the reason of brevity we often refer to linear, semilinear, or quasilinear (elliptic or parabolic optimal control) problems, although being aware that "linear", "semilinear", or "quasilinear", and "elliptic", or "parabolic" only refers to the underlyling partial differential equation and not the entire optimization problem, of course. We follow the well-known convention in PDE-constrained optimization to write y, u, and p (possibly with indices, bars, hats etc.) for states, controls, and adjoint states. This causes a minor clash of notation, because "p" will also be used to

denote an integrability exponent — however, it will always become clear from the respective context whether "p" refers to a function space object or a real number.

In equations we utilize the notation "lhs $\lesssim$ rhs", if there is $c > 0$ such that "lhs $\leq c \cdot$ rhs" and the actual value of c does not matter or is clear in the respective context.

Given a set X and a subset $S \subset X$ we denote by $\mathbf{1}_S \colon X \to \mathbb{R}$, $\mathbf{1}_S(x) = 1$ for $x \in S$, $\mathbf{1}_S(x) = 0$ for $x \in X \setminus S$, the indicator function of S. Given another set Y and functions $f, g \colon X \to Y$ we use the abbreviation

$$\{f = g\} := \{x \in X \colon f(x) = g(x)\}.$$

By "$\rightrightarrows$" we indicate set-valued maps. The euclidean (i.e. the ℓ^2-)norm on $\mathbb{R}^d$ is denoted by $|\cdot|_2$, and the ℓ^1-norm by $|\cdot|_1$. We use the notation $B_r(x)$ for the open euclidean balls in $\mathbb{R}^d$ with center $x \in \mathbb{R}^d$ and radius $r > 0$.

Finally, we use the set-valued sign function defined by $\operatorname{sign}(z) = \{\pm 1\}$ for $z \gtrless 0$ and $\operatorname{sign}(0) = [-1, 1]$.

Normed spaces and linear operators

For an overview on linear functional analysis we refer the reader to, e.g., the summaries in the introductory chapters of [**12, 240**], or to the extensive monographies [**290, 104, 241**].

Let X, Y be (real) Banach spaces with norm $\|\cdot\|_X$, $\|\cdot\|_Y$, respectively. By $\mathcal{L}(X, Y)$ we denote the Banach space of bounded linear maps $X \to Y$, equipped with the operator norm

$$\|A\|_{\mathcal{L}(X,Y)} := \sup_{x \in X \setminus \{0\}} \frac{\|Ax\|_Y}{\|x\|_X}.$$

We use the short notation $\mathcal{L}(X) := \mathcal{L}(X, X)$ for the bounded linear maps acting on X. The notation $X \hookrightarrow Y$ indicates that $X \subset Y$ with the inclusion being a bounded linear map. If this embedding is dense, we use the notation $X \hookrightarrow_d Y$. The topological dual $\mathcal{L}(X, \mathbb{R})$ of X is denoted by X^*, and by $\langle \cdot, \cdot \rangle_{X^*, X}$ we refer to the respective duality pairing. If $\|\cdot\|_X$ is induced by a scalar product, we also write $\langle \cdot, \cdot \rangle_X$ for this scalar product.

If $A \colon X \to Y$ is a linear operator, we denote by A^* its adjoint $Y^* \to X^*$. Moreover, if $A \colon X \to Y$ is closed, we denote by $\operatorname{Dom}_X(A)$ the domain of A, equipped with the graph norm

$$\|x\|_{\operatorname{Dom}_X(A)} := \|x\|_X + \|Ax\|_Y, \qquad x \in \operatorname{Dom}_X(A).$$

As usual, $R(z, A) = (z - A)^{-1}$ denotes the resolvent of an operator A.

Given a sequence $(x_n)_n \subset X$ and $x \in X$, we denote by $x_n \to x$ and $x_n \rightharpoonup x$, norm- and weak convergence of x_n to x, respectively. An embedding that is additionally a compact linear map is denoted by $X \hookrightarrow_c Y$. Finally, we use the notation

$$\mathbb{B}_r^X(x) := \{z \in X \colon \|z - x\|_X < r\},$$

for the open ball in X with radius $r > 0$ and center $x \in X$, and $\operatorname{cl}_X(S)$ for the closure of a set $S \subset X$ with respect to the norm-topology on X. If the topology becomes clear from the context we also write $\overline{S}$ instead of $\operatorname{cl}_X(S)$. The closure of S with respect to the weak topology on X is denoted by weak-$\operatorname{cl}_X(S)$.

Interpolation theory

Let X and Y be (real) Banach spaces such that $X \hookrightarrow Z$, $Y \hookrightarrow Z$ for some locally convex space Z; in that case we call (X, Y) an interpolation couple. Given $\theta \in (0,1)$ and $r \in (0,\infty)$ we denote by

$$[X,Y]_\theta, \qquad \text{and} \qquad (X,Y)_{\theta,r}$$

the complex and real interpolation spaces between X and Y; for the precise definition hereof and basic functorial properties we refer the reader to, e.g., [**267**] or [**12**, Chapter I.2].

Spaces of continuous, Hölder continuous, and smooth functions

Given a topological space E and a Banach space X, we denote by $C(E,X)$ the space of bounded and continuous functions $E \to X$, equipped with the supremum norm

$$\|\varphi\|_{C(E,X)} := \sup_{s\in E} \|\varphi(s)\|_X.$$

If, in addition, E is a metric space with metric d_E, and $\sigma \in (0,1]$ we introduce the space of X-valued, bounded, σ-Hölder continuous functions on E by

$$C^{0,\sigma}(E,X) := \{\varphi\colon E \to X\colon \|\varphi\|_{C(E,X)} + |\varphi|_{C^{0,\sigma}(E,X)} < \infty\}$$

with $|\varphi|_{C^{0,\sigma}(E,X)} := \sup_{s,t\in E, s\neq t} \frac{\|\varphi(s)-\varphi(t)\|_X}{d_E(s,t)^\sigma}$ and equipp this space with norm

$$\|\cdot\|_{C^{0,\sigma}(E,X)} := \|\cdot\|_{C(E,X)} + |\cdot|_{C^{0,\sigma}(E,X)}.$$

For $\sigma = 1$ we obtain the space of bounded, Lipschitz continuous X-valued functions on E. We refer the reader to, e.g., [**12**, Chapter II.1.1] for these definitions.

For $X = \mathbb{R}$ we omit X in the notation and write $C(E)$ and $C^{0,\sigma}(E)$ instead of $C(E,\mathbb{R})$ and $C^{0,\sigma}(E,\mathbb{R})$. Let now E be a locally compact Hausdorff space. By $C_c(E)$ we denote the subspace of $C(E)$ consisting of continuous functions $E \to \mathbb{R}$ with compact support, and by $C_0(E)$ the closure of $C_c(E)$ in $C(E)$. The space of (signed) regular Borel measures on E is defined by $\mathcal{M}(E) := (C_0(E))^*$ and equipped with the canonical total variation norm. The support of a regular Borel measure $\mu \in \mathcal{M}(E)$ will be denoted by $\operatorname{supp}(\mu)$; see, e.g., [**242**, Chapter 6] for the details.

Finally, if $E \subset \mathbb{R}^d$ is open, $k \in \mathbb{N}$, $\sigma \in (0,1]$ we denote by $C^{k,\sigma}(E)$ the space of functions $E \to \mathbb{R}$ that are k-times differentiable with the k-th derivatives being bounded and σ-Hölder continuous.

Lebesgue, Sobolev, and Bessel potential spaces

For an introduction to measure theory and Lebesgue spaces we refer the reader to [**242**], and for the theory of Sobolev spaces in general to, e.g., [**127**]. Introducing Sobolev and Bessel potential spaces —in particular such with boundary conditions— on rough domains is a bit delicate. Therefore, we recall the main steps from [**25**, Section 2.3] to which we also refer for further details and references.

Function spaces are first introduced on $\mathbb{R}^d$ and then on open subsets of $\mathbb{R}^d$ by restriction. Given an integrability exponent $p \in [1,\infty]$ we denote by $p' = (1-p^{-1})^{-1}$ always the corresponding dual exponent. For $s \in (0,\infty)$, $p \in (1,\infty)$

we refer by $L^p(\mathbb{R}^d)$ and $H^{s,p}(\mathbb{R}^d)$ to the Lebesgue and Bessel potential spaces, respectively, on $\mathbb{R}^d$; see, e.g., [**25**, Section 2.3] or [**127**, Definition 1.3.1.3] for the standard definitions. Given a so-called $(d-1)$-regular set $D \subset \mathbb{R}^d$ [**25**, Definition 2.3] and $s \in (\frac{1}{p}, 1+\frac{1}{p})$ we define

$$H^{s,p}_D(\Omega) := \{\varphi \in H^{s,p}_D(\Omega)\colon\ \mathrm{tr}_D\,\varphi = 0\ \mathcal{H}^{d-1}\text{-a.e. on } D\},$$

where tr_D denotes the trace in the sense of Jonsson and Wallin and $\mathcal{H}^{d-1}$ the $(d-1)$-dimensional Hausdorff measure on D. The respective function spaces on subsets of $\mathbb{R}^d$ can now be defined by pointwise restriction: for open $E \subset \mathbb{R}^d$ and $(d-1)$-regular $D \subset \overline{E}$ we set

$$H^{s,p}(E) := \{\varphi|_E\colon\ \varphi \in H^{s,p}(\mathbb{R}^d)\},$$
$$H^{s,p}_D(E) := \{\varphi|_E\colon\ \varphi \in H^{s,p}_D(\mathbb{R}^d)\},$$

equipped with the quotient norms

$$\|\varphi\|_{H^{s,p}(E)} := \inf\{\|\psi\|_{H^{s,p}(\mathbb{R}^d)}\colon\ \psi \in H^{s,p}(\mathbb{R}^d), \psi|_E = \varphi\},$$
$$\|\varphi\|_{H^{s,p}_D(E)} := \inf\{\|\psi\|_{H^{s,p}(\mathbb{R}^d)}\colon\ \psi \in H^{s,p}_D(\mathbb{R}^d), \psi|_E = \varphi\}.$$

Finally, for $s < 0$ and $p \in (1,\infty)$ we set $H^{s,p}(E) = (H^{-s,p'}(E))^*$ and $H^{s,p}_D(E) = (H^{-s,p'}_D(E))^*$, where the duality pairing extends the $L^2(E)$-scalar product.

Since the function spaces introduced so far are defined by restriction of the respective spaces on $\mathbb{R}^d$, the well-known Sobolev embeddings hold true for them. Nevertheless, the definition by restriction is rather abstract. For $s \in \mathbb{Z}$, $s \geq -1$, and appropriate conditions on the underlying domain the spaces $H^{s,p}$ coincide with the well-known classical Sobolev spaces $W^{s,p}$. We give a brief summary hereof in the following. First, recall that in the classical way the Sobolev space $W^{1,p}_D(E)$, $p \in (1,\infty)$, with homogeneous Dirichlet boundary condition on D can by defined as the closure of

$$C^\infty_D(E) := \{\varphi\colon E \to \mathbb{R} \text{ is infinitely often differentiable, } \mathrm{supp}(\varphi) \cap D = \emptyset\}$$

w.r.t. the norm

$$\|\varphi\|_{W^{1,p}} := \left(\|\varphi\|^p_{L^p(E)} + \|\nabla\varphi\|^p_{L^p(E)}\right)^{\frac{1}{p}}.$$

Of course, $W^{1,p}_D(E)$ is also equipped with this norm. Moreover, one defines

$$W^{-1,p}_D(E) := (W^{1,p'}_D(E))^*.$$

For the equivalence of this definition to the second classical definition based on L^p-integrability of weak derivatives we refer the reader to [**221**]. We recall the following result [**25**, Proposition B.3] concerning the relation of Bessel potantial and Sobolev spaces.

Proposition 0.1. *If $E \subset \mathbb{R}^d$ is open, $D \subset \overline{E}$ is $(d-1)$-regular, and E satisfies a uniform Lipschitz condition around $\overline{E \setminus D}$, then it holds*

$$W^{1,p}_D(E) = H^{1,p}_D(E), \quad \textit{with equivalent norms.}$$

Similarly, one usually defines the Sobolev space (without boundary conditions) $W^{k,p}(E)$, $k \in \mathbb{N}$, $p \in (1,\infty)$, as the closure of $C_{\emptyset}^{\infty}(E)$ w.r.t. the $W^{k,p}(E)$-norm

$$\|\varphi\|_{W^{k,p}} := \left(\sum_{\ell=0}^{k} \sum_{|\alpha|=\ell} \|D_\alpha \varphi\|_{L^p}^p \right)^{\frac{1}{p}}.$$

For so-called (ϵ,δ)-domains, see, e.g., [**235**, Definition 5] for this notion, it is well-known that these Sobolev spaces again coincide with the Bessel potential spaces.

Proposition 0.2. *Let $E \subset \mathbb{R}^d$ be an (ϵ,δ)-domain. Then, it holds*

$$W^{k,p}(E) = H^{k,p}(E), \qquad \text{with equivalent norms,}$$

for $k \in \mathbb{Z}$, $k \geq -1$.

This is an immediate consequence of the following extension property for (ϵ,δ)-domains that is also noteworthy on its own and the well-known relation $H^{k,p}(\mathbb{R}^d) = W^{k,p}(\mathbb{R}^d)$ for $k \in \mathbb{N}$.

Theorem 0.3 ([**235**], Theorem 8). *Let $E \subset \mathbb{R}^d$ be an (ϵ,δ)-domain and $p \in [1,\infty]$. Then, there exists a so-called degree independent Sobolev extension operator, i.e. a bounded linear operator $\iota\colon L^p(E) \to L^p(\mathbb{R}^d)$ such that $(\iota\varphi)|_E = \varphi$ for all $\varphi \in L^p(E)$ and the restrictions*

$$\iota|_{W^{k,p}(E)}\colon W^{k,p}(E) \to W^{k,p}(\mathbb{R}^d),$$

are well-defined and bounded linear for each $k \in \mathbb{N}$.

Bochner-Lebesgue and Bochner-Sobolev spaces

Let $I \subset \mathbb{R}$ be an interval, $p \in [1,\infty]$, and X a Banach space. By $L^p(I,X)$ we denote the Bochner-Lebesgue space with respect to the Lebesgue measure $\mathrm{d}t$ on I; see, e.g., [**240**, Chapter 1.5] for the definition. By $W^{1,p}(I,X)$ we denote the Bochner-Sobolev space

$$W^{1,p}(I,X) := \{\varphi \in L^p(I,X)\colon \partial_t \varphi \in L^p(I,X)\},$$

where ∂_t denotes the derivative in the sense of distributions; see for instance [**240**, Chapter 7.1] or [**12**, Chapter III.1]. We equipp this space with norm

$$\|\varphi\|_{W^{1,p}(I,X)} := \|\varphi\|_{L^p(I,X)} + \|\partial_t \varphi\|_{L^p(I,X)}.$$

Moreover, given Banach spaces $Y \hookrightarrow_d X$ with dense embedding, and $s \in (1,\infty)$ we introduce the maximal regularity space by

$$\mathbb{W}^{1,s}(I,(X,Y)) := W^{1,s}(I,X) \cap L^s(I,Y),$$

equipped with norm $\|\cdot\|_{\mathbb{W}^{1,s}(I,(X,Y))} := \|\cdot\|_{W^{1,s}(I,X)} + \|\cdot\|_{L^s(I,Y)}$.

Maximal parabolic regularity

Let $Y \hookrightarrow_d X$ be Banach spaces and fix some interval $I = (0,T) \subset \mathbb{R}$, $T > 0$. Moreover, let $A\colon I \to \mathcal{L}(Y,X)$ be bounded and measurable such that $A(t)$ is a closed operator in X for each $t \in I$. In particular, we can regard A as as a bounded linear map $A\colon L^s(I,Y) \to L^s(I,X)$ by defining $(Aw)(t) = A(t)w(t)$ for $w \in L^s(I,Y)$.

Definition 0.4 ([**15**]). A is said to have (nonautonomous) maximal parabolic regularity on $L^s(I,X)$, $s\in(1,\infty)$, if for every $f\in L^s(I,X)$ and $w_0\in(X,Y)_{1/s',s}$ there exists a unique solution $w\in\mathbb{W}^{1,s}(I,(X,Y))$ to the equation

$$\partial_t w + Aw = f \text{ in } L^s(I,X), \qquad w(0)=w_0 \text{ in } (X,Y)_{1/s',s}.$$

For the well-definedness of $w(0)$ we refer to, e.g., Proposition 1.1. We denote by $\mathcal{MR}^s(I,(X,Y))$ the set of all operators having maximal parabolic regularity on $L^s(I,X)$. For equivalent formulations we refer to, e.g., Proposition 1.2 and [**15**, Proposition 3.1]. If A is autonomous, i.e. $A(t)\equiv A$ for all $t\in I$, maximal regularity of A on $L^{s_0}(I,X)$ for one $s_0\in(1,\infty)$ is equivalent to maximal regularity of A on all $L^s(I,X)$, $s\in(1,\infty)$ and $T>0$, cf., e.g., Proposition 1.3, and we say that A has maximal parabolic regularity on X. We use the notation $\mathcal{MR}(X,Y)$ for the set of all such operators.

Convex Analysis

Let X be a Banach space and $f\colon X\to\mathbb{R}$ be a convex function. It is well-known that the directional derivatives

$$f'(x,v) := \lim_{t\searrow 0}\frac{f(x+tv)-f(x)}{t},$$

are well-defined for any $x,v\in X$. The subgradient $\partial f(x)\subset X^*$ of f at x is defined by

$$\partial f(x) := \{\lambda\in X^*\colon f(y)-f(x)\geq\langle\lambda,y-x\rangle_{X^*,X} \quad \forall y\in X\}$$

and it holds $f'(x,v)=\sup_{\lambda\in\partial f(x)}\langle\lambda,v\rangle_{X^*,X}$; see, e.g., [**106**, Chapter I.5] or [**36**, Chapter 2.4] for these definitions. In this sense, we denote by $\|\cdot\|'_{L^1}(u,v)$ or $|\cdot|'_1(x,y)$ the directional derivatives of the L^1- or the ℓ^1-norm at u in direction v or at x in direction y, respectively.

From [**36**, Definition 2.54] let us recall the following definitions. Given a convex set $C\subset X$ and $x\in C$ we introduce the radial cone

$$\mathcal{R}_C(x) := \{\alpha(y-x)\colon \alpha>0,\ y\in C\},$$

and the tangent cone (contingent cone)

$$\begin{aligned}\mathcal{T}_C(x) &:= \{v\in X\colon \exists (v_k)_k\subset X\ \exists (t_k)_k\subset(0,\infty)\\ &\qquad\qquad \text{s.t. } v_k\to v,\ t_k\searrow 0 \text{ and } x+t_kv_k\in C\}\\ &= \operatorname{cl}_X\mathcal{R}_C(x),\end{aligned}$$

where the last equality is due to convexity of C; cf. [**36**, Proposition 2.55]. Moreover, we recall that C is called polyhedric if for all $x\in C$ and $\varphi\in X^*$ it holds

$$\mathcal{T}_C(x)\cap\ker\varphi = \operatorname{cl}_X(\mathcal{R}_C(x)\cap\ker\varphi);$$

note that the classic definition of polyhedricity, see , e.g., [**36**, Definition 3.51] or [**287**, Definition 3.1.1], can be simplified in the above way due to [**287**, Lemma 4.1].

Part I

Analysis of the control problems

Chapter 1

Preliminaries: the control-constrained problem

As explained in the introduction, the goal of this first part of the thesis is the analysis of two variants of the following control-constrained optimal control problem from [**35**]:

$$\text{(P)} \qquad \begin{cases} \min J(y,u) := \dfrac{1}{2}\|y - y_d\|^2_{L^2(I\times\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(\Lambda)}, \\ \quad \text{s.t.} \quad u \in U_{\mathrm{ad}} \\ \quad\quad\; \text{and (Eq)}. \end{cases}$$

The quasilinear parabolic state equation (Eq) will be given by

$$\text{(Eq)} \qquad \begin{cases} \partial_t y + \mathcal{A}(y)y = Bu & \text{in } L^s(I, W^{-1,p}_{\Gamma_D}), \\ y(0) = y_0 & \text{in } (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})_{1/s',s}, \end{cases}$$

where the quasilinear differential operator $\mathcal{A}$ is defined by $\mathcal{A}(y) := -\nabla \cdot \xi(y)\mu\nabla$. Boundary conditions are incorporated in the right-hand side of (Eq) and the respective function spaces. Our assumptions on the control space $L^s(\Lambda)$ and the control operator B will allow to cover different types of control mechanisms. For the precise definitions and assumptions we refer the reader to Section 1.2 below, where also examples will be given.

Throughout this thesis, the set of admissible controls $U_{\mathrm{ad}} \subset L^s(\Lambda)$ will be given by so-called box-constraints, i.e. by pointwise inequality-constraints from below and above. Again, the precise formulation will be provided below. The objective functional J consists of two summands: the quadratic L^2-tracking-type functional $y \mapsto \|y - y_d\|^2_{L^2(I\times\Omega)}$ and the control cost term $u \mapsto \frac{\alpha}{2}\|u\|^2_{L^2(\Lambda)}$. Consequently, by formulating (P) we aim at finding an admissible control $u \in U_{\mathrm{ad}}$ such that the associated state y is close to the function y_d, the so-called desired state, w.r.t. the $L^2(I\times\Omega)$-norm while keeping also track of the "cost" of the control u, measured in the squared $L^2(\Lambda)$-norm and weighted by the parameter $\alpha > 0$. In the context of inverse problems, such a cost term is also known as Tikhonov regularization term which is the reason why we refer to α as the Tikhonov parameter in the following. Let us briefly mention that replacing the tracking-type part of J by a different, sufficiently regular functional usually does not cause major problems, in general. The presence of a Tikhonov regularization term with $\alpha > 0$, however, is crucial for many results. The so-called bang-bang case $\alpha = 0$ exhibits quite different structural properties and usually requires different arguments as will be explained in more detail in Section 1.5.3.

The setting described above in (P), i.e. minimization of a quadratic L^2-tracking-type functional with L^2-control cost subject to box-constraints on the control-variable, is a typcial problem type in PDE-constrained optimization; cf., e.g., [**270, 156**]. For the case of a quasilinear parabolic state equation it has recently been analyzed in [**35**] and [**45**]; our problem (P) is identical to the one under consideration in [**35**]. Both papers [**35**] and [**45**] deal with existence of optimal controls and first- and second-order optimality conditions for optimal control problems of quasilinear parabolic PDEs with box-constraints on the control. They mainly differ w.r.t. the concrete form of the state equation and the assumptions on the underlying data.

In Chapter 2 we will be concerned with additional constraints on the state variable (problem (P^{st})) and in Chapter 3 we will add certain nonsmooth cost terms to the functional (problems ($\mathrm{P}_k^{\mathrm{sp}}$)). In Chapter 4 we come back to (P) and prove convergence of the SQP method. In all three chapters, the results on (P) obtained in [**216, 35, 45**] will serve as a fundament for our analysis. Therefore, we provide a summary hereof in the present chapter. Since we collect this material for later reference, our selection of the results is slightly biased: we cite the first- and second-order analysis exclusively from [**35**] although analogous results have been obtained in [**45**] as well. From [**45**] we only cite a result concerning improved regularity of solutions of the state equation. Moreover, we completely omit the stability analysis of the control problem w.r.t. the coefficient function ξ from [**35**, Section 5].

Let us note that the assumptions from [**35**] are minimal in the sense that they include certain nonsmooth domains and coefficient functions and mixed boundary conditions as they were formulated in [**216**] for the analysis of the underlying equation. Such a setting, which is also referred to as "rough", is particulary interesting because real-world problems often come along some kind of irregular data. Whenever possible, we will keep these minimal assumptions from [**35**] in our analysis. Therefore, we state and discuss them in Section 1.2 and will rely on them in most parts of the thesis. This is the reason why we often refer to [**35**] only although similar results have been obtained in [**45**] as well. Nevertheless, problem-specific difficulties will sometimes force us to apply certain results that are not available in the rough setting of [**35**]. In that case, we will switch to stronger, rather classical assumptions on the data in the flavour of [**45**] that in return allow to utilize much stronger regularity results obtained in [**45**]. In particular, we recall these strengthened assumptions on the equation together with the resulting regularity theory for solutions of this equation from [**45**] in Section 1.6. An overview on the relation of the different regularity assumptions used troughout this thesis together with the dependence of the main results on these assumptions is provided at the end of this chapter in Figure 1.1 and Table 1.1.

The structure of this chapter can be summarized as follows: since we make repeatedly use of the concept of maximal parabolic regularity we start with a summary of related important results in Section 1.1. In Section 1.2 we state and discuss the minimal assumptions on (P) from [**35**]. Results from [**216, 35**] concerning the solvability of the state equation (Eq) and differentiability of the associated solution map in this rough setting are summarized in Section 1.3. In Section 1.4 we briefly recall the analysis of the so-called adjoint equation from [**35**] and state

two extensions hereof from [**167, 169**]. Subsequently, we summarize the first- and second-order optimality conditions for the purely control-constrained instance of (P) from [**35**] in Section 1.5. At the end of the chapter we finally recall the improved regularity analysis for solutions of a variant of (Eq) under the additional assumptions from [**45**].

1.1. Maximal parabolic regularity

Maximal regularity spaces and the concept of (nonautonomous) maximal parabolic regularity provide the framework for the analysis of the underyling time-dependent PDEs in [**35, 216**]. Consequently, they will also play an important role in almost every part of our analysis. Therefore, we summarize some basic facts that will be used throughout the thesis.

Let $Y \hookrightarrow_d X$ be Banach spaces and fix some interval $I = (0,T) \subset \mathbb{R}$, $T > 0$. First, we recall from [**13**, Theorem 3] and [**12**, Theorem III.4.10.2], see also [**14**, formula (1.2)], the following embeddings of the maximal regularity spaces.

Proposition 1.1. *For $s \in [1,\infty)$ it holds:*

1. $\mathbb{W}^{1,s}(I,(X,Y)) \hookrightarrow L^q(I,(X,Y)_{\theta,1})$, *if* $1 \geq \frac{1}{q} > \theta - \frac{1}{s'} > 0$,
2. $\mathbb{W}^{1,s}(I,(X,Y)) \hookrightarrow C(\overline{I},(X,Y)_{1/s',s})$,
3. $\mathbb{W}^{1,s}(I,(X,Y)) \hookrightarrow C^{0,\sigma}(I,(X,Y)_{\theta,1})$, *if* $0 \leq \sigma < \frac{1}{s'} - \theta$.

If the embedding $Y \hookrightarrow X$ is compact, the first and the third embedding are compact as well.

The second embedding hereof implies that for each $t \in \overline{I}$ the trace map $\mathrm{tr}_t\colon \mathbb{W}^{1,s}(I,(X,Y)) \to (X,Y)_{1/s',s}$ is a well-defined bounded linear map.

The notion of (nonautonomous) maximal parabolic regularity has been introduced in Chapter 0. We start with one of the equivalent formulations from [**14**, Proposition 2.1] that is of particular interest to us.

Proposition 1.2. *Given $A \in L^1(I,\mathcal{L}(Y,X)) \cap \mathcal{L}(\mathbb{W}^{1,s}(I,(X,Y)),L^s(I,X))$ the following properites are equivalent:*

- $A \in \mathcal{MR}^s(I,(X,Y))$,
- $(\partial_t + A, \mathrm{tr}_0)\colon \mathbb{W}^{1,s}(I,(X,Y)) \to L^s(I,X) \times (X,Y)_{1/s',s}$ *is a topological isomorphism.*

As mentioned in Chapter 0, maximal parabolic regularity in the autonomous case is independent of the time interval and the integrability of the right-hand side. For completeness, we state the precise result from [**15**, Remark 6.1].

Proposition 1.3. *Let $A \in \mathcal{L}(Y,X)$ define a closed operator in X. Then the following conditions are equivalent:*

- $A \in \mathcal{MR}^s(I,(X,Y))$ *for some particular* $s \in (1,\infty)$ *and* $T > 0$,
- $A \in \mathcal{MR}^s(I,(X,Y))$ *for <u>all</u>* $s \in (1,\infty)$ *and* $T > 0$.

Finally, we will need the following result that allows to deduce maximal parabolic regularity of certain nonautonomous and autonomous operators. This is particularly usefull in the analysis of the linearized state equation or the adjoint equation; cf., Lemma 1.12, Proposition 1.17, and Lemma 2.24.

Proposition 1.4.

1. ([**15**], Theorem 7.1) *Let $\vartheta, \theta \in (0,1)$ and $0 \le \frac{1}{\rho} < \min(1-\theta, \frac{1}{s})$. It holds*

$$C(I, \mathcal{MR}(X,Y)) + L^\infty(I, \mathcal{L}(Y, (X,Y)_{\vartheta,\infty})) + L^\rho(I, \mathcal{L}((X,Y)_{\theta,\infty}, X)) \\ \subset \mathcal{MR}^s(I, (X,Y)).$$

2. ([**15**], Proposition 7.1) *If $A \in \mathcal{MR}^s(I,(X,Y)) \cap C(\overline{I}, \mathcal{L}(Y,X))$, then $A(t) \in \mathcal{MR}(X,Y)$ for each $t \in \overline{I}$.*
3. ([**231**], Corollary 3.4) *It holds*

$$C(I, \mathcal{MR}(X,Y)) + L^s(I, \mathcal{L}((X,Y)_{1/s',s}, X)) \subset \mathcal{MR}^s(I,(X,Y)).$$

Herein, when referring to continuous functions $I \to \mathcal{M}(X,Y)$, we equipp the set $\mathcal{M}(X,Y) \subset \mathcal{L}(Y,X)$ with the topology inherited from $\mathcal{L}(Y,X)$.

1.2. Minimal Assumptions

In the following we summarize the minimal assumptions required for the analysis of the state equation (Eq) from [**216, 35**]. As explained in the introduction, "minimal" refers to the fact that rather irregular data are allowed. Assumptions 1.5, 1.6 and 1.8 are close to the Assumptions 1-4 formulated in [**35**] but at first we forego those parts that refer to the improved regularity analysis from [**35**] on Bessel potential spaces. Instead, we stick to the original setting of [**216**] and formulate the improved regularity assumptions separately. We will explain this in more detail below Example 1.9 after having stated the assumptions.

At the end of this chapter, in Figure 1.1, we show how the different regularity settings used in this thesis are related to each other. Moreover, in Table 1.1 the required regularity assumptions for each of the main results in this thesis are summarized.

We start with the assumptions on the underlying domain, its boundary, and the boundary conditions imposed.

Assumption 1.5. $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, is a bounded domain with boundary $\partial\Omega$. $\Gamma_N \subset \partial\Omega$ is relatively open and denotes the Neumann boundary part, whereas $\Gamma_D = \partial\Omega \setminus \Gamma_N$ denotes the part of $\partial\Omega$ where homogeneous Dirichlet boundary conditions are prescribed. Let $\Omega \cup \Gamma_N$ be regular in the sense of Gröger [**128**] such that every chart map in the definition of regularity in the sense of Gröger can be chosen volume-preserving. The time interval $I = (0,T)$ with $T > 0$ is fixed. We denote the space-time cylinder by $Q := I \times \Omega$.

For an alternative, geometric characterization of regularity in the sense of Gröger we refer the reader to [**133**, Section 5]. Assumption 1.5 is fulfilled for any domain with a Lipschitz boundary ("strong Lipschitz domain", [**127**, Definition 1.2.1.1]) in the case $\Gamma_N = \emptyset$ or $\Gamma_N = \partial\Omega$; cf. [**134**, Remark 3.3]. However, there are also domains without Lipschitz boundary fulfilling this assumption, e.g., a pair of crossing beams in 3D [**134**, Section 7.3].

Function spaces on Ω are defined according to Chapter 0. Regarding the definition of Sobolev and Bessel potential spaces, with and without Dirichlet boundary conditions on Γ_D, we point out that Assumption 1.5 in particular implies the geometric conditions from Chapter 0 with $E = \Omega$ and $D = \Gamma_D$; see, e.g., [**35**, Appendix

A] for a proof. Moreover, the assumptions of Propositions 0.1 and 0.2 are satisfied, cf., e.g., [**105**, Lemma 2.2.20] for the fact that Ω is an (ϵ, δ)-domain, and hence integer Sobolev and Bessel potential spaces coincide. Since the spatial domain will be fixed we will omit it in our notation in the following, and write, e.g., L^p, $W^{1,p}_{\Gamma_D}$, $H^{-s,p}_{\Gamma_D}$, etc., instead of $L^p(\Omega)$, $W^{1,p}_{\Gamma_D}(\Omega)$, $H^{-s,p}_{\Gamma_D}(\Omega)$.

Next, we state our assumptions regarding the coefficient functions.

Assumption 1.6.

1. The function $\xi\colon \mathbb{R} \to \mathbb{R}$ is twice differentiable with ξ'' being Lipschitz continuous on bounded subsets of $\mathbb{R}$. Let $\mu\colon \Omega \to \mathbb{R}^{d\times d}$ be measurable, symmetric, and uniformly bounded and coercive in the following sense:

$$0 < \mu_\bullet := \inf_{x\in\Omega} \inf_{z\in\mathbb{R}^d\setminus\{0\}} \frac{z^T\mu(x)z}{z^Tz}, \qquad \mu^\bullet := \sup_{x\in\Omega} \sup_{1\le i,j\le d} |\mu_{i,j}(x)| < \infty.$$

   We assume a coercivity condition $0 < \xi_\bullet \le \xi \le \xi^\bullet$ for ξ as well. With this we define

$$\langle \mathcal{A}(y)\varphi, \psi\rangle_{L^2(I,W^{1,2}_{\Gamma_D})} := \int_I \int_\Omega \xi(y)\mu\nabla\varphi\nabla\psi \,\mathrm{d}x\mathrm{d}t, \qquad \varphi,\psi \in L^2(I, W^{1,2}_{\Gamma_D}).$$

2. We assume that there is $p \in (d, 4)$ such that

$$-\nabla\cdot\mu\nabla + 1\colon W^{1,p}_{\Gamma_D} \to W^{-1,p}_{\Gamma_D}$$

   is a topological isomorphism and fix this choice of p.

Assumptions 1.5 and 1.6 certainly impose nontrivial conditions on the geometry of the domain, the elliptic operator $-\nabla\cdot\mu\nabla + 1$, and the boundary conditions. Hence, we mention the following examples; cf. also [**35**, Remarks 2.1 and 2.3] and [**168**, Example 2.3].

Example 1.7.

1. If Ω is a bounded domain with Lipschitz boundary, $\Gamma_N = \emptyset$ or $\Gamma_N = \partial\Omega$, and $\mu\colon \Omega \to \mathbb{R}^{d\times d}$ is symmetric-valued and uniformly continuous, Assumption 1.5.2 is fulfilled with some $p > 3$; see [**107**, Theorem 3.12, Remark 3.17]. Therefore, Assumptions 1.5 and 1.6 cover the classical regular setting of domains with Lipschitz boundary in dimensions $d = 2, 3$ with pure Dirichlet or Neumann boundary conditions and symmetric, uniformly continuous coefficient μ.
2. According to the well-known work [**128**] the isomorphism property from Assumption 1.6.2 for some $p > 2$ is a consequence of Assumption 1.5 for any coefficient μ fulfilling Assumption 1.6.1. This is also true under more general assumptions on the domain; see [**132**]. Hence, for space dimension $d = 2$ Assumption 1.6.2 is guaranteed for a broad range of nonsmooth domains, mixed boundary conditions, and nonsmooth μ.
3. It is well-known that for mixed boundary conditions Assumption 1.6.2 can only be expected to hold for some $p < 4$, in general; cf. the famous counterexample in [**254**]. In [**97**] for instance several real-world constellations in dimension $d = 3$ have been described that fulfill Assumptions 1.5 and 1.6, e.g., two crossing beams in 3D with constant μ and pure homogeneous Dirichlet or Neumann boundary conditions.

Finally, we introduce the space of controls and how they enter the state equation.

Assumption 1.8. Let $s > 2$ be fixed such that $\frac{1}{s} < \frac{1}{2}(1 - \frac{d}{p})$ holds. For a measure space (Λ, ρ) we define the control space $U := L^s(\Lambda)$ and the admissible set

$$U_{\mathrm{ad}} = \{u \in L^s(\Lambda)\colon u_{\mathrm{a}}(x) \le u(x) \le u_{\mathrm{b}}(x) \quad \text{for } \rho\text{-a.a. } x \in \Lambda\}$$

with $u_{\mathrm{a}}, u_{\mathrm{b}} \in L^\infty(\Lambda)$, $u_{\mathrm{a}} \le u_{\mathrm{b}}$ ρ-almost everywhere. The control operator

$$B\colon U \to L^s(I, W_{\Gamma_D}^{-1,p})$$

is bounded linear and admits a bounded linear extension

$$B\colon L^2(\Lambda) \to L^2(I, W_{\Gamma_D}^{-1,p}).$$

Finally, the initial condition $y_0 \in (W_{\Gamma_D}^{-1,p}, W_{\Gamma_D}^{1,p})_{1-1/s,s}$, the desired state $y_d \in L^\infty(I, L^p)$, and an L^2-Tikhonov parameter $\alpha > 0$ are fixed.

The constants p and s are fixed from now on until further notice. Note that in Assumption 1.8 we only suppose B to be continuous from $L^s(\Lambda)$ to $L^s(I, W_{\Gamma_D}^{-1,p})$, instead from $L^\infty(\Lambda)$ to $L^s(I, W_{\Gamma_D}^{-1,p})$ as in [**35**]. Moreover, we have changed regularity of the desired state from $y_d \in L^\infty(I, L^2)$ to $y_d \in L^\infty(I, L^p)$. Nevertheless, this does not destroy applicability of the assumption to the full range of situations described in [**35**, Section 2.2]. For convenience we repeat these examples as stated in [**168**, Example 2.2].

Example 1.9.

1. Distributed control: It holds $\Lambda = Q$, i.e. $U = L^s(I \times \Omega)$, and B is the identity map $L^s(Q) \to L^s(I, W_{\Gamma_D}^{-1,p})$. Denoting the outer normal unit vector of $\partial\Omega$ by n_Ω, the state equation reads

$$\begin{aligned} \partial_t y + \mathcal{A}(y)y &= u && \text{on } Q,\\ n_\Omega \cdot \xi(y)\mu\nabla y &= 0 && \text{on } I \times \Gamma_N,\\ y &= 0 && \text{on } \Gamma_D. \end{aligned}$$

2. Neumann boundary control ($d = 2$): We choose $\Lambda = I \times \Gamma_N$, i.e. $U = L^s(I \times \Gamma_N)$, and $B = \mathrm{tr}^*$ where tr: $L^{s'}(I, W_{\Gamma_D}^{1,p'}) \to L^{s'}(I \times \Gamma_N)$ denotes the trace map. With this the state equation reads

$$\begin{aligned} \partial_t y + \mathcal{A}(y)y &= 0 && \text{on } Q,\\ n_\Omega \cdot \xi(y)\mu\nabla y &= u && \text{on } I \times \Gamma_N,\\ y &= 0 && \text{on } \Gamma_D. \end{aligned}$$

   Of course, adding a sufficiently regular, fixed, nonhomogeneous distributed source is possible. The same is true for a sufficiently regular, fixed, nonhomogeneous Neumann boundary source in Example 1.9.1.
3. The following setting for the controls, often referred to as "purely time-dependent controls", has been introduced in [**91**] in the case of a semilinear parabolic state equation. We fix $b_1, \dots, b_m \in W_{\Gamma_D}^{-1,p}$, that we may imagine as fixed actuators, set $U = L^s(I, \mathbb{R}^m)$, and define $Bu := \sum_{i=1}^m u_i b_i$. If, for

instance, $b_i = \mathrm{tr}^* f_i$ with $f_i \in L^s(\Gamma_N)$ where tr: $W^{1,p'}_{\Gamma_D} \to L^{s'}(\Gamma_N)$ denotes the trace map on Γ_N, we obtain as state equation:

$$\begin{aligned} \partial_t y + \mathcal{A}(y)y &= 0 && \text{on } Q, \\ n_\Omega \cdot \xi(y)\mu\nabla y &= \sum_{i=1}^m u_i f_i && \text{on } I \times \Gamma_N, \\ y &= 0 && \text{on } \Gamma_D. \end{aligned}$$

Note that this approach allows control action on the Neumann boundary also in space dimension $d = 3$, but now with a-priori fixed actuators. A similar construction applies to actuators $b_i \in L^s(\Omega)$.

Besides its relevance in practice, the purely time-dependent control setting has also several advantages w.r.t. the theoretical analysis. We will explain this in more detail as soon as we will restrict ourselves to this setting, e.g., in Section 2.4, Chapter 3 and Chapter 4.

As announced at the beginning of this section, let us briefly comment on the difference of Assumptions 1.5, 1.6 and 1.8 compared to the assumptions in [**35**]. In fact, note that except for mininmal changes (symmetry of μ, slightly increased integrability of y_d) Assumptions 1.5, 1.6 and 1.8 are identical to Assumptions 1-3 of [**35**], i.e. the suppositions w.r.t. domain, coefficients, and boundary conditions essentially remain unchanged. We only modify the assumptions w.r.t. the initial condition and regularity of the right-hand side of (Eq). Assumption 4 in [**35**], cf. Assumption 1.10 below, is related to the improved regularity analysis of the state equation on Bessel potential spaces. As pointed out in [**35**, Section 3] this analysis is not required for the first- and second-order analysis of Sections 3.1 and 4.1-4.3 of [**35**], except for [**35**, Proposition 4.7], a result concerning the so-called adjoint state. We will recall the respective results from [**35**] in the remaining part of this chapter. If not stated otherwise we will only rely on those results that are obtained completely within the $W^{-1,p}_{\Gamma_D}$-$W^{1,p}_{\Gamma_D}$-setting described in our Assumptions 1.5, 1.6 and 1.8, cf. also [**216**, Theorem 5.3], and do not include the improved regularity assumptions of [**35**]. For completeness and for later reference, however, we also state the following strengthened version of Assumption 1.8 that is identical to [**35**, Assumption 4]. It enables to obtain improved regularity of the states on the Bessel potential spaces $H^{-\zeta,p}_{\Gamma_D}$ which will also be recalled in this chapter.

Assumption 1.10. Let $\zeta \in (0,1)$ and $s \in (2,\infty)$ satisfy

$$\max\left(1 - \frac{1}{p}, \frac{d}{p}\right) < \zeta, \qquad \frac{1}{s} < \min\left(\frac{1}{2} - \frac{\zeta}{2}, \frac{\zeta}{2} - \frac{d}{2p}\right).$$

For a measure space (Λ, ρ) we define the control space $U = L^s(\Lambda)$ and the admissible set by

$$U_{\mathrm{ad}} = \{u \in L^s(\Lambda) \colon u_{\mathrm{a}}(x) \le u(x) \le u_{\mathrm{b}}(x) \quad \text{for } \rho\text{-a.a. } x \in \Lambda\},$$

with $u_{\mathrm{a}}, u_{\mathrm{b}} \in L^\infty(\Lambda)$, $u_{\mathrm{a}} \le u_{\mathrm{b}}$ ρ-almost everywhere. The control operator

$$B \colon L^s(\Lambda) \to L^s(I, H^{-\zeta,p}_{\Gamma_D}),$$

is bounded linear and admits a bounded linear extension

$$B \colon L^2(\Lambda) \to L^2(I, W^{-1,p}_{\Gamma_D}).$$

Finally the initial condition $y_0 \in (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))_{1/s',s}$, the desired state $y_d \in L^\infty(I, L^p)$, and a Tikhonov-parameter $\alpha > 0$ are fixed.

We note that Assumption 1.10 allows for all constellations described in Example 1.9 with the only difference that the purely time-dependent control case now requires actuators $b_i \in H^{-\zeta,p}_{\Gamma_D}$, $i = 1, ..., m$. Herein, the assumptions on ζ ensure that $H^{-\zeta,p}_{\Gamma_D}$ also contains boundary integrals or certain distributional sources; cf. [**134**, Theorem 6.9].

1.3. Solutions of the state equation

In this section we summarize results from [**216, 35**] concerning the existence of solutions to (P). Moreover, we precisely introduce the so-called control to state map, recall results concerning its differentiability and state formulas for the derivatives of the reduced functional. In particular, we explain briefly how the regularity assumptions are related to these differentiability results.

First, let us make precise in which sense solutions to (Eq) have to be understood: $y \in \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ is said to be a solution of the equation

$$\begin{cases} \partial_t y + \mathcal{A}(y)y = v, & \text{in } L^s(I, W^{-1,p}_{\Gamma_D}), \\ y(0) = y_0, & \text{in } (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})_{1/s',s}, \end{cases} \tag{1.1}$$

if and only if

$$\langle \partial_t y, \varphi\rangle_{W^{-1,p}_{\Gamma_D}, W^{1,p'}_{\Gamma_D}} + \int_\Omega \xi(y(t))\mu\nabla y(t)\nabla\varphi \,\mathrm{d}x = \langle v(t), \varphi\rangle_{W^{-1,p}_{\Gamma_D}, W^{1,p'}_{\Gamma_D}}$$

for all $\varphi \in W^{1,p'}_{\Gamma_D}$ and almost all $t \in I$, and $y(0) = y_0$ in $(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})_{1/s',s}$. For the well-definedness of $y(0) \in (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})_{1/s',s}$ we refer to Proposition 1.1.

Existence and uniqueness of solutions to (1.1) has first been obtained in [**216**, Corollary 5.8] for a slightly different equation with elliptic operator $-\nabla\cdot\xi(y)\mu\nabla + 1$ instead of $-\nabla\cdot\xi(y)\mu\nabla$ and an additional, uniformly bounded, zero-order semilinear term. In [**35**] this result has been adapted to our precise setting.

Theorem 1.11 ([**35**], Proposition 3.5). *Under Assumptions 1.5, 1.6 and 1.8 the solution map $\tilde{S}$ of the equation* (1.1)*, defined by $y := \tilde{S}(v)$ if and only if* (1.1) *holds, is a well-defined map*

$$\tilde{S}\colon L^s(I, W^{-1,p}_{\Gamma_D}) \to \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})).$$

By composition of $\tilde{S}$ with B we obtain the so-called control-to-state map

$$S\colon L^s(\Lambda) \to \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})), \qquad u \mapsto \tilde{S}(Bu),$$

which is the solution map of (Eq). For the applicability of the results from [**216**] to the slightly different equation in [**35**] we refer the reader, e.g., to Appendix A of [**168**] the respective details have been summarized.

Given $y \in \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ we recall from [**35**] the notation for certain terms related to the derivatives of the nonlinearity, stated in weak form:

$$\langle \mathcal{A}'(y)v, \varphi\rangle := \int_Q \xi'(y)v\mu\nabla y\nabla\varphi \,\mathrm{d}x\mathrm{d}t,$$

$$\langle \mathcal{A}''(y)[v_1,v_2],\varphi\rangle := \int_Q (\xi'(y)(v_1\mu\nabla v_2 + v_2\mu\nabla v_1) + \xi''(y)v_1v_2\mu\nabla y)\nabla\varphi \,\mathrm{d}x\mathrm{d}t,$$

with $v, v_1, v_2 \in \mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ and a test function $\varphi \in L^{s'}(I, W^{1,p'}_{\Gamma_D})$. It is possible to relax the regularity requirements on v, v_1, v_2, as done, e.g., in the proof of the following differentiability properties of $\tilde{S}$ [**35**, Proposition 4.4 and Lemma 4.5].

Lemma 1.12. *Let Assumptions 1.5, 1.6 and 1.8 be satisfied.*

1. *The map $\tilde{S}\colon L^s(I,W^{-1,p}_{\Gamma_D}) \to \mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ is twice continuously Fréchet differentiable with derivatives $\tilde{S}'(v)h = z$ and $\tilde{S}''(v)[h_1,h_2]$ given by the unique solutions of*

$$\partial_t z + \mathcal{A}(y)z + \mathcal{A}'(y)z = h, \qquad z(0) = 0, \tag{1.2}$$

$$\partial_t w + \mathcal{A}(y)w + \mathcal{A}'(y)w = \mathcal{A}''(y)[\tilde{S}'(v)h_1, \tilde{S}'(u)h_2], \qquad w(0) = 0 \tag{1.3}$$

   for $y = \tilde{S}(v)$, respectively.
2. *The nonautonomous operator $\mathcal{A}(y) + \mathcal{A}'(y)$ exhibits maximal parabolic regularity on $L^r(I,W^{-1,p}_{\Gamma_D})$ for $r \in (1,s]$. It holds*

$$\tilde{S}'(v) \in \mathcal{L}(L^r(I,W^{-1,p}_{\Gamma_D}), \mathbb{W}^{1,r}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$$

   for all $v \in L^s(I,W^{-1,p}_{\Gamma_D})$, $r \in (1,s]$.

At this point, some comments regarding the relation between the choice of the time integrability exponent s in Assumption 1.8 and differentiability of $\tilde{S}$ seem to be appropriate. Differentiablity of $\tilde{S}$ is proven in [**35**] by an application of the implicit function theorem; cf. the proof of [**35**, Lemma 4.5]. Consequently, it is crucial to ensure that the map

$$\mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \to L^s(I,W^{-1,p}_{\Gamma_D}), \quad y \mapsto \partial_t + \mathcal{A}(y)y,$$

is twice continuously Fréchet differentiable. As we have pointed out in the introduction on behalf of the example (0.2), differentiability of superposition operators between function spaces is a nontrivial topic; see, e.g., [**270**, Chapter 4.3] or the extensive monography [**16**]. In the present case, differentiability of the superposition operator $y \mapsto \xi(y)$ as $L^\infty(Q)$-valued map is needed. Thus, s has to be chosen in such a way that there is an embedding

$$\mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow L^\infty(Q).$$

This results in exactly the choice for s as in Assumption 1.8; in fact, there is even $\kappa > 0$ such that

$$\mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow_c C^{0,\kappa}(\overline{Q}); \tag{1.4}$$

cf. [**35**, Proposition 3.3]. In particular, we point out that considering controls in $L^s(\Lambda)$ with $s > 2$ is necessary to ensure differentiability of the control-to-state map.

For convenience of the reader, let us state the concrete Bochner-Sobolev embedding that led to (1.4) for later reference.

Lemma 1.13 ([**35**], Proposition 3.3). *Let Assumption 1.5 hold and $p > d$. If $1 > \frac{1}{r} \geq \frac{1}{2} - \frac{d}{2p}$, then*

$$\mathbb{W}^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow_c L^q(I, C^{0,\kappa})$$

with $\frac{1}{q} > \frac{1}{r} - \frac{1}{2} + \frac{d}{2p}$ and $\kappa = \kappa(q) > 0$. If $\frac{1}{r} < \frac{1}{2} - \frac{d}{2p}$, then

$$\mathbb{W}^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow_c C^{0,\rho}(I, C^{0,\kappa})$$

with $\rho = \frac{1}{2} - \frac{d}{2p} - \frac{1}{r} - \frac{\kappa}{2}$ and $\kappa > 0$ sufficiently small.

Let us now introduce the reduced objective functional

$$j\colon L^s(\Lambda) \to \mathbb{R},\ u \mapsto J(S(u), u). \tag{1.5}$$

From [**35**, Lemma 4.6] we recall that the reduced functional j is twice continuously Fréchet differentiable on $L^s(\Lambda)$ with gradient

$$\nabla j(u) = B^* \tilde{S}'(Bu)^*(y - y_d) + \alpha u, \qquad u \in L^s(\Lambda). \tag{1.6}$$

As typical in PDE-constrained optimization, the abbreviation $p := \tilde{S}'(Bu)^*(y - y_d)$, the so-called adjoint state, which is well-defined in $L^{r'}(I, W^{1,p'}_{\Gamma_D})$, $r' \in [s', \infty)$, due to Lemma 1.12, is introduced and one writes $\nabla j(u) = B^*p + \alpha u$. In this notation, the second derivative of j at $u \in L^s(\Lambda)$ in direction $v \in L^s(\Lambda)$ is given by

$$j''(u)v^2 = \alpha \|v\|^2_{L^2(\Lambda)} + \int_Q \left((1 - \xi''(y)\nabla p \mu \nabla y) z_v^2 - 2\xi'(y) z_v \nabla p \mu \nabla z_v\right) \mathrm{d}x\mathrm{d}t, \tag{1.7}$$

with $z_v = S'(u)v$; cf. [**35**, Proposition 4.10].

We conclude this summary on the analysis of (Eq) with the improved regularity result on Bessel potential spaces that holds under the slightly stronger regularity Assumption 1.10.

Theorem 1.14 ([**35**], Theorem 3.20). *Let Assumptions 1.5, 1.6 and 1.10 hold. For each control $u \in L^s(\Lambda)$ the associated state y has additional regularity*

$$y \in \mathbb{W}^{1,s}(I, (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla \cdot \mu \nabla))).$$

Moreover, the operator $-\nabla \cdot \xi(y)\mu\nabla$ has nonautonomous maximal parabolic regularity on $H^{-\zeta,p}_{\Gamma_D}$.

We further note that there is $\kappa > 0$ such that

$$\mathbb{W}^{1,s}(I, (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla \cdot \mu \nabla))) \hookrightarrow_c C^{0,\kappa}(I, W^{1,p}_{\Gamma_D}); \tag{1.8}$$

cf. [**35**, Corollary 3.7].

1.4. The adjoint equation

In the previous section we have expressed the gradient of the reduced functional of (P) with the help of the so-called adjoint state; cf. (1.6). Expressing the adjoint state, so far only given by the rather abstract definition $p = \tilde{S}'(Bu)^*(y - y_d)$, in terms of the solution of a certain backward parabolic PDE, the so-called adjoint equation, is a standard technique in optimal control of parabolic PDEs. In essence, one has to prove that $\tilde{S}(Bu)^*$ can be identified with the solution operator of the

adjoint equation. Roughly speaking, since $\tilde{S}'(Bu)$ is the solution map of a linear nonautonomous parabolic PDE, cf. Lemma 1.12, one has to observe that the adjoint of the solution operator of such a linear (forward) parabolic PDE is given by the solution operator of an appropriate backward parabolic PDE; see, e.g., [**270**, Chapter 3.6 and 5.5] for a detailed exposition in the case of linear and semilinear control problems.

In this section we recall the corresponding result for the quasilinear parabolic case from [**35**]. Moreover, we provide two different extensions hereof that we will rely on during our further analysis. We start with the result from [**35**], dealing with the adjoint equation in the $W^{-1,p'}_{\Gamma_D}$-$W^{1,p'}_{\Gamma_D}$-setting.

Proposition 1.15 ([**35**], Proposition 4.7). *Let $y \in C(\bar{I}, W^{1,p}_{\Gamma_D})$ and $r' \in [s', \infty)$. For all $w \in L^{r'}(I, W^{-1,p'}_{\Gamma_D})$ and $w_T \in (W^{-1,p'}_{\Gamma_D}, W^{1,p'}_{\Gamma_D})_{1/r,r'}$, the equation*

$$\begin{cases} -\partial_t p + \mathcal{A}(y)^* p + \mathcal{A}'(y)^* p = w, & \text{in } L^{r'}(I, W^{-1,p'}_{\Gamma_D}), \\ p(T) = w_T, & \text{in } (W^{-1,p'}_{\Gamma_D}, W^{1,p'}_{\Gamma_D})_{1/r,r'}, \end{cases}$$

has a unique solution $p \in \mathbb{W}^{1,r'}(I, (W^{-1,p'}_{\Gamma_D}, W^{1,p'}_{\Gamma_D}))$, and the solution map $(w, w_T) \mapsto p$ is bounded linear.

In particular, under Assumptions 1.5, 1.6 and 1.10 the adjoint state p introduced in the previous section exhibits regularity

$$p \in \mathbb{W}^{1,r'}(I, (W^{-1,p'}_{\Gamma_D}, W^{1,p'}_{\Gamma_D})), \qquad r' \in [s', \infty),$$

and satisfies the equation

$$-\partial_t p + \mathcal{A}(y)^* + \mathcal{A}'(y)^* p = y - y_d, \qquad p(T) = 0,$$

in the sense of distributions.

Since the application of Proposition 1.15 requires $y \in C(\bar{I}, W^{1,p}_{\Gamma_D})$, we only get improved regularity of the adjoint state if the required regularity for the state is guaranteed, e.g., by Theorem 1.14. The latter, however, relies on the additional regularity assumptions from [**35**]; cf. Assumption 1.10. Obviously, the result does not allow to obtain $L^\infty(I, W^{1,p'}_{\Gamma_D})$-regularity of the adjoint state. Since the latter will be required during the analysis of the SQP method in Chapter 4, we discuss the adjoint equation on the Bessel potential spaces $H^{-\zeta,p}_D$. This is our first extension of Proposition 1.15.

Proposition 1.16 ([**167**], Theorem 7.2). *Let $y \in C(\bar{I}, W^{1,p}_{\Gamma_D})$ and $r' \in (1, \infty)$. For all $w \in L^{r'}(I, H^{-\zeta,p}_{\Gamma_D})$ and $w_T \in (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla \cdot \mu \nabla))_{1/r,r'}$, the equation*

$$\begin{cases} -\partial_t p + \mathcal{A}(y)^* p + \mathcal{A}'(y)^* p = w, & \text{in } L^{r'}(I, H^{-\zeta,p}_{\Gamma_D}), \\ p(T) = w_T, & \text{in } (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla \cdot \mu \nabla))_{1/r,r'}, \end{cases}$$

has a unique solution $p \in \mathbb{W}^{1,r'}(I, (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla \cdot \mu \nabla)))$, and the solution map $(w, w_T) \mapsto p$ is bounded linear.

In particular, under Assumptions 1.5, 1.6 and 1.10 the adjoint state p exhibits regularity

$$p \in \mathbb{W}^{1,r'}(I,(H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))), \qquad r' \in [s',\infty),$$

and satisfies the equation

$$-\partial_t p + \mathcal{A}(y)^* + \mathcal{A}'(y)^* p = y - y_d, \qquad p(T) = 0,$$

in the sense of distributions.

For $r' = s$ this allows to conclude $C(\bar{I}, W^{1,p}_{\Gamma_D})$-regularity of the adjoint state by an application of embedding (1.8). This is even more than $L^\infty(I, W^{1,p'}_{\Gamma_D})$-regularity of the adjoint state that we will rely on in Chapter 4; cf. the remarks below Theorem 4.11. The following proof from [**167**] is a shorter and more elegant version of our original argument, communicated by H. Meinlschmidt (FAU Erlangen).

Proof. The starting point of the argument is the observation that the map

$$(-\partial_t + \mathcal{A}(y)^* + \mathcal{A}'(y)^*, \mathrm{tr}_T)\colon \mathbb{W}^{1,r'}(I,(W^{-1,p'}_{\Gamma_D}, W^{1,p'}_{\Gamma_D})) \to L^{r'}(I, W^{-1,p'}_{\Gamma_D}) \times (W^{-1,p'}_{\Gamma_D}, W^{1,p'}_{\Gamma_D})_{1/r,r'} \tag{1.9}$$

is a topological isomorphism for each $r' \in (1,\infty)$; this is a consequence of Proposition 1.15 and Proposition 1.2. We will consider restrictions of this isomorphism to smaller spaces of more regular functions. First, a short computation verifies that

$$\mathcal{A}'(y)^*|_{L^{r'}(I,W^{1,p}_{\Gamma_D})}\varphi = \xi'(y)\mu\nabla y\nabla\varphi, \qquad \varphi \in L^{r'}(I, W^{1,p}_{\Gamma_D}),$$

and that the conditions on ζ and p ensure $L^{p/2} \hookrightarrow H^{-\zeta,p}_{\Gamma_D}$. We choose $\frac{1+\zeta}{2} < \theta < 1$ such that $\frac{1}{r'} > 1 - \vartheta$ and obtain with the help of Proposition 1.1.1 and [**35**, Proposition 3.6] that

$$\mathbb{W}^{1,r'}(I,(H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))) \hookrightarrow_c L^{r'}(I,(H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))_{\theta,1}) \hookrightarrow L^{r'}(I, W^{1,p}_{\Gamma_D}). \tag{1.10}$$

Together, the operator

$$\begin{aligned}\mathcal{A}'(y)^*\colon \mathbb{W}^{1,r'}(I,(H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))) &\hookrightarrow_c L^{r'}(I, W^{1,p}_{\Gamma_D}) \to L^{r'}(I, L^{p/2}) \hookrightarrow L^{r'}(I, H^{-\zeta,p}_{\Gamma_D}),\\ z &\mapsto \xi'(y)\mu\nabla y\nabla z,\end{aligned}$$

is compact as it can be expressed as the composition of bounded linear operators of which one is the compact embedding (1.10). From Theorem 1.14 in combination with Proposition 1.2 and Proposition 1.4 we know that the map

$$(-\partial_t + \mathcal{A}(y)^*, \mathrm{tr}_T)\colon \mathbb{W}^{1,r'}(I,(H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))) \to L^{r'}(I, H^{-\zeta,p}_{\Gamma_D}) \times (H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla)_{1/r,r'}$$

is a topological isomorphism. Consequently, the sum

$$(-\partial_t + \mathcal{A}(y)^* + \mathcal{A}'(y)^*, \mathrm{tr}_T)\colon \mathbb{W}^{1,r'}(I,(H^{-\zeta,p}_{\Gamma_D},\mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla\cdot\mu\nabla))) \tag{1.11}$$

$$\to L^{r'}(I, H^{-\zeta,p}_{\Gamma_D}) \times (H^{-\zeta,p}_{\Gamma_D}, \mathrm{Dom}_{H^{-\zeta,p}_{\Gamma_D}}(-\nabla \cdot \mu\nabla)_{1/r,r'}$$

is a Fredholm operator of index 0. Since it is the restriction of the isomorphism (1.9), its kernel is trivial, and hence (1.11) is a topological isomorphism itself. □

In contrary to the above two results on the adjoint equation, the following result on regularity of the adjoint states on the L^p-$\mathrm{Dom}_{L^p}(-\nabla \cdot \mu\nabla)$-scale only relies on Assumptions 1.5, 1.6 and 1.8. Together with its proof it has been provided by L. Bonifacius (Munich) and J. Rehberg (WIAS, Berlin) in personal communication.

Proposition 1.17 ([**169**], Proposition 3.2)**.** *Given $y \in C(\overline{Q})$, the operator $-\xi(y)\nabla \cdot \mu\nabla$ has nonautonomous maximal parabolic regularity on $L^r(I, L^q)$ for each $q \in [2, \infty)$. In particular, under Assumptions 1.5, 1.6 and 1.8 the adjoint state p introduced in the previous section exhibits the regularity*

$$p \in \mathbb{W}^{1,r}(I, (L^p, \mathrm{Dom}_{L^p}(-\nabla \cdot \mu\nabla))), \qquad r \in (1, \infty),$$

and satisfies the equation

$$-\partial_t p - \xi(y)\nabla \cdot \mu\nabla p = y - y_d, \qquad p(T) = 0,$$

in the sense of distributions. Moreover, there is an embedding

$$\mathbb{W}^{1,r}(I, (L^p, \mathrm{Dom}_{L^p}(-\nabla \cdot \mu\nabla))) \hookrightarrow C^{0,\sigma}(I, W^{1,p}_{\Gamma_D}) \tag{1.12}$$

with some $\sigma > 0$ provided that $r \in (2, \infty)$.

Proof. Due to our assumptions on ξ, μ, Ω, and Γ_D we can apply [**140**, Proposition 5.4] to obtain maximal parabolic regularity of each autonomous operator $-\xi(y(t))\nabla \cdot \mu\nabla$, $t \in I$. Due to $y \in C(\overline{Q})$, cf. Theorem 1.11 and (1.4), the nonautonomous operator $t \mapsto -\xi(y(t))\nabla \cdot \mu\nabla$ is continuous as a map $I \to \mathcal{L}(\mathrm{Dom}_{L^q}(-\nabla \cdot \mu\nabla), L^q)$, from which we conclude nonautonomous maximal parabolic regularity of $-\xi(y)\nabla \cdot \mu\nabla$ on L^q by application of Proposition 1.4.1.

Next, we prove that $\mathcal{A}(y)^* + \mathcal{A}'(y)^* = -\xi(y)\nabla \cdot \mu\nabla$ on $\mathrm{Dom}_{L^q}(-\nabla \cdot \mu\nabla)$ for $q \in [2, \infty)$ and any $y \in W^{1,p}_{\Gamma_D}$. Let $z \in \mathrm{Dom}_{L^q}(-\nabla \cdot \mu\nabla)$ and $\psi \in C^\infty_c(\Omega)$. Since $\psi\xi(y) \in W^{1,p}_{\Gamma_D}$ has compact support and $\mu\nabla z$ has weak divergence in L^q we obtain:

$$\begin{aligned}\int_\Omega (-\xi(y)\nabla \cdot \mu\nabla z)\psi \mathrm{d}x &= \int_\Omega (-\nabla \cdot \mu\nabla z)\psi\xi(y)\mathrm{d}x = \int_\Omega \mu\nabla z\nabla(\psi\xi(y))\mathrm{d}x \\ &= \langle z, (\mathcal{A}(y) + \mathcal{A}'(y))\psi\rangle_{W^{1,p'}_{\Gamma_D}, W^{-1,p}_{\Gamma_D}} = \langle (\mathcal{A}(y) + \mathcal{A}'(y))^* z, \psi\rangle_{W^{-1,p'}_{\Gamma_D}, W^{1,p}_{\Gamma_D}}.\end{aligned}$$

The left-hand side thereof is well-defined for every $\psi \in L^{q'}$, and hence the claimed identity follows from density of $C^\infty_c(\Omega)$ in $L^{q'}$. Finally, we can proceed similar to the proofs of [**35**, Lemma 4.6 and Proposition 4.7] to improve regularity of $p = \tilde{S}'(u)^*(y - y_d)$ and to obtain the adjoint equation. It remains to show (1.12). According to [**134**, Lemma 6.6] it holds

$$(L^p, \mathrm{Dom}_{L^p}(-\nabla \cdot \mu\nabla))_{\tau,1} \hookrightarrow W^{1,p}_{\Gamma_D}, \qquad \tau \in \left(\frac{1}{2}, 1\right).$$

Hence we can apply standard Bochner-Sobolev embeddings; see, e.g., Proposition 1.1. □

1.5. Optimality conditions for pure control-constraints

In this section we summarize the analysis of the control-constrained problem (P) from [**35**]. As explained before Assumption 1.10, the respective results can be obtained under Assumptions 1.5, 1.6 and 1.8 and do not require the slightly stronger Assumption 1.10. The overal structure of this section resembles the structure of Chapters 2 and 3 and is typical for the analysis of nonlinear PDE-constrained optimization problems; cf. [**270, 156**]: first, one has to prove well-posedness of the problem by showing existence of a globally optimal control. Second, local solutions of the problem are characterized by first-order necessary optimality conditions. Third, second-order sufficient conditions and, if possible, complementing second-order necessary conditions are derived.

Before going into the details, let us mention that analogous first- and second-order results for distributed control of the slightly different equation (1.22) instead of (Eq) have been obtained in [**45**].

1.5.1. Existence of optimal controls. Let us start with the well-posedness of (P), i.e. existence of optimal controls.

Proposition 1.18 ([**216**], Proposition 6.4, [**35**], Lemma 4.1)**.** *Under Assumptions 1.5, 1.6 and 1.8 the optimal control problem* (P) *admits at least one globally optimal control.*

The proof utilizes standard arguments in the calculus of variations, cf., e.g., [**270, 156**], and can be found in [**216**]. U_{ad} is bounded in $L^s(\Lambda)$ and hence weakly sequentially compact and weak sequential lower semicontinuity of the reduced functional j is a consequence of weak-to-strong continuity of the control-to-state map as map $L^s(\Lambda) \to C(\overline{Q})$; cf. [**216**, Proposition 6.1]. For the adaptation of the results from [**216**] to the slightly different equation (Eq) we refer the reader to Appendix A of [**168**] where the respective details have been summarized. In fact, weak-to-strong continuity of the control-to-state map as map $L^s(\Lambda) \to C(\overline{Q})$ is much more than actually needed for the present purpose; however, we will use the full strength of [**216**, Proposition 6.1] when proving existence of optimal control in the state-constrained case in Chapter 2.

1.5.2. First-order conditions. Since (P) is a nonconvex problem, in general, optimality conditions deal with so-called local solutions of (P). As in [**35**] we call $\bar{u} \in U_{\mathrm{ad}}$ a local solution to (P) (in the sense of $L^2(\Lambda)$) if there is some $r > 0$ such that $j(u) \geq j(\bar{u})$ holds for all $u \in \mathbb{B}_r^{L^2(\Lambda)}(\bar{u}) \cap U_{\mathrm{ad}}$. Of course, any global solution to (P) is also a local solution.

For any local solution of (P) the following first-order necessary optimality conditions (FONs) have to hold.

Theorem 1.19 ([**35**], Lemma 4.8)**.** *Let Assumptions 1.5, 1.6 and 1.8 be satisfied and let $\bar{u} \in U_{\mathrm{ad}}$ be a local solution to* (P) *with associated state $\bar{y} = S(\bar{u}) \in \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$. Then, there exists the unique adjoint state $\bar{p} \in \mathbb{W}^{1,r}(I, (L^p, \mathrm{Dom}_{L^p}(-\nabla \cdot \mu \nabla)))$, $r \in (2, \infty)$, such that*

$$\left\{ \begin{array}{rl} \partial_t \bar{y} + \mathcal{A}(\bar{y})\bar{y} = B\bar{u}, & \text{in } L^s(I, W^{-1,p}_{\Gamma_D}), \\ y(0) = y_0, & \text{in } (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})_{1/s',s}, \end{array} \right. \tag{1.13}$$

$$\begin{cases} -\partial_t \bar{p} + \mathcal{A}(\bar{y})^* \bar{p} + \mathcal{A}'(\bar{y})^* = \bar{y} - y_d, & \text{in } L^r(I, L^p), \\ p(T) = 0, & \text{in } (L^p, \mathrm{Dom}_{L^p}(-\nabla \cdot \mu\nabla))_{1/r',r}, \end{cases} \tag{1.14}$$

$$j'(\bar{u})(u - \bar{u}) = \langle B^*\bar{p} + \alpha\bar{u}, u - \bar{u}\rangle_{L^2(\Lambda)} \geq 0, \qquad \forall u \in U_{\mathrm{ad}}. \tag{1.15}$$

This is proven in exactly the same way as in [**35**] utilizing Proposition 1.17 for improved regularity of the adjoint state. For an introduction to the main ideas we refer the reader to, e.g., [**156**, Chapter 1.7.2]. Since the set of admissible controls U_{ad} is convex, the main difficulty during the derivation of first-order necessary optimality conditions for (P) is verifying differentiability of the reduced functional. For a highly nonlinear problem like (P) this requires considerable work; cf. our summary of the respective results from [**35**] in Section 1.3. The state-constrained or sparse cases, (P^{st}) and ($\mathrm{P}_k^{\mathrm{sp}}$), respectively, exhibit additional intrinsic difficulties because a constraint qualification is needed or the functional is not differentiable, respectively; we will address these issues in more detail in Chapters 2 and 3.

We note that one does not require $\alpha > 0$ for the derivation of these first-order conditions. The generalization to the bang-bang case $\alpha = 0$ is straightforward. Moreover, for $\alpha > 0$ it is well-known that the variational inequality (1.15) can be expressed equivalently as

$$\bar{u} = \mathrm{Proj}_{U_{\mathrm{ad}}}(-\alpha^{-1}B^*\bar{p}) = \min\left(u_{\mathrm{b}}, \max(u_{\mathrm{a}}, -\alpha^{-1}B^*\bar{p})\right), \tag{1.16}$$

where min() and max() have to be understood pointwise on Λ. Similarly, in the bang-bang case $\alpha = 0$ it holds

$$\bar{u} = u_a \quad \text{on } \{B^*\bar{p} > 0\} \qquad \text{and } \bar{u} = u_b \quad \text{on } \{B^*\bar{p} < 0\},$$

i.e. the optimal control jumps between the control-bounds u_a and u_b which explains the notion "bang-bang". For the derivation of these pointwise reformulations we refer the reader to, e.g., [**270**, Chapter 2.8].

Before addressing second-order conditions, let us as in [**168**, Example 2.9] state the specific form of the variational inequality (1.15) for the three variants of B discussed in Example 1.9.

Example 1.20.

1. In the case of distributed control we obtain B^* to be the identity map $L^{s'}(I, W^{1,p'}_{\Gamma_D}) \to L^{s'}(Q)$ and (1.15) reads
$$\int_Q (\bar{p} + \alpha\bar{u})(u - \bar{u})\mathrm{d}x\mathrm{d}t \geq 0, \qquad \forall u \in U_{\mathrm{ad}} \subset L^s(Q).$$
2. For Neumann boundary control ($d = 2$) B^* is the trace map $L^{s'}(I, W^{1,p'}_{\Gamma_D}) \to L^{s'}(I \times \Gamma_N)$ and we obtain
$$\int_{I\times\Gamma_N} (\bar{p}|_{I\times\Gamma_N} + \alpha\bar{u})(u - \bar{u})\mathrm{d}s\mathrm{d}t \geq 0 \qquad \forall u \in U_{\mathrm{ad}} \subset L^s(I \times \Gamma_N).$$
3. We obtain
$$\begin{aligned} B^*\bar{p} &= \left(t \mapsto \langle b_i, \bar{p}(t)\rangle_{W^{-1,p}_{\Gamma_D}, W^{1,p'}_{\Gamma_D}}\right)_{i=1}^m \\ &= \left(t \mapsto \int_{\Gamma_N} f_i \bar{p}(t)|_{\Gamma_N}\mathrm{d}s\right)_{i=1}^m \in L^{s'}(I, \mathbb{R}^m), \end{aligned}$$

and

$$\sum_{i=1}^{m} \int_I \left(\int_{\Gamma_N} f_i \cdot \bar{p}(t)|_{\Gamma_N} \mathrm{d}s + \alpha \bar{u}_i(t) \right) (u_i(t) - \bar{u}_i(t)) \mathrm{d}t \geq 0$$

for all $u \in U_{\mathrm{ad}} \subset L^s(I, \mathbb{R}^m)$ in the case of purely time-dependent controls.

1.5.3. Second-order conditions. Let us now turn to second-order conditions. As in finite dimensions there are second-order necessary (SNCs) and second-order sufficient (SSCs) conditions. The first ones are (together with FONs) necessarily fulfilled at a local solution of (P) and consequently characterize such minimizers, while the second ones serve as sufficient criteria for local optimality: since (P) is nonconvex, some $u \in U_{\mathrm{ad}}$ that satisfies the first-order necessary optimality conditions does not need to be a local solution of (P), in general. If, however, u satisfies both FONs and a SSC then u is indeed a local minimizer of (P).

On the surface, this seems to be completely analogous to the finite dimensional theory, but, as we explained in the introduction on behalf of Example (0.2), the topic of second-order conditions in infinite dimensions confronts us with phenomena that are not known from finite dimensions. In Section 1.3 we have explained that the reduced functional of (P) is differentiable w.r.t. the $L^s(\Lambda)$-norm with $s > 2$ as in Assumption 1.8, but not necessarily w.r.t. the $L^2(\Lambda)$-norm. Moreover, it is well-known that coercivity of $j''(\bar{u})$ (see (1.7)) can only be expected to hold w.r.t. $L^2(\Lambda)$ but not w.r.t. $L^s(\Lambda)$[1]. Such a property, i.e. that differentiability and coercivity of the second derivative only hold w.r.t. different norms, often occures in PDE-constrained optimization and is usually referred to as two-norm discrepancy [**174, 73, 71**]. One might expect that a similar norm gap as in (0.2) occures during the formulation of second-order conditions for (P), i.e. that coercivity of $j''(\bar{u})$ w.r.t. $L^2(\Lambda)$ only allows to conclude local optimality w.r.t. $L^s(\Lambda)$, but not w.r.t. $L^2(\Lambda)$. We refer the reader to, e.g., the survey [**73**] for more details on second-order optimality conditions in PDE-constrained optimization.

The second-order conditions proven in both [**35**] and [**45**] avoid the two-norm gap. Their proof is based on results of the paper [**71**] in which second-order conditions in the presence of a two-norm discrepancy, but without two-norm gap, are obtained in an abstract setting that resembles the typical structure of optimal control problems with PDEs and pure control-constraints[2]. We mention that similar arguments were also used in [**70**] in the context of abstract finite element errors. The abstract setting from [**71**] applies to Neumann boundary or Dirichlet boundary control of semilinear elliptic PDEs, distributed control of semilinear parabolic PDEs [**71**], and, as proven in [**35**], to the quasilinear parabolic problem (P).

In essence, due to [**71**] the main work in the proof of SNCs/SSCs for the control-constrained instance of (P) consists of checking the respective properties of the reduced functional of the control problem under consideration. For later reference we summarize the respective results from [**35**] that rely on a careful and detailed analysis of the PDEs appearing in the formulas for j' and j''.

---

[1] In fact, it has been proven in [**35**, Theorem 4.13] that the positivity condition $j''(\bar{u})v^2 > 0$ for all $v \in C_{\bar{u}}$ with the so-called critical cone $C_u \subset L^2(\Lambda)$ already implies $L^2(\Lambda)$-coercivity of $j''(\bar{u})$ on $C_{\bar{u}}$; cf. Proposition 2.9

[2] We will state the precise assumptions of this framework as part 1 of our Assumption 2.7 in Chapter 2 when formulating an extension of [**71**].

Proposition 1.21. *Under Assumptions 1.5, 1.6 and 1.8 the reduced functional satisfies the conditions (A1) and (A2) from* [**71**]*: the reduced functional j: $L^s(\Lambda) \to \mathbb{R}$ is twice continuously Fréchet differentiable, and for every $u \in U_{\mathrm{ad}}$ the first and second derivatives of j at u extend to continuous linear and bilinear forms on $L^2(\Lambda)$, respectively. Moreover, for any sequences $(u_k)_k \subset U_{\mathrm{ad}}$ and $(v_k)_k \subset L^2(\Lambda)$ such that $u_k \to \bar{u}$ strongly in $L^2(\Lambda)$ and $v_k \rightharpoonup v$ weakly in $L^2(\Lambda)$ it holds:*

1. $j'(\bar{u})v = \lim_{k\to\infty} j'(u_k)v_k$,
2. $j''(\bar{u})v \le \liminf_{k\to\infty} j''(u_k)v_k^2$,
3. *if $v = 0$, there is $c > 0$ such that*

$$c \liminf_{k\to\infty} \|v_k\|^2_{L^2(\Lambda)} \le \liminf_{k\to\infty} j''(u_k)v_k^2.$$

The proof hereof can be found in [**35**, Propositions 4.9 and 4.10] and is independent of the improved regularity analysis of the adjoint state in [**35**, Proposition 4.7]. We point out that L^∞-boundedness of U_{ad} is essential for this result. Moreover, the presence of an L^2-Tikhonov cost term, i.e. $\alpha > 0$, is crucial for condition 3. For the concrete formulas for j' and j'' we refer to (1.6) and (1.7).

The following no-gap second-order optimality conditions without two-norm gap for (P) have been obtained in [**35**]. Once the respective assumptions from [**71**] on the reduced functional have been checked, cf. Proposition 1.21, they are an immediate consequence of the abstract results [**71**, Theorems 2.2 and 2.3].

Theorem 1.22 ([**35**],Theorems 4.12 and 4.14)**.** *Let Assumptions 1.5, 1.6 and 1.8 hold. If $\bar{u} \in U_{\mathrm{ad}}$ is a local solution of* (P)*, it holds*

$$j''(\bar{u})v^2 \ge 0 \qquad \forall v \in C_{\bar{u}} \tag{1.17}$$

with $C_{\bar{u}} := \mathrm{cl}_{L^2(\Lambda)}(\mathcal{R}_{U_{\mathrm{ad}}}(\bar{u})) \cap \ker j'(\bar{u})$. Conversely, if some $\bar{u} \in U_{\mathrm{ad}}$ satisfies the first-order condition (1.15) *and*

$$j''(\bar{u})v^2 > 0 \qquad \forall v \in C_{\bar{u}} \setminus \{0\} \tag{1.18}$$

there are $\epsilon, \delta > 0$ such that

$$j(u) \ge j(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|^2_{L^2(\Lambda)} \qquad \forall u \in \mathbb{B}_\epsilon^{L^2(\Lambda)}(\bar{u}) \cap U_{\mathrm{ad}}. \tag{1.19}$$

In particular, $\bar{u}$ is a strict local solution of (P)*.*

Before we put this result into context, let us state the following important observation.

Proposition 1.23 ([**35**], Theorem 4.13)**.** *Under the assumptions of the previous theorem, the positivity condition* (1.18) *is equivalent to the coercivity condition*

$$j''(\bar{u})v^2 \ge \gamma\|v\|^2_{L^2(\Lambda)} \qquad \forall v \in C_{\bar{u}}, \tag{1.20}$$

with some $\gamma > 0$.

Now, to make the context of Theorem 1.22 more clear, let us briefly comment on the notions "no-gap" and "without two-norm gap" characterizing these second-order conditions. In particular, one has to observe that the two formulations refer to different gaps.

We start with the notion "two-norm gap" which has already been explained on behalf of Example 0.2 in the introduction. Unlike in Example 0.2, there is *no* two-norm gap in the above second-order sufficient conditions. Although j is only differentiable w.r.t. the $L^s(\Lambda)$-norm with $s > 2$, cf. the comments below Lemma 1.12, the positivity condition (1.18) and the —due to Proposition 2.9— equivalent coercivity condition (1.20) w.r.t. the $L^2(\Lambda)$-norm allow to deduce the quadratic growth condition (1.19), and hence strict local optimality, on an $L^2(\Lambda)$-neighbourhood. As for Proposition 1.21, L^∞-boundedness of $U_{\rm ad}$ and $\alpha > 0$ are crucial for Theorem 1.22; without L^∞-boundedness of $U_{\rm ad}$ a two-norm gap would be inferred. Second-order conditions for the bang-bang case, i.e. $\alpha = 0$, usually require completely different techniques; cf., e.g., [**78, 84, 59, 286**]. For quasilinear parabolic problems this is, as far as we now, an open problem. We will comment on the specific problems of bang-bang optimal control in the context of quasilinear parabolic problems in more detail in a slightly different setting in Section 3.3.3.

Let us now explain the notion of no-gap second-order conditions, which means the following: the gap between necessary and sufficient second-order conditions is minimal, i.e. the same cone of directions $C_{\bar u}$ appears in (1.17) and (1.20). This is due to the fact that the abstract second-order conditions from [**71**] are of no-gap-type if the admissible set is polyhedric; see, e.g., Chapter 0 for the definition. In the present case, $U_{\rm ad}$ is indeed polyhedric and the critical cone $C_{\bar u}$ may be expressed equivalently as follows:

$$C_{\bar u} = {\rm cl}_{L^2(\Lambda)}(\mathcal{R}_{U_{\rm ad}}(\bar u) \cap \ker j'(\bar u)) = \{v \in L^2(\Lambda)\colon \text{(1.21) holds}\},$$

with

$$\text{(1.21)}\quad v \geq 0 \text{ on } \{\bar u = u_{\rm a}\},\ v \leq 0 \text{ on } \{\bar u = u_{\rm b}\},\ \text{ and } v \equiv 0 \text{ on } \{\bar u = -\alpha^{-1}B^*\bar p\};$$

cf. [**35**, Proposition 4.11]. For recent no-gap second-order conditions for control-constrained optimal control of a nonsmooth quasilinear elliptic PDE we refer the reader to [**87**].

In Chapters 2 and 3 we will prove second-order sufficient conditions for problems with additional state-constraints and second-order necessary and sufficient conditions for sparse problems. To put our own work into the context of the abovementioned state of the art results concerning pure control-constraints, let us note that the second-order conditions proven in this thesis avoid the two-norm gap, too. Our second-order conditions for sparse problems in Chapter 3 are even of no-gap-type as well. Whether this is also true for the second-order sufficient conditions for state-constrained problems in Chapter 2 is not clear because we did not obtain complementing necessary conditions. We will explain this in more detail at the end of Section 2.2 where also literature concerned with no-gap optimality conditions in the case of state-constraints will be discussed. To prove our second-order results, we follow a similar approach as in [**35, 45**]. We develop extensions of the abstract framework from [**71**] that allow us to handle the respective problems. In particular, when applying our abstract results to the concrete problems ($\mathrm{P}^{\rm st}$) and ($\mathrm{P}_k^{\rm sp}$) we will make use of Proposition 1.21 to check the conditions of the reduced functional (Chapter 2) or of the smooth part of the reduced functional (Chapter 3). Finally, let us mention that the analysis of the SQP method in Chapter 4 will rely on assuming certain second-order sufficient conditions that need to be stronger than those from Theorem 1.22. In fact, one has to assume coercivity of $j''(\bar u)$ on a

subspace of $L^2(\Lambda)$ strictly larger than the critical cone $C_{\bar{u}}$. However, we will put particular emphasis on keeping this subspace as small as possible.

1.6. Improved regularity of the state

We conclude this chapter by mentioning an improved regularity result for the solutions of the state equation. In [**45**] a problem similar to (P) but with slightly different assumptions w.r.t. the domain, the boundary conditions, and the coefficient functions has been discussed. In particular, first-order necessary as well as second-order necessary and sufficient optimality conditions analogous to those from [**35**] summarized in the previous sections have been derived.

The authors of [**45**] consider a more regular setting for the underlying equation than in [**35, 216**], but in return they obtain improved regularity results for the states in $W^{2,p}$-spaces, include a zero-order semilinear term and allow for possibly unbounded nonlinearities. In the same setting they also derive finite element discretization errors for the state equation in [**46**]. We will need such higher regularity for the states when addressing second-order sufficient optimality conditions for a problem with additional pointwise state-constraints in Chapter 2. Therefore, we provide a summary of the respective regularity result. A version hereof and the underlying assumptions adapted to our particular setting will be stated in Theorem 2.20.

Theorem 1.24 ([**45**], Theorem 2.3). *Let $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, be a domain with $C^{1,1}$-boundary, $s,p \in [2,\infty)$ such that $\frac{1}{s} + \frac{d}{p} < 2$, and let the coefficient functions $a\colon \overline{Q} \times \mathbb{R} \to \mathbb{R}$ and $a_0\colon Q \times \mathbb{R} \to \mathbb{R}$ have the following properties:*

1. *a is continuous and*
   - *there there is $a_\bullet > 0$ such that*

     $$a(x,t,y) \geq a_\bullet \qquad \forall (t,x,y) \in \overline{Q} \times \mathbb{R},$$

   - *and for each $M > 0$ there is $L_M > 0$ such that*

     $$|a(t,x_2,y_2) - a(t,x_1,y_1)| \leq L_M(|x_2 - x_1| + |y_2 - y_1|)$$

     holds for all $x_1, x_2 \in \overline{\Omega}$, $t \in I$, and $y_1, y_2 \in \mathbb{R}$ such that $|y_1|, |y_2| \leq M$,
2. *a_0 is a Carathéodory function, i.e. measurable w.r.t. the first two arguments and continuous w.r.t. the third argument,*
   - *monotone nondecreasing w.r.t. y,*
   - *for each $M > 0$ there is $L_{0,M} > 0$ such that*

     $$|a_0(t,x,y_2) - a_0(t,x,y_1)| \leq L_{0,M}|y_2 - y_1|$$

     holds for all $(t,x) \in Q$ and all $y_1, y_2 \in \mathbb{R}$ such that $|y_1|, |y_2| \leq M$,
   - *and $a_0(\cdot,\cdot,0) \in L^{2s}(I, L^p)$.*

Then, for each $u \in L^{2s}(I, L^p)$ *and* $y_0 \in C(\Omega) \cap (W^{-1,2p}_{\Gamma_D}, W^{1,2p}_{\Gamma_D})_{1-1/(2s),2s} \cap (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})_{1/s',s}$, *the equation*

$$
(1.22)\quad \begin{cases} \partial_t y(t,x) - \nabla \cdot a(t,x,y(t,x))\nabla y(t,x) \\ \qquad\qquad + a_0(t,x,y(t,x)) = u(t,x) & \forall (t,x) \in Q, \\ \qquad\qquad y(t,x) = 0 & \forall (t,x) \in I \times \partial\Omega, \\ \qquad\qquad y(t,x) = y_0(x) & \forall x \in \Omega, \end{cases}
$$

has a unique solution $y \in \mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$.

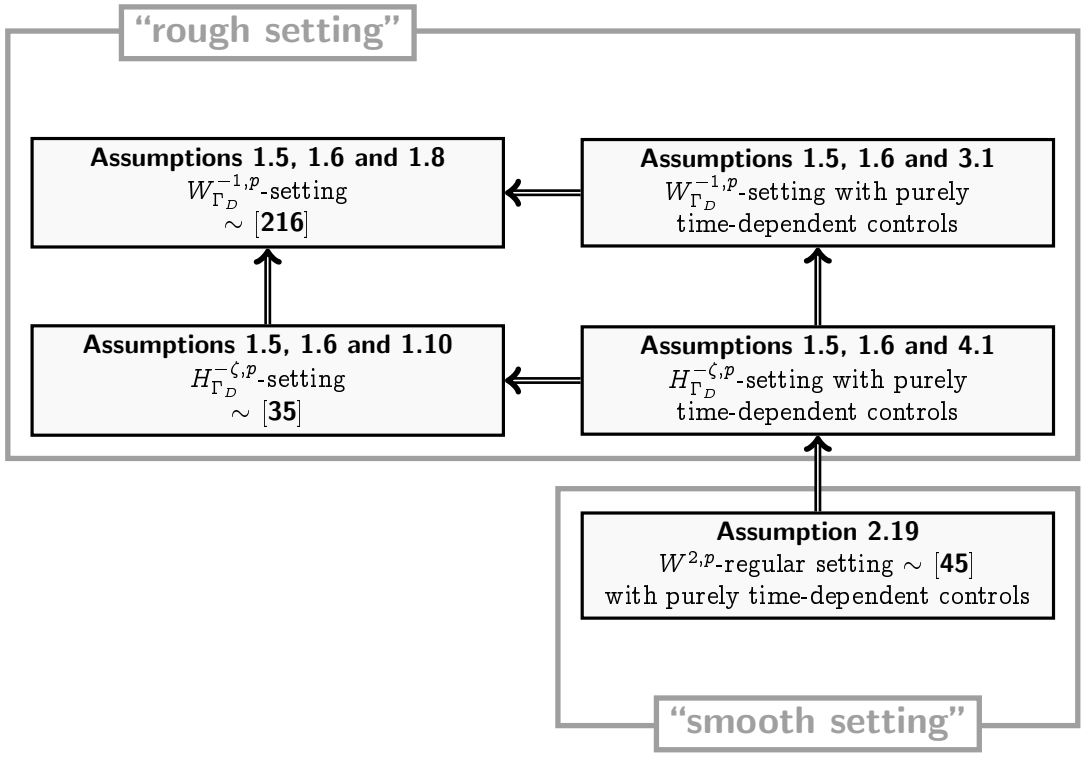


Figure 1.1. Relation of the different regularity assumptions.

	Assumption						
Results	1.5	1.6	1.8	1.10	3.1	4.1	2.19
State equation on $W_{\Gamma_D}^{-1,p}$ (Theorem 1.11)	×	×	×				
State equation on $H_{\Gamma_D}^{-\zeta,p}$ (Theorem 1.14)	×	×		×			
State equation on L^p with $W^{2,p}$-regularity (Theorem 2.20)							×
Adjoint equation on $W_{\Gamma_D}^{-1,p'}$ (Proposition 1.15)	×	×		×			
Adjoint equation on $H_{\Gamma_D}^{-\zeta,p}$ (Proposition 1.16)	×	×		×			
Adjoint equation on L^p (Proposition 1.17)	×	×	×				
FONs for pure control-constraints (Theorem 1.19)	×	×	×				
SNCs/SSCs for pure control-constraints (Theorem 1.22)	×	×	×				
FONs for state-constraints (Theorems 2.4 and 2.17)	×	×	×				
SSCs for averaged state-constraints (Theorem 2.18)	×	×	×				
SSCs for pointwise state-constraints (Theorem 2.29)							×
FONs/SNCs/SSCs for sparse optimal control (Theorems 3.3 and 3.4)	×	×			×		
Convergence of the SQP method (Theorems 4.15 and 4.25)	×	×				×	
A-posteriori POD errors approach I* (Theorem 5.6)							×
A-posteriori POD errors approach II* (Theorem 5.9)	×	×		×			

Table 1.1. Regularity assumptions required for the main results. We only list assumptions related to the regularity setting for the state equation; additional assumptions such as, e.g., the properties of additional state-constraints or the existence of a linearized Slater point (Theorem 2.4) etc., are not considered here. (*As will be explained in detail at the beginning of Section 5.2.2 and right before Theorem 5.9, our results on a-posteriori POD error estimates actually do not rely on these assumptions. However, they rely on quantities of which we believe that they can only be expected to be independent of discretization under these assumptions and the resulting regularity results.)

Chapter 2

Pointwise constraints on the state

This chapter is based on the work [**168**] by I. Neitzel and the author. We modify problem (P) from Chapter 1 by adding pointwise constraints on the state-variable. More precisely, we consider the following model problem:

$$(\mathrm{P}^{\mathrm{st}}) \qquad \begin{cases} \min J(y,u) := \dfrac{1}{2}\|y - y_d\|^2_{L^2(I\times\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(\Lambda)}, \\ \quad \text{s.t.} \quad u \in U_{\mathrm{ad}},\ y \in Y_{\mathrm{ad}}, \\ \quad\quad\ \text{and (Eq)}. \end{cases}$$

The set of admissible states Y_{ad} will be clarified in each section, and is given either by pointwise in space and time inequality-constraints, i.e.

$$Y_{\mathrm{ad}} = \left\{ y \in C(\overline{Q}):\ \ y_{\mathrm{a}}(t,x) \le y(t,x) \le y_{\mathrm{b}}(t,x)\ \forall (t,x) \in \overline{Q} \right\},$$

or, if we require a weaker type of constraints for our analysis, by pointwise in space and averaged in time bounds of type

$$Y_{\mathrm{ad}} = \left\{ y \in L^1(I, C(\overline{\Omega})):\ \ y_{\mathrm{a}}(x) \le \int_I y(t,x)\ \mathrm{d}t \le y_{\mathrm{b}}(x)\ \forall x \in \overline{\Omega} \right\}.$$

We prove existence of optimal controls and derive first-order necessary optimality conditions (FONs) under the rather general assumptions on the state equation from Section 1.2 and pointwise in space and time state-constraints. Under additional assumptions we provide second-order sufficient optimality conditions (SSCs). For the rough setting from Section 1.2 we restrict the analysis to pointwise in space and averaged in time state-constraints. Pointwise in space and time state-constraints are discussed for a more regular state equation as in Section 1.6 and purely time-dependent controls utilizing the improved regularity analysis from [**45**].

As we have explained in the introduction, PDE-constrained optimiziation problems with pointwise state-constraints are as interesting as challenging. The state equation under consideration in the present work models, e.g., heat conduction in material with temperature-dependent thermal conductivity and therefore state-constrained problems are certainly conceivable, e.g., in applications in which the temperature of an object has to be kept between prescribed bounds. Since the field of optimal control with pointwise state-constraints gained much attention in the last years the following literature overview focusses on optimality conditions and excludes topics less related to the present work. For a brief summary of literature concerned with discretizational and numerical aspects we refer the reader to Section 2.5 at the end of this chapter.

In order to derive first-order optimality conditions for problems with pointwise state-constraints one usually needs continuity of the states in order to fulfill a Slater-type constraint qualification. This results in the presence of regular Borel measures, i.e. objects of low regularity, as Lagrange multipliers in the KKT system; see, e.g., [**40, 41**] for early work concerned with linear and semilinear elliptic problems. For an alternative approach with even slightly less regular multipliers in a convex setting we refer to [**248**]. Moreover, we mention that under certain conditions improved regularity of the Lagrange multipliers has been obtained for linear elliptic problems [**63**], and that recent work [**83**] on parabolic problems avoids the usage of a Slater-type condition by a reformulation of the problem involving a variational inequalty. The case of mixed control-state-constraints, in which one imposes pointwise bounds on terms like, e.g., $\epsilon u + y$ with $\epsilon \in \mathbb{R}$, is known to be different because in that case the Lagrange multipliers are given by integrable functions under natural assumptions; see, e.g., [**18**] for semilinear parabolic problems or [**239**] and the references therein.

For control problems with nonlinear PDEs, SSCs are important because FONs are not sufficient in general. Regarding an overview on second-order optimality conditions in PDE-constrained optimization we refer the reader to, e.g., the survey [**73**] and the references therein. State of the art no-gap second-order conditions for quasilinear parabolic optimal control problems in the presence of pure control-constraints [**35, 45**] have been summarized and put into context in Section 1.5.3. Let us briefly point out a few more aspects particularly related to our work. The first publication addressing SSCs for parabolic problems in the presence of additional state-constraints is, to the best of our knowledge, [**113**]. As already mentioned in the introduction, a specific difficulty arising in the second-order analysis of PDE-constrained optimization is the two-norm discrepancy [**174, 71**]: differentiability of the reduced functional and coercivity of its second derivative often only hold w.r.t. different norms. In any case, a careful regularity analysis of the underlying PDE is necessary. Especially in the presence of state-constraints this often leads to certain restrictions or additional assumptions that are necessary in order to guarantee the required regularity. Let us briefly recall some examples from the literature: the derivation of SSCs for semilinear elliptic problems usually requires the restriction to space dimension $d \leq 3$ in the case of distributed and to dimension $d \leq 2$ in the case of Neumann boundary control [**49, 133**]. With the same techniques, distributed control of semilinear parabolic PDEs can be handeled for space dimension $d = 1$ only [**49**], but the restriction on the space dimension can be lifted if one in return considers purely time-dependent controls [**91**]. A careful analysis utilizing the concept of maximal parabolic regularity has been used in [**189**] to prove SSCs for distributed control of semilinear parabolic PDEs in space dimension $d = 2$ and 3. However, these SSCs are different from those of [**91, 49**] because they are postulated w.r.t. a larger cone of directions that includes some first-order information. Alternatively, the regularity requirements of the problem type can be weakened by switching from pointwise in space and time state-constraints to certain averaged-type [**113**] or finitely many [**37, 75**] state-constraints. In particular, averaged-type state-constraints turned out to be a fertile concept, also in other areas of PDE-constrained optimization, whenever dealing with purely pointwise

state-constraints is not possible for theoretical reasons; see, e.g., [**211, 204**] for their application in the context of discretization error estimates.

Regarding second-order necessary optimality conditions (SNCs) for pure state-constraints we only mention [**184**], as well as both SNCs and SSCs with emphasis on a possibly small gap between them in [**258, 185**] for the different setting of pointwise mixed control-state-constraints. Finally, we cite [**84, 287, 288**] for SNCs and SSCs in an abstract optimization-theoretic setting. More literature on SNCs will be given at the end of Section 2.2.

As far as we know, there are only few results on state-constrained optimal control problem with quasilinear parabolic PDEs [**109, 110, 214, 215, 168**]. As already pointed out in the introduction, the nontrivial existence and regularity theory for solutions of the quasilinear PDE and its linearizations poses the main difficulty of such problems. Hence, combining them with additional state-constraints —that are known to require a deeper regularity analysis than purely control-constrained problems, in general— is both interesting and challenging. The early papers [**109, 110**] address existence of optimal controls and FONs for a problem with averaged in space and pointwise in time, or finitely many state-constraints of integral-type, respectively. Optimal control of the thermistor problem, a coupled system consisting of a quasilinear parabolic and a nonlinear elliptic equation, with pointwise in space and time state-constraints is addressed in [**214, 215**].

In this chapter we present results that have been obtained in [**168**]. First, we establish existence of optimal controls and FONs for (P^{st}) in the presence of pointwise in space and time state-constraints, extending the results on the purely control-constrained problem (P) from [**35, 45**] that have been summarized in Chapter 1. In particular, let us emphasize again that the underlying regularity assumptions on the problem data (see Section 1.2) are fairly general: our first-order results for pointwise in space and time state-constraints hold in a rough setting, including certain nonsmooth domains, nonsmooth coefficient functions, and mixed boundary conditions, and allow for different types of control mechanisms, such as distributed control ($d = 2, 3$), Neumann boundary control ($d = 2$ only), or purely time-dependent controls with actuator functions from $W_{\Gamma_D}^{-1,p}$ ($d = 2, 3$); cf. Examples 1.7 and 1.9. This is different from, e.g., the more restrictive regularity assumptions in [**109, 110**]. In essence, to obtain first-order conditions we have to apply the classical abstract Theorem 0.1 to our concrete setting; the appropriate choice of the underlying function spaces and having at hand suitable regularity results for the respective PDEs are the main difficulty. Fortunately, we can build on results from [**35, 216**] that we have summarized in Chapter 1.

The second part of our results from [**168**] deals with SSCs under different additional assumptions. As one may expect with a view to the literature overview provided above, some restrictions cannot be avoided when aiming at SSCs for a state-constrained optimal control problem with a time-dependent, highly nonlinear PDE. Nevertheless, it is one of the goals of our work to investigate how far the analysis of (P^{st}) can be performed within the rough setting, before regularity requirements force us to switch to a different setup. Essentially, for both choices of Y_{ad} stated at the beginning of this chapter we prove SSCs under natural assumptions: first, we keep the rough setting and weaken the pointwise in space and time state-constraints towards such that are pointwise in space but only integral w.r.t.

time. As we already pointed out above, switching to such averaged-type state-constraints is a well-known trick [**113, 211, 204**]. However, due to the regularity results available in the rough setting we have to average w.r.t. time, instead of taking averages w.r.t. space, which is diffent from the aforementioned results from the literature. Second, we keep the pointwise in space and time state-constraints and, in return, strengthen our regularity assumptions towards a classical, smooth setting in the flavour of [**45**]. In both cases, a careful investigation of the state equation and its linearizations is crucial crucial for obtaining our results.

Let us point out another important feature of our work. As we have explained in the introduction chapter, we pay particular attention to avoiding the so-called two-norm gap when addressing SSCs in this thesis. The phenomena two-norm discrepancy and two-norm gap, that are typical for the second-order analysis of PDE-constrained optimization problems, have been illustrated on behalf of Example (0.2) in the introduction chapter; see also our remarks in Section 1.5.3. In our results we can avoid a two-norm gap in the formulation of SSCs although a two-norm discrepancy is present in the problem formulation. Recall from our summary in Chapter 1 that the proof of SNCs and SSCs for the purely control-constrained problem (P) in [**35, 45**] is based on an application of abstract results of the paper [**71**]. There, second-order conditions avoiding the two-norm gap despite the presence of two norms in the problem data were proven in an abstract framework that resembles the structure of control-constrained problems. We extend the abstract framework from [**71**] in such a way that, as a consequence on the concrete level, we obtain SSCs for both the pointwise in space and time and the pointwise in space and averaged in time state-constrained instace of (P^{st}). Doing so, we make use of techniques from [**71**] and abstract ideas known from, e.g., [**91, 49**]. Let us note that the chosen abstract approach has the advantage that it also facilitates the application to other problem settings: as a byproduct, e.g., our abstract result allows to reformulate the SSCs for semilinear parabolic problems from [**91**], that originally contained a two-norm gap, without this norm gap.

This chapter is structured as follows: In Section 2.1 we introduce the problem setting, prove existence of optimal controls, and derive FONs in this rather general context. In Section 2.2 we prove SSCs for an abstract optimization problem extending the result from [**71**]. In Section 2.3 we explain why our abstract result from Section 2.2 does not apply to the model problem as stated in Section 2.1. Then, we prove SSCs without two-norm gap for a modified version of our model problem where the regularity assumptions remain unchanged but the pointwise state-constraints are replaced by averaged in time state-constraints. In Section 2.4 we come back to pointwise state-constraints and prove SSCs for this situation, but now assuming a more regular setting as in Section 1.6 and purely time-dependent controls. In the final section we provide some numerical examples.

2.1. Existence of optimal controls and first-order necessary conditions

The problem (P^{st}) under consideration has been stated at the beginning of this chapter. Note that compared to the control-constrained problem (P) from the previous chapter we do not change the underlying state equation (Eq). Also, the regularity Assumptions 1.5, 1.6 and 1.8 on the underlying data remain unchanged

until further notice. For the newly introduced state-constraints we choose the following setting, which remains unchanged until noted otherwise, too.

Assumption 2.1.

1. The set of admissible states is given by
$$Y_{\mathrm{ad}} = \{y \in C(\overline{Q}) : y_{\mathrm{a}}(t,x) \le y(t,x) \le y_{\mathrm{b}}(t,x)\ \forall (t,x) \in \overline{Q}\},$$
with bounds $y_{\mathrm{a}}, y_{\mathrm{b}} \in C(\overline{Q})$ satisfying $y_{\mathrm{a}}(t,x) < y_{\mathrm{b}}(t,x)$ for all $(t,x) \in \overline{Q}$, $y_{\mathrm{a}}(t,x) \le 0 < y_{\mathrm{b}}(t,x)$ for all $(t,x) \in \overline{I}\times\Gamma_D$, and $y_{\mathrm{a}}(0,x) < y_0(x) < y_{\mathrm{b}}(0,x)$ for $x \in \overline{\Omega}$. We allow for $y_{\mathrm{a}} \equiv -\infty$ or $y_{\mathrm{b}} \equiv +\infty$.
2. There is a feasible point, i.e. there is $(y,u) \in Y_{\mathrm{ad}} \times U_{\mathrm{ad}}$ such that y and u fulfill the state equation (Eq).

A compatibility condition like Assumption 2.1.2 is difficult to check a-priori on behalf on the problem data but it cannot be avoided, in general. Roughly speaking, we need to assume that the set of controls satisfying all constraints imposed in (P^{st}) is not empty. For the construction of a feasible control-state pair in the particular case of a linear elliptic PDE, distributed control, and pure state-constraints we refer the reader to, e.g., [**63**, Theorem 2.3]. For our setting we are not aware of analogous results in the presence of additional control-constraints or a general control operator B.

2.1.1. Existence of optimal controls. For completeness, we start with the proof of existence of a minimizer for (P^{st}). The respective results for the case with control-constraints only from [**35, 216**] have been summarized in Section 1.5. Moreover, let us note that an analogous result for the state-constrained thermistor problem has already been obtained in [**214**].

Proposition 2.2 ([**168**], Theorem 2.6)**.** *Let Assumptions 1.5, 1.6, 1.8, and 2.1 hold. Then there exists a globally optimal control $\bar{u} \in U_{\mathrm{ad}}$ for the optimal control problem* (P^{st}).

Proof. This is analogous to Proposition 1.18 with two exceptions: first, existence of an infimizing sequence now needs to be ensured by assuming the existence of a feasible point (Assumption 2.1.2). Second, we need the full strength of [**216**, Proposition 6.4] where weak-to-strong continuity of the control to state map as map $L^s(\Lambda) \to C(\overline{Q})$ has been shown in order to see that a subsequence of the corresponding sequence of states converges in $C(\overline{Q})$. Since Y_{ad} is closed in $C(\overline{Q})$ the limit is still in Y_{ad}, i.e. it fulfills the state-constraints. □

For comments on how Proposition 2.2 and all further results in this chapter change in the case without or with only unilateral control-constraints we refer the reader to Section 6 of our paper [**168**].

2.1.2. First-order necessary optimality conditions. We closely follow [**168**, Section 2.4] in this section. We characterize local solutions of (P^{st}) that fulfill a Slater-type constraint qualification by FONs. The need for such a constraint qualification is the main difference to the case without state-constraints summarized in Theorem 1.19. As we have explained in the introduction, the following main difficulty is well-known in state-constrained optimal control of PDEs; cf. e.g. [**40, 41**]. To use a Slater-type condition, we have to ensure nonempty interior of Y_{ad} in the

respective space. Since $Y_{\rm ad}$ is defined by pointwise inequality-constraints, this excludes states in $L^q(Q)$, $1 \le q < +\infty$. We have to consider them in $C(\overline{Q})$, which infers regular Borel measures, i.e. the corresponding dual objects, in the KKT system. To apply the abstract Theorem 0.1 for optimization problems in Banach spaces to our problem $(\mathrm{P}^{\rm st})$ we formulate an additional assumption.

Assumption 2.3. Let $\bar u \in U_{\rm ad}$ be an $L^2(\Lambda)$-local solution to $(\mathrm{P}^{\rm st})$ with associated state $\bar y = S(\bar u) \in Y_{\rm ad}$, i.e. there is $\epsilon > 0$ such that $j(u) \ge j(\bar u)$ for all $u \in \mathbb{B}_\epsilon^{L^2(\Lambda)}(\bar u) \cap U_{\rm ad}$ fulfilling $S(u) \in Y_{\rm ad}$. Further, assume that the following linearized Slater condition is fulfilled at $\bar u$: there is $u_{\rm Sl} \in U_{\rm ad}$ such that $\bar y + S'(\bar u)(u_{\rm Sl} - \bar u) \in \mathring{Y}_{\rm ad}$, i.e.

$$y_{\rm a}(t,x) < \bar y(t,x) + S'(\bar u)(u_{\rm Sl} - \bar u)(t,x) < y_{\rm b}(t,x) \quad \forall (t,x) \in \overline{Q}.$$

Since the L^s-norm is stronger than the L^2-norm every $L^2(\Lambda)$-local solution is in particular also an $L^s(\Lambda)$-local solution. In general, it will be extremely difficult to check only on behalf of the problem data whether a local or global solution of $(\mathrm{P}^{\rm st})$ satisfies the linearized Slater condition. However, assuming a constraint qualification can usually not be avoided when aiming at first-order necessary optimality conditions. For the constructive verification of the linearized Slater condition in the special case of optimal control of a semilinear elliptic equation with pure state-constraints we refer to, e.g., [**222**, Lemma 2.11].

Theorem 2.4 ([**168**], Theorem 2.8)**.** *Under Assumptions 1.5, 1.6, 1.8, 2.3, and 2.1.1 there exists a regular Borel measure $\bar\lambda \in \mathcal{M}(\overline{Q}) = C(\overline{Q})^*$ on $\overline{Q}$ and the so-called adjoint state $\bar p \in L^{r'}(I, W^{1,p'}_{\Gamma_D})$, $r' \in (1, \frac{2p}{p+d})$, such that the optimality system*

$$\begin{cases} \partial_t \bar y + \mathcal{A}(\bar y)\bar y = B\bar u, \\ \bar y(0) = y_0, \end{cases} \tag{2.1}$$

$$\begin{cases} -\partial_t \bar p + \mathcal{A}(\bar y)^* \bar p + \mathcal{A}'(\bar y)^* \bar p = \bar y - y_d + \bar\lambda, \\ \bar p(T) = 0, \end{cases} \tag{2.2}$$

$$\langle \bar\lambda, y - \bar y\rangle_{\mathcal{M}(\overline{Q}), C(\overline{Q})} \le 0 \qquad \text{for all } y \in Y_{\rm ad}, \tag{2.3}$$

$$\langle B^*\bar p + \alpha \bar u, u - \bar u\rangle_{L^{s'}(\Lambda), L^s(\Lambda)} \ge 0 \qquad \text{for all } u \in U_{\rm ad} \tag{2.4}$$

is satisfied. The so-called adjoint equation (2.2) *has to be understood in the sense outlined in the proof below; cf. also Remark 2.6.*

For the concrete form of B^* and the respective variational inequality (2.4) we refer the reader to Example 1.20.

Proof. In Theorem 0.1 choose $U = L^s(\Lambda)$, $Z = C(\overline{Q})$, $g = S$, $K = U_{\rm ad}$ and $C = Y_{\rm ad}$. Note that the embedding $\mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow C(\overline{Q})$, cf. (1.4), ensures that the control-to-state operator maps $L^s(\Lambda)$ into $C(\overline{Q})$. It holds

$$\tilde S'(B\bar u) \in \mathcal{L}(L^r(I, W^{-1,p}_{\Gamma_D}), \mathbb{W}^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})), \qquad r \in (1, s];$$

cf. Lemma 1.12.2. Employing the second embedding from Lemma 1.13, i.e. $\mathbb{W}^{1,r}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D}))\hookrightarrow C(\overline{Q})$ for $r\in(\frac{2p}{p-d},\infty)$, we obtain

$$\tilde{S}'(B\bar{u})\in\mathcal{L}(L^r(I,W^{-1,p}_{\Gamma_D}),C(\overline{Q}))$$

for those r, and consequently

$$\tilde{S}'(B\bar{u})^*\in\mathcal{L}(\mathcal{M}(\overline{Q}),L^{r'}(I,W^{1,p'}_{\Gamma_D})),\qquad r'\in\left(1,\frac{2p}{p+d}\right). \tag{2.5}$$

Following the usual adjoint technique in optimal control, see e.g. [**270**, Chapter 6.2.1], we introduce the adjoint state $\bar{p}:=\tilde{S}'(B\bar{u})^*(\bar{y}-y_d+\bar{\lambda})$. Note that $\bar{p}$ is well-defined in this way and exhibits the regularity stated in the theorem due to (2.5) and $\bar{y}-y_d+\bar{\lambda}\in\mathcal{M}(\overline{Q})$. The adjoint equation (2.2) has to be understood purely formal, in the very-weak/adjoint sense. We discuss this further in Remark 2.6 below. Combining equation (1.6) for the reduced gradient of our particular setting with the abstract variational inequality (0.4) and the definition of $\bar{p}$ yields (2.4). □

Some remarks are in order to put this into context. The most obvious difference of Theorem 2.4 compared to the corresponding result for the control-constrained problem (P) in Theorem 1.19 is the so-called complementary slackness condition (2.3). For completeness, we state the following well-known observation.

Remark 2.5. The complementary slackness condition (2.3) can be rewritten in a more illustrative way. The Jordan decomposition $\bar{\lambda}=\bar{\lambda}^+-\bar{\lambda}^-$ into non-negative measures $\bar{\lambda}^+,\bar{\lambda}^-\geq 0$, cf. [**242**, Chapter 6], satisfies

$$\begin{aligned}\operatorname{supp}\bar{\lambda}^+&\subset\{(t,x)\in\overline{Q}\colon\ \bar{y}(t,x)=y_{\mathrm{b}}(t,x)\},\\ \operatorname{supp}\bar{\lambda}^-&\subset\{(t,x)\in\overline{Q}\colon\ \bar{y}(t,x)=y_{\mathrm{a}}(t,x)\}.\end{aligned}$$

For a proof we refer to, e.g., [**63**, Proposition 2.5]. We note that the adaptation from the elliptic setting in [**63**] is straightforward because except for the bounds defining Y_{ad} no specific problem data of (P^{st}) appear in (2.3).

Next, let us point out some issues related to the appearance of the Borel measure $\bar{\lambda}$ on the right-hand side of the adjoint equation (2.2). These problems are typical for parabolic control problems with pointwise in space and time state-constraints.

Remark 2.6. The adjoint equation (2.2) has to be understood purely formal. In general, it is not guaranteed that $\bar{p}\in L^{r'}(I,W^{1,p'}_{\Gamma_D})$ has a distributional time derivative or a well-defined trace on $\{T\}\times\Omega$. Hence, (2.2) really only serves as a more illustrative and intuitive notation for the precise definition of $\bar{p}$ given by $\bar{p}=\tilde{S}'(B\bar{u})^*(\bar{y}-y_d+\bar{\lambda})$. The notation as backward parabolic PDE is motivated by the fact that $\tilde{S}'(B\bar{u})^*$ restricted to the spaces $L^{r'}(I,W^{-1,p'}_{\Gamma_D})$, $r'\in(1,\infty)$, can be identified with the solution map of the respective backward nonautonomous parabolic PDE; cf. Section 1.4. Moreover, the presence of mixed boundary conditions in the state equation does not pose additional difficulties, see, e.g., [**189, 133, 162**], in particular because the support of $\bar{\lambda}$ is disjoint from $\overline{I}\times\Gamma_D$ and $\{0\}\times\overline{\Omega}$; cf. Remark 2.5 and Assumption 2.1.1.

Because $\bar\lambda$ is, in general, only a Borel measure, we cannot improve regularity of the adjoint state $\bar p$ along the lines of [**35**, Proposition 4.7], cf. Proposition 1.15. However, we mention that improved regularity for adjoint states in state-constrained optimal control has been obtained under additional assumptions and with different techniques in case of linear and semilinear elliptic [**63**] and parabolic [**83**] PDEs.

The fact that the adjoint equation has to be understood purely formal as pointed out in Remark 2.6 is not a too severe problem. Recall, e.g., from our summary in Chapter 1 that the second-order analysis for the control-constrained problem (P) in [**35**] has been carried out completely in the $W^{-1,p}_{\Gamma_D}$-$W^{1,p}_{\Gamma_D}$-setting, i.e. in a setting where in [**35**] no adjoint equation was available; cf. in particular our comments below Proposition 1.15.

The main obstruction in our further analysis arises from the extremely low regularity of $\bar p$, which is a typical issue when dealing with state-constrained problems: in fact, due to $\frac{2p}{p+d} < 2$ and $p' < 2$ Theorem 2.4 shows rather poor temporal and spatial regularity for $\bar p$. This difficulty has to be overcome during the analysis of second-order optimality conditions for (P^{st}) as we will outline in Sections 2.3 and 2.4. To do so, we will either have to modify the type of state-constraints (Assumption 2.13) or assume a more regular setting for the state equation (Assumption 2.19).

2.2. An abstract result on second-order sufficient conditions

The presentation of this section follows Section 3 of the underlying paper [**168**]. We extend the abstract framework of [**71**] towards inclusion of state-constraints, i.e. we give SSCs for an abstract optimization problem similar to the one from Theorem 0.1, but now enriched with two norms as typical for PDE-constrained optimization. However, we prove SSCs that avoid the two-norm gap. The framework is developed having in particular the setting and the arguments from [**91**] in mind. We start by introducing the abstract problem

$$(\mathrm{AP}^{\mathrm{st}}) \qquad \min j(u) \qquad \text{s.t. } u \in K, \quad g(u) \in C,$$

with the assumptions given below. The suppositions on the real-valued functional j and the underlying spaces U_2, U_∞, respectively, are identical to those from [**71**] and thus fulfilled for the functional of our quasilinear parabolic problem as we have seen in Proposition 1.21. The second-order conditions for the control-constrained problem (P) summarized in Section 1.5.3 have been proven in [**35, 45**] by an application of the abstract framework from [**71**]. Now, we extend this work in such a way that, on the concrete level, we are able to obtain SSCs for the state-constrained problem (P^{st}).

More precisely, we extend [**71**] towards the inclusion of a state-constraint-like constraint of type "$g(u) \in C$", cf. ($\mathrm{AP}^{\mathrm{st}}$), that is formulated in a further Banach space Z. For instance, choosing g to be the control-to-state map allows to handle state-constraints. Since the set $K \cap g^{-1}(C)$ is nonconvex in general, this situation is not covered by the results of [**71**]. Further, j and g are differentiable w.r.t. the U_∞-norm, but not necessarily w.r.t. the weaker U_2-norm. We have in mind the case $U_2 = L^2(\Sigma, m)$ and $U_\infty = L^p(\Sigma, m)$ with some $p \in (2, \infty]$ for a measure

space $(\Sigma, \mathrm{d}m)$. The presence of such two norms is typical for PDE-constrained optimization as we have explained in Section 1.5.3.

Let us briefly put our result into context. As far as we know, other results on SSCs for state-constraints without two-norm gap required differentiability of j and g w.r.t. L^2; cf. [**258**, Section 4], [**49**, Theorem 4.3] — an assumption that we can now avoid. In particular we can state SSCs for the same semilinear parabolic optimal control problem as in [**91**], but without two-norm gap; see Example 2.10 below. In [**84**] both SNCs and SSCs for certain optimization problems in infinite dimensions are proven. The results rely on the concept of a directional curvature functional for the (possibly nonconvex) admissible set. The authors state that it is possible to include cases with two-norm discrepancy (see Remark 4.6.iv), but the special case of the present chapter and [**71**], in which such a discrepancy appears but can be avoided in the formulation of second-order conditions, is not addressed. Further, the explicit computation of the directional curvature term in the presence of pointwise state-constraints is left as topic of further research. We believe that our approach, explicitly tailored to situations as, e.g., (P^{st}), [**91**], and [**71**], respectively, is of independent interest.

Assumption 2.7. Let U_2 be a Hilbert space and U_∞ a Banach space such that there is a continuous embedding $U_\infty \hookrightarrow U_2$. With $\|\cdot\|_2$ and $\|\cdot\|_\infty$ we denote the corresponding norms. Moreover, $\langle\cdot,\cdot\rangle_2$ is the duality product in $U_2^* \times U_2$. Further, let Z be a Banach space with norm $\|\cdot\|_Z$ and duality pairing $\langle\cdot,\cdot\rangle_{Z^*,Z}$.

1. Let $\emptyset \neq K \subset U_\infty$ be convex and $A \supset K$ be open in U_∞. We fix $\bar u \in K$. The functional $j\colon A \to \mathbb{R}$ is twice continuously Fréchet differentiable w.r.t. the norm $\|\cdot\|_\infty$.
   - 1a. The derivatives of j taken w.r.t. the space U_∞ extend to continuous linear respectively bilinear forms on U_2, i.e.

     $$j'(u) \in \mathcal{L}(U_2, \mathbb{R}) \qquad \text{and} \quad j''(u) \in \mathcal{L}(U_2 \otimes U_2, \mathbb{R}) \quad \text{for } u \in A.$$

   - 1b. Let $(u_k)_k \subset K$, $(v_k)_k \subset U_2$ be arbitrary sequences such that $u_k \to \bar u$ strongly w.r.t. the U_2-norm and $v_k \rightharpoonup v$ weakly in U_2 as $k \to \infty$. Then it holds:
     - 1bi. $j'(\bar u)v = \lim_{k\to\infty} j'(u_k)v_k$,
     - 1bii. $j''(\bar u)v^2 \le \liminf_{k\to\infty} j''(u_k)v_k^2$,
     - 1biii. if $v = 0$, there is $c > 0$ such that

       $$c \liminf_{k\to\infty} \|v_k\|_2^2 \le \liminf_{k\to\infty} j''(u_k)v_k^2.$$

2. Let $g\colon A \to Z$ be twice continuously Fréchet differentiable w.r.t. $\|\cdot\|_\infty$.
   - 2a. The derivatives of g taken w.r.t. U_∞ extend to continuous linear respectively bilinear forms on U_2, i.e.

     $$g'(u) \in \mathcal{L}(U_2, Z), \qquad \text{and} \quad g''(u) \in \mathcal{L}(U_2 \otimes U_2, Z) \qquad \text{for } u \in A.$$

   - 2b. Let $(u_k)_k \subset K$, $(v_k)_k \subset U_2$ be arbitrary sequences such that $u_k \to \bar u$ strongly w.r.t. the U_2-norm and $v_k \rightharpoonup v$ weakly in U_2 as $k \to \infty$. Then it holds:
     - 2bi. $g'(u_k)v_k \rightharpoonup g'(\bar u)v$ weakly in Z
     - 2bii. $g''(u_k)v_k^2 \rightharpoonup g''(\bar u)v^2$ weakly in Z

The following is our main abstract result and extends [**71**, Theorem 2.3] towards the inclusion of a state-constraint-like constraint of type "$g(u) \in C$". We denote by $\mathcal{R}_S(x)$ and $\mathcal{T}_S(x)$ the radial cone and the tangent cone of a closed convex set S in a Banach space X at some $x \in S$; see Chapter 0 for the definition.

Theorem 2.8 ([**168**], Theorem 3.2). *Let Assumption 2.7 hold. Let $C \subset Z$ be a closed convex set and let $\bar{u} \in K$, $g(\bar{u}) \in C$, and $\bar{\zeta} \in Z^*$ fulfill the following properties:*

$$\langle j'(\bar{u}) + g'(\bar{u})^*\bar{\zeta}, u - \bar{u}\rangle_2 \geq 0 \qquad \forall u \in K, \tag{2.6}$$

$$\langle \bar{\zeta}, z - g(\bar{u})\rangle_{Z^*,Z} \leq 0 \qquad \forall z \in C, \tag{2.7}$$

i.e. the KKT conditions for the problem $(\mathrm{AP}^{\mathrm{st}})$. *Assume further that it holds*

$$j''(\bar{u})v^2 + \langle \bar{\zeta}, g''(\bar{u})v^2\rangle_{Z^*,Z} > 0 \qquad \forall v \in C_{\bar{u},\bar{\zeta}} \setminus \{0\} \tag{2.8}$$

with

$$C_{\bar{u},\bar{\zeta}} := \mathrm{cl}_{U_2}(\mathcal{R}_K(\bar{u})) \cap \{v \in U_2 :\ j'(\bar{u})v = 0,\ \langle g'(\bar{u})^*\bar{\zeta}, v\rangle_2 = 0,\ g'(\bar{u})v \in \mathcal{T}_C(g(\bar{u}))\}.$$

Then, there are $\epsilon, \delta > 0$ such that the quadratic growth condition

$$j(u) \geq j(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|_2^2$$

holds for all $u \in K$ that satisfy $\|u - \bar{u}\|_2 \leq \epsilon$ and $g(u) \in C$; in particular, $\bar{u}$ is an U_2-local minimizer for $(\mathrm{AP}^{\mathrm{st}})$.

In the theorem and its proof we make extensive use of the continuation properties from Assumption 2.7.1a and 2.7.2a. In formula (2.6), for instance, $g'(\bar{u})^*\bar{\zeta} \in U_2^*$ is well-defined because of $g'(\bar{u}) \in \mathcal{L}(U_2, Z)$ by Assumption 2.7.2a. We follow the the proof of [**71**, Theorem 2.3], and abstract the techniques of several similar results in this context; see, e.g., [**49, 258, 185**], and, in particular, [**91**].

Proof. Assume the contrary, i.e. that there exist $(u_k)_k \subset K$ such that

$$\|u - u_k\|_2 < \frac{1}{k}, \quad j(u_k) < j(\bar{u}) + \frac{1}{2k}\|u_k - \bar{u}\|_2^2, \quad \text{and} \quad g(u_k) \in C.$$

Define $\rho_k := \|u_k - \bar{u}\|_2$ and $v_k := \frac{1}{\rho_k}(u_k - \bar{u})$. Since $(v_k)_k \subset U_2$ is bounded by definition and U_2 is a Hilbert space we can assume w.l.o.g. that $v_k \rightharpoonup v$ with some $v \in U_2$. We prove $v \in C_{\bar{u},\bar{\zeta}}$ in four steps:

Step A. From weak convergence and (2.6) we derive immediately:

$$\begin{aligned}\langle j'(\bar{u}) + g'(\bar{u})^*\bar{\zeta}, v\rangle_2 &= \lim_{k\to\infty} \langle j'(\bar{u}) + g'(\bar{u})^*\bar{\zeta}, v_k\rangle_2 \\ &= \lim_{k\to\infty} \frac{1}{\rho_k}\langle j'(\bar{u}) + g'(\bar{u})^*\bar{\zeta}, u_k - \bar{u}\rangle_2 \geq 0.\end{aligned}$$

Step B. To show $\langle g'(\bar{u})^*\bar{\zeta}, v\rangle_2 \leq 0$ observe that

$$\langle \bar{\zeta}, g'(u_k^\theta)v_k\rangle_{Z^*,Z} = \frac{1}{\rho_k}\langle \bar{\zeta}, g'(u_k^\theta)(u_k - \bar{u})\rangle_{Z^*,Z} = \frac{1}{\rho_k}\langle \bar{\zeta}, g(u_k) - g(\bar{u})\rangle_{Z^*,Z} \overset{(2.7)}{\leq} 0$$

with some $u_k^\theta := \theta_k u_k + (1 - \theta_k)\bar{u}$, $(\theta_k)_k \subset [0, 1]$, originating from the mean value theorem. Utilizing Assumption 2.7.2bi we obtain

$$\langle g'(\bar{u})^*\bar{\zeta}, v\rangle_2 = \langle \bar{\zeta}, g'(\bar{u})v\rangle_{Z^*,Z} = \lim_{k\to\infty} \langle \bar{\zeta}, g'(u_k^\theta)v_k\rangle_{Z^*,Z} \leq 0.$$

Similarly we obtain for arbitrary but fixed $\eta \in Z^*$ that

$$\langle \eta, \frac{1}{\rho_k}(g(u_k) - g(\bar{u}))\rangle_{Z^*,Z} = \langle \eta, g'(u_k^{\theta,\eta})v_k\rangle_{Z^*,Z} \to \langle \eta, g'(\bar{u})v\rangle_{Z^*,Z}$$

due to Assumption 2.7.2bi, i.e. $g'(\bar{u})v \in \text{weak-cl}_Z(\mathcal{R}_C(g(\bar{u}))) = \mathcal{T}_C(g(\bar{u}))$, which is shown as follows: since C is assumed to be closed and convex, we infer from [**36**, Proposition 2.55] that $\mathcal{T}_C(g(\bar{u})) = \text{cl}_Z(\mathcal{R}_C(g(\bar{u})))$. The radial cone $\mathcal{R}_C(g(\bar{u}))$ is convex due to convexity of C, and hence its (strong) closure in Z is equal to its weak closure $\text{weak-cl}_Z(\mathcal{R}_C(g(\bar{u})))$; see [**36**, Theorem 2.23.ii] for instance.

Step C. As in the proof of [**71**, Theorem 2.3] we find with help of the mean value theorem that $j'(\bar{u})v \le 0$ holds. Together with step B we obtain $\langle j'(\bar{u}) + g'(\bar{u})^*\bar{\zeta}, v\rangle_2 \le 0$ and therefore with step A:

$$\langle j'(\bar{u}) + g'(\bar{u})^*\bar{\zeta}, v\rangle_2 = 0.$$

Step D. Now, by step B we have $j'(u)v = -\langle g'(\bar{u})^*\bar{\zeta}, v\rangle_2 \ge 0$, which implies together with $j'(\bar{u})v \le 0$ that $j'(\bar{u})v = 0$. Finally it follows by step C that $\langle g'(\bar{u})^*\bar{\zeta}, v\rangle_2 = 0$.

As in [**71**] one can show that $v \in \text{cl}_{U_2}(\mathcal{R}_K(\bar{u}))$ and hence it follows from steps A-D that $v \in C_{\bar{u},\bar{\zeta}}$. Using our assumption and Taylor expansion we find

$$\frac{\rho_k^2}{2k} > j(u_k) - j(\bar{u}) = j'(\bar{u})(u_k - \bar{u}) + \frac{1}{2}j''(u_k^\theta)(u_k - \bar{u})^2$$

with some $u_k^\theta = \theta_k u_k + (1-\theta_k)\bar{u}$, $(\theta_k) \subset [0,1]$. Exploiting (2.6) and (2.7) it follows

$$\begin{aligned}
\frac{\rho_k^2}{2k} &\overset{(2.6)}{>} -\langle \bar{\zeta}, g'(\bar{u})(u_k - \bar{u})\rangle_{Z^*,Z} + \frac{1}{2}j''(u_k^\theta)(u_k - \bar{u})^2 \\
&= -\langle \bar{\zeta}, g(u_k) - g(\bar{u})\rangle_{Z^*,Z} + \langle \bar{\zeta}, g(u_k) - g(\bar{u}) - g'(\bar{u})(u_k - \bar{u})\rangle_{Z^*,Z} \\
&\qquad + \frac{1}{2}j''(u_k^\theta)(u_k - \bar{u})^2 \\
&\overset{(2.7)}{\ge} \langle \bar{\zeta}, g(u_k) - g(\bar{u}) - g'(\bar{u})(u_k - \bar{u})\rangle_{Z^*,Z} + \frac{1}{2}j''(u_k^\theta)(u_k - \bar{u})^2 \\
&= \frac{1}{2}\rho_k^2(\langle \bar{\zeta}, g''(\tilde{u}_k^\theta)v_k^2\rangle_{Z^*,Z} + j''(u_k^\theta)v_k^2),
\end{aligned}$$

where we used the mean value theorem for the last equality with some $\tilde{u}_k^\theta = \tilde{\theta}_k u_k + (1-\tilde{\theta}_k)\bar{u} \in K$, $(\tilde{\theta}_k) \subset [0,1]$. From

$$\frac{1}{k} > \langle \bar{\zeta}, g''(\tilde{u}_k^\theta)v_k^2\rangle_{Z^*,Z} + j''(u_k^\theta)v_k^2$$

and $u_k^\theta \to \bar{u}$, $\tilde{u}_k^\theta \to \bar{u}$ in U_2, $v_k \rightharpoonup v$ weakly in U_2 we find with Assumptions 2.7.1bii and 2.7.2bii:

$$j''(\bar{u})v^2 + \langle \bar{\zeta}, g''(\bar{u})v^2\rangle_{Z^*,Z} \le 0.$$

Since (2.8) and $v \in C_{\bar{u},\bar{\zeta}}$ hold, we conclude $v = 0$. Using Assumptions 2.7.1biii at (♣) and 2.7.2bii at (★) we finally arrive at

$$0 < c = c\liminf_{k\to\infty}\|v_k\|_2^2 \overset{(\clubsuit)}{\le} \liminf_{k\to\infty} j''(u_k^\theta)v_k^2 \le \liminf_{k\to\infty}\Big(\frac{1}{k} - \langle \bar{\zeta}, g''(\tilde{u}_k^\theta)v_k^2\rangle_{Z^*,Z}\Big) \overset{(\bigstar)}{=} 0,$$

which is the desired contradiction. □

Before addressing applications to optimal control of quasilinear parabolic problems, let us put our result into context.

In [**35**, Theorem 4.13] it has been proven that in the purely control-constrained case the positivity condition (1.18) is actually equivalent to the seemingly stronger coercivity condition (1.20). We have summarized this in Proposition 2.9 in Section 1.5. Such an observation is typical for PDE-constrained optimization problems in the presence of an L^2-Tikhonov regularization term; cf., e.g., [**73**, Theorem 4.11] for a semilinear parabolic problem. Since one of the read threads of this thesis is concerned with the two-norm discrepancy, let us point out that this property also holds true in the present abstract setting.

Proposition 2.9. *Under the assumptions of Theorem 2.8 the positivity condition* (2.8) *implies that there is* $\gamma > 0$ *such that*

$$j''(\bar{u})v^2 + \langle \bar{\zeta}, g''(\bar{u})v^2 \rangle_{Z^*,Z} \geq \gamma \|v\|_2^2 \qquad \forall v \in C_{\bar{u},\bar{\zeta}}.$$

The proof uses similar techniques as the proof of [**73**, Theorem 4.11] and relies crucially on Assumption 2.7.1biii, i.e. on the presence of the L^2-Tikhonov term.

Proof. Assume the contrary, i.e. that there exist $(v_k)_k \subset C_{\bar{u},\bar{\zeta}}$, $\|v_k\|_2 = 1$, such that

$$j''(\bar{u})v_k^2 + \langle \bar{\zeta}, g''(\bar{u})v_k^2 \rangle_{Z^*,Z} < \frac{1}{k}. \tag{2.9}$$

W.l.o.g. we can assume $v_k \rightharpoonup v$ weakly in U_2 for some $v \in U_2$. With the same techniques as in the proof of Theorem 2.8 we see that actually $v \in C_{\bar{u},\bar{\zeta}}$ holds. Moreover, taking inferior limits on both sides of (2.9) and using Assumption 2.7.1bii and 2bii we obtain

$$\begin{aligned} j''(\bar{u})v^2 + \langle \bar{\zeta}, g''(\bar{u})v^2 \rangle_{Z^*,Z} &\leq \liminf_{k\to\infty} j''(\bar{u})v_k^2 + \lim_{k\to\infty} \langle \bar{\zeta}, g''(\bar{u})v^2 \rangle_{Z^*,Z} \\ &\leq \liminf_{k\to\infty} \frac{1}{k} = 0. \end{aligned}$$

Due to (2.8) we conclude $v = 0$. Therefore, again taking inferior limits on both sides of (2.9) and now utilizing Assumption 2.7.1biii and 2bii yields the desired contradiction $0 < c \leq 0$. □

Note that Proposition 2.9 highlights the strength of Theorem 2.8: the positivity condition (2.8) implies coercivity of $j''(\bar{u}) + \langle \bar{\zeta}, g''(\bar{u})[\cdot]^2 \rangle_{Z^*,Z}$ w.r.t. the U_2-norm on $C_{\bar{u},\bar{\zeta}}$, but differentiability of j and g only hold w.r.t. the U_∞-norm. Nevertheless, the first-order conditions (2.6) and (2.7) together with (2.8) are sufficient for local optimality of $\bar{u}$ w.r.t. the U_2-norm.

Let us now indicate how Theorem 2.8 allows to extend a well-known result for semilinear parabolic problems.

Example 2.10 ([**168**], Example 3.3)**.** Let us recall the following simplified version of the semilinear parabolic optimal control problem with pointwise state-constraints and purely time-dependent controls from [**91**]:

$$
(2.10)\qquad \begin{cases} \min\limits_{u,y} J(y,u) := \dfrac{1}{2}\|y-y_d\|_{L^2(I\times\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(I,\mathbb{R}^m)} \\ \quad \text{s.t.} \qquad u \in U_{ad} := \{v \in L^2(I,\mathbb{R}^m):\quad u_{\mathrm{a}} \le v(t) \le u_{\mathrm{b}} \quad \forall t \in I\}, \\ \qquad\qquad \partial_t y - \nabla\cdot\mu\nabla y + d(y) = \sum\limits_{i=1}^m u_i b_i, \\ \qquad\qquad\qquad\qquad \partial_n y = 0, \\ \qquad\qquad\qquad\qquad y(0) = y_0, \\ \quad \text{and} \qquad g(y) \le 0 \quad \text{pointwise on } I\times\Omega. \end{cases}
$$

Here, $-\nabla\cdot\mu\nabla$ denotes an elliptic second-order differential operator with L^∞-coefficients, and d and g respectively are functions $\mathbb{R}\to\mathbb{R}$ that induce sufficiently smooth superposition operators $L^\infty(Q)\to L^\infty(Q)$. Further, d is assumed to be monotone increasing. The actuator functions $b_1,\dots,b_m \in L^\infty(\Omega)$ are fixed.

The reader may easily verify along the lines of [**91, 71**] that this control problem fits into the framework of Assumption 2.7. Hence, Theorem 2.8 allows to reformulate the second-order sufficient conditions obtained in [**91**, Theorem 5] with L^∞-replaced by L^2-neighbourhoods, i.e. a two-norm gap can be avoided. Despite that in our formulation of (2.10) we have replaced the general objective functional from [**91**] by a standard quadratic one for reasons of shortness, Theorem 2.8 is still applicable to the class of more general functionals described in [**91**]; see also [**71**]. For the precise assumptions and a detailed discussion of the full model problem we refer the reader to [**91**].

The following example, although of artificial nature, illustrates that the assumptions in the formulation of Theorem 2.8 are necessary. Necessity of the assumptions on j is addressed in [**71**] and hence we only concentrate on the assumptions on g.

Example 2.11 ([**168**], Example 3.5)**.** With $U_\infty = L^\infty([0,1])$, $U_2 = L^2([0,1])$, $Z = C([0,1])$ we consider

$$
(\mathrm{E})\qquad \begin{cases} \min\limits_{u\in L^2([0,1])} j(u) := \displaystyle\int_0^1 u(t)^2\mathrm{d}t \\ \quad \text{s.t.} \qquad -1 \le u(t) \le 1, \\ \qquad\qquad [g(u)](t) \ge t \ \forall t\in[0,1], \end{cases}
$$

with $[g(u)](t) := \int_0^t (1-\cos(\frac{\pi}{2}u(s)))\mathrm{d}s$. Note that j satisfies Assumption 2.7 and observe that g: $L^2([0,1])\to C([0,1])$ is well-defined. Yet, since the superposition operator associated to the cosine function is known to be Fréchet differentiable on $L^\infty([0,1])$, but not on $L^2([0,1])$, we only have at hand twice Fréchet differentiablity of g as map $L^\infty([0,1])\to C([0,1])$.

One verifies that $\bar u \equiv 1$ is feasible for (E), and satisfies the FONs (2.6) and (2.7) with $\bar\zeta = -\frac{2}{\pi}\delta_1 \in C([0,1])^*$. Herein, δ_1 denotes the Dirac measure concentrated at $t=1$. The coercivity condition (2.8) is trivially satisfied at $(\bar u,\bar\zeta)$, because

$C_{\bar{u},\bar{\zeta}} = \{0\}$. Further, the second derivative of the functional at $\bar{u}$ is even $L^2([0,1])$-coercive, but any $u_n \in L^2([0,1])$ defined by $u_n(t) = -1$ for $t \in [0, \frac{1}{n}]$ and $u_n(t) = 1$, else, is also feasible for (E) and satisfies $j(u_n) = j(\bar{u})$. Together with $u_n \to \bar{u}$ w.r.t. the $L^2([0,1])$-norm, this shows that a quadratic growth condition around $\bar{u}$ cannot hold.

The reason is that Theorem 2.8 cannot be applied, because Assumption 2.7.2 fails to hold. Choose $v_n := n^{\frac{1}{2}} \mathbf{1}_{(0,\frac{1}{n})}$, then it holds $v_n \rightharpoonup 0$ weakly in $L^2([0,1])$, but for $\hat{u}_n := \frac{1}{2}(u_n + \bar{u})$ we obtain $\hat{u}_n \to \hat{u}$ strongly in $L^2([0,1])$ and

$$\langle \delta_1, g''(\hat{u}_n) v_n^2 \rangle = \frac{\pi^2}{4} \int_0^1 \cos(\frac{\pi}{2} \hat{u}_n(t)) v_n^2(t) \mathrm{d}t = \frac{\pi^2}{4} \not\to 0 = \langle \delta_1, g''(\bar{u}) v^2 \rangle$$

which disproves Assumption 2.7.2bii. However, we note that due to continuous Fréchet differentiability of g w.r.t. $L^\infty([0,1])$ an argument summarized in Remark 2.12 below applies: $\bar{u}$ is an $L^\infty([0,1])$-local but not an $L^2([0,1])$-local solution of (E).

In fact, we note that under slightly weaker assumptions than in Assumption 2.7 a modified version of Theorem 2.8, now with two-norm gap, can be obtained. This version is of interest, e.g., for (P^{st}) in the case without control-constraints. We refer the reader to [**168**, Section 6] for the details.

Remark 2.12 ([**168**], Remark 3.4)**.** Let us for a moment replace convergence $u_k \to \bar{u}$ in U_2 in Assumption 2.7 by the stronger convergence $u_k \to \bar{u}$ in V, where $(V, \|\cdot\|_V)$ is a Banach space such that $V \hookrightarrow U_\infty$ and $K \subset V$. The proof of Theorem 2.8 still shows that a quadratic growth condition of type $j(u) \geq j(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|_2^2$ holds, but now only for those $u \in K$ that fulfill $\|u - \bar{u}\|_V < \epsilon$ and $g(u) \in C$, i.e. there is a two-norm gap in the quadratic growth condition. Consequently, $\bar{u}$ is at best a V-local minimizer for ($\mathrm{AP}^{\mathrm{st}}$), which corresponds —on the abstract level— for $V = U_\infty$ to the result of [**91**]; cf. also [**71**, Theorem 1.3] and the references given there.

We conclude this section by pointing out an open problem. An important property of the SSCs in [**71**, Theorem 2.3] is their minimal gap to corresponding SNCs if the admissible set K is polyhedric. Positivity of $j''(\bar{u})$ on a certain cone $C_{\bar{u}} \subset U_2$ is —together with FONs— a sufficient optimality condition for $\bar{u}$, while nonnegativity of $j''(\bar{u})$ on the *same* cone is necessarily implied by local optimality of $\bar{u}$ [**71**, Theorem 2.2]. The second-order conditions for the control-constrained quasilinear parabolic problems from [**35, 45**] have this property as already explained in Section 1.5. Obtaining SNCs for ($\mathrm{AP}^{\mathrm{st}}$), however, seems to be a challenging topic and is beyond the scope of our work. Indeed, proving SNCs for an optimal control problem with infinitely many state-constraints is known to be highly difficult in general, and in the survey [**73**] the respectiv theory is described as *"widely open"*; cf. [**73**, Section 7.5]. To the best of our knowledge, the only contributions so far pertain to SNCs in the case of pure state-constraints for semilinear elliptic or the stationary Navier-Stokes equations [**184**], or to the case of mixed control-state-constraints and semilinear elliptic PDEs; see, e.g., [**258, 185**]. Apart from that there are several results on second-order conditions for problems with finitely many state-constraints; see for instance [**37, 66, 57, 67, 43**]. In this context we

also mention abstract optimization-theoretic results [**84, 287, 288**] with different applications to PDE-constrained optimization.

2.3. Second-order sufficient conditions for averaged in time state-constraints

This section roughly follows [**168**, Section 4] and contains the first part of our discussion of SSCs for (P^{st}) with state-constraints. We replace the pointwise in space and time state-constraints by averaged in time state-constraints; see Assumption 2.13 below. For the resulting modified model problem we prove SSCs avoiding the two-norm gap while keeping the rather low regularity requirements on the state equation from Assumptions 1.5, 1.6 and 1.8. Consequently, our results apply to the full range of situations described in Section 1.2, yet with additional state-constraints.

The first part of Assumption 2.7, referring to the unchanged state equation and the objective functional, has already been verified in [**35**, Section 4.3]; we have summarized this in Proposition 1.21. Therefore, the remaining work is to check Assumption 2.7.2. This requires a careful regularity analysis of the derivatives of the control-to-state map. The results of this analysis also highlight the obstructions that prevent us from applying Theorem 2.8 under Assumptions 1.5, 1.6, 1.8 and 2.1 directly and therefore motivate the introduction of averaged in time state-constraints. In particular, the analysis of the quasilinear problem (P^{st}) is quite different from the discussion of the semilinear problem mentioned in Example 2.10 due to the more complicated structure of derivatives of the nonlinearity in the differential operator. This yields slightly better regularity results in the case of semilinear PDEs.

2.3.1. Averaged in time state-constraints. The regularity Assumptions 1.5, 1.6 and 1.8 on (P^{st}) remain unchanged. However, throughout Section 2.3 we will be concerned with the following, modified type of state-constraints.

Assumption 2.13.

1. The set of admissible states is
$$Y_{\mathrm{ad}} = \left\{ y \in L^1(I, C(\overline{\Omega})) : y_{\mathrm{a}}(x) \leq \int_I y(t,x)\mathrm{d}t \leq y_{\mathrm{b}}(x) \quad \forall x \in \overline{\Omega} \right\},$$
with bounds $y_{\mathrm{a}}, y_{\mathrm{b}} \in C(\overline{\Omega})$ satisfying $y_{\mathrm{a}}(x) < y_{\mathrm{b}}(x)$ for all $x \in \overline{\Omega}$ and $y_{\mathrm{a}}(x) < 0 < y_{\mathrm{b}}(x)$ for all $x \in \Gamma_D$. We allow for $y_{\mathrm{a}} \equiv -\infty$ or $y_{\mathrm{b}} \equiv \infty$.
2. There is a feasible point, i.e. there is $(y,u) \in Y_{\mathrm{ad}} \times U_{\mathrm{ad}}$ such that y and u fulfill the state equation (Eq).

Intuitively, this means, e.g., in the case of controling temperature, keeping the average temperature over the time interval at each point of an object in a certain desired range. Of course, it is also possible to take the average w.r.t. a subinterval $I_{\mathrm{obs}} \subset I$ only. In Section 2.5 we will illustrate the influence of averaging w.r.t. time and of the choice of I_{obs} on behalf of some numerical examples. Moreover, to get closer to the original pointwise in time formulation, one may also consider averaging on a finite number of subintervals of I separately. Since these two modifications differ from the basic setting described in Assumption 2.19 only by technicalities, we will not address them further during our analysis. Averaged-type instead of purely pointwise constraints are common in the literature, e.g., averaged in space

and pointwise in time bounds on the state [**113, 211, 204, 109**] or its gradient [**205**]. Interestingly, the regularity results available for our problem type in the rough setting require us to average w.r.t. time, instead of w.r.t. space as in the aforementioned publications; see the explanation below Proposition 2.14 for the details.

To motivate the setting chosen in Assumption 2.13, let us already state the following result concerned with properties of the control-to-state map and its derivatives. It will serve as the main step towards the proof of SSCs for $(\mathrm{P}^{\mathrm{st}})$ under Assumption 2.13. We postpone the proof to Section 2.3.3 below.

Proposition 2.14 ([**168**], Proposition 4.6)**.** *Let Assumptions 1.5, 1.6 and 1.8 hold and fix $u \in L^s(\Lambda)$.*

1a. *The first derivative $S'(u)$ of the control-to-state map extends to a continuous linear map from $L^2(\Lambda)$ to $L^q(I, C(\overline{\Omega}))$ for any $q \in (1, \frac{2p}{d})$.*

1b. *The second derivative $S''(u)$ extends to a continuous bilinear map from $L^2(\Lambda) \times L^2(\Lambda)$ to $\mathbb{W}^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ for any $r \in (1, \frac{2p}{p+d})$.*

2. *Let $(u_k)_k \subset L^s(\Lambda)$ converge to $\bar{u}$ strongly in $L^s(\Lambda)$ and $(v_k)_k \subset L^2(\Lambda)$ converge to some v weakly in $L^2(\Lambda)$. Then it follows $S'(u_k)v_k \to S'(\bar{u})v_k$ strongly in $L^q(I, C(\overline{\Omega}))$ and $S''(u_k)v_k^2 \rightharpoonup S''(\bar{u})v^2$ weakly in $W^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ for q and r as in part 1a and 1b.*

Having in mind these continuity and continuation properties for S and its derivatives, assume that we want to apply Theorem 2.8 to $(\mathrm{P}^{\mathrm{st}})$ in case of pointwise in space and time state-constraints (Assumption 2.1). Consequently, we have to verify Assumption 2.7 for $U_\infty = L^s(\Lambda)$, $U_2 = L^2(\Lambda)$, $K = U_{\mathrm{ad}}$, $Z = C(\overline{Q})$, $C = Y_{\mathrm{ad}}$, j being the reduced functional and $g = S$ being the control-to-state map of $(\mathrm{P}^{\mathrm{st}})$. We would have to show that $S'(u)$ extends to a bounded linear map $L^2(\Lambda) \to C(\overline{Q})$, and that $S''(u)$ extends to a continuous bilinear map $L^2(\Lambda) \times L^2(\Lambda) \to C(\overline{Q})$, for any fixed $u \in U_{\mathrm{ad}}$. This, however, already fails to hold for the first derivative: from Lemma 1.12 we know that the extension $S'(u)\colon L^2(\Lambda) \to \mathbb{W}^{1,2}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ is the best possible we can expect. Unfortunately, there is no embedding $\mathbb{W}^{1,2}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow C(\overline{Q})$. Due to $\frac{2p}{p+d} < 2$ in Proposition 2.14.1b, the situation is even worse for $S''(u)$. Similarly, the application to averaged in space and pointwise in time state-constraints [**211**], i.e.

$$Y_{\mathrm{ad}} = \left\{ y\colon \ y_{\mathrm{a}}(t) \le \int_\Omega y(t,x)\omega(x)\mathrm{d}x \le y_{\mathrm{b}}(t) \ \forall t \in I \right\}$$

with continuous functions $y_{\mathrm{a}}, y_{\mathrm{b}} \in C(I)$ and a weight function $\omega \in L^\infty$ is not possible. Utilizing Proposition 2.14.1b we would require an embedding

$$\mathbb{W}^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow C(I, L^1)$$

for some $r \in (1, \frac{2p}{p+d})$ in order to verify Assumption 2.7.2bii. Unfortunately, such an embedding cannot be true. However, the embedding

$$\mathbb{W}^{1,r}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow L^1(I, W^{1,p}_{\Gamma_D}) \hookrightarrow L^1(I, C(\overline{\Omega}))$$

is obvious. Therefore, averaging in time —instead of averaging in space— seems to be reasonable, resulting in the formulation of Assumption 2.13.

Having motivated the chosen setting we start our analysis of $(\mathrm{P}^{\mathrm{st}})$.

2.3.2. Existence of optimal controls and first-order optimality conditions. For completeness, we proceed in the standard way and start with proving well-posedness of the problem and first-order conditions. Except for minor modifications due to the different structure of the state-constraints this is rather similar to Section 2.1.2.

Existence of an optimal control for $(\mathrm{P}^{\mathrm{st}})$ with averaged in time state-constraints is proven completely analogous to Proposition 2.2. We therefore only state the result.

Proposition 2.15 ([**168**], Theorem 4.2)**.** *Let Assumptions 1.5, 1.6, 1.8 and 2.13 hold. Then, there exists a globally optimal control $\bar{u} \in U_{\mathrm{ad}}$ for the optimal control problem $(\mathrm{P}^{\mathrm{st}})$.*

To address FONs we first need to formulate a suitable constraint qualification.

Assumption 2.16. Under Assumption 2.13.1 let $\bar{u} \in U_{\mathrm{ad}}$ be an $L^2(\Lambda)$-local solution to $(\mathrm{P}^{\mathrm{st}})$ with associated state $\bar{y} = S(\bar{u}) \in Y_{\mathrm{ad}}$ such that the following linearized Slater condition is fulfilled at $\bar{u}$: there is $u_{\mathrm{Sl}} \in U_{\mathrm{ad}}$ such that $\bar{y} + S'(\bar{u})(u_{\mathrm{Sl}} - \bar{u}) \in \mathring{Y}_{\mathrm{ad}}$, i.e.

$$y_{\mathrm{a}}(x) < \int_\Omega [\bar{y}(t,x) + S'(\bar{u})(u_{\mathrm{Sl}} - \bar{u})(t,x)]\,\mathrm{d}t < y_{\mathrm{b}}(x), \quad \forall x \in \overline{\Omega}.$$

As in Section 2.1.2 the proof of the following result is based on the abstract KKT conditions from Theorem 0.1.

Theorem 2.17 ([**168**], Theorem 4.4)**.** *Under Assumptions 1.5, 1.6, 1.8 and 2.16 and Assumption 2.13.1 there exists a regular Borel measure $\bar{\nu} \in \mathcal{M}(\overline{\Omega}) = C(\overline{\Omega})^*$ on $\overline{\Omega}$ and the adjoint state $\bar{p} \in L^{r'}(I, W^{1,p'}_{\Gamma_D})$, $r' \in (1,\infty)$, such that the optimality system*

$$\begin{cases} \partial_t \bar{y} + \mathcal{A}(\bar{y})\bar{y} = B\bar{u}, \\ \bar{y}(0) = y_0, \end{cases} \tag{2.11}$$

$$\begin{cases} -\partial_t \bar{p} + \mathcal{A}(\bar{y})^*\bar{p} + \mathcal{A}'(\bar{y})^*\bar{p} = \bar{y} - y_d + \mathrm{d}t \otimes \bar{\nu}, \\ \bar{p}(T) = 0, \end{cases} \tag{2.12}$$

$$\mathrm{supp}(\bar{\nu}^+) \subset \{\int_I \bar{y}(t,\cdot)\mathrm{d}t = y_{\mathrm{b}}\}, \quad \mathrm{supp}(\bar{\nu}^-) \subset \{\int_I \bar{y}(t,\cdot)\mathrm{d}t = y_{\mathrm{a}}\}, \tag{2.13}$$

$$\langle B^*\bar{p} + \alpha\bar{u}, u - \bar{u}\rangle_{L^{s'}(\Lambda),L^s(\Lambda)} \geq 0 \qquad \text{for all } u \in U_{\mathrm{ad}}, \tag{2.14}$$

is satisfied. Here, $\bar{\nu} = \bar{\nu}^+ - \bar{\nu}^-$ denotes the Jordan decomposition of $\bar{\nu}$, cf. Remark 2.5, and (2.12) has to be understood in the sense outlined in the proof.

As in Theorem 2.4, pointwise in space and averaged in time state-constraints infer a complementary slackness condition in the FONs. In Theorem 2.17, however, we have already rephrased this condition as (2.13) follwing Remark 2.5. The overal structure of the following proof is analogous to the proof of Theorem 2.4. Hence, we put particular emphasis on the handling of the new type of averaged in time state-constraints introduced in Assumption 2.13.

Proof. In Theorem 0.1 we choose $Z = C(\overline{\Omega})$ and $g := \iota \circ A \circ S$: , where S: $L^s(\Lambda) \to \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ is the control-to-state map, ι: $W^{1,p}_{\Gamma_D} \hookrightarrow C(\overline{\Omega})$ the Sobolev embedding, and A: $\varphi \mapsto (x \mapsto \int_I \varphi(t,x)\mathrm{d}t)$ is averaging w.r.t. time, which is a bounded linear map $L^r(I, W^{1,p}_{\Gamma_D}) \to W^{1,p}_{\Gamma_D}$ for any $r \in (1,\infty)$. The choice $r = s$ shows that g is well-defined from $L^s(\Lambda)$ into $C(\overline{\Omega})$. From $A \in \mathcal{L}(L^r(I, W^{1,p}_{\Gamma_D}), W^{1,p}_{\Gamma_D})$ we conclude $A^* \in \mathcal{L}(W^{-1,p'}_{\Gamma_D}, L^{r'}(I, W^{-1,p'}_{\Gamma_D}))$, and ι^* is the embedding $\mathcal{M}(\overline{\Omega}) \hookrightarrow W^{-1,p'}_{\Gamma_D}$. For a test function $\psi \in L^r(I, W^{1,p}_{\Gamma_D})$ we compute

$$\langle A^*\iota^*\bar\nu, \psi\rangle_{L^{r'}(I,W^{-1,p'}_{\Gamma_D}),L^r(I,W^{1,p}_{\Gamma_D})} = \langle \iota^*\bar\nu, A\psi\rangle_{\mathcal{M}(\overline{\Omega}),C(\overline{\Omega})} = \int_{\overline{\Omega}} A\psi \mathrm{d}\bar\nu = \int_{\bar Q} \psi \mathrm{d}t\mathrm{d}\bar\nu,$$

i.e. $A^*\iota^*\bar\nu = \mathrm{d}t \otimes \bar\nu \in L^{r'}(I, W^{-1,p'}_{\Gamma_D})$ for each $r' \in (1,\infty)$. Together with

$$\tilde S'(B\bar u) \in \mathcal{L}(L^r(I, W^{-1,p}_{\Gamma_D}), L^r(I, W^{1,p}_{\Gamma_D})),$$

for $r \in (1, s]$, which follows from Lemma 1.12.2, we find

$$\tilde S'(B\bar u)^* A^*\iota^* \in \mathcal{L}(\mathcal{M}(\overline{\Omega}), L^{r'}(I, W^{1,p'}_{\Gamma_D}))$$

for $r' \in [s', \infty)$. This shows that $\bar p = \tilde S'(B\bar u)^*(\bar y - y_d + A^*\iota^*\bar\nu) \in L^{r'}(I, W^{1,p'}_{\Gamma_D})$, $r' \in (1,\infty)$, is well-defined. Equation (2.12) has to be understood in this sense. Finally, a short computation shows that (2.14) holds. □

Let us now briefly comment on the differences of this result from the corresponding Theorem 2.4 in the case of pointwise in space and time state-constraints. In the introduction of this chapter we have explained that averaged-type state-constraints are often introduced to cope with regularity issues, i.e. when regularity results available for the problem type under consideration do not suffice to handle purely pointwise state-constraints in a satisfying manor. Our specific setting of averaged in time state-constraints has been motivated below Proposition 2.14. In Theorem 2.17, the advantage of averaged-type state-constraints becomes visible, too. Unlike for pointwise in space and time state-constraints, cf. our remarks at the end of Section 2.1.2, now the regularity of the adjoint state can be improved as pointed out in [**168**, Remark 4.5]: let instead of Assumption 1.8 the enhanced regularity Assumption 1.10 hold that allows to apply Theorem 1.14 and Proposition 1.15. Since it holds $A^*\iota^*\bar\nu = \mathrm{d}t \otimes \bar\nu \in L^{r'}(I, W^{-1,p'}_{\Gamma_D})$ for any $r' \in (1,\infty)$ we can apply Proposition 1.15 to obtain improved regularity

$$\bar p \in \mathbb{W}^{1,r'}(I, (W^{-1,p'}, W^{1,p'}_{\Gamma_D})), \qquad r' \in [s', \infty),$$

for the adjoint state from Theorem 2.17 in this case. In particular, the adjoint equation (2.12) even holds in the distributional sense in the respective spaces. This is different from Theorem 2.4 where the adjoint equation had to be understood purely formal; cf. Remark 2.6. The reason for this different behaviour is the special structure of the right-hand side in (2.12) which originates from the averaged-type state-constraints.

2.3.3. Regularity of the derivatives of the control-to-state map: proof of Proposition 2.14. We now give the proof of Proposition 2.14, i.e. the proof of [**168**, Proposition 4.6]. We need to perform a detailed analysis w.r.t. regularity, continuity, and extension properties of the derivatives of the control-to-state map

S. In particular, recall the definition of S and its derivatives stated in Lemma 1.12. As already pointed out at the end of Section 1.2 our Assumptions 1.5, 1.6 and 1.8 suffice to apply those results from [**35**] summarized in Chapter 1 that we will use in the following.

We start with the proof of part 1 of Proposition 2.14. We know $S'(u) \in \mathcal{L}(L^2(\Lambda), \mathbb{W}^{1,2}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})))$ for all $u \in L^s(\Lambda)$, cf. Lemma 1.12. Hence, 1a. follows from

$$\mathbb{W}^{1,2}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \hookrightarrow_c L^q(I, C(\overline{\Omega})), \qquad q \in \left(1, \frac{2p}{d}\right),$$

which is the first embedding from Lemma 1.13. Since this embedding is compact, the mapping $S'(u) \in \mathcal{L}(L^2(\Lambda), L^q(I, C(\overline{\Omega})))$ is also compact. For 1b. it suffices due to Lemma 1.12 to show for $r \in (1, \frac{2p}{p+d})$ and $z_1, z_2 \in \mathbb{W}^{1,2}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$ that

$$\|\mathcal{A}''(y)[z_1,z_2]\|_{L^r(I,W^{-1,p}_{\Gamma_D})} \quad \lesssim \quad \|z_1\|_{\mathbb{W}^{1,2}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D}))}\|z_2\|_{\mathbb{W}^{1,2}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D}))}$$

holds, where $y = S(u)$. This, however, follows from the definition of $\mathcal{A}''$, Hölder's inequality and the aforementioned embedding.

Let us now prove part 2. In the proof of [**35**, Proposition 4.9] it has been shown that

$$\tilde{S}'(Bu_k) \to \tilde{S}'(B\bar{u}) \quad \text{in } \mathcal{L}\left(L^r(I, W^{-1,p}), \mathbb{W}^{1,r}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))\right)$$

as long as $r \le \frac{2p}{p-d}$; see Section 2.1.2 for the meaning of $\tilde{S}$. In particular,

$$S'(u_k) \to S'(\bar{u}) \quad \text{in } \mathcal{L}\left(L^2(\Lambda), \mathbb{W}^{1,2}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))\right)$$

is true, from which we conclude the first statement of part 2. For the second derivative we write

$$\begin{aligned} S''(u_k)v_k^2 - S''(\bar{u})v^2 = &\left(\tilde{S}'(Bu_k) - \tilde{S}'(B\bar{u})\right)\mathcal{A}''(y_k)[S'(u_k)v_k]^2 \\ &+ \tilde{S}'(B\bar{u})\left(\mathcal{A}''(y_k)[S'(u_k)v_k]^2 - \mathcal{A}''(\bar{y})[S'(\bar{u})v]^2\right), \end{aligned}$$

with $y_k = S(u_k)$ and $\bar{y} = S(\bar{u})$. Convergence of the operators above is in particular true for $r \in (1, \frac{2p}{p+d})$. Hence, it suffices to show that $\mathcal{A}''(y_k)[S'(u_k)v_k]^2 \rightharpoonup \mathcal{A}''(\bar{y})[S'(\bar{u})v]^2$ weakly in $L^r(I, W^{-1,p}_{\Gamma_D})$, which follows by Hölder's inequality and the previous results. This completes the proof of the proposition.

The following comments are an extended version of [**168**, Remark 4.7]. Utilizing improved regularity of the states (Theorem 1.14) obtained in [**35**] together with considering the linearized state equation on Bessel potential spaces instead of $W^{-1,p}_{\Gamma_D}$ would not improve the situation significantly, as can be seen along the lines of the above proof. Moreover, the appearance of differential operators in the $\mathcal{A}''$-term in the second derivative of the control-to-state map, cf. (1.3), and hence in the second derivative of the Lagrangian of (P^{st}), makes it impossible to repeat the approach of [**189**]; cf. in particular [**189**, Proposition 3.8]: for the semilinear equation discussed in [**189**] all terms in the second derivative of the nonlinearity are of order zero, which allows to get along with less regularity for the linearized state equation than in our case.

Let us point out this interesting detail a bit more precisely because it highlights the differences between semilinear and quasilinear problems: the second derivative of the control-to-state map, $w = S''(\bar u)v^2$, is given by the solution of the equation

$$\partial_t w - \nabla\cdot\xi(\bar y)\mu\nabla w - \nabla\cdot\xi'(\bar y)w\mu\nabla\bar y = \underbrace{-\nabla\cdot\big(2\xi'(\bar y)z\mu\nabla z + \xi''(\bar y)z^2\mu\nabla\bar y\big)}_{\mathcal{A}''(\bar y)[z,z]}$$
$$w(0) = 0,$$

where $z = S'(\bar u)v$ and $\bar y = S(\bar u)$; cf. Lemma 1.12. The right-hand side of this equation is a distributional object and not a measurable function on Q, in general. Consequently, the regularity that can be expected for w is rather limited. We have seen this in the above proof of Proposition 2.14 when we had to estimate the norm of $\mathcal{A}''(y)[z_1, z_2]$ in $L^r(I, W_{\Gamma_D}^{-1,p})$ in terms of z_1 and z_2.

This is clearly different from the semilinear parabolic case: let, e.g., G denote the control-to-state map associated with the problem (2.10) from Example 2.10. Hereby, it does not matter that this problem has been formulated with purely time-dependent controls. Standard computations, see, e.g., [**270**, Chapter 5], show that the second derivative $w = G''(\bar u)v^2$ is now given by the solution of

$$\partial_t w - \nabla\cdot\mu\nabla w + d'(\bar y)w = -d''(\bar y)z^2$$
$$w(0) = 0,$$

where $z = G'(\bar u)v$ and $\bar y = G(\bar u)$. Unlike for the quasilinear case, the right-hand side hereof is given by a measurable function on Q, i.e. having at hand the same regularity for the first derivatives $S'(\bar u)v$ and $G'(\bar u)v$ the regularity of the second derivative will always be much lower in the quasilinear case than in the semilinear case. This effect is due to the presence of differential operators in $\mathcal{A}''$, i.e. in the term originating from the second derivative of the nonlinearity in (Eq), and hence due to the quasilinear structure of this equation.

2.3.4. Second-order sufficient conditions. Using the previously obtained auxiliary results we formulate SSCs for (P^{st}). As already pointed out, the proof relies on Theorem 2.8. For convenience, we introduce the regular part $\hat p$ of the adjoint state $\bar p$ defined by the following equation

$$-\partial_t\hat p + \mathcal{A}(\bar y)^*\hat p + \mathcal{A}'(\bar y)^*\hat p = \bar y - y_d, \qquad \hat p(T) = 0, \tag{2.15}$$

Note that this allows us to express the first derivative of the reduced functional j as $j'(\bar u)v = \langle B^*\hat p + \alpha\bar u, v\rangle_{L^2(\Lambda)}$; cf. Section 1.3.

Theorem 2.18 ([**168**], Theorem 4.8). *Let Assumptions 1.5, 1.6 and 1.8 and Assumption 2.13.1 hold, and let $\bar u \in U_{\mathrm{ad}}$, $\bar y = S(\bar u) \in Y_{\mathrm{ad}}$, $\bar\nu \in \mathcal{M}(\overline\Omega)$ fulfill the optimality system* (2.11)-(2.14) *from Theorem 2.17. We define the critical cone by*

$$C_{\bar u,\bar\nu} := \{v \in L^2(\Lambda)\colon \text{(2.16)-(2.18) } \textit{hold}\}$$

with

$$\int_\Lambda (\alpha\bar u + B^*\hat p)v = 0, \qquad \int_\Omega\int_I z_v(t,x)\mathrm{d}t\mathrm{d}\bar\nu = 0, \tag{2.16}$$

$$\begin{aligned} &\int_I z_v(t,\cdot)\mathrm{d}t \geq 0 \quad \text{on } \{\int_I \bar{y}(t,\cdot)\mathrm{d}t = y_\mathrm{a}\},\\ &\int_I z_v(t,\cdot)\mathrm{d}t \leq 0 \quad \text{on } \{\int_I \bar{y}(t,\cdot)\mathrm{d}t = y_\mathrm{b}\}, \end{aligned} \tag{2.17}$$

$$v \leq 0 \quad \text{on } \{\bar{u} = u_\mathrm{b}\}, \quad v \geq 0 \quad \text{on } \{\bar{u} = u_\mathrm{a}\}, \tag{2.18}$$

where $\bar{p}$ and $\hat{p}$ are defined by (2.12) *and* (2.15)*, respectively, and $z_v = S'(\bar{u})v$. If*

$$\alpha\|v\|^2_{L^2(\Lambda)} + \int_Q ((1-\xi''(\bar{y})\mu\nabla\bar{y}\nabla\bar{p})z_v^2 - 2\xi'(\bar{y})z_v\mu\nabla z_v\nabla\bar{p})\mathrm{d}x\mathrm{d}t > 0 \tag{2.19}$$

holds for all $v \in C_{\bar{u},\bar{p}} \setminus \{0\}$, there are $\epsilon, \delta > 0$ such that the quadratic growth condition

$$j(u) \geq j(\bar{u}) + \frac{\delta}{2}\|u-\bar{u}\|^2_{L^2(\Lambda)}$$

holds for all $u \in U_\mathrm{ad}$ such that $\|u-\bar{u}\|_{L^2(\Lambda)} < \epsilon$ and $y_\mathrm{a}(x) \leq \int_I S(u)(t,x)\mathrm{d}t \leq y_\mathrm{b}(x)$ for all $x \in \overline{\Omega}$. In particular, $\bar{u}$ is an local solution of (P^st) *w.r.t. the $L^2(\Lambda)$-topology.*

The setting of pointwise in space and averaged in time state-constraints from Assumption 2.13 has already been motivated in Section 2.3.1. That the application of Theorem 2.8 to this setting goes through under Assumptions 1.5, 1.6 and 1.8 retrospectively completes this motivation.

Proof. We apply Theorem 2.8 with $U_\infty = L^s(\Lambda)$, $U_2 = L^2(\Lambda)$, $K = U_\mathrm{ad}$, $Z = C(\overline{\Omega})$, and $C = \{z \in C(\overline{\Omega})\colon y_\mathrm{a} \leq z \leq y_\mathrm{b} \text{ on } \overline{\Omega}\}$. The properties for the reduced functional j, $j(u) = J(S(u),u)$, required in Assumption 2.7 have already been checked in [**35**, Theorem 4.14]; cf. also Proposition 1.21. Note that the average-in-time map A is linear and continuous from both $L^q(I,C(\overline{\Omega}))$ and $\mathbb{W}^{1,r}(I,(W^{-1,p},W^{1,p})) \hookrightarrow L^r(I,C(\overline{\Omega}))$ into $C(\overline{\Omega})$ for any $q,r \geq 1$. Hence, extension and continuity properties for the derivatives of $g := A\circ S$ in Assumption 2.7.2 immediately follow from Proposition 2.14. Hereby, observe that convergence of $(u_k)_k \subset U_\mathrm{ad}$ to $\bar{u}$ w.r.t. $L^2(\Lambda)$ implies, due to $L^\infty(\Lambda)$-boundedness of U_ad, also convergence w.r.t. $L^s(\Lambda)$ by the Riesz-Thorin interpolation theorem. Therefore, application of Proposition 2.14 is possible. □

Problem (P^st) with averaged in time state-constraints is slightly easier than (P^st) with pointwise in space and time state-constraints from an analytical point of view; this is exactly the reason for considering averaged in time instead of purely pointwise state-constraints. Nevertheless, Theorem 2.18 still illustrates the full strength of Theorem 2.8. To prove C^2-differentiability of the control-to-state map we need controls in $L^s(\Lambda)$ with $s \gg 1$ as in Assumptions 1.5, 1.6 and 1.8 because already existence of solutions to (Eq) relies on such an assumption; cf. Theorem 1.11. Hence, C^2-differentiability, and even well-definedness, of the reduced functional j is guaranteed on $L^s(\Lambda)$ but not necessarily on $L^2(\Lambda)$. For the same reason, a similar situation holds for $g := A\circ S$. It is clear that g is well-defined and C^2-differentiable on $L^s(\Lambda)$. The question whether g is even well-defined on $L^2(\Lambda)$ is not clear. However, according to Proposition 2.9 the positivity condition (2.19)

is equivalent to

$$\alpha\|v\|^2_{L^2(\Lambda)} + \int_Q ((1-\xi''(\bar y)\mu\nabla\bar y\nabla\bar p)z_v^2 - 2\xi'(\bar y)z_v\mu\nabla z_v\nabla\bar p)\mathrm{d}x\mathrm{d}t \geq \gamma\|v\|^2_{L^2(\Lambda)}$$

for all $v \in C_{\bar u,\bar\nu}$ with some $\gamma > 0$, i.e. coercivity of the expression in (2.19) holds w.r.t. $L^2(\Lambda)$ and not w.r.t. $L^s(\Lambda)$. Although the problem necessarily requires us to refer to two nonequivalent norms, a two-norm gap in the formulation of Theorem 2.18 can be avoided. This is the main benefit and novelty of Theorem 2.8; see the introduction chapter or Section 1.5.3 for detailed explanations regarding the two-norm discrepancy and two-norm gaps. Finally, let us emphasize that unlike for first-order necessary conditions we now need the presence of L^2-Tikhonov regularization, i.e. $\alpha > 0$, in Assumption 1.8. This is due to the fact that $\alpha > 0$ ensures that Assumption 2.7.1biii is satisfied; cf. Proposition 1.21 and the respective comments in Section 1.5.3.

2.4. Second-order sufficient conditions for pointwise state-constraints

In the previous section of this chapter we relaxed the type of state-constraints while keeping the regularity assumptions for the equation unchanged. Now, we proceed the other way round and strengthen the regularity assumptions and restrict ourselves to purely time-dependent controls. In return, we establish SSCs for (P^{st}) with pointwise in space and time state-constraints as introduced in Section 2.1. The content of this section is based on Section 5 of our paper [**168**].

We replace Assumptions 1.5, 1.6 and 1.8 by a slightly smoother setting that allows to use the stronger regularity result obtained in [**45**] and summarized in Theorem 1.24. Based on this we derive a result analogous to Proposition 2.14 in the L^p-$W^{2,p}$-setting that finally allows to apply Theorem 2.8 also in the case of pointwise in space and time state-constraints.

Before going into the details, let us briefly motivate the setting chosen in [**168**, Section 5] in an informal way. In Section 2.3.1 we have pointed out that in order to apply Theorem 2.8 to (P^{st}) with pointwise in space and time state-constraints we need to guarantee that, among further conditions, at least

$$S'(\bar u) \in \mathcal{L}\left(L^2(\Lambda), C(\overline{Q})\right) \tag{2.20}$$

holds. According to Assumption 1.8 it holds $B \in \mathcal{L}(L^2(\Lambda), L^2(I, W^{-1,p}_{\Gamma_D}))$ and, hence, due to Lemma 1.12 $S'(\bar u) \in \mathcal{L}\left(L^2(\Lambda), \mathbb{W}^{1,2}(I, (W^{-1,p}_{\Gamma_D}, W^{-1,p}_{\Gamma_D}))\right)$ which is not sufficient for (2.20). Hence, we need to improve the regularity setting. To do so, let us for the moment just assume that $B \in \mathcal{L}(L^2(\Lambda), L^2(I, H^{-\theta,p}_{\Gamma_D}))$ with some $\theta \in [0,1]$. Moreover, assume that for the linearized state equation (1.3) a regularity result in the $H^{-\theta,p}_{\Gamma_D}$-$H^{2-\theta,p}_{\Gamma_D}$-setting is available, i.e. that

$$S'(\bar u) \in \mathcal{L}\left(L^2(\Lambda), \mathbb{W}^{1,2}(I, (H^{-\theta,p}_{\Gamma_D}, H^{2-\theta,p}_{\Gamma_D}))\right) \tag{2.21}$$

holds. Utilizing the Bochner-Sobolev embedding from Proposition 1.1.2 we can prove (2.20) if we can show

$$(H^{-\theta,p}_{\Gamma_D}, H^{2-\theta,p}_{\Gamma_D})_{1/2,2} \hookrightarrow C(\overline{\Omega}).$$

To compute this interpolation space, let us, in addition, assume that there is an simultaneous extension operator for $H^{-\theta,p}_{\Gamma_D} \to H^{-\theta,p}_{\Gamma_D}(\mathbb{R}^d)$ and $H^{2-\theta,p}_{\Gamma_D} \to H^{2-\theta,p}_{\Gamma_D}(\mathbb{R}^d)$. In that case the retraction-coretraction theorem [**267**, Theorem 1.2.4], interpolation results for function spaces on $\mathbb{R}^d$, and standard Sobolev embeddings yield

$$(H^{-\theta,p}_{\Gamma_D}, H^{2-\theta,p}_{\Gamma_D})_{1/2,2} \hookrightarrow H^{\vartheta,p}_{\Gamma_D} \hookrightarrow C(\overline{\Omega})$$

as long as $\vartheta < 1-\theta$ and $\vartheta > \frac{d}{p}$. Since $\frac{d}{p} < 1$ might be arbitrarily close to 1, we need to choose $\theta = 0$, i.e. $H^{-\theta,p}_{\Gamma_D} = L^p$ and $H^{2-\theta,p}_{\Gamma_D} = W^{2,p} \cap W^{1,p}_{\Gamma_D}$.

For the rest of this section we will make the above reasoning mathematically precise. Let us briefly indicate how this results in the formulation of Assumption 2.19 below: first of all, we need to ensure $B \in \mathcal{L}(L^2(\Lambda), L^2(I, L^p))$ which is done by choosing the purely time-dependent control setting as described in Example 1.9.3 with fixed actuators $b_i \in L^p$. Making a regularity result for the linearized state equation in the L^p-$W^{2,p} \cap W^{1,p}_{\Gamma_D}$-setting available is more delicate; in particular, this means that the appearing elliptic operators $\mathcal{A}(y) + \mathcal{A}'(y)$ have to be well-defined as maps $W^{2,p} \to L^p$ which can only be expected to hold in a rather smooth setting as introduced below. Fortunately, we can build on the analysis from [**45**] and a corresponding regularity result for the state equation in the L^p-$W^{2,p}$-setting that has already been summarized in Theorem 1.24. For brevity and clarity we do not consider the equation from Theorem 1.24 in its full generality, where, e.g., an additional semilinear term is allowed. Instead, we keep the structure of (Eq) unchanged and enforce the assumptions on the underlying data in such a way that the results from [**45**] now apply to (Eq) as well. Finally, we note that the existence of a simultaneous extension operator for L^p and $W^{2,p}$ does not require additional assumptions once Ω is assumed to be a sufficiently smooth domain.

2.4.1. Regularity assumptions for the state equation. From now on we consider (P^{st}) with pointwise in space and time state-constraints (Assumption 2.1) under the following additional assumptions.

Assumption 2.19.

1. $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ is a bounded domain with $C^{1,1}$-boundary, and homogeneous Dirichlet boundary conditions hold on the entire boundary, i.e. $\Gamma_D = \partial\Omega$.
2. Let Assumption 1.6 on μ and ξ hold and assume in addition that μ is Lipschitz continuous as map $\Omega \to \mathbb{R}^{d\times d}$.
3. Choose $p > d$ and $s > 2$ such that $\frac{1}{s} < \frac{1}{2}(1 - \frac{d}{p})$. The set of admissible controls is given by

   $$U_{\mathrm{ad}} := \{u \in L^{2s}(I, \mathbb{R}^m) : u_{\mathrm{a}} \le u \le u_{\mathrm{b}} \text{ on } I\}$$

   with control-bounds $u_{\mathrm{a}}, u_{\mathrm{b}} \in L^\infty(I, \mathbb{R}^m)$, and for fixed actuator functions $b_i \in L^p$, $i = 1, \dots, m$, we define

   $$B\colon L^{2s}(I, \mathbb{R}^m) \to L^{2s}(I, L^p),\ u \mapsto \sum_{i=1}^m u_i b_i.$$

   The initial value y_0 for the state equation fulfills

   $$y_0 \in (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})_{1-1/s,s} \cap (W^{-1,2p}_{\Gamma_D}, W^{1,2p}_{\Gamma_D})_{1-1/(2s),2s} \cap C(\Omega),$$

and the desired state has regularity $y_d \in L^\infty(I, L^p)$. We fix $\alpha > 0$.

As explained above, Assumption 2.19 is the adaptation of the assumptions from [**45**], see Theorem 1.24 for a summary, applied to the setting described in Assumptions 1.5, 1.6 and 1.8. The reason for the integrability exponent $2s$ w.r.t. time in part 3 of the assumption is of rather technical nature and will be explained in Remark 2.25 below. Unlike in [**45**] we have to restrict ourselves to purely time-dependent controls as introduced in [**91**]. The reason has already been indicated above and is the following; cf. also [**91**, Remark 2]: when switching from controls in $U_\infty = L^{2s}(I, \mathbb{R}^m)$ to controls in $U_2 = L^2(I, \mathbb{R}^m)$, only time integrability decreases, but the spatial regularity of the right-hand sides of the PDEs is not affected. This is crucial for obtaining the required regularity for the derivatives of the control-to-state map. From the applied point of view, having only finitely many predefined actuators to influence a system might also seem reasonable. However, note that L^p-regularity (unlike $W_{\Gamma_D}^{-1,p}$-regularity in Example 1.9.3) of the fixed actuator functions now excludes any possibility of boundary control. Moreover, compared to the rough regularity setting considered in this thesis so far, Assumption 2.19 is a really strong assumption. Nevertheless, as we have pointed out in the introduction of this chapter, such kind of restriction is exactly what we have to expect when aiming at second-order results for control problems with nonlinear, time-dependent PDEs in the presence of pointwise in space and time state-constraints.

To make the relation of Assumption 2.19 to the previous rough regularity setting for (Eq), i.e. Assumptions 1.5, 1.6 and 1.8, clear, let us recall from [**45**, p. 609] the following observation: $C^{1,1}$-smoothness of $\partial\Omega$ combined with pure homogeneous Dirichlet boundary conditions on $\Gamma_D = \partial\Omega$ and Lipschitz-continuity of μ implies that

$$-\nabla \cdot \mu\nabla + 1\colon\ W_{\Gamma_D}^{1,q} \to W_{\Gamma_D}^{-1,q} \tag{2.22}$$

is a topological isomorphism for any $q \in (1, \infty)$. Consequently, Assumption 2.19 is indeed a (heavily) tightened version of Assumptions 1.5, 1.6 and 1.8.

2.4.2. Improved regularity of the state. From Theorem 1.24 we obtain immediately the following regularity result that will be the cornerstone of our further analysis.

Theorem 2.20 ([**45**], Theorem 2.3)**.** *Under Assumption 2.19 the control-to-state map S introduced in Section 1.3 is well-defined from $L^{2s}(I, \mathbb{R}^m)$ to $\mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W_{\Gamma_D}^{1,p}))$.*

From Section 1.3, formulas (1.4) and (1.8), recall that under Assumptions 1.5, 1.6 and 1.8 or Assumptions 1.5, 1.6 and 1.10 $C^{0,\sigma}(\overline{Q})$- or $C^{0,\kappa}(\overline{I}, W_{\Gamma_D}^{1,p})$-regularity, respectively, has been proven in [**35**] for solutions of (Eq). The above regularity result due to [**45**] yields considerably more regularity as the following corollary shows.

Corollary 2.21 ([**168**], Corollary 5.4)**.** *Under Assumption 2.19 there are some $\rho, \kappa > 0$ such that*

$$\mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W_{\Gamma_D}^{1,p})) \hookrightarrow_c C^{0,\rho}(I, C^{1,\kappa}).$$

The proof of this embedding makes use of a similar interpolation argument as we have sketched at the beginning of this section during the motivation of the chosen setting. Now, it can be made precise because the existence of a simultaneous Sobolev extension operator is guaranteed for domains with $C^{1,1}$-boundary.

Proof. Choose $\frac{1}{2}(1+\frac{d}{p}) < \theta < 1 - \frac{1}{s}$ and set $\rho = 1 - \frac{1}{s} - \theta > 0$. By Proposition 1.1.3 it holds

$$\mathbb{W}^{1,s}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D})) \hookrightarrow_c C^{0,\rho}(I,(L^p,W^{2,p})_{\theta,1}).$$

Further, it is well-known that $(L^p,W^{2,p})_{\theta,1} \hookrightarrow [L^p,W^{2,p}]_\theta$. Since Ω is in particular a domain with Lipschitz boundary, there is a bounded linear extension operator $L^p \to L^p(\mathbb{R}^d)$ that restricts to a bounded extension operator $W^{2,p} \to W^{2,p}(\mathbb{R}^d)$; we have summarized this result from [**235**] in Theorem 0.3. Thus, a standard argument utilizing the retraction-coretraction theorem ([**267**, Theorem 1.2.4], [**12**, Proposition I.2.3.2]) shows that it suffices to prove $[L^p(\mathbb{R}^d),W^{2,p}(\mathbb{R}^d)]_\theta \hookrightarrow C^{1,\kappa}(\mathbb{R}^d)$. The latter follows from $[L^p(\mathbb{R}^d),W^{2,p}(\mathbb{R}^d)]_\theta = H^{2\theta,p}(\mathbb{R}^d)$ [**267**, Theorem 4.3.2.2] and standard Sobolev embeddings on $\mathbb{R}^d$ with $\kappa = 2\theta - \frac{d}{p} - 1 > 0$ [**267**, Theorem 2.8.1]. □

2.4.3. Improved regularity of the derivatives of the control-to-state map. We provide an improved version of Lemma 1.12 under the strengthened regularity Assumption 2.19. This is the regularity result for the linearized state equation in the L^p-$W^{2,p}$-setting that corresponds to (2.21) in our motivation of the chosen setting at the beginning of this section.

The improved regularity of the state from Theorem 2.20 is the crucial point because we can show that the domain of $-\nabla\cdot\xi(y(t))\mu\nabla$ in L^p is independent of $t\in I$ for $y\in C^{0,\rho}(I,C^{1,\kappa})$. Hence, it is possible to show that $\mathcal{A}(y)$ and $\mathcal{A}(y)+\mathcal{A}'(y)$ exhibit maximal parabolic regularity on L^p-spaces, which finally allows to prove the desired regularity result analogous to Lemma 1.12 and Proposition 2.14. The approach is similar to [**35**] with the essential difference that the weaker regularity $y\in\mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D}))$ for the states in [**35**, Section 3.2] suffices to show constant domains and maximal parabolic regularity on $H^{-\zeta,p}_{\Gamma_D}$ for certain $\zeta\in(0,1)$ close to 1, but not on L^p; cf. the proof of [**35**, Proposition 3.17]. However, an analysis carried out on $H^{-\zeta,p}_{\Gamma_D}$ will not suffice for the derivation of SSCs for (P^{st}) in case of pointwise in space and time state-constraints as we have pointed out in Section 2.3.3.

The following observation is rather trivial in our case. We state it due to its importance for the following results.

Lemma 2.22 ([**168**], Lemma 5.5)**.** *Under Assumption 2.19 let $\eta\in W^{1,\infty}$ with $\eta\geq\eta_\bullet>0$ on Ω. Then it holds:*

1. $\mathrm{Dom}_{L^p}(-\nabla\cdot\eta\mu\nabla+1)\cong\mathrm{Dom}_{L^p}(-\nabla\cdot\mu\nabla+1)=W^{2,p}\cap W^{1,p}_{\Gamma_D}$, *i.e.* $-\nabla\cdot\eta\mu\nabla+1$ *is a topological isomorphism* $W^{2,p}\cap W^{1,p}_{\Gamma_D}\to L^p$.
2. *The map*

$$W^{1,\infty}\to\mathcal{L}(W^{2,p}\cap W^{1,p}_{\Gamma_D},L^p),\qquad \eta\mapsto-\nabla\cdot\eta\mu\nabla$$

is bounded linear.

Similar results have been obtained in [**134**, Lemmas 6.5, 6.7, Corollary 6.8] if (Eq) is considered on Bessel potential spaces instead of L^p.

Proof. 1. This follows from [**127**, Theorem 2.4.2.5] for instance.

2. It holds $-\nabla \cdot \eta\mu\nabla \in \mathcal{L}(W^{2,p} \cap W^{1,p}_{\Gamma_D}, L^p)$ for any $\eta \in W^{1,\infty}$, with linear dependence on η. A short computation shows

$$\|-\nabla \cdot \eta\mu\nabla\varphi\|_{L^p} \lesssim \|\eta\|_{W^{1,\infty}} \|\varphi\|_{W^{2,p}\cap W^{1,p}_{\Gamma_D}}$$

for any $\eta \in W^{1,\infty}$, $\varphi \in W^{2,p} \cap W^{1,p}_{\Gamma_D}$, which verifies boundedness. □

The following lemma is a first step towards the analysis of the linearized state equation on L^p, where linearization takes place at some y that exhibits the regularity obtained in Theorem 2.20 for solutions of (Eq). The linearized state equation is given by the parabolic PDE associated to the nonautonomous linear parabolic operator $\mathcal{A}(y) + \mathcal{A}'(y)$. Regularity of the first summand of this operator, i.e. $\mathcal{A}(y)$, is provided by the following lemma; the whole operator will be addressed in Lemma 2.24.

Lemma 2.23 ([**168**], Lemma 5.6). *Let Assumption 2.19 hold and fix $y \in \mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$. The nonautonomous linear parabolic operator $\mathcal{A}(y)$ exhibits maximal parabolic regularity on $L^r(I, L^p)$, $r \in (1,\infty)$, i.e. the solution map $(w, w_0) \mapsto z$ of the equation*

$$\partial_t z + \mathcal{A}(y)z = w, \qquad z(0) = w_0,$$

is linear and bounded as a map

$$L^r(I, L^p) \times (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})_{1/r',r} \to \mathbb{W}^{1,r}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})).$$

Moreover, the corresponding operators norms are bounded uniformly for y coming from a bounded set in $\mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$.

The proof relies on the same technique as in [**35**, Theorem 3.20]. Nevertheless, the present situation is slightly easier than in [**35**], because the additional regularity assumptions ensure that the domains of $\mathcal{A}(y(t))$ in L^p stay independent of t.

Proof. We apply [**35**, Lemma D.1]; see also [**230**, Corollary 14]. First, note that L^p is an UMD space; see, e.g., [**12**, Section III.4.4] for the definition. Uniform resolvent estimates and uniform $\mathcal{R}$-sectoriality for $A(t) := -\nabla \cdot \xi(y(t))\mu\nabla$ on L^p have already been established; see formulas (3.16) and Lemma 3.12 in [**35**]. Note that uniformity already holds for y's coming from a bounded set in $C^{0,\alpha}(\overline{Q})$, which is a much weaker assumption than in the present case. It remains to check the so-called Acquistapace-Terreni condition on L^p. The latter was done in [**35**] on the Bessel potential spaces $H^{-\zeta,p}_{\Gamma_D}$ for appropriate $\zeta \in (0,1)$ but not on L^p. As in the proof of [**35**, Proposition 3.18] we write with help of the resolvent calculus:

$$\begin{aligned}(A(t)+1)R(z, A(t)+1)[(A(t)+1)^{-1} - (A(s)+1)^{-1}] \\ = R(z, A(t)+1)[A(t) - A(s)](A(s)+1)^{-1}.\end{aligned}$$

From Lemma 2.22.2 it follows that

$$\|A(t) - A(s)\|_{\mathcal{L}(W^{2,p}\cap W^{1,p}_{\Gamma_D}, L^p)} \le c\|\xi(y)\|_{C^\rho(I, W^{1,\infty})}|t-s|^\rho$$

with $c > 0$ independent of y. Employing formula (3.16) from [**35**] there is $\theta \in (0, \pi/2)$ such that $\|R(z, A(t)+1)\|_{\mathcal{L}(L^p)} \le c|z|^{-1}$ for all $z \in \mathbb{C} \setminus \overline{\Sigma_\theta}$, $t \in I$, with $\Sigma_\theta = \{z \in \mathbb{C} \setminus \{0\}\colon |\arg z| < \theta\}$, and finally it follows again from Lemma 2.22 that

$$\|(A(s)+1)^{-1}\|_{\mathcal{L}(L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})} \le c\|\xi(y)\|_{L^\infty(I, W^{1,\infty})}$$

with a constant c independent of y. Together, this shows the Acquistapace-Terreni condition,

$$\|(A(t)+1)R(z, A(t)+1)[(A(t)+1)^{-1} - (A(s)+1)^{-1}]\|_{\mathcal{L}(L^p)} \le C|t-s|^\rho |z|^{-1}$$

for all $z \in \mathbb{C} \setminus \overline{\Sigma_\theta}$, $t, s \in I$, with the constant $C > 0$ depending on the $C^\rho(I, W^{1,\infty})$-norm of y. Therefore, C can be chosen uniformly for y's coming from a bounded set in $\mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$ due to Corollary 2.21. □

Now, we consider maximal parabolic regularity for the linearized state equation. This extends Lemma 1.12, i.e. [**35**, Proposition 4.4], or [**45**, Theorem 3.2], where maximal parabolic regularity on $W^{-1,p}$ has been dealt with.

Lemma 2.24 ([**168**], Lemma 5.7)**.** *Let Assumption 2.19 hold, and fix $y \in \mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$. For any $r \in (1, s]$ and $f \in L^r(I, L^p)$, the linearized state equation*

$$\partial_t w + \mathcal{A}(y)w + \mathcal{A}'(y)w = f, \qquad w(0) = 0,$$

has a unique solution $w \in \mathbb{W}^r(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$. The nonautonomous operator $\mathcal{A}(y) + \mathcal{A}'(y)$ has maximal parabolic regularity on $L^r(I, L^p)$ for $r \in (1, s]$.

Proof. Maximal parabolic regularity of $\mathcal{A}(y)$ on $L^r(I, L^p)$, $r \in (1, \infty)$, has been shown in Lemma 2.23. Corollary 2.21 together with Lemma 2.22 implies continuity of the map

$$I \to \mathcal{L}(W^{2,p} \cap W^{1,p}_{\Gamma_D}, L^p),\ t \mapsto -\nabla \cdot \xi(y(t))\mu\nabla,$$

from which we conclude that each autonomous operator $-\nabla \cdot \xi(y(t))\mu\nabla \in \mathcal{L}(W^{2,p} \cap W^{1,p}_{\Gamma_D}, L^p)$, $t \in I$, has in fact maximal parabolic regularity on L^p. This follows from Proposition 1.4.2. Regarding the second summand, $\mathcal{A}'(y)$, we observe that the map

$$I \to \mathcal{L}(W^{1,\infty}, L^p),\ t \mapsto (\psi \mapsto -\nabla \cdot \xi'(y(t))\psi\mu\nabla y(t))$$

is L^s-integrable w.r.t. time: this follows from the continuity of the map

$$W^{1,\infty} \to \mathcal{L}(W^{2,p} \cap W^{1,p}_{\Gamma_D}, L^p),\ \eta \mapsto -\nabla \cdot \eta\mu\nabla,$$

see Lemma 2.22.2, together with $\xi'(y) \in L^\infty(I, W^{1,\infty})$, the continuity of the product on $W^{1,\infty} \times W^{1,\infty}$, and the fact that $y \in L^s(I, W^{2,p} \cap W^{1,p}_{\Gamma_D})$. Hence, we have just shown

$$\mathcal{A}'(y) = (t \mapsto (\psi \mapsto -\nabla \cdot \xi'(y)\psi\mu\nabla y)) \in L^s(I, \mathcal{L}(W^{1,\infty}, L^p)) \hookrightarrow L^s(I, \mathcal{L}((L^p, W^{2,p})_{\theta,\infty}, L^p))$$

with some $1 - 1/s > \theta > \hat\theta > \frac{1}{2} + \frac{d}{2p}$. Hereby, we made use of the embedding $(L^p, W^{2,p})_{\theta,\infty} \hookrightarrow [L^p, W^{2,p}]_{\hat\theta} \hookrightarrow W^{1,\infty}$; cf. the proof of Corollary 2.21. From Proposition 1.4.1 we conclude maximal parabolic regularity of $\mathcal{A}(y) + \mathcal{A}'(y)$ on

$L^r(I,L^p)$ for $r\in(1,s)$. Similar to the proof of [**35**, Proposition 4.4] we invoke Proposition 1.4.3 to get maximal parabolic regularity on $L^s(I,L^p)$. □

We point out that Lemma 2.24 and Theorem 2.20 do *not* allow immediately to conclude differentiability of the solution map of (1.1) from $L^r(I,L^p)$ to $\mathbb{W}^{1,r}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$. Of course, for $\frac{1}{r}<1-\frac{d}{2p}$, e.g., $r=2$, there is an embedding $\mathbb{W}^{1,r}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))\hookrightarrow C(\overline{Q})$, which can be shown with a similar argument as for Corollary 2.21. Hence, the map

$$F\colon \mathbb{W}^{1,r}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))\times L^r(I,L^p)\to L^r(I,L^p)\times(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D})_{1/r',r},\\ (y,v)\mapsto(\partial_t y+\mathcal{A}(y)-v,y(0)-y_0),$$

is continuously Fréchet differentiable. Further, for $r\in(1,s]$ the partial derivative $\partial_y F(y,v)$ is even continuously invertible; cf. Lemma 2.24. Nevertheless, the fact that prevents us from application of the implicit function theorem is that we would first require a *well-defined* solution map $v\mapsto y(v)$ associated with $F(y,v)=0$, and we do not have such a map at hand. To obtain solutions to (1.1) in $\mathbb{W}^{1,s}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$ we need right-hand sides $v\in L^{2s}(I,L^p)$ and not in $L^s(I,L^p)$; see Theorem 2.20. For $v\in L^s(I,L^p)$ we do not know whether there exists some $y\in\mathbb{W}^{1,s}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$ such that $F(y,v)=0$. On the other hand, $\partial_y F(y,v)$ cannot be invertible from $\mathbb{W}^{1,s}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$ to $L^{2s}(I,L^p)$, because invertibility of $\partial_y F(y,v)$ holds between $\mathbb{W}^{1,r}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$ and $L^r(I,L^p)$, $r\in(1,s]$; cf. Lemma 2.24.

Remark 2.25. Double time integrability on the right-hand side of (1.1) in Theorem 2.20 is due to the technique applied in the proof of [**45**, Theorem 2.3]. For a short outline we refer the reader to the proof of Lemma 2.27 below.

The following lemma is the first step towards an analogue to Proposition 2.14.1 under Assumption 2.19. Particularly, the regularity of the $\mathcal{A}''$-term appearing in the second derivative of the control-to-state map can be essentially improved in the present case. Even in this highly regular setting $\mathcal{A}''(y)w^2$ is from $L^r(I,W^{-1,p}_{\Gamma_D})$, i.e. a distributional object in general, which illustrates the difficulty of this term. We also refer the reader to the end of Section 2.3.3 where we have compared this situation to the semilinear parabolic case.

Lemma 2.26 ([**168**], Lemma 5.9)**.** *For $y\in\mathbb{W}^{1,s}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$, $w\in\mathbb{W}^{1,2}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))$ it holds*

$$\|\mathcal{A}''(y)w^2\|_{L^r(I,W^{-1,p}_{\Gamma_D})}\le c_{y,r}\|w\|^2_{\mathbb{W}^{1,2}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))}$$

for $r\in(1,\infty)$. The constant $c_{y,r}$ can be chosen uniformly w.r.t. y coming from a bounded set in $L^\infty(I,W^{1,p})$.

Proof. This follows from the definition of $\mathcal{A}''$ and Hölder's inequality. We have to make use of the embeddings $\mathbb{W}^{1,2}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))\hookrightarrow C(\overline{Q})$ and $\mathbb{W}^{1,2}(I,(L^p,W^{2,p}\cap W^{1,p}_{\Gamma_D}))\hookrightarrow L^q(I,W^{1,p})$ for every $q\in(1,\infty)$, that can be shown similarly as in Corollary 2.21. □

The following lemma is the last auxiliary result before we will be able to verify the assumptions of Theorem 2.8 in the proposition thereafter.

Lemma 2.27 ([**168**], Lemma 5.10). *The solution map of the state equation* (1.1) *is continuous from* $L^{2s}(I, L^p)$ *to* $\mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D}))$.

This result is implicitely, but not explicitly contained in [**45**]. There, differentiability and, consequently, continuity of the control-to-state map have been addressed in the $W^{-1,p}_{\Gamma_D}$-setting; cf. [**45**, Theorem 3.2]. As outlined after Lemma 2.24, argueing via the implicit function theorem is not possible here. To prove continuous dependence we go through the steps in [**45**] tracking continuous dependence of the quantities under consideration.

Proof. It is well-known that $L^{2s}(I, L^p) \hookrightarrow L^{2s}(I, W^{-1,2p}_{\Gamma_D})$ and that solution map of (1.1) is continuous (in fact, even C^2) from $v \in L^{2s}(I, W^{-1,2p}_{\Gamma_D})$ to $y = y(v) \in \mathbb{W}^{1,2s}(I, (W^{-1,2p}, W^{1,2p}_{\Gamma_D}))$. Hereby, existence of a solution is clear by [**45**, Theorem 2.1] and differentiability of the solution map follows using the implicit function theorem similarly as in the proof of [**45**, Theorem 3.2]: the required invertibility property is assured by maximal parabolic regularity of $\mathcal{A}(y) + \mathcal{A}'(y)$ on $L^{2s}(I, W^{-1,2p}_{\Gamma_D})$, which is proven similarly as in the proof of [**35**, Proposition 4.4] with s and p replaced by $2s$ and $2p$, respectively; cf. also the similar proof of Lemma 2.24 in the L^p-setting. Next, following the main idea in the proof of [**45**, Theorem 2.3] we rewrite equation (1.1) as

$$\partial_t z - \xi \nabla \cdot \mu \nabla z = v + \nabla \xi \cdot \mu \nabla y,$$

with $y = y(v) \in \mathbb{W}^{1,2s}(I, (W^{-1,2p}, W^{1,2p}_{\Gamma_D}))$ being the solution to (1.1) and $\xi = \xi(y(v))$. It is clear that the right-hand side measured in $L^s(I, L^p)$ depends continuously on ξ in $L^{2s}(I, W^{1,2p})$ and on y in $L^{2s}(I, W^{1,2p})$, respectively, i.e. on v in $L^{2s}(I, W^{-1,2p})$ by the above consideration. Further, due to the embedding

$$\mathbb{W}^{1,2s}(I, (W^{-1,2p}, W^{1,2p}_{\Gamma_D})) \hookrightarrow C(\overline{Q})$$

also $\xi = \xi(y(v))$ depends continuously in $C(\overline{Q})$ on v. Finally, the map

$$C(\overline{Q}) \to \mathcal{L}(\mathbb{W}^{1,s}(I, (L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})), L^s(I, L^p)), \qquad \xi \mapsto \partial_t - \xi \nabla \cdot \mu \nabla$$

is continuous. Therefore, using [**45**, Lemma 2.4] the solution $z = y$ depends continuously on ξ in $C(\overline{Q})$ and y in $\mathbb{W}^{1,2s}(I, (W^{-1,2p}, W^{1,2p}_{\Gamma_D}))$, and thus on v in $L^{2s}(I, L^p)$. □

The following proposition is our analogue to Proposition 2.14 for the present section. It provides the main steps in checking Assumption 2.7 for the setting described by Assumption 2.19, and therefore forms the main part of the proof of our second main result, SSCs for (P^{st}) with pointwise in space and time state-constraints, below. Retrospectively, this result completes the motivation of Assumption 2.19 that we have provided at the beginning of Section 2.4.

Proposition 2.28 ([**168**], Proposition 5.11). *Under Assumption 2.19 the control-to-state map is twice continuously Fréchet differentiable as map*

$$L^s(I, \mathbb{R}^m) \to \mathbb{W}^{1,s}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$$

and the following continuation and continuity properties hold for the respective derivatives:

1. *For any $u \in L^{2s}(I,\mathbb{R}^m)$, $S'(u)$ and $S''(u)$ extend to continuous linear and bilinear forms on $L^2(I,\mathbb{R}^m)$ with values in $C(\overline{Q})$, respectively.*
2. *Let $(u_k)_k \subset L^{2s}(I,\mathbb{R}^m)$ converge to $\bar u$ strongly in $L^{2s}(I,\mathbb{R}^m)$ and $(v_k)_k \subset L^2(I,\mathbb{R}^m)$ converge weakly in $L^2(I,\mathbb{R}^m)$ to some v. Then $S'(u_k)v_k \rightharpoonup S'(\bar u)v_k$ and $S''(u_k)v_k^2 \rightharpoonup S''(\bar u)v^2$ weakly in $C(\overline{Q})$.*

The proof has similar structure as the one of Proposition 2.14. Nevertheless, we give some details due to the importance of the result.

Proof. Differentiability of the control-to-state map and the formulas for the respective derivatives follow from Lemma 1.12. Note that Assumption 2.19 indeed suffices to invoke this result as has been pointed out at the end of Section 2.4.1. The extension property for the first derivative follows from Lemma 2.24 with $r = 2$ and the first embedding in the proof of Lemma 2.26. For the continuation of the second derivative combine the continuation property for $S'(u)$ with Lemmas 2.24 and 2.26 and the second embedding from Lemma 1.13. It remains to check the continuity properties: as an auxiliary result, we first show that

$$S'(u_k) \to S'(\bar u) \quad \text{in } \mathcal{L}(L^r(I,\mathbb{R}^m), \mathbb{W}^{1,r}(I,(L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})))$$

for any $r \in (1,\infty)$. To do so, it suffices, by continuity of operator inversion, to show convergence

$$\mathcal{A}(y_k) + \mathcal{A}'(y_k) \to \mathcal{A}(\bar y) + \mathcal{A}'(\bar y) \quad \text{in } \mathcal{L}(\mathbb{W}^{1,r}(I,(L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})), L^r(I,L^p)).$$

This can be done using Lemma 2.27, Hölder's inequality, and

$$\mathbb{W}^{1,r}(I,(L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})) \hookrightarrow L^q(I, W^{1,\infty})$$

for some q such that $\frac{1}{q} + \frac{1}{s} \le \frac{1}{r}$, which can be shown by the same technique as for Corollary 2.21. Having at hand this auxiliary result, the continuity property for the first derivative follows similarly as in the proof of Proposition 2.14. For the second derivative we also argue similarly as in the proof of Proposition 2.14: due to the second embedding from Lemma 1.13 for $r > \frac{2p}{p-d}$ it suffices to show that

$$\tilde S'(Bu_k) \to \tilde S'(B\bar u) \quad \text{in } \mathcal{L}(L^r(I, W^{-1,p}_{\Gamma_D}), \mathbb{W}^{1,r}(I,(W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})))$$

and $\mathcal{A}''(y_k)[w_k]^2 \rightharpoonup \mathcal{A}''(y)[w]^2$ weakly in $L^r(I, W^{-1,p}_{\Gamma_D})$. We leave the details to the reader. □

2.4.4. Second-order sufficient conditions. Now, we can apply Theorem 2.8 to (P^{st}) under Assumptions 2.1 and 2.19 and formulate SSCs for (P^{st}) with pointwise in space and time state-constraint. Compared to Theorem 2.18 we crucially rely on the improved regularity results due to the strengthened regularity Assumption 2.19.

Theorem 2.29 ([**168**], Theorem 5.12). *Let Assumption 2.19 and Assumption 2.1.1 hold, and let $\bar u \in L^{2s}(I,\mathbb{R}^m)$, $\bar y \in \mathbb{W}^{1,s}(I,(L^p, W^{2,p} \cap W^{1,p}_{\Gamma_D})) \cap Y_{\mathrm{ad}}$ and $\bar\lambda \in \mathcal{M}(\overline{Q})$ fulfill the FONs (2.1)-(2.4) from Theorem 2.4. We define the cone of critical directions by*

$$C_{\bar u,\bar\lambda} := \{v \in L^2(I,\mathbb{R}^m)\colon (2.23) - (2.25)\ \textit{hold}\},$$

with

$$\int_I (\alpha \bar{u}(t) + B^* \hat{p}(t))^T v(t) \mathrm{d}t = 0, \qquad \int_{\overline{Q}} z_v \mathrm{d}\bar{\lambda} = 0, \tag{2.23}$$

$$z_v(t,x) \le 0 \quad \text{on } \{\bar{y} = y_{\mathrm{b}}\},\ z_v(t,x) \ge 0 \quad \text{on } \{\bar{y} = y_{\mathrm{a}}\}, \tag{2.24}$$

$$v_i(t) \le 0 \quad \text{if } \bar{u}_i(t) = u_{\mathrm{b},i}(t), \quad v_i(t) \ge 0 \quad \text{if } \bar{u}_i(t) = u_{\mathrm{a},i}(t), \tag{2.25}$$

where $\hat{p}$ is defined by (2.15) *and $z_v = S'(\bar{u})v$. If*

$$\alpha \|v\|^2_{L^2(I,\mathbb{R}^m)} + \int_Q ((1 - \xi''(\bar{y})\mu \nabla \bar{y} \nabla \bar{p}) z_v^2 - 2\xi'(\bar{y}) z_v \mu \nabla z_v \nabla \bar{p}) \mathrm{d}x\mathrm{d}t \ > \ 0 \tag{2.26}$$

holds for all $v \in C_{\bar{u},\bar{\lambda}} \setminus \{0\}$, then $\bar{u}$ is an $L^2(I,\mathbb{R}^m)$-local minimizer for (P^{st}) *and there are $\epsilon, \delta > 0$ such that the quadratic growth condition*

$$j(u) \ge j(\bar{u}) + \frac{\delta}{2} \|u - \bar{u}\|^2_{L^2(I,\mathbb{R}^m)}$$

holds for all $u \in U_{\mathrm{ad}}$ that satisfy $\|u - \bar{u}\|_{L^2(I,\mathbb{R}^m)} < \epsilon$ and $S(u) \in Y_{\mathrm{ad}}$.

Proof. We apply Theorem 2.8 with $U_\infty = L^s(I,\mathbb{R}^m)$, $U_2 = L^2(I,\mathbb{R}^m)$, $Z = C(\overline{Q})$, $K = U_{\mathrm{ad}}$, $C = Y_{\mathrm{ad}}$. As already observed in the proof of Theorem 2.18, the assumptions on the reduced functional j from Assumption 2.7.1 have been verified in [**35**]; this has been summarized in Proposition 1.21. Assumption 2.7.2 for $g = S$ is fulfilled due to Proposition 2.28. The crucial point is as in the proof of Theorem 2.18 to observe that due to L^∞-boundedness of U_{ad} convergence w.r.t. the $L^2(I,\mathbb{R}^m)$-norm implies both $L^{2s}(I,\mathbb{R}^m)$- and $L^s(I,\mathbb{R}^m)$-convergence. □

To put this result into context, let us note that the problem formulation requires two different norms, as it was the case also in Example 0.2 in the introduction. Reduced functional and control-to-state map are well-defined and C^2-Fréchet on $L^s(I,\mathbb{R}^m)$ with some $s \gg 2$ but not necessarily on $L^2(I,\mathbb{R}^m)$. The positivity condition (2.26) implies the coercivity condition

$$\alpha \|v\|^2_{L^2(I,\mathbb{R}^m)} + \int_Q ((1 - \xi''(\bar{y})\mu \nabla \bar{y} \nabla \bar{p}) z_v^2 - 2\xi'(\bar{y}) z_v \mu \nabla z_v \nabla \bar{p}) \mathrm{d}x\mathrm{d}t \ge \gamma \|v\|^2_{L^2(I,\mathbb{R}^m)}$$

for all $v \in C_{\bar{u},\bar{\lambda}}$ and some $\gamma > 0$; cf. Proposition 2.9. Thus, differentiability of j and S and coercivity of the expression in (2.26) hold w.r.t. different norms. However, unlike in Example 0.2 or some previous results on SSCs for parabolic problems, e.g., [**91**], it is possible to state the quadratic growth condition in Theorem 2.29 only referring to the $L^2(I,\mathbb{R}^m)$-norm: similarly to Theorem 2.18 and Example 2.10 the occurrence of a two-norm gap can be avoided. As for Theorem 2.18 we point out that the L^∞-boundedness of U_{ad} and the assumption $\alpha > 0$ are crucial for Theorem 2.29.

2.5. Numerical illustration

We conclude this chapter by some numerical examples. As explained in the introduction of this chapter or in Section 2.3.1, considering state-constraints that are averaged w.r.t. time (instead of averaged w.r.t. space) is a new feature of our work [**168**]. Therefore, we now provide a numerical illustration of this type of constraint. In particular, we consider averaging on different subintervals $I_{\mathrm{obs}} \subset I$ and compare this with the case of pointwise in space and time state-constraints.

Our computations indicate that standard numerical schemes that have already successfully been applied to other PDE-constrained optimization problems with state-constraints also allow to handle the problem class under consideration in this chapter. The theoretical analysis of the discretization scheme or the regularization method applied to the state-constraints are beyond the scope of this thesis and not addressed.

2.5.1. Computational approach. In order to solve the state-constrained problem (P^{st}) numerically, we employ Moreau-Yosida regularization; see, e.g., [**176, 147**]. In essence, instead of (P^{st}) we solve the regularized problem

$$(\mathrm{P}^{\mathrm{st}}_\beta) \qquad \begin{cases} \min\limits_{y,u} J(y,u) := \dfrac{1}{2}\|y - y_d\|^2_{L^2(I\times\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(\Lambda)} + \dfrac{\beta}{2}P(y), \\ \text{subject to} \quad u \in U_{\mathrm{ad}} \quad \text{and} \quad (\mathrm{Eq}), \end{cases}$$

where $\beta = \beta_{\mathrm{final}} > 0$ is large and $P(y)$ is the following functional that penalizes constraint-violation of y:

$$P(y) := \int_Q \max(0, y(t,x) - y_{\mathrm{b}}(t,x))^2 \mathrm{d}x\mathrm{d}t \quad \text{for Assumption 2.1 or}$$

$$P(y) := \int_\Omega \max\left(0, \int_I y(t,x)\mathrm{d}t - y_{\mathrm{b}}(x)\right)^2 \mathrm{d}x \quad \text{for Assumption 2.13.}$$

For simplicity, we have restricted ourselves to unilateral bounds from above and formulated the regularized problem without discretization; the discretization of ($\mathrm{P}^{\mathrm{st}}_\beta$) is addressed below. Basically, ($\mathrm{P}^{\mathrm{st}}_\beta$) is a straightforward adaptation of Problem (P_γ) from [**147**, p1137], which was formulated for linear state equations, to our problem setting. Problem ($\mathrm{P}^{\mathrm{st}}_\beta$) itself is solved by a semismooth Newton method (SSN), cf., e.g., [**274, 275**], with Armijo linesearch in order to decrease the projection formula residual sufficiently in each Newton iteration. If the latter fails, we perform a gradient descent step with Armijo linesearch instead. Since ($\mathrm{P}^{\mathrm{st}}_\beta$) with large $\beta = \beta_{\mathrm{final}}$ cannot be solved without a good initial guess we follow the approach described in [**147**, Section 4]: first, we solve ($\mathrm{P}^{\mathrm{st}}_\beta$) with $\beta = 0$ as described above and obtain its approximate solution $\bar{u}_0$. After that, we solve for $k = 1, \ldots, N$ the problem ($\mathrm{P}^{\mathrm{st}}_\beta$) with $\beta = \beta_{\mathrm{final}}^{\frac{k-1}{N-1}}$; as initial guess in the semismooth Newton method at each $k = 1, \ldots, N$ we utilize the previously obtained approximate solution $\bar{u}_{k-1}$. We stop the semismooth Newton method after at most 15 iterations or if the residual is smaller than 10^{-3} for $k = 0$ and $k = N$ or after at most 8 iterations or if the residual is smaller than β^{-1} for $k = 1, \ldots, N-1$.

Let us briefly comment on how the implementation could be refined utilizing techniques that have successfully been applied in the literature to different, mostly linear elliptic, model problems. For a more elaborate coupling of inner and outer iterations we refer the reader to, e.g., [**146**]. As alternatives to Moreau-Yosida regularization we mention, e.g., barrier methods [**247**], or Lavrentiev [**220**] and virtual control regularization [**190**]. For a proof of convergence of Moreau-Yosida and Lavrentiev regularization in the case of semilinear parabolic equations we refer the reader to [**223**]. Moreover, to enhance efficiency, the usage of appropriate preconditioners [**226**] and reasonable coupling of regularization and discretization [**144, 158, 154**] could be applied.

2.5.2. Example problem and results. We consider the following specification of (P^{st}) in two space dimensions: let $\Omega = [0,1]^2 \setminus ([\frac{1}{3},1] \times [\frac{2}{3},1] \cup [\frac{2}{3},1] \times [\frac{1}{2},1])$, $\Gamma_D = \emptyset$, $T = 1$, $\gamma = 2 \cdot 10^{-2}$, $y_0 \equiv 0$, $\mu \equiv \mathrm{id}_2$,

$$y_d(t,x) := \max(0, \min(1, 3t-1)), \qquad \text{and} \quad \xi(y) := \frac{3}{4} + \frac{1}{2(1+\exp(-10(y-\frac{1}{2})))}.$$

Note that Assumptions 1.5 and 1.6 are satisfied; cf. Example 1.7.2. In Assumption 1.8 we choose $L^s(\Lambda) = L^s(I)$ and $Bu := ub$ with $b \in W^{-1,p}_{\Gamma_D}$ given by $b(\varphi) := \int_{\partial\Omega} \varphi \mathrm{d}s$. For simplicity, we omit control-constraints. We consider the following six different situations:

- no state-constraints,
- pointwise in space and time state-constraints with upper bound $y_{\mathrm{b}} \equiv 1$,
- averaged in time state-constraints of type

$$\int_{I_{\mathrm{obs}}} y(t,\cdot)\mathrm{d}t \le y_{\mathrm{b}} \quad \text{on } \Omega,$$

with

$$\begin{aligned} I_{\mathrm{obs}} &= \left[\frac{2}{3}, 1\right], & y_{\mathrm{b}} &\equiv \frac{1}{3}, \\ I_{\mathrm{obs}} &= \left[\frac{1}{2}, 1\right], & y_{\mathrm{b}} &\equiv \frac{11}{24}, \\ I_{\mathrm{obs}} &= \left[\frac{1}{2}, \frac{5}{6}\right], & y_{\mathrm{b}} &\equiv \frac{7}{24}, \\ \text{or} \quad I_{\mathrm{obs}} &= [0,1], & y_{\mathrm{b}} &\equiv \frac{1}{2}. \end{aligned}$$

Here, the respective upper bounds for the averaged-type constraints are chosen to be the integrals of y_d over I_{obs}.

The numerical experiments are implemented in python utilizing FEniCS and mshr [**9, 203**] for finite element discretization. Each problem ($\mathrm{P}^{\mathrm{st}}_\beta$) is discretized as follows: the state and the adjoint equation are discretized by continuous, piecewise linear finite elements in space (1076 DoF, maximum cell diameter $h_{\max} \approx 4.24 \cdot 10^{-2}$) and by piecewise constant, discontinuous ansatz functions in time which result in an implicit Euler time stepping (555 equidistant time steps). The purely time-dependent controls are discretized by piecewise constant in time functions. Let us note that this coincides with the variational discretization concept [**149**]. For the solution of the nonlinear equations appearing in each time step of the solution of the state equation we use the built-in nonlinear solver of FEniCS.

As pointed out in the introduction chapter, we are not aware of literature dealing with discretization error estimates for the control problem (P^{st}). The estimates for the state equation from [**46**, Theorem 3.11 and Corollary 3.14] made us choose the above number of time steps in order to roughly balance space and time discretization errors for the state equation. Regarding discretization error estimates for PDE-constrained optimization problems with state-constraints, we mention the survey [**157**], and exemplarily literature on linear elliptic [**93, 188, 63**], semilinear elliptic [**42, 58, 155, 222**], parabolic [**94, 211, 115, 83**], and semilinear parabolic problems [**204**].

To solve (P^{st}) approximately, we apply the pathfollowing scheme described in Section 2.5.1 with $N = 6$ and $\beta_{\mathrm{final}} = 10^5$ emplyoing the above explained discretization of ($\mathrm{P}^{\mathrm{st}}_{\beta}$). Figures 2.1 and 2.2 provide an overview over the convergence behaviour of the outer and the inner iterations; hereby, we only show the outer iterations $k = 1, ..., 6$. The residuals during the inner semismooth Newton iterations are displayed in Figure 2.1. Since penalization-based regularization approaches produce infeasible iterates, in general, i.e. the iterates violate the state-constraints, we show the respective constraint-violations of the outer iterates in Figure 2.2. The optimal controls for all six cases defined above are displayed in Figures 2.3 to 2.5 together with the range of values of the associated optimal states. Some differences between the pointwise in space and time and the averaged in time and pointwise in space state-constraints are clearly visible. This is due to the fact that averaging w.r.t. time allows to compensate exceeding a certain pointwise bound for a while by staying strictly below this bound for some other time in I_{obs}. Hence, the choice of the observation interval I_{obs} is crucial in the averaged-type case.

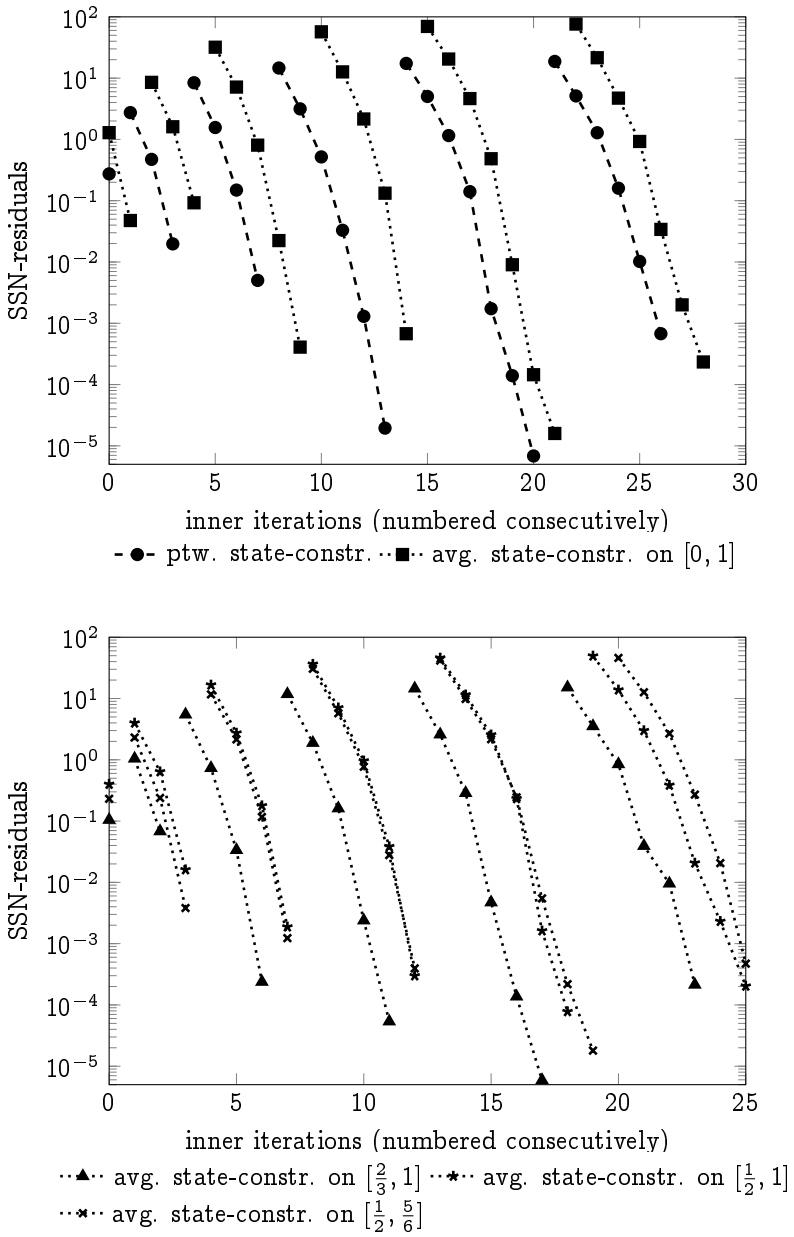


Figure 2.1. L^2-norm of the residuals in the semismooth Newton method (inner iterations, numbered consecutively) during the outer iterations $k = 0, 1, ..., 5$.

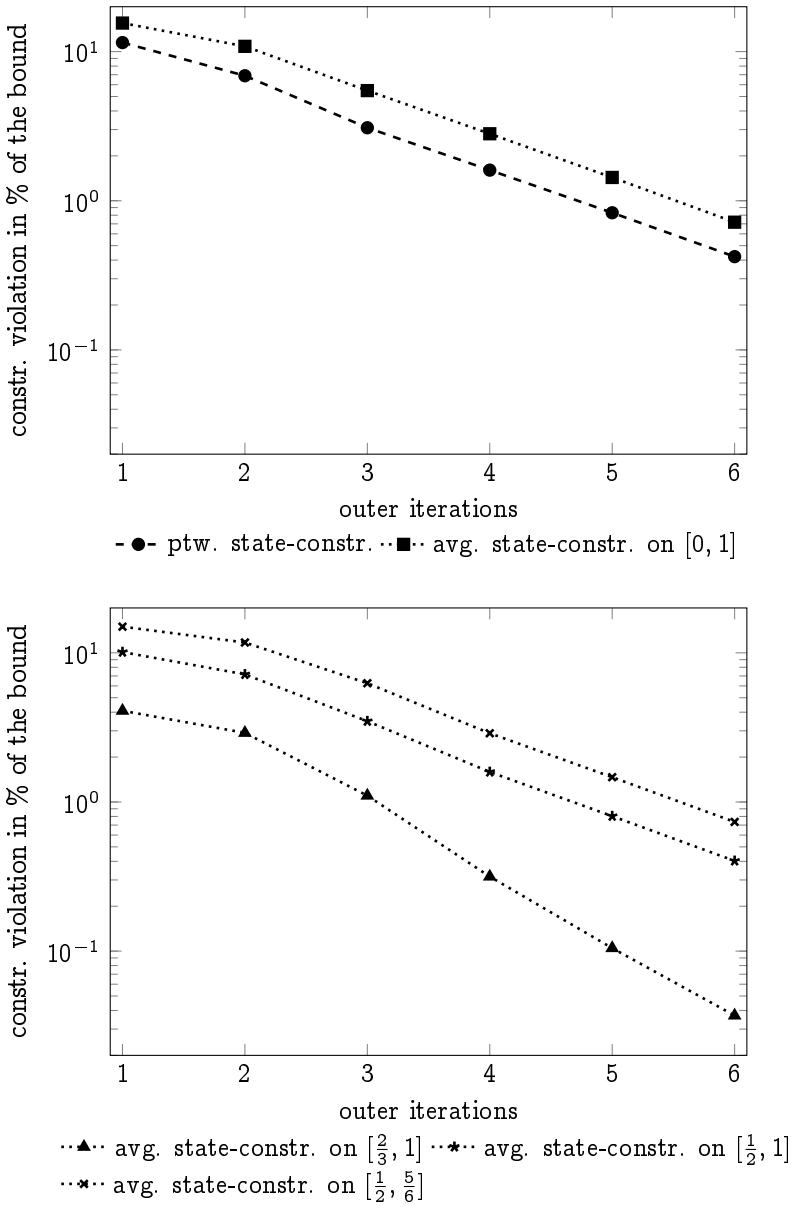
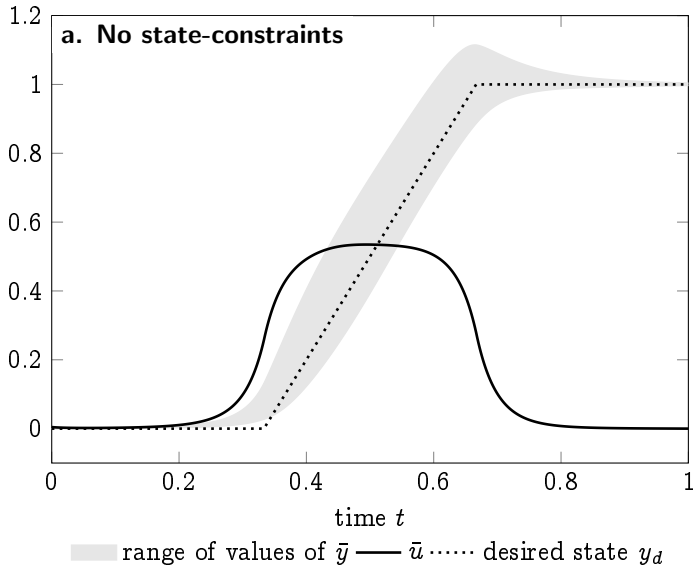


Figure 2.2. Constraint violation of the (outer) control iterates u_k, $k = 0, \ldots, 5$, measured in percent of the respective upper bound.



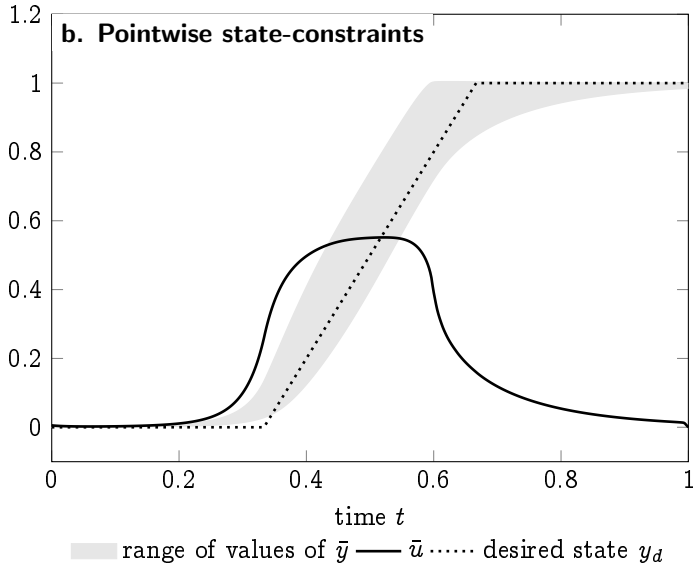
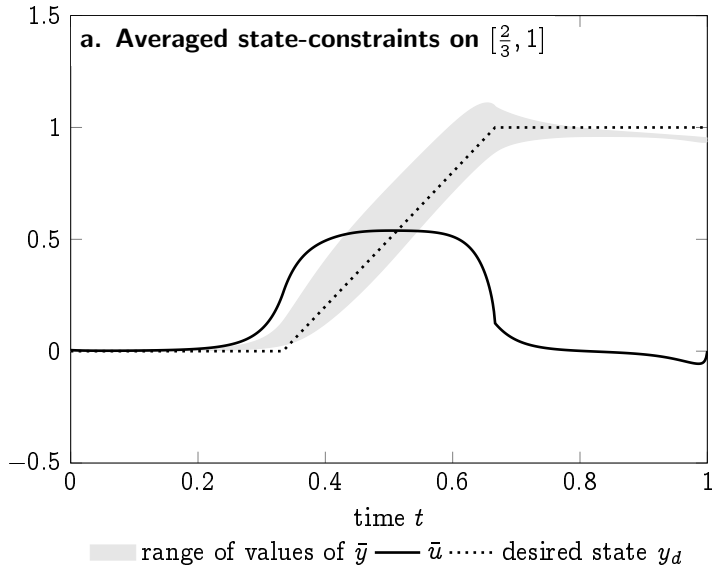


Figure 2.3. a. No state-constraints and **b.** pointwise in space and time state-constraints.



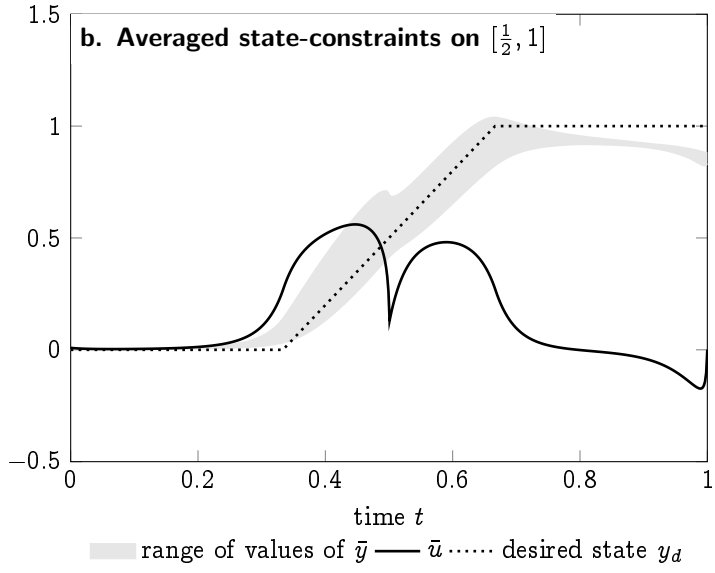
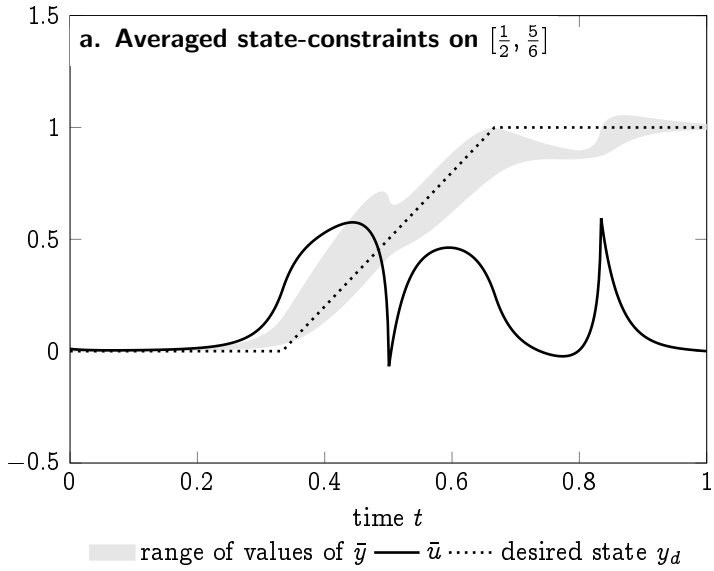


Figure 2.4. Averaged state-constraints on **a.** $I_{\text{obs}} = [\frac{2}{3}, 1]$ and **b.** $I_{\text{obs}} = [\frac{1}{2}, 1]$.



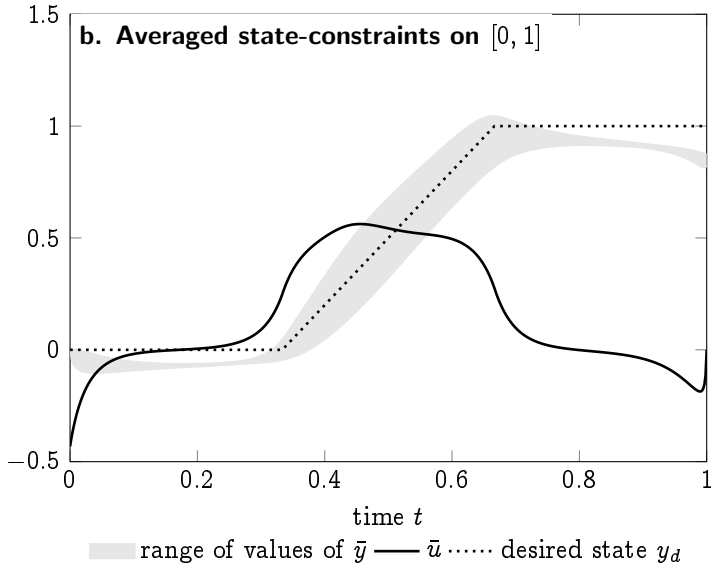


Figure 2.5. Averaged state-constraints on **a.** $I_{\text{obs}} = [\frac{1}{2}, \frac{5}{6}]$ and **b.** $I_{\text{obs}} = [0, 1]$.

Chapter 3

Sparse purely time-dependent optimal control

After having addressed additional constraints on the state-variable in Chapter 2, we now turn to considering (P) again with pure control-constraints, but in return with a modified functional. More precisely, this chapter, which is based on the preprint [**169**] by I. Neitzel and the author, is devoted to sparse optimal control. Hereby, sparsity is enforced by adding certain penalization terms to the objective functional. We focus on the purely time-dependent control setting, i.e. controls depending on time only, but not on space, cf. [**91**] or Example 1.9.3, a particularly interesting setting that will be motivated in detail below.

We obtain results that we expect from known results for the linear and semilinear case; see, e.g., [**138, 54**]. More precisely, we derive first-order necessary optimality conditions and associated sparsity patterns as well as second-order necessary and sufficient optimality conditions for problems of the following type:

$$(\mathrm{P}_k^{\mathrm{sp}}) \qquad \begin{cases} \min\limits_{y,u} J(y,u) := \dfrac{1}{2}\|y - y_d\|^2_{L^2(I\times\Omega)} + \dfrac{\alpha}{2}\displaystyle\sum_{i=1}^m \|u_i\|^2_{L^2(I)} + \beta j_k(u), \\ \qquad \text{s.t.} \qquad u \in U_{\mathrm{ad}} \quad \text{and (Eq)}. \end{cases}$$

Herein, the assumptions on the state equation, the set of admissible controls, the first two summands of J, as well as the $L^2(I,\mathbb{R}^m)$-Tikhonov parameter $\alpha > 0$ are the same as introduced in Chapter 1. The new parameter $\beta > 0$ weighs the sparsity-enforcing penalization/cost term $j_k\colon L^2(I,\mathbb{R}^m) \to \mathbb{R}$. For $k \in \{1,\dots,7\}$, the latter is given by one of the following functionals that are adaptations of the classical (directional) sparsity-enforcing penalizers [**138**] to the purely time-dependent setting:

$$j_1(u) := \sum_{i=1}^m \|u_i\|_{L^1(I)},$$

$$j_2(u) := \sum_{i=1}^m \|u_i\|_{L^2(I)}, \qquad j_3(u) := \int_I \left(\sum_{i=1}^m |u_i(t)|^2\right)^{\frac{1}{2}} \mathrm{d}t,$$

$$j_4(u) := \left(\sum_{i=1}^m \|u_i\|^2_{L^1(I)}\right)^{\frac{1}{2}}, \quad j_5(u) := \left(\int_I \left(\sum_{i=1}^m |u_i(t)|\right)^2 \mathrm{d}t\right)^{\frac{1}{2}},$$

$$j_6(u) := \frac{1}{2}\sum_{i=1}^m \|u_i\|^2_{L^1(I)}, \qquad j_7(u) := \frac{1}{2}\int_I \left(\sum_{i=1}^m |u_i(t)|\right)^2 \mathrm{d}t.$$

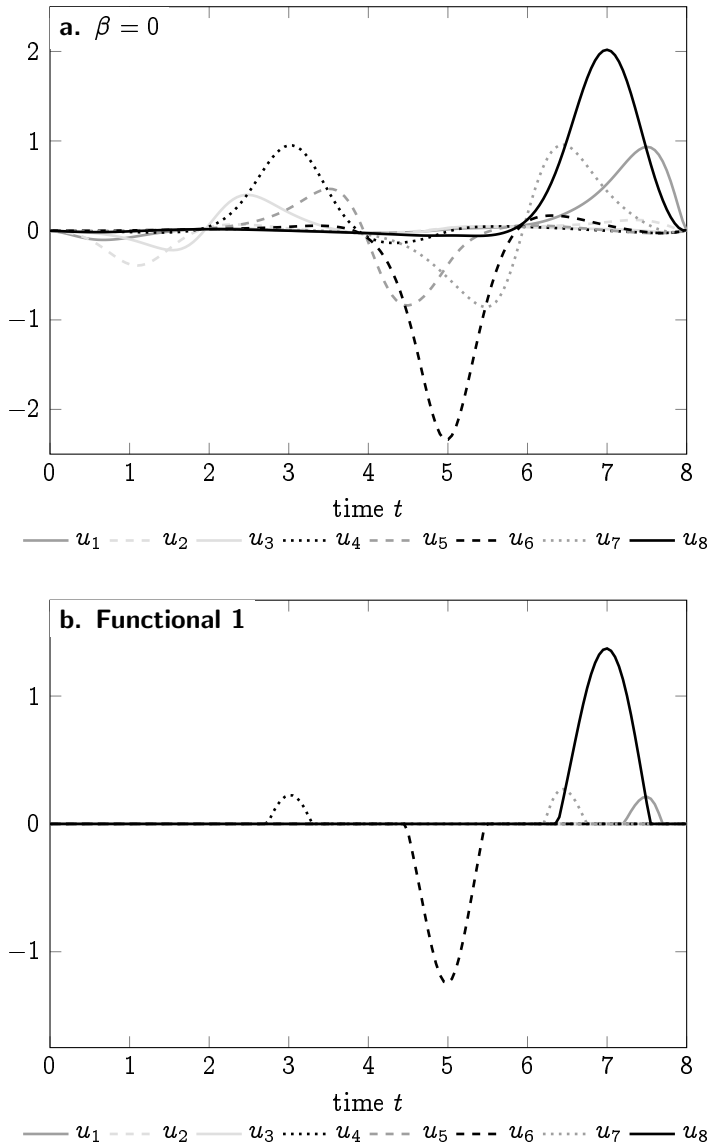


Figure 3.1. Optimal controls **a.** without sparsity and **b.** for functional j_1.

We start with a numerical illustration from [**169**] in order to give the reader an impression why problems involving the cost terms j_1-j_7 are of interest. For details on the underlying model problem, its discretization, and its numerical solution we refer to Section 3.4.2 at the end of this chapter. Let us just say that this problem may be imagined as a very abstract counterpart of the following setting: in a room (which corresponds to the domain Ω) there are eight air conditioning devices with a fixed position (that correspond to fixed actuators $b_i \in L^s$ in Example 1.9.3).

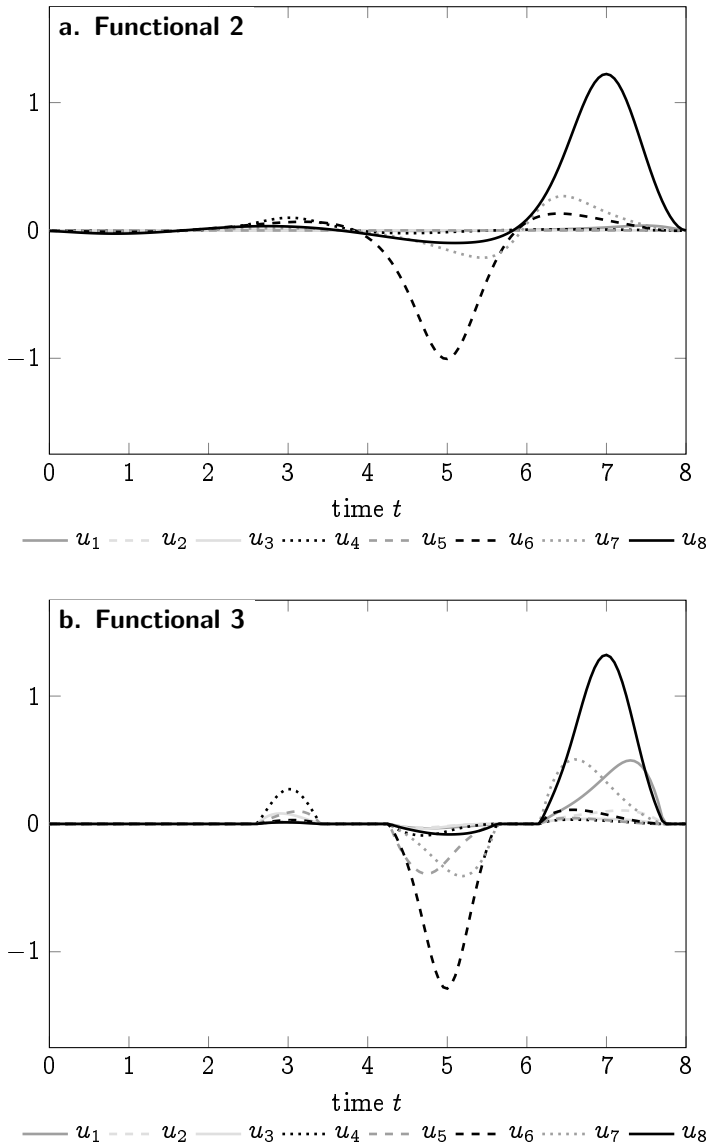


Figure 3.2. Optimal controls **a.** for functional j_2 and **b.** for functional j_3.

The purely time-dependent controls u_i, $i = 1, ..., 8$, are the intensities of the i-th device, i.e. $u_i(t) > 0$ or $u_i(t) < 0$ means heating or cooling by the i-th device at time t with the respective intensity given by $u_i(t)$, while $u_i(t) = 0$ simply means that the i-th device is turned off at time t. In the respective control problem we aim at making the temperature in the room follow a given desired trajectory over time by regulating the intensity of the air conditioning devices apropriately.

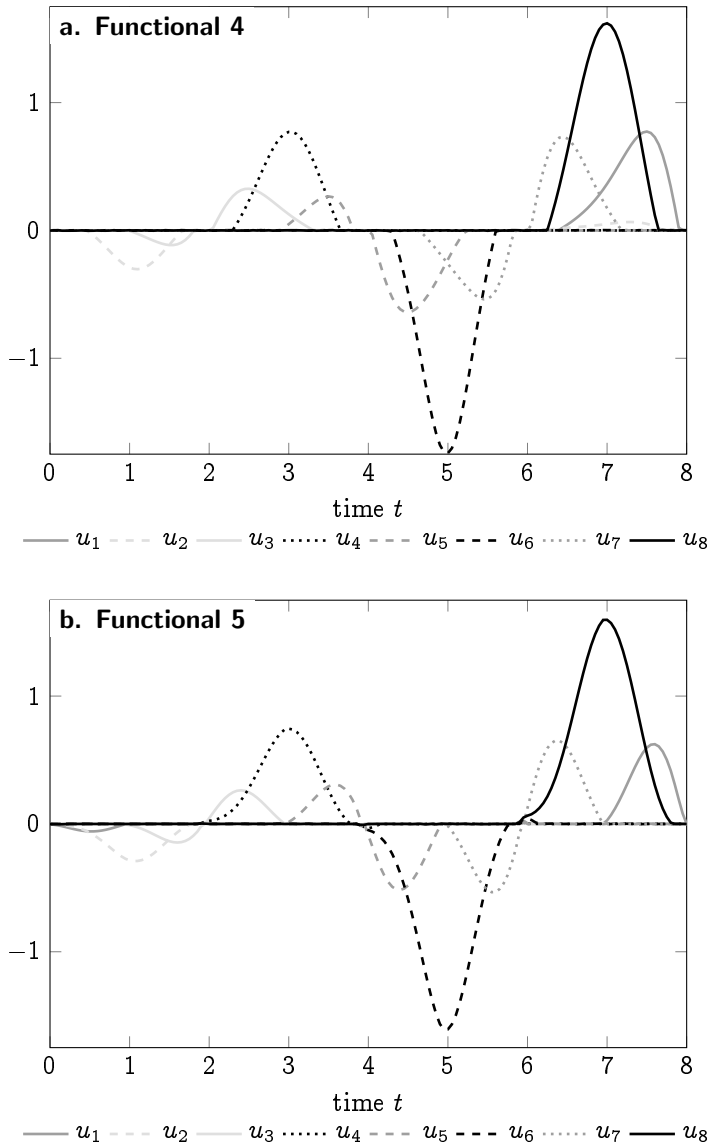


Figure 3.3. Optimal controls **a.** for functional j_4 and **b.** for functional j_5.

Figure 3.1.a shows how the solution of such a problem may look like for $\beta = 0$, i.e. in the classical setting without sparsity considered in Chapter 1: it can be seen that all devices are turned on at every time point, although some devices, e.g., device no. 2 or 3, seem to be used with much less intensity than others. Moreover, at some time points, e.g. at the beginning of the time interval, the intensity of all devices is rather low. From an economic point of view one may therefore be

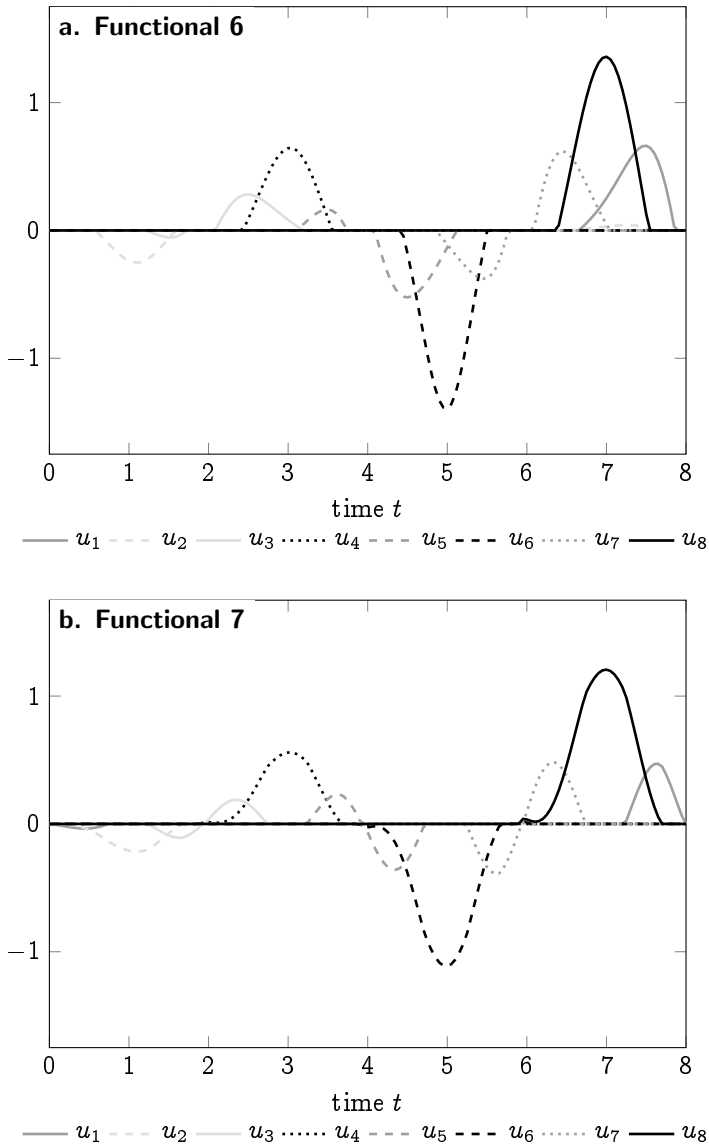


Figure 3.4. Optimal controls **a.** for functional j_6 and **b.** for functional j_7.

tempted to avoid the usage of air conditioning at the beginning of the time interval and/or to dispense with devices no. 2 or 3 at all.

The solution looks completely different when adding one of the functionals j_1-j_7 to the objective functional of the problem: functional j_1 selects five devices that are used to control the temperature of the room while the remaining three devices are not used at all. Even those devices that are used, are only turned on for certain rather small time intervals. In particular, no device is turned on at the beginning of

the time interval and devices no. 2 and 3 are not used at all. Functional j_2 selects a certain subset of the devices, too, but now each of these devices is constantly used for the whole time interval under consideration as can be seen in Figure 3.2.a. In particular, these devices are also turned on at the beginning of the time interval with extremly low intensity. Figure 3.2.b illustrates that functional j_3 allows to determine a small subset of time points at which all devices are turned on, while no air conditioning takes place at all other time points at all. We see, e.g., that all devices stay turned off at the beginning of the time interval. Consequently, the functionals j_1-j_3 allow in some sense to incorporate the abovementioned economic considerations into the optimization problem.

The interpretation of functionals j_4-j_7 within the narrative of optimal airconditioning is less obvious, but could be interesting in different contexts. In the case of functional j_4, cf. Figure 3.3.a, each device is turned on only for a small subset of timepoints. This set of time points is different for each device and seems to be given by those time points where the respective device is most needed. In Figure 3.3.b it can be seen that at each time point exactly two devices are turned on. In this sense, functional j_5 seems to select for each time point only a small number of devices that are turned on simultanuously. The solutions of $(\mathrm{P}_k^{\mathrm{sp}})$ associated with j_6 and j_7 (Figure 3.4) look quite similar as those with j_4 and j_5 from a structural point of view; we will come back to this lateron.

Having thus motivated the consideration of $(\mathrm{P}_k^{\mathrm{sp}})$, let us now give a short overview on literature concerned with sparse optimal control. Starting with the pioneering work of Stadler [**261**] on sparse optimal control of linear elliptic equations, there have been many contributions on this topic in the recent past. For a broader overview we refer the reader, e.g., to the survey article [**44**], and focus on literature related to the present work in the following. Regarding literature following the original idea of Stadler to enforce sparsity by adding an L^1-penalization term to the objective functional we mention, e.g., [**289, 53, 52, 64, 72, 65, 74, 260, 59, 286**]. These publications refer to different types of PDEs and cover several aspects, including first- and second-order optimality conditions, discretization error estimates, and additional state-constraints. When considering parabolic PDEs it might be favorable to obtain a space-time sparsity profile of the optimal control in which space- and time-variable are treated in a different way. This leads to the concept of directional sparsity introduced in [**138**] for linear PDEs. An extension towards the setting of polar coordinates has been discussed in [**137**] in the case of linear elliptic problems, and finite element discretization error estimates have been derived for linear parabolic [**62**] and semilinear parabolic [**61**] problems. We mention in particular that first- and second-order optimality conditions for directionally sparse optimal control of semilinear parabolic PDEs have been obtained in [**54, 61**], for instance. The specific difficulty herein arises from the fact that sparsity-enforcing penalizers are convex, but nonsmooth, whereas the remaining part of the objective functional is smooth, but —due to nonlinearity of the state equation— nonconvex. Moreover, as typical for optimal control of nonlinear PDEs, the so-called two-norm discrepancy [**174, 71**] occurs: differentiability and coercivity of the second derivative of the smooth part of the objective functional hold only w.r.t. different norms. Let us already note that the second-order sufficient conditions from [**54**, Theorems 4.3 and 4.8] and [**61**, Theorem 4.2] nevertheless avoid introducing a two-norm gap.

As alternative approaches to enforce sparsity in PDE-constrained opimization we finally mention, e.g., control in measure spaces [**86, 47, 229, 55, 56, 76**], directional sparsity with measure spaces [**48, 192**], or L^0-penalization [**178, 77**].

As far as we know, the first contribution addressing sparse optimal control of quasilinear PDEs is our preprint [**169**]. Consequently, this chapter, which is based on the results of [**169**], contributes both to the fields of optimal control of quasilinear PDEs and sparse optimal control. We extend the first- and second-order analysis for sparse optimal control of semilinear parabolic PDEs from [**54, 61**] to problems with a quasilinear parabolic state equation. Here, we focus on the purely time-dependent control setting, which can be motivated as follows: first, such a setting may be advantageous in applications as one may imagine with a view to the abovementioned artificial setting. Among the real-world examples given in the introduction of [**91**] we mention, e.g., optimal cooling of steel profiles by controlling the intensities of the finite number of nozzles that spray water on the profile. Second, while space-time sparsity patterns for distributed optimal control problems have already been under detailed consideration in [**54**], the purely time-dependent control setting has not been addressed systematically in the context of directional sparsity before [**169**]. Nevertheless, our results would also apply to the classical directionally sparse distributed control setting as in, e.g., [**54**]. Finally, the chosen purely time-dependent control setup allows to include control by fixed Neumann boundary sources up to dimension 3, whereas distributed Neumann boundary control is only possible up to dimension 2; cf. Example 1.9 or [**35**, Example 2.4]. Let us also mention that the discretization of control problems with purely time-dependent controls is usually slightly easier to handle than the discretization of problems with distributed controls; cf. Section 3.4.2.

With our work, we combine two challenging aspects, namely sparsity-enforcing penalization and a quasilinear state equation. In the presence of L^2-Tikhonov regularization we are able to carry out a full first- and second-order analysis, the latter one avoiding the two-norm gap. Let us emphasize that with respect to the state equation we again rely on the rough regularity setting of [**216, 35**] as introduced in Section 1.2. Similarly as done in Chapter 2 for the analysis of the state-constrained problem (P^{st}), we pursue an abstract approach in the flavour of [**71**] and work out the abstract core of existing arguments for second-order conditions from [**53, 54, 61**], which may also facilitate the application to other problems. Due to the different nature of our nonlinearity, the second-order analysis from [**54, 59**] cannot be carried over to our setting in the bang-bang case $\alpha = 0$, i.e. the case without L^2-Tikhonov regularization; cf. Section 3.3.3 below. This illustrates that the transfer of techniques from semilinear to quasilinear problems is by no means trivial.

In this chapter we provide an extensive analysis of directional sparsity for purely time-dependent controls. The interesting effects of the different j_k on the solution of ($\mathrm{P}_k^{\mathrm{sp}}$) has already been illustrated numerically above. Throughout this chapter we carry out a detailed analysis of the respective control problems. In particular, the structural properties (often called "sparsity patterns") of the optimal controls observed in the numerical examples turn out to be a direct consequence of the respective first-order necessary optimality conditions for ($\mathrm{P}_k^{\mathrm{sp}}$). Besides the functionals j_1-j_5, whose structure corresponds to those already discussed in [**54**],

we also propose and analyze the functionals j_6 and j_7, that have —to the best of our knowledge— not been dealt with in the context of PDE-constrained optimization before [**169**]. These two functionals are interesting, because their sparsity patterns are similar to those of j_4 and j_5, respectively, while they are advantegeous compared to j_4 and j_5 from a numerical point of view because their proximity operator is computable.

In order to reduce redundancy, we formulate as many of the results and arguments as possible on an abstract, general level from which the concrete results can be obtained afterwards. This is analogous to Chapter 2 where our abstract theorem on SSCs with a state-constraint-like constraint, cf. Theorem 2.8, allowed to prove SSCs both for averaged in time and pointwise in space and time state-constraints on the concrete level; cf. Theorems 2.18 and 2.29. Consequently, we state our main results concerning $(\mathrm{P}_k^{\mathrm{sp}})$ right at the beginning in Section 3.1 and postpone their proofs to the remaining part of the chapter. For convenience of the reader, let us already point out some key features of our arguments. In two important steps of our arguments we introduce abstract settings that allow to handle seemingly different situations with a common argument.

The first abstraction takes place in Section 3.2.1: we prove first- and second-order optimality conditions for an abstract optimization problem in Banach spaces whose functional is given by the sum of a smooth and a nonsmooth functional, both satisfying certain assumptions. For the rest of this chapter, we show that these abstract results apply to the concrete instances of $(\mathrm{P}_k^{\mathrm{sp}})$, $k = 1, \dots, 7$. The main work is to check that both smooth and nonsmooth part of the (reduced) functional of $(\mathrm{P}_k^{\mathrm{sp}})$ satisfy the required assumptions. For the smooth part, given by the first two summands of J in $(\mathrm{P}_k^{\mathrm{sp}})$, this has already been shown in [**35**]; see Proposition 1.21. The assumptions on the nonsmooth part, i.e. for $u \mapsto \beta j_k(u)$, are verified in Section 3.2.2. To avoid checking the conditions for all seven cases of j_k separately, we apply the second step of abstraction: it turns out that the seven functionals can be reduced to four generic cases for which we then prove the required results utilizing techniques known from the literature. Hereby, it is important to note that the conditions on the smooth and the nonsmooth part of the functional can be checked independently of each other. Consequently, the presence of a quasilinear parabolic state equation does not cause additional problems when applying results on the nonsmooth part of the functional that have been derived earlier in the context of optimal control of different underlying equations. In particular, our results on the abstract level may also serve as kind of a template for analogous concrete first- and second-order results on other problems, e.g., with a different underlying state equation.

3.1. First- and second-order optimality conditions

Throughout this chapter we will again rely on the rough regularity setting described by Assumptions 1.5, 1.6 and 1.8 in Chapter 1. We only slightly enforce Assumption 1.8 by restricting ourselves to purely time-dependent controls.

Assumption 3.1. Let s, y_d, y_0, and α be chosen as in Assumption 1.8. In addition, it holds $L^s(\Lambda) = L^s(I, \mathbb{R}^m)$, the control operator is given by

$$B\colon L^s(I,\mathbb{R}^m) \to L^s(I, W_{\Gamma_D}^{-1,p}), \quad u \mapsto \sum_{i=1}^m u_i b_i,$$

where $b_i \in W_{\Gamma_D}^{-1,p}$, $i = 1, \dots, m$, are fixed actuator functions, and the set of admissible controls is given by

$$U_{\mathrm{ad}} := \{u \in L^s(I,\mathbb{R}^m)\colon u_{\mathrm{a},i} \le u_i \le u_{\mathrm{b},i} \text{ a.e. on } I,\ i = 1, \dots, m\}$$

with control bounds $u_{\mathrm{a}}, u_{\mathrm{b}} \in L^\infty(I,\mathbb{R}^m)$, $u_{\mathrm{a},i} \le u_{\mathrm{b},i}$ a.e. on I for $i = 1, \dots, m$. Moreover, we fix the cost parameter $\beta > 0$.

Let us start by mentioning the main results of our analysis of $(\mathrm{P}_k^{\mathrm{sp}})$ from [**169**]. The proofs of the first- and second-order results are quite technical, rely on auxiliary material from Section 3.2, and are therefore postponed to Section 3.3. As in Section 1.5 our presentation follows the standard procedure of the analysis of an optimal control problem. Of course, we have to prove well-posedness of $(\mathrm{P}_k^{\mathrm{sp}})$ first.

Proposition 3.2. *Under Assumptions 1.5, 1.6 and 3.1 the optimal control problem* $(\mathrm{P}_k^{\mathrm{sp}})$ *admits at least one globally optimal control.*

Due to convexity and continuity of j_k and boundedness of U_{ad}, existence of an optimal control for $(\mathrm{P}_k^{\mathrm{sp}})$ is guaranteed by [**216**, Proposition 6.4]. Since there is no difference to Proposition 1.18 we do not give more details.

Regarding first-order necessary optimality conditions for an $L^2(I,\mathbb{R}^m)$-local solution to $(\mathrm{P}_k^{\mathrm{sp}})$ we will obtain the following result that also characterizes the different (directional) sparsity patterns resulting from the different functionals j_k, $k = 1, \dots, 7$.

Theorem 3.3 ([**169**], Theorem 1.4)**.** *Let Assumptions 1.5, 1.6, 1.8 and 3.1 hold and let $\bar u$ be a local solution to $(\mathrm{P}_k^{\mathrm{sp}})$ w.r.t. the $L^2(I,\mathbb{R}^m)$-topology. Then, there exists a unique, so-called adjoint state $\bar p \in \mathbb{W}^{1,r}(I, (L^p, \mathrm{Dom}_{L^p}(-\nabla\cdot\mu\nabla)))$, $r \in (1,\infty)$, fulfilling* (1.14) *and a unique $\bar\lambda \in \partial j_k(\bar u)$ (see formulas* (3.20)-(3.26)*), such that the variational inequality*

$$\int_I (B^*\bar p + \alpha\bar u + \beta\bar\lambda)^T(u - \bar u)\mathrm{d}t \ge 0 \qquad \forall u \in U_{\mathrm{ad}} \tag{3.1}$$

is satisfied. In particular, in the respective cases $k = 1, \dots, 7$ the optimal control $\bar u$ exhibits the following sparsity patterns, that will be described on more detail below:

$$\bar u_i(t) = 0 \qquad \textit{if and only if}$$

$k = 1$: $|(B^*\bar p)_i(t)| \le \beta$,

$k = 2$: $\|(B^*\bar p)_i\|_{L^2(I)} \le \beta$,

$k = 3$: $|B^*\bar p(t)|_2 \le \beta$,

$k = 4$: $|(B^*\bar p)_i(t)| \le \beta\gamma_i$ *with* $\gamma_i = \frac{\|\bar u_i\|_{L^1(I)}}{|(\|\bar u_i\|_{L^1(I)})_i|_2}$ *if* $\bar u \ne 0$ *and* $\gamma_i = 1$ *otherwise,*

$k = 5$: $|(B^*\bar p)_i(t)| \le \beta\gamma(t)$ *with* $\gamma(t) = \frac{|\bar u(t)|_1}{\||u(\cdot)|_1\|_{L^2(I)}}$ *if* $\bar u \ne 0$ *and* $\gamma(t) = 1$ *otherwise,*

$k=6$: $|(B^*\bar{p})_i(t)| \leq \beta\gamma_i$ *with* $\gamma_i = \|\bar{u}_i\|_{L^1(I)}$ *if* $\bar{u} \neq 0$ *and* $\gamma_i = 1$ *otherwise,*
$k=7$: $|(B^*\bar{p})_i(t)| \leq \beta\gamma(t)$ *with* $\gamma(t) = |\bar{u}(t)|_1$ *if* $\bar{u} \neq 0$ *and* $\gamma(t) = 1$ *otherwise.*

In essence, this result consists of two parts: the optimality condition (3.1) characterizing an optimal control $\bar{u}$ and conclusions on the structure of $\bar{u}$ that can be drawn from (3.1) and the structure of $\bar{\lambda}$ in the different cases $k = 1, ..., 7$.

Let us note that the first part of Theorem 3.3 roughly resembles other first-order results in this thesis; see, e.g., Theorem 1.19 for the control-constrained problem (P) or Theorems 2.4 and 2.17 for (P^{st}) with state-constraints. In particular, the gradient of the first two summands of J, i.e. the smooth part of J, is given by $B^*\bar{p} + \alpha\bar{u}$, as before. The only difference compared to Theorem 1.19 originates from the newly introduced, nonsmooth term βj_k. Since the latter is not differentiable, but convex, its subgradient $\beta\bar{\lambda}$ instead of a gradient appears in (3.1).

Next, some detailed comments on the sparsity patterns seem to be in order. In an application one usually has to determine a suitable β experimentally in order to ensure that the support of the corresponding solution of ($\mathrm{P}_k^{\mathrm{sp}}$) has roughly the desired size. Heuristically, choosing larger β decreases the support of the associated $\bar{u}$ and for sufficiently large β it may hold $\bar{u} \equiv 0$. However, since all the quantities $\bar{u}, \bar{y}, \bar{p}, \bar{\lambda}, \gamma$ in the above theorem are coupled, the actual size of the support of $\bar{u}$, i.e. the concrete amount of sparsity of $\bar{u}$, depending on the size of β is difficult to predict a priori. Nevertheless, on a qualitative instead of a quantitative level we can describe the different sparsity patterns as follows. Hereby, we refer to the b_i from Assumption 3.1 as "actuators" (="device" in the numerical example provided in the introduction) and say that an actuator b_i is "active" (="turned on" in the introduction) at time t if $u_i(t) \neq 0$.

- j_1 — **"Sparsity":** This approach ensures sparsity of both the number of actuators b_i and the time intervals at which they are active. However, there is no further structure in this sparsity.
- j_2 — **"Sparse time-global selection of actuators":** This approach selects a subset of the actuators that are allowed to be active. All other actuators are not used. The activity intervals of those actuators used are not sparse, in general.
- j_3 — **"Sparsity in time of any control action":** Any actuator, and then possibly all actuators, can become active only on a subset of I that is sparse.
- j_4 — **"Sparse activity-time for each actuator":** An actuator i becomes active at some time point t only if its contribution at time point t is above a threshold depending on i. Therefore, the time of activity of each actuator i is sparse with a sparsity pattern depending on i.
- j_5 — **"Sparse selection of actuators at each time":** At each time point t, an actuator i can be active only if its contribution is above a threshold depending on t. Therefore, at each time point t a certain sparse subset of actuators is selected to become active at t.
- j_6 **and** j_7**:** These functionals result in similar sparsity patterns as j_4 and j_5, but with different thresholds that weight the components differently. More precisely, in the cases $k = 4$ and $k = 6$ an actuator b_i only becomes active at time point t if and only if $|(B^*\bar{p})_i(t)| > \beta\gamma_i$ holds. The thresholds γ_i,

however, are different; it holds:

$$\gamma_i = \frac{\|\bar{u}_i\|_{L^1(I)}}{|(\|\bar{u}_i\|_{L^1(I)})_i|_2} \quad \text{for } k=4, \qquad \text{and } \gamma_i = \|\bar{u}_i\|_{L^1(I)} \quad \text{for } k=6.$$

Similarly, in the cases $k=5$ and $k=7$ an actuator b_i becomes active at time point t if and only if $|(B^*\bar{p})_i(t)| > \beta\gamma(t)$ with

$$\gamma(t) = \frac{|\bar{u}(t)|_1}{\||u(\cdot)|_1\|_{L^2(I)}} \quad \text{for } k=5, \qquad \text{and } \gamma(t) = |\bar{u}(t)|_1 \quad \text{for } k=7.$$

Note that γ_i in the case $k=4$ may be interpreted as the fraction of the cost term $|(\|\bar{u}_i\|_{L^1(I)})_i|_2$ that is due to the i-th actuator. Similarly, in the case $k=5$ we can imagine $\gamma(t)$ as the fraction of the overal cost $\||u(\cdot)|_1\|_{L^2(I)}$ that is consumed at time point t. In this sense, the thresholds for $k=4,5$ are more intuitive than those for $k=6,7$.

Each of these possibilities may be of interest in certain applications. A numerical illustration together with some rough ideas why the respective sparsity patterns could be interesting in practice has been given in Figures 3.1 to 3.4 in the introduction of this chapter; for the details of the numerical implementation of this example we refer to Section 3.4.2. At this point, let us just emphasize that functionals j_6 and j_7 have an advantage compared to j_4 and j_5 from the perspective of fast numerical implementation, while the latter are superior in terms of interpretability. The respective details will be explained in Section 3.4.

We note that analogous sparsity patterns are also obtained for $\alpha = 0$; cf. the results of Section 3.2.2. Until this point of our analysis we could indeed allow both for $\alpha > 0$ or $\alpha = 0$, to which we refer as the regular or the bang-bang case, respectively.

For the formulation of second-order conditions, however, we have to restrict the analysis to the regular case, as will be explained in Section 3.3.3. In the following we slightly change the notation compared to the previous chapters. For the rest of this chapter, we denote by f the smooth part of the functional of $(\mathrm{P}_k^{\mathrm{sp}})$ (which equals the reduced functional j from Chapters 1 and 2) and by $\hat{J} = f + \beta j_k$ the reduced functional of $(\mathrm{P}_k^{\mathrm{sp}})$; a detailed definition will be given at the beginning of Section 3.3 below.

The following theorem is our second main result of this chapter.

Theorem 3.4 ([**169**], Theorem 1.5). *Let Assumptions 1.5, 1.6, 1.8 and 3.1 hold. If $\bar{u} \in U_{\mathrm{ad}}$ is a local solution to $(\mathrm{P}_k^{\mathrm{sp}})$ such that the reduced functional $\hat{J}$ fulfills*

$$\hat{J}(u) \geq \hat{J}(\bar{u}) + \frac{c}{2}\|u-\bar{u}\|^2_{L^2(I,\mathbb{R}^m)} \qquad \forall u \in U_{\mathrm{ad}} \ s.t. \ \|u-\bar{u}\|_{L^2(I,\mathbb{R}^m)} < \epsilon \tag{3.2}$$

with some $c \geq 0$ and $\epsilon > 0$, it holds:

$$f''(\bar{u})v^2 + \beta j_k''(\bar{u}, v^2) \geq c\|v\|^2_{L^2(I,\mathbb{R}^m)} \qquad \forall v \in C_{\bar{u}} \tag{3.3}$$

with

$$C_{\bar{u}} = \big\{ v \in L^2(I,\mathbb{R}^m): \ v \geq 0, \ \text{if } \bar{u} = u_{\mathrm{a}}, \quad v \leq 0, \ \text{if } \bar{u} = u_{\mathrm{b}}, \\ f'(\bar{u})v + \beta j_k'(\bar{u}, v) = 0 \big\}$$

and f', f'', j_k', and j_k'' given by (1.6), (1.7), (3.27)-(3.33), *and* (3.34)-(3.40), *respectively. Conversely, let $\bar{u} \in U_{\mathrm{ad}}$ satisfy the first-order necessary optimality condition* (3.1) *and*

$$f''(\bar{u})v^2 + \beta j_k''(\bar{u}, v^2) > 0 \qquad \forall v \in C_{\bar{u}} \setminus \{0\}. \tag{3.4}$$

Then, there are $\epsilon, c > 0$ such that the quadratic growth condition (3.2) *holds true. In particular, $\bar{u}$ is a strict local solution to* ($\mathrm{P}_k^{\mathrm{sp}}$) *w.r.t. the $L^2(I, \mathbb{R}^m)$-topology.*

On the surface, the overal structure of Theorem 3.4 is analogous to the corresponding result from [**35**] for the problem (P) without sparsity-enforcing penalization summarized in Theorem 1.22: nonnegativity of the second derivative of the reduced functional on a certain cone of directions is necessary for local optimality and positivity of this second derivative on the same cone is sufficient for strict local optimality. A closer look, however, reveals that unlike in Theorem 1.22 actually there is no second derivative of the reduced functional $\hat{J}$ in Theorem 3.4. In (3.3) and (3.4), only f'', i.e. the second derivative of the smooth part f of $\hat{J}$, is a true Fréchet derivative, while for the nonsmooth part a surrogate for the second derivative, denoted by j_k'', is used. Similarly, also the definition of the critical cone is slightly different from Theorem 1.22: instead of using Fréchet derivatives, i.e. linear forms, we now need to refer to the directional derivative of j_k at $\bar{u}$ that is not linear w.r.t. the direction, in general.

Next, let us mention some similarities of Theorem 3.4 with the second-order results Theorems 1.22, 2.18 and 2.29 encountered in this thesis so far. First, we note that necessary and sufficient optimality conditions in Theorem 3.4 have minimal gap: SNCs and SSCs refer to the same cone of directions. This is analogous to the result for the problem (P) without sparsity; cf. our comments in Section 1.5.3. For the state-constrained problem (P^{st}) we had to leave the question whether our SSCs are of no-gap-type as an open question; see the end of Section 2.2. Second, as also observed in [**54**], the positivity condition (3.4) and the coercivity condition (3.3) are equivalent for $\alpha > 0$. This is analogous to the situation in Theorem 1.22 as well as to the the observation in Proposition 2.9 in the case of state-constraints. For no-gap second-order conditions for bang-bang problems, i.e. the case $\alpha = 0$, with a semilinear elliptic state equation and L^1-penalization we refer the reader to the recent paper [**286**].

As in earlier second-order results in this thesis, we encounter the two-norm discrepancy, but avoid inferring a two-norm gap in the SSCs; a brief explanation hereof has been given in the introduction on behalf of Example 0.2 and in Section 1.5.3. In the situation of Theorem 3.4 differentiability of f and coercivity of $f'' + \beta j_k''$ hold w.r.t. the $L^s(I, \mathbb{R}^m)$- and the $L^2(I, \mathbb{R}^m)$-norm, respectively, i.e. w.r.t. different, nonequivalent norms. Nevertheless, a two-norm gap can be avoided in Theorem 3.4 since the quadratic growth condition (3.2) resulting from (3.4) holds on an $L^2(I, \mathbb{R}^m)$-neighbourhood of $\bar{u}$. This is consistent with corresponding second-order conditions for directionally sparse optimal control of semilinear parabolic PDEs; cf. [**54**, Theorems 4.3 and 4.8] and [**61**, Theorem 4.2].

After having stated and explained our main results, let us briefly point out the structure of the remaining part of this chapter containing their proofs. In Section 3.2.1 we discuss optimality conditions for an abstract problem in the flavour

of [**71**] or Section 2.2 by abstracting the main ideas from [**53, 54, 61**]. A more concrete instance hereof is dealt with in Section 3.2.2. There, we show that the previous results apply to a certain class of optimization problems on Lebesgue spaces with four different sparsity-enforcing penalization terms, and analyze the resulting sparsity patterns of their solutions. Here, we rely again on [**53, 54, 61**]. In Section 3.3 we finally prove our main results by applying the framework of Section 3.2.2 to ($\mathrm{P}_k^{\mathrm{sp}}$). Details on the numerical experiments already presented in the introduction of this chapter are provided in Section 3.4.

3.2. First- and second-order optimality conditions on an abstract level

This section closely follows [**169**, Section 2] and prepares the proofs of our main theorems in Section 3.3. As a first step, we analyze in Section 3.2.1 first- and second-order optimality conditions for a purely abstract optimization problem whose functional is given by the sum of a smooth, but nonconvex, and a convex, but nonsmooth term. More precisely, we extend the abstract framework for smooth functionals from [**71**] in an appropriate way. As explained in Section 1.5, the abstract framework from [**71**] resembles the typical structure of control problems with smooth functional and pure control-constraints. Now, we extend this towards the inclusion of nonsmooth, but convex summands that satisfy certain properties that are typical for sparsity promoting cost terms as, e.g., j_1-j_7. The results are obtained utilizing the techniques of [**53, 54, 61**], and may therefore also be viewed as a summary of these earlier results on an abstract level. This strategy is analogous to Chapter 2 where we extented the work [**71**] in such a way that we could, on the concrete level, derive second-order sufficient conditions in the presence of state-constraints. After proving second-order conditions in this abstract setting, we make the problem under consideration a bit more concrete in Section 3.2.2 and deal with optimization problems on Lebesgue spaces with directional sparsity-enforcing penalization terms. Following [**53, 54, 61**] we verify that these problems fit into the framework of Section 3.2.1 and analyze the corresponding sparsity patterns of the solutions. Hereby, one has to note that the properties of the nonsmooth cost term are independent of the state equation of the respective optimal control problem. This allows to transfer results on sparsity-enforcing penalization terms from the literature concerned with problems with a different state equation.

3.2.1. Optimality conditions for an abstract nonsmooth and nonconvex problem. Let us define the following optimization problem:

$$(\mathrm{AP}^{\mathrm{sp}}) \qquad \min_{u\in K} \hat{J}(u) := f(u) + g(u),$$

where K is a closed, convex set in a Banach space U_∞ and f and g are real-valued functionals such that f is smooth, but possibly nonconvex, and g is convex and Lipschitz but not necessarily smooth. Since we have in mind the concrete situation of $\hat{J}$ being the reduced functional of a sparse PDE-constrained optimal control problem, we include a two-norm discrepancy; cf. Section 1.5.3: coercivity of second derivatives can be expected only w.r.t. a weaker norm in a Hilbert space $U_2 \supset U_\infty$. As in [**71**] and Section 2.2, we usually expect U_2 to be an L^2-space in applications, while U_∞ is an L^s-space with $s \in (1,\infty]$. The precise setting is described in detail below. In particular, our assumptions on the smooth functional

f are identical to those in [**71**] and Assumption 2.7.1, and cover a broad range of functionals arising from PDE-constrained optimization as has been pointed out in Section 1.5.3. Hence, we generalize the result from [**71**] for the smooth case $\hat{J} = f$ to the inclusion of a nonsmooth summand g in the functional.

Our approach differs from the similar abstract approach in [**286**] by several technical aspects. Our setting includes the presence of two nonequivalent norms, but instead of working with a Banach space and its predual as in [**286**] we restrict ourselves to formulating optimality conditions w.r.t. the Hilbert space U_2. Moreover, unlike in [**286**] we do not include the convex constraint "$u \in K$" as indicator function in the nonsmooth part of the functional. We will show in Section 3.2.2 that our assumptions on g are typically fulfilled by penalizers promoting directional sparsity, while the applications discussed in [**286**] are primarily concerned with different, nonuniformly convex integral functionals. The proofs and assumptions of this section are inspired by [**71**] and well-known techniques employed in particular in [**53, 54, 61**].

For convenience of the reader we state the full assumptions on $(\mathrm{AP}^{\mathrm{sp}})$ although the first part is identical to Assumption 2.7.1.

Assumption 3.5. Let U_2 be a Hilbert space and U_∞ a Banach space such that $U_\infty \hookrightarrow U_2$. With $\|\cdot\|_\infty$, $\|\cdot\|_2$, and $\langle\cdot,\cdot\rangle_2$ we denote the corresponding norms and the duality pairing on $U_2^* \times U_2$. Let $\emptyset \neq K \subset U_\infty$ be convex and $A \supset K$ be open in U_∞. We fix $\bar{u} \in K$.

1. The functional $f\colon A \to \mathbb{R}$ is assumed to be twice continuously Fréchet differentiable w.r.t. $\|\cdot\|_\infty$ and to fulfill the following properties:
   - 1a. The derivatives of f taken w.r.t. the space U_∞ extend to continuous linear and bilinear forms on U_2, i.e.
   $$f'(u) \in \mathcal{L}(U_2, \mathbb{R}) \quad \text{and} \quad f''(u) \in \mathcal{L}(U_2 \otimes U_2, \mathbb{R}), \qquad u \in A.$$
   - 1b. Let $(u_k)_k \subset K$, $(v_k)_k \subset U_2$ be arbitrary sequences such that $u_k \to \bar{u}$ strongly w.r.t. the U_2-norm and $v_k \rightharpoonup v$ weakly in U_2 as $k \to \infty$. Then it holds:
     - 1bi. $f'(\bar{u})v = \lim_{k\to\infty} f'(u_k)v_k$,
     - 1bii. $f''(\bar{u})v^2 \leq \liminf_{k\to\infty} f''(u_k)v_k^2$,
     - 1biii. if $v = 0$, there is some $c > 0$ such that
     $$c \liminf_{k\to\infty} \|v_k\|_2^2 \leq \liminf_{k\to\infty} f''(u_k)v_k^2.$$
2. The functional $g\colon U_2 \to \mathbb{R}$ is assumed to be convex and Lipschitz continuous. By g' and ∂g we denote its directional derivatives and subgradient and introduce the following sets
   $$D_{\bar{u}} := \mathrm{cl}_{U_2}\left(\{v \in \mathcal{R}_K(\bar{u})\colon\ f'(\bar{u})v + g'(\bar{u}, v) = 0\}\right),$$
   $$C_{\bar{u}} := \mathcal{T}_K(\bar{u}) \cap \{v \in U_2\colon\ f'(\bar{u})v + g'(\bar{u}, v) = 0\},$$
   where $\mathcal{R}_K(\bar{u})$ and $\mathcal{T}_K(\bar{u})$ denote the radial and tangent cone of K at $\bar{u}$; see Chapter 0 for the definitions. Moreover, let $g''(\bar{u}, \cdot)\colon U_2 \to \mathbb{R}$ denote a continuous quadratic form such that:
   - 2a. If $v \in D_{\bar{u}}$ there is a sequence $(v_k)_k \subset U_2$ such that $f'(\bar{u})v_k + g'(\bar{u}, v_k) = 0$, $v_k \to v$ in U_2, $u_k := \bar{u} + t_k v_k \in K$, $t_k \searrow 0$, $u_k \to \bar{u}$ in

U_∞, and

$$g''(\bar u, v^2) \geq \lim_{k\to\infty} \frac{2}{t_k^2}\left(g(\bar u + t_k v_k) - g(\bar u) - t_k g'(\bar u, v_k)\right).$$

2b. If $(v_k) \subset U_2$, $(t_k) \subset \mathbb{R}_{>0}$ such that $t_k \searrow 0$, $v_k \rightharpoonup v$ weakly in U_2 with $v \in C_{\bar u}$, $g'(\bar u, v_k) \to g'(\bar u, v)$, and $\bar u + t_k v_k \in K$, it holds

$$g''(\bar u, v^2) \leq \liminf_{k\to\infty} \frac{2}{t_k^2}\left(g(\bar u + t_k v_k) - g(\bar u) - t_k g'(\bar u, v_k)\right).$$

We start with a discussion of first-order necessary optimality conditions.

Theorem 3.6 ([**169**], Theorem 2.2). *Let Assumption 3.5.1a hold and suppose that $\bar u$ is a local minimizer of $(\mathrm{AP}^{\mathrm{sp}})$ w.r.t. the U_2-topology. Then there is $\bar\lambda \in \partial g(\bar u)$ such that*

$$\langle f'(\bar u) + \bar\lambda, u - \bar u\rangle_2 \geq 0, \qquad \forall u \in K. \tag{3.5}$$

The proof works completely analogous to for instance the proof of [**54**, Theorem 2.1]. For convenience we provide the main steps. Hereby, note that both differentiability of f and convexity of g play a crucial role.

Proof. Given $u \in K$, it holds $\hat J(\bar u + t(u - \bar u)) - \hat J(\bar u) \geq 0$ for all sufficiently small $t \in (0,1)$, due to local optimality of $\bar u$ and convexity of K. From convexity of g we infer $g(u) - g(\bar u) \geq t^{-1}[g(\bar u + t(u-\bar u)) - g(\bar u)]$ for all $t \in (0,1)$ and together with differentiability of f we therefore obtain

$$f'(\bar u)(u - \bar u) + g(u) - g(\bar u) \geq \liminf_{t\searrow 0} t^{-1}[\hat J(\bar u + t(u-\bar u)) - \hat J(\bar u)] \geq 0 \quad \forall u \in K.$$

In particular, $\bar u$ is a solution of the following optimization problem on U_2:

$$\min_{u\in K}\ f'(\bar u)u + g(u).$$

The map $f'(\bar u) + g\colon U_2 \to \mathbb{R}$ is convex and continuous and hence the claim follows from standard results of convex analysis; see, e.g., [**106**, Proposition I.5.6]. □

Before addressing second-order optimality conditions, some comments on $D_{\bar u}$ and $C_{\bar u}$ seem to be appropriate.

Lemma 3.7 ([**169**], Lemma 2.3). *Let $\bar u$ and $\bar\lambda$ satisfy (3.5). Then $C_{\bar u}$ is a closed, convex cone in U_2. Moreover, it holds $D_{\bar u} \subset C_{\bar u}$ and $g'(\bar u, v) = \langle \bar\lambda, v\rangle_2$ for all $v \in C_{\bar u}$.*

We can follow, e.g., [**53**, Proposition 3.4] to prove this.

Proof. Closedness and the cone-property of $C_{\bar u}$ as well as the inclusion $D_{\bar u} \subset C_{\bar u}$ are obvious consequences of the definition. Herein, note that it holds $\mathcal{T}_K(\bar u) = \mathrm{cl}_{U_2}(\mathcal{R}_K(\bar u))$; cf. [**36**, Proposition 2.55]. Moreover, (3.5) implies

$$f'(\bar u)v + g'(\bar u, v) \overset{(\star)}{\geq} \langle f'(\bar u) + \bar\lambda, v\rangle_2 \geq 0 \qquad \forall v \in \mathcal{T}_K(\bar u)$$

and that equality in $(\star)$, and hence $g'(\bar u, v) = \langle \bar\lambda, v\rangle_2$, holds for $v \in C_{\bar u}$. For $v_1, v_2 \in C_{\bar u}$ and $t \in (0,1)$ we conclude, utilizing convexity of $g(\bar u, \cdot)$ and convexity of $\mathcal{T}_K(\bar u)$, that

$$0 \leq f'(\bar u)(tv_1 + (1-t)v_2) + g'(\bar u, tv_1 + (1-t)v_2)$$

$$\leq t[f'(\bar{u})v_1 + g'(\bar{u}, v_1)] + (1-t)[f'(\bar{u})v_2 + g'(\bar{u}, v_2)] \leq 0,$$

i.e. $tv_1 + (1-t)v_2 \in C_{\bar{u}}$. This proves convexity of $C_{\bar{u}}$. □

Due to the fact that $g'(\bar{u}, \cdot)$ is not a linear form on U_2, we cannot apply the concept of polyhedricity; see Chapter 0 for the definition. This is different from the smooth case in [**71**], where $D_{\bar{u}} = C_{\bar{u}}$ holds for polyhedric K. In fact, we do not know whether this equality still holds true in our abstract nonsmooth setting. Since in the following sufficient conditions will be formulated on the cone $C_{\bar{u}}$, and necessary conditions on the possibly smaller cone $D_{\bar{u}}$, we do not obtain no-gap second-order conditions for the fully abstract setting. However, for sparse optimization problems on Lebesgue spaces equality holds, cf. Section 3.2.2, because Assumption 3.5.2a can be verified with $D_{\bar{u}}$ replaced by $C_{\bar{u}}$ in these cases. The following necessary second-order optimality condition is the abstract version of the first part of Theorem 3.4.

Theorem 3.8 ([**169**], Theorem 2.4)**.** *Let Assumption 3.5.1 and 2a hold and suppose that there are $c \geq 0$ and $r > 0$ such that*

$$\hat{J}(u) \geq \hat{J}(\bar{u}) + \frac{c}{2}\|u - \bar{u}\|_2^2 \qquad \forall u \in K \cap \mathbb{B}_r^{U_2}(\bar{u}).$$

Then, it holds

$$f''(\bar{u})v^2 + g''(\bar{u}, v^2) \geq c\|v\|_2^2 \qquad \forall v \in D_{\bar{u}}.$$

Proof. Fix $v \in D_{\bar{u}}$. Due to Assumption 3.5.2a there is a sequence $(v_k)_k \subset U_2$, such that $v_k \to v$ strongly in U_2, $u_k = \bar{u} + t_k v_k \in K$, $f'(\bar{u})v + g'(\bar{u}, v) = 0$, $t_k \searrow 0$, and $u_k \to \bar{u}$ in U_∞. It holds

$$\frac{c}{2}t_k^2\|v_k\|_2^2 = \frac{c}{2}\|u_k - \bar{u}\|_2 \leq \hat{J}(u_k) - \hat{J}(\bar{u}) = f(u_k) - f(\bar{u}) + g(u_k) - g(\bar{u})$$
$$= f'(\bar{u})(u_k - \bar{u}) + \frac{1}{2}f''(u_k^\theta)(u_k - \bar{u})^2 + (g(u_k) - g(\bar{u}) - g'(\bar{u}, u_k - \bar{u})) + g'(\bar{u}, u_k - \bar{u})$$

by assumption and Taylor expansion of f at $\bar{u}$ with some $u_k^\theta := (1-\theta_k)\bar{u} + \theta_k u_k$, $\theta_k \in [0,1]$. Exploiting that $f'(\bar{u})(u_k - \bar{u}) + g'(\bar{u}, u_k - \bar{u}) = t_k[f'(\bar{u})v_k + g'(\bar{u}, v_k)] = 0$ and dividing by t_k^2 yields:

$$\frac{c}{2}\|v_k\|_2^2 \leq \frac{1}{2}f''(u_k^\theta)v_k^2 + \frac{1}{t_k^2}\left(g(\bar{u} + t_k v_k) - g(\bar{u}) - t_k g'(\bar{u}, v_k)\right). \tag{3.6}$$

Taking inferior limits on both sides of (3.6) and utilizing Assumption 3.5.1.bii and 2a concludes the proof. □

Obviously, the above proof also keeps valid with slightly modified assumptions. If Assumption 3.5.2a is weakened by replacing limits by inferior limits, we need to strengthen Assumption 3.5.1 by demanding additional continuity of f'' as map $U_\infty \to \mathcal{L}(U_2 \otimes U_2, \mathbb{R})$.

Next, we state and prove sufficient optimality conditions, that correspond —on the abstract level— to the second part of Theorem 3.4.

Theorem 3.9 ([**169**], Theorem 2.5)**.** *Let Assumption 3.5 hold and suppose that $\bar{u} \in K$ and $\bar{\lambda} \in \partial g(\bar{u})$ satisfy the first-order necessary optimality condition*

in Theorem 3.6. If in addition

$$f''(\bar{u}) + g''(\bar{u}, v^2) > 0 \qquad \forall v \in C_{\bar{u}} \setminus \{0\}$$

holds, there are $c, r > 0$ *such that*

$$\hat{J}(u) \geq \hat{J}(\bar{u}) + \frac{c}{2}\|u - \bar{u}\|_2^2 \qquad \forall u \in K \cap \mathbb{B}_r^{U_2}(\bar{u}).$$

In particular, $\bar{u}$ *is a* U_2*-local solution of* $(\mathrm{AP}^{\mathrm{sp}})$.

Proof. We argue by contradiction, following, e.g., the well-known approach in [**71, 54, 61**]. If the statement of the theorem is not true there is a sequence $(u_k)_k \subset K$ such that $u_k \to \bar{u}$ in U_2 and

$$\frac{1}{2k}\|u_k - \bar{u}\|_2^2 > \hat{J}(u_k) - \hat{J}(\bar{u}) = f(u_k) - f(\bar{u}) + g(u_k) - g(\bar{u}). \tag{3.7}$$

We set $t_k := \|u_k - \bar{u}\|_2$, $v_k := t_k^{-1}(u_k - \bar{u})$ and assume w.l.o.g. that $v_k \rightharpoonup v \in U_2$ weakly. The contradiction is achieved in three steps I.-III.

Step I. First, we show that $v \in C_{\bar{u}}$. It clearly holds $v \in \text{weak-cl}_{U_2}(\mathcal{R}_K(\bar{u}))$. Since $\mathcal{R}_K(\bar{u})$ is convex due to convexity of K it follows that the weak and strong closure of $\mathcal{R}_K(\bar{u})$ coincide, cf. [**36**, Theorem 2.23ii], from which we deduce $v \in \mathcal{T}_K(\bar{u})$. From the first-order necessary optimality condition $\langle f'(\bar{u}) + \bar{\lambda}, u_k - \bar{u}\rangle_2 \geq 0$ together with $f'(\bar{u}) + \bar{\lambda} \in U_2^*$ and weak convergence of $(v_k)_k$ we immediately conclude $\langle f'(\bar{u}) + \bar{\lambda}, v\rangle_2 \geq 0$. The subgradient property therefore implies $f'(\bar{u})v_k + g'(\bar{u}, v_k) \geq 0$ and $f'(\bar{u})v + g'(\bar{u}, v) \geq 0$. Applying Taylor expansion to f at $\bar{u}$ we obtain from (3.7)

$$\frac{t_k^2}{2k} > \hat{J}(u_k) - \hat{J}(\bar{u}) \geq f'(u_k^\theta)(u_k - \bar{u}) + g(u_k) - g(\bar{u}),$$

where $u_k^\theta = (1 - \theta_k)\bar{u} + \theta_k u_k$, $\theta_k \in [0, 1]$. Dividing by $t_k > 0$, this leads to

$$f'(u_k^\theta)v_k + g'(\bar{u}, v_k) \leq f'(u_k^\theta)v_k + t_k^{-1}[g(u_k) - g(\bar{u})] \leq \frac{t_k}{2k} \to 0 \quad \text{as } k \to \infty.$$

Taking the inferior limits on both sides hereof and using Assumption 3.5.1bi for the first summand on the left-hand side we obtain

$$f'(\bar{u})v + g'(\bar{u}, v) = f'(\bar{u})v + \liminf_{k\to\infty} t_k^{-1}[g(\bar{u} + t_k v_k) - g(\bar{u})] \leq 0,$$

where we have used that $g'(\bar{u}, \cdot)$ is convex and continuous as in [**61**, Proof of Theorem 4.2]. Hence, this shows $v \in C_{\bar{u}}$. Moreover, we have $-f'(\bar{u})v_k \leq g'(\bar{u}, v_k) \leq \frac{t_k}{2k} - f'(u_k^\theta)v_k$, and hence $g'(\bar{u}, v_k) \to -f'(\bar{u})v = g'(\bar{u}, v)$ as $k \to \infty$.

Step II. Next, we prove $v = 0$. Again, we apply Taylor expansion to f in (3.7) and obtain with some $\tilde{u}_k^\theta = (1 - \tilde{\theta}_k)\bar{u} + \tilde{\theta}_k u_k$, $\tilde{\theta}_k \in [0, 1]$:

$$\begin{aligned}\frac{t_k^2}{2k} &> f'(\bar{u})(u_k - \bar{u}) + \frac{1}{2}f''(\tilde{u}_k^\theta)(u_k - \bar{u})^2 + g'(\bar{u}, u_k - \bar{u}) \\ &\qquad + (g(u_k) - g(\bar{u}) - g'(\bar{u}, u_k - \bar{u})) \\ &\geq \frac{1}{2}f''(\tilde{u}_k^\theta)(u_k - \bar{u})^2 + (g(u_k) - g(\bar{u}) - g'(\bar{u}, u_k - \bar{u}))\,.\end{aligned}$$

Here, we have used the first-order necessary condition in the second inequality. Dividing by t_k^2 and taking the inferior limits on both sides yields:

$$\begin{aligned}0 &\geq \liminf_{k\to\infty} \left(\frac{1}{2} f''(\tilde{u}_k^\theta) v_k^2 + \frac{1}{t_k^2} \left(g(\bar{u} + t_k v_k) - g(\bar{u}) - t_k g'(\bar{u}, v_k) \right) \right) \\ &\geq \liminf_{k\to\infty} \frac{1}{2} f''(\tilde{u}_k^\theta) v_k^2 + \liminf_{k\to\infty} \frac{1}{t_k^2} \left(g(\bar{u} + t_k v_k) - g(\bar{u}) - t_k g'(\bar{u}, v_k) \right) \\ &\geq \frac{1}{2} f''(\bar{u}) v^2 + \frac{1}{2} g''(\bar{u}, v^2),\end{aligned}$$

where we have applied Assumptions 3.5.1bii and 2b in the last step. Due to $v \in C_{\bar{u}}$ it follows from the assumption of the theorem that $v = 0$.

Step III. In this final step we arrive at the desired contradiction. From Assumption 3.5.1biii and $\|v_k\|_2 = 1$ we infer from the above considerations:

$$\begin{aligned}0 < c \leq \liminf_{k\to\infty} f''(\tilde{u}_k^\theta) v_k^2 &\leq \liminf_{k\to\infty} \left(\frac{1}{2k} - \frac{1}{t_k^2} \left(g(\bar{u} + t_k v_k) - g(\bar{u}) - t_k g'(\bar{u}, v_k) \right) \right) \\ &\leq - \limsup_{k\to\infty} \frac{1}{t_k^2} \left(g(\bar{u} + t_k v_k) - g(\bar{u}) - t_k g'(\bar{u}, v_k) \right).\end{aligned}$$

Since the term inside the limes superior is always nonnegative due to convexity of g we arrive at the desired contradiction $0 < c \leq 0$. □

The crucial observation in the final step of the proof of Theorem 3.9 is the inequality

$$\liminf_{k\to\infty} f''(\tilde{u}_k^\theta) v_k^2 \leq - \limsup_{k\to\infty} \frac{1}{t_k^2} \left(g(\bar{u} + t_k v_k) - g(\bar{u}) - t_k g'(\bar{u}, v_k) \right).$$

Assumption 3.5.1biii ensures positivity of the left-hand side, while convexity of g implies nonpositivity of the right-hand side. Without Assumption 3.5.1biii, we would only have nonnegativity of the left-hand side, which does not suffice to achieve a contradiction, unless the right-hand side could be shown to be negative. However, we think that the latter can only hold for g being strongly convex at $\bar{u}$ w.r.t. the U_2-norm. A strongly convex function g, however, is the sum of a convex function and a U_2-Tikhonov term $u \mapsto \frac{\gamma}{2}\|u\|_2^2$. Such a Tikhonov term is smooth and would ensure Assumption 3.5.1biii when being shifted to f. This is the reason why the application of Theorem 3.9 will be restricted to the regular case in the subsequent sections. Note that this is the same as in the proofs of Theorem 2.8 or [**71**, Theorem 2.3]: in both results, the final contradiction in the proof by contradiction is achieved by utilizing Assumption 3.5.1biii. Consequently, not only Theorem 3.4 but all second-order results dealt with in this thesis exclusively pertain to the regular case since they are based either on [**71**] or extensions hereof. As will be explained in Section 3.3.3, we expect that a different type of argument is needed for the the bang-bang case.

To conclude the section, we mention that similarly as in Remark 2.12 a variant of Theorem 3.9 with norm gap can also be obtained. Let strong convergence $u_k \to \bar{u}$ in Assumption 3.5.1b hold only w.r.t. another Banach space V such that $V \hookrightarrow U_\infty$, e.g., $V = U_\infty$. Under this weaker supposition, the quadratic growth condition in Theorem 3.9 holds true in a V-neighbourhood of $\bar{u}$, and, consequently, $\bar{u}$ is a V-local solution to ($\mathrm{AP}^{\mathrm{sp}}$).

3.2.2. Optimality conditions for directionally sparse optimization on Lebesgue spaces. In this section we incorporate directionally sparse optimization problems on Lebesgue spaces into the framework established before. We replace $(\mathrm{AP}^{\mathrm{sp}})$ by the following slightly more concrete model problem that contains $(\mathrm{P}_k^{\mathrm{sp}})$ as an instance: given a complete, finite measure space (Λ, ρ) we consider

$$(\mathrm{LP}_X) \qquad \min_{u\in U_{\mathrm{ad}}} \hat{J}(u) := \underbrace{h(u) + \frac{\alpha}{2}\|u\|^2_{L^2(\Lambda)}}_{=:f(u)} + \underbrace{\beta j_X(u)}_{=:g(u)}, \qquad X \in \{A, B, C, D\},$$

with the following four typical (directional) sparsity-enforcing functionals j_X defined on $L^2(\Lambda)$:

A. $j_A = \|\cdot\|_{L^1(\Lambda)}$,
B. $j_B = \|\cdot\|_{L^1(\Lambda_1, L^2(\Lambda_2))}$,
C. $j_C = \|\cdot\|_{L^2(\Lambda_1, L^1(\Lambda_2))}$,
D. $j_D = \frac{1}{2}\|\cdot\|^2_{L^2(\Lambda_1, L^1(\Lambda_2))}$,

where in the cases B-D, (Λ, ρ) is given by the product measure space of two complete, finite measure spaces (Λ_1, ρ_1) and (Λ_2, ρ_2). Moreover, let $\alpha \geq 0$, $\beta > 0$ and $f\colon L^s(\Lambda) \to \mathbb{R}$ fulfill Assumption 3.5.1 with $U_\infty = L^s(\Lambda)$, $s > 2$, $U_2 = L^2(\Lambda)$, and $K = U_{\mathrm{ad}} := \{u \in L^s(\Lambda)\colon u_{\mathrm{a}} \leq u \leq u_{\mathrm{b}}\ \rho\text{-a.e. on } \Lambda\}$ with $u_{\mathrm{a}}, u_{\mathrm{b}} \in L^\infty(\Lambda)$. In the following we will analyze first-order optimality conditions for (LP_X) together with the resulting sparsity patterns of the minimizers. Moreover, we will verify that functionals j_X, $X \in \{A, B, C, D\}$, fit into the framework of Section 3.2.1 which allows to formulate second-order necessary and sufficient conditions for (LP_X) provided that h fulfills the appropriate assumptions.

Before we put the content of this section into the context of the literature, a few comments motivating the formulation of (LP_X) are in order. Note that functionals j_1-j_7 from Section 3.1 are included in this setting by an appropriate choice of Λ_1 and Λ_2; see Table 3.1 in Section 3.3.1 below. Let us explain this briefly on behalf of, e.g., the functional j_2. First, one has to observe that the space $L^2(I, \mathbb{R}^m)$ can be written equivalently as $L^2(\Lambda)$ with $\Lambda = \Lambda_1 \times \Lambda_2$, where (Λ_1, ρ_1) is the m-element set equipped with the counting measure and (Λ_2, ρ_2) is the interval I equipped with the Lebesgue measure. Next, a short computation shows that j_2 can be expressed as $u \mapsto \|u\|_{L^1(\Lambda_1, L^2(\Lambda_2))}$ in this setting. In essence, one has to observe that measurable functions on Λ_1 can be identified with vectors in $\mathbb{R}^m$ and that the $L^1(\Lambda_1)$-norm corresponds to the ℓ^1-norm on $\mathbb{R}^m$ under this identification; the variable $x_1 \in \Lambda_1$ corresponds to the index i enumerating the actuators in Assumption 3.1 and the variable $x_2 \in \Lambda_2$ corresponds to the time t. Thus, j_2 is a particular instance of the abstract functional j_B introduced above.

In Section 3.3, i.e. in the proof of our main results for $(\mathrm{P}_k^{\mathrm{sp}})$, the smooth functional h will be given by the first summand of the reduced functional of $(\mathrm{P}_k^{\mathrm{sp}})$. Nevertheless, the particular choice of h does not matter for the arguments of the present section. This is the reason why we can easily adapt proofs from earlier literature concerned with semilinear elliptic or parabolic problems to the present setting. Again, let us briefly illustrate the case $k = 2$. The formulas (3.10) and (3.11) for the directional derivatives and the subgradient of j_B that we will obtain in this section can easily be translated back to j_2 in formulas (3.28) and (3.21). Similarly, by translating back Proposition 3.11.1 on first-order conditions for (P_B)

to the setting of j_2 we prove the case $k = 2$ in Theorem 3.3. Here, it is important to observe that the choice $h(u) = \frac{1}{2}\|S(u) - y_d\|^2_{L^2(Q)}$ in (P_B), where S denotes the solution map of the quasilinear equation (Eq), is possible because it has been already been shown in [**35**] that h exhibits the required properties; cf. our summary of the results from [**35**] in Chapter 1. Let us also sketch how to obtain second-order optimality conditions. To do so, we apply the abstract Theorems 3.8 and 3.9 with $U_2 = L^2(I, \mathbb{R}^m)$, $U_\infty = L^s(I, \mathbb{R}^m)$, $K = U_{\mathrm{ad}}$, $f(u) = h(u) + \frac{\alpha}{2}\|u\|^2_{L^2(I,\mathbb{R}^m)} = \frac{1}{2}\|S(u) - y_d\|^2_{L^2(Q)} + \frac{\alpha}{2}\|u\|^2_{L^2(I,\mathbb{R}^m)}$ and $g(u) = \beta j_2(u)$. That under Assumptions 1.5, 1.6 and 3.1 the smooth part of the functional, f, satisfies the respective assumptions has again already been proven in [**35**]; see Proposition 1.21. The assumptions on the nonsmooth part, g, will be verified in Proposition 3.11.2 of this section for $g = \beta j_B$, and hence, by the above reasoning, for $g = \beta j_2$. Let us in particular point out that in all arguments the smooth and the nonsmooth part of the appearing functionals can be handeled completely independent of each other as long as both parts exhibit the required assumptions. This is the reason why earlier results obtained in the context of optimal control of linear or semilinear equations can easily be transferred to the present case.

At a first look, one may wonder whether argueing in such an abstract way is worth the effort coming along with this. In fact, the benefits are the following: first, instead of checking assumptions for all seven j_k we only need to check four generic cases in this section. For instance, once the abstract functional j_B has been analyzed, the results immediately apply to the two concrete functionals j_2 and j_3; in order to deal with j_3 one has to interchange the choice of Λ_1 and Λ_2 compared to j_2. Further, replacing the m-element set equipped with the counting measure by Ω equipped with the Lebesgue measure one can easily address the case of distributed directionally sparse optimal control as in, e.g., [**54**]. Second, argueing in an abstract way reveals that the actual choice of the state equation (entering only via the functional h) does not matter as long as the smooth part f of the functional stil exhibits the typical properties. Therefore, the abstract setting is also intended to facilitate the application towards problems with a different state equation.

Let us now put the content of this section into the context of the literature. Except for case D, all these results have already been obtained in [**261, 138, 54, 61**] dealing with linear and semilinear problems in a more concrete setting. There, Λ is a domain in $\mathbb{R}^d$ or a space-time cylider, equipped with the Lebesgue measure, and h is a smooth tracking-type functional originating from optimal control of a linear or semilinear PDE. For functional j_3 in the context of optimal control of an ordinary differential equation we refer the reader to [**251**]. As pointed out above, the proofs also apply to our abstract setting because they are actually independent of the concrete structure of h. Nevertheless, for convenience of the reader we repeat these results in our notation. We also mention that discrete analoga of functionals A and B are well-known in the machine learning as "lasso" [**264**] and "group lasso" [**291**].

To the best of our knowledge, our preprint [**169**] contains the first analysis of case D in the context of PDE-constrained optimization. It can be motivated by the successfull use of analogous functionals in the discrete setting, e.g., the so-called "exclusive lasso" [**39**] in machine learning, or the sparse regression problem in [**187**].

In particular, j_D results in sparsity patterns similar to j_C, but, as well-known in the discrete case, j_D unlike j_C allows the application of proximal algorithms; cf. Section 3.4.1. Therefore, we may view case D as an alternative to case C that also deserves a theoretical analysis.

In the following we repeatedly make use of the fact that $h'(\bar{u}) \in L^2(\Lambda)^*$ can be identified with its Riesz-representative $\nabla h(u) \in L^2(\Lambda)$. Above, we introduced the functionals j_X, $X \in \{A, B, C, D\}$, to be defined on the Hilbert space $U_2 = L^2(\Omega)$. By $\tilde{j}_X$ we refer to their respective, straightforward extensions to $L^1(\Lambda)$ in the case A, to $L^1(\Lambda_1, L^2(\Lambda_2))$ in the case B, or to $L^2(\Lambda_1, L^1(\Lambda_2))$ in the cases C and D.

Functional A. Both, directional derivatives and subdifferential of j_A, are well-known; see for instance [**53**]. Note that the proofs that originally pertain to Lebesgue spaces on open sets of $\mathbb{R}^d$ apply to our slightly more general setting without changes. Therefore, for some $u, v \in L^1(\Lambda)$ the directional derivatives of $\tilde{j}_A$: $L^1(\Lambda) \to \mathbb{R}$ are given by

$$\tilde{j}_A'(u, v) = \int_{\Lambda^+(u)} v \mathrm{d}\rho - \int_{\Lambda^-(u)} v \mathrm{d}\rho + \int_{\Lambda^0(u)} |v| \mathrm{d}\rho, \tag{3.8}$$

where we use the notation $\Lambda^0(u) := \{x \in \Lambda\colon u(x) = 0\}$ and $\Lambda^{\pm}(u) := \{x \in \Lambda\colon u(x) \gtrless 0\}$. The subdifferential is given by

$$\partial \tilde{j}_A(u) = \{\lambda \in L^\infty(\Lambda)\colon \ \lambda(x) \in \operatorname{sign}(u(x)) \quad \rho\text{-a.e. on } \Lambda\}\,. \tag{3.9}$$

Here, recall that $L^1(\Lambda)^* = L^\infty(\Lambda)$ for any σ-finite measure space (Λ, ρ); cf. [**242**, Satz 6.16]. However, note that we will actually be concerned with $j_A = \tilde{j}_A \circ \iota$ where ι: $L^2(\Lambda) \hookrightarrow L^1(\Lambda)$ denotes the canonical embedding. It is obvious, that the formulas for the directional derivatives remain true for j_A. Regarding the subdifferential, recall that by the chain rule, see, e.g., [**106**, Proposition 5.7], it holds $\partial j_A(u) = \partial(\tilde{j}_A \circ \iota)(u) = \iota^* \partial \tilde{j}_A(\iota u) = \iota^* \partial \tilde{j}_A(u)$, which implies that the above characterization of the subdifferential is also valid for j_A because ι^* acts as the embedding $L^\infty(\Lambda) \hookrightarrow L^2(\Lambda)$.

Proposition 3.10 ([**169**], Proposition 2.6)**.** *Let $\bar{u} \in U_{\mathrm{ad}}$ be a local solution to* (LP_A).

1. *If $\alpha > 0$, it holds ρ-a.e. on Λ:*

$$\begin{aligned} \bar{u}(x) = 0 \quad &\Leftrightarrow \quad |\nabla h(\bar{u})(x)| \le \beta, \\ \bar{\lambda} &= \operatorname{Proj}_{[-1,1]}\left(-\beta^{-1} \nabla h(\bar{u})\right), \\ \bar{u} &= \operatorname{Proj}_{[u_{\mathrm{a}}, u_{\mathrm{b}}]}\left[-\alpha^{-1}\left(\nabla h(\bar{u}) + \beta \bar{\lambda}\right)\right]. \end{aligned}$$

   If $\alpha = 0$, it holds for ρ-a.e. on Λ:

$$\begin{aligned} |\nabla h(\bar{u})(x)| < \beta \quad &\Rightarrow \quad \bar{u}(x) = 0, \\ \nabla h(\bar{u})(x) > \beta \quad &\Rightarrow \quad \bar{u}(x) = u_{\mathrm{a}}(x), \\ \nabla h(\bar{u})(x) < -\beta \quad &\Rightarrow \quad \bar{u}(x) = u_{\mathrm{b}}(x). \end{aligned}$$

2. *$g = \beta j_A$ satisfies the properties of Assumption 3.5.2 with $D_{\bar{u}}$ replaced by $C_{\bar{u}}$ and $g''(\bar{u}, v^2) \equiv 0$ for all $v \in L^2(\Lambda)$.*

Proof. The first-order conditions and the analysis of the sparsity pattern can be found in [**53**, Corollary 3.2]. Regarding the second statement, Assumption 3.5.2a is verified in the proof of [**53**, Theorem 3.7], while Assumption 3.5.2b is an immediate consequence of the convexity of j_A. Note that the properties of the nonsmooth cost term in [**53**] are actually independent of those of the smooth part of the functional, which is the only part of the functional specifically related to the state equation. Hence, the transfer of the techniques from the semilinear elliptic setting in [**53**] does not cause problems. □

Functional B. This functional has been discussed in [**54, 61**] for the special case that Λ_1 is a domain in $\mathbb{R}^d$ and Λ_2 is an interval, both equipped with the Lebesgue measure. We refer the reader to [**251**] for the particular case j_3 in the context of an optimal control problem with ODE-constraints. The results and their proofs also apply to our setting. Using the notation

$$\Lambda_1^0(u) = \{x_1 \in \Lambda_1 \colon \ \|u(x_1,\cdot)\|_{L^2(\Lambda_2)} = 0\}$$

we obtain the directional derivatives and subgradients of $\tilde{j}_B \colon L^1(\Lambda_1, L^2(\Lambda_2)) \to \mathbb{R}$; cf. [**54**, Proposition 2.8]:

(3.10)
$$\begin{aligned}\tilde{j}_B'(u,v) = &\int_{\Lambda_1^0(u)} \|v(x_1,\cdot)\|_{L^2(\Lambda_2)} \mathrm{d}\rho_1(x_1) \\ &+ \int_{\Lambda_1 \setminus \Lambda_1^0(u)} \frac{1}{\|u(x_1,\cdot)\|_{L^2(\Lambda_2)}} \int_{\Lambda_2} u(x_1,x_2) v(x_1,x_2) \mathrm{d}\rho_2(x_2) \mathrm{d}\rho_1(x_1),\end{aligned}$$

(3.11)
$$\partial \tilde{j}_B(u) = \{\lambda \in L^\infty(\Lambda_1, L^2(\Lambda_2)) \colon \lambda(x_1,\cdot) \begin{cases} \in \overline{\mathbb{B}_1^{L^2(\Lambda_2)}(0)}, & \text{if } x_1 \in \Lambda_1^0(u), \\ = \frac{u(x_1,\cdot)}{\|u\|_{L^1(\Lambda_1, L^2(\Lambda_2))}}, & \text{if } x_1 \notin \Lambda_1^0(u), \end{cases} \quad \rho_1\text{-a.e. on } \Lambda_1\}.$$

Here, note that $L^1(\Lambda_1, L^2(\Lambda_2))^* = L^\infty(\Lambda_1, L^2(\Lambda_2))$; cf. [**104**, Theorem 8.18.3]. As for case A, we obtain the representation of the subdifferential of j_B on $L^2(\Lambda)$ by an application of the chain rule.

Proposition 3.11 ([**169**], Proposition 2.7)**.** *Let $\bar{u} \in U_{\mathrm{ad}}$ be a local solution to (LP_B).*

1. *If $\alpha > 0$ it holds ρ_1-a.e. on Λ_1 or ρ-a.e. on Λ, respectively:*

$$\begin{aligned} &\|u(x_1,\cdot)\|_{L^2(\Lambda_2)} = 0 \quad \Leftrightarrow \quad \|\nabla h(\bar{u})(x_1,\cdot)\|_{L^2(\Lambda_2)} \le \beta, \\ &\bar{\lambda}(x_1,x_2) = \begin{cases} -\beta^{-1} \nabla h(\bar{u})(x_1,x_2) & \text{if } x_1 \in \Lambda_1^0(u), \\ \frac{u(x_1,x_2)}{\|\bar{u}(x_1,\cdot)\|_{L^2(\Lambda_2)}} & \text{if } x_1 \in \Lambda_1 \setminus \Lambda_1^0(u), \end{cases} \\ &\bar{u} = \mathrm{Proj}_{[u_{\mathrm{a}}, u_{\mathrm{b}}]} \left(-\alpha^{-1} \left(\nabla h(\bar{u}) + \beta \bar{\lambda} \right) \right). \end{aligned}$$

   If $\alpha = 0$ it holds ρ_1-a.e. on Λ_1:

$$\begin{aligned} \|\nabla h(\bar{u})(x_1,\cdot)\|_{L^2(\Lambda_2)} < \beta \quad &\Rightarrow \quad u(x_1,\cdot) \equiv 0, \\ \bar{u}(x_1,\cdot) \equiv 0 \quad &\Rightarrow \quad \|\nabla h(\bar{u})(x_1,\cdot)\|_{L^2(\Lambda_2)} \le \beta. \end{aligned}$$

2. *$g = \beta j_B$ satisfies Assumption 3.5.2 with $D_{\bar u}$ replaced by $C_{\bar u}$,*

$$g''(\bar u, v^2) := \beta \int_{\Lambda_1 \setminus \Lambda_1^0(\bar u)} \frac{1}{\|\bar u(x_1,\cdot)\|_{L^2(\Lambda_2)}} \Big[\|v(x_1,\cdot)\|^2_{L^2(\Lambda_2)} - \Big(\int_{\Lambda_2} \frac{\bar u(x_1,x_2) v(x_1,x_2)}{\|\bar u(x_1,\cdot)\|_{L^2(\Lambda_2)}} \mathrm{d}\rho_2(x_2) \Big)^2 \Big] \mathrm{d}\rho_1(x_1)$$

for $\bar u \neq 0$, and $g''(0, v^2) \equiv 0$ otherwise.

Proof. For the first part, see [**54**, Corollary 2.9]. For the second part, Case III in the proof of [**54**, Theorem 3.3] and [**61**, Section 4] prove Assumption 3.5.2a and 2b. As already observed in the case A, the specific semilinear parabolic state equation under consideration in [**54, 61**] does not influence the properties of the nonsmooth cost term. Therefore, the adaptation of the cited results is possible. □

Functional C. This functional has been addressed in [**54**] for the special case that Λ_1 is an interval and Λ_2 is a domain in $\mathbb{R}^d$, both equipped with the Lebesgue measure. The proof, however, also applies to our setting and we obtain expressions for the directional derivatives and the subdifferential of $\tilde j_C$: $L^2(\Lambda_1, L^1(\Lambda_2)) \to \mathbb{R}$ as follows; cf. [**54**, Proposition 2.4]:

$$\tilde j_C'(u, v) = \frac{1}{\|u\|_{L^2(\Lambda_1, L^1(\Lambda_2))}} \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_2)}(u(x_1,\cdot), v(x_1,\cdot)) \mathrm{d}\rho_1(x_1). \tag{3.12}$$

Regarding the subdifferential, first note that $L^2(\Lambda_1, L^1(\Lambda_2))^* = L^2_{\text{w-}\star}(\Lambda_1, L^\infty(\Lambda_2))$, where the latter denotes the space of weak-$\star$ measurable functions $\Lambda_1 \to L^\infty(\Lambda_2)$, equipped with the $L^2(\Lambda_1, L^\infty(\Lambda_2))$-norm; cf. [**104**, Theorem 8.20.3]. The subdifferential is given by

$$\partial \tilde j_C(u) = \Big\{ \lambda \in L^2_{\text{w-}\star}(\Lambda_1, L^\infty(\Lambda_2)) : \lambda(x_1,x_2) \in \operatorname{sign}(u(x_1,x_2)) \frac{\|u(x_1,\cdot)\|_{L^1(\Lambda_2)}}{\|u\|_{L^2(\Lambda_1,L^1(\Lambda_2))}} \quad \rho\text{-a.e. on } \Lambda \Big\}. \tag{3.13}$$

As for A and B, the formulas for the directional derivatives also stay true for j_C instead of $\tilde j_C$. If we denote by ι the embedding $L^2(\Lambda_1 \times \Lambda_2) \hookrightarrow L^2(\Lambda_1, L^1(\Lambda_2))$, it follows by the chain rule that $\partial j_C(u) = \iota^* \partial \tilde j_C(u)$, where ι^* is the embedding $L^2_{\text{w-}\star}(\Lambda_1, L^\infty(\Lambda_2)) \hookrightarrow L^2(\Lambda_1 \times \Lambda_2)$. Note that this embedding is a consequence of the separability of $L^2(\Lambda_2)$, Pettis' measurability theorem [**104**, Theorem 8.15.2], and Fubini's theorem.

Proposition 3.12 ([**169**], Proposition 2.8). *Let $\bar u \in U_{\text{ad}}$ be a local solution to (LP_C) and define*

$$\bar\gamma(x_1) = \frac{\|\bar u(x_1,\cdot)\|_{L^1(\Lambda_2)}}{\|\bar u\|_{L^2(\Lambda_1,L^1(\Lambda_2))}} \text{ if } \bar u \neq 0 \quad \text{and } \gamma(x_1) = 1 \text{ else.}$$

1. *If $\alpha > 0$ it holds for ρ-a.a. $x \in \Lambda$:*

$$\begin{aligned} &\bar u(x_1,x_2) = 0 \quad \Leftrightarrow \quad |\nabla h(\bar u)(x_1,x_2)| \le \beta \bar\gamma(x_1), \\ &\bar\lambda(x_1,x_2) = \operatorname{Proj}_{[-\bar\gamma(x_1),\bar\gamma(x_1)]} \left(-\beta^{-1} \nabla h(\bar u)(x_1,x_2) \right), \\ &\bar u(x_1,x_2) = \operatorname{Proj}_{[u_{\mathrm{a}},u_{\mathrm{b}}]} \left[-\alpha^{-1} \left(\nabla h(\bar u)(x_1,x_2) + \beta \bar\lambda(x_1,x_2) \right) \right]. \end{aligned}$$

If $\alpha = 0$ it holds for ρ-a.a. $x \in \Lambda$:

$$\begin{aligned} |\nabla h(\bar{u})(x)| < \beta\gamma(x_1) \quad &\Rightarrow \quad \bar{u}(x) = 0, \\ \nabla h(\bar{u})(x) > \beta\gamma(x_1) \quad &\Rightarrow \quad \bar{u}(x) = u_{\mathrm{a}}(x), \\ \nabla h(\bar{u})(x) < -\beta\gamma(x_1) \quad &\Rightarrow \quad \bar{u}(x) = u_{\mathrm{b}}(x). \end{aligned}$$

2. *If in addition $\nabla h(\bar{u}) \in L^\infty(\Lambda)$ holds, then $g = \beta j_C$ fulfills Assumption 2.7.2 with $D_{\bar{u}}$ replaced by $C_{\bar{u}}$,*

$$g''(\bar{u}, v^2) = \frac{\beta}{\|\bar{u}\|_{L^2(\Lambda_1, L^1(\Lambda_2))}} \left(\int_{\Lambda_1} \|\cdot\|'^2_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot), v(x_1,\cdot)) \mathrm{d}\rho_1(x_1) - j'_C(\bar{u}, v)^2 \right)$$

for $\bar{u} \neq 0$, and $g''(0, v^2) \equiv 0$ otherwise.

Proof. As for the cases A and B, the concrete type of the state equation in [**54**] does not influence the properties of the nonsmooth cost term. Therefore, we can easily adapt the respective arguments: for the first part, see [**54**, Corollary 2.6]. For the second part, note that Assumption 3.5.2a is verified in Case II of the proof of [**54**, Theorem 3.3], while 2b is obtained as follows: for t_k, v_k as in Assumption 3.5.2b it follows from [**54**, Lemma 4.7] that

$$\liminf_{k\to\infty} \frac{2}{t_k^2} [j_C(\bar{u} + t_k v_k) - j_C(\bar{u}) - t_k j'_C(\bar{u}, v_k)] \geq \liminf_{k\to\infty} j''_C(\bar{u}, v_k^2). \tag{3.14}$$

From [**54**, Lemma 4.6] we know that $v_k \rightharpoonup v$ in $L^2(\Lambda)$, $v \in C_{\bar{u}}$, and $j'_C(\bar{u}, v_k) \to j'_C(\bar{u}, v)$ implies that the sequence of functions $x_1 \mapsto \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot), v_k(x_1,\cdot))$ converges weakly in $L^2(\Lambda_1)$ to $x_1 \mapsto \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot), v(x_1,\cdot))$. Consequently, by weak lower semicontinuity of the $L^2(\Lambda_1)$-norm it holds

$$\begin{aligned} \liminf_{k\to\infty} \int_{\Omega_1} \|\cdot\|'^2_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot), v_k(x_1,\cdot)) \mathrm{d}\rho_1(x_1) \\ \geq \int_{\Omega_1} \|\cdot\|'^2_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot), v(x_1,\cdot)) \mathrm{d}\rho_1(x_1). \end{aligned}$$

Due to $j'_C(\bar{u}, v_k) \to j'_C(\bar{u}, v)$ and the formula for j''_C stated in the proposition, we conclude $\liminf_{k\to\infty} j''_C(\bar{u}, v_k^2) \geq j''(\bar{u}, v^2)$, and together with (3.14) the claim follows. □

The additional assumption $\nabla h(\bar{u}) \in L^\infty(\Lambda)$ is only required to verify Assumption 3.5.2a, i.e. for second-order necessary optimality conditions.

Functional D. Discrete versions of j_D are well-known in the machine learning community, cf., e.g., [**187, 39**]. We now provide an analysis of j_D in the present infinite dimensional setting as done for j_A-j_C. First, a short computation shows for any $u, v \in L^2(\Lambda_1, L^1(\Lambda_2))$ and $t > 0$:

(3.15)

$$\begin{aligned} \tilde{j}_D(u+tv) - \tilde{j}_D(u) &= \frac{1}{2} \int_{\Lambda_1} \left(\|u(x_1,\cdot) + tv(x_1,\cdot)\|^2_{L^1(\Lambda_2)} - \|u(x_1,\cdot)\|^2_{L^1(\Lambda_2)} \right) \mathrm{d}\rho_1(x_1) \\ &= \int_{\Lambda_1} \|u(x_1,\cdot)\|_{L^1(\Lambda_2)} \cdot \left(\|u(x_1,\cdot) + tv(x_1,\cdot)\|_{L^1(\Lambda_2)} - \|u(x_1,\cdot)\|_{L^1(\Lambda_2)} \right) \mathrm{d}\rho_1(x_1) \end{aligned}$$

$$+\frac{1}{2}\int_{\Lambda_1}\left(\|u(x_1,\cdot)+tv(x_1,\cdot)\|_{L^1(\Lambda_2)}-\|u(x_1,\cdot)\|_{L^1(\Lambda_2)}\right)^2\mathrm{d}\rho_1(x_1).$$

Dividing by t and sending $t\searrow 0$ yields:

$$\tilde{j}_D'(u,v)=\int_{\Lambda_1}\|\cdot\|'_{L^1(\Lambda_2)}(u(x_1,\cdot),v(x_1,\cdot))\|u(x_1,\cdot)\|_{L^1(\Lambda_2)}\mathrm{d}\rho_1(x_1). \tag{3.16}$$

Consequently, we obtain

$$\begin{aligned}\partial\tilde{j}_D(u)=\big\{&\lambda\in L^2_{\text{w-}\star}(\Lambda_1,L^\infty(\Lambda_2)):\\ &\lambda(x_1,x_2)\in\operatorname{sign}(u(x_1,x_2))\|u(x_1,\cdot)\|_{L^1(\Lambda_2)}\quad\rho\text{-a.e. on }\Lambda\big\}\end{aligned} \tag{3.17}$$

and as for A-C these formulas remain true for j_D instead of $\tilde{j}_D$.

Proposition 3.13 ([**169**], Proposition 2.9). *Let $\bar{u}\in U_{\mathrm{ad}}$ be a local solution to (LP_D) and define*

$$\bar{\gamma}(x_1)=\|\bar{u}(x_1,\cdot)\|_{L^1(\Lambda_2)}\ \textit{if}\ \bar{u}\neq 0\qquad\textit{and}\ \gamma(x_1)=1\ \textit{else.}$$

1. *If $\alpha>0$ it holds for ρ-a.a. $x\in\Lambda$:*

$$\begin{aligned}&\bar{u}(x_1,x_2)=0\quad\Leftrightarrow\quad|\nabla h(\bar{u})(x_1,x_2)|\le\beta\bar{\gamma}(x_1),\\ &\bar{\lambda}(x_1,x_2)=\operatorname{Proj}_{[-\bar{\gamma}(x_1),\bar{\gamma}(x_1)]}\left(-\beta^{-1}\nabla h(\bar{u})(x_1,x_2)\right),\\ &\bar{u}(x_1,x_2)=\operatorname{Proj}_{[u_{\mathrm{a}},u_{\mathrm{b}}]}\left[-\alpha^{-1}\left(\nabla h(\bar{u})(x_1,x_2)+\beta\bar{\lambda}(x_1,x_2)\right)\right].\end{aligned}$$

   If $\alpha=0$ it holds for ρ-a.a. $x\in\Lambda$:

$$\begin{aligned}|\nabla h(\bar{u})(x)|<\beta\gamma(x_1)\quad&\Rightarrow\quad\bar{u}(x)=0,\\ \nabla h(\bar{u})(x)>\beta\gamma(x_1)\quad&\Rightarrow\quad\bar{u}(x)=u_{\mathrm{a}}(x),\\ \nabla h(\bar{u})(x)<-\beta\gamma(x_1)\quad&\Rightarrow\quad\bar{u}(x)=u_{\mathrm{b}}(x).\end{aligned}$$

2. *$g=\beta j_D$ fulfills Assumption 3.5.2 with $D_{\bar{u}}$ replaced by $C_{\bar{u}}$,*

$$g''(\bar{u},v^2)=\beta\int_{\Lambda_1}\|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot),v(x_1,\cdot))^2\mathrm{d}\rho_1(x_1)$$

   for $\bar{u}\neq 0$, and $g''(0,v^2)\equiv 0$ otherwise.

Proof. Although functional j_D has not been under consideration in [**54**] we will make use of some techniques and intermediate results from [**54**] in the following. As explained for the cases A-C, they can be applied in our setting because the analysis of the nonsmooth cost terms in [**54**] is actually independent of the concrete choice of the smooth functional. Part one is verified along the lines of the proof of [**54**, Corollary 2.6] utilizing the above formula (3.17) for the subgradient. Regarding part two, we start with the verification of Assumption 3.5.2a with $D_{\bar{u}}=C_{\bar{u}}$. Let $v\in C_{\bar{u}}$ and $\bar{u}\neq 0$. As in the case II of the proof of [**54**, Theorem 3.3] we define $v_k\in L^2(\Lambda)$ by $v_k(x)=0$ if $u(x)\in[u_{\mathrm{a}}(x),u_{\mathrm{a}}(x)+k^{-1})\cup(-k^{-1},0)\cup(0,k^{-1})\cup(u_{\mathrm{b}}(x)-k^{-1},u_{\mathrm{b}}(x)]$, and $v_k(x)=\operatorname{Proj}_{[-k,k]}(v(x))$, otherwise, and observe that $v_k\to v$ in $L^2(\Lambda)$, and $\bar{u}+tv_k\in U_{\mathrm{ad}}$ if $0<t<k^{-1}$. Moreover, it follows directly from the definition of v_k that

$$\begin{aligned}\|\bar{u}(x_1,\cdot)+tv_k(x_1,\cdot)\|_{L^1(\Lambda_2)}=\|\bar{u}(x_1,\cdot)\|_{L^1(\Lambda_2)}\\ +t\|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1,\cdot),v_k(x_1,\cdot))\end{aligned} \tag{3.18}$$

holds ρ_1-a.e. on Λ_1 for $0 < t < k^{-2}$. With similar arguments as in [**54**] it can be shown that $f'(\bar{u})v_k + \beta j_D'(\bar{u}, v_k) = 0$, i.e. $v_k \in C_{\bar{u}}$. From (3.18) and (3.15) we conclude for those $0 < t < k^{-2}$ that

$$\frac{2}{t^2}[j_D(\bar{u} + tv_k) - j_D(\bar{u}) - tj_D'(\bar{u}, v_k)] = \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_1)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot))^2 \mathrm{d}\rho_1(x_1).$$

Finally, we take $0 < t_k < k^{-2}$ and conclude due to $v_k \to v$ strongly in $L^2(\Lambda)$:

$$\begin{aligned}\lim_{k\to\infty} \frac{2}{t_k^2}[j_D(\bar{u} + t_k v_k) - j_D(\bar{u}) - t_k j_D'(\bar{u}, v_k)] \\ = \lim_{k\to\infty} \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_1)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot))^2 \mathrm{d}\rho_1(x_1) \\ = \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_1)}(\bar{u}(x_1, \cdot), v(x_1, \cdot))^2 \mathrm{d}\rho_1(x_1) = j_D''(\bar{u}, v^2).\end{aligned}$$

Hence, we have verified Assumption 3.5.2a. Next, let v_k, t_k be as in Assumption 3.5.2b. First, recall from the proof of [**54**, Lemma 4.7] that

$$\|\bar{u}(x_1, \cdot) + t_k v_k(x_1, \cdot)\|_{L^1(\Lambda_2)} \geq \left| \|\bar{u}(x_1, \cdot)\|_{L^1(\Lambda_2)} + t_k \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot)) \right|$$

which implies

$$\begin{aligned}&\frac{1}{2} \int_{\Lambda_1} \|\bar{u}(x_1, \cdot) + t_k v_k(x_1, \cdot)\|^2_{L^1(\Lambda_2)} \mathrm{d}\rho_1(x_1) \\ &\geq \frac{1}{2} \int_{\Lambda_1} \left(\|\bar{u}(x_1, \cdot)\|_{L^1(\Lambda_2)} + t_k \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot)) \right)^2 \mathrm{d}\rho_1(x_1) \\ &= t_k \int_{\Lambda_1} \|\bar{u}(x_1, \cdot)\|_{L^1(\Lambda_2)} \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot)) \mathrm{d}\rho_1(x_1) \\ &\qquad + \frac{t_k^2}{2} \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot))^2 \mathrm{d}\rho_1(x_1) \\ &\qquad\qquad + \frac{1}{2} \int_{\Lambda_1} \|\bar{u}(x_1, \cdot)\|^2_{L^1(\Lambda_2)} \mathrm{d}\rho_1(x_1)\end{aligned}$$

and hence

$$\begin{aligned}(3.19)\quad &\frac{2}{t_k^2}[j_D(\bar{u} + t_k v_k) - j_D(\bar{u}) - t_k j_D'(\bar{u}, v_k)] \\ &\qquad \geq \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot))^2 \mathrm{d}\rho_1(x_1).\end{aligned}$$

Along the lines of the proof of [**54**, Lemma 4.6] we obtain that $v_k \rightharpoonup v$, $v \in C_{\bar{u}}$, and $j_D'(\bar{u}, v_k) \to j_D'(\bar{u}, v)$ imply

$$\|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v_k(x_1, \cdot)) \rightharpoonup \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v(x_1, \cdot)) \qquad \text{weakly in } L^2(\Lambda_1).$$

Thus, we conclude from (3.19):

$$\begin{aligned}\liminf_{k\to\infty} \frac{2}{t_k^2}[j_D(\bar{u} + t_k v_k) - j_D(\bar{u}) - t_k j_D'(\bar{u}, v_k)] \\ \geq \int_{\Lambda_1} \|\cdot\|'_{L^1(\Lambda_2)}(\bar{u}(x_1, \cdot), v(x_1, \cdot))^2 \mathrm{d}\rho_1(x_1) = j_D''(\bar{u}, v^2),\end{aligned}$$

		(Λ_1, ρ_1)	(Λ_2, ρ_2)	x_1	x_2
$k=1$	case A	$(I, \mathrm{d}t)$	$(\{\bullet\}^m, \text{counting measure})$	t	i
$k=2$	case B	$(\{\bullet\}^m, \text{counting measure})$	$= (I, \mathrm{d}t)$	i	t
$k=3$	case B	$(I, \mathrm{d}t)$	$(\{\bullet\}^m, \text{counting measure})$	t	i
$k=4$	case C	$(\{\bullet\}^m, \text{counting measure})$	$(I, \mathrm{d}t)$	i	t
$k=5$	case C	$(I, \mathrm{d}t)$	$(\{\bullet\}^m, \text{counting measure})$	t	i
$k=6$	case D	$(\{\bullet\}^m, \text{counting measure})$	$(I, \mathrm{d}t)$	i	t
$k=7$	case D	$(I, \mathrm{d}t)$	$(\{\bullet\}^m, \text{counting measure})$	t	i

Table 3.1. Reduction of functionals j_k, $k = 1, ..., 7$, to the four generic cases A-D from Section 3.2.2. By $\{\bullet\}^m$ we denote the m-element set.

i.e. we have verified Assumption 3.5.2b. □

3.3. Proofs of the main results

Finally, we can prove Theorems 3.3 and 3.4 for $(\mathrm{P}_k^{\mathrm{sp}})$ from Section 3.1. As in [**169**, Section 3] they are obtained by application of the abstract results of Sections 3.2.1 and 3.2.2 to

$$f(u) := \underbrace{\frac{1}{2}\|S(u) - y_d\|^2_{L^2(Q)}}_{=:h(u)} + \frac{\alpha}{2}\|u\|^2_{L^2(I,\mathbb{R}^m)}, \quad g(u) := \beta j_k(u),$$

$$\hat{J}(u) = f(u) + g(u).$$

Here, recall that S denotes the control-to-state map introduced in Section 1.3. Before proving Theorems 3.3 and 3.4 we summarize the required auxiliary results on f and g. That f satisfies Assumption 3.5.1 with $U_\infty = L^s(I, \mathbb{R}^m)$, $U_2 = L^2(I, \mathbb{R}^m)$ and $K = U_{\mathrm{ad}}$ has been proven in [**35**] and summarized in Proposition 1.21. The respective properties for g are discussed in the next subsection.

3.3.1. Auxiliary results for the nonsmooth part g of the functional $\hat{J}$. Regarding the nonsmooth part g of the functional, it suffices to observe that the seven possibilities for j_k given in the introduction can be reduced to the four generic cases A-D from Section 3.2.2. The respective choices of Λ_1, Λ_2 and the identification of the variables x_1, x_2 (notation of Section 3.2.2) with the indices i and time t are summarized in Table 3.1. The details have been explained on behalf of $k = 2$ at the beginning of Section 3.2.2.

Therefore, we can translate the results for j_X, $X \in \{A, B, C, D\}$ from Section 3.2.2 back to j_k for $k = 1, ..., 7$. For reference, we state concrete formulas for the directional derivatives, the subgradients, and the surrogates for the second derivative.

Subgradients.

$$\partial j_1(u) = \left\{\lambda \in L^2(I, \mathbb{R}^m)\colon \lambda_i(t) \in \operatorname{sign}(u_i(t)) \text{ for a.a. } t \in I,\ i = 1, ..., m\right\}, \tag{3.20}$$

$$\partial j_2(u) = \left\{ \lambda \in L^2(I, \mathbb{R}^m)\colon \lambda_i \begin{cases} \in \overline{\mathbb{B}_1^{L^2(I)}(0)}, & \text{if } u_i \equiv 0, \\ = \frac{u_i}{|(\|u_i\|_{L^2(I)})_i|_1}, & \text{else.} \end{cases} \right\}, \tag{3.21}$$

$$\partial j_3(u) = \left\{ \lambda \in L^2(I,\mathbb{R}^m)\colon \lambda(t) \begin{cases} \in \overline{B_1(0)}, & \text{if } u(t) = 0, \\ = \frac{u(t)}{\|\|u\|_2\|_{L^1(I)}}, & \text{else.} \end{cases} \right\}, \tag{3.22}$$

$$\partial j_4(u) = \left\{ \lambda \in L^2(I,\mathbb{R}^m)\colon \lambda_i(t) \in \operatorname{sign}(u_i(t)) \frac{\|u_i\|_{L^1(I)}}{|(\|u_i\|_{L^1(I)})_i|_2} \right\}, \tag{3.23}$$

$$\partial j_5(u) = \left\{ \lambda \in L^2(I,\mathbb{R}^m)\colon \lambda_i(t) \in \operatorname{sign}(u_i(t)) \frac{|u(t)|_1}{\||u|_1\|_{L^2(I)}} \right\}, \tag{3.24}$$

$$\partial j_6(u) = \left\{ \lambda \in L^2(I,\mathbb{R}^m)\colon \lambda_i(t) \in \operatorname{sign}(u_i(t)) \|u_i\|_{L^1(I)} \right\}, \tag{3.25}$$

$$\partial j_7(u) = \left\{ \lambda \in L^2(I,\mathbb{R}^m)\colon \lambda_i(t) \in \operatorname{sign}(u_i(t)) |u(t)|_1 \right\}. \tag{3.26}$$

Directional derivatives.

$$j_1'(u,v) = \sum_{i=1}^m \|\cdot\|_{L^1(I)}'(u_i,v_i), \tag{3.27}$$

$$j_2'(u,v) = \sum_{\{i\colon \|u_i\|_{L^2(I)}=0\}} \|v_i\|_{L^2(I)} + \sum_{\{i\colon \|u_i\|_{L^2(I)}\neq 0\}} \int_I \frac{u_i(t)}{\|u_i\|_{L^2(I)}} v_i(t) \mathrm{d}t, \tag{3.28}$$

$$j_3'(u,v) = \int_{\{t\colon u(t)=0\}} |v(t)|_2 \mathrm{d}t + \int_{\{t\colon u(t)\neq 0\}} \frac{u(t)^T}{|u(t)|_2} v(t) \mathrm{d}t, \tag{3.29}$$

$$j_4'(u,v) = \sum_{i=1}^m \|\cdot\|_{L^1(I)}'(u_i,v_i) \frac{\|u_i\|_{L^1(I)}}{|(\|u_i\|_{L^1(I)})_i|_2}, \tag{3.30}$$

$$j_5'(u,v) = \int_I |\cdot|_1'(u(t),v(t)) \frac{|u(t)|_1}{\||u|_1\|_{L^2(I)}} \mathrm{d}t, \tag{3.31}$$

$$j_6'(u,v) = \sum_{i=1}^m \|\cdot\|_{L^1(I)}'(u_i,v_i) \|u_i\|_{L^1(I)}, \tag{3.32}$$

$$j_7'(u,v) = \int_I |\cdot|_1'(u(t),v(t)) |u(t)|_1 \mathrm{d}t. \tag{3.33}$$

Surrogates for the second derivatives. It holds

$$j_1''(u,v^2) = 0, \tag{3.34}$$

$$j_2''(u,v^2) = \sum_{\{i\colon u_i\neq 0\}} \frac{1}{\|u_i\|_{L^2(I)}} \left[\|v_i\|_{L^2(I)}^2 - \left(\int_I \frac{u_i(t) v_i(t)}{\|u_i\|_{L^2(I)}} \mathrm{d}t \right)^2 \right], \tag{3.35}$$

$$j_3''(u,v^2) = \int_{\{t\colon u(t)\neq 0\}} \frac{1}{|u(t)|_2} \left[|v(t)|_2^2 - \left(\frac{u(t)^T v(t)}{|u(t)|_2} \right)^2 \right], \tag{3.36}$$

$$j_4''(u,v^2) = \frac{1}{j_4(u)} \left(\sum_{i=1}^m \|\cdot\|_{L^1(I)}'(u_i,v_i) - j_4'(u,v)^2 \right), \tag{3.37}$$

$$j_5''(u,v^2) = \frac{1}{j_5(u)} \left(\int_I |\cdot|_1'(u(t),v(t)) \mathrm{d}t - j_5'(u,v)^2 \right), \tag{3.38}$$

$$j_6''(u,v^2) = \sum_{i=1}^m \|\cdot\|_{L^1(I)}'(u_i,v_i)^2, \tag{3.39}$$

$$j_7''(u,v^2) = \int_I |\cdot|_1'(u(t),v(t))^2 \mathrm{d}t, \tag{3.40}$$

for $u \neq 0$ and $j_k(0,v^2) = 0$ for all $v \in L^2(\Lambda)$ and $k = 1,...,7$.

3.3.2. Main results: Proofs of Theorems 3.3 and 3.4. We follow Section 3.3 of our preprint [**169**]. First, we prove Theorem 3.3 that states first-order optimality conditions and the resulting sparsity patterns of the optimal control. As already pointed out, we obtain this result by straightforward application of Propositions 3.10, 3.11, 3.12, and 3.13.

Proof of Theorem 3.3. Each $(\mathrm{P}_k^{\mathrm{sp}})$, $k \in \{1,...,7\}$, can be understood as realization of (LP_X) for some $X \in \{A,B,C,D\}$; cf. Section 3.3.1. First-order optimality conditions and sparsity patterns for the latter have been obtained in Section 3.2.2. The formula for the gradient of the smooth part of the functional has been stated in (1.6) and Proposition 1.17. □

Next, we prove Theorem 3.4 on second-order optimality conditions. Again, the proof is short because the main work has already been done in Theorems 3.8 and 3.9, and the verification of the corresponding assumptions in Sections 1.5 and 3.2.2.

Proof of Theorem 3.4. Apply Theorems 3.8 and 3.9 to $U_\infty = L^s(I,\mathbb{R}^m)$, $U_2 = L^2(I,\mathbb{R}^m)$, f, g, and h specified above, and $K = U_{\mathrm{ad}}$. Assumption 3.5.1 and 2 with $D_{\bar u}$ replaced by $C_{\bar u}$ have been verified in Proposition 1.21 and Section 3.2.2. The additional requirement $\nabla h(\bar u) = B^*\bar p \in L^\infty(I,\mathbb{R}^m)$ in the case B follows from Proposition 1.17. □

Here, L^∞-boundedness of U_{ad} and $\alpha > 0$ are crucial for the verification of Assumption 3.5.1 as has been pointed out below Proposition 1.21. When omitting control-constraints, the results of this chapter need to be modified in roughly the same way as described in Section 6 of our paper [**168**] in the case of the state-constrained problem $(\mathrm{P}^{\mathrm{st}})$. Thereby, Remark 2.12 needs to be replaced by the observation formulated at the end of Section 3.2.1.

3.3.3. Limitations of the approach: The bang-bang case. We conclude the theoretical part of this chapter by the outlook to the bang-bang case $\alpha = 0$ from [**169**, Section 3.4] that illustrates the limits of our second-order analysis. In fact, the present approach cannot be extended to the bang-bang case. As we will see in the following, the main obstruction is not due to sparsity but due to the highly nonlinear structure of the underlying PDE-constraint (Eq).

Regarding necessary optimality conditions, a short computation shows that $C_{\bar u} = \{0\}$ holds for $\bar u$ satisfying the first-order optimality conditions of (P_1). Hence, the first part of the statement of Theorem 3.8 is still true, but trivial for $\alpha = 0$. For sufficient optimality conditions, Assumption 3.5.1biii is crucial. It is well-known that this property for the smooth part of the functional can only be expected in the case of $\alpha > 0$, or a similar so-called Legendre-Clebsh condition; cf., e.g., formulas (3.3) and (5.3) in [**71**]. As explained at the end of Section 3.2.1, it seems impossible to avoid this assumption on f by exploiting properties of g. Recall from Section 1.3

that in the case $\alpha = 0$ the second derivative of the smooth part of the functional reads as follows:

$$f''(\bar{u})v^2 = \int_Q \left([1 - \xi''(\bar{y})\nabla\bar{p} \cdot \mu\nabla y] z_v^2 - 2\xi'(\bar{y}) z_v \nabla\bar{p} \cdot \mu\nabla z_v \right) \mathrm{d}x\mathrm{d}t, \ v \in L^2(I, \mathbb{R}^m),$$

with $\bar{y} = S(\bar{u})$, $z_v := S'(\bar{u})v$ and $\bar{p}$ as in Lemma 1.12 and Theorem 3.3. Assuming appropriate higher regularity for $\bar{p}$ this can be transformed into

$$f''(\bar{u})v^2 = \int_Q (1 - \xi'(\bar{y})\nabla \cdot \mu\nabla\bar{p}) z_v^2 \mathrm{d}x\mathrm{d}t.$$

The approach in [**54, 59**] for the bang-bang case is based on a second-order sufficient optimality condition of the type

$$f''(\bar{u})v^2 + \beta j_k(\bar{u}, v^2) \geq c\|z_v\|^2_{L^2(Q)}$$

with some $c > 0$ and all directions v from a certain cone. For such a condition to hold in our setting, we expect that $1 - \xi'(\bar{y})\nabla \cdot \mu\nabla\bar{p} \in L^\infty(Q)$, and therefore $\nabla \cdot \mu\nabla\bar{p} \in L^\infty(Q)$ has to hold, which is a very strong assumption. Moreover, to follow the arguments of [**54, 59**] we would need certain continuity properties like

$$|f''(u)v^2 - f''(\bar{u})v^2| \lesssim \|u - \bar{u}\|_{L^s(I,\mathbb{R}^m)} \|z_v\|^2_{L^2(Q)}.$$

Consequently, $\nabla \cdot \mu\nabla p \in L^\infty(Q)$ would be required to depend continuously on u, which is out of reach, even in a highly smooth setup and with controls measured in $L^\infty(I, \mathbb{R}^m)$ instead of $L^s(I, \mathbb{R}^m)$.

On the other hand, to apply the approach for L^1-penalized semilinear elliptic bang-bang problems from [**286**], we would have to guarantee that there is a bounded bilinear extension $f''(\bar{u})$: $\mathcal{M}(I, \mathbb{R}^m) \times \mathcal{M}(I, \mathbb{R}^m) \to \mathbb{R}$, to the space $\mathcal{M}(I, \mathbb{R}^m)$ of $\mathbb{R}^m$-valued regular Borel measures on I. Moreover, appropriate higher regularity for the adjoint state would be needed. This is, of course, more delicate for parabolic problems than for elliptic ones and the additional problems when switching from semilinear to quasilinear problems have been highlighted at the end of Section 2.3.3. Moreover, it is not clear whether the results obtained in [**286**] for L^1-penalization also hold for directional sparsity functionals. Nevertheless, we point out that the aforementioned obstructions do not originate from sparsity, but primarily from the the structure of f'' that is due to the underlying quasilinear parabolic equation. Hence, we would be confronted with roughly the same problems when aiming at a generalization of the second-order conditions from [**35, 45**] that have been summarized in Theorem 1.22 towards the bang-bang case, even without additional sparsity functional.

This shows that it is by no means obvious that techniques successfully applied to semilinear parabolic or semilinear elliptic problems can be transferred to the quasilinear parabolic case. We have to leave this as an interesting open problem.

3.4. Numerical illustration

This section closely follows [**169**, Section 4]. We provide the announced details regarding the numerical example used for the illustration of different sparsity patterns in the introduction of this chapter. This will be done in Section 3.4.2. Before doing so, we give a concise overview over the fast proximal method in Section 3.4.1, that is used to solve $(\mathrm{P}_k^{\mathrm{sp}})$ for $k = 1\text{-}3, 6, 7$. For $k = 4, 5$ we apply a subgradient

method to ($\mathrm{P}_k^{\mathrm{sp}}$). As expected, the fast proximal method turns out to be much faster than subgradient descent in our numerical experiments. This illustrates that j_6 and j_7 may be avantageous compared to j_4 and j_5 from a numerical perspective.

3.4.1. Proximity operators and fast proximal methods. Proximal algorithms, see, e.g., [**26, 225**], have been applied successfully in different areas, e.g., image processing, and machine learning, but also in PDE-constrained optimization of elliptic and parabolic PDEs with L^1-penalization [**249, 250**]. For further references we refer to the introductions of [**249, 250**].

This class of algorithms specifically applies to problems of type (LP_X), cf. Section 3.2.2, consisting of a nonconvex, but smooth, and a convex, but nonsmooth summand in the functional. In the context of sparse optimal control we are aware of possibly faster methods, e.g., certain Newton-type methods in function space [**261, 138, 228, 209**], or algorithms on the discrete level, e.g., [**48**, Section 6] and [**138**, Section 4]. However, proximal methods usually have the advantage that they are relatively easy to implement and, compared to second-order methods, less intrusive. Let us recall from, e.g., [**249**, Algorithm 2] the basic concept of the so-called fast proximal method, formulated on behalf of (LP_X). Given a fixed step size $L > 0$, an initial guess u^0, and $t^0 = 1$, set $v^0 = u^0$ and for $\ell = 1, 2, 3, \dots$

$$u^\ell = \mathrm{Prox}_{\frac{\beta}{L}}\left(v^{\ell-1} - \frac{1}{L}\nabla f(v^{\ell-1})\right),$$
$$t^\ell = \frac{1}{2}\left(1 + \sqrt{1 + 4(t^{\ell-1})^2}\right), \qquad v^\ell = u^\ell + \frac{t^{\ell-1} - 1}{t^\ell}(u^\ell - u^{\ell-1}),$$

until the current iterate u^ℓ reaches a desired optimality criterion. Here, we denote by

$$\mathrm{Prox}_\tau(v) := \mathrm{argmin}_{u \in U_{\mathrm{ad}}}\left(\frac{1}{2}\|u - v\|^2_{L^2(\Lambda)} + \tau j_X(u)\right) \tag{3.41}$$

the so-called proximity operator; see, e.g., [**225**] or [**24**, Chapter 24] for an overview. It is a crucial condition for the applicability of proximal algorithms to (LP_X), that the nonsmooth part of the functional, j_X, is "proximable", i.e. we have to know how to compute (3.41) efficiently. We briefly address this issue using the notation of Section 3.2.2. In case A, it is well-known, see, e.g., [**249**, Lemma 4.3] or [**228**, Section 3.3.2], that j_A is proximable with

$$[\mathrm{Prox}_\tau(v)](x) = \begin{cases} \min(v(x) - \tau, u_{\mathrm{b}}(x)) & \text{if } v(x) > \tau, \\ 0 & \text{if } |v(x)| \le \tau, \\ \max(v(x) + \tau, u_{\mathrm{a}}(x)) & \text{if } v(x) < -\tau, \end{cases} \qquad \text{for a.a. } x \in \Lambda.$$

For case B and $U_{\mathrm{ad}} = L^2(\Lambda)$ it holds

$$[\mathrm{Prox}_\tau(v)](x_1, x_2) = \max\left(0, 1 - \frac{\tau}{\|v(x_1, \cdot)\|_{L^2(\Lambda_2)}}\right) v(x_1, x_2), \qquad \text{for a.a. } x \in \Lambda;$$

cf., e.g., [**228**, Section 3.3.2]. We refer the reader to [**225**, Section 6.5.4] or [**187**, Theorem 3] for the same formula in the discrete case. For the proximity operator in the case B with $U_{\mathrm{ad}} := \{u \in L^2(\Lambda)\colon u \ge 0 \text{ a.e.}\}$, we refer the reader to [**228**, Section 3.3.2]. We are not aware of an explicit formula for the proximity operator in the case of bilateral box-constraints. To the best of our knowledge, the functional

of case C is not proximable. For case D and $U_{\mathrm{ad}} = L^2(\Lambda)$, however, a formula for the proximity operator is well-known for the discrete analogue, see, e.g., [**187**, Theorem 3], and the adaptation to our setting is not difficult. We obtain:

$$[\mathrm{Prox}_\tau(v)](x_1,x_2) = \mathrm{sign}(v(x_1,x_2))\max(0,|v(x_1,x_2)| - \sigma(x_1)) \quad \text{for a.a. } x \in \Lambda,$$

where $\sigma(x_1) \in \mathbb{R}$ has to satisfy for every $x_1 \in \Lambda_1$:

$$\|\max(0,|v(x_1,\cdot)| - \sigma(x_1))\|_{L^1(\Lambda_2)} = \frac{\sigma(x_1)}{2\tau}. \tag{3.42}$$

The efficient computation of proximity operators for case C, as well as for cases B and D in the presence of box-constraints, is certainly of interest, but beyond the scope of this thesis. Also, we do not address the convergence analysis of the fast proximal method in our precise setting.

3.4.2. Results. We consider the following specification of $(\mathrm{P}_k^{\mathrm{sp}})$: we choose $\Omega = B_1(0) \subset \mathbb{R}^2$, $T = 8$, $\mu \equiv I_2 \in \mathbb{R}^{2\times 2}$, $\Gamma_D = \partial\Omega$, $\alpha = \frac{\pi}{25}\cdot 10^{-2}$, $\beta = 10^{-2}$, $y_0 \equiv 0$, $\xi(s) := \frac{1}{2} + \frac{1}{1+\exp(-20s)}$, and $y_d(t,x) := -\frac{t}{4}\sin(\frac{\pi}{2}t)\exp(-36|x-m(t)|_2^2)$ with $m(t) = \frac{2}{3}(\sin(\frac{\pi}{4}t),\cos(\frac{\pi}{4}t))^T$. The eight control actuators are given by $\langle b_i,\varphi\rangle := \int_\Omega \mathbf{1}_{B_{\frac{1}{5}}(m(i-1))}(x)\varphi(x)\mathrm{d}x$, $i = 1,\dots,8$. Consequently, the state equation reads as follows:

$$\begin{aligned}
\partial_t y(t,x) - \nabla\cdot\left(\frac{1}{2} + \frac{1}{1+\exp(-20y(t,x))}\right)\nabla y(t,x) &= \sum_{i=1}^{8} u_i(t)\cdot \mathbf{1}_{B_{\frac{1}{5}}(m(i-1))}(x), \\
&\qquad \text{on } [0,8]\times B_1(0), \\
y(t,x) &= 0, \qquad \text{on } [0,8]\times\partial B_1(0), \\
y(0,x) &\equiv 0, \qquad \text{on } B_1(0).
\end{aligned}$$

We omit control-constraints and set $U_{\mathrm{ad}} = L^s(I,\mathbb{R}^m)$.

Space is discretized with the help of FEniCS and mshr [**9, 203**] using piecewise linear finite elements with 3324 DoF and mesh size $h_{\max} \approx 5.0\cdot 10^{-2}$. For time discretization we use the implicit Euler scheme with 160 equidistant time steps; more precisely: state and adjoint state are discretized piecewise linear in space and piecewise constant in time, while the controls —whose "spatial" components are vectors in $\mathbb{R}^m$ in our setting and, consequently, do not need to be discretized— are $\mathbb{R}^m$-valued piecewise constant in time. Hence, the chosen discretization coincides with the variational discretization concept [**149**]. Moreover, discretization and optimization commute ("optimize then discretize = discretize then optimize"); in particular, the discretization of the subgradient of j_k yields the subgradient of the discretization of j_k. Note that introducing constant (w.r.t. time) control bounds $u_a, u_b \in \mathbb{R}^m$ would not change this sitation. However, we point out that dealing with distributed controls (instead of purely time-dependent ones) would be more intricate because then the spatial component of the controls requires discretization (or variational discretization) as well. For issues related to the handling of the control-variable in the discretization of sparse optimal control problems and corresponding error estimates we refer the reader to the overview in [**228**, Section 4.5.3], or to, e.g., [**61, 62**] for semilinear parabolic or [**53, 52**] for semilinear elliptic problems.

For $k = 1\text{-}3, 6, 7$, i.e. cases A, B, and D, we solve the discretized counterpart of $(\mathrm{P}_k^{\mathrm{sp}})$ by the fast proximal method described above with step size $L = 10^{-2}$. In the case D, equation (3.42) is solved by bisection. Regarding the computation of the proximity operators note that commutativity of optimization and discretization pointed out above applies to the minimization problem (3.41) defining the proximity operator vice versa. Since for the case C we are not aware of a proximity operator, we solve the optimal control problem for $k = 4, 5$ by a classical subgradient descent method, cf. [**208**, Chapter II.2.1.2] for instance, with the step size in iteration ℓ given by $\frac{10}{\sqrt{\ell}}$. The initial guess in all cases is $u_i^0 \equiv 0.1$ and the nonlinear problem at each time step of the solution of the discretized state equation is solved by the built-in nonlinear solver provided by FEniCS.

For the solutions of $(\mathrm{P}_k^{\mathrm{sp}})$ we refer to Figures 3.1 to 3.4 at the beginning of the chapter. The different sparsity patterns described below Theorem 3.3 are clearly visible. In Figure 3.5 we illustrate the convergence speed of the fast proximal and the subgradient method. We display the $L^2(I, \mathbb{R}^m)$-norm of the residuals r^ℓ of the control iterates u^ℓ, i.e.

$$r^\ell := u^\ell + \alpha^{-1}(p^\ell + \beta\lambda^\ell),$$

where p^ℓ and λ^ℓ denote the adjoint state and the subgradient of j_k associated with u^ℓ; cf. the optimality conditions in Theorem 3.3. As expected, the fast proximal method converges much faster than the subgradient method. In our opinion, this indicates that replacing functionals j_4, j_5 by j_6, j_7, respectively, may be worth considering in applications in order to allow the application of fast proximal methods.

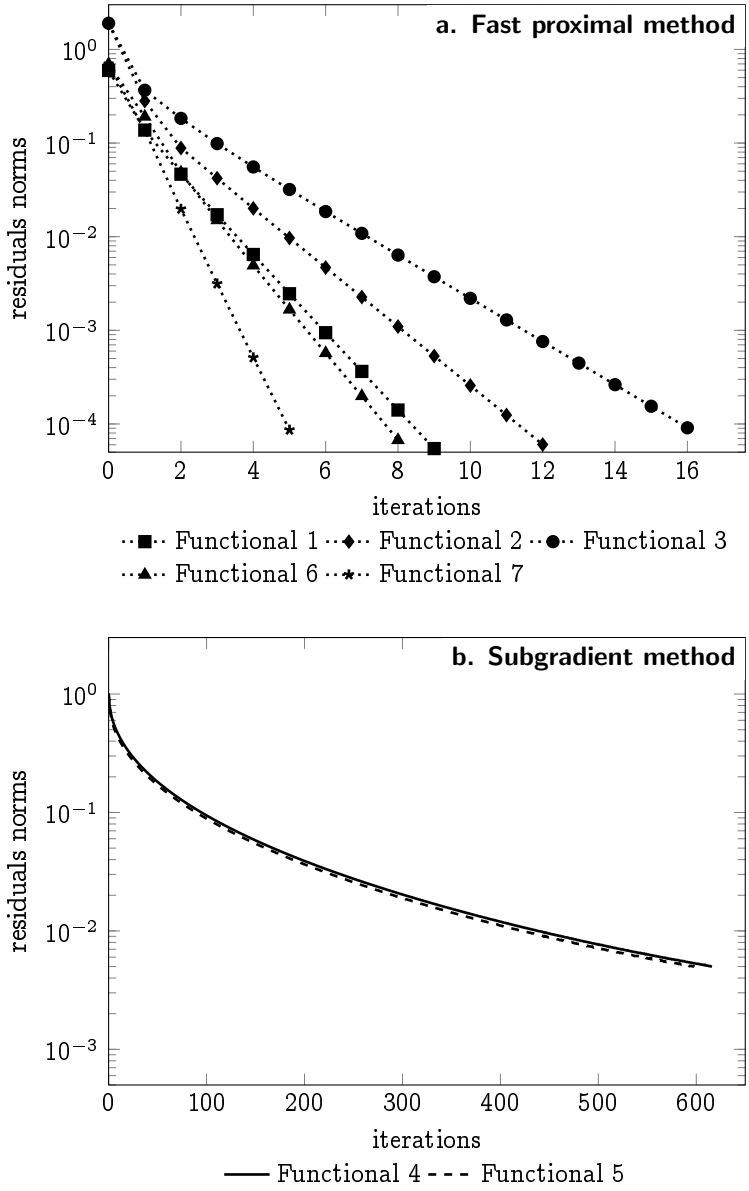


Figure 3.5. Convergence behaviour of **a.** the fast proximal method and **b.** the subgradient method.

Part II

Algorithms and model order reduction

Chapter 4

Convergence of the SQP method

In our numerical examples for state-constrained or sparse optimal control in the previous chapters we used standard computational schemes without addressing their theoretical analysis. In the present chapter based on the authors paper [**167**] (joint work with I. Neitzel) we now turn to such a type of question and analyze the convergence of the so-called SQP method for the control-constrained problem (P) from Chapter 1 in function space.

Sequential quadratic programming (SQP) methods form a prominent class of state of the art algorithms for the efficient numerical solution of nonlinear optimal control problems. Instead of solving the nonlinear control problem directly, a sequence of linear quadratic optimal control problems, i.e. problems with quadratic objective functional and a linear PDE-constraint, is solved iteratively. The idea behind this approach is that, in general, such linear quadratic problems are easier to solve than the original problem with a nonlinear PDE-constraint, e.g., by application of the well-understood primal dual active set strategy [**31, 145**]. The analysis of SQP methods for nonlinear optimal control problems has been addressed in several publications; see, e.g., [**268, 114**] for semilinear parabolic equations, [**143, 152, 285**] for optimal control of the time-dependent Navier-Stokes equations, [**125, 124**] for semilinear elliptic problems with mixed control-state-constraints, [**271, 281**] for optimal control of Burgers equation, and [**135**] for optimal control of a phase field equation. For an overview concerning the origins of SQP methods in the context of PDE-constrained optimization we also refer to the introduction of [**114**]. As further second-order methods for the solution of nonlinear optimal control problems we mention the Newton method, the semismooth Newton method, and versions of the primal dual active set strategy; see, e.g., [**152, 153, 126, 148, 177, 275**].

In this chapter we present our convergence analysis of the SQP method for optimal control problems governed by quasilinear parabolic PDEs in function space from [**167**]. As the most closely contributions from the literature we mention [**276, 293, 294**] and [**108**, Chapter 8], respectively. The papers [**276, 293, 294**] deal with trust-region and trust-region SQP methods for optimal control of general nonlinear PDE. The main difference to our work is that discretization is included and convergence of adaptive multilevel algorithms is proven whereas we stick to the undiscretized function space setting. In return, we are able to prove locally superlinear convergence around local minima fulfilling certain second-order conditions avoiding the two-norm gap, whereas the three aforementioned publications establish global convergence to a point fulfilling first-order optimality conditions but

without explicit rate. In [**108**] optimal control of the so-called chemotaxis system, a quasilinear system of parabolic equations, is considered. Besides the difference in the structure of the underlying PDE-constraint another important difference to our work is the following: convergence of the SQP method in [**108**] is proven assuming a rather strong second-order sufficient condition. This corresponds to our interim result in Section 4.4.1, whereas our main focus for the rest of this chapter is on the interplay of weaker second-order conditions and the notation of locality in the SQP method. We will discuss this in more detail below.

To conclude our literature overview, let us briefly mention some topics related to SQP methods that we do not address in this chapter and that have not been cited so far. First, there are globalization strategies, e.g., by linesearch w.r.t. certain value functionals [**141, 142**]. Also, there are slightly different concepts for the update of the adjoint state-variable; see, e.g., [**142, 280, 245**]. Moreover, the SQP ansatz can be combined with so-called Augmented Lagrangian approaches; cf., e.g., [**175, 245**] and the references therein. Finally, some results on mesh-independence have been obtained in [**279, 280**]. For literature on the combination of SQP methods with model order reduction we refer the reader to the introduction of the next chapter. In Section 5.3 we will also provide numerical experiments concerned with the combination of an SQP-type method and model order reduction by proper orthogonal decomposition.

Let us now put the results presented in this chapter into context. Many of our arguments in [**167**] are similar to those known from other, earlier publications. However, we believe that our analysis is of interest for particularly two reasons.

First, we demonstrate that the results on optimal control of quasilinear parabolic PDEs obtained in [**35**] and summarized in Chapter 1 allow to derive convergence of the SQP method. As we have seen in the previous chapters of this thesis, dealing with quasilinear problems is quite different from semilinear ones and sometimes much more involved. This makes the choice of the correct function spaces in the analysis of the SQP method particularly complicated and we believe that it is not clear a-priori that —in the end— the arguments from literature concerned with, e.g., semilinear parabolic PDEs apply to the present model problem as well.

Second, most results addressing convergence of the SQP method have been published before the introduction of a framework for second-order sufficient conditions without two-norm gap in [**71**]. As pointed out in the introduction on behalf of Example (0.2) or in Section 1.5.3, issues related with the so-called two-norm discrepancy and two-norm gaps in second-order optimality conditions are typical in PDE-constrained optimization. They accompanied us through all chapters of this thesis so far. In Section 1.5.3 we have explained that it has been shown in [**35**] that our model problem (P) fits into the framework of [**71**] and hence — unlike in earlier publications on second-order conditions for nonlinear parabolic problems— second-order conditions without two-norm gap can be formulated; cf. Theorem 1.22. In the convergence analysis of SQP methods, SSCs are known to play a central role, too. Consequently, it is natural to revisit the convergence theory of the SQP method in view of these new second-order result without two-norm gap. Let us explain this in a bit more detail: unlike in finite dimensions, it has been observed that the set of admissible controls of the linear quadratic subproblems of the SQP method needs to be restricted in the infinite dimensional setting

for proving convergence. Hereby, the restriction is related to the SSCs assumed to hold at the (local) solution of (P) to which we want to prove convergence of the SQP method. Roughly speaking, assuming stronger SSCs allows to weaken the restriction of the subproblems and the other way round: strongest possible second-order sufficient conditions, i.e.

$$j''(\bar{u})v^2 \geq c\|v\|^2_{L^2(\Lambda)} \qquad \forall v \in L^2(\Lambda)$$

with some $c > 0$, e.g., allow to avoid any restriction of the subproblems as we will see in Section 4.4.1. However, such conditions can hardly be expected to hold, in general, since there is a huge gap to the corresponding necessary optimality condition that are guaranteed to hold at $\bar{u}$; cf. Theorem 1.22. Weaker SSCs (their concrete type will be introduced lateron in Section 4.1.2), however, were known to require restriction of the quadratic subproblems: either the controls have to be fixed on certain parts of the active set of $\bar{u}$ or the subproblems need to be restricted to L^∞-balls around $\bar{u}$; see, e.g., [**114, 268**] for semilinear parabolic problems. Such a restriction to L^∞-balls around $\bar{u}$ has some similarity with the conclusion that could be drawn at that time from these SSCs w.r.t. the original nonlinear problem, i.e. L^2-quadratic growth of the reduced functional L^∞-locally around $\bar{u}$; cf., e.g., [**270**, Theorems 5.17 and 5.18]. Meanwhile, new results on SSCs following [**71**] allow to avoid this two-norm gap for certain problem types, i.e. the same SSCs actually imply quadratic growth of the reduced objective functional L^2-locally (instead of L^∞-locally) around the optimal control. As proven in [**35**] this is also the case for the quasilinear parabolic problem (P); cf. Theorem 1.22. Therefore, one may wonder, whether it is possible to replace L^∞-neighbourhoods from previous convergence proofs for the SQP method by L^2-neighbourhoods. The answer of this question is not straightforward due to the fact that convergence of the SQP method is established by showing convergence of a generalized Newton method for a certain generalized (set-valued) equation. In order to obtain a differentiable map in this generalized equation we still need to measure controls in a norm stronger than the L^2-norm; this is roughly the same issue that has been pointed out below Lemma 1.12. In contrast, the regularity property (analogous to the invertibility of the Hessian in the classical Newton method) relies on the L^2-coercivity property due to the second-order sufficient conditions. Thus, the presence of two norms in the arguments cannot be avoided, in general, and we need, again, to adapt some ideas from [**71**] to give an answer to this question in Section 4.4.3, which is our main result.

The chapter is organized as follows and keeps the main structure of many contributions concerning the analysis of SQP methods; cf. in particular [**268, 285, 114, 167**]. In Section 4.1 we introduce the precise problem setting under consideration and outline the idea of the SQP method together with appropriate second-order sufficient conditions. To prepare the analysis of the convergence properties of the SQP method, we provide some auxiliary results that are specifically related to our quasilinear parabolic model problem in Section 4.2. After that, we follow the standard argument to prove convergence of the SQP method in Sections 4.3 and 4.4. We utilize the connection to the Josephy-Newton method for a generalized equation originating from the first-order optimality conditions. Convergence

of this Newton method is proven in Section 4.3 and the interpretation of the iterates as the solutions of certain linear quadratic optimization problems is topic of Section 4.4. Assuming strongest possible second-order sufficient conditions we formulate our first main result in Section 4.4.1. The remaining two theoretical Sections 4.4.2 and 4.4.3 of this chapter are devoted to the analysis of the generalized Newton and the SQP method under weaker second-order assumptions. In particular we are able to replace the L^∞-neighbourhoods in the results of [**268, 285**] by L^2-neighbourhoods in our final result in Section 4.4.3. For a detailed overview of this part of the chapter we refer to the introduction of Section 4.4. Finally, we give short numerical examples that illustrate our theoretical findings in Section 4.5.

4.1. Optimality conditions and the SQP method

In this section we briefly state our assumptions and explain the basic idea of the SQP method. We roughly follow Section 3 of our paper [**167**]. First, we formulate the first-order necessary optimality conditions for (P) from Theorem 1.19 as a generalized equation and observe that application of a Newton-type method to this generalized equation results in a system of equations that can be interpreted as first-order necessary optimality condition for a certain linear quadratic optimization problem. To identify solutions of the equation with minimizers of the linear quadratic problem we need to impose certain second-order conditions on the original problem that ensure convexity of the linear quadratic problem.

Let us start with the precise problem formulation. The model problem for this chapter is an instance of the control-constrained problem (P) as introduced in Chapter 1. We will rely on Assumptions 1.5, 1.6 and 1.10, i.e. we utilize the improved regularity assumptions on Bessel potential spaces from [**35**]. Moreover, we have to restrict ourselves to the purely time-dependent control setting, i.e. we need to enforce Assumption 1.10 in the following way.

Assumption 4.1. Let ζ, s, y_d, y_0, and α be chosen as in Assumption 1.10. In addition, it holds $L^s(\Lambda) = L^s(I, \mathbb{R}^m)$, the control operator is given by

$$B\colon L^s(I,\mathbb{R}^m) \to L^s(I, H^{-\zeta,p}_{\Gamma_D}), \quad u \mapsto \sum_{i=1}^m u_i b_i,$$

where $b_i \in H^{-\zeta,p}_{\Gamma_D}$, $i = 1, \dots, m$, are fixed actuator functions, and the set of admissible controls is given by

$$U_{\mathrm{ad}} := \{u \in L^s(I,\mathbb{R}^m)\colon u_{\mathrm{a},i} \le u_i \le u_{\mathrm{b},i} \text{ a.e. on } I,\ i = 1, \dots, m\},$$

with control bounds $u_{\mathrm{a}}, u_{\mathrm{b}} \in L^\infty(I, \mathbb{R}^m)$, $u_{\mathrm{a},i} \le u_{\mathrm{b},i}$ a.e. on I for $i = 1, \dots, m$.

The reason for restricting ourselves to the Bessel potential space setting in this chapter is that in Section 4.1.1 below we are going to formulate the adjoint equation as an equation in the $W^{1,p'}_{\Gamma_D}$-$W^{1,p'}_{\Gamma_D}$-setting as in Proposition 1.15. The latter result relies on improved regularity for the states from Theorem 1.14 proven in [**35**] and hence requires the chosen Bessel potential space setting. A more detailed explanation, in particular regarding the restriction to purely time-dependent controls, will be provided lateron in Section 4.3.1 below Theorem 4.11.

From Section 1.3 recall the definition of $\mathcal{A}'$ and $\mathcal{A}''$. Moreover, let us introduce some additional notation that will be useful in the context of the following analysis:

For reasons of shortness we will sometimes write the state equation (Eq) as

$$e(y,u) := (\partial_t y + \mathcal{A}(y)y - Bu, \quad \operatorname{tr}_0 y - y_0) = 0 \tag{4.1}$$

with the twice Fréchet differentiable map

$$e\colon \mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})) \times L^s(\Lambda) \to L^s(I,W^{-1,p}_{\Gamma_D}) \times (W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})_{1/s',s}.$$

By $\mathcal{L}_{\mathrm{P}}(y,u,p) := J(y,u) - \langle p, e_1(y,u)\rangle$ we denote the Lagrangian of (P).

4.1.1. First-order optimality conditions as generalized equation. Utilizing the the analysis of the adjoint equation in the $W^{-1,p'}_{\Gamma_D}$-$W^{1,p'}_{\Gamma_D}$-setting summarized in Proposition 1.15, we can write the first-order necessary optimality conditions for (P) from Theorem 1.19 equivalently as the following generalized equation:

$$0 \in F(y,u,p) + N(y,u,p) \tag{GE}$$

with the maps

$$F(y,u,p) := \begin{pmatrix} \partial_t y + \mathcal{A}(y)y - Bu \\ \operatorname{tr}_0 y - y_0 \\ -\partial_t p + \mathcal{A}(y)^* p + \mathcal{A}'(y)^* p - (y - y_d) \\ \operatorname{tr}_T p \\ \alpha u + B^* p \end{pmatrix}$$

and $$N(y,u,p) := \big(\{0\},\{0\},\{0\},\{0\},N_{U_{\mathrm{ad}}}(u)\big)^T,$$

where $N_{U_{\mathrm{ad}}}(u)$ denotes the normal cone of the closed convex set U_{ad} at the point $u \in L^s(\Lambda)$, i.e. $N_{U_{\mathrm{ad}}}(u) = \big\{v \in L^s(\Lambda)\colon \langle v, w-u\rangle_{L^2(\Lambda)} \le 0 \text{ for all } w \in U_{\mathrm{ad}}\big\}$. To make the definition of F and N precise, F is understood as map $F\colon X_s \to Z_s$ with

$$X_s := \mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})) \times L^s(\Lambda) \times \mathbb{W}^{1,s}(I,(W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D}))$$

and

$$\begin{aligned} Z_s := L^s(I,W^{-1,p}_{\Gamma_D}) &\times (W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})_{1/s',s} \times L^s(I,W^{-1,p'}_{\Gamma_D}) \\ &\times (W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D})_{1/s',s} \times L^s(\Lambda). \end{aligned}$$

Accordingly, N is understood as set-valued map $X_s \rightrightarrows Z_s$. We equip X_s and Z_s with the canonical norms

$$\begin{aligned} \|(y,u,p)\|_{X_s} &:= \|y\|_{\mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D}))} + \|u\|_{L^s(\Lambda)} + \|p\|_{\mathbb{W}^{1,s}(I,(W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D}))}, \\ \|(f,y_0,g,p_T,r)\|_{Z_s} &:= \|f\|_{L^s(I,W^{-1,p}_{\Gamma_D})} + \|y_0\|_{(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})_{1/s',s}} + \|g\|_{L^s(I,W^{-1,p'}_{\Gamma_D})} \\ &\quad + \|p_T\|_{(W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D})_{1/s',s}} + \|r\|_{L^s(\Lambda)}. \end{aligned}$$

Having chosen these spaces, the following result holds that retrospectively also motivates the chosen setting.

Lemma 4.2 ([**167**], Lemma 3.1). *$F\colon X_s \to Z_s$ is continuously Fréchet differentiable, $N\colon X_s \rightrightarrows Z_s$ has closed graph.*

Proof. Differentiability has been used implicitely by [**35**, Lemma 4.5] where the differentiability of the control to state map is shown by the implicit function theorem. The closed graph property is standard. □

For the reason why (GE) needs to be defined with L^s-integrable control functions u instead of L^2-integrable controls we refer the reader to our comment below Lemma 1.12. As explained there, we require time integrability $s \gg 2$ as in Assumption 1.8 in order to ensure differentiability of the superposition operators associated with ξ and ξ', and hence differentiability of F.

Sometimes we will need the following subspaces X_∞ and Z_∞ of X_s, Z_s:

$$\begin{aligned} X_\infty &:= \mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})) \times L^\infty(\Lambda) \times \mathbb{W}^{1,s}(I,(W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D})),\\ Z_\infty &:= L^s(I,W^{-1,p}_{\Gamma_D}) \times (W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D})_{1/s',s} \times L^s(I,W^{-1,p'}_{\Gamma_D})\\ &\qquad \times (W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D})_{1/s',s} \times L^\infty(\Lambda), \end{aligned}$$

equipped with the canonical norms similarly as above. Note that changing from X_s, Z_s to X_∞, Z_∞ means nothing more than replacing the $L^s(\Lambda)$-factors by $L^\infty(\Lambda)$-factors, i.e. considering controls in the L^∞- instead of the L^s-norm. The same result as before holds.

Lemma 4.3 ([**167**], Lemma 3.3)**.** *F: $X_\infty \to Z_\infty$ is continuously Fréchet differentiable and N: $X_\infty \rightrightarrows Z_\infty$ has closed graph.*

Due to Lemma 4.2 we can formulate the ansatz of the SQP method in its abstract form as the Josephy-Newton method for generalized equations; see [**180, 100, 10**], or [**156**, chapter 2]. Given a current iterate $(y_k,u_k,p_k) \in X_s$, solve

$$0 \in F(y_k,u_k,p_k) + F'(y_k,u_k,p_k)(y-y_k,u-u_k,p-p_k) + N(y,u,p) \tag{4.2}$$

to obtain the new iterate $(y_{k+1},u_{k+1},p_{k+1}) \in X_s$. Writing down the full system of equations for (4.2) we find:

$$\begin{cases} \partial_t y + \mathcal{A}(y_k)y + \mathcal{A}'(y_k)y = Bu + \mathcal{A}'(y_k)y_k, \\ \operatorname{tr}_0 y = y_0, \end{cases} \tag{4.3}$$

$$\begin{cases} -\partial_t p + \mathcal{A}(y_k)^*p + \mathcal{A}'(y_k)^*p = y - y_d - \mathcal{A}''(y_k)[y-y_k,\cdot]^*p_k, \\ \operatorname{tr}_T p = 0, \end{cases} \tag{4.4}$$

$$0 \in \alpha u + B^*p + N_{U_{\mathrm{ad}}}(u). \tag{4.5}$$

Obviously, the current control-iterate u_k has canceled out, which implies that the next iterate (y_k,u_k,p_k) depends on y_k and p_k but not on u_k. This is due to the structure of our model problem. Note that the first two equations (4.3) are equivalent to the linearized state equation

$$0 = e(y_k,u_k) + e_y(y_k,u_k)(y-y_k) + e_u(y_k,u_k)(u-u_k). \tag{4.6}$$

A standard computation shows that

$$\frac{1}{2}\mathcal{L}''_{\mathrm{P}}(y_k,u_k,p_k)[(y-y_k,u-u_k)]^2 + J'(y_k,u_k)(y-y_k,u-u_k) \tag{4.7}$$

is equal (up to addition of constants) to the expression

$$J_k(y,u) := \frac{1}{2}\|y-y_d\|^2 + \frac{\alpha}{2}\|u\|^2 - \frac{1}{2}\langle p_k, \mathcal{A}''(y_k)[y-y_k,y-y_k]\rangle \tag{4.8}$$

that finally fulfills: the system of equations (4.3)-(4.5) is the formal optimality system of the optimal control problem

$$\text{(QP)} \qquad \begin{cases} \min\limits_{y,u} J_k(y,u) \\ \text{s.t.} \quad u \in U_{\text{ad}} \quad \text{and (4.3).} \end{cases}$$

This is the classical formulation of the SQP method as sequence of quadratic problems to solve. Note that our computations were completely formal in the sense that we do not know whether (QP) is convex or not. Hence, we cannot say whether there is a unique minimizer of (QP) or whether the optimality system (4.3)-(4.5) is a sufficient characterization for this minimizer. This issue will be addressed in the following section utilizing the assumption of second-order sufficient conditions.

4.1.2. Second-order sufficient conditions and the SQP method. Depending on second-order sufficient conditions (SSCs) for (P) we have to restrict the admissible set for (QP) to ensure convexity. For technical reasons it is not possible to work with SSCs on the critical cone as considered in Theorem 1.22; this will be explained in more detail below. Instead, the set of directions w.r.t. which we demand coercivity of $j''(\bar u)$ needs to be enlarged. In fact, SSCs related to the so-called strongly active sets turned out to be suitable assumptions for the analysis of SQP methods; see, e.g., [**268, 114, 285**], which work with the same assumption as we do. More precisely, we need to state the following assumptions on the locally optimal control $\bar u$ which we want to find with help of the SQP method.

Assumption 4.4. From now on let $\bar u \in U_{\text{ad}}$ be a fixed L^2-local minimizer for (P), i.e. there is $r > 0$ such that

$$u \in U_{\text{ad}} \ \text{ and } \ \|u - \bar u\|_{L^2(I,\mathbb{R}^m)} < r \qquad \Longrightarrow \qquad j(u) \geq j(\bar u).$$

Let $\bar y$ and $\bar p$ be the state and adjoint state associated with $\bar u$; cf. Theorem 1.19. For $\sigma \in [0,\infty]$ we define the σ-active set of $\bar u$ as

$$A^\sigma(\bar u) := \{x \in \Lambda\colon\ |\alpha \bar u + B^* \bar p|(x) > \sigma\}$$

and the corresponding subspace

$$C^\sigma(\bar u) := \{v \in L^2(\Lambda)\colon\ v = 0 \text{ on } A^\sigma(\bar u)\}$$

of directions vanishing on $A^\sigma(\bar u)$. We assume that the following second-order sufficient condition for (P) is satisfied at $\bar u$: there is a fixed $\sigma \in [0,\infty]$ (whether we allow the case $\sigma = 0$ or not will be stated in our further results) such that there exists $\delta > 0$ such that

$$\text{(SSC-}\sigma\text{)} \qquad \begin{cases} \mathcal{L}''_{\text{P}}(\bar y, \bar u, \bar p)[(y,u)]^2 \geq \delta \|u\|^2_{L^2(\Lambda)} \\ \text{for all} \quad (y,u) \in \mathbb{W}^{1,2}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D})) \times L^2(\Lambda) \quad \text{s.t.} \\ \qquad u \in C^\sigma(\bar u), \\ \qquad e_y(\bar y, \bar u) y + e_u(\bar y, \bar u) u = 0. \end{cases}$$

Let us put this assumption into context. First, note that it holds

$$\mathcal{L}''_{\text{P}}(\bar y, \bar u, \bar p)[(y,u)]^2 = j''(\bar u) u^2$$

for y and u as in (SSC-σ), which follows from a short computation utilizing the definition of the Lagrangian. Consequently, one may express (SSC-σ) equivalently as

$$j''(\bar{u})v^2 \geq \delta \|v\|^2_{L^2(\Lambda)} \qquad \forall v \in C^\sigma(\bar{u}).$$

The formulation in terms of the second derivative of the Lagrangian, however, is more convenient in the context of SQP methods. Due to $C_{\bar{u}} \subset C^\sigma(\bar{u})$ for every $\sigma \geq 0$ it is obvious that condition (SSC-σ) is stronger than the second-order sufficient condition (1.20); cf. Theorem 1.22 and Proposition 2.9. Consequently, (SSC-σ) implies strict local optimality of $\bar{u}$ together with the local quadratic growth condition (1.19) on the reduced functional. Of course, the gap of (SSC-σ) w.r.t. the necessary second-order optimality condition from Theorem 1.22 is larger. In particular, local quadratic growth of the reduced functional does not allow to deduce (SSC-σ), in general.

That it is not possible to work directly with the SSCs from Theorem 1.22 established in [**35**], that are minimal in the sense that they have minimal gap to the corresponding SNCs, has two technical, but important reasons: first, we require the coercivity condition in (SSC-σ) to hold on a vector space instead of just a cone in the proof of the L^2-stability result in Section 4.3.1. Second, in Section 4.4.2 we will make use of the fact that strongly active sets behave well under small perturbations for $\sigma > 0$.

Given $\sigma \in [0, \infty]$ that will always become clear from the context we introduce the modified admissible set as

$$U^\sigma_{\mathrm{ad}} := U_{\mathrm{ad}} \cap (\bar{u} + C^\sigma(\bar{u})) = \{u \in U_{\mathrm{ad}} \colon u = \bar{u} \text{ on } A^\sigma(\bar{u})\} \tag{4.9}$$

and define the corresponding restricted quadratic problem as follows:

$$\text{(QP-}\sigma\text{)} \qquad \begin{cases} \min\limits_{y,u} J_k(y,u) \\ \text{s.t.} \quad u \in U^\sigma_{\mathrm{ad}} \quad \text{and (4.3).} \end{cases}$$

Using the relation of J_k to the second derivative of the Lagrangian of (P) (see (4.7) and (4.8)) it is clear that (QP-σ) is a linear quadratic and, under Assumption 4.4, strictly coercive control problem, at least for $(y_k, u_k, p_k) = (\bar{y}, \bar{u}, \bar{p})$. Therefore, it is strictly convex in this case which will be crucial for the convergence analysis of the SQP method.

In general, increasing σ means imposing a stronger SSC and results in larger U^σ_{ad}, i.e. in less restriction of U_{ad}. As already explained in the introduction of this chapter, strongest possible second-order conditions, i.e. coercivity of $\mathcal{L}''_{\mathrm{P}}$ on the whole space $L^2(\Lambda)$ form a special case since they allow to avoid technical restrictions of the SQP subproblems. We will refer to this by $\sigma = \infty$. In that case it holds $C^\infty(\bar{u}) = L^2(\Lambda)$ and $U^\infty_{\mathrm{ad}} = U_{\mathrm{ad}}$; see, e.g., [**125, 124, 108, 135**] for such an assumption in the context of SQP methods. In Section 4.4.1 we state our main theorem for this special case: local superlinear convergence of the SQP method can be shown with the quadratic subproblems given by (QP).

To avoid confusion, let us emphasize that these strongest possible SSCs are indeed quite strong but not unreasonably strong: if $j''(\bar{u})$ is assumed to be coercive w.r.t. all directions, one may imagine j to behave locally at $\bar{u}$ like a strictly convex function. The following example shows that this does *not* imply that minimality of

$\bar{u}$ is trivial in the sense that $\bar{u}$ is also a local minimizer of j without the constraint $u \in U_{\mathrm{ad}}$:

$$\left\{ \begin{array}{c} \min\limits_{u \in L^2(0,1)} j(u) := \displaystyle\int_0^1 \left(u(t) + \frac{1}{2}u(t)^2 + \frac{1}{3}u(t)^3 \right) \mathrm{d}t \\ \text{s.t. } 0 \le u(t) \le 1 \text{ a.e. on } (0,1). \end{array} \right. \tag{4.10}$$

Indeed, $\bar{u} \equiv 0$ is global the solution of (4.10), but not even a local minimizer of j if the pointwise inequality-constraint is omitted. Nevertheless, j is twice continuously Fréchet differentiable as map $L^\infty(0,1) \to \mathbb{R}$ and it holds $j''(\bar{u})v^2 = \|v\|^2_{L^2(0,1)}$ for all $v \in L^2(0,1)$, i.e. strongest possible SSCs hold at $\bar{u}$.

During our analysis of the SQP method, a restriction of U_{ad} in the formulation of the quadratic subproblems cannot be avoided when relying on weaker SSCs, i.e. such with $\sigma \in [0, \infty)$. Nevertheless, it will be the topic of the remaining part of Section 4.4 to keep this as small and as natural as possible.

4.2. Auxiliary results

Before going into the details of the convergence analysis for the SQP method we collect some auxiliary results in the following section. We follow [**167**, Section 4].

More precisely, we analyze a property of B^*, provide detailed estimates for the $\mathcal{A}''$-term under two different assumptions, and prove stability properties for the gradient and the Hessian of the reduced functional associated with the quadratic problems (QP).

4.2.1. A property of the control operator. Recall from Assumption 1.8 the definition of the control operator in the case of purely time-dependent controls. Obviously, B is continuous from $L^2(\Lambda)$ to $L^2(I, W_{\Gamma_D}^{-1,p})$ and therefore its adjoint B^* is defined on $L^2(I, W_{\Gamma_D}^{1,p'})$ with values in $L^2(\Lambda)$. To derive the L^∞-stability result from the L^2-stability result in Section 4.3.1 we need to perform a bootstrapping argument that requires us to know how B^* behaves restricted to a space of more regular functions.

To simplify notation, let B: $L^s(I, \mathbb{R}) \to L^s(I, H^{-\zeta,p})$ be defined by $u \mapsto u \cdot b_1$ with only a single fixed control function $b_1 \in H_{\Gamma_D}^{-\zeta,p}$. Of course, this yields

$$(B^*v)(t) = \langle b_1, v(t) \rangle_{W_{\Gamma_D}^{-1,p}, W_{\Gamma_D}^{1,p'}} \qquad \text{for every } v \in L^2(I, W_{\Gamma_D}^{1,p'}).$$

It is obvious that B maps $L^r(\Lambda)$ into $L^r(I, H_{\Gamma_D}^{-\zeta,p})$ for $r \in [2, \infty]$. To obtain $B^*v \in L^q(\Lambda)$, we have to ensure that $v \in L^q(I, H_{\Gamma_D}^{\zeta,p'})$ holds. As earlier in this thesis, the argument relies on Bochner-Sobolev embeddings. We therefore need the following lemma on the appearing interpolation spaces.

Lemma 4.5 ([**167**], Lemma 4.2). *It holds*

$$(W_{\Gamma_D}^{-1,q}, W_{\Gamma_D}^{1,q})_{\theta,1} \hookrightarrow H_{\Gamma_D}^{2\theta-1,q}$$

for $0 < \theta < 1$ *and* $q \in (1, \infty)$ *as long as* $2\theta - 1 \notin \{1/q, -1/q'\}$.

Proof. This is a direct consequence of [**123**, Theorem 3.5]. □

Now, set $\theta := (\zeta+1)/2$. For $r \in (1,\infty)$ there are two possibilities: if $\theta < 1-1/r$, then it holds for $0 \le \rho < 1 - 1/r - \theta$:

$$\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'})) \hookrightarrow C^{0,\rho}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'})_{\theta,1}) \hookrightarrow C^{0,\rho}(I,H_{\Gamma_D}^{\zeta,p'}),$$

i.e. B^* is continuous from $\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'}))$ to $L^\infty(\Lambda)$. Otherwise, if $\theta > 1-1/r$, then we obtain $q \ge 1$ such that $1/q > \theta - (1-1/r) > 0$ and

$$\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'})) \hookrightarrow L^q(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'})_{\theta,1}) \hookrightarrow L^q(I,H_{\Gamma_D}^{\zeta,p'}),$$

which means that B^* maps $\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'}))$ to $L^q(\Lambda)$. For the two embeddings we refer, e.g., to Proposition 1.1. We will come back to this in Section 4.3.1: given an estimate on the control in $L^r(\Lambda)$, we have estimates for linearized state and adjoint state in $\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))$ and $\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'}))$, respectively. Application of B^* either yields an estimate for the control in L^q with some $q > r$ or in L^∞ if r already was large enough.

4.2.2. Estimates for $\mathcal{A}''$. Recall the definition of $\mathcal{A}''$ from Section 1.3. In Section 2.3.3 we have pointed out that this term originating from the second derivative of the nonlinearity causes significant additional difficulties compared to the semilinear case. Our analysis of the SQP method requires a careful estimation of this term as well.

We start with the following result that we will require during the proof of the L^2- and L^∞-stability results in Section 4.3.1.

Lemma 4.6 ([**167**], Lemma 4.3)**.** *It holds*

$$\|\mathcal{A}''(y)[v,\cdot]^*p\|_{L^r(I,W_{\Gamma_D}^{-1,p'})} \le C(\xi,\mu,y)\|p\|_{L^\infty(I,W_{\Gamma_D}^{1,p'})}\|y\|_{L^\infty(I,W_{\Gamma_D}^{1,p})}\|v\|_{L^r(I,W_{\Gamma_D}^{1,p})}.$$

The constant C can be chosen uniformly with respect to y for y's coming from a bounded subset of $\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))$.

Proof. Estimate $\langle\mathcal{A}''(y)[v,\cdot]^*p,w\rangle = \langle\mathcal{A}''(y)[v,w],p\rangle$ for an arbitrary testfunction $w \in L^{r'}(I,W_{\Gamma_D}^{1,p})$ utilizing Hölder's inequality. □

In Lemma 4.6 we bound the norm of $\mathcal{A}''(\bar y)[v,\cdot]^*\bar p$ in the space $L^r(I,W_{\Gamma_D}^{-1,p'})$ against the norm of v in the space $\mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{1,p},W_{\Gamma_D}^{1,p}))$ for each $r \in [2,s]$ by estimating $\langle\mathcal{A}''(y)[v,w],p\rangle$ with arguments $v \in L^r(I,W_{\Gamma_D}^{1,p})$ and $w \in L^{r'}(I,W_{\Gamma_D}^{1,p})$. This generality will be necessary in the bootstrapping argument in the proof of the L^∞-stability, which was already mentioned in the previous Section 4.2.1. As explained in the remark after Lemma 4.7, this requires bounds for y in $L^\infty(I,W_{\Gamma_D}^{1,p})$ and p in $L^\infty(I,W^{1,p'})$.

In the next section, however, we will require an estimate of $\langle\mathcal{A}''(y)[v,w],p\rangle$ directly (and not of $\mathcal{A}(y)''[v,\cdot]^*p$) which allows us to use the arguments v and w from the space $\mathbb{W}^{1,2}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))$ in Lemma 4.7. In that case we can exploit more regularity of v and w, which allows to relax the assumptions on y and p.

Lemma 4.7 ([**167**], Lemma 4.4)**.** *It holds*

$$|\langle\mathcal{A}''(y)[v,w],p\rangle| \le C(\xi,\mu,y)\|y\|_{L^s(I,W_{\Gamma_D}^{1,p})}\|p\|_{L^s(I,W_{\Gamma_D}^{1,p'})}$$

$$\cdot \|v\|_{\mathbb{W}^{1,2}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))}\|w\|_{\mathbb{W}^{1,2}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))}.$$

The constant C can be chosen uniformly with respect to y for y's coming from a bounded subset of $\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))$.

Proof. The proof works similar as for Lemma 4.6, but now we try to exploit more regularity of v and w. Using the first embedding from Lemma 1.13 and the definition of s in Assumption 4.1 we find

$$\mathbb{W}^{1,2}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p})) \hookrightarrow_c L^q(I,L^\infty)$$

with some $q \in (2,\infty)$ satisfying $\frac{1}{q}+\frac{1}{s} \leq \frac{1}{2}$. Now, an application of Hölder's inequality (the temporal integrability exponents match due to the conditions on q and s) yields the desired result. The uniform choice of the constant with respect to y follows from the boundedness of ξ and its derivatives on bounded sets of $\mathbb{R}$ and the compactness of the embedding $\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p})) \hookrightarrow_c C(\overline{Q})$. □

As in [**167**, Remarks 4.5 and 4.6] let us emphasize that the difference in the regularities assumed for y and p in the two lemmas is essential. Lemma 4.6 will be applied in Section 4.3.1 for $y=\bar{y}$ and $p=\bar{p}$ only, i.e. the required regularity is guaranteed by Proposition 1.17 for $\bar{p}$ and Section 1.3 for $\bar{y}$. In Section 4.2.3 we will have to apply Lemma 4.7 for $y=y_k$, $p=p_k$ with y_k,p_k being iterates of the SQP method, i.e. y_k and p_k are solutions of the linearized state and adjoint equation. Hence, the regularity requirements for Lemma 4.7 are met, but not immediately those of Lemma 4.6. Moreover, we note that Lemma 4.6 cannot be improved. The limiting factor is the summand

$$\int_Q \xi'(y)w\mu\nabla p\nabla v \mathrm{d}x\mathrm{d}t,$$

which has to be estimated for $v \in \mathbb{W}^{1,r}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))$ and $w \in L^{r'}(I,W_{\Gamma_D}^{1,p})$, $r \in [2,s]$. The function w has $L^{r'}$-integrability in time and L^∞-integrability in space, whereas ∇v has L^r-integrability in time and L^p-integrability in space, which is the best we can expect from the assumptions each. This implies that we require $p \in L^\infty(I,W_{\Gamma_D}^{1,p'})$ in order to be able to estimate the above integral.

4.2.3. Derivatives associated to (QP)**.** In this section we provide results on the first and second derivatives of the reduced objective functionals associated to the quadratic subproblems (QP). We will apply them in Section 4.4.3 briefly before obtaining our main result.

Recall from Section 4.1.1 the definition of the space X_s. By $j_k\colon L^2(\Lambda)\to\mathbb{R}$ we denote the reduced functional associated with the linear quadratic optimal control problem (QP) at $(y_k,u_k,p_k)\in X_s$. In particular, note that j_k'' is constant because j_k is a quadratic functional, which makes us write j_k'' instead of $j_k''(v)$ for some v because $v\mapsto j_k''(v)[\cdot,\cdot]$ is constant and hence independent of such v.

Proposition 4.8 ([**167**], Proposition 4.7)**.** *Let Assumptions 1.5, 1.6, 4.1 and 4.4 be satisfied. Then, it holds uniformly in $u\in L^2(\Lambda)$*

$$|(j_k''-j''(\bar{u}))u^2| \lesssim \Big(\|y_k-\bar{y}\|_{\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p}))}$$

$$+ \|p_k - \bar{p}\|_{\mathbb{W}^{1,s}(I,(W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D}))}\Big) \|u\|^2_{L^2}$$

as $y_k \to \bar{y}$, $p_k \to \bar{p}$ *in the above norms.*

Proof. Recall by (4.7) that $j_k'' \cdot u^2 = \mathcal{L}_{\mathrm{P}}''(y_k, u_k, p_k)(y,u)^2$ with

$$e_y(y_k,u_k)y + e_u(y_k,u_k)u = 0 \tag{4.11}$$

holds. We expand this as

$$\begin{aligned}\mathcal{L}_{\mathrm{P}}''(y_k,u_k,p_k)(y,u)^2 = &\underbrace{\mathcal{L}_{\mathrm{P}}''(\bar{y},\bar{u},\bar{p})(\tilde{y},u)^2}_{=:(\mathrm{I})}\\ &\underbrace{-\left(\mathcal{L}_{\mathrm{P}}''(\bar{y},\bar{u},\bar{p})(\tilde{y},u)^2 - \mathcal{L}_{\mathrm{P}}''(\bar{y},\bar{u},\bar{p})(y,u)^2\right)}_{=:(\mathrm{II})}\\ &\underbrace{-\left(\mathcal{L}_{\mathrm{P}}''(\bar{y},\bar{u},\bar{p}) - \mathcal{L}_{\mathrm{P}}''(y_k,u_k,p_k)\right)(y,u)^2}_{=:(\mathrm{III})}\end{aligned} \tag{4.12}$$

with $\tilde{y} \in W^{1,2}(I, W^{-1,p}_{\Gamma_D}) \cap L^2(I, W^{1,p}_{\Gamma_D})$ defined by

$$e_y(\bar{y},\bar{u})\tilde{y} + e_u(\bar{y},\bar{u})u = 0. \tag{4.13}$$

From the definition of the Lagrangian we know $(\mathrm{I}) = j''(\bar{u})u^2$. Hence, it remains to show that the contribution of (II) and (III) gets uniformly small as claimed above. By definition we have

$$\begin{aligned}(\mathrm{II}) &= \underbrace{\|\tilde{y}\|^2 - \|y\|^2}_{=:(\mathrm{IIa})} - \underbrace{\langle \bar{p}, \mathcal{A}''(\bar{y})\tilde{y}^2 - \mathcal{A}''(\bar{y})y^2\rangle}_{=:(\mathrm{IIb})},\\ (\mathrm{III}) &= \langle p_k, \mathcal{A}''(y_k)y^2\rangle - \langle \bar{p}, \mathcal{A}''(\bar{y})y^2\rangle\\ &= \underbrace{\langle p_k - \bar{p}, \mathcal{A}''(y_k)y^2\rangle}_{=:(\mathrm{IIIa})} + \underbrace{\langle \bar{p}, (\mathcal{A}''(y_k) - \mathcal{A}''(\bar{y}))y^2\rangle}_{=:(\mathrm{IIIb})},\end{aligned}$$

wherein the summands

$$(\mathrm{IIa}) = \langle \tilde{y} + y, \tilde{y} - y\rangle \qquad \text{and} \qquad (\mathrm{IIb}) = \langle \bar{p}, \mathcal{A}''(\bar{y})[\tilde{y} + y, \tilde{y} - y]\rangle \tag{4.14}$$

can be estimated using the boundedness of the solution operator of the linearized state equation [**35**, Proposition 4.4] (see Lemma 1.12) and applying Lemma 4.7 and a similar argument as in the proof of Lemma 4.7; in particular recall our comments at the end of Section 4.2.2. In the same way one can treat (III) as well. □

For the gradient of j_k we obtain a similar result.

Proposition 4.9 ([**167**], Proposition 4.8)**.** *If* $(y_k, u_k, p_k) \to (\bar{y}, \bar{u}, \bar{p})$ *in* X_s, $v_k \to \bar{u}$ *in* L^s, *it holds*

$$\nabla j_k(v_k) \to \nabla j(\bar{u}) \quad \textit{strongly in } L^2(\Lambda).$$

Proof. We split

$$\nabla j_k(v_k) - \nabla j(\bar{u}) = \underbrace{\nabla j_k(v_k) - \nabla j(v_k)}_{=:(\mathrm{A})} + \underbrace{\nabla j(v_k) - \nabla j(\bar{u})}_{=:(\mathrm{B})}$$

and estimate both summands. For some $v \in U_{\mathrm{ad}}$, e.g., $v = v_k$, introducing the following quantities will be helpful:

$$\begin{aligned} y(v) &\quad \text{state associated to } v \text{ w.r.t. (P)},\\ p(v) &\quad \text{adjoint state associated to } v \text{ w.r.t. (P)},\\ y_k(v) &\quad \text{state associated to } v \text{ w.r.t. (QP)},\\ p_k(v) &\quad \text{adjoint state associated to } v \text{ w.r.t. (QP)}. \end{aligned}$$

Regarding (B) we know from [**35**, Proposition 4.9] that

$$\|\nabla j(v_k) - \nabla j(\bar{u})\|_{L^2(\Lambda)} \le \alpha \|v_k - \bar{u}\|_{L^2} + \|B^*(p(v_k) - p(\bar{u}))\|_{L^2} \to 0$$

holds as $v_k \to \bar{u}$ in L^s because the adjoint states $p(v_k)$ converge to $\bar{p}$ in $L^s(I, W^{1,p'}_{\Gamma_D})$. To estimate (A), first note that the states $y_k(v_k)$ of the quadratic problem converge to $\bar{y} = y(\bar{u})$ in $\mathbb{W}^{1,2}(I, (W^{-1,p}_{\Gamma_D}, W^{1,p}_{\Gamma_D}))$; this is shown using the convergence of the solution operators of the linearized state equation [**35**, Proposition 4.9] (see Lemma 1.12). Utilizing similar techniques as before the desired result follows after some straightforward computations. We omit the details. □

4.3. Generalized Newton method on U_{ad}^{σ}

This section closely follows [**167**, Section 5]. Following the standard arguments, see, e.g., [**269, 268, 114, 11, 125, 124, 285, 143**], we show that the Newton-Josephy method applied to a modified version of the generalized equation (GE), see Section 4.1.1, converges.

The main challenge here is to verify that —under the correct choice of spaces and with the help of suitable auxiliary results that have been achieved in the previous section— arguments known from the literature for different problem types apply to the quasilinear case as well. That this is nontrivial becomes apparent especially in the proof of Theorem 4.11: here, the available regularity results for the linearized state equation and the adjoint equation fit together exactly as needed; cf. also our remarks below this result. Proving convergence of the generalized Newton method is a central step toward showing convergence of the SQP method. The iterates of the generalized Newton method will be interpreted as iterates of the SQP method in Section 4.4.

From formula (4.9) in Section 4.1.2 recall the definition of the modified admissible set U_{ad}^{σ} for some $\sigma \in [0, \infty]$. We consider the generalized equation with this modified admissible set, i.e. we replace (GE) by

$$0 \in F(y, u, p) + N^{\sigma}(y, u, p), \tag{GE-σ}$$

where U_{ad} is replaced by U_{ad}^{σ} in the definition of the normal cone map N, i.e.

$$N^{\sigma}(y, u, p) := \left(\{0\}, \quad \{0\}, \quad \{0\}, \quad \{0\}, \quad N_{U_{\mathrm{ad}}^{\sigma}}(u)\right)^T,$$

where $N_{U_{\mathrm{ad}}^{\sigma}}(u)$ denotes the normal cone of U_{ad}^{σ} at u. The map F: $X_s \to Z_s$ as well as the spaces X_s, Z_s, see Section 4.1.1 for their definitions, do not change.

To prove convergence of the generalized Newton method, strong regularity in the sense of Robinson [**234**] (see also [**156**, Definition 2.5]) has to be shown at an optimal point $(\bar{y}, \bar{u}, \bar{p}) \in X_s$: there exist $\delta, \epsilon > 0$ such that for all perturbations $d \in$

Z_s, $\|d\|_{Z_s} < \delta$ there exists a unique solution $(y,u,p) \in X_s$, $\|(y,u,p)-(\bar{y},\bar{u},\bar{p})\|_{X_s} < \epsilon$ to the generalized equation

$$\text{(GE-}\sigma\text{-D)} \qquad d \in F(\bar{y},\bar{u},\bar{p}) + F'(\bar{y},\bar{u},\bar{p})(y-\bar{y},u-\bar{u},p-\bar{p}) + N^\sigma(y,u,p)$$

with Lipschitz continuous solution map $d \mapsto (y,u,p)$.

Translating back this generalized equation for (y,u,p) into an optimal control problem yields

$$\text{(QP-}\sigma\text{-D)} \qquad \begin{cases} \min\limits_{y,u} \dfrac{1}{2}\|y-y_d\|^2 + \dfrac{\alpha}{2}\|u\|^2 - \dfrac{1}{2}\langle \bar{p}, \mathcal{A}''(\bar{y})[y-\bar{y}]^2\rangle \\ \qquad\qquad + \langle d_T, \operatorname{tr}_T y\rangle - \langle d_u, u\rangle + \langle d_p, y\rangle \\ \text{s.t.} \quad u \in U_{\text{ad}}^\sigma \\ \text{and} \quad \begin{pmatrix} d_y \\ d_0 \end{pmatrix} = e_y(\bar{y},\bar{u})(y-\bar{y}) + e_u(\bar{y},\bar{u})(u-\bar{u}), \end{cases}$$

where $d = (d_y, d_0, d_p, d_T, d_u) \in Z_s$ is a given perturbation with components coming from the corresponding spaces. Note that (GE-σ-D) is indeed the first-order necessary and (due to convexity) sufficient optimality condition for (QP-σ-D) because (QP-σ-D) is convex since only linear perturbation terms have been added to the convex objective function from (QP-σ). The perturbation in the corresponding affine linear state equation is only a constant and does not destroy convexity as well.

4.3.1. Stability of the quadratic problems (QP-σ)**.** We fix $d_0 = 0$ and $d_T = 0$, i.e. we assume that initial and final conditions are met exactly during the application of the SQP method, which is reasonable from the numerical point of view.

Proposition 4.10 ([**167**], Proposition 5.1)**.** *Let Assumptions 1.5, 1.6, 4.1 and 4.4 with some $\sigma \in [0,\infty]$ hold. Denote with $(y^i,u^i,p^i) \in X_s$, $i=1,2$, the solution of* (QP-σ-D) *for arbitrary perturbation vectors $d^i \in Z_s$. Then it holds*

$$\|u^2-u^1\|_{L^2}^2 \lesssim \|d_u^2-d_u^1\|_{L^2}^2 + \|d_y^2-d_y^1\|_{L^2(I,W^{-1,p})}^2 + \|d_p^2-d_p^1\|_{L^2(I,W^{-1,p'})}^2.$$

The hidden constant depends on the data of (P) *and $(\bar{y},\bar{u},\bar{p})$, but not on d^i.*

To enhance clarity, we state the KKT system of the perturbed problems, that can easily be derived from (GE-σ-D) using (4.2) and (4.3)-(4.5), before starting the proof:

$$\text{(4.15)} \qquad \begin{cases} \partial_t y^i + \mathcal{A}(\bar{y})y^i + \mathcal{A}'(\bar{y})y^i = Bu^i + \mathcal{A}'(\bar{y})\bar{y} + d_y^i, \\ y^i(0) = y_0, \\ -\partial_t p^i + \mathcal{A}(\bar{y})^* p^i + \mathcal{A}'(\bar{y})^* p^i = y^i - y_d - \mathcal{A}''(\bar{y})[y^i-\bar{y},\cdot]^*\bar{p} + d_p^i, \\ p^i(T) = 0, \\ d_u^i \in \alpha u^i + B^* p^i + N_{U_{\text{ad}}^\sigma}(u^i). \end{cases}$$

In the following we use the short notation $\Delta_y := y^2 - y^1$, $\Delta_u := u^2 - u^1$, $\Delta_p := p^2 - p^1$ (and similarly for d_y, d_u, d_p). From (4.15) we derive:

$$\partial_t \Delta_y + \mathcal{A}(\bar{y})\Delta_y + \mathcal{A}'(\bar{y})\Delta_y = B\Delta_u + \Delta_{d_y}, \tag{4.16}$$

$$-\partial_t \Delta_p + \mathcal{A}(\bar{y})^*\Delta_p + \mathcal{A}'(\bar{y})^*\Delta_p = \Delta_y - \mathcal{A}''(\bar{y})[\Delta_y, \cdot]^*\bar{p} + \Delta_{d_p}, \tag{4.17}$$

with vanishing initial and final condition: $\Delta_y(0) = 0$ and $\Delta_p(T) = 0$.

Proof. The proof is similar to the one in [**114**] and relies on the linear quadratic structure of (QP-σ-D). Therefore, we omit the details. Let us just note that the regularity results Lemma 1.12 and Proposition 1.15 for the linearized state equation and the adjoint equation proven in [**35**] have to be applied and that terms containing $\mathcal{A}''$ are estimated with the help of Lemma 4.6. □

This shows L^2-stability of the quadratic problems (QP-σ) with respect to perturbations measured in corresponding norms. Utilizing a standard bootstrapping argument as, e.g., in [**269**] we can show the corresponding L^s- and L^∞-stability result.

Theorem 4.11 ([**167**], Theorem 5.2). *Let Assumptions 1.5, 1.6, 4.1 and 4.4 with some $\sigma \in [0, \infty]$ hold. Then, for the (y^i, u^i, p^i), $i = 1, 2$, from the previous proposition we have*

$$\|u^2 - u^1\|_{L^s} \lesssim \|d_u^2 - d_u^1\|_{L^s} + \|d_y^2 - d_y^1\|_{L^s(I, W^{-1,p})} + \|d_p^2 - d_p^1\|_{L^s(I, W^{-1,p'})},$$
$$\|u^2 - u^1\|_{L^\infty} \lesssim \|d_u^2 - d_u^1\|_{L^\infty} + \|d_y^2 - d_y^1\|_{L^s(I, W^{-1,p})} + \|d_p^2 - d_p^1\|_{L^s(I, W^{-1,p'})},$$

and

$$\|(y^1, u^1, p^1) - (y^2, u^2, p^2)\|_{X_s} \lesssim \|d^1 - d^2\|_{Z_s},$$
$$\|(y^1, u^1, p^1) - (y^2, u^2, p^2)\|_{X_\infty} \lesssim \|d^1 - d^2\|_{Z_\infty}.$$

In particular, the generalized equation (GE-σ) *is strongly regular in the sense of Robinson at its solution $(\bar{y}, \bar{u}, \bar{p})$ with respect to the spaces X_s, Z_s and X_∞, Z_∞.*

Proof. Again, the proof follows the techniques from [**114, 269**]. From the projection formula $u^i = \mathrm{Proj}_{U_{\mathrm{ad}}^{\sigma}}\left(-\frac{1}{\alpha}(B^*p^i - d_u^i)\right)$, $i = 1, 2$, we infer that

$$|\Delta_u| \le \frac{1}{\alpha}\left(|B^*\Delta_p| + |\Delta_{d_u}|\right)$$

holds pointwise on Λ. Thus, we can bound Δ_u in the $L^q(\Lambda)$-norm if we can bound $B^*\Delta_p$ and Δ_{d_u} in the $L^q(\Lambda)$-norm. We apply a bootstrapping argument that relies on the property of B^* from Section 4.2.1. Assume that we already know

$$\|\Delta_u\|_{L^r} \lesssim \|\Delta_{d_u}\|_{L^r} + \|\Delta_{d_y}\|_{L^r(I, W^{-1,p})} + \|\Delta_{d_p}\|_{L^r(I, W^{-1,p'})}$$

for some $r \in [2, s)$. Using the regularity theory of the linearized state equation and the adjoint equation for (4.16) and (4.17) we conclude

$$\|\Delta_p\|_{L^r(I, W_{\Gamma_D}^{-1,p'})} \lesssim \|\Delta_{d_u}\|_{L^r} + \|\Delta_{d_y}\|_{L^r(I, W^{-1,p})} + \|\Delta_{d_p}\|_{L^r(I, W^{-1,p'})};$$

hereby, we make use of Lemma 1.12 and Proposition 1.15 that have been established in [**35**]. At this point we need the full strength of Lemma 4.6 in order to estimate

the $\mathcal{A}''$-terms for different $r \in [2, s]$. Note that $\bar{p} \in L^\infty(I, W^{1,p'})$ holds due to Proposition 1.16. Our discussion of B^* from Section 4.2.1 shows that either

$$(\zeta + 1)/2 < 1 - 1/r, \text{ which implies } \|B^*\Delta_p\|_{L^\infty} \lesssim \|\Delta_p\|_{L^r(I,W^{-1,p'})}$$

or

$$(\zeta + 1)/2 > 1 - 1/r, \text{ which implies } \|B^*\Delta_p\|_{L^q} \lesssim \|\Delta_p\|_{L^r(I,W^{-1,p'})}$$

with some q fulfilling $1/q > 1/r + (\zeta - 1)/2$ holds. In the first case it follows

$$\|\Delta_u\|_{L^\infty} \lesssim \|\Delta_{d_u}\|_{L^\infty} + \|\Delta_{d_y}\|_{L^s(I,W^{-1,p}_{\Gamma_D})} + \|\Delta_{d_p}\|_{L^s(I,W^{-1,p'}_{\Gamma_D})}$$

and we are done. In the second case we have

$$\|\Delta_u\|_{L^q} \lesssim \|\Delta_{d_u}\|_{L^q} + \|\Delta_{d_y}\|_{L^q(I,W^{-1,p})} + \|\Delta_{d_p}\|_{L^q(I,W^{-1,p'})}$$

and we repeat the procedure with $r = q$ as long as the first case holds, which is clearly the case due to Assumption 4.1 if $r = s$ is reached. Note that $(\zeta - 1)/2 < 0$ is fixed and that we can avoid q being equal to the exceptional cases of Lemma 4.5 due to the strict inequality that allows small perturbations. □

Instead of using Proposition 1.16, one could also apply Proposition 1.17 in order to obtain sufficient regularity for $\bar{p}$ in the above proof. This does not make any difference because both results ensure that $\bar{p} \in C(\bar{I}, W^{1,p}_{\Gamma_D}) \hookrightarrow L^\infty(I, W^{1,p'}_{\Gamma_D})$, which suffices to apply Lemma 4.6.

Nevertheless, we have to point out that Assumption 4.1 cannot be relaxed in a straightforward way. We rely both on the purely time-dependent control setting and the Bessel potential space setting in the above proof of Theorem 4.11. As already explained in Section 4.1, the formulation of the adjoint equation of (P) as an equation in the $W^{-1,p'}_{\Gamma_D}$-$W^{1,p'}_{\Gamma_D}$-setting in Section 4.1.1 requires the Bessel potential space setting. In the above proof we utilize the specific mapping properties of B^* from Section 4.2.1 in combination with the regularity results for the linearized adjoint equation in the $W^{-1,p'}_{\Gamma_D}$-$W^{1,p'}_{\Gamma_D}$-setting and Lemma 4.6 on the $\mathcal{A}''$-term. In fact, this is the part of our analysis where the difficulties specifically associated to the underlying quasilinear parabolic equation become most visible. A generalization towards the full setting of Assumption 1.8 or Assumption 1.10 is not immediately clear. The main difficulty would lie in keeping the arguments for Proposition 4.10 and Theorem 4.11 working. Let us, e.g., consider the case of distributed control, i.e. $\Lambda = Q$ and $B = \mathrm{id}\colon L^s(Q) \to L^s(I, H^{-\zeta,p}_{\Gamma_D})$ in Assumption 1.10. In that case, B^* is also given by identity map and in order to to obtain $L^\infty(Q)$-estimates for Δ_u as in the proof of Theorem 4.11 we would need to consider the adjoint states in a space that embeds into $L^\infty(Q)$. The latter certainly excludes dealing with the adjoint states in the $W^{-1,p'}_{\Gamma_D}$-$W^{1,p'}_{\Gamma_D}$-setting applied so far. However, it is not possible just to choose a more regular function space for the adjoint states and to consider the adjoint equation in the setting of, e.g., Proposition 1.17 instead without additional modifications of the overal setting. This is due to the fact that in that case also an improved version of Lemma 4.6 and, consequently, improved regularity for the states would be needed. For the same reason it is also unclear whether dealing with the adjoint equation in the setting from Proposition 1.17 instead of Proposition 1.15 from the beginning is a reasonable alternative.

4.3.2. Convergence of the generalized Newton method. Invoking a general result on the convergence of generalized Newton methods, see, e.g., [**156**, Theorem 2.19], our previous results allow to derive the following theorem.

Theorem 4.12 ([**167**], Theorem 5.4). *Let Assumptions 1.5, 1.6, 4.1 and 4.4 with some $\sigma \in [0, \infty]$ hold.*

1. *Then there is a radius $r_{\mathrm{Newton}} > 0$ such that for any triple $(y_0, u_0, p_0) \in X_s$ fulfilling*
$$(y_0, u_0, p_0) \in \mathbb{B}^{X_s}_{r_{\mathrm{Newton}}}((\bar{y}, \bar{u}, \bar{p}))$$
the sequence of iterates generated by the Newton-Josephy method for equation (GE-σ) with (y_0, u_0, p_0) as start is well-defined, stays in the ball $\mathbb{B}^{X_s}_{r_{\mathrm{Newton}}}((\bar{y}, \bar{u}, \bar{p}))$, and converges q-superlinearly to $(\bar{y}, \bar{u}, \bar{p})$ in X_s.
2. *The same result as in part 1 holds with X_∞ instead of X_s.*

Proof. The proof is standard; see, e.g., [**268, 114, 285, 143, 124, 125**]. □

4.4. Convergence of the SQP method

So far, the well-definedness of the iterates in Theorem 4.12 is only ensured by some generalized implicit function theorem and the strong regularity of (GE-σ) at $(\bar{y}, \bar{u}, \bar{p})$. Convexity of the quadratic subproblems (QP-σ) is up to now only known in the case $(y_k, u_k, p_k) = (\bar{y}, \bar{u}, \bar{p})$, i.e. the relation of possible minimizers of (QP-σ) and solutions of (GE-σ) is unclear at the moment.

In this section we closely follow [**167**, Section 6], provide an extended analysis of the generalized Newton method for (GE), and interprete the Newton iterates as solutions of some linear quadratic optimal control problems. Roughly speaking, we can prove convergence of the SQP method with the subproblems being restricted either to U^σ_{ad} (see Section 4.1.2 for the definition) or to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}^{L^2}_\rho(\bar{u})}$ with some $\rho > 0$. Hereby, note that for theoretical reasons it is not possible to avoid such technical restrictions completely, even in finite dimensions; cf. the example given in [**114**, Section 6]. In the infinite dimensional case an additional difficulty arises as pointed out in [**268**, final Remark]: unlike in finite dimensions we cannot assume that the possibly infinite set of active constraints is correctly identified after the first iteration and therefore technical restrictions encoding some a-priori knowledge on the correct active set have to be imposed.

In order to make the flow of the argumentation more clear, we give a short summary of this section.

In a first step (Section 4.4.1) we consider the quadratic problems restricted to U^σ_{ad}, i.e. the set of those controls from U_{ad} that coincide with the optimal control $\bar{u}$ on the σ-active set of $\bar{u}$. The main argument here is that the quadratic problems sufficiently close to the true KKT triple get strictly convex when restricted to U^σ_{ad}. Hence, their unique solution is characterized by the corresponding first-order necessary optimality condition, which coincides with the generalized equation originating from the Newton method discussed in Section 4.3.

The assumption to restrict to U^σ_{ad} can be slightly relaxed in the case that (SSC-σ) holds for a positive σ: the quadratic subproblems have to be restricted to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}^{L^2}_\rho(\bar{u})}$ with some radius $\rho > 0$, as shown in Section 4.4.3, and the generalized Newton method for (GE) converges locally, even without further restrictions; see

Section 4.4.2. As we have pointed out in the introduction of this chapter, the fact that the restriction of the quadratic subproblems can be done in terms of L^2-balls around $\bar{u}$ distinguishes our results from [**167**] presented in this chapter from earlier results where restriction to L^∞-balls had been considered. Our results are obtained by by careful application of the SSCs. The main steps of the argument are as follows: first, we establish convergence of the generalized Newton method for the corresponding set-valued equation (GE) in Section 4.4.2, Theorem 4.20. Thereby, the proof of strong regularity is the crucial part and essentially relies on the observation that L^2-local quadratic growth and L^2-local uniqueness of critical points implied by SSCs for certain quadratic problems also stays valid uniformly under perturbation (Proposition 4.17). This and the fact that the set of strongly active points behaves sufficiently well under perturbation (Lemma 4.16) allows to carry over results on U_{ad}^σ to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$ in Corollary 4.18.

At this point, let us recall that the proof of second-order conditions for (P) in [**35**] relied on the results of [**71**]; cf. our comments in Section 1.5.3. Moreover, in Chapters 2 and 3 we applied and extended techniques from [**71**] to address second-order conditions for the state-constrained problem (P^{st}) and the sparse problems ($\mathrm{P}_k^{\mathrm{sp}}$). Now, an extension of the results from [**71**] again plays a central role in the formulation and the proof of Proposition 4.17 that is decisive for formulating the quadratic problems with restriction onto L^2- instead of L^∞-balls.

Finally, in Section 4.4.3 the iterates of the generalized Newton method are identified with the solutions of the quadratic subproblems; see Proposition 4.24. We start with the iterates of the SQP method with subproblems restricted to U_{ad}^σ from Section 4.4.1. Using perturbation arguments analogous to those from Section 4.4.2 it is shown that sufficiently close to the true KKT triple these iterates can also be obtained as unique solution of the quadratic subproblems on $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$ with appropriate $\rho > 0$, or as the unique local solution of the global quadratic subproblem that is contained in the aforementioned set; see Proposition 4.24.

4.4.1. The SQP method on U_{ad}^σ. In this section we relate the iterates of the Newton method from Section 4.3 to solutions of (QP-σ); see Section 4.1.2 for the definition of U_{ad}^σ and (QP-σ). To do so, we will show that the formal optimality conditions for (QP-σ) encoded in the Newton equations for (GE-σ) are indeed sufficient optimality conditions for (QP-σ). Following again [**268, 114, 285**] this is done by showing strict convexity for (QP-σ) for (y_k, u_k, p_k) sufficiently close to $(\bar{y}, \bar{u}, \bar{p})$. We prove convergence of the SQP method under the technical restriction to replace U_{ad} by U_{ad}^σ. Assuming strongest possible SSCs, i.e. $U_{\mathrm{ad}} = U_{\mathrm{ad}}^\sigma$, this yields our first main result.

Recall the definition of the space X_s from Section 4.1.1. The following result corresponds to [**268**, Lemma 6.2, Corollary 6.3].

Proposition 4.13 ([**167**], Proposition 6.1). *Let Assumptions 1.5, 1.6, 4.1 and 4.4 with some $\sigma \in [0, \infty]$ be satisfied. Then, the linear quadratic SQP problem* (QP-σ) *is a strictly convex optimization problem as long as (y_k, u_k, p_k) is sufficiently close to $(\bar{y}, \bar{u}, \bar{p})$ in X_s.*

Proof. The optimization problems (QP-σ) are of linear quadratic type. To show strict convexity, it suffices to show coercivity but the latter is an immediate consequence of the second-order sufficient condition (SSC-σ) and the uniform estimate from Proposition 4.8. □

Now, we can show locally superlinear convergence of the SQP method with quadratic problems on U_{ad}^{σ}.

Theorem 4.14 ([**167**], Theorem 6.2). *Let the assumptions of Theorem 4.12 be fulfilled.*

1. *There is a radius $r_{\mathrm{SQP}\text{-}\sigma} > 0$ such that for any start triple $(y_0, u_0, p_0) \in X_s$ fulfilling*
$$(y_0, u_0, p_0) \in \mathbb{B}^{X_s}_{r_{\mathrm{SQP}\text{-}\sigma}}((\bar{y}, \bar{u}, \bar{p}))$$
the sequences of iterates generated by the generalized Newton method applied to (GE-σ) *or generated by the SQP method with quadratic subproblems* (QP-σ) *are both well-defined, coincide, stay in the ball $\mathbb{B}^{X_s}_{r_{\mathrm{SQP}\text{-}\sigma}}((\bar{y}, \bar{u}, \bar{p}))$, and converge superlinearly to $(\bar{y}, \bar{u}, \bar{p})$ in X_s.*
2. *The statement analogous to part 1 with X_s replaced by X_∞ is true, too.*
3. *There is a radius $\tilde{r}_{\mathrm{SQP}\text{-}\sigma} > 0$ such that the SQP method with quadratic subproblems* (QP-σ) *and initial iterate (y_0, u_0, p_0) with*
$$\|y_0 - \bar{y}\|_{\mathbb{W}^{1,s}(I,(W^{-1,p}_{\Gamma_D},W^{1,p}_{\Gamma_D}))} + \|p_0 - \bar{p}\|_{\mathbb{W}^{1,s}(I,(W^{-1,p'}_{\Gamma_D},W^{1,p'}_{\Gamma_D}))} \leq \tilde{r}_{\mathrm{SQP}\text{-}\sigma}$$
converges superlinearly in X_s and X_∞ to $(\bar{y}, \bar{u}, \bar{p})$. In particular, we can choose
$$u_0 \in U_{\mathrm{ad}}, \qquad \|u_0 - \bar{u}\|_{L^2(\Lambda)} \quad \textit{sufficiently small},$$
$$y_0, p_0 \ \textit{state and adjoint state associated to } u_0.$$

Proof. For parts 1 and 2 the proof works analogous to that of [**268**, Theorem 6.4]. For part 3 note that (QP-σ) is actually independent of the current control iterate u_k, cf. also the remark after (4.5), which shows the first statement in part 3. Since U_{ad} is bounded in L^∞ and $s > 2$ by Assumption 4.1 it holds
$$\|u_0 - \bar{u}\|_{L^s} \leq C\|u_0 - \bar{u}\|_{L^2}^{2/s} \qquad \forall u_0 \in U_{\mathrm{ad}}$$
by the Riesz-Thorin interpolation theorem; cf. also the remark after the next theorem. Here, $C > 0$ is a constant depending on the L^∞-bound of U_{ad} only, i.e. on u_{a} and u_{b}. From this we conclude by continuity
$$\|(y_0, u_0, p_0) - (\bar{y}, \bar{u}, \bar{p})\|_{X_s} \lesssim \|u_0 - \bar{u}\|_{L^2}^{2/s},$$
which shows the second statement of part 3. □

In Section 4.1.2 we have pointed out that (SSC-σ) with $\sigma = \infty$, i.e. coercivity of the second derivative of the Lagrangian on the whole space $L^2(\Lambda)$, is the strongest possible SSC. Under this assumption we are now able to state our first main result. Note that it is possible to formulate all "closeness" required for convergence of the SQP method with respect to L^2-norms.

Theorem 4.15 ([**167**], Theorem 6.3)**.** *Let Assumptions 1.5, 1.6 and 4.1 be fulfilled and let the second-order sufficient condition* (SSC-σ) *from Assumption 4.4 hold on the whole space $L^2(\Lambda)$ (i.e. $\sigma = \infty$). Then the SQP method for* (P) *started in $(y_0, u_0, p_0) \in X_s$,*

$$u_0 \in U_{\mathrm{ad}}, \qquad \|u_0 - \bar{u}\|_{L^2(\Lambda)} \quad \textit{sufficiently small},$$
$$y_0, p_0 \textit{ state and adjoint state associated to } u_0,$$

converges superlinearly in X_s and X_∞ to $(\bar{y}, \bar{u}, \bar{p})$.

Proof. Use Theorem 4.14.3 together with $U_{\mathrm{ad}}^\sigma = U_{\mathrm{ad}}$. □

That the topologies generated by the L^2- and the L^s-norm ($s > 2$) coincide on an L^∞-bounded set by the Riesz-Thorin interpolation theorem, is a well-known fact. However, this observation is a key argument for many proofs concerning second-order conditions without two-norm gap; see, e.g., [**71**, Proposition 3.4] and our comments below Proposition 1.21 that is a crucial step towards Theorems 1.22, 2.18, 2.29 and 3.4. Nevertheless, let us emphasize that the application of the framework from [**71**], cf. Assumption 2.7.1 or Assumption 3.5.1, is not exclusively limited to such situations; see, e.g., [**71**, Section 4]. In Theorem 4.14.3 and Theorem 4.15 above, we made use of this technique to tighten the unsatisfying gap between the quadratic growth condition for j implied by (SSC-σ) —this growth condition holds L^2-locally— and the L^s-local convergence of the SQP method.

4.4.2. The generalized Newton method on U_{ad} and $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$. Before showing convergence of the SQP method restricted to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$ we first consider convergence of the Newton method for the associated generalized equation.

Doing so, we follow arguments from [**285**] but similar results are also contained in [**114, 268**]. Our analysis, however, differs from these references due to L^2-locality instead of L^∞-locality in the statements of Proposition 4.17. An analogous technique will be utilized afterwards in Section 4.4.3 to prove also convergence of the SQP method under certain localization conditions.

In the following we consider the perturbed generalized equation

$$\text{(GE-D)} \qquad d \in F(\bar{y}, \bar{u}, \bar{p}) + F'(\bar{y}, \bar{u}, \bar{p})(y - \bar{y}, u - \bar{u}, p - \bar{p}) + N(y, u, p).$$

Note that we now use the normal cone map N associated with the true set of admissible controls U_{ad} instead of the normal cone map N^σ associated with the modified admissible set U_{ad}^σ that was used for the definition of (GE-σ-D) in the previous sections. Furthermore, note that (GE-D) can be understood as generalized equation both in the spaces X_s, Z_s and X_∞, Z_∞; for the definition of these spaces see Section 4.1.1. As before, the generalized equation (GE-D) is the formal optimality system of the following perturbed optimal control problem:

$$\text{(QP-D)} \quad \begin{cases} \min\limits_{y,u} \dfrac{1}{2}\|y - y_d\|^2 + \dfrac{\alpha}{2}\|u\|^2 - \dfrac{1}{2}\langle \bar{p}, \mathcal{A}''(\bar{y})[y - \bar{y}]^2\rangle - \langle d_u, u\rangle + \langle d_p, y\rangle \\ \text{s.t.} \quad u \in U_{\mathrm{ad}} \quad \text{and} \quad \begin{pmatrix} d_y \\ 0 \end{pmatrix} = e_y(\bar{y}, \bar{u})(y - \bar{y}) + e_u(\bar{y}, \bar{u})(u - \bar{u}). \end{cases}$$

The reduced objective function for (QP-D) will be denoted by j_d. Note that we did not discuss properties of this optimization problem so far. Further, we introduce the following notation for the strongly active sets w.r.t. (QP-D):

$$A_d^\sigma(u) := \{x \in \Lambda\colon\ |\nabla j_d(u)|(x) = |B^*p + \alpha u - d_u|(x) > \sigma\},$$
$$A^\sigma(u) := A_0^\sigma(u), \qquad \text{i.e. } d = 0 \text{ in the definition above.}$$

Here, p denotes the adjoint state associated with u w.r.t. (QP-D) with perturbation vector d; see (4.15). As the notation $A^\sigma(u)$ indicates, $A_0^\sigma(\bar u)$ coincides with the strongly active set for $\bar u$ defined in Assumption 4.4.

In Section 4.3 we observed that under Assumptions 1.5, 1.6, 4.1 and 4.4 the restricted optimal control problem (QP-σ-D), i.e. problem (QP-D) restricted to U_{ad}^σ, is strictly convex and admits a unique solution $(\bar y_d, \bar u_d, \bar p_d)$. This holds true for arbitrarily large perturbation vectors d. In particular, the map $d = (d_y, d_p, d_u) \mapsto (\bar y_d, \bar u_d, \bar p_d)$ was shown to be Lipschitz from Z_∞ to X_∞ in Theorem 4.11, say with modulus $L' > 0$. It follows that the mapping

$$\begin{aligned} &Z_\infty \to L^\infty(\Lambda),\\ &\quad d \mapsto \alpha\bar u_d + B^*\bar p_d - d_u = \nabla j_d(\bar u_d) \end{aligned} \tag{4.18}$$

is Lipschitz as well, say with modulus $L > 0$. Of course, even the map $Z_s \to X_s$, $d \mapsto (\bar y_d, \bar u_d, \bar p_d)$ is Lipschitz continuous as shown in Theorem 4.11, which implies that $d \mapsto \alpha\bar u_d + B^*\bar p_d - d_u$ is Lipschitz continuous from Z_s to $L^s(\Lambda)$. Unfortunately, we will rely on L^∞-estimates in the following.

Assuming that (SSC-σ) holds for some $\sigma \in (0, \infty)$ we can draw some immediate conclusions from the Lipschitz continuity of (4.18) as done in [**285**, Corollaries 5.3 and 5.4].

Lemma 4.16 ([**167**], Lemma 6.6). *Let Assumptions 1.5, 1.6, 4.1 and 4.4 with some $\sigma \in (0, \infty)$ hold and suppose that $\|d\|_{Z_\infty} < \frac{\sigma}{2L}$.*

1. *It holds $A^\sigma(\bar u) \subset A_d^{\sigma/2}(\bar u_d)$ and the signs of $\nabla j_d(\bar u_d)$ and $\nabla j_0(\bar u)$ coincide on $A^\sigma(\bar u)$.*
2. *The solution $(\bar y_d, \bar u_d, \bar p_d)$ of* (QP-σ-D) *is a solution of* (GE-D) *as well, i.e. it holds*

$$\langle \alpha\bar u_d + B^*\bar p_d - d_u, u - \bar u_d\rangle_{L^2(\Lambda)} \geq 0 \qquad \forall u \in U_{\text{ad}}.$$

Proof. This works completely analogous to [**285**]. □

Lemma 4.16 shows that the solution of (QP-σ-D) that depends Z_∞-X_∞-Lipschitz on d is a solution of (GE-D) as well if the perturbation d is small enough in Z_∞. To establish strong regularity of (GE) w.r.t. the spaces X_∞, Z_∞ from this result we have to prove that this solution is locally unique. This is done by proving that $(\bar y_d, \bar u_d, \bar p_d)$ is not only a global solution of (QP-σ-D) but even a local solution of (QP-D) fulfilling a quadratic growth condition on a ball around $(\bar y_d, \bar u_d, \bar p_d)$ with radius independent of d. As announced in the introduction of this section, the following result is, together with its analogue Proposition 4.23 in Section 4.4.3 below, the cornerstone of this section.

Proposition 4.17 ([**167**], Proposition 6.7). *Let the assumptions of Lemma 4.16 be satisfied.*

1. *Then there exist* $0 < \tilde{\epsilon} < \frac{\sigma}{2L}$ *and* $\tilde{\rho}, \eta > 0$, *such that* $(\bar{y}_d, \bar{u}_d, \bar{p}_d)$, *i.e. the solution of* (QP-σ-D), *is also an* L^2*-local solution of* (QP-D) *satisfying the quadratic growth condition*

$$j_d(u) \geq j_d(\bar{u}_d) + \eta \|u - \bar{u}_d\|_{L^2}^2$$

for $\|u - \bar{u}_d\|_{L^2(\Lambda)} \leq \tilde{\rho}$, $u \in U_{\mathrm{ad}}$, *as long as* $\|d\|_{Z_\infty} < \tilde{\epsilon}$.

2. *There are* $0 < \hat{\epsilon} \leq \tilde{\epsilon}$, $0 < \hat{\rho} \leq \tilde{\rho}$ *such that* $(\bar{y}_d, \bar{u}_d, \bar{p}_d)$ *is the only stationary*[1] *point for* (QP-D) *in* $\overline{\mathbb{B}_{\hat{\rho}}^{L^2}(\bar{u}_d)}$.

The first statement of this proposition corresponds to [**285**, Theorem 5.5] with the L^∞-ball around $\bar{u}_d$ replaced by an L^2-ball. To establish quadratic growth L^∞-locally around $\bar{u}_d$, one could follow the direct proof of [**270**, Theorem 5.17]. Avoiding the two-norm gap —which is our aim— can be done following ideas from [**71**, Theorem 2.3], see also [**273**, Theorem 3.22], utilizing a proof by contradiction. Different extensions of the abstract techniques of [**71**] have already been used in Chapters 2 and 3 of this thesis to derive second-order conditions for state-constrained or sparse control problems without two-norm gap. Note that for every single perturbation $d \in Z_\infty$, both properties in the proposition are directly implied by [**71**, Theorem 2.3 and Corollary 2.6]. The crucial point here is to guarantee that the radii of the respective balls can be chosen independently of the choice of the perturbation d as long as $\|d\|_{Z_\infty}$ is small enough.

Proof. 1. For the proof of part 1 we extended the technique of the proof of [**71**, Theorem 2.3] to our needs. First, note that due to the quadratic structure of (QP-D) it holds $j_d'''(\bar{u}_d)[v_1, v_2] = j'''(\bar{u})[v_1, v_2]$. In particular, j_d''' is independent of d. We are going to argue by contradiction and assume the contrary of our claim: there are sequences $(d_n)_n \subset Z_\infty$, $(h_n)_n \subset L^2(\Lambda)$ with $\|d_n\|_{Z_\infty} < \frac{1}{n}$, $\|h_n\|_{L^2} < \frac{1}{n}$ and $\bar{u}_{d_n} + h_n \in U_{\mathrm{ad}}$ such that

$$j_{d_n}(\bar{u}_{d_n} + h_n) - j_{d_n}(\bar{u}_{d_n}) < \frac{1}{n}\|h_n\|_{L^2}^2. \tag{4.19}$$

Define $v_n := \frac{h_n}{\|h_n\|_{L^2}}$ and $\rho_n := \|h_n\|_{L^2}$. It holds $d_n = (d_{y,n}, d_{p,n}, d_{u,n}) \to 0$ strongly in Z_∞, which implies $\bar{u}_{d_n} \to \bar{u}$ and $\nabla j_{d_n}(\bar{u}_{d_n}) \to \nabla j(\bar{u})$ strongly in $L^\infty(\Lambda)$. Due to $\|v_n\|_{L^2} = 1$ for all $n \in \mathbb{N}$ we can w.l.o.g. assume that $v_n \rightharpoonup v_*$ weakly in $L^2(\Lambda)$ for some $v_* \in L^2(\Lambda)$.

Step I. We prove $j'(\bar{u})v_* = 0$. We have

$$\begin{aligned} j'(\bar{u})v_* &= \langle \text{strong-}\lim_{n\to\infty} \nabla j_{d_n}(u_{d_n}), \text{weak-}\lim_{n\to\infty} v_n \rangle_{L^2} \\ &= \lim_{n\to\infty} \langle \nabla j_{d_n}(u_{d_n}), v_n \rangle_{L^2} \geq 0 \end{aligned} \tag{4.20}$$

because $\langle \nabla j_{d_n}(u_{d_n}), v_n \rangle_{L^2} = \frac{1}{\rho_n} \langle \nabla j_{d_n}(u_{d_n}), h_n \rangle_{L^2} \geq 0$ holds for every n due to $\bar{u}_{d_n} + h_n \in U_{\mathrm{ad}}$ and Lemma 4.16.2, for which we can assume w.l.o.g. that $\|d_n\|_{Z_\infty} < \frac{\sigma}{2L}$. Further, by the mean value theorem there are $\theta_n \in (0,1)$ such that

$$\frac{j_{d_n}(u_{d_n} + \rho_n v_n) - j_{d_n}(\bar{u}_{d_n})}{\rho_n} = \langle \nabla j_{d_n}(\bar{u}_{d_n} + \theta_n \rho_n v_n), v_n \rangle_{L^2}.$$

[1] We call (y, u, p) stationary for (QP-D) if (y, u, p) fulfills the first-order necessary conditions for (QP-D).

Due to the structure of (QP-D) —see, e.g., (4.16), (4.17) and use regularity results as in the proof of Theorem 4.11— we know that $\nabla j_{d_n}(\bar{u}_{d_n} + \theta_n \rho_n v_n) \to \nabla j(\bar{u})$ strongly in $L^2(\Lambda)$, which implies

$$\frac{j_{d_n}(u_{d_n} + \rho_n v_n) - j_{d_n}(\bar{u}_{d_n})}{\rho_n} \to j'(\bar{u})v_* \qquad \text{as } n \to \infty. \tag{4.21}$$

On the other hand, it holds by assumption (4.19) that

$$\frac{j_{d_n}(u_{d_n} + \rho_n v_n) - j_{d_n}(\bar{u}_{d_n})}{\rho_n} < \frac{1}{\rho_n} \cdot \frac{1}{n}\|h_n\|_{L^2}^2 = \frac{\rho_n}{n} \to 0,$$

which together with (4.21) yields $j'(\bar{u})v_* \leq 0$ first, and then together with (4.20):

$$j'(\bar{u})v_* = 0. \tag{4.22}$$

Step 2. We want to show that $v_* = 0$ if $|\nabla j(\bar{u})| > 0$. To do so, we show $v_* \geq 0$ if $\nabla j(\bar{u}) > 0$ and $v_* \leq 0$ if $\nabla j(\bar{u}) < 0$, which implies together with Step 1 the desired property. For $\sigma' > 0$ arbitrary define $A^{\sigma',a}(\bar{u}) := \{x \in \Lambda : \nabla j(\bar{u}) > \sigma'\}$. As in the proof of Lemma 4.16 we conclude that $\nabla j_{d_n}(\bar{u}_{d_n}) > 0$ on $A^{\sigma',a}(\bar{u})$ for all sufficiently large n, which implies $h_n, v_n \geq 0$ on $A^{\sigma',a}(\bar{u})$ for all such n. Because weak convergence in L^2 preserves signs a.e. we conclude $v_* \geq 0$ on $A^{\sigma',a}(\bar{u})$. Since $\sigma' > 0$ was arbitrary it follows $v_* \geq 0$ whenever $\nabla j(\bar{u}) > 0$, as stated. The case $\nabla j(\bar{u}) < 0$ is handeled in the same way.

Step 3. In Step 2 we have shown that $v_* \in C^0(\bar{u}) \subset C^\sigma(\bar{u})$ holds; for the definition of $C^0(\bar{u})$ and $C^\sigma(\bar{u})$ see Assumption 4.4. In this final step we will arrive at the final contradiction. First observe that by our assumption

$$\begin{aligned}\frac{\rho_n^2}{n} = \frac{1}{n}\|h_n\|_{L^2}^2 > j_{d_n}(\bar{u}_{d_n} + h_n) - j_{d_n}(\bar{u}_{d_n}) &\overset{(\bigstar)}{=} j'_{d_n}(\bar{u}_{d_n})h_n + \frac{1}{2}j''(\bar{u})h_n \\ &\overset{(\blacksquare)}{\geq} \frac{\rho_n^2}{2} j''(\bar{u})v_n^2,\end{aligned}$$

where we used the linear quadratic structure of (QP-D) at ($\bigstar$) and the first-order optimality condition at ($\blacksquare$). It follows

$$j''(\bar{u})v_*^2 \leq \liminf_{n\to\infty} j''(\bar{u})v_n^2 \leq \liminf_{n\to\infty} \frac{2}{n} = 0, \tag{4.23}$$

where the first inequality comes from the weak lower semicontinuity of $j''(\bar{u})$; see Proposition 1.21.2. Since $v_* \in C^\sigma(\bar{u})$ we can apply (SSC-σ) and conclude from (4.23) that $v_* = 0$. Using Proposition 1.21.3 at ($\blacktriangle$) we obtain

$$c = c \liminf_{n\to} \|v_n\|_{L^2}^2 \overset{(\blacktriangle)}{\leq} \liminf_{n\to\infty} j''(\bar{u})v_n^2 \overset{(4.23)}{=} 0$$

which is the desired contradiction.

2. The second part of the proposition is shown similarly adapting the proof of [**71**, Corollary 2.6]. Since one of the read threads of this thesis is concerned with second-order conditions, the two-norm gap, and related issues, we nevertheless give the details. Again we assume the contrary, i.e. that there are sequences $(d_n)_n \subset Z_\infty$, $(u_n)_n \subset U_{ad}$, such that $d_n \to 0$ strongly in Z_∞ and

$$\|u_n - \bar{u}_{d_n}\|_{L^2} \leq \frac{1}{n} \qquad \text{and} \qquad j'_{d_n}(u_n)(v - u_n) \geq 0 \quad \forall v \in U_{ad}.$$

W.l.o.g. we can assume $\|u_n - \bar{u}_{d_n}\|_{L^2} \leq \tilde{\rho}$ and $\|d_n\|_{Z_\infty} < \tilde{\epsilon}$ and apply the first part of the proposition:

$$\begin{aligned} j_{d_n}(u_n) &\geq j_{d_n}(\bar{u}_{d_n}) + \eta\|u_n - \bar{u}_{d_n}\|_{L^2}^2 \\ &= j_{d_n}(u_n) + j'_{d_n}(u_n)(\bar{u}_{d_n} - u_n) + \frac{1}{2}j''_{d_n}(u_n)[\bar{u}_{d_n} - u_n]^2 + \eta\|u_n - \bar{u}_{d_n}\|_{L^2}^2 \\ &\overset{(\star)}{\geq} j_{d_n}(u_n) + \frac{1}{2}j''(\bar{u})[\bar{u}_{d_n} - u_n]^2 + \eta\|u_n - \bar{u}_{d_n}\|_{L^2}^2 \end{aligned}$$

Note that at $(\star)$ we use our assumption that u_n is a stationary point for (QP-D) and that $j''_{d_n} = j''(\bar{u})$ holds. We conclude that for $v_n := \frac{1}{\rho_n}(u_n - \bar{u}_{d_n})$ and $\rho_n := \|u_n - \bar{u}_{d_n}\|_{L^2} < \frac{1}{n}$ it holds:

$$\frac{1}{2}j''(\bar{u})v_n^2 \leq -\eta. \tag{4.24}$$

Since $(v_n)_n \subset L^2(\Lambda)$ is bounded we can assume w.l.o.g. that $v_n \rightharpoonup v_*$ weakly in $L^2(\Lambda)$ with some $v_* \in L^2(\Lambda)$. Due to optimality of $\bar{u}_{d_n}$ and $u_n \in U_{ad}$ we have

$$j'_{d_n}(\bar{u}_{d_n})v_n = \frac{1}{\rho_n}j'_{d_n}(\bar{u}_{d_n})(u_n - \bar{u}_{d_n}) \geq 0$$

and by stationarity of u_n

$$-j'_{d_n}(u_n)v_n = \frac{1}{\rho_n}j'_{d_n}(u_n)(\bar{u}_{d_n} - u_n) \geq 0.$$

As in Step 1 in the proof of the first part it follows $j'(\bar{u})v_* = 0$. Further, the same argument as in Step 2 shows that $v_* \in C^\sigma(\bar{u})$. Hence we get the desired contradiction

$$0 \overset{\text{(SSC-}\sigma\text{)}}{\leq} \frac{1}{2}j''(\bar{u})v_*^2 \overset{(\clubsuit)}{\leq} \liminf_{n\to\infty} \frac{1}{2}j''(\bar{u})v_n^2 \leq -\eta < 0,$$

where we used Proposition 1.21.2 at $(\clubsuit)$. □

Given a radius $\rho > 0$ we introduce another modification of the perturbed linear quadratic problem (QP-D)

$$\text{(QP-D-}\rho\text{)} \qquad \begin{cases} \min\limits_{y,u} \dfrac{1}{2}\|y - y_d\|^2 + \dfrac{\alpha}{2}\|u\|^2 - \dfrac{1}{2}\langle \bar{p}, \mathcal{A}''(\bar{y})[y - \bar{y}]^2\rangle - \langle d_u, u\rangle \\ \qquad\qquad + \langle d_p, y\rangle \\ \text{s.t. } u \in U_{\text{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})} \\ \qquad \text{and } \begin{pmatrix} d_y \\ 0 \end{pmatrix} = e_y(\bar{y}, \bar{u})(y - \bar{y}) + e_u(\bar{y}, \bar{u})(u - \bar{u}) \end{cases}$$

for which the following result holds.

Corollary 4.18 ([**167**], Corollary 6.8). *Let the assumptions of Lemma 4.16 be satisfied.*

1. *There are $\epsilon, \rho > 0$ such that for $\|d\|_{Z_\infty} < \epsilon$ the triple $(\bar{y}_d, \bar{u}_d, \bar{p}_d)$, i.e. the unique solution of* (QP-σ-D)*, is also the unique solution of* (QP-D-ρ).
2. *There are $\epsilon, \tau > 0$ such that for $\|d\|_{Z_\infty} < \epsilon$ the control $\bar{u}_d$ is the unique solution of* (GE-D) *that is contained in the set $U_{\text{ad}} \cap \overline{\mathbb{B}_\tau^{L^2}(\bar{u})}$.*

A result similar to part 2 —but with L^∞- instead of L^2-balls— was proven in [**114**, Theorem 5.4] using a different argument that relies on strongly active sets and continuity of (4.18).

Proof. 1. Choose $\rho = \frac{2\tilde{\rho}}{3}$ and $\epsilon < \min\left(\tilde{\epsilon}, \frac{\tilde{\rho}}{3C}\right)$, where $C > 0$ is the Z_∞-L^2-Lipschitz constant for the map $d \mapsto \bar{u}_d$; cf. Theorem 4.11 for the Lipschitz continuity. Then, it holds in particular $\|d\|_{Z_\infty} < \tilde{\epsilon}$, i.e. the previous Proposition applies, and

$$\bar{u}_d \in U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})} \subset U_{\mathrm{ad}} \cap \overline{\mathbb{B}_{\tilde{\rho}}^{L^2}(\bar{u}_d)}$$

for all $\|d\|_{Z_\infty} < \epsilon$. Since $\bar{u}_d$ is the unique minimizer of (QP-D) restricted to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_{\tilde{\rho}}^{L^2}(\bar{u}_d)}$ by quadratic growth (Proposition 4.17.1) and this minimizer is contained in the smaller set $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$, we finally proved that $\bar{u}_d$ is the unique minimizer of (QP-D) restricted to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$, i.e. the unique minimizer of (QP-D-ρ).

2. Similarly as for part 1. Now make use of Proposition 4.17.2. □

We introduce another variation of (GE),

$$0 \in F(y,u,p) + N^\rho(y,u,p), \tag{GE-ρ}$$

with the set-valued map $N^\rho(y,u,p) := \left(\{0\},\{0\},\{0\},\{0\},N_{U_{\mathrm{ad}}\cap\overline{\mathbb{B}_\rho^{L^2}(\bar{u})}}(u)\right)^T$, where $N_{U_{\mathrm{ad}}\cap\overline{\mathbb{B}_\rho^{L^2}(\bar{u})}}(u)$ denotes the normal cone of the closed convex set $U_{\mathrm{ad}}\cap\overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$ at some point u. The first part of the following result is similar to [**285**, Corollary 5.6] and the second part to the observation on top of p. 240 in [**114**].

Theorem 4.19 ([**167**], Theorem 6.9)**.** *Let the assumptions of Lemma 4.16 be fulfilled. It holds:*

1. *The generalized equation* (GE) *is strongly regular at* $(\bar{y},\bar{u},\bar{p})$ *w.r.t. the spaces* X_∞, Z_∞.
2. *There is a* $\rho > 0$ *such that the generalized equation* (GE-ρ) *is strongly regular at* $(\bar{y},\bar{u},\bar{p})$ *w.r.t. the spaces* X_∞, Z_∞.

Proof. Both statements are consequences of Corollary 4.18 and Theorem 4.11. The first part is proven in the same way as in [**285**]. We have to use that the L^∞-norm is stronger than the L^2-norm. For the second part note that for all u in the L^2-interior of the ball $\overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$, i.e. in particular for all u sufficiently close to $\bar{u}$ in the L^∞-norm, the equality $N_{U_{\mathrm{ad}}\cap\overline{\mathbb{B}_\rho^{L^2}(\bar{u})}}(u) = N_{U_{\mathrm{ad}}}(u)$ holds, as already mentioned in [**114**]. □

The following result is an immediate consequence of an abstract result [**156**, Theorem 2.19] and Theorem 4.19. The closed graph property for the normal cone map N^ρ is standard.

Theorem 4.20 ([**167**], Theorem 6.10)**.** *Let Assumptions 1.5, 1.6, 4.1, and 4.4 with some* $\sigma \in (0,\infty)$ *hold. For any* (y_0,p_0) *sufficiently close to* $(\bar{y},\bar{p})$ *in the space*

$$\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p},W_{\Gamma_D}^{1,p})) \times \mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p'},W_{\Gamma_D}^{1,p'}))$$

it holds:

1. *The sequence of iterates generated by the Newton-Josephy method for* (GE) *with initial iterate* (y_0, u_0, p_0) *is well-defined and converges superlinearly in* X_∞ *to* $(\bar{y}, \bar{u}, \bar{p})$.
2. *The same holds true for the sequence of iterates generated by the Newton-Josephy method for* (GE-ρ) *with* ρ *from Theorem 4.19.2.*

From Lemma 4.16 on we had to consider perturbations in Z_∞, i.e. we had to measure the control in $L^\infty(\Lambda)$. This is the reason why have to show strong regularity only in Z_∞, X_∞ and not in Z_s, X_s as we did before. That we impose no condition on u_0 is due to the fact that the Newton update equations for (GE) and (GE-ρ) are independent of the current u-iterate u_k; see the comment after equation (4.5).

4.4.3. The SQP method on $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$. Finally, we investigate how the iterates of the generalized Newton method from Theorem 4.20 can be computed by solving linear quadratic optimal control problems restricted to $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar{u})}$. For analogous results in the case of semilinear equations (but with L^∞- instead of L^2-balls) we refer the reader to [**268**] and [**114**].

Lemma 4.21 ([**167**], Lemma 6.11)**.** *Let the assumptions of Theorem 4.20 hold. Let* $(y_k, u_k, p_k) \in X_\infty$ *be a given triple and consider the restricted quadratic subproblem* (QP-σ) *associated with this triple. There exists a* X_∞*-neighbourhood* V_1 *of* $(\bar{y}, \bar{u}, \bar{p})$ *such that the map*

$$(y_k, u_k, p_k) \mapsto (y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$$

is well-defined on V_1 *and Lipschitz continuous, where* $(y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$ *denotes the unique solution of* (QP-σ).

Proof. Existence and uniqueness of a solution to (QP-σ) is established in Proposition 4.13 for (y_k, u_k, p_k) sufficiently close to $(\bar{y}, \bar{u}, \bar{p})$. Define $\tilde{V}$ to be such a neighbourhood of $(\bar{y}, \bar{u}, \bar{p})$. To see Lipschitz continuity, note that $(y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$ is a solution of the parametrized generalized equation

$$\begin{aligned} 0 &\in G((y_k, u_k, p_k), (y, u, p)) + N^\sigma(y, u, p) \\ &:= F(y_k, u_k, p_k) + F'(y_k, u_k, p_k)(y - y_k, u - u_k, p - p_k) + N^\sigma(y, u, p) \end{aligned}$$

—with (y_k, u_k, p_k) being the parameter— and that

$$\begin{aligned} 0 &\in G((\bar{y}, \bar{u}, \bar{p}), (y, u, p)) + N^\sigma(y, u, p) \\ &= F(\bar{y}, \bar{u}, \bar{p}) + F'(\bar{y}, \bar{u}, \bar{p})(y - \bar{y}, u - \bar{u}, p - \bar{p}) + N^\sigma(y, u, p) \end{aligned}$$

is strongly regular at its solution $(\bar{y}, \bar{u}, \bar{p})$ according to Theorem 4.11. Further, G and G', i.e. F and F', depend continuously on (y_k, u_k, p_k), because F: $X_\infty \to Z_\infty$ is continuously differentiable; cf. Lemma 4.3. Hence, [**156**, Theorem 2.18] guarantees well-definedness and Lipschitz continuity of $(y_k, u_k, p_k) \mapsto (y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$ from X_∞ to X_∞ on a sufficiently small neighbourhood $\hat{V}$ of $(\bar{y}, \bar{u}, \bar{p})$. Now, $V_1 := \tilde{V} \cap \hat{V}$ yields the desired neighbourhood. □

In the previous lemma we have shown in particular that

$$\begin{aligned} X_\infty &\to L^\infty(\Lambda) \\ (y_k, u_k, p_k) &\mapsto \nabla j_k(u_{k+1}^\sigma) = \alpha u_{k+1}^\sigma + B^* p_{k+1}^\sigma \end{aligned} \tag{4.25}$$

is Lipschitz continuous on the X_∞-neighbourhood V_1 of $(\bar{y}, \bar{u}, \bar{p})$. By j_k we denoted the reduced functional of (QP-σ) and p_{k+1}^σ is the adjoint state w.r.t. (QP-σ) associated to the control u_{k+1}^σ; see equations (4.3), (4.4). The same argument as for Lemma 4.16 now shows the following result.

Lemma 4.22 ([**167**], Lemma 6.12). *Let the assumptions of Theorem 4.20 hold. There is a X_∞-neighbourhood V_2 of $(\bar{y}, \bar{u}, \bar{p})$ such that for all $(y_k, u_k, p_k) \in V_2$ the solution $(y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$ of* (QP-σ) *satisfies the first-order necessary optimality conditions of* (QP).

Proof. State- and adjoint equation of (QP) and (QP-σ) coincide. We only have to show that $(y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$ fulfills the variational inequality of (QP) as well; this works completely analogous to Lemma 4.16 replacing (4.18) by (4.25). □

Now, we can show the following result that is similar to Proposition 4.17.

Proposition 4.23 ([**167**], Proposition 6.13). *Let the assumptions of Theorem 4.20 hold. Then, there is a X_∞-neighbourhood V_3 of $(\bar{y}, \bar{u}, \bar{p})$ and there are $\rho, \eta > 0$ such that for all triples $(y_k, u_k, p_k) \in V_3$ the unique solution $(y_{k+1}, u_{k+1}, p_{k+1}) := (y_{k+1}^\sigma, u_{k+1}^\sigma, p_{k+1}^\sigma)$ of* (QP-σ)

1. *is an L^2-local solution of* (QP) *satisfying the quadratic growth condition*

$$j_k(u) \geq j_k(u_{k+1}) + \eta \|u - u_{k+1}\|_{L^2}^2$$

   for all $u \in U_{\mathrm{ad}}$ such that $\|u - u_{k+1}\|_{L^2(\Lambda)} \leq \rho$.
2. *is the only stationary point for* (QP) *in $\overline{\mathbb{B}_\rho^{L^2}(u_{k+1})}$.*

Proof. We proceed as in the proofs of Proposition 4.17.1 and 2 and argue by contradiction. Instead of j_{d_n} and $\bar{u}_{d_n}$ we have to consider j_k and u_{k+1}. We only mention the essential ingredients that keep all the previous arguments working:

i. For any sequence $(w_k) \subset U_{\mathrm{ad}}$ such that $w_k \to \bar{u}$ in $L^2(\Lambda)$ it holds

$$\nabla j_k(w_k) \to \nabla j(\bar{u}) \quad \text{strongly in } L^2(\Lambda).$$

   This was shown in Proposition 4.9; use the Riesz-Thorin interpolation theorem as explained at the end of Section 4.4.1 to obtain the required L^s-convergence $w_k \to \bar{u}$ from the given L^2-convergence.

ii. If $u_k \to \bar{u}$ strongly in L^2 and $v_k \rightharpoonup v_*$ weakly in L^2 we have

$$j''(\bar{u}) v_*^2 \leq \liminf_{k \to \infty} j_k''(u_k) v_k^2.$$

   This can be shown as follows: one uses boundedness of $(v_k)_k$, Proposition 4.8, weak lower semicontinuity of j'' (see Proposition 1.21.2), and finds

$$\liminf_k j_k''(u_k) v_k^2 \;\geq\; \liminf_k \underbrace{\left(j_k''(u_k) - j''(u_k)\right) v_k^2}_{\to 0 \text{ uniformly in } v_k} + \liminf_k j''(u_k) v_k^2 \;\geq\; j''(\bar{u}) v_*^2.$$

iii. If $v_* = 0$ in part ii, then

$$c \liminf_{k\to\infty} \|v_k\|_{L^2}^2 \le \liminf_{k\to\infty} j_k''(u_k)v_k^2$$

with some $c > 0$. This is shown by the same argument as above utilizing the results of [**35**] summarized in Proposition 1.21 as auxiliary results.

□

Next, we obtain the following result with the same argument as for Corollary 4.18.

Proposition 4.24 ([**167**], Proposition 6.14)**.** *Let the assumptions of Theorem 4.20 hold.*

1. *There is a X_∞-neighbourhood V_4 of $(\bar y, \bar u, \bar p)$ and a radius $\rho > 0$ such that for all $(y_k, u_k, p_k) \in V_4$ the next SQP iterate $(y_{k+1}, u_{k+1}, p_{k+1})$ given by the unique solution of (QP-σ) is also the unique solution of (QP) with admissible set $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar u)}$.*
2. *There is a X_∞-neighbourhood V_5 of $(\bar y, \bar u, \bar p)$ and a radius $\rho > 0$, such that for all $(y_k, u_k, p_k) \in V_5$ the next SQP iterate $(y_{k+1}, u_{k+1}, p_{k+1})$ given by the unique solution of (QP-σ) is also the unique L^2-local solution of the global quadratic problem (QP) that is contained in $U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar u)}$.*

Now we can state our main result of this chapter.

Theorem 4.25 ([**167**], Theorem 6.15)**.** *Let Assumptions 1.5, 1.6, 4.1, and 4.4 with some $\sigma \in (0, \infty)$ hold. Then there are radii $\rho > 0$, $r_{\mathrm{SQP}} > 0$ such that for any initial guess*

$$(y_0, p_0) \in \mathbb{W}^{1,s}(I, (W_{\Gamma_D}^{-1,p}, W_{\Gamma_D}^{1,p})) \times \mathbb{W}^{1,s}(I, (W_{\Gamma_D}^{-1,p'}, W_{\Gamma_D}^{1,p'}))$$

fulfilling

$$\|y_0 - \bar y\|_{\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p}, W_{\Gamma_D}^{1,p}))} + \|p_0 - \bar p\|_{\mathbb{W}^{1,s}(I,(W_{\Gamma_D}^{-1,p'}, W_{\Gamma_D}^{1,p'}))} \le r_{\mathrm{SQP}}$$

the sequence of iterates generated by the successive solution of the SQP subproblems

$$(\mathrm{QP}(\rho, y_k, p_k)) \quad \begin{cases} \min\limits_{y,u} J_k(y,u) := \dfrac{1}{2}\|y - y_d\|^2 + \dfrac{\alpha}{2}\|u\|^2 - \dfrac{1}{2}\langle p_k, \mathcal{A}''(y_k)[y - y_k]^2\rangle \\ \text{s.t.} \quad u \in U_{\mathrm{ad}} \cap \overline{\mathbb{B}_\rho^{L^2}(\bar u)} \\ \qquad \text{and} \quad \begin{cases} \partial_t y + \mathcal{A}(y_k)y + \mathcal{A}'(y_k)y = Bu + \mathcal{A}'(y_k)y_k, \\ y(0) = y_0 \end{cases} \end{cases}$$

converges superlinearly in X_∞ to $(\bar y, \bar u, \bar p)$.

A possible choice of y_0, p_0 are state y_0 and adjoint state p_0 associated to some control $u_0 \in U_{\mathrm{ad}}$ w.r.t. (P) if $\|u_0 - \bar u\|_{L^2}$ is chosen small enough.

Proof. Combine Proposition 4.24 with Theorem 4.12. □

Note in particular that we tightened the gap between the L^2-local growth condition originating from the second-order sufficient condition (SSC-σ) in Assumption 4.4 and the closeness conditions in the SQP method. This is different from, e.g., [**114, 268, 285**] where closeness had been formulated with respect to L^∞. Now, in Theorem 4.25 above all required closeness can be formulated with respect to the L^2-norm.

4.5. Numerical examples

In the final section of this chapter we illustrate our theoretical results by the numerical examples. We follow the presentation in [**167**, Section 8]. First, we test with so-called manufactured solution examples, i.e. optimal control problems with analytically known solution; see, e.g., [**270**, Section 2.9] for the construction of such examples. Second, we test with an example based on real-world parameters; cf. Section 4.5.2. Let us note that our focus is clearly on illustrating our convergence results and not on computational efficiency, which can be seen, e.g., in the much too accurate solution of the quadratic subproblems that is usually unnecessary and expensive. Approaches to speed up the numerical solution of (P) by combining SQP(-type) methods with model order reduction will be addressed experimentally in Section 5.3, on behalf of the model problem from Section 4.5.2.

We implemented the SQP algorithm in python using `FEniCS` [**9, 203**] for the finite element discretization of the problem. States and adjoint states are discretized piecewise constant w.r.t. time and by piecewise linear finite elements w.r.t. space. The purely time-dependent controls are discretized piecewise constant w.r.t. time. Since the control bounds are constant w.r.t. time, this coincides with the well-known variational discretization concept [**149**]. The details of the respective discretization will be given for each problem below. Following the approach of [**143**], the algorithm implemented consists of three nested loops: the outermost iteration is given by the SQP method. The quadratic subproblem of each SQP iteration is solved iteratively by application of the semismooth Newton method (SSN); see, e.g., [**275**]. Finally, the innermost loop consists of the iterative solution of the Newton update equation by the CG method in every semismooth Newton iteration.

In order to solve the quadratic subproblems accurately enough we choose the relative tolerance for SSN to be 10^{-5}, i.e. the solver of the quadratic subproblems either terminates when the L^2-norm of projection residual (of the subproblem) is reduced by at least 10^{-5} or the maximum of 20 SSN iterations is reached. To avoid problems in case of already very small initial residual norms, the SSN iteration also ends when the residual norm gets smaller than 10^{-12} (absolute tolerance). Similarly, the CG method terminates if the intial CG-residual is decreased by factor at least 10^{-2}. To enhance stability, SSN is combined with Armijo linesearch with the squared L^2-norm of the projection residual (of the subproblem) as merit function.

As observed in, e.g., [**114, 285**] the restriction of the quadratic subproblems to L^∞- or —in our case— L^2-balls is only required to prove convergence of the algorithm in function space. Fortunately, we can omit this additional constraint in practice and solve the quadratic subproblems on U_{ad} without loosing convergence,

i.e. the subproblems in our implementation are given by (QP); cf. the end of Section 4.1.1.

In all three examples the initial guess for the SQP method is $(y_0, u_0, p_0) := (0, 0, 0)$. To measure optimality of some iterate u_k we compute the L^2-norm of the residual of the projection formula

$$\mathrm{res}_k := \left\| u_k - \mathrm{Proj}_{U_{\mathrm{ad}}}\left(-\gamma^{-1} B^* p(u_k)\right)\right\|_{L^2}, \tag{4.26}$$

where the adjoint state $p(u_k)$ associated to u_k w.r.t. (P) is computed using the implicit Euler scheme for the time discretization of the nonlinear state equation. The nonlinear equations appearing at each timestep during the solution of the state equation are solved by the built-in nonlinear solver of FEniCS. Convergence of the SQP method is measured by the increments

$$\begin{aligned}
\mathrm{incr}_k^{\infty} &:= \|y_{k+1} - y_k\|_{L^\infty} + \|u_{k+1} - u_k\|_{L^\infty} + \|p_{k+1} - p_k\|_{L^\infty}, \\
\mathrm{incr}_k^{2} &:= \|y_{k+1} - y_k\|_{L^2(I, H^1_{\Gamma_D})} + \|y_{k+1} - y_k\|_{W^{1,2}(I, H^{-1}_{\Gamma_D})} + \|u_{k+1} - u_k\|_{L^\infty} \\
&\qquad + \|p_{k+1} - p_k\|_{L^2(I, H^1_{\Gamma_D})} + \|p_{k+1} - p_k\|_{W^{1,2}(I, H^{-1}_{\Gamma_D})}.
\end{aligned}$$

Note that we do not compute the norm of the increments with respect to the norms appearing in Theorem 4.25 because we do not have the abstract exponents p, s at hand in a practical context. To illustrate our theoretical results, we show for all examples both increments and residuals for different discretizations. Convergence behaviour of the SQP method uniform with respect to sufficiently fine discretization strongly indicates convergence in function space.

4.5.1. Manufactured solution examples.

4.5.1.1. *Example 1.* For $I = [0, 1]$ and $\Omega = [0, 1]$ we consider the problem

$$\begin{cases}
\min\limits_{y,u} J(y,u) := \dfrac{1}{2}\|y - y_d\|^2_{L^2(Q)} + 10^{-3} \cdot \|u\|^2_{L^2([0,1])} \\
\text{s.t.} \quad u \in \left\{ v \in L^2([0,1]) : -\dfrac{9}{10} \le v(x) \le \dfrac{\sqrt{2}}{2} \quad \text{a.e.} \right\}, \\
\qquad \text{and} \quad \begin{cases} \partial_t y - \nabla \cdot \xi(y) \nabla y = b \cdot u + f \quad \text{on } Q, \\ y = 0 \quad \text{on } I \times \partial\Omega, \\ y(0) = \sin(\pi x_1) \end{cases}
\end{cases} \tag{4.27}$$

and choose

$$\begin{aligned}
\bar{y}(t,x) &= \cos(2\pi t)\sin(\pi x), \\
\bar{p}(t,x) &= \frac{1}{100}\sin(2\pi t)\sin(\pi x), \\
b(x) &= \mathbf{1}_{[1/3,2/3]}(x), \\
\xi(z) &= \frac{1}{2} + \frac{1}{1 + \exp(-5z)}.
\end{aligned}$$

	Level 1	Level 2	Level 3
$h_{\max}$	$7.95 \cdot 10^{-2}$	$3.98 \cdot 10^{-2}$	$1.99 \cdot 10^{-2}$
N_t	158	632	2529
DoF FEM	421	1640	6549

Table 4.1. Discretization levels for Example 2 (Manufactured Solution in 2D, Section 4.5.1.2).

With the help of `Wolfram Mathematica` we compute the remaining quantities y_d, f, $\bar{u}$ such that the optimality system for (4.27) is fulfilled. In particular, it holds

$$\bar{u}(t) = \min\left(\frac{\sqrt{2}}{2}, \max\left(-\frac{9}{10}, -\frac{10}{\pi}\sin(2\pi t)\right)\right).$$

Note that all our theoretical results remain true for a problem of type (4.27) since addition of the term f to the model problem (P) does not change its structural properties.

Discretization of spatial functions is done with piecewise linear finite elements on a equidistant partition of $\Omega = [0,1]$ into N_h subintervals. For time discretization we apply an implicit Euler discretization with $N_t = N_h^2$ equidistant time steps, whereby the number of time steps is chosen in order to roughly balance spatial and temporal discretization errors of the state equation; cf. [**46**]. The behaviour of the increments incr_k^∞ and incr_k^2 during the SQP iteration is shown in Table 4.3, whereas L^2-residuals and errors of the SQP iterates with respect to the interpolated true KKT triple are shown in Table 4.4. Note that increments (Table 4.3.a and b) and their decrease factors (Table 4.3.c and d) indicate superlinear convergence and behave uniform with respect to the different discretization levels, which illustrates superlinear convergence in function space. Also, the residuals (Table 4.4.a) and errors (Table 4.4.b-f) seem to behave uniformly, at least until their convergence stagnates due to the limited accuracy given by discretization.

4.5.1.2. *Example 2.* For $I = [0,1]$ and $\Omega = [0,1]^2$ we consider a problem of the same structure as the 1D manufactured solution example (4.27) but now with

$$\begin{aligned} y_0(x) &= \sin(\pi x_1)\sin(\pi x_2), \\ \bar{y}(t,x) &= \cos(2\pi t)\sin(\pi x_1)\sin(\pi x_2), \\ \bar{p}(t,x) &= \frac{1}{100}\sin(2\pi t)\sin(\pi x_1)\sin(\pi x_2), \\ b(x) &= \pi^2 \cdot \mathbf{1}_{[1/3,2/3]^2}(x) \end{aligned}$$

and the regularization parameter $\alpha = 2 \cdot 10^{-3}$ in (4.27) replaced by $\alpha = 10^{-2}$. As before, the remaining quantities are computed utilizing `Wolfram Mathematica` and the optimal control is given by

$$\bar{u}(t) = \min\left(\frac{\sqrt{2}}{2}, \max\left(-\frac{9}{10}, -\sin(2\pi t)\right)\right).$$

Discretization of spatial functions is now done with piecewise linear finite elements on a triangular mesh generated by `mshr`, the mesh-generation tool of `FEniCS`,

	Level 1	Level 2	Level 3
$h_{\max}$	$5.90 \cdot 10^{-1}$	$2.84 \cdot 10^{-1}$	$1.84 \cdot 10^{-1}$
N_t	115	495	1182
DoF FEM	548	2848	10073

Table 4.2. Discretization levels for Example 3 (Section 4.5.2).

with maximum element diameter $h_{\max}$. For time discretization we apply an implicit Euler discretization with N_t equidistant time steps, whereby as for Example 1 the size of time steps $N_t^{-1} \approx h_{\max}^2$ is chosen in order to roughly balance spatial and temporal discretization errors of the state equation. Maximum element diameter and number of time steps of the three different discretization levels used in our numerical experiments can be found in Table 4.1. In Table 4.5 we display increments and their decrease rates during the SQP iteration. Similarly to the 1D manufactured solution example these quantities behave uniform with respect to different discretization levels, which illustrates convergence in function space. Moreover, residuals (Table 4.6.a) and errors of the iterates with respect to the interpolated true KKT triple (Table 4.6.b-f) show uniform behaviour until stagnation due to the respective discretization occurs.

4.5.2. Example 3. This final example is chosen to demonstrate that our assumptions also cover an example with real-world parameters. We consider the following problem related to heat conduction in a block of silicon modelled according to [**253**]:

$$\begin{cases} \min\limits_{y,u} J(y,u) := \dfrac{1}{2}\|y - y_d\|^2_{L^2(Q)} + 10^{-2} \cdot \|u\|^2_{L^2(I)} \\ \text{s.t.} \quad u \in \left\{v \in L^2(I) : 2.9 \le v(t) \le 10\right\} \\ \qquad \text{and} \quad \begin{cases} \partial_t y - \nabla \cdot \xi(y) \nabla y = 0 \quad \text{on } Q, \\ \xi(y)\partial_{n_\Omega} y + \alpha y = \alpha u \quad \text{on } I \times \partial\Omega, \\ y(0) = 10. \end{cases} \end{cases} \tag{4.28}$$

The spatial domain is

$$\Omega = [-2,2] \times [-0.5,0.5] \times [-1,0] \cup [-0.5,0.5] \times [-2,2] \times [0,1] \subset \mathbb{R}^3$$

and consists of two crossed beams. For a picture of the domain we refer the reader to Figure 5.5 in the next chapter. The time interval is $I = [0,T] = [0,40]$, the desired state

$$y_d(t,x) = 10 - \frac{71}{400}t,$$

the nonlinearity

$$\xi(y) := \frac{1}{a + by + cy^2}, \qquad a = 0.0818292, \quad b = 0.4255118, \quad c = 0.0450061,$$

and $\alpha = 0.0146647$. In order to make ξ formally fulfill Assumption 1.6 we can choose a twice continously differentiable, uniformly bounded from below and above continuation of the above ξ outside the relevant values of y.

Measuring temperature in units of 100 Kelvin [K], length in 0.1 meters [m] and time in 60 seconds [s], the state equation of (4.28) describes the evolution of the

temperature y of a block Ω of silicon with initial temperature 1000K when the temperature of the surrounding air is given by the control-variable u. Hence, the optimal control problems aims at finding the optimal temperature trajectory for the ambient air in order to cool down the block Ω following the desired temperature trajectory y_d from 1000K to room temperature 290K. Density, specific heat, and temperature-dependent thermal conductivity are taken from [**253**, Chapters 2.5 and 4.3] and rescaled according to the abovementioned units. For the heat transfer coefficient between silicon and air (forced convection) we guess the value $40\mathrm{Wm}^{-2}\mathrm{K}^{-1}$ which results in the value given for α.

As pointed out in Example 1.7.3, the domain under consideration fulfills our assumptions although not being a domain with Lipschitz boundary. The Robin boundary condition in (4.28) is not covered by our assumptions but, since it differs from Neumann boundary conditions only by a linear term, this can be tackled by straightforward modifications of our arguments; cf. also [**214, 215**].

Again, we use `FEniCS` and `mshr` for the implementation. Space is discretized by piecewise linear finite elements on tetrahedral meshes with maximal cell diameter $h_{\max}$ while time is discretized by N_t equidistant implicit Euler time steps; see Table 4.2 for the different discretization levels. As before, $h_{\max}$ and N_t are chosen in such a way that spatial and temporal discretization errors of the state equation are of roughly the same order. The numerically determined optimal control and associated optimal state are shown in Figure 4.1.a. Due to the three-dimensionality of the problem we were not able to choose discretization as fine as in the previous examples and therefore the behaviour the increments (Table 4.7) and residuals (Table 4.8) is not as illustrative as in 1D or 2D.

Figure 4.1.b shows an enlarged section of the control iterates near the change from inactive to active set at $t \approx 17.1$. It can be seen that, once the correct active set is identified after the third iteration, convergence is so fast that there is no visible difference between the further iterates. This might be seen as an illustration of the importance of detection of the correct active sets in infinite dimensions that has been discussed at the beginning of Section 4.4. The small kinks in the plots at the border between active and inactive set are due to the fact that time discretization (size of timesteps $\tau \approx 3.38 \cdot 10^{-2}$) does not resolve the active/inactive sets exactly.

	a. Increments incr_k^∞			b. Increments incr_k^2		
k	$N_h = 32$	$N_h = 64$	$N_h = 128$	$N_h = 32$	$N_h = 64$	$N_h = 128$
0	2.15e+00	2.15e+00	2.15e+00	3.54e+00	3.54e+00	3.54e+00
1	1.89e+00	1.89e+00	1.89e+00	2.22e+00	2.22e+00	2.22e+00
2	1.46e-01	1.46e-01	1.46e-01	1.54e-01	1.54e-01	1.54e-01
3	9.00e-05	9.29e-05	9.16e-05	9.81e-05	1.01e-04	1.00e-04
4	4.96e-10	4.98e-10	5.15e-10	5.13e-10	5.17e-10	5.33e-10
5	1.52e-15	2.73e-15	4.75e-15	9.27e-15	1.82e-14	3.64e-14

	c. Decrease of increments $\frac{\text{incr}_{k+1}^\infty}{\text{incr}_k^\infty}$			d. Decrease of increments $\frac{\text{incr}_{k+1}^2}{\text{incr}_k^2}$		
k	$N_h = 32$	$N_h = 64$	$N_h = 128$	$N_h = 32$	$N_h = 64$	$N_h = 128$
0	8.79e-01	8.79e-01	8.79e-01	6.27e-01	6.28e-01	6.27e-01
1	7.74e-02	7.71e-02	7.73e-02	6.96e-02	6.93e-02	6.94e-02
2	6.15e-04	6.37e-04	6.27e-04	6.36e-04	6.58e-04	6.48e-04
3	5.51e-06	5.36e-06	5.62e-06	5.23e-06	5.10e-06	5.33e-06
4	3.06e-06	5.47e-06	9.22e-06	1.81e-05	3.52e-05	6.82e-05

Table 4.3. Increments during the SQP method applied to Example 1 (Manufactured Solution in 1D, Section 4.5.1.1).

	a. Residuals res_k			b. Error in the control $\|u_k - \bar{u}\|_{L^\infty}$		
k	$N_h = 32$	$N_h = 64$	$N_h = 128$	$N_h = 32$	$N_h = 64$	$N_h = 128$
0	7.60e-01	7.56e-01	7.58e-01	9.00e-01	9.00e-01	9.00e-01
1	9.24e-01	9.23e-01	9.26e-01	1.61e+00	1.61e+00	1.61e+00
2	1.13e-01	1.13e-01	1.13e-01	1.70e-01	1.48e-01	1.43e-01
3	3.29e-05	3.85e-05	3.78e-05	3.48e-02	1.59e-02	6.09e-03
4	3.60e-06	3.29e-07	2.51e-07	3.48e-02	1.60e-02	6.04e-03
5	3.60e-06	3.29e-07	2.51e-07	3.48e-02	1.60e-02	6.04e-03
6	3.60e-06	3.29e-07	2.51e-07	3.48e-02	1.60e-02	6.04e-03

	c. Error in the state $\|y_k - \bar{y}\|_{L^\infty}$			d. Error in the state $\|y_k - \bar{y}\|_W$		
k	$N_h = 32$	$N_h = 64$	$N_h = 128$	$N_h = 32$	$N_h = 64$	$N_h = 128$
0	1.00e+00	1.00e+00	1.00e+00	2.61e+00	2.61e+00	2.61e+00
1	2.56e-01	2.57e-01	2.56e-01	5.63e-01	5.69e-01	5.68e-01
2	6.11e-03	6.68e-03	6.65e-03	1.50e-02	1.40e-02	1.40e-02
3	1.85e-03	6.31e-04	2.06e-04	4.73e-03	1.23e-03	4.29e-04
4	1.85e-03	6.32e-04	2.05e-04	4.73e-03	1.23e-03	4.28e-04
5	1.85e-03	6.32e-04	2.05e-04	4.73e-03	1.23e-03	4.28e-04
6	1.85e-03	6.32e-04	2.05e-04	4.73e-03	1.23e-03	4.28e-04

	e. Error in the adjoint state $\|p_k - \bar{p}\|_{L^\infty}$			f. Error in the adjoint state $\|p_k - \bar{p}\|_W$		
k	$N_h = 32$	$N_h = 64$	$N_h = 128$	$N_h = 32$	$N_h = 64$	$N_h = 128$
0	1.00e-02	1.00e-02	1.00e-02	2.61e-02	2.61e-02	2.61e-02
1	2.15e-02	2.16e-02	2.15e-02	4.42e-02	4.47e-02	4.46e-02
2	1.10e-03	1.09e-03	1.07e-03	2.28e-03	2.22e-03	2.15e-03
3	1.32e-04	4.29e-05	1.12e-05	3.36e-04	9.45e-05	2.50e-05
4	1.31e-04	4.27e-05	1.15e-05	3.36e-04	9.41e-05	2.48e-05
5	1.31e-04	4.27e-05	1.15e-05	3.36e-04	9.41e-05	2.48e-05
6	1.31e-04	4.27e-05	1.15e-05	3.36e-04	9.41e-05	2.48e-05

Table 4.4. Residuals and errors of the iterates during the SQP method applied to Example 1 (Manufactured Solution in 1D, Section 4.5.1.1). We use the abbreviation $\|\cdot\|_W := \|\cdot\|_{L^2(I,H^1_{\Gamma_D})} + \|\cdot\|_{W^{1,2}(I,H^{-1}_{\Gamma_D})}$.

	a. Increments incr_k^∞			b. Increments incr_k^2		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	2.15e+00	2.16e+00	2.16e+00	2.75e+00	2.67e+00	2.61e+00
1	1.04e+00	1.05e+00	1.06e+00	1.10e+00	1.11e+00	1.11e+00
2	2.57e-02	2.38e-02	2.22e-02	2.70e-02	2.52e-02	2.35e-02
3	6.32e-06	7.45e-06	7.99e-06	6.09e-06	7.21e-06	7.86e-06
4	1.84e-11	2.03e-12	1.93e-12	1.90e-11	1.10e-12	1.04e-12
	c. Decrease of increments $\frac{\text{incr}_{k+1}^\infty}{\text{incr}_k^\infty}$			d. Decrease of increments $\frac{\text{incr}_{k+1}^2}{\text{incr}_k^2}$		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	4.82e-01	4.88e-01	4.89e-011	4.00e-01	4.17e-01	4.26e-01
1	2.49e-02	2.26e-02	2.10e-02	2.46e-02	2.26e-02	2.11e-02
2	2.45e-04	3.13e-04	3.60e-04	2.25e-04	2.87e-04	3.34e-04
3	2.91e-06	2.72e-07	2.42e-07	3.13e-06	1.53e-07	1.32e-07

Table 4.5. Increments during the SQP method applied to Example 2 (Manufactured Solution in 2D, Section 4.5.1.2).

	a. Residuals res_k			b. Error in the control $\|u_k - \bar{u}\|_{L^\infty}$		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	7.60e-01	7.50e-01	7.43e-01	9.00e-01	9.00e-01	9.00e-01
1	8.05e-01	7.99e-01	7.94e-01	8.23e-01	8.13e-01	8.12e-01
2	2.21e-02	2.36e-02	2.45e-02	1.33e-01	9.08e-02	5.54e-02
3	5.72e-06	1.39e-06	1.13e-06	1.37e-01	8.09e-02	4.48e-02
4	6.88e-06	3.80e-07	1.05e-06	1.37e-01	8.09e-02	4.48e-02
5	6.88e-06	3.80e-07	1.05e-06	1.37e-01	8.09e-02	4.48e-02
	c. Error in the state $\|y_k - \bar{y}\|_{L^\infty}$			d. Error in the state $\|y_k - \bar{y}\|_W$		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	1.00e+00	1.00e+00	1.00e+00	1.81e+00	1.72e+00	1.67e+00
1	2.33e-01	2.40e-01	2.45e-01	2.82e-01	2.93e-01	2.93e-01
2	3.02e-02	1.62e-02	1.02e-02	4.15e-02	2.06e-02	1.13e-02
3	2.93e-02	1.49e-02	7.48e-03	4.06e-02	1.93e-02	9.27e-03
4	2.93e-02	1.49e-02	7.47e-03	4.06e-02	1.93e-02	9.27e-03
5	2.93e-02	1.49e-02	7.47e-03	4.06e-02	1.93e-02	9.27e-03
	e. Error in the adjoint state $\|p_k - \bar{p}\|_{L^\infty}$			f. Error in the adjoint state $\|p_k - \bar{p}\|_W$		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	1.00e-02	1.00e-02	1.00e-02	1.81e-02	1.72e-02	1.67e-02
1	7.66e-03	8.72e-03	9.12e-03	1.17e-02	1.22e-02	1.22e-02
2	1.08e-03	5.52e-04	3.10e-04	1.39e-03	6.26e-04	3.15e-04
3	1.06e-03	6.07e-04	3.20e-04	1.46e-03	6.71e-04	3.18e-04
4	1.06e-03	6.07e-04	3.20e-04	1.46e-03	6.71e-04	3.18e-04
5	1.06e-03	6.07e-04	3.20e-04	1.46e-03	6.71e-04	3.18e-04

Table 4.6. Residuals and errors of the iterates during the SQP method applied to Example 2 (Manufactured Solution in 2D, Section 4.5.1.2). We use the abbreviation $\|\cdot\|_W := \|\cdot\|_{L^2(I,H^1_{\Gamma_D})} + \|\cdot\|_{W^{1,2}(I,H^{-1}_{\Gamma_D})}$.

	a. Increments incr_k^∞			b. Increments incr_k^2		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	1.71e+01	1.72e+01	1.72e+01	1.91e+02	1.85e+02	1.75e+02
1	1.16e+00	1.15e+00	1.17e+00	1.08e+01	1.09e+01	1.07e+01
2	2.24e-02	2.52e-02	2.58e-02	1.42e-01	1.50e-01	1.49e-01
3	5.23e-06	1.15e-05	6.77e-06	3.78e-05	4.51e-05	4.01e-05
4	2.78e-11	2.40e-10	3.37e-10	4.05e-11	2.69e-10	3.74e-10
5	-	1.03e-13	2.80e-13	-	1.29e-12	2.12e-12

	c. Decrease of increments $\frac{\mathrm{incr}_{k+1}^\infty}{\mathrm{incr}_k^\infty}$			d. Decrease of increments $\frac{\mathrm{incr}_{k+1}^2}{\mathrm{incr}_k^2}$		
k	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
0	6.79e-02	6.70e-02	6.80e-02	5.68e-02	5.88e-02	6.10e-02
1	1.93e-02	2.19e-02	2.21e-02	1.31e-02	1.38e-02	1.40e-02
2	2.33e-04	4.57e-04	2.62e-04	2.65e-04	3.01e-04	2.68e-04
3	5.31e-06	2.09e-05	4.98e-05	1.07e-06	5.96e-06	9.33e-06
4	-	4.30e-04	8.31e-04	-	4.81e-03	5.67e-03

Table 4.7. Increments during the SQP method applied to Example 3 (Section 4.5.2).

	Residuals res_k		
k	Level 1	Level 2	Level 3
0	2.58e+01	2.59e+01	2.59e+01
1	1.15e+01	1.16e+01	1.16e+01
2	1.69e+00	1.72e+00	1.72e+00
3	4.23e-04	4.28e-04	4.28e-04
4	2.09e-08	1.10e-09	1.44e-08
5	2.09e-08	1.06e-09	1.45e-08
6	-	1.06e-09	1.45e-08

Table 4.8. Residuals during the SQP method applied to Example 3 (Section 4.5.2).

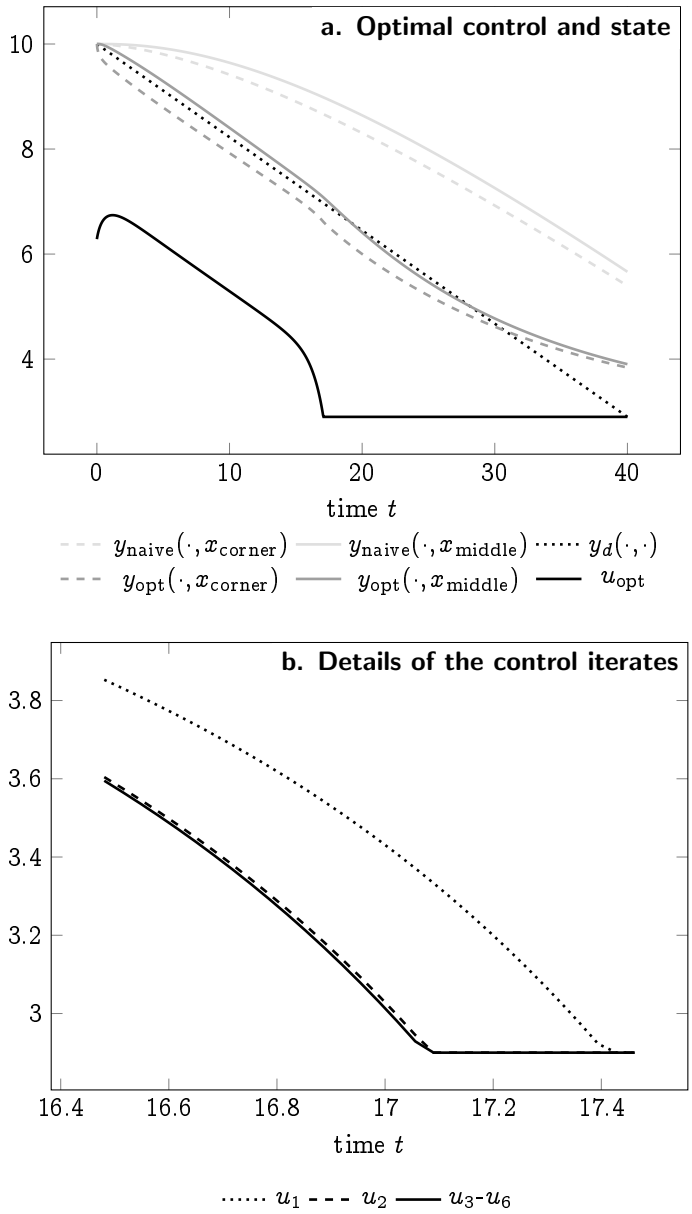


Figure 4.1. Example 3 (Section 4.5.2) on the finest discretization level: **a.** Optimal control and optimal state y_{opt} evaluated at the points $x_{\text{corner}} = (-0,5,2,0)$ and $x_{\text{middle}} = (0,0,0)$. For comparison we also display the state y_{naive} associated with the "naive" first guess $u_{\text{naive}}(t) := 10 - \frac{71}{400}t$ at the same points and the desired trajectory y_d. **b.** Control iterates during the SQP method on a certain subinterval of I.

Chapter 5

Model order reduction by proper orthogonal decomposition

In the previous chapter we have presented several numerical examples for the SQP method with focus on illustrating its convergence in function space. This final chapter of the thesis is devoted to the important problem of making these (or similar) computations faster and more efficient by utilizing so-called model order reduction (MOR). More precisely, we apply proper orthogonal decomposition (POD) to the semidiscrete in space counterpart of the state equation (Eq) and prove corresponding a-posteriori error estimates for the POD error. This is joint work with I. Neitzel and has been published in [**166**]. Moreover, we present numerical experiments in which we achieve a significant reduction of computing time during the numerical solution of the entire control problem by combining POD-MOR with SQP or SQP-type methods on a heuristic level.

The numerical solution of optimal control problems governed by PDEs is a so-called many-query scenario: the same linear or nonlinear equation has to be solved repeatedly for different right-hand sides or parameters. Such a task is highly time consuming, in particular in the case that the underlying equation is time-dependent or the spatial domain is three-dimensional which drastically increases the number of degrees of freedom. A typical way out of this problem is to apply model order reduction (MOR). The aim of MOR is to replace the high-dimensional original model by a suitable model with less degrees of freedom, the so-called reduced-order model. We mention, e.g., the recent monography series [**29, 27, 28**]. A prominent method of MOR for parabolic PDEs is the so-called proper orthogonal decomposition (POD) method; see, e.g., the survey [**121**] contained in [**27**]. This approach uses so-called snapshots of the dynamical system under consideration to construct a low-dimensional subspace of, e.g., a high-dimensional finite element space. More generally speaking, projection of a high-dimensional dynamical or parametric system onto smaller dimensional spaces leads to so-called reduced basis methods (RB); see, e.g., the monographies [**139, 233**]. Regarding applications of POD-MOR to optimal control we mention the survey articles [**131**] and [**246, 30**] on linear quadratic and nonlinear control problems, respectively.

Both POD and RB approaches are data-driven: the quality of the reduced-order model crucially depends on the data, i.e. the snapshots, used for the generation of the reduced-order model. More precisely, the low-dimensional subspaces generated from the data need to be in some sense capable of expressing the original, high-dimensional trajectory of the system sufficiently well. This issue poses one of

the main difficulties in the context of POD/RB methods and the following questions naturally arise when applying POD/RB-MOR, in particular in the context of PDE-constrained optimization: how can we estimate and/or control the additional errors arising from MOR in our computations? How can we couple MOR and numerical optimization when solving an optimal control problem numerically? These both questions are closely related to each other and have been subject to intensive research in the last years. Since there is a huge amount of literature about POD/RB-MOR we have to restrict ourselves to an incomplete literature overview focussed on the abovementioned two questions.

We start our literature overview on the first question with literature on a-priori error estimates for POD-MOR of equations. Such estimates have been obtained, e.g., in an abstract setting for linear and semilinear parabolic equations [**193, 79**], fluid dynamic equations [**194, 255**], linear wave equations [**79, 136**], for a nonlinear elliptic-parabolic system related to the modelling of lithium ion batteries [**198**], an equation related to laser surface hardening [**163**], and a partial integro-differential equation related to finance [**244**]. Due to the data-driven character of POD-MOR these a-priori error estimates contain quantities that are not explicitely computable in advance. Therefore, a-posteriori error estimates are needed in practical computations. We mention, e.g., such estimates in the RB-setting for quasilinear elliptic and parabolic equations related to magneto(quasi)statics [**150, 151, 186**], time-discrete nonlinear parabolic equations [**103**], and parametrized nonlinear parabolic equations related to lithium ion batteries [**173**]. Regarding control problems, a-priori POD error estimates have been obtained, e.g., for optimal control of linear parabolic equations [**160**] or for certain infinite horizon problems [**3**]. As for POD/RB-MOR of equations, a-posteriori error estimates are needed in applications. As far as we know, the first a-posteriori POD error estimates in optimal control have been obtained in [**272**] for linear quadratic parabolic problems. Meanwhile, the underlying perturbation approach has been extended to linear quadratic parabolic problems with mixed control-state-constraints [**130**], semilinear parabolic problems [**182**], or bilinear elliptic control problems [**181**]. A different, residual-based approach has been applied to semilinear parabolic control problems [**236**] and parameter optimization problems with linear elliptic and parabolic PDEs [**232**]. Parameter optimization problems with evolution equations or a nonlinear elliptic-parabolic system have been under consideration in [**96, 200**].

Let us now turn to the second question stated above, i.e. the coupling of POD-MOR and numerical optimization, which is an active area of research. Following the pioneering work on linear quadratic problems [**272**], the perturbation approach has directly been applied to, e.g., semilinear parabolic problems [**182**] or linear quadratic problems involving the Helmholtz equation [**265**]. Moreover, this idea has been applied to the linear quadratic subproblems during the application of the SQP method to bilinear elliptic [**181**] and semilinear parabolic [**116**] problems. Let us also mention that iterated solution of the POD-reduced optimal control problem under successive enrichment of the snapshot set with the states associated with the control iterates has been suggested in [**1**] for a flow control problem. Several trust-region frameworks for the coupling of POD/RB-MOR and numerical optimization have been developed for various problem types. In [**19, 38, 232**] the local model function in the trust-region algorithm is given by the reduced-order counterpart

of the original objective functional. While [**19, 252, 243, 38**] use balls, i.e. the classical choice, as trust-regions, the trust-region in [**232**] is defined by sublevel sets of a-posteriori MOR error estimates for the reduced functional. In [**236**] the second-order Taylor polynomial of the POD-reduced functional is used as local model function and the trust-regions are given by balls again. In [**183, 21**] the reduced-order counterpart of a nonconforming dual corrected version of the functional serves as local model and the trust-regions are chosen as in [**232**]. Finally, a completely different idea, so-called optimality system POD, has been developed in [**195, 282**], and a-priori convergence of this method in the case of linear parabolic problems has been proven in [**191**].

A related aspect is the interplay of POD/RB-MOR and discretization. We mention recent work on the combination of POD and adaptive space discretization for parametrized PDEs [**277**], evolution equations [**118**], incompressible flows [**119**], and optimal control of the Cahn-Hilliard system [**120**]. Adaptive selection of the snapshots has been addressed in [**5, 4**] for linear parabolic control problems. Moreover, the optimal location of snapshots has been discussed for time-dependent [**196**] or parametric [**199**] equations. Adaptive refinement of the discretization together with adaptive updates of the reduced-order model during numerical optimization has been proposed in [**38**] for control problems governed by the Navier-Stokes equations. Moreover, the balancing of POD and discretization errors for linear parabolic control problems has been studied in [**129**].

Finally, let us exemplarily cite recent applications of POD/RB-MOR to multiobjective optimization [**171, 172, 22**], mixed-integer optimal control problems [**20**], robust optimization [**6**], and model predictive control [**210, 8, 197**].

Let us now come back to the content of this chapter that consists of two contributions to model order reduction in the context of quasilinear parabolic PDEs and related control problems. The first one is concerned with POD-MOR for the equation (Eq); we achieve both theoretical and numerical results. Our second contribution is of purely numerical and experimental nature and deals with the numerical solution of the control problem (P) utilizing POD-MOR.

Our results on POD-MOR and corresponding a-posteriori POD/RB error estimates for (Eq) from [**166**] are the topic of Section 5.2. As a motivation for these results one may have in mind recent approaches for the coupling of numerical optimization and MOR like [**236, 232, 183, 21**] that require a-posteriori error estimation for the state and the adjoint equation. We fix a spatial discretization and consider the semidiscrete (in space) counterpart of (Eq) as our reference. Consequently, our a-posteriori error estimates include errors arising from RB/POR-MOR, so-called EIM-hyperreduction of the nonlinearity, and time discretization. The incorporation of time discretization errors in the a-posteriori error estimates is different from, e.g., [**103, 151, 232**] and has the advantage that it may prevent the choice of unnecessarily accurate reduced-order models below the time discretization error in practice. Let us point out further important features of our work. First of all, the application of POD/RB-MOR, which is a projection-based and hence linear technique, to nonlinear problems is known to be challenging, in general, and so-called hyperreduction techniques, such as, e.g., empirical interpolation (EIM), are required to allow for an efficient handling of the nonlinear terms within the reduced-order model. The particular structure of the nonlinearity in (Eq) adds

another difficulty: in fact, the presence of a nonmonotone nonlinearity is different from other publications concerned with POD/RB a-posteriori error estimates for nonlinear PDEs [**151, 150, 236**] and also poses the main challenge in our a-posteriori error analysis. In [**166**] we followed two different ideas to cope with the missing monotonicity. The two approaches result in error estimates of different structure and, interestingly, they are, on a heuristic level, related to the regularity setting assumed for (Eq) and the corresponding regularity results for the solutions; this will be explained in detail at the beginning of Section 5.2.2 and right before Theorem 5.9. For the sake of brevity, we will only discuss the details of the first approach in this thesis and restrict ourselves to stating the result of the second one. In the numerical examples we finally compare the results of both approaches and observe a significant speedup of the numerical solution of (Eq) due to POD-MOR.

In order to reproduce the results from [**236, 232, 183**] for our model problem, we would have to derive error estimates for the adjoint equation, too. We believe that the latter is particularly challenging as will be explained in more detail in Section 5.2.5. This issue motivates our second contribution that will be the topic of Section 5.3: we present numerical experiments concerned with a straightforward, heuristic, and estimate-free POD-Newton-SQP method that has some similarities to other ideas that target the combination of SQP-like methods with POD-MOR [**181, 116, 38**]. Our numerical findings indicate that this approach allows for a significant speedup of about factor 10 in the numerical solution of (P) and may be less sensitive w.r.t. initialization than methods relying on a single, a-priori fixed POD basis. In essence, for a Newton-like variant of the SQP method we solve the linear quadratic subproblems utilizing POD-reduction with a POD basis determined from the current full-order state and adjoint state. Consequently, we only need to apply MOR in a linear quadratic setting. Compared to a more elaborate, certified coupling between POD-MOR and numerical optimization, as, e.g., in the abovementioned literature, this purely heuristic approach has the advantage that we do not require a-posteriori error estimates for the POD error of the state and the adjoint equation. However, we cannot carry out a convergence analysis and do not reach the same level of efficiency as in [**116, 236, 232**].

This chapter is organized as follows. We start with a brief introduction to proper orthogonal decomposition in Section 5.1. Section 5.2 is devoted to POD-MOR for (Eq) and corresponding a-posteriori error estimates: first of all, we precisely introduce the semidiscrete (in space) counterpart of (Eq) that serves a reference object in the following. A-posteriori POD/RB error estimates are derived in Section 5.2.2. Additional hyperreduction of the nonlinearity by empirical interpolation (EIM) and the incorporation of the respective errors into the error estimates are addressed in Section 5.2.3. In Section 5.2.4 we illustrate our results concerning POR-MOR of (Eq) numerically and put them into the context of optimal control in Section 5.2.5. The final Section 5.3 contains our numerical experiments concerning the POD-reduced solution of the entire control problem.

5.1. Proper orthogonal decomposition

For convenience of the reader we provide a short introduction to the main ideas of proper orthogonal decomposition in this section. Proper orthogonal decomposition (POD) goes back to applications in fluid dynamics [**256**] and has successfully

been applied in different context. In particular, the same or similar techniques related to the mathematical concept of singular value decomposition (SVD; see, e.g., [**290**, Theorem VI.3.6]) are also known as principal component analysis (PCA) in the data science community [**227, 170, 179**]. We give a short, and therefore necessarily incomplete, introduction to the main ideas in the following. For the remaining part of this section we follow the exposition in [**194**, Section 3.1] and refer the reader to, e.g., [**131**] for more details.

Let V be a separable Hilbert space and suppose that $y_1, \dots, y_n \in V$, the so-called snapshots, together with some weights $\alpha_1, \dots, \alpha_n > 0$ are given. The snapshots generate a finite-dimensional subspace $S := \operatorname{span}\{y_1, \dots, y_n\} \subset V$ with dimension $\dim S \leq n$. A POD basis of rank ℓ, $1 \leq \ell \leq \dim S$, is given by ℓ orthonormal vectors $\psi_1, \dots, \psi_\ell \in V$ that solve the following optimization problem:

$$\left\{ \begin{array}{l} \displaystyle\min_{(\psi_1,\dots,\psi_\ell) \in \bigtimes_{i=1}^{\ell} V} \sum_{j=1}^{n} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_V \psi_i \right\|_V^2 \\ \text{subject to} \quad \langle \psi_i, \psi_j \rangle_V = \delta_{ij}, \quad 1 \leq i \leq \ell,\ 1 \leq j \leq i. \end{array} \right. \tag{5.1}$$

Due to

$$\sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_V \psi_i = \operatorname{argmin}_{z \in \operatorname{span}\{\psi_1,\dots,\psi_\ell\}} \|y_j - z\|_V,$$

the POD basis of rank ℓ allows for the best simultanuous approximation of the snapshots (weighted by the α_j) among all orthonormal bases of length ℓ in V. To compute such a POD basis the following linear operators are introduced:

$$\begin{aligned} &\mathcal{K}\colon \mathbb{R}^n \to V, && x \mapsto \sum_{j=1}^{n} \alpha_j x_j y_j, \\ &\mathcal{K}^*\colon V \to \mathbb{R}^n, && v \mapsto (\langle v, y_j \rangle_V)_{i=1,\dots,n}, \\ &\mathcal{R}\colon V \to V, && v \mapsto \mathcal{K}\mathcal{K}^* v = \sum_{j=1}^{n} \alpha_j \langle v, y_j \rangle_V y_j, \\ &\mathcal{Q}\colon \mathbb{R}^n \to \mathbb{R}^n, && \mathcal{Q} = \mathcal{K}^*\mathcal{K} = (\langle y_i, y_j \rangle_V)_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}. \end{aligned}$$

Here, we equipp $\mathbb{R}^n$ with the scalar product $\langle x^1, x^2 \rangle := \sum_{j=1}^{n} \alpha_j x_j^1 x_j^2$. We note that $\mathcal{R}$ is bounded, selfadjoint, nonnegative, and has finite-dimensional range $\operatorname{range}(\mathcal{R}) = S$. Consequently, it is a compact operator and there are an orthonormal basis $(\psi_i)_{i \in \mathbb{N}}$ for V and eigenvalues $\lambda_1 \geq \dots \geq \lambda_{\dim S} > 0$, $\lambda_i = 0$ for $i > \dim S$, such that $S = \operatorname{span}\{\psi_1, \dots, \psi_{\dim S}\}$ and

$$\mathcal{R}\psi_i = \lambda_i \psi_i \qquad \forall i \in \mathbb{N}.$$

It turns out that these eigenvectors ψ_i and the POD basis of rank ℓ for the snapshots $y_1, \dots, y_n$ are closely related.

Proposition 5.1. *For any $1 \leq \ell \leq \dim S$, the POD basis of rank ℓ for the snapshots $y_1, \dots, y_n$ is given by the first ℓ eigenvectors $\psi_1, \dots, \psi_\ell$ of $\mathcal{R}$. Moreover, the following estimate holds true:*

$$\sum_{j=1}^{b} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_V \psi_i \right\|_V^2 = \sum_{i=\ell+1}^{\dim S} \lambda_i.$$

The error estimate given in this proposition is often used to determine the length of a POD basis. Assume that we are given a tolerance $\epsilon_{\mathrm{POD}} > 0$ that sets an upper bound on how much "information" contained in the snapshots we want give up when projecting onto a POD basis of rank ℓ. Since it holds $\sum_{j=1}^{n} \alpha_j \|y_j\|_V^2 = \sum_{i=1}^{\dim S} \lambda_i$, one may therefore determine the length of the POD basis as the minimal ℓ that satisfies

$$\sum_{i=\ell+1}^{\dim S} < \epsilon_{\mathrm{POD}} \cdot \sum_{i=1}^{\dim S} \lambda_i.$$

For practical computations, the following relation of the POD basis to the eigensystem of the matrix $\mathcal{Q}$ and to the singular systems of $\mathcal{K}$ and $\mathcal{K}^*$ is of interest. The n eigenvalues (in descending order) of the the symmetric, positive semidefinite matrix $\mathcal{Q} \in \mathbb{R}^{n \times n}$ coincide with the first n eigenvalues of $\mathcal{R}$. Moreover, if $v_1, ..., v_n \in \mathbb{R}^n$ denote the corresponding, pairwise orthonormal eigenvectors of $\mathcal{Q}$, the following relations hold true:

$$v_i = \frac{1}{\sqrt{\lambda_i}} \mathcal{K}^* \psi_i, \qquad \psi_i = \frac{1}{\sqrt{\lambda_i}} \mathcal{K} v_i, \qquad \sum_{i=1}^{\dim S} \lambda_i = \mathrm{trace}(\mathcal{Q}).$$

In particular, the eigenvalues and eigenvectors of $\mathcal{R}$ can be computed with the help of $\mathcal{Q}$ and the other way round. Moreover, the singular value decomposition (SVD) of $\mathcal{K}$ is given by

$$\mathcal{K} x := \sum_{i=1}^{n} \sqrt{\lambda_i} x^T v_i \psi_i,$$

and an analogous expression holds true for $\mathcal{K}^*$. Consequently, the POD basis can also be computed as the first ℓ singular vectors of $\mathcal{K}$. All four approaches, i.e. computation of the eigensystem of $\mathcal{Q}$ or $\mathcal{R}$ and computation of the SVD of $\mathcal{K}$ or $\mathcal{K}^*$ can be used in practice; the approach based on the eigensystem of $\mathcal{Q}$ is particularly well-known as "method of snapshots". In the case of rapidly decaying eigenvalues, however, SVD-based approaches may be advantageous since the singular values of $\mathcal{K}$ and $\mathcal{K}^*$ drop below machine precision later than the eigenvalues of $\mathcal{R}$ or $\mathcal{Q}$ do.

Let us briefly comment on the relation to the so-called continuous version of POD which helps to explain the meaning of the weights α_j. Let $I = (0, T)$ be an interval with a partition $0 = t_0 < t_1 < ... < t_{n-1} < t_n = T$ and $\alpha_j = t_j - t_{j-1}$ for $j = 1, ..., n$. If we define the piecewise constant trajectory $y \in L^2(I, V)$ by $y = \sum_{j=1}^{n} \mathbf{1}_{(t_{j-1}, t_j]} y_j$, we can rewrite (5.1) equivalently as

$$\left\{ \begin{array}{l} \displaystyle\min_{(\psi_1, ..., \psi_\ell) \in \times_{i=1}^{\ell} V} \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i \rangle_V \psi_i \right\|_V^2 \mathrm{d}t \\ \text{subject to} \quad \langle \psi_i, \psi_j \rangle_V = \delta_{ij}, \quad 1 \le i \le \ell, 1 \le j \le i. \end{array} \right. \tag{5.2}$$

Therefore, in the case of time-dependent problems in which the snapshots $y_1, ..., y_n$ may be viewed as given by snapshots of a trajectory $I \to V$, the weights α_j will be related to quadrature weights related to the integral on I. The observation that leads to the continuous version of POD is the following: the problem (5.2) can also be considered for an arbitrary (i.e. not necessarily piecewise constant) trajectory $y \in L^2(I, V)$. In that case, $\mathbb{R}^n$ has to be replaced in the above considerations by the infinite-dimensional Hilbert space $L^2(I)$; retrospectively, this explains why

$\mathbb{R}^n$ has been equipped with the weighted scalar product introduced above. One obtains, e.g., the following expressions for $\mathcal{K}$ and $\mathcal{R}$:

$$\mathcal{K}\colon L^2(I) \to V, \qquad \varphi \mapsto \int_I \varphi(t) y(t) \mathrm{d}t,$$

$$\mathcal{R}\colon V \to V, \qquad v \mapsto \mathcal{K}\mathcal{K}^* v = \int_I \langle v, y(t)\rangle_V y(t) \mathrm{d}t.$$

Hilbert-Schmidt theory still allows to prove existence of eigensystems and singular systems as before, with the main difference that now infinitely many eigen- and singularvalues will be positive, in general. For details on this and on the transition from finitely many snapshots to continuous POD we refer the reader to, e.g., [**194**, Section 3.2] or [**131**].

5.2. POD-MOR for the state equation

In this section we present the results obtained in [**166**] on POD/RB-MOR for the state equation and corresponding a-posteriori POD/RB error estimates. We consider the equation

$$\begin{cases} \partial_t y + \mathcal{A}(y) y = f, & \text{in } L^s(I, W_{\Gamma_D}^{-1,p}), \\ y(0) = y_0, & \text{in } (W_{\Gamma_D}^{-1,p}, W_{\Gamma_D}^{1,p})_{1/s',s}, \end{cases} \tag{5.3}$$

with $f \in L^s(I, W_{\Gamma_D}^{-1,p})$, $y_0 \in (W_{\Gamma_D}^{-1,p}, W_{\Gamma_D}^{1,p})_{1/s',s}$ and rely on Assumptions 1.5, 1.6 and 1.8. Recall that results on existence and regularity of solutions to this equation have been collected in Chapter 1; see in particular Theorems 1.11, 1.14 and 1.24 and Theorem 2.20 in Chapter 2. Moreover, let us point out that troughout this section we are concerned with this equation only and not with an optimal control problem. Our main results on a-posteriori POD/RB error estimates for (5.3) are Theorems 5.6 and 5.9 in Section 5.2.3 below and a numerical illustration will be given in Section 5.2.4. Finally, we will put our results into the context of optimal control in Section 5.2.5. In the following we always refer to POD/RB because our arguments and results apply to general RB methods although we clearly have the particular case of POD in mind.

Before going into the details, let us briefly comment on the nonmonotone structure of the nonlinearity in (5.3). The main difficulty as well as the main novelty in the following a-posteriori error analysis arise from this fact. Recall that a nonlinear operator $\mathcal{N}\colon X \to X^*$ on a Banach space X is called monotone if

$$\langle \mathcal{N}(x) - \mathcal{N}(y), x - y\rangle_{X^*,X} \geq 0 \qquad \forall x, y \in X,$$

and strongly monotone if there exists a constant $c > 0$ such that

$$\langle \mathcal{N}(x) - \mathcal{N}(y), x - y\rangle_{X^*,X} \geq c\|x - y\|_X^2 \qquad \forall x, y \in X;$$

cf., e.g., [**240, 292**] for this notion and its application in the theory of nonlinear PDEs. It has turned out that exploitation of strong monotonicity of the nonlinear terms is also an important step in the derivation of RB a-posteriori error estimates for semilinear parabolic [**236**], quasilinear elliptic [**150**], and quasilinear parabolic [**151**] PDEs. Note that the quasilinear nonlinearities in [**151, 150, 186**] refer to problems from magneto(quasi)statics and depend on the gradient of the solution. The nonlinear operator $H_{\Gamma_D}^1 \to H_{\Gamma_D}^{-1}$ under consideration in the present thesis,

however, is given by the map $y \mapsto \mathcal{A}(y)y = -\nabla \cdot \xi(y)\mu\nabla y$, and hence it depends also on the solution y and not on its gradient ∇y only. The counterexample [**111**, Example 8.18] shows that this changes the structure of the nonlinearity. It cannot be expected to be monotone, and therefore it is essentially different from those considered in, e.g., [**151, 150, 236**]. In fact, the main difficulty in the derivation of RB a-posteriori error estimates will be to find a workaround for the missing strong monotonicity of our nonlinearity. In our paper [**166**] we presented two different approaches to do so; their relation will be discussed in more detail below. We will work out the first approach in detail in Section 5.2.2 while for the second one we will only provide a summary of the main result and numerical examples.

5.2.1. Semidiscretization in space, truth solution, and POD/RB-MOR. We follow Section 2.2 of our paper [**166**] and introduce the semidiscrete (in space) counterpart of (5.3). Its solution will serve as the so-called truth-solution, i.e. the reference solution to which the a-posteriori error estimates will refer to. In particular, this means that we do not address spatial discretization errors. Moreover, we also introduce the POD/RB-reduced counterpart of the semidiscrete (in space) equation.

Let V^h be an $H^1_{\Gamma_D}$-conforming finite element space on Ω and $I^h\colon H^1_{\Gamma_D} \to V^h$ be an appropriate interpolation operator. We may introduce the semidiscrete (in space) counterpart of (5.3) as follows: Find $y^h \in \mathbb{W}^{1,2}(I,((V^h)^*,V^h))$ such that

$$(5.3_h) \qquad \begin{cases} \langle \partial_t y^h(t) + \mathcal{A}(y^h(t))y^h(t), \varphi^h\rangle_{H^{-1}_{\Gamma_D},H^1_{\Gamma_D}} = \langle f(t),\varphi^h\rangle_{H^{-1}_{\Gamma_D},H^1_{\Gamma_D}} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall t \in I, \varphi^h \in V^h, \\ y^h(0) = I^h y_0. \end{cases}$$

Due to finite-dimensionality of V^h, equation (5.3_h) results in a system of ordinary differential equations (ODEs) for the coefficients of y^h w.r.t. some basis of V^h. This allows to discuss existence of solutions to (5.3_h), even without using all parts of Assumptions 1.5, 1.6 and 1.8.

Proposition 5.2 ([**166**], Proposition 2.2)**.** *Assume that V^h is a finite-dimensional subspace of $H^1_{\Gamma_D}(\Omega)\cap C(\overline{\Omega})$. For any right-hand side $f \in L^2(I,H^{-1}_{\Gamma_D})$ and initial value $I^h y_0 \in V^h$ there exists a unique solution $y^h\colon I \to V^h$ of (5.3_h) such that*

$$\|y^h\|_{L^\infty(I,L^2)\cap L^2(I,H^1_{\Gamma_D})} + \|y^h\|_{\mathbb{W}^{1,2}(I,((V^h)^*,V^h))} \le C\left(\|f\|_{L^2(I,H^{-1}_{\Gamma_D})} + \|I^h y_0\|_{L^2}\right),$$

with a constant $C > 0$ independent of V^h, f and $I^h y_0$.

For the rest of the chapter we will refer to y^h as the semidiscrete (in space) solution to (5.3), or, shorter, the truth-solution. By y we will denote the continuous in space and time solution to (5.3), short: the true solution.

A finite element error analysis for y^h in the setting of Assumptions 1.5, 1.6 and 1.8 without additional suppositions has not been carried out in the existing literature. Under the regularity assumptions of [**45**], cf. Section 1.6 or Assumption 2.19, error estimates for the equation from [**45**], cf. (1.22), have been obtained in [**46**]. We refer to the references therein or to the introduction of [**99**] for an overview of further results under different additional assumptions. Let us only

mention two results: an $L^\infty(I,L^2)\cap L^2(I,H^1)$-quasi-bestapproximation result has been derived in [**101**] with a technique slightly related to the one applied in Section 5.2.2. In [**99**] for instance pointwise error estimates have been obtained.

Proof of Proposition 5.2. Let $(\varphi_i)_{i=1,\dots,N_h}$, $N_h = \dim V^h$, be a basis for V^h. Writing $y^h(t) = \sum_{i=1}^{N_h} \mathbf{x}_i(t)\varphi_i$, the coefficient vector $\mathbf{x}(t)\in\mathbb{R}^{N_h}$ fulfills

$$\mathbf{M}\partial_t \mathbf{x}(t) + \mathbf{A}(\mathbf{x}(t))\mathbf{x}(t) = \mathbf{f}(t), \qquad \mathbf{x}(0) = \mathbf{x}_0, \tag{5.4}$$

with

$$\begin{aligned}
\mathbf{M} &:= \left(\langle \varphi_i, \varphi_j\rangle_{L^2}\right)_{i,j},\\
\mathbf{A}(\mathbf{z}) &:= \left(\int_\Omega \xi\left(\sum_{n=1}^{N_h}\mathbf{z}_n\varphi_n\right)\mu\nabla\varphi_i\nabla\varphi_j dx\right)_{i,j},\\
\mathbf{f}(t) &:= \left(\langle f(t),\varphi_i\rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1}\right)_i,
\end{aligned}$$

and $\mathbf{x}_0$ being the coefficient vector of $I^h y_0$. Using continuity of the basis functions as well as global Lipschitz continuity and boundedness of ξ, one easily verifies that (5.4) satisfies the assumptions of [**262**, Lemma 5.7]. Therefore, existence and uniqueness of a solution $\mathbf{x}\in W^{1,2}(I,\mathbb{R}^{N_h})$ to (5.4) follows, which implies existence and uniqueness of a solution $y^h\in W^{1,2}(I,V^h)$ to (5.3_h). The estimate follows by standard techniques: testing (5.3_h) with $\varphi^h = y^h$ and integrating by parts yields the first summand of the estimate. The second summand is obtained by testing with an arbitray $\varphi^h\in L^2(I,V^h)$. □

Note that the embedding $V^h\hookrightarrow C(\overline{\Omega})$ is crucial in the previous argument. It ensures Lipschitz continuity of $\mathbf{A}(\cdot)$ and therefore existence of a solution to (5.4) via a generalized Picard-Lindelöf principle. If V^h is a classical continuous Lagrange finite element space on a polygonal (polyhedral) domain Ω equipped with a triangular (tetrahedral) mesh, this assumption is obviously fulfilled. However, we would like to point out that except for the assumptions from Proposition 5.2 we do not rely on further details of spatial discretization.

Performing expensive computations in the high-dimensional space V^h can be avoided by application of the reduced basis approach. We replace the finite element space V^h by a much smaller ℓ-dimensional subspace $V^{h,\ell}\subset V^h$ that is related to the physical properties of the system and might be determined by the well-known POD approach; cf. Section 5.1. Although our arguments do not rely on the particular choice of $V^{h,\ell}$ and therefore also cover general RB methods, we clearly have in mind $V^{h,\ell}$'s obtained by POD and also restrict our numerical experiments in Section 5.2.4 to this case. Having at hand a reduced ansatz space $V^{h,\ell}\subset V^h$ and a suitable projection $P^{h,\ell}\colon V^h\to V^{h,\ell}$, usually the L^2- or H^1-orthogonal projection, we introduce the reduced-order counterpart of (5.3_h) as follows: find $y^{h,\ell}\in\mathbb{W}^{1,2}(I,((V^{h,\ell})^*,V^{h,\ell}))$ such that

$$\left\{\begin{aligned}
\langle\partial_t y^{h,\ell}(t) + \mathcal{A}(y^{h,\ell}(t))y^{h,\ell}(t),\varphi^{h,\ell}\rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} &= \langle f(t),\varphi^{h,\ell}\rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1}\\
&\qquad \forall t\in I, \varphi^{h,\ell}\in V^{h,\ell},\\
y^{h,\ell}(0) &= P^{h,\ell}I^h y_0.
\end{aligned}\right. \tag{5.3$_{h,\ell}$}$$

A reader familiar with ROM techniques may already have noticed that the nonlinear term in $(5.3_{h,\ell})$ does not allow for efficient evaluation within the reduced-order model. We have to overcome this issue by so-called hyperreduction techniques, e.g., the empirical interpolation method (EIM), which will be addressed in Section 5.2.3.

5.2.2. A-posteriori POD/RB error estimates. We now state and prove our first main contribution of this chapter: a-posteriori error estimates for (5.3_h) including both reduced-order and time discretization errors. For the reason of clarity we exclude hyperreduction for the nonlinearity at this point, and address this issue subsequently. This section is based on [**166**, Section 3].

We roughly follow the ansatz of [**236**], where a semilinear equation with monotone nonlinearity has been discussed. To overcome the difficulties arising from the fact that our nonlinearity is not monotone, we exploit $L^\infty(I, W^{1,\infty})$-regularity of the truth-solution y^h and obtain explicit estimates of classical structure in terms of the error in the initial condition and the $(V^h)^*$-residual of the discrete solution under consideration. As a semidiscrete in space solution, y^h obviously exhibits the required regularity for any fixed (spatial) discretization level. However, since the error estimates will depend on the value of the $L^\infty(I, W^{1,\infty})$-norm of y^h it is desirable to have uniform bounds for this norm for all sufficiently fine spatial discretization levels. We believe that we can only expect such a uniform bound if the continuous in space and time solution of (5.3) exhibits $L^\infty(I, W^{1,\infty})$-regularity, which is guaranteed, e.g., in the smooth setting from [**45**], i.e. under Assumption 2.19; cf. Theorem 1.24 and Corollary 2.21. For the reason of brevity, we focus on the details of this approach in the present thesis. The main result of our second, alternative approach from [**166**] that exploits less regularity of the truth-solution will be summarized in Theorem 5.9 at the end of Section 5.2.3 below.

We start by fixing the following notation and assumptions.

Assumption 5.3.

1. Assume that $V^h \subset H^1_{\Gamma_D} \cap C(\overline{\Omega}) \cap W^{1,\infty}(\Omega)$ is an N_h-dimensional conforming finite element space, and $V^{h,\ell}$ a ℓ-dimensional subspace of V^h. By $y^h \in \mathbb{W}^{1,2}(I, ((V^h)^*, V^h))$ we denote the truth-solution, i.e. the unique solution to (5.3_h).
2. Moreover, let $y^{h,\ell} \in \mathbb{W}^{1,2}(I, ((V^{h,\ell})^*, V^{h,\ell}))$ be arbitrary. By $\mathrm{e}_y^{h,\ell} := y^{h,\ell} - y^h$ we denote the error with respect to the truth-solution.
3. We assume that ξ is globally Lipschitz continuous and denote the global Lipschitz constant of ξ by $|\xi'|_\infty$.

We have in mind the following situation: $y^{h,\ell}$ is the solution of a time-discrete counterpart of $(5.3_{h,\ell})$, and we want to estimate how good $y^{h,\ell}$ approximates the truth-solution y^h. Note that in order to ensure that $y^{h,\ell}$ meets the regularity requirements of Assumption 5.3 we have to choose a time discretization for $(5.3_{h,\ell})$ that results in sufficiently regular solutions, e.g., the Crank-Nicolson scheme in its CG1-DG0 Petrov-Galerkin form. Time-discrete solutions of $(5.3_{h,\ell})$ obtained by discontinuous Galerkin time discretization, e.g., backward/implicit Euler, do not fulfill Assumption 5.3. Since discontinuous time discretization might be of particular interest in the context of PDE-constrained optimization we will outline an approach to overcome this restriction in Remark 5.8.

We start with some prelimiary calculations and follow the residual-based ansatz of [**236**] as far as possible without modification, i.e. up to the point where strong monotonicity of the nonlinearity would be required. First, we introduce the residual of $y^{h,\ell}$ by

$$\mathrm{r}_y^{h,\ell}(t) := \partial_t y^{h,\ell}(t) + \mathcal{A}(y^{h,\ell}(t))y^{h,\ell}(t) - f(t) \in (V^{h,\ell})^* \hookrightarrow H_{\Gamma_D}^{-1}, \quad t \in I. \tag{5.5}$$

To keep notation short we will omit the argument "t" in the following. A short computation utilizing (5.3_h) shows that

$$\begin{aligned}\langle \mathrm{r}_y^{h,\ell}, \varphi^h \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} &= \langle \partial_t \mathrm{e}_y^{h,\ell}, \varphi^h \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} \\ &\quad + \langle \mathcal{A}(y^{h,\ell})y^{h,\ell} - \mathcal{A}(y^h)y^h, \varphi^h \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1}\end{aligned} \tag{5.6}$$

holds for all $\varphi^h \in V^h$. We consider V^h as a vector space canonically equipped with the $H_{\Gamma_D}^1$-norm. Therefore, its dual $(V^h)^*$ is canonically equipped with the following norm:

$$\|\ell_h\|_{(V^h)^*} := \sup_{0 \neq \psi_h \in V^h} \frac{\langle \ell_h, \psi_h \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1}}{\|\psi_h\|_{H_{\Gamma_D}^1}} = \sup_{0 \neq \psi_h \in V^h} \frac{\ell_h(\psi_h)}{\|\psi_h\|_{H_{\Gamma_D}^1}}. \tag{5.7}$$

Note that this norm is not equal to the $H_{\Gamma_D}^{-1}$-norm, because we only test with elements ψ_h from V^h in (5.7). For later use we state the following observation.

Lemma 5.4 ([**166**], Lemma 3.2)**.** *Let Assumption 5.3 hold. Then, the function $I \to \mathbb{R}$, $t \mapsto \|\mathrm{r}_y^{h,\ell}(t)\|_{(V^h)^*}^2$ is well-defined a.e. on I and L^2-integrable.*

Proof. This follows from the definition of $\mathrm{r}_y^{h,\ell}$ and the regularity assumed for $y^{h,\ell}$. □

Plugging in $\varphi^h = \mathrm{e}_y^{h,\ell}(t)$ for every fixed t in (5.6), and using the classical integration by parts formula from [**240**, Remark 7.5] we obtain

$$\frac{d}{dt}\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \langle \mathcal{A}(y^{h,\ell})y^{h,\ell} - \mathcal{A}(y^h)y^h, \mathrm{e}_y^{h,\ell} \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} = \langle \mathrm{r}_y^{h,\ell}, \mathrm{e}_y^{h,\ell} \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1}. \tag{5.8}$$

Note that the second summand on the left-hand side of (5.8) causes problems in our case: If the nonlinearity $y \mapsto \mathcal{A}(y)y$ was strongly monotone, we could proceed as done in [**236**] for a semilinear term and estimate as follows:

$$\langle \mathcal{A}(y^{h,\ell})y^{h,\ell} - \mathcal{A}(y^h)y^h, \mathrm{e}_y^{h,\ell} \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} \geq c|\mathrm{e}_y^{h,\ell}|_{H_{\Gamma_D}^1}^2.$$

However, as pointed out at the beginning of Section 5.2 such an estimate cannot be expected to hold true. We cannot even bound the term under consideration from below by zero. Therefore, we have to proceed in a different way and split the problematic term into a coercive part and a remainder as follows:

$$\begin{aligned}&\langle \mathcal{A}(y^{h,\ell})y^{h,\ell} - \mathcal{A}(y^h)y^h, y^{h,\ell} - y^h \rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} \\ &= \int_\Omega (\xi(y^{h,\ell})\mu\nabla y^{h,\ell} - \xi(y^h)\mu\nabla y^h)\nabla(y^{h,\ell} - y^h)\mathrm{d}x \\ &= \int_\Omega \xi(y^{h,\ell})\mu\nabla \mathrm{e}_y^{h,\ell}\nabla \mathrm{e}_y^{h,\ell}\mathrm{d}x + \int_\Omega (\xi(y^{h,\ell}) - \xi(y^h))\mu\nabla y^h \nabla \mathrm{e}_y^{h,\ell}\mathrm{d}x\end{aligned}$$

$$\geq \xi_\bullet \mu_\bullet |\mathrm{e}_y^{h,\ell}|^2_{H^1_{\Gamma_D}} + \int_\Omega (\xi(y^{h,\ell}) - \xi(y^h)) \mu \nabla y^h \nabla \mathrm{e}_y^{h,\ell} \mathrm{d}x.$$

Plugging this into (5.8) yields

$$\begin{aligned}(5.9)\quad \frac{d}{dt}\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|^2_{L^2} + \xi_\bullet \mu_\bullet |\mathrm{e}_y^{h,\ell}|_{H^1_{\Gamma_D}} \\ \leq \langle \mathrm{r}_y^{h,\ell}, \mathrm{e}_y^{h,\ell} \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} - \int_\Omega (\xi(y^{h,\ell}) - \xi(y^h)) \mu \nabla y^h \nabla \mathrm{e}_y^{h,\ell} \mathrm{d}x,\end{aligned}$$

i.e. except for the remainder term that we have shifted to the right-hand side we have preserved a similar structure as in [**236**]. In [**166**] formula (5.9) serves as the common basis for our two different approaches for obtaining a-posteriori error estimates. The main challenge in both cases is to estimate the second summand on the right-hand side in (5.9) in such a way that Gronwall's Lemma or a similar comparison principle can be applied to the resulting inequality. As already explained at the beginning of this section, we will only explain our first approach from [**166**] in detail in this thesis. It is closer to [**236**] than the second one, and relies on $L^\infty(I, W^{1,\infty})$-regularity of the truth-solution. An analogous regularity assumption on the true solution y and a similar estimate of the nonlinear term as below have been used in [**101**] in the context of finite element errors.

Theorem 5.5 ([**166**], Theorem 3.3)**.** *Let Assumptions 1.5, 1.6, 1.8 and 5.3 hold, and let $c_\infty > 0$ be such that*

$$|y^h(t)|_{W^{1,\infty}} \leq c_\infty \qquad \forall t \in I.$$

Moreover, let $\varepsilon, \eta > 0$ be chosen such that

$$\eta + \varepsilon |\xi'|_\infty \mu^\bullet c_\infty = \xi_\bullet \mu_\bullet$$

and define $\beta := 2\left(\frac{1}{2\varepsilon}|\xi'|_\infty \mu^\bullet c_\infty + \xi_\bullet \mu_\bullet\right)$. Then, the following a-posteriori error estimates for $y^{h,\ell}$ hold true:

$$(5.10)\qquad \|\mathrm{e}_y^{h,\ell}(t)\|^2_{L^2} \leq e^{\beta t}\|y^{h,\ell}(0) - y^h(0)\|^2_{L^2} + \eta^{-1}\int_0^t e^{\beta(t-s)}\|\mathrm{r}_y^{h,\ell}(s)\|^2_{(V^h)^*}\mathrm{d}s,$$

$$\begin{aligned}(5.11)\quad \|\mathrm{e}_y^{h,\ell}\|^2_{L^2(I,L^2)} \leq \beta^{-1}\left(e^{\beta T} - 1\right)\|y^{h,\ell}(0) - y^h(0)\|^2_{L^2} \\ + \eta^{-1}\beta^{-1}\int_0^T (e^{\beta(T-t)} - 1)\|\mathrm{r}_y^{h,\ell}(t)\|^2_{(V^h)^*}\mathrm{d}t.\end{aligned}$$

$$\begin{aligned}(5.12)\quad \|\mathrm{e}_y^{h,\ell}\|^2_{L^2(I,H^1_{\Gamma_D})} \leq \xi_\bullet^{-1}\mu_\bullet^{-1}e^{\beta T}\|y^{h,\ell}(0) - y^h(0)\|^2_{L^2} \\ + \xi_\bullet^{-1}\mu_\bullet^{-1}\eta^{-1}\int_0^T e^{\beta(T-t)}\|\mathrm{r}_y^{h,\ell}(t)\|^2_{(V^h)^*}\mathrm{d}t.\end{aligned}$$

Proof. We proceed with the above computations. Starting with the estimate (5.9) we bound the remaining term of the nonlinearity in the following way:

$$(5.13)\qquad \left|\int_\Omega (\xi(y^{h,\ell}) - \xi(y^h))\mu\nabla y^h \nabla \mathrm{e}_y^{h,\ell}\mathrm{d}x\right| \leq |\xi'|_\infty \mu^\bullet c_\infty \|\mathrm{e}_y^{h,\ell}\|_{L^2}\|\mathrm{e}_y^{h,\ell}\|_{H^1_{\Gamma_D}}.$$

Using $W^{1,\infty}$-regularity for y^h we can estimate one of the $\mathrm{e}_y^{h,\ell}$-factors in the L^2-norm, which would not be possible assuming $W^{1,p}$-regularity for u_h with some

finite p only. With the help of Young's inequality we arrive at

$$\frac{d}{dt}\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \xi_\bullet\mu_\bullet\|\mathrm{e}_y^{h,\ell}\|_{H^1_{\Gamma_D}}^2 \leq \xi_\bullet\mu_\bullet\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \langle \mathrm{r}_y^{h,\ell}, \mathrm{e}_y^{h,\ell}\rangle_{H^{-1}_{\Gamma_D},H^1_{\Gamma_D}} + |\xi'|_\infty\mu^\bullet c_\infty\left(\frac{1}{2\varepsilon}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \frac{\varepsilon}{2}\|\mathrm{e}_y^{h,\ell}\|_{H^1_{\Gamma_D}}^2\right)$$

with some $\varepsilon > 0$. Another application of Young's inequality yields

$$\frac{d}{dt}\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \xi_\bullet\mu_\bullet\|\mathrm{e}_y^{h,\ell}\|_{H^1_{\Gamma_D}}^2 \leq \left(\frac{1}{2\varepsilon}|\xi'|_\infty\mu^\bullet c_\infty + \xi_\bullet\mu_\bullet\right)\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \frac{1}{2\eta}\|\mathrm{r}_y^{h,\ell}\|_{(V^h)^*}^2 + \left(\frac{\eta}{2} + \frac{\varepsilon}{2}|\xi'|_\infty\mu^\bullet c_\infty\right)\|\mathrm{e}_y^{h,\ell}\|_{H^1_{\Gamma_D}}^2$$

with $\eta > 0$. Now, we choose η, ε as in the statement of the theorem and obtain

$$\frac{d}{dt}\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \frac{1}{2}\xi_\bullet\mu_\bullet\|\mathrm{e}_y^{h,\ell}\|_{H^1_{\Gamma_D}}^2 \leq \beta\cdot\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \frac{1}{2\eta}\|\mathrm{r}_y^{h,\ell}\|_{(V^h)^*}^2, \tag{5.14}$$

where we will use from now on the abbreviation $\beta = 2\left(\frac{1}{2\varepsilon}|\xi'|_\infty\mu^\bullet c_\infty + \xi_\bullet\mu_\bullet\right)$ to enhance readability. With the help of Gronwall's Lemma [**102**, Corollary 2] we obtain an a-posteriori estimate for the $L^\infty(I, L^2)$-error from this:

$$\|\mathrm{e}_y^{h,\ell}(t)\|_{L^2}^2 \leq \|P_n I^h y_0 - I^h y_0\|_{L^2}^2 e^{\beta t} + \eta^{-1}\int_0^t \|\mathrm{r}_y^{h,\ell}(s)\|_{(V^h)^*}^2 e^{\beta(t-s)}\mathrm{d}s.$$

The second summand thereof is integrated using integration by parts, i.e.

$$\int_0^T e^{\beta t}\left(\int_0^t e^{-\beta s}\|\mathrm{r}_y^{h,\ell}(s)\|_{(V^h)^*}^2\mathrm{d}s\right)\mathrm{d}t = \beta^{-1}\int_0^T (e^{\beta(T-t)} - 1)\|\mathrm{r}_y^{h,\ell}(t)\|_{(V^h)^*}^2\mathrm{d}t,$$

and together with the first summand we obtain the $L^2(I, L^2)$-estimate (5.11). As in [**236**] the $L^2(I, H^1)$-estimate (5.12) is obtained from (5.14) by integrating with respect to time over I and using (5.11). □

To conclude this section, let us note that by a slight modification of this approach it is possible to exploit less regularity of y^h. The price to pay is that the constants in the resulting estimates cannot be computed explicitely, in general. More precisely, the unknown referee of [**166**] suggested the following modification based on the Ladyzhenskaya-Gagliardo-Nirenberg inequality; cf. the end of [**166**, Section 3.2]: if $d = 2$ and $y^h \in L^\infty(I, W^{1,4})$, estimate (5.13) can be replaced by

$$\left|\int_\Omega (\xi(y^{h,\ell}) - \xi(y^h))\mu\nabla y^h\nabla\mathrm{e}_y^{h,\ell}\mathrm{d}x\right| \leq |\xi'|_\infty\mu^\bullet\|\mathrm{e}_y^{h,\ell}\|_{L^4}|y^h|_{W^{1,4}}|\mathrm{e}_y^{h,\ell}|_{H^1}.$$

In order to apply Gronwall's Lemma as before one has to observe

$$\begin{aligned}\|\mathrm{e}_y^{h,\ell}\|_{L^4}|y^h|_{W^{1,4}}|\mathrm{e}_y^{h,\ell}|_{H^1} &\leq \frac{1}{2\epsilon}\|\mathrm{e}_y^{h,\ell}\|_{L^4}^2|y^h|_{W^{1,4}}^2 + \frac{\epsilon}{2}|\mathrm{e}_y^{h,\ell}|_{H^1}^2\\ &\leq \frac{C_{\mathrm{LGN},2}^2}{2\epsilon}|y^h|_{W^{1,4}}^2\|\mathrm{e}_y^{h,\ell}\|_{L^2}|\mathrm{e}_y^{h,\ell}|_{H^1} + \frac{\epsilon}{2}|\mathrm{e}_y^{h,\ell}|_{H^1}^2,\end{aligned}$$

where Young's inequality with parameter $\epsilon > 0$ has been used in the first, and the 2D-Ladyzhenskaya-Gagliardo-Nirenberg interpolation inequality,

$$\|\varphi\|_{L^4} \leq C_{\mathrm{LGN},2} \|\varphi\|_{L^2}^{\frac{1}{2}} |\varphi|_{W^{1,4}}^{\frac{1}{2}} \qquad \forall \varphi \in W^{1,4},$$

in the second step. Finally, a second application of Young's inequality, again with parameter ϵ, allows to obtain

$$\|\mathrm{e}_y^{h,\ell}\|_{L^4} |y^h|_{W^{1,4}} |\mathrm{e}_y^{h,\ell}|_{H^1} \leq \frac{C_{\mathrm{LGN},2}^4}{8\epsilon^3} |y^h|_{W^{1,4}}^4 \|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \epsilon |\mathrm{e}_y^{h,\ell}|_{H^1}^2.$$

In dimension $d = 3$ one has to apply

$$\|\varphi\|_{L^4} \leq C_{\mathrm{LGN},3} \|\varphi\|_{L^2}^{\frac{1}{4}} |\varphi|_{W^{1,4}}^{\frac{3}{4}} \qquad \forall \varphi \in W^{1,4}$$

and Young's inequality with exponents 4 and $\frac{4}{3}$ to arrive at

$$\|\mathrm{e}_y^{h,\ell}\|_{L^4} |y^h|_{W^{1,4}} |\mathrm{e}_y^{h,\ell}|_{H^1} \leq \frac{C_{\mathrm{LGN},3}^8}{64\epsilon^8} |y^h|_{W^{1,4}}^8 \|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \left(\frac{\epsilon}{2} + \frac{3\epsilon^{\frac{4}{3}}}{4} \right) |\mathrm{e}_y^{h,\ell}|_{H^1}^2.$$

In both cases one can proceed similarly as in the proof before in order to obtain estimates for $\mathrm{e}_y^{h,\ell}$ of structure analogous to those in Theorem 5.5. The main difference is that the constants $C_{\mathrm{LGN},2}, C_{\mathrm{LGN},3} > 0$, whose exact values are unknown in general, enter the estimates. For explicit upper bounds of these constants we refer the reader to, e.g., [**2**, Theorem 7.3] and the references therein.

5.2.3. A-posteriori error estimates including EIM. It is a well-known issue in RB methods that the evaluation of nonlinear terms such as $\xi(y)$ requires access to the full number of degrees of freedom. Since the reasoning behind MOR is to avoid such computations within the full model, alternatives have to be found. In order to allow for a so-called efficient offline-online splitting, the evaluation of nonlinearities in the reduced-order model for (5.3) needs to be done by methods of hyperreduction, e.g., the empirical interpolation method (EIM, [**23**]). In this section we follow Section 4 of our paper [**166**], describe a very basic version of the latter technique applied to our model problem, and show how the additional errors can be incorporated in the a-posteriori error estimates of Theorem 5.5 using the same technique as in [**151, 150**].

First, we introduce EIM as far as required for our purpose and as concise as possible; for details see, e.g., [**23, 283**]. In order to present the main idea without technicalities, we stick to the continuous setting and omit space discretization; the generalization to finite element spaces with a nodal basis is straightforward. Given so-called snapshots $y_1, \ldots, y_N \in C(\overline{\Omega})$, and a tolerance $\mathrm{tol}_{\mathrm{EIM}} > 0$, determine via a greedy procedure some functions $\Xi_1, \ldots, \Xi_m \in C(\overline{\Omega})$ and interpolation points $x_1, \ldots, x_m \in \overline{\Omega}$ such that

$$\xi(y_i(x_j)) = \sum_{k=1}^{m} c_{i,k} \Xi_k(x_j), \qquad i = 1, \ldots, N, \quad j = 1, \ldots, m,$$

implies $\left\| \xi(y_i) - \sum_{k=1}^m c_{i,k} \Xi_k \right\|_{L^\infty} \leq \mathrm{tol}_{\mathrm{EIM}}$. For some $w \in C(\overline{\Omega})$ we define the EIM approximation of $\xi(w)$ as

$$\xi_m^{\mathrm{EIM}}(w) = \sum_{k=1}^{m} c_k \Xi_k,$$

where $c \in \mathbb{R}^m$ solves the linear $m \times m$-system $\xi(w(x_j)) = \sum_{k=1}^m c_k \Xi_k(x_j)$, $j = 1, \dots, m$. With this we may introduce a RB-EIM-reduced counterpart of (5.3) as

$(5.3_{h,\ell,m})$

$$\begin{cases} \langle \partial_t y^{h,\ell,m}(t), \varphi^{h,\ell} \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} \\ \quad + \langle \mathcal{A}^{\mathrm{EIM}}_m(y^{h,\ell,m}(t)) y^{h,\ell,m}(t), \varphi^{h,\ell} \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} = \langle f(t), \varphi^{h,\ell} \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall t \in I, \varphi^{h,\ell} \in V^{h,\ell}, \\ \qquad\qquad\qquad\qquad\qquad\qquad y^{h,\ell,m}(0) = P^{h,\ell} I^h y_0, \end{cases}$$

where $\mathcal{A}^{\mathrm{EIM}}_m$ denotes the EIM-reduced version of the nonlinear differential operator defined by

$$\langle \mathcal{A}^{\mathrm{EIM}}_m(y)\varphi, \psi \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} := \int_\Omega \xi^{\mathrm{EIM}}_m(y) \mu \nabla \varphi \nabla \psi \mathrm{d}x, \qquad \varphi, \psi \in H^1_{\Gamma_D}.$$

Note that there is an efficient online evaluation of $\mathcal{A}^{\mathrm{EIM}}_m$ because the stiffness matrices associated to the operators $-\nabla \cdot \Xi_k \mu \nabla$ can be precomputed in the offline-phase. Therefore, we only have to deal with m and ℓ degrees of freedom when dealing with $\mathcal{A}^{\mathrm{EIM}}_m$. In the following we will denote the EIM error by

$$\Delta^{\mathrm{EIM}}_m(y) := \|\xi(y) - \xi^{\mathrm{EIM}}_m(y)\|_{L^\infty}.$$

Results addressing a-priori convergence of EIM can be found, e.g., in [**23, 122, 207**]. For sophisticated algorithmic coupling of model order reduction and hyperreduction we refer to [**103, 257**] for instance. Other kinds of hyperreduction include, e.g., discrete empirical interpolation (DEIM, [**80**]), or dynamic mode decomposition (DMD, [**7**]).

Now, we extend the results from Section 5.2.2 by incorporating also EIM errors. It is clear that $\mathcal{A}(y^{h,\ell})y^{h,\ell}$, and therefore $\mathrm{r}^{h,\ell}_y$, cannot be computed efficiently during the online phase due to the fact that the assembly of the stiffness matrix for $\mathcal{A}(y^{h,\ell})$ would require us to use the full number of degrees of freedom. Hence, the estimates of Theorem 5.5 cannot be evaluated efficiently in the online phase. Consequently, evaluation of $\mathrm{r}^{h,\ell}_y$ has to be avoided. Instead, given an arbitrary trajectory $y^{h,\ell}$ fulfilling Assumption 5.3 (not necessarily a solution of $(5.3_{h,\ell,m})$), we introduce the EIM-reduced residual $\mathrm{r}^{h,\ell,m}_y$ of $y^{h,\ell}$ as

$$\mathrm{r}^{h,\ell,m}_y(t) := \partial_t y^{h,\ell}(t) + \mathcal{A}^{\mathrm{EIM}}_m(y^{h,\ell}(t)) y^{h,\ell}(t) - f(t) \in (V^h)^*, \quad t \in I. \tag{5.15}$$

It is obvious, that $\mathrm{r}^{h,\ell,m}_y$ allows an efficient online evaluation. It remains to show how the error $y^{h,\ell} - y^h$ to the truth-solution can be estimated in terms of $\mathrm{r}^{h,\ell,m}_y$ instead of $\mathrm{r}^{h,\ell}_y$. Since all changes in the arguments already known from Section 5.2.2 are straightforward utilizing the estimates (5.16) and (5.17) below, we omit the details and only state the results. As before, we will we omit the argument "t" in the following. A short computation as in Section 5.2.2 shows that the error $\mathrm{e}^{h,\ell}_y := y^{h,\ell} - y^h$ fulfills

$$\begin{aligned} \langle \mathrm{r}^{h,\ell,m}_y, \varphi_h \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} &= \langle \partial_t \mathrm{e}^{h,\ell}_y, \varphi_h \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} \\ &\quad + \langle \mathcal{A}(y^{h,\ell}) y^{h,\ell} - \mathcal{A}(y^h) y^h, \varphi_h \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}} \\ &\quad + \langle \mathcal{A}^{\mathrm{EIM}}_m(y^{h,\ell}) y^{h,\ell} - \mathcal{A}(y^{h,\ell}) y^{h,\ell}, \varphi_h \rangle_{H^{-1}_{\Gamma_D}, H^1_{\Gamma_D}}. \end{aligned}$$

As before, it follows

$$\begin{aligned}(5.16)\quad \frac{d}{dt}\frac{1}{2}\|\mathrm{e}_y^{h,\ell}\|_{L^2}^2 + \xi_\bullet\mu_\bullet|\mathrm{e}_y^{h,\ell}|_{H^1}^2 &\le \langle \mathrm{r}_y^{h,\ell,m}, \mathrm{e}_y^{h,\ell}\rangle_{H_{\Gamma_D}^{-1},H_{\Gamma_D}^1} \\ &\quad - \int_\Omega (\xi(y^{h,\ell}) - \xi(y^h))\mu\nabla y^h\nabla \mathrm{e}_y^{h,\ell}\,\mathrm{d}x \\ &\quad - \int_\Omega (\xi_m^{\mathrm{EIM}}(y^{h,\ell}) - \xi(y^{h,\ell}))\mu\nabla y^{h,\ell}\nabla \mathrm{e}_y^{h,\ell}\,\mathrm{d}x.\end{aligned}$$

The second summand on the right-hand side can be estimated as in Section 5.2.2. The third summand is estimated as follows:

$$\begin{aligned}(5.17)\quad &\left|\int_\Omega (\xi_m^{\mathrm{EIM}}(y^{h,\ell}) - \xi(y^{h,\ell}))\mu\nabla y^{h,\ell}\nabla \mathrm{e}_y^{h,\ell}\,\mathrm{d}x\right| \\ &\qquad \le \Delta_m^{\mathrm{EIM}}(y^{h,\ell})\mu^\bullet |y^{h,\ell}|_{H^1}\|\mathrm{e}_y^{h,\ell}\|_{H^1} \\ &\qquad \le \frac{1}{2\delta}\Delta_m^{\mathrm{EIM}}(y^{h,\ell})\mu^\bullet |y^{h,\ell}|_{H^1}^2 + \frac{1}{2}\delta\Delta_m^{\mathrm{EIM}}(y^{h,\ell})\mu^\bullet \|\mathrm{e}_y^{h,\ell}\|_{H^1}^2,\end{aligned}$$

where $\delta > 0$ is the parameter in Young's inequality. With this, we are ready to state the modified version of Theorem 5.5.

Theorem 5.6 ([**166**], Theorem 4.1)**.** *Let Assumptions 1.5, 1.6, 1.8 and 5.3 hold, and let $c_\infty > 0$ be such that*

$$|y^h(t)|_{W^{1,\infty}} \le c_\infty \qquad \forall t \in I.$$

Given $y^{h,\ell}$, choose $\varepsilon, \eta, \delta > 0$ such that

$$\eta + \varepsilon|\xi'|_\infty\mu^\bullet c_\infty + \delta\Delta_m^{\mathrm{EIM}}\mu^\bullet = \xi_\bullet\mu_\bullet,$$

is satisfied with the EIM error $\Delta_m^{\mathrm{EIM}} := \sup_{t\in I}\Delta_m^{\mathrm{EIM}}(y^{h,\ell}(t))$. Moreover, we introduce the constant $\beta := 2\left(\frac{1}{2\varepsilon}|\xi'|_\infty\mu^\bullet c_\infty + \xi_\bullet\mu_\bullet\right)$. Then, the following a-posteriori error estimates for $y^{h,\ell}$ hold true:

$$\begin{aligned}(5.18)\quad \|\mathrm{e}_y^{h,\ell}(t)\|_{L^2}^2 &\le e^{\beta t}\|y^{h,\ell}(0) - y^h(0)\|_{L^2}^2 \\ &\quad + \int_0^t e^{\beta(t-s)}\left(\eta^{-1}\|\mathrm{r}_y^{h,\ell,m}(s)\|_{(V^h)^*}^2\right. \\ &\qquad \left. + \delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet |y^{h,\ell}(s)|_{H^1}^2\right)\mathrm{d}s,\end{aligned}$$

$$\begin{aligned}(5.19)\quad \|\mathrm{e}_y^{h,\ell}\|_{L^2(I,L^2)}^2 &\le \beta^{-1}\left(e^{\beta T} - 1\right)\|y^{h,\ell}(0) - y^h(0)\|_{L^2}^2 \\ &\quad + \beta^{-1}\int_0^T (e^{\beta(T-t)} - 1)\left(\eta^{-1}\|\mathrm{r}_y^{h,\ell,m}(t)\|_{(V^h)^*}^2\right. \\ &\qquad \left. + \delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet |y^{h,\ell}(t)|_{H^1}^2\right)\mathrm{d}t,\end{aligned}$$

$$\begin{aligned}(5.20)\quad \|\mathrm{e}_y^{h,\ell}\|_{L^2(I,H_{\Gamma_D}^1)}^2 &\le \xi_\bullet^{-1}\mu_\bullet^{-1}e^{\beta T}\|y^{h,\ell}(0) - y^h(0)\|_{L^2}^2 \\ &\quad + \xi_\bullet^{-1}\mu_\bullet^{-1}\int_0^T e^{\beta(T-t)}\left(\eta^{-1}\|\mathrm{r}_y^{h,\ell,m}(t)\|_{(V^h)^*}^2\right. \\ &\qquad \left. + \delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet |y^{h,\ell}(t)|_{H^1}^2\right)\mathrm{d}t.\end{aligned}$$

We also fix the following simplified estimates, that are less sharp but exhibit a favorable structure. They are weighted sums of the initial L^2-error, the L^2-$(V^h)^*$-norm of the residual, and the EIM error. This allows to determine the optimal choice of the parameters $\varepsilon, \eta, \delta$ for these simpler estimates.

Corollary 5.7 ([**166**], Corollary 4.2). *Under the assumptions of the previous theorem it holds:*

$$\begin{aligned}
\|\mathrm{e}_y^{h,\ell}(t)\|_{L^2}^2 &\leq e^{\beta t}\|y^{h,\ell}(0) - y^h(0)\|_{L^2}^2 + e^{\beta t}\eta^{-1}\int_0^t \|\mathrm{r}_y^{h,\ell,m}(s)\|_{(V^h)^*}^2 \,\mathrm{d}s \\
&\qquad + \delta^{-1} e^{\beta t} \Delta_m^{\mathrm{EIM}}\mu^\bullet \int_0^t |y^{h,\ell}(s)|_{H^1}^2 \mathrm{d}s, \\
\|\mathrm{e}_y^{h,\ell}\|_{L^2(I,L^2)}^2 &\leq \beta^{-1}\left(e^{\beta T}-1\right)\|y^{h,\ell}(0) - y^h(0)\|_{L^2}^2 \\
&\quad + \beta^{-1}(e^{\beta T}-1)\left(\eta^{-1}\|\mathrm{r}_y^{h,\ell,m}\|_{L^2(I,(V^h)^*)}^2 \right. \\
&\qquad \left. + \delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet |y^{h,\ell}|_{L^2(I,H^1)}^2\right), \\
\|\mathrm{e}_y^{h,\ell}\|_{L^2(I,H^1_{\Gamma_D})}^2 &\leq e^{\beta T}\xi_\bullet^{-1}\mu_\bullet^{-1}\|y^{h,\ell}(0) - y^h(0)\|_{L^2}^2 \\
&\quad + e^{\beta T}\xi_\bullet^{-1}\mu_\bullet^{-1}\left(\eta^{-1}\|\mathrm{r}_y^{h,\ell,m}\|_{L^2(I,(V^h)^*)}^2 \right. \\
&\qquad \left. + \delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet |y^{h,\ell}|_{L^2(I,H^1)}^2\right).
\end{aligned}$$

Note that the error estimates in the results above hold true without additional assumptions on the size of residuals and EIM errors. In practice, however, in order to obtain an accurate reduced-order model and corresponding small error estimates, it will be necessary to construct RB and EIM bases in such a way that the size of residuals and EIM errors is balanced appropriately. We do not address this issue here and refer the reader for instance to [**103, 257**].

Moreover, let us point out that the EIM error $\Delta_m^{\mathrm{EIM}}(y^{h,\ell})$ at $y^{h,\ell}$ cannot be computed without referring to the full number of degrees of freedom; however, the computation of $\|\xi(y^{h,\ell}) - \xi_m^{\mathrm{EIM}}(y^{h,\ell})\|_{L^\infty}$ in the full degrees of freedom is still much cheaper than the computation of the respective full stiffness matrices associated with the nonlinear elliptic operator that would be required for the computation of $\mathrm{r}_y^{h,\ell}$. In contrast, note that the H^1-seminorm of $y^{h,\ell}$ required in Theorems 5.6 and 5.9 admits efficient online evaluation, because it is induced by a bilinear form whose matrix w.r.t. the basis of $V^{h,\ell}$ can be precomputed and saved. Similarly, also the weight matrices for the evaluation of the EIM-reduced residual can be precomputed and saved in the offline phase.

Let us briefly outline a possibility to relax Assumption 5.3.2 in order to allow error estimation also for a discontinuous in time trajectory.

Remark 5.8 ([**166**], Remark 4.4). Let $y^{h,\ell}$, e.g., be given as

$$y^{h,\ell} := \mathbf{1}_{\{0\}} Y_0^{h,\ell} + \sum_{k=1}^{N_t} \mathbf{1}_{(t_{k-1},t_k]} Y_k^{h,\ell}, \qquad Y_k^{h,\ell} \in V^{h,\ell}, \quad k = 0, \ldots, N_t,$$

for a partition $0 = t_0 < t_1 < \ldots < t_{N_t-1} < t_{N_t} = T$. Such $y^{h,\ell}$ might be obtained by applying the backward Euler method in its DG0-formulation to $(5.3_{h,\ell,m})$. Since our error estimates do not apply directly to $y^{h,\ell}$ due to discontinuity w.r.t. time,

we replace $y^{h,\ell}$ by its piecewise linear and continuous w.r.t. time interpolation $\hat{y}^{h,\ell}$ w.r.t. the same partition defined by $\hat{y}^{h,\ell}(t_k) := y^{h,\ell}(t_k) = Y_k^{h,\ell}$ for $k = 0, \dots, N_t$. Obviously, Theorems 5.6 and 5.9 apply to $\hat{y}^{h,\ell}$ and in order to obtain an estimate for the overall error we need to add the interpolation error $\hat{y}^{h,\ell} - y^{h,\ell}$. The latter can be computed explicitely:

$$\|y^{h,\ell} - \hat{y}^{h,\ell}\|^2_{L^\infty(I,L^2)} \leq \max_{1\leq k\leq N_t} \|Y_k^{h,\ell} - Y_{k-1}^{h,\ell}\|^2_{L^2},$$

$$\|y^{h,\ell} - \hat{y}^{h,\ell}\|^2_{L^2(I,L^2)} \leq \sum_{k=1}^{N_t} \frac{1}{3}(t_k - t_{k-1}) \|Y_k^{h,\ell} - Y_{k-1}^{h,\ell}\|^2_{L^2},$$

$$\|y^{h,\ell} - \hat{y}^{h,\ell}\|^2_{L^2(I,H^1)} \leq \sum_{k=1}^{N_t} \frac{1}{3}(t_k - t_{k-1}) \|Y_k^{h,\ell} - Y_{k-1}^{h,\ell}\|^2_{H^1}.$$

The appearance of such jump terms is what we may expect for an a-posteriori error for a discontinuous in time trajectory. Note that compared to classical a-posteriori error estimates for discontinuous in time methods, see [**202, 263**] for instance, we do not assume that $y^{h,\ell}$ is the solution of a discrete in time analogue to $(5.3_{h,\ell})$.

We conclude this section by mentioning the second main result from our paper [**166**] without proof; for the respective details we refer the interested reader to [**166**, Section 3.3]. Recall from the beginning of Section 5.2.2 that the a-posteriori error estimates presented in this section so far depend on the $L^\infty(I, W^{1,\infty})$-norm of the truth-solution. In the rough regularity setting of Assumptions 1.5, 1.6 and 1.8 or Assumptions 1.5, 1.6 and 1.10 we expect that the respective constant c_∞ in Theorem 5.6 depends on the space discretization because the continuous in space and time solution of (5.3) does not exhibit $L^\infty(I, W^{1,\infty})$-regularity, in general. More precisley, we expect that c_∞ tends to infinity as discretization gets finer. Therefore, we proposed in [**166**, Sections 3.3] also a second approach that is motivated by the intention to exploit less regularity of the truth-solution, more precisely: $L^\infty(I, W^{1,p})$-regularity for some $p > d$ only. For continuous in space and time solutions of (5.3) this regularity is guaranteed in the Bessel potential space setting from [**35**], i.e. under Assumptions 1.5, 1.6 and 1.10; cf. Theorem 1.14 and (1.8). The price to pay for exploiting less regularity of y^h is that we do no longer obtain an explicit formula for the error estimate. Instead, the evaluation of the estimate now requires the solution of an ODE. Moreover, for technical reasons we require additional assumptions on time regularity of the residual and the size of the initial error.

Theorem 5.9 ([**166**], Theorem 4.3)**.** *Let Assumptions 1.5, 1.6, 1.8 and 5.3 hold, and let $p > d$ and $c_p > 0$ be such that*

$$|y^h(t)|_{W^{1,p}} \leq c_p \qquad \forall t \in I.$$

Moreover, we assume that the initial error does not vanish, i.e. $\|y^{h,\ell}(0) - y^h(0)\|_{L^2} > 0$, and that $t \mapsto \|\mathrm{r}_y^{h,\ell,m}(t)\|^2_{V_h^*}$ is piecewise continuous on I. Choose $\varepsilon, \eta, \delta > 0$ such that

$$\xi_\bullet \mu_\bullet = \eta + \varepsilon \cdot \mu^\bullet (2\xi^\bullet)^{1-\frac{2}{q}} |\xi'|_\infty^{\frac{2}{q}} c_p + \delta \Delta_m^{\mathrm{EIM}} \mu^\bullet$$

is satisfied for the EIM error $\Delta_m^{\mathrm{EIM}} = \sup_{t\in I} \Delta_m^{\mathrm{EIM}}(y^{h,\ell}(t))$. *Given the constants* $\alpha = 2\xi_\bullet\mu_\bullet$, $\beta = \varepsilon^{-1}\mu^\bullet(2\xi^\bullet)^{1-\frac{2}{q}}|\xi'|_\infty^{\frac{2}{q}}c_p$, *and* $r = 1-\frac{2}{p}$, *let* $\varphi\colon I\to[0,\infty)$ *be the solution to the ODE*

$$\varphi'(t) = \alpha\varphi(t) + \beta\varphi(t)^r + \eta^{-1}\|\mathrm{r}_y^{h,\ell,m}(t)\|_{V_h^*}^2 + \delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet|y^{h,\ell}(t)|_{H^1}^2, \qquad t\in I,$$
$$\varphi(0) = \|\mathrm{e}_y^{h,\ell}(0)\|_{L^2}^2.$$

Then, the following a-posteriori error estimates for $y^{h,\ell}$ *hold true:*

$$\|\mathrm{e}_y^{h,\ell}(t)\|_{L^2}^2 \le \varphi(t), \quad \forall t\in I, \qquad \|\mathrm{e}_y^{h,\ell}\|_{L^2(I,L^2)}^2 \le \int_0^T \varphi(s)\mathrm{d}s, \tag{5.21}$$

$$\begin{aligned}\|\mathrm{e}_y^{h,\ell}\|_{L^2(I,H^1_{\Gamma_D})}^2 \le \frac{1}{\xi_\bullet\mu_\bullet}\Bigg(&\|y^{h,\ell}(0)-y^h(0)\|_{L^2}^2 + \eta^{-1}\|\mathrm{r}_y^{h,\ell,m}\|_{L^2(I,V_h^*)}^2\\ &+\delta^{-1}\Delta_m^{\mathrm{EIM}}\mu^\bullet|y^{h,\ell}|_{L^2(I,H^1)}^2 + \alpha\int_0^T\varphi(s)\mathrm{d}s\\ &+\beta\int_0^T\varphi(s)^{2/q}\mathrm{d}s\Bigg).\end{aligned} \tag{5.22}$$

5.2.4. Numerical illustration. This section closely follows [**166**, Section 5]. We illustrate and compare the quality of our a-posteriori POD/RB-EIM error estimates numerically for three prototypical test problems. By "approach I" we refer to the estimates from Theorem 5.6 and by "approach II" to those from Theorem 5.9. Although our results apply to general RB methods, our particular focus is on POD-MOR as explained at the beginning of this chapter. Therefore, we restrict ourselves to reduced ansatz spaces $V^{h,\ell}$ spanned by a POD basis of rank ℓ in our numerical tests.

The two-dimensional underlying domain $\Omega = [0,1]^2$ and the time interval $I = [0,1]$ are the same in all three test problems. We fix two euclidean discs $C_1 = B_{\frac{1}{5}}\left(\frac{1}{4},\frac{1}{4}\right)$ and $C_2 = B_{\frac{1}{5}}\left(\frac{3}{4},\frac{3}{4}\right)$, and the three boundary parts $\Gamma_1 = \{x\in\partial\Omega\colon x_2 = 1\}$, $\Gamma_2 = \{x\in\partial\Omega\colon x_1 = 0, x_2 < \frac{1}{2}\}$, $\Gamma_3 = \{x\in\partial\Omega\colon x_1 = 1, x_2 < \frac{1}{2}\}$. The nonlinearity is given by

$$\xi(y) = \frac{3}{4} + \frac{1}{2(1+e^{5y})}.$$

We introduce the three test problems P1-P3 by equipping the equation

$$\partial_t y - \nabla\cdot\xi(y)\nabla y = 10\sin(2\pi t)\mathbf{1}_{C_1} - 10\cos(2\pi t)\mathbf{1}_{C_2}$$

with the following boundary and initial conditions:

- **P1.** *Pure homogeneous Dirichlet boundary conditions* and zero initial condition.
- **P2.** *Pure homogeneous Neumann boundary conditions* and zero initial condition.
- **P3.** *Mixed boundary conditions:* homogeneous Dirichlet boundary condition $y = 0$ on $I\times\Gamma_1$, nonhomogeneous Neumann conditions $\xi(y)\partial_n y = \sin(2\pi t)$ on $I\times\Gamma_2$, and $\xi(y)\partial_n y = -\cos(2\pi t)$ on $I\times\Gamma_3$, and natural boundary condition $\partial_n y = 0$ on the remaining part of the boundary. The initial condition is given by $[y(0)](x_1,x_2) := \frac{1}{10}(1-x_1)$.

Space and time discretization. All computations are done utilizing FEniCS [**9, 203**] and piecewise linear finite elements on a mesh generated by mshr, the mesh-generation tool of FEniCS, with $N_h = 5769$ degrees of freedom and maximum cell diameter $h_{\max} \approx 2.1 \cdot 10^{-2}$. The POD basis is generated with snapshots coming from an (implicit) Crank-Nicolson solution of the equation with $N_t = 2500$ time steps ("reference solution"). Hereby, the appearing nonlinear equations are solved by the built-in nonlinear solver of FEniCS. The same set of snapshots is also used to generate the EIM approximation of the nonlinearity in a standard greedy procedure with L^∞-tolerance 10^{-6} independent of the number of POD basis functions, i.e. we do not balance accuracy of POD and EIM approximation. The POD-EIM-reduced equation is again solved utilizing the (implicit) Crank-Nicolson scheme with $N_t = 2500$ time steps, whereby the nonlinear algebraic equations appearing in every time step are solved by a standard Newton method that is initialized with a semiimplicit Euler step as first guess ("reduced solution"). Approximate true $L^2(I, L^2)$-, $L^\infty(I, L^2)$, and $L^2(I, H^1)$-errors are computed with respect to a further numerical solution that is computed on the same finite element mesh, but with a four times higher number of time steps than for the snapshot generation ("truth-solution"). Finally, to ensure comparability between the different test problems and norms, all errors and estimates are relative errors, i.e. the absolute error or error estimate is divided by the corresponding norm of the truth-solution.

Estimation of the required parameters. Parameters like $\xi_\bullet, \mu_\bullet, |\xi'|_\infty$ etc. are known from the problem data. The solution-dependent parameters are found as follows: the norms of y^h are computed exactly based on the truth-solution in order to give the possibility to determine whether our estimates are sharp or not under the exact data. In real applications we would have to estimates those norms appropriately. The quality of the error estimates —as absolute values— can heavily deteriorate in case of "safe" (i.e. large) estimates for the parameters. The same might happen in case of just inconvenient problem data due to the exponential terms in the estimates. However, we would like to point out that one might still hope in such a case that the relative behavior of the estimates, i.e. whether they decrease/increase by some factor, provides some information on the quality of the reduced-order model. Although we compute the EIM error Δ_m^{EIM} as defined in Section 5.2.3 by accessing the full number of degrees of freedom, we did not observe significant time consumption for this. We believe that this is due to the fact that evaluation of $\xi(y^{h,\ell})$ in the full model is much cheaper than assembling the corresponding stiffness matrices in the full model.

Estimates for approach I (Theorem 5.6). For approach I we determine the parameters $\varepsilon, \eta, \delta$ in such a way that the simpler estimates for the $L^2(I, H^1)$-error in Corollary 5.7 become optimal, and plug in the same parameters into the estimates from Theorem 5.6. Integrals with respect to time are evaluated using Gauss quadrature of order 2 on every subinterval given by the time steps. To improve readability, we omit the estimates of Corollary 5.7 in our plots. In fact, they are not much worse than those of Theorem 5.6; we refer the interested reader to the diagramms in [**166**] that also display the estimates from Corollary 5.7.

Estimates for approach II (Theorem 5.9). In order to obtain meaningful results we had to use relatively large values for the integrability exponent p, e.g., $p = 16$. Therefore, choosing p according to the requirements of [**35**], cf. Chapter 1, i.e.

only slightly larger than d in general, seems to be difficult. Moreover, the following parameters turned out to be a good choice:

$$\eta = \frac{1}{10}(1 - \Delta_m^{\mathrm{EIM}})\xi_\bullet\mu_\bullet, \qquad \varepsilon = \frac{9}{10}\frac{1 - \Delta_m^{\mathrm{EIM}}}{\mu^\bullet(2\xi^\bullet)^{1-\frac{2}{q}}|\xi'|_\infty^{\frac{2}{q}}c_p}, \qquad \delta = \frac{\xi_\bullet\mu_\bullet}{\mu^\bullet}.$$

Note that optimization of the parameters as in approach I is not possible because we do not have an explicit formula at hand. The ODE for the evaluation of φ is solved utilizing the backward difference formulae solver (BDF) within the `solve_ivp`-routine from `scipy.integrate`, with relative tolerance `rtol`$=10^{-6}$, and absolute tolerance `atol`$=10^{-3} \cdot \|y^h(0) - y^{h,\ell}(0)\|_{L^2}^2$. The maximal allowed step size is the same as the size of time steps in the reduced-order model. We found that among other methods (Runge-Kutte with 2/3 and 4/5 stages, Radau) this choice delivered the best results. However, it is clear that the numerical approximation of φ is challenging (in particular, for small p or small initial values), which might influence the reliability of the results.

Figures 5.1 to 5.4 show the results of our experiments. It can be seen that approach I yields the better results the smoother the truth-solution is. Test problems P1 and P2 (homogeneous boundary conditions, Figures 5.1 and 5.2) perform better than the problem with mixed boundary conditions (Test problem P3, Figure 5.3). Moreover, we observe that the a-posteriori error estimates of both approaches start stagnating at about the same point at which also the true errors stagnate due to time discretization. Indeed, in Figure 5.4 it can be seen that this stagnation comes from stagnation of the residual norm at roughly the same magnitude as the size of time steps. This indicates that from that point on the overall accuracy of the reduced-order model cannot be improved further by increasing the number of basis functions; see also, e.g., [**129**] for balancing of POD-MOR and time discretization errors for linear quadratic parabolic optimal control problems.

How much approach II depends on the choice of the exponent p can be seen in Figure 5.2.b. The estimates stagnate very early for small p, i.e. approach II unfortunately does not yield reasonable results in that case. For large p the estimates seem to get closer to the values of approach I. In this sense one might interpret approach II as a modification of approach I that trades strength of the required assumption (bigger p means stronger assumption) against quality of results (smaller p means less meaningful results and numerical instability).

For the computing times observed in our numerical experiments we refer to Table 5.1. The evaluation of the POD-EIM-reduced model is about 25- to 100-times faster than the evaluation of the full model. We believe that even higher speedups might be possible in case of finer finite element discretization. Compared to the computing time for the full model, evaluation of the a-posteriori error estimates from approach I is quite cheap. Evaluation of the POD-EIM-reduced model together with computation of an error estimate still yields a speedup of factor at least 10. As expected, evaluation of the estimates from approach II needs slightly more time.

5.2.5. A-posteriori POD/RB errors in the context of optimal control. Extending the comments in [**166**, Section 2.3], let us come back to the optimal control

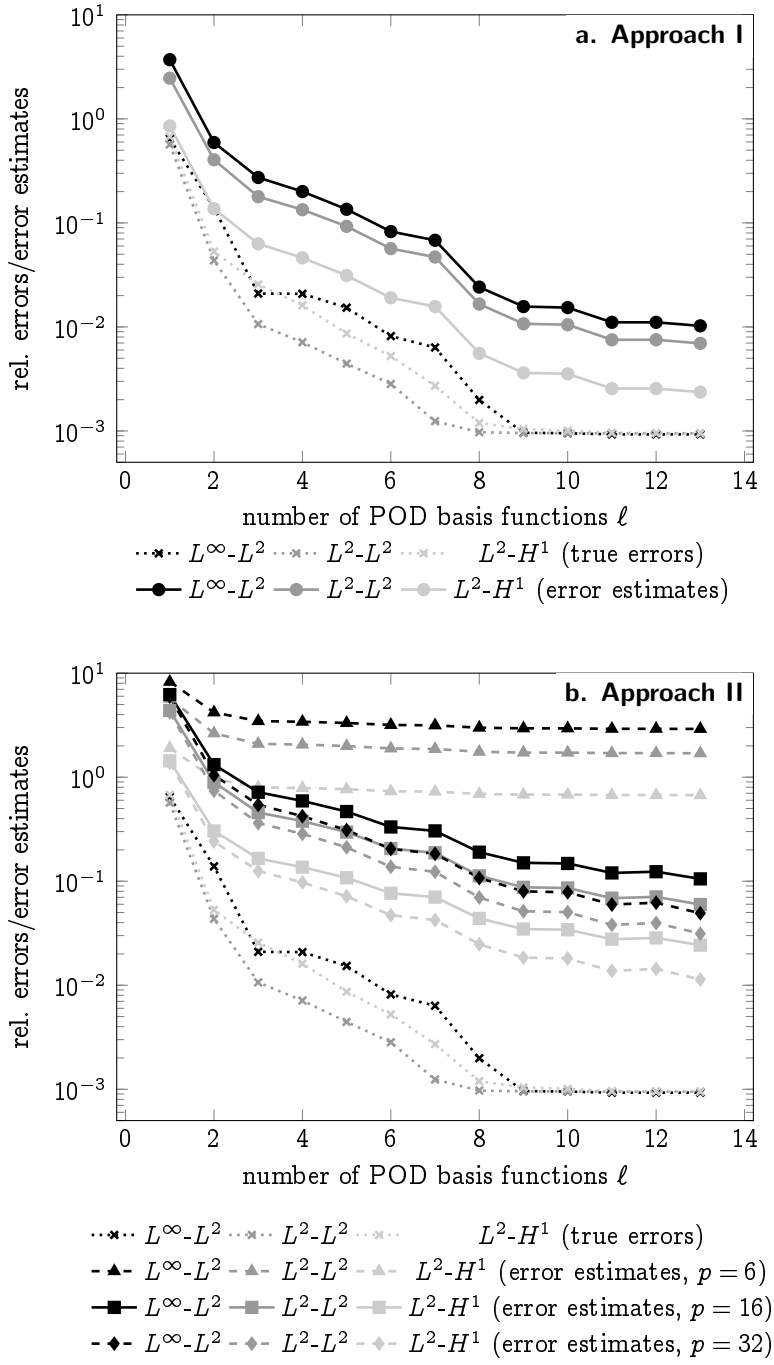


Figure 5.1. Test problem P1 (homogeneous Dirichlet boundary conditions): **a.** estimates from approach I (Theorem 5.6 with optimized parameters) and **b.** estimates from approach II (for $p \in \{6, 16, 32\}$). Approximate true errors w.r.t. the truth-solution are included in dotted lines.

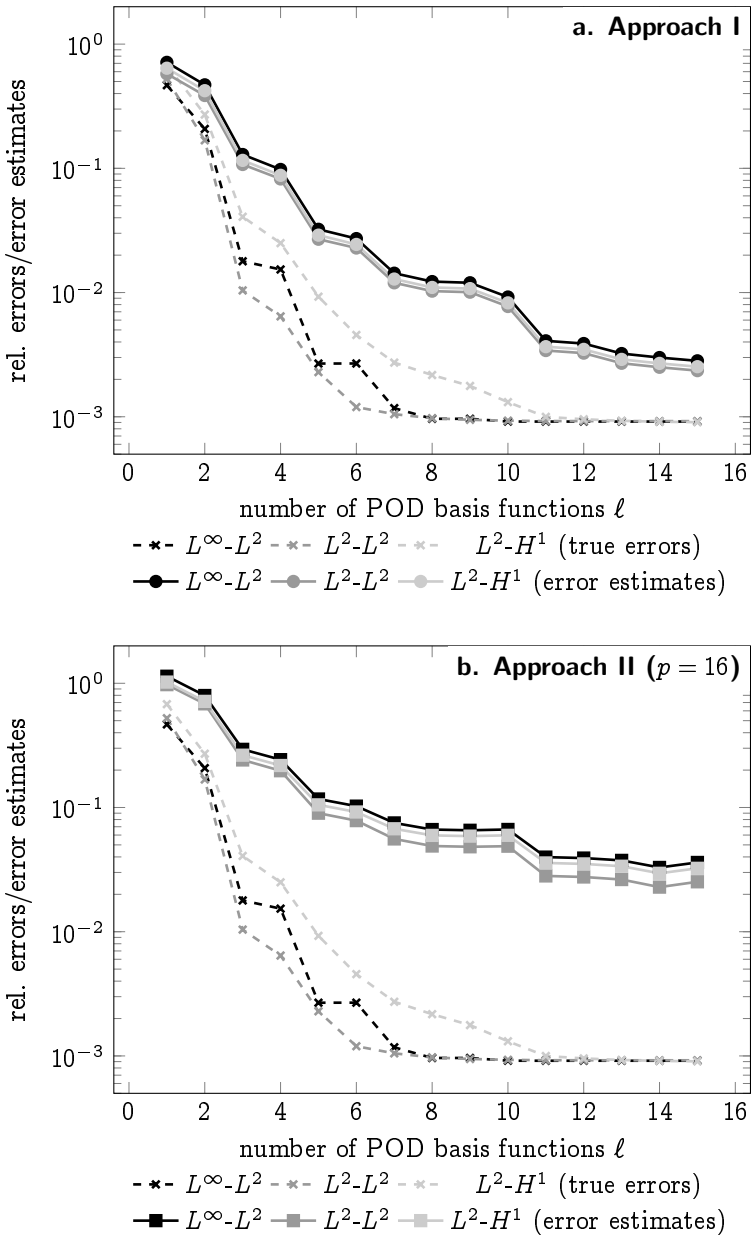


Figure 5.2. Example P2 (homogeneous Neumann boundary conditions): **a.** estimates from approach I (Theorem 5.6 with optimized parameters) and **b.** estimates from approach II for $p = 16$. Approximate true errors w.r.t. the truth-solution are included in dotted lines.

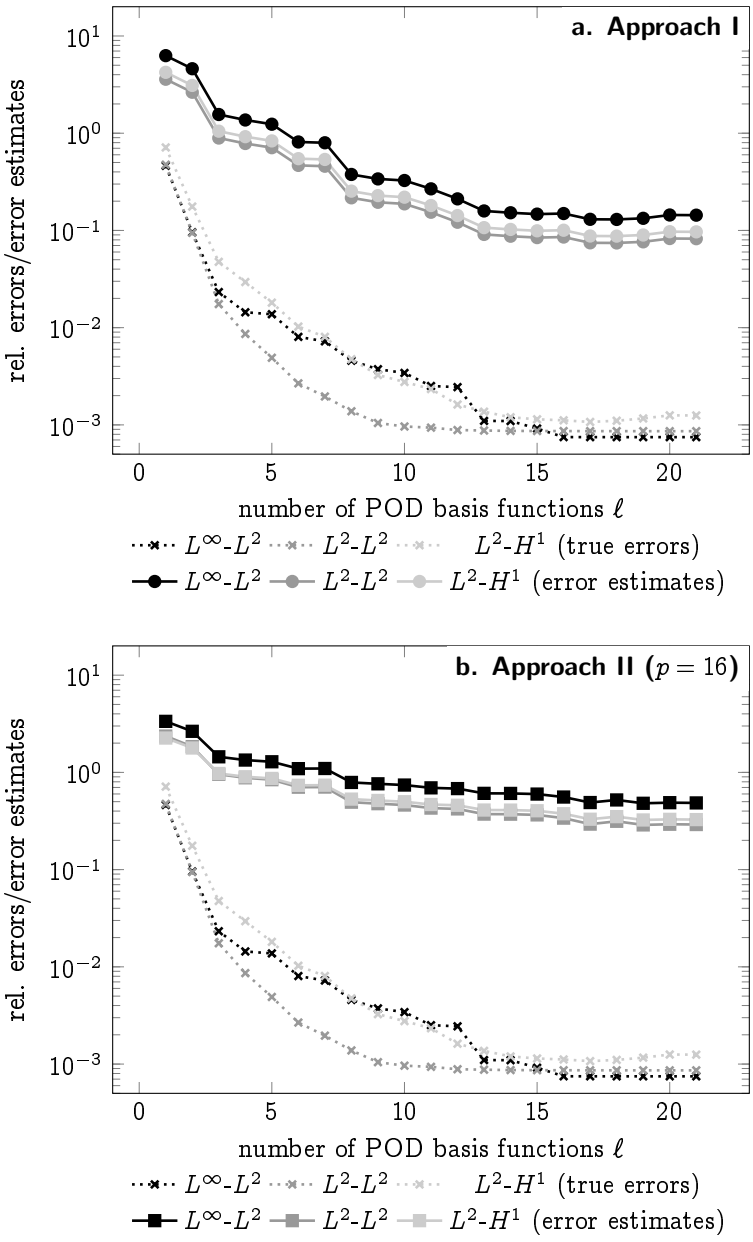


Figure 5.3. Example P3 (mixed boundary conditions): **a.** estimates from approach I (Theorem 5.6 with optimized parameters) and **b.** estimates from approach II for $p = 16$. Approximate true errors w.r.t. the truth-solution are included in dotted lines.

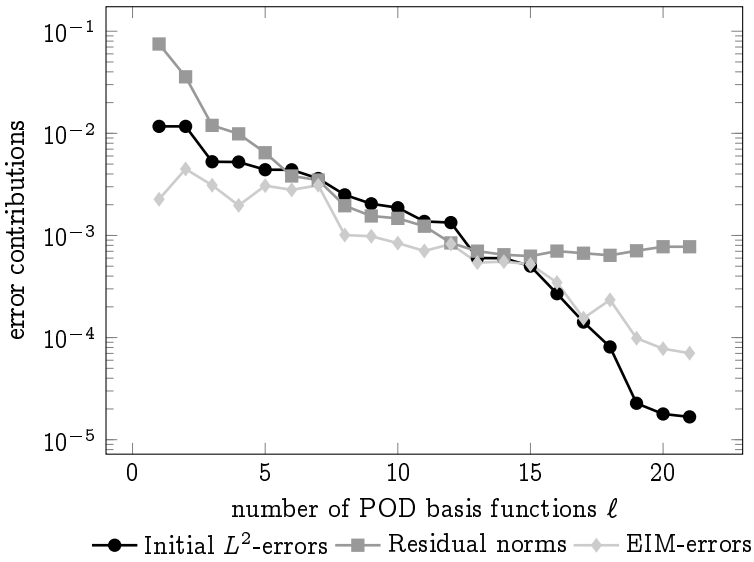


Figure 5.4. Error contributions in Example P3 (mixed boundary conditions): residual norms $\|\mathrm{r}_y^{h,\ell,m}\|_{L^2(I,(V^h)^*)}$, initial errors $\|\mathrm{e}_y^{h,\ell}(0)\|_{L^2}$, and EIM errors Δ^{EIM} for $y^{h,\ell,m}$.

Computing times for	Example P1	Example P2	Example P3
Number of EIM basis functions	28	36	49
Setup EIM-reduced model	50-57%	70%	99-157%
POD-EIM-reduced model	1% (0.9%)	1% (1.1%)	1-4% (1.6%)
Approach I (Corollary 5.7)	2-3% (3%)	2-4% (4%)	3-6% (5%)
Approach I (Theorem 5.6)	3-6% (6%)	3-9% (8%)	5-12% (10%)
Approach II	3-15% (9-15%)	4-22% (18%)	6-15% (11%)

Table 5.1. Computing times for the setup of the EIM-reduction of the nonlinearity, the evaluation of the POD-EIM-reduced model, and the error estimates, respectively. 100% correspond to the time that is required to compute the snapshots ("reference solution"). We show the range of times observed in the experiments from Figures 5.1 to 5.3, and in brackets we give the time observed for $\ell = 13$ POD basis functions.

problem (P) introduced in Chapter 1. We consider the particular purely time-dependent control setting from Assumption 3.1; see also Example 1.9.3. Fixing a space discretization for the underlying state equation as described in Section 5.2

results in a semidiscrete (in space) counterpart (P_h) of (P):

$$
(\mathrm{P}_h)\quad \begin{cases} \min\limits_{y^h,u}\ J(y^h,u) \\ \text{s.t. } u\in U_{\mathrm{ad}} \\ \text{and} \begin{cases} \langle \partial_t y^h(t),\varphi^h\rangle_{H^{-1}_{\Gamma_D},H^1_{\Gamma_D}} \\ \quad +\langle \mathcal{A}(y^h(t))y^h(t),\varphi^h\rangle_{H^{-1}_{\Gamma_D},H^1_{\Gamma_D}} = \sum\limits_{i=1}^m \langle b_i,\varphi^h\rangle_{H^{-1}_{\Gamma_D},H^1_{\Gamma_D}} u_i(t) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall t\in I,\varphi^h\in V^h, \\ y^h(0)=I^h y_0. \end{cases} \end{cases}
$$

Again, fixing a concrete space discretization we may consider the resulting semidiscrete problem (P_h) and its solution as reference object ("truth-solution") as we did with (5.3_h) and its solution. Proposition 5.2 shows that (P_h) is well-defined because the solution map of the underlying semidiscrete equation is well-defined.

At this point, recall the remarks we made about discretization of optimal control problems in the introduction chapter of this thesis. Due to the purely time-dependent control structure of (P), the semidiscrete (in space) problem (P_h) may be regarded as variational discretization [**149**] of (P) when only space is discretized. We mention, e.g., [**236**] for the same semidiscrete (in space) setting in the context of a-posteriori POD errors for an optimal control problem governed by a semilinear parabolic PDE, or [**92**] for semidiscrete (in space) finite element error estimates for optimal control of the instationary Navier-Stokes equations. To obtain a fully discrete problem, there are different possibilities to choose an appropriate time discretization. Our POD/RB a-posteriori error estimates in particular apply to CG1 time discretization, but an extension to, e.g., DG0 (implicit Euler) discretization has been sketched in Remark 5.8.

In numerical algorithms for the solution of the semidiscrete control problem (P_h), we may have to evaluate the semidiscrete reduced functional $j^h(u) := J(y^h(u),u)$ for several control functions u where $y^h(u)$ denotes the solution of the underlying state equation in (P_h). Since repeated evaluation of j^h is costly, POD/RB-MOR together with EIM can be applied to this equation. Therefore, and due to additional time discretization, we only have the possibility to compute an approximate solution $y^{h,\ell,m} = y^{h,\ell,m}(u)$, i.e. a solution of a time-discrete counterpart of $(5.3_{h,\ell,m})$, instead of the respective truth-solution $y^h(u)$. A short computation shows that the resulting error in the reduced functional can be estimated, e.g., as follows:

$$
\begin{aligned}
|J(y^{h,\ell,m},u)-J(y^h,u)| \le \left[\frac{1}{2}\|y^{h,\ell,m}-y^h\|_{L^2(I,L^2)} + \|y^{h,\ell,m}-y_d\|_{L^2(I,L^2)}\right] \\
\cdot \|y^{h,\ell,m}-y^h\|_{L^2(I,L^2)}.
\end{aligned}
$$

Consequently, for each of the $L^2(I,L^2)$-estimates for the solutions of (5.3_h) from Theorems 5.6 and 5.9 we immediately obtain a corresponding a-posteriori error estimate for the reduced functional of (P_h) . Retrospectively, this may be regarded as motivation for the results presented in the section. Note that the above estimate for the functional error differs from the one obtained in [**232**, Theorems 4 and 9] because we do not utilize adjoint information.

Let us conclude our discussion of a-posteriori POD/RB error estimates for the state equation by pointing out some difficulties related to the open problem of such estimates for the whole optimal control problem. We already saw that our error estimates for the semidiscrete state equation allow to estimate inexactness in the POD/RB-reduced evalutation of the reduced objective functional j^h of the semidiscrete control problem (P_h). As far as we know, all recent approaches for the adaptive coupling of MOR and numerical optimization for a nonlinear problem like (P_h), see, e.g., [**38, 232, 236, 183, 21**], require a-posteriori POD/RB error estimates both for the evaluation of j^h and its gradient. For the latter, in essence an a-posteriori POD/RB error estimate w.r.t. the $L^2(I, H^1_{\Gamma_D})$-norm for the adjoint equation of (P_h) would be needed. We refer to, e.g., [**232, 236**] for such estimates in the case of different, linear and semilinear parabolic model problems. An extension of these results to our quasilinear parabolic problem seems to be a challenging task. Let us briefly explain the main obstruction. The coefficients of the POD/RB-reduced adjoint equation depend on the POD/RB-reduced state. For the POD/RB error of the state we have at hand a-posteriori errors w.r.t. the $L^\infty(I, L^2)$- and the $L^2(I, H^1_{\Gamma_D})$-norm so far, cf. Theorem 5.6, but not, e.g., w.r.t. the $L^\infty(Q)$-norm. The latter, however, would be a typical norm to measure inexactnes in the coefficients of a PDE. Consequently, obtaining a-posteriori POD/RB error estimates for the adjoint equation poses the following additional difficulty: we have to deal with inexactnes in the coefficients inferred by an inexact state, and we can only estimate this inexactnes in the coefficients with the help of existing a-posteriori POD/RB error estimates for the state. If possible at all, we believe that this will require rather strong assumptions. Therefore, we think that as a first step it could be worth considering a model problem with a slightly easier equation instead in which the quasilinear nonlinearity is of integral (nonlocal) type as in, e.g., [**82**] or [**81**, Chapter 12.2].

5.3. POD for the optimal control problem

In the previous section we have presented a-posteriori POD/RB error estimates for the state equation of our control problem (P) that have been obtained in [**166**]. Further, we explained that state of the art techniques for the adaptive coupling of POD/RB-MOR and numerical optimization would require us to have at hand also such estimates for the adjoint equation. This topic, however, is an open problem whose difficulty essentially arises from the combination of the linear nature of POD/RB methods with the highly nonlinear structure of our problem.

Nevertheless, POD-MOR allows to reduced the computational costs of solving (P) numerically by a significant amount. In our numerical experiments presented in the following we observe, e.g., a speedup factor of about 10. Hereby, instead of utilizing a fully certified method [**38, 232, 236, 183, 21**] based on a-posteriori error control we roughly follow earlier ideas on the combination of POD-MOR and SQP-type optimization algorithms from [**116, 38**] and formulate a purely heuristic, estimate-free method. Consequently, our contribution in this section is of purely computational nature and we do not address a theoretical analysis of the applied method. In some sense one may view this section as experimental continuation and extension of our analysis of the SQP method in Chapter 4.

We consider two approaches for the coupling of POD-MOR and SQP(-type) methods that will be described precisely in Section 5.3.1. The first one, called "POD-SQP" from now on, has been introduced in [**181**]. In a nutshell, one initially determines a POD basis and reduces all computations of the classical SQP method w.r.t. this basis. The second approach, called "POD-Newton-SQP" in the following, has —to the best of our knowledge— not been considered in the literature so far, although it has some similarties to different known ideas as we will explain in more detail below. In essence, for a Newton-like modification of the SQP method, called "Newton-SQP" from know on, we solve the linear quadratic subproblems with the help of POD-MOR. Hereby, in each Newton-SPQ-iteration a new POD basis is determined from the corresponding full-order solutions of the nonlinear state and the adjoint equation.

Of course, we expect that POD-SQP is more efficient because a larger share of the computations is performed within the reduced-order model than in POD-Newton-SQP. On the other hand, our numerical results in Section 5.3.2 indicate that due to the update of the POD basis in each iteration POD-Newton-SQP is less sensitive w.r.t. the initialization than POD-SQP. Regarding computing times, POD-Newton-SQP achieves a reduction of the computational costs of roughly factor 10 compared to Newton-SQP in our experiments.

5.3.1. POD-SQP, Newton-SQP, and POD-Newton-SQP. Let us now give some more details on the numerical schemes under consideration. As done for the SQP method in Chapter 4, we state all algorithms in function space. Of course, all computations formulated in functions spaces have to be carried out utilizing an appropriate space and time discretization in practice. We will not address this aspect in more detail.

We start with POD-SQP, an ansatz for the coupling of POD-MOR and the SQP method that goes back to [**181**] where it was applied to bilinear elliptic problems. We follow the variant in [**116**, Algorithm 7.2] concerned with semilinear parabolic problems, see also [**117**], and describe a simplified version hereof applied to our setting in Algorithm 1. Let us briefly summarize the simplifications in the formulation of this algorithm compared to [**116**]. First, we have fixed the length of the POD basis while the number of basis functions in [**116**] is chosen adaptively in each SQP iteration. Moreover, [**116**] also accounts for inexactness in the solutions of the subproblems. Nevertheless, the essence of the approach, i.e. applying an SQP method with the subproblems being POD-reduced w.r.t. an a-priori determined and fixed basis, is kept.

In order to propose a slightly different coupling of POD-MOR with an SQP-type optimization algorithm, let us as an intermediate step first explain the main idea of what we call Newton-SQP method in the following. We describe it in Algorithm 2. Herein, note that the SQP subproblem (QP) at (y_k, u_k, p_k) in line 7 of Algorithm 2 is equal to

$$u_{k+1} := \operatorname{argmin}_{u \in U_{\mathrm{ad}}} j'(u_k)(u - u_k) + \frac{1}{2} j''(u_k)(u - u_k)^2,$$

because y_k and p_k are state and adjoint state associated with u_k. The difference between the Newton-SQP and the SQP method is that in the Newton-SQP method we do not linearize state and adjoint equation; see lines 2 and 3 of Algorithm 2.

Algorithm 1: POD-reduced SQP method ("POD-SQP")

Input: initial guesses (y_0, u_0, p_0) for the KKT tripple, control $u_{\mathrm{sg}} \in U_{\mathrm{ad}}$ for the snapshot generation, accuracy parameter $\epsilon_{\mathrm{POD}} > 0$ for the generation of a POD basis, optimality tolerance $\epsilon_{\mathrm{SQP}} > 0$

Output: Approximation of the optimal state, control, and adjoint state

Compute the state $y_{\mathrm{sg}} = S(u_{\mathrm{sg}})$;
Compute the the adjoint state $p_{\mathrm{sg}} = \tilde{S}'(Bu_{\mathrm{sg}})^*(y_{\mathrm{sg}} - y_d)$;
Compute a POD basis $\mathcal{B}$ from y_{sg} and p_{sg} with accuracy $\epsilon_{\mathrm{POD}} > 0$;
for $k = 1, 2, 3, \ldots$ **do**
    Set up the POD-reduced (w.r.t. basis $\mathcal{B}$) SQP subproblem (QP) at $(y_{k-1}, u_{k-1}, p_{k-1})$;
    Solve this problem to obtain the next iterates (y_k, u_k, p_k);
    Compute $\mathrm{incr}_k = \|y_k - y_{k-1}\|_{L^\infty} + \|u_k - u_{k-1}\|_{L^\infty} + \|p_k - p_{k-1}\|_{L^\infty}$;
    if $\mathrm{incr}_k < \epsilon_{\mathrm{SQP}}$ **then**
        return y_k, u_k, p_k

This is similar to [**38**, Algorithm 5.2] or [**276**, Algorithm 3.3]. Solving the nonlinear parabolic state equation (instead of its linearization) is not a major issue since there are efficient nonlinear solvers available in many finite element software libraries. We call this approach Newton-SQP because it reduces to the classical Newton method for solving $\nabla j(\bar{u}) \stackrel{!}{=} 0$ in the special case without control-constraints. One may keep this in mind as a heuristic motivation of this approach.

Algorithm 2: Newton-SQP method

Input: initial guess $u_0 \in U_{\mathrm{ad}}$ for the control, optimality tolerance $\epsilon_{\mathrm{Newton}} > 0$

Output: Approximation of the optimal state, control, and adjoint state

for $k = 0, 1, 2, \ldots$ **do**
    Compute the state $y_k = S(Bu_k)$;
    Compute the adjoint state $p_k = \tilde{S}'(Bu_k)^*(y_k - y_d)$;
    Compute $\mathrm{res}_k = \|u_k - \mathrm{Proj}_{U_{\mathrm{ad}}}(-\gamma^{-1}B^*p_k)\|_{L^2(\Lambda)}$;
    if $\mathrm{res}_k < \epsilon_{\mathrm{Newton}}$ **then**
        return y_k, u_k, p_k
    Solve the SQP subproblem (QP) at (y_k, u_k, p_k) to obtain the next control iterate u_{k+1};

Including POD-MOR into Algorithm 2 is rather straightforward. Our suggestion inspired by [**38**] is formulated in Algorithm 3. Note that this approach still requires the solution of the nonlinear state and the adjoint equation in the full number of degrees of freedom. POD-MOR is only applied to the linear quadratic subproblems. This has the advantage that we do not require hyperreduction

Algorithm 3: POD-Newton-SQP method

Input: initial guess $u_0 \in U_{\text{ad}}$ for the control, optimality tolerance $\epsilon_{\text{Newton}} > 0$, accuracy parameter $\epsilon_{\text{POD}} > 0$ for the generation of a POD basis

Output: Approximation of the optimal state, control, and adjoint state

1 **for** $k = 0, 1, 2, \ldots$ **do**
2 Compute the state $y_k = S(Bu_k)$;
3 Compute the adjoint state $p_k = \tilde{S}'(Bu_k)^*(y_k - y_d)$;
4 Compute $\text{res}_k = \|u_k - \text{Proj}_{U_{\text{ad}}}(-\gamma^{-1}B^*p_k)\|_{L^2(\Lambda)}$;
5 **if** $\text{res}_k < \epsilon_{\text{Newton}}$ **then**
6 **return** y_k, u_k, p_k
7 Determine a POD basis $\mathcal{B}_k$ from y_k, p_k with accuracy $\epsilon_{\text{POD}} > 0$;
8 Solve the POD-reduced (w.r.t. $\mathcal{B}_k$) SQP subproblem (QP) at (y_k, u_k, p_k) to obtain the next control iterate u_{k+1};

techniques, cf. Section 5.2.3, to cope with nonlinearities. Also, we hope that computing a POD basis from the solution of the nonlinear state equation and the adjoint equation helps to capture the true dynamics of the underlying nonlinear problem. Since the solution of the quadratic subproblems is responsible for a major part of the overall computing time of the Newton-SQP method, POD-Newton-SQP still acchieves a good amount of reduction of the computational costs as we will see in our numerical examples. Moreover, we point out that in the case without control-constraints the proposed method essentially reduces to an inexact Newton method.

Let us briefly comment on differences and similarities of the POD-Newton-SQP method compared to the two most closely related approaches [**116**] and [**38**]. Except for the incorporation of inexact solution of the subproblems and (slightly heuristic) a-posteriori POD error estimation, we have already sketched the main idea of [**116**, Algorithm 7.2] at the beginning of this section. The main differences are that in the POD-Newton-SQP, first, we solve the nonlinear state and adjoint equation in each iteration, and, second, we determine a new POD basis in each iteration. In [**116**], a POD basis of maximal length is only generated once at the beginning and then (with adaptively chosen number of basis functions) used for the solution of all subproblems of the subsequently applied SQP method. The most obvious difference to [**38**] is, of course, that in [**38**, Algorithm 5.2] FEM discretization is adaptively refined during the algorithm while discretization is a priori fixed in our case. Moreover, we do not employ a trust-region framework. The approach in [**38**] and our POD-Newton-SQP have in common that they require solutions of the nonlinear state equation and the adjoint equation — however, in [**38**] this may also be done within the reduced-order model, while we perform this step always in the full number of degrees of freedom. Finally, although POD-Newton-SQP is not a trust-region algorithm, let us point out that the local model function in line 8 of Algorithm 3 differs from the ones used in [**19, 38, 236, 232,**

183, 22]; cf. the introduction of this chapter: since the subproblem in POD-Newton-SQP is given by the reduced-order counterpart of the SQP subproblem with full-order state and adjoint state, our local model is given by the reduced-order counterpart of the second-order Taylor polynomial of the full-order nonlinear objective functional.

5.3.2. Numerical examples. We test and compare the three algorithms described above on behalf of Example 3 from Section 4.5.2 utilizing discretization level 2 in Table 4.2. We present three numerical experiments. The first two are intended to illustrate the different influence of initialization in POD-SQP and POD-Newton-SQP. The third one compares Newton-SQP and POD-Newton-SQP for different accuracies of the POD-models and is also concerned with the respective computing times.

We implemented POD-SQP, Newton-SQP, and POD-Newton-SQP in python utilizing `fenics` and `mshr` [**9, 203**] for the finite element discretization. The appearing linear quadratic subproblems are solved similar to Section 4.5 by a semismooth Newton method that is now stopped after at most 10 iterations or if the initial residual (of the linear quadratic problem) is reduced by a factor $< 10^{-2}$. Again, stability of the semismooth Newton method is enhanced by Armijo linesearch, and the Newton update equations are solved by CG with at most 100 iterations and relative tolerance 10^{-2}. POD bases are determined by the method of snapshots from the snapshots both of the state- and the adjoint state-variable with equal weights $\alpha_j = 1$ and accuracy ϵ_{POD}; see Section 5.1 for an explanation hereof. As for the SQP method in Chapter 4 we illustrate the convergence behaviour of the algorithms by referring to the L^2-residuals of the iterates, defined and computed as in (4.26).

Experiment 1. To have a reference for comparison, we first address a specific aspect of the behaviour of the POD-SQP method. More precisely, we focus on the question how the choice of u_{sg} influences the outcome. Therefore, we choose a very high accuracy $\epsilon_{\mathrm{POD}} = 10^{-14}$ for the POD basis and, consequently, we do not compare computing times. We consider five different constellations:

Run 0. $u_{\mathrm{sg}}^{(0)}(t) = \bar{u}(t)$,
Run 1. $u_{\mathrm{sg}}^{(1)}(t) = y_d(t, \cdot)$,
Run 2. $u_{\mathrm{sg}}^{(2)}(t) = u_{\mathrm{a}}(t)$,
Run 3. $u_{\mathrm{sg}}^{(3)}(t) = y_d(T - t, \cdot)$,
Run 4. $u_{\mathrm{sg}}^{(4)}(t) = 2.9 + 7.1 \cdot \frac{1+\cos\left(\frac{2\pi}{40}t\right)}{2}$.

Here, we choose the initial guess (y_0, p_0) in the SQP method to be state and adjoint state associated with $u_{\mathrm{sg}}^{(k)}$ for Run 1-4 and $y_0 \equiv 0$, $p_0 \equiv 0$ for Run 0. One may view Run 0 as the POD-reduced version of the computations from Section 4.5.2 with the "optimal" POD basis. Figure 5.7 displays the behaviour of the POD-SQP algorithm for these different u_{sg}'s and Table 5.2 shows the respective POD ranks. As expected, it can be seen that the final accuracy (in terms of the residual) depends on the choice of the control u_{sg} used for the generation of the POD basis. The best results are achieved in Run 0 by using the optimal control $\bar{u}$ for the generation of the snapshots; of course, this only of theoretical interest because in applications $\bar{u}$ is not known in advance. To give the reader an impression how a

	Run 0	Run 1	Run 2	Run 3	Run 4
POD-ranks	17	9	18	18	14

Table 5.2. Experiment 1: POD ranks for the POD-SQP method for different u_{sg}'s.

	POD-ranks at iteration							
	0	1	2	3	4	5	6	7
Run 1	9	18	17	17	17	17	17	17
Run 2	18	23	17	17	17	17	17	17
Run 3	18	22	17	17	17	17	17	17
Run 4	14	24	17	17	17	17	17	17

Table 5.3. Experiment 2: POD ranks during the POD-Newton-SQP method for different initialization.

typical POD basis in the case of the present example may look like, we show the first five basis functions of the POD basis from Run 0 in Figure 5.5. The rapid decay of the corresponding eigenvalues in Figure 5.6 may be regarded as a heuristic explanation why POD-MOR allows to obtain good results for the present problem.

Remarkably, there does not seem to be an obvious, intuitive relation between u_{sg} and the final accuracy for Runs 1-4. The rather intuive guess for u_{sg} in Run 1 produces much worse results than the somehow contraintuitive choices for u_{sg} in Runs 3 and 4. This illustrates an important issue in the application of the POD-SQP algorithm: since there is no update of the POD basis during the SQP iteration, the choice of this single POD basis is crucial. However, finding such an appropriate basis without having a-priori knowledge on the solution of the problem and its characteristics is not straightforward.

Experiment 2. In our next experiment we demonstrate that the POD-Newton-SQP method is less sensitive w.r.t. the initialization because the POD basis is updated in each iteration. For comparison with Experiment 1 we choose the same high accuracy $\epsilon_{\mathrm{POD}} = 10^{-14}$ and repeat Run 1-4 from Experiment 1 for the POD-Newton-SQP method, now with four different choices for the initial guess $u_0^{(k)} = u_{\mathrm{sg}}^{(k)}$, $k = 1, ..., 4$. Figure 5.8 displays residuals and increments of the respective iterations; the respective POD ranks are shown in Table 5.3. Now, the final accuracy does not depend on the initialization, which is different to the behaviour of the POD-SQP method shown in Figure 5.7. Nevertheless, we have to note that this advantage is payed by performing a larger amount of computations outside the ROM than done in the POD-SQP method.

Experiment 3. Finally, we test the POD-Newton-SQP method for different ϵ_{POD} and compare it to the full-order Newton-SQP algorithm. We choose $\epsilon_{\mathrm{Newton}} = 10^{-5}$ and perform at most 8 Newton iterations. In Figure 5.9 we display the convergence behaviour of the two methods. The Newton-SQP method reaches the desired accuracy $\|\mathrm{res}_k\|_{L^2(\Lambda)} < \epsilon_{\mathrm{Newton}}$ after three steps. In Figure 5.9.a it can be seen that the POD-reduced version of the Newton-SQP method roughly behaves the same until the residuals stagnate at some level depending on the chosen accuracy of the POD-model. The increments (see Figure 5.9.b) still decrease which may

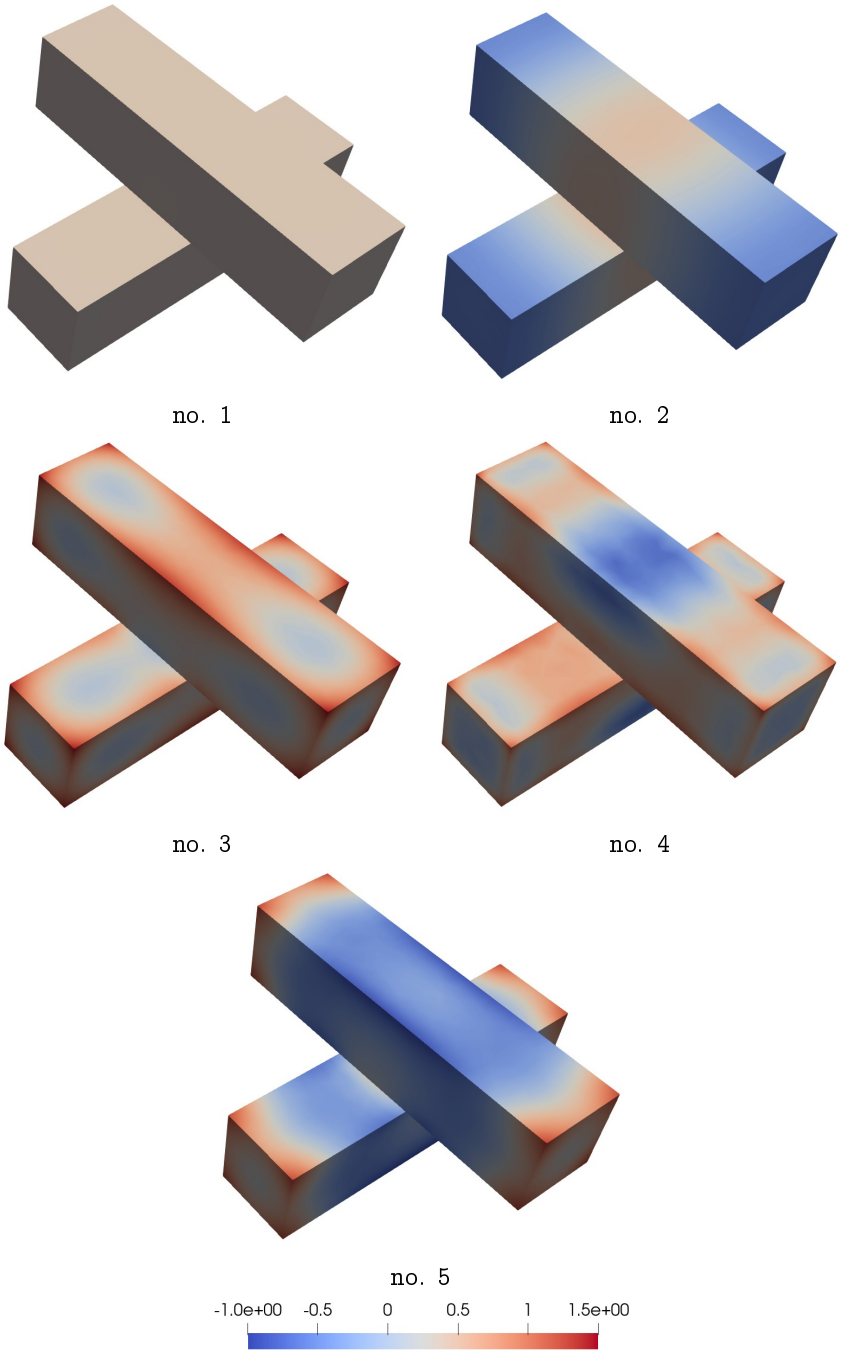


Figure 5.5. Experiment 1: The first five POD basis functions at $\bar{u}$. Of course, the plots can only show the values of the basis functions on the surface of Ω. For the corresponding eigenvalues we refer to Figure 5.6.

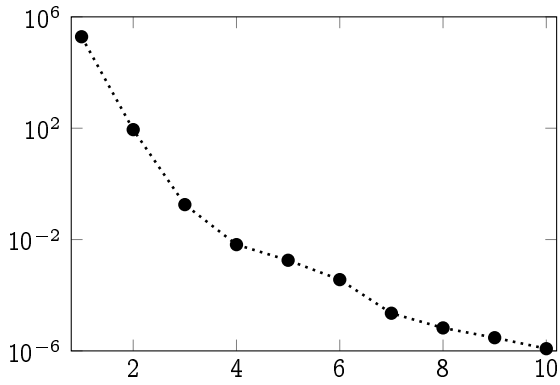


Figure 5.6. Experiment 1: POD eigenvalues at $\bar{u}$. Since we computed the POD basis with weights $\alpha_j = 1$ the eigenvalues have to be scaled by the factor $N_t^{-2} \approx 4 \cdot 10^{-6}$ in order to obtain approximations for the eigenvalues related to the continuous version of POD; cf. Section 5.1.

	a. POD ranks at iteration								**b. reduction of computing time** compared to Newton-SQP
ϵ_{POD}	0	1	2	3	4	5	6	7	all iterations (iterations 0-3 only)
10^{-6}	2	3	2	2	2	2	2	2	89% (94%)
10^{-7}	2	3	3	3	3	3	3	3	89% (94%)
10^{-8}	3	5	5	5	5	5	5	5	88% (93%)
10^{-9}	3	6	6	6	6	6	6	6	88% (93%)
10^{-10}	4	7	7	7	7	7	7	7	88% (93%)
10^{-11}	5	9	9	9	9	9	9	9	88% (93%)

Table 5.4. Experiment 3: **a.** POD ranks during POD-Newton-SQP for different POD accuracies. **b.** Reduction of the respective computing times of POD-Newton-SQP vs Newton-SQP. We compare the time needed for 3 iterations of Newton-SQP with the time for all 8 iterations of POD-Newton-SQP ("all iterations") or the times needed for the first 3 iterations of each method ("iterations 0-3 only").

indicate that the POD-Newton-SQP still converges, but only to some suboptimal control close to $\bar{u}$. However, note that it holds $\|\bar{u}\|_{L^2(I)} \approx 26.8$ for our problem, i.e. the final residuals for $\epsilon_{\text{POD}} = 10^{-6}$ and $\epsilon_{\text{POD}} = 10^{-7}$ are about 6% and 0.05% of $\|\bar{u}\|_{L^2(I)}$. Consequently, already POD-MOR with these rather low accuracies and corresponding low numbers of basis functions (see Table 5.4) yields good results. Indeed, as can be seen in Table 5.4.a, POD-MOR allows to reduce the numbers of degrees of freedom in the quadratic subproblems drastically; hereby, note that the underlying finite element space has 2848 degrees of freedom. Hence, POD-MOR acchieves a heavy reduction of the computing times by a factor of about 10; cf. Table 5.4.b. We think that this reduction will stay of comparable size or will even increase when choosing a finer finite element discretization.

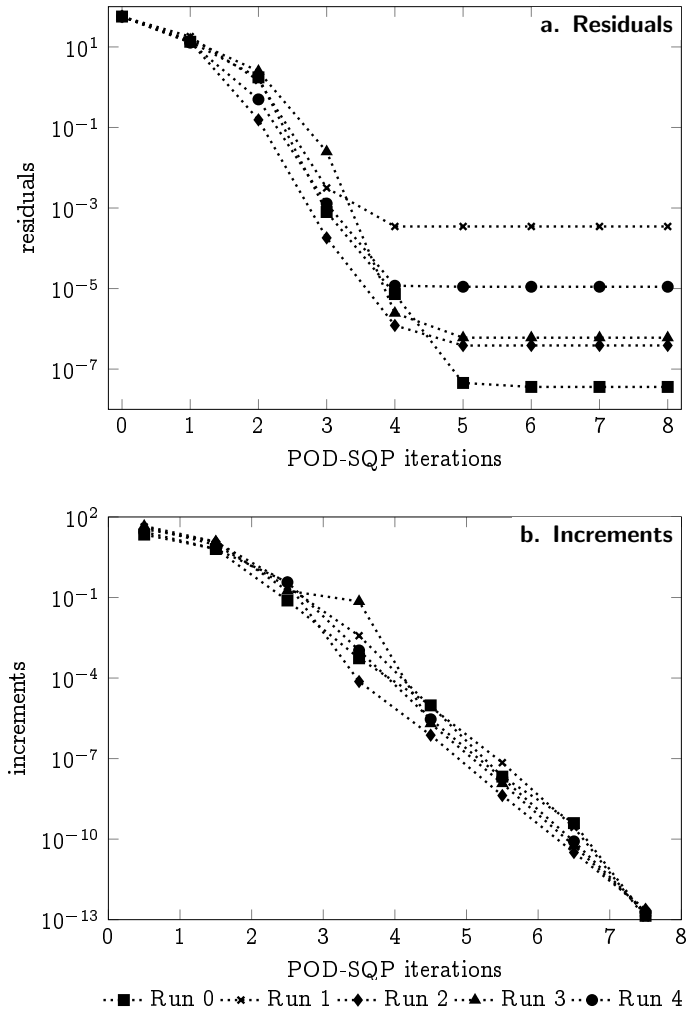


Figure 5.7. Experiment 1: **a.** L^2-norms of the residuals and **b.** L^∞-increments during the iterations of the POD-SQP method.

In order to conclude this section, let us mention some directions in which the implementation of the POD-Newton-SQP method might be refined. First, choosing separate POD bases for state and adjoint state or different POD accuracies in different (outer) iterations is possible. Second, for certain linear quadratic optimal control problems there is a well-known a-posteriori POD error estimate for the control-variable; see, e.g., [**272**]. Hereby, it is essential that the Hessian of the reduced problem is constant and coercive. The underlying idea has been generalized toward subproblems of the SQP method [**116**, Algorithm 7.1] or the nonlinear problem itself [**182, 266**] in the case of a semilinear parabolic state equation. Since in these cases the Hessian of the respective reduced problems cannot be supposed

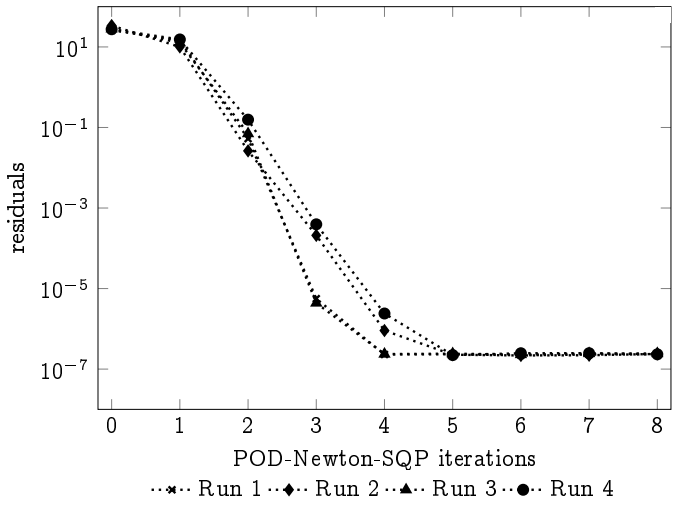


Figure 5.8. Experiment 2: L^2-norms of the residuals during the iterations of the POD-Newton-SQP method for different initialization.

to be uniformly coercive in general, these approaches are of slightly heuristic nature. For this reason we did not follow them in the present work. However, we have to note that, despite this issue on the theoretical side, the numerical performance observed in [**116, 182**] is outstanding and might be promising also for our problem type. Finally, one could also refine the implementation by combining POD-Newton-SQP with a trust-region framework as in, e.g., [**19**].

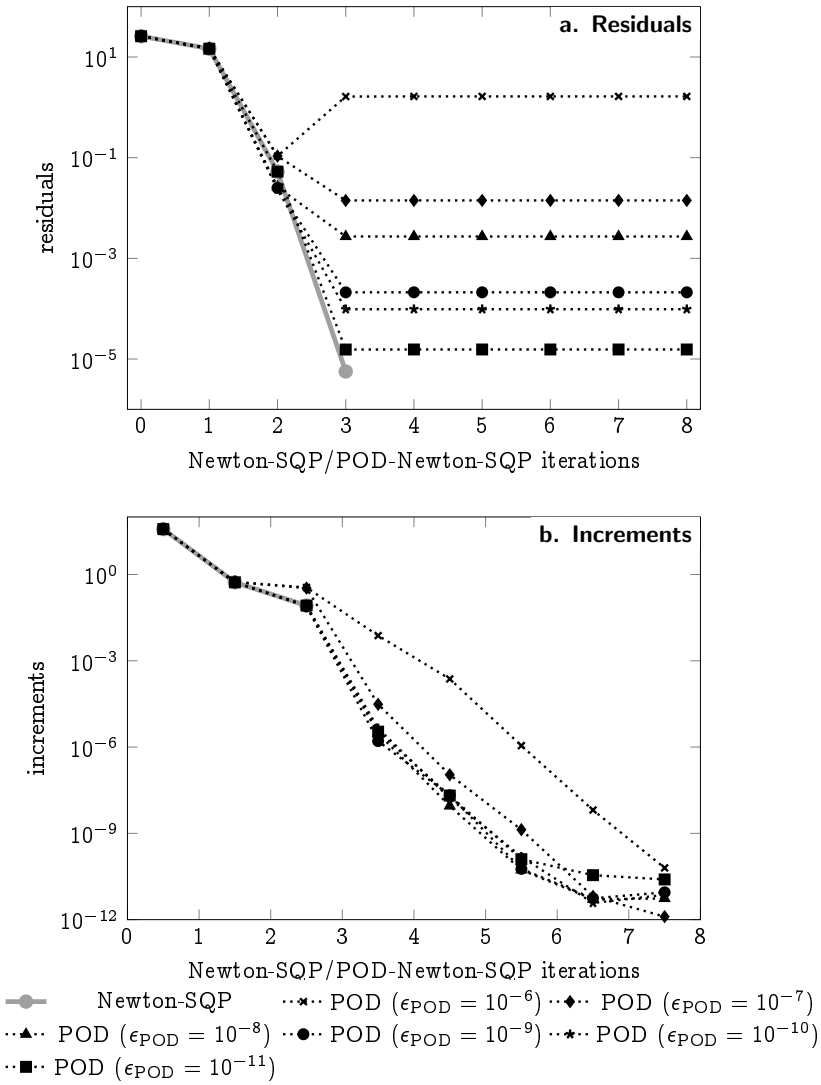


Figure 5.9. Experiment 3: **a.** L^2-norms of the residuals and **b.** L^∞-increments during the iterations of Newton-SQP and POD-Newton-SQP.

Conclusion and outlook

> Kein Buch wird jemals fertig; während wir daran arbeiten, lernen wir immer gerade genug, um seine Unzulänglichkeit zu sehen, wenn wir es der Öffentlichkeit übergeben.
>
> K. Popper, *Die offene Gesellschaft und ihre Feinde, Band I, Der Zauber Platons*[1]

In this thesis we have extensively discussed our recent contributions to optimal control of quasilinear parabolic PDEs [**167, 166, 168, 169**]. Following [**168**] and [**169**] we derived first- and second-order optimality conditions for problems with pointwise state-constraints and for problems with sparsity-enforcing penalization, respectively. Moreover, we presented the convergence analysis of the SQP method in function space from [**167**] and a-posteriori error estimates for POD/RB-model order reduction of the state equation from [**166**]. Finally, we demonstrated numerically that such reduction techniques also allow to speed up the solution of the entire control problem significantly. As the title indicates, we may regard our work in particular as the extension of earlier work on optimal control of quasilinear parabolic PDEs [**35, 216, 45**] towards the abovementioned additional aspects. However, it has to be pointed out that we also built on earlier work in the fields of state-constrained or sparse optimal control, the convergence analysis of optimization algorithms, and model order reduction for nonlinear problems. Therefore, the results presented in this thesis certainly contribute to these areas, too.

Let us briefly revisit the two read threads announced at the end of the introduction chapter. The first one, problems related to the highly nonlinear structure of the state equation, accompanied us through every chapter. In Chapter 2 we estimated the second derivative of the nonlinearity very carefully in order to guarantee that the second derivative of the control-to-state map exhibits appropriate extension properties required during the proof of second-order sufficient conditions; see the comments at the end of Section 2.3.3 and above Lemma 2.26. Similar regularity issues specifically due to the quasilinear parabolic PDE prevented us from addressing second-order conditions for the so-called bang-bang case in Chapter 3 as

[1] Preface to the first american edition 1950, quoted from: K. R. Popper, Gesammelte Werke in deutscher Sprache, Band 5, Mohr Siebeck, Tübingen, 2003.

we pointed out in detail in Section 3.3.3. Also, in our analysis of the SQP method in Chapter 4 the careful estimation of the second derivative of the nonlinearity combined with matching estimates for the linearized state equation and the adjoint equation was crucial. We explained this in particular in Section 4.2.2 and in the remark below Theorem 4.11. Finally, the nonmonotone structure of the nonlinearity turned out to be a major difficulty and novelty when aiming at POD/RB error estimates for the state equation as we saw in Section 5.2; cf. particularly the remark below Lemma 5.4.

The second recurring topic, the two-norm discrepancy and optimality conditions with/without norm gap, was of particular importance in Chapters 2 to 4. Our second-order results —Theorems 2.18 and 2.29 for problems with state-constraints, and Theorem 3.4 for sparse problems— avoid a norm gap although a two-norm discrepancy appears in the problem formulation. The basis of our analysis is [**71**] where second-order conditions without a two-norm gap despite the presence of two norms in the problem formulation have been proven for certain control-constrained problems on behalf of an abstract framework. In Sections 2.2 and 3.2 of this thesis, we extended the abstract framework from [**71**] in two different directions: second-order sufficient conditions in the case of additional state-constraints, and second-order conditions for problems with certain nonsmooth terms in the functional. To do so, we abstracted and extended techniques known from the literature; cf. the references given in Chapters 2 and 3. Nevertheless, finding appropriate abstract formulations that in particular allow to handle the concrete quasilinear parabolic problems (P^{st}) and ($\mathrm{P}_k^{\mathrm{sp}}$) was not trivial. Finally, we encountered the the topic of second-order sufficient conditions in Chapter 4 again, as explained in the introduction of this chapter. We were able to avoid another norm gap when restricting the SQP subproblems to L^2- instead of L^∞-balls around the optimal control. To do so, we essentially extended the results of [**71**] in Sections 4.4.2 and 4.4.3 towards a family of perturbed problems satisfying suitable assumptions; see in particular Proposition 4.17.

Let us now give an outlook to ongoing and closely related work that has not been included in this thesis. Currently, two results in collaboration with L. Bonifacius (Munich), H. Meinlschmidt (University of Erlangen) and I. Neitzel are in preparation. The first one addresses improved regularity of the state equation on the $[W_D^{-1,p}, L^p]$-scale. The second one deals with an optimal control problem with additional constraints on the gradient of the state, governed by a quasilinear parabolic PDE including a gradient term.

The main goal of the improved regularity analysis ([**165**], joint work with H. Meinlschmidt and I. Neitzel) is global-in-time existence of solutions to a quasilinear parabolic equation similar to the one from [**216**], cf. (0.1), on the whole scale of Bessel potential spaces $[W^{-1,p}, L^p]_\theta$, $\theta \in [0,1]$. Roughly speaking, our aim is to close the gap between [**216**] ($\theta = 0$), [**35**] ($\theta > 0$, but close to 0), and [**45**] ($\theta = 1$), while keeping the respective regularity assumptions w.r.t. the domain, the coefficient functions, and the boundary conditions as general as possible. This requires, e.g., a careful analysis of maximal parabolic regularity of certain nonautonomous parabolic operators with Hölder continuous coefficients on Bessel potential spaces and the computation of interpolation spaces of various function

spaces under rather low regularity assumptions. Let us note that the resulting improved regularity theory for (Eq) would allow, e.g., to simply the arguments in Section 2.4 a bit. Moreover, as one may already have expected, pure homogeneous Neumann boundary conditions can be allowed in Assumption 2.19.1, too.

The second part of our ongoing work ([**34**], joint work with L. Bonifacius, H. Meinlschmidt, I. Neitzel) deals with optimal control of a quasilinear parabolic PDE in the presence of gradient-constraints and may therefore be viewed as evident continuation of our results on constraints on the state in this thesis. First of all, a typical difficulty related to pointwise constraints on the gradient of the state is the following: proving first-order conditions in that case usually requires having states with continuous gradients; cf., e.g., [**51**] dealing with semilinear elliptic equations. Of course, obtaining such regularity for the solutions of a quasilinear parabolic PDE is nontrivial and requires rather strong assumptions, e.g., in the flavour of [**45**]. Consequently, when aiming at results in a rough regularity setting as in [**35**] one has to switch to averaged-type gradient constraints; cf., e.g., [**206, 60, 205**]. Moreover, we include another issue in our considerations: when adding a semilinear term of order one, e.g., a quadratic gradient term, to the state equation (Eq), we encounter the problem that we cannot prove existence of global-in-time solutions to the resulting equation, in general. Interestingly, this can be overcome by imposing certain bounds on the gradient of the solution; see for instance [**214**] for a related idea. Besides applications requiring bounds on the gradient of the state, this can be seen as a further motivation for the consideration of problems with gradient-constraints.

Finally, let us mention sparse optimal control of quasilinear PDEs as a promising area of future research. Besides sparsity-enforcing penalization, which we have considered in Chapter 3 of this thesis, approaches that rely on measure spaces are particulary interesting. In general, sparse optimal control in measure spaces is known to be challenging already for linear parabolic [**48, 192, 56**] or semilinear elliptic problems [**55**]. As a first step in the direction of quasilinear control problems in measure spaces the author has addressed sparse optimal control of a quasilinear elliptic PDE in measure spaces in the preprint [**164**]. There are several difficulties associated with this problem type. First, one has to prove well-posendness of the optimal control problem by ensuring existence, uniqueness, and sufficient regularity of solutions to the state equation for controls of very low regulartiy. This is a particular problem for parabolic equations, cf., e.g., [**56**] where the states do not have $L^2(Q)$-regularity, or nonlinear equations [**55**]. Second, in the case of a nonlinear state equation the investigation of differentiability properties of the control-to-state map becomes particulary challenging. This is due to the fact that differentiability of the nonlinear terms has to be addressed in appropriate, sufficiently regular function spaces, while solutions to PDEs with measure right-hand sides tend to have low regularity; see, e.g., [**55**]. A further difficulty arises from the presence of the total variation-norm of the underlying measure space in the objective functional, which makes the control problem a nonsmooth problem. In [**164**] we apply the so-called Kirchhoff transform, a nonlinear superposition operator that transforms the quasilinear elliptic equation into a linear one, to cope with these issues. In particular, we are able to prove well-posedness of the state

equation, obtain existence of optimal controls, and derive first- and second-order optimality conditions.

Bibliography

If I have seen further it is by standing on y^{e} shoulders of giants.

I. Newton, letter to R. Hooke, February 05, 1676[2]

1. K. Afanasiev and M. Hinze, *Adaptive control of a wake flow using proper orthogonal decomposition*, Shape optimization and optimal design (Cambridge, 1999), Lecture Notes in Pure and Appl. Math., vol. 216, Dekker, New York, 2001, pp. 317–332.
2. A. Ahmad Ali, K. Deckelnick, and M. Hinze, *Global minima for semilinear optimal control problems*, Comput. Optim. Appl. **65** (2016), no. 1, 261–288.
3. A. Alla, M. Falcone, and S. Volkwein, *Error analysis for POD approximations of infinite horizon problems via the dynamic programming approach*, SIAM J. Control Optim. **55** (2017), no. 5, 3091–3115.
4. A. Alla, C. Gräßle, and M. Hinze, *A residual based snapshot location strategy for POD in distributed optimal control of linear parabolic equations*, IFAC-PapersOnLine **49** (2016), no. 8, 13–18.
5. ———, *A posteriori snapshot location for POD in optimal control of linear parabolic equations*, ESAIM Math. Model. Numer. Anal. **52** (2018), no. 5, 1847–1873.
6. A. Alla, M. Hinze, P. Kolvenbach, O. Lass, and S. Ulbrich, *A certified model reduction approach for robust parameter optimization with PDE constraints*, Adv. Comput. Math. **45** (2019), no. 3, 1221–1250.
7. A. Alla and J. N. Kutz, *Nonlinear model order reduction via dynamic mode decomposition*, SIAM J. Sci. Comput. **39** (2017), no. 5, B778–B796.
8. A. Alla and S. Volkwein, *Asymptotic stability of POD based model predictive control for a semilinear parabolic PDE*, Adv. Comput. Math. **41** (2015), no. 5, 1073–1102.
9. M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, *The fenics project version 1.5*, Archive of Numerical Software **3** (2015), no. 100.
10. W. Alt, *The Lagrange-Newton method for infinite-dimensional optimization problems*, Numer. Funct. Anal. Optim. **11** (1990), no. 3-4, 201–224.
11. W. Alt, R. Griesse, N. Metla, and A. Rösch, *Lipschitz stability for elliptic optimal control problems with mixed control-state constraints*, Optimization **59** (2010), no. 5-6, 833–849.
12. H. Amann, *Linear and quasilinear parabolic problems. Vol. I*, Monographs in Mathematics, vol. 89, Birkhäuser Boston, Inc., Boston, MA, 1995, Abstract linear theory.
13. ———, *Linear parabolic problems involving measures*, RACSAM. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. **95** (2001), no. 1, 85–119.

[2]quoted from: R. S. Westfall, The Life of Isaac Newton, Cambridge University Press, Cambridge, 1993, p106.

14. _____, *Nonautonomous parabolic equations involving measures*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **306** (2003), no. Kraev. Zadachi Mat. Fiz. i Smezh. Vopr. Teor. Funktsii. 34, 16–52, 229.
15. _____, *Maximal regularity for nonautonomous evolution equations*, Adv. Nonlinear Stud. **4** (2004), no. 4, 417–430.
16. J. Appell and P. P. Zabrejko, *Nonlinear superposition operators*, Cambridge Tracts in Mathematics, vol. 95, Cambridge University Press, Cambridge, 1990.
17. N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl. **23** (2002), no. 2, 201–229.
18. N. Arada and J.-P. Raymond, *Optimal control problems with mixed control-state constraints*, SIAM J. Control Optim. **39** (2000), no. 5, 1391–1407.
19. E. Arian, M. Fahl, and E.W. Sachs, *Trust-region proper orthogonal decomposition for flow control*, Technical Report 2000-25, ICASE, 2000.
20. F. Bachmann, D. Beermann, J. Lu, and S. Volkwein, *POD-based mixed-integer optimal control of the heat equation*, J. Sci. Comput. **81** (2019), no. 1, 48–75.
21. S. Banholzer, T. Keil, L. Mechelli, M. Ohlberger, F. Schindler, and S. Volkwein, *An adaptive projected Newton non-conforming dual approach for trust-region reduced basis approximation of PDE-constrained parameter optimization*, Preprint, arXiv:2012.11653v1 [math.NA], To appear in Pure and Applied Functional Analysis, 2021.
22. S. Banholzer, E. Makarov, and S. Volkwein, *POD-based multiobjective optimal control of time-variant heat phenomena*, Numerical mathematics and advanced applications—ENUMATH 2017, Lect. Notes Comput. Sci. Eng., vol. 126, Springer, Cham, 2019, pp. 881–888.
23. M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, *An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations*, C. R. Math. Acad. Sci. Paris **339** (2004), no. 9, 667–672.
24. H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, second ed., CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, Cham, 2017, With a foreword by Hédy Attouch.
25. S. Bechtel and M. Egert, *Interpolation theory for Sobolev functions with partially vanishing trace on irregular open sets*, J. Fourier Anal. Appl. **25** (2019), no. 5, 2733–2781.
26. A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. **2** (2009), no. 1, 183–202.
27. P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. Schilders, and L. M. Silveira (eds.), *Model order reduction: Volume 2: Snapshot-based methods and algorithms*, De Gruyter, 2020.
28. P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. Schilders, and L. M. Silveira (eds.), *Model order reduction: Volume 3: Applications*, De Gruyter, 2020.
29. P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. Schilders, and L. M. Silveira (eds.), *Model order reduction: Volume 1: System- and data-driven methods and algorithms*, De Gruyter, 2021.
30. P. Benner, E. Sachs, and S. Volkwein, *Model order reduction for PDE constrained optimization*, Trends in PDE constrained optimization, Internat. Ser. Numer. Math., vol. 165, Birkhäuser/Springer, Cham, 2014, pp. 303–326.
31. M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim. **37** (1999), no. 4, 1176–1194.
32. L. Blank and J. Meisinger, *Optimal control of a quasilinear parabolic equation and its time discretization*, Preprint, arXiv:2102.02616, 2021.
33. C. Boehm and M. Ulbrich, *A semismooth Newton-CG method for constrained parameter identification in seismic tomography*, SIAM J. Sci. Comput. **37** (2015), no. 5, S334–S364.
34. L. Bonifacius, F. Hoppe, H. Meinlschmidt, and I. Neitzel, *Optimal control of quasilinear parabolic PDE with gradient terms and pointwise constraints on the gradient of the state*, Work in Progress, 2022.
35. L. Bonifacius and I. Neitzel, *Second order optimality conditions for optimal control of quasilinear parabolic equations*, Math. Control Relat. Fields **8** (2018), no. 1, 1–34.

36. J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*, Springer Series in Operations Research, Springer-Verlag, New York, 2000.
37. J. F. Bonnans and H. Zidani, *Optimal control problems with partially polyhedric constraints*, SIAM J. Control Optim. **37** (1999), no. 6, 1726–1741.
38. S. Bott, *Adaptive SQP Method with Reduced Order Models for Optimal Control Problems with Constraints on the State Applied to the Navier-Stokes Equations*, Dissertation, Technische Universität Darmstadt, 2016.
39. F. Campbell and G. I. Allen, *Within group variable selection through the Exclusive Lasso*, Electronic Journal of Statistics **11** (2017), no. 2, 4220 – 4257.
40. E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim. **24** (1986), no. 6, 1309–1318.
41. ———, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim. **31** (1993), no. 4, 993–1006.
42. ———, *Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints*, ESAIM Control Optim. Calc. Var. **8** (2002), 345–374.
43. ———, *Necessary and sufficient optimality conditions for elliptic control problems with finitely many pointwise state constraints*, ESAIM Control Optim. Calc. Var. **14** (2008), no. 3, 575–589.
44. ———, *A review on sparse solutions in optimal control of partial differential equations*, SeMA J. **74** (2017), no. 3, 319–344.
45. E. Casas and K. Chrysafinos, *Analysis and optimal control of some quasilinear parabolic equations*, Math. Control Relat. Fields **8** (2018), no. 3-4, 607–623.
46. ———, *Numerical analysis of quasilinear parabolic equations under low regularity assumptions*, Numer. Math. **143** (2019), no. 4, 749–780.
47. E. Casas, C. Clason, and K. Kunisch, *Approximation of elliptic control problems in measure spaces with sparse solutions*, SIAM J. Control Optim. **50** (2012), no. 4, 1735–1752.
48. ———, *Parabolic control problems in measure spaces with sparse solutions*, SIAM J. Control Optim. **51** (2013), no. 1, 28–63.
49. E. Casas, J. C. de los Reyes, and F. Tröltzsch, *Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints*, SIAM J. Optim. **19** (2008), no. 2, 616–643.
50. E. Casas and V. Dhamo, *Optimality conditions for a class of optimal boundary control problems with quasilinear elliptic equations*, Control Cybernet. **40** (2011), no. 2, 457–490.
51. E. Casas and L. A. Fernández, *Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state*, Appl. Math. Optim. **27** (1993), no. 1, 35–56.
52. E. Casas, R. Herzog, and G. Wachsmuth, *Approximation of sparse controls in semilinear equations by piecewise linear functions*, Numer. Math. **122** (2012), no. 4, 645–669.
53. ———, *Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional*, SIAM J. Optim. **22** (2012), no. 3, 795–820.
54. ———, *Analysis of spatio-temporally sparse optimal control problems of semilinear parabolic equations*, ESAIM Control Optim. Calc. Var. **23** (2017), no. 1, 263–295.
55. E. Casas and K. Kunisch, *Optimal control of semilinear elliptic equations in measure spaces*, SIAM J. Control Optim. **52** (2014), no. 1, 339–364.
56. ———, *Parabolic control problems in space-time measure spaces*, ESAIM Control Optim. Calc. Var. **22** (2016), no. 2, 355–370.
57. E. Casas and M. Mateos, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim. **40** (2002), no. 5, 1431–1454.
58. ———, *Uniform convergence of the FEM. Applications to state constrained control problems*, Comput. Appl. Math. **21** (2002), no. 1, 67–100.
59. ———, *Critical cones for sufficient second order conditions in PDE constrained optimization*, SIAM J. Optim. **30** (2020), no. 1, 585–603.
60. E. Casas, M. Mateos, and J.-P. Raymond, *Pontryagin's principle for the control of parabolic equations with gradient state constraints*, Nonlinear Anal. **46** (2001), no. 7, Ser. A: Theory Methods, 933–956.

61. E. Casas, M. Mateos, and A. Rösch, *Finite element approximation of sparse parabolic control problems*, Math. Control Relat. Fields **7** (2017), no. 3, 393–417.
62. ———, *Improved approximation rates for a parabolic control problem with an objective promoting directional sparsity*, Comput. Optim. Appl. **70** (2018), no. 1, 239–266.
63. 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. **20** (2014), no. 3, 803–822.
64. E. Casas, C. Ryll, and F. Tröltzsch, *Sparse optimal control of the Schlögl and FitzHugh-Nagumo systems*, Comput. Methods Appl. Math. **13** (2013), no. 4, 415–442.
65. ———, *Second order and stability analysis for optimal sparse control of the FitzHugh-Nagumo equation*, SIAM J. Control Optim. **53** (2015), no. 4, 2168–2202.
66. E. Casas and F. Tröltzsch, *Second-order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim. **39** (1999), no. 2, 211–227.
67. ———, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim. **13** (2002), no. 2, 406–431.
68. ———, *First- and second-order optimality conditions for a class of optimal control problems with quasilinear elliptic equations*, SIAM J. Control Optim. **48** (2009), no. 2, 688–718.
69. ———, *Numerical analysis of some optimal control problems governed by a class of quasilinear elliptic equations*, ESAIM Control Optim. Calc. Var. **17** (2011), no. 3, 771–800.
70. ———, *A general theorem on error estimates with application to a quasilinear elliptic optimal control problem*, Comput. Optim. Appl. **53** (2012), no. 1, 173–206.
71. E. Casas and F. Tröltzsch, *Second order analysis for optimal control problems: improving results expected from abstract theory*, SIAM J. Optim. **22** (2012), no. 1, 261–279.
72. E. Casas and F. Tröltzsch, *Second-order and stability analysis for state-constrained elliptic optimal control problems with sparse controls*, SIAM J. Control Optim. **52** (2014), no. 2, 1010–1033.
73. ———, *Second order optimality conditions and their role in PDE control*, Jahresber. Dtsch. Math.-Ver. **117** (2015), no. 1, 3–44.
74. ———, *State-constrained semilinear elliptic optimization problems with unrestricted sparse controls*, Math. Control Relat. Fields **10** (2020), no. 3, 527–546.
75. E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim. **38** (2000), no. 5, 1369–1391.
76. E. Casas, B. Vexler, and E. Zuazua, *Sparse initial data identification for parabolic PDE and its finite element approximations*, Math. Control Relat. Fields **5** (2015), no. 3, 377–399.
77. E. Casas and D. Wachsmuth, *First and second order conditions for optimal control problems with an L^0 term in the cost functional*, SIAM J. Control Optim. **58** (2020), no. 6, 3486–3507.
78. E. Casas, D. Wachsmuth, and G. Wachsmuth, *Sufficient second-order conditions for bang-bang control problems*, SIAM J. Control Optim. **55** (2017), no. 5, 3066–3090.
79. D. Chapelle, A. Gariah, and J. Sainte-Marie, *Galerkin approximation with proper orthogonal decomposition: new error estimates and illustrative examples*, ESAIM Math. Model. Numer. Anal. **46** (2012), no. 4, 731–757.
80. S. Chaturantabut and D. C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput. **32** (2010), no. 5, 2737–2764.
81. M. Chipot, *Elements of nonlinear analysis*, Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks], Birkhäuser Verlag, Basel, 2000.
82. M. Chipot and B. Lovat, *Some remarks on nonlocal elliptic and parabolic problems*, Proceedings of the Second World Congress of Nonlinear Analysts, Part 7 (Athens, 1996), vol. 30, 1997, pp. 4619–4627.
83. C. Christof and B. Vexler, *New regularity results and finite element error estimates for a class of parabolic optimal control problems with pointwise state constraints*, ESAIM Control Optim. Calc. Var. **27** (2021), Paper No. 4, 39.
84. C. Christof and G. Wachsmuth, *No-gap second-order conditions via a directional curvature functional*, SIAM J. Optim. **28** (2018), no. 3, 2097–2130.

85. ———, *On second-order optimality conditions for optimal control problems governed by the obstacle problem*, Optimization **70** (2021), no. 10, 2247–2287.
86. C. Clason and K. Kunisch, *A duality-based approach to elliptic control problems in non-reflexive Banach spaces*, ESAIM Control Optim. Calc. Var. **17** (2011), no. 1, 243–266.
87. C. Clason, V. H. Nhu, and A. Rösch, *No-gap second-order optimality conditions for optimal control of a non-smooth quasilinear elliptic equation*, ESAIM Control Optim. Calc. Var. **27** (2021), Paper No. 62, 35.
88. ———, *Optimal control of a non-smooth quasilinear elliptic equation*, Math. Control Relat. Fields **11** (2021), no. 3, 521–554.
89. C. Clason, V.H. Nhu, and A. Rösch, *Error estimates for the numerical approximation of a non-smooth quasilinear elliptic control problem*, Preprint, arXiv:2203.16865 [math.OC], 2022.
90. J. C. De los Reyes, *Numerical PDE-constrained optimization*, SpringerBriefs in Optimization, Springer, Cham, 2015.
91. J. C. de Los Reyes, P. Merino, J. Rehberg, and F. Tröltzsch, *Optimality conditions for state-constrained PDE control problems with time-dependent controls*, Control Cybernet. **37** (2008), no. 1, 5–38.
92. K. Deckelnick and M. Hinze, *Semidiscretization and error estimates for distributed control of the instationary Navier-Stokes equations*, Numer. Math. **97** (2004), no. 2, 297–320.
93. ———, *Convergence of a finite element approximation to a state-constrained elliptic control problem*, SIAM J. Numer. Anal. **45** (2007), no. 5, 1937–1953.
94. ———, *Variational discretization of parabolic control problems in the presence of pointwise state constraints*, J. Comput. Math. **29** (2011), no. 1, 1–15.
95. P. Deuflhard, A. Schiela, and M. Weiser, *Mathematical cancer therapy planning in deep regional hyperthermia*, Acta Numer. **21** (2012), 307–378.
96. M. A. Dihlmann and B. Haasdonk, *Certified PDE-constrained parameter optimization using reduced basis surrogate models for evolution problems*, Comput. Optim. Appl. **60** (2015), no. 3, 753–787.
97. K. Disser, H.-C. Kaiser, and J. Rehberg, *Optimal Sobolev regularity for linear second-order divergence elliptic operators occurring in real-world problems*, SIAM J. Math. Anal. **47** (2015), no. 3, 1719–1746.
98. K. Disser, A. F. M. ter Elst, and J. Rehberg, *Hölder estimates for parabolic operators on domains with rough boundary*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **17** (2017), no. 1, 65–79.
99. M. Dobrowolski, *L^∞-convergence of linear finite element approximation to nonlinear parabolic problems*, SIAM J. Numer. Anal. **17** (1980), no. 5, 663–674.
100. A. L. Dontchev, *Local analysis of a Newton-type method based on partial linearization*, The mathematics of numerical analysis (Park City, UT, 1995), Lectures in Appl. Math., vol. 32, Amer. Math. Soc., Providence, RI, 1996, pp. 295–306.
101. J. Douglas, Jr. and T. Dupont, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal. **7** (1970), 575–626.
102. S. S. Dragomir, *Some Gronwall type inequalities and applications*, Nova Science Publishers, Inc., Hauppauge, NY, 2003.
103. M. Drohmann, B. Haasdonk, and M. Ohlberger, *Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation*, SIAM J. Sci. Comput. **34** (2012), no. 2, A937–A969.
104. R. E. Edwards, *Functional analysis. Theory and applications*, Holt, Rinehart and Winston, New York-Toronto-London, 1965.
105. M. Egert, *On Kato's conjecture and mixed boundary conditions*, Dissertation, Technische Universität Darmstadt, 2015.
106. I. Ekeland and R. Témam, *Convex analysis and variational problems*, english ed., Classics in Applied Mathematics, vol. 28, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999, Translated from the French.
107. J. Elschner, J. Rehberg, and G. Schmidt, *Optimal regularity for elliptic transmission problems including C^1 interfaces*, Interfaces Free Bound. **9** (2007), no. 2, 233–252.

108. H. Feldhordt, *Boundary control of a chemotaxis system*, Dissertation, Universität Duisburg-Essen, 2017.
109. L. A. Fernández, *State constrained optimal control for some quasilinear parabolic equations*, Optimal control of partial differential equations (Chemnitz, 1998), Internat. Ser. Numer. Math., vol. 133, Birkhäuser, Basel, 1999, pp. 145–156.
110. ———, *Integral state-constrained optimal control problems for some quasilinear parabolic equations*, Nonlinear Anal. **39** (2000), no. 8, Ser. A: Theory Methods, 977–996.
111. J. Francu, *Monotone operators. A survey directed to applications to differential equations*, Apl. Mat. **35** (1990), no. 4, 257–301.
112. Matthias Gerdts, *Optimal control of ODEs and DAEs*, De Gruyter Textbook, Walter de Gruyter & Co., Berlin, 2012.
113. H. Goldberg and F. Tröltzsch, *Second-order sufficient optimality conditions for a class of nonlinear parabolic boundary control problems*, SIAM J. Control Optim. **31** (1993), no. 4, 1007–1025.
114. H. Goldberg and F. Tröltzsch, *On a Lagrange-Newton method for a nonlinear parabolic boundary control problem*, Optim. Methods Softw. **8** (1998), no. 3-4, 225–247.
115. W. Gong and M. Hinze, *Error estimates for parabolic optimal control problems with control and state constraints*, Comput. Optim. Appl. **56** (2013), no. 1, 131–151.
116. C. Gräßle, *POD based inexact SQP methods for optimal control problems governed by a semilinear heat equation*, Diplomarbeit, Universität Konstanz, 2014.
117. C. Gräßle, M. Gubisch, S. Metzdorf, S. Rogg, and S. Volkwein, *POD basis updates for nonlinear PDE control*, at - Automatisierungstechnik **65** (2017), no. 5.
118. C. Gräßle and M. Hinze, *POD reduced-order modeling for evolution equations utilizing arbitrary finite element discretizations*, Adv. Comput. Math. **44** (2018), no. 6, 1941–1978.
119. C. Gräßle, M. Hinze, J. Lang, and S. Ullmann, *POD model order reduction with space-adapted snapshots for incompressible flows*, Adv. Comput. Math. **45** (2019), no. 5-6, 2401–2428.
120. C. Gräßle, M. Hinze, and N. Scharmacher, *POD for optimal control of the Cahn-Hilliard system using spatially adapted snapshots*, Numerical mathematics and advanced applications—ENUMATH2017, Lect. Notes Comput. Sci. Eng., vol. 126, Springer, Cham, 2019, pp. 703–711.
121. C. Gräßle, M. Hinze, and S. Volkwein, *Model order reduction by proper orthogonal decomposition*, in Model Order Reduction, Volume 2: Snapshot-Based Methods and Algorithms, De Gruyter, 2020, pp. 47–96.
122. M. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera, *Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations*, M2AN Math. Model. Numer. Anal. **41** (2007), no. 3, 575–605.
123. J. A. Griepentrog, K. Gröger, H.-C. Kaiser, and J. Rehberg, *Interpolation for function spaces related to mixed boundary value problems*, Math. Nachr. **241** (2002), 110–120.
124. R. Griesse, N. Metla, and A. Rösch, *Convergence analysis of the SQP method for nonlinear mixed-constrained elliptic optimal control problems*, ZAMM Z. Angew. Math. Mech. **88** (2008), no. 10, 776–792.
125. R. Griesse, N. Metla, and A. Rösch, *Local quadratic convergence of SQP for elliptic optimal control problems with mixed control-state constraints*, Control Cybernet. **39** (2010), no. 3, 717–738.
126. R. Griesse and S. Volkwein, *A primal-dual active set strategy for optimal boundary control of a nonlinear reaction-diffusion system*, SIAM J. Control Optim. **44** (2005), no. 2, 467–494.
127. P. Grisvard, *Elliptic problems in nonsmooth domains*, Monographs and Studies in Mathematics, vol. 24, Pitman (Advanced Publishing Program), Boston, MA, 1985.
128. K. Gröger, *A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann. **283** (1989), no. 4, 679–687.
129. M. Gubisch, I. Neitzel, and S. Volkwein, *A-posteriori error estimation of discrete POD models for PDE-constrained optimal control*, Model reduction of parametrized systems, MS&A. Model. Simul. Appl., vol. 17, Springer, Cham, 2017, pp. 213–234.
130. M. Gubisch and S. Volkwein, *POD a-posteriori error analysis for optimal control problems with mixed control-state constraints*, Comput. Optim. Appl. **58** (2014), no. 3, 619–644.

131. ———, *Proper orthogonal decomposition for linear-quadratic optimal control*, Model reduction and approximation, Comput. Sci. Eng., vol. 15, SIAM, Philadelphia, PA, 2017, pp. 3–63.
132. R. Haller-Dintelmann, A. Jonsson, D. Knees, and J. Rehberg, *Elliptic and parabolic regularity for second-order divergence operators with mixed boundary conditions*, Math. Methods Appl. Sci. **39** (2016), no. 17, 5007–5026.
133. R. Haller-Dintelmann, C. Meyer, J. Rehberg, and A. Schiela, *Hölder continuity and optimal control for nonsmooth elliptic problems*, Appl. Math. Optim. **60** (2009), no. 3, 397–428.
134. R. Haller-Dintelmann and J. Rehberg, *Maximal parabolic regularity for divergence operators including mixed boundary conditions*, J. Differential Equations **247** (2009), no. 5, 1354–1396.
135. M. Heinkenschloss and F. Tröltzsch, *Analysis of the Lagrange-SQP-Newton method for the control of a phase field equation*, Control Cybernet. **28** (1999), no. 2, 177–211.
136. S. Herkt, M. Hinze, and R. Pinnau, *Convergence analysis of Galerkin POD for linear second order evolution equations*, Electron. Trans. Numer. Anal. **40** (2013), 321–337.
137. R. Herzog, J. Obermeier, and G. Wachsmuth, *Annular and sectorial sparsity in optimal control of elliptic equations*, Comput. Optim. Appl. **62** (2015), no. 1, 157–180.
138. R. Herzog, G. Stadler, and G. Wachsmuth, *Directional sparsity in optimal control of partial differential equations*, SIAM J. Control Optim. **50** (2012), no. 2, 943–963.
139. J. S. Hesthaven, G. Rozza, and B. Stamm, *Certified reduced basis methods for parametrized partial differential equations*, SpringerBriefs in Mathematics, Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2016, BCAM SpringerBriefs.
140. M. Hieber and J. Rehberg, *Quasilinear parabolic systems with mixed boundary conditions on nonsmooth domains*, SIAM J. Math. Anal. **40** (2008), no. 1, 292–305.
141. M. Hintermüller, *On a globalized augmented Lagrangian-SQP algorithm for nonlinear optimal control problems with box constraints*, Fast solution of discretized optimization problems (Berlin, 2000), Internat. Ser. Numer. Math., vol. 138, Birkhäuser, Basel, 2001, pp. 139–153.
142. M. Hintermüller and M. Hinze, *Globalization of SQP-methods in control of the instationary Navier-Stokes equations*, M2AN Math. Model. Numer. Anal. **36** (2002), no. 4, 725–746.
143. M. Hintermüller and M. Hinze, *A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier-Stokes system subject to control constraints*, SIAM J. Optim. **16** (2006), no. 4, 1177–1200.
144. M. Hintermüller and M. Hinze, *Moreau-Yosida regularization in state constrained elliptic control problems: error estimates and parameter adjustment*, SIAM J. Numer. Anal. **47** (2009), no. 3, 1666–1683.
145. M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim. **13** (2002), no. 3, 865–888 (2003).
146. M. Hintermüller and K. Kunisch, *Feasible and noninterior path-following in constrained minimization with low multiplier regularity*, SIAM J. Control Optim. **45** (2006), no. 4, 1198–1221.
147. ———, *PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative*, SIAM J. Optim. **20** (2009), no. 3, 1133–1156.
148. M. Hintermüller, S. Volkwein, and F. Diwoky, *Fast solution techniques in constrained optimal boundary control of the semilinear heat equation*, Control of coupled partial differential equations, Internat. Ser. Numer. Math., vol. 155, Birkhäuser, Basel, 2007, pp. 119–147.
149. M. Hinze, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Comput. Optim. Appl. **30** (2005), no. 1, 45–61.
150. M. Hinze and D. Korolev, *Reduced basis methods for quasilinear elliptic PDEs with applications to permanent magnet synchronous motors*, Model reduction of complex dynamical systems, Internat. Ser. Numer. Math., vol. 171, Birkhäuser/Springer, Cham, 2021, pp. 307–326.
151. ———, *A space-time certified reduced basis method for quasilinear parabolic partial differential equations*, Adv. Comput. Math. **47** (2021), no. 3, Paper No. 36, 26.
152. M. Hinze and K. Kunisch, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim. **40** (2001), no. 3, 925–946.
153. ———, *Second order methods for boundary control of the instationary Navier-Stokes system*, ZAMM Z. Angew. Math. Mech. **84** (2004), no. 3, 171–187.

154. M. Hinze and C. Meyer, *Variational discretization of Lavrentiev-regularized state constrained elliptic optimal control problems*, Comput. Optim. Appl. **46** (2010), no. 3, 487–510.
155. ———, *Stability of semilinear elliptic optimal control problems with pointwise state constraints*, Comput. Optim. Appl. **52** (2012), no. 1, 87–114.
156. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*, Mathematical Modelling: Theory and Applications, vol. 23, Springer, New York, 2009.
157. M. Hinze and A. Rösch, *Discretization of optimal control problems*, Constrained optimization and optimal control for partial differential equations, Internat. Ser. Numer. Math., vol. 160, Birkhäuser/Springer Basel AG, Basel, 2012, pp. 391–430.
158. M. Hinze and A. Schiela, *Discretization of interior point methods for state constrained elliptic optimal control problems: optimal error estimates and parameter adjustment*, Comput. Optim. Appl. **48** (2011), no. 3, 581–600.
159. M. Hinze and F. Tröltzsch, *Discrete concepts versus error analysis in PDE-constrained optimization*, GAMM-Mitt. **33** (2010), no. 2, 148–162.
160. M. Hinze and S. Volkwein, *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition*, Comput. Optim. Appl. **39** (2008), no. 3, 319–345.
161. D. Hömberg, K. Krumbiegel, and N. Togobytska, *Optimal control of multiphase steel production*, J. Math. Ind. **9** (2019), Paper No. 6, 32.
162. D. Hömberg, C. Meyer, J. Rehberg, and W. Ring, *Optimal control for the thermistor problem*, SIAM J. Control Optim. **48** (2009/10), no. 5, 3449–3481.
163. D. Hömberg and S. Volkwein, *Control of laser surface hardening by a reduced-order approach using proper orthogonal decomposition*, Math. Comput. Modelling **38** (2003), no. 10, 1003–1028.
164. F. Hoppe, *Sparse optimal control of a quasilinear elliptic PDE in measure spaces*, Submitted, Available as INS-Preprint No. 2202 under https://ins.uni-bonn.de/media/public/publication-media/INSPreprint2202.pdf, 2022.
165. F. Hoppe, H. Meinlschmidt, and I. Neitzel, *Global-in-time solutions for quasilinear parabolic PDEs on the $[W_D^{-1,p}, L^p]$-scale*, Work in Progress, 2022.
166. F. Hoppe and I. Neitzel, *A-posteriori reduced basis error-estimates for a semi-discrete in space quasilinear parabolic PDE*, Comput. Optim. Appl. (2021).
167. ———, *Convergence of the SQP method for quasilinear parabolic optimal control problems*, Optim. Eng. **22** (2021), no. 4, 2039–2085.
168. ———, *Optimal Control of Quasilinear Parabolic PDEs with State-Constraints*, SIAM J. Control Optim. **60** (2022), no. 1, 330–354.
169. ———, *Purely time-dependent optimal control of quasilinear parabolic PDEs with sparsity enforcing penalization*, In Revision at COCV, Available as INS-Preprint No. 2201 under https://ins.uni-bonn.de/media/public/publication-media/2201.pdf, 2022.
170. H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology **24** (1933), no. 6, 417–441.
171. L. Iapichino, S. Trenz, and S. Volkwein, *Reduced-order multiobjective optimal control of semilinear parabolic problems*, Numerical mathematics and advanced applications—ENUMATH 2015, Lect. Notes Comput. Sci. Eng., vol. 112, Springer, [Cham], 2016, pp. 389–397.
172. L. Iapichino, S. Ulbrich, and S. Volkwein, *Multiobjective PDE-constrained optimization using the reduced-basis method*, Adv. Comput. Math. **43** (2017), no. 5, 945–972.
173. L. Iapichino, S. Volkwein, and A. Wesche, *A-posteriori error analysis for lithium-ion concentrations in batteries utilizing the reduced-basis method*, Math. Comput. Model. Dyn. Syst. **22** (2016), no. 4, 362–379.
174. A. D. Ioffe, *Necessary and sufficient conditions for a local minimum. III. Second order conditions and augmented duality*, SIAM J. Control Optim. **17** (1979), no. 2, 266–288.
175. K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim. **34** (1996), no. 3, 874–891.
176. ———, *Semi-smooth Newton methods for state-constrained optimal control problems*, Systems Control Lett. **50** (2003), no. 3, 221–228.

177. ———, *The primal-dual active set method for nonlinear optimal control problems with bilateral constraints*, SIAM J. Control Optim. **43** (2004), no. 1, 357–376.
178. ———, *Optimal control with $L^p(\Omega)$, $p \in [0,1)$, control cost*, SIAM J. Control Optim. **52** (2014), no. 2, 1251–1275.
179. I. T. Jolliffe, *Principal component analysis*, second ed., Springer Series in Statistics, Springer-Verlag, New York, 2002.
180. N. H. Josephy, *Newton's Method for Generalized Equations*, Tech. report, 1979.
181. Kahlbacher, M. and Volkwein, S., *POD a-posteriori error based inexact SQP method for bilinear elliptic optimal control problems*, ESAIM: M2AN **46** (2012), no. 2, 491–511.
182. E. Kammann, F. Tröltzsch, and S. Volkwein, *A posteriori error estimation for semilinear parabolic optimal control problems with application to model reduction by POD*, ESAIM Math. Model. Numer. Anal. **47** (2013), no. 2, 555–581.
183. T. Keil, L. Mechelli, M. Ohlberger, F. Schindler, and S. Volkwein, *A non-conforming dual approach for adaptive trust-region reduced basis approximation of PDE-constrained parameter optimization*, ESAIM Math. Model. Numer. Anal. **55** (2021), no. 3, 1239–1269.
184. B. T. Kien, V. H. Nhu, and A. Rösch, *Second-order necessary optimality conditions for a class of optimal control problems governed by partial differential equations with pure state constraints*, J. Optim. Theory Appl. **165** (2015), no. 1, 30–61.
185. B. T. Kien, V. H. Nhu, and N. H. Son, *Second-order optimality conditions for a semilinear elliptic optimal control problem with mixed pointwise constraints*, Set-Valued Var. Anal. **25** (2017), no. 1, 177–210.
186. D. Korolev, *Reduced basis methods for quasilinear PDEs and applications to electrical machines*, Dissertation, Universität Koblenz-Landau, 2021.
187. M. Kowalski, *Sparse regression using mixed norms*, Appl. Comput. Harmon. Anal. **27** (2009), no. 3, 303–324.
188. K. Krumbiegel, C. Meyer, and A. Rösch, *A priori error analysis for linear quadratic elliptic Neumann boundary control problems with control and state constraints*, SIAM J. Control Optim. **48** (2010), no. 8, 5108–5142.
189. K. Krumbiegel and J. Rehberg, *Second order sufficient optimality conditions for parabolic optimal control problems with pointwise state constraints*, SIAM J. Control Optim. **51** (2013), no. 1, 304–331.
190. K. Krumbiegel and A. Rösch, *A virtual control concept for state constrained optimal control problems*, Comput. Optim. Appl. **43** (2009), no. 2, 213–233.
191. K. Kunisch and M. Müller, *Uniform convergence of the POD method and applications to optimal control*, Discrete Contin. Dyn. Syst. **35** (2015), no. 9, 4477–4501.
192. K. Kunisch, K. Pieper, and B. Vexler, *Measure valued directional sparsity for parabolic optimal control problems*, SIAM J. Control Optim. **52** (2014), no. 5, 3078–3108.
193. K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math. **90** (2001), no. 1, 117–148.
194. ———, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM J. Numer. Anal. **40** (2002), no. 2, 492–515.
195. ———, *Proper orthogonal decomposition for optimality systems*, M2AN Math. Model. Numer. Anal. **42** (2008), no. 1, 1–23.
196. ———, *Optimal snapshot location for computing POD basis functions*, M2AN Math. Model. Numer. Anal. **44** (2010), no. 3, 509–529.
197. K. Kunisch, S. Volkwein, and L. Xie, *HJB-POD-based feedback design for the optimal control of evolution problems*, SIAM J. Appl. Dyn. Syst. **3** (2004), no. 4, 701–722.
198. O. Lass and S. Volkwein, *POD Galerkin schemes for nonlinear elliptic-parabolic systems*, SIAM J. Sci. Comput. **35** (2013), no. 3, A1271–A1298.
199. ———, *Adaptive POD basis computation for parametrized nonlinear systems using optimal snapshot location*, Comput. Optim. Appl. **58** (2014), no. 3, 645–677.
200. ———, *Parameter identification for nonlinear elliptic-parabolic systems with application in lithium-ion battery modeling*, Comput. Optim. Appl. **62** (2015), no. 1, 217–239.
201. J.-L. Lions, *Optimal control of systems governed by partial differential equations*, Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170, Springer-Verlag, New York-Berlin, 1971.

202. W. Liu, H. Ma, T. Tang, and N. Yan, *A posteriori error estimates for discontinuous Galerkin time-stepping method for optimal control problems governed by parabolic equations*, SIAM J. Numer. Anal. **42** (2004), no. 3, 1032–1061.
203. A. Logg, K.-A. Mardal, G. N. Wells, et al., *Automated solution of differential equations by the finite element method*, Springer, 2012.
204. F. Ludovici, I. Neitzel, and W. Wollner, *A priori error estimates for state-constrained semilinear parabolic optimal control problems*, J. Optim. Theory Appl. **178** (2018), no. 2, 317–348.
205. F. Ludovici and W. Wollner, *A priori error estimates for a finite element discretization of parabolic optimization problems with pointwise constraints in time on mean values of the gradient of the state*, SIAM J. Control Optim. **53** (2015), no. 2, 745–770.
206. U. Mackenroth, *On parabolic distributed optimal control problems with restrictions on the gradient*, Appl. Math. Optim. **10** (1983), no. 1, 69–95.
207. Y. Maday, O. Mula, and G. Turinici, *Convergence analysis of the generalized empirical interpolation method*, SIAM J. Numer. Anal. **54** (2016), no. 3, 1713–1731.
208. M. M. Mäkelä and P. Neittaanmäki, *Nonsmooth optimization*, World Scientific Publishing Co., Inc., River Edge, NJ, 1992, Analysis and algorithms with applications to optimal control.
209. F. Mannel and A. Rund, *A hybrid semismooth quasi-Newton method for nonsmooth optimal control with PDEs*, Optim. Eng. **22** (2021), no. 4, 2087–2125.
210. L. Mechelli and S. Volkwein, *POD-based economic model predictive control for heat-convection phenomena*, Numerical mathematics and advanced applications—ENUMATH 2017, Lect. Notes Comput. Sci. Eng., vol. 126, Springer, Cham, 2019, pp. 663–671.
211. D. Meidner, R. Rannacher, and B. Vexler, *A priori error estimates for finite element discretizations of parabolic optimization problems with pointwise state constraints in time*, SIAM J. Control Optim. **49** (2011), no. 5, 1961–1997.
212. D. Meidner and B. Vexler, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems. I. Problems without control constraints*, SIAM J. Control Optim. **47** (2008), no. 3, 1150–1177.
213. ———, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems. II. Problems with control constraints*, SIAM J. Control Optim. **47** (2008), no. 3, 1301–1329.
214. H. Meinlschmidt, C. Meyer, and J. Rehberg, *Optimal control of the thermistor problem in three spatial dimensions, Part 1: Existence of optimal solutions*, SIAM J. Control Optim. **55** (2017), no. 5, 2876–2904.
215. ———, *Optimal control of the thermistor problem in three spatial dimensions, Part 2: Optimality conditions*, SIAM J. Control Optim. **55** (2017), no. 4, 2368–2392.
216. H. Meinlschmidt and J. Rehberg, *Hölder-estimates for non-autonomous parabolic problems with rough data*, Evol. Equ. Control Theory **5** (2016), no. 1, 147–184.
217. J. Merger, A. Borzí, and R. Herzog, *Optimal control of a system of reaction-diffusion equations modeling the wine fermentation process*, Optimal Control Appl. Methods **38** (2017), no. 1, 112–132.
218. C. Meyer and P. Philip, *Optimizing the temperature profile during sublimation growth of SiC single crystals: control of heating power, frequency, and coil position*, Crystal Growth & Design **5** (2005), no. 3, 1145–1156.
219. C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim. **43** (2004), no. 3, 970–985.
220. C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl. **33** (2006), no. 2-3, 209–228.
221. N. G. Meyers and J. Serrin, $H = W$, Proc. Nat. Acad. Sci. U.S.A. **51** (1964), 1055–1056.
222. I. Neitzel, J. Pfefferer, and A. Rösch, *Finite element discretization of state-constrained elliptic optimal control problems with semilinear state equation*, SIAM J. Control Optim. **53** (2015), no. 2, 874–904.
223. I. Neitzel and F. Tröltzsch, *On convergence of regularization methods for nonlinear parabolic optimal control problems with control and state constraints*, Control Cybernet. **37** (2008), no. 4, 1013–1043.

224. I. Neitzel and B. Vexler, *A priori error estimates for space-time finite element discretization of semilinear parabolic optimal control problems*, Numer. Math. **120** (2012), no. 2, 345–386.
225. N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends® in Optimization **1** (2014), no. 3, 127–239.
226. J. W. Pearson, M. Stoll, and A. J. Wathen, *Preconditioners for state-constrained optimal control problems with Moreau-Yosida penalty function*, Numer. Linear Algebra Appl. **21** (2014), no. 1, 81–97.
227. K. Pearson F.R.S., *LIII. On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2** (1901), no. 11, 559–572.
228. K. Pieper, *Finite element discretization and efficient numerical solution of elliptic and parabolic sparse control problems*, Dissertation, Technische Universität München, 2015.
229. K. Pieper and B. Vexler, *A priori error analysis for discretization of sparse elliptic optimal control problems in measure space*, SIAM J. Control Optim. **51** (2013), no. 4, 2788–2808.
230. P. Portal and Ž. Štrkalj, *Pseudodifferential operators on Bochner spaces and an application*, Math. Z. **253** (2006), no. 4, 805–819.
231. J. Prüss, *Maximal regularity for evolution equations in L_p-spaces*, Conf. Semin. Mat. Univ. Bari (2002), no. 285, 1–39 (2003).
232. E. Qian, M. Grepl, K. Veroy, and K. Willcox, *A certified trust region reduced basis approach to PDE-constrained optimization*, SIAM J. Sci. Comput. **39** (2017), no. 5, S434–S460.
233. A. Quarteroni, A. Manzoni, and F. Negri, *Reduced basis methods for partial differential equations*, Unitext, vol. 92, Springer, Cham, 2016, An introduction, La Matematica per il 3+2.
234. S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res. **5** (1980), no. 1, 43–62.
235. L. G. Rogers, *Degree-independent Sobolev extension on locally uniform domains*, J. Funct. Anal. **235** (2006), no. 2, 619–665.
236. S. Rogg, S. Trenz, and S. Volkwein, *Trust-Region POD using A-Posteriori Error Estimation for Semilinear Parabolic Optimal Control Problems*, Technical Report, Konstanzer Schriften in Mathematik, 2017.
237. A. Rösch, *Error estimates for parabolic optimal control problems with control constraints*, Z. Anal. Anwendungen **23** (2004), no. 2, 353–376.
238. ———, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw. **21** (2006), no. 1, 121–134.
239. A. Rösch and F. Tröltzsch, *On regularity of solutions and Lagrange multipliers of optimal control problems for semilinear equations with mixed pointwise control-state constraints*, SIAM J. Control Optim. **46** (2007), no. 3, 1098–1115.
240. T. Roubicek, *Nonlinear partial differential equations with applications*, International Series of Numerical Mathematics, vol. 153, Birkhäuser Verlag, Basel, 2005.
241. W. Rudin, *Functional analysis*, McGraw-Hill Book Co., New York-Düsseldorf-Johannesburg, 1973, McGraw-Hill Series in Higher Mathematics.
242. ———, *Reelle und komplexe Analysis*, R. Oldenbourg Verlag, Munich, 1999, Translated from the third English (1987) edition by Uwe Krieg.
243. E. W. Sachs, M. Schneider, and M. Schu, *Adaptive trust-region POD methods in PIDE-constrained optimization*, Trends in PDE constrained optimization, Internat. Ser. Numer. Math., vol. 165, Birkhäuser/Springer, Cham, 2014, pp. 327–342.
244. E. W. Sachs and M. Schu, *A priori error estimates for reduced order models in finance*, ESAIM Math. Model. Numer. Anal. **47** (2013), no. 2, 449–469.
245. E. W. Sachs and S. Volkwein, *Augmented Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates*, SIAM J. Numer. Anal. **40** (2002), no. 1, 233–253.
246. ———, *POD-Galerkin approximations in PDE-constrained optimization*, GAMM-Mitt. **33** (2010), no. 2, 194–208.
247. A. Schiela, *Barrier methods for optimal control problems with state constraints*, SIAM J. Optim. **20** (2009), no. 2, 1002–1031.
248. ———, *State constrained optimal control problems with states of low regularity*, SIAM J. Control Optim. **48** (2009), no. 4, 2407–2432.

249. A. Schindele and A. Borzì, *Proximal methods for elliptic optimal control problems with sparsity cost functional*, Applied Mathematics **7** (2016), 967–992.
250. ———, *Proximal schemes for parabolic optimal control problems with sparsity promoting cost functionals*, Internat. J. Control **90** (2017), no. 11, 2349–2367.
251. C. Schneider and G. Wachsmuth, *Regularization and discretization error estimates for optimal control of ODEs with group sparsity*, ESAIM Control Optim. Calc. Var. **24** (2018), no. 2, 811–834.
252. M. Schu, *Adaptive Trust-Region POD Methods and Their Applications in Finance*, Dissertation, Universität Trier, 2012.
253. S. Selberherr, *Analysis and simulation of semiconductor devices*, Springer-Verlag, 1984.
254. E. Shamir, *Regularization of mixed second-order elliptic problems*, Israel J. Math. **6** (1968), 150–168.
255. J. R. Singler, *New POD error expressions, error bounds, and asymptotic results for reduced order models of parabolic PDEs*, SIAM J. Numer. Anal. **52** (2014), no. 2, 852–876.
256. L. Sirovich, *Turbulence and the dynamics of coherent structures. parts I-III.*, Quart. Appl. Math. **45** (1987), no. 3, 561–590.
257. K. Smetana and M. Ohlberger, *Hierarchical model reduction of nonlinear partial differential equations based on the adaptive empirical projection method and reduced basis techniques*, ESAIM Math. Model. Numer. Anal. **51** (2017), no. 2, 641–677.
258. N. H. Son, B. T. Kien, and A. Rösch, *Second-order optimality conditions for boundary control problems with mixed pointwise constraints*, SIAM J. Optim. **26** (2016), no. 3, 1912–1943.
259. E. D. Sontag, *Mathematical control theory*, second ed., Texts in Applied Mathematics, vol. 6, Springer-Verlag, New York, 1998.
260. J. Sprekels and F. Tröltzsch, *Sparse optimal control of a phase field system with singular potentials arising in the modeling of tumor growth*, ESAIM Control Optim. Calc. Var. **27** (2021), no. suppl., Paper No. S26, 27.
261. G. Stadler, *Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices*, Comput. Optim. Appl. **44** (2009), no. 2, 159–181.
262. L.-M. Susu, *Analysis and optimal control of a damage model with penalty*, Dissertation, Technische Universität Dortmund, 2017.
263. V. Thomée, *Galerkin finite element methods for parabolic problems*, second ed., Springer Series in Computational Mathematics, vol. 25, Springer-Verlag, Berlin, 2006.
264. R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288.
265. T. Tonn, K. Urban, and S. Volkwein, *Comparison of the reduced-basis and POD a posteriori error estimators for an elliptic linear-quadratic optimal control problem*, Math. Comput. Model. Dyn. Syst. **17** (2011), no. 4, 355–369.
266. S. Trenz, *POD-Based A-posteriori Error Estimation for Control Problems Governed by Nonlinear PDEs*, Dissertation, Universität Konstanz, 2017.
267. H. Triebel, *Interpolation theory, function spaces, differential operators*, second ed., Johann Ambrosius Barth, Heidelberg, 1995.
268. F. Tröltzsch, *On the Lagrange-Newton-SQP method for the optimal control of semilinear parabolic equations*, SIAM J. Control Optim. **38** (1999), no. 1, 294–312.
269. ———, *Lipschitz stability of solutions of linear-quadratic parabolic control problems with respect to perturbations*, Dynam. Contin. Discrete Impuls. Systems **7** (2000), no. 2, 289–306.
270. F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*, Vieweg Studium, Vieweg+Teubner Verlag, 2010.
271. F. Tröltzsch and S. Volkwein, *The SQP method for control constrained optimal control of the Burgers equation*, ESAIM Control Optim. Calc. Var. **6** (2001), 649–674.
272. ———, *POD a-posteriori error estimates for linear-quadratic optimal control problems*, Comput. Optim. Appl. **44** (2009), no. 1, 83–115.
273. F. Tröltzsch and D. Wachsmuth, *Second-order sufficient optimality conditions for the optimal control of Navier-Stokes equations*, ESAIM Control Optim. Calc. Var. **12** (2006), no. 1, 93–119.
274. M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim. **13** (2002), no. 3, 805–842 (2003).

275. ———, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, MOS-SIAM Series on Optimization, vol. 11, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011.
276. S. Ulbrich and J. C. Ziems, *Adaptive multilevel trust-region methods for time-dependent PDE-constrained optimization*, Port. Math. **74** (2017), no. 1, 37–67.
277. S. Ullmann, M. Rotkvic, and J. Lang, *POD-Galerkin reduced-order modeling with adaptive finite element snapshots*, J. Comput. Phys. **325** (2016), 244–258.
278. A. Unger and F. Tröltzsch, *Fast solution of optimal control problems in the selective cooling of steel*, ZAMM Z. Angew. Math. Mech. **81** (2001), no. 7, 447–456.
279. S. Volkwein, *Mesh-independence for an augmented Lagrangian-SQP method in Hilbert spaces*, SIAM J. Control Optim. **38** (2000), no. 3, 767–785.
280. ———, *Mesh-independence of Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates*, Optim. Methods Softw. **17** (2002), no. 1, 77–111.
281. ———, *Lagrange-SQP techniques for the control constrained optimal boundary control for the Burgers equation*, Comput. Optim. Appl. **26** (2003), no. 3, 253–284.
282. ———, *Optimality system POD and a-posteriori error analysis for linear-quadratic problems*, Control Cybernet. **40** (2011), no. 4, 1109–1124.
283. ———, *Proper orthogonal decomposition: Theory and reduced-order modelling*, Lecture Notes, University of Konstanz (2012).
284. N. von Daniels, M. Hinze, and M. Vierling, *Crank-Nicolson time stepping and variational discretization of control-constrained parabolic optimal control problems*, SIAM J. Control Optim. **53** (2015), no. 3, 1182–1198.
285. D. Wachsmuth, *Analysis of the SQP-method for optimal control problems governed by the nonstationary Navier-Stokes equations based on L^p-theory*, SIAM J. Control Optim. **46** (2007), no. 3, 1133–1153.
286. D. Wachsmuth and G. Wachsmuth, *Second-order conditions for non-uniformly convex integrands: quadratic growth in L^1*, J. Nonsmooth Anal. Optim. **3** (2022), Paper No. 8733, 36.
287. G. Wachsmuth, *A guided tour of polyhedric sets: basic properties, new results on intersections, and applications*, J. Convex Anal. **26** (2019), no. 1, 153–188.
288. ———, *No-gap second-order conditions under n-polyhedric constraints and finitely many nonlinear constraints*, J. Convex Anal. **27** (2020), no. 2, 735–753.
289. G. Wachsmuth and D. Wachsmuth, *Convergence and regularization results for optimal control problems with sparsity functional*, ESAIM Control Optim. Calc. Var. **17** (2011), no. 3, 858–886.
290. D. Werner, *Funktionalanalysis*, extended ed., Springer-Verlag, Berlin, 2000.
291. M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68** (2006), no. 1, 49–67.
292. E. Zeidler, *Nonlinear functional analysis and its applications. II/B*, Springer-Verlag, New York, 1990, Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.
293. J. C. Ziems, *Adaptive multilevel inexact SQP-methods for PDE-constrained optimization with control constraints*, SIAM J. Optim. **23** (2013), no. 2, 1257–1283.
294. J. C. Ziems and S. Ulbrich, *Adaptive multilevel inexact SQP methods for PDE-constrained optimization*, SIAM J. Optim. **21** (2011), no. 1, 1–40.