Technische Universität München
Lehrstuhl für Mathematische Optimierung

Adaptive Numerical Solution of State Constrained Optimal Control Problems

Olaf Benedix

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Peter Rentrop
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Boris Vexler
	2.	Univ.-Prof. Dr. Thomas Apel
		(Universität der Bundeswehr München)

Die Dissertation wurde am 14. 06. 2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 11. 07. 2011 angenommen.

Contents

1. Introduction

The central subject of interest of this thesis is the class of optimal control problems with partial differential equations and additional state constraints. The focus lies especially on the construction of numerical solution algorithms to find an approximate solution to such a problem, and the effectiveness of such algorithms.

The central problem class features many different ingredients, over which we give a short overview here. The general problem form considered is

$$(P)\begin{cases} \min J(q,u) & q \in Q, u \in X \\ u = S(q) & \\ G(u) \geq 0 & . \end{cases} \tag{1.1}$$

Here, u is called the *state function*, searched for in the *state space* X, and q the *control variable*, searched for in the *control space* Q. In the field of optimal control, X is usually considered a function space. The domain of the state functions might be a spatial domain $\Omega \subset \mathbb{R}^n$ ($n \in \{2,3\}$) or in the case of time-dependent problems a domain in time and space $I \times \Omega$ with a given time interval $I = (0,T)$. The operator S is called *control-to-state operator*. It represents the solution operator of a partial differential equation, which in turn is called the *state equation*. In this thesis, elliptic and parabolic state equations are considered. The problem (P) is then called elliptic, or parabolic optimal control problem (OCP), respectively. The functional $J\colon Q \times X \to \mathbb{R}$ is called the *cost functional*, and the function G is the *constraint function* for the state. With all these ingredients present, (P) is called a *state constrained optimal control problem*. Without the condition $G(u) \geq 0$ one would speak of an *unconstrained optimal control problem*, which can be regarded as the basis class of optimal control problems.

Unconstrained optimal control problems have been of interest in applied mathematics for some time now. A lot of practical problems, their origin ranging from civil engineering via optics to chemical engineering and biological applications, can be modeled as optimal control problems with partial differential equations. This is not surprising, since most technical processes allow for user input after the initial setup, and guiding the system's output to a user-determined configuration is a natural desire as well. Also understandable is the possible need for bounds on input and output variables. For most technical problems, only certain amounts of input are possible, and concerning the output, certain states might lead to catastrophic scenarios that must be avoided at all cost.

This thesis deals with state constrained problems, which can be motivated in different ways. From the viewpoint of the field of optimization, (P) can be seen as an optimization problem on $Q \times X$, with a partial differential equation as an equality constraint, and a pointwise inequality constraint. The motivation to consider this problem class becomes possibly clearer from an applicational point of view. Suppose that a scientific or technical process of interest

is described by a partial differential equation. For notational purposes the quantities which are considered influencable are gathered in the control variable q. On the other hand, the quantities that are regarded as descriptive of the process' status, are gathered in the state function u. We think of the partial differential equation in such a way that u is the solution depending on q, and thus write formally $u = S(q)$. The quest is now to find the pair (q, u) of a control q and corresponding state $u = S(q)$ that is the most favorable to the user. By means of the condition $G(u) \geq 0$ with a properly modeled function G the user can rule out some pairs completely. Amongst the remaining pairs, favorability is determined by a given functional $J(q, u)$. This functional is modeled in such a way that a more favourable pair (q, u) is mapped to a smaller value of J.

Optimal control problems with partial differential equations have been subject of investigation for some time, see [63] for an early main work considering elliptic, parabolic and hyperbolic optimal control problems. Numerical methods to solve these OCPs are usually comprised of two steps, a discretization by the finite element method, and the solution of a discrete optimal control problem. These steps are connected in an overall algorithm, so that a more or less sophisticated sustained refinement of the discretization in the former step leads to optimal solutions in the latter step which converge to the solution of the continous problem (P).
For elliptic OCPs without additional constraints solution methods have been discussed in [32, 38] and many following publications. A priori discretization error estimations have been derived for a number of settings and discretization methods. The most basic result considers a distributed linear-quadratic optimal control problem on a convex domain of computation. By using discretizations with linear finite elements described by uniform meshes with *discretization parameter* h, the order of convergence of the finite element solutions q_h to the exact one can be proven to be h^2 in the L^2-norm if either the variational discretization concept or a postprocessing step is used, see [51, 71].
Parabolic problems pose more difficulties even for proving the existence of optimal solutions, see [35, 63, 92]. Solution techniques have been developed and for the most basic case of distributed control, $Q = L^2(I \times \Omega)$, the linear-quadratic optimal control problem discretized by linear finite elements in space, uniform with discretization parameter h, and the dG(0)-method in time, uniform with discretization parameter k, the convergence order of $h^2 + k$ has been established for the controls q_{kh} to q in the L^2-norm. A proof can be found in [67], as a special case of a more general result allowing for finite elements of different order.

A neighboring problem class frequently under consideration is the class of control constrained optimal control problems. Here it is the control q that is required to fulfill the pointwise constraint, $G(q) \geq 0$, rather than the state u. The presence of this additional constraint may reduce the regularity of the optimal solution of the OCP. This, in turn, reduces the order of convergence of the numerical solution. An overview over different situations can be found in [64]. A counter-measure to speed up the performance, or even restore the full convergence order is the construction of locally refined meshes, that take the structure of the problem into account. A widely used approach is the use of adaptive methods, where the discretization error is estimated a posteriori on a coarse starting grid, and expressed in local contributions. By the principle of equilibrating these error contributions, a local refinement algorithm is set up. One example, where the a posteriori estimation assesses the error in the natural norms of the involved spaces, can be found in, e.g., [46, 62]. A different approach is called *goal oriented*

adaptivity, here the error in terms of a functional of interest, for example the cost functional, is estimated, see, e.g., [94].

In the problem class (1.1) that is in this thesis' center of attention, major care has to be put on the state constraint. In comparison to unconstrained optimal control problems, the introduction of the state constraint has the direct effect of a reduced regularity of the optimal solution, see, e.g., [17]. This has further consequences for the construction of solution algorithms for (1.1).

Consider first the solution of one discretized optimal control problem only. A direct approach, yielding exploitable optimality conditions by incorporating the state constraint by the Lagrange formalism, shows that the Lagrange multiplier, denoted by μ_h, is in general a regular Borel measure. This means that a direct numerical treatment of this problem needs to face the handling of Borel measures, and a simple transfer of the methods for unconstrained optimal control problems is not possible. The method of choice to solve the discretized OCPs will be a primal-dual active set method, introduced in [14].
An alternative approach is the regularization of the problem (P) on the continous level. This means the construction of problems (P_γ), with $\gamma \in \mathbb{R}$ being the *regularization parameter*, whose solution exhibits the higher regularity, but is close to the original solution in the sense that the regularized solutions converge to the original solution with $\gamma \to \infty$. These regularized problems can subsequently be numerically solved with methods for unconstrained OCPs. This approach leaves the question of how to balance the driving of $\gamma \to \infty$ and $h \to 0$ (and possibly $k \to 0$) to achieve maximum effectivity of the method. Concrete choices of regularization are the Moreau-Yosida-regularization, see [49, 54], and barrier methods, see [85, 86]. In this thesis, the latter method is investigated.

Apart from the optimization on one discretization level, consider now the process of refinement of the discretization. A second consequence of the reduced regularity is again the reduction of the achievable order of convergence of the discretization error in terms of $h \to 0$, $k \to 0$ if uniform discretization is used, see, e.g., [26, 69]. Thus it is desirable to set up a mesh refinement strategy to improve the convergence. In this thesis, a goal-oriented a posteriori error estimator is derived by the dual weighted residual method, see [7] for an overview. For problem (1.1) the error estimator is dissected into the contributions, if they are present,

$$\eta := \eta_h + \eta_k + \eta_d + \eta_\gamma, \tag{1.2}$$

which are the spatial discretization error, temporal discretization error, control discretization error, and regularization error. Each of these contributions is then further split up into cellweise or intervall-wise contributions, where applicable. The so obtained error indicators are used in the execution of the mesh refinement strategy. In the construction of a comprehensive solution algorithm for state constrained optimal control problems one must be aware that regularities and convergence orders can also be reduced due to other phenomena, for example boundary properties like reentrant corners or edges, or nonsmooth boundaries, or singularities in the data, like jumping coefficients. If the spatial location of these singularities is known, they can be treated with a priori knowledge, e.g. mesh grading techniques, see, e.g., [89] and the references therein. The spatial location of the singularities due to the state constraints however is a priori unknown. Therefore it is advantageous to use a posteriori techniques.

An additional goal of investigation in this thesis is the treatment of a large-scale real-world problem, which originates from the field of civil engineering. For structures made of concrete,

the time span of the first few days after the concrete pour is called young concrete phase. During that phase the concrete solidifies, and its thermomechanical properties develop. Amongst others, tensile strength is built up with time. The exact progression depends on the temperature field inside the structure, which is changing due to chemically produced heat and heat outflow. These phenomena in turn depend on initial and boundary conditions and material parameters. Counteracting to this tensile strength, tensile stresses are building up. Should at any point the tensile stress exceed the strength, the concrete will crack, which is seen as an event that is to be avoided. The goal is to choose boundary conditions and material in such a way that no cracks occur.
This practical problem can be modeled as an optimal control problem. The prohibition of cracks translates as a constraint on the state variable. The state equation is a parabolic equation coupled with one ordinary differential equation in every spatial point. So the problem does not fit in the category (1.1) strictly speaking, but on the other hand the additional ordinary differential equation can be treated by standard methods. Since the concrete structures are generally large-scale and nonconvex, a goal-oriented discretization is required. Thus the numerical treatment of this problem by the methods developed in this thesis assures the necessary effectivity.

Summarizing, the goal of the thesis is the efficient numerical solution of optimal control problems with elliptic or parabolic state equation and pointwise state constraints, with all aspects mentioned above to be taken into consideration. Preferably large problem classes are set up. The two numerical solution strategies, amounting to the primal-dual active set strategy and to an interior-point method, are developed for a given discretization. For both these strategies a posteriori error estimators are derived, and used to guide an adaptive mesh refinement algorithm.

The thesis is divided into the following chapters:
In Chapter 2 the necessary basic notation is introduced, as well as the basic form of elliptic and parabolic state constrained optimal control problems. This includes the general formulation of elliptic and parabolic state equations, state constraints and cost functionals. Assumptions that allow for the proof of existence and uniqueness of optimal solutions are given. Common concepts of numerical solution strategies are introduced: for unconstrained optimal control problems the process of deriving optimality conditions, discretization and optimization methods are described, serving as a starting point for the development of the equivalent steps in the treatment of state constrained problems. An overview of methods to include the state constraints is then given, and the strategy for a posteriori error estimation and the adaptive mesh-refinement algorithm is laid out.
In Chapter 3, these general concepts are concretized for elliptic optimal control problems with state constraints. The state equation is formulated in a precise setting. For a large problem class the unique existence of optimal solutions is proven. Necessary optimality conditions of first order are given. The finite element discretization is described in detail, a continous Galerkin method of order s is used to discretize the state space. For this discretized problem the optimization is carried out with the primal-dual active set method. The a posteriori error estimator is derived, consisting of the spatial part η_h and the control discretization part η_d. With subsequent splitting into the respective cellwise error indicators, the adaptive mesh refinement algorithm is set up. Finally the interior point method as alternate optimization method is introduced briefly.

Chapter 4 deals with parabolic state constrained optimal control problems. Giving a precise setting again allows to prove existence of an optimal solution and necessary optimality conditions for it. The regularization by barrier functions is introduced, and optimality conditions of the regularized problem are derived. Further, the finite element discretization in space and time is carried out, using a continous Galerkin method of order s in space as before, and the discontinous Galerkin method in time. The interior point optimization algorithm is applied. The a posteriori error estimator derived distinguishes between the influences of regularization, η_γ, and temporal, η_k, spatial, η_h, and control discretization, η_d. Implementational aspects are considered in Chapter 5. A combination of all ingredients to a comprehensive solution algorithm is given, considerations on the choice of parameters are made. Improvements of subalgorithms in special situations are discussed.
Numerical experiments to validate the theoretical results and the advised optimization algorithms of this thesis are carried out in Chapter 6. Combinations of elliptic and parabolic, linear and nonlinear test problems with different structures of the active set are considered. The efficiency of the error estimator itself, and its parts $\eta_\gamma, \eta_k, \eta_h, \eta_d$ is evaluated. Also the adaptive refinement strategy driven by the local error indicators is compared to the uniform refinement strategy by the respective convergence rates.
Chapter 7 contains the application of the methods discussed in this thesis to the real-world application of optimal control of young concrete thermo-mechanical properties. The model functions for the different physical phenomena are introduced and the possibilities how to assemble an optimal control problem are shown. The unique solvability of the state equation and the existence of an optimal control is proven. Finally, several numerical examples are considered and solved by the methods developed in this thesis.

2. Basic Concepts in Optimal Control

2.1. Problem setting

2.1.1. Notation

In the following, we will introduce the basic notation used throughout the thesis, and describe the considered problem class in a rather abstract formulation. Let $\Omega \subset \mathbb{R}^n, n \in \{2,3\}$ denote a spatial domain with Lipschitz boundary $\partial\Omega =: \Gamma$. For a point $x \in \Gamma$ let $n(x)$ denote the outer unit normal vector of Ω, if it exists. By $L^p(\Omega)$, $W^{m,p}(\Omega)$, and $H^m(\Omega)$ with $1 \leq p \leq \infty, m \in \mathbb{R}$ we denote the usual Lebesgue and Sobolev spaces. The space of continous functions on $\bar{\Omega}$ with continous derivatives up to m-th order, $m = 0, 1, \ldots$ is denoted by $C^m(\bar{\Omega})$, and the dual space to $C^0(\bar{\Omega}) = C(\bar{\Omega})$ is identified with the space of regular Borel measures $\mathcal{M}(\Omega)$.
The considered time interval is denoted by $I = (0,T) \subset \mathbb{R}$. For any Banach space Z and time interval $[t_1, t_2]$ the Lebesgue and Sobolev spaces of time dependent, Z-valued functions are denoted by $L^p([t_1,t_2],Z), W^{m,p}([t_1,t_2],Z), H^m([t_1,t_2],Z)$. For a proper definition of these spaces including Bochner integrals, see, e. g., [99]. If $[t_1,t_2] = I$, the interval can be omitted in the previous notation, and we just write $L^p(Z), W^{m,p}(Z), H^m(Z)$. Again we identify $C(\bar{I} \times \bar{\Omega})^* = \mathcal{M}(I \times \Omega)$. The following convention concerning the evaluation of space and time dependent functions is used: a function $v \in C(I \times \Omega)$ can be interpreted as an abstract function $v\colon [0,T] \to C(\Omega)$, so that it is possible to write both $v(t,x)$ (a number) and $v(t)$ (a $C(\Omega)$-function) without being ambiguous.
All function spaces can be endowed with a subscript to prescribe a homogenous Dirichlet boundary condition; the subscript 0 indicates the boundary condition is prescribed on the whole boundary. If the condition is to be applied to a part $\Gamma_1 \subset \Gamma$ only, the subscript Γ_1 is used.

Let V, H, R be Hilbert spaces equipped with scalar products $(\cdot,\cdot)_V$, $(\cdot,\cdot)_H$, and $(\cdot,\cdot)_R$, respectively, such that V is continously and densely embedded into H. With the dual spaces V^* and H^* the Gelfand triple $V \hookrightarrow H \hookrightarrow V^*$ is formed, assuming an identification of H with H^* is possible. This makes it possible to represent functionals in V^* with their effect in the duality pairing $\langle\cdot,\cdot\rangle_{V^*,V}$ by the effect in inner products $(\cdot,\cdot)_H$. Abbreviations for the most commonly used scalar and duality products are

$$
\begin{aligned}
&(.,.) := (.,.)_H, && (v,w)_I := \int_I (v(t), w(t))_H \,\mathrm{d}t, \quad v,w \in L^2(I,H),\\
&(\cdot,\cdot)_\Omega := (\cdot,\cdot)_{L^2(\Omega)}, && (\cdot,\cdot)_{I\times\Omega} := (\cdot,\cdot)_{L^2(I\times\Omega)},\\
&(\cdot,\cdot)_\Gamma := (\cdot,\cdot)_{L^2(\Gamma)}, && (\cdot,\cdot)_{I\times\Gamma} := (\cdot,\cdot)_{L^2(I\times\Gamma)},\\
&\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_{\mathcal{M}(\Omega),C(\bar{\Omega})}, && \langle\cdot,\cdot\rangle_I := \langle\cdot,\cdot\rangle_{\mathcal{M}(I\times\Omega),C(\bar{I}\times\bar{\Omega})}.
\end{aligned}
$$

Throughout this thesis, parlance and notation will be differentiated depending on the type of state equation S represents. The first case considers stationary problems, where the state equation is an elliptic partial differential equation. After the general introduction in this section, Chapter 3 is devoted to the study of elliptic optimal control problems with state constraints. Here, the domain of the state functions $u \in X$ is $\bar{\Omega}$. The specific choice of the state space X is done in Chapter 3, as it depends on the properties of (1.1). However, one basic regularity requirement that needs to be fulfilled simply because of the presence of the pointwise state constraints, is the continuity of the states on the whole domain. This property is used as a starting point for the derivation of optimality conditions, see [92, section 6.1], as it assures that the cone of non-negative functions has interior points. Thus we require

$$X \subset C(\bar{\Omega}) \qquad \text{for elliptic OCPs.} \tag{2.1}$$

As a second case, time-dependent problems are considered, with the state equation being of parabolic type. The detailed treatment is done in Chapter 4. As the state is now time and space dependent, the domain of the state functions $u \in X$ is $\bar{I} \times \bar{\Omega}$. Including the continuity of the state function on the computational space-time domain, the state space has to be chosen according to

$$X \subset C(\bar{I} \times \bar{\Omega}) \qquad \text{for parabolic OCPs.} \tag{2.2}$$

For the choice of the control space no additional regularity requirements are made. The domain of the control functions $q \in Q$ is a subset of $\bar{\Omega}$ for elliptic optimal control problems, and a subset of $\bar{I} \times \bar{\Omega}$ for parabolic optimal control problems. The actual choice depends on the problem structure, specifically the way in which q enters the state equation. In the case of parameter control, it is also possible to choose $Q \subset \mathbb{R}^k$ or $Q \subset L^2(\mathbb{R}^k)$ as a subspace, respectively.

2.1.2. State equation

The state equation is frequently introduced in different formulations. The classical form employs a differential operator that will be denoted by $\mathcal{A}$ here. We will first introduce the state equation for the elliptic case. Let a differential operator of second order

$$\mathcal{A}\colon Q \times V \to V^* \tag{2.3}$$

and a right hand side $f \in V^*$ be given. They form the state equation in weak formulation:

$$\mathcal{A}(q, u) = f. \tag{2.4}$$

Remark 2.1. Thinking of classical situations in PDE analysis, the natural spaces employed in the formulation in general contain discontinous functions. For example, the classical Poisson problem is formulated in the space $H_0^1(\Omega)$. We can not choose this space as state space, since this choice would not fulfill (2.1). Instead, the classical formulation with the natural space denoted by V is set up, as done in (2.3). Then X is chosen as a subspace of V in an way that secures continuity of the states.

The weak formulation (2.4), being an equation in V^*, can be concretized by testing with all functions $\varphi \in V$. As mentioned before, the right hand side is hereby represented as a scalar product in H. Introducing the form

$$a\colon Q \times V \times V \to \mathbb{R}, \quad a(q,u)(\varphi) := \langle \mathcal{A}(q,u), \varphi \rangle_{V^*,V} \tag{2.5}$$

the weak formulation of the state equation is given as

$$a(q,u)(\varphi) = (f,\varphi) \quad \forall \varphi \in V. \tag{2.6}$$

Remark 2.2. In the notation $a(\cdot)(\cdot)$ the two pairs of parentheses are meant to indicate any dependence of the function a on the argument(s) in the first parenthesis, but a linear dependence on the argument(s) in the second one.

Two common examples for the state equation and the choices of the involved spaces are considered next:

Example 2.1. In distributed control, q may directly enter the right hand side of the partial differential equation. As linear state equation we might consider

$$\begin{aligned} -\Delta u(x) &= q(x) \quad \forall x \in \Omega \\ u(x) &= 0 \qquad \forall x \in \Gamma \end{aligned} \tag{2.7}$$

so that the choice $Q = L^2(\Omega)$, $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, $a(q,u)(\varphi) = (\nabla u, \nabla \varphi)_\Omega - (q,\varphi)_\Omega$, and $f = 0$ fits into the framework.

Example 2.2. An example of a boundary control problem uses the control q entering the right hand side of a Neumann boundary condition. We then speak of *Neumann control*. As linear state equation we might consider

$$\begin{aligned} -\Delta u(x) + u(x) &= f(x) \quad \forall x \in \Omega \\ \partial_n u(x) &= q(x) \quad \forall x \in \Gamma \end{aligned} \tag{2.8}$$

with a given function $f \in L^2(\Omega)$ so that the choice $Q = L^2(\Gamma)$, $V = H^1(\Omega)$, $H = L^2(\Omega)$, and $a(q,u)(\varphi) = (\nabla u, \nabla \varphi)_\Omega + (u,\varphi)_\Omega - (q,\varphi)_\Gamma$ fits into the framework.

In order to choose the state space X in accordance with (2.1) and Remark 2.1 we make the assumption that the actual regularity of the state is better than $u \in V$. This assumption is justified in many practical situations, and demonstrated in the previous two examples. In Example 2.1 $u \in H^2(\Omega)$, and in Example 2.2 $u \in H^{\frac{3}{2}}(\Omega)$ can be shown in the case of convex polyhedral domains.

Assumption 2.1. *For every $q \in Q$, every state $u \in V$ solving the state equation* (2.6) *has actually the regularity*

$$u \in W^{1,p}(\Omega) \qquad \text{with some } p > n. \tag{2.9}$$

This assumption assures the desired regularity $u \in C(\bar{\Omega})$ by utilizing the limiting case in the well known embedding theorem

$$W^{1,p}(\Omega) \hookrightarrow C(\bar{\Omega}) \quad \forall p > n.$$

The final choice for the state space is thus

$$X := V \cap W^{1,p}(\Omega) \qquad \text{state space for elliptic OCPs.} \tag{2.10}$$

This choice yields a consequence for the differential operator $\mathcal{A}$: in (2.3), we had introduced the operator as $\mathcal{A}\colon Q \times V$, but Assumption 2.1 implies that a definition of $\mathcal{A}\colon Q \times X$ would have sufficed. To find out the consequences of a restriction of the domain of $\mathcal{A}$, assume functions $q \in Q, u \in W^{1,p}(\Omega)$ given and consider e.g. from Example 2.1 the term

$$a(q,u)(v) = \int_\Omega (\nabla u \cdot \nabla v - qv)\,.$$

This term is well-defined for any function $v \in W^{1,p'}(\Omega)$, such that in this case $\mathcal{A}(q,u) \in (W^{1,p'}(\Omega))^*$ can be allowed. This motivates the following assumption for the general case:

Assumption 2.2. *The restriction of $\mathcal{A}$ to states that actually possess the regularity $u \in X$ restrains the image of $\mathcal{A}$ according to*

$$\mathcal{A}\colon Q \times X \to Z^* := (W^{1,p'}(\Omega))^*. \tag{2.11}$$

Accordingly we assume $f \in Z^*$.

The space $Z := W^{1,p'}(\Omega)$ is called *dual space*. With the according redefinition of a,

$$a\colon Q \times X \times Z \to \mathbb{R}, \quad a(q,u)(\varphi) := \langle \mathcal{A}(q,u), \varphi \rangle_{Z^*,Z}, \tag{2.12}$$

the formulation of the *elliptic state constrained optimal control problem* reads

$$(P_{ell}) \begin{cases} \min J(q,u) & q \in Q, u \in X \\ a(q,u)(\varphi) = (f,\varphi) & \forall \varphi \in Z \\ G(x,u(x)) \geq 0 & \forall x \in \Omega \end{cases}. \tag{2.13}$$

Next, parabolic state equations are considered. The usual way to formulate a parabolic state equation in weak form is

$$\begin{aligned} \partial_t u(t) + \mathcal{A}(q(t),u(t)) &= f(t) \quad \forall t \in I, \\ u(0) &= u_0(q(0)). \end{aligned} \tag{2.14}$$

To incorporate the potential time dependency of the control variable, the construction of Q is done in the following way: The spatial layers of the controls $q(t)$ are elements of the space R. The control space Q can then be chosen as a subspace of $L^2(I,R)$. Thus, in (2.14) the differential operator of second order is defined as $\mathcal{A}\colon R \times V \to V^*$ first, and u_0 is a given operator that allows the control to enter the initial condition. The states u are also time dependent, so as a basis for the definition of X consider

$$W(I,V) := \{v \in L^2(I,V)\colon \partial_t v \in L^2(I,V^*)\}. \tag{2.15}$$

Similar to the elliptic case, an assumption on the regularity of the state and the range of the differential operator need to be made:

Assumption 2.3. *For every $q \in Q$, every state $u \in W(I,V)$ solving the state equation (2.14) has actually the regularity*

$$u \in L^s(I, W^{1,p}(\Omega)) \cap W^{1,s}(I, (W^{1,p'}(\Omega))^*) \tag{2.16}$$

for a number $p > n$ like above, and $s > \frac{2p}{p-n} > 2$.

This assures the continuity $u \in C(\bar{I} \times \bar{\Omega})$, as the embedding

$$L^s(I, W^{1,p}(\Omega)) \cap W^{1,s}(I, (W^{1,p'}(\Omega))^*) \hookrightarrow C(\bar{I} \times \bar{\Omega})$$

holds for coefficients satisfying $s > \frac{2p}{p-n}$, proven in [2, 91]. Thus the state space for parabolic OCPs is chosen as

$$X = W(I,V) \cap L^s(I, W^{1,p}(\Omega)) \cap W^{1,s}(I, (W^{1,p'}(\Omega))^*). \tag{2.17}$$

Similar to the elliptic case, the necessary regularity of functions v for the term

$$\int_I (\partial_t u\, v + \nabla u \cdot \nabla v)\, \mathrm{d}t$$

to be well-defined is used to motivate

Assumption 2.4. *The restriction of $\mathcal{A}$ to states that actually possess the regularity $u(t) \in W^{1,p}(\Omega)$ restrains the image of $\mathcal{A}$ according to*

$$\mathcal{A}\colon R \times (V \cap W^{1,p}(\Omega)) \to (W^{1,p'}(\Omega))^*. \tag{2.18}$$

Accordingly we assume $f(t) \in (W^{1,p'}(\Omega))^*$.

The necessary temporal regularity is brought in via the weak formulation: Defining the space

$$Z := L^{s'}(I, W^{1,p'}(\Omega)) \cap W^{1,s'}(I, (W^{1,p})^*), \tag{2.19}$$

the forms

$$\bar{a}\colon R \times (V \cap W^{1,p}(\Omega)) \times W^{1,p'}(\Omega) \to \mathbb{R}, \quad \bar{a}(q(t), u(t))(\varphi) = \langle \mathcal{A}(q(t), u(t)), \varphi \rangle_{W^{1,p}(\Omega), W^{1,p'}(\Omega)}$$

and

$$a\colon Q \times X \times Z \to \mathbb{R}, \quad a(q,u)(\varphi) = \int_I \bar{a}(q(t), u(t))(\varphi(t))\, \mathrm{d}t$$

are well-defined. Allowing for right hand sides $f \in Z^*$ and initial conditions $u_0\colon R \to V \cap W^{1,p}(\Omega)$ allows to set up the weak formulation of the state equation in the following form

$$(\partial_t u, \varphi)_I + a(q,u)(\varphi) + (u(0), \varphi(0)) = (f, \varphi)_I + (u_0(q), \varphi(0)) \quad \forall \varphi \in Z. \tag{2.20}$$

All together, the formulation of the *parabolic state constrained optimal control problem* reads

$$(P_{par}) \begin{cases} \min J(q,u) & q \in Q, u \in X \\ (\partial_t u, \varphi)_I + a(q,u)(\varphi) + (u(0), \varphi(0)) = (f, \varphi)_I + (u_0(q), \varphi(0)) & \forall \varphi \in Z \\ G(t, x, u(t,x)) \geq 0 & \forall t \in [0,T], x \in \bar{\Omega}. \end{cases} \tag{2.21}$$

Example 2.3. For the choice $R = L^2(L^2(\Omega))$, $Q = R$, $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ a state equation representing *distributed control* is

$$\begin{cases} \partial_t u(t,x) - \Delta u(t,x) = q(t,x) & \text{in } I \times \Omega, \\ u(t,x) = 0 & \text{on } \Gamma \times [0,T], \\ u(0,x) = 0 & \text{on } \Omega \end{cases} \tag{2.22}$$

so that $a(q,u)(\varphi) = \int_I \int_\Omega (\nabla u(t,x) \cdot \nabla \varphi(t,x) - q(t,x)) \,\mathrm{d}x \,\mathrm{d}t$. We can see that Assumption 2.4 is justified, since for any $u \in X$ the number $a(q,u)(\varphi)$ is well-defined for any $\varphi \in Z$, because the lower regularity of φ is countered by the higher regularity in the definition of X, now also in time.

The last point of this section covers both elliptic and parabolic problems again. In order to formulate an optimal control problem of form (1.1) the state equation needs to possess a unique solution u for every control q. The abstract forms (2.6) and (2.20) do not imply unique solvability. Furthermore, the approach that will be chosen for the evaluation of optimality conditions requires S to be twice differentiable.

Assumption 2.5. *The control-to-state operator*

$$S \colon Q \to X, \qquad S(q) = u$$

is well-defined and twice continously differentiable.

This assumption can be proven for large classes of state equations, see the specific chapters for elliptic and parabolic problems.

Remark 2.3. The theory of optimal control of hyperbolic equations differs substantially from the one for elliptic or parabolic OCPs and is also not as advanced yet. In this thesis, hyperbolic equations will not be considered. Basic theory can be found for example in [63, 72, 73] or the survey article [102] and the references therein. For numerical treatment, see [37, 40, 56], amongst others.

2.1.3. State constraints

Throughout this thesis, the state constraint is given in abstract form by the function G, whose domain depends on the type of the optimal control problem as follows

$$G \colon \bar{\Omega} \times \mathbb{R} \to \mathbb{R} \qquad \text{or} \qquad G \colon \bar{I} \times \bar{\Omega} \times \mathbb{R} \to \mathbb{R}$$

and is often represented by the pointwise formulation

$$G(x,u(x)) \geq 0 \quad \forall x \in \bar{\Omega} \qquad \text{or} \qquad G(t,x,u(t,x)) \geq 0 \quad \forall t \in \bar{I}, x \in \bar{\Omega}. \tag{2.23}$$

An alternative formulation makes use of the *admissible set*, that is

$$X_{ad} = \{u \in X \colon G(x,u(x)) \geq 0 \, \forall x \in \bar{\Omega}\} \text{ or } X_{ad} = \{u \in X \colon G(t,x,u(t,x)) \geq 0 \, \forall t \in \bar{I}, x \in \bar{\Omega}\}. \tag{2.24}$$

The state constraint then simply reads

$$u \in X_{ad}. \tag{2.25}$$

States $u \in X_{ad}$ are called *admissible*. The notion of admissibility of controls is not used in this thesis, as it refers to constraints of the control variable by an explicitly given set $Q_{ad} \subset Q$. In order to execute the error estimation process later, we make the following assumption:

Assumption 2.6. *The constraint function G is twice differentiable in the last variable, the control u. Furthermore, G is continous in the remaining variables.*

This assures that the concatenation $G(\cdot, u(\cdot))$ is a continous function, $G(\cdot, u(\cdot)) \in C(\bar{\Omega})$ or $G(\cdot, u(\cdot)) \in C(\bar{I} \times \bar{\Omega})$, respectively. This observation retrospectively justifies the formulation $G(u) \geq 0$ in (1.1), as we can now identify the term $G(u)$ with a continous function from $C(\bar{\Omega})$ or $C(\bar{I} \times \bar{\Omega})$. The Assumption 2.6 is also useful since it guarantees the closedness of X_{ad} in X, which is proven next:

Lemma 2.7. *Let G be continous. Then the set X_{ad} is closed in X.*

Proof. We give the proof only for the elliptic case, the parabolic case can be proved in an analogous way. Consider a sequence $u_n \to u$ in X. Since $X \hookrightarrow C(\bar{\Omega})$ there also holds $u_n \to u$ in $C(\bar{\Omega})$, so that there exists a constant $M > 0$ such that

$$\|u\|_{C(\bar{\Omega})} < M, \qquad \|u_n\|_{C(\bar{\Omega})} < M \quad \forall n \in \mathbb{N}.$$

Since now $G\colon \bar{\Omega} \times [-M, M] \to \mathbb{R}$ is uniformly continous in the second variable there holds $\|G(\cdot, u_n(\cdot)) - G(\cdot, u(\cdot))\|_{C(\bar{\Omega})} \to 0$ or

$$G(\cdot, u_n(\cdot)) \to G(\cdot, u(\cdot)) \qquad \text{in } C(\bar{\Omega}).$$

Since $G(\cdot, u_n(\cdot)) \geq 0$ on $\bar{\Omega}$ it follows that $G(\cdot, u(\cdot)) \geq 0$ on $\bar{\Omega}$ giving the claim of the lemma. □

Furthermore we require

Assumption 2.8. *The admissible set X_{ad} is convex.*

In practical applications this assumption is often justified, as it means that convex combinations of admissible states are admissible themselves.
The next lemma makes use of the fact that functions with $G(\cdot, u(\cdot)) = 0$ for some points are still included in X_{ad}, a formulation of $G(u) > 0$ in the definition of X_{ad} above could lead to a non-closed set.

Frequently state constraints are given explicitly, without the use of the function G. We will give a few examples of common forms of state constraints next, but in the remainder of the thesis the abstract notation involving G will be kept.

- The one-sided pointwise state constraint

$$u(x) \leq u_b(x) \quad \forall x \in \bar{\Omega} \qquad \text{or} \qquad u_a(x) \leq u(x) \quad \forall x \in \bar{\Omega}$$

  with given functions $u_a, u_b \colon \Omega \to \mathbb{R}$ in the elliptic case. Equivalently

$$u(x,t) \leq u_b(x,t) \quad \forall (x,t) \in \Omega \times [0,T] \qquad \text{or} \qquad u_a(x,t) \leq u(x,t) \quad \forall (x,t) \in \Omega \times [0,T]$$

  in the parabolic case with given functions $u_a, u_b \colon \Omega \times [0,T] \to \mathbb{R}$.

- Generalizing the abstract formulation (2.23), more than one constraint can be incorporated by using a function $G \colon \bar{\Omega} \times \mathbb{R} \to \mathbb{R}^k$. As the additional constraints can be treated analog to the first distributed constraint, we will for the sake of simpler notation restrict ourselves to one constraint.

- two-sided constraints

$$u_a(x) \leq u(x) \leq u_b(x) \text{ or } u_a(x,t) \leq u(x,t) \leq u_b(x,t)$$

  as a special case of the previous one, that frequently occurs in practical applications.

These types of constraints fulfill the assumptions discussed above. Constraints on the state that are not considered in this thesis include

- constraints on the gradient, like $\|\nabla u\| \leq C_G$ with $C_G > 0$ a given number, or constraints that involve other differential operators. E.g. for gradient state constraints in elliptic optimal control problems see [25, 100].

- state constraints that are posed only on a subset of the domain, e.g.

$$u(x) \leq u_b(x) \quad \forall x \in \Omega_1 \subset \Omega,$$

  where Ω_1 has a positive distance to the boundary of the domain, $\operatorname{dist}(\Omega_1, \partial\Omega) \geq d > 0$. This makes it possible to prove higher regularity of the state near the boundary, which can be utilized in the error estimation process, see [57, 58].

- constraints in single points

$$G(x_i, u(x_i)) \geq 0 \text{ or } G(t_i, x_i, u(t_i, x_i)) \geq 0 \qquad \forall i = 1, \ldots, l$$

  for some given points $x_i \in \bar{\Omega}$ and possibly $t_i \in \bar{I}$. See, e.g., [19, 68].

- constraints on the control variable, or mixed constraints, like for distributed control

$$q(x) + u(x) \leq u_b(x).$$

2.1.4. Cost functional

In order to formulate an optimal control problem we assume a cost functional $J\colon Q \times X \to \mathbb{R}$ to be given. (For practical purposes it suffices to have $J\colon Q \times X_{ad} \to \mathbb{R}$ given.) While for the purpose of this thesis J will be left in this abstract form, we remark that in many practical applications, and thus in many scientific articles, J admits a special structure,

$$J(q,u) = J_1(q) + J_2(u),$$

it is assumed to be the sum of *control costs* $J_1(q)$ und *state costs* $J_2(u)$. A common representative of this structure is the tracking type functional: Given a function $u_d \in X$ the OCP can be interpreted as the task of guiding the state u as close to the *desired state* u_d as possible. So the aim is to find a control q such that the distance $\|u - u_d\|^2$ is as small as possible. The utilized norm is here often the L^2-norm.
In order to secure the coercivity of j, often a regularization term $\|q\|_Q^2$ is added, weighed by a typically small factor $\alpha > 0$. So the most commonly used cost functional takes the form

$$J(q,u) = \frac{1}{2}\|u - u_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|q\|_Q^2 \quad \text{or} \quad J(q,u) = \frac{1}{2}\|u - u_d\|^2_{L^2(\Omega\times I)} + \frac{\alpha}{2}\|q\|_Q^2,$$

respectively, called *tracking type functional*.

For parabolic problems another practically interesting functional is *end time control*: u is controlled to reach a desired profile in the end time point, so that we choose

$$J(q,u) = \|u(T) - u_d\|_H^2 + \frac{\alpha}{2}\|q\|_Q^2$$

with a given $u_d \in H$.

The common approaches in the numerical solution process build on optimality conditions that require differentiability of the cost functional. Throughout the thesis it will be thus assumed, that J is Frechet differentiable. For the error estimation process higher differentiability is required, the necessary assumptions on the cost functional will be indicated at the appropriate places.

Further assumptions on J which assure the existence of a solution of (P) will be discussed in the following section. Let us just anticipate that a cost functional of tracking type possesses all necessary properties. This section is concluded by the an alternative description of problem (1.1), that utilizes the *reduced cost functional*: provided the unique solvability of the state equation, define

$$j\colon Q \to \mathbb{R} \qquad j(q) := J(q, S(q)). \tag{2.26}$$

Then the optimization problem (1.1) can be represented in the reduced form:

$$(P_{red}) \begin{cases} \min j(q) & q \in Q \\ S(q) \in X_{ad} & \end{cases}. \tag{2.27}$$

2.2. Existence and uniqueness of optimal solutions

In this section we will discuss conditions under which a solution of the optimal control problem (1.1) exists, and is unique. For the proof, some assumptions on the cost functional J and the control-to-state operator S are made. Due to the general formulation of (1.1), these assumptions may seem unnatural at first, however they are motivated by frequently considered concrete realizations of the general problem class.

The first question is for the continuity of the control-to-state operator. This property is desirable, as one would hesitate to call a problem with a noncontinous assignment between the control and the state of the system a "control problem". In the proof of existence, a stronger assumption on S is needed.

Assumption 2.9. *Let $q_n \rightharpoonup q$ converge weakly in Q. Then it holds that*

$$S(q_n) \rightharpoonup S(q) \qquad \text{in the sense of } X$$
$$S(q_n) \to S(q) \text{ in the sense of } L^2(\Omega) \text{ or } L^2(I \times \Omega), \text{ respectively.}$$

A strong convergence $S(q_n) \to S(q)$ in X is unrealistic for frequent state equations, but Assumption 2.9 can often be shown.

The second quantity to consider is the cost functional. We will need the following properties for our considerations, see, e.g., [23] for the notation:
A functional $f\colon Q \to \mathbb{R}$ is said to be weakly lower semicontinous, if for any sequence $(q_n) \subset Q$ holds

$$q_n \rightharpoonup q \text{ in } Q \qquad \Longrightarrow \qquad \liminf_{n\to\infty} f(q_n) \geq f(q), \tag{2.28}$$

and it is said to be coercive over Q, if

$$\exists\, \alpha > 0, \beta \in \mathbb{R}: \quad f(q) \geq \alpha \|q\|_Q + \beta \qquad \forall q \in Q. \tag{2.29}$$

If the cost functional can be dissected into control cost and state cost, we make the assumption:

Assumption 2.10. *The cost functional takes the form*

$$J = J_1(q) + J_2(u).$$

The functional J_1 is continous from Q to $\mathbb{R}$ and convex, and J_2 is continous from $L^2(\Omega)$ to $\mathbb{R}$.

In the case where J_1 is a regularization term $J_1(q) = \alpha\|q\|_Q^2,\ \alpha > 0$, see Section 2.1.4, and J_2 is bounded from below, the reduced cost functional $j = J_1 + J_2 \circ S$ is coercive.
A further assumption for the formulation of a meaningful OCP that possesses an optimal solution is the following:

Assumption 2.11. *There exists a control $q^* \in Q$ such that $S(q^*) \in X_{ad}$.*

Then we can prove the following general existence theorem:

Theorem 2.12. *Consider the abstract optimization problem in formulation* (1.1)*, with the spaces Q and X as discussed in Section 2.1. Let $S\colon Q \to X$ be properly defined and continous according to Assumption 2.9. The admissible set X_{ad} shall be closed and fulfill Assumptions 2.8 and 2.11. Let J be a functional according to Assumption 2.10 with a corresponding reduced functional j that is coercive. Then there exists a globally optimal solution $\bar{q}$ to* (1.1)*.*

Proof. Since there exists an admissible control, and j is bounded from below due to (2.29) it follows that there exists an infimum value of the cost functional

$$\bar{j} := \inf_{q\in Q\colon\, S(q)\in X_{ad}} J(q, S(q)) > -\infty. \tag{2.30}$$

Consequently there exists a sequence $q_n \in Q$ such that $S(q_n) \in X_{ad}$ and $j(q_n) \to \bar{j}$. Coercivity of j gives for some $K > 0, n_0 \in \mathbb{N}$

$$\|q_n\|_Q < K \qquad \forall n > n_0,$$

such that we can extract from q_n a weakly convergent subsequence, for simplicity here also denoted by q_n, with $q_n \rightharpoonup \bar{q}$. This control $\bar{q}$ is a candidate for the optimal solution.
Consider the sequence of associated states $u_n = S(q_n)$. Assumption 2.9 gives $u_n \rightharpoonup S(\bar{q}) =: \bar{u}$. For the next step, from [92, Theorem 2.11] it is concluded that since X_{ad} is convex and closed in X, it is also weakly sequentially closed. The definition of this property is that every weak limit of $u_n \in X_{ad}$ is itself in X_{ad}, so it is shown that $\bar{u} \in X_{ad}$.

Again due to Assumption 2.9 this gives $u_n \to \bar{u}$ in $L^2(\Omega)$ or $L^2(I \times \Omega)$. With Assumption 2.10 this yields instantly convergence of the values $J_2(u_n)$, and the weak lower semicontinuity of J_1 implied by the same assumption then gives

$$J(\bar{q}, \bar{u}) = J_1(\bar{q}) + J_2(\lim_{n\to\infty} u_n) \le \liminf_{n\to\infty} J_1(q_n) + \lim_{n\to\infty} J_2(u_n) = \lim_{n\to\infty} J(q_n, u_n) = \bar{j}.$$

□

In order to prove uniqueness, an additional assumption needs to be made, e.g. by using the property of strong convexity of the functional $j\colon Q \to \mathbb{R}$ over Q, which means that

$$j(\lambda q_1 + (1-\lambda)q_2) < \lambda j(q_1) + (1-\lambda) j(q_2) \qquad \forall\, \lambda \in (0,1)\ \forall q_1 \neq q_2 \in Q. \tag{2.31}$$

Theorem 2.13. *Consider the situation of Theorem 2.12. Let additionally j be strongly convex. Then the optimal control $\bar{q}$ is unique.*

Proof. Assume $\bar{q}_1 \neq \bar{q}_2$ are solutions of (1.1), $\lambda \in (0,1)$ arbitrary. This would lead to the contradiction

$$j(\lambda \bar{q}_1 + (1-\lambda)\bar{q}_2) < \lambda j(\bar{q}_1) + (1-\lambda) j(\bar{q}_2) = \bar{j}.$$

□

In the following part of this thesis, dealing with the numerical approaches, locally optimal solutions are searched, i.e. controls $\bar{q} \in Q$ with $S(\bar{q}) \in X_{ad}$ such that

$$\exists \text{ neighborhood } Q_0 \text{ of } \bar{q} \text{ s.t. } j(\bar{q}) \le j(q) \quad \forall q \in Q_0 \text{ with } S(q) \in X_{ad}, \tag{2.32}$$

as these can be characterized well and in an accessible form.

2.3. Discretization and optimization algorithms for problems without pointwise constraints

The upcoming section will give an overview over the adaptive numerical solution of optimal control problems without pointwise constraints. The methods of optimization and discretization widely employed to these problems are not directly transferable to the state constrained problem (1.1). But they form the basis for the development of such algorithms, which will be derived in Chapter 3 (elliptic problems) and Chapter 4 (parabolic problems).

The class of problems without additional pointwise constraints central to this section is

$$(\bar{P}) \begin{cases} \min J(q,u) & q \in Q, u \in V \\ u = S(q) & . \end{cases} \tag{2.33}$$

The concrete formulation of its elliptic variant uses the form $a\colon Q \times V \times V \to \mathbb{R}$ as defined in (2.5). However the omittance of the pointwise state constraints removes the necessity of securing continous state functions at this point. Thus the space V can be left as the state space and the formulation of the unconstrained elliptic optimal control problem is

$$(\bar{P}_{ell}) \begin{cases} \min J(q,u), & q \in Q, u \in V \\ a(q,u)(\varphi) = (f,\varphi) & \forall \varphi \in V. \end{cases} \tag{2.34}$$

Similarly, in the unconstrained parabolic optimal control problem, the state space is chosen as $W(I,V)$, such that the problem as a whole reads

$$(\bar{P}_{par}) \begin{cases} \min J(q,u), & q \in Q, u \in W(I,V) \\ (\partial_t u, \varphi)_I + a(q,u)(\varphi) + (u(0), \varphi(0)) = (f,\varphi)_I + (u_0(q), \varphi(0)) & \forall \varphi \in W(I,V). \end{cases} \tag{2.35}$$

While the derivation of the optimality conditions is discussed in many sources, e.g., [92], the approach for the evaluation of derivatives is detailed, e.g., in [65].

2.3.1. Optimality conditions

A first-order optimality condition can be shown easily.

Lemma 2.14. *If $\bar{q} \in Q$ is a locally optimal solution of the problem* (2.34) *or* (2.35)*, and the reduced cost functional $j(q) = J(q, S(q))$ is Gateaux differentiable in the point $\bar{q}$, then there holds*

$$j'(\bar{q})(\delta q) = 0 \quad \forall \delta q \in Q$$

Proof. See [65]. □

This result can not be carried over to the state constrained problem (2.27). The reason is that the proof considers for every direction $\delta q \in Q$ the points $\bar{q} + \lambda \delta q$. For state constrained problems, there may be directions δq such that the point $\bar{q} + \lambda \delta q$ is not feasible for any λ from an interval $(0, \lambda_0)$ with some $\lambda_0 > 0$. The proof could only be transferred under the additional

assumption that $\bar{q}$ is feasible together with a neighborhood. This assumption would lose out on the crucial situation of an active state constraint.

Since in general the problem (P) is non-convex, in order to prove first-order optimality conditions, so called Karush-Kuhn-Tucker conditions, a constraint qualification is needed. In the following it is assumed that the *constraint qualification of Kurcyusz and Zowe* holds. A general formulation of this condition and its application in different settings can be found in [92, Section 6.1.2]. In the context of the unconstrained optimal control problem here it can be formulated as follows:

Assumption 2.15. *Let $\bar{q} \in Q$ be a locally optimal solution of the problem* (2.34) *or* (2.35)*. Then the operator $S'(\bar{q})$ is a surjective operator.*

Remark 2.4. For some types of semilinear elliptic optimal control problems, Assumption 2.15 can be proven regardless, see the example in [92, Page 250].

The optimality condition then reads:

Lemma 2.16. *Let $(\bar{q}, \bar{u})$ be a locally optimal point of the unconstrained optimal control problem* (2.34) *or* (2.35)*, and let Assumption 2.15 be fulfilled. Then with the Lagrange functional defined in the elliptic case as*

$$\bar{\mathcal{L}} \colon Q \times V \times V \to \mathbb{R}, \quad \bar{\mathcal{L}}(q,u,z) := J(q,u) + (f,z) - a(q,u)(z) \tag{2.36}$$

and in the parabolic case as

$$\begin{aligned} &\bar{\mathcal{L}} \colon Q \times W(I,V) \times W(I,V) \to \mathbb{R}, \\ &\quad \bar{\mathcal{L}}(q,u,z) := J(q,u) + (f - \partial_t u, z)_I - a(q,u)(z) + (u_0(q) - u(0), z(0)) \end{aligned} \tag{2.37}$$

the following first-order necessary optimality condition holds: There exists an adjoint state $\bar{z} \in X$, such that

$$\bar{\mathcal{L}}'_z(\bar{q},\bar{u},\bar{z})(\varphi) = 0 \quad \forall \varphi \in V \text{ (elliptic) or } \forall \varphi \in W(I,V) \text{ (parabolic)} \tag{2.38a}$$

$$\bar{\mathcal{L}}'_u(\bar{q},\bar{u},\bar{z})(\varphi) = 0 \quad \forall \varphi \in V \text{ (elliptic) or } \forall \varphi \in W(I,V) \text{ (parabolic)} \tag{2.38b}$$

$$\bar{\mathcal{L}}'_q(\bar{q},\bar{u},\bar{z})(\psi) = 0 \quad \forall \psi \in Q. \tag{2.38c}$$

Proof. The existence of the adjoint state is detailed in [101]. The display of the conditions using the Lagrange functional tightens the notation, see also [92] for a discussion of the formal Lagrange principle. □

It is also possible to examine the existence of optimality conditions of second order, see, e.g., [92], but in this thesis the numerical approach and optimization algorithms rely on the first-order necessary optimality conditions.

2.3.2. Evaluation of derivatives

In the last section, Lemma 2.14 gave the optimality condition $j'(\bar{q})(\delta q) = 0 \quad \forall \delta q \in Q$ as a starting point for the solution of the unconstrained problem

$$(\bar{P}) \qquad \Leftrightarrow \qquad (\bar{P}_{red}) \qquad \min j(q), \quad q \in Q \tag{2.39}$$

So during an iterative algorithm to find $\bar{q}$, we need to be able to evaluate the first derivative $j'(q)(\delta q)$ for the current iterate q in any direction δq. We use the quantities from the Lagrange approach to find a suitable representation. Thus for the current choice of q we ensure that

$$u = S(q) \quad \text{is fixed as the solution of the state equation}$$

during the course of the algorithm. Analog, for the current q and $u = S(q)$, the solution of the dual equation

$$\bar{\mathcal{L}}'_u(q, u, z)(\varphi) = 0 \quad \forall \varphi \in V \tag{2.40}$$

is denoted by z and called dual or adjoint state. By $T\colon Q \to V$ we denote the operator mapping q to its associated dual state, and in the implementation we ensure that

$$z = T(q) \quad \text{is fixed as the solution of the adjoint equation}$$

during the course of the algorithm. With these choices, the state equation is equivalent to

$$\bar{\mathcal{L}}'_z(q, u, z)(\varphi) = 0 \quad \forall \varphi \in V. \tag{2.41}$$

Thus we get the following representations for the reduced cost functional and its first derivatives:

$$j(q) = \bar{\mathcal{L}}(q, u, z), \tag{2.42}$$

$$j'(q)(\delta q) = \bar{\mathcal{L}}'_q(q, u, z)(\delta q). \tag{2.43}$$

The latter can be expressed explicitly by

$$j'(q)(\delta q) = J'_q(q, u)(\delta q) - a'_q(q, u)(\delta q, z) \qquad \text{(elliptic case), and} \tag{2.44}$$

$$j'(q)(\delta q) = J'_q(q, u)(\delta q) - a'_q(q, u)(\delta q, z) + (u'_0(q)(\delta q), z(0)) \qquad \text{(parabolic case).} \tag{2.45}$$

This representation is advantageous since the evaluation of the directional derivative of j in the point q in an arbitrary number of directions δq requires only one solution of a differential equation, as the adjoint equation (2.38b) does not depend on the direction δq.

Later in the process of solving the nonlinear equation $j'(\bar{q})(\delta q) = 0$ by the Newton method it is necessary to evaluate second derivatives of j. More specifically, after discretization the system of equations $\nabla^2 j(q)\delta q = -\nabla j(q)$ is typically very large. By using matrix-free methods the full assembling of the Hessian matrix is avoided. Instead, the evaluation of matrix-vector products for given directions δq is needed. This is equivalent to the evaluation of $j''(q)(\delta q, \tau q)$ for one given δq and all directions τq. In the derivation of a favorable representation, we start with equation (2.43) and add the terms $\bar{\mathcal{L}}'_u(q, u, z)(v)$ and $\bar{\mathcal{L}}'_z(q, u, z)(w)$ to its right hand side. They are both zero for any $v, w \in V$ due to the choices $u = S(q)$, $z = T(q)$. The resulting equation,

$$j'(q)(\delta q) = \bar{\mathcal{L}}'_q(q, u, z)(\delta q) + \bar{\mathcal{L}}'_u(q, u, z)(v) + \bar{\mathcal{L}}'_z(q, u, z)(w),$$

is differentiated in direction τq. Using the notation

$$\tau u = S'(q)\tau q \quad \text{and} \quad \tau z = T'(q)\tau q, \tag{2.46}$$

this gives the representation

$$\begin{aligned} j''(q)(\delta q, \tau q) = \quad & \bar{\mathcal{L}}''_{qq}(q,u,z)(\delta q, \tau q) + \bar{\mathcal{L}}''_{qu}(q,u,z)(\delta q, \tau u) + \bar{\mathcal{L}}''_{qz}(q,u,z)(\delta q, \tau z) \\ & + \bar{\mathcal{L}}''_{uq}(q,u,z)(v, \tau q) + \bar{\mathcal{L}}''_{uu}(q,u,z)(v, \tau u) + \bar{\mathcal{L}}''_{uz}(q,u,z)(v, \tau z) \\ & + \bar{\mathcal{L}}''_{zq}(q,u,z)(w, \tau q) + \bar{\mathcal{L}}''_{zu}(q,u,z)(w, \tau u), \end{aligned} \tag{2.47}$$

which holds for all $v, w \in V$. We can show that it is possible to choose one $v \in V$ in such a way that

$$\bar{\mathcal{L}}''_{qz}(q,u,z)(\delta q, \varphi) + \bar{\mathcal{L}}''_{uz}(q,u,z)(v, \varphi) = 0 \quad \forall \varphi \in V,$$

since by differentiation of (2.41) in direction δq this equation is true for the choice

$$v = S'(q)\delta q =: \delta u.$$

In an analogous way we can show that it is possible to choose one $w \in V$ in such a way that

$$\bar{\mathcal{L}}''_{qu}(q,u,z)(\delta q, \varphi) + \bar{\mathcal{L}}''_{uu}(q,u,z)(v, \varphi) + \bar{\mathcal{L}}''_{zu}(q,u,z)(w, \varphi) = 0 \quad \forall \varphi \in V,$$

since by differentiation of (2.40) in direction δq this equation is true for the choice

$$w = T'(q)\delta q =: \delta z.$$

The remaining terms determine the representation

$$j''(q)(\delta q, \tau q) = \bar{\mathcal{L}}''_{qq}(q,u,z)(\delta q, \tau q) + \bar{\mathcal{L}}''_{uq}(q,u,z)(\delta u, \tau q) + \bar{\mathcal{L}}''_{zq}(q,u,z)(\delta z, \tau q) \tag{2.48}$$

To summarize the procedure in explicit form, the evaluation of $j''(q)(\delta q, \tau q)$ for one given direction δq and possibly many given directions τq is performed as follows: In the implementation, for the current iterate q we have calculated $u = S(q), z = T(q)$. Then, in the elliptic case

- Given δq, compute δu by solving the *tangent equation*, which is

$$a'_u(q,u)(\delta u, \varphi) = -a'_q(q,u)(\delta q, \varphi) \qquad \forall \varphi \in V. \tag{2.49}$$

- Given $\delta q, \delta u$, compute δz by solving the *additional adjoint equation*, which is

$$\begin{aligned} a'_u(q,u)(\varphi, \delta z) = \quad & J''_{qu}(q,u)(\delta q, \varphi) + J''_{uu}(q,u)(\delta u, \varphi) \\ & - a''_{uu}(q,u)(\delta u, \varphi, z) - a''_{qu}(q,u)(\delta q, \varphi, z) \qquad \forall \varphi \in V. \end{aligned} \tag{2.50}$$

- Calculate $j''(q)(\delta q, \tau q)$ by

$$\begin{aligned} j''(q)(\delta q, \tau q) = \quad & J''_{qq}(q,u)(\delta q, \tau q) + J''_{uq}(q,u)(\delta u, \tau q) \\ & - a''_{qq}(q,u)(\delta q, \tau q, z) - a''_{uq}(q,u)(\delta u, \tau q, z) - a'_q(q,u)(\tau q, \delta z). \end{aligned} \tag{2.51}$$

In the parabolic case, the equations are as follows:

- Tangent equation: given δq, compute δu by solving

$$(\partial_t \delta u, \varphi)_I + a'_u(q,u)(\delta u, \varphi) + (\delta u(0), \varphi(0)) = -a'_q(q,u)(\delta q, \varphi) + (u'_0(q)(\delta q), \varphi(0)) \forall \varphi \in V. \tag{2.52}$$

- Additional adjoint equation: given $\delta q, \delta u$, compute δz by solving

$$\begin{aligned} &- (\varphi, \partial_t \delta z)_I + a'_u(q,u)(\varphi, \delta z) + (\varphi(T), \delta z(T)) = \\ &- a''_{uu}(q,u)(\delta u, \varphi, z) - a''_{qu}(q,u)(\delta q, \varphi, z) + J''_{uu}(q,u)(\delta u, \varphi) + J''_{qu}(q,u)(\delta q, \varphi) \forall \varphi \in V. \end{aligned} \tag{2.53}$$

  Note that this equation runs backward in time.

- Calculate $j''(q)(\delta q, \tau q)$ by

$$\begin{aligned} j''(q)(\delta q, \tau q) &= J''_{qq}(q,u)(\delta q, \tau q) + J''_{uq}(q,u)(\delta u, \tau q) - a''_{qq}(q,u)(\delta q, \tau q, z) \\ &- a''_{uq}(q,u)(\delta u, \tau q, z) - a'_q(q,u)(\tau q, \delta z) + (u'_0(q)(\tau q), \delta z(0)) + (u''_0(q)(\delta q, \tau q), z(0)). \end{aligned} \tag{2.54}$$

2.3.3. Discretization

In further preparation of the construction of approximate solution algorithms of the optimal control problems, a discretization of the involved infinite dimensional objects is carried out. For analytical purposes it is convenient to execute the discretization sequentially, yielding optimal control problems on different levels of discretization. This section is used to explain this idea and introduce the used notation in the context of unconstrained optimal control problems. Detailed extensions for the treatment of state constrained OCPs are done in Sections 3.3 and 4.3.
The stepwise discretization of the state and control spaces is as follows:

- The starting point of all considerations, the problem (P) introduced in (1.1), is the continous problem. It is concretized as elliptic problem in (2.13) or as parabolic problem in (2.21).

- For parabolic problems, a semidiscretization in time is performed. This corresponds to the dissection of the time interval $\bar{I}$ into subintervals I_m by the choice of time points $0 = t_0 < t_1 < \ldots, t_M = T$, and the construction of the semidiscretized state space $\tilde{X}_k$, which contains those functions that are polynomials in time if restricted to one of the intervals I_m. The related discretization parameter is a function $k\colon \bar{I} \to \mathbb{R}$ taking at every time point $t \in [0,T]$ the value that is the length of the interval I_m that contains t. It is used as a subscript in all the related quantities.
  Allowing for state functions from $\tilde{X}_k$ in the formulation of the optimal control problem results in the semidiscrete problem (P_k).

- For both elliptic and parabolic problems, a discretization in space is performed. For elliptic problems, this corresponds to the choice of a finite dimensional subspace $X_h \subset X$. The associated mesh $\mathcal{T}_h$ is a dissection of $\bar{\Omega}$ into spatial elements; the space X_h contains those globally continous functions that are polynomials on every element. The related discretization parameter h is the function $h\colon \bar{\Omega} \to \mathbb{R}$ taking at every spatial point the

value of the diameter of the spatial element that contains this point. The set of nodes of the mesh $\mathcal{T}_h$ is denoted by $\mathcal{N}_h$, and their number by N_h.
For parabolic problems, two different approaches are considered. The first one utilizes the same spatial discretization at every time point, so it uses one mesh $\mathcal{T}_h$ and one related function $h\colon \bar{\Omega} \to \mathbb{R}$ like before. Consequently, the space $\tilde{X}_{kh} \subset \tilde{X}_k$ contains those functions whose restriction to any temporal interval is a globally continous, elementwise polynomial function. The second approach allows for different meshes $\mathcal{T}_h^m$ on each of the subintervals I_m and in the initial point t_0. These are related to $M+1$ functions $h_i\colon \bar{\Omega} \to \mathbb{R}$, $i = 0 \dots M$. The discrete state space $\tilde{X}_{kh}$ is made up of those functions whose restriction to the kth temporal interval, or $t = t_0$, is globally continous and polynomial on every spatial element from precisely the kth mesh. Denote by N_m the number of nodes of the mesh $\mathcal{T}_h^m$, and by N_{tot} and $N_{\max}$ the sum and maximum over the respective numbers for all meshes.
Allowing for state functions from X_h or $\tilde{X}_{kh}$ results in the problem (P_h) or (P_{kh}), respectively.

- In the case of an infinite dimensional control space Q, the control space needs to be discretized as well. Even if it is already finite dimensional, it can be worthwhile to choose a smaller subspace. Since the control space is kept abstract, one cannot describe the discretization process more precisely than by introduction of a finite dimensional subspace $Q_d \subset Q$. One can at least give a few hints concerning common situations. If Q consists of functions with domain $\bar{\Omega}$ or $\bar{I} \times \bar{\Omega}$, like X, a discretization analog to the one of X, has the advantage that some residual term in the a posteriori error estimator vanishes. A coarser control can sometimes be useful as well. In parameter control problems, where Q is already discrete, we simply set $Q_d = Q$.
  Utilizing discrete controls allows finally to formulate the fully discrete problem $(P_{h,d})$, or $(P_{k,h,d})$, respectively.
  Alternatively, for some problem classes the discretization of even an infinite dimensional control space can be avoided if the variational discretization concept is used, see Remark 3.4.

For simpler notation, an overall discretization parameter σ will be used as a collective quantity for all possible discretization procedures of a concrete problem. Comparing with above, it can take the values $\sigma = (h)$, $\sigma = (k,h)$, $\sigma = (h,d)$ or $\sigma = (k,h,d)$. Thus, the optimal solution of the fully discretized problem is always denoted by (q_σ, u_σ).

On these levels of discretization, in order to formulate the optimal control problems, it does not suffice to replace the function spaces. It is also necessary to discretize the state equation, as it is not guaranteed that $S(q) \in X_h$ or $S(q) \in X_{kh}$ for $q \in Q$ or $q \in Q_d$. In the respective chapters the discrete state equations will be introduced. This is equivalent to the introduction of discrete solution operators for the according levels of discretization:

$$\begin{aligned} & & S_k\colon Q \to \tilde{X}_k, \\ S_h\colon Q \to X_h \quad & \text{or} \quad & S_{kh}\colon Q \to X_{kh}. \end{aligned}$$

We introduce S_σ to refer to the highest level of discretization, so $S_\sigma := S_h$ for elliptic, and $S_\sigma := S_{kh}$ for parabolic problems.

Also the state constraint is discretized by a function evaluation in finitely many points, e.g., the mesh points of the discretization of the state. In the respective chapters the constraint $G(\cdot, u_\sigma(\cdot)) \geq 0$ for infinitely many points x or (t, x) is discretized by a constraint $G_\sigma(\cdot, u_\sigma(\cdot)) \geq 0$ in finitely many points.

The cost functional does not need to be discretized explicitly, but is discretized indirectly by the insertion of the discrete state into the functional. The discretized reduced cost functionals are defined as

$$j_k \colon Q \to \mathbb{R}, \quad j_k := J(q, S_k(q)) \tag{2.55}$$

$$j_h \colon Q \to \mathbb{R}, \quad j_h := J(q, S_h(q)) \quad \text{or} \quad j_{kh} \colon Q \to \mathbb{R}, \quad j_{kh} := J(q, S_{kh}(q)) \tag{2.56}$$

Analog to the notation before, j_σ always refers to the highest level of discretization.

2.3.4. Optimization methods for unconstrained problems

It has not been discussed yet at which point in the solution process of (1.1) the discretization is applied. It is possible to discretize (P) directly and then apply a finite dimensional optimization algorithm, which is called the *discretize-then-optimize* approach. Or one can apply optimization theory to (P) and discretize later, when optimality conditions have been found, the *optimize-then-discretize* approach. This decision is also connected to the utilized optimization method. Since some of the optimization algorithms can be formulated only for the discrete problem, the derivation of all algorithms for the comprehensive solution will be made using the discretize-then-optimize approach formally. However it can be shown that when a Galerkin type discretization is employed, and the state and adjoint variables are discretized by the same method, the two approaches lead to the same discrete optimality system, [52, Section 3.2], so that the discrimination between these two approaches does not need to be pursued in this thesis from now on.

The explanation of the optimization method for unconstrained problems, that will be used as a basis for the development of algorithms to solve (1.1), will be carried out for the discretize-then-optimize approach. The discretization will hereby be left abstract, we assume to be given the discrete spaces $Q_d \subset Q$ and $X_h \subset X$ or $X_{kh} \subset X$. It can however be anticipated here that a Galerkin finite element discretization will be used later. So for the task of this subsection, to find an optimal solution of the discretized version of (2.33), no additional optimality conditions and related equations need to be derived. Due to the analog structure, the relations from Section 2.3.2 stay valid, just by replacing the spaces Q and X by their discrete counterparts.

Remark 2.5. The restriction of $X_{kh} \subset X$ is only for simplicity of the expressions in this section and does not allow for discontinous Galerkin discretization in time. This type of time discretization causes additional jump terms in the state, adjoint and additional adjoint equations, the explicit formulation can be found in [65]. For state constrained OCPs in the main part of this thesis, dG methods will be considered and the corresponding terms will be derived in Chapter 4.

The algorithm is based on the Newton method to solve the nonlinear equation $j'_\sigma(q) = 0$, which is the first order optimality condition for the considered problem

$$\begin{cases} \min J(q,u), & q \in Q_d, u \in X_h \text{ or } X_{kh} \\ u = S_\sigma(q) \end{cases} \qquad \Leftrightarrow \qquad \min j_\sigma(q), \quad q \in Q_d. \tag{2.57}$$

As indicated before, in order to exploit the representations of Section 2.3.2 it needs to be ensured that during the course of the optimization algorithm the state and adjoint variable are set as the solutions of the discrete state and adjoint equations, represented by the operators $u = S_\sigma(q)$ and $z = T_\sigma(q)$, with the current iterate q. In the following we describe how to calculate one Newton step for the equation $j'_\sigma(\bar{q})(\delta q) = 0 \quad \forall \delta q \in Q_d$. This means, given the current iterate $q \in Q_d$ we search for the direction $\delta q \in Q_d$ in which the step is taken, i.e. the next iterate is determined by

$$q + \lambda \delta q,$$

where $\lambda = 1$ is chosen for a full Newton step, or a $\lambda < 1$ is determined, e.g., by a line search method. The full Newton step δq is determined by

$$j''_\sigma(q)(\delta q, \tau q) = -j'_\sigma(q)(\tau q) \qquad \forall \tau q \in Q_d, \tag{2.58}$$

or equivalently for all vectors τq from a basis of Q_d.

To set up the equations representing the necessary quantities as matrices and vectors to be used in an implementation, introduce the gradient $\nabla j_\sigma(q) \in Q_d$ and the Hessian $\nabla^2 j_\sigma(q) \colon Q_d \to Q_d$ by the usual Riesz representation formulas

$$\begin{aligned} (\nabla j_\sigma(q), \tau q)_Q &= j'_\sigma(q)(\tau q) && \forall \tau q \in Q_d \\ (\nabla^2 j_\sigma(q)\delta q, \tau q)_Q &= j''_\sigma(q)(\delta q, \tau q) && \forall \delta q, \tau q \in Q_d. \end{aligned}$$

Next, we want to express the gradient and Hessian by means of a basis $\{\tau q_i\}_{i=1}^{\dim Q_d}$. Denote the coefficient vector of the gradient with respect to that basis by $\mathbf{f}$, such that

$$\nabla j_\sigma(q) = \sum_{j=1}^{\dim Q_d} \mathbf{f}_j \tau q_j.$$

It follows with

$$(\nabla j_\sigma(q), \tau q_i) = \sum_{j=1}^{\dim Q_d} \mathbf{f}_j (\tau q_j, \tau q_i) \qquad \text{that} \qquad \mathbf{G}\mathbf{f} = \left(j'_\sigma(q)(\tau q_i)\right)_{i=1}^{\dim Q_d}, \tag{2.59}$$

where $\mathbf{G}$ is the Gramian matrix with the entries $(\tau q_j, \tau q_i)$ at the (i,j)-th position. The vector $(j'_\sigma(q)(\tau q_i))_{i=1}^{\dim Q_d}$ can be evaluated by the right hand sides of (2.44) or (2.45), respectively.

Next, the full Newton step δq is also expressed by its coefficient vector, denoted $\mathbf{d}$, such that

$$\delta q = \sum_{j=1}^{\dim Q_d} \mathbf{d}_j \tau q_j.$$

Its definition equation

$$(\nabla^2 j_\sigma(q)\delta q, \tau q_i) = -(\nabla j_\sigma(q), \tau q_i) \quad i = 1, 2, \ldots, \dim Q_d. \tag{2.60}$$

thus becomes

$$\sum_{j=1}^{\dim Q_d} \mathbf{d}_j (\nabla^2 j_\sigma(q)\tau q_j, \tau q_i) = -(\nabla j_\sigma(q), \tau q_i) \quad i = 1, 2, \ldots, \dim Q_d,$$

so that $\mathbf{d}$ is determined by

$$\mathbf{K}\mathbf{d} = (\nabla^2 j_\sigma(q)\delta q, \tau q_i)_{i=1}^{\dim Q_d} = (j''_\sigma(q)(\delta q, \tau q_i))_{i=1}^{\dim Q_d} = -\mathbf{G}\mathbf{f}, \tag{2.61}$$

where $\mathbf{K}$ is the matrix with the entries $j''_\sigma(q)(\tau q_j, \tau q_i)$ at the (i, j)-th position. The entries of $\mathbf{K}$ can be evaluated by the right hand sides of (2.51) or (2.54), respectively. This allows finally to set up the linear system used to determine the Newton step as

$$\mathbf{H}\mathbf{d} = -\mathbf{f} \tag{2.62}$$

where the matrix $\mathbf{H} = \mathbf{G}^{-1}\mathbf{K}$ as the coefficient matrix of the Hessian $\nabla^2 j_\sigma(q)$ is symmetric. The execution of the Newton algorithm with the explicit buildup of this matrix $\mathbf{H}$ and the following exact solution of the linear system (2.62) is called an *exact Newton method*. If $\dim Q_d$ is very large, this computation is very costly due to $\mathbf{H}$ typically not being sparse, and can be avoided by solving (2.62) iteratively by a method that utilizes only products of of the matrix with a vector, e.g. the CG method. Hereby a product of the form $\nabla^2 j_\sigma(q)\delta q$ is represented by its coefficient vector $\mathbf{h}$ where

$$\mathbf{G}\mathbf{h} = (j''_\sigma(q)(\delta q, \tau q_i))_{i=1}^{\dim Q_d} \tag{2.63}$$

similar as before. This approach including the approximative solution of (2.62) is an example of an *inexact Newton method*. For more considerations on different solvers see [65].

Assembling all the parts introduced above, we obtain Algorithm 2.1 for finding the optimal solution of problem (2.57). Techniques well established for Newton type algorithms, like line search, and a stopping criterion based on the norm of the coefficient vector $\mathbf{f}$, complete the algorithm.

Algorithm 2.1. Newton-type optimization for an unconstrained optimal control problem

input data: current triple q^0, u^0, z^0
where there is secured $u^0 = S_\sigma(q^0), z^0 = T_\sigma(q^0)$
parameter: TOL_N, TOL_L
Set counter $i = 0$.
repeat
  Compute $\mathbf{f}$ as vector representation of $\nabla j_\sigma(q^i)$ by (2.59)
  Compute $\mathbf{d}$ as vector representation of the Newton update δq
  by solving $\mathbf{Hd} = -\mathbf{f}$ iteratively, e.g., by CG method with tolerance TOL_L
  for any product $\mathbf{H\tilde{d}}$ the CG algorithm requests **do**
    With $\widetilde{\delta q}$ being the direction represented by $\tilde{\mathbf{d}}$
    Compute $\widetilde{\delta u}$ by the tangent equation of the current problem
    i.e. (2.49) or (2.52)
    Compute $\widetilde{\delta z}$ by the additional adjoint equation of the current problem
    i.e. (2.50) or (2.53)
    Evaluate (2.51) or (2.54) to get right hand side of (2.63)
    Get $\mathbf{h} = \mathbf{H\tilde{d}}$ by solving (2.63)
  Determine step length λ^i by line search
  (might involve repeated solution of the state equation)
  Set $q^{i+1} = q^i + \lambda^i \delta q$
  Compute $u^{i+1} = S_\sigma(q^{i+1})$
  Compute $z^{i+1} = T_\sigma(q^{i+1})$
  $i = i + 1$
until $|\nabla j_\sigma(q^i)| \leq TOL_N$
output data: q^i, u^i, z^i

2.4. Treatment of inequality constraints

In this section we will give a raw plan on the necessary steps to include state constraints into the analytic and algebraic framework laid out in Section 2.3. The basic equations exploited in the set up were the optimality conditions (2.38). Thus first the equivalent Karush-Kuhn-Tucker conditions for state constrained problems are derived. The evaluation of the Kurcyusz-Zowe constraint qualification according to [92] yields a more comprehensive condition:

Assumption 2.17. *Let $\bar{q}$ be a locally optimal control for (P_{ell}). Additionally to Assumption 2.15, there exists a control $\hat{q} \in Q$ such that $S(\bar{q}) + S'(\bar{q})(\hat{q} - \bar{q}) \in \mathrm{int}(X_{ad})$.*

It is also called a *local Slater condition*, and has the meaning that the resulting function $S(\bar{q}) + S'(\bar{q})(\hat{q} - \bar{q})$ has no active points. The equivalence of the local Slater condition and the Kurcyusz-Zowe constraint qualification relies on the fact that the set used to formulate the inequality constraints, i.e. X_{ad}, possesses interior points, see [92]. It was therefore crucial to set up the state space X as a space of continous functions, a demand that was made at the very beginning in (2.1). Without this property the KKT conditions, which are the basis of the numerical solution algorithms and the error estimation and adaptivity process, would not stand.

The compact representation of the optimality conditions uses the Lagrange functionals, defined by

$$\mathcal{L}\colon Q \times X \times Z \times \mathcal{M}(\Omega) \to \mathbb{R}, \quad \mathcal{L}(q,u,z,\mu) := J(q,u) + (f,z) - a(q,u)(z) - \langle \mu, G(u)\rangle \tag{2.64}$$

in the elliptic case and

$$\begin{aligned} &\mathcal{L}\colon Q \times X \times Z \times \mathcal{M}(I\times\Omega) \to \mathbb{R}, \\ &\mathcal{L}(q,u,z,\mu) := J(q,u) + (f - \partial_t u, z)_I - a(q,u)(z) + (u_0(q) - u(0), z(0)) - \langle \mu, G(u)\rangle \end{aligned} \tag{2.65}$$

in the parabolic case. With the use of the following notation for Borel measures $\mu \in \mathcal{M}(\Omega)$,

$$\mu \geq 0 \qquad \Leftrightarrow \qquad \langle \mu, f\rangle \geq 0 \quad \forall f \in C(\bar{\Omega}) \text{ with } f(x) \geq 0 \text{ in } \Omega, \tag{2.66}$$

the optimality conditions take the following form:

Lemma 2.18. *Let $(\bar{q},\bar{u})$ be a locally optimal point of the state constrained optimal control problem* (2.13) *or* (2.21)*, and let Assumption 2.17 hold. Then there exists an adjoint state $\bar{z} \in Z$, and a multiplier $\bar{\mu}$ with $\bar{\mu} \in \mathcal{M}(\Omega)$ in the elliptic case and $\bar{\mu} \in \mathcal{M}(I\times\Omega)$ in the parabolic case, such that*

$$\begin{aligned} \mathcal{L}'_z(\bar{q},\bar{u},\bar{z},\bar{\mu})(\varphi) &= 0 \quad \forall \varphi \in Z \\ \mathcal{L}'_u(\bar{q},\bar{u},\bar{z},\bar{\mu})(\varphi) &= 0 \quad \forall \varphi \in X \\ \mathcal{L}'_q(\bar{q},\bar{u},\bar{z},\bar{\mu})(\psi) &= 0 \quad \forall \psi \in Q \\ \langle \bar{\mu}, G(\bar{u})\rangle = 0, &\qquad \bar{\mu} \geq 0. \end{aligned} \tag{2.67}$$

The proof of existence is again done in [101], for the representation using $\mathcal{L}$ compare [92, Section 6.2].
Some recent results on second-order sufficient optimality conditions for state constrained elliptic problems can be found in [20].

The next steps in the transfer of the ideas from unconstrained problems, the evaluation of the derivatives and the discretization, are very specific to the type of the state equation and thus discussed in the specific chapters. For the treatment of state constraints in the optimization process, a wide variety of approaches has been developed. An overview will be given in the following.

The first method to be outlined is the primal-dual **active set** (PDAS) method. Note that the loss of regularity, reflected by the introduction of the measure μ, has a direct effect on the choice of the optimization method, see [14] and the references therein. Also, for the continous state constrained optimal control problems the PDAS method can not be established as an analog to the control constrained case. Instead, the method is formulated for the discretized problems.
The description is reduced to the variables q_d and μ_h, thus u_h and z_h are required to be coupled to q_d and μ_h by the discrete state and adjoint equations. For elliptic problems, the basic idea is as follows: given an actual control q_d and multiplier μ_h we alternate between the following

steps: First determine the active set, that is the set of points where the state constraints are exactly fulfilled or plain violated

$$A = \{x_i \in \mathcal{N}_h : G_h(u_h(x_i)) + c \cdot \mu_i \leq 0\}, \tag{2.68}$$

with some constant $c > 0$. Then, calculate a new pair (q_d,μ_h) by solving the minimization subproblem

$$(P_E) \begin{cases} \min J(S_h(q), q) & q \in Q_d \\ G_h(S_h(q)(x_i)) = 0 & \forall x_i \in A \end{cases},$$

which requires an equality constraint to be fulfilled on the active set. The repetition of the two steps is stopped if two successively computed active sets are equal. The PDAS method will be described in detail in Section 3.4 for the solution of elliptic problems. It is equivalent to a semismooth Newton method. Thus for one given discretization the PDAS method converges superlinearly, but considering the repeated solution of the discrete optimal control problems on adaptively refined meshes, there holds no mesh-independence. For a detailed discussion, see [47, 49].
It should be noted that (P_E) is only guaranteed to have a solution for some kinds of optimal control problems. In general this does not hold, e.g., for boundary control problems. This question is discussed in detail in Section 3.4.

Remark 2.6. A method that incorporates the state constraint directly without utilizing a Lagrange multiplier, for a smaller class of problems, is introduced in [50]. It utilizes a level set approach.

Another class of methods to incorporate state constraints are **regularization methods**. Here the state constrained problem is altered in such a way, that the solution regains the original regularity and can thus be calculated by known methods. An example is the **barrier method**, see, e.g., [85, 97], where a regularization term is added to the cost functional: Instead of (2.13), the problem

$$(P_\gamma) \begin{cases} \min J_\gamma(q, u) := J(q, u) + b_\gamma(u) & q \in Q, u \in X \\ u = S(q) \end{cases} \tag{2.69}$$

is considered with a regularization parameter $\gamma > 0$ and a barrier functional $b_\gamma(u)$ that is chosen in such a way that it is small for values bounded away from the constraint, but goes to infinity as the state function approaches the constraint. Clearly, this problem belongs to the class (2.33) and can thus be solved by the techniques described before. Driving $\gamma \to \infty$ lets the solution of (P_γ), denoted by (q_γ, u_γ), approach the solution (q, u) of the state constrained problem. On the other hand, the regularization introduces an additional error, that has to be accounted for in the error estimation process.
A barrier method will be applied to elliptic state constrained OCPs in Section 3.6, and studied in depth when solving parabolic problems in Chapter 4. Naturally, also the problems (P_γ) are discretized to solve them approximately. The solutions of the discrete regularized problems are denoted by $(q_{\gamma\sigma}, u_{\gamma\sigma})$.

In contrast to barrier methods, in **Moreau-Yosida regularization**, the regularized problem is given by

$$(P_\gamma) \begin{cases} \min J(q, u) + \frac{1}{2\gamma} \int\limits_\Omega |(\bar{\lambda} - \gamma G(u))^+|^2 & q \in Q, u \in X \\ u = S(q) \end{cases} \tag{2.70}$$

in the elliptic case and the obvious analog for parabolic problems. In (P_γ), $\gamma > 0$ is the regularization parameter, $\bar{\lambda}$ a given square-integrable function, and $(\cdot)^+$ is short for $\max(0, \cdot)$. Here, infeasible iterates are allowed, as penalization is only done on violation of the bounds, and not on approaching the bounds. Driving $\gamma \to \infty$ lets the solutions of (P_γ) approach that of (P); see, e.g., [74] for parabolic problems. Path following methods, that describe how fast the iteration $\gamma \to \infty$ can be done, are discussed e.g. in [48].

Some different regularization methods depend on the actual structure of the problem, for example they require distributed control, which means q is defined on the same domain as u. In **Lavrentiev regularization**, the regularized problem is given by

$$(P_\varepsilon) \begin{cases} \min J(q,u) & q \in Q, u \in X \\ u = S(q) & \\ G(u) + \varepsilon q \geq 0 & \end{cases}, \tag{2.71}$$

with the regularization parameter $\varepsilon > 0$. The condition $G(u) + \varepsilon q \geq 0$ is a *mixed state-control constraint*. Optimal control problems with that type of constraint exhibit solutions of full regularity, in particular the Lagrange multiplier is an L^2-function, see, e.g., [75, 93], thus enabling the use of optimization algorithms derived from the optimality system in a way similar to the proceeding for optimal control problems without further inequality constraints. Driving $\varepsilon \to 0$ lets the solutions $(q_\varepsilon, u_\varepsilon)$ of (P_ε) approach the solution (q, u) of (P); see, e.g., [74] for parabolic problems. Strategies how to drive $\varepsilon \to 0$ have been considered, e.g., in [21].

In contrast to this situation, for boundary control it is not possible to add u and q as the domains of the state and the control function are different. A regularization can still be done by the **virtual control** concept, which was developed in [59]. The virtual control v is introduced as a new quantity, defined on Ω, or $I \times \Omega$. The regularized problem is then given by

$$(P_\varepsilon) \begin{cases} \min J(q,u) + \frac{\Psi(\varepsilon)}{2}\|v\|_{L^2(\Omega)} & q \in Q, u \in X, v \in L^2(\Omega) \\ u = S(q) + \hat{S}(\Phi(\varepsilon)v) & \\ G(u) + \xi(\varepsilon)v \geq 0 & \end{cases} \tag{2.72}$$

with some functions $\Psi(\varepsilon), \Phi(\varepsilon), \xi(\varepsilon)$, and the operator $\hat{S}$ is the solution operator of the partial differential equation that is obtained by equipping the original state equation with homogenous boundary conditions and using the argument of $\hat{S}$ as the distributed right hand side.

2.5. A posteriori error estimation and adaptive algorithm

The optimization methods introduced in the last section can be used to solve a discretized version of the optimal control problem (1.1). Any such discrete problem is described by a set T containing the following elements: the spatial meshes $\mathcal{T}_h$ or $\mathcal{T}_h^m$, $m = 0 \dots M$, for parabolic problems the time intervals I_m, $m = 1 \dots M$, and the discrete control space Q_d. If a barrier method is used, the regularization parameter γ is needed to describe the discretized problem as well.

In an algorithm to find the best possible approximation of $(\bar{q}, \bar{u})$, under restriction of the computational effort or tolerance, not only one such discrete problem is solved. Instead,

a sequence of such problems is solved, described by the discretizations $T^{(i)}$ and possibly parameters $\gamma^{(i)}$, $\quad i = 0, 1, \ldots$, in the following fashion: the starting discretization $T^{(0)}$, and possibly $\gamma^{(0)}$, are given. Now the discrete problem is solved and the respective error is estimated. The information generated from this estimation is used to create a refined discretization $T^{(i+1)}$, and possibly a new $\gamma^{(i+1)}$. This process is repeated until the overall error is estimated to be smaller than a given tolerance.

The easiest strategy for the refinement step $T^{(i)} \to T^{(i+1)}$ is *uniform refinement*: in the temporal discretization the time intervals are bisected into two equal parts, and in the spatial discretization every spatial element K is dissected into an appropriate number of elements of equivalent size, by bisection of each edge. Since in this refinement strategy no information from the problem itself is used, it can not be expected that it decreases the error in the fastest possible way.

Looking for a different strategy to decrease the error, one important question is: in which quantity is the error measured. This determines the quantity that is decreased well, and about evenly distributed over the cells or time intervals. A possible approach is assessing the error measured in the natural norms of the spaces in question, i.e.

$$\|\bar{u} - u_\sigma\|_X \text{ and } \|\bar{q} - q_\sigma\|_Q.$$

But since it is the minimization of J which determines the success of the computations, i.e. the convergence and its rate in terms of effort, one has a strong cause to estimate the discretization error with regard to the cost functional. The error estimators developed in this thesis for state constrained problems will follow this principle of goal oriented error estimation. This approach, estimating

$$J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma) \approx \eta,$$

has been developed in [8] considering unconstrained elliptic optimal control problems. It has since been successfully developed to be applied to parabolic problems [66], and problems including control constraints [44, 94] or state constraints [13, 41]. A somewhat more general concept is to estimate the error in a given functional $I \colon Q \times X \to \mathbb{R}$, called *quantity of interest*. This estimation of $I(\bar{q}, \bar{u}) - I(q_\sigma, u_\sigma)$ can be motivated by physical considerations, when the quantity that the user is actually interested in is not the one that is to be minimized. For unconstrained and control constrained problems this concept is developed in [11, 12, 65].

The error estimator η consists of all of the following shares, or a selection of it:

- η_k: For parabolic problems, a discretization in the time variable is necessary, the estimate of the temporal discretization error is denoted by η_k.
- η_h: A discretization in the space variable is always necessary, the estimate of the spatial discretization error is denoted by η_h.
- η_d: If the control space is discretized, $Q_d \subset Q$, the introduced control discretization error is estimated by η_d. On the other hand, should $Q_d = Q$, which can happen for finite dimensional control spaces only, then η_d does not occur.
- η_γ: If a regularization method is used, the introduced regularization error is estimated by η_γ. On the other hand, if we use the PDAS method, then η_γ does not occur.

Some of these can be dissected further; localized according to temporal and spatial influence:

η_k: For parabolic problems, the temporal discretization error consists of estimates of the error on the subintervals I_m,

$$\eta_k = \sum_{m=1}^{M} \eta_k^m. \tag{2.73}$$

η_h: The spatial discretization error consists of estimates of the error on the cells K. If there is only one mesh $\mathcal{T}_h$ we have

$$\eta_h = \sum_{K \in \mathcal{T}_h} \eta_{h,K}. \tag{2.74}$$

In the case of dynamic discretization in space the localization is

$$\eta_h = \sum_{m=0}^{M} \sum_{K \in \mathcal{T}_h^m} \eta_{h,K}^m. \tag{2.75}$$

Should either temporal or spatial discretization be chosen for refinement, the refinement can now be done locally guided by this localized error estimator. For the control discretization a similar construction is possible, if Q is also distributed in time and/or space.
The overall strategy is displayed in Algorithm 2.2. The parameters $c_1, c_2, c_3 \in (0,1), c_\gamma \in (1,\infty)$ can be used to fine tune the behavior of the algorithm, but should be chosen with care, as to allow for a sufficient distance in the distinction of the cases. Where not indicated otherwise, in the numerical experiments the values $c_1 = 0.6, c_2 = 0.8, c_3 = 0.9, c_\gamma = 3.16$ were used.

Before carrying out this a posteriori strategy, the question arises what improvement can be expected compared to a uniform refinement strategy. For reference, consider the optimal control problem with linear elliptic state equation, distributed control and tracking type cost functional, on a two-dimensional domain. For uniform discretization with discretization parameter h, the convergence rate of $\|\bar{q} - q_\sigma\|_{L^2(\Omega)}$, indicative of the one for the cost functional $J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma)$, is h^2 for the problem without additional pointwise constraints, see [64]. The inclusion of state constraints reduces the order to $h^{1-\varepsilon}$, see [26].
This simplest example shows already an order reduction. Therefore the goal of the adaptive refinement strategy is to improve the convergence order, or at least improve the convergence by a constant factor.

Similarly to state constraints, other types of singularities can cause convergence order reduction. Consider for example singularities due to reentrant corners or edges in nonconvex domains. In [5] the convergence order is improved by the creation of non-uniform meshes, albeit the utilized techniques use a priori information as opposed to the a posteriori approach used in this thesis.
In the case of optimal control problems it can be argued that the inclusion of the state constraints leads to the lower regularity of the optimal solution and thus the lower convergence rates via irregular data in the dual equation. Considering the sole finite element approximation of such a partial differential equation with irregular data, without connection to an optimal control problem, convergence order reduction can be countered by the use of graded meshes. In Appendix A such a partial differential equation is considered. Its finite element approximation

Algorithm 2.2. Error equilibration algorithm

input data: the old discretization $T = (\mathcal{T}_h, Q_d)$ (elliptic)
or $T = \left((I_m)_{m=1}^M, \mathcal{T}_h, Q_d\right)$ (parabolic)
or $T = \left((I_m)_{m=1}^M, (\mathcal{T}_h^m)_{m=0}^M, Q_d\right)$ (parabolic, dynamic)
and possibly the old regularization parameter γ
parameters: c_1, c_2, c_3, c_γ
Evaluate the relevant error estimators of $\eta_\gamma, \eta_k, \eta_h, \eta_d$ according to

- (3.52), (3.53) (elliptic OCP)
- (4.41), (4.45), (4.46) (parabolic OCP)

calculate relative contributions: with $\bar{\eta}_{tot} = |\eta_d| + |\eta_h| + |\eta_k| + |\eta_\gamma|$ these are $\bar{\eta}_1 = \frac{|\eta_d|}{\bar{\eta}_{tot}}, \bar{\eta}_2 = \frac{|\eta_h|}{\bar{\eta}_{tot}}, \bar{\eta}_3 = \frac{|\eta_k|}{\bar{\eta}_{tot}}, \bar{\eta}_4 = \frac{|\eta_\gamma|}{\bar{\eta}_{tot}}$
if the maximum relative contribution from $\{\bar{\eta}_1, \ldots, \bar{\eta}_4\}$ is $> c_1$ **then**
    choose the relevant structure for refinement
else if the two largest relative contributions from $\{\bar{\eta}_1, \ldots, \bar{\eta}_4\}$ combined are $> c_2$ **then**
    choose the two relevant structures for refinement
else if the three largest relative contributions from $\{\bar{\eta}_1, \ldots, \bar{\eta}_4\}$ combined are $> c_3$ **then**
    choose the three relevant structures for refinement
else
    choose all four structures for refinement
Refinement process. Set $\bar{T} = T$, $\bar{\gamma} = \gamma$.
if spatial discretization is chosen for refinement **then**
    refine $\mathcal{T}_h \Rightarrow \bar{\mathcal{T}}_{\bar{h}}$, see Algorithm 3.3 (elliptic)
    or $\mathcal{T}_h \Rightarrow \bar{\mathcal{T}}_{\bar{h}}$, see Algorithm 4.2 (parabolic)
    or $(\mathcal{T}_h^m)_{m=0}^M \Rightarrow \left(\bar{\mathcal{T}}_{\bar{h}}^m\right)_{m=0}^M$, see Algorithm 4.2 (parabolic, dynamic)
if temporal discretization is chosen for refinement **then**
    refine $(I_m)_{m=1}^M \Rightarrow \left(\bar{I}_m\right)_{m=1}^{\bar{M}}$, see Algorithm 4.3 (parabolic)
    or refine $(I_m)_{m=1}^M \Rightarrow \left(\bar{I}_m\right)_{m=1}^{\bar{M}}$, with $\left(\bar{\mathcal{T}}_{\bar{h}}^m\right)_{m=0}^M \Rightarrow \left(\bar{\mathcal{T}}_{\bar{h}}^m\right)_{m=0}^{\bar{M}}$ see Algorithm 4.3 (parabolic, dynamic)
if γ is chosen for refinement **then**
    set $\bar{\gamma} = c_\gamma \gamma$
if Q_d is chosen for refinement **then**
    refine $Q_d \Rightarrow \bar{Q}_d$ as described in Section 3.5 or Section 4.5
output data: the new discretization $\bar{T}$, the new regularization parameter $\bar{\gamma}$

on a family of uniform meshes would lead to convergence order h^2 for a regular right hand side, but only $h^{1-\varepsilon}$ for irregular data. Then a family of meshes $\mathcal{T}_h$ is constructed that is not uniform but *graded*, such that the convergence order is restored to $h^2|\ln(h)|^{\frac{3}{2}}$. The existence of such meshes justifies the expectation to restore the convergence order also for state constrained optimal control problems. The intention of this thesis is the creation of such meshes using a posteriori techniques, as the location of the singularity caused by the state constraint is a priori unknown.

Returning to the more general thoughts from the beginning of this section. To set up a fair comparison of the quality of different strategies solving (P) approximately, we need a

measure of the computational effort invested to reach a certain error level $J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma)$, or $J(\bar{q}, \bar{u}) - J(q_{\gamma\sigma}, u_{\gamma\sigma})$ if the solution method utilizes regularization as well as discretization. If no regularization of the problem is involved, the number of degrees of freedom of the discretization may be an acceptable measure for the complexity of the discrete problem. It is plausible that at least asymptotically this does not skew the comparison.

This does not hold when using regularization. One could, without changing the number of degrees of freedom, increase the regularization parameter γ reducing the error but increasing the necessary computational effort, as the problem gets harder to solve due to a larger condition number of the discrete problem. As this increase in relation to the one caused by increasing the degrees of freedom is unknown, it cannot be accounted for. To achieve a more reasonable comparison, one may

- compare only computational times. This has its drawbacks as it is implementation dependent and requires the numerical tests to be carried out on a closed system to avoid fluctuations in computational power,
- leave γ constant and only compare efficiency of the other refinements,
- couple the increase of γ to the number of degrees of freedom.

One could also investigate computational efficiency in terms not only of elapsed time, but also of needed storage space. We think however that the developments in the computer industry in the last decades have made the issue of limited storage space almost disappear, so we do not investigate this further. For some application problems like climate models it might however be a valid concern.

3. Elliptic Optimal Control Problems with State Constraints

In this chapter elliptic optimal control problems with pointwise state constraints are considered. With the notation from Section 2.1.2, such a problem takes the form

$$(P_{ell}) \begin{cases} \min J(q,u), & q \in Q, u \in X \\ a(q,u)(\varphi) = (f,\varphi) & \forall \varphi \in Z \\ G(u) \geq 0 \end{cases} \tag{3.1}$$

For a large class of semilinear elliptic state equations unique existence of a solution of the state equation is shown, and conditions under which a local optimal solution of (3.1) exists, and obeys first order Karush-Kuhn-Tucker optimality conditions are given.
For the numerical solution of any elliptic optimal control problem with a locally optimal point obeying these, the finite-element-discretization of problem (P_{ell}) will be executed, and two optimization algorithms will be discussed. For problems with distributed control, a primal-dual active set method can be used. Here, the Lagrange multiplier needs to be introduced into the implementation, so it is required to deal with Borel measures in the program code. For the aim of producing efficient meshes, an a posteriori error estimator is derived and utilized in an adaptive refinement algorithm. Alternatively, (P_{ell}) can be regularized and a sequence of regularized problems can be solved by an interior point algorithm.

3.1. Analysis of the state equation

From Section 2.1.2, recall the definitions of

$$X = V \cap W^{1,p}(\Omega) \qquad \text{with some } p > n, \tag{3.2}$$

such that the state space X contains continous functions from the weak solution space, and

$$Z = W^{1,p'}(\Omega) \qquad \text{where } \frac{1}{p} + \frac{1}{p'} = 1. \tag{3.3}$$

Linking the control and state space by Assumptions 2.1 and 2.2, the elliptic state equation has been formulated as

$$a(q,u)(\varphi) = (f,\varphi) \quad \forall \varphi \in X, \tag{3.4}$$

using the semilinear form

$$a \colon Q \times X \times Z \to \mathbb{R}.$$

The interpretation that the given space V with its properties, like the satisfaction of Dirichlet boundary conditions, is used to construct the state space X, could be transferred to the control space: the given space R is regarded as the spatial part of the control. For elliptic OCPs the distinction does not make a difference, we can simply set

$$Q = R \qquad \text{control space for elliptic OCPs,} \tag{3.5}$$

for parabolic problems this will be different. This procedure allows to include different possible choices of Q, especially different control domains. The distinction between control domains gives rise to the labeling of certain classes of elliptic control problems, some of which are given in the following examples by model equations in this framework.

Example 3.1. 1. For the choice $Q = L^2(\Omega)$, the following equation is an example for *distributed control*:

$$\begin{aligned} -\Delta u &= q \quad \text{in } \Omega \\ u|_\Gamma &= 0 \quad \text{on } \Gamma \end{aligned}$$

The weak formulation is obtained by $a(q,u)(\varphi) := (\nabla u, \nabla\varphi) - (q,\varphi)$ with state space $X = W_0^{1,p}(\Omega)$.

2. For the choice $Q = L^2(\Gamma)$, the following equation is an example for *boundary control*:

$$\begin{aligned} -\Delta u + u^3 &= 0 \quad \text{in } \Omega \\ \partial_n u|_\Gamma &= q \quad \text{on } \Gamma \end{aligned}$$

The weak formulation is obtained by $a(q,u)(\varphi) := (\nabla u, \nabla\varphi) + (u^3,\varphi) - (q,\varphi)_\Gamma$ with state space $X = W^{1,p}(\Omega)$.

3. For the choice $Q = \mathbb{R}^k$, so that the control space is in fact k-dimensional, the following equation is an example for *parameter control*:

$$\begin{aligned} -\Delta u &= \sum_{i=1}^{k} q_i f_i \quad \text{in } \Omega \\ u|_\Gamma &= 0 \quad \text{on } \Gamma, \end{aligned}$$

where the $f_i \in L^2(\Omega)$ are given functions. The weak formulation is obtained by $a(q,u)(\varphi) := (\nabla u, \nabla\varphi) - \sum_{i=1}^k q_i(f_i,\varphi)$ with state space $X = W_0^{1,p}(\Omega)$.

The first property that has to be ensured for a meaningful formulation of a problem of class (3.1) is the unique solvability of the state equation with the necessary regularity. For a large class of semilinear problems the proof will be given here.

Example 3.2. Let $\Omega \subset \mathbb{R}^2$ be a polygonal Lipschitz domain, its boundary seperated into $\Gamma = \Gamma_1 \cup \Gamma_2$ with $|\Gamma_1| > 0$. Let a linear and continous operator $B\colon Q \to L^2(\Omega)$ be given, and the differential operator

$$\bar{A}u(x) = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u(x)\right) \tag{3.6}$$

where the coefficients a_{ij} can be arranged in a symmetric matrix $A(x) = (a_{ij}(x))$ with the entries $a_{ij} \in L^\infty(\Omega)$ satisfying for some $\alpha_0 > 0$ the condition

$$\sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \geq \alpha_0|\xi|^2 \quad \forall \xi \in \mathbb{R}^2 \text{ and a.e. in } \Omega.$$

Denote by $\partial_{\nu_{\bar{\mathcal{A}}}}(x)$ the conormal derivative to the operator $\bar{\mathcal{A}}$ defined for $x \in \Gamma$ as the directional derivative in the direction $\nu_{\bar{\mathcal{A}}}(x) := A(x) \cdot n(x)$. Further, let the given functions d and b describing the nonlinearity be measurable with respect to the first argument, and $d(x,\cdot)$ and $b(x,\cdot)$ monotone increasing and three times differentiable on $\mathbb{R}$ with respect to the second argument for each fixed $x \in \Omega$ or $x \in \Gamma_2$ respectively. Furthermore, b and d are assumed to be bounded of order two with respect to the first variable, this means there exists a constant $K > 0$ such that

$$|d(x,0)| + |d_u(x,0)| + |d_{uu}(x,0)| \leq K \quad \text{a.e. in } \Omega \tag{3.7}$$

and analog for b on all spatial points of Γ_2. The semilinear elliptic state equation is then given as

$$\begin{aligned} \bar{\mathcal{A}}u(x) + d(x,u(x)) &= (Bq)(x) && \forall x \in \Omega, \\ u(x) &= 0 && \forall x \in \Gamma_1, \\ \partial_{\nu_{\mathcal{A}}} u(x) + b(x,u(x)) &= 0 && \forall x \in \Gamma_2. \end{aligned} \tag{3.8}$$

The mapping to an $L^2(\Omega)$-function on the right-hand side of (3.8) allows for several types of control to be realized in this problem class. Possibilities include

- parameter control, by chosing $R = Q = \mathbb{R}^k$, the operator $Bq := \sum_{i=1}^{k} q_i b_i$ with some given functions $b_i \in L^2(\Omega)$, and
- distributed control, by chosing $R = Q = L^2(\Omega), B = \mathrm{id}$.

The semilinear state equation can be expressed in the standard notation by the semilinear form

$$a(q,u)(\varphi) = (A\nabla u, \nabla\varphi) + (d(\cdot,u),\varphi) + \langle b(\cdot,u),\varphi\rangle_{\Gamma_2} - (Bq,\varphi) \tag{3.9}$$

and the choice of the right-hand side $f = 0$ in (3.4).

Lemma 3.1. *In the setting of Example 3.2, for every $q \in Q$ there exists a unique weak solution $u \in V := H^1_{\Gamma_1}(\Omega)$ of the state equation. Moreover, there holds $u \in W^{1,p}(\Omega)$ for some $p > 2$.*

Proof. Unique existence in $H^1_{\Gamma_1}(\Omega)$ follows by standard arguments for monotone operators. Next, by following the steps in [92, Theorem 4.7, 4.8], we prove that $u \in C(\bar{\Omega})$. It remains to prove $u \in W^{1,p}(\Omega)$.
The solution u fulfills the linear elliptic equation

$$\begin{cases} \bar{\mathcal{A}}u(x) = \bar{f}(x) & \text{in } \Omega, \\ u(x) = 0 & \text{on } \Gamma_1, \\ \partial_{\nu_{\mathcal{A}}} u(x) = g(x) & \text{on } \Gamma_2, \end{cases}$$

where $\bar{f}(x) = Bq(x) - d(x, u(x))$ and $g(x) = -b(x, u(x))$. By the properties of the nonlinearity functions b and d and the continuity of u we obtain $\bar{f} \in L^2(\Omega)$. Using a trace theorem we get that $u \in H^{\frac{1}{2}}(\Gamma_2) \cap C(\bar{\Gamma}_2)$. Then, we obtain due to the Lipschitz-continuity of $b(\cdot, \cdot)$ with respect to the second argument that $g \in H^{\frac{1}{2}}(\Gamma_2)$. This implies by [39, Theorem 4.4.4.13, Corollary 4.4.4.14] that for all $s < 2$

$$u - \sum c_i \psi_i \in W^{2,s}(\Omega),$$

where $c_i \in \mathbb{R}$ and the ψ_i are functions describing the singular behaviour of u at the corners of the domain Ω. It can be directly checked, that $\psi_i \in W^{1,p}(\Omega)$ holds with some $p > 2$. This, together with the fact that $W^{2,s}(\Omega) \hookrightarrow W^{1,p}(\Omega)$, completes the proof. □

This ensures the well-definedness of the control-to-state operator S. The operator is also known to be twice continously Fréchet differentiable, which can be shown as in [92].

3.2. Optimality conditions

In Theorem 2.12 conditions were formulated for the existence of an optimal solution. Given a concrete optimal control problem, it can usually not immediately be checked whether the involved assumptions hold, especially Assumption 2.9 on the state equation. For an example problem class, the necessary steps will be proven here. Therefore, the semilinear state equation (3.8) from the last section is considered, only for simplicity of notation with a simpler boundary condition. To set up the optimal control problem, a cost functional and a state constraint are considered that fulfill the assumptions of Theorem 2.12. For this, a function φ is introduced which enters the state cost part of the cost functional, and which fulfills the following conditions:

Assumption 3.2. *Let $\varphi\colon \Omega \times \mathbb{R}$, $(x, u(x)) \mapsto \varphi(x, u(x))$ be a function that is nonnegative, measurable with respect to the spatial variable x for every real u and twice differentiable with respect to u for almost all $x \in \Omega$. Let φ fulfill the boundedness condition of order 2 analog to* (3.7) *and the local Lipschitz condition*

$$\exists L(M) : |\varphi_y(x, y_1) - \varphi_y(x, y_2)| \leq L(M)|y_1 - y_2| \qquad \text{a.e. in } \Omega, \forall y_1, y_2 \in [-M, M].$$

Theorem 3.3. *Consider the problem*

$$\begin{aligned} \min J(q, u) := \int_\Omega \varphi(x, u(x))\, dx + \tfrac{\alpha}{2}\|q\|_Q^2 \qquad & q \in Q = L^2(\Omega), u \in X = W^{1,p}(\Omega) \\ \bar{\mathcal{A}}u(x) + d(x, u(x)) = q(x) \quad & \text{in } \Omega, \\ u(x) = 0 \quad & \text{on } \Gamma_1, \\ \partial_{\nu_{\mathcal{A}}} u(x) = 0 \quad & \text{on } \Gamma_2. \\ G(u) \geq 0 \quad & \text{in } \Omega, \end{aligned} \tag{3.10}$$

with the quantities $\bar{\mathcal{A}}, d, \Gamma_1, \Gamma_2$ fulfilling the same assumptions as in Lemma 3.1. The function φ is assumed to have the properties according to Assumption 3.2, also $\alpha > 0$. The admissible set X_{ad} induced by the constraint function G is assumed to be closed in X and fulfill Assumptions 2.8 and 2.11. Then, problem (3.10) *admits an optimal solution.*

Proof. Due to Lemma 3.1, $S\colon Q \to X$ is well-defined. As in Theorem 2.12, the boundedness of J implies the existence of

$$\bar{j} := \inf_{q\colon S(q)\in X_{ad}} J(q, S(q)),$$

which in turn gives a sequence (q_n) with $j(q_n) \to \bar{j}$. Due to the regularization term $\frac{\alpha}{2}\|q\|_Q^2$ this sequence must be bounded by some constant,

$$\|q_n\|_Q < K \quad \forall n > n_0,$$

implying the existence of a weakly convergent subsequence, which is again denoted by q_n, so that $q_n \rightharpoonup \bar{q}$. Setting $u_n := S(q_n)$, the maximum-norm a priori estimation [92, Theorem 4.8],

$$\|S(q)\|_{L^\infty(\Omega)} \le c_S(\|q\|_{L^2(\Omega)} + 1),$$

gives a bound for $\|u_n\|_{L^\infty(\Omega)} \le M = c_S(K+1) \quad \forall n > n_0$. Now, consider $z_n := d(x, u_n(x))$. Due to the properties of d, the z_n are bounded in $L^\infty(\Omega)$ too, see [92, p. 156]. Thus the z_n are bounded in $L^2(\Omega)$ as well, so we can choose a weakly convergent subsequence, w.l.o.g. again denoted by (z_n), $z_n \rightharpoonup \bar{z}$ in $L^2(\Omega)$. Thus the u_n fulfill the equation

$$\begin{aligned} \bar{\mathcal{A}} u_n &= q_n - z_n \text{ in } \Omega \\ u_n &= 0 \text{ on } \Gamma_1, \\ \partial_{\nu_{\mathcal{A}}} u_n &= 0 \text{ on } \Gamma_2. \end{aligned}$$

with the right hand side $q_n - z_n$ converging weakly in $L^2(\Omega)$ to $\bar{q} - \bar{z}$. Since this equation is linear, it is known that its solution operator is linear and continous from $L^2(\Omega)$ to $H^1(\Omega)$. As a linear operator the solution operator is weakly continous, which yields the convergence

$$u_n \rightharpoonup \bar{u} \qquad \text{in } H^1(\Omega),$$

and since $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$ also

$$u_n \to \bar{u} \qquad \text{in } L^2(\Omega).$$

Note, that in contrast to the proof of Theorem 2.12 we do not know yet that $\bar{u} = S(\bar{q})$. But now, including the boundedness of u_n in $C(\bar{\Omega})$ [92, Lemma 4.9] proves that

$$\|d(\cdot, u_n) - d(\cdot, \bar{u})\|_{L^2(\Omega)} \le L(M)\|u_n - \bar{u}\|_{L^2(\Omega)},$$

such that

$$d(\cdot, u_n) \to d(\cdot, \bar{u}) \qquad \text{in } L^2(\Omega).$$

Considering the weak form of the state equation,

$$\int_\Omega \sum_{i,j=1}^{2} \left(a_{ij} \frac{\partial}{\partial x_j} u_n \right) \frac{\partial}{\partial x_i} v \,\mathrm{d}x + \int_\Omega d(\cdot, u_n) v \,\mathrm{d}x = \int_\Omega q_n v \,\mathrm{d}x,$$

for any $v \in H^1_{\Gamma_1}(\Omega)$ the proven properties $u_n \rightharpoonup \bar{u}$ in $H^1(\Omega)$, $u_n \to \bar{u}$ in $L^2(\Omega)$, $\|u_n\|_{L^\infty(\Omega)} \le M$ allow to take the single expressions to the limit to conclude

$$\int_\Omega \sum_{i,j=1}^{2} \left(a_{ij} \frac{\partial}{\partial x_j} \bar{u} \right) \frac{\partial}{\partial x_i} v \,\mathrm{d}x + \int_\Omega d(\cdot, \bar{u}) v \,\mathrm{d}x = \int_\Omega \bar{q} v \,\mathrm{d}x,$$

which means $\bar{u} = S(\bar{q})$. As in Theorem 2.12 the closedness in X and convexity of X_{ad} secures $\bar{u} \in X_{ad}$. Finally due to the properties of φ the functional $u \mapsto \int_\Omega \varphi(x, u(x))$ is Lipschitz-continous on the set of all $u \in L^2(\Omega)$ with $\|u\|_{L^\infty(\Omega)} \leq M$, see again [92, Lemma 4.9], which in turn secures that $J(\bar{q}, \bar{u}) = \bar{j}$ and concludes the proof. □

We will now go on to characterize local optima of the general problem class (3.1). The general derivation has been done in Lemma 2.18 already, but for a better overview the conditions are stated again specially for the elliptic case and with minimum preconditions. Remember the Lagrange functional used to formulate the KKT conditions is defined on $\mathcal{L}\colon Q \times X \times Z \times \mathcal{M}(\Omega) \to \mathbb{R}$ by

$$\mathcal{L}(q, u, z, \mu) := J(q, u) - a(q, u)(z) + (f, z) - \langle \mu, G(u) \rangle. \tag{3.11}$$

The optimality conditions are as follows:

Theorem 3.4. *Consider the problem* (3.1)*. Let S and G be one time Fréchet differentiable, and Assumptions 2.11 and 2.17 be fulfilled. Let the point $(\bar{q}, \bar{u}) \in Q \times X$ be locally optimal for the problem* (3.1)*. Then there exist an adjoint state $\bar{z} \in Z$ and a Lagrangian multiplier $\bar{\mu} \in \mathcal{M}(\Omega)$ so that the following optimality system holds for $\bar{x} = (\bar{q}, \bar{u}, \bar{z}, \bar{\mu})$:*

$$\mathcal{L}'_z(\bar{x})(\varphi) = 0 \qquad \forall \varphi \in Z \tag{3.12a}$$
$$\mathcal{L}'_u(\bar{x})(\varphi) = 0 \qquad \forall \varphi \in X \tag{3.12b}$$
$$\mathcal{L}'_q(\bar{x})(\xi) = 0 \qquad \forall \xi \in Q \tag{3.12c}$$
$$\langle \bar{\mu}, G(\bar{u}) \rangle = 0 \text{ and } \quad \bar{\mu} \geq 0. \tag{3.12d}$$

Consider the equations from Theorem 3.4 in detail. Writing (3.12a) in an explicit way yields the state equation in weak form again,

$$a(\bar{q}, \bar{u})(\varphi) = (f, \varphi) \quad \forall \varphi \in Z. \tag{3.13}$$

Concerning condition (3.12b), the *adjoint equation*, the explicit formulation is given by

$$a'_u(\bar{q}, \bar{u})(\varphi, \bar{z}) = J'_u(\bar{q}, \bar{u})(\varphi) - \langle \bar{\mu}, G'(\bar{u})\varphi \rangle \quad \forall \varphi \in X, \tag{3.14}$$

The adjoint equation is central to the theory of state constrained optimal control problems. Since $\bar{\mu}$ is in general a Borel measure, this equation dictates the low regularity of the adjoint state. This, in turn, makes the full regularity of $X \subset W^{1,p}(\Omega)$ necessary for the test functions; they can in general not be chosen from a larger set. This point is illustrated in the following example.

Example 3.3. Consider the linear-quadratic distributed optimal control problem with given functions $u_d, f \in L^2(\Omega)$, $u_b \in C(\bar{\Omega})$ with $u_b > 0$:

$$\begin{aligned} \min J(q, u) = \tfrac{1}{2}\|u - u_d\|^2_{L^2(\Omega)} + \tfrac{\alpha}{2}\|q\|^2_{L^2(\Omega)}, &\quad q \in Q = L^2(\Omega), u \in X = W^{1,p}_0(\Omega) \\ a(q, u)(\varphi) := (\nabla u, \nabla \varphi) - (q, \varphi) = (f, \varphi) &\quad \forall \varphi \in X \\ G(u) := u_b - u \geq 0 & \end{aligned}$$

The adjoint equation in weak form according to (3.14) then reads: Find a $z \in Z$ such that

$$(\nabla z, \nabla \varphi) = (u - u_d, \varphi) + \langle \mu, \varphi \rangle \quad \forall \varphi \in X.$$

In strong form, this can formally be written as

$$\begin{aligned} -\Delta z &= u - u_d + \mu \quad \text{in } \Omega \\ z|_\Gamma &= 0 \end{aligned} \tag{3.15}$$

which is a Poisson equation for z with a right-hand side that is not in $H^{-1}(\Omega)$. It can be proven that the solution has the regularity $z \in W^{1,p'}(\Omega)$ for all $p' < \frac{n}{n-1}$, see [18]. Thus for the state the regularity $u \in W^{1,p}(\Omega)$ is required with some $p > n$ to guarantee that the term

$$\int_\Omega \nabla u(x) \cdot \nabla z(x) \, \mathrm{d}x,$$

contained in the Lagrange functional, is well-defined.

The explicit formulation of condition (3.12c) gives the *gradient equation*

$$J_q'(\bar{q}, \bar{u})(\xi) = a_q'(\bar{q}, \bar{u})(\xi, \bar{z}) \quad \forall \xi \in Q. \tag{3.16}$$

The conditions (3.12d) can be expressed equivalently by the variational inequality

$$\langle \bar{\mu}, \varphi - G(\bar{u}) \rangle \geq 0 \qquad \forall \varphi \in C(\bar{\Omega}), \varphi \geq 0. \tag{3.17}$$

3.3. Finite element discretization

Next the discretizations used for the elliptic problem are described. Here we have two levels of discretization, the spatial discretization indicated by the subscript h, and the disretization of the control space, indicated by the subscript d, such that $\sigma = (h, d)$.

The discretization of the spatial state variable is done using a continous Galerkin finite element method of order s, with $s \in \mathbb{N}$, $s \geq 1$, in short cG(s). The discretization of the control variable has to be kept more abstract since different structures of Q are possible, a few examples for typical situations will be discussed.

3.3.1. Discretization of the state variable

Concerning the state variable, the discretization of the state space is described by a mesh on the computational domain Ω. Let us assume here that Ω is indeed polygonal - otherwise a polygonal approximation Ω_h would need to be considered that approaches Ω in the refinement limit $h \to 0$. Details can be found, e.g., in [16].

The mesh on the considered discretization level is denoted by $\mathcal{T}_h$, and is composed of cells K. These are the spatial domains of the finite elements. We use (nondegenerate) quadrilaterals

in twodimensional, and hexahedrals in threedimensional domains. The vertices of all cells - counted only once if several cells share one vertex - are also called nodes, making up the set $\mathcal{N}_h$, and their number is denoted by N_h. We denote the diameter of each cell K by h_K and set the diameter of the mesh as the function

$$h\colon \bar{\Omega} \to \mathbb{R}, \quad x \mapsto h_K \text{ if } x \in K. \tag{3.18}$$

We will now describe what properties of the triangulation we expect. The property of *regularity* of the mesh $\mathcal{T}_h$ means

- domain exploitation: $\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} \bar{K}$
- void of overlaps: $\forall K_1, K_2 \in \mathcal{T}_h : K_1 \cap K_2 = \emptyset \quad \Leftrightarrow \quad K_1 \neq K_2$
- face adaption: $\forall K \in \mathcal{T}_h$: every face of K is either a subset of the boundary Γ or equal to a face of a different cell.

To ease the construction of the intended local refinement, we are not demanding regularity by the strict definition above, but one exception is made: for every face of a cell, we will allow for a minimum number of hanging nodes. In 2D this is one hanging node, which has to be in the midpoint of the face. In 3D, the construction requires five hanging nodes, one in the midpoint of the face plus one in the midpoint of each of the four edges. The consequence of this is that faces with hanging nodes are equal to the faces of two (in 2D) or four (in 3D) neighboring cells of equal size. An example configuration can be seen in Figure 3.1. A further

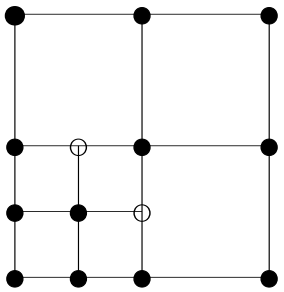


Figure 3.1.. Mesh structure - regular nodes (filled) and hanging nodes (empty) in a 2D mesh

demand that is added: every spatial mesh has to obey a patch-wise structure. That means that the mesh $\mathcal{T}_h$ can be interpreted as the global refinement of a coarser mesh $\mathcal{T}_{2h}$. In other words, any cell together with three (in 2D) or seven (in 3D) neighboring cells forms a patch which is the common coarser cell from $\mathcal{T}_{2h}$. This property will be utilized in the construction of computable error indicators. The mesh in Figure 3.1 does not have this property, but the one in Figure 3.2 has. Next we introduce the basis functions in every cell used to define the finite element space. The functions to build a finite element of order s on the cell $K \in \mathcal{T}_h$ are obtained by a transformation from the reference cell $\hat{K} = (0,1)^n$. Since K is nondegenerate there exists an affine bilinear transformation function $T_K : \hat{K} \to K$. The space of functions

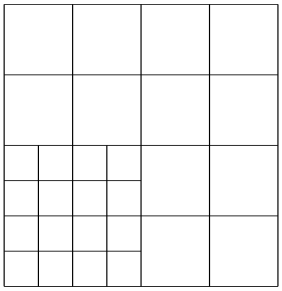


Figure 3.2.. Mesh structure - patched mesh in 2D

on the reference cell for a Lagrange element of order s is the space of polynomials with each coordinate up to power s,

$$\mathcal{Q}_s(\hat{K}) := \operatorname{span}\left\{\prod_{i=1}^{n} x_i^{\alpha_i} \,|\, \alpha_i \in \{0,1,\ldots,s\}\right\}.$$

The transformation then yields

$$\mathcal{Q}_s(K) := \left\{v_h : K \to \mathbb{R} \,|\, v_h \circ T_K \in \mathcal{Q}_s(\hat{K})\right\}.$$

as the set of FEM functions on the cell K. This finally gives the FE space

$$X_h^s := \{v_h \in V \cap C(\bar{\Omega}) \,|\, v_h|_K \in \mathcal{Q}_s(K) \quad \forall K \in \mathcal{T}_h\}. \tag{3.19}$$

Note that the function value of finite element functions in hanging nodes is determined by point-wise interpolation. Hanging nodes thus do not carry a degree of freedom, and are not accounted for in the set $\mathcal{N}_h$. Prescribing the value this way secures global continuity. With the definition of X_h^s the semidiscrete state equation can be formulated as

$$a(q, u_h)(\varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in X_h^s. \tag{3.20}$$

Its solution operator is denoted by $S_h \colon Q \to X_h^s$.

As the intention is the approximate solution on a sequence of refined meshes $(\mathcal{T}^{(i)}), \quad i = 1, 2 \ldots,$ some remarks on the mesh refinement process $\mathcal{T}^{(i)} \to \mathcal{T}^{(i+1)}$ are made. An important property of the refinement process is obviously that it preserves all the desired properties of the previous mesh. Assume we are given one mesh and one set of cells $\tilde{\mathcal{T}}_h \subset \mathcal{T}_h$ marked for refinement by the evaluation of an error estimator. A refinement of one cell means replacing it with four (in 2D) or eight (in 3D) cells of half the dimensions of the old one. This creates one regular node (in the midpoint of the old cell) and four (in 2D) or 18 (in 3D) nodes that may be hanging nodes or regular ones depending on the neighboring cells.
In general it does not suffice to refine only the cells marked by the error estimator, as this would violate some of the conditions posed above. Together with one cell marked for refinement all the cells from the same patch need to be refined in order to keep the patch structure. Also, to avoid multiple hanging nodes per face, we can not allow for neighboring cells of more of

one size level difference. Should this occur during refinement, the larger cell has to be marked for refinement additionally, and consequently its whole patch. This process must be repeated until every face has at most one hanging node.

Furthermore let us stress the fact that this way of refinement leads a to quasi-uniform family of meshes $(\mathcal{T}^{(i)})$. Remember, such a family of quadrilaterals is called *shape-regular*, if there exists a constant $\kappa > 0$ such that

$$h_K \leq \kappa h'_K \qquad \forall K \in \mathcal{T}^{(i)} \quad \forall i = 1, 2, \ldots$$

where h'_K denotes the smallest diameter of any side (in 2D) or face (in 3D) of K. However, in the context of a posteriori error estimation and adaptivity it is not necessary to demand $(\mathcal{T}^{(i)})$ to be quasi-uniform, this means families where there exists a $\bar{\kappa}$, such that

$$h \leq \bar{\kappa} h'_K \qquad \forall K \in \mathcal{T}^{(i)} \quad \forall i = 1, 2, \ldots,$$

as we could easily keep refining the mesh in one subdomain, and leave it unrefined in another, causing the ratio of largest and smallest cell diameter to grow arbitrarily.

3.3.2. Discretization of Lagrange multiplier and state constraint

The spatial discretization, described by the mesh $\mathcal{T}_h$, can also be used to motivate a discretization of the space $\mathcal{M}(\Omega)$ and of the state constraint. Let δ_{x_i} denote the Dirac measure concentrated at the node x_i. We then define the discrete multiplier space as

$$\mathcal{M}_h := \left\{ \mu_h = \sum_{i=1}^{N_h} \mu_i \delta_{x_i} \colon \mu_i \in \mathbb{R}, x_i \in \mathcal{N}_h \right\}. \tag{3.21}$$

For a discrete multiplier the positivity can be easily checked by

$$\mathcal{M}_h \ni \mu_h \geq 0 \quad \Leftrightarrow \quad \mu_i \geq 0 \quad \forall i \in \{0, 1, \ldots, N_h\}. \tag{3.22}$$

Further a discretization of the constraint $G(x, u(x)) \geq 0$ is necessary, since it must be fulfilled in infinitely many points. In some common cases this constraint has an equivalent formulation in finitely many points, in general an approximation needs to be introduced by

$$G_h(x_i, u_h(x_i)) \geq 0 \quad \forall x_i \in \mathcal{N}_h \tag{3.23}$$

with an appropriately chosen function $G_h \colon \bar{\Omega} \times \mathbb{R} \to \mathbb{R}$.

Consider the special case of a one-sided state constraint, e.g. $G(x, u(x)) = u_b(x) - u(x)$, a discretization of the state variable with linear or bilinear finite elements, and let the upper boundary u_b be in this FE space, or simply be a constant function. Using the coordinates of u_h, $u_h = \sum u_i \varphi_i$, and u_b in the nodal basis, $u_b = \sum u_{bi} \varphi_i$, it can be easily shown that the equivalence

$$u_h(x) \leq u_b(x) \quad \forall x \in \bar{\Omega} \qquad \Leftrightarrow \qquad u_h(x_i) \leq u_b(x_i) \quad \forall x_i \in \mathcal{N}_h$$

holds true. This means we can set $G_h = G$ and write $G_h(u_h) \geq 0$ as abbreviation for (3.23) like before.

Had u_b not been an FE function, a possible approximation would have been the use of the function $u_{bh} := \sum u_b(x_i)\varphi_i$ in the definition

$$G_h(x_i, u_h(x_i)) = u_{bh}(x_i) - u_h(x_i) \tag{3.24}$$

of the discretized constraint function.

Another example for the discretization of the state constraint, this time for $\mathcal{Q}_2$-elements, has been done in [24]. In the general case with abstract G we assume an appropriate G_h can be found.

The discrete admissible set can now be defined as

$$X_{ad,h} := \{u \in X_h^s \colon G_h(x_i, u_h(x_i)) \geq 0 \, \forall x_i \in \mathcal{N}_h\},$$

and the spatially discretized elliptic control problem reads

$$(P_h) \begin{cases} \min J(q, u_h) & q \in Q, u_h \in X_h^s \\ a(q, u_h)(\varphi_h) = (f, \varphi_h) & \forall \varphi_h \in X_h^s \\ G_h(u_h) \geq 0 & . \end{cases} \tag{3.25}$$

3.3.3. Discretization of the control variable

The discretization of the control variable, or the choice of a finite dimensional subspace

$$Q_d \subset Q \tag{3.26}$$

can not be described in such a detailed manner, as Q is an abstract space. The general case has to be left to the user. For the examples considered in Section 3.1, possibilities are discussed here:

Example 3.4. For distributed control, as introduced in Example 3.1, no. 1, the space $Q = L^2(\Omega)$ can either be discretized like the state space by a cG(s) method. Alternatively cellwise constant functions can be employed, induced by the same mesh. This would mean setting

$$Q_d = \{v \in Q : v|_K = const \quad \forall K \in \mathcal{T}_h\}$$

and is a dG(0) method. Other choices are possible if a specific problem suggests it, for example a different mesh could be used in the discretization process. But for the numerical examples in this thesis only one mesh is used for the discretization of both state and control variable.

Example 3.5. For Neumann control, as in Example 3.1, no. 2, also the mesh $\mathcal{T}_h$ can be used to induce a discretization of $Q = L^2(\Gamma)$: A cG(s_d) finite element space on Ω is set up as described before, and the traces of those functions on the boundary make up the discrete control space:

$$Q_d = \left\{\gamma(v_h) \in C(\Gamma) : v_h \in V_h^{s_d}\right\}.$$

For Dirichlet control the discussion is more involved, as there are several possibilities to establish a weak formulation and choose an appropriate control space to begin with, see, e.g., [60] for a discussion of this. Also specially designed boundary element methods can be used, see [78].

In the case of parameter control as in Example 3.1, no. 3, Q is already finite to begin with, so it suffices to set $Q_d = Q$.

Remark 3.1. For certain optimal control problems there also exist solution techniques that require no discretization of the control. For this so called variational discretization concept, see [51] for elliptic, and [27] for parabolic problems.

3.3.4. Discrete optimality conditions

Employing the discrete spaces defined before, and using the combination of subscripts $\sigma = (h, d)$, the fully discrete problem is introduced as

$$(P_\sigma) \begin{cases} \min J(q_\sigma, u_\sigma) & q_\sigma \in Q_d, u_\sigma \in X_h^s \\ a(q_\sigma, u_\sigma)(\varphi_\sigma) = (f, \varphi_\sigma) & \forall \varphi_\sigma \in X_h^s \\ G_\sigma(u_\sigma) \geq 0 & . \end{cases} \tag{3.27}$$

Analog to the continous problem, we need the following assumption for the proof of existence of an optimal solution:

Assumption 3.5. *There exists a control $q_d^* \in Q_d$ such that $S_h(q_d^*) \in X_{ad,h}$.*

In some situations the assumption can be proven for h small enough, see [70].

Theorem 3.6. *Consider problem* (3.27)*, and let Assumption 3.5 hold. Then, there exists an optimal control $\bar{q}_\sigma$.*

The proof can be done like in the continous case. Again we need a local Slater condition to be fulfilled.

Assumption 3.7. *Let $\bar{q}_\sigma$ denote a locally optimal control. There exists a control $\hat{q}_d \in Q_d$ such that $S_h(\bar{q}_\sigma) + S_h'(\bar{q}_\sigma)(\hat{q}_d - \bar{q}_\sigma) \in \operatorname{int}(X_{ad,h})$.*

The optimality conditions can be formulated using the Lagrangian $\mathcal{L}$ as follows:

Theorem 3.8. *Let $(\bar{q}_\sigma, \bar{u}_\sigma)$ be locally optimal for the discrete problem* (3.27)*. Then there exist an adjoint state $\bar{z}_\sigma \in X_h^s$ and a discrete multiplier $\bar{\mu}_\sigma \in \mathcal{M}_h$ such that the following condition holds in the point $\bar{x}_\sigma = (\bar{q}_\sigma, \bar{u}_\sigma, \bar{z}_\sigma, \bar{\mu}_\sigma) \in Q_d \times X_h^s \times X_h^s \times \mathcal{M}_h$:*

$$\mathcal{L}_z'(\bar{x}_\sigma)(\varphi_\sigma) = 0 \quad \forall \varphi_\sigma \in X_h^s \tag{3.28a}$$

$$\mathcal{L}_u'(\bar{x}_\sigma)(\varphi_\sigma) = 0 \quad \forall \varphi_\sigma \in X_h^s \tag{3.28b}$$

$$\mathcal{L}_q'(\bar{x}_\sigma)(\xi_\sigma) = 0 \quad \forall \xi_\sigma \in Q_d \tag{3.28c}$$

$$\langle \bar{\mu}_\sigma, G_\sigma(\bar{u}_\sigma) \rangle = 0, \quad \bar{\mu}_\sigma \geq 0 \tag{3.28d}$$

The proof is analog to the continous case. Again we write the equations from Theorem 3.8 in explicit form. We obtain the discrete state equation

$$a(\bar{q}_\sigma, \bar{u}_\sigma)(\varphi_\sigma) = (f, \varphi_\sigma) \quad \forall \varphi_\sigma \in X_h^s, \tag{3.29}$$

the discrete adjoint equation

$$a'_u(\bar{q}_\sigma, \bar{u}_\sigma)(\varphi_\sigma, \bar{z}_\sigma) = J'_u(\bar{q}_\sigma, \bar{u}_\sigma)(\varphi_\sigma) - \langle \bar{\mu}_\sigma, G'_\sigma(\bar{u}_\sigma)\varphi_\sigma \rangle \quad \forall \varphi_\sigma \in X_h^s, \tag{3.30}$$

and the discrete gradient equation

$$J'_q(\bar{q}_\sigma, \bar{u}_\sigma)(\xi_\sigma) = a'_q(\bar{q}_\sigma, \bar{u}_\sigma)(\xi_\sigma, \bar{z}_\sigma) \quad \forall \xi_\sigma \in Q_d. \tag{3.31}$$

3.4. Optimization with the primal-dual active set method

In order to find a solution algorithm for (3.27), note that (P_σ) describes a fairly large problem class. Thus it can not be expected that there exists a numerical method that solves all problem instances contained in (3.27) at all. Even if such a method exists, it cannot be expected that it solves all the problems equally well. Methods that utilize special features of a subclass of problems will usually do better.

In the upcoming section the method of direct treatment of the state constraints by the primal-dual active set (PDAS) method will be introduced. The use of this method is well established, but it is applicable only to a subset of the problems included in (3.27). A method that can be applied to the complete problem class is described in Section 3.6.

The primal-dual active set method is based on the partition of the set $\mathcal{N}_h$ into an *active* and an *inactive set*. If the active set of the optimal solution

$$A_{\text{exact}} := \{x_i \in \mathcal{N}_h \colon G_\sigma(\bar{u}_\sigma(x_i)) = 0\}. \tag{3.32}$$

would be known, then the optimal control could be determined by the solution of an equality-constrained optimal control problem. This corresponds to an optimal control problem on the inactive set

$$\mathcal{I}_{\text{exact}} = \mathcal{N}_h \setminus A_{\text{exact}}.$$

that can be solved with Newton-type methods. Naturally, A_{exact} is unknown to us. It is thus approximated by a sequence of sets $(A_i) \subset \mathcal{N}_h$, where A_0 is an arbitrary starting set, and the others are gained by the recursion of the following two steps:

- Given A_i, solve the following auxiliary problem

$$(P_E) \begin{cases} \min j_\sigma(q_\sigma), & q_\sigma \in Q_d \\ G_\sigma(S_\sigma(q_\sigma))|_{A_i} = 0 \end{cases}$$

  This is an optimal control problem with additional equality constraints in $|A_i|$ points with the explicit formulation

$$G_\sigma(x_j, S_\sigma(q_\sigma)(x_j)) = 0 \qquad \forall x_j \in A_i.$$

  The Lagrange multiplier associated with these constraints is denoted by $\mu^{i+1} \in \mathcal{M}_h$, the jth component μ_j^{i+1} corresponds to the point $x_j \in \mathcal{N}_h$.

- With the solution of (P_E), denoted by q_σ^{i+1}, and the according multiplier μ^{i+1}, the next active set A_{i+1} then corresponds to the state $u_\sigma^{i+1} = S_\sigma(q_\sigma^{i+1})$ and is given by

$$A_{i+1} := \{x_j \in \mathcal{N}_h : G_\sigma(x_j, u_\sigma^{i+1}(x_j)) + c \cdot \mu_j^{i+1} \leq 0\}, \tag{3.33}$$

  with some constant $c > 0$.

This iteration yields a sequence of sets A_i and controls q_σ^{i+1} .The method has converged when $A_i = A_{i+1}$ for some i.

The detailed explanation of the solution of (P_E) will be done below. Let us first address the point which problems of type (3.1) can be solved by the PDAS method. In this algorithm it is not inherently clear whether the first step, the solution of (P_E), is well-defined. For some types of state equations it might be impossible to find a control so that the corresponding state fulfills the constraint with equality on the prescribed set A_i. There are several examples of elliptic OCPs the PDAS method can be applied to, see e.g. [13, 15]. A sufficient condition for the well-definedness of the algorithm in the general framework (3.1) is obviously that the discrete control-to-state operator is surjective, or

$$S_\sigma(Q_d) = X_h^s.$$

The statement of a weaker condition, this means an a priori specification of the range of S_σ, is hardly possible even for a given realization of (P_σ). As a non-rigorous rule of thumb one may say though, that a larger control space leads to the permissibility of the PDAS algorithm more often than a smaller one. For the standard problems from Example 3.1, the rule of thumb favors distributed over boundary over parameter control. This limitation of the range of operation of the PDAS method is intrinsic to state constrained OCPs, in contrast to control constrained ones.

Next, the solution of (P_E) will be detailed. As it has equality constraints only, the algorithm will be built up with strategies similar to those from Section 2.3.4. Again, the solution $(q_\sigma^{i+1}, \mu^{i+1})$ is approximated by a sequence of controls and multipliers

$$(q_k^{i+1}, \mu_k^{i+1}), \quad k = 0, 1, \ldots. \tag{3.34}$$

As starting values the last values from the last PDAS step are chosen, $q_0^{i+1} := q^i, \mu_0^{i+1} := \mu^i$. Then it suffices to describe one step $k \to k+1$ in the sequence (3.34). For simplicity of notation, assume q_σ and μ_σ are the current iterates, and the current active set is denoted by A. The task is to find the update $(\delta q, \delta\mu)$ to advance in the sequence (3.34). Like before, the method is developed in reduced form, but now reduced to the control q_σ and the multiplier μ_σ. The state u_σ and adjoint state z_σ, are fixed as the solutions of the discrete state and adjoint equations (3.29) and (3.30), represented by the solution operators of these equations,

$$u_\sigma = S_\sigma(q_\sigma) \qquad \text{and} \qquad z_\sigma = T_\sigma(q_\sigma, \mu_\sigma).$$

For the equality constrained problem (P_E), denote its Lagrangian by

$$M(q_\sigma, \mu_\sigma) := j_\sigma(q_\sigma) - \langle \mu_\sigma, G_\sigma(u_\sigma) \rangle_A, \tag{3.35}$$

where $\langle \cdot, \cdot \rangle_A$ with $A \subset \mathcal{N}_h$ is defined for discrete measures of the form $\mu = \sum_{x_i \in \mathcal{N}_h} \mu_i \delta_{x_i}$ and functions $f \in C(\bar{\Omega})$ as

$$\langle \mu, f \rangle_A := \sum_{x_i \in A} \mu_i f(x_i).$$

The optimality conditions are as before

$$M'(q_\sigma, \mu_\sigma) = 0 \quad \Leftrightarrow \quad M'_q(q_\sigma, \mu_\sigma)(\delta q) = M'_\mu(q_\sigma, \mu_\sigma)(\delta \mu) = 0 \quad \forall \delta q \in Q_d, \delta \mu \in \mathcal{M}_h.$$

The evaluation of these directional derivatives for given directions $\delta q \in Q_d, \delta \mu \in \mathcal{M}_h$ is done as follows:

$$\begin{aligned} M'_q(q_\sigma, \mu_\sigma)(\delta q) &= j'(q_\sigma)(\delta q) - \langle \mu_\sigma, G'_\sigma(u_\sigma) S'_\sigma(q_\sigma) \delta q \rangle_A \\ &= J'_q(q_\sigma, u_\sigma)(\delta q) + J'_u(q_\sigma, u_\sigma)(\delta u) - \langle \mu_\sigma, G'_\sigma(u_\sigma) \delta u \rangle_A \\ &= J'_q(q_\sigma, u_\sigma)(\delta q) + a'_u(q_\sigma, u_\sigma)(\delta u, z_\sigma) \\ &= J'_q(q_\sigma, u_\sigma)(\delta q) - a'_q(q_\sigma, u_\sigma)(\delta q, z_\sigma), \end{aligned} \tag{3.36}$$

where, as before, $\delta u = S'_\sigma(q_\sigma)\delta q$ is given as solution of the discrete tangent equation

$$a'_u(q_\sigma, u_\sigma)(\delta u, \varphi_\sigma) = a'_q(q_\sigma, u_\sigma)(\delta q, \varphi_\sigma) \quad \forall \varphi_\sigma \in X_h^s, \tag{3.37}$$

which is obtained by total derivation of the discrete state equation.
The other directional derivative is

$$M'_\mu(q_\sigma, \mu_\sigma)(\delta \mu) = -\langle \delta \mu, G_\sigma(u_\sigma) \rangle_A. \tag{3.38}$$

Like before, the equation $M'(q_\sigma, \mu_\sigma) = 0$ will be solved using a Newton-type method. The necessary second derivatives are evaluated as follows:

$$\begin{aligned} M''_{qq}(q_\sigma, \mu_\sigma)(\delta q, \tau q) &= \frac{\partial}{\partial q}(J'_q(q_\sigma, u_\sigma)(\tau q) - a'_q(q_\sigma, u_\sigma)(\tau q, z_\sigma))(\delta q) \\ &= J''_{qq}(q_\sigma, u_\sigma)(\delta q, \tau q) + J''_{uq}(q_\sigma, u_\sigma)(\delta u, \tau q) - a''_{qq}(q_\sigma, u_\sigma)(\delta q, \tau q, z_\sigma) \\ &\quad - a''_{uq}(q_\sigma, u_\sigma)(\delta u, \tau q, z_\sigma) - a'_q(q_\sigma, u_\sigma)(\tau q, T'_{\sigma,q}(q_\sigma, \mu_\sigma)\delta q), \\ M''_{\mu q}(q_\sigma, \mu_\sigma)(\delta \mu, \tau q) &= - a'_q(q_\sigma, u_\sigma)(\tau q, T'_{\sigma,\mu}(q_\sigma, \mu_\sigma)\delta \mu), \\ M''_{q\mu}(q_\sigma, \mu_\sigma)(\delta q, \tau \mu) &= - \langle \tau \mu, G'_\sigma(u_\sigma)\delta u \rangle_A, \\ M''_{\mu\mu}(q_\sigma, \mu_\sigma)(\delta \mu, \tau \mu) &= 0. \end{aligned}$$

The two terms involving T'_σ are treated as follows: total derivation of the dual equation yields the term $T'_{\sigma,q}(q_\sigma, \mu_\sigma)\delta q + T'_{\sigma,\mu}(q_\sigma, \mu_\sigma)\delta \mu$. This motivates the definition of

$$\delta z := T'_{\sigma,q}(q_\sigma, \mu_\sigma)\delta q + T'_{\sigma,\mu}(q_\sigma, \mu_\sigma)\delta \mu,$$

which is obtained for given $\delta q, \delta u, \delta \mu$ as solution of the discrete additional adjoint equation:

$$\begin{aligned} &a'_u(q_\sigma, u_\sigma)(\varphi_\sigma, \delta z) = -a''_{uu}(q_\sigma, u_\sigma)(\delta u, \varphi_\sigma, z_\sigma) - a''_{qu}(q_\sigma, u_\sigma)(\delta q, \varphi_\sigma, z_\sigma) \\ &+ J''_{qu}(\delta q, \varphi_\sigma) + J''_{uu}(\delta u, \varphi_\sigma) - \langle \mu_\sigma, G''_\sigma(u_\sigma)(\delta u, \varphi_\sigma) \rangle - \langle \delta \mu, G'_\sigma(u_\sigma)(\varphi_\sigma) \rangle \quad \forall \varphi_\sigma \in X_h^s. \end{aligned} \tag{3.39}$$

The absolute second derivative can thus be evaluated as

$$\begin{aligned} M''(q_\sigma,\mu_\sigma)((\delta q,\delta\mu),(\tau q,\tau\mu)) = \quad & J''_{qq}(q_\sigma,u_\sigma)(\delta q,\tau q) + J''_{uq}(q_\sigma,u_\sigma)(\delta u,\tau q) \\ & - a''_{qq}(q_\sigma,u_\sigma)(\delta q,\tau q,z_\sigma) - a''_{uq}(q_\sigma,u_\sigma)(\delta u,\tau q,z_\sigma) \\ & - a'_q(q_\sigma,u_\sigma)(\tau q,\delta z) - \langle \tau\mu, G'_\sigma(u_\sigma)\delta u\rangle. \end{aligned} \tag{3.40}$$

This formulation is indeed favorable, since, like in Section 2.3.4, the repeated evaluation of this term for one given direction $(\delta q,\delta\mu)$ and many directions $(\tau q,\tau\mu)$ requires only the solution of two partial differential equations (assuming q,u,z are given):

- the tangent equation with δq to calculate δu,
- the additional adjoint equation with $\delta q,\delta u,\delta\mu$ to calculate δz.

The procedure of solving the equation $M'(q_\sigma,\mu_\sigma)=0$ can be done analog to Section 2.3.4. Given bases (τq_j), $j=1\dots\dim(Q_d)$ of Q_d and $(\tau\mu_j)$, $j=1\dots\dim(\mathcal{M}_h)$ of $\mathcal{M}_h$, the utilized directions are counted in this order:

$$(\tau q,\tau\mu)_j = \begin{cases} \tau q_j\colon & 1\le j\le \dim(Q_d) \\ \tau\mu_{j-\dim(Q_d)}\colon & \dim(Q_d)+1\le j\le \dim(Q_d)+\dim(\mathcal{M}_h) \end{cases}.$$

Thus the gradient ∇M is written as

$$\nabla M(q_\sigma,\mu_\sigma) = \sum_{j=1}^{\dim(Q_d)+\dim(\mathcal{M}_h)} f_j(\tau q,\tau\mu)_j,$$

where its coefficient vector $\mathbf{f}$ is determined by

$$(M'(q_\sigma,\mu_\sigma)((\tau q,\tau\mu)_i))_{i=1}^{\dim(Q_d)+\dim(\mathcal{M}_h)} = (\nabla M(q_\sigma,\mu_\sigma),(\tau q,\tau\mu)_i)_{i=1}^{\dim(Q_d)+\dim(\mathcal{M}_h)} = \mathbf{Gf} \tag{3.41}$$

with the Gramian matrix $\mathbf{G}$.
The full Newton step $(\delta q,\delta\mu)$, determined by $M''(q_\sigma,\mu_\sigma)((\delta q,\delta\mu),(\tau q,\tau\mu)) = -M'(q_\sigma,\mu_\sigma)(\tau q,\tau\mu)$ is represented by

$$(\delta q,\delta\mu) = \sum_{j=1}^{\dim(Q_d)+\dim(\mathcal{M}_h)} d_j(\tau q,\tau\mu)_j,$$

with its coefficient vector $\mathbf{d}$. Utilizing these quantities the formal buildup of the system of equations

$$\mathbf{Kd} = -\mathbf{Gf}$$

can be done as before, with $\mathbf{K}$ being the matrix with entries

$$M''(q_\sigma,\mu_\sigma)((\tau q,\tau\mu)_j,(\tau q,\tau\mu)_i)$$

at the (i,j)-th position. However due to its origin in the Lagrangian (3.35) the system matrix $\mathbf{H} := \mathbf{G}^{-1}\mathbf{K}$ is not positive definite, but exhibits a saddle point structure. The solution of the system of equations can be achieved, e.g., by a GMRes method, see [83]. Analog to the representation (2.63), products of the form $\nabla^2 M(q_\sigma,\mu_\sigma)(\delta q,\delta\mu)$ to be used within the GMRes method can be evaluated by

$$\mathbf{Gh} = (M''(q_\sigma,\mu_\sigma)(\delta q,\delta\mu)(\tau q,\tau\mu)_i)_{i=1}^{\dim(Q_d\times\mathcal{M}_h)} \tag{3.42}$$

such that $\mathbf{h}$ is the coefficient vector of the product. After the determination of $\mathbf{d}$ and thus $(\delta q, \delta\mu)$ one can use the full Newtin step $q_{i+1}^{k+1} = q_{i+1}^{k} + \delta q$ and $\mu_{i+1}^{k+1} = \mu_{i+1}^{k} + \delta\mu$, or it can be necessary to include a globalization technique, such as a line search method, to determine a $\lambda^i \in (0, 1]$ to set $q_{i+1}^{k+1} = q_{i+1}^{k} + \lambda^i \delta q$ and $\mu_{i+1}^{k+1} = \mu_{i+1}^{k} + \lambda^i \delta\mu$. An overview over the whole PDAS method for state constrained ellipctic OCPs is given in Algorithm 3.1. An analog formulation is possible for parabolic problems.

Algorithm 3.1. Primal-dual active set method for state constrained ellipctic OCPs

input data: control q^0, multiplier μ^0,
parameter: TOL_N, TOL_L
solve $u^0 = S_\sigma(q^0)$, $z^0 = T_\sigma(q^0, \mu^0)$
determine the active set A^0 by (3.33)
set $i = 0$
repeat
  Solve (P_E) $(q^i, u^i, z^i, \mu^i, A^i, TOL_N, TOL_L)$, see Algorithm 3.2.
  this yields $q^{i+1}, u^{i+1}, z^{i+1}, \mu^{i+1}$
  determine the active set A^{i+1} by (3.33)
  set i:=i+1
until $A^i = A^{i-1}$
output data: $\bar{q} := q^i, \bar{u} := u^i, \bar{z} := z^i, \bar{\mu} := \mu^i$

Algorithm 3.2. Newton-type optimization for PDAS

input data: current functions q^0, u^0, z^0, μ^0, active set A
parameter: TOL_N, TOL_L
Set counter $i = 0$.
repeat
  Compute $\mathbf{f}$ as vector representation of $\nabla M(q^i, \mu^i)$ by (3.41)
  Compute $\mathbf{d}^\top$ as vector representation of the Newton update $(\delta q, \delta\mu)^\top$
  by solving $\mathbf{H}\mathbf{d}^\top = -\mathbf{f}$ iteratively, e.g. by GMRes method with tolerance TOL_L
  for any product $\mathbf{H}\tilde{\mathbf{d}}$ the GMRes algorithm requests **do**
    With $(\widetilde{\delta q}, \widetilde{\delta\mu})$ being the direction represented by $\tilde{\mathbf{d}}$
    Compute $\widetilde{\delta u}$ by (3.37)
    Compute $\widetilde{\delta z}$ by (3.39)
    Evaluate (3.40) to get right hand side of (3.42)
    Get $\mathbf{h} = \mathbf{H}\tilde{\mathbf{d}}$ by solving (3.42)
  Determine step length λ^i by line search
  Set $q^{i+1} = q^i + \lambda^i \delta q$, $\mu^{i+1} = \mu^i + \lambda^i \delta\mu$
  Solve $u^{i+1} = S_\sigma(q^{i+1})$, $z^{i+1} = T_\sigma(q^{i+1}, \mu^{i+1})$
  $i = i + 1$
until $|\nabla M(q^i, \mu^i)| \leq TOL_N$
output data: q^i, u^i, z^i, μ^i

3.5. A posteriori error estimator and adaptivity

At this stage, given a mesh $\mathcal{T}_h$ and a control discretization Q_d, a discrete approximation (q_σ, u_σ) to a locally optimal solution $(\bar{q}, \bar{u})$ of (3.1) can be computed. We now turn to the subject of estimating the error that this approximation has caused in terms of the cost functional. Thus the aim is to derive an error estimator

$$\eta \approx J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma).$$

The first result represents the error in terms of derivatives of the Lagrangian. Note that in contrast to the derivation of the numerical solution algorithm before, it is now required that the control-to-state operator S is three times Gateaux differentiable.

Lemma 3.9. *Let $\bar{x} = (\bar{q}, \bar{u}, \bar{z}, \bar{\mu}) \in Q \times X \times Z \times \mathcal{M}(\Omega)$ be a point satisfying the first-order necessary optimality condition* (3.12)*, and let $x_\sigma = (q_\sigma, u_\sigma, z_\sigma, \mu_\sigma) \in Q_d \times X_h \times X_h \times \mathcal{M}_h$ be a discrete point satisfying the corresponding discrete optimality condition* (3.28) *with the Lagrange functional $\mathcal{L}$ being three times Gateaux differentiable. Then it holds for the discretization error with respect to the cost functional*

$$J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma) = \frac{1}{2}\mathcal{L}'(\bar{x})(\bar{x} - x_\sigma) + \frac{1}{2}\mathcal{L}'(x_\sigma)(\bar{x} - x_\sigma) + \langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle + \mathcal{R}, \quad (3.43)$$

where $\mathcal{R}$ is a term of third order, $\mathcal{R} = \mathcal{O}(\|\bar{x} - x_h\|^3)$.

Proof. For the points $\bar{x}$ and x_σ, the application of the respective optimality conditions to the definition of the Lagrangian (3.11) shows that there holds

$$\mathcal{L}(\bar{x}) = J(\bar{q}, \bar{u}) \quad \text{and} \quad \mathcal{L}(x_\sigma) = J(q_\sigma, u_\sigma) + \langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle. \quad (3.44)$$

Following the proof of the respective theorem in [10], an evaluation of the occuring integral with the trapezoidal rule, using the abbreviation $e := \bar{x} - x_\sigma$, yields

$$\begin{aligned} J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma) &= \mathcal{L}(\bar{x}) - \mathcal{L}(x_\sigma) + \langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle \\ &= \int_0^1 \mathcal{L}'(x_\sigma + se)(e)\,\mathrm{d}s + \langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle \\ &= \frac{1}{2}\mathcal{L}'(\bar{x})(\bar{x} - x_\sigma) + \frac{1}{2}\mathcal{L}'(x_\sigma)(\bar{x} - x_\sigma) + \langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle + \mathcal{R} \end{aligned}$$

where the remainder term is given as

$$\mathcal{R} = \frac{1}{2}\int_0^1 \mathcal{L}'''(x_\sigma + se)(e, e, e) \cdot s \cdot (s - 1)\,\mathrm{d}s.$$

□

The utilization of the continous and discrete optimality conditions gives the next step in the derivation.

Lemma 3.10. *In the situation of Lemma 3.9, there holds the error representation formula*

$$\begin{aligned} J(\bar{q},\bar{u}) - J(q_\sigma,u_\sigma) = \; & \frac{1}{2}\{J'_u(\bar{q},\bar{u})(\bar{u}-u_\sigma) - a'_u(\bar{q},\bar{u})(\bar{u}-u_\sigma,\bar{z}) - a(q_\sigma,u_\sigma)(\bar{z}-\tilde{z}_\sigma) \\ & +(f,\bar{z}-\tilde{z}_\sigma) + J'_q(q_\sigma,u_\sigma)(\bar{q}-\tilde{q}_\sigma) - a'_q(q_\sigma,u_\sigma)(\bar{q}-\tilde{q}_\sigma,z_\sigma) \\ & +J'_u(q_\sigma,u_\sigma)(\bar{u}-u_\sigma) - a'_u(q_\sigma,u_\sigma)(\bar{u}-u_\sigma,z_\sigma)\} + \mathcal{R} + \mathcal{R}_2 \\ & +\langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle \end{aligned} \tag{3.45}$$

where $\tilde{q}_\sigma \in Q_d$ and $\tilde{z}_\sigma \in X_h^s$ can be arbitrarily chosen, and $\mathcal{R}_2$ is a quadratic remainder term detailed below.

Proof. Starting from equation (3.43), the terms to be considered from the derivative of $\mathcal{L}$ in the continous optimal point $\bar{x}$ are:

$$\mathcal{L}'_z(\bar{x})(\bar{z}-z_\sigma) = 0 \qquad \text{due to optimality condition (3.12a)}, \tag{3.46a}$$
$$\mathcal{L}'_u(\bar{x})(\bar{u}-u_\sigma) = 0 \qquad \text{due to optimality condition (3.12b)}, \tag{3.46b}$$
$$\mathcal{L}'_q(\bar{x})(\bar{q}-q_\sigma) = 0 \qquad \text{due to optimality condition (3.12c)}, \tag{3.46c}$$
$$\mathcal{L}'_\mu(\bar{x})(\bar{\mu}-\mu_\sigma) = -\langle \bar{\mu}-\mu_\sigma, G(\bar{u})\rangle. \tag{3.46d}$$

For the discrete optimal point x_σ the following terms occur:

$$\mathcal{L}'_z(x_\sigma)(\bar{z}-z_\sigma) = -a(q_\sigma,u_\sigma)(\bar{z}-z_\sigma) + (f,\bar{z}-z_\sigma), \tag{3.47a}$$
$$\mathcal{L}'_q(x_\sigma)(\bar{q}-q_\sigma) = J'_q(q_\sigma,u_\sigma)(\bar{q}-q_\sigma) - a'_q(q_\sigma,u_\sigma)(\bar{q}-q_\sigma,z_\sigma), \tag{3.47b}$$
$$\mathcal{L}'_u(x_\sigma)(\bar{u}-u_\sigma) = J'_u(q_\sigma,u_\sigma)(\bar{u}-u_\sigma) - a'_u(q_\sigma,u_\sigma)(\bar{u}-u_\sigma,z_\sigma) - \langle \mu_\sigma, G'(u_\sigma)(\bar{u}-u_\sigma)\rangle, \tag{3.47c}$$
$$\mathcal{L}'_\mu(x_\sigma)(\bar{\mu}-\mu_\sigma) = -\langle \bar{\mu}-\mu_\sigma, G(u_\sigma)\rangle. \tag{3.47d}$$

Using the discrete state and gradient equations (3.29) and (3.31) in the right hand sides of (3.47a) and (3.47b) any arbitrary discrete functions $\tilde{z}_h \in X_h^s$, $\tilde{q}_d \in Q_d$ can be inserted:

$$\mathcal{L}'_z(x_\sigma)(\bar{z}-z_\sigma) = -a(q_\sigma,u_\sigma)(\bar{z}-\tilde{z}_\sigma) + (f,\bar{z}-\tilde{z}_\sigma), \tag{3.48a}$$
$$\mathcal{L}'_q(x_\sigma)(\bar{q}-q_\sigma) = J'_q(q_\sigma,u_\sigma)(\bar{q}-\tilde{q}_\sigma) - a'_q(q_\sigma,u_\sigma)(\bar{q}-\tilde{q}_\sigma,z_\sigma). \tag{3.48b}$$

Take a step back to get an overview over the terms that are summed up for the representation of $J(\bar{q},\bar{u}) - J(q_\sigma,u_\sigma)$ via (3.43). It comprises of the right hand sides of the equations (3.46a) through (3.46c), which are zero, and those of (3.46d), (3.48a), (3.48b), (3.47c), and (3.47d). All summands that do not involve any measures can be transferred straight to the claim of the lemma in (3.45). The terms with Lagrange multipliers are summed up and treated further. The following terms remain:

$$-\langle \mu_\sigma, G'(u_\sigma)(\bar{u}-u_\sigma)\rangle - \langle \bar{\mu}-\mu_\sigma, G(\bar{u})\rangle - \langle \bar{\mu}-\mu_\sigma, G(u_\sigma)\rangle \tag{3.49}$$

By using the complementarity conditions (3.12d) and (3.28d), and a Taylor expansion on two terms, the term (3.49) is transformed to

$$\begin{aligned} & \langle \mu_\sigma, G(\bar{u}) - G'(u_\sigma)(\bar{u}-u_\sigma)\rangle - \langle \bar{\mu}, G(u_\sigma)\rangle \\ =& \langle \mu_\sigma, G(u_\sigma) - \mathcal{R}_2^1\rangle - \langle \bar{\mu}, G(\bar{u}) + G'(\bar{u})(u_\sigma-\bar{u}) + \mathcal{R}_2^2\rangle \\ =& \langle \bar{\mu}, G'(\bar{u})(\bar{u}-u_\sigma)\rangle + \mathcal{R}_2, \end{aligned} \tag{3.50}$$

where the remainder terms from the Taylor expansion are

$$\mathcal{R}_2^1 = \frac{1}{2}\int_0^1 G''(u_\sigma + s(\bar{u} - u_\sigma))(\bar{u} - u_\sigma, \bar{u} - u_\sigma)s(1-s)\,\mathrm{d}s,$$

$$\mathcal{R}_2^2 = \frac{1}{2}\int_0^1 G''(\bar{u} + s(u_\sigma - \bar{u}))(\bar{u} - u_\sigma, \bar{u} - u_\sigma)s(1-s)\,\mathrm{d}s,$$

such that the sum $\mathcal{R}_2 = \mathcal{R}_2^1 + \mathcal{R}_2^2$ is quadratic in $\|\bar{u} - u_\sigma\|$. This last term in (3.50), without the remainder $\mathcal{R}_2$, is finally replaced by utilizing the adjoint equation (3.14), achieving the term

$$J_u'(\bar{q}, \bar{u})(\bar{u} - u_\sigma) - a_u'(\bar{q}, \bar{u})(\bar{u} - u_\sigma, \bar{z}).$$

Summing up all contributions yields the claim of the lemma. □

Remark 3.2. Due to the general formulation of the state constraint using the function G some unusual terms appear in the error representation. In the common situation that G is linear, the remainder term $\mathcal{R}_2$ disappears.
Also since the discretization of G by G_σ is left abstract, the term $\langle \mu_\sigma, G(u_\sigma) - G_\sigma(u_\sigma)\rangle$ can not be simplified. If for example an upper state constraint $G(u) = u_b - u$ is present with the approximation G_σ as introduced in Section 3.3.2, see (3.24), the term reduces to $\langle \mu_\sigma, u_b - u_{bh}\rangle$. Since μ_σ is comprised of point evaluations in gridpoints, this term is zero.

This motivates the following assumption with the intention to omit the term

$$\langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle$$

in the error estimator:

Assumption 3.11. *Let the approximation of G by G_σ be of such a quality that the term $\langle \mu_\sigma, G_\sigma(u_\sigma) - G(u_\sigma)\rangle$ is of not larger order than the remainder terms $\mathcal{R}_2, \mathcal{R}_3$.*

For more complicated state constraints it might be necessary to construct a computable estimator for this term.

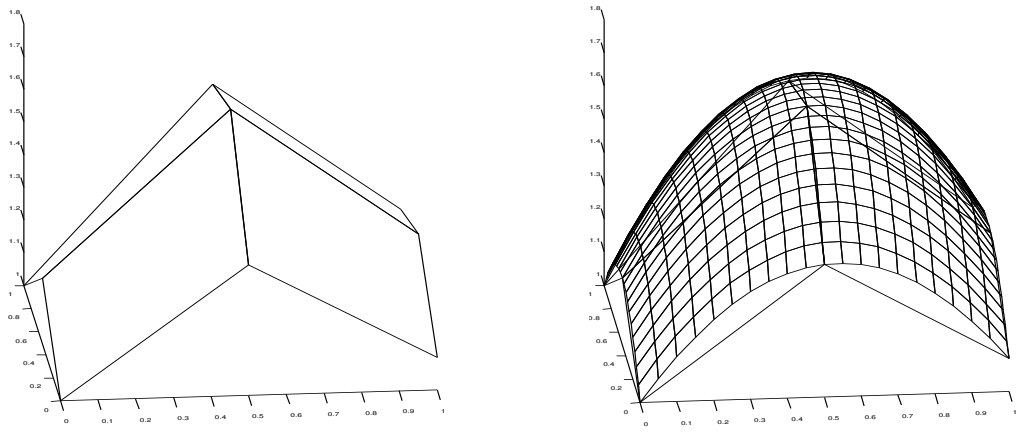
The error representation (3.45) still contains the continous solution $\bar{q}, \bar{u}, \bar{z}$. To define computable error estimators, that only contain the quantities $q_\sigma, u_\sigma, z_\sigma$, we employ some interpolation operators to get suitable approximations. The technique of interpolation in higher order finite element spaces has been used successfully in a posteriori error estimation. We use operators

$$P_h \colon X_h^s \to \hat{X}_h^s, P_q \colon Q_d \to \hat{Q}_d \tag{3.51}$$

where $\hat{X}_h^s$ and $\hat{Q}_d$ are suitable finite element spaces such that $P_h u_\sigma$ and $P_d q_\sigma$ are assumed to be good approximations to $\bar{u} - u_\sigma$ and $\bar{q} - q_\sigma$. As an example of such an operator we discuss an operator that can be used for quantities that are spatially discretized by the $cG(1)$ method. Remember that the mesh $\mathcal{T}_h$ is assumed to have a patch structure. We use

$$P_h = I_{2h}^{(2)} - id,$$

where $I_{2h}^{(2)} u_\sigma$ interpolates the bilinear function u_σ into the space of biquadratic finite elements on the patches. Figure 3.3 illustrates this interpolation.



(a) function from X_h^1, restricted to one patch

(b) the same function including its biquadratic interpolant

Figure 3.3.. Biquadratic interpolation on a patch in 2D

Remark 3.3. The use of the operators P_h for estimation of local approximation errors can be rigorously justified only for smooth solutions $\bar{q}, \bar{u}, \bar{z}$ employing super-convergence effects. Since in the state constrained case the adjoint solution $\bar{z}$ and consequently the control variable $\bar{q}$ possess in general only reduced regularity, this justification could be debated. Nevertheless, we expect a useful behaviour of the proposed error estimator, since the operator P_h is defined locally and the regions where the adjoint state $\bar{z}$ is not smooth are usually strongly localized.

Now by the approximations

$$\bar{q} - q_\sigma \approx P_d q_\sigma, \quad \bar{u} - u_\sigma \approx P_h u_\sigma, \quad \bar{z} - z_\sigma \approx P_h z_\sigma,$$

the computable error estimators can be formulated. An approximation of Lagrange multipliers is not necessary. The discretization error of the state space is estimated by

$$\begin{aligned}\eta_h :=& \frac{1}{2}\left(J_u'(q_\sigma, u_\sigma)(P_h u_\sigma) - a_u'(q_\sigma, u_\sigma)(P_h u_\sigma, z_\sigma) - a(q_\sigma, u_\sigma)(P_h z_\sigma) + (f, P_h z_\sigma)\right. \\ &\left. + J_u'((P_d + \mathrm{id})q_\sigma, (P_h + \mathrm{id})u_\sigma)(P_h u_\sigma) - a_u'((P_d + \mathrm{id})q_\sigma, (P_h + \mathrm{id})u_\sigma)(P_h u_\sigma, (P_h + \mathrm{id})z_\sigma)\right)\end{aligned} \tag{3.52}$$

and the discretization error of the control space is estimated by

$$\eta_d := \frac{1}{2}\left(J_q'(q_\sigma, u_\sigma)(P_d q_\sigma) - a_q'(q_\sigma, u_\sigma)(P_d q_\sigma, z_\sigma)\right). \tag{3.53}$$

Their sum makes up the total discretization error.

$$\eta = \eta_h + \eta_d \tag{3.54}$$

Remark 3.4. The residual of the gradient equation,

$$J_q'(q_\sigma, u_\sigma)(\bar{q} - \tilde{q}_\sigma) - a_q'(q_\sigma, u_\sigma)(\bar{q} - \tilde{q}_\sigma, z_\sigma),$$

can sometimes be shown to be zero, leading also to the estimator $\eta_d = 0$. Examples for this situation are the case $Q = Q_d$, or the situation discussed in [13], where for distributed control and $Q = L^2(\Omega)$ it was secured that $V_h \subset Q_d$. Then, a coupling of the discretization of Q to the one of X in the refinement strategy is necessary.

Remark 3.5. The origin of the interpolation operators in the first argument of J'_u and a'_u is that these terms originate from the term $\langle \mu, G'_\sigma(\bar{u})(\bar{u} - u_\sigma)\rangle$, which has been replaced by the dual equation. One cannot approximate this term directly by $\langle \mu_\sigma, G'_\sigma(u_\sigma)(I_{2h}^{(2)} u_\sigma - u_\sigma)\rangle$ since μ_σ acts only on the nodes $x_i \in \mathcal{N}_h$, where the term $I_{2h}^{(2)} u_\sigma - u_\sigma$ is zero. Another possibility is the definition of a different operator, in [41] an operator is employed which uses function evaluations in the midpoints of element faces.

Remark 3.6. Under additional regularity assumptions on active sets and problem data, a more thorough analysis is conducted in [45]. There, the multiplier $\bar{\mu}$ can be decomposed into a regular L^2-part on the active set and a singular part concentrated on the boundary between the active and inactive sets, which is used in the construction of the error estimator. Also the structure of the active set is taken into account, the article allows for active sets with nonempty interior in $\mathbb{R}^n$ and active sets that are just curves in Ω.

In the construction of an overall approximative solution algorithm for (3.1), after the solution of the discrete problem on $T = (\mathcal{T}_h, Q_d)$ and an estimate of the discretization error with respect to the cost functional is obtained, a new, refined, discretization has to be established unless some stopping criterion is met. Several strategies will be described here. The construction process of the refined discretization $\bar{T} = (\bar{\mathcal{T}}_{\bar{h}}, \bar{Q}_d)$ consists of decisions on two levels:

1. Which structure is chosen for refinement. A general framework is given in Algorithm 2.2, for elliptic problems the decision reduces to
   - only $\mathcal{T}_h$, or
   - only Q_d, or
   - both $\mathcal{T}_h$ and Q_d.
2. For every structure to be refined, the method of refinement can be chosen as
   - uniform, or
   - local, by using error indicators.

Although these decisions are independent from each other, the strategy's description is often given by just one word. Frequently used strategies that will also be used in the test computations here, are

- *Global refinement.* Both $\mathcal{T}_h$ and Q_d are refined uniformly. For the state space this means that the mesh $\bar{\mathcal{T}}_{\bar{h}}$ is obtained from the mesh $\mathcal{T}_h$ by replacing every cell by 2^n equivalent ones by evenly partition. The same is done with the control space, if it is distributed in space. In the case of boundary control, evenly partition the face elements.
  This basic „strategy“ uses no information from the error estimation at all. Its use is not recommended except for extremely simple problems, and is more thought of as a comparison strategy to measure the success of the other strategies against.

One could advise a discriminated variant of the global refinement by comparing the error estimators for the state and control contributions, and uniformly refine only the structure, whose error estimator has the larger absolute value.

- *Adaptive refinement.* First, the contributions of the error estimator η_h and η_d are used to decide which structure is to refine according to Algorithm 2.2.
  First assume that the state space is chosen for refinement. The local refinement strategy is based on a localization of the estimated error η_h. The estimate needs to be split up into cellwise contributions, local error indicators. For the refinement of the state space, the localization of η_h should not be obtained by taking the explicit formulation of (3.52) and evaluation of the respective integrals over the cell in question instead of Ω. This would lead to an overestimation of the error as the residual terms exhibit a strongly oscillatory behavior, see [10]. Instead, the localization can be achieved by two strategies: an integration by parts in space, or a filtering technique, which both secure the correct local order of convergence. Details are described in [65, Section 6.4.2]. Both these procedures yield the cellwise error indicators

$$\eta_h = \sum_{K \in \mathcal{T}_h} \eta_{h,K}, \tag{3.55}$$

  but there exist also strategies to gain nodewise indicators. For the construction of the new mesh $\overline{\mathcal{T}_h}$ from the error indicators $\eta_{h,K}$ a number of standard strategies exist. The natural idea is to order the cells according to the absolute values of their error indicators, starting from the largest,

$$|\eta_{h,K_1}| \geq |\eta_{h,K_2}| \geq \dots,$$

  and find one index i, up to which the corresponding cells are marked for refinement. The strategy to determine the index i is explained in detail in [65, section 6.5]. It is found as the argument minimizing

$$\mathcal{E}(i)\mathcal{N}(i)^{\delta}, \tag{3.56}$$

  where $\mathcal{E}(i)$ is a prediction of the discretization error on the refined mesh, $\mathcal{N}(i)$ is the number of degrees of freedom of that mesh, and δ depends on the degree of the polynomials used in the FE space, and the dimension of the respective domain to be discretized. The details of the refinement of the spatial discretization are described in Algorithm 3.3.
  If, on the other hand, the control space is chosen for refinement, a localization of the error estimator η_d can be used equivalently.

- *Coupled adaptive strategy.* For distributed control, where the state and control functions are defined on the same domain, in this strategy the discretization of the control is coupled to the discretization of the state. Only the localization of η_h is used to refine the mesh $\mathcal{T}_h$ locally, which induces both state and control discretization.

Algorithm 3.3. Local refinement of the spatial discretization for elliptic OCPs

1: **input data**: mesh $\mathcal{T}_h$
2: evaluate localization $\eta_h = \sum_{K \in \mathcal{T}_h} |\eta_{h,K}|$
3: sort $\eta_{h,K}$ by their absolute value: $|\eta_{h,K_1}| \geq |\eta_{h,K_2}| \geq \dots$
4: find the index $i = \arg \min_{1 \leq i \leq \mathcal{N}_h} \mathcal{E}(i)\mathcal{N}(i)^\delta$
5: mark cells $K_1, \dots, K_i$ for refinement
6: together with any marked cell, mark also all the cells from its patch, to keep the patch structure
7: refine all marked cells by evenly partition
8: **repeat**
9: iterate over all cells:
10: if current cell K_i has neighboring cell such that K_i has more than one hanging node on the shared face: refine K_i, together with its patch
11: **until** no such pair of cells exists any more
12: **output data**: refined mesh $\bar{\mathcal{T}}_{\bar{h}}$

3.6. Regularization and interior point method

The solution of (P) as described in Section 3.4 is only possible for some classes of problems. In the following, a method is presented that formally allows the numerical solution of all problems of type (1.1). The problem is regularized by replacement of the pointwise state constraint by a penalty functional in the cost functional, weighted by a decreasing function of a regularization parameter $\gamma > 0$. This unconstrained optimal control problem possesses a solution of increased regularity that can be approximated by usual methods. The convergence of the solution of the regularized problems to the unregularized solution can be proven for certain problem classes, see, e.g., [85].
The application of this approach introduces a new source of error, the regularization error $J(q,u) - J(q_\gamma, u_\gamma)$. This is not necessarily a drawback, if it can be kept small, that means equilibrated to the other error contributions. This finally poses the question of how to choose γ in comparison to the discretization parameter h, which in turn needs some kind of error estimation a priori or a posteriori to run a path following strategy. That is because a naive coupling $\gamma = \gamma(h)$ without taking the problem structure into account would be not very helpful as a γ too small causes a too large regularization error; and a γ too big makes the problem harder to solve and may lead out of the preferrable Newton convergence radius.
Looking ahead to parabolic problems, the decision between the solution methods, here primal-dual active set and regularization, will reappear. However, since the computational domain is increased by the temporal dimension it will be harder to find practical problems there that allow the application of the PDAS method and are still not too involved numerically. Thus the focus in Chapter 4 is on the regularization approach allowing us in the following to keep this section short for the treatment of elliptic problems.

Define a penalty functional, e.g. by

$$b_\gamma(u) = \gamma \int_\Omega - \ln(G(u(x)))\,\mathrm{d}x. \tag{3.57}$$

Alternatives are discussed in Section 4.1. Then the penalized cost functional

$$J_\gamma(q,u) = J(q,u) + b_\gamma(u) \tag{3.58}$$

is used to set up the regularized optimal control problems

$$(P_\gamma) \begin{cases} \min J_\gamma(q,u) & q \in Q, u \in X \\ a(q,u)(\varphi) = (f,\varphi) & \forall \varphi \in X \end{cases} . \tag{3.59}$$

The numerical solution of these problems can be done like in Section 2.3. The solution of a discretized variant of (P_γ) is denoted by $(q_{\gamma\sigma}, u_{\gamma\sigma})$. The error estimator for

$$J(\bar{q},\bar{u}) - J(q_{\gamma\sigma}, u_{\gamma\sigma}) \approx \eta = \eta_h + \eta_d + \eta_\gamma,$$

used to guide the simultaneous adaptive refinement and driving of $\gamma \to \infty$ can be derived equivalently to the parabolic case, which will be derived in the following chapter.
In a setting similar to the one of this thesis, the derivation of an estimator for the error $J(\bar{q},\bar{u}) - J_\gamma(q_{\gamma\sigma}, u_{\gamma\sigma})$ can be found in [100].

4. Parabolic Optimal Control Problems with State Constraints

This chapter is devoted to parabolic optimal control problems with state constraints. With the notation from Section 2.1.2, such a problem takes the form

$$(P_{par}) \begin{cases} \min J(q,u) & q \in Q, u \in X \\ (\partial_t u, \varphi)_I + a(q,u)(\varphi) + (u(0), \varphi(0)) = (f,\varphi)_I + (u_0(q), \varphi(0)) & \forall \varphi \in X. \\ G(u) \geq 0 \end{cases} \tag{4.1}$$

Remember the choices $Q \subset L^2(I,R)$ as a subspace, and

$$X = W(I,V) \cap L^s(I, W^{1,p}(\Omega)) \cap W^{1,s}(I, (W^{1,p'}(\Omega))^*)$$

with some $p > n$ and $s > \frac{2p}{p-n}$. Again we shortly discuss properties of the state equation, give conditions under which local optima exist, and obey first order Karush-Kuhn-Tucker optimality conditions.
As the PDAS method considered before is only possible for a limited class of optimal control problems with state constraints, we concentrate on a regularization method for the solution of state constrained parabolic problems. Regularized problems (P_γ), which are problems without state constraints, are used to approximate the state constrained problem (P_{par}). The discretization of these problems will be done by a discontinous Galerkin method, dG(r) in time, and in space a continous Galerkin method cG(s) like before.
As optimization method an interior point algorithm will be used. Due to the absence of inequality constraints, it does not contain Borel measures. The regularization causes an additional error, which needs to be accounted for in the a posteriori error estimation process.

4.1. Continous setting and optimality conditions

In Section 2.1.2, the parabolic state equation has been formulated as

$$(\partial_t u, \varphi)_I + a(q,u)(\varphi) + (u(0), \varphi(0)) = (f,\varphi)_I + (u_0(q), \varphi(0)) \quad \forall \varphi \in X. \tag{4.2}$$

There, the link between the control space Q and the state space X was established by the semilinear form

$$a \colon Q \times X \times Z \to \mathbb{R}$$

under Assumption 2.3 and Assumption 2.4. Again, this general formulation allows for different possible choices of the control space and, unlike the elliptic case, Q as a space of time dependent controls is set up using the spatial function space R. A few examples for these control types are:

Example 4.1. 1. *Distributed control*, where the distribution is in space and time. By the choice $R = L^2(\Omega), Q = L^2(L^2(\Omega))$ the following equation is set up:

$$\begin{aligned} \partial_t u - \Delta u &= q \quad \text{in } I \times \Omega \\ u|_{I\times\Gamma} &= 0 \quad \text{on } I \times \Gamma \\ u(0) &= 0 \quad \text{on } \{0\} \times \Omega \end{aligned}$$

2. *Boundary control*, by the choice $R = L^2(\Gamma), Q = L^2(L^2(\Gamma))$ the following equation is set up:

$$\begin{aligned} \partial_t u - \Delta u + u^3 &= 0 \quad \text{in } I \times \Omega \\ \partial_n u|_{I\times\Gamma} &= q \quad \text{on } I \times \Gamma \\ u(0) &= 0 \quad \text{on } \{0\} \times \Omega \end{aligned}$$

3. Distributed *initial control*, so that the control does only depend on the spatial, but not the temporal point. One can choose $R = L^2(\Omega), Q = \mathcal{P}_0(I, R)$, so there holds $Q \subset L^2(R)$ still. The following equation is set up:

$$\begin{aligned} \partial_t u - \Delta u + u^3 &= 0 \quad \text{in } I \times \Omega \\ \partial_n u|_{I\times\Gamma} &= 0 \quad \text{on } I \times \Gamma \\ u(0) &= q \quad \text{on } \{0\} \times \Omega \end{aligned}$$

4. *Parameter control*, by the choice $R = \mathbb{R}^k, Q = \mathcal{P}_0(I, R)$, so that the control space is in fact k-dimensional, the following equation is set up:

$$\begin{aligned} \partial_t u - \Delta u &= \sum_{i=1}^{k} q_i f_i \quad \text{in } I \times \Omega \\ \partial_n u|_{I\times\Gamma} &= 0 \quad \text{on } I \times \Gamma \\ u(0) &= 0 \quad \text{on } \{0\} \times \Omega \end{aligned}$$

with given functions $f_i \in L^2(L^2(\Omega))$.

5. *Parameter control with time dependent parameters*, similar to the last equation, but the parameters are timedependent in general. By the choice $R = \mathbb{R}^k, Q = L^2(\mathbb{R}^k)$, the following equation is set up:

$$\begin{aligned} \partial_t u - \Delta u &= \sum_{i=1}^{k} q_i(t) f_i(x) \quad \text{in } I \times \Omega \\ \partial_n u|_{I\times\Gamma} &= 0 \quad \text{on } I \times \Gamma \\ u(0) &= 0 \quad \text{on } \{0\} \times \Omega \end{aligned}$$

with given functions $f_i \in L^2(\Omega)$.

Remark 4.1. The state equations presented in Example 4.1 were given to illustrate the variety of choices that lie in the general introduction of $Q \subset L^2(I,R)$ as a subspace. Setting up an optimal control problem that can be analyzed with the help of a Lagrange multiplier by the Karush-Kuhn-Tucker theory requires a control-to-state operator $S\colon Q \to C(\bar{I} \times \bar{\Omega})$ with range in the continous states, as has been argued before. For parabolic problems this is frequently problematic, as it may put severe restrictions to the spatial dimension n, see [75]. A possible remedy is the introduction of additional constraints, specifically upper and lower L^∞-constraints on the control variable. In the following, the continuity of the states shall be assumed.

As in the elliptic case, we assume the unique solvability of the state equation according to Assumption 2.5, as a proof is possible for concrete realizations of (4.2), but not in the most general setting. The same holds true for the existence of an optimal control.

Next, first order optimality conditions are formulated. Although the discretization of the parabolic optimal control problem will not be based on formulation (4.1), we will still utilize the following conditions in the error estimation process. The measure space employed from now on is

$$\mathcal{M}(I \times \Omega) = (C(\bar{I} \times \bar{\Omega}))^*,$$

and the Lagrangian is now defined on $\mathcal{L}\colon Q \times X \times Z \times \mathcal{M}(I \times \Omega) \to \mathbb{R}$ by

$$\mathcal{L}(q,u,z,\mu) = J(q,u) + (f - \partial_t u, z)_I - a(q,u)(z) + (u_0(q) - u(0), z(0)) - \langle \mu, G(u) \rangle. \quad (4.3)$$

The KKT conditions that have been proven in Lemma 2.18 already are for a better overview stated again specially for the parabolic case and with minimum preconditions:

Theorem 4.1. *Let the point $(\bar{q},\bar{u}) \in Q \times X$ be locally optimal for the problem (4.1). Let S and G be one time Fréchet differentiable, and Assumptions 2.11 and 2.17 be valid. Then there exist an adjoint state $\bar{z} \in Z$ and a Lagrangian multiplier $\bar{\mu} \in \mathcal{M}(I \times \Omega)$ so that the following optimality system holds for $\bar{x} = (\bar{q},\bar{u},\bar{z},\bar{\mu})$:*

$$\mathcal{L}'_z(\bar{x})(\varphi) = 0 \qquad \forall \varphi \in Z \quad (4.4a)$$
$$\mathcal{L}'_u(\bar{x})(\varphi) = 0 \qquad \forall \varphi \in X \quad (4.4b)$$
$$\mathcal{L}'_q(\bar{x})(\psi) = 0 \qquad \forall \psi \in Q \quad (4.4c)$$
$$\langle \bar{\mu}, G(\bar{u}) \rangle = 0, \quad \bar{\mu} \geq 0. \quad (4.4d)$$

The explicit formulation of the optimality conditions is as follows: Equation (4.4a) gives the state equation again:

$$(\partial_t \bar{u}, \varphi)_I + a(\bar{q},\bar{u})(\varphi) + (\bar{u}(0), \varphi(0)) = (f,\varphi)_I + (u_0(\bar{q}), \varphi(0)) \quad \forall \varphi \in Z \quad (4.5)$$

The evaluation of (4.4b) gives the formulation of the adjoint equation

$$(\partial_t \varphi, \bar{z})_I + a'_u(\bar{q},\bar{u})(\varphi,\bar{z}) + (\varphi(0), \bar{z}(0)) = J'_u(\bar{q},\bar{u})(\varphi) - \langle \bar{\mu}, G'(\bar{u})(\varphi) \rangle \quad \forall \varphi \in X \quad (4.6)$$

The gradient equation (4.4c) is expressed by

$$J'_q(\bar{q},\bar{u})(\psi) - a'_q(\bar{q},\bar{u})(\psi,\bar{z}) + (u'_0(\bar{q})(\psi), \bar{z}(0)) = 0 \quad \forall \psi \in Q. \quad (4.7)$$

Like in the elliptic case, the complementarity conditions (4.4d) can be expressed equivalently by the variational inequality

$$\langle \bar{\mu}, \varphi - G(\bar{u}) \rangle \geq 0 \qquad \forall \varphi \in C(\bar{I} \times \bar{\Omega}), \varphi \geq 0. \tag{4.8}$$

Analog to the treatment of elliptic optimal control problems it is possible to solve parabolic ones by the PDAS method, provided the structural assumption that the auxiliary problem (P_E) can always be solved. But since this method can be transferred directly from the elliptic case, we will refrain from covering this method here.

4.2. Regularization

Instead, we consider a regularization method, as has been introduced in Section 2.4. The considered penalty functionals are defined as follows. For a given order $o \geq 1$ the polynomial or logarithmic penalty functional is

$$\begin{aligned} b_\gamma(u) &:= \int_{\Omega \times I} -\gamma \ln(G(u)) \,\mathrm{d}(x,t) \qquad \text{for } o = 1, \\ b_\gamma(u) &:= \int_{\Omega \times I} \frac{1}{o-1} \gamma^o (G(u))^{1-o} \,\mathrm{d}(x,t) \qquad \text{for } o > 1, \end{aligned} \tag{4.9}$$

and depends on the regularization parameter $\gamma > 0$. The derivative of the penalty functional is thus

$$b'_\gamma(u) = \int_{\Omega \times I} -\gamma^o (G(u(t,x)))^{-o} G'(u) \,\mathrm{d}(x,t). \tag{4.10}$$

For every regularization parameter $\gamma > 0$ the regularized parabolic optimal control problem is formulated by

$$(P_\gamma) \begin{cases} \min J_\gamma(q_\gamma, u_\gamma) := J(q_\gamma, u_\gamma) + b_\gamma(u_\gamma) \qquad q_\gamma \in Q, u_\gamma \in W \\ (\partial_t u_\gamma, \varphi)_I + a(q_\gamma, u_\gamma)(\varphi) + (u_\gamma(0), \varphi(0)) = (f, \varphi)_I + (u_0(q_\gamma), \varphi(0)) \quad \forall \varphi \in W, \end{cases} \tag{4.11}$$

where the state space is chosen like for unconstrained problems,

$$W = W(I, V).$$

A problem (P_γ) can be solved by methods for unconstrained problems, which will be detailed below. The intention is to solve a sequence of these problems (P_{γ_i}) with $\gamma_i \to \infty$ such that the solutions of these problems converge to the solution of the constrained problem. For some classes of optimal control problems and penalty functionals this property has been proven, see, e.g., [85]. Of course the question arises whether a later implementation should really solve the problems (P_{γ_i}) with good accuracy before increasing γ, or whether a few steps in the respective approximative solution algorithm are sufficient.

For problem (4.11), the Lagrange functional is defined by $\mathcal{L}_\gamma \colon Q \times W \times W \to \mathbb{R}$

$$\mathcal{L}_\gamma(q_\gamma, u_\gamma, z_\gamma) = J(q_\gamma, u_\gamma) + (f - \partial_t u_\gamma, z_\gamma)_I - a(q_\gamma, u_\gamma)(z_\gamma) + (u_0(q_\gamma) - u_\gamma(0), z_\gamma(0)) + b_\gamma(u_\gamma). \tag{4.12}$$

The optimality conditions can now be derived according to Lemma 2.16:

Theorem 4.2. *Let the point $(q_\gamma, u_\gamma) \in Q \times W$ be locally optimal for the problem (4.11). Then there exists an adjoint state $z_\gamma \in W$ such that the following optimality system holds for $x_\gamma = (q_\gamma, u_\gamma, z_\gamma)$:*

$$\mathcal{L}'_{\gamma,z}(x_\gamma)(\varphi) = 0 \quad \forall \varphi \in W \tag{4.13a}$$
$$\mathcal{L}'_{\gamma,u}(x_\gamma)(\varphi) = 0 \quad \forall \varphi \in W \tag{4.13b}$$
$$\mathcal{L}'_{\gamma,q}(x_\gamma)(\psi) = 0 \quad \forall \psi \in Q \tag{4.13c}$$

The explicit formulations are given as follows: The state equation is

$$(\partial_t u_\gamma, \varphi)_I + a(q_\gamma, u_\gamma)(\varphi) + (u_\gamma(0), \varphi(0)) = (f, \varphi)_I + (u_0(q_\gamma), \varphi(0)) \quad \forall \varphi \in W. \tag{4.14}$$

The formal derivation of the adjoint equation gives

$$(\partial_t \varphi, z_\gamma)_I + a'_u(q_\gamma, u_\gamma)(\varphi, z_\gamma) + (\varphi(0), z_\gamma(0)) = J'_u(q_\gamma, u_\gamma)(\varphi) + b'_\gamma(u_\gamma)(\varphi) \quad \forall \varphi \in W,$$

first. For implementational reasons, this equation should be transformed, so that it does not contain a terminal but an initial condition of the differential equation running backwards in time. Usually this is done by integration in parts of the term $(\partial_t \varphi, z_\gamma)_I$, so that the formulation of the adjoint equation becomes

$$-(\varphi, \partial_t z_\gamma)_I + a'_u(q_\gamma, u_\gamma)(\varphi, z_\gamma) + (\varphi(T), z_\gamma(T)) = J'_u(q_\gamma, u_\gamma)(\varphi) + b'_\gamma(u_\gamma)(\varphi) \quad \forall \varphi \in W. \tag{4.15}$$

The gradient equation is given by

$$J'_q(q_\gamma, u_\gamma)(\psi) - a'_q(q_\gamma, u_\gamma)(\psi, z_\gamma) + (u'_0(q_\gamma)(\psi), z_\gamma(0)) = 0 \quad \forall \psi \in Q. \tag{4.16}$$

4.3. Finite element discretization in space and time

For the discretization of a parabolic optimal control problem, discretizations in time and space need to be performed. In this order, the levels of discretization are indicated by the subscripts k for the temporal, h for the spatial, and d for the control space discretizations, such that $\sigma = (k, h, d)$.

The discretization of the spatial state variable is again done by a Galerkin finite element method of order s, with $s \in \mathbb{N}$, $s \geq 1$, in short cG(s). For the time variable the discontinous Galerkin method dG(r) is used. The discretization of the control variable is kept abstract.

First the regularized continous problem (4.11) is semidiscretized in time. For that, assume we are given a set of $M + 1$ time points

$$0 = t_0 < t_1 < t_2 < \cdots < t_{M-1} < t_M = T.$$

The subintervals defined by $I_m = (t_{m-1}, t_m] \subset I$ are the ones used to define the spaces. Their lengths are denoted by $k_m := t_m - t_{m-1}$, and analog to the spatial discretization parameter from Section 3.3 we define the temporal discretization parameter k as a function on I by setting

$$k|_{I_m} = k_m \, \forall m = 1, \ldots, M.$$

For the discontinous Galerkin method of order r, the space

$$\tilde{X}_k^r := \left\{ v_k \in L^2(I,H) | v_k|_{I_m} \in \mathcal{P}_r(I_m, X), m = 1,2,\ldots,m \text{ and } v_k(0) \in H \right\} \tag{4.17}$$

is employed. The following derivation is possible for arbitrary $r \in \mathbb{N}_0$, but in the numerical experiments later on only $r = 0$ is used, which is equivalent to a variant of the implicit Euler method.
Next, for any discontinous function the notation for function values at the left and right endpoint of the time intervals, and the jump in between is introduced by

$$v_{k,m}^+ := \lim_{t \searrow 0} v_k(t_m + t), \quad v_{k,m}^- := \lim_{t \searrow 0} v_k(t_m - t), \quad [v_k]_m = v_{k,m}^+ - v_{k,m}^-$$

The semidiscretized state equation then reads: For a $q_k \in Q$ find $u_k \in \tilde{X}_k^r$ so that

$$\sum_{m=1}^{M} (\partial_t u_k, \varphi)_{I_m} + a(q_k, u_k)(\varphi) + \sum_{m=0}^{M-1} ([u_k]_m, \varphi_m^+) + (u_{k,0}^-, \varphi_0^-) = (f,\varphi)_I + (u_0(q_k), \varphi_0^-) \, \forall \varphi \in \tilde{X}_k^r. \tag{4.18}$$

The solution operator of this equation is denoted by

$$S_k \colon Q \to \tilde{X}_k^r$$

With this, the semidiscretized optimal control problem is given by

$$(P_{\gamma k}) \begin{cases} \min J_\gamma(q_{\gamma k}, u_{\gamma k}), & q_{\gamma k} \in Q, u_{\gamma k} \in \tilde{X}_k^r \\ S_k(q_{\gamma k}) = u_{\gamma k} \end{cases}, \tag{4.19}$$

and the Lagrangian associated with the discontinous Galerkin discretization in time $\tilde{\mathcal{L}}_\gamma \colon Q \times \tilde{X}_k^r \times \tilde{X}_k^r$ is

$$\begin{aligned} \tilde{\mathcal{L}}_\gamma(q_{\gamma k}, u_{\gamma k}, z_{\gamma k}) = J(q_{\gamma k}, u_{\gamma k}) + (f, z_{\gamma k})_I - \sum_{m=1}^{M} (\partial_t u_{\gamma k}, z_{\gamma k})_{I_m} - a(q_{\gamma k}, u_{\gamma k})(z_{\gamma k}) \\ - \sum_{m=0}^{M-1} ([u_{\gamma k}]_m, z_{\gamma k,m}^+) + (u_0(q_{\gamma k}) - u_{\gamma k}(0), z_{\gamma k}(0)) + b_\gamma(u_{\gamma k}). \end{aligned} \tag{4.20}$$

The next discretization level is the spatial discretization. This is done by a continous Galerkin method of order s similar to in the elliptic case, but on every time section $\{t_0\}$ and $I_m, m = 1 \ldots M$ one spatially discretized space needs to be specified. One possible choice is to use the same mesh $\mathcal{T}_h$ with its according space X_h^s, like in (3.19), on every interval. The finite element space would then be chosen as

$$\tilde{X}_{kh}^{r,s} = \left\{ v_{kh} \in L^2(I,H) : v_{kh}|_{I_m} \in \mathcal{P}_r(I_m, X_h^s) \, \forall m = 1,2,\ldots,M \text{ and } v_{kh}(0) \in X_h^s \right\} \tag{4.21}$$

Another possibility is to allow for different meshes $\mathcal{T}_h^m$ for every time section $m = 0 \ldots M$. This *dynamic* spatial discretization then employs the $M+1$ spaces $X_h^{s,m}$ implied by these meshes, and uses the finite element space

$$\tilde{X}_{kh}^{r,s} = \left\{ v_{kh} \in L^2(I,H) : v_{kh}|_{I_m} \in \mathcal{P}_r(I_m, X_h^{s,m}) \forall m = 1,2,\ldots,M \text{ and } v_{kh}(0) \in X_h^{s,0} \right\} \tag{4.22}$$

The fully discretized state equation then reads: For given $q_{\gamma kh} \in Q$ find $u_{\gamma kh} \in \tilde{X}_{kh}^{r,s}$ so that

$$\begin{aligned}\sum_{m=1}^{M} (\partial_t u_{\gamma kh}, \varphi)_{I_m} + a(q_{\gamma kh}, u_{\gamma kh})(\varphi) + \sum_{m=0}^{M-1} ([u_{\gamma kh}]_m, \varphi_m^+) + (u_{\gamma kh,0}^-, \varphi_0^-) \\ = (f, \varphi)_I + (u_0(q_{\gamma kh}), \varphi_0^-) \, \forall \varphi \in \tilde{X}_{kh}^{r,s}.\end{aligned} \tag{4.23}$$

The solution operator of this equation is denoted by

$$S_{kh} \colon Q \to \tilde{X}_{kh}^{r,s},$$

so the temporally and spatially discretized problem can be written as

$$(P_{\gamma kh}) \begin{cases} \min J_\gamma(q_{\gamma kh}, u_{\gamma kh}), & q_{\gamma kh} \in Q, u_{\gamma kh} \in \tilde{X}_{kh}^{r,s} \\ S_{kh}(q_{\gamma kh}) = u_{\gamma kh} \end{cases} . \tag{4.24}$$

On the last level, the control variable has to be discretized by the choice of a finite dimensional $Q_d \subset Q$. Due to the abstract nature of Q in the setting of this section we can as usual not give a concrete form, but will discuss a few examples.

Example 4.2. For distributed control, as introduced in Example 4.1, the space $Q = L^2(I, L^2(\Omega))$ can again be discretized like the state space W. So utilizing the same time mesh and space mesh(es) and the application of the dG(r) method in time and the cG(s) method in space. It can sometimes make sense to discretize the control on a coarser time mesh than the state. The time points of the control discretization would the be a subset of the time points of the state discretization.

Example 4.3. For initial control, where $Q = L^2(\Omega)$, it seems reasonable to utilize the same discretization that has been used in t_0 for the state equation, so set $Q_d = V_h^{s_d}$ or $Q_d = V_h^{s_d,0}$ with some appropriate polynomial degree s_d.

Example 4.4. For control by time dependent parameters, where $R = \mathbb{R}^k$ and $Q = L^2(R^k)$, one can use the same time points $\{t_i\}_{i=0}^M$ as before and set $Q_d = \{q \in Q \colon q|_{I_m} \in \mathcal{P}_{r_d}(I_m, R^k)\}$ with some appropriate polynomial degree r_d.

With the combination of the subscripts $\sigma = (k, h, d)$, the fully discretized problem then reads

$$(P_{\gamma\sigma}) \begin{cases} \min J_\gamma(q_{\gamma\sigma}, u_{\gamma\sigma}), & q_{\gamma\sigma} \in Q_d, u_{\gamma\sigma} \in \tilde{X}_{kh}^{r,s} \\ S_\sigma(q_{\gamma\sigma}) = u_{\gamma\sigma} \end{cases} . \tag{4.25}$$

Utilizing the reduced cost functional $j_{\gamma\sigma} \colon Q \to \mathbb{R}$ given by $j_{\gamma\sigma}(q) = J_\gamma(q, S_\sigma(q))$ the fully discretized problem in reduced form is formulated as

$$(P_{\gamma\sigma,red}) \qquad \min j_{\gamma\sigma}(q_{\gamma\sigma}), \qquad q_{\gamma\sigma} \in Q_d. \tag{4.26}$$

4.4. Optimization by interior point method

In this section the numerical solution of the fully discretized regularized problem (4.25) for one given discretization $\tilde{X}_{kh}^{r,s}, Q_d$ and regularization parameter γ will be described. As it is an optimal control problem without additional constraints, it can be solved by the Newton method as layed out in Section 2.3.4, with only a few adaptions.

One difference is that starting value for the control q^0 can not be chosen arbitrarily, but it has to be an admissible control for $(P_{\gamma\sigma})$, which means $J_\gamma(q^0, S_\sigma(q^0)) < \infty$. This means also, that Assumption 2.11 which secures the existence of an admissible control for (P), or the extension of this assumption to a discretization of (P), is not sufficient. The reason is that for this admissible control the constraint could be active on a set with nonzero measure so that $b_\gamma(u_{\gamma\sigma}) = \infty$. Thus a new assumption on the existence of an admissibe control, and a constraint qualification is necessary:

Assumption 4.3. *There exists a control $q_d^* \in Q_d$ such that $J_\gamma(q_d^*, S_\sigma(q_d^*)) < \infty$.*

Assumption 4.4. *Let $q_{\gamma\sigma} \in Q_d$ be a locally optimal solution of the problem $(P_{\gamma\sigma})$. Then the operator $S'_\sigma(q_{\gamma\sigma})$ is a surjective operator.*

Given an admissible starting control q^0 the iteration $q^i \to q^{i+1}$ follows the strategy from Section 2.3.4. Therefore the derivation of the computable first and second derivatives of $j_{\gamma\sigma}$ utilizes the Lagrangian $\tilde{\mathcal{L}}_\gamma$ from (4.20) instead of $\bar{\mathcal{L}}$ from (2.37) within the approach from Section 2.3.2. So consider the optimality conditions for problem $(P_{\gamma\sigma})$ which, according to Lemma 2.16 can under Assumption 4.3 and Assumption 4.4 be formulated using the Lagrange functional $\tilde{\mathcal{L}}_\gamma$ as

$$\tilde{\mathcal{L}}'_{\gamma,z}(q_{\gamma\sigma}, u_{\gamma\sigma}, z_{\gamma\sigma})(\varphi) = \tilde{\mathcal{L}}'_{\gamma,u}(q_{\gamma\sigma}, u_{\gamma\sigma}, z_{\gamma\sigma})(\varphi) = \tilde{\mathcal{L}}'_{\gamma,q}(q_{\gamma\sigma}, u_{\gamma\sigma}, z_{\gamma\sigma})(\psi) = 0 \,\forall \varphi \in \tilde{X}_{kh}^{r,s}, \forall \psi \in Q_d$$

Analog to the derivation in [9], the optimality conditions are expressed explicitly, and the following equations are derived in explicit form:

- the discrete state equation (4.23), determining $u_{\gamma\sigma}$ for given $q_{\gamma\sigma} \in Q_d$.
- the discrete adjoint equation: for given $q_{\gamma\sigma} \in Q_d$, $u_{\gamma\sigma} \in \tilde{X}_{kh}^{r,s}$ determine $z_{\gamma\sigma} \in \tilde{X}_{kh}^{r,s}$ by solving

$$\begin{aligned} -\sum_{m=1}^{M} (\varphi, \partial_t z_{\gamma\sigma})_{I_m} - \sum_{m=1}^{M-1} (\varphi_m^-, [z_{\gamma\sigma}]_m) + (\varphi(T), z_{\gamma\sigma}(T)) + a'_u(q_{\gamma\sigma}, u_{\gamma\sigma})(\varphi, z_{\gamma\sigma}) \\ = J'_u(q_{\gamma\sigma}, u_{\gamma\sigma})(\varphi) + b'_\gamma(u_{\gamma\sigma})(\varphi) \qquad \forall \varphi \in \tilde{X}_{kh}^{r,s}. \end{aligned} \tag{4.27}$$

- the discrete tangent equation, which his obtained by total differentiation of the state equation, determining $\delta u \in \tilde{X}_{kh}^{r,s}$ for a given direction $\delta q \in Q_d$ by solving

$$\begin{aligned} \sum_{m=1}^{M} (\partial_t \delta u, \varphi)_{I_m} + a'_u(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta u, \varphi) + \sum_{m=0}^{M-1} ([\delta u]_m, \varphi_m^+) + (\delta u(0), \varphi(0)) \\ = -a'_q(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta q, \varphi) + (u'_0(q)(\delta q), \varphi(0)) \quad \forall \varphi \in \tilde{X}_{kh}^{r,s}. \end{aligned} \tag{4.28}$$

- the discrete additional adjoint equation, which is obtained by total differentiation of the dual equation, determining $\delta z \in \tilde{X}_{kh}^{r,s}$ for given $\delta q \in Q_d$ and $\delta u \in \tilde{X}_{kh}^{r,s}$ by solving

$$\begin{aligned} -\sum_{m=1}^{M} (\varphi, \partial_t \delta z)_{I_m} + a'_u(q_{\gamma\sigma}, u_{\gamma\sigma})(\varphi, \delta z) - \sum_{m=1}^{M-1} (\varphi_m^-, [\delta z]_m) + (\varphi(T), \delta z(T)) \\ = -a''_{uu}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta u, \varphi, z_{\gamma\sigma}) - a''_{qu}(q, u)(\delta q, \varphi, z_{\gamma\sigma}) + J''_{uu}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta u, \varphi) \\ + J''_{qu}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta q, \varphi) + b''_{\gamma}(u_{\gamma\sigma})(\delta u, \varphi) \quad \forall \varphi \in \tilde{X}_{kh}^{r,s}. \end{aligned} \tag{4.29}$$

With these equations, like in Section 2.3.4 the first and second derivatives can be calculated as follows:

- for any given direction $\delta q \in Q_d$ calculate $j'_\gamma(q_{\gamma\sigma})(\delta q)$ as

$$\begin{aligned} j'_{\gamma\sigma}(q_{\gamma\sigma})(\delta q) = &\quad \tilde{\mathcal{L}}'_{\gamma,q}(q_{\gamma\sigma}, u_{\gamma\sigma}, z_{\gamma\sigma})(\delta q) \\ = &\quad J'_q(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta q) - a'_q(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta q, z_{\gamma\sigma}) + (u'_0(q_{\gamma\sigma})(\delta q), z_{\gamma\sigma}(0)) \end{aligned} \tag{4.30}$$

- for any given directions $\delta q, \tau q \in Q_d$ calculate $j''_\gamma(q_{\gamma\sigma})(\delta q, \tau q)$ as

$$\begin{aligned} j''_\gamma(q_{\gamma\sigma})(\delta q, \tau q) = &J''_{qq}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta q, \tau q) + J''_{uq}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta u, \tau q) - a''_{qq}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta q, \tau q, z_{\gamma\sigma}) \\ &- a''_{uq}(q_{\gamma\sigma}, u_{\gamma\sigma})(\delta u, \tau q, z_{\gamma\sigma}) - a'_q(q_{\gamma\sigma}, u_{\gamma\sigma})(\tau q, \delta z) \\ &+ (u'_0(q_{\gamma\sigma})(\tau q), \delta z(0)) + (u''_0(q_{\gamma\sigma})(\delta q, \tau q), z_{\gamma\sigma}(0)). \end{aligned} \tag{4.31}$$

Thus, Algorithm 2.1 can be utilized to solve $(P_{\gamma\sigma})$ as the necessary evaluations of $j'_{\gamma\sigma}$ and $j''_{\gamma\sigma}$ and differential equations are provided. The solution of $(P_{\gamma\sigma})$ for a given discretization σ and regularization parameter γ is presented in Algorithm 4.1. The incorporation into a comprehensive algorithm for the solution of (P_{par}) needs to detail more steps, for example secure the admissibility of q^0, and manage the increasing of γ and refinement of the discretization. In preparation of this algorithm, in the following section the necessary error estimators will be derived.

Algorithm 4.1. Interior point optimization method for state constrained parabolic OCPs

1: **input data**: q^0, γ
2: **parameter**: TOL_N, TOL_L
3: Solve $u^0 = S_\sigma(q^0)$ by (4.23), $z^0 = T_\sigma(q^0)$ by (4.28)
4: check for admissibility: make sure $b_\gamma(u^0) < \infty$
5: set up the problem

$$\begin{cases} \min J(q_{\gamma\sigma}, u_{\gamma\sigma}) + b_\gamma(u_{\gamma\sigma}) & q_{\gamma\sigma} \in Q_d, u_{\gamma\sigma} \in \tilde{X}_{hk}^{r,s} \\ u_{\gamma\sigma} = (S_\sigma(q_{\gamma\sigma})) \end{cases}$$

6: Solve by Newton method $(q^0, u^0, z^0, TOL_N, TOL_L)$, see Algorithm 2.1.
7: this yields $\bar{q}, \bar{u}, \bar{z}$.
8: **output data**: $\bar{q}, \bar{u}, \bar{z}$

4.5. A posteriori error estimator and adaptivity

In order to estimate the error with respect to the cost functional J caused by the regularization and discretization of problem (4.1), this error is dissected in the following way:

$$\begin{aligned} J(\bar{q}, \bar{u}) - J(q_{\gamma\sigma}, u_{\gamma\sigma}) &= J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) \\ &\quad + J(q_\gamma, u_\gamma) - J(q_{\gamma\sigma}, u_{\gamma\sigma}) \\ &= J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) \\ &\quad + J_\gamma(q_\gamma, u_\gamma) - J_\gamma(q_{\gamma\sigma}, u_{\gamma\sigma}) + b_\gamma(u_{\gamma\sigma}) - b_\gamma(u_\gamma) \end{aligned} \tag{4.32}$$

The influences of the steps of numerical treatment are seperated by the following contributions:

$$\begin{aligned} \eta_\gamma &:\approx J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) \\ \eta_k &:\approx J_\gamma(q_\gamma, u_\gamma) - J_\gamma(q_{\gamma k}, u_{\gamma k}) + b_\gamma(u_{\gamma k}) - b_\gamma(u_\gamma) \\ \eta_h &:\approx J_\gamma(q_{\gamma k}, u_{\gamma k}) - J_\gamma(q_{\gamma hk}, u_{\gamma hk}) + b_\gamma(u_{\gamma hk}) - b_\gamma(u_{\gamma k}) \\ \eta_d &:\approx J_\gamma(q_{\gamma hk}, u_{\gamma hk}) - J_\gamma(q_{\gamma\sigma}, u_{\gamma\sigma}) + b_\gamma(u_{\gamma\sigma}) - b_\gamma(u_{\gamma hk}) \end{aligned} \tag{4.33}$$

Here, η_γ is the regularization error estimator, η_k the temporal error estimator, η_h the spatial error estimator, and η_d like before the estimator for the control discretization error. Their combination gives

$$\eta := \eta_h + \eta_k + \eta_d + \eta_\gamma. \tag{4.34}$$

Remark 4.2. The fully discretized problem (4.25) which is solved numerically approximates a local minimizer of J_γ. Thus the term $b_\gamma(u_{\gamma\sigma}) - b_\gamma(u_\gamma)$ from the representation (4.32) can be viewed as an error in a quantity of interest. This means an error estimator for this term could be derived as in [65], using methods for a general quantity of interest. This approach requires the solution of an additional linear-quadratic optimal control problem. To avoid this numerical effort the error estimator will be derived utilizing the available information on b and its dissection in (4.33).

First the regularization error is approached.

Lemma 4.5. *Let $(\bar{q}, \bar{u}) \in Q \times X$ be a local optimal solution of the original problem (4.1), and $(q_\gamma, u_\gamma) \in Q \times W$ a local optimal solution of the regularized problem (4.11), with the Lagrange functional $\mathcal{L}$ being three times Gateaux differentiable. Then the following representation formula for the regularization error holds:*

$$J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) = \frac{1}{2}\langle b'_\gamma(u_\gamma), G(\bar{u}) - G(u_\gamma)\rangle - \frac{1}{2}\langle \bar{\mu}, G(u_\gamma)\rangle + \mathcal{R}_{reg}, \tag{4.35}$$

where the remainder term $\mathcal{R}_{reg}$ is of third order.

Proof. Together with the adjoint states $\bar{z}$ and z_γ and multiplier $\bar{\mu}$ from Theorems 4.1 and 4.2, we set

$$\bar{x} = (\bar{q}, \bar{u}, \bar{z}, \bar{\mu}) \qquad \text{and} \qquad x_\gamma = (q_\gamma, u_\gamma, z_\gamma, b'_\gamma(u_\gamma)).$$

Since $(\bar{q}, \bar{u})$ and (q_γ, u_γ) both satisfy the state equation (4.2), and $\bar{x}$ satisfies the complementarity condition (4.4d), it follows that

$$J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) = \mathcal{L}(\bar{x}) - \mathcal{L}(x_\gamma) - \langle b'_\gamma(u_\gamma), G(u_\gamma)\rangle \tag{4.36}$$

With the procedure from the proof of Lemma 3.9, the formulation of the difference $\mathcal{L}(\bar{x}) - \mathcal{L}(x_\gamma)$ as an integral and its evaluation with the trapezoidal rule, we get

$$J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) = \frac{1}{2}\mathcal{L}'(\bar{x})(\bar{x} - x_\gamma) + \frac{1}{2}\mathcal{L}'(x_\gamma)(\bar{x} - x_\gamma) - \langle b'_\gamma(u_\gamma), G(u_\gamma)\rangle + \mathcal{R}_{reg}, \tag{4.37}$$

where the remainder term takes the form

$$\mathcal{R}_{reg} = \frac{1}{2}\int_0^1 \mathcal{L}'''(s\bar{x} + (1-s)x_\gamma)(\bar{x} - x_\gamma, \bar{x} - x_\gamma, \bar{x} - x_\gamma)s(s-1)\,\mathrm{d}s. \tag{4.38}$$

The evaluation of the partial derivatives in the q, u and z coordinate gives zero both for $\mathcal{L}'(\bar{x})$ and $\mathcal{L}'(x_\gamma)$ due to the optimality conditions (4.4) and (4.13). There remain the terms

$$\begin{aligned}\frac{1}{2}\mathcal{L}'_\mu(\bar{x})(\bar{\mu} - b'_\gamma(u_\gamma)) + \frac{1}{2}\mathcal{L}'_\mu(x_\gamma)(\bar{\mu} - b'_\gamma(u_\gamma)) &= -\frac{1}{2}\langle \bar{\mu} - b'_\gamma(u_\gamma), G(\bar{u})\rangle - \frac{1}{2}\langle \bar{\mu} - b'_\gamma(u_\gamma), G(u_\gamma)\rangle \\ &= \frac{1}{2}\langle b'_\gamma(u_\gamma), G(\bar{u})\rangle - \frac{1}{2}\langle \bar{\mu} - b'_\gamma(u_\gamma), G(u_\gamma)\rangle,\end{aligned}$$

using the complementarity condition (4.4d). Adding the remaining summand $-\langle b'_\gamma(u_\gamma), G(u_\gamma)\rangle$ from (4.37) proves the assertion. □

In order to define a computable error estimator from (4.35), in [100, section 4.1] it is argued that the convergence of $b'_\gamma(u_\gamma)$ to $\bar{\mu}$ for $\gamma \to \infty$ in the sense of $\mathcal{M}(I \times \Omega)$, [85], justifies the approximation of $\langle \bar{\mu}, G(u_\gamma)\rangle$ by $\langle b'_\gamma(u_\gamma), G(u_\gamma)\rangle$. This yields the intermediary approximation

$$J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) \approx \frac{1}{2}\langle b'_\gamma(u_\gamma), G(\bar{u})\rangle - \langle b'_\gamma(u_\gamma), G(u_\gamma)\rangle. \tag{4.39}$$

Two different ways to treat this expression further are discussed in [100], either using the sign of $\langle b'_\gamma(u_\gamma), G(\bar{u})\rangle$ or the convergence of $u_\gamma \to \bar{u}$. Either possibility results in an estimator of the form

$$J(\bar{q}, \bar{u}) - J(q_\gamma, u_\gamma) \approx -c_0\langle b'_\gamma(u_{\gamma\sigma}), G(u_{\gamma\sigma})\rangle \tag{4.40}$$

with the constant $c_0 \in \{0.5, 1\}$. The choice of a constant $c_0 \in [0.5, 1]$ can also be argued for in the following example.

Example 4.5. In the case of an upper state constraint, where the state u is bounded from above by a given function $\psi \in C^2(\Omega)$

$$u \leq \psi \text{ on } \Omega \times I \quad \Leftrightarrow \quad G(u) := \psi - u \geq 0,$$

the regularization error estimator in representation (4.39) takes the form

$$-\int\limits_{\Omega\times I} \underbrace{\frac{\psi - u_\gamma}{\gamma^o(\psi - u_\gamma)^o}}_{=:B} \,\mathrm{d}(x,t) + \frac{1}{2}\int\limits_{\Omega\times I} \underbrace{\frac{\psi - u}{\gamma^o(\psi - u_\gamma)^o}}_{=:A} \,\mathrm{d}(x,t)$$

So, on the active set, there holds $\psi - u = 0$, and, for large values of γ, $\psi - u_\gamma$ is small due to the regularization. Thus we expect $|A| \ll |B|$.
On the inactive set, $\psi - u$ is in general large, and $\psi - u_\gamma$ also. Thus we expect $|A| \approx |B|$. Altogether, comparing points from the active and the inactive set, we expect the value of either function A and B in these points to be much smaller on the inactive than on the active set. This leads to the following extreme cases:

- if the active set is large, then $|\int B \,\mathrm{d}(x,t)| \gg |\int A \,\mathrm{d}(x,t)|$, and the choice $c_0 = 1$ in (4.40) is justified.

- if the active set is small, for instance it consists only of a point, then $|\int B \,\mathrm{d}(x,t)| \approx |\int A \,\mathrm{d}(x,t)|$, and the choice $c_0 = 0.5$ in (4.40) is justified.

So we define the regularization error estimator as

$$\eta_\gamma := -c_0 \langle b'_\gamma(u_{\gamma\sigma}), G(u_{\gamma\sigma})\rangle \tag{4.41}$$

with a constant $c_0 \in [0.5, 1]$.

Considering from (4.33) the parts including $J_\gamma()$, we can see that the evaluation is done in points that are optimal solutions of problem (4.11) on the different levels of discretization, and the evaluated functional J_γ is also that of this problem. As this problem is an optimal control problem without further inequality constraints, we can use the methods from [65, section 6.2] to find a suitable representation. For shorter notation define the residuals

$$\begin{aligned}
\tilde{\rho}^u(q,u)(\varphi) &= \tilde{\mathcal{L}}'_{\gamma,z}(q,u,z)(\varphi),\\
\tilde{\rho}^z(q,u,z)(\varphi) &= \tilde{\mathcal{L}}'_{\gamma,u}(q,u,z)(\varphi),\\
\tilde{\rho}^q(q,u,z)(\varphi) &= \tilde{\mathcal{L}}'_{\gamma,q}(q,u,z)(\varphi).
\end{aligned}$$

Lemma 4.6. *Let (q_γ, u_γ), $(q_{\gamma k}, u_{\gamma k})$, $(q_{\gamma kh}, u_{\gamma kh})$, $(q_{\gamma\sigma}, u_{\gamma\sigma})$ be stationary points of $\mathcal{L}_\gamma$ or $\tilde{\mathcal{L}}_\gamma$, respectively, which are assumed to be three times Gateaux differentiable functionals. Then, with arbitrary $\hat{z}_{\gamma k}, \hat{u}_{\gamma k} \in \tilde{X}_k$, $\hat{z}_{\gamma kh}, \hat{u}_{\gamma kh} \in \tilde{X}^{r,s}_{kh}$, $\hat{q}_{\gamma\sigma} \in Q_d$ the following representation formulas*

hold,

$$\begin{aligned}
J_\gamma(q_\gamma,u_\gamma)-J_\gamma(q_{\gamma k},u_{\gamma k}) = &\ \frac{1}{2}\tilde\rho^u(q_{\gamma k},u_{\gamma k})(z_\gamma-\hat z_{\gamma k})+\frac{1}{2}\tilde\rho^z(q_{\gamma k},u_{\gamma k},z_{\gamma k})(u_\gamma-\hat u_{\gamma k})+\mathcal{R}_k\\
J_\gamma(q_{\gamma k},u_{\gamma k})-J_\gamma(q_{\gamma kh},u_{\gamma kh}) = &\ \frac{1}{2}\tilde\rho^u(q_{\gamma kh},u_{\gamma kh})(z_{\gamma k}-\hat z_{\gamma kh})\\
&\ +\frac{1}{2}\tilde\rho^z(q_{\gamma kh},u_{\gamma kh},z_{\gamma kh})(u_{\gamma k}-\hat u_{\gamma kh})+\mathcal{R}_h\\
J_\gamma(q_{\gamma kh},u_{\gamma kh})-J_\gamma(q_{\gamma\sigma},u_{\gamma\sigma}) = &\ \frac{1}{2}\tilde\rho^q(q_{\gamma\sigma},u_{\gamma\sigma},z_{\gamma\sigma})(q_{\gamma kh}-\hat q_{\gamma\sigma})+\mathcal{R}_d,
\end{aligned}\tag{4.42}$$

where the remainder terms $\mathcal{R}_k,\mathcal{R}_h,\mathcal{R}_d$ are of third order and take a form analog to $\mathcal{R}_{reg}$ in (4.38)*.*

The proof is completely analog to the one in [65].

Comparing the already treated terms with the plan laid out in (4.33) it is found that the summands missing a proper representation are the six involving the penalty functional b_γ. These are treated by the following Taylor expansion:

$$\begin{aligned}
b_\gamma(u_{\gamma k})-b_\gamma(u_\gamma)&\approx -b'_\gamma(u_{\gamma k})(u_\gamma-u_{\gamma k})\\
b_\gamma(u_{\gamma kh})-b_\gamma(u_{\gamma k})&\approx -b'_\gamma(u_{\gamma kh})(u_{\gamma kh}-u_{\gamma k})\\
b_\gamma(u_{\gamma\sigma})-b_\gamma(u_{\gamma kh})&\approx -b'_\gamma(u_{\gamma\sigma})(u_{\gamma\sigma}-u_{\gamma kh})
\end{aligned}\tag{4.43}$$

The next step towards the definition of computable error estimators is the approximation of the weights from (4.42) and (4.43). Like in the elliptic case, in Section 3.5, higher order interpolation is used. We use linear operators P_k,P_h,P_d to approximate

$$\begin{aligned}
u_\gamma-\hat u_{\gamma k}\approx P_k u_{\gamma k}\qquad & z_\gamma-\hat z_{\gamma k}\approx P_k z_{\gamma k}\quad q_{\gamma kh}-\hat q_{\gamma\sigma}\approx P_d q_{\gamma\sigma}\\
u_{\gamma k}-\hat u_{\gamma kh}\approx P_h u_{\gamma kh}\quad & z_{\gamma k}-\hat z_{\gamma kh}\approx P_h z_{\gamma kh},
\end{aligned}$$

and analog for the weights from (4.43). We follow [65] further to choose the operators: Naturally the operator P_k should depend on the degree r of the dG(r)-method of temporal discretization. For the implemented dG(0)-method the operator $P_k=I_k^{(1)}-\mathrm{id}$ is chosen, where $I_k^{(1)}\colon \tilde X_k^0\to X_k^1$ is an interpolation operator into the space of continous and piecewise linear functions in time, explicitly given by

$$I_k^{(1)}v|_{I_m}=v_{m-1}^-+\frac{t-t_{m-1}}{t_m-t_{m-1}}(v_m^--v_{m-1}^-)\qquad\text{for } v\in\tilde X_k^0.\tag{4.44}$$

Considering the spatial operator P_h, if the spatial discretization is done by the cG(1) method, we can use the operator $I_{2h}^{(2)}$ from Section 3.5 and extend it to time dependent functions by setting

$$(I_{2h}^{(2)}v_{hk})(t):=I_{2h}^{(2)}v_{hk}(t).$$

Then we choose $P_h=I_{2h}^{(2)}-\mathrm{id}$. The combination all these considerations, (4.33), the error representation (4.42), (4.43), and the interpolation operators, gives rise to the definition of the temporal error estimator

$$\eta_k:=\frac{1}{2}\tilde\rho^u(q_{\gamma\sigma},u_{\gamma\sigma})(P_k z_{\gamma\sigma})+\frac{1}{2}\tilde\rho^z(q_{\gamma\sigma},u_{\gamma\sigma},z_{\gamma\sigma})(P_k u_{\gamma\sigma})-b'_\gamma(u_{\gamma\sigma})(P_k u_{\gamma\sigma}),\tag{4.45}$$

and the spatial error estimator

$$\eta_h := \frac{1}{2}\tilde{\rho}^u(q_{\gamma\sigma}, u_{\gamma\sigma})(P_h z_{\gamma\sigma}) + \frac{1}{2}\tilde{\rho}^z(q_{\gamma\sigma}, u_{\gamma\sigma}, z_{\gamma\sigma})(P_h u_{\gamma\sigma}) - b'_\gamma(u_{\gamma\sigma})(P_h u_{\gamma\sigma}). \tag{4.46}$$

Remark 4.3. To define a computable control error estimator, in the current error representation

$$\eta_d \approx \frac{1}{2}\tilde{\rho}^q(q_{\gamma\sigma}, u_{\gamma\sigma}, z_{\gamma\sigma})(P_d q_{\gamma\sigma}) - b'_\gamma(u_{\gamma\sigma})(u_{\gamma\sigma} - u_{\gamma kh})$$

the weight $u_{\gamma\sigma} - u_{\gamma kh}$ remains to be approximated. Since the refinement strategy employed for the numerical calculations in this thesis does not utilize an extra control discretization, but instead couples the discretization of the control variable to that of the state, an application of a control error estimator was not necessary and no approximation of the term has been considered so far. This could be done by at least two possibilities.
One is the treatment of the term $b_\gamma(u_{\gamma\sigma}) - b_\gamma(u_\gamma)$ as an additional quantity of interest, as discussed in Remark 4.2.
Alternatively, one could possibly use the transformation

$$\begin{aligned} b'_\gamma(u_{\gamma\sigma})(u_{\gamma\sigma} - u_{\gamma kh}) &= b'_\gamma(u_{\gamma\sigma})(S_{kh}(q_{\gamma\sigma}) - S_{kh}(q_{\gamma kh})) \\ &\approx b'_\gamma(u_{\gamma\sigma})(S'_{kh}(q_{\gamma\sigma})(q_{\gamma\sigma} - q_{\gamma kh})) \\ &\approx b'_\gamma(u_{\gamma\sigma})(S'_{kh}(q_{\gamma\sigma})(P_d q_{\gamma\sigma})) =: b'_\gamma(u_{\gamma\sigma})(\delta u_d), \end{aligned}$$

where δu_d denotes the solution of the tangent equation (4.28) with the direction $\delta q = P_d q_{\gamma\sigma}$.

Analog to the elliptic case, the error estimates $\eta_\gamma, \eta_k, \eta_h$ can now be used within an adaptive refinement process. The strategies now deal with the four components $\gamma, (I_m)_{m=1}^M, \mathcal{T}_h$ or $(\mathcal{T}_h^m)_{m=0}^M$ and Q_d instead of two. Still, the first decision to make is which structure(s) are refined. Using the error estimates, Algorithm 2.2 can be used to choose a subset of structures to be treated. Alternatively, for test strategies used in numerical experiments, it may be desirable to

- fix all structures to be chosen
- fix a certain subset of the structures to be always refined
- fix a subset to be used within Algorithm 2.2 to choose the structures to be refined.

If the regularization is chosen, then simply γ is increased by a given factor.
For the other components, a second decision has to be made, whether the refinement is to be conducted globally, or locally, by using local error indicators.
For the spatial refinement, the procedure is analog to the elliptic case. For the local refinement of a non-dynamic spatial discretization, the cell- or nodewise error indicators

$$\eta_h = \sum_{K \in \mathcal{T}_h} \eta_{h,K},$$

can be obtained like before. For a dynamic spatial discretization the procedure has to be done for all time steps, leading to a breakdown into indicators

$$\eta_h = \sum_{m=0}^{M} \sum_{K \in \mathcal{T}_h} \eta_{h,K}^m,$$

details can be found in [87]. So for one given $m \in \{0, \dots, M\}$ the indicators $\{\eta_{h,K}^m : K \in \mathcal{T}_h\}$ can be used to refine the mesh $\mathcal{T}_h^m$ as before.
Considering the temporal refinement, global refinement means the obvious bisection of every interval from $(I_m)_{m=1}^M$. For local refinement the localization

$$\eta_k = \sum_{m=1}^{M} \eta_k^m$$

is evaluated, again according to [87]. A choice of a subset of intervals to be refined can be obtained analog to the spatial refinement. Every chosen interval I_m is dissected by including the point $\frac{1}{2}(t_{m-1}+t_m)$ to the set of time points. In the case of dynamic discretization also a new spatial mesh has to be introduced. In the implementation, a copy of the mesh corresponding to the right end point of the interval to be refined is chosen.
The refinement of the control space, if necessary, can be done equivalently by means of a localization of the error estimator η_d.

Like before, these partial aspects can be assembled to an overall refinement strategy. A *global* strategy would refine always all components globally. The fully *adaptive* strategy consists of the application of the error equilibration strategy Algorithm 2.2 which in turn uses Algorithm 4.2 and Algorithm 4.3 as local refinement strategies for spatial and temporal refinement, if needed. Intermediate versions between the fully global and adaptive ones can be set up as well.

Algorithm 4.2. Local refinement of the spatial discretization for parabolic OCPs

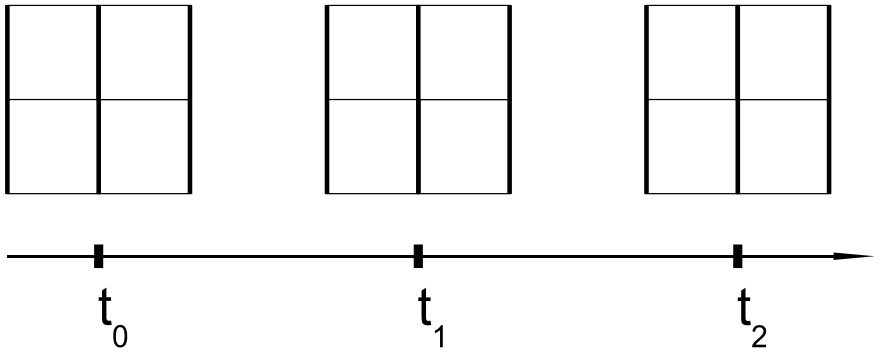
1: **input data**: mesh $\mathcal{T}_h$ or $(\mathcal{T}_h^m)_{m=0}^M$ (dynamic)
2: evaluate localization $\eta_h = \sum_{K\in\mathcal{T}_h} \eta_{h,K}$ or $\eta_h = \sum_{m=0}^{M} \sum_{K\in\mathcal{T}_h} \eta_{h,K}^m$
3: sort these eror indicators by their absolute values:
$|\eta_{h,K_1}| \geq |\eta_{h,K_2}| \geq \dots$ (nondynamic)
analog for every time point m independently (dynamic)
4: find the index $i = \arg\min_{1\leq i\leq \mathcal{N}_h} \mathcal{E}(i)\mathcal{N}(i)^\delta$ (nondynamic)
analog find an index i_m for every time point m independently (dynamic)
5: mark cells $K_1 \dots K_i$ for refinement (dynamic), or $K_1 \dots K_{i_m}$ for every time point m
6: refine the marked cells by evenly partition, together with all the cells from the same patches
7: **repeat**
8: iterate over all cells $K \in \mathcal{T}_h$, or time levels $m = 0 \dots M$ and cells $K \in \mathcal{T}_h$ (dynamic):
9: if the current cell K has a neighboring cell such that it has more than one hanging node on the shared face: refine K, together with its patch
10: **until** no such pair of cells exists any more
11: **output data**: mesh $\bar{\mathcal{T}}_{\bar{h}}$ or $\left(\bar{\mathcal{T}}_{\bar{h}}^m\right)_{m=0}^M$ (dynamic)

Remark 4.4. When following the temporal course of a numerical solution obtained with dynamic discretization, it can occur that, possibly restricted to a part of the domain Ω, the spatial discretization at a later time point is coarser than it has been at an earlier time point. Sometimes this behavior is referred to as coarsening, as the calculation has „started out“ with a fine spatial discretization and has „progressed“ to a coarser one.

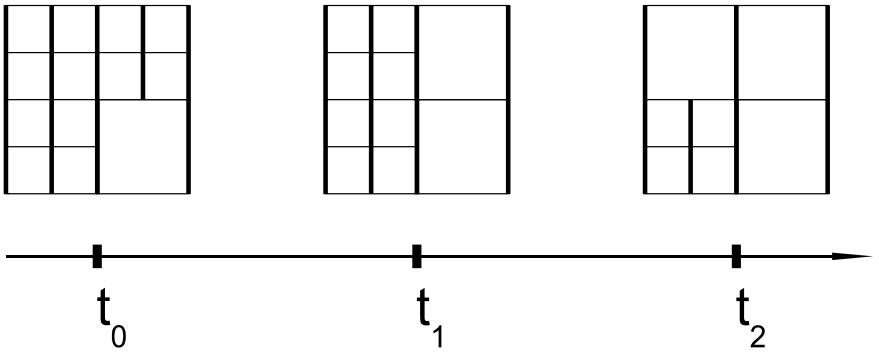
Algorithm 4.3. Local refinement of the temporal discretization for parabolic OCPs

1: **input data**: $(I_m)_{m=1}^{M}$, and possibly $(\mathcal{T}_h^m)_{m=0}^{M}$ (dynamic)
2: denote the time points separating the intervals I_m by $0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = T$
3: evaluate localization $\eta_k = \sum_{m=1}^{M} \eta_k^m$.
4: sort η_k^m by their absolute values: $|\eta_k^{m_1}| \geq |\eta_k^{m_2}| \geq \ldots$
5: find the index i analog to (3.56)
6: **for** every interval I_{m_j} with $j \leq i$ **do**
7: insert midpoint $\frac{1}{2}(t_{m_j-1} + t_{m_j})$ into set of time points
8: if dynamic discretization: construct new spatial FE space connected to this new time point, copy the mesh $\mathcal{T}_h^{m_j}$
9: get new intervals $\left(\bar{I}_m\right)_{m=1}^{\bar{M}}$ from set of time points
10: **output data**: $\left(\bar{I}_m\right)_{m=1}^{\bar{M}}$, and possibly $\left(\bar{\mathcal{T}}_{\bar{h}}^m\right)_{m=0}^{\bar{M}}$

This notion does however not fit into the tighter sense of the concept of coarsening a discretization, as the spatial discretizations at two subsequent time points on the same discretization level T_i are not derived from one another. They rather stem from the spatial discretizations at the two time points used at the former discretization level T_{i-1}. We only allow for refinement in this step, $V_h^{s,m,(i)} \supset V_h^{s,m,(i-1)}$, but not coarsening. Consequently it follows that $X_{kh}^{(i)} \supset X_{kh}^{(i-1)}$, which would not hold for true coarsening. Figure 4.1 illustrates this point about the word *coarsening* within dynamic discretization.



(a) Starting discretization: the meshes corresponding to times t_0 (left), t_1 (middle) and t_2 (right) are equal.



(b) Dynamically refined discretization: the meshes corresponding to $t_n,\ n = 0, 1, 2$, have been obtained by refinement of the respective meshes in Figure 4.1(a). As no nodes, or degrees of freedom, have been removed, no coarsening has taken place, although viewing only the meshes in Figure 4.1(b) in their timely order could suggest otherwise.

Figure 4.1.. Refinement of a dynamic spatial discretization

5. Aspects of Implementation

This chapter deals with implementational issues of the algorithms proposed before. While the main ingredients have been derived and discussed thoroughly, some practical points remain to be clarified. Also, alternatives to some aspects within the complete algorithm will be discussed.

5.1. Complete algorithm

Until now, the single aspects of the solution process have been discussed seperately from each other. A composition of these ingredients into one general optimization algorithm for the

Algorithm 5.1. Optimization algorithm - general

---

input data: a problem of type (1.1)
parameter: tolerances TOL_E, TOL_C for error and computational time, TOL_N, TOL_L
Set i=0
Choose starting discretization $T_0 = (\mathcal{T}_h^{(0)}, Q_d^{(0)})$ (elliptic)
or $T_0 = \left(\left(I_m^{(0)} \right)_{m=1}^{M}, \mathcal{T}_h^{(0)}, Q_d^{(0)} \right)$ (parabolic)
or $T_0 = \left(\left(I_m^{(0)} \right)_{m=1}^{M}, \left(\mathcal{T}_h^{m,(0)} \right)_{m=0}^{M}, Q_d^{(0)} \right)$ (parabolic, dynamic).
If optimization method $\neq$ PDAS, choose starting regularization parameter γ_0
repeat
  Set up the fully discretized problem (P_σ) (3.27) or $(P_{\gamma\sigma})$ (4.25)
  with discretization T_i implying the spaces $X_h^{(i)}$ or $\tilde{X}_{kh}^{s,(i)}$, $\mathcal{M}_h^{(i)}, Q_d^{(i)}$.
  Choose starting control $q_0^{(i)} \in Q_d^{(i)}$. If PDAS, choose $\mu_0^{(i)} \in \mathcal{M}_h^{(i)}$.
  Solve with PDAS$(q_0^{(i)}, \mu_0^{(i)}, TOL_N, TOL_L)$, see Algorithm 3.1
  or solve with IP$(q_0^{(i)}, \gamma_i, TOL_N, TOL_L)$, see Algorithm 4.1
  This yields discrete solution $\bar{u}^{(i)}, \bar{z}^{(i)} \in X_h^{(i)}$ or $\tilde{X}_{kh}^{s,(i)}$, $\bar{q}^{(i)} \in Q_d^{(i)}$
  and possibly $\bar{\mu}^{(i)} \in \mathcal{M}_h^{(i)}$
  Evaluate a posteriori error estimator (3.54) or (4.34) resp., giving $\eta^{(i)}$
  if $(|\eta^{(i)}| \leq TOL_E)$ OR (computational time $\geq TOL_C$) **then**
    BREAK
  Use equilibration strategy with input T_i, γ_i, see Algorithm 2.2
  This yields T_{i+1}, γ_{i+1}.
  Set $i = i + 1$.
until false
output data: $(\bar{q}^{(i)}, \bar{u}^{(i)})$ as approximate solution of (1.1)

---

efficient approximate solution of problem (1.1) from start to finish is conducted in Algorithm 5.1. In the following the execution of step 7, the choice of the starting control, will be concretized, as the general formulation in Algorithm 5.1 may leave questions. A good choice of the starting value of a control $q_0^{(i)}$ on the discretization level i shall possess two advantageous properties: it has to be admissible, and it is preferrably in close proximity to the mesh-optimal control, to allow for immediate superlinear convergence of the optimization algorithm.

On the first discretization level, T_0, the existence of an admissible control is secured by Assumptions 3.5 and 4.3. Mathematical information on the mesh-optimal control has not been retrieved yet, so one has to pass on the proximity property, unless information from an applicational background can be utilized to guess an acceptable control.
On the subsequent discretization levels $T_i, i \geq 1$, it is possible to get a starting control by interpolation of the mesh-optimal control $\bar{q}^{(i-1)}$ from the last level. Naturally it can be expected to be close to the optimal control on the new discretization. So if $I_q^{(i)} \colon Q_d^{(i-1)} \to Q_d^{(i)}$ is an interpolation operator, we choose

$$q_0^{(i)} := I_q^{(i)} \bar{q}^{(i-1)} \tag{5.1}$$

as the starting value. For parameter optimization, where Q is finite dimensional to begin with, this step does not apply ($I_q^{(i)} = \mathrm{id}$).

Example 5.1. In the case of a spatially distributed control, the operator $I_q^{(i)} \colon Q_d^{(i-1)} \to Q_d^{(i)}$ could be chosen as the identity mapping on the linear finite element functions. This is achieved by the following construction: For every node $x_j \in \mathcal{N}^{(i)}$, the value of the control on the new level is set to

$$I_q^{(i)} \bar{q}^{(i-1)}(x_j) = \begin{cases} \bar{q}^{(i-1)}(x_j) \colon & x_j \in \mathcal{N}^{(i-1)} \\ \frac{1}{|\mathcal{N}_j^{(i)}|} \sum\limits_{x_k \in \mathcal{N}_j^{(i)}} \bar{q}^{(i-1)}(x_k) \colon & x_j \notin \mathcal{N}^{(i-1)} \end{cases}, \tag{5.2}$$

where $\mathcal{N}_j^{(i)}$ denotes the set of neighboring nodes of x_j, these are the closest nodes of the „parent“ patch, the refinement of which defined x_j.

This proposed choice of starting control does however not solve the problem of admissibility: in general the value $I_q^{(i)} \bar{q}^{(i-1)}$ is not admissible. If the optimization method is PDAS, then this is not a problem, an iterior-point method however requires an admissible starting control. This problem occurs in analog form for control constrained OCPs, but can there be solved easily by the projection of $I_q^{(i)} \bar{q}^{(i-1)}$ on the admissible set: for control constrained problems Q_{ad} is explicitly given, so that the construction of a projection onto the discretized set $Q_{ad,h}$ is usually a simple task. This is contrary to state constrained problems, where the set of admissible controls is not given explicitly. In fact, the only exploitable information on the interior of the admissible set is in general the value $q_0^{(0)}$ on the first discretization level. This may be utilized to construct a feasible control as follows:
Assume the interpolation of $q_0^{(0)}$ in the set $Q_d^{(i)}$ yields an admissible control, that will be denoted by $\mathring{q}^{(i)}$. Since $\mathring{q}^{(i)}$ may be outside the fast convergence neighborhood of the exact solution, a convex linear combination of $I_q^{(i)} \bar{q}^{(i-1)}$ and $\mathring{q}^{(i)}$ close to the former is taken as a strong candidate for a close, but admissible starting control. So set

$$q_{0,k}^{(i)} := (1 - \lambda_k) I_q^{(i)} \bar{q}^{(i-1)} + \lambda_k \mathring{q}^{(i)}, \qquad k = 0, 1, \ldots \tag{5.3}$$

until for some K the control $q_{0,K}^{(i)}$ is admissible, then choose $q_0^{(i)} := q_{0,K}^{(i)}$ as starting control. The λ_k can be chosen as $k \cdot c$ with a small constant $0 < c \ll 1$. As numerical experience shows, with advancing refinement the violation of the state constraint decreases. Since it is advantageous to choose the final λ_k as small as possible, it is also possible to make the choice $\lambda_k = k \cdot c^{(i)}$ with a level-dependent (decreasing) factor $c^{(i)}$: If for one discretization level T_i the linear combination with $k = 1$ gave an admissible control (but $k = 0$ did not), then decrease the factor, e.g. by setting $c^{(i+1)} = \frac{1}{2}c^{(i)}$.
A finer tuning of the values of $c^{(i)}, \lambda_k$ etc. is usually not worth the effort, as the admissibility test of the resulting controls is usually too expensive.

5.2. Implementation of Borel measures

In the framework of Algorithm 5.1 the fully discretized optimization problems may be solved by different methods. If the primal-dual active set method is chosen, the Lagrange multiplier μ needs to be introduced as a new system variable into the computational treatment.

In the overall agorithm in step 7 a starting value $\mu_0^{(i)}$ has to be chosen just like for the control $q_0^{(i)}$ as considered in Section 5.1. On the first mesh T_0, the consideration for the multiplier applies analog: since there is no prior knowledge, one may just take any value, unless technical background suggests otherwise. For the subsequent levels, since admissibility is not a requirement for the starting solution for the PDAS algorithm the correction part related to (5.3) can be omitted and one just sets $q_0^{(i)} := I_q^{(i)} \bar{q}^{(i-1)}$.

The analog setting of a multiplier

$$\mu_0^{(i)} := I_\mu^{(i)} \bar{\mu}^{(i-1)} \tag{5.4}$$

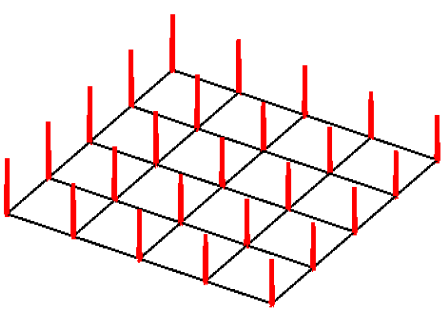
requires some care in the construction of the interpolation operator $I_\mu^{(i)}$. The choice $I_\mu^{(i)} = I_q^{(i)}$ as the interpolation of the multiplier analog to (5.2) may be easy to implement as the operators work on nodal vectors regardless. But since $I_q^{(i)}$ is the identity operator on the space of linear finite element functions, this choice seems unnatural. A possible alternative to this interpolation is the use of the operator

$$I_\mu^{(i)} \bar{\mu}_j^{(i-1)} = \begin{cases} \bar{\mu}_j^{(i-1)}: & x_j \text{ is node of } T_{i-1} \\ 0: & \text{else} \end{cases}, \tag{5.5}$$

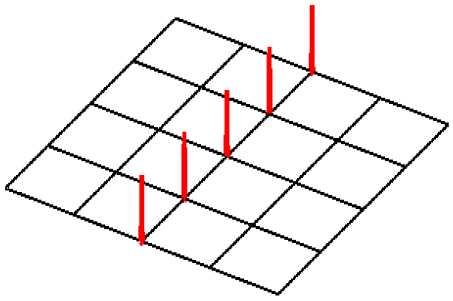
which is the identity operator on the space of discrete Borel measures $\mathcal{M}_h^{(i-1)}$.

The decision whether to choose interpolation according to (5.2) or according to (5.5) thus depends on the structure of the multiplier. This can be explained by considering the effect of these two operators on the discretizations of two prototypical examples of multipliers - the first is a constant function, the second is a line measure - see Figure 5.1 for the discretizations on a coarse grid. Global refinement of the mesh and subsequent interpolation of these multipliers using (5.2) leads to the situation depicted in Figure 5.2. While the support of the regular part is realized correctly, the line measure part is hugely overapproximated, it exhibits three times as many active nodes as necessary. This does not hinder the convergence of the algorithm, but the convergence speed is slowed down considerably. The situation when using the interpolation (5.5)

is depicted in Figure 5.3. This method underestimates the support of both the regular and the measure part of the multiplier. Still, numerical experience shows interpolation by (5.5) to work considerably better. This seems plausible, as the introduced mismatch of the support is smaller, also active sets are usually approached from the outside.

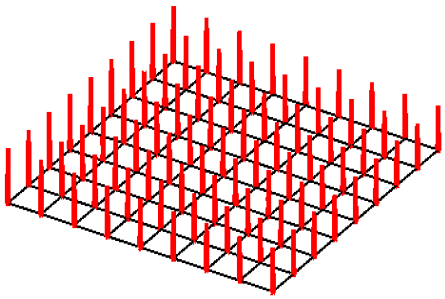


(a) discrete measure $\mu_h \in \mathcal{M}_h$ as discretization of the constant function $\mu \in L^2(\Omega)$

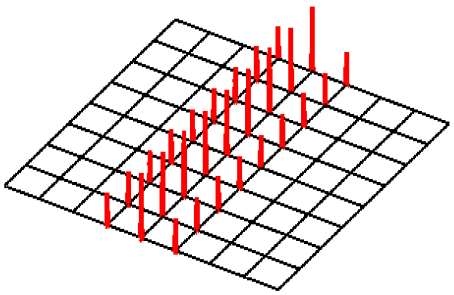


(b) discrete measure $\mu_h \in \mathcal{M}_h$ as discretization of the line measure $\mu \notin L^2(\Omega)$

Figure 5.1.. Discrete Borel measures. The height of the bar in node x_i represents the value of the related coefficient μ_i.

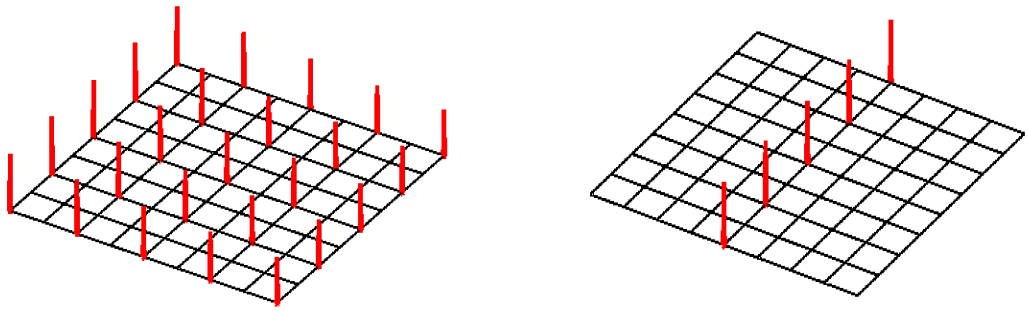


(a) applied to the discretization of the constant function $\mu \in L^2(\Omega)$



(b) applied to the discretization of the line measure $\mu \notin L^2(\Omega)$

Figure 5.2.. Result of the interpolation according to (5.2) applied to the measures from Figure 5.1.



(a) applied to the discretization of the constant function $\mu \in L^2(\Omega)$

(b) applied to the discretization of the line measure $\mu \notin L^2(\Omega)$

Figure 5.3.. Result of the interpolation according to (5.5) applied to the measures from Figure 5.1.

5.3. Possible modifications of the standard algorithm

In this section possible modifications of Algorithm 5.1 will be discussed. In the buildup of this algorithm, in the previous chapters the following typical procedures were discussed:

- (adaptive) discretization methods,
- methods to solve a nonlinear (optimization) problem, and
- methods to solve the linear subproblems generated by the nonlinear method.

Different methods that realize each of these specific steps are available. For example, the discretization can be governed by

- a uniform refinement strategy, or
- an adaptive refinement strategy
  - using the dynamic mesh approach, or
  - using only one spatial mesh.

The optimization method may be

- the primal dual active set method, or
- the interior point method.

For linear solvers there is a large amount of possibilities, in this thesis only the CG and GMRes methods were mentioned.
The setup of these procedures to form the complete method in this thesis is as follows: The steps are executed in a nested loop in the following order

1. adaptive discretization,

2. $\rightarrow$ nonlinear optimization,

3. $\longrightarrow$ linear subproblems.

An alternative to this nested iteration is discussed in [42, 85]. Switching the order of the loops to

1. nonlinear optimization,

2. $\rightarrow$ adaptive discretization,

3. $\longrightarrow$ linear subproblems

leads to the application of adaptivity for linear problems, but requires the use of the nonlinear solver in function space.

Another modification concerns problems with a two-sided state constraint. So consider the problem

$$\begin{cases} \min J(q,u) & q \in Q, u \in X \\ u = S(q) & \\ u_a(x) \leq u(x) \leq u_b(x) & \forall x \in \bar{\Omega} \text{ or } \bar{\Omega} \times \bar{I} \end{cases} \tag{5.6}$$

where we assume the active sets of the state constraints to be separated. This can be achieved by securing

$$u_b > u_a \text{ on } \bar{\Omega} \text{ or } \bar{\Omega} \times \bar{I}.$$

The standard conversion of the two-sided constraint like indicated in Section 2.1.3 would utilize two multipliers, $\mu_a \in \mathcal{M}(\Omega)$ or $\mathcal{M}(I \times \Omega)$ associated to the partial constraint $u_a(x) \leq u(x)$, and $\mu_b \in \mathcal{M}(\Omega)$ or $\mathcal{M}(I \times \Omega)$ associated to $u(x) \leq u_b(x)$.
An improvement of this approach can be achieved by observing the known sign of the multipliers: As optimality conditions require $\mu_a, \mu_b \geq 0$, and the implementation dictates the positive sign for all components at every computational step, a common multiplier can be defined by

$$\mu := \mu_b - \mu_a. \tag{5.7}$$

Now the implementation can be done with the single multiplier μ instead of both μ_a and μ_b, thus saving computational effort. A component of μ with a negative value corresponds to a point where the lower bound is active, a positive component corresponds to the upper bound. This construction works due to the separation of the active sets.

The third modification of the solution process is the transformation of nonlinear boundaries to constant ones. Consider the most general problem (1.1). If the inverse of the constraint function, G^{-1}, exists and is monotone decreasing, then it can be easily shown that the relations

$$G(u) \geq 0 \quad \Leftrightarrow \quad u - G^{-1}(0) \leq 0$$

are equivalent and thus the transformation

$$\hat{u} := G^{-1}(0) - u \tag{5.8}$$

leads to the following problem that is equivalent to (P):

$$\begin{cases} \min J(q, G^{-1}(0) - \hat{u}) & q \in Q, \hat{u} \in X \\ \hat{u} = G^{-1}(0) - S(q) & \\ \hat{u} \geq 0 & . \end{cases}$$

A similar construction is possible for a monotone increasing inverse of the constraint function, with

$$\hat{u} := u - G^{-1}(0)$$

being the appropriate transformation.
The advantage of this formulation is that due to the constant bounds the discretization of the constraint function G becomes trivial.

Example 5.2. For the single upper bound $u \leq u_b$ the transformation is

$$\hat{u} := u_b - u.$$

The problem is then transformed as follows:

$$\begin{cases} \min J(q, u) & q \in Q, u \in X \\ u = S(q) & \\ u \leq u_b & \end{cases} \quad \Rightarrow \quad \begin{cases} \min J(q, u_b - \hat{u}) & q \in Q, \hat{u} \in X \\ \hat{u} = u_b - S(q) & \\ \hat{u} \geq 0 & \end{cases}$$

Example 5.3. For the two-sided constraint $u_a \leq u \leq u_b$, the transformation introduced in (5.8) can not be applied directly. Still, the transformation

$$\hat{u} := \frac{u - u_a}{u_b - u_a},$$

obtained in a similar way, leads to the following equivalent problem,

$$\begin{cases} \min J(q, u) & q \in Q, u \in X \\ u = S(q) & \\ u_a \leq u \leq u_b & \end{cases} \quad \Rightarrow \quad \begin{cases} \min J(q, \hat{u}(u_b - u_a) + u_a) & q \in Q, \hat{u} \in X \\ \hat{u} = \frac{S(q) - u_a}{u_b - u_a} & \\ 0 \leq \hat{u} \leq 1 & , \end{cases}$$

which obtains the goal of smooth constraint functions as well.

5.4. Considerations derived from practical problems

In Chapter 7 the methods developed in this thesis are applied to the large-scale practical problem of the control of structural properties developed during the hydration phase of young concrete. A wide range of different practical problems has been treated by modelling the task as an optimal control problem. Studying the utilized solution processes reveals aspects that have not been emphasized yet. In the following, we give a few examples of practical optimal control problems, and links to publications:

- In hypothermia cancer treatment, see [96], the computational domain is a part of the patient's body. The forward operator includes the heat as well as electric field equations, optimization is done over parameters. State constraints arise naturally, as sound tissue has to be protected against too much heat.

- The problem of optimal glass cooling, investigated, e.g., in [22] has the goal to guide the temperature of a glass melt to room temperature by adjusting the furnace temperature, which is modelled as boundary control. The control aspires to guide the temperature of the glass to a given temperature profile, which is chosen for a minimum of unwanted stresses, that are building up during the cooling process.

- In the surface hardening of steel [65], a laser beam is moved along the surface of a workpiece, inducing heat and with that the formation of austenite, causing the hardening effect. The aim is to control the laser energy such that a desired hardening profile is reached. The control-to-state operator consists of the heat equation coupled with an ordinary differential equation in every spatial point that describes the formation of austenite depending on the temperature.

An uncritical application of the solution process to practical problems like these as described until now will solve those problems, but it can be improved easily by taking some further aspects of the practical problem into account. These can comprise of the following:

The choice of the starting mesh $\mathcal{T}_h^{(0)}$ at the beginning of the first discretization process holds the opportunity to improve the numerical behavior via the following consideration: The spatial extension of the computational domain can be enormous. This leads to large discrete problems even on a relatively coarse discretization level. It is thus unfavorable to choose an equidistant mesh as $\mathcal{T}_h^{(0)}$. Typically in practical problems the solution exhibits structures on very different spatial or temporal size scales so that a uniform discretization would either result in huge discrete problems or the loss of fine-scale information. Instead, based on an a priori understanding of the physical process in question, a starting mesh can be designed that resolves the structure of the solution well with minimal effort.

Another aspect is that in practical problems the computational domain Ω is often nonconvex. This can lead to additional loss of regularity of the solution, and so reduce the accuracy and convergence speed of the algorithms explained here. To counteract this phenomenon, a priori mesh grading might be used. This means that the starting mesh is constructed to have a finer discretization in a neighborhood of the edge or corner in question, i.e. the cells' diameter depend on the distance to the reentrant edge or corner. For some basic problem classes, this dependency takes a form analog to (A.4), where the distance to the point a corresponds to the distance to the critical structure. A detailed analysis of this concept for elliptic problems in two and three dimensions can be found in [4–6]. A grading of the mesh towards the origin can then be obtained by applying the transformation

$$T(x) = x\|x\|^{\frac{1}{\mu}-1}$$

to all points in a neighborhood of the critical structure, where $\mu \in (0, 1]$ denotes the grading parameter.

In many practical applications, more than one physical quantity is needed to describe the state of the system. This corresponds to a state variable with more than one component, $u\colon \bar{\Omega} \to \mathbb{R}^{n_s}$ or $u\colon \bar{I} \times \bar{\Omega} \to \mathbb{R}^{n_s}$. For reasons of simplicity, this is not explicitly treated in this thesis. For simple problems it suffices to treat the other components like the first one. Especially the a posteriori error estimation is not affected by a greater number of components. Only for more difficult problems it may be necessary to use more involved techniques; also this affects mainly questions of discretization and implementation. For flow control problems, for example, it can be necessary to use stabilization techniques to obtain a useable discretization.

Furthermore, in practical problems it often occurs that uncertainties on the input data are relatively large. Therefore it may be unrealistic to expect computational results that are in very accurate accordance to reality observations. Then sometimes in the numerical algorithms the focus is shifted away from high order accuracy, convergence rates and related. Instead the expectations on speedup of calculations in the first refinement steps are increased. This approach can be supported by adaptive mesh refinement based on a posteriori error estimation.

6. Numerical Results

In this chapter some example problems will be considered that fit into the general framework (3.1) or (4.1), respectively. They will be solved numerically by the algorithms developed in the previous chapters. The employed optimization method on one level of discretization for the elliptic problems is the PDAS method described in Section 3.4. For the parabolic problems, a regularization by the barrier approach and subsequently the optimization by the interior point method is applied, as described in Section 4.4. In one example, both optimization methods are applied to make a comparison between them.

The generality of the setting allows for many different constellations. Thus example problems of different types are chosen to illustrate the possible differences. Specifically, linear and nonlinear problems, elliptic and parabolic ones, and such with known and unknown exact solutions are considered. Additionally, attention is payed to the structure of the optimal solution; this means the active sets can e.g. be points, lines or twodimensional sets.

One goal of the numerical experiments is to evaluate the quality of the error estimators derived previously. Two aspects will be evaluated: the first is how good the estimated value η matches the overall error in the cost functional. For this the effectivity index

$$I_{\text{eff}} = \frac{J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma)}{\eta} \tag{6.1}$$

is defined. It should be evaluated for every error estimation for the mesh-optimal solution on every considered level of discretization. To evaluate I_{eff}, the value of the cost functional is needed, which is only known if the optimal solution $(\bar{q}, \bar{u})$ is known. If $J(\bar{q}, \bar{u})$ is not available, it is replaced by a precalculated value $J^* := J(q_{\sigma^*}, u_{\sigma^*})$, where $(q_{\sigma^*}, u_{\sigma^*})$ is the optimal solution to a discrete problem that is on every level finer discretized than the problems I_{eff} is to be evaluated for, so that the small inaccuracy does not have a sizeable influence. To judge the quality of the error estimator, a proximity of I_{eff} to 1 would be best. Strictly this can be expected only for the limit $h \to 0$ ($k \to 0, \gamma \to \infty$). So if the discretization is relatively coarse, relative to the difficulty of the problem in question, sizable deviations from this value cannot be excluded at all; especially changes in the sign of the error are often an indicator for these deviations.

Another indicator for the quality of the a posteriori error estimation is the effect of the local refinement of the discretizations, respectively the error equilibration. Solving the numerical problems with different discretization strategies, for example once on a series of uniformly refined discretizations, and once including the adaptation, we are able to compare the numerical effort relative to the cost functional error.

The implementation of the solution methods and error estimation were all done in the optimization toolkit RoDoBo [80] in connection with the finite element library Gascoigne [36]. For visualization we used the visualization tool VisuSimple [95].

6.1. Elliptic problem with known exact solution

As the first example problem the following elliptic optimal control problem governed by a linear state equation is considered:

$$(Ex_1)\quad \begin{cases} \text{Minimize} & J(q,u) = \dfrac{1}{2}\|u-u_d\|^2_{L^2(\Omega)} + \dfrac{1}{2}\|q\|^2_{L^2(\Omega)}, \\ -\Delta u = q+f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_1, \\ \partial_n u = 0 & \text{on } \Gamma_2, \\ G(u) \geq 0 & \text{in } \bar{\Omega}, \end{cases}$$

where $\Omega = (0,1)^2$ is the unit square, and the parts of the boundary are

$$\Gamma_1 = \{\, x = (x_1,x_2) \in \partial\Omega \mid x_1 = 0 \,\} \quad \text{and} \quad \Gamma_2 = \partial\Omega \setminus \Gamma_1.$$

To integrate this problem into the framework provided in Chapter 2 the spaces are chosen as $Q = R = L^2(\Omega)$, $V = H^1_{\Gamma_1}(\Omega) = \{\, v \in H^1(\Omega) \mid v|_{\Gamma_1} = 0 \,\}$, $X = W^{1,p}(\Omega) \cap H^1_{\Gamma_1}(\Omega)$. The functions G, f, u_d represent the data of this problem, and will be chosen in such a way that the optimal solution can be calculated explicitly and exhibits some interesting features.

The goal is to construct an optimal solution that fulfills the state constraint exactly with $G(\bar{u}) = 0$ on the set $\left\{ (x_1,x_2) \in \bar{\Omega} \;\middle|\; x_1 \geq s \right\}$ and with $G(\bar{u}) > 0$ on the rest of the domain. So the active and the inactive sets of the optimal solution are seperated by the line $\{x_1 = s\}$ with some parameter $s \in (0,1)$ to be chosen later. As described in [49] from this construction follows a structure of the multiplier $\bar{\mu}$ as the sum of a regular and a line measure part, that is $\bar{\mu} = \bar{\mu}_1 + \bar{\mu}_2$ with $\bar{\mu}_1 \in \mathcal{M}(\Omega) \setminus L^2(\Omega)$ and $\bar{\mu}_2 \in L^2(\Omega)$. The representation

$$\langle \bar{\mu}_1, \varphi \rangle = c_\mu \int_0^1 \varphi(s,x_2)\, dx_2 \;\forall \varphi \in C(\bar{\Omega}), \quad \bar{\mu}_2 = \begin{cases} 0, & x_1 < s \\ b, & x_1 \geq s. \end{cases} \tag{6.2}$$

is employed with a constant c_μ to be determined later and $b > 0$ to be chosen freely.

To construct an optimal solution with the properties described above, upper state constraints are chosen, i.e. $G(u) = u_b - u$, and an ansatz for the optimal state is made demanding that

- $\bar{u}$ is not depending on x_2,
- the restriction of $\bar{u}$ to the active set is a polynomial of degree 4,
- the restriction of $\bar{u}$ to the inactive set is a polynomial of degree 3,
- the transition over the boundary between these sets is C^2, but the third derivative has a jump there,

- $\bar{u}$ fulfills the boundary conditions of (Ex_1).

This determines $\bar{u}$ up to a constant $m \leq s^{-3}$ which is left as a parameter that can later be used to work out the influence of the measure part of the error estimator. The optimal state takes the form

$$\bar{u}(x_1, x_2) = \begin{cases} \frac{x_1^3}{s^3} - 3\frac{x_1^2}{s^2} + 3\frac{x_1}{s}, & x_1 < s \\ -\frac{3m}{4(1-s)}(x_1 - s)^4 + m(x_1 - s)^3 + 1, & x_1 \geq s. \end{cases}$$

The constraint function u_b is thus determined on the active set, and needs to be continued on the inactive set which can be done by any function larger than $\bar{u}$. Considering the boundary conditions on u, the choice

$$u_b(x_1, x_2) = \begin{cases} 1, & x_1 < s \\ -\frac{3m}{4(1-s)}(x_1 - s)^4 + m(x_1 - s)^3 + 1, & x_1 \geq s \end{cases}$$

is made. Setting further

$$u_d(x_1, x_2) := \bar{u} + 2 = \begin{cases} \frac{x_1^3}{s^3} - 3\frac{x_1^2}{s^2} + 3\frac{x_1}{s} + 2, & x_1 < s \\ -\frac{3m}{4(1-s)}(x_1 - s)^4 + m(x_1 - s)^3 + 3, & x_1 \geq s \end{cases}$$

for easier calculations and incorporating the adjoint and the gradient equation with the choice

$$f(x_1, x_2) = \begin{cases} \frac{6}{s^2} - 6mx_1 + x_1(x_1 - 2) + b(1 - s)x_1, & x_1 < s \\ (1 - r)x_1^2 + (b - \frac{18ms}{1-s} - 2 - 6m)x_1 + \frac{6}{s^2} - rs^2, & x_1 \geq s, \end{cases}$$

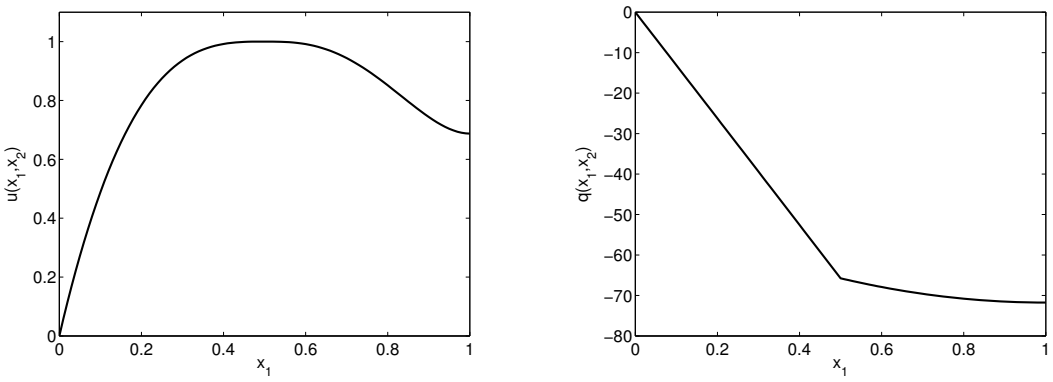
where the abbreviation $r = \frac{b}{2} - \frac{9m}{1-s}$ is used, the missing parts of the exact solution $(\bar{q}, \bar{u}, \bar{z}, \bar{\mu})$ turn out to be

$$\bar{q}(x_1, x_2) = \begin{cases} -x_1(x_1 - 2) - (\frac{6}{s^3} - 6m)x_1 - b(1 - s)x_1, & x_1 < s \\ -x_1(x_1 - 2) - \frac{6}{s^2} + 6ms + \frac{b}{2}x_1^2 - bx_1 + \frac{b}{2}s^2, & x_1 \geq s, \end{cases}$$

$$\bar{z} = -\bar{q},$$

and $\bar{\mu}$ is set according to (6.2) with $c_\mu = \left(\frac{6}{s^3} - 6m\right)$. Remember this construction leaves three parameters $s \in (0, 1), m < s^{-3}, b > 0$ entering the data u_b, u_d, f and the optimal solution of the problem (Ex_1). A visualization of the optimal solution $(\bar{q}, \bar{u})$ for one choice of the parameters can be seen in Figure 6.1. Also observe that the construction of the structure especially of $\bar{\mu}$ is achieved with smooth data $u_b, u_d \in C^2(\Omega), f \in C^1(\Omega)$.

The optimization method used in the following numerical solution of (Ex_1) is PDAS. The starting discretization is always a mesh of 4×4 congruent quadratic cells, used for the discretization of the state and the control space. The refinement strategies used to create a new mesh after a mesh-optimal solution on an old mesh has been found, that will be compared here, are the uniform refinement and the adaptive refinement given by a localization of the error estimator η_h (3.52) in Section 3.5. As the discretization of the control space is tied to that of the state space, η_d is zero and thus omitted, see also Remark 3.4.



(a) optimal state $\bar{u}$ (b) optimal control $\bar{q}$

Figure 6.1.. Optimal solution of (Ex_1) for $(s, m, b) = (0.5, -10, 50)$, dependence on x_1 only, as the functions are x_2-independent.

We present the results of the computations for two sets of parameters (s, m, b). The choice of the parameter $m = -2$ secures that the multiplier part of the error representation (3.52), i.e., $J'_u((P_d + \mathrm{id})q_d, (P_h + \mathrm{id})u_h)(P_h u_h) - a'_u((P_d + \mathrm{id})q_d, (P_h + \mathrm{id})u_h)(P_h u_h, (P_h + \mathrm{id})z_h)$, has a significant size compared to the other parts. For significantly larger m the effects of the reduced regularity originating from the state constraints would be negligible, and the behavior of the numerical solution would be as expected with experimental order of convergence with respect to h being almost exactly 2 and effectivity index almost exactly 1. The choice of the parameter b is less significant, we choose $b = 50$.

The first choice of the remaining parameter is $s = 0.125$. As the exact solution is known, the optimal value of the cost functional $J(\bar{q}, \bar{u}) = 74244.18954366\ldots$ is used to evaluate the discretization error and the effectivity index on every mesh. The results can be seen in Table 6.1, for every discretization level the number of degrees of freedom N, the discretization error $J(\bar{q}, \bar{u}) - J(q_\sigma, u_\sigma)$ and the efficiency index I_{eff} are displayed. The efficiency indices show that an accurate error estimation is observed after the second refinement step. The visualization of the relation between the remaining two quantities in Figure 6.2(a) shows an advantage of the local refinement strategy compared to the uniform strategy in the discretization error relative to the degrees of freedom. As the choice $s = 0.125$ means that the line where the measure $\bar{\mu}_1$ is concentrated is always a grid line, the second test is made with the choice $s = 0.3$. The evaluation of the discretization error and the effectivity index using the new optimal value of $J(\bar{q}, \bar{u}) = 3044.536619\ldots$ in Table 6.2 shows accurate error estimation in most cases. However, the localization of the estimator guides the local refinement process to more efficient meshes, as can be seen in Figure 6.2(b). An example plot of such a mesh is displayed in Figure 6.3. A refinement of the region around $\{x_1 = s\}$ is observed.

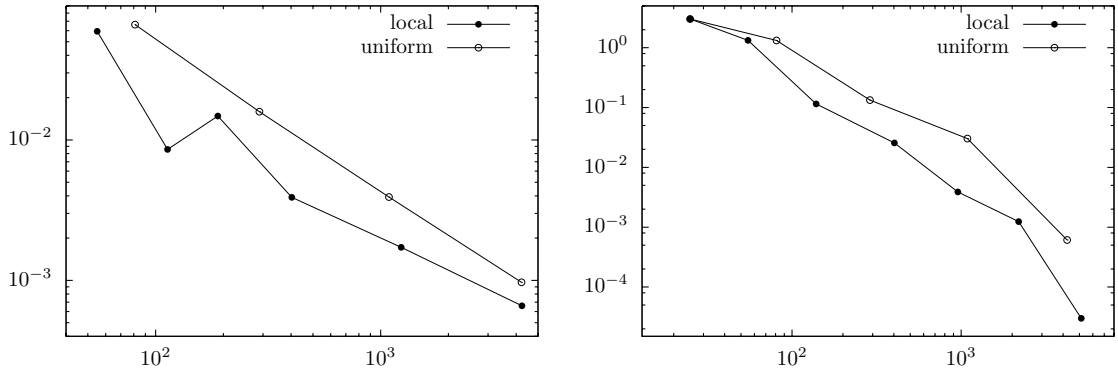
Table 6.1.. Development of discretization errors and of the effectivity indices for $s = 0.125$ for (Ex_1)

(a) adaptive refinement

N	$J(\bar{q},\bar{u}) - J(q_\sigma, u_\sigma)$	I_{eff}
25	1.37e+03	4.54
55	-5.93e-02	0.00
113	-8.56e-03	0.41
189	-1.48e-02	0.94
403	-3.90e-03	0.94
1233	-1.72e-03	0.93
4241	-6.60e-04	0.96

(b) uniform refinement

N	$J(\bar{q},\bar{u}) - J(q_\sigma, u_\sigma)$	I_{eff}
25	1.37e+03	4.54
81	-6.62e-02	0.00
289	-1.59e-02	0.98
1089	-3.92e-03	0.96
4225	-9.70e-04	0.97



(a) $s = 0.125$ (b) $s = 0.3$

Figure 6.2.. Discretization errors vs degrees of freedom for Ex_1

Table 6.2.. Development of discretization errors and of the effectivity indices for $s = 0.3$ for (Ex_1)

(a) adaptive refinement

N	$J(\bar{q},\bar{u}) - J(q_\sigma, u_\sigma)$	I_{eff}
25	3.02e+00	0.65
55	1.33e+00	8.74
139	1.15e-01	1.71
403	2.56e-02	-4.68
955	-3.88e-03	0.96
2185	-1.24e-03	0.78
5125	-2.99e-05	0.81

(b) uniform refinement

N	$J(\bar{q},\bar{u}) - J(q_\sigma, u_\sigma)$	I_{eff}
25	3.02e+00	0.65
81	1.32e+00	8.03
289	1.33e-01	1.52
1089	3.03e-02	-0.45
4225	6.10e-04	1.23

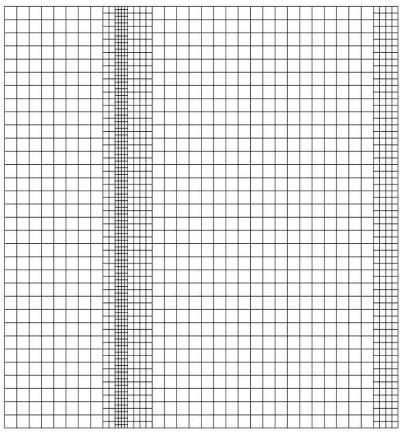


Figure 6.3.. An example of a locally refined mesh for $s = 0.3$ for Ex_1

6.2. Elliptic problem with unknown exact solution

The second example problem takes a form similar to (Ex_1), but this time the data are chosen in such a way that the active set has a curved boundary so it can not be matched by the spatial discretization. Consider on the unit square $\Omega = (0,1)^2$

$$(Ex_2) \quad \begin{cases} \text{Minimize} & J(q,u) = \dfrac{1}{2}\|u - u_d\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\|q\|^2_{L^2(\Omega)}, \\ -\Delta u = q & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma, \\ G(u) \geq 0 & \text{in } \bar{\Omega}, \end{cases}$$

with the data $\alpha = 0.1$, $G(u) = u_b - u$ (upper state constraint) with $u_b = 0.01$, and

$$u_d = 10(\sin(2\pi x_1) + x_2).$$

For this problem, which has been considered in [49], the exact solution is not available, so for the following investigations the approximate optimal value $J^* = 41.62230492265025$ is used, which was computed on a fine mesh with $N = 66049$ nodes. The behavior of the error and effectivity index when using PDAS as optimization method, and adaptive refinement can be seen in Table 6.3(a). Additionally the value of the error estimator η_h itself is displayed, which may seem redundant at this point, but can be compared in magnitude to the estimator contributions from the following tests.

There, (Ex_2) is solved by regularization (order $o = 2$) and interior point method. For this approach two different refinement strategies are considered. In both, the discretization error η_h and the regularization error η_γ are estimated, and the result of this estimation determines whether a new spatial mesh is used, or the regularization parameter is increased, or both, as described in Algorithm 2.2. The difference is in the creation of the new spatial meshes, it can again be created by global, or adaptive refinement of the old mesh. The results can be seen in Table 6.3(b) and Table 6.3(c). The comparison of the error convergence relative

Table 6.3.. Development of discretization errors and of the effectivity indices for (Ex_2)

(a) solution with PDAS

N	η_h	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
9	-9.45e-03	-5.94e-01	62.80
25	-9.04e-03	-1.07e-02	1.18
69	-3.73e-03	-3.19e-03	0.86
97	-3.06e-03	-2.59e-03	0.85
271	-7.60e-04	-8.41e-04	1.11
789	-2.37e-04	-2.50e-04	1.06
2783	-6.46e-05	-6.49e-05	1.01
9817	-1.75e-05	-1.75e-05	1.00

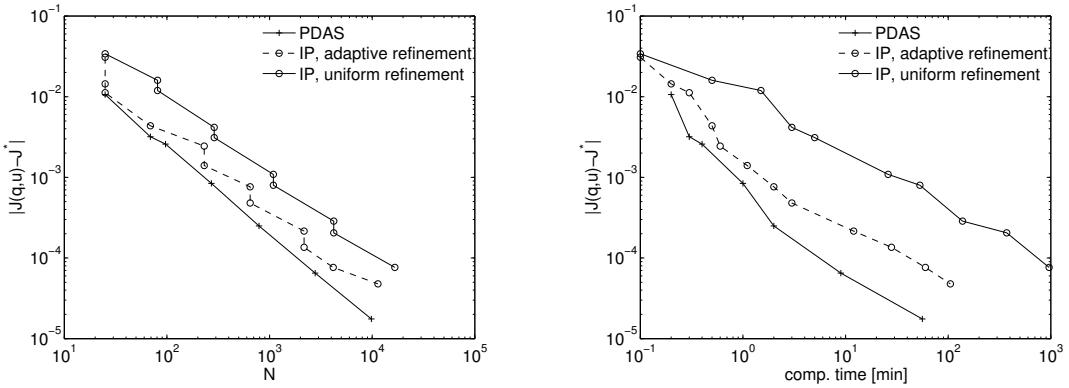
(b) solution with IP, global refinement by component

N	γ	η_h	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	1.0e+02	-1.47e-03	-3.08e-02	-3.22e-02	-3.408e-02	1.06
81	3.2e+02	-2.10e-03	-6.71e-03	-8.81e-03	-1.598e-02	1.81
81	1.0e+03	-2.60e-03	-1.56e-03	-4.16e-03	-1.195e-02	2.87
289	1.0e+03	-5.83e-04	-1.68e-03	-2.27e-03	-4.170e-03	1.84
289	3.2e+03	-6.47e-04	-4.39e-04	-1.09e-03	-3.112e-03	2.87
1089	3.2e+03	-1.79e-04	-4.51e-04	-6.31e-04	-1.088e-03	1.73
1089	1.0e+04	-1.88e-04	-1.23e-04	-3.11e-04	-7.990e-04	2.57
4225	1.0e+04	-4.69e-05	-1.26e-04	-1.73e-04	-2.868e-04	1.66
4225	3.2e+04	-4.83e-05	-3.51e-05	-8.34e-05	-2.051e-04	2.46
16641	3.2e+04	-1.21e-05	-3.58e-05	-4.79e-05	-7.631e-05	1.60

(c) solution with IP, adaptive refinement

N	γ	η_h	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	1.0e+02	-4.82e-03	-2.91e-02	-3.40e-02	-3.084e-02	0.91
25	3.2e+02	-4.51e-03	-5.61e-03	-1.01e-02	-1.443e-02	1.43
25	1.0e+03	-4.43e-03	-1.16e-03	-5.59e-03	-1.122e-02	2.01
69	1.0e+03	-3.00e-03	-1.50e-03	-4.50e-03	-4.362e-03	0.97
231	1.0e+03	-5.75e-04	-1.66e-03	-2.24e-03	-2.444e-03	1.09
231	3.2e+03	-5.89e-04	-4.27e-04	-1.02e-03	-1.403e-03	1.38
647	3.2e+03	-2.36e-04	-4.41e-04	-6.77e-04	-7.635e-04	1.13
647	1.0e+04	-2.46e-04	-1.20e-04	-3.66e-04	-4.807e-04	1.31
2169	1.0e+04	-9.35e-05	-1.23e-04	-2.17e-04	-2.153e-04	0.99
2169	3.2e+04	-9.58e-05	-3.38e-05	-1.30e-04	-1.357e-04	1.05
4173	3.2e+04	-4.21e-05	-3.51e-05	-7.72e-05	-7.630e-05	0.99
11379	3.2e+04	-1.42e-05	-3.57e-05	-4.98e-05	-4.780e-05	0.96

to the number of degrees of freedom is displayed in Figure 6.4(a). The adaptive IP strategy has an advantage over the uniform one. This comparison should be expanded however, as for the regularization approach the number of degrees of freedom is not the only influence on the numerical effort: the regularization parameter γ is as well. Thus in Figure 6.4(b) the computational time is displayed as the quantity to evaluate the error level against. It can also



(a) depending on the degrees of freedom (b) depending on computational time

Figure 6.4.. Convergence of the error for (Ex_2)

be seen that the PDAS method produces better results than the two IP strategies, however this might be problem-dependent.

6.3. Nonlinear elliptic problem

Example problem (Ex_3) on the unit square $\Omega = (0,1)^2$ has a nonlinear state equation and two-sided state constraints:

$$(Ex_3) \quad \begin{cases} \text{Minimize} & J(q,u) = \frac{1}{2}\|u - u_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|q\|^2_{L^2(\Omega)}, \\ -\Delta u + u^3 = q + f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma, \\ u_a \leq u \leq u_b & \text{in } \bar{\Omega}, \end{cases}$$

with $\alpha = 0.001$, $f = 0$, $u_b = 0$, and

$$u_d = 16x(1-x)^2(x-y) + \frac{3}{5}, \quad u_a = -0.08 - 4\left(x - \frac{1}{4}\right)^2 - 4\left(y - \frac{27}{32}\right)^2.$$

Again, no exact solution is avaliable so the approximate optimal value $J^* = 0.2506264253907605$ is used. The numerical tests show that the active set A^+ corresponding to the upper bound is a set with non-zero two-dimensional volume and the active set A^- corresponding to the lower bound contains apparently only one point. The development of the discretization errors and

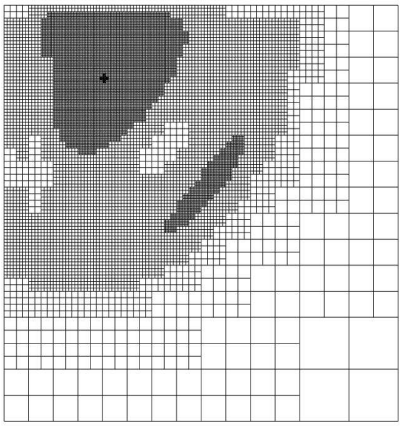


Figure 6.5.. Example of a locally refined mesh for (Ex_3)

the effectivity indices for the numerical solutions calculated by the PDAS method are given in Table 6.4 for both uniform and local mesh refinement. The comparison of both refinement strategies with respect to the required number of degrees of freedom to reach a given error tolerance is done in Figure 6.6. A typical locally refined mesh is shown in Figure 6.5.

Table 6.4.. Development of discretization errors and of the effectivity indices for (Ex_3)

(a) local refinement

N	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	5.38e-04	-1.41
41	-1.16e-04	0.43
99	-4.48e-05	0.33
245	-2.68e-05	0.60
541	-1.04e-05	0.56
1459	-6.04e-06	0.89
4429	-1.54e-06	0.83
13107	-5.01e-07	0.89

(b) uniform refinement

N	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	5.38e-04	-1.41
81	-1.58e-04	0.62
289	-6.18e-05	0.87
1089	-1.58e-05	0.87
4225	-3.99e-06	0.89
16641	-7.45e-07	0.66

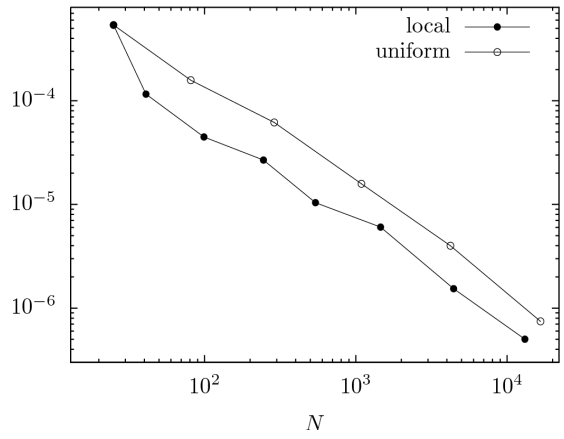


Figure 6.6.. Convergence of the error for (Ex_3)

6.4. Parabolic problem

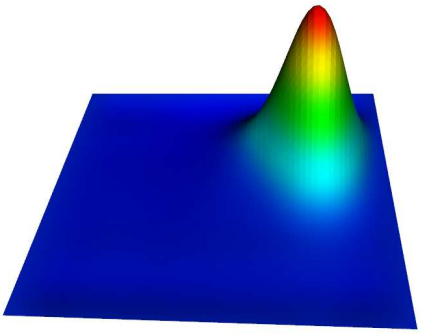
As a time-dependent problem the following example with a linear parabolic state equation is considered:

$$(Ex_4)\quad \begin{cases} \text{Minimize} & J(q,u) = \dfrac{1}{2}\|u-u_d\|^2_{L^2(\Omega\times I)} + \dfrac{\alpha}{2}\|q\|^2_Q, \\ u_t - \Delta u = q & \text{in } (0,T)\times\Omega, \\ u(t,x) = 0 & \forall t\in(0,T), x\in\Gamma, \\ u(0,x) = 0 & \text{on } \Omega, \\ G(u)\geq 0 & \text{in } \bar{I}\times\bar{\Omega}. \end{cases}$$

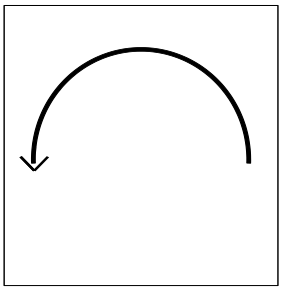
The domain is $\Omega = (0,1)^2$, and the end time is $T = 1$ so that $I = (0,1)$. The integration into the theoretical framework utilizes the spaces $R = L^2(\Omega)$, $Q = L^2(L^2(\Omega)), V = H^1_0(\Omega), X = W(I,V)\cap L^s(I,W^{1,p}(\Omega))\cap W^{1,s}(I,(W^{1,p'}(\Omega))^*)$. The problem data are $\alpha = 0.001$, upper state constraints $G(u) = u_b - u$ with $u_b = 0.1$, and

$$u_d = t\sin^6((2tx+(1-2t)x^4)\pi)\sin^6(((2t-1)^2y - 4t(t-1)y^4)\pi).$$

This function exhibits a growing peak, see Figure 6.7, in other words $\sup_{x\in\bar{\Omega}} u_d(t,x)$ is increasing. With the present choice of the upper state constraint u_b, this leads to the following structure of the active set: from a certain time interval starting in $t = 0$ there are no active points, at time $t = u_b$ the state constraint becomes active in one point, and after that the constraint is active in a set with nonempty two-dimensional interior. For the determination of an approximate solution of (Ex_4) a regularization according to Section 4.2 is considered, with a starting regularization parameter of $\gamma = 100$ and order $o = 1$. By the interior point method, see Section 4.4, the discrete optimal solution is found. The starting discretization always consists of an evenly partition of I into 6 subintervals, and a spatial mesh consisting of 4×4 congruent quadratic cells is used in every time point. The results for this setup with different



(a) Plot of $u_d(t = \frac{1}{6}, \cdot)$



(b) Path of the peak from $t = 0$ to $t = 1$. The peak grows in time.

Figure 6.7.. Structure of u_d for (Ex_4)

refinement strategies are displayed in Table 6.5 and Figure 6.8. As the exact optimal solution is not available, the approximate optimal value $J^* = 1.19205981 \cdot 10^{-3}$ is used in the calculation of the I_{eff}.

The simplest strategy of global refinement of all components at the same time seemingly leads to convergence just fine, but apparently its performance suffers from the fact that the error contributions relating to the estimators $\eta_h, \eta_k, \eta_\gamma$ are of different order of size. The second strategy separates the regularization error from the other contributions. Since this is the error connected to the state constraint, we investigate the effect of discarding any knowledge about η_γ in the refinement strategy: the discrimination which component to refine is made only between η_h and η_k, while γ is increased regardless. While this is a small improvement of the first strategy, the third, full adaptive refinement strategy increases the convergence order considerably.

A more detailed investigation of the importance of the estimator η_γ and its use in the refinement strategy is displayed in Figure 6.9. Here the comparison is taken between two strategies: Firstly, the most involved strategy is considered, this is the evaluation of all estimator components and their use in the error equilibration algorithm, see Section 2.5. Alternatively, the use of η_γ is omitted, but instead γ is always increased by a constant factor c_γ instead. The evaluation is done for the values $c_\gamma \in \{1.5, 10, 31.6\}$. Looking at Figure 6.9(a) it can be seen that $c_\gamma = 1.5$ leads to a too slow convergence of the error relative to the degrees of freedom. Obviously this guess for c_γ is too small, the regularization error is not decreased fast enough. The other choices for c_γ seem to have no disadvantage compared to the involved strategy. The disadvantage for a too large c_γ can be recognized with a comparison of the error relative to the computational time, see Figure 6.9: As the convergence properties of the numerical methods deteriorate with growing γ, the numerical effort needs to be increased to solve the discrete problems, leading to an increase in overall computational time.

Lastly, within the adaptive refinement strategy a comparison is made between the dynamic and the nondynamic spatial discretization approach. The results, this time obtained with a barrier functional of order $o = 2$, can be seen in Table 6.6 and, together with a plot of one mesh, in Figure 6.4. In the first few discretization levels no difference occurs, as in the

Table 6.5.. Results for (Ex_4) with $o = 1$ for the simpler refinement strategies

(a) Global refinement of all components at the same time

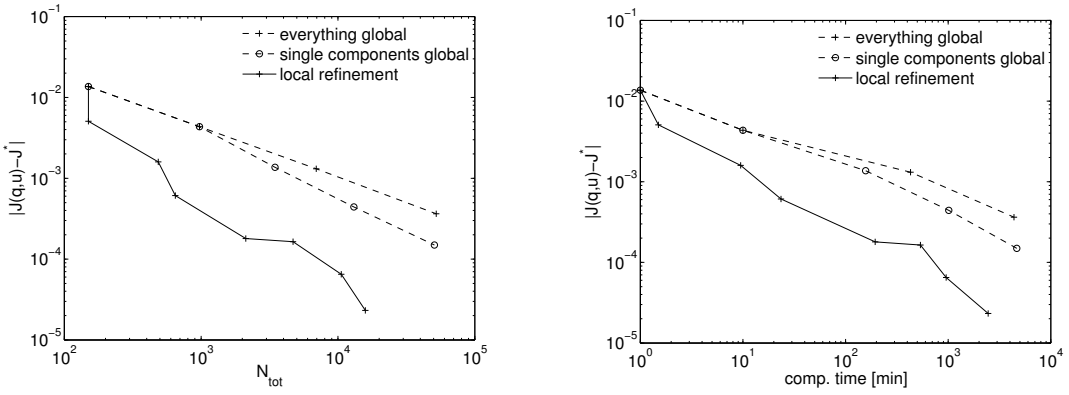
N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	6	1.0e+02	-7.54e-04	-5.41e-04	-1.00e-02	-1.13e-02	-1.367e-02	1.21
81	12	3.2e+02	-9.98e-05	-3.50e-05	-3.16e-03	-3.30e-03	-4.352e-03	1.32
289	24	1.0e+03	-1.87e-04	6.76e-05	-1.01e-03	-1.12e-03	-1.312e-03	1.17
1089	48	3.2e+03					-3.635e-04	

(b) Global refinement, comparison between η_h and η_k

N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	6	1.0e+02	-7.54e-04	-5.41e-04	-1.00e-02	-1.13e-02	-1.367e-02	1.21
81	12	3.2e+02	-9.98e-05	-3.50e-05	-3.16e-03	-3.30e-03	-4.352e-03	1.32
289	12	1.0e+03	-5.04e-05	1.11e-05	-1.01e-03	-1.04e-03	-1.369e-03	1.31
1089	12	3.2e+03	-1.50e-04	-6.57e-07	-3.17e-04	-4.68e-04	-4.416e-04	0.94
4225	12	1.0e+04					-1.494e-04	

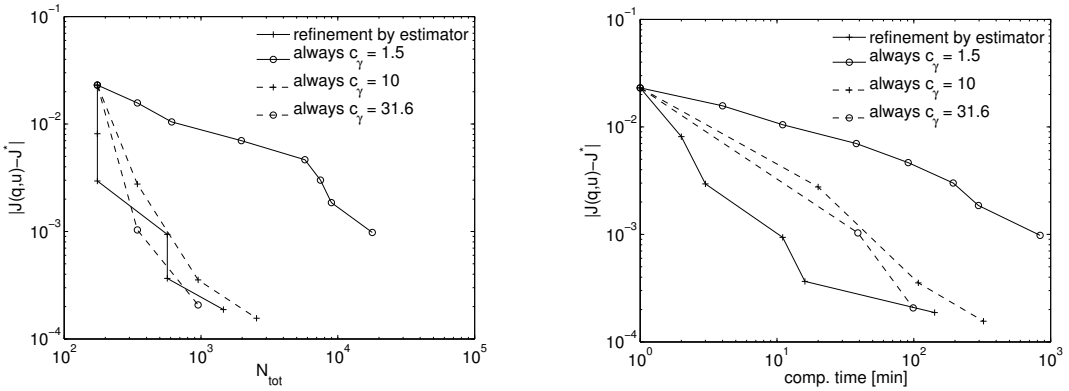
(c) Adaptive refinement, comparison between η_h, η_k and η_γ. Nondynamic discretization.

N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
25	6	1.0e+02	-7.54e-04	-5.41e-04	-1.00e-02	-1.13e-02	-1.367e-02	1.21
25	6	3.2e+02	-1.01e-03	-3.62e-04	-3.16e-03	-4.54e-03	-5.072e-03	1.12
81	6	1.0e+03	-2.58e-04	-5.79e-04	-1.01e-03	-1.84e-03	-1.595e-03	0.87
81	8	3.2e+03	-2.22e-04	-4.01e-05	-3.17e-04	-5.79e-04	-6.119e-04	1.06
265	8	1.0e+04	-1.53e-03	-5.45e-05	-1.02e-04	-1.69e-03	-1.796e-04	1.06
587	8	1.0e+04	-9.50e-05	-3.13e-05	-1.02e-04	-2.27e-04	-1.642e-04	0.72
1321	8	3.2e+04	-6.88e-06	-1.32e-05	-3.17e-05	-5.18e-05	-6.492e-05	1.25
1321	12	1.0e+05	-2.06e-05	-1.33e-04	-1.02e-05	-1.64e-04	-2.322e-05	0.14



(a) depending on the degrees of freedom (b) depending on computational time

Figure 6.8.. Convergence of the error for (Ex_4) for different refinement strategies



(a) depending on the degrees of freedom

(b) depending on computational time

Figure 6.9.. Convergence of the error for (Ex_4) for different values of c_γ

Table 6.6.. Results for (Ex_4) with $o = 2$ for the adaptive spatial refinement strategy

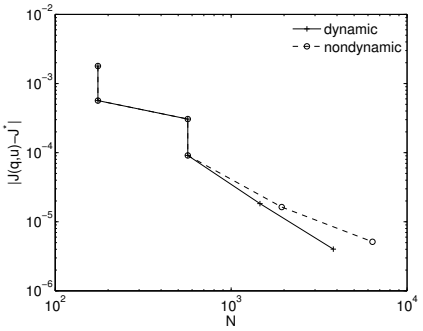
(a) nondynamic version

N_{tot}	γ	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
175	1.0e+01	-6.820e-02	1.40
175	3.2e+01	-1.055e-02	1.55
175	1.0e+02	-1.798e-03	1.30
175	3.2e+02	-5.705e-04	0.55
567	3.2e+02	-3.097e-04	1.90
567	1.0e+03	-9.484e-05	1.37
1939	3.2e+03	-1.994e-05	1.31
6363	1.0e+04	-8.729e-06	1.23

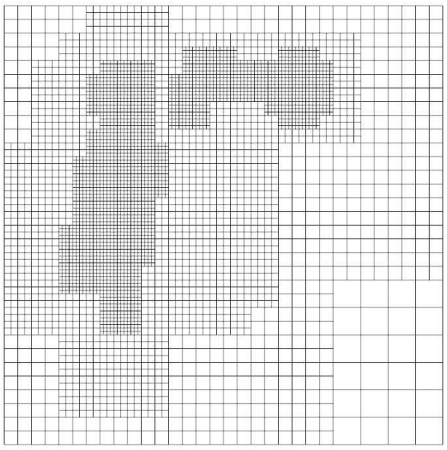
(b) dynamic version

N_{tot}	γ	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
175	1.0e+01	-6.820e-02	0.54
175	3.2e+01	-1.055e-02	0.54
175	1.0e+02	-1.798e-03	0.59
175	3.2e+02	-5.705e-04	0.63
567	3.2e+02	-3.097e-04	0.64
567	1.0e+03	-9.484e-05	0.68
1459	3.2e+03	-2.192e-05	0.72
3809	1.0e+04	-7.623e-06	0.60

dynamic version the spatial error indicators are close enough to each other that the same spatial refinement is chosen on every temporal interval. When the dynamic discretization kicks in in level 6, the dynamic version produces slightly smaller functional errors.



(a) Convergence of the error



(b) Mesh for (Ex_4), created by nondynamic refinement strategy

Figure 6.10.. Comparison of the dynamic and the nondynamic approach of the spatial discretization for (Ex_4)

7. Optimal Control of Young Concrete Thermo-Mechanical Properties

In this chapter the results of the previous work will be applied to a large-sized real-world problem. Large-sized means that a large number of degrees of freedom is needed for the coarsest sensible discretization, which is due to the presence of physical phenoment of different order of magnitude. The spatial domain Ω is a three-dimensional, non-convex set with characteristic structures of different length scales. Similarly, in the considered time interval $[0, T]$ chemical processes of a very small duration take place that need to be resolved. The state variable consists of two components so that the control-to-state operator represents the solution of a system of differential equations. The control variable is comprised of parameters as well as a component that is distributed in time.

7.1. Problem introduction

In the field of civil engineering, especially construction, the decision process on how to execute the building of any structure needs to include different aspects, e.g. stability, practicability, security, legal issues and so on. The means to fulfill these criteria are obviously all interconnected and influencing each other. Finally, they all have different costs, so that the task is to minimize the overall cost in compliance with the above criteria.

A classical problem within this set is the control of the properties of a young concrete structure. The word "young" refers to the time span beginning just after the pour of the concrete, where it is a liquid paste, until the solidification is complete and the concrete has reached maximum strength. The entity of chemical processes is often referred to as *hydration*. This process, including the following mechanical hardening, takes usually a few days. The driving force of these is the heat development, as an exothermal chemical reaction takes place during the solidification. The thermal expansion and following contraction, possibly under external restrains, leads to internal tensile stresses, that may decrease the possible workload the structure can sustain when in use later, or even cause the concrete to crack. Measures that are usually taken to decrease the stresses include

- varying the concrete recipe, that is the mixing ratio of the ingredients,
- changing the choice of ingredients in the concrete recipe, e.g. changing the type of cement, or using additives,
- manipulating the temperature of the raw material before the pour, that is the initial condition,

- manipulating the heat exchange at the boundary, that is shifting the stripping point and after stripping the heating or cooling of the blank surface.

In the context of optimal control, these measures correspond to the control variable q. The task is to control the stresses, or certain derived quantities according to user specification, with minimal cost. Due to the financial magnitude of the problems addressed, it is no doubt worthwhile to utilize computational methods to decrease the costs.

The chemical and physical processes taking place in the young concrete phase are well investigated. For the use of scientific computation, models of different complexity and accuracy have been developed. For an overview over the general field, see [28, 43, 82, 90] and the references therein. Publications that deal with partial aspects are [30, 34, 84, 98] for the investigation of the heat produced during the hydration, [61] for the study of creep phenomena, and [1] for the investigation of the influence of the moisture content. Aspects of the stochastic distribution of material properties are dealt with in [55]. In the following sections, a scenario of a concrete hydration problem will be specified, and a suitable model chosen.

In practical use, these models are mainly used for simulation computations. That means, the user chooses a constellation, that is the values for the control q are assigned by user experience. Then one forward simulation is carried out. If the resulting state does not violate the constraints, it is usually accepted. If it violates the constraints, or the user has the feeling that the solution is not „good enough“ q is changed, again by user experience. An example for the course of action can be found in [77], further descriptions in [28, 81]. To the author's knowledge the problem has never been investigated from the viewpoint of mathematical optimization. The imperative to minimize costs (under the above security and stability constraints) on the other hand does point strongly to using optimization, at least if the computational effort can be limited reasonably. This is not clear due to the large size of the simulation problem alone, and demands an efficient discretization. In section Section 7.5 several classes of optimal control problems will be formulated for the young concrete hydration problem. These are state constrained parabolic optimal control problems, so that the techniques from the previous chapters will be applied to some instances of the problem in the last section.

7.2. Modelling the involved quantities

Current models require at least the two quantities temperature and maturity to characterize the state of the concrete. They will be denoted by $y(t,x)$ and $\tau(t,x)$ in the spatial point $x \in \bar{\Omega}$ at time $t \in [0,T]$, respectively. Staying in the framework of the previous chapters, we set

$$u(t,x) = (y(t,x), \tau(t,x)) \tag{7.1}$$

as the state of the optimal control problem. More state variables that can be used in broader models like moisture content, stresses etc. will not be considered here.

Basis of the forward operator $S(q) = u$ will be the heat equation

$$\begin{aligned} c\rho y_t(t,x) - \lambda \Delta y(t,x) &= \dot{Q}(t,x) && \text{in } (0;T] \times \Omega \\ y(0,x) &= y_0 && \text{in } \Omega \\ \frac{\partial}{\partial n} y(t,x) &= \sigma\left(\bar{y}(t,x) - y(t,x)\right) && \text{on } (0;T] \times \Gamma, \end{aligned} \tag{7.2}$$

where the internal heat source $\dot{Q}(t,x)$ is composed of the heat internally produced by the chemical reaction, and the possible decrease by some water cooling device. While $\bar{y}(t,x)$ is considered a given external temperature profile, the heat capacity c, density ρ, and heat conductivity λ are material parameters and thus potentially subject to the control measures, as well as the initial temperature y_0 and the heat transfer coefficient σ. In the following the considered user influences will be precisely modelled, specifying the influence of q on these quantities.

The direct influence of the user on the concrete composition is as follows: We assume the concrete recipe specified by the partial densities ρ_i of its ingredients. The set of ingredients is fixed for our purposes, see table Table 7.1(a), but could of course be extended. Also the type of ingredients like cement species, additives and so on, is fixed and has to be made by user experience. The partial densities are now to be assigned to the control variable q, but we notice that not all four partial densities can be manipulated independently from each other, as they must fulfill a volume condition: denoting the partial volumes by V_i and the overall volume by V, the relation

$$\sum_i V_i = V \quad \Leftrightarrow \quad \sum_i \frac{\rho_i}{\rho_{g,i}} = 1, \tag{7.3}$$

has to be fulfilled, with $\rho_{g,i}$ denoting the bulk densities of the ingredients, see Table B.2(a) for example data. Thus one degree of freedom is lost, and one of the partial densities can not be considered a component of the control variable. The partial density of aggregate ρ_4 is chosen to be that one, the others are assigned to the control component with the same index,

$$q_i := \rho_i, \quad i = 1 \ldots 3,$$

so that the remaining partial density can be expressed by

$$\rho_4 = \rho_{g,4} \left(1 - \frac{q_1}{\rho_{g,1}} - \frac{q_2}{\rho_{g,2}} - \frac{q_3}{\rho_{g,3}}\right). \tag{7.4}$$

The composition of the mixture influences the heat equation (7.2) via the material parameters.

Table 7.1.. Partial densities and other components of the control variable

(a) partial densities

$q_1 = \rho_1$	partial density of (blast-furnace) cement
$q_2 = \rho_2$	partial density of fly ash
$q_3 = \rho_3$	partial density of water
ρ_4	partial density of aggregate

(b) other

$q_4 = y_0$	initial temperature
$q_5 = t_0$	stripping point
$q_6 = w(t)$	water cooling rate

For some of them, the connection is directly known. The density $\rho(q)$ of the mixture is given by

$$\rho(q) = \sum_{i=1}^{3} \rho_{N,i} q_i \qquad \text{with } \rho_{N,i} := \left(1 - \frac{\rho_{g,4}}{\rho_{g,i}}\right), \tag{7.5}$$

the thermal conductivity $\lambda(q)$ and the heat capacity $c(q)$ of the mixture are simply the means of the respective values of the ingredients, weighted with their partial densities:

$$\lambda(q) = \lambda_4 \frac{\rho_{g,4}}{\rho(q)} + \sum_{i=1}^{3} \left(\lambda_i - \lambda_4 \frac{\rho_{g,4}}{\rho_{g,i}}\right) \frac{q_i}{\rho(q)}, \tag{7.6}$$

$$c(q) = c_4 \frac{\rho_{g,4}}{\rho(q)} + \sum_{i=1}^{3} \left(c_i - c_4 \frac{\rho_{g,4}}{\rho_{g,i}}\right) \frac{q_i}{\rho(q)}. \tag{7.7}$$

Here λ_i and c_i denote the heat conductivity and capacity of the single ingredients. Example data, which will also be used in the numerical tests later, can be found in Table B.2(a).

The second way of user influence on the technical process is the manipulation of the initial temperature, we can directly set

$$y_0 = q_4 \tag{7.8}$$

to be a constant. While it would be mathematically possible to treat non-constant initial temperatures $y_0(x)$, when considering young concrete the mixing of the ingredients before the pour leads to an even temperature of the material throughout the domain.

The third user influence measure to consider is concerned with the heat exchange of the concrete stucture with the environment. While the heat exchange coefficient can be a function $\sigma(t)$ in general, in engineering practise it is likely to be a piecewise constant function. In the utilized model it is assumed that σ is constant with a given value for some time after the pour, that is the time span when the formwork is applied to the construction. After stripping the formwork at some time t_0, σ takes a different value. The stripping point itself, however, is user controllable, and thus included in the control variable as

$$t_0 = q_5. \tag{7.9}$$

So the heat exchange coefficient is given as

$$\sigma(t) = \begin{cases} \sigma_0 : & t \leq q_5 \\ \sigma_1 : & t > q_5 \end{cases}. \tag{7.10}$$

From equation (7.2) there remains one term to be considered, the internal heat source $\dot{Q}(t,x)$. Two phenomena take part in this, the chemically produced heat $\dot{Q}_c(t,x)$ due to the chemical reactions within the concrete, and the heat deducted by a possible water cooling device $\dot{Q}_p(t,x)$, so that

$$\dot{Q}(t,x) = \dot{Q}_c(t,x) + \dot{Q}_p(t,x). \tag{7.11}$$

Many publications are devoted to the study of the chemical heat $\dot{Q}_c(x,t)$; for the model employed in this thesis the considerations from [55] were used as a starting point. In the literature dealing with these models it is common to introduce the degree of hydration

$$\alpha(t,x) := \frac{Q_c(t,x)}{Q_\infty} \tag{7.12}$$

as a new variable. There, Q_∞ is a material constant, the choice of which will be discussed along with other material constants later in this section.
One more variable has to be introduced first. The reason is that the chemical heat source model needs to reflect two different effects. For one, the chemical reaction rate depends on the temperature y, increasing with y. But it also depends on the leftover raw material, which can be indicated by the heat that has been produced until that time point. This second effect is commonly incorporated into the model by the use of a quantity called maturity or effective age, here denoted by τ. The interpretation of the maturity is to trace the heat development back to one calibration configuration. This means $\alpha(\tau)$ describes the progress of the reaction in a test scenario that can be achieved under controlled conditions. For these tests usually adiabatic boundary conditions are chosen.

For both the course of $\alpha(\tau)$ and $\tau(t,\cdot)$, a number of models have been discussed in the civil engineering literature, for an overview see e.g. [29]. A common form for the maturity is $\tau\left(t,x,y(\cdot,\cdot)\right) = \int_0^t g\left(y(\theta,x)\right)\,\mathrm{d}\theta$ with an appropriate function $g(\cdot)$. One maturity, which was introduced by Saul [84], is

$$\tau\left(t,x,y(\cdot,\cdot)\right) = \int_0^t \frac{y(\theta,x)+10}{30}\,\mathrm{d}\theta. \tag{7.13}$$

This model does not incorporate material parameters, thus it is independent of the concrete recipe. A more involved approach which can be motivated by chemical reaction kinetics is the maturity of Freiesleben Hansen et al. from [34],

$$\tau\left(t,x,y(\cdot,\cdot)\right) = \int_0^t \exp\left(\frac{A}{R}\left(\frac{1}{293}-\frac{1}{273+y(\theta,x)}\right)\right)\,\mathrm{d}\theta. \tag{7.14}$$

In this formula R is the universal gas constant and A the activation energy. The activation energy of the hydration reactions can generally depend on the temperature. But according to [55, (5.22)], an activation energy constant in the temperature is applicable to a large class of cements (containing "German cements"), so for simplicity we are assuming a constant activation energy, given by

$$\frac{A}{R} = 5050K\cdot c_{SL} - 2950K, \tag{7.15}$$

and c_{SL} depends on the type of cement only, see [55, Chapter 5] and Table B.2(b). The adiabatic reaction progression can be modeled by Wesche's proposal in [98] as

$$\alpha = \alpha\left(\tau\right) = e^{a_W \tau^{b_W}}, \tag{7.16}$$

where $a_W, b_W < 0$ are material parameters. Another very common model was introduced by Jonasson [30]:

$$\alpha = \alpha\left(\tau\right) = e^{a_J\left[\log\left(1+\frac{\tau}{\tau_k}\right)\right]^{b_J}} \tag{7.17}$$

where $a_J, b_J < 0$ and $\tau_k > 0$ are material parameters. When using this model, experiments find only a small range of values the parameter a_J takes, so some sources set a_J to the approximate

value $a_J = -1$ to begin with (see [43, Section 2.3.2.5] or [79, formula 4.1 and 4.2 in Section 4.1.2]).

In the model of the chemical heat source it remains to specify the values of the material parameters. For optimization problems with constant concrete composition, these parameters $Q_\infty, a_W, b_W, b_J, \tau_k$ can be assumed to be known constants. The values chosen for numerical tests can be found in Table B.1. If however the concrete composition is subject to the control variable, then they have to be regarded as $Q_\infty(q), a_W(q), b_W(q), b_J(q), \tau_k(q)$. Unfortunately, no analytic relation of these parameters to the concrete composition is known. So until more in-depth research is carried out, the test measurements from [55, Appendices C,D] are used. These give the values of the material parameters for a number of standard concrete recipes. These data points are used as reference points for a parameter fitting approach. For the numerical tests in this thesis, the linear models

$$Q_\infty(q) = m_{Q_\infty,0} + \sum_{i=1}^{4} m_{Q_\infty,i}\rho_i, \tag{7.18}$$

$$a_W(q) = m_{aW,0} + \sum_{i=1}^{4} m_{aW,i}\rho_i, \tag{7.19}$$

$$b_W(q) = m_{bW,0} + \sum_{i=1}^{4} m_{bW,i}\rho_i, \tag{7.20}$$

$$b_J(q) = m_{bJ,0} + \sum_{i=1}^{4} m_{bJ,i}\rho_i, \tag{7.21}$$

$$\tau_k(q) = m_{\tau_k,0} + \sum_{i=1}^{4} m_{\tau_k,i}\rho_i, \tag{7.22}$$

were used. The parameters $m_{aW,i}, m_{bW,i}, m_{bJ,i}, m_{\tau_k,i}, m_{Q_\infty,i}$ are hereby found by linear fitting of the data from [55], see Appendix B and especially Table B.4 for example data.

With this, the chemical heat source is modelled. It remains to consider $\dot{Q}_p(t,x)$ as the deducted heat of a possible water cooling system, which is the last considered method of user influence. It is thought of as one pipe of radius $\bar{r}$ going straight through the concrete structure. Water of temperature y_c colder than the concrete is pumped through the pipe at a rate of $w(t)$ extracting heat energy from the concrete. The modelling is taken from [53]. For simplicity, we do not model the pipe as boundary with according Robin boundary conditions, but use the distributed term $\dot{Q}_p$ in the right hand side of (7.2). The amount of extracted energy can be controlled by the flow rate of the water $w(t)$, which can be adjusted over time. Thus $\dot{Q}_p$ depends on the time point t, but since the cooling water heats up as it runs through the pipe, $\dot{Q}_p$ also depends on the spatial position x. For an easier description we consider $\bar{x}$ as lengthwise coordinate of the pipe, such that $\bar{x}$ takes values between $\bar{x} = 0$ at the inflow and $\bar{x} = \bar{l}$ at the outflow, see Figure 7.1 for an illustration. A mapping $\bar{x} \to x$, which gives the lengthwise position of any point x inside the pipe, is easily obtained. The deducted heat is modelled as follows:

$$\dot{Q}_p(t,x,w(t)) = \frac{-2\sigma_W(w(t))}{\bar{r}} \left(y - y_c - \frac{2\pi\bar{r}\bar{x}(x)\sigma_W(w(t))(y-y_c)}{w(t)\rho_W c_W} \right), \tag{7.23}$$

$$\text{with} \qquad \sigma_W(w(t)) = \left(350 + 210\frac{\sqrt{w(t)/\pi/\frac{m}{s}}}{\bar{r}}\right)\frac{W}{Km^2}, \tag{7.24}$$

following the considerations in [53, section 8.3].

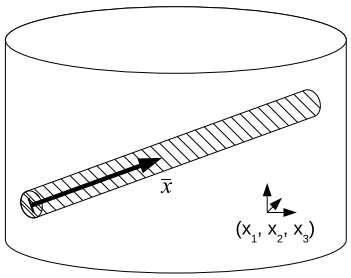


Figure 7.1.. Cooling pipe inside a concrete body. While any point within the body can be described by its spatial coordinates $x = (x_1, x_2, x_3)$, every point inside the pipe, which is simplified as a one-dimensional object, can also be described by its distance from the inflow $\bar{x}$.

To integrate all the models into one state equation, not only the temperature y but also the maturity τ is considered as a component of the state variable. This is benefitial due to the dependence (7.13) or (7.14). Also the term $\dot{Q}_c$ is replaced via (7.12) and the chain rule by

$$\dot{Q}_c = Q_\infty \frac{\partial \alpha}{\partial \tau}\frac{\partial \tau}{\partial t}$$

and the two partial derivatives in this expression are denoted by h and g. So, with explicit marking of the dependencies of all functions on q, y, τ, but suppressed time and space-coordinates, the state equation reads: For a given $q \in \mathbb{R}^5 \times L^\infty(\bar{I})$, find a $u = (y, \tau) \in W(I, H^1(\Omega))^2$ such that

$$\begin{aligned}
\tau_t &= g(y) && \text{in } (0;T]\times\Omega \\
c(q)\rho(q)y_t - \lambda(q)\Delta y &= Q_\infty(q)g(y)\,h(\tau,q) - \dot{Q}_p(q) && \text{in } (0;T]\times\Omega \\
\tau(0,x) &= 0 && \text{in } \Omega \\
y(0,x) &= y_0(q) && \text{in } \Omega \\
\frac{\partial}{\partial n}y &= \sigma(q)\,(\bar{y} - y) && \text{on } (0;T]\times\partial\Omega,
\end{aligned} \tag{7.25}$$

where the functions g, h are chosen according to the models discussed above as

$$g(y) = \frac{y+10}{30} \qquad \text{(Saul's maturity) or} \tag{7.26}$$

$$g(y) = \exp\left(\frac{A}{R}\left(\frac{1}{293} - \frac{1}{273+y}\right)\right) \qquad \text{(maturity of Fr. Hansen et al.),} \tag{7.27}$$

and

$$h(\tau, q) = a_W(q) b_W(q)\, e^{a_W(q)\tau^{b_W(q)}} \tau^{b_W(q)-1} \qquad \text{(model of Wesche) or} \tag{7.28}$$

$$h(\tau, q) = -\frac{b_J(q)}{\tau + \tau_k(q)} e^{-\left[\log\left(1+\frac{\tau}{\tau_k(q)}\right)\right]^{b_J(q)}} \left[\log\left(1+\frac{\tau}{\tau_k(q)}\right)\right]^{b_J(q)-1} \qquad \text{(model of Jonasson).} \tag{7.29}$$

7.3. State equation

In this section we will study the properties of the state equation. For an arbitrary, but constant control $q \in Q$ it can be written as

$$\begin{aligned} \tau_t &= g(y) && \text{in } (0;T] \times \Omega \\ c\rho y_t - \lambda \Delta y &= Q_\infty g(y)\, h(\tau) - \dot{Q}_p && \text{in } (0;T] \times \Omega \\ \tau(0,x) &= 0 && \text{in } \Omega \\ y(0,x) &= y_0 && \text{in } \Omega \\ \frac{\partial}{\partial n} y &= \sigma\,(\bar{y} - y) && \text{on } (0;T] \times \partial\Omega. \end{aligned} \tag{7.30}$$

with given constants $c > 0, \rho > 0, Q_\infty > 0, y_0$, and $\sigma \in L^\infty([0,T]), \bar{y}, \dot{Q}_p \in L^\infty([0,T] \times \Gamma)$. This is a parabolic partial differential equation for y coupled with an ordinary differential equation for τ in every point $x \in \bar{\Omega}$, which is not in the functional analytic setting of the previous chapters. To apply the central concepts of this thesis, first the existence and uniqueness of a solution of (7.30) will be shown. For the proof some properties of the functions g and h are necessary:

Assumption 7.1. *The model functions are continously differentiable, $g, h \in C^1(\mathbb{R}_+)$, and possess the following properties: For g there holds either*

- *g is affine linear, so $g(y) = C_1 y + C_2 \leq C(1+y)$* **(case 1)**, *or*
- *g is a bounded function with bounded derivative, so $|g(y)| + |g'(y)| \leq C_3$* **(case 2)**.

For h there holds:

- *h and its derivative are bounded, so $|h(\tau)| + |h'(\tau)| \leq C$.*

Remark 7.1. It can be checked that the models (7.26)-(7.29) used in this thesis fulfill Assumption 7.1 as long as the material parameters have their natural sign $a, b \leq 0, \tau_k, A, R > 0$.

Further, define the set

$$K := \{v \in L^2(I \times \Omega) : \|v\|_{W(0,T)} \leq B\},$$

where B is a constant independent of y, τ and will be determined in the course of the following lemmas. K is then a compact, convex and nonempty subset of $L^2(I \times \Omega)$ due to the compact embedding of $W(0,T)$ into $L^2(I \times \Omega)$.

Theorem 7.2. *The system of equations* (7.30) *with the parameters given as above, and the functions q, h fulfilling Assumption 7.1, exhibits a solution $u = (y, \tau) \in K \times H^1(L^2(\Omega))$. Also it holds that $\|y\|_{L^\infty(\Omega \times I)} \leq C_0$ with a constant $C_0 > 0$ independent from y and τ.*

This theorem will be proven by applying Schauder's fixed point theorem to a fixed point operator D which will be built up step-wise over the course of the following lemmas.

Lemma 7.3. *The Nemyzki-operator $D_1 \colon y \mapsto g(y)$ maps the set K into $L^2(I \times \Omega)$, and is continous from $L^2(I \times \Omega)$ to $L^2(I \times \Omega)$.*

Proof. To prove that $g(y) \in L^2(I \times \Omega)$ for all $y \in K$: distinguish two cases by the properties of g according to Assumption 7.1:

- **Case 1**: If g is affine linear, then

$$\|g(y)\|^2_{L^2(I\times\Omega)} = \iint\limits_{I\times\Omega} (g(y))^2 dxdt \leq \iint\limits_{I\times\Omega} C^2(1+y)^2 dxdt \leq \tilde{C},$$

  with $\tilde{C}$ depending on B, C, I, Ω.

- **Case 2**: Since g is bounded,

$$\|g(y)\|^2_{L^2(I\times\Omega)} \leq C_3^2 \iint\limits_{I\times\Omega} dxdt \leq \tilde{C},$$

  with $\tilde{C}$ depending on C, I, Ω.

Continuity of D_1 follows from continous differentiability of g:
$\|g(y) - g(y_n)\|_{L_2(I\times\Omega)} \leq C\|y - y_n\|_{L_2(I\times\Omega)}$ (Lipschitz-condition) so that a converging sequence $y_n \overset{L^2(I\times\Omega)}{\to} y$ has converging values $g(y_n) \overset{L^2(I\times\Omega)}{\to} g(y)$. □

Lemma 7.4. *The operator $D_2 \colon g \mapsto \tau$, where τ ist the unique function solving the ordinary differential equation*

$$\begin{aligned} \tau_t &= g \\ \tau(0) &= 0, \end{aligned}$$

maps the set $L^2(I \times \Omega)$ into $H^1(L^2(\Omega)) \subset L^2(I \times \Omega)$, and is continous from $L^2(I \times \Omega)$ to $L^2(I \times \Omega)$.

Proof. by basic properties of the integral. □

Lemma 7.5. *The Nemyzki-operator $D_3 \colon \tau \mapsto h(\tau)$ maps the set $L^2(I \times \Omega)$ into the set $\{h \in L^\infty(I \times \Omega) : \|h\|_{L^\infty(I\times\Omega)} \leq C_4\} \subset L^2(I \times \Omega)$, and is continous from $L^2(I \times \Omega)$ to $L^2(I \times \Omega)$.*

Proof. like in Lemma 7.3. □

Lemma 7.6. *Let the operators $F_1, F_2 \colon L^2(I \times \Omega) \to L^2(I \times \Omega)$ be continous, and let their images $F_1(L^2(I \times \Omega)), F_2(L^2(I \times \Omega)) \subset \{u : \|u\|_{L^\infty(I\times\Omega)} \leq C\}$. Then the product operator $F \colon (g, h) \to F_1(g) \cdot F_2(h)$ is continous from $L^2(I \times \Omega) \times L^2(I \times \Omega)$ to $L^2(I \times \Omega)$.*

Proof. F maps to $L^2(I \times \Omega)$ indeed, as

$$\|F(g,h)\|_{L^2(I\times\Omega)} = \|F_1(g)F_2(h)\|_{L^2(I\times\Omega)} \leq \|F_1(g)\|_{L^\infty(I\times\Omega)}\|F_2(h)\|_{L^2(I\times\Omega)}.$$

For a sequence $(g_n, h_n) \to (g, h)$, that means $g_n \overset{L^2(I\times\Omega)}{\to} g$, and $h_n \overset{L^2(I\times\Omega)}{\to} h$, the continuity of F_1, F_2 provides $F_1(g_n) \overset{L^2(I\times\Omega)}{\to} F_1(g)$ and $F_2(h_n) \overset{L^2(I\times\Omega)}{\to} F_2(h)$. Then

$$\begin{aligned}
&\|F(g_n,h_n) - F(g,h)\|_{L^2(I\times\Omega)} \\
&\qquad \leq \|F(g_n,h_n) - F(g_n,h)\|_{L^2(I\times\Omega)} + \|F(g_n,h) - F(g,h)\|_{L^2(I\times\Omega)} \\
&\qquad = \|F_1(g_n)(F_2(h_n) - F_2(h))\|_{L^2(I\times\Omega)} + \|F_2(h)(F_1(g_n) - F_1(g))\|_{L^2(I\times\Omega)} \\
&\qquad \leq \|F_1(g_n)\|_{L^\infty(I\times\Omega)}\|F_2(h_n) - F_2(h)\|_{L^2(I\times\Omega)} \\
&\qquad + \|F_2(h)\|_{L^\infty(I\times\Omega)}\|F_1(g_n) - F_1(g)\|_{L^2(I\times\Omega)} \\
&\qquad \to 0 \text{ as } n \to \infty
\end{aligned}$$

proves the continuity of F. □

Lemma 7.7. *For $l \in L^\infty(I \times \Omega)$ with $\|l\|_{L^\infty(I\times\Omega)} \leq C$ there exists a unique solution $y \in W(0,T), y \in L^\infty(I \times \Omega)$ to either of the problems*

$$\left\{\begin{array}{r} c\rho y_t - \lambda\Delta y = Q_\infty l \text{ in } \Omega \\ y(0) = y_0 \\ \lambda\frac{\partial}{\partial n}y = \sigma(\bar{y} - y) \text{ on } \Gamma \end{array}\right. \quad \text{and} \quad \left\{\begin{array}{r} c\rho y_t - \lambda\Delta y = Q_\infty(C_1 y + C_2)l \text{ in } \Omega \\ y(0) = y_0 \\ \lambda\frac{\partial}{\partial n}y = \sigma(\bar{y} - y) \text{ on } \Gamma \end{array}\right. ,$$

and it holds $\|y\|_{W(0,T)} + \|y\|_{L^\infty(I\times\Omega)} \leq C\|l\|_{L^\infty(I\times\Omega)} \leq C^2 =: B$ with B only depending on $c, \rho, \lambda, Q_\infty, y_0, \sigma, \bar{y}, C$.

Proof. E.g. like in [76, Proposition 2.1]. □

Lemma 7.8. *The operator $D_4 \colon l \mapsto y$, where y solves*

$$\begin{aligned} c\rho y_t - \lambda\Delta y &= Q_\infty l \text{ in } \Omega \\ y(0) &= y_0, \\ \lambda\frac{\partial}{\partial n}y &= \sigma(\bar{y} - y) \text{ on } \Gamma \end{aligned} \tag{7.31}$$

maps $L^\infty(I \times \Omega)$ into $K \subset L^2(I \times \Omega)$, and is continous from $L^2(I \times \Omega)$ to $L^2(I \times \Omega)$.

Proof. Well-definedness of the operator was proven in the last lemma. For continuity, consider a sequence $L^\infty(I \times \Omega) \supset (l_n) \overset{L^2}{\to} l \subset L^\infty(I \times \Omega)$. Let $y_n, y \in W(0,T) \subset L^2(I \times \Omega)$ be the according solutions $y_n = D_4(l_n), y = D_4(l)$. We have to show $y_n \overset{L^2}{\to} y$. Subtracting the

equations (7.31) for y and y_n, we find $y - y_n$ solves problem (7.31) with $Q_\infty(l - l_n)$ as right hand side, which by Lemma 7.7 gives us

$$\|y - y_n\|_{W(0,T)} \leq C\|l - l_n\|_{L^2(I\times\Omega)}$$

thus proving the lemma. □

Lemma 7.9. *The operator $\tilde{D}_4 \colon l \mapsto y$, where y solves*

$$\begin{aligned} c\rho y_t - \lambda\Delta y &= Q_\infty(C_1 y + C_2)l \text{ in } \Omega \\ y(0) &= y_0, \\ \lambda\frac{\partial}{\partial n}y &= \sigma(\bar{y} - y) \text{ on } \Gamma \end{aligned} \tag{7.32}$$

maps $L^\infty(I\times\Omega)$ into $K \subset L^2(I\times\Omega)$, and is continous from $L^2(I\times\Omega)$ to $L^2(I\times\Omega)$.

Proof. Again, well-definedness of the operator was proven in Lemma 7.7. With the sequence $\{u : \|u\|_{L^\infty(I\times\Omega)} \leq C\} \supset (l_n) \overset{L^2}{\to} l \subset L^\infty(I\times\Omega)$ as before, we set $y_n = \tilde{D}_4(l_n), y = \tilde{D}_4(l)$, having to show $y_n \overset{L^2}{\to} y$. Subtracting the equations (7.32) for y and y_n, and substituting $z_n = e^{-Lt}y_n, z = e^{-Lt}y$, with L to be determined later, yields

$$c\rho(z_t - z_{n,t}) + c\rho L(z - z_n) - \lambda\Delta(z - z_n) = Q_\infty(l(C_1 y + C_2) - l_n(C_1 y_n + C_2))e^{-Lt} \tag{7.33}$$

$$\begin{aligned} z_n(0) = z(0) &= e^{-Lt}y_0, \\ \lambda\frac{\partial}{\partial n}z_n &= \sigma(e^{-Lt}\bar{y} - z_n) \text{ on } \Gamma \\ \lambda\frac{\partial}{\partial n}z &= \sigma(e^{-Lt}\bar{y} - z) \text{ on } \Gamma. \end{aligned}$$

Testing this equation with $z - z_n$ and integrating over $\Omega\times[0,T]$ leads to

$$\begin{aligned} &\frac{c\rho}{2}(\|z(T) - z_n(T)\|^2_{L^2(\Omega)} - \underbrace{\|z(0) - z_n(0)\|^2_{L^2(\Omega)}}_{=0}) + c\rho L\|z - z_n\|^2_{L^2(I\times\Omega)} \\ &+ \lambda\|\nabla(z - z_n)\|^2_{L^2(I\times\Omega)} + \int_\Gamma \sigma(z - z_n)^2 = Q_\infty \iint_{I\times\Omega} (l(C_1 y + C_2) - l_n(C_1 y_n + C_2))e^{-Lt}(z - z_n) \end{aligned}$$

To estimate the L^2-norm of $z - z_n$ it suffices to consider the third term on the left hand side. This is legit, as the other terms are positive. In the following we estimate the right hand side:

$$\begin{aligned} &c\rho L\|z - z_n\|^2_{L^2(I\times\Omega)} \leq Q_\infty \iint_{I\times\Omega} (C_2(l - l_n) + C_1(ly - l_n y + l_n y - l_n y_n))e^{-Lt}(z - z_n) \\ &= Q_\infty \iint_{I\times\Omega} (C_2(l - l_n)e^{-Lt}(z - z_n) + C_1 y e^{-Lt}(z - z_n)(l - l_n) + C_1 l_n(z - z_n)^2) \\ &\leq Q_\infty(C_2\|(l - l_n)e^{-Lt}\|_{L^2(I\times\Omega)}\|(z - z_n)\|_{L^2(I\times\Omega)} + \\ &C_1\|(l - l_n)\|_{L_2(I\times\Omega)}\|z - z_n\|_{L_2(I\times\Omega)}\|y\|_{L^\infty(I\times\Omega)} + C_1\|l_n\|_{L^\infty(I\times\Omega)}\|z - z_n\|^2_{L_2(I\times\Omega)}) \\ &\leq Q_\infty C\|z - z_n\|^2_{L_2(I\times\Omega)} + Q_\infty C\|(l - l_n)\|^2_{L_2(I\times\Omega)} \end{aligned}$$

Choosing L large enough gives the desired convergence. □

Now the proof of Theorem 7.2 can be given:

Proof. **Case 1**: g is affine linear. Then we define the operator $D\colon K \to K$ by

$$y \mapsto \tilde{D}_4 \circ D_3 \circ D_2 \circ D_1(y).$$

Case 2: g is bounded. Then we define the operator $D\colon K \to K$ by

$$y \mapsto D_4 \circ F(D_1(y), D_3 \circ D_2 \circ D_1(y)).$$

In either case, the previous lemmas provide the well-definedness of D and its continuity from $L^2(I \times \Omega)$ to $L^2(I \times \Omega)$. Since K is nonempty, convex and compact, the application of Schauder's fixed point theorem, see, e.g., [31, Chapter 9.2, Theorem 3], to D yields that a solution $y \in K$ exists. Subsequent application of the previous lemmas gives additionally $\tau \in H^1(L^2(\Omega))$ and $y \in L^\infty(I \times \Omega)$, with $\|y\|_{L^\infty(I\times\Omega)} \leq B$. □

Theorem 7.10. *Under the assumptions made in Theorem 7.2, the solution $u = (y, \tau)$ of the state equation* (7.30) *is unique.*

Proof. Let (y_1, τ_1) and (y_2, τ_2) be two solutions of (7.30). We denote

$$u = y_1 - y_2, \qquad \mu = \tau_1 - \tau_2.$$

By a few calculations it can be shown that there holds:

$$\begin{cases} \mu_t = g'(y^*)u & \text{in } (0,T] \times \Omega \\ \mu(0) = 0 & \text{on } \Omega \\ c\rho u_t - \lambda \Delta u = Q_\infty(g(y_1)h'(\tau^*)\mu + h(\tau_2)g'(y^*)u) & \text{in } (0,T] \times \Omega \\ u(0) = 0 & \text{on } \Omega \\ \lambda \dfrac{\partial}{\partial n} u = -\sigma u & \text{on } (0,T] \times \Gamma \end{cases}, \tag{7.34}$$

with some $y^*(t) \in (y_1(t), y_2(t)), \tau^*(t) \in (\tau_1(t), \tau_2(t))$. With a constant $L > 0$, that will be specified later, consider the functions $w(t) = u(t)e^{-Lt}$ and $\nu(t) = \mu(t)e^{-Lt}$. Then, (7.34) transforms to

$$\begin{cases} \nu_t + L\,\nu = g'(y^*)w & \text{in } (0,T] \times \Omega \\ \nu(0) = 0 & \text{on } \Omega \\ c\rho w_t + c\rho L w - \lambda \Delta w = Q_\infty(g(y_1)h'(\tau^*)\nu + h(\tau_2)g'(y^*)w) & \text{in } (0,T] \times \Omega \\ w(0) = 0 & \text{on } \Omega \\ \lambda \dfrac{\partial}{\partial n} w = -\sigma w & \text{on } (0,T] \times \Gamma \end{cases} \tag{7.35}$$

Testing the first equation with ν and integrating from 0 to T we obtain:

$$\frac{1}{2}\|\nu(T)\|^2 + L\,\|\nu\|^2_{L^2(L^2(\Omega))} = \int_0^T (g'(y^*)w(t), \nu(t))\,dt \leq \frac{C_1}{2L}\|w\|^2_{L^2(L^2(\Omega))} + \frac{L}{2}\|\nu\|^2_{L^2(L^2(\Omega))}.$$

Hence,

$$\|\nu\|_{L^2(L^2(\Omega))} \leq \frac{\sqrt{C_1}}{L} \|w\|_{L^2(L^2(\Omega))}.$$

Testing the third equation in (7.35) with w and integrating from 0 to T we obtain:

$$\begin{aligned}\frac{c\rho}{2}\|w(T)\|^2 + c\rho L \|w\|^2_{L^2(L^2(\Omega))} + \lambda\|\nabla w\|^2_{L^2(L^2(\Omega))} + \sigma\|w\|^2_{L^2(L^2(\Gamma))} \\ = Q_\infty \underbrace{\int_0^T \int_\Omega h(\tau_2) g'(y^*) w(t)^2 \, dx \, dt}_{\leq C_1 C_2 \|w\|^2_{L^2(L^2(\Omega))}} + Q_\infty \underbrace{\int_0^T \int_\Omega g(y_1) h'(\tau^*) \, \nu(t) \, w(t) \, dx \, dt}_{\leq \frac{C_1(1+C_0)C_2^{3/2}}{L} \|w\|^2_{L^2(L^2(\Omega))}}.\end{aligned}$$

Note, that we used $g(y_1) \leq g(0) + C_1 y_1 \leq C_1(1 + C_0)$ due to the boundedness of the derivative of g. Therefore, there holds:

$$\begin{aligned}\frac{c\rho}{2}\|w(T)\|^2 + \left(c\rho L - C_1 C_2 - \frac{C_1(1+C_0)C_2^{3/2}}{L} \right) \|w\|^2_{L^2(L^2(\Omega))} + \lambda\|\nabla w\|^2_{L^2(L^2(\Omega))} \\ + \sigma\|w\|^2_{L^2(L^2(\Gamma))} \leq 0.\end{aligned}$$

Choosing L large enough we conclude $w = 0$. Therefore $y_1 = y_2$ and $\tau_1 = \tau_2$. □

7.4. Optimization problems

In civil engineering there is not just one prototypical optimal control problem to be found for young concrete. Instead, the demands on the structure may differ in their nature. The formulation of an optimal control problem of the form

$$\begin{cases} \min J(q, u) \\ u = (y, \tau) = S(q) \quad \text{by (7.30)} \\ G(u) \geq 0 \end{cases} \tag{7.36}$$

has to reflect the precise situation on the applicant's side. In this section, a number of common choices of cost functionals and state constraints will be discussed.

Also note that the introduction of

$$q = (q_1, \ldots, q_6) = (\rho_1, \rho_2, \rho_3, y_0.t_0, w) \in \mathbb{R}^5 \times L^2(I)$$

is to be seen as a „maximum" control, but in practise not always all of these control measures may be possible or desired. One can easily exclude some component(s) of $(q_1, \ldots, q_6)$ from the formulation of (7.30) by inserting a constant value.

Remark 7.2. Although no control constraints were formulated explicitly, these are in a certain sense still present for technical reasons: for every component in q there are upper and lower bounds beyond which the physical quantities become meaningless: partial densities are bound between zero and the corresponding bulk densities, and so on. It may be necessary to reflect this in the implementation.

The formulation of the state constraint or the cost functional may utilize some physical quantities that can be derived from the state variable. These are mechanical properties that are developed during the solidification phase. One is the dergee of hydration itself:

$$\alpha(t,x) = \alpha(\tau(t,x)) \qquad \text{by the model (7.16) or (7.17)}$$

as an indicator for the progress of the hydration. Other model functions use the maturity, see, e.g., [55], to approximate the tensile strength f_{ct}

$$f_{ct}(t,x) = f_{ct,\infty}\left(\frac{\alpha(\tau(t,x)) - \alpha_0}{1-\alpha_0}\right)^{\gamma_1},$$

the compressive strength f_{cc}

$$f_{cc}(t,x) = f_{cc,\infty}\left(\frac{\alpha(\tau(t,x)) - \alpha_0}{1-\alpha_0}\right)^{\gamma_2},$$

and Young's modulus

$$E(t,x) = E_{\infty}\left(\frac{\alpha(\tau(t,x)) - \alpha_0}{1-\alpha_0}\right)^{\gamma_3}.$$

The final values $f_{ct,\infty}, f_{cc,\infty}, E_{\infty}$ and the exponents $\gamma_1, \gamma_2, \gamma_3$ are constants. Typical values for the final values are in the range of $f_{ct,\infty} = 2.5\,\text{MPa}$, $f_{cc,\infty} = 40\,\text{MPa}$, $E_{\infty} = 30\,\text{GPa}$, and for the exponents $\gamma_1 = 1, \gamma_2 = \frac{3}{2}, \gamma_3 = \frac{1}{2}$.

7.4.1. State constraint

The simplest pointwise state constraint is bounding the temperature from above. This demand may be necessary explicitly since above a certain temperature range the chemical reactions in the hydration change, putting the construction at risk. This is formulated as

$$y(t,x) \le y_{max} \qquad \Leftrightarrow \qquad G(t,x,u(t,x)) := y_{max} - y(t,x) \ge 0,$$

with, e.g., $y_{max} = 70°C$.

A meaningful constraint of the strength of the concrete could be demanding a minimum value for the tensile strength (compressive strength is seldom a problem) at every time point or in the endpoint only. This results in the formulation

$$f_{ct}(t,x) \ge f_{ct,min}(t) \qquad \Leftrightarrow \qquad G := f_{ct,\infty}\left(\frac{\alpha(\tau(t,x)) - \alpha_0}{1-\alpha_0}\right)^{\gamma_1} - f_{ct,min}(t) \ge 0 \text{ or}$$

$$f_{ct}(T,x) \ge f_{ct,min} \qquad \Leftrightarrow \qquad G := f_{ct,\infty}\left(\frac{\alpha(\tau(T,x)) - \alpha_0}{1-\alpha_0}\right)^{\gamma_1} - f_{ct,min} \ge 0$$

with $f_{ct,min}(t)$ or $f_{ct,min}$ given.

A frequent constraint is a criterium for freedom of cracks. Since the development of the temperature inside the structure causes the building up of tensions, the structures are often at risk of cracking. Although these tensions could be approximated by solving the equations of linear thermo-elasticity, and so the cracking predicted, the effort to solve the additional

partial differential equations is frequently avoided by using temperature criteria instead. One criterium that is often used states that if the maximum temperature difference within the structure is $15K$ or lower, then no cracks do occur. From physical considerations it is often a priori known where the coldest and warmest point of the structure is going to be. Naming these points x_1, x_2, the constraint is

$$y(t, x_1) - y(t, x_2) \leq 15K \quad \forall t \in [0, T].$$

This criterium can not be written with a constraint function G as in Section 2.1.3. But also in this case an analog approach is possible by defining

$$G := 15 - y(t, x_1) + y(t, x_2)$$

7.4.2. Cost functional

Here are now some suggestions for contributions that may be chosen to be used as summands in the definition of a cost functional.

An obvoius suggestion is to take the term *cost* functional literally, and have

- $F_1(q_1, q_2, q_3)$ describe the actual material costs of cement, fly ash, water and additives,
- $F_2(q_4)$ describe the heating or cooling costs of the raw material,
- $F_3(q_5)$ describe the costs for the application of the formwork,
- $F_4(q_6)$ describe the costs for operating the water cooling device.

These suggestions amount to control costs.

Additionally, or instead, state costs can play a role in the sense of real monetary costs. For example failure to reach an agreed upon goal in terms of the tensile strength may result in having to pay a fine, the amount of which depends upon the time span the realization of the minimum value was delayed. State costs can also gradually reward or penalize present properties of the state on a user-defined scale, for example the consideration of tensile strength as a property that is more advantageous the higher its value is, can lead to a summand

$$F_5(f_{ct}(u)) \qquad \text{with a monotonically decreasing function } F_5.$$

The cost functional can also be used for the weak, or regularized, fulfilling of state constraints, or an approach that weighs the fulfilling of the state constraints against decreasing of the cost functional.

7.5. Examples and numerical results

7.5.1. Control of initial temperature and heat transfer

In this section an optimal control problem of young concrete hydration is considered that is motivated as follows: Assume that a large part of a construction process is fixed for external reasons, and only the initial temperature y_0 of the ingredients and the heat exchange coefficient σ can be chosen freely; the latter, e.g., by adjusting the thickness of the formwork. These are commonly used methods to influence the temperature distribution. The executing company may now have a standard construction procedure leading to some preferred values for y_0 and σ, and changing these values induces costs. If now the fulfillment of an additional state constraint is demanded which the standard procedure would violate, the question is to find values for y_0, σ such that the state constraint is fulfilled at minimal cost.

Thus the parameter control problem is modelled by setting

$$q = (q_1, q_2) = (y_0, \sigma) \in \mathbb{R}^2 =: Q.$$

as the control variable with the components q_1 denoting the initial temperature in $^\circ C$ and q_2 the heat exchange coefficient in $\frac{kJ}{m^2 K h}$. The state equation is then given by

$$\begin{aligned} \tau_t &= g(y) && \text{in } (0;T] \times \Omega \\ c\rho y_t - \lambda \Delta y &= Q_\infty g(y)\, h(\tau) && \text{in } (0;T] \times \Omega \\ \tau(0,x) &= 0 && \text{in } \Omega \\ y(0,x) &= q_1 && \text{in } \Omega \\ \frac{\partial}{\partial n} y &= q_2\,(\bar{y} - y) && \text{on } (0;T] \times \Gamma. \end{aligned} \tag{7.37}$$

For the following numerical tests the models (7.13) and (7.16) are chosen for the chemical heat source, the remaining material parameters $c, \rho, \lambda, Q_\infty$ are set according to the reference concrete recipe in Table B.1, and $\bar{y} = 20^\circ C$ is chosen as exterior temperature. The considered temporal interval has a length of $T = 48\,h$ and the spatial domain Ω is the wall illustrated in Figure 7.2. Since heat is only produced in the part of the domain that is labeled as „new concrete“ $g(y)$ and $h(\tau)$ are set to zero on the „foundation“ part of the domain. The objective is then to solve the optimal control problem with upper temperature constraints

$$(Ex_5) \begin{cases} \min 5(q_1 - 20)^2 + 5(q_2 - 8)^2, & q \in Q, \\ S(q) = (y, \tau) & \text{according to (7.37)}, \\ y(t,x) \le 72^\circ C & . \end{cases} \tag{7.38}$$

Note that the problem data have indeed been chosen in such a way that the optimal control of the unrestricted problem $(q_1, q_2) = (20, 8)$ violates the state constraints.

The computations were carried out with the interior point method with a barrier functional of order $o = 2$, starting regularization parameter $\gamma = 0.3$, a starting temporal discretization with $M = 12$ equidistant time steps, and an equidistant spatial discretization with $N_m = 135$ in every time step.

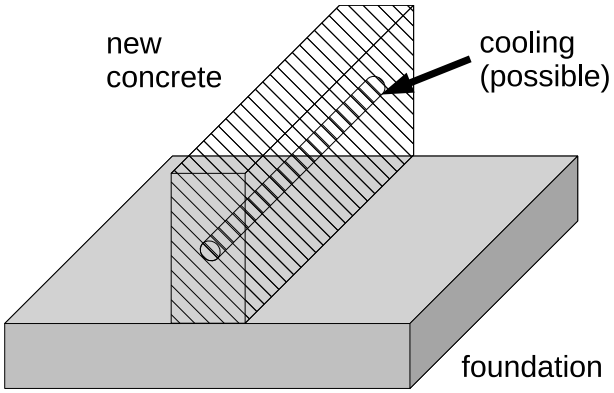


Figure 7.2.. Computational domain for (Ex_5) and (Ex_7) consisting of a foundation of old concrete (solid grey), where no heat is produced and a wall of fresh concrete (shaded). In (Ex_7) the latter includes a cooling pipe.

First consider the temporal discretization only. Leaving the spatial discretization and the regularization parameter γ constant, the global temporal refinement is compared to the adaptive refinement driven by the error estimator η_k from (4.45). The results are displayed in Table 7.2 where the estimated optimal value of the cost functional is $J^* = 1937.8$ and was obtained on a finer temporal discretization than those used for the table. The temporal meshes created by the adaptive process are depicted in Figure 7.3 up to the level where $M = 236$. The efficiency

Table 7.2.. Results for (Ex_5), temporal refinement only

(a) Global refinement

M	η_k	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
12	8.16e+02	9.808e+02	1.20
24	5.93e+02	7.130e+02	1.20
48	3.57e+02	3.711e+02	1.04
96	1.84e+02	1.877e+02	1.02
192	9.32e+01	9.427e+01	1.01
384	4.68e+01	4.719e+01	1.01
768	2.34e+01	2.357e+01	1.01

(b) Adaptive refinement

M	η_k	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
12	8.16e+02	9.808e+02	1.20
18	6.14e+02	7.549e+02	1.23
32	3.65e+02	4.298e+02	1.18
60	2.07e+02	2.377e+02	1.15
120	1.18e+02	1.258e+02	1.07
236	6.30e+01	6.478e+01	1.03
462	3.25e+01	3.283e+01	1.01
920	1.65e+01	1.646e+01	1.00

indices being close to 1 point to a good quality of the error estimation. The adaptive strategy yields a marginally better error convergence than the global one, see also Figure 7.5(a).
Next, only the spatial discretization is subject to investigation. Starting from the initial discretization again, the global spatial refinement strategy and the adaptive strategy using error estimator (4.46) are compared. The evaluation of the error with $J^* = 593.0$ calculated on the finest level, displayed in Table 7.3, shows however also no significant improvement of

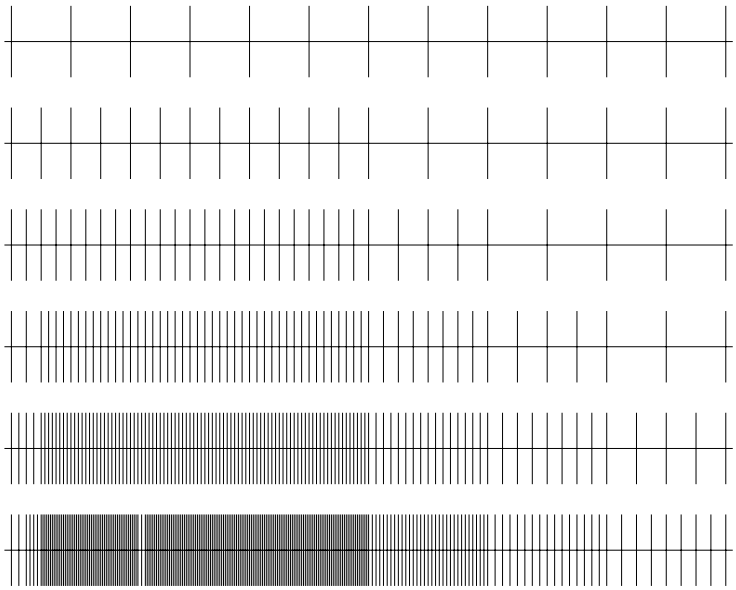


Figure 7.3.. Adaptive temporal refinement for (Ex_5)

the functional error convergence. A locally refined mesh created during the process is shown in Figure 7.4. Finally the fully adaptive algorithm is applied to the problem. Here the error

Table 7.3.. Results for (Ex_5), spatial refinement only

(a) Global refinement

N_{max}	η_h	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
135	-2.48e+03	-3.640e+02	0.15
765	-1.91e+01	-1.019e+01	0.53
5049	-5.34e+00	-2.578e+00	0.48
36465		-5.855e-01	

(b) Adaptive, nondynamic refinement

N_{max}	η_h	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
135	-2.48e+03	-3.640e+02	0.15
765	-1.91e+01	-1.019e+01	0.53
3011	-5.04e+00	-3.289e+00	0.65
13709	-1.02e+00	-1.840e+00	1.81

estimators $\eta_\gamma, \eta_k, \eta_h$ are evaluated and Algorithm 2.2 is used to determine which discretization is to be refined adaptively. The results are displayed in Table 7.4, showing again a good efficiency index. In Figure 7.5(b) the convergence of the error is displayed, together with the errors from the temporal and spatial refinement. For better visual comparability of the convergence rate the starting values of these errors have been normed so that the plots share a starting point. It can be seen that Figure 7.5(b) indicates a better convergence rate for the complete algorithm.

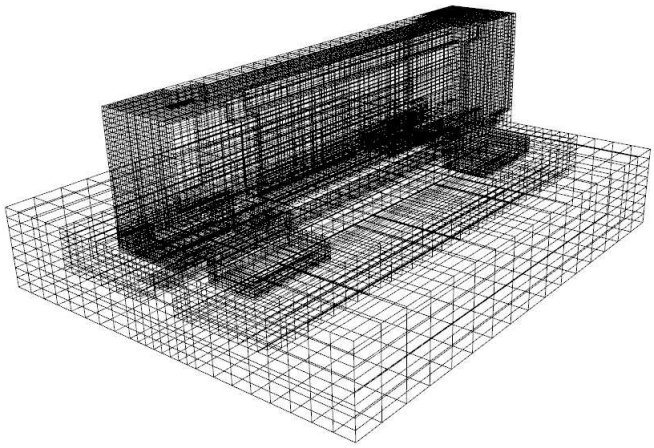
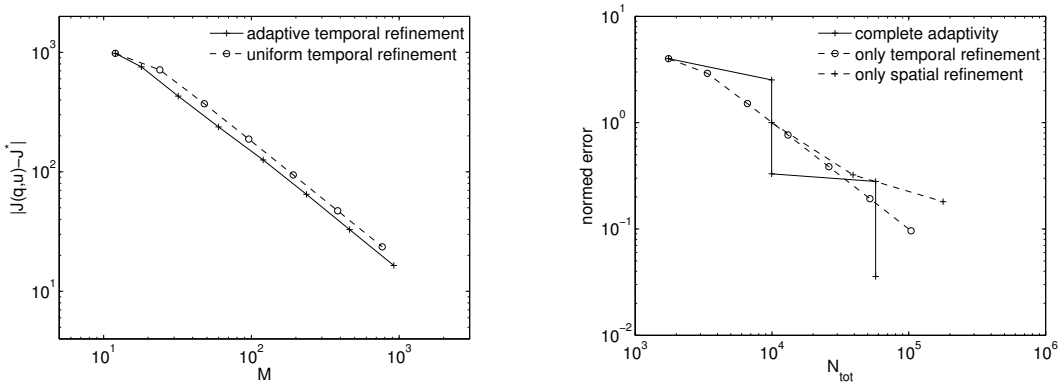


Figure 7.4.. Locally refined mesh for (Ex_5)

Table 7.4.. Results for (Ex_5), complete strategy

N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
135	12	3.0e-01	-2.48e+03	8.16e+02	-7.56e+02	-2.42e+03	-9.570e+02	0.40
765	12	3.0e-01	-1.91e+01	2.32e+02	-8.65e+02	-6.52e+02	-6.032e+02	0.93
765	12	9.5e-01	-3.16e+02	3.37e+02	-8.85e+01	-6.89e+01	-7.891e+01	1.15
3011	18	9.5e-01	2.81e-01	1.27e+01	-9.74e+01	-8.43e+01	-6.695e+01	0.79
3011	18	3.0e+00					-8.544e+00	



(a) Refinement of the temporal discretization

(b) Complete adaptive strategy

Figure 7.5.. Convergence of the error for (Ex_5) for different discretization strategies.

7.5.2. Control of the concrete recipe

The following example is concerned with the control of the concrete recipe. Changing the composition of the concrete mix is a frequent way to manipulate the temperature development, tensile strength or other quantities. According to the models set up in Section 7.2, the control recipe is controlled by the three partial densities of cement, fly ash, and water. Thus the parameter control problem is modelled by setting

$$q = (q_1, q_2, q_3) = (\rho_1, \rho_2, \rho_3) \in \mathbb{R}^3 =: Q.$$

The state equation is then given by

$$\begin{aligned} \tau_t &= g(y) && \text{in } (0;T] \times \Omega \\ c(q)\rho(q)y_t - \lambda(q)\Delta y &= Q_\infty(q) g(y)\, h(\tau, q) && \text{in } (0;T] \times \Omega \\ \tau(0,x) &= 0 && \text{in } \Omega \\ y(0,x) &= y_0 && \text{in } \Omega \\ \frac{\partial}{\partial n} y &= \sigma\,(\bar{y} - y) && \text{on } (0;T] \times \Gamma, \end{aligned} \tag{7.39}$$

with the material models from Section 7.2, specifically (7.5), (7.6), (7.7) for density, thermal conductivity and heat capacity. For the chemical heat source the models (7.13) and (7.16) are chosen, the occurring material parameters are modelled by (7.18), (7.19), (7.20) using the example data from Appendix B. For the following numerical tests the remaining input parameters are chosen as $\bar{y} = 20°C, y_0 = 15°C, \sigma = 20\frac{kJ}{m^2 Kh}$. The considered temporal interval

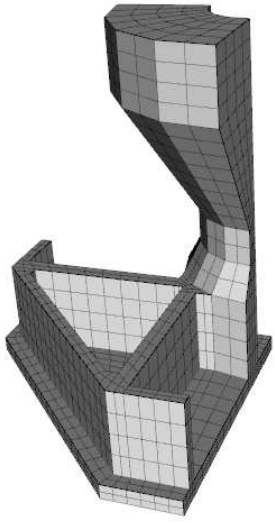


Figure 7.6.. Computational domain for (Ex_6), due to symmetry only one quarter of the platform needs to be considered.

has a length of $T = 48\,h$ and the spatial domain Ω is the platform illustrated in Figure 7.6. The

objective is then to solve the optimal control problem with temperature difference constraints

$$(Ex_6)\begin{cases} \min 500(q_1-400)^2+250(q_2-60)^2+10(q_3-160)^2, & q\in Q,\\ S(q)=(y,\tau) \quad \text{according to (7.39)}, & (7.40)\\ |y(t,x_1)-y(t,x_2)|\le 15\,K \quad \forall t\in[0,T], \end{cases}$$

where $x_1, x_2 \in \bar{\Omega}$ are two given points. The idea behind the chosen values in (Ex_6) is to mimick a frequent problem: The control minimizing the cost functional $(400, 60, 160)$ has a high cement content. This may be cheap, since due to quick heat release the concrete structure is quickly completed. But a too quick heat release can weaken the structure by inducing cracks. This shall be avoided by demanding that the temperature difference constraint

$$|y(t,x_1)-y(t,x_2)|\le 15\,K$$

to be fulfilled. The points x_1 and x_2 are the ones which are going to exhibit the coldest and warmest temperatures, which in these types of concrete constructions are known pretty well a priori. For the example problem (Ex_6) the constants are chosen in such a way that $(400, 60, 160)$ does not fulfill the temperature constraint, and the objective is to find the cheapest control that does.

The computations were carried out with the interior point method with a barrier functional of order $o = 2$, starting regularization parameter $\gamma = 10$, a starting temporal discretization with $M = 24$ equidistant time steps, and a spatial discretization with $N_m = 476$ in every time step. First consider the temporal discretization only. Leaving the spatial discretization and the regularization parameter γ constant, the global temporal refinement is compared to the adaptive refinement driven by the error estimator η_k from (4.45). The results are displayed in Table 7.5 where the estimated optimal cost functional value is $J^* = 8213.3$ and was obtained on a finer temporal discretization than those used for the table. Again, a good quality of the

Table 7.5.. Results for (Ex_6), temporal refinement only

(a) Global refinement

M	η_k	$J^*-J(q_\sigma,u_\sigma)$	I_{eff}
24	-1.40e+03	-2.517e+03	1.80
48	-6.98e+02	-7.953e+02	1.14
96	-3.75e+02	-4.050e+02	1.08
192	-1.97e+02	-2.013e+02	1.02
384	-1.03e+02	-1.038e+02	1.01

(b) Adaptive refinement

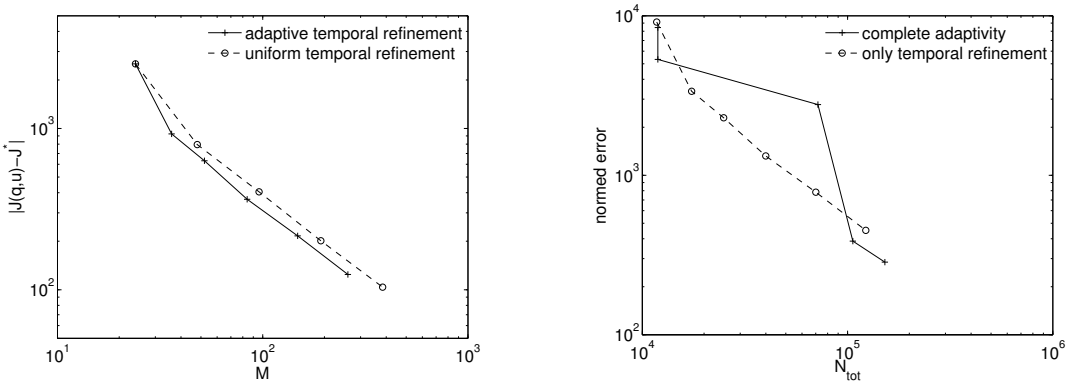
M	η_k	$J^*-J(q_\sigma,u_\sigma)$	I_{eff}
24	-1.40e+03	-2.517e+03	1.80
36	-7.59e+02	-9.258e+02	1.22
52	-5.13e+02	-6.312e+02	1.23
84	-3.34e+02	-3.641e+02	1.09
148	-2.08e+02	-2.158e+02	1.04
260	-1.22e+02	-1.244e+02	1.02

error estimation is obtained. The adaptive strategy leads to a slightly better error convergence than the global one, see also Figure 7.7(a).

Next, the fully adaptive algorithm is applied to the problem. Here the error estimators $\eta_\gamma, \eta_k, \eta_h$ are evaluated and Algorithm 2.2 is used to determine which discretization is to be refined adaptively. The results can be seen in Table 7.6. In Figure 7.5(b) the convergence of the error is displayed, together with the errors from the temporal refinement.

Table 7.6.. Results for (Ex_6), complete strategy

N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
476	24	1.0e+01	-3.26e+03	-1.40e+03	-9.17e+03	-1.383e+04	-8.436e+03	0.61
476	24	3.1e+01	-2.91e+03	-1.21e+03	-7.23e+02	-4.843e+03	-5.327e+03	1.10
2868	24	3.1e+01	-9.28e+01	-1.15e+03	-7.67e+02	-2.010e+03	-2.774e+03	1.38
2868	36	1.0e+02	-9.60e+01	-3.28e+02	-9.89e+01	-5.229e+02	-3.863e+02	0.83
2868	52	1.0e+02					-2.851e+02	



(a) Refinement of the temporal discretization

(b) Complete adaptive strategy

Figure 7.7.. Convergence of the error for (Ex_6) for different discretization strategies.

7.5.3. Control of the flow rate of a water cooling system

This section deals with the control of a water cooling system. The control variable set as

$$q = (q_6) = w(t) \in L^2(I) =: Q$$

gives the deduced heat by equations (7.23) and (7.24). The models (7.13) and (7.16) are chosen for the chemical heat source, and the material parameters set as constants according to a reference concrete recipe

$$\begin{aligned} c = 1000\tfrac{J}{kg\,K}, \quad \rho = 2000\tfrac{kg}{m^3}, \quad \lambda = 2.143\tfrac{W}{m\,K}, \\ Q_\infty = 293.2\tfrac{kJ}{kg}, \qquad a_w = -11, \qquad b_w = -1. \end{aligned}$$

Further, the values for the occuring temperatures are chosen as $y_0 = 15°C, \bar{y} = 20°C, y_c = 10°C$ and the heat transfer coefficient $\sigma = 8.33\frac{W}{m^2\,K}$. So the state equation is

$$\begin{aligned} \tau_t &= g(y) && \text{in } (0;T] \times \Omega \\ c\rho y_t - \lambda \Delta y &= Q_\infty g(y)\, h(\tau) - \dot{Q}_p(q) && \text{in } (0;T] \times \Omega \\ \tau(0,x) &= 0 && \text{in } \Omega \\ y(0,x) &= y_0 && \text{in } \Omega \\ \frac{\partial}{\partial n} y &= \sigma\,(\bar{y} - y) && \text{on } (0;T] \times \Gamma. \end{aligned} \tag{7.41}$$

The considered time interval has a length of $T = 96\,h$. The domain Ω is a wall that is erected on a foundation, see Figure 7.2. Since heat is only produced in the part of the domain that is labeled as „new concrete“ in Figure 7.2, $g(y)$ and $h(\tau)$ are set to zero on the „foundation“ part of the domain. The objective is then to solve the optimal control problem with upper temperature constraints

$$(Ex_7)\begin{cases} \min\|q\|^2_{L^2(0,T)}, & q \in Q, \\ S(q) = (y,\tau) & \text{according to (7.41)}, \\ y(t,x) \leq 57°C & . \end{cases} \tag{7.42}$$

Note that the data of (Ex_7) are chosen in such a way that $q \equiv 0$ is not a feasible control, that means no cooling would violate the temperature constraint. Thus the most efficient cooling profile is searched that obeys the temperature constraint.
The computations were carried out with an interior point method with order $o = 2$, see Section 4.4, starting regularization parameter $\gamma = 5$, a starting temporal discretization with $M = 6$ equidistant time steps, and an equidistant spatial discretization with $N_m = 765$ in every time step.
First compare the global refinement strategy that refines all components uniformly with the fully adaptive strategy, that first chooses the component(s) with substantial error contribution according to Algorithm 2.2, and then refines these locally. In the spatial discretization, the non-dynamic approach is used first. The results, using the estimated value $J^* = 39.683440762$, can be seen in Table 7.7 and Figure 7.8(a) The error estimation yields efficiency indices not

Table 7.7.. Results for (Ex_7) for simultanous spatial and temporal refinement

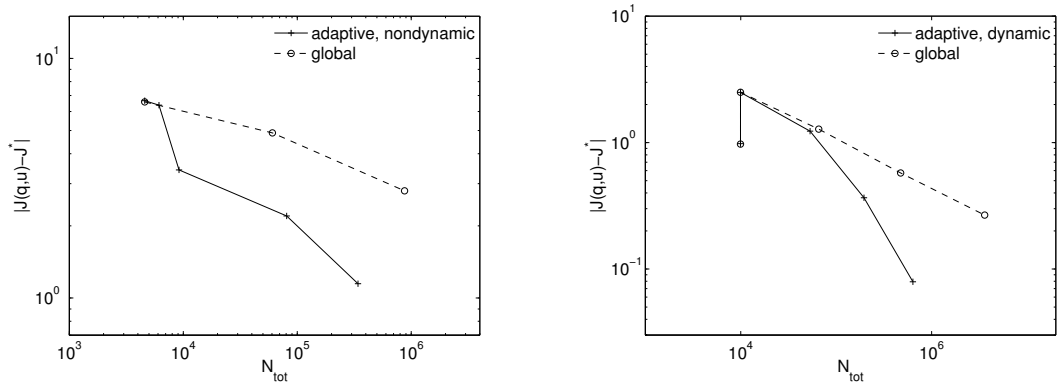
(a) global refinement of spatial and temporal discretization

N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
765	6	5.0e+00	9.36e-01	4.05e+00	-5.58e+00	-5.94e-01	6.763e+00	-11.39
5049	12	1.6e+01	1.01e+00	5.73e+00	-5.87e-01	6.15e+00	5.071e+00	0.82
36465	24	5.0e+01	4.21e-01	3.98e+00	-6.36e-02	4.34e+00	2.983e+00	0.69

(b) adaptive, non-dynamic refinement of spatial and temporal discretization

N_{max}	M	γ	η_h	η_k	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
765	6	5.0e+00	1.01e+00	3.99e+00	-5.58e+00	-0.58e-01	6.846e+00	-11.80
765	8	1.6e+01	1.65e+00	5.05e+00	-5.78e-01	6.12e+00	6.563e+00	1.07
765	12	1.6e+01	2.48e+00	2.78e+00	-6.00e-01	4.67e+00	3.606e+00	0.77
5049	16	1.6e+01	1.50e+00	1.29e+00	-6.13e-01	2.17e+00	2.384e+00	1.10
18965	18	1.6e+01					1.330e+00	

far from 1. The adaptive strategy leads to a considerably faster convergence of the error. During the repeated use of the error equilibration algorithm Algorithm 2.2, all the components (regularization, spatial, temporal discretization) have been refined at least once. This allows for the following consideration: Compare the second line of Table 7.7(a) and the third line of Table 7.7(b). At the same regularization parameter, the adaptive strategy uses a temporal discretization with the same number of subintervals as the global strategy. But although the adaptive spatial discretization uses fewer nodes than the global one, its discretization error is



(a) Global vs. nondynamic adaptive strategy, all components can be refined

(b) Global vs. dynamic strategy, fixed temporal discretization

Figure 7.8.. Convergence of the error for (Ex_7) for different discretization strategies.

lower. The reason for this is that the $M = 12$ subintervals created by the adaptive strategy are not equidistant, but instead resolve the time period better where large temperatures occur, thus reducing the temporal discretization error greatly. The development of the temporal discretization can be seen in Figure 7.9.

In a second test, the dynamic approach of the spatial discretization is to be investigated. Since the tests showed a strong focus of the effects on the time discretization around the point where the maximum temperature is reached, the comparison is done by using a constant temporal discretization with $M = 12$ equidistant time steps that is not refined during the process. The process of error equilibration is now executed with the spatial discretization and regularization errors of a semidiscrete problem (P_k), resulting in only the error estimators η_h, η_γ being used. The estimated optimal cost functional value, now that for (P_k), is $J^* = 35.85$. The results are displayed in Table 7.8 and Figure 7.8(b). The dynamic discretization strategy yields faster convergence than the global one. In fact, a better order of convergence is achieved. Two examples for grids from the dynamic discretization approach can be seen in Figure 7.10, representing the meshes with the most and the fewest nodes. To investigate the distribution of the number of spatial nodes over time, a numerical test with a constant temporal discretization with $M = 48$ equidistant intervals was carried out. The results are graphically displayed in Figure 7.11(a). One notes a large number in the first time step, that can be attributed to the initial condition singularity, the mismatch between y_0 and $\bar{y}$. Apart from this one time point, the numbers are fairly low and change only moderately from time step to time step, except for a spike in the middle of the time interval around $t = 30\,h$. Note that the need for a finer spatial discretization occurs some time after the temperature maximum, compare Figure 7.11(b). Remember that this time point of maximum temperature was the one where the temporal discretization needed to be refined according to the error estimator. This difference is somewhat surprising and illustrates the fact that even for practical problems intuition is not always right when dealing with the question where the local refinement is to be executed.

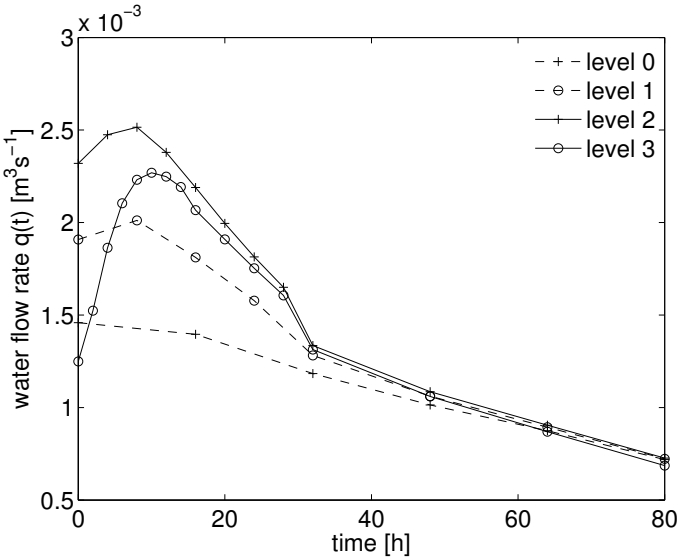


Figure 7.9.. Development of the mesh-optimal water flow rate when using adaptive discretization. Note the refinement of time intervals around $t = 16\,h$ only.

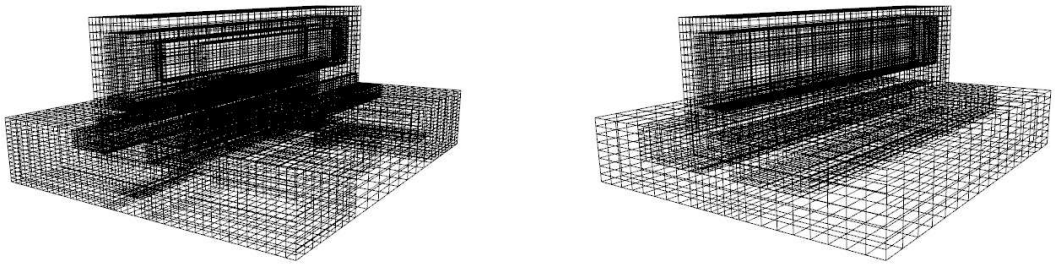
Table 7.8.. Results for (Ex_7) for spatial refinement only, $M = 12$

(a) global refinement of the spatial discretization

N_{tot}	N_{max}	γ	η_h	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
9945	765	5.0e+00	1.75e+00	-5.80e+00	-4.05e+00	-9.806e-01	-0.24
9945	765	1.6e+00	1.69e+00	-5.81e-01	1.11e+00	2.498e+00	2.25
65637	5049	1.6e+00	1.09e+00	-5.87e-01	5.04e-01	1.279e+00	2.54
474045	36465	5.0e+01	2.90e-01	-5.94e-02	2.31e-01	5.753e-01	2.49
3597165	276705	5.0e+01				2.672e-01	

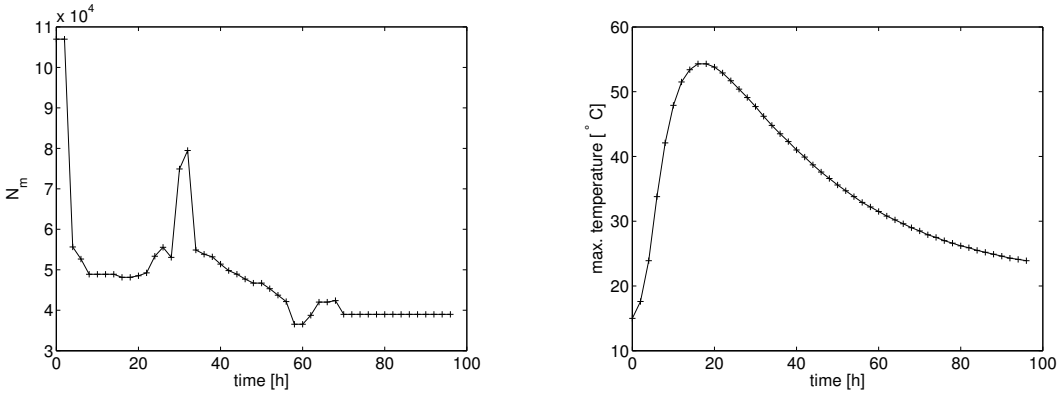
(b) adaptive, dynamic refinement of the spatial discretization

N_{tot}	N_{max}	γ	η_h	η_γ	η	$J^* - J(q_\sigma, u_\sigma)$	I_{eff}
9945	765	5.0e+00	1.75e+00	-5.80e+00	-4.05e+00	-9.806e-01	-0.24
9945	765	1.6e+00	1.69e+00	-5.81e-01	1.11e+00	2.498e+00	2.25
53309	5049	1.6e+01	1.08e+00	-5.88e-01	4.90e-01	1.233e+00	2.51
195433	19957	5.0e+01	2.77e-01	-5.97e-02	2.18e-01	3.661e-01	1.68
636469	72311	5.0e+01				7.926e-02	



(a) at time $t = 24\,h$, with $N_3 = 72311$ (b) at time $t = 96\,h$, with N_{12}=39451

Figure 7.10.. Two spatial discretization meshes for (Ex_7)



(a) Number of nodes N_m over time (b) For comparison: maximum temperature inside the structure $T_{\max}$ over time

Figure 7.11.. Distribution of the number of nodes of the spatial discretization on the time intervals when using dynamic spatial discretization, but constant temporal discretization for (Ex_7)

8. Summary

This thesis was concerned with the development of efficient numerical solution strategies for elliptic and parabolic optimal control problems (OCPs) with pointwise state constraints.

The main analytical difficulty was hereby caused by the reduced regularity induced by the state constraint. This needed to be accounted for in the functional analytic setting of the problem, theorems on the existence and uniqueness of optimal solutions and the derivation of optimality conditions.

Two optimization strategies were proposed for the numerical solution of the OCPs at hand. The first is a primal-dual active set method that can be applied to the optimality system, reduced to the control and multiplier variables, directly. This method was described in detail for elliptic problems. A disadvantage is that it is only applicable to a certain class of OCPs. The second optimization strategy is an interior point algorithm applied to a regularized variant of the original problem. It was presented extensively for parabolic problems. The introduction of an additional regularization parameter can here be seen as a disadvantage.

For the numerical solution of the problems, the governing equations were discretized by Galerkin finite element methods. If the PDAS optimization method is used, this leads to the consideration of discrete Borel measures, which poses an additional difficulty in the implementation. The main point of the thesis here was however the choice of efficient discretizations. To that end, estimators for the error with respect to the cost functional were developed, based on the DWR method. Their contributions, potentially spatial, temporal, control and regularization error estimators, were used in an error equilibration algorithm. Furthermore, localizations of the temporal and spatial estimators were used in an adaptive algorithm creating locally refined meshes. By these means an improvement of the convergence speed of the numerical solution was to be achieved.

The efficiency of the developed algorithms was illustrated on several numerical examples. Especially promising in practical regard is the application to the optimal control of young concrete thermo-mechanical properties.

Acknowledgments

The work for this thesis was carried out during my time as a research assistant at the Johann Radon Institute for Computational and Applied Mathematics (RICAM) in Linz, Austria and later at the the Technische Universität München, Germany. For the financial support within the project „Numerical analysis and discretization strategies for optimal control problems with singularities“ I would like to thank the Austrian Science Fund (FWF, Project No. P18971-N18) and the Deutsche Forschungsgemeinschaft DFG with its priority program 1253 „Optimierung mit partiellen Differentialgleichungen“.

For the scientific part of this work I would like to express my gratitude to my supervisor Prof. Dr. Boris Vexler for the suggestion of this interesting topic, his advice, help, motivation, patience and critical remarks.
Furthermore I am very grateful to Prof. Dr. Thomas Apel for reviewing this thesis, and for sharing his knowledge with me during the research.

A special credit is due to the developers of the software packages used for the numerical experiments in this thesis. Over the past years, many people have contributed to RoDoBo, Gascoigne and VisuSimple, they all deserve my thanks.

Furthermore I would like to thank my colleagues in Linz and in Munich for the collaboration on mathematical topics and for the enjoyable time spent together inside and outside of the workplace.

Last but not least I thank my parents and my sister for the support they gave me over the years in every kind of way.

A. Convergence order for the Laplace equation with irregular data

This appendix considers the use of graded meshes in the approximative solution of elliptic differential equations with irregular right hand side. The quintessential content of Appendix A has been published in [3], in the creation of which the author of this thesis was also involved. In contrast to the often considered L^2-right hand sides, permitting less regular ones reduces also the regularity of the solution and thus the approximation order on uniform meshes. The use of graded meshes proves to be a remedy. Thus, following the discussion in Section 2.5, motivates a similar procedure for the solution of optimal control problems with additional state constraints.

Consider the elliptic boundary value problem

$$-\Delta u = \delta_a \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega, \tag{A.1}$$

with a convex polygonal domain $\Omega \subset \mathbb{R}^2$, and δ_a denoting the Dirac measure concentrated in the point $a \in \operatorname{int}(\Omega)$. Since this problem does not have an $H^1(\Omega)$-solution, consider the solution u in the space

$$W_0^{1,q}(\Omega) := \{v \in W^{1,q}(\Omega) \,:\, v = 0 \text{ on } \partial\Omega \text{ in the sense of } L^q(\partial\Omega)\},$$

$q \in [1, 2)$, defined via

$$(\nabla u, \nabla v) = v(a) \quad \forall v \in W_0^{1,q'}(\Omega) \tag{A.2}$$

where $q' > 2$ satisfies $1/q + 1/q' = 1$. If (A.2) is approximated by the finite element method by

$$(\nabla u_h, \nabla v_h) = v_h(a) \quad \forall v_h \in V_h. \tag{A.3}$$

where V_h is the space of linear finite element functions corresponding to a mesh $\mathcal{T}_h$ from a family of quasi-uniform triangulations, then the error of the finite element approximation in the L^2-Norm converges only with order h^1, [88], as opposed to h^2 that would be obtained for a regular right hand side in (A.1). The goal of this section is to prove that the use of specially designed meshes improves the convergence order to almost the original rate, precisely $h^2|\ln h|^{3/2}$.

Thus let $(\mathcal{T}_h)$ be a family of shape-regular triangulations of Ω that is *graded* with grading parameter $\mu = \frac{1}{2}$ towards the point $a \in \operatorname{int}(\Omega)$, i.e. for every cell $T \in \mathcal{T}_h$ the cell diameter h_T depends on the distance r_T of the cell T from the point a by

$$h_T \sim \begin{cases} h r_T^{1/2} : & r_T > 0 \\ h^2 : & r_T = 0 \end{cases} \tag{A.4}$$

W.l.o.g assume that

$$h \leq h_0 < 1$$

holds in order to ensure that $\ln h$ does not change sign. The domain Ω is split into the sets

$$\Omega_0 = \bigcup_{r_T=0} T \text{ and } \Omega_1 = \Omega \setminus \Omega_0$$

and an element $T^* \in \Omega_0$ chosen. Its diameter is $h_* \sim h^2$.

The main result will be derived from a theorem that considers the application of these graded meshes to problems with regular right hand side. Thus consider the Poisson problem with a right-hand side $f \in L^2(\Omega)$,

$$-\Delta z = f \text{ in } \Omega, \quad z = 0 \text{ on } \partial\Omega, \tag{A.5}$$

and state the following theorem:

Theorem A.1. *Let $f \in L^2(\Omega)$, $z \in H_0^1(\Omega) \cap H^2(\Omega)$ be the solution of problem* (A.5) *and z_h be a finite element approximation of z in the space of linear finite elements V_h using a mesh that is graded according to condition* (A.4)*. Then the a priori estimate*

$$|(z - z_h)(a)| \leq ch^2 |\ln h|^{3/2} \|z\|_{H^2(\Omega)}$$

holds for all $h \leq h_0$.

With this result the main result can be proven quickly:

Corollary A.2. *Let u be the solution of* (A.1) *and $u_h \in V_h$ its finite element approximation defined via* (A.3) *on a family of meshes that are graded according to condition* (A.4)*. Then the a priori estimate*

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^2 |\ln h|^{3/2}$$

holds for all $h \leq h_0$.

Proof. Denoting the error by $e := u - u_h$, we define the function $v \in H_0^1(\Omega)$ as the solution of

$$(\nabla v, \nabla \varphi) = (e, \varphi) \quad \forall \varphi \in H_0^1(\Omega),$$

i.e. the weak solution of the boundary value problem

$$-\Delta v = e \text{ in } \Omega, \quad v = 0 \text{ on } \partial\Omega.$$

Note that $v \in H^2(\Omega) \hookrightarrow W^{1,p}(\Omega)$ holds for any $p < \infty$. Its finite element approximation $v_h \in V_h$ is defined by

$$(\nabla v_h, \nabla \varphi_h) = (e, \varphi_h) \quad \forall \varphi_h \in V_h.$$

With these auxiliary quantities we can estimate $\|e\|_{L^2(\Omega)}$ by utilizing Theorem A.1

$$\begin{aligned}
\|u - u_h\|_{L^2(\Omega)}^2 &= \|e\|_{L^2(\Omega)}^2 = (e, u) - (e, u_h) \\
&= (\nabla v, \nabla u) - (\nabla v, \nabla u_h) \\
&= v(a) - v_h(a) = (v - v_h)(a) \\
&\leq ch^2 |\ln h|^{3/2} \|\nabla^2 v\|_{L^2(\Omega)} \\
&\leq ch^2 |\ln h|^{3/2} \|e\|_{L^2(\Omega)}.
\end{aligned}$$

Dividing this inequality by $\|u - u_h\|_{L^2(\Omega)}$ gives the desired result. □

As first step in the proof of Theorem A.1 define the weight function $\sigma : \Omega \to \mathbb{R}$ by

$$\sigma(x) := (|x-a|^2 + h_*^2)^{1/2} \tag{A.6}$$

The following properties of σ can be proven by calculation.

Lemma A.3. *For the function σ defined in* (A.6) *the inequalities*

$$\begin{aligned} |\sigma| + |\nabla\sigma| &\le c \\ |\nabla^2\sigma| &\le c\sigma^{-1} \\ \sigma^{-1}(x) &\le \begin{cases} h_*^{-1} & \text{if } \ x \in \{T \in \mathcal{T}_h \,:\, r_T = 0\}, \\ cr_T^{-1} & \text{if } \ x \in \{T \in \mathcal{T}_h \,:\, r_T > 0\} \end{cases} \end{aligned} \tag{A.7}$$

are valid.

For functions with elementwise H^2-regularity the notation $\nabla_h v \in L^2(\Omega)$ and $\nabla_h^2 v \in L^2(\Omega)$ given through

$$\nabla_h v|_T = \nabla v|_T \quad \text{ and } \quad \nabla_h^2 v|_T = \nabla^2 v|_T$$

will be used. The nodal interpolant of a function $v \in H_0^1(\Omega) \cap C(\bar{\Omega})$ is denoted by $\mathcal{I}_h v \in V_h$. We show the following estimate of a weighted interpolation error.

Lemma A.4. *For any function v from the set*

$$\{v \in H_0^1(\Omega) \cap C(\bar{\Omega}) \,:\, v \in H^2(T)\ \forall T \in \mathcal{T}_h\}$$

the estimate

$$\|\sigma^{-1/2}\nabla(v - \mathcal{I}_h v)\|_{L^2(\Omega)} \le ch\|\nabla_h^2 v\|_{L^2(\Omega)}$$

holds on meshes of type (A.4)*. For functions $v \in H_0^1(\Omega) \cap H^2(\Omega)$ this results in*

$$\|\sigma^{-1/2}\nabla(v - \mathcal{I}_h v)\|_{L^2(\Omega)} \le ch\|\nabla^2 v\|_{L^2(\Omega)}.$$

Proof. One can calculate by using (A.7)

$$\begin{aligned} \|\sigma^{-1/2}\nabla(v - \mathcal{I}_h v)\|_{L^2(\Omega)}^2 &= \sum_{T\subset\Omega_0} \int_T \sigma^{-1}|\nabla(v - \mathcal{I}_h v)|^2 + \sum_{T\subset\Omega_1} \int_T \sigma^{-1}|\nabla(v - \mathcal{I}_h v)|^2 \\ &\le \sum_{T\subset\Omega_0} ch_*^{-1}h_*^2\|\nabla^2 v\|_{L^2(T)}^2 + \sum_{T\subset\Omega_1} cr_T^{-1}h_T^2\|\nabla^2 v\|_{L^2(T)}^2 \\ &\le \sum_{T\subset\Omega} ch^2\|\nabla^2 v\|_{L^2(T)}^2. \end{aligned}$$

This proves the assertion. □

Lemma A.5. *For any function $v \in H_0^1(\Omega) \cap H^2(\Omega)$ the inequality*

$$\|\nabla(v - \mathcal{I}_h v)\|_{L^2(\Omega)} \le c\|\sigma\nabla^2 v\|_{L^2(\Omega)}$$

holds provided the mesh is graded according to (A.4)*.*

Proof. With the help of the function σ we can estimate the element size on the two subdomains. On Ω_0 there follows directly

$$h_*^2 \le \sigma^2(x) \qquad \forall x \in \Omega_0. \tag{A.8}$$

On Ω_1 one has $\sigma(x) \ge r_T$ and $\sigma(x) \ge h_*$. Since there holds $h_T \sim h r_T^{1/2}$ the relation $h_T^2 \sim h^2 r_T \sim h_* r_T$ is used to conclude

$$h_T^2 \le c\sigma^2(x) \qquad \forall x \in \Omega_1. \tag{A.9}$$

Now we can estimate

$$\|\nabla(v - \mathcal{I}_h v)\|_{L^2(\Omega)}^2 \le c \sum_T \int_T h_T^2 |\nabla^2 v|^2 = c \sum_{T\subset\Omega_0} \int_T h_*^2 |\nabla^2 v|^2 + c \sum_{T\subset\Omega_1} \int_T h_T^2 |\nabla^2 v|^2.$$

With the estimates (A.8), (A.9) one can continue with

$$\|\nabla(v - \mathcal{I}_h v)\|_{L^2(\Omega)}^2 \le c \sum_T \int_T \sigma^2 |\nabla^2 v|^2 = c\|\sigma \nabla^2 v\|_{L^2(\Omega)}^2,$$

and the assertion is proved. □

Lemma A.6. *Let the function $y \in H_0^1(\Omega) \cap H^2(\Omega)$ be the solution of the boundary value problem*

$$-\Delta y = w \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega \tag{A.10}$$

with a given right-hand side $w \in L^2(\Omega)$. Then for $h \le h_0$ the estimate

$$\|\sigma \nabla^2 y\|_{L^2(\Omega)} \le c|\ln h| \|\sigma w\|_{L^2(\Omega)}$$

holds, where σ is the weight function defined in (A.6).

Proof. Set $\xi := x - a$ and denote by ξ_1, ξ_2 its components. By the chain rule it holds

$$\|\xi_i \nabla^2 y\|_{L^2(\Omega)} \le \|\nabla^2(\xi_i y)\|_{L^2(\Omega)} + c\|\nabla y\|_{L^2(\Omega)}, \qquad i = 1, 2.$$

With the definition of σ and the a priori estimate $\|\nabla^2 y\|_{L^2(\Omega)} \le c\|\Delta y\|_{L^2(\Omega)}$ this yields

$$\begin{aligned}
\|\sigma \nabla^2 y\|_{L^2(\Omega)}^2 &= \sum_{i=1}^2 \|\xi_i \nabla^2 y\|_{L^2(\Omega)}^2 + h_*^2 \|\nabla^2 y\|_{L^2(\Omega)}^2 \\
&\le \sum_{i=1}^2 \left(\|\nabla^2(\xi_i y)\|_{L^2(\Omega)}^2 + c\|\nabla y\|_{L^2(\Omega)}^2 \right) + c h_*^2 \|\Delta y\|_{L^2(\Omega)}^2.
\end{aligned}$$

With the use of $h_* \le \sigma$ we continue

$$\begin{aligned}
\|\sigma \nabla^2 y\|_{L^2(\Omega)}^2 &\le c \sum_{i=1}^2 \|\Delta(\xi_i y)\|_{L^2(\Omega)}^2 + c\|\nabla y\|_{L^2(\Omega)}^2 + c\|\sigma \Delta y\|_{L^2(\Omega)}^2 \\
&\le c \sum_{i=1}^2 \|\xi_i \Delta y\|_{L^2(\Omega)}^2 + c\|\nabla y\|_{L^2(\Omega)}^2 + c\|\sigma w\|_{L^2(\Omega)}^2 \\
&\le c\|\sigma \Delta y\|_{L^2(\Omega)}^2 + c\|\nabla y\|_{L^2(\Omega)}^2 + c\|\sigma w\|_{L^2(\Omega)}^2 \\
&\le c\|\sigma w\|_{L^2(\Omega)}^2 + c\|\nabla y\|_{L^2(\Omega)}^2,
\end{aligned} \tag{A.11}$$

where we have used inequality (A.7) and the definition (A.10) of y. It remains to show that $\|\nabla y\|_{L^2(\Omega)} \leq |\ln h| \|\sigma w\|_{L^2(\Omega)}$. Start with the estimation

$$\|\nabla y\|^2_{L^2(\Omega)} = |(\Delta y, y)| \leq \|\sigma \Delta y\|_{L^2(\Omega)} \|\sigma^{-1} y\|_{L^2(\Omega)} = \|\sigma w\|_{L^2(\Omega)} \|\sigma^{-1} y\|_{L^2(\Omega)}. \tag{A.12}$$

The last factor will be estimated by using its representation in polar coordinates (r, θ) with respect to a. In the following we use the observation

$$\sigma(r) = \left(r^2 + h_*^2\right)^{\frac{1}{2}} \quad \Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}r}(\ln \sigma(r) - \ln \sigma(0)) = \frac{r}{\sigma^2} \tag{A.13}$$

and the inequality

$$\left| \frac{\ln \sigma(r) - \ln \sigma(0)}{r} \right| \leq \frac{c}{\sigma} |\ln h| \quad \text{for } h \leq h_0, \tag{A.14}$$

which is proved later. Furthermore for simplicity of notation we replace the integration domain Ω by a disc of radius $R = \operatorname{diam}(\Omega) \geq 1$ with the center in a, such that this disc contains Ω. We continue the function y with $y = 0$ outside the domain Ω such that this extension of the domain does not change the value of any quantities involved. With the observation (A.13), partial integration with respect to the radius r, and estimate (A.14) one can conclude

$$\begin{aligned}
\|\sigma^{-1} y\|^2_{L^2(\Omega)} &= \int_\Omega \sigma^{-2} y^2 \,\mathrm{d}x = \int_0^{2\pi} \int_0^R r \sigma^{-2} y^2 \,\mathrm{d}r \,\mathrm{d}\theta \\
&= \int_0^{2\pi} \int_0^R \frac{|\ln \sigma(r) - \ln \sigma(0)|}{r} r \, 2y \partial_r y \,\mathrm{d}r \,\mathrm{d}\theta \\
&\leq \int_0^{2\pi} \int_0^R \frac{c}{\sigma} |\ln h| r |y \partial_r y| \,\mathrm{d}r \,\mathrm{d}\theta \\
&\leq c |\ln h| \int_0^{2\pi} \int_0^R \sigma^{-1} r |y| |\nabla y| \,\mathrm{d}r \,\mathrm{d}\theta \\
&\leq c |\ln h| \|\sigma^{-1} y\|_{L^2(\Omega)} \|\nabla y\|_{L^2(\Omega)}.
\end{aligned}$$

Dividing by $\|\sigma^{-1} y\|_{L^2(\Omega)}$ yields

$$\|\sigma^{-1} y\|_{L^2(\Omega)} \leq c |\ln h| \|\nabla y\|_{L^2(\Omega)}.$$

Inserting this into equation (A.12) and dividing by $\|\nabla y\|_{L^2(\Omega)}$ yields

$$\|\nabla y\|_{L^2(\Omega)} \leq c |\ln h| \|\sigma w\|_{L^2(\Omega)}$$

and thus with (A.11) the claim of the lemma.

It remains to prove inequality (A.14). To this end, we distinguish the cases $r > h_*$ and $r \leq h_*$ and begin with the case $r > h_*$. Since $\sigma(r)$ is strictly monotone and positive the function

$|\ln \sigma(r)|$ takes its maximum at the left or right boundary of $[0, R]$. For $h \le h_0$ these values can be estimated by

$$|\ln \sigma(0)| = |\ln h_*| \le c|\ln h| \text{ and} \tag{A.15}$$

$$|\ln \sigma(R)| = |\ln \sqrt{R^2 + h_*^2}| \le c|\ln h|, \tag{A.16}$$

since $\ln \sqrt{R^2 + h_*^2} \le \ln \sqrt{R^2 + h_0^2} = c|\ln h_0| \le c|\ln h|$ for $c = \ln \sqrt{R^2 + h_0^2}/|\ln h_0|$. Thus it follows

$$|\ln \sigma(r) - \ln \sigma(0)| \le 2 \max_{0 \le r \le R} |\ln \sigma(r)| \le c|\ln h|,$$

again for $h \le h_0$. Since it is $1/r \le c/\sigma$ the inequality (A.14) is proved.

For the case $r \le h_*$ we can conclude by the mean value theorem

$$\left| \frac{\ln \sigma(r) - \ln \sigma(0)}{r} \right| \le \max_{0 \le s \le h_*} |(\ln \sigma)'(s)| = \max_{0 \le s \le h_*} \frac{s}{\sigma(s)^2}.$$

As the last function is monotonically increasing on $[0, h_*]$ it takes its maximum at the end of the interval. This means by using $h_* \le \sigma(r) \le \sqrt{2} h_*$

$$\left| \frac{\ln \sigma(r) - \ln \sigma(0)}{r} \right| \le \frac{h_*}{2h_*^2} \le \frac{\sqrt{2}}{2} \frac{1}{\sigma}$$

and inequality (A.14) is also proved in this case. □

For our further considerations we introduce a regularized Dirac function by

$$\delta^h := \begin{cases} |T^*|^{-1} \operatorname{sign}(z - z_h) & \text{in } T^*, \\ 0 & \text{elsewhere,} \end{cases}$$

where z is the solution of (A.5) and z_h is the corresponding finite element approximation from Theorem A.1. Notice that $\delta^h \in L^2(\Omega)$. The corresponding regularized Green function $g^h \in H_0^1(\Omega) \cap H^2(\Omega)$ is defined by

$$-\Delta g^h = \delta^h \text{ in } \Omega, \qquad g^h = 0 \text{ on } \partial\Omega. \tag{A.17}$$

Also, consider the function $g_h^h \in V_h$ as the Ritz projection of g^h onto V_h, i.e.,

$$(\nabla g_h^h, \nabla \varphi_h) = (\nabla g^h, \nabla \varphi_h) \quad \forall \varphi_h \in V_h. \tag{A.18}$$

Lemma A.7. *For the regularized Green function g^h defined in* (A.17) *the estimate*

$$\|\sigma \nabla^2 g^h\|_{L^2(\Omega)} \le c\,|\ln h|^{1/2}$$

holds for $h \le h_0$.

Proof. The assertion follows from setting $\rho = h_*$ in [33, Theorem B4]. In this paper, a $C^{1,1}$-domain Ω is considered but this assumption is not necessary for the result of this lemma. □

Lemma A.8. *For the regularized Green function g^h and its Ritz projection g_h^h defined in* (A.17) *and* (A.18)*, respectively, the estimate*

$$\|\sigma^{-1}(g^h - g_h^h)\|_{L^2(\Omega)} \le c|\ln h|^{3/2}$$

holds for $h \le h_0$.

Proof. We introduce the abbreviation $e_g := g^h - g_h^h$ and consider the auxiliary equation

$$-\Delta y = \frac{\sigma^{-2} e_g}{\|\sigma^{-1} e_g\|_{L^2(\Omega)}} \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega.$$

Its weak form can be written as

$$(\nabla y, \nabla \varphi) = \frac{(\sigma^{-1} e_g, \sigma^{-1}\varphi)}{\|\sigma^{-1} e_g\|_{L^2(\Omega)}} \quad \forall \varphi \in H_0^1(\Omega).$$

The choice $\varphi = e_g$ yields

$$\|\sigma^{-1} e_g\|_{L^2(\Omega)} = (\nabla e_g, \nabla y) = (\nabla e_g, \nabla(y - \mathcal{I}_h y)) \le \|\nabla e_g\|_{L^2(\Omega)} \|\nabla(y - \mathcal{I}_h y)\|_{L^2(\Omega)}. \quad \text{(A.19)}$$

For the first term of the right-hand side we use Lemma A.5 with the choice $v = g^h$ and conclude with the result from Lemma A.7

$$\|\nabla e_g\|_{L^2(\Omega)} \le c \|\nabla(g^h - \mathcal{I}_h g^h)\|_{L^2(\Omega)} \le c \|\sigma \nabla^2 g^h\|_{L^2(\Omega)} \le c |\ln h|^{1/2}. \quad \text{(A.20)}$$

For the second term on the right-hand side of inequality (A.19) we write with the Lemmas A.5 and A.6

$$\|\nabla(y - \mathcal{I}_h y)\|_{L^2(\Omega)} \le c \|\sigma \nabla^2 y\|_{L^2(\Omega)} \le c |\ln h| \left\| \sigma \frac{\sigma^{-2} e_g}{\|\sigma^{-1} e_g\|} \right\|_{L^2(\Omega)} = c |\ln h|. \quad \text{(A.21)}$$

Inequality (A.19) yields together with estimates (A.20) and (A.21) the assertion of this lemma. □

Lemma A.9. *For the regularized Green function g^h and its Ritz projection g_h^h defined in* (A.17) *and* (A.18)*, respectively, the inequality*

$$\|\nabla_h^2(\sigma(g^h - g_h^h))\| \le c |\ln h|^{3/2}$$

is satisfied for $h \le h_0$.

Proof. We use again the abbreviation $e_g := g^h - g_h^h$, apply the product rule on every element $T \in \mathcal{T}_h$ and get

$$\nabla^2(\sigma e_g)|_T = (\nabla^2 \sigma) e_g|_T + 2 \nabla \sigma|_T \cdot \nabla e_g|_T + \sigma(\nabla^2 e_g)|_T.$$

This results with Lemma A.3 in the estimate

$$\|\nabla_h^2(\sigma e_g)\|_{L^2(\Omega)}^2 \le c \left(\|\sigma^{-1} e_g\|_{L^2(\Omega)}^2 + \|\nabla e_g\|_{L^2(\Omega)}^2 + \|\sigma(\nabla_h^2 e_g)\|_{L^2(\Omega)}^2 \right). \quad \text{(A.22)}$$

The first term of the right-hand side of this inequality is estimated in Lemma A.8, giving a contribution of $c|\ln h|^3$. The second term is estimated in (A.20). Since the equality $\nabla^2(g_h^h|_T) = 0$ holds for linear elements on every element T it follows for the third term with application of Lemma A.7

$$\|\sigma(\nabla_h^2 e_g)\|_{L^2(\Omega)}^2 = \|\sigma \nabla^2 g^h\|_{L^2(\Omega)}^2 \le c |\ln h|. \quad \text{(A.23)}$$

This means, Lemma A.8 yields together with the inequalities (A.22), (A.20) and (A.23) the assertion. □

Lemma A.10. *For the regularized Green function g^h and its Ritz projection g_h^h defined in* (A.17) *and* (A.18) *the inequality*

$$\|\sigma^{1/2}\nabla(g^h - g_h^h)\|_{L^2(\Omega)} \leq ch|\ln h|^{3/2}$$

holds for $h \leq h_0$.

Proof. We use the abbreviation $e_g := g^h - g_h^h$. With the product rule we observe

$$\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 = (\nabla e_g, \sigma\nabla e_g) = (\nabla e_g, \nabla(\sigma e_g)) - (\nabla e_g, e_g\nabla\sigma). \tag{A.24}$$

For the first term of the right hand side we apply the Galerkin orthogonality and estimate

$$\begin{aligned}
(\nabla e_g, \nabla(\sigma e_g)) &= (\nabla e_g, \nabla(\sigma e_g - \mathcal{I}_h(\sigma e_g))) \\
&= (\sigma^{1/2}\nabla e_g, \sigma^{-1/2}\nabla(\sigma e_g - \mathcal{I}_h(\sigma e_g))) \\
&\leq \frac{1}{4}\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 + \|\sigma^{-1/2}\nabla(\sigma e_g - \mathcal{I}_h(\sigma e_g))\|_{L^2(\Omega)}^2 \\
&\leq \frac{1}{4}\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 + ch^2\|\nabla_h^2(\sigma e_g)\|_{L^2(\Omega)}^2 \\
&\leq \frac{1}{4}\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 + ch^2|\ln h|^3
\end{aligned} \tag{A.25}$$

where we have used Lemmas A.4 and A.9 in the last two steps, respectively. For estimating the second term of the right hand side of (A.24) we consider another auxiliary equation,

$$-\Delta y = \frac{e_g}{\|e_g\|_{L^2(\Omega)}} \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega.$$

Utilizing the weak form of this equation with e_g as the test function, and later on Lemma A.4, we can write

$$\begin{aligned}
\|e_g\|_{L^2(\Omega)} &= (\nabla e_g, \nabla y) = (\nabla e_g, \nabla(y - \mathcal{I}_h y)) \\
&\leq \|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}\|\sigma^{-1/2}\nabla(y - \mathcal{I}_h y)\|_{L^2(\Omega)} \\
&\leq \|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)} ch\|\nabla^2 y\|_{L^2(\Omega)} \\
&\leq ch\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}
\end{aligned} \tag{A.26}$$

since the L^2-norm of $e_g/\|e_g\|_{L^2(\Omega)}$ is one. With this result the second term of the right-hand side of (A.24) can be estimated with the help of Lemma A.3 as

$$\begin{aligned}
(\nabla e_g, e_g\nabla\sigma) &= (\sigma^{1/2}\nabla e_g, \sigma^{-1/2}e_g\nabla\sigma) \\
&\leq \|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}\|\sigma^{-1/2}e_g\nabla\sigma\|_{L^2(\Omega)} \\
&\leq \frac{1}{8}\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 + c\|\sigma^{-1/2}e_g\|_{L^2(\Omega)}^2 \\
&\leq \frac{1}{8}\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 + c(e_g, \sigma^{-1}e_g) \\
&\leq \frac{1}{8}\|\sigma^{1/2}\nabla e_g\|_{L^2(\Omega)}^2 + c\|e_g\|_{L^2(\Omega)}\|\sigma^{-1}e_g\|_{L^2(\Omega)}.
\end{aligned}$$

With estimate (A.26) and Lemma A.8 one can conclude

$$\begin{aligned}(\nabla e_g, e_g \nabla\sigma) &\leq \frac{1}{8}\|\sigma^{1/2}\nabla e_g\|^2_{L^2(\Omega)} + ch|\ln h|^{3/2}\|\sigma^{1/2}\nabla e_g\| \\ &\leq \frac{1}{4}\|\sigma^{1/2}\nabla e_g\|^2_{L^2(\Omega)} + ch^2|\ln h|^3 \end{aligned} \tag{A.27}$$

by applying Young's inequality in the last step. With equation (A.24) the assertion follows from inequalities (A.25) and (A.27). □

Now Theorem A.1 can be proven.

Proof. Let T^* denote an element that contains a, and set $\tilde{e} := z - z_h$. By using the nodal interpolant $\mathcal{I}_h$ estimate

$$\begin{aligned}|(z - z_h)(a)| &\leq \max_{T^*}|\tilde{e}| \\ &\leq \max_{T^*}|z - \mathcal{I}_h z| + \max_{T^*}|\mathcal{I}_h\tilde{e}| \\ &\leq \max_{T^*}|z - \mathcal{I}_h z| + c|T^*|^{-1}\int_{T^*}|\mathcal{I}_h\tilde{e}|\,\mathrm{d}x \\ &\leq \max_{T^*}|z - \mathcal{I}_h z| + c|T^*|^{-1}\left(\int_{T^*}|z - \mathcal{I}_h z|\,\mathrm{d}x + \int_{T^*}|\tilde{e}|\,\mathrm{d}x\right) \\ &\leq c\max_{T^*}|z - \mathcal{I}_h z| + c|T^*|^{-1}\int_{T^*}|\tilde{e}|\,\mathrm{d}x \\ &\leq ch_*\|\nabla^2 z\|_{L^2(T^*)} + c|T^*|^{-1}\int_{T^*}|\tilde{e}|\,\mathrm{d}x. \end{aligned} \tag{A.28}$$

Since $h_* \sim h^2$ it remains to estimate $|T^*|^{-1}\int_{T^*}|\tilde{e}|\,\mathrm{d}x$. To this end, we consider the auxiliary problem (A.17). From the weak form of this boundary value problem it is easy to see that

$$(\nabla g^h, \nabla\tilde{e}) = (\delta^h, \tilde{e}) = |T^*|^{-1}\int_{T^*}|\tilde{e}|\,\mathrm{d}x \tag{A.29}$$

is the term left to consider. With the Ritz projection g_h^h defined in (A.18) we can write

$$\begin{aligned}(\nabla g^h, \nabla\tilde{e}) &= (\nabla(z - z_h), \nabla g^h) \\ &= (\nabla(z - z_h), \nabla(g^h - g_h^h)) \\ &= (\nabla(z - \mathcal{I}_h z), \nabla(g^h - g_h^h)) \\ &\leq \|\sigma^{-1/2}\nabla(z - \mathcal{I}_h z)\|_{L^2(\Omega)}\|\sigma^{1/2}\nabla(g^h - g_h^h))\|_{L^2(\Omega)}, \end{aligned} \tag{A.30}$$

using Galerkin orthogonality. The application of Lemmas A.4 and A.10 yields together with equation (A.29) the assertion. □

B. Utilized data for the models of the material properties of concrete

In the physical models established for the hydration of young concrete in Section 7.2 a number of material dependent quantities were utilized. Depending on the context, some of them may be constants within the optimization problem at hand, or a dependence on the control variable may be present. For the solution of a practical problem, the values of these quantities must be determined according to the used material. The purpose of this appendix is to present realistic example data for the material dependent quantities. These were also used in the numerical tests presented in Section 7.5.

In the case that the concrete composition is not subject to the control variable, the values according to Table B.1 were used. If the specific composition of the concrete matters, the used

Table B.1.. Standard material parameters utilized for constant concrete composition

parameter	value	parameter	value	parameter	value
c	$1.0\frac{kJ}{kg\,K}$	ρ	$2000\frac{kg}{m^3}$	λ	$2.143\frac{W}{m\,K}$
Q_∞	$293.2\frac{kJ}{kg}$	a_W	-11	b_W	-1
c_{SL}	1.4	b_J	-1	τ_k	24

bulk values for density, thermal conductivity and heat capacity can be found in Table B.2(a). The values of c_{SL} for different types of cement presented in Table B.2(b) have been taken from [55, Table 5.6]. When the quantities $Q_\infty, b_J, \tau_k, a_W, b_W$ are needed in dependence of the

Table B.2.. Used data for material properties

(a) bulk properties for the ingredients

ingredient	index i	$\rho_{g,i}/\frac{kg}{m^3}$	$\lambda_i/\frac{W}{m\,K}$	$c_i/\frac{kJ}{kg\,K}$
cement	1	3000	1.3	0.80
fly ash	2	2300	1.3	0.75
water	3	1000	0.6	4.18
aggregate	4	2600	3.0	0.80

(b) c_{SL} for different cements

type of cement	c_{SL}
ENCI CEM I *(diff. types)*	1.25
ENCI CEM II/B-V 32.5 R	1.25
ENCI CEM III/B 42.5 LHHS	1.65
ENCI CEM III/B 42.5 LHHS +	1.60
ENCI CEM III/A 52.5	1.40
ENCI CEM V/A 42.5	1.40

concrete recipe, see Section 7.2, the models (7.18) through (7.22) are used. In these models the constants $m_{\cdot,i}$ are unknown until now. To determine them the following data is used: In [55, Appendix C,D] some concrete recipes named ICO-03 through ICO-07 containing cement type

CEM III are listed together with their material properties, see Table B.3. By insertion of the

Table B.3.. Data for modelling the dependence of material properties on the concrete recipe

(a) recipes considered in [55]

name	ρ_1	ρ_2	ρ_3	ρ_4
ICO-03	270	90	156	1836
ICO-04	390	0	183	1786
ICO-05	280	100	152	1826
ICO-06	270	80	143	1876
ICO-07	240	110	150	1872

(b) measured data according to [55]

name	Q_∞	b_J	τ_k	a_W	b_W
ICO-03	379	-1.37	20.14	-10.2	-0.65
ICO-04	361	-0.83	26.35	-4.6	-0.39
ICO-05	385	-0.86	21.91	-16.7	-0.84
ICO-06	379	-1.01	25.44	-6.55	-0.49
ICO-07	401	-0.92	28.03	-6.40	-0.47

recipes and measured data into the models (7.18) through (7.22) the parameters $m_{\cdot,i}$ can be determined. In the following this is demonstrated for Q_∞. Collecting the measured data in a vector and the recipe data in a matrix by

$$\mathbf{Q}_\infty = \begin{pmatrix} 379 \\ 361 \\ 385 \\ 379 \\ 401 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1 & 270 & 90 & 156 & 1836 \\ 1 & 390 & 0 & 183 & 1786 \\ 1 & 280 & 100 & 152 & 1826 \\ 1 & 270 & 80 & 143 & 1876 \\ 1 & 240 & 110 & 150 & 1872 \end{pmatrix},$$

and the missing parameters in the vector $\mathbf{m}_{\mathbf{Q}_\infty} = (m_{Q_\infty,0}\, m_{Q_\infty,1}\, m_{Q_\infty,2}\, m_{Q_\infty,3}\, m_{Q_\infty,4})^\top$, then model (7.18) is expressed by the linear system $\mathbf{R}\mathbf{m}_{\mathbf{Q}_\infty} = \mathbf{Q}_\infty$, which has the unique solution

$$\mathbf{m}_{\mathbf{Q}_\infty} = (-1.7173e+03 \quad 7.9182e-01 \quad 1.2679e+00 \quad 1.5762e+00 \quad 8.2926e-01)^\top.$$

The quadratic structure of this linear system of equations is occurring by chance here in the sense that there were just as many recipes considered in [55] as the number of considered ingredients plus one. In general with more measurements one would expect a higher accuracy of the model, and the then rectangular, overdetermined system of equations would be solved in the sense of least squares by solving the normal equations. The necessary measurements are rare, especially since only data corresponding to the same type of cement should be used according to experience.
For the other material parameters the procedure described above yields the parameters $m_{\cdot,i}$ displayed in Table B.4.

Table B.4.. Model parameters utilized in (7.18) through (7.22)

parameter	$m_{\cdot,0}$	$m_{\cdot,1}$	$m_{\cdot,2}$	$m_{\cdot,3}$	$m_{\cdot,4}$
Q_∞	-1.7173e+03	7.9182e-01	1.2679e+00	1.5762e+00	8.2926e-01
b_J	-7.4728e+01	4.4821e-02	4.5562e-02	2.6197e-02	2.8905e-02
τ_k	-1.0075e+03	4.1591e-01	4.2298e-01	5.8434e-01	4.2816e-01
a_W	-3.7871e+02	-1.0769e-01	-1.7945e-01	4.3651e-01	1.8826e-01
b_W	-2.1095e+01	2.9982e-04	-2.1576e-03	1.8871e-02	9.5937e-03

List of Figures

List of Tables

List of Algorithms

Bibliography

[1] T. Adam. *Ein Modell zur Beschreibung der Hydratation von Beton in Abhängigkeit vom Feuchtegehalt.* Ph.D. thesis, Technische Universität Darmstadt, 2006.

[2] H. Amann. *Linear and quasilinear parabolic problems.* Birkhäuser Verlag, 1995.

[3] T. Apel, O. Benedix, D. Sirch, and B. Vexler. A priori mesh grading for an elliptic problem with Dirac right-hand side. *SINUM* 49(3), pp. 992–1005, 2011.

[4] T. Apel, A. Rösch, and D. Sirch. L^∞-error estimates on graded meshes with application to optimal control. *SIAM J. Control Optim.* 48, pp. 1771–1796, 2009.

[5] T. Apel, A. Rösch, and G. Winkler. Optimal control in nonconvex domains: a priori discretization error estimates. *Calcolo* 44, pp. 137–158, 2007.

[6] T. Apel, D. Sirch, and G. Winkler. Error estimates for control constrained optimal control problems: Discretization with anisotropic finite element meshes 2008. Submitted to Math. Program.

[7] R. Becker, M. Braack, D. Meidner, R. Rannacher, and B. Vexler. Adaptive Finite Element Methods for PDE-Constrained Optimal Control Problems. In *Reactive Flows, Diffusion and Transport*, edited by R. Rannacher. Springer Verlag, Berlin, 2006.

[8] R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concepts. *SIAM J. Control Optim.* 39(1), pp. 113–132, 2000.

[9] R. Becker, D. Meidner, and B. Vexler. Efficient numerical solution of parabolic optimization problems by finite element methods. *Optimization Methods and Software* 22(5), pp. 813–833, 2007.

[10] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation. In *Acta Numerica 2001*, edited by A. Iserles, pp. 1–102. Cambridge University Press, 2001.

[11] R. Becker and B. Vexler. A posteriori error estimation for finite element discretizations of parameter identification problems. *Numer. Math.* 96(3), pp. 435–459, 2004.

[12] R. Becker and B. Vexler. Mesh refinement and numerical sensitivity analysis for parameter calibration of partial differential equations. *J. Comp. Physics* 206(1), pp. 95–110, 2005.

[13] O. Benedix and B. Vexler. A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints. *Comput. Optim. Appl.* 44(1), pp. 3–25, 2009.

[14] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch. A comparison of interior point methods and a Moreau-Yosida based active set strategy for constrained optimal control problems. *SIAM J. Optim.* 11(2), pp. 495–521, 2000.

[15] M. Bergounioux and K. Kunisch. Primal-dual active set strategy for state constrained optimal control problems. *Computational Optimization and Applications* 22, pp. 193–224, 2002.

[16] D. Braess. *Finite Elemente.* Springer Verlag, Berlin, Heidelberg, New York, 1992.

[17] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.* 24, pp. 1309–1318, 1986.

[18] E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM J. Control Optim.* 34, pp. 933–1006, 1993.

[19] E. Casas. Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints. *ESAIM: COCV* 8, pp. 345–374, 2002.

[20] E. Casas and F. Tröltzsch. Recent advances in the analysis of pointwise state-constrained elliptic optimal control problems. *ESAIM: COCV* 16, pp. 581–600, 2010.

[21] S. Cherednichenko, K. Krumbiegel, and A. Rösch. Error estimates for the Lavrentiev regularization of elliptic optimal control problems. *Inverse Problems* 24(5), p. 055003, 2008.

[22] D. Clever and J. Lang. Optimal control of radiative heat transfer in glass cooling with restrictions on the temperature gradient. *Optimal Control Applications and Methods* 2011. Accepted.

[23] B. Dacorogna. *Direct Methods in the Calculus of Variations.* Springer, Berlin, 1989.

[24] J. C. de los Reyes, C. Meyer, and B. Vexler. Finite element error analysis for state-constrained optimal control of the Stokes equations. *Control and Cybernetics* 37(2), pp. 251–284, 2008.

[25] K. Deckelnick, A. Günther, and M. Hinze. Finite element approximation of elliptic control problems with constraints on the gradient. *Numer. Math.* 111, pp. 335–350, 2009.

[26] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state constrained elliptic control problem. *SIAM J. Numer. Anal.* 35, pp. 1937–1953, 2007.

[27] K. Deckelnick and M. Hinze. Variational discretization of parabolic control problems in the presence of pointwise state constraints. *Journal of Computational Mathematics* 29, pp. 1–16, 2011.

[28] K. S. Dewald. *Ein Ansatz zur stoffgerechten Bemessung von offenen Becken.* Ph.D. thesis, Universität Duisburg-Essen, 2006.

[29] B. Eierle. *Berechnungsmodelle für rißgefährdete Betonbauteile unter frühem Temperaturzwang.* Ph.D. thesis, TU München, 2000.

[30] J. Eric Jonasson. *Modelling of Temperature, Moisture and Stresses in Young Concrete.* Ph.D. thesis, Luleå University of Technology, 1994.

[31] L. C. Evans. *Partial differential equations.* American Mathematical Society, Providence, R.I., 2002.

[32] R. S. Falk. Approximation of a class of optimal control problems with order of convergence estimates. *J. Math. Anal. Appl.* 44, pp. 28–47, 1973.

[33] J. Frehse and R. Rannacher. Eine L^1-Fehlerabschätzung für diskrete Grundlösungen in der Methode der finiten Elemente. In *Finite Elemente. Tagungsband des Sonderforschungsbereichs 72*, edited by J. Frehse, R. Leis, and R. Schaback, volume 89 of *Bonner Mathematische Schriften*, pp. 92–114. Bonn, 1976.

[34] P. Freiesleben Hansen and E. J. Pedersen. Maleinstrument til kontrol af betons haerding. *Nordisk Beton* 1, pp. 21–55, 1977.

[35] A. V. Fursikov. *Optimal control of distributed systems.* American Mathematical Society, Providence, 2000.

[36] Gascoigne. The finite element toolkit Gascoigne. `http://www.gascoigne.uni-hd.de`.

[37] M. Gerdts, G. Greif, and H. Pesch. Numerical optimal control of the wave equation: optimal boundary control of a string to rest in finite time. *Math. Comput. Simulation* 79(4), pp. 1020–1032, 2008.

[38] T. Geveci. On the approximation of the solution of an optimal control problem governed by an elliptic equation. *Math. Model. Numer. Anal.* 13, pp. 313–328, 1979.

[39] P. Grisvard. *Singularities in Boundary Value Problems.* Springer-Verlag, Masson, Paris, Berlin, 1992.

[40] M. Gugat and V. Grimm. Optimal boundary control of the wave equation with pointwise control constraints. *Comput. Optim. Appl* 2009. Published online.

[41] A. Günther and M. Hinze. A posteriori error control of a state constrained elliptic control problem. *J. Numer. Math.* 16, pp. 307–322, 2008.

[42] A. Günther and A. Schiela. Interior point methods in function space for state constraints - inexact Newton and adaptivity. *DFG-SPP1253 preprint: Nr. SPP1253-08-06* 2009.

[43] A.-W. Gutsch. *Stoffeigenschaften jungen Betons - Versuche und Modelle.* Ph.D. thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig, 1998.

[44] M. Hintermüller and R. H. Hoppe. Goal-oriented adaptivity in control constrained optimal control of partial differential equations. *SIAM J. Control Optim.* 47(4), pp. 1721–1743, 2008.

[45] M. Hintermüller and R. H. Hoppe. Goal-oriented adaptivity in pointwise state constrained optimal control of partial differential equations. *SIAM J. Control Optim.* 48(8), pp. 5468–5487, 2010.

[46] M. Hintermüller, R. H. Hoppe, Y. Iliash, and M. Kieweg. An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. *ESAIM Control Optim. Calc. Var.* 14, pp. 540–560, 2008.

[47] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization* 13(3), pp. 865–888, 2003.

[48] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.* 17(1), pp. 159–18, 2006.

[49] M. Hintermüller and K. Kunisch. Stationary optimal control problems with pointwise state constraints. In *Numerical PDE Constrained Optimization*, edited by T. Barth, M. Griebel, D. Keyes, R. Nieminen, D. Roose, and T. Schlick, volume 72 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, 2009.

[50] M. Hintermüller and W. Ring. A level set approach for the solution of a state constrained optimal control problem. *Numer. Math.* 98(1), pp. 135–166, 2004.

[51] M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Comput. Optim. Appl.* 30(1), pp. 45–61, 2005.

[52] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer Science + Business Media B.V., 2009.

[53] J. Huckfeldt. *Thermomechanik hydratisierenden Betons – Theorie, Numerik und Anwendung*. Ph.D. thesis, TU Carolo – Wilhelmina, Braunschweig, 1993.

[54] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems and Control Letters* 50, pp. 221–228, 2003.

[55] M. Krauss. *Probabilistischer Nachweis der Wirksamkeit von Maßnahmen gegen frühe Trennrisse in massigen Betonbauteilen*. Ph.D. thesis, Technische Universität Carolo-Wilhelmina zu Braunschweig, 2004.

[56] A. Kröner, K. Kunisch, and B. Vexler. Semismooth Newton methods for optimal control of the wave equation with control constraints. *SIAM Journal on Control and Optimization* 49(2), pp. 830 – 858, 2011.

[57] K. Krumbiegel, C. Meyer, and A. Rösch. A priori error analysis for state constrained boundary control problems. Part I: Control discretization. Technical report, 2009. Weierstrass Institute for Applied Analysis and Stochastics, WIAS Preprint 1393.

[58] K. Krumbiegel, C. Meyer, and A. Rösch. A priori error analysis for state constrained boundary control problems. Part II: Full discretization. Technical report, 2009. Weierstrass Institute for Applied Analysis and Stochastics, WIAS Preprint 1394.

[59] K. Krumbiegel and A. Rösch. A virtual control concept for state constrained optimal control problems. *Comput. Optim. Appl* 43(2), pp. 213–233, 2009.

[60] K. Kunisch and B. Vexler. Constrained Dirichlet boundary control in L^2 for a class of evolution equations. *SIAM J. Control Optim.* 40(5), pp. 1726–1753, 2007.

[61] M. Laube. *Werkstoffmodell zur Berechnung von Temperaturspannungen in massigen Betonbauteilen im jungen Alter.* Ph.D. thesis, TU Carolo - Wilhelmina, Braunschweig, 1990.

[62] R. Li, W. Liu, H. Ma, and T. Tang. Adaptive finite element approximation for distributed elliptic optimal control problems. *SIAM J. Control Optim.* 41(5), pp. 1321–1349, 2002.

[63] J. L. Lions. *Optimal control of systems governed by partial differential equations.* Springer Verlag, Berlin, 1971.

[64] K. Malanowski. Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems. *Appl. Math. Optim.* 8, pp. 69–95, 1981.

[65] D. Meidner. *Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems.* Ph.D. thesis, Ruprecht-Karls-Universität Heidelberg, 2008.

[66] D. Meidner and B. Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.* 46(1), pp. 116–142, 2007.

[67] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. Part I: Problems without control constraints. *SIAM J. Control Optim.* 47(3), pp. 1150–1177, 2008.

[68] P. Merino, F. Tröltzsch, and B. Vexler. Error estimates for the finite element approximation of a semilinear elliptic control problem with state constraints and finite dimensional control space. *Mathematical Modelling and Numerical Analysis* 44(1), pp. 167–188, 2010.

[69] C. Meyer. Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints. *Control and Cybernetics* 37, pp. 51–85, 2008.

[70] C. Meyer and M. Hinze. Stability of infinite dimensional control problems with pointwise state constraints. *WIAS, Preprint 1236* 2007.

[71] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems. *SIAM J. Control Optim.* 43, pp. 970–985, 2004.

[72] B. S. Mordukhovich and J.-P. Raymond. Dirichlet boundary control of hyperbolic equations in the presence of state constraints. *Appl. Math. Optim.* 49, pp. 145–157, 2004.

[73] B. S. Mordukhovich and J.-P. Raymond. Neumann boundary control of hyperbolic equations with pointwise state constraints. *SIAM Journal on Control and Optimization* 43(4), pp. 1354–1372, 2005.

[74] I. Neitzel and F. Tröltzsch. On convergence of regularization methods for nonlinear parabolic optimal control problems with control and state constraints. *Control and Cybernetics* 37(4), 2008. To appear.

[75] I. Neitzel and F. Tröltzsch. On regularization methods for the numerical solution of parabolic control problems with pointwise state constraints. *ESAIM COCV* 15(2), pp. 426–453, 2009.

[76] I. Neitzel and B. Vexler. A priori error estimates for space-time finite element discretization of semilinear parabolic optimal control problems. *DFG-SPP1253 preprint: Nr. SPP1253-107* 2010.

[77] L. Nietner and D. Schmidt. Temperatur- und Festigkeitsmodellierungen durch Praxiswerkzeuge – Grundlage dauerhafter Betonteile. *Beton- und Stahlbetonbau* (12/03), pp. 738–746, 2003.

[78] G. Of, T. X. Phan, and O. Steinbach. Boundary element methods for Dirichlet boundary control problems. *Math. Methods Appl. Sci.* 33, pp. 2187–2205, 2010.

[79] P. Onken and F. S. Rostásy. *Wirksame Betonzugfestigkeit im Bauwerk bei früh einsetzendem Temperaturzwang*, volume 449 of *Deutscher Ausschuss für Stahlbeton (DafStb) im DIN, Deutsches Institut für Normung e.V.* Beuth Verlag, Berlin.

[80] RoDoBo. RoDoBo: A C++ library for optimization with stationary and nonstationary PDEs with interface to Gascoigne [36]. `http://www.rodobo.uni-hd.de`.

[81] F. S. Rostásy. Risskontrolle bei massigen Betonbauteilen - Stand der Technik, neue Wege und offene Fragen. In *Risskontrolle massiger Betonbauteile. Bauwerk, Werkstoff, Simulation*, number 153 in Schriftenreihe des iBMB. Braunschweig, 2001.

[82] F. S. Rostásy and M. Krauss. *Frühe Risse in massigen Betonbauteilen – Ingenieurmodelle für die Planung von Gegenmaßnahmen*, volume 520 of *Deutscher Ausschuss für Stahlbeton (DafStb) im DIN, Deutsches Institut für Normung e.V.* Beuth Verlag, Berlin, 2001.

[83] Y. Saad. *Iterative methods for sparse linear systems.* PWS Publ. Co., Boston, 1996.

[84] A. G. A. Saul. Principles underlying the steam curing of concrete at atmospheric pressure. *Magazine of Concrete Research* 2, pp. 127–140, 1951.

[85] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. Optim.* 20(2), pp. 1002–1031, 2009.

[86] A. Schiela and W. Wollner. Barrier methods for optimal control problems with convex nonlinear gradient constraint. *SIAM J. Optim.* 2009. Accepted.

[87] M. Schmich and B. Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM J. Sci. Comput.* 30(1), pp. 369–393, 2008.

[88] R. Scott. Finite element convergence for singular data. *Numer. Math.* 21, pp. 317–327, 1973.

[89] D. Sirch. *Finite Element Error Analysis for PDE-constrained Optimal Control Problems: The Control Constrained Case Under Reduced Regularity.* Ph.D. thesis, Technische Universität München, 2010.

[90] R. Springenschmid, editor. *Thermal cracking in concrete at early ages*. Spon, London u.a., 1995. Proceedings of the international symposium held by RILEM at the Technical Univ. of Munich, Oct. 10 - 12, 1994.

[91] H. Triebel. *Interpolation theory, function spaces, differential operators*. Barth, Heidelberg, 1995.

[92] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen*. Friedr. Vieweg & Sohn Verlag, Wiesbaden, 2005.

[93] F. Tröltzsch. Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM J. Optim.* 15(2), pp. 616–634, 2005.

[94] B. Vexler and W. Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM J. Control Optim.* 47(1), pp. 509–534, 2008.

[95] VisuSimple. VisuSimple: An interactive VTK-based visualization and graphics/mpeg-generation program. `http://www.visusimple.uni-hd.de`.

[96] M. Weiser. Optimization and identification in regional hyperthermia. *Int. J. Appl. Electromagn. and Mech.* 30, pp. 265–275, 2009.

[97] M. Weiser, T. Gänzler, and A. Schiela. A control reduced primal interior point method for pde constrained optimization. *Comp. Opt. Appl.* 41(1), pp. 127–145, 2008.

[98] K. Wesche. Baustoffkennwerte zur Berechnung von Temperaturfeldern in Betonbauteilen. In *Liber Amicorum opgedragen aan F.G. Riessauw ter gelegenheid van zijn zeventigste verjaardag 17 april 1982*. Gent, 1982.

[99] J. Wloka. *Partial Differential Equations*. Cambridge University Press, Cambridge, 1987.

[100] W. Wollner. A posteriori error estimates for a finite element discretization of interior point methods for an elliptic optimization problem with state constraints. *Comput. Optim. Appl* 47(1), pp. 133–159, 2010.

[101] J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optimization* 5, pp. 49–62, 1979.

[102] E. Zuazua. Propagation, observation and control of waves approximated by finite difference methods. *SIAM Rev.* 47(2), pp. 197–243, 2005.