Technische Universität München

Fakultät für Mathematik

Lehrstuhl für Optimalsteuerung

Finite element discretization and efficient numerical solution of elliptic and parabolic sparse control problems

Konstantin Pieper

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. Michael Ulbrich
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Boris Vexler
	2.	Univ.-Prof. Dr. Karl Kunisch Karl-Franzens-Universität, Graz, Österreich
	3.	Prof. Dr. Eduardo Casas Renteria Universidad de Cantabria, Santander, Spanien

Die Dissertation wurde am 26.02.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 31.03.2015 angenommen.

Abstract

This thesis is concerned with the numerical analysis of sparse control problems for elliptic and parabolic state equations. A focus is set on controls which are measures in space, where the instationary problem formulation favors fixed-in-space point sources. A general optimization framework based on a sparsity-preserving regularization and a semismooth Newton method is developed. A priori error estimates for a suitable finite element discretization of two model problems are derived. An algorithm for adaptive mesh refinement is proposed.

Zusammenfassung

Diese Arbeit befasst sich mit der numerischen Analyse von Optimalsteuerungsproblemen mit „Sparsity“ für ellitische und parabolische Zustandsgleichungen. Im Fokus liegen Kontrollen, die Maße im Ort sind, wobei die instationäre Formulierung feststehende Punktquellen favorisiert. Ein Optimierungsansatz wird entwickelt, der auf geeigneter Regularisierung und einer semiglatten Newton-Methode basiert. A priori Fehlerabschätzungen für die Finite Elemente Diskretisierung von zwei Modellproblemen werden hergeleitet. Ein Algorithmus zur adaptiven Gitterverfeinerung wird vorgestellt.

Acknowledgments

First of all, I would like to thank my supervisor Boris Vexler for suggesting this interesting dissertation topic and for giving the essential directions towards its development. Most of the results of this thesis would not have been possible without his guidance. I am also grateful for his general support: his trust in my abilities, the encouragement to attend international conferences, the opportunity (not the obligation) to do teaching, and the freedom to pursue additional research interests. Furthermore, I want to thank my second supervisor Karl Kunisch for providing new impulses for the development of this thesis during my stays in Graz; for sharing his vast expertise and especially for his readiness for spontaneous, detailed discussions. Additionally, I am thankful for the successful collaboration in another project. I also want to thank my mentor Dominik Meidner for always being ready for questions and discussions.

I gratefully acknowledge the funding received towards my Ph.D. from the German Research Foundation (DFG) through the *International Research Training Group (IGDK) 1754 Munich – Graz "Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures"*, which is co-funded by the Austrian Science Fund (FWF). I am especially grateful to the IGDK 1754 and its members for the opportunities concerning collaboration, the frequent exchange of ideas, and the possibilities for further studies, such as compact courses and the ability to invite external lecturers.

Furthermore, the following persons have provided valuable input for the development of this thesis. In the first place, I want to thank Philip Trautmann for countless discussions on convex and infinite dimensional analysis and Armin Rund for many debates on the practical aspects of optimization methods. I am grateful to Michael Ulbrich for helpful input on semismooth Newton methods (especially for pointing out the concept of the normal map during a very productive flight to Belgrade) and to Andre Milzarek for introducing me to the mathematical beauty of the proximal map. I also want to thank Anton Schiela for his excellent compact course. Finally, I am grateful to Andreas Springer for all the time we spent discussing various mathematical and technical subjects related to this work.

Special thanks go to my friends and colleagues from the IGDK and the institutes in Munich and Graz for making my time at work more enjoyable. Especially, I want to mention Bao, Bernhard, Behzad, David, Felix, Jelena, Linus, Lukas, Max, Moritz, Marco, Olena and Philip (in particular, for the very pleasant instances of the "Students' Workshop"), and Alana, Andreas, Anne-Céline, Axel, Boris, Daniel, Dominik, Ira, Lucas, Lukas, Olaf, and Tom. Furthermore, I want to thank Armin, Christian and Philip for making me feel at home during my stays in Graz.

Finally, I am grateful to my family for their never-ending support, encouragement, and care. You have always believed in me.

Contents

1. Introduction

In this work we consider finite element discretizations and efficient numerical solution methods for sparse optimal control problems subject to elliptic and parabolic partial differential equations (PDE). *Sparse* optimal control problems are problems with (spatially and/or temporally) distributed controls in combination with control cost (or regularization) terms that favor solutions which are supported only on a "small" set; see below. More precisely, we mean problems of the type

$$\begin{aligned} \min_{u \in U_{\text{ad}},\, y \in Y} \quad & J(y) + \psi(u), \\ \text{subject to} \quad & e(y,u) = 0. \end{aligned} \tag{$\mathcal{P}$}$$

Here, J is a smooth tracking-type functional for the state variable y, defined on the state space Y, and e is an (elliptic or parabolic) state equation, coupling the state to the control variable u. The control is searched for in the convex control set U_{ad} and ψ is a sparsity inducing functional. In general, the cost or regularization term ψ will be a convex, but *not* a strictly convex functional. The canonical example is the L^1 norm of a function or the total variation norm of a measure. Problems of this type arise in different contexts:

- For a cost functional
$$\psi(u) = \int_\Omega u(x)\,\mathrm{d}x$$
with $u \geq 0$, we have a linear dependence of the cost on the control u. In some applications, this is a more appropriate cost functional than an L^2-type norm; see, e.g., [VM06; BCS13].
- The L^1 norm, defined on a bounded domain Ω as
$$\psi(u) = \int_\Omega |u(x)|\,\mathrm{d}x,$$
and its appropriate generalization to measures, the total variation norm, given by
$$\psi(u) = \int_\Omega \mathrm{d}|u|(x),$$
are known to induce *sparsity*. This means that the optimal solutions of ($\mathcal{P}$) will be supported only on a possibly very small set (and will be zero everywhere else). In particular, in the case where the control is searched for in a space of measures, we can obtain a sum of point sources as the optimal solution. This makes such functionals useful as regularization terms in the context of inverse problems (see, e.g., [SW09; BP13; CFG13; CFG14; ACG15]) and actuator placement problems (see, e.g., [Sta09; CK11b; Bru+12]). Recently, optimal control with sparsity has also been proposed for PDEs arising as mean-field limits of systems of ODEs; see [FS14].

In practice, to compute solutions of ($\mathcal{P}$), we have to replace the partial differential equation e by an appropriate discrete approximation and replace the solution spaces U_{ad} and Y by finite dimensional spaces. Since we consider PDEs of elliptic or parabolic type, finite elements are

a canonical method of choice (especially, since we will deal with highly irregular data). We obtain the discrete problem

$$\begin{aligned} \min_{u\in U_{\mathrm{ad}}^{\sigma},\, y\in Y^{\sigma}} \quad & J_\sigma(y)+\psi_\sigma(u), \\ \text{subject to} \quad & e_\sigma(y,u)=0, \end{aligned} \tag{$\mathcal{P}_\sigma$}$$

with an additional discretization parameter $\sigma > 0$. A strong focus of this work will be the derivation and analysis of discretization concepts for two concrete problem settings (elliptic and parabolic; see below). On the one hand, we will be concerned with a priori estimates for $\sigma \to 0$, and, on the other hand, we derive an adaptive finite element method based on *goal-oriented* a posteriori estimates (for the elliptic problem).

The missing strict convexity of ψ poses difficulties not only in the theoretical analysis, but also in the numerical and algorithmic treatment. For the computation of solutions to $(\mathcal{P})$, we will consider an auxiliary problem with an additional Hilbert-space regularization term. It is given by

$$\begin{aligned} \min_{u\in H\cap U_{\mathrm{ad}},\, y\in Y} \quad & J(y)+\psi(u)+\frac{\gamma}{2}\|u\|_H^2, \\ \text{subject to} \quad & e(y,u)=0, \end{aligned} \tag{$\mathcal{P}_\gamma$}$$

where H is a Hilbert space (such as, e.g., $L^2(\Omega)$). Typically, the optimal solutions of $(\mathcal{P})$ are not contained in H, such that the solutions of $(\mathcal{P}_\gamma)$ have higher regularity. Additionally, the strong convexity of the regularized problem enables us to give a very general optimization framework, which is based on a semismooth Newton method. For fixed $\gamma > 0$, it can be formulated and analyzed in a function space setting. Therefore, we can expect that appropriate concrete realizations of the corresponding algorithms will show *mesh-independence* in practice (the number of steps of the algorithm will be essentially independent of the number of degrees of freedom of the discretization). To compute a solution of the original problem $(\mathcal{P})$, we apply a continuation method in the parameter γ. We derive a priori estimates of the regularization error for two model problems. Based on that, we also develop an a posteriori estimation strategy, which is employed in the adaptive algorithm. Let us point out that the additional regularization preserves some of the structural properties of the optimal solutions, since $(\mathcal{P}_\gamma)$ still contains the unmodified (generally nonsmooth) term ψ. In particular, the solutions of the regularized problem will inherit the sparsity property.

For the theoretical analysis of the discretization and regularization error we will mainly focus on two characteristic model problems with measure valued controls. The first one is a tracking-type problem for the Poisson equation given by

$$\begin{aligned} \min_{u\in\mathcal{M}(\Omega_c),\, y\in Y} \quad & \frac{1}{2}\|y-y_d\|_{L^2(\Omega_o)}^2+\alpha\int_{\Omega_c}\mathrm{d}|u|, \\ \text{subject to} \quad & \begin{cases} -\Delta y=\chi_{\Omega_c}u & \text{in } \Omega, \\ \qquad u=0 & \text{on } \partial\Omega. \end{cases} \end{aligned}$$

The control is searched for in the space of finite Radon measures $\mathcal{M}(\Omega_c)$ on the (relatively closed) control set $\Omega_c \subset \Omega$. Note that this space contains controls of the specific form $u=\sum_{n=1}^{N} u_n\delta_{x_n}$, where $x_n \in \Omega_c$ and $u_n \in \mathbb{R}$ for $n = 1,\dots,N$, which is a common model for pointwise control (see, e.g., [Chr81; BMR91; Lio92]). The control cost term is given by the total variation norm $\int_{\Omega_c}\mathrm{d}|u| = \|u\|_{\mathcal{M}(\Omega_c)}$ and the tracking functional is formulated on the observation domain Ω_o. The second model problem (with a parabolic state equation) is, in a sense, the canonical

generalization of the elliptic problem. It reads

$$\min_{u\in\mathcal{M}(\Omega_c,L^2(I)),\, y\in Y} \quad \frac{1}{2}\|y-y_d\|^2_{L^2(I,L^2(\Omega_o))} + \alpha\int_{\Omega_c} \mathrm{d}|u|,$$
$$\text{subject to} \quad \begin{cases} \partial_t y - \Delta y = \chi_{\Omega_c} u & \text{in } I\times\Omega, \\ y = 0 & \text{on } \partial\Omega, \\ y(0) = y_0 & \text{in } \Omega. \end{cases}$$

Here, the state equation is the linear heat equation with zero Dirichlet boundary conditions. Again, the control is searched for in a space of measures supported on the control set Ω_c. However, in the parabolic case, we consider the space of vector valued measures $\mathcal{M}(\Omega_c, L^2(I))$. The prototypical example for elements of this space are point sources with time-dependent coefficients of the form $u(t) = \sum_{n=1}^N u_n(t)\delta_{x_n}$ for $t \in I$, where $x_n \in \Omega_c$ and $u_n \in L^2(I)$ for $n = 1, \dots, N$.

Let us give some further motivation for the choice of vector measures and the particular form of the cost term in the parabolic problem. It is motivated on the one hand by the concept of *directional sparsity* proposed by Herzog, Stadler, and Wachsmuth [HSW12], where and an additional quadratic $L^2(I \times \Omega_c)$ term is contained in the objective. In fact, for the regularized problem $(\mathcal{P}_\gamma)$ we will recover exactly their problem formulation. It favors controls that are zero on "stripes" of the parabolic cylinder; see section 5.4. On the other hand, the cost term is motivated by the concept of *joint sparsity* in finite dimensions; see, e.g., Fornasier and Rauhut [FR08] and the references given therein. In their setting, a regularization with an $\ell^1(\ell^2)$ norm is considered. On the discrete level, for the problem $(\mathcal{P}_\sigma)$, we recover exactly this cost term; see section 5.2. In parallel to the development of this thesis, sparse controls in the space of vector measures have also been proposed in other contexts. In Kunisch, Trautmann, and Vexler [KTV14] an optimal control problem for the wave equation is considered. The problem formulation is motivated by an application to a seismic inverse source location problem. Furthermore, Henneke [Hen15] proposes the use of sparsity with vector measures in combination with appropriate control operators for the optimal control of bilinear quantum systems. In this case, sparsity is applied to a time-frequency representation of the control.

This thesis is structured as follows. In chapter 2 we provide a theoretical framework that allows for the mathematical discussion of the elliptic and parabolic problem formulation mentioned above. In each case, we will discuss a slightly more general problem setting, which encompasses more general elliptic operators and boundary conditions than the Laplace operator with homogeneous Dirichlet boundary conditions above. It also allows for positivity constraints on the control and contains the case of boundary control and observation. The elliptic and parabolic solution theory is based on known results and transposition arguments. Furthermore, we derive optimality conditions and discuss the sparsity properties of the optimal solutions. We introduce and discuss the regularized problem $(\mathcal{P}_\gamma)$ and provide some preparatory results for the analysis of the regularization error.

Chapter 3 is concerned with optimization methods for the regularized problem $(\mathcal{P}_\gamma)$. Here, we consider a general problem setting (based on a reduced cost functional) that also allows for the discussion of some nonlinear control problems with convex cost terms. We develop the optimization framework based on a reformulation of the optimality conditions with the *normal map*. We show superlinear convergence of a semismooth Newton method, which is based mainly on known results, and prove global convergence of a related first order optimization method. Since a reformulation with the normal map has not been extensively studied in the infinite

dimensional semismooth Newton literature, we include a comparison to other (more commonly employed) reformulations and highlight what we consider to be the advantages of the proposed approach.

The finite element discretization of the elliptic model problem introduced above is discussed in chapter 4. We provide an asymptotic a priori error analysis that improves earlier results obtained for the same problem. Under additional assumptions, we prove a higher regularity result for the optimal solutions, which excludes point sources as optimal solutions. This allows for an (even further) improved error estimate in three spatial dimensions. Furthermore, we derive an a priori error estimate for the error in the cost functional due to regularization. A discretization concept for the regularized problem is also presented and a corresponding error estimate is derived. The discretization of the regularized problem is designed to reproduce the original discrete problem in the limiting case for $\gamma \to 0$. Many of the results of this chapter have already appeared in similar form in [PV13].

Similarly, the parabolic model problem and its appropriate finite element discretization are discussed in chapter 5. We provide a corresponding a priori error analysis that seems to be optimal (at least in some aspects). An estimate for the regularization error of the objective is also provided. Additionally, we apply the problem formulation to an inverse source location problem to give a practical motivation for the proposed approach. Most of the results of this chapter have already appeared in similar form in [KPV14].

Chapter 6 is devoted to an adaptive algorithm with local mesh refinement for the solution of the elliptic problem. Here, the problem is first regularized and then discretized. A heuristic a posteriori estimation strategy of the regularization error based on an asymptotic model is introduced and error estimates for the discretization error of the objective functional based on the *dual-weighted-residual* approach are derived. The adaptive strategy is based on a balancing of both error contributions. The effectivity of the error estimates is evaluated in numerical experiments and practical results are presented. We also compare the discretization concept of this chapter with the one presented in chapter 4.

2. Theoretical framework

In this section we will discuss a theoretical framework to consider a reasonably large class of sparse control problems with measure valued controls. We will take care to introduce a general setting that allows us to treat the elliptic case and the parabolic case in a unified way and that is extensible towards more general problem settings. In the elliptic case, we essentially follow the ideas in Kunisch and Clason [CK11a; CK11b] and Bredies and Pikkarainen [BP13]. In the parabolic setting, we provide a problem formulation that favors controls which consists of a fixed measure in space with time-dependent "coefficients". Most of the corresponding theory has already appeared in Kunisch, Pieper, and Vexler [KPV14]. We also refer to Casas, Clason, and Kunisch [CCK13] for a related parabolic control problem.

This chapter is organized in the following way. In section 2.1 we provide an abstract problem setting and framework for the discussion of existence of solutions and optimality conditions. Section 2.2 is devoted to the elliptic problem setting and section 2.3 to the parabolic setting. In each case, we introduce the corresponding measure space and discuss well-posedness and regularity of the state equation based on known regularity results and the method of transposition. Furthermore, we derive optimality conditions. In the parabolic case, we compare the proposed problem formulation with vector measures to the formulation from [CCK13], where the optimal solutions have different sparsity properties. In section 2.4 we briefly explain the relation to a convex duality approach for the analysis of the given problem formulation (which is used in [CK11a; CK11b]) and draw a parallel to state constrained optimization. The Hilbert space regularization of the problem is introduced in section 2.5. Finally, we analyze the regularized problem and provide estimates for the error that arises due to regularization. We show that the solutions of the regularized problem converge to solutions of the original problem formulation for vanishing regularization parameter.

2.1. Problem setting

On an abstract level, we consider the constrained minimization problem

$$\begin{aligned} \min_{u\in\mathcal{M},\, y\in Y} \quad & J(y) + \psi(u), \\ \text{subject to} \quad & e(y,u) = 0 \quad \text{in } W^*. \end{aligned} \tag{$\mathcal{P}$}$$

As announced in the introduction, we search for a control u in the Banach space $\mathcal{M}$ and the state y in the Banach space Y. The state and control are coupled with the state equation e, which is formulated as a linear equation in the dual space of a third Banach space W.

2.1.1. Conditions for a well-posed problem

To ensure that $(\mathcal{P})$ has a solution, we state the following general assumptions:

- The space $\mathcal{M}$ is formed as the dual space of another, *separable* Banach space $\mathcal{C}$; the *predual* space. We have the identification

$$\mathcal{M} \cong \mathcal{C}^*.$$

  However, $\mathcal{M}$ will generally not be reflexive, i.e., we have $\mathcal{C} \subsetneq \mathcal{M}^*$. We denote the duality pairing of $u \in \mathcal{M}$ and $\varphi \in \mathcal{C}$ by $\langle u, \varphi \rangle = \langle u, \varphi \rangle_{\mathcal{M},\mathcal{C}}$.

  We will frequently work with the notion of weak-$*$ convergence in $\mathcal{M}$. Recall that a sequence $\{ u_n \}_n \subset \mathcal{M}$ for $n \in \mathbb{N}$ converges in the weak-$*$ sense towards $u \in \mathcal{M}$ (denoted by $u_n \rightharpoonup^* u$ in $\mathcal{M}$) if

$$\langle u_n, \varphi \rangle \to \langle u, \varphi \rangle \quad \text{for } n \to \infty \quad \text{for all } \varphi \in \mathcal{C}.$$

- The (extended real valued) functional $\psi \colon \mathcal{M} \to \mathbb{R} \cup \{ +\infty \}$ is convex, proper (not constantly equal to $+\infty$), and (sequentially) weak-$*$ lower semicontinuous. Furthermore, it is bounded from below and radially unbounded, i.e., we require for any sequence $\{ u_n \}_n \subset \mathcal{M}$ that

$$\|u_n\|_{\mathcal{M}} \to \infty \quad \text{implies} \quad \psi(u_n) \to \infty \qquad (\text{for } n \to \infty).$$

- The functional $J \colon Y \to \mathbb{R}$ is continuously differentiable (continuously Fréchet differentiable) and bounded from below.
- The spaces Y and W are reflexive. The state equation, which can be written as

$$\langle e(y, u), \varphi \rangle_{W^*,W} = 0 \quad \text{for all } \varphi \in W,$$

  admits for all $u \in \mathcal{M}$ a unique state solution $y \in Y$. The corresponding solution operator

$$S \colon \mathcal{M} \to Y, \quad S(u) = y$$

  is affine linear, bounded and (sequentially) weak-$*$ to strong continuous: we require

$$S(u_n) \to S(u) \ \text{ in } Y \quad \text{for all} \quad u_n \rightharpoonup^* u \ \text{ in } \mathcal{M}.$$

These assumptions will be verified for the concrete problems below.

Since we have supposed the existence of a solution operator for the state equation, we can define a reduced cost functional. We define the smooth part of the reduced cost functional $f \colon \mathcal{M} \to \mathbb{R}$ as

$$f(u) = J(S(u))$$

and the full reduced cost functional $j \colon \mathcal{M} \to \mathbb{R} \cup \{ \infty \}$ for $(\mathcal{P})$ as

$$j(u) = f(u) + \psi(u) = J(S(u)) + \psi(u).$$

Since the control-to-state mapping is continuous, it follows that f is continuous, as well.

Proposition 2.1. *The functional $f \colon \mathcal{M} \to \mathbb{R}$ is (sequentially) weak-$*$ continuous and the reduced cost functional $j \colon \mathcal{M} \to \mathbb{R}$ is (sequentially) weak-$*$ lower semicontinuous.*

Since $\mathcal{M}$ is the dual space of the separable space $\mathcal{C}$, the unit ball in $\mathcal{M}$ can be endowed with a norm (different from the Banach space norm) that induces the weak-$*$ topology on the unit ball of $\mathcal{M}$. As a consequence of this result and the Banach-Alaoglu theorem, any bounded sequence in $\mathcal{M}$ admits a weak-$*$ convergent subsequence (which is the sequential version of the Banach-Alaoglu theorem; see, e.g., [Bre11, Corollary 3.30]).

Theorem 2.2 (Banach-Alaoglu)**.** *Let $\{u_n\}_n$ for $n \in \mathbb{N}$ be a bounded sequence in $\mathcal{M}$. Then there exists a subsequence n_k for $k \in \mathbb{N}$ and a $u \in \mathcal{M}$ such that $u_{n_k} \rightharpoonup^* u$ in $\mathcal{M}$ for $k \to \infty$.*

Based on this and the general assumptions, we can give the following standard existence result based on the direct method.

Theorem 2.3. *The problem $(\mathcal{P})$ possesses at least one global solution*

$$(\bar{u}, \bar{y}) = (\bar{u}, S(\bar{u})) \in \mathcal{M} \times Y.$$

Proof. We give a proof for the sake of completeness. Since the functionals J and ψ are bounded from below and ψ is proper, we have

$$\inf_{u \in \mathcal{M}} j(u) = \hat{\jmath} \in \mathbb{R}.$$

Take any minimizing sequence $\{ u_n \}_n \subset \mathcal{M}$ for $n \in \mathbb{N}$. For n large enough we have

$$\psi(u_n) \leq j(u_n) = j(\tilde{u}) + 1$$

for some fixed $\tilde{u} \in \mathcal{M}$ where $\psi(\tilde{u})$ is finite. By the radial unboundedness of ψ, it follows that there exists a C independent of n, such that

$$\|u_n\|_{\mathcal{M}} \leq C \quad \text{for all } n \in \mathbb{N}.$$

With Theorem 2.2, we obtain a subsequence $\{ u_n \}_n$ (denoted again with the same index), such that $u_n \rightharpoonup^* \bar{u}$ in $\mathcal{M}$ for some $\bar{u} \in \mathcal{M}$. It follows with the lower semicontinuity of j from Proposition 2.1 that

$$j(\bar{u}) \leq \liminf_{n \to \infty} j(u_n) = \hat{\jmath},$$

which concludes the proof. □

2.1.2. Optimality conditions

To obtain optimality conditions for $(\mathcal{P})$, we recall some concepts from convex analysis (see, e.g., [ET99]) and nonlinear functional analysis (see, e.g., [Cla13; BS00]). We introduce the subdifferential of the convex functional ψ at the point $u \in \mathcal{M}$ as the set

$$\partial\psi(u) = \{ w \in \mathcal{C} \mid \langle \tilde{u} - u, w \rangle + \psi(u) \leq \psi(\tilde{u}) \quad \text{for all } \tilde{u} \in \mathcal{M} \}. \tag{2.1}$$

Note, that we define the subdifferential as a subset of the predual space $\mathcal{C}$ (which is natural if we endow $\mathcal{M}$ with the weak-$*$ topology; cf. [ET99]). Furthermore, we denote the directional derivative of f in direction $\delta u \in \mathcal{M}$ by $f'(u)(\delta u)$. By the general assumptions and the chain rule, it follows that f is Gâteaux differentiable in an optimal solution $\bar{u}$. We will verify below that the derivative of f is represented by a function in the predual space. Therefore, we identify

$$f'(\bar{u})(\delta u) = \langle \nabla f(\bar{u}), \delta u \rangle, \tag{2.2}$$

where $\nabla f(\bar{u}) \in \mathcal{C}$ is the gradient of f at $\bar{u}$. We obtain the following optimality condition.

Proposition 2.4. *Let $\bar{u} \in \mathcal{M}$ be an optimal solution from Theorem 2.3. We have the variational inequality*

$$-f'(\bar{u})(u - \bar{u}) + \psi(\bar{u}) \leq \psi(u) \quad \text{for all } u \in \mathcal{M}.$$

With (2.2), this can be expressed as $-\nabla f(\bar{u}) \in \partial\psi(\bar{u})$.

Proof. We give the elementary proof for the sake of completeness: By minimality of $\bar{u}$ and convexity of ψ it holds for every $u \in \mathcal{M}$ and $\lambda > 0$ that

$$0 \leq j(\bar{u} + \lambda(u - \bar{u})) - j(\bar{u}) \leq f(\bar{u} + \lambda(u - \bar{u})) - f(\bar{u}) + \lambda(\psi(\bar{u}) - \psi(u)).$$

Dividing by λ and letting $\lambda \to 0$, we obtain the result. □

We will compute the gradient of f with the help of the *adjoint* equation, which is done separately in the elliptic and parabolic setting in section 2.2 and section 2.3. Furthermore, we characterize the subdifferential of ψ to obtain optimality conditions. The characteristic example for the function ψ in this work is given by the norm

$$\psi(\cdot) = \|\cdot\|_{\mathcal{M}}.$$

In general, ψ will be positively homogeneous (of degree one), which means that $\psi(\lambda u) = |\lambda|\psi(u)$ for any $\lambda \in \mathbb{R}$ and $u \in \mathcal{M}$. It can be easily verified with the definition that all convex, positively homogeneous (of degree one) functions fulfill the triangle inequality, i.e., $\psi(u+v) \leq \psi(u)+\psi(v)$ for all u and v in $\mathcal{M}$. The subdifferential of such functions can be characterized with the following standard result.

Proposition 2.5. *Let $\psi(\cdot)$ be positively homogeneous (of degree one). For any $u \in \mathcal{M}$ and $w \in \mathcal{C}$, the inclusion $w \in \partial\psi(u)$ is equivalent to the conditions*

$$\sup_{\delta u \in \mathcal{M}, \psi(\delta u) \leq 1} \langle \delta u, w \rangle \leq 1 \quad \text{and} \quad \langle u, w \rangle = \psi(u).$$

Proof. For completeness, we prove the implication in one direction. The other direction is proved similarly. We take $\delta u \in \mathcal{M}$ arbitrary and set $\tilde{u} = u + \delta u$ in the definition of the subdifferential (2.1) to obtain

$$\langle \delta u, w \rangle \leq \psi(u + \delta u) - \psi(u) \leq \psi(\delta u)$$

with the triangle inequality. For δu with $\psi(\delta u) \leq 1$ this yields the first condition. For the second one we insert $\tilde{u} = \lambda u$ for $\lambda \in \mathbb{R}$ in the definition of the subdifferential to obtain $(1 - \lambda)\langle u, w \rangle \leq (|\lambda| - 1)\psi(u)$. From the choice $\lambda = 0$ and $\lambda = 2$ we obtain the second condition. □

Corollary 2.6. *For $\psi(\cdot) = \|\cdot\|_{\mathcal{M}}$, the inclusion $w \in \partial\psi(u)$ is equivalent to the conditions*

$$\|w\|_{\mathcal{C}} \leq 1 \quad \text{and} \quad \langle u, w \rangle = \psi(u).$$

Proof. It holds $\sup_{\delta u \in \mathcal{M}, \|\delta u\|_{\mathcal{M}} \leq 1} \langle \delta u, w \rangle = \|w\|_{\mathcal{C}}$, which is a consequence of the Hahn-Banach theorem; see, e.g., [Bre11, Corollary 1.4]. □

Remark 2.1. Proposition 2.5 can also be derived as a corollary from the equivalent characterization of $w \in \partial\psi(u)$ as

$$\psi(u) + \psi^*(w) = \langle w, u \rangle,$$

where $\psi^* \colon \mathcal{C} \to \mathbb{R}$ is the convex conjugate of ψ; see [ET99, Proposition 5.1]. For a positively homogeneous (of degree one) functional, the convex conjugate is given by the convex indicator function of the subdifferential of ψ at zero.

For the rest of this chapter, we will additionally assume that ψ is positively homogeneous (of degree one).

2.2. Elliptic problem setting

In the following, we present the required theory for the discussion of the elliptic problem. We denote by $\Omega \subset \mathbb{R}^d$ for $d \in \{2, 3\}$ a bounded domain, by $\Gamma \subset \partial\Omega$ a (relatively) open subset of the boundary (the set $\partial\Omega \setminus \Gamma$ is closed). Γ will be the part of the boundary where we impose Neumann conditions, and $\Gamma_D = \partial\Omega \setminus \Gamma$ will be the part where we impose Dirichlet conditions. We are mostly interested in controls of the form

$$u = \sum_{n=1}^{N} u_n \delta_{x_n}$$

with $u_n \in \mathbb{R}$ and $x_n \in \Omega \cup \Gamma$ arbitrary. We do not fix either the position of the x_n or the number of points N. However, the space of linear combinations of Dirac delta functions is not suitable for existence of an optimal solution in the general case (since it lacks the property from Theorem 2.2). Therefore, we consider the more general space of finite Radon measures. In other problem settings, we are interested in positive controls and the realistic cost term is linear, as mentioned in the introduction. Therefore, the controls are bounded in a L^1 norm, which is not enough to guarantee existence in L^1 (it is not reflexive). Again, to ensure existence of a solution in the general case, one can enlarge the solution space to the finite Radon measures.

2.2.1. Radon measures

For the elliptic problem, the control space is a subspace of the finite (signed) Borel measures $\mathcal{M}(\bar{\Omega})$ on the compact set $\bar{\Omega}$; see, e.g., [Rud87; Els11]. We recall some of the basic properties of this space. Any $u \in \mathcal{M}(\bar{\Omega})$ can be considered as a countably additive set function $u \colon \mathcal{B}(\bar{\Omega}) \to \mathbb{R}$, where $\mathcal{B}(\bar{\Omega})$ denotes the Borel sets on $\bar{\Omega}$. For every $u \in \mathcal{M}(\bar{\Omega})$, we define the corresponding *total variation measure* $|u| \colon \mathcal{B}(\bar{\Omega}) \to \mathbb{R}^+$ as

$$|u|(B) = \sup \Big\{ \sum_{n=1}^{\infty} |u(B_n)| \;\Big|\; B_n \in \mathcal{B}(\bar{\Omega}) \text{ disjoint partition of } B \Big\}.$$

The total variation measure is always positive and we denote the corresponding space of positive measures by $\mathcal{M}^+(\bar{\Omega})$. The total variation norm is now given as the total variation measure of the whole set, i.e., we set

$$\|u\|_{\mathcal{M}(\bar{\Omega})} = |u|(\bar{\Omega}) \quad \text{for all } u \in \mathcal{M}(\bar{\Omega}).$$

We endow $\mathcal{M}(\bar{\Omega})$ with this norm, which makes it a Banach space. For any bounded, continuous function $\varphi\colon \bar{\Omega} \to \mathbb{R}$, we can define the duality pairing between $u \in \mathcal{M}(\bar{\Omega})$ and φ as the integral

$$\langle u, \varphi \rangle = \int_{\bar{\Omega}} \varphi(x) \, \mathrm{d}u(x).$$

Furthermore, with the given duality pairing, the space $\mathcal{M}(\bar{\Omega})$ can be identified with the dual space of $\mathcal{C}(\bar{\Omega})$, the space of bounded, continuous functions: by the Riesz representation theorem (see, e.g., [Bre11, Theorem 4.31]), it follows that all bounded linear functionals on $\mathcal{C}(\bar{\Omega})$ can be represented as elements of $\mathcal{M}(\bar{\Omega})$. We identify

$$\mathcal{M}(\bar{\Omega}) \cong \mathcal{C}(\bar{\Omega})^*,$$

where $\mathcal{C}(\bar{\Omega})$ is endowed as usual with the supremum norm

$$\|\varphi\|_{\mathcal{C}(\bar{\Omega})} = \sup_{x \in \bar{\Omega}} |\varphi(x)|.$$

By this identification, we also obtain the alternative characterization of the total variation norm

$$\|u\|_{\mathcal{M}(\bar{\Omega})} = \sup_{\varphi \in \mathcal{C}(\bar{\Omega}),\, \|\varphi\|_{\mathcal{C}(\bar{\Omega})} \leq 1} \langle u, \varphi \rangle \quad \text{for all } u \in \mathcal{M}(\bar{\Omega}).$$

Note that finite Borel measures on subsets of $\mathbb{R}^d$ are regular (and therefore Radon measures; see, e.g., [Rud87, Theorem 2.18]). Thereby, the support of the measure $u \in \mathcal{M}(\bar{\Omega})$, which is defined as

$$\operatorname{supp} u = \operatorname{supp} |u| = \bar{\Omega} \setminus \Big(\bigcup \{\, B \text{ open} \mid |u|(B) = 0 \,\} \Big),$$

is a closed set (and therefore $\operatorname{supp} u \in \mathcal{B}(\bar{\Omega})$). As a consequence of the *Lebesgue-Radon-Nikodym* theorem, for any $u \in \mathcal{M}(\bar{\Omega})$ there exists a unique function $\operatorname{sgn} u \in L^1(\bar{\Omega}, |u|)$ (the space of integrable functions w.r.t. the measure $|u|$), such that

$$\mathrm{d}u = \operatorname{sgn} u \, \mathrm{d}|u| \quad \text{and} \quad |\operatorname{sgn} u(x)| = 1 \quad \text{for } x \in \bar{\Omega} \quad |u|\text{-almost everywhere.}$$

The expression $\mathrm{d}u = \operatorname{sgn} u \, \mathrm{d}|u|$ is an short-hand notation for $\int \varphi \, \mathrm{d}u = \int \varphi \operatorname{sgn} u \, \mathrm{d}|u|$ for all $\varphi \in \mathcal{C}(\bar{\Omega})$. Furthermore, with the Jordan decomposition theorem we can split any signed measure u into the uniquely defined positive and negative part

$$u = u^+ - u^- \quad \text{for } u^+, u^- \in \mathcal{M}^+(\bar{\Omega}),$$

with u^+ and u^- of minimal norm. The positive and negative part are given in terms of the sign function as $\mathrm{d}u^+ = (\operatorname{sgn} u)^+ \, \mathrm{d}|u|$ and $\mathrm{d}u^- = (\operatorname{sgn} u)^- \, \mathrm{d}|u|$, where $(\cdot)^+$ and $(\cdot)^-$ denote the positive and negative part of a function, respectively.

In accordance with the elliptic problem given below, which includes control on the domain and on the Neumann boundary, we consider the space of Radon measures on the control set, which is a subset of $\Omega \cup \Gamma$. In general, we require the control set

$$\Omega_c \subseteq \Omega \cup \Gamma,$$

to be relatively closed in $\Omega \cup \Gamma$, i.e., it shall hold $\Omega_c = \bar{\Omega}_c \cap (\Omega \cup \Gamma)$. It is clear that if we construct the space of Radon measures on the subset Ω_c, we obtain the same object as if we restrict the space $\mathcal{M}(\bar{\Omega})$ to the control set, i.e., we can identify

$$\mathcal{M}(\Omega_c) \cong \{\, u|_{\Omega_c} \mid u \in \mathcal{M}(\bar{\Omega}) \,\}.$$

Since Ω_c is generally not compact, we obtain additional zero boundary conditions for the predual space of continuous functions. Here, we can identify

$$\mathcal{M}(\Omega_c) \cong \mathcal{C}_0(\Omega_c)^*,$$

where $\mathcal{C}_0(\Omega_c)$ is the space of continuous and bounded functions which are zero on $\partial\Omega \setminus \Gamma$, i.e., we define $\mathcal{C}_0(\Omega_c)$ as the closure of $\mathcal{C}_c(\Omega_c)$ with respect to the supremum norm, where $\mathcal{C}_c(\Omega_c)$ denotes the compactly supported continuous functions on Ω_c. Note that this identification is compatible with the previous one, since $\mathcal{C}_0(B) = \mathcal{C}(B)$ holds for any compact set B.

2.2.2. Elliptic equations with measure data

In the following, we briefly outline the known regularity theory that is necessary for the discussion of the elliptic problem (see also Casas [Cas86]). For a given measure $u \in \mathcal{M}(\Omega \cup \Gamma)$, we will consider the convection-diffusion problem with measure valued data given (formally) by

$$\begin{aligned} -\nabla \cdot (\kappa \nabla y) + \beta \cdot \nabla y + c_0 y &= u|_\Omega \quad \text{in } \Omega, \\ n \cdot (\kappa \nabla y) + c_1 y &= u|_\Gamma \quad \text{on } \Gamma, \\ y &= 0 \qquad \text{on } \partial\Omega \setminus \Gamma. \end{aligned} \tag{2.3}$$

We make the general assumptions that $\partial\Omega$ is Lipschitz continuous (in the sense of [Gri85, Definition 1.2.1.1]) and $\kappa \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, $\beta \in W^{1,\infty}(\Omega, \mathbb{R}^d)$, $c_0 \in L^\infty(\Omega)$, $c_1 \in L^\infty(\Gamma)$. Moreover, the diffusion tensor κ takes values in the symmetric matrices and is uniformly elliptic, i.e., there exists an $\varepsilon > 0$, such that $\xi \cdot \kappa(x)\xi \geq \varepsilon |\xi|_2^2$ for all $x \in \Omega$ and $\xi \in \mathbb{R}^d$. Furthermore, we make the following standard assumptions to ensure unique solvability: we require $c_0 - \nabla \cdot \beta/2 \geq 0$ in Ω and $c_1 + n \cdot \beta/2 \geq 0$ on Γ. Moreover, in the case where $\Gamma = \partial\Omega$, not both of these expressions can be essentially zero.

We denote by $W^{1,p}(\Omega)$ for $p \in [1, \infty]$ the usual Sobolev spaces; see [AF03]. For the solution of the mixed boundary value problem, we introduce the Sobolev spaces $W_0^{1,p}(\Omega \cup \Gamma)$ with zero Dirichlet conditions on $\Gamma_D = \partial\Omega \setminus \Gamma$ as the closure of $\{\, \varphi|_\Omega \mid \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^d \setminus \Gamma_D) \,\}$ in $W^{1,p}(\Omega)$. We also define $W^{-1,p'}(\Omega\cup\Gamma)$ for $1/p+1/p' = 1$ as the corresponding dual spaces and abbreviate $H_0^1(\Omega \cup \Gamma) = W_0^{1,2}(\Omega \cup \Gamma)$ and $H^{-1}(\Omega \cup \Gamma) = W^{-1,2}(\Omega \cup \Gamma)$. We denote for any p and p' as above the extended $L^2(\Omega)$ duality pairing by

$$\langle f, v\rangle = \langle f, v\rangle_{W^{-1,p'}, W_0^{1,p}} \quad \text{for } f \in W^{-1,p'}(\Omega \cup \Gamma) \text{ and } v \in W_0^{1,p}(\Omega \cup \Gamma).$$

Furthermore, we assume that $\Omega\cup\Gamma$ is regular in the sense of Gröger; see [Grö89, Definition 2] or the alternative characterization in [HD+09]. Here, we are mostly interested in the special case of pure Neumann/Robin or pure Dirichlet conditions, where either $\Gamma = \partial\Omega$ or $\Gamma = \emptyset$. In this case, regularity follows immediately from the Lipschitz assumption on $\partial\Omega$, so we will not go into the details. We point out that it holds

$$W_0^{1,p}(\Omega \cup \Gamma) = \begin{cases} W_0^{1,p}(\Omega) & \text{for } \Gamma = \emptyset, \\ W^{1,p}(\Omega) & \text{for } \Gamma = \partial\Omega, \end{cases} \tag{2.4}$$

where the space $W_0^{1,p}(\Omega)$ is defined as usual. For the weak formulation of (2.3) we define the bilinear form

$$a(y, \varphi) = (\kappa \nabla y, \nabla \varphi) + (\beta \cdot \nabla y, \varphi) + (c_0 y, \varphi) + (c_1 y, \varphi)_\Gamma$$

for y and φ in $H_0^1(\Omega \cup \Gamma)$. By $(\cdot,\cdot)$ and $(\cdot,\cdot)_\Gamma$ we denote the $L^2(\Omega)$ and $L^2(\Gamma)$ inner product, respectively. The form a can be continuously extended for $u \in W_0^{1,s}(\Omega\cup\Gamma)$ and $\varphi \in W_0^{1,s'}(\Omega\cup\Gamma)$ for any $s \in (1,\infty)$, where $1/s + 1/s' = 1$. This is a consequence of Hölders inequality and the Sobolev trace theorem. Furthermore we recall the definition of the duality pairing

$$\langle u, \varphi \rangle = \int_\Omega \varphi \,\mathrm{d}u + \int_\Gamma \varphi \,\mathrm{d}u$$

for $u \in \mathcal{M}(\Omega \cup \Gamma)$ and $\varphi \in \mathcal{C}_0(\Omega \cup \Gamma)$. With the Sobolev embedding, the space $W_0^{1,s'}(\Omega \cup \Gamma)$ is continuously embedded into the Hölder continuous functions for $s' > d$. Therefore, we have

$$W_0^{1,s'}(\Omega \cup \Gamma) \hookrightarrow \mathcal{C}_0(\Omega \cup \Gamma) \quad \text{for} \quad s' > d.$$

Thereby, we can give the following weak formulation for (2.3). In the following, we fix Sobolev indices

$$s < \frac{d}{d-1} \quad \text{and } s' > d \quad \text{with} \quad \frac{1}{s} + \frac{1}{s'} = 1.$$

Definition 2.1. For given $u \in \mathcal{M}(\Omega \cup \Gamma)$, we call $y \in W_0^{1,s}(\Omega \cup \Gamma)$ a weak solution for (2.3) if it fulfills

$$a(y, \varphi) = \langle u, \varphi \rangle \quad \text{for all} \quad \varphi \in W_0^{1,s'}(\Omega \cup \Gamma). \tag{2.5}$$

However, it is well-known that the solutions of the weak formulation (2.5) are not unique in general; see Prignet [Pri95]. There are different techniques to obtain a unique solution (all of which lead to the same result in the present setting); see the overview and comparison in [MPS11]. We will use the method of duality, which goes back to Stampacchia [Sta65], and apply known regularity results from the literature. First, we consider the dual equation for $w \in H_0^1(\Omega \cup \Gamma)$ given as

$$a(\varphi, w) = \langle f, \varphi \rangle \quad \text{for all } \varphi \in H_0^1(\Omega \cup \Gamma) \tag{2.6}$$

for a given $f \in H^{-1}(\Omega \cup \Gamma)$. The existence of a unique solution of this equation follows from classical arguments. With the divergence theorem and the chain rule we can rewrite the non-symmetric part of the bilinear form as

$$(\beta \cdot \nabla y, \varphi) = \frac{1}{2}\left[(\beta \cdot \nabla y, \varphi) - (y, \beta \cdot \nabla \varphi) - (\nabla \cdot \beta y, \varphi) + ((n \cdot \beta) y, \varphi)_\Gamma\right].$$

In this form, it is evident that a is coercive. We have

$$a(y, y) = (\kappa \nabla y, \nabla y) + ((c_0 - \nabla \cdot \beta/2) y, y) + ((c_1 + n \cdot \beta/2) y, y)_\Gamma.$$

By our assumptions, we obtain the $H_0^1(\Omega \cup \Gamma)$ ellipticity of a. Now, the existence of a unique solution follows with the Lax-Milgram theorem.

Then, we suppose that f is additionally in $W^{-1,s'}(\Omega \cup \Gamma)$ and apply the following regularity result due to Griepentrog and Recke [GR01]; see also Droniou [Dro00] for a similar result and [Tro87, Theorem 3.6.(i)] for the case of a $\mathcal{C}^1$ domain with open and closed Γ.

Theorem 2.7 (Elliptic regularity [GR01, Theorem 6.3])**.** *For $f \in W^{-1,s'}(\Omega \cup \Gamma)$ with $s' > d$, the unique solution $w \in W_0^{1,2}(\Omega \cup \Gamma)$ to (2.6) is Hölder continuous up to the boundary, i.e., there exists a $\beta > 0$ such that $w \in \mathcal{C}^\beta(\bar{\Omega})$. Moreover, we have the a priori estimate*

$$\|w\|_{\mathcal{C}^\beta(\bar{\Omega})} \leq C_{s'} \|f\|_{W^{-1,s'}(\Omega\cup\Gamma)}$$

We denote the corresponding solution operator of (2.6) by $S_{\mathrm{dual}}\colon W^{-1,s'}(\Omega\cup\Gamma)\to\mathcal{C}_0(\Omega\cup\Gamma)$. With this *predual* solution operator we can construct the solution of the state equation.

Definition 2.2. We call $y\in W_0^{1,s}(\Omega\cup\Gamma)$ the solution by duality of (2.3) if it fulfills

$$\langle f,y\rangle=\langle u,S_{\mathrm{dual}}(f)\rangle\quad\text{for all } f\in W^{-1,s'}(\Omega\cup\Gamma). \tag{2.7}$$

In other words, we set $y=S_{\mathrm{dual}}^*(u)$, which directly implies that (2.7) has a unique solution.

Proposition 2.8. *For $u\in\mathcal{M}(\Omega\cup\Gamma)$, there exists a unique solution $y\in W_0^{1,s}(\Omega\cup\Gamma)$ (for every $s<d/(d-1)$) in the sense of Definition 2.2 with the corresponding estimate*

$$\|y\|_{W_0^{1,s}(\Omega\cup\Gamma)}\le C_s\|u\|_{\mathcal{M}(\Omega\cup\Gamma)}.$$

Furthermore, this solution also solves the weak formulation (2.5).

Proof. Since $\langle\cdot,y\rangle=\langle u,S_{\mathrm{dual}}(\cdot)\rangle$ is a continuous functional on $W^{-1,s'}(\Omega\cup\Gamma)$ by Theorem 2.7 and $W_0^{1,s}(\Omega\cup\Gamma)$ is reflexive, y can be identified with this functional and is therefore uniquely determined. For any given $w\in W_0^{1,s'}(\Omega\cup\Gamma)$ we define the functional $f_w\in W^{-1,s'}(\Omega\cup\Gamma)$ by $\langle f_w,\cdot\rangle=a(\cdot,w)$. Clearly, it holds $w=S_{\mathrm{dual}}(f_w)$. By density of $H_0^1(\Omega\cup\Gamma)$ in $W_0^{1,s}(\Omega\cup\Gamma)$ the weak formulation of the dual problem (2.6) also holds on this larger space and we have

$$a(\varphi,w)=\langle f_w,\varphi\rangle\quad\text{for all }\varphi\in W_0^{1,s}(\Omega\cup\Gamma).$$

Therefore, we have $a(y,w)=\langle y,f_w\rangle=\langle u,w\rangle$ and (2.5) is verified. □

We denote the corresponding solution operator by $S\colon\mathcal{M}(\Omega\cup\Gamma)\to W_0^{1,s}(\Omega\cup\Gamma)$. It can be easily verified with the definition of the solution by duality that it is weak-$*$ to weak continuous. Furthermore, the Sobolev embedding $W^{1,s}(\Omega)\hookrightarrow W^{1,s-\varepsilon}(\Omega)$ is compact for $\varepsilon>0$; therefore it is weak-$*$ to strong continuous with values in $W_0^{1,s-\varepsilon}(\Omega\cup\Gamma)$. Since the limiting case $s=d/(d-1)$ is excluded, we can without restriction omit the additional ε to obtain the following result.

Proposition 2.9. *The solution operator $S\colon\mathcal{M}(\Omega\cup\Gamma)\to W_0^{1,s}(\Omega\cup\Gamma)$ is weak-$*$ to strong continuous.*

Following [MPS11], we can employ an alternative, operator theoretic formulation of the solution by duality: we introduce the domain of the main part of the elliptic differential operator as

$$D_{s'}=\mathrm{dom}_{W^{-1,s'}(\Omega\cup\Gamma)}(\nabla\cdot\kappa\nabla)=\{\,v\in H_0^1(\Omega\cup\Gamma)\mid\nabla\cdot\kappa\nabla v\in W^{-1,s'}(\Omega\cup\Gamma)\,\},$$

which is endowed with the graph norm $\|v\|_{D_{s'}}=\|\nabla\cdot\kappa\nabla v+v\|_{W^{-1,s'}(\Omega\cup\Gamma)}$ for $v\in D_{s'}$. By construction, $(\nabla\cdot\kappa\nabla+\mathrm{Id})\colon D_{s'}\to W^{-1,s'}(\Omega\cup\Gamma)$ is an isomorphism. It should be noted that $W_0^{1,s'}(\Omega\cup\Gamma)\subset D_{s'}$ with continuous (but generally *not* dense) embedding. Observe that the bilinear form a can be canonically extended for $y\in W_0^{1,s}(\Omega\cup\Gamma)$ and $\varphi\in D_{s'}$ for certain $s'>d$.

Proposition 2.10. *Let $s'<2d/(d-2)$ (equivalently, $s>2d/(d+2)$). Then the bilinear form a can be continuously extended for $y\in W_0^{1,s}(\Omega\cup\Gamma)$ and $\varphi\in D_{s'}$.*

Proof. For the main part of the differential operator, this follows directly from the definition of $D_{s'}$. For the lower-order terms, we use for $\varphi \in D_{s'}$ the continuous embedding $D_{s'} \hookrightarrow H_0^1(\Omega \cup \Gamma)$ and the Sobolev trace and embedding theorems; therefore we need $s' < \infty$ for $d = 2$ and $s' \leq 6$ for $d = 3$. □

Furthermore, from Theorem 2.7 we deduce the continuous embedding

$$D_{s'} \hookrightarrow C^{\beta}(\bar{\Omega}).$$

Therefore, the pairing $\langle u, \varphi \rangle$ can be extended in the same way.

Proposition 2.11. *With $s' \in (d, 2d/(d-2))$, the solution by duality is given as the unique solution of*

$$a(y, \varphi) = \langle u, \varphi \rangle \quad \textit{for all } \varphi \in D_{s'}.$$

Furthermore, under additional suppositions on the smoothness of the coefficients, the space $D_{s'}$ is again given by $W_0^{1,s'}(\Omega \cup \Gamma)$.

Theorem 2.12. *(Elliptic regularity: smooth case) Suppose that one of the following conditions is fulfilled:*

- *$\partial\Omega$ is of class $\mathcal{C}^{1,\beta}$, $\Gamma = \partial\Omega$ or $\Gamma = \emptyset$, and $\kappa \in \mathcal{C}^{\beta}(\bar{\Omega}, \mathbb{R}^{d \times d})$.*
- *$\partial\Omega$ is Lipschitz, $\Gamma = \emptyset$, κ is the identity, $\beta \equiv 0$, $c_0 \equiv 0$, and $c_1 \equiv 0$ ($A = -\Delta$ is the negative Laplacian with zero Dirichlet boundary conditions).*

Then there exists a $s' > d$, such that for all $f \in W^{-1,s'}(\Omega \cup \Gamma)$ the solution w of (2.6) lies in $W_0^{1,s'}(\Omega \cup \Gamma)$, together with the a priori estimate

$$\|w\|_{W_0^{1,s'}(\Omega \cup \Gamma)} \leq C_{s'} \|f\|_{W^{-1,s'}(\Omega \cup \Gamma)}.$$

Proof. The first result can be found in, e.g., [Tro87, Theorem 1.6.(iv)]. The second result is due to Jerison and Kenig [JK95]. □

We will need this higher regularity for the error estimates in chapter 4. In the situation of Theorem 2.12, the weak formulation (2.5) guarantees uniqueness.

Corollary 2.13. *Suppose that one of the conditions of Theorem 2.12 is fulfilled. Then we can identify*

$$D_{s'} \cong W_0^{1,s'}(\Omega \cup \Gamma)$$

with equivalent norms.

2.2.3. Elliptic optimization problem

The elliptic problem is given in the general case as

$$\begin{aligned} \min_{u \in \mathcal{M}(\Omega_c),\, y \in W_0^{1,s}(\Omega)} \quad & J(y) + \psi(u), \\ \text{subject to} \quad & a(y, \varphi) = \langle \chi_{\Omega_c} u, \varphi \rangle \quad \text{for all } \varphi \in D_{s'}. \end{aligned} \tag{2.8}$$

Since the control is supported only on the control set Ω_c, we introduce $\chi_{\Omega_c}\colon \mathcal{M}(\Omega_c) \to \mathcal{M}(\Omega \cup \Gamma)$ as the canonical extension by zero operator. The standard example for the convex functional ψ is the norm

$$\psi(u) = \alpha \|u\|_{\mathcal{M}(\Omega_c)} = \int_{\Omega_c} \alpha \,\mathrm{d}|u|.$$

More generally, we can also consider the weighted norm

$$\psi(u) = \int_{\Omega_c} \hat{\alpha}(x) \,\mathrm{d}|u|(x)$$

for a continuous weight function $\hat{\alpha}\colon \bar{\Omega}_c \to \mathbb{R}^+$ which fulfills $\inf_{x\in\bar{\Omega}_c} \hat{\alpha}(x) > 0$. To enforce positivity of the measure, we can add a convex indicator function to ψ, i.e., we consider

$$\psi(u) = \alpha \|u\|_{\mathcal{M}(\Omega_c)} + \mathbb{I}_{\mathcal{M}^+(\Omega_c)}(u),$$

where $\mathbb{I}_{\mathcal{M}^+(\Omega_c)}(u) = 0$ for $u \in \mathcal{M}^+(\Omega_c)$ and $\mathbb{I}_{\mathcal{M}^+(\Omega_c)}(u) = +\infty$ otherwise. It easy to verify that each of these ψ fulfills the assumptions from section 2.1 with $\mathcal{M} = \mathcal{M}(\Omega_c)$.

Proposition 2.14. *All of the functionals ψ given above are convex, proper, weak-$*$ (sequentially) lower semicontinuous, and radially unbounded.*

Proof. The convexity can be checked with the definition and we have $\psi(0) = 0$. Radial unboundedness follows from

$$\psi(u) \geq \inf_{x\in\bar{\Omega}_c} \hat{\alpha}(x) \|u\|_{\mathcal{M}(\Omega_c)} \quad \text{for all } u \in \mathcal{M}(\Omega_c).$$

Weak-$*$ lower semicontinuity of the norm $\|\cdot\|_{\mathcal{M}(\Omega_c)}$ follows by the Banach-Alaoglu theorem (the norm ball is weak-$*$ compact). For the weighted norm we take a sequence $u_n \rightharpoonup^* u$ and define the weighted measures $\mathrm{d}\tilde{u}_n = \hat{\alpha}\,\mathrm{d}u_n$ and $\mathrm{d}\tilde{u} = \hat{\alpha}\,\mathrm{d}u$. We have $\langle \tilde{u}_n - \tilde{u}, \varphi\rangle = \langle u_n - u, \hat{\alpha}\varphi\rangle \to 0$ for $n \to \infty$. By construction it holds that $\int_{\Omega_c} \hat{\alpha}\,\mathrm{d}|u_n| = \|\tilde{u}_n\|_{\mathcal{M}(\Omega_c)}$ (the same holds for u). Now, we apply the previous result. To prove lower semicontinuity of the indicator function, we note that $u \geq 0$ in the sense of measures is equivalent to $\langle u, \varphi\rangle \geq 0$ for all $\varphi \in \mathcal{C}(\Omega_c)$ with $\varphi \geq 0$. □

For the functional $J\colon W^{1,s}(\Omega) \to \mathbb{R}$ we will mostly consider linear quadratic tracking functionals of the form

$$J(y) = \frac{1}{2}\|C_{\mathrm{obs}} y - y_d\|^2_{L^2(\Omega_o)}$$

for a subset $\Omega_o \subset \bar{\Omega}$ and an appropriate observation operator $C_{\mathrm{obs}}\colon W^{1,s}(\Omega) \to L^2(\Omega_o)$. With the continuous Sobolev embedding

$$W^{1,s}(\Omega) \hookrightarrow L^2(\Omega) \quad \text{for } s \geq \frac{2d}{d+2},$$

and since the interval $[2d/(d+2), d/(d-1))$ is not empty for $d \in \{2, 3\}$, we can consider distributed tracking on an open subset $\Omega_o \subset \Omega$. We define by $C_{\mathrm{obs}} = \chi_{\Omega_o}\colon W^{1,s}(\Omega) \to L^2(\Omega_o)$ the corresponding observation and restriction operator. We can also consider the case of boundary observation; cf. section 5.6. Then, Ω_o is chosen as a relatively open subset of Γ. With the trace theorem and the Sobolev embedding we have the chain of continuous embeddings

$$W^{1,s}(\Omega) \hookrightarrow W^{1-1/s,s}(\Gamma) \hookrightarrow L^q(\Gamma) \quad \text{for } q = \frac{d-s}{s(d-1)}.$$

Due to the restriction $s < d/(d-1)$ we can choose an arbitrary $q < \infty$ in the case $d = 2$ and $q < 2$ for $d = 3$. Therefore, at least in the two dimensional case, a boundary tracking term in L^2 is covered by the general analysis and we define again C_{obs} as the canonical embedding and restriction.

2.2.4. Optimality conditions

We define the Lagrange function for $u \in \mathcal{M}(\Omega_c)$, $y \in W_0^{1,s}(\Omega \cup \Gamma)$, and $p \in D_{s'}$ as

$$\mathcal{L}(u, y, p) = J(y) - a(y, p) + \langle \chi_{\Omega_c} u, p \rangle.$$

By construction, for arbitrary $p \in D_{s'}$ it holds that $f(u) = \mathcal{L}(u, y, p)$ for any $u \in \mathcal{M}(\Omega_c)$ and $y = S(u)$. Now, we fix any $\hat{u} \in \mathcal{M}(\Omega_c)$ and $\hat{y} = S(\hat{u})$. By the linearity of $y = S(u)$ and the chain rule, it follows directly that

$$f'(\hat{u})(\delta u) = \mathcal{L}'_y(\hat{u}, \hat{y}, \hat{p})(\delta y) + \mathcal{L}'_u(\hat{u}, \hat{y}, \hat{p})(\delta u) \quad \text{for any } \delta u \in \mathcal{M}(\Omega_c)$$

where $\delta y = S(\delta u)$. Now, we choose $\hat{p} \in D_{s'}$ as the solution of the corresponding adjoint equation

$$\mathcal{L}'_y(\hat{u}, \hat{y}, \hat{p})(\varphi) = J'(\hat{y})(\varphi) - a(\varphi, \hat{p}) = 0 \quad \text{for all} \quad \varphi \in W_0^{1,s}(\Omega \cup \Gamma),$$

which yields

$$f'(\hat{u})(\delta u) = \mathcal{L}'_u(\hat{u}, \hat{y}, \hat{p})(\delta u) = \langle \chi_{\Omega_c} \delta u, \hat{p} \rangle \quad \text{for all} \quad \delta u \in \mathcal{M}(\Omega_c).$$

Since $p \in D_{s'} \hookrightarrow C_0(\Omega \cup \Gamma)$, this gives the gradient of f. As a corollary of this representation and Proposition 2.4, we obtain the following optimality condition.

Theorem 2.15. *Let $(\bar{u}, \bar{y}) = (\bar{u}, S(\bar{u}))$ be the optimal solution of (2.8) from Theorem 2.3. There exists a unique adjoint state $\bar{p} \in D_{s'}$ solving the adjoint equation*

$$a(\varphi, \bar{p}) = J'(\bar{y})(\varphi) \quad \textit{for all } \varphi \in W_0^{1,s}(\Omega \cup \Gamma), \tag{2.9}$$

which fulfills the variational inequality

$$-\langle \chi_{\Omega_c}(u - \bar{u}), \bar{p} \rangle + \psi(\bar{u}) \leq \psi(u) \quad \textit{for all } u \in \mathcal{M}(\Omega_c). \tag{2.10}$$

Furthermore, we can characterize the subdifferential of ψ for each of the presented cases.

Theorem 2.16. *Let $\bar{u}$ be an optimal solution of (2.8) and $\bar{p}$ be the corresponding adjoint state.*
i) For $\psi(u) = \int_{\Omega_c} \hat{\alpha}(x) \, \mathrm{d}|u|(x)$ the variational inequality (2.10) can be equivalently expressed as

$$\begin{aligned} |\bar{p}(x)| &\leq \hat{\alpha}(x) \quad \textit{for } x \in \Omega_c, \\ \operatorname{supp} \bar{u}^+ &\subset \{\, x \in \Omega_c \mid \bar{p}(x) = -\hat{\alpha}(x) \,\}, \\ \textit{and } \operatorname{supp} \bar{u}^- &\subset \{\, x \in \Omega_c \mid \bar{p}(x) = \hat{\alpha}(x) \,\}. \end{aligned}$$

ii) For $\psi(u) = \int_{\Omega_c} \hat{\alpha}(x) \, \mathrm{d}u(x) + \mathbb{I}_{\mathcal{M}^+(\Omega_c)}(u)$ the variational inequality (2.10) can be equivalently expressed as

$$\begin{aligned} \bar{p}(x) &\geq -\hat{\alpha}(x) \quad \textit{for } x \in \Omega_c, \\ \textit{and } \operatorname{supp} \bar{u} &\subset \{\, x \in \Omega_c \mid \bar{p}(x) = -\hat{\alpha}(x) \,\}. \end{aligned}$$

Proof. We only show how conditions from i) and ii) can be derived from the variational inequality; the reverse direction can be proved similarly. We apply the first part of Proposition 2.5, which is

$$\sup_{\psi(\delta u)\leq 1} -\langle \delta u, \bar{p}\rangle \leq 1,$$

and compute the expression on the left-hand side in each case. By choosing for each $x \in \Omega_c$ the scaled Dirac delta function $\delta u = \pm 1/\hat{\alpha}(x)\,\delta_x$ in the case i) and $\delta u = -1/\hat{\alpha}(x)\,\delta_x$ in the case ii), we derive the first part of the result. With the second part of Proposition 2.5, we have in both cases that

$$\int_{\Omega_c} \operatorname{sgn}\bar{u}(x)\bar{p}(x)\,\mathrm{d}|\bar{u}|(x) = \int_{\Omega_c} \bar{p}(x)\,\mathrm{d}\bar{u}(x) = \psi(\bar{u}) = \int_{\Omega_c} \hat{\alpha}(x)\,\mathrm{d}|\bar{u}|(x).$$

We define the function $f(x) = \hat{\alpha}(x) - \operatorname{sgn}\bar{u}(x)\bar{p}(x)$ for $x \in \Omega_c$. We have $f(x) \geq 0$ for $x \in \Omega_c$ $|\bar{u}|$-almost everywhere due to the pointwise bounds on $\bar{p}$ and $\int_{\Omega_c} f(x)\,\mathrm{d}|\bar{u}|(x) = 0$ with the previous inequality. This implies that $f(x) = 0$ for all $x \in \Omega_c$ $|\bar{u}|$-almost everywhere. With continuity of $\bar{p}$ we can now derive the conditions on the support of $\bar{u}^+$ and $\bar{u}^-$. □

2.3. Parabolic problem setting

In the following we discuss a class of parabolic problems. We use the same notation and make the same assumptions on Ω, Γ, and Ω_c as in the previous section. Additionally, we denote the time interval by $I = (0, T)$ for some $T > 0$. In the parabolic setting, we are interested in controls of the form

$$u(t) = \sum_{n=1}^{N} u_n(t)\delta_{x_n}$$

with time-dependent coefficients $u_n \in L^2(I)$ and $x_n \in \Omega \cup \Gamma$ arbitrary. Again, we do not fix either the coefficient functions, the location of the points x_n, or the number N; we only require the points to be independent of time. To ensure that the resulting problem is well-posed, we again enlarge the solution space. In this case a space of vector measures will be the appropriate choice.

2.3.1. Vector measures

The space of vector measures can defined as the space of the countably additive set functions $u\colon \mathcal{B}(\bar{\Omega}) \to L^2(I)$ on the Borel sets with values in the Hilbert space $L^2(I)$; see, e.g., [Lan83, Section 12.3] or [Lan93, Section VII.4]. For $u \in \mathcal{M}(\bar{\Omega}, L^2(I))$ the total variation measure $|u| \in \mathcal{M}^+(\bar{\Omega})$ is defined as

$$|u|(B) = \sup\Big\{ \sum_{n=1}^{\infty} \|u(B_n)\|_{L^2(I)} \mid B_n \in \mathcal{B}(\bar{\Omega}) \text{ disjoint partition of } B\Big\}$$

for each $B \in \mathcal{B}(\bar{\Omega})$ and by $|u|(\bar{\Omega})$ we denote the total variation of u. It is easy to see that we have

$$\|u(B)\|_{L^2(I)} \leq |u|(B) \tag{2.11}$$

for all $B \in \mathcal{B}(\bar{\Omega})$. The space of vector measures with finite total variation is denoted by $\mathcal{M}(\bar{\Omega}, L^2(I))$. Endowed with the total variation norm $\|u\|_{\mathcal{M}(\bar{\Omega},L^2(I))} = \||u|\|_{\mathcal{M}(\bar{\Omega})} = |u|(\bar{\Omega})$ it

is a Banach space. The support of the vector measure u is defined as before with the total variation measure as

$$\operatorname{supp} u = \operatorname{supp}|u|.$$

For each $u \in \mathcal{M}(\bar{\Omega}, L^2(I))$ we can define a *polar decomposition*, which consist of the total variation measure $|u|$ and the function $u' \in L^1(\bar{\Omega}, |u|, L^2(I))$ (the space of functions with values in $L^2(I)$, which are integrable w.r.t $|u|$), such that

$$\mathrm{d}u = u' \,\mathrm{d}|u|, \tag{2.12}$$

which is a short form of $\int \varphi \,\mathrm{d}u = \int \varphi u' \,\mathrm{d}|u|$ in $L^2(I)$ for all $\varphi \in \mathcal{C}(\bar{\Omega})$. The function u' is the *Radon-Nikodym derivative* of u with respect to $|u|$; see [Lan83, Corollary 12.4.2] or [DU77, Corollary IV.1.4]. Certainly, u is absolutely continuous with respect to $|u|$ due to (2.11). In fact we even have $u' \in L^\infty(\bar{\Omega}, |u|, L^2(I))$ with

$$\begin{aligned} &\|u'\|_{L^\infty(\bar{\Omega},|u|,L^2(I))} \leq 1, \\ &\text{and} \quad \|u'(x)\|_{L^2(I)} = 1 \quad \text{for } x \in \bar{\Omega} \quad |u|\text{-almost everywhere.} \end{aligned} \tag{2.13}$$

The first property is a consequence of

$$\|\int_B u' \,\mathrm{d}|u|\|_{L^2(I)} = \|\int_B \mathrm{d}u\|_{L^2(I)} = \|\mu(B)\|_{L^2(I)} \leq |u|(B),$$

which implies that the $|u|$-average of u' lies in the unit ball of $L^2(I)$. By the averaging lemma [Lan83, Theorem 11.5.15] this implies $\|u'(x)\|_{L^2(I)} \leq 1$ $|u|$-almost everywhere. The second property is implicitly contained in [Lan83, Theorem 12.4.1]. By $\mathcal{C}(\bar{\Omega}, L^2(I))$ we denote the space of bounded continuous functions on $\bar{\Omega}$ with values in $L^2(I)$, endowed with the supremum norm

$$\|v\|_{\mathcal{C}(\bar{\Omega},L^2(I))} = \sup_{x\in\bar{\Omega}} \|v(x)\|_{L^2(I)}.$$

With the duality pairing defined for $u \in \mathcal{M}(\bar{\Omega}, L^2(I))$ and $v \in \mathcal{C}(\bar{\Omega}, L^2(I))$ as

$$\langle u, v\rangle = \int_{\bar{\Omega}} (u'(x), v(x))_{L^2(I)} \,\mathrm{d}|u|(x)$$

we have a natural injection into the dual space $\mathcal{M}(\bar{\Omega}, L^2(I)) \hookrightarrow \mathcal{C}(\bar{\Omega}, L^2(I))^*$. In fact this is an isometric isomorphism. The identification

$$\mathcal{M}(\bar{\Omega}, L^2(I)) \cong \mathcal{C}(\bar{\Omega}, L^2(I))^*$$

for a more general setting is known as Singer's representation theorem; see, e.g., [Hen96; Mez09].

As in the elliptic setting, the control set Ω_c is a relatively closed subset of $\Omega \cup \Gamma$, and we have the canonical embedding

$$\mathcal{M}(\Omega_c, L^2(I)) \hookrightarrow \mathcal{M}(\bar{\Omega}, L^2(I))$$

with the identification

$$\mathcal{M}(\Omega_c, L^2(I)) \cong \{\, u \in \mathcal{M}(\bar{\Omega}, L^2(I)) \mid \operatorname{supp} u \subseteq \Omega_c \,\}.$$

Let $\mathcal{C}_c(\Omega_c, L^2(I))$ be the space of continuous functions on Ω_c with values in $L^2(I)$ which are compactly supported in Ω_c. Then we define $\mathcal{C}_0(\Omega_c, L^2(I))$ as the closure of $\mathcal{C}_c(\Omega_c, L^2(I))$ with respect to the supremum norm. We note that this is equivalent to

$$\mathcal{C}_0(\Omega_c, L^2(I)) = \{\, \varphi \in \mathcal{C}(\bar{\Omega}_c, L^2(I)) \mid \varphi(x) = 0 \text{ for } x \in \partial\Omega \setminus \Gamma \,\}.$$

With the pairing defined for $u \in \mathcal{M}(\Omega_c, L^2(I))$ and $v \in \mathcal{C}_0(\Omega_c, L^2(I))$ as before we identify

$$\mathcal{M}(\Omega_c, L^2(I)) \cong \mathcal{C}_0(\Omega_c, L^2(I))^*.$$

There is another canonical space of functions which are L^2 in time and continuous in space; the Lebesgue-Bochner space $L^2(I, C_0(\Omega_c))$. This space is strictly smaller, i.e., we have the continuous embedding

$$L^2(I, \mathcal{C}_0(\Omega_c)) \hookrightarrow \mathcal{C}_0(\Omega_c, L^2(I)),$$

which can be directly verified. It has the (larger) dual space $L^2_w(I, \mathcal{M}(\Omega_c))$; we will address this space and compare it to the space of vector measures in section 2.3.5.

2.3.2. Parabolic equations with measure data

In the following, we briefly outline the necessary results from parabolic regularity theory for the discussion of the parabolic optimization problem. For a pointwise control problem, the following regularity results have been obtained by Droniou and Raymond [DR00] (in a more general setting). For equations with general measures on the parabolic cylinder, see also Casas [Cas97], Raymond and Zidani [RZ98], and Amann [Ama05]. For a given vector measure $u \in \mathcal{M}(\Omega \cup \Gamma, L^2(I))$, we will consider the parabolic equation given (formally) by

$$\begin{aligned} \partial_t y - \nabla \cdot (\kappa \nabla y) &= u|_\Omega && \text{in } I \times \Omega, \\ n \cdot (\kappa \nabla y) &= u|_\Gamma && \text{on } I \times \Gamma, \\ y &= 0 && \text{on } I \times (\partial\Omega \setminus \Gamma), \\ y(0) &= 0 && \text{in } \Omega. \end{aligned} \tag{2.14}$$

We define a suitable solution to equation (2.14) with the method of transposition (as in the elliptic setting). We make the same regularity assumptions on Ω and Γ and κ as in section 2.2.2. As before, we fix Sobolev indices

$$s < \frac{d}{d-1} \quad \text{and } s' > d \quad \text{with} \quad \frac{1}{s} + \frac{1}{s'} = 1.$$

We recall from section 2.2.2 the definition of the spaces $W_0^{1,s}(\Omega \cup \Gamma)$ and the domain $D_{s'}$. We define the elliptic operator

$$A \colon W_0^{1,s}(\Omega \cup \Gamma) \to D_{s'}^*$$

by the weak formulation

$$\langle Ay, \varphi \rangle = a(y, \varphi) = (\kappa \nabla y, \nabla \varphi) \quad \text{for } y \in W_0^{1,s}(\Omega \cup \Gamma) \quad \text{and } \varphi \in D_{s'}.$$

We can extend the definition canonically to $y \in L^2(I, W_0^{1,s}(\Omega \cup \Gamma))$ and $\varphi \in L^2(I, D_{s'})$ (which are the usual Lebesgue-Bochner spaces). We denote corresponding extension of the form a and the operator A again by the same symbol. By $\langle \cdot\,, \cdot \rangle_I$ we denote the extended $L^2(I \times \Omega)$ duality pairing between $L^2(I, W_0^{1,s}(\Omega \cup \Gamma))$ and its dual $L^2(I, W^{-1,s'}(\Omega \cup \Gamma))$.

At first, we consider for a given right hand side $f \in L^2(I, W^{-1,s'}(\Omega \cup \Gamma))$ the dual equation to (2.14), which is the backwards in time parabolic equation

$$\begin{aligned} -\partial_t w + A^* w &= f \quad \text{in } L^2(I, W^{-1,s'}(\Omega \cup \Gamma)), \\ w(T) &= 0, \end{aligned} \tag{2.15}$$

where the time derivative is interpreted as the distributional derivative [Ama95, Chapter III.1]. We apply a result on maximal parabolic regularity by Haller-Dintelmann and Rehberg [HDR09] to characterize the solutions of (2.15).

Theorem 2.17 (Theorem 5.4 in [HDR09])**.** *Suppose that $s' \in [2, 2d/(d-2))$. Then the unique solution w to* (2.15) *lies in the space*

$$X^{s'} = L^2(I, D_{s'}) \cap H^1(I, W^{-1,s'}(\Omega \cup \Gamma))$$

with the corresponding a priori estimate

$$\|w\|_{X^{s'}} \leq C_s \|f\|_{L^2(I, W^{-1,s'}(\Omega \cup \Gamma))}. \tag{2.16}$$

We denote the corresponding solution operator by $w = S_{\text{dual}}(f)$. With Theorem 2.17 it is an isomorphism on the spaces

$$S_{\text{dual}} \colon L^2(I, W^{-1,s'}(\Omega \cup \Gamma)) \to X^{s'}.$$

Moreover, since $D_{s'} \hookrightarrow \mathcal{C}^\beta(\Omega)$ for some $\beta > 0$, we additionally obtain the embedding

$$X^{s'} \hookrightarrow L^2(I, \mathcal{C}_0(\Omega \cup \Gamma)).$$

It should be noted that since $L^2(I, \mathcal{C}_0(\Omega \cup \Gamma)) \hookrightarrow \mathcal{C}_0(\Omega \cup \Gamma, L^2(I))$ the adjoint solution operator maps into the predual space. A very weak solution of the state equation (2.14) can now be given in the following way: consider dual exponents $s' \in (d, 2d/(d-2))$. For any $u \in \mathcal{M}(\Omega \cup \Gamma, L^2(I))$ the state $y \in L^2(I, W_0^{1,s}(\Omega \cup \Gamma))$ is defined as the solution of the very weak formulation

$$\langle y, -\partial_t \varphi + A^* \varphi \rangle_I = (y_0, \varphi(0)) + \langle u, \varphi \rangle \quad \text{for all } \varphi \in X^{s'} \quad \text{with } \varphi(T) = 0. \tag{2.17}$$

With the embedding $X^{s'} \hookrightarrow X^2$ and the trace theorem $X^2 \hookrightarrow \mathcal{C}(I, L^2(\Omega))$ (see, e.g., [Ama95, Theorem III 4.10.2]), the point evaluation $\varphi(0)$ is well defined and continuous in $X^{s'}$. Therefore for any $f \in L^2(I, W^{-1,s'}(\Omega \cup \Gamma))$ and the corresponding $w = S_{\text{dual}}(f)$ we obtain from the definition (2.15) that

$$\langle f, y \rangle_I = (y_0, w(0)) + \langle u, w \rangle \leq C \left(\|u\|_{\mathcal{M}(\Omega \cup \Gamma, L^2(I))} + \|y_0\|_{L^2(\Omega)} \right) \|w\|_{X^{s'}}.$$

By reflexivity of the space $L^2(I, W_0^{1,s}(\Omega \cup \Gamma))$ we now see that the very weak formulation has a unique solution and with (2.16) we obtain

$$\|y\|_{L^2(I, W_0^{1,s}(\Omega \cup \Gamma))} \leq C_s \left(\|u\|_{\mathcal{M}(\Omega \cup \Gamma, L^2(I))} + \|y_0\|_{L^2(\Omega)} \right). \tag{2.18}$$

By choosing appropriate test functions in (2.17) we derive that $\partial_t y = -Ay + \chi_{\Omega_c} u$ holds in the distributional sense and we obtain $\partial_t y \in L^2(I, W_0^{-1,s}(\Omega \cup \Gamma))$. With this and the integration by parts formula (see, e.g., [Ama05, Proposition 5.1]) it follows

$$\langle y(0), \varphi(0) \rangle = -\langle \partial_t y, \varphi \rangle_I - \langle y, \partial_t \varphi \rangle_I = (y_0, \varphi(0)) \leq \|y_0\|_{L^2(\Omega)} \|\varphi(0)\|_{L^2(\Omega)} \tag{2.19}$$

for all $\varphi \in H^1(I, W_0^{1,s'}(\Omega \cup \Gamma))$ with $\varphi(T) = 0$. We can choose φ with $\varphi(0) \in W_0^{1,s'}(\Omega \cup \Gamma)$ arbitrarily and conclude $y(0) = y_0$ by density. Finally, we obtain the following result.

Proposition 2.18. *The state equation, given in the weak formulation*

$$\begin{aligned} \langle \partial_t y, \varphi \rangle_I + a(y, \varphi) &= \langle u, \varphi \rangle \quad \text{for all } \varphi \in L^2(I, D_{s'}), \\ y(0) &= y_0, \end{aligned} \tag{2.20}$$

with $s' > d$ possesses a unique solution y in the space

$$Y^s = L^2(I, W_0^{1,s}(\Omega \cup \Gamma)) \cap H^1(I, W^{-1,s}(\Omega \cup \Gamma)),$$

where $1 \leq s < d/(d-1)$, with the corresponding estimate

$$\|y\|_{Y^s} \leq C_s \left(\|u\|_{\mathcal{M}(\Omega \cup \Gamma, L^2(I))} + \|y_0\|_{L^2(\Omega)} \right). \tag{2.21}$$

Proof. We take the unique solution $y \in Y^s$ of the very weak formulation (2.17), which fulfills (2.21) by (2.18) and the representation of the time derivative. The regularity for all $s < d/(d-1)$ is a consequence of the Sobolev embedding theorem. We argue that y is also a solution to the weak formulation (2.20) by applying integration parts in (2.17) to obtain

$$\langle u, \varphi \rangle = \langle y, -\partial_t \varphi \rangle_I + a(y, \varphi) - (y_0, \varphi(0)) = \langle \partial_t y, \varphi \rangle_I + a(y, \varphi)$$

for all $\varphi \in X^{s'}$ with $\varphi(T) = 0$. Since the space $\{ \varphi \in X^{s'} \mid \varphi(T) = 0 \}$ is dense in $L^2(I, D_{s'})$ the solution y fulfills (2.20), which proves existence for (2.20). Conversely, uniqueness of the solution to the weak formulation (2.20) follows by $X^{s'} \subset L^2(I, D_{s'})$. □

Remark 2.2. The weak formulation (2.20) holds also for test functions φ from the subspace $L^2(I, W_0^{1,s'}(\Omega \cup \Gamma))$. However, if we restrict the test space in this way, we lose uniqueness of the solution in the general case; cf. section 2.2.2.

We denote the corresponding solution operator for the state equation by $y = S(y_0, u) = S(u)$.

Proposition 2.19. *The solution operator $S \colon \mathcal{M}(\Omega \cup \Gamma, L^2(I)) \to Y^s$ is weak-$*$ to weak continuous. Furthermore, it is weak-$*$ to strong continuous with values in $L^2(I, W_0^{1,s}(\Omega \cup \Gamma))$.*

Proof. Since u_n is bounded in $\mathcal{M}(\Omega \cup \Gamma, L^2(I))$, the sequence $y_n = S(u_n)$ is bounded in Y^s with Proposition 2.18. Thus, it contains a weakly converging subsequence (denoted again by y_n) with $y_n \rightharpoonup \hat{y}$ for some $\hat{y} \in Y^s$. By taking the limit in (2.20), we see that $\hat{y} = S(\hat{u})$. The result follows since this argument can be repeated if we start from any subsequence of u_n. The second statement follows from the compact embedding

$$Y^s \hookrightarrow L^2(I, W_0^{1,s-\varepsilon}(\Omega \cup \Gamma)) \quad \text{for } \varepsilon > 0,$$

which follows with the Aubin-Lions lemma; see, e.g., [Lio69, Théorème I.5.1]. □

In the following we implicitly restrict the parameter s to the interval $(2d/(d+2), d/(d-1))$ when we use the spaces Y^s and $X^{s'}$, unless explicitly mentioned otherwise.

2.3.3. Parabolic optimization problem

With these preparations we can now state the precise problem formulation in the parabolic setting:

$$\begin{aligned} \min_{u\in\mathcal{M}(\Omega_c,L^2(I)),\, y\in Y^s} \quad & J(y)+\psi(u), \\ \text{subject to} \quad & \begin{cases} \langle \partial_t y,\varphi\rangle + a(y,\varphi) = \langle \chi_{\Omega_c} u,\varphi\rangle & \text{for all } \varphi\in L^2(I,D_{s'}), \\ y(0)=y_0. \end{cases} \end{aligned} \tag{2.22}$$

Again, χ_{Ω_c} denotes the canonical extension by zero

$$\chi_{\Omega_c}\colon \mathcal{M}(\Omega_c,L^2(I)) \to \mathcal{M}(\Omega\cup\Gamma,L^2(I)).$$

For the cost functional J we consider again the quadratic tracking functional

$$J(y)=\frac{1}{2}\|C_{\mathrm{obs}}y-y_d\|^2_{L^2(I\times\Omega_o)}$$

on the observation region $I\times\Omega_o$. With the properties of the solution operator from Proposition 2.19 we can choose Ω_o to be an open subset of Ω in the both the two and three dimensional case. In the two dimensional case, we can additionally consider a boundary tracking term; cf. section 2.2.3. The convex term ψ is either given by a multiple of the norm

$$\psi(u)=\alpha\|u\|_{\mathcal{M}(\Omega_c,L^2(I))}=\int_{\Omega_c}\alpha\,\mathrm{d}|u|(x),$$

or by a norm plus a convex indicator function to realize positivity constraints for the time-dependent coefficient,

$$\psi(u)=\alpha\|u\|_{\mathcal{M}(\Omega_c,L^2(I))}+\mathbb{I}_{\mathcal{M}^+(\Omega_c,L^2(I))}(u).$$

We say that a vector measure u is positive (non-negative) if the corresponding functional on $\mathcal{C}_0(\Omega_c,L^2(I))$ is non-negative: we have $u\in\mathcal{M}^+(\Omega_c,L^2(I))$ if it holds

$$\langle u,\varphi\rangle\geq 0 \quad \text{for all } \varphi\in\mathcal{C}_0(\Omega_c,L^2(I)) \quad \text{with } \varphi(t,x)\geq 0 \quad \text{for } (t,x)\in I\times\Omega_c.$$

It can be verified that $\mathcal{M}^+(\Omega_c,L^2(I))$ consists exactly of the measures for which the polar decomposition $\mathrm{d}u=u'\,\mathrm{d}|u|$ with $|u|\in\mathcal{M}^+(\Omega_c)$ and $u'\in L^\infty(\Omega_c,|u|,L^2(I))$ yields a positive function $u'(t,x)\geq 0$ for $(t,x)\in I\times\Omega_c$.

Proposition 2.20. *The two functionals ψ given above are convex, proper, weak-$*$ (sequentially) lower semicontinuous, and radially unbounded.*

Proof. The arguments are straightforward and analogous to the ones given in the elliptic case; cf. Proposition 2.14. □

2.3.4. Optimality conditions

Similar to the the elliptic case, we define the Lagrange function for $u\in\mathcal{M}(\Omega_c,L^2(I))$, $y\in Y^s$, and $p\in X^{s'}$ as

$$\mathcal{L}(u,y,p)=J(y)-\langle\partial_t y,p\rangle_I-a(y,p)+\langle\chi_{\Omega_c}u,p\rangle+(y(0)-y_0,p(0)).$$

We proceed as in the elliptic case. For any $\hat{u} \in \mathcal{M}(\Omega_c, L^2(I))$ and $\hat{y} = S(y_0, \hat{u})$ we derive

$$f'(\hat{u})(\delta u) = \mathcal{L}'_y(\hat{u}, \hat{y}, \hat{p})(\delta y) + \mathcal{L}'_u(\hat{u}, \hat{y}, \hat{p})(\delta u) \quad \text{for any } \delta u \in \mathcal{M}(\Omega_c)$$

where $\delta y = S(0, \delta u)$. Again, we choose $\hat{p} \in X^{s'}$ as the solution of the corresponding adjoint equation

$$\mathcal{L}'_y(\hat{u}, \hat{y}, \hat{p})(\varphi) = J'(\hat{y})(\varphi) - \langle \partial_t \varphi, \hat{p} \rangle_I - a(\varphi, \hat{p}) + (\varphi(0), \hat{p}(0)) = 0 \quad \text{for all } \varphi \in Y^s.$$

We can apply integration by parts in time to obtain the equivalent formulation

$$-\langle \varphi, \partial_t \hat{p} \rangle_I + a(\varphi, \hat{p}) + (\varphi(T), \hat{p}(T)) = J'(\hat{y})(\varphi) \quad \text{for all } \varphi \in Y^s.$$

By choosing special test functions φ, we obtain that this condition is equivalent to $\hat{p}(T) = 0$ and $-\partial_t \hat{p} + A^* \hat{p} = J'(\hat{y}) \in L^2(I, W^{-1,s'}(\Omega))$; cf. Proposition 2.18. With the unique solution $\hat{p} \in X^{s'}$ of the adjoint equation (see Theorem 2.17), we can now express the gradient of f at $\hat{u}$ as

$$f'(\hat{u})(\delta u) = \mathcal{L}'_u(\hat{u}, \hat{y}, \hat{p})(\delta u) = \langle \chi_{\Omega_c} \delta u, \hat{p} \rangle \quad \text{for all } \delta u \in \mathcal{M}(\Omega_c).$$

As a corollary of this representation and Proposition 2.4, we obtain the following optimality condition.

Theorem 2.21. *For any optimal solution $(\bar{u}, \bar{y}) = (\bar{u}, S(y_0, \bar{u}))$ of (2.22), there exists a unique adjoint state $\bar{p} \in X^{s'}$ solving the adjoint equation*

$$\begin{aligned} -\langle \varphi, \partial_t \bar{p} \rangle + a(\varphi, \bar{p}) &= J'(\bar{y})(\varphi) \quad \text{for all } \varphi \in L^2(I, W^{1,s}(\Omega \cup \Gamma)), \\ \bar{p}(T) &= 0, \end{aligned} \tag{2.23}$$

and the subgradient condition

$$-\langle \chi_{\Omega_c}(u - \bar{u}), \bar{p} \rangle + \psi(\bar{u}) \leq \psi(u) \quad \text{for all } u \in \mathcal{M}(\Omega_c, L^2(I)). \tag{2.24}$$

From the variational inequality (2.24) we can derive additional properties of the optimal control.

Theorem 2.22. *Let $\bar{u}$ be an optimal solution of (2.22), $\bar{u}' \, \mathrm{d}|\bar{u}| = \mathrm{d}\bar{u}$ its polar decomposition, and $\bar{p}$ be the corresponding adjoint state.*

i) For $\psi(u) = \alpha \|u\|_{\mathcal{M}(\Omega_c, L^2(I))}$ the variational inequality (2.24) can be equivalently expressed as

$$\|\bar{p}(x)\|_{L^2(I)} \leq \alpha \quad \text{for all } x \in \Omega_c, \tag{2.25}$$

$$\operatorname{supp}|\bar{u}| \subset \{ x \in \Omega_c \mid \|\bar{p}(x)\|_{L^2(I)} = \alpha \}, \tag{2.26}$$

$$\alpha \bar{u}'(x) + \bar{p}(x) = 0 \quad \text{for } x \in \Omega_c \quad |\bar{u}|\text{-almost everywhere.} \tag{2.27}$$

ii) For $\psi(u) = \alpha \|u\|_{\mathcal{M}(\Omega_c, L^2(I))} + \mathbb{I}_{\mathcal{M}^+(\Omega_c, L^2(I))}(u)$ the variational inequality (2.10) can be equivalently expressed as

$$\|\bar{p}^-(x)\|_{L^2(I)} \leq \alpha \quad \text{for all } x \in \Omega_c, \tag{2.28}$$

$$\operatorname{supp}|\bar{u}| \subset \{ x \in \Omega_c \mid \|\bar{p}^-(x)\|_{L^2(I)} = \alpha \}, \tag{2.29}$$

$$\alpha \bar{u}'(x) - \bar{p}^-(x) = 0 \quad \text{for } x \in \Omega_c \quad |\bar{u}|\text{-almost everywhere.} \tag{2.30}$$

Here, $\bar{p}^- \in \mathcal{C}_0(\Omega_c, L^2(I))$ denotes the negative part of $\bar{p}$ given as $\bar{p}^-(t,x) = \bar{p}(t,x)^- = -\min(0, \bar{p}(t,x))$ for $(t,x) \in I \times \Omega_c$.

Proof. We first show i). By Proposition 2.5 we obtain

$$\sup_{\alpha\|\delta u\|_{\mathcal{M}(\Omega_c,L^2(I))}\leq 1} -\langle \chi_{\Omega_c}\delta u, \bar{p}\rangle = \frac{1}{\alpha}\|\bar{p}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} \leq 1.$$

The first equality follows from the inequality $|\langle \chi_{\Omega_c}\delta u, \bar{p}\rangle| \leq \|\delta u\|_{\mathcal{M}(\Omega_c,L^2(I))}\|\bar{p}\|_{\mathcal{C}_0(\Omega_c,L^2(I))}$ and the fact that equality holds for a special choice of δu (choose $\delta u = -1/\alpha\, p(\hat{x})\delta_{\hat{x}}$, where $\hat{x} \in \Omega_c$ is a point such that $\|\bar{p}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} = \|\bar{p}(\hat{x})\|_{L^2(I)}$ is achieved.) This already shows (2.25). With the second condition from Proposition 2.5 we obtain

$$-\langle \chi_{\Omega_c}\bar{u}, \bar{p}\rangle = \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))} = \int_{\Omega_c} \alpha \,\mathrm{d}|\bar{u}|.$$

Applying the polar decomposition and reordering we obtain

$$\int_{\Omega_c} \Big(\alpha + (\bar{u}'(x), \bar{p}(x))_{L^2(I)}\Big)\,\mathrm{d}|\bar{u}|(x) = 0. \tag{2.31}$$

With the Cauchy-Schwarz inequality and the conditions (2.13) and (2.25) we obtain

$$-(\bar{u}'(x), \bar{p}(x))_{L^2(I)} \leq \|\bar{u}'(x)\|_{L^2(I)}\|\bar{p}(x)\|_{L^2(I)} \leq \alpha \tag{2.32}$$

for $x \in \Omega_c$ $|\bar{u}|$-almost everywhere, which means that the integrand in (2.31) is non-negative. Therefore it must be zero almost everywhere, i.e., it follows

$$-(\bar{u}'(x), \bar{p}(x))_{L^2(I)} = \alpha \quad \text{for } x \in \Omega_c \quad |\bar{u}|\text{-almost everywhere.}$$

Considering again (2.32), we see that equality can only hold if the conditions

$$\|\bar{p}(x)\|_{L^2(I)} = \alpha \quad \text{and} \quad \bar{p}(t,x) = -\alpha\bar{u}'(t,x)$$

hold for $|\bar{u}|$-almost all $x \in \Omega_c$ and almost every $t \in I$. This proves (2.27). From the first identity, we can derive (2.26) using basic measure theoretic arguments: define the function $f\colon \Omega_c \to \mathbb{R}^+$, $f(x) = \alpha - \|\bar{p}(x)\|_{L^2(I)}$, which is positive and continuous due to $\bar{p} \in X^{s'} \hookrightarrow \mathcal{C}_0(\Omega_c, L^2(I))$. Furthermore, it fulfills $\int_{\Omega_c} f(x)\,\mathrm{d}|\bar{u}|(x) = 0$. We can easily argue that $\operatorname{supp}|\bar{u}|$ must be a subset of the zero set $\{\, x \in \Omega_c \mid f(x) = 0 \,\}$, for instance by a contradiction argument.

To show ii), we take an arbitrary $\delta u \in \mathcal{M}^+(\Omega_c, L^2(I))$ with $\alpha\|\delta u\|_{\mathcal{M}(\Omega_c,L^2(I))} \leq 1$ and compute

$$-\langle \chi_{\Omega_c}\delta u, \bar{p}\rangle = -\langle \chi_{\Omega_c}\delta u, \bar{p}^+\rangle + \langle \chi_{\Omega_c}\delta u, \bar{p}^-\rangle \leq \frac{1}{\alpha}\|\bar{p}^-\|_{\mathcal{C}_0(\Omega_c,L^2(I))},$$

using the positivity of δu, where the positive part $\bar{p}^+ = \bar{p} + \bar{p}^-$ is defined similarly as the negative part. Again, equality is achieved for a special choice of δu and we obtain (2.28) by Proposition 2.5 as before. As in the previous case, we obtain (2.31). In this case, the integrand in (2.31) is positive as well, due to

$$-(\bar{u}'(x), \bar{p}(x))_{L^2(I)} = -(\bar{u}'(x), \bar{p}^+(x))_{L^2(I)} + (\bar{u}'(x), \bar{p}^-(x))_{L^2(I)} \leq \|\bar{u}'(x)\|_{L^2(I)}\|\bar{p}^-(x)\|_{L^2(I)} \leq \alpha$$

for $x \in \Omega_c$ $|\bar{u}|$-almost everywhere. Here, we have used again the positivity of u' and $\bar{p}^+$. Due to equality in (2.31) we derive

$$-(\bar{u}'(x), \bar{p}^-(x))_{L^2(I)} = \alpha \quad \text{for } x \in \Omega_c \quad |\bar{u}|\text{-almost everywhere,}$$

which implies (2.30) and (2.29) as before. □

We can identify a characteristic special case.

Corollary 2.23. *Suppose that equality in* (2.25) *is only achieved in a finite collection of points,*

$$\{\, x \in \Omega_c \mid \|\bar{p}(x)\|_{L^2(I)} = \alpha \,\} = \{\, x_n \,\}_{n=1,\dots,N}.$$

Then $\bar{u}$ is given by a sum of point sources: for some coefficients $c_n \geq 0$ for $n \in \{\, 1 \dots N \,\}$ we have

$$\bar{u} = \frac{1}{\alpha} \sum_{n=1}^{N} c_n \, \bar{p}(\cdot, x_n) \, \delta_{x_n}.$$

Proof. We infer from (2.26) that $|\bar{u}| = \sum_{n=1}^{N} c_n \delta_{x_n}$ for some positive coefficients $c_n > 0$. Then the time dependent coefficients $u_n \in L^2(I)$ are given due to (2.27) by the formula $u_n(t) = c_n \bar{u}'(t, x_n) = -c_n/\alpha \, \bar{p}(t, x_n)$. □

By (2.12), the optimal controls of (2.22) have a "sparsity pattern" which is independent of time. To emphasize this point, we will compare the problem formulation to a different one, which does not have this property.

2.3.5. Comparison to another problem formulation

Up to now we have always considered the controls $u \in \mathcal{M}(\Omega_c, L^2(I))$ as objects depending on the spatial variable $x \in \Omega_c$. Now we want to switch the point of view to consider u as a variable of $t \in I$. For a given $u \in \mathcal{M}(\Omega_c, L^2(I))$ and its polar decomposition $\mathrm{d}u = u' \, \mathrm{d}|u|$, we can define the control $u(t) \in \mathcal{M}(\Omega_c)$ at time t by

$$\mathrm{d}u(t) = u'(t) \, \mathrm{d}|u| \quad \text{for } t \in I \quad \text{almost everywhere.} \tag{2.33}$$

Due to (2.13) the polar decomposition u' is an element of the Hilbert space $L^2(\Omega_c, |u|, L^2(I))$, which is isometrically isomorphic to $L^2(I, L^2(\Omega_c, |u|))$ with Fubini's theorem.

From this point of view, it is natural to directly consider controls $I \ni t \mapsto u(t) \in \mathcal{M}(\Omega_c)$, such that $\|u(t)\|_{\mathcal{M}(\Omega_c)}$ is square integrable in time. Since $\mathcal{M}(\Omega_c)$ is not separable, it is necessary to distinguish between weakly and strongly measurable functions. We recall from [CCK13] and the references therein the definition of the Bochner space $L^2_w(I, \mathcal{M}(\Omega_c))$ (of weakly measurable $\mathcal{M}(\Omega_c)$-valued functions which are square integrable in time) and the identification of the dual

$$L^2_w(I, \mathcal{M}(\Omega_c)) \cong L^2(I, \mathcal{C}_0(\Omega_c))^*.$$

We have the continuous embedding

$$\mathcal{M}(\Omega_c, L^2(I)) \hookrightarrow L^2_w(I, \mathcal{M}(\Omega_c)), \tag{2.34}$$

which follows from the (dense) embedding $L^2(I, \mathcal{C}_0(\Omega_c)) \hookrightarrow \mathcal{C}_0(\Omega_c, L^2(I))$. Therefore, for each $u \in \mathcal{M}(\Omega_c, L^2(I))$ the expression $u(t) \in \mathcal{M}(\Omega_c)$ for $t \in I$ is also defined with the help of this embedding. We can verify that this is compatible with the definition given in (2.33) independently of the equivalence representations chosen for $u' \colon I \to L^2(\Omega_c, |u|)$ and $u \colon I \to \mathcal{M}(\Omega_c)$. It is obvious that the inclusion (2.34) is strict, i.e., it holds

$$\mathcal{M}(\Omega_c, L^2(I)) \subsetneq L^2_w(I, \mathcal{M}(\Omega_c)).$$

We will give an example below.

Let us compare the conditions from Theorem 2.22 with the optimality system obtained by Casas, Clason, and Kunisch [CCK13] for the problem

$$\min_{u\in L^2_w(I,\mathcal{M}(\Omega_c))} J(S(y_0,u)) + \alpha\|u\|_{L^2_w(I,\mathcal{M}(\Omega_c))}. \tag{2.35}$$

For problem (2.35) the optimality condition (see [CCK13, Theorem 3.3]) implies that for almost all $t \in I$ it holds

$$\operatorname{supp}|u(t)| \subseteq \{\, x \in \Omega_c \mid |p(t,x)| = \|p(t)\|_{\mathcal{C}_0(\Omega_c)} \,\},$$

which means that the support of $u(t)$ is variable over time. Note, that this implies significantly lower regularity for problem (2.35) in comparison with problem (2.22) under consideration. For instance, a regularity result such as $\bar{u} \in \mathcal{C}(\bar{I}, \mathcal{M}(\Omega_c))$ (cf. Theorem 5.8) cannot be expected. Indeed, it is false for problem (2.35). We just have to consider as an example the measure $u \in L^2_w(I, \mathcal{M}(\bar{\Omega}))$, $u \notin \mathcal{M}(\bar{\Omega}, L^2(I))$ with $I = \Omega = (0,1)$ defined by

$$u(t) = g(t)\delta_t, \tag{2.36}$$

with a nontrivial smooth function g with $g(1) = 0$. Here, the Dirac delta function moves in space as time increases. Such controls can actually be found as optimal solutions of (2.35): choosing A as the negative Laplacian with homogeneous Neumann boundary conditions it is possible to construct a desired state y_d such that the optimal solution of (2.35) is given by (2.36). The construction is analogous to the one which will be given in section 5.5.

2.4. An approach with convex duality

In both of the examples that have been presented above, the smooth part of the objective is of linear quadratic type. More precisely, we can consider most of the given concrete examples as instances of the problem

$$\min_{u\in\mathcal{M}} \quad \frac{1}{2}\|S_{\text{obs}}u - y_d\|_V^2 + \alpha\|u\|_{\mathcal{M}}, \tag{2.37}$$

where V is a Hilbert space and $S_{\text{obs}}\colon \mathcal{M} \to V$ is a linear control-to-observation operator. In the concrete elliptic case, we can form $S_{\text{obs}} = C_{\text{obs}} \circ S$ as the concatenation of the solution operator S and the observation operator C_{obs}. In the parabolic setting, $S = S(y_0, \cdot)$ is only affine linear for $y_0 \neq 0$. We can set the observation operator to $S_{\text{obs}} = C_{\text{obs}} \circ S(0,\cdot)$ and replace y_d by $y_d + S(y_0,\cdot)$ to bring it in the form (2.37). The problem (2.37) is convex and due to its specific structure a dual problem can be identified with the *Fenchel duality* theorem; see [ET99, Theorem III.4.1]. This is the approach chosen by Clason and Kunisch [CK11a; CK11b] in the context of an optimal control problem with an elliptic equation as in section 2.2. There, a dual problem is derived, which can be equivalently given as

$$\begin{aligned} \min_{v\in V} \quad & \frac{1}{2}\|v + y_d\|_V^2, \\ \text{subject to} \quad & \|S^*_{\text{obs}}(v)\|_{\mathcal{C}} \leq \alpha. \end{aligned} \tag{2.38}$$

Here, $S^*_{\text{obs}}\colon V \to \mathcal{C}$ is the predual operator of S_{obs} with $(S_{\text{obs}}u, v)_V = \langle u, S^*_{\text{obs}}v\rangle$ for all $u \in \mathcal{M}$, $v \in V$. It can be identified as $S^*_{\text{obs}} = S_{\text{dual}} \circ C^*_{\text{obs}}$. The optimal solutions of (2.37) and (2.38) are

linked via the relation $\bar{v} = S_{\text{obs}}\bar{u} - y_d$ and the optimal solution $\bar{u}$ plays the role of the Lagrange multiplier for the constraint in (2.38). The optimality conditions from Theorem 2.16 can then be derived or interpreted as complementarity conditions for the constraint on $\bar{p} = S^*_{\text{obs}}(\bar{v})$ and the multiplier $\bar{u}$. We remark that a similar analysis would likely also be valid for the parabolic problem discussed in section 2.3.

Here, we have chosen a direct primal approach with the subdifferential as outlined in section 2.1. This is motivated by the fact that this approach can be more easily extended to nonlinear problem settings, where either S is not (affine) linear or J is not convex. In fact, a similar approach is used by Casas and Kunisch [CK14] for a problem with a semilinear elliptic equation. We mention the approach with the dual problem (2.38), since it gives an important connection to *state constrained* optimization; see, e.g., [Cas86]. The analysis of sparse control problems with measures profits from the advanced level of research in this area. For instance, for the numerical realization we will introduce a regularized problem in section 2.5. In terms of the dual problem, this regularization corresponds to a quadratic penalty of the constraints; see [CK11a; CK11b]. Due to this parallel we are able to adapt techniques from state constrained optimization (see, e.g., [HK06a; HSW14]) for an improved analysis of the regularization error. However, in the construction of the optimization algorithms in section 3 (where we can work with L^2 controls and bypass the regularity theory for measure valued controls) we will take special care to also handle the general case, where Fenchel-duality would not be directly applicable.

2.5. Hilbert space regularization

For the algorithmic realization we consider a regularized problem. In the regularized problem, the control is searched for in the Hilbert space H. We generally identify H with its dual space $H \cong H^*$ and we abbreviate the inner product and norm in H by

$$(u, v) = (u, v)_H \quad \text{and} \quad \|u\| = \|u\|_H.$$

We suppose that the space $\mathcal{C}$ is *densely* embedded into H. Furthermore, we require that the inner product in H is compatible with the duality pairing between $\mathcal{M} = \mathcal{C}^*$ and $\mathcal{C}$, such that it holds

$$\langle u, \varphi \rangle = (u, \varphi) \quad \text{for all } u \in H, \text{ and } \varphi \in \mathcal{C}.$$

Thereby, we have the chain of continuous embeddings

$$\mathcal{C} \hookrightarrow H \cong H^* \hookrightarrow \mathcal{M}.$$

Recall that since $\mathcal{C} \hookrightarrow H$ is dense, the second embedding is injective.

Now we introduce the (Tikhonov-)regularized version of $(\mathcal{P})$. For a positive regularization parameter $\gamma > 0$ we consider the problem

$$\begin{aligned} \min_{u \in H,\, y \in Y} \quad & J(y) + \psi(u) + \frac{\gamma}{2}\|u\|^2, \\ \text{subject to} \quad & e(y, u) = 0 \quad \text{in } W^*. \end{aligned} \tag{$\mathcal{P}_\gamma$}$$

We make the same suppositions on J, ψ and e as before. In particular, since H is continuously embedded into $\mathcal{M}$, we can reuse the control-to-state mapping from above by concatenating it with the embedding. We obtain the solution operator

$$S \colon H \to Y, \quad S(u) = y,$$

denoted again by the same symbol. However, since the space H is smaller than $\mathcal{M}$, S maps H to a space of higher regularity than Y, which we will use in the concrete applications. For $u \in H$, we denote the reduced objective of $(\mathcal{P}_\gamma)$ by

$$j_\gamma(u) = j(u) + \frac{\gamma}{2}\|u\|^2 = J(S(u)) + \psi(u) + \frac{\gamma}{2}\|u\|^2,$$

and define the corresponding smooth part as

$$f_\gamma(u) = f(u) + \frac{\gamma}{2}\|u\|^2 = J(S(u)) + \frac{\gamma}{2}\|u\|^2.$$

We defer a concrete description of the regularized problems to chapters 4 and 5. Let us only point out that for the elliptic problem we will choose $H = L^2(\Omega_c)$ and for the parabolic problem we will choose $H = L^2(\Omega_c, L^2(I)) = L^2(I \times \Omega_c)$. Furthermore, it holds that

$$\begin{aligned} \|u\|_{\mathcal{M}(\bar{\Omega})} &= \|u\|_{L^1(\Omega)} \quad \text{for } u \in L^2(\Omega), \\ \|u\|_{\mathcal{M}(\bar{\Omega}, L^2(I))} &= \|u\|_{L^1(\Omega, L^2(I))} \quad \text{for } u \in L^2(I \times \Omega), \end{aligned}$$

which provides a different interpretation of the considered functionals ψ in the regularized setting.

2.5.1. Existence and optimality conditions

Under the general conditions from section 2.1 the regularized problem $(\mathcal{P}_\gamma)$ is also well posed. Since the objective functional contains the squared norm of H, any minimizing sequence is bounded in H. Furthermore, any weakly converging sequence in H converges also in the weak-$*$ sense in $\mathcal{M}$. Thereby, we obtain the following result (in the same way as in Theorem 2.3).

Proposition 2.24. *The problem $(\mathcal{P}_\gamma)$ possesses at least one global solution*

$$(\bar{u}_\gamma, \bar{y}_\gamma) = (\bar{u}_\gamma, S(\bar{u}_\gamma)) \in H \times Y.$$

Similarly to Proposition 2.4, we can obtain a necessary optimality condition for $(\mathcal{P}_\gamma)$. In the regularized setting, we define the subdifferential of ψ as in (2.1) as a subset of H. We work with Gâteaux differentiability in the Hilbert space H, and identify

$$(\nabla f(u), \delta u) = f'(u)(\delta u) \quad \text{for } u \in H.$$

Due to the compatibility of the duality pairing and the inner product, we obtain the same expressions for the gradient of f. The gradient of f_γ is given by

$$\nabla f_\gamma(u) = \nabla f(u) + \gamma u \quad \text{for } u \in H.$$

Thereby, we obtain the following necessary conditions as in Proposition 2.4.

Proposition 2.25. *Let $\bar{u}_\gamma$ be an optimal solution of $(\mathcal{P}_\gamma)$. It holds*

$$-(\nabla f(\bar{u}_\gamma) + \gamma \bar{u}_\gamma, u - \bar{u}_\gamma) + \psi(\bar{u}_\gamma) \leq \psi(u) \quad \textit{for all } u \in H. \tag{2.39}$$

Alternatively, this can be expressed as $-\nabla f(\bar{u}_\gamma) - \gamma \bar{u}_\gamma \in \partial\psi(\bar{u})$.

We can again derive more concrete conditions using the specific structure of ψ. In the original (unregularized) problem setting, we could only obtain information about the support of $|\bar{u}|$, cf. sections 2.2.4 and 2.3.4. In the regularized setting, we can derive an explicit formula for $\bar{u}_\gamma$ in terms of the optimal gradient $\nabla f(\bar{u}_\gamma)$. It is easy to see that (2.39) can be alternatively expressed as

$$\bar{u}_\gamma = \operatorname*{argmin}_{u \in H} \left[(\nabla f(\bar{u}_\gamma), u) + \frac{\gamma}{2}\|u\|^2 + \psi(u) \right],$$

which has a unique solution due to strong convexity. This can be elegantly expressed with the help of the *proximal map* of ψ, which we will discuss in section 3.1. We also refer to the examples in section 3.3.2 and the optimality conditions given for the concrete examples in chapters 4 and 5.

2.5.2. Regularization error

Define for $\gamma > 0$ the value function v as

$$v(\gamma) = j_\gamma(\bar{u}_\gamma) = j(\bar{u}_\gamma) + \frac{\gamma}{2}\|\bar{u}_\gamma\|^2 = J(S(\bar{u}_\gamma)) + \psi(\bar{u}_\gamma) + \frac{\gamma}{2}\|\bar{u}_\gamma\|^2.$$

We derive an estimate for the error introduced due to regularization in terms of the cost functional

$$j(\bar{u}) = \inf_{u \in \mathcal{M}} j(u) = v(0).$$

Note that the value of the objective in a globally optimal solution from 2.3 is independent of the specific solution, which may not be uniquely determined. In the context of Moreau-Yosida regularization of state constrained optimization problems, it is known that the value function is differentiable and concave; see Hintermüller and Kunisch [HK06a, Proposition 4.1]. In fact, we can obtain with similar techniques (see also [HK06b; Sch09; Sch13]) that

$$v'(\gamma) = \frac{1}{2}\|\bar{u}_\gamma\|^2 \geq 0, \tag{2.40}$$

$$v''(\gamma) \leq 0 \tag{2.41}$$

for almost all $\gamma \in (0, \infty)$. Similar results are also well-known in the context of Tikhonov-regularization of inverse problems; cf., e.g., [IK92; JLS09; LR10; KKV11].

Proposition 2.26. *The value function $v\colon [0,\infty) \to [j(\bar{u}), \infty)$ is (locally) Lipschitz-continuous, concave and differentiable for $\gamma \in (0,\infty)$ almost everywhere with derivatives as in* (2.40) *and* (2.41).

Proof. By comparing functional values we can verify that v is concave. In fact, for any $\gamma_0 > 0$, $\gamma > 0$, and the convex combination $\gamma_\theta = \theta\,\gamma_0 + (1-\theta)\,\gamma$ we have

$$\begin{aligned}\theta\, v(\gamma_0) + (1-\theta)\, v(\gamma) &\leq \theta\, j_{\gamma_0}(\bar{u}_{\gamma_\theta}) + (1-\theta)\, j_\gamma(\bar{u}_{\gamma_\theta}) \\ &= j(\bar{u}_{\gamma_\theta}) + \theta\frac{\gamma_0}{2}\|\bar{u}_{\gamma_\theta}\|^2 + (1-\theta)\frac{\gamma}{2}\|\bar{u}_{\gamma_\theta}\|^2 = j_{\gamma_\theta}(\bar{u}_{\gamma_\theta}) = v(\gamma_\theta)\end{aligned}$$

by minimality of $j_{\gamma_0}(\bar{u}_{\gamma_0})$ and $j_\gamma(\bar{u}_\gamma)$. The fact that v is concave directly implies that it is locally Lipschitz-continuous; see, e.g., [BC11, Proposition 8.28]. Furthermore it is differentiable almost everywhere by Rademacher's theorem; see, e.g., [EG92, Section 3.1.2]. Existence of the

second derivatives almost everywhere and formula (2.41) follows by the Alexandrov theorem; see, e.g., [EG92, Section 6.4]. It remains to compute the form of the first derivative as in (2.40). Therefore we consider that for every $\varepsilon > 0$ the difference quotient is bounded from above and below by

$$\frac{1}{\varepsilon}\left(v(\gamma_0+\varepsilon)-v(\gamma_0)\right) \leq \frac{1}{\varepsilon}\left(j_{\gamma_0+\varepsilon}(\bar{u}_{\gamma_0})-j_{\gamma_0}(\bar{u}_{\gamma_0})\right) = \frac{1}{2}\|\bar{u}_{\gamma_0}\|^2$$
$$= \frac{1}{\varepsilon}\left(j_{\gamma_0}(\bar{u}_{\gamma_0})-j_{\gamma_0-\varepsilon}(\bar{u}_{\gamma_0})\right) \leq \frac{1}{\varepsilon}\left(v(\gamma_0)-v(\gamma_0-\varepsilon)\right)$$

Therefore, letting $\varepsilon \to 0$ we obtain

$$\mathrm{d}v(\gamma_0, 1) \leq \frac{1}{2}\|\bar{u}_{\gamma_0}\|^2 \leq -\,\mathrm{d}v(\gamma_0, -1),$$

where $\mathrm{d}v(\gamma_0, \pm 1)$ denotes the directional derivative of v in positive and negative direction (which exist, since v is concave). This implies $v'(\gamma) = 1/2\,\|\bar{u}_\gamma\|^2$ for $\gamma > 0$ almost everywhere. □

To show continuity of the value function at zero, which implies the convergence of the regularized functional values for $\gamma \to 0$, we can apply a very general argument. For this we need a special approximating sequence in H of the optimal solution $\bar{u} \in \mathcal{M}$.

Assumption 2.1. Suppose that for $\bar{u} \in \mathcal{M}$ there exists a sequence $\{u_n\}_{n\in\mathbb{N}} \subset H$ with $\psi(u_n) \to \psi(\bar{u})$ and $u_n \rightharpoonup^* \bar{u}$ in $\mathcal{M}$ for $n \to \infty$.

The existence of a sequence with the second property follows in the concrete settings by the weak-$*$ density of H in $\mathcal{M}$. The convergence of the functional values is less obvious. For the spaces $\mathcal{M} = \mathcal{M}(\bar{\Omega})$ and $H = L^2(\bar{\Omega})$ and the functional $\psi(\cdot) = \alpha\|\cdot\|_{\mathcal{M}(\bar{\Omega})}$ we refer to [Bre11, Problem 24]. We give the analogous argument for vector measures in Appendix A.1.

Proposition 2.27. *Suppose that Assumption 2.1 holds. Then the value function is continuous at zero with limit* $v(0) = \lim_{\gamma\to 0^+} v(\gamma) = j(\bar{u})$.

Proof. Take any optimal $\bar{u} \in \mathcal{M}$ and the sequence $\{u_n\}_{n\in\mathbb{N}} \subset H$ from Assumption 2.1 with $\psi(u_n) \to \psi(\bar{u})$ and $u_n \rightharpoonup^* \bar{u}$ in $\mathcal{M}$ for $n \to \infty$. Then we have for all $\gamma > 0$ and all $n \in \mathbb{N}$ that

$$v(0) = j(\bar{u}) \leq j(\bar{u}_\gamma) \leq j(\bar{u}_\gamma) + \frac{\gamma}{2}\|\bar{u}_\gamma\|^2 = v(\gamma) \leq j(u_n) + \frac{\gamma}{2}\|u_n\|^2 \tag{2.42}$$

by optimality of $\bar{u}$ and $\bar{u}_\gamma$. By the convergence of the u_n we have

$$j(u_n) = f(u_n) + \psi(u_n) \to f(\bar{u}) + \psi(\bar{u}) = j(\bar{u}) \quad \text{for } n \to \infty$$

since f is weak-$*$ continuous in $\mathcal{M}(\Omega_c)$. Therefore, for any given $\varepsilon > 0$ we can choose $n_\varepsilon \in \mathbb{N}$ large enough such that

$$j(\bar{u}) \leq j(u_{n_\varepsilon}) \leq j(\bar{u}) + \frac{\varepsilon}{2}.$$

Combining this with (2.42) results in

$$j(\bar{u}) \leq j(u_{n_\varepsilon}) + \frac{\gamma}{2}\|u_{n_\varepsilon}\|^2 \leq j(\bar{u}) + \frac{\varepsilon}{2} + \frac{\gamma}{2}\|u_{n_\varepsilon}\|^2.$$

Note that this inequality holds independently of $\gamma > 0$. Choosing now $\gamma_\varepsilon \leq \varepsilon/\|u_{n_\varepsilon}\|^2$ we obtain $j(\bar{u}) \leq v(\gamma_\varepsilon) \leq j(\bar{u}) + \varepsilon$ with arbitrary $\varepsilon > 0$, which concludes the proof. □

As a consequence of the continuity of the value function, we obtain the (subsequential) weak-$*$ convergence of the regularized solutions $\bar{u}_\gamma$ to an optimal solution $\bar{u}$ of the original problem formulation for $\gamma \to 0$.

Theorem 2.28. *Consider a sequence of solutions $\bar{u}_\gamma \in H$ of problem $(\mathcal{P}_\gamma)$ for $\gamma \to 0$. There exists a subsequence γ_n for $n \in \mathbb{N}$ and a solution $\bar{u} \in \mathcal{M}$ of $(\mathcal{P})$ such that*

$$\bar{u}_{\gamma_n} \rightharpoonup^* \bar{u} \quad \text{in } \mathcal{M}, \quad \text{for } n \to \infty.$$

If $(\mathcal{P})$ has a unique solution $\bar{u}$, we have $\bar{u}_\gamma \rightharpoonup^* \bar{u}$ for any sequence $\gamma \to 0$.

Proof. As in Theorem 2.3, the values of $\psi(\bar{u}_\gamma)$ are bounded by

$$\psi(\bar{u}_\gamma) \leq j_\gamma(\bar{u}_\gamma) \leq J(S(0))$$

due to the optimality of $\bar{u}_\gamma$ and $\psi(0) = 0$, and we can select a subsequence γ_n such that it holds $\bar{u}_{\gamma_n} \rightharpoonup^* \hat{u}$ in $\mathcal{M}$ for some $\hat{u} \in \mathcal{M}$. Due to the weak-$*$ lower semicontinuity of j it holds $j(\hat{u}) \leq \liminf_{n\to\infty} j_{\gamma_n}(\bar{u}_{\gamma_n}) = v(0) = j(\bar{u})$ with Proposition 2.27. Therefore $\hat{u}$ is an optimal solution for $(\mathcal{P})$. If the solution is unique, the convergence of the whole sequence follows since the argument can be repeated if we start from an arbitrary subsequence of $\bar{u}_\gamma$. □

Furthermore, the error due to regularization can be expressed as an integral over the regularization term.

Corollary 2.29. *Suppose that Assumption 2.1 holds. We have the error representation*

$$j_\gamma(\bar{u}_\gamma) - j(\bar{u}) = v(\gamma) - v(0) = \int_0^\gamma \frac{1}{2}\|\bar{u}_\sigma\|^2 \,\mathrm{d}\sigma.$$

Proof. Since v is locally Lipschitz continuous on $(0, \infty)$ by Proposition 2.26, and due to $v'(\gamma) = 1/2\,\|\bar{u}_\gamma\|^2$ almost everywhere, we obtain for $0 < \varepsilon < \gamma$ that $v(\gamma) - v(\varepsilon) = \int_\varepsilon^\gamma v'(\sigma)\,\mathrm{d}\sigma = \int_\varepsilon^\gamma 1/2\,\|\bar{u}_\sigma\|^2\,\mathrm{d}\sigma$. Letting $\varepsilon \to 0$ and applying Proposition 2.27 yields the result. □

The preceding arguments were based mainly on the global optimality of the $\bar{u}_\gamma$ and little on the concrete structure of the problem. In a convex setting, which we will address in the next section, it is possible to show more.

2.5.3. Regularization error in the convex case

In this chapter, we will additionally assume that J is convex. Note that this is the case for a quadratic tracking functional. Since the solution operator S was already assumed to be affine linear, the reduced objective functional j is therefore convex, as well. This directly implies that $(\mathcal{P}_\gamma)$ has a unique solution, since the reduced cost functional

$$j_\gamma(\cdot) = j(\cdot) + \frac{\gamma}{2}\|\cdot\|^2$$

is even strongly convex (see, e.g., [BC11, Corollary 11.8]).

Proposition 2.30. *Assume that J is convex. Then, the solution $\bar{u}_\gamma$ of $(\mathcal{P}_\gamma)$ is unique.*

Then we obtain the well-known result that the unique optimal solution $\bar{u}_\gamma$ depends Lipschitz continuously on the regularization parameter γ; see also [HK06a; HK06b; JLS09; LR10; WW11].

Proposition 2.31. *Assume that J is convex. We have for all $\gamma > 0$ and $\rho > 0$ that*

$$\|\bar{u}_\gamma - \bar{u}_\rho\| \leq \frac{|\rho - \gamma|}{\gamma} \|\bar{u}_\rho\|$$

Proof. We insert the optimal solutions for $\bar{u}_\gamma$ and $\bar{u}_\rho$ into the variational inequality from Proposition 2.25 for ρ and γ, respectively, which yields

$$\begin{aligned} -(\nabla f(\bar{u}_\rho) + \rho \bar{u}_\rho, \bar{u}_\gamma - \bar{u}_\rho) + \psi(\bar{u}_\gamma) \leq \psi(\bar{u}_\rho), \\ -(\nabla f(\bar{u}_\gamma) + \gamma \bar{u}_\gamma, \bar{u}_\rho - \bar{u}_\gamma) + \psi(\bar{u}_\rho) \leq \psi(\bar{u}_\gamma). \end{aligned}$$

Adding both inequalities and rearranging results in

$$\begin{aligned} &(\gamma \bar{u}_\gamma - \rho \bar{u}_\rho, \bar{u}_\gamma - \bar{u}_\rho) + (\nabla f(\bar{u}_\gamma) - \nabla f(\bar{u}_\rho), \bar{u}_\gamma - \bar{u}_\rho) \\ &\qquad = \gamma \|\bar{u}_\gamma - \bar{u}_\rho\|^2 + (\gamma - \rho)(\bar{u}_\rho, \bar{u}_\gamma - \bar{u}_\rho) + (\nabla f(\bar{u}_\gamma) - \nabla f(\bar{u}_\rho), \bar{u}_\gamma - \bar{u}_\rho) \leq 0. \end{aligned}$$

Since f is convex, $\nabla f \colon H \to H$ is a monotone operator (see, e.g., [BC11, Proposition 17.10]), and the last term is positive and can be dropped (for a linear quadratic tracking term as in (2.37) we even have $(\nabla f(\bar{u}_\gamma) - \nabla f(\bar{u}_\rho), \bar{u}_\gamma - \bar{u}_\rho) = \|S_{\mathrm{obs}}(\bar{u}_\gamma - \bar{u}_\rho)\|^2$). This yields

$$\gamma \|\bar{u}_\gamma - \bar{u}_\rho\|^2 \leq (\rho - \gamma)(\bar{u}_\rho, \bar{u}_\gamma - \bar{u}_\rho) \leq |\rho - \gamma| \|\bar{u}_\rho\| \|\bar{u}_\gamma - \bar{u}_\rho\|,$$

and we divide by $\|\bar{u}_\gamma - \bar{u}_\rho\|$ to conclude the proof. □

It follows that the value function is continuously differentiable with Lipschitz continuous derivative, which implies that (2.40) holds for all $\gamma > 0$.

Proposition 2.32. *Assume that J is convex. The value function v is continuously differentiable with Lipschitz continuous second derivative. It holds*

$$0 \leq -v''(\gamma) \leq \frac{1}{\gamma} \|\bar{u}_\gamma\|^2 \quad \text{for } \gamma > 0 \quad \text{almost everywhere.}$$

Proof. Since $\gamma \to 1/2 \, \|u_\gamma\|^2$ is continuous, v' has a continuous representative. We have

$$\begin{aligned} v'(\gamma) - v'(\rho) &= \frac{1}{2} \left(\|\bar{u}_\gamma\|^2 - \|\bar{u}_\rho\|^2 \right) = \frac{1}{2} (\bar{u}_\gamma - \bar{u}_\rho, \bar{u}_\gamma + \bar{u}_\rho) \\ &\leq \frac{1}{2} \|\bar{u}_\gamma - \bar{u}_\rho\| \|\bar{u}_\gamma + \bar{u}_\rho\| \leq \frac{|\gamma - \rho|}{2\gamma} \|\bar{u}_\gamma + \bar{u}_\rho\| \|\bar{u}_\rho\| \leq \frac{|\gamma - \rho|}{2\gamma} \left(\|\bar{u}_\gamma\| \|\bar{u}_\rho\| + \|\bar{u}_\rho\|^2 \right) \end{aligned}$$

with Proposition 2.31. Dividing by $|\gamma - \rho|$, letting $\rho \to \gamma$, and using concavity of the value function yields the result. □

The previous structural results for the value function will be used for an asymptotic a priori estimate of the regularization error of the functional in sections 4.5.1 and 5.4.1. Furthermore, we will use them for an a posteriori estimate of the same error in section 6.2.

2.5.4. Computation of the second derivatives

In chapter 3 we will discuss a second order optimization algorithm for the solution of $(\mathcal{P}_\gamma)$ based on a semismooth Newton method for the reduced cost functional j_γ. Therefore, we require the gradient and second derivatives of the smooth part $f(\cdot) = J(S(\cdot))$, while the nonsmooth part ψ is treated differently. More precisely, we need evaluations of the Hessian at $u \in H$ in directions $\delta u \in H$. The Hessian is defined as usual by

$$(\varphi, \nabla^2 f(u)\delta u) = f''(u)(\varphi, \delta u) \quad \text{for all } \varphi \in H,$$

where f'' is the second derivative of f. To ensure that f is two times continuously differentiable, we additionally assume that:

- The functional $J\colon Y \to \mathbb{R}$ is twice continuously (Fréchet) differentiable.

In the general setting of section 2.1, where $S\colon u \mapsto y$ is affine linear, the derivative of S is given by $S'(\cdot) = S(\cdot) - S(0)$ and by the chain rule we obtain for $u \in H$ that

$$\begin{aligned} \nabla f(u) &= (S')^* J'(S(u)) \\ \nabla^2 f(u)\delta u &= (S')^* J''(S(u))S'(\delta u) \quad \text{for all } \delta u \in H. \end{aligned}$$

As usual, $(S')^*\colon Y^* \to H^* \cong H$ denotes the adjoint of S'. We have already seen in the elliptic and parabolic case that the gradient $\nabla f(u)$ can be expressed with the solution of the *adjoint* equation. In the same way, for a given $\delta u \in H$ the Hessian product $\nabla^2 f(u)\delta u$ can be expressed with an auxiliary *tangent* and a *second adjoint* equation; see below.

A more general setting

However, in the Hilbert space setting, there is no need to restrict attention to problems with linear state equation. The optimization algorithms we will discuss in chapter 3 are applicable to a much larger class of optimization problems. There, we can work under the following more general assumptions on the state equation:

- Define $U_{\text{ad}} = \operatorname{dom}\psi = \{\, u \in H \mid \psi(u) < \infty \,\}$. For all $u \in U_{\text{ad}}$, the state equation $e(y, u) = 0$ in W has a unique solution $y = S(u) \in Y$. The corresponding solution operator $S\colon H \to Y$ is weak to strong continuous. The operator

  $$e(\cdot, \cdot)\colon Y \times H \to W^*$$

  is twice continuously (Fréchet) differentiable on a neighborhood of $Y \times U_{\text{ad}}$ and the partial derivative

  $$e'_y(y, u)\colon Y \to W^*$$

  is an isomorphism for all $(y, u) \in Y \times U_{\text{ad}}$.

In this setting, we can still show existence of an optimal solution with the same arguments as in Proposition 2.24. Furthermore, we can compute algorithmically useful expressions for the gradient and the second derivative (in the sense of continuous Fréchet differentiability).

Remark 2.3. The ideas behind the following computations are well-known; cf. also [Ulb11, Appendix A.1]. We only sketch the derivation for the sake of completeness. We also mention that the conditions given above are still very restrictive and fail to cover many interesting problems, especially in a parabolic setting. However, the results of the computations below remain valid

for a much larger class of problems. We also refer to Lions [Lio71], Tröltzsch [Trö10b] and Ito and Kunisch [IK08, Chapter 5] for different conditions and settings where similar results can be derived.

We fix some $\hat{u} \in U_{\text{ad}}$ and denote $\hat{y} = S(\hat{u})$. There exists an open ball $\mathcal{N}(\hat{u}) \subset H$ around $\hat{u}$, such that the solution operator

$$S\colon \mathcal{N}(\hat{u}) \to Y \quad \text{with } e(S(u), u) = 0 \quad \text{for } u \in H,$$

is well-defined and continuously differentiable. This follows with the implicit function theorem; see, e.g., [Die69, Theorem 10.2.1]. Since e is twice differentiable, this property transfers to $S\colon \mathcal{N}(\hat{u}) \to Y$; see, e.g., [Die69, Theorem 10.2.3]. This allows to compute the first and second derivative of f at u. As before, we define the Lagrange function as

$$\mathcal{L}(u, y, p) = J(u) - \langle e(y, u), p\rangle \quad \text{for } (u, y, p) \in U \times Y \times W.$$

As in sections 2.2.4 and 2.3.4, we have for any $p \in W$ that

$$f(u) = \mathcal{L}(u, S(u), p) \quad \text{for all } u \in \mathcal{N}(\hat{u}).$$

By the chain rule it follows for arbitrary $p \in W$ and any $\delta u \in H$ that

$$f'(\hat{u})(\delta u) = \mathcal{L}'_u(\hat{u}, \hat{y}, p)(\delta u) + \mathcal{L}'_y(\hat{u}, \hat{y}, p)(\delta y)$$

where $\delta y = S'(\hat{u})(\delta u)$ is the solution of the tangent equation, given by

$$\langle e'_y(\hat{y}, \hat{u})(\delta y), \varphi\rangle = -\langle e'_u(\hat{y}, \hat{u})(\delta u), \varphi\rangle \quad \text{for all } \varphi \in W. \tag{2.43}$$

Now, we choose $\hat{p} \in W$ as the solution of the adjoint equation $\mathcal{L}'_y(\hat{u}, \hat{y}, p) = 0$, which is given by

$$\langle \varphi, e'_y(\hat{y}, \hat{u})^* p\rangle = J'(\hat{y})(\varphi) \quad \text{for all } \varphi \in Y. \tag{2.44}$$

Since $e'_y(\hat{y}, \hat{u})\colon Y \to W^*$ is an isomorphism, the same holds for the transpose $e'_y(\hat{y}, \hat{u})^*\colon W \to Y^*$, and $\hat{p}$ is uniquely determined. This leads to the representation

$$f'(\hat{u})(\varphi) = \mathcal{L}'_u(\hat{u}, \hat{y}, \hat{p})(\varphi) = (e'_u(\hat{y}, \hat{u})(\varphi), \hat{p}) \quad \text{for any } \varphi \in H. \tag{2.45}$$

Therefore it holds $\nabla f(\hat{u}) = e'_u(\hat{y}, \hat{u})^* \hat{p}$. With this, we can again derive the result of Proposition 2.25. To obtain a representation for the second derivative, we first argue that the mapping $(u, y) \mapsto p$, where p is the corresponding solution of (2.44), is continuously differentiable. Again, this follows by the implicit function theorem. By differentiating (2.45) in direction δu with the chain rule, it follows now that

$$\begin{aligned} f''(\hat{u})(\varphi, \delta u) &= \mathcal{L}''_{uu}(\hat{u}, \hat{u}, \hat{p})(\varphi, \delta u) + \mathcal{L}''_{uy}(\hat{u}, \hat{u}, \hat{p})(\varphi, \delta y) + \mathcal{L}''_{up}(\hat{u}, \hat{u}, \hat{p})(\varphi, \delta p) \\ &= \langle e''_{uu}(\hat{y}, \hat{u})(\varphi, \delta u), \hat{p}\rangle + \langle e''_{uy}(\hat{y}, \hat{u})(\varphi, \delta y), \hat{p}\rangle + \langle e'_u(\hat{y}, \hat{u})(\varphi), \delta p\rangle \end{aligned} \tag{2.46}$$

where δy solves the tangent equation (2.43) and δp solves the second adjoint equation

$$\begin{aligned} &\langle \varphi, e'_y(\hat{y}, \hat{u})^* \delta p\rangle \\ &\qquad = J''(\hat{y})(\varphi, \delta u) - \langle e'_{yu}(\hat{y}, \hat{u})(\varphi, \delta u), \hat{p}\rangle - \langle e'_{yy}(\hat{y}, \hat{u})(\varphi, \delta y), \hat{p}\rangle \quad \text{for all } \varphi \in Y. \end{aligned} \tag{2.47}$$

We see that the gradient of f can be derived by the solution of one auxiliary equation. Similarly, a product of the Hessian with any given vector can be computed with the help of two additional equations. This is going to be important for the iterative solution of the linear system of the Newton method described in chapter 3.

The same formulas can also be derived for the discrete versions of the control problems, which will be discussed in chapter 4 and chapter 6 for the elliptic and chapter 5 for the parabolic problem. Then, the computations are based on a discrete version of the Lagrange function. Furthermore, in the parabolic setting we usually perform integration by parts in time to obtain a more convenient interpretation of the expression $e_y'(\hat{y}, \hat{u})^*$; cf. section 2.3.4. For a more detailed exposition and information on the efficient realization in the context of parabolic equations we also refer to [BMV07; Mei08].

3. Algorithmic framework

In this chapter we discuss the theoretical and practical aspects concerning the numerical solution of the optimization problems considered in this thesis. The methods will be based on a reduced cost functional and are not be specific to the convex examples discussed in chapter 2. We employ a reformulation of the optimality system based on the *normal map* (due to Robinson [Rob92]), which is different from the reformulation commonly employed in the infinite dimensional semismooth Newton literature (cf., e.g., [HIK03; Ulb11] or [Sta09; HSW12] specifically for sparse control problems). In the context of variational inequalities and generalized equations in the finite dimensional context, optimization algorithms based on the normal map are well studied: we mention for instance the Newton-Robinson method [Rob94] and the PATH algorithm [Ral94; DF95]. In the context of semismooth Newton methods, reformulations based on the normal map are also known; see, e.g., [Ulb11, p. 9]. A systematic development of an approach with the normal map, specifically for the minimization of the reduced cost functional of problems of the type $(\mathcal{P}_\gamma)$, appears not to have been done yet. In this chapter we will develop a corresponding framework. In many places we can use existing theory: for instance, concerning the locally superlinear convergence of the semismooth Newton method, we can apply the known results by Ulbrich [Ulb02; Ulb11], Hintermüller, Ito, and Kunisch [HIK03], and Schiela [Sch08]. However, differences arise in the structure of the linear system and in the form of the iterates. For instance, in the presence of constraints, the iterates of a method based on the normal map are admissible, whereas the standard approach generally produces infeasible iterates.

This chapter is organized as follows. In section 3.1 we introduce some notation and concepts from convex analysis and introduce the nonsmooth reformulation of the optimality condition based on the normal map. The theoretical framework allows for a very general class of minimization problems consisting of a smooth and a convex part. Section 3.2 introduces some of the necessary theory of semismoothness and semismooth Newton methods that we will need in the following. Here, we essentially follow [Ulb11]. Furthermore, we analyze the structure of the Newton system and show how it can be reduced to a *symmetric* system. Finally, we discuss conditions for the bounded invertibility of the Newton operator. In section 3.3, we discuss the necessary theory of semismoothness of superposition operators that we need for the applications discussed in chapter 2. Here, we essentially follow [Sch08]. Besides, we apply the theory to the concrete functionals ψ introduced in chapter 2. In particular, we verify the assumptions made in the previous sections for the case of sparsity and directional sparsity. We also include the classical case of box-constraints and discuss the case of directional sparsity with positivity constraints. In section 3.4 we give details on the algorithmic realization. An iterative solution strategy based on the method of conjugate gradients for the "symmetrized system" is discussed. Since the semismooth Newton method is in general only locally convergent, we discuss a globalization strategy and prove global convergence for a first order optimization scheme based on the normal map (similar to the projected/proximal gradient method). Motivated by that, we also introduce a heuristic globalization strategy for the semismooth Newton method based on the reduced cost functional and the truncated conjugate gradients approach due to

Steihaug [Ste83]. Here, the reformulation with the normal map is essential. In section 3.6 we systematically compare the normal map formulation to other common formulations. We find that for linear quadratic, convex problems, all considered methods are similar (i.e., equivalent under an appropriate interpretation). For problems with nonlinear state equations we explain the differences that arise in each formulation.

3.1. Nonsmooth reformulation of the optimality condition

Let H be a separable Hilbert space (which is an appropriate L^2-space in all of our applications). We generally abbreviate the inner product in H by $(\cdot,\cdot)$ and the norm by $\|\cdot\|$. We consider the problem

$$\min_{u\in H} \quad j_\gamma(u) = f(u) + \psi(u) + \frac{\gamma}{2}\|u\|^2 \tag{$\mathcal{P}_\gamma$}$$

for a fixed parameter $\gamma \geq 0$. Throughout this chapter we make the following general assumptions.

- $\psi\colon H \to \mathbb{R}\cup\{\infty\}$ is a convex, proper, and lower semicontinuous functional. By $U_{\mathrm{ad}} = \operatorname{dom}\psi = \{\, u\in H \mid \psi(u) < \infty \,\}$ we denote the admissible set.
- $f\colon H \to \mathbb{R}$ is a twice continuously Fréchet-differentiable functional. It is typically of the form $f(u) = J(S(u))$ for a C^2 functional $J\colon Y \to \mathbb{R}$, a Banach space Y, and a C^2 solution operator $S\colon H \to Y$; cf. section 2.5.4.

Remark 3.1. Since the optimization algorithm will only produce admissible controls, we will only ever insert elements $u \in U_{\mathrm{ad}}$ into the functional f. Therefore, it is be possible to consider functionals f that are only defined on a neighborhood of U_{ad}, which is important for some applications. Here, we require f to be defined on all of H to improve readability. Besides, this is the case for the problems from chapter 2.

In this general setting the existence of a minimizer can be proved with classical arguments.

Proposition 3.1. *Let $\gamma > 0$ and f and ψ be bounded from below. Then $(\mathcal{P}_\gamma)$ possesses a minimizer $\bar u \in H$.*

Proposition 3.2. *Let $U_{\mathrm{ad}} = \operatorname{dom}\psi = \{\, u\in H \mid \psi(u) < \infty \,\}$ be a bounded subset of H. Then $(\mathcal{P}_\gamma)$ possesses a minimizer $\bar u \in U_{\mathrm{ad}}$.*

In the following, the case $\gamma = 0$ is included mainly on a formal level. For most of the concrete combinations of f and ψ considered in this thesis, problem $(\mathcal{P}_\gamma)$ is not well posed for $\gamma = 0$, i.e., the optimal solution cannot be found in the Hilbert space H. In the case that the optimal solution lies in H, for instance in the presence of control constraints, it is (in principle) possible to apply the optimization algorithms in this chapter. Most of the convergence analysis will however only be applicable for $\gamma > 0$. For convenience of notation, we abbreviate the smooth part of the functional by

$$f_\gamma(\cdot) = f(\cdot) + \frac{\gamma}{2}\|\cdot\|^2.$$

This implies

$$\nabla f_\gamma(u) = \nabla f(u) + \gamma u \quad \text{for all } u \in U$$

As an optimality condition, we obtain the following standard result; cf. also Proposition 2.25.

Proposition 3.3. *Suppose that $\bar{u}$ is a minimizer of $(\mathcal{P}_\gamma)$. Then we have*

$$-\nabla f_\gamma(u) \in \partial\psi(\bar{u}).$$

Since the subdifferential is not easily accessible algorithmically, we will reformulate the inclusion property from Proposition 3.3 as an equation. For this we introduce the proximal mapping (alternatively called Prox-operator or proximity mapping), which is due to Moreau [Mor65].

Definition 3.1. We define the proximal map $P_c\colon H \to H$ of the convex functional ψ for the constant $c > 0$ as

$$P_c(q) = u = \operatorname*{argmin}_{\tilde{u}\in H} \left[\frac{c}{2}\|\tilde{u} - q\|^2 + \psi(\tilde{u})\right]. \tag{3.1}$$

Remark 3.2. Usually, the proximal map is defined for $c = 1$ and the resulting operator is denoted by Prox_ψ; cf. [BC11, Definition 12.23]. The operator P_c can then be obtained as the proximal map of ψ/c, i.e., we have

$$P_c(q) = \mathrm{Prox}_{\psi/c}(q) = \operatorname*{argmin}_{\tilde{u}\in H} \left[\frac{1}{2}\|\tilde{u} - q\|^2 + \frac{1}{c}\,\psi(\tilde{u})\right].$$

We prefer P_c here, for convenience of notation.

Due to the simple structure of the minimization problem (3.1) it can often be solved analytically (this is the case for all concrete ψ considered in this thesis). Furthermore, its numerical computation can be realized in an efficient way. We recall some of the general properties hereafter.

Proposition 3.4. *The proximal mapping* (3.1) *is a well defined, bounded (nonlinear) operator $P_c\colon H \to U_{\mathrm{ad}} \subset H$. Furthermore:*

(i) For all $q, u \in H$ we have the equivalence

$$u = P_c(q) \;\Leftrightarrow\; c(q - u) \in \partial\psi(u).$$

(ii) P_c is "firmly nonexpansive", i.e., for all $q, \tilde{q} \in H$ we have

$$\|P_c(q) - P_c(\tilde{q})\|^2 \le (P_c(q) - P_c(\tilde{q}), q - \tilde{q}).$$

(iii) P_c is Lipschitz-continuous with constant one, i.e., for all $q, \tilde{q} \in H$ we have

$$\|P_c(q) - P_c(\tilde{q})\| \le \|q - \tilde{q}\|.$$

(iv) P_c is a monotone operator, i.e., for all $q, \tilde{q} \in H$ we have

$$(P_c(q) - P_c(\tilde{q}), q - \tilde{q}) \ge 0.$$

Proof. In the proofs of these standard results, which we sketch for completeness, we mainly follow Bauschke and Combettes [BC11]. First, we see that the minimization problem in (3.1) has a unique solution due to strong convexity of the functional

$$u \mapsto \frac{c}{2}\|u - q\|^2 + \psi(u).$$

Since $U_{\mathrm{ad}} = \operatorname{dom}\psi$ is not empty, the minimizer must lie in U_{ad}. Property (i) follows directly from the equivalent characterization of the minimizer of the convex problem with the subdifferential as in Proposition 3.3; cf. [BC11, Proposition 16.34]. For properties (ii), (iii) and (iv) we take $q, \tilde{q} \in H$ arbitrary. By writing out the subgradient condition (i) for $u = P_c(q)$ and $\tilde{u} = P_c(\tilde{q})$ we have

$$\psi(u) + c(q - u, \tilde{u} - u) \leq \psi(\tilde{u}),$$
$$\psi(\tilde{u}) + c(\tilde{q} - \tilde{u}, u - \tilde{u}) \leq \psi(u).$$

Adding both inequalities, dividing by $c > 0$, and rearranging, we obtain

$$\|\tilde{u} - u\|^2 \leq (q - \tilde{q}, u - \tilde{u}).$$

This is the firm nonexpansiveness property (ii); cf. [BC11, Proposition 12.27]. The remaining two items are direct consequences: with Cauchy-Schwarz on the right-hand side we obtain (iii) and due to positivity of the left-hand side we obtain (iv). □

With the help of the proximal mapping it is possible to reformulate the subdifferential inclusion property from Proposition 3.3 as an equality. We are going to rely on the concept of the *normal map*, which is due to Robinson [Rob92].

Definition 3.2. For any $c > 0$ we define the normal map of ∇f_γ and ψ as

$$G(q) = c(q - P_c(q)) + \nabla f_\gamma(P_c(q)). \tag{3.2}$$

With the help of Proposition 3.4.(i) we see that $G(q) = 0$ implies the stationarity condition from Proposition 3.3 for $u = P_c(q)$. In fact, we obtain the following result.

Proposition 3.5. *Suppose that $\bar{u} \in H$ is an optimal solution of $(\mathcal{P}_\gamma)$. Then there exists an $\bar{q} \in H$, such that $\bar{u} = P_c(\bar{q})$ and $G(\bar{q}) = 0$. Moreover, we have*

$$G(q) = 0 \;\Leftrightarrow\; -\nabla f_\gamma(P_c(q)) \in \partial\psi(P_c(q)).$$

In other words, $G(q) = 0$ with $u = P_c(q)$ is equivalent to the stationarity condition from Proposition 3.3.

Proof. Let $\bar{u}$ be an optimal solution of $(\mathcal{P}_\gamma)$. According to Proposition 3.3, we have $-\nabla f_\gamma(\bar{u}) \in \partial\psi(\bar{u})$. We set

$$\bar{q} = \bar{u} - \frac{1}{c}\nabla f_\gamma(\bar{u}).$$

We obtain $c(\bar{q} - \bar{u}) = -\nabla f_\gamma(\bar{u}) \in \partial\psi(\bar{u})$, which directly implies that $\bar{u} = P_c(\bar{q})$ with Proposition 3.4.(i). By construction, we have that $G(\bar{q}) = 0$. □

We will base the optimization method on finding a zero of the map G. An instructive interpretation of this can be given as follows. Moreau's identity (see, e.g, [BC11, Theorem 14.3]) tells us that we can always decompose a variable $q \in H$ into

$$q = P_c(q) + P_c^*(q) = u + u^*,$$

where $P_c^* = \mathrm{Prox}_{(\psi/c)^*}$ is the proximal map of the convex conjugate $(\psi/c)^* \colon H \to \mathbb{R} \cup \{\infty\}$ defined as

$$(\psi/c)^*(u^*) = \sup_{u \in H}\left[(u^*, u) - \frac{1}{c}\psi(u)\right] = \frac{1}{c}\psi^*(cu^*).$$

By using $q = u + u^*$ as an optimization variable, we combine the primal iterate u and dual iterate u^*. It generates both the iterate u and the current candidate cu^* for the subgradient of ψ. The first part of the stationarity condition, which is given by $cu^* \in \partial\psi(u)$, is always fulfilled according to Proposition 3.4.(i). In the optimum, we additionally obtain with Proposition 3.5 that

$$c\bar{u}^* = c(\bar{q} - \bar{u}) = -\nabla f_\gamma(\bar{u}),$$

which is the second part of the stationarity condition. This is equivalent to $G(\bar{q}) = 0$.

The important special case

For the function space analysis of the following semismooth Newton method we will have to suppose $\gamma > 0$. In this case, we will always choose the parameter $c = \gamma$. By this choice, the term $\gamma P_\gamma(q)$ cancels and we obtain

$$G(q) = \gamma q + \nabla f(P_\gamma(q)). \tag{3.3}$$

In this formulation, we can directly obtain the optimal $\bar{q}$ as a multiple of the gradient of f in the optimum.

Proposition 3.6. *Suppose $\gamma > 0$ and set $c = \gamma$. The optimal variable $\bar{q}$ with $\bar{u} = P_\gamma(\bar{q})$ from Proposition 3.5 is given by*

$$\bar{q} = -\frac{1}{\gamma}\nabla f(\bar{u}). \tag{3.4}$$

In other words, we obtain the "proximal" formula

$$\bar{u} = P_\gamma(\bar{q}) = P_\gamma\left(-\frac{1}{\gamma}\nabla f(\bar{u})\right),$$

giving the optimal control $\bar{u}$ in terms of the proximal map of the gradient of f.

In many applications (as in chapter 2) the gradient of f can be represented in terms of the adjoint state $p = p(u)$ as $\nabla f(u) = B^*p$, where B is a bounded linear operator. In most cases, the adjoint state p will have higher regularity than u, which results in higher regularity for $\bar{q}$ due to (3.4). Moreover, we can interpret the method as operating on the variable $q = -1/\gamma\, B^*p$. This draws a parallel to the "control-reduced approach" used by, e.g., Schiela [Sch08], where the control is eliminated from the optimality system with the projection formula $u = P_\gamma(-1/\gamma\, B^*p)$. We will explore the similarities and differences to other reformulations used for semismooth Newton methods in section 3.6.

3.2. Newton method framework

We will apply a Newton-type method to find a zero of the equation

$$G(q) = 0. \tag{3.5}$$

However, the proximal mapping is not differentiable in general (specifically not in the cases considered in this thesis). Therefore, we work with the concept of semismoothness, which is a generalization of differentiability for certain nonsmooth functions.

3.2.1. Semismoothness calculus

In the context of finite dimensional optimization, a generalized differential for locally Lipschitz continuous functions can be derived from Clarke's generalized Jacobian. We refer the reader to the overview in Qi and Sun [QS99], or Ulbrich [Ulb11, Section 2.1]. The construction makes use of Rademacher's theorem and cannot be transferred easily to the infinite dimensional case. In the abstract Banach space setting, we define semismoothness as a relation between an operator F and another object $\mathrm{D}F$, the candidate for the generalized derivative. In this section we mainly follow Ulbrich [Ulb11, Section 3.2]. Note, that in contrast to [Ulb11] we do not discuss multi-valued $\mathrm{D}F$, mainly to simplify notation. The following approach can be seen as the single-valued special case of the multi-valued approach (and therefore the results from [Ulb11] are applicable in this setting). The motivation for the definition is to provide a minimal requirement that allows to show superlinear convergence of Newton's method.

Definition 3.3. Let V_1, V_2 be Banach spaces and $F\colon V_1 \to V_2$. Furthermore, let $\mathrm{D}F\colon V_1 \to \mathcal{B}(V_1, V_2)$. We say that F is semismooth at the point $v \in V_1$ with respect to $\mathrm{D}F$ (F is $\mathrm{D}F$-semismooth) if we have

$$\lim_{\|\delta v\|_{V_1}\to 0} \frac{1}{\|\delta v\|_{V_1}} \|F(v+\delta v) - F(v) - \mathrm{D}F(v+\delta v)\delta v\|_{V_2} = 0.$$

In this case, we refer to $\mathrm{D}F$ as a generalized derivative/differential for F.

For a semismooth operator equation $F(v) = 0$ with boundedly invertible $\mathrm{D}F(\cdot)$, superlinear convergence of Newton's method can be shown.

Theorem 3.7. *Suppose that we have an operator $F\colon V_1 \to V_2$, a generalized differential $\mathrm{D}F\colon V_1 \to \mathcal{B}(V_1, V_2)$, and a $\bar{v} \in V_1$ with*

- *$F(\bar{v}) = 0$,*
- *F is semismooth at $\bar{v}$ with respect to the generalized differential $\mathrm{D}F$,*
- *the generalized derivatives $\mathrm{D}F(\cdot)$ are boundedly invertible in a neighborhood $\mathcal{N}_1(\bar{v})$ of $\bar{v}$ in V_1 with a uniform bound*

$$\|\mathrm{D}F(v)^{-1}\|_{V_2\to V_1} \le M \quad \text{for all } v \in \mathcal{N}_1(\bar{v}).$$

Then there exists a neighborhood $\mathcal{N}_2(\bar{v})$ of $\bar{v}$ in V_1, such that for all $v_0 \in \mathcal{N}_2(\bar{v})$ the Newton iterates v_n, defined by

$$v_{n+1} = v_n - \mathrm{D}F(v_n)^{-1}F(v_n),$$

converge to $\lim_{n\to\infty} v_n = \bar{v}$. Furthermore, the convergence is superlinear, i.e., we have

$$\|v_{n+1} - \bar{v}\|_{V_1} \le \lambda_n \|v_n - \bar{v}\|_{V_1}$$

for a sequence $0 \le \lambda_n \to 0$ for $n \to \infty$.

Proof. Let us reproduce the proof from Hintermüller, Ito, and Kunisch [HIK03, Theorem 1.1], since it is short and instructive. By definition, the Newton iterates fulfill

$$v_{n+1} - \bar{v} = -\mathrm{D}F(v_n)^{-1}\left[F(v_n) - F(\bar{v}) - \mathrm{D}F(v_n)(v_n - \bar{v})\right].$$

We set $e_n = v_n - \bar{v}$. With the uniform bound on the inverses of $\mathrm{D}F(v_n)$, we obtain

$$\|e_{n+1}\|_{V_1} \leq M\|F(\bar{v} + e_n) - F(\bar{v}) - \mathrm{D}F(\bar{v} + e_n)e_n\|_{V_2}$$

for all $e_n \in \mathcal{N}_1(\bar{v})$. Due to semismoothness of F at $\bar{v}$, by choosing $e_0 = v_0 - \bar{v}$ sufficiently small, there exists a $\rho \in [0, 1)$, such that

$$\|F(\bar{v}) - F(\bar{v} + e) - \mathrm{D}F(\bar{v} + e)e\|_{V_2} \leq \frac{\rho}{M}\|e\|_{V_1}$$

for all $e \in V_1$ with $\|e\|_{V_1} \leq \|e_0\|_{V_1} = \|v_0 - \bar{v}\|_{V_1}$. It follows $\|e_1\|_{V_1} \leq \rho\|e_0\|_{V_1}$. Furthermore, we obtain $\|e_n\|_{V_1} \leq \rho^n\|e_0\|_{V_1}$ by induction. Therefore, we have $v_n \to \bar{v}$ for $n \to \infty$. Using again the semismoothness of F, we can choose the constant ρ arbitrarily small for sufficiently high n, which is the superlinear convergence. □

Remark 3.3. The concept of pointwise semismoothness as given in Definition 3.3 requires some caution. For any given operator F and a given point $\bar{v}$, it is always possible to construct a special $\mathrm{D}F^{\bar{v}}$ such that $F(\bar{v} + \delta v) - F(\bar{v}) - \mathrm{D}F^{\bar{v}}(\bar{v} + \delta v)\delta v = 0$ for all $\delta v \in V_1$ (this would entail termination of a corresponding Newton method after the first step). For algorithmic purposes, we are interested in semismoothness with respect to a generalized differential $\mathrm{D}F$ which can be chosen a priori, without knowledge of the point $\bar{v}$.

Let us derive a suitable generalized differential for G. First, we can directly see that continuously Fréchet-differentiable functions are semismooth (see [Ulb11, Proposition 3.4]).

Proposition 3.8. *Let $F\colon V_1 \to V_2$ be continuously differentiable. Then F is semismooth for all $v \in V_1$ with generalized derivative $\mathrm{D}F = F'$, where F' is the Fréchet derivative.*

Furthermore, we have the following chain rule (see [Ulb11, Proposition 3.8]).

Proposition 3.9. *Let $F_1\colon V_1 \to V_2$ be semismooth at v_1 with generalized differential $\mathrm{D}F_1$ and let $F_2\colon V_2 \to V_3$ be semismooth at $v_2 = F_1(v_1)$ with generalized differential $\mathrm{D}F_2$. Furthermore, let F_1 be Lipschitz continuous at v_1 and let the generalized derivatives $\mathrm{D}F_2$ be uniformly bounded near v_2. Then, the composition $F = F_2 \circ F_1\colon V_1 \to V_3$ is semismooth at v_1 with generalized derivative $\mathrm{D}F$ defined as $\mathrm{D}F(v) = \mathrm{D}F_2(F_1(v))\mathrm{D}F_1(v)$ for $v \in V_1$.*

Our goal is to apply this chain rule with $F_1 = P_c$ and $F_2 = \nabla f$ to construct a generalized differential for $\nabla f \circ P_c$, which appears in the definition of G. The nontrivial aspect is to find a suitable generalized derivative $\mathrm{D}P_c\colon H \to \mathcal{B}(H)$ for P_c. It is well known that, in the infinite dimensional context, we cannot expect semismoothness of P_c, when regarded as an operator from H to H (we will see this in section 3.3.1). In the concrete examples discussed in section 3.3, which are superposition operators, we have to impose a norm gap between the image and preimage space. For now, we formulate this as the following general assumption.

Assumption 3.1. Suppose that there exists a Banach space $H_{\mathrm{sub}} \subset H$, which is continuously embedded in H, and a generalized derivative $\mathrm{D}P_c\colon H \to \mathcal{B}(H)$ such that P_c, when regarded as an operator $H_{\mathrm{sub}} \to H$, is semismooth with respect to $\mathrm{D}P_c$. In other words, we have for all $q \in H_{\mathrm{sub}}$ that

$$\lim_{\|\delta q\|_{H_{\mathrm{sub}}} \to 0} \frac{1}{\|\delta q\|_{H_{\mathrm{sub}}}}\|P_c(q + \delta q) - P_c(q) - \mathrm{D}P_c(q + \delta q)\delta q\| = 0.$$

The construction of $\mathrm{D}P_c$ and the appropriate choice of H_{sub} are discussed in section 3.3.2 in the context of concrete examples. The important aspect of Assumption 3.1 is that we can infer semismoothness of G with respect to the canonical candidate for the generalized derivative.

Proposition 3.10. *Suppose that Assumption 3.1 holds. Then G, when regarded as an operator $G\colon H_{\mathrm{sub}} \to H$ is semismooth for all $q \in H_{\mathrm{sub}}$ with the generalized derivative*

$$\mathrm{D}G(q) = c(\mathrm{Id} - \mathrm{D}P_c(q)) + \nabla^2 f_\gamma(P_c(q))\mathrm{D}P_c(q).$$

Proof. We apply the chain rule from Proposition 3.9 to $\nabla f_\gamma \circ P_c$. For that, recall that P_c is Lipschitz continuous and the derivatives $\nabla^2 f(\cdot)$ are uniformly bounded near $u = P_c(q)$ due to continuous Fréchet differentiability. Furthermore, it is clear from the definition that sums of semismooth functions are semismooth with respect to the canonical generalized derivative. □

To be able to apply Theorem 3.7 we would have to suppose that $\mathrm{D}G(\cdot)^{-1}\colon H \to H_{\mathrm{sub}}$, which is an unrealistic assumption for any nontrivial subspace $H_{\mathrm{sub}} \subset H$, as we will see in section 3.2.2. In fact, it is clearly violated in most cases. Consider, e.g., the trivial case $f, \psi \equiv 0$, where we obtain $\mathrm{D}G(\cdot) \equiv \gamma\,\mathrm{Id}$.

The important special case

For the superlinear convergence proof it will be essential to require $\gamma > 0$. Consequently, we choose $c = \gamma$. Recall that now, G simplifies to

$$G(q) = \gamma q + \nabla f(P_\gamma(q)).$$

Since P_γ appears here only behind ∇f, we can get rid of the norm gap for the semismoothness property of G by imposing a smoothing condition on ∇f.

Assumption 3.2. Let $H_{\mathrm{sub}} \subset H$ be the subspace from Assumption 3.1. We assume that $\nabla f(H) \subset H_{\mathrm{sub}}$ and that ∇f is also continuously Fréchet-differentiable as an operator $\nabla f\colon H \to H_{\mathrm{sub}}$.

Proposition 3.11. *Suppose that $c = \gamma > 0$ and that Assumptions 3.1 and 3.2 hold. Then G can be regarded as an operator $G\colon H_{\mathrm{sub}} \to H_{\mathrm{sub}}$, which is semismooth for all $q \in H_{\mathrm{sub}}$ with the generalized derivative*

$$\mathrm{D}G(q) = \gamma\,\mathrm{Id} + \nabla^2 f(P_\gamma(q))\mathrm{D}P_\gamma(q).$$

Proof. We apply again Proposition 3.9 to $F_1 = P_\gamma\colon H_{\mathrm{sub}} \to H$ and $F_2 = \nabla f\colon H \to H_{\mathrm{sub}}$. □

To apply Theorem 3.7, we need to discuss invertibility of the Newton matrices $\mathrm{D}G(\cdot)$, which we will address in the next section.

3.2.2. Newton system and quadratic model

In the following we will consider the Newton update equation based on the reformulation (3.2), i.e., the solvability of the linear equation

$$\mathrm{D}G(q)\,\delta q = -G(q) \tag{3.6}$$

for the Newton update δq. The operator $\mathrm{D}G(q)$ is generally not symmetric, which is typical for semismooth Newton methods. However, solving the linear equation (3.6) can be reduced to the solution of a symmetric system. To this purpose we need to introduce some notation and concepts.

First, we are going to formulate some additional hypotheses on the generalized differential $\mathrm{D}P_c$ of the proximal map P_c, which will be fulfilled for all concrete examples; see section 3.3.2.

Assumption 3.3. For all $q \in H$ the generalized derivative $\mathrm{D}P_c(q)\colon H \to H$ has the properties:

(i) $\mathrm{D}P_c(q)$ is a bounded operator on H with $\|\mathrm{D}P_c(q)\|_{H\to H} \leq 1$.

(ii) $\mathrm{D}P_c(q)$ is a positive semidefinite operator, i.e., $(\mathrm{D}P_c(q)\delta q, \delta q) \geq 0$ for all $\delta q \in H$.

(iii) $\mathrm{D}P_c(q)$ is a self adjoint operator, i.e., $\mathrm{D}P_c(q)^* = \mathrm{D}P_c(q)$.

Let us give some motivation for these assumptions. Recall that P_c is monotone and Lipschitz continuous with constant one; see Proposition 3.4. If the directional derivative of P_c at $q \in H$ in direction δq exists, we obtain directly that $(\mathrm{d}P_c(q;\delta q), \delta q) \geq 0$ and $\|\mathrm{d}P_c(q;\delta q)\| \leq \|\delta q\|$. Moreover, the generalized derivative $\mathrm{D}P_c$ is related to the second derivatives of the convex functional ψ. Consider (for simplicity) the case where ψ is twice differentiable at the point $u = P_c(q)$. With the implicit function theorem it follows that P_c is differentiable at q and we have the identity $\nabla^2\psi(u) = c(\nabla P_c(q)^{-1} - \mathrm{Id})$. This implies that $\nabla P_c(q)$ is a symmetric operator. We will not go into further detail here. In the following, Assumption 3.3 is regarded as a restriction on the choice of the generalized differential, which will be verified for each of the concrete examples.

Remark 3.4. For the semismoothness concept in finite dimensions, where the canonical candidate for the generalized differential is derived from Clarke's generalized Jacobian in a systematic way, all of these assumptions can be proven as theorems; see Milzarek [Mil15]. There, the generalized derivative of the proximal map can be related to a generalized Hessian; see [HUSN84] for a definition. Specifically, if P_c is differentiable at the point q it holds $\nabla P_c(q) = q - 1/c\,\nabla^2\Psi_c(q)$, where $\Psi_c(q) = c/2\|P_c(q) - q\|^2 + \psi(P_c(q))$ is the Moreau envelope of ψ. This follows with the help of the well-known formula $\nabla\Psi_c(q) = c(q - P_c(q))$; see, e.g., [BC11, Proposition 12.29]. Since the generalized differential $\mathrm{D}P_c$ is derived from ∇P_c (which exists almost everywhere), the properties from Assumption 3.3 can be shown. We remark that, if a generalized differential for a superposition operator is derived from the finite dimensional one via superposition, the properties from Assumption 3.3 transfer, as we will see in section 3.3.

We briefly sketch the idea behind the computation of the Newton update before giving a detailed rigorous argument below. Therefore, we fix a $q \in H$ and abbreviate

$$T = \mathrm{D}P_c(q)$$

for convenience of notation. Furthermore we denote $u = P_c(q)$. As discussed in the previous section we consider for (3.2) the generalized derivative given by

$$\begin{aligned}\mathrm{D}G(q) &= c\,(\mathrm{Id} - T) + \nabla^2 f_\gamma(u)T \\ &= c\,\mathrm{Id} + \left(\nabla^2 f_\gamma(u) - c\,\mathrm{Id}\right)T.\end{aligned} \tag{3.7}$$

Note, that $\mathrm{D}G(q)$ is in general *not* a symmetric operator. However, if we multiply the Newton system (3.6) by the self-adjoint operator $T = \mathrm{D}P_c(q)$ from the left, we obtain the system

$$T\mathrm{D}G(q)\,\delta\tilde{q} = -TG(q), \tag{3.8}$$

where the symmetric operator on the left-hand side is given by

$$T\mathrm{D}G(q) = c\left(T - T^2\right) + T\nabla^2 f_\gamma(u)T.$$

Note, that the equation (3.8) corresponds to the stationarity condition for the quadratic problem

$$\begin{aligned}\min_{v\in H_T} Q_q(v) &= (TG(q), v) + \frac{1}{2}(v, T\mathrm{D}G(q)v)\\ &= (G(q), Tv) + \frac{1}{2}(Tv, \nabla^2 f_\gamma(u)Tv) + \frac{c}{2}(Tv, (\mathrm{Id} - T)v).\end{aligned} \tag{3.9}$$

We will see that, under some conditions to be specified below, we can solve (3.8) for a $\delta\tilde{q}$ in an appropriate space H_T (which still guarantees $T\delta\tilde{q} \in H$). We then observe that it holds $T\delta\tilde{q} = T\delta q$. By taking another look at the full equation (3.6) we derive the representation for the full Newton step δq as

$$\delta q = -\frac{1}{c}\left[\left(\nabla^2 f_\gamma(u) - c\,\mathrm{Id}\right)T\delta\tilde{q} + G(q)\right]. \tag{3.10}$$

Let us make the meaning of (3.8) and (3.10) precise in the following. To discuss solvability for $\delta\tilde{q}$, we now introduce the previously mentioned solution space $H_T = H_{\mathrm{D}P_c(q)}$. Let us first recall that for the bounded, self-adjoint operator T we have the equalities

$$\operatorname{Ker} T = (\operatorname{Ran} T)^\perp \quad \text{and} \quad \overline{\operatorname{Ran} T} = (\operatorname{Ker} T)^\perp,$$

linking the range and the kernel of T. In general, $T = \mathrm{D}P_c(q)$ will have a nontrivial kernel and its range will not be closed; cf. the discussion in the context of concrete examples in section 3.3.2. We proceed to define the space H_T as the Hilbert space induced by the inner product derived from the symmetric operator T.

Definition 3.4. Define the symmetric and positive semi-definite form $(\cdot,\cdot)_T = (\cdot, T\,\cdot)$ and the associated seminorm $\|\cdot\|_T = \sqrt{(\cdot,\cdot)_T}$. The space H_T is given as

$$H_T = \overline{\left({}^{H}\!/_{\operatorname{Ker} T}\right)}^{\|\cdot\|_T},$$

which is the closure of the quotient space ${}^{H}\!/_{\operatorname{Ker} T}$ w.r.t. the T-norm.

Proposition 3.12. *The bilinear form $(\cdot,\cdot)_T$, extended in the canonical way to the quotient space ${}^{H}\!/_{\operatorname{Ker} T}$, is symmetric and positive definite. $\|\cdot\|_T$ is a norm on ${}^{H}\!/_{\operatorname{Ker} T}$. Consequently, H_T, endowed with the inner product $(\cdot,\cdot)_T$, is a Hilbert space.*

Proposition 3.13. *The operator $T\colon H \to H$ extends in a natural way to an operator $T\colon H_T \to H$ (denoted with the same symbol), such that*

$$\|T\|_{H_T\to H} \le 1 \quad \textit{and} \quad T(v + \operatorname{Ker} T) = Tv \quad \textit{for all } v \in H.$$

The elementary proofs of these results are given in Appendix A.2.

By Proposition 3.13, the Newton operator $\mathrm{D}G(q)$ can be also considered as an operator from H_T to itself, which is now *self-adjoint*. The same holds if we regard it as an operator on the space of equivalence classes ${}^{H}\!/_{\operatorname{Ker} T}$ or the orthogonal complement $(\operatorname{Ker} T)^\perp$. The proof of invertibility and the practical solution strategy will be based on this observation. Next, we discuss conditions under which the operator $\mathrm{D}G(q)$ is invertible.

3.2.3. Invertibility of the Newton operator

With these technical prerequisites we can discuss solvability of the Newton equation. We consider $\mathrm{D}G(q)$ as given in (3.7), abbreviate $T = \mathrm{D}P_c(q)$, and suppose that Assumption 3.3 holds for all $q \in H$. Note that a uniform bound on the norm of the inverses $\mathrm{D}G(\cdot)^{-1}$ is needed to show convergence of the semismooth Newton method with Theorem 3.7. Therefore, it is important to explicitly mark the dependency on the point $q \in H$ in the following estimates.

The convex case

Let us assume first that f is given by

$$f(u) = \frac{1}{2}\|Su - y_d\|_Y^2$$

for some linear bounded operator $S\colon H \to Y$, mapping from H to the Hilbert space Y. The Hessian of f is therefore given as $\nabla^2 f(u) = S^*S$ for all $u \in H$, where $S^*\colon Y \to H$ is the Hilbert-space adjoint of S. More generally, we can consider

$$f(u) = J(Su)$$

for a convex C^2 functional $J\colon Y \to \mathbb{R}$. Then, the Hessian is given by $\nabla^2 f(u) = S^*J''(Su)S$ for all $u \in H$. The important observation is that in this case, the functional f is convex and therefore the Hessian $\nabla^2 f(\cdot)$ is positive semidefinite. If we suppose additionally that $\gamma > 0$, the Hessian $\nabla^2 f_\gamma(\cdot) = \gamma + \nabla^2 f(\cdot)$ is even positive definite. Under these conditions, the Newton operator (3.7) is invertible for any $q \in H$. To make the invertibility useful in the context of Theorem 3.7 and Proposition 3.11 we need to work with the subspace H_{sub} from Assumption 3.2. Since ∇f is assumed to be Fréchet differentiable as an operator from H to H_{sub}, the Hessian of f has a smoothing property, i.e., it can be regarded as an operator

$$\nabla^2 f(u)\colon H \to H_{\mathrm{sub}}.$$

With this observation, we obtain the following result.

Lemma 3.14. *Assume that f is convex and that $c = \gamma > 0$. Furthermore, suppose that Assumption 3.2 holds. Then, for all $q \in H$ the Newton operator $\mathrm{D}G(q)\colon H \to H$ as given in* (3.7) *is boundedly invertible. Furthermore, we have $\mathrm{D}G(q)^{-1}(H_{\mathrm{sub}}) \subset H_{\mathrm{sub}}$ with the estimate*

$$\|\mathrm{D}G(q)^{-1}\|_{H_{\mathrm{sub}}\to H_{\mathrm{sub}}} \leq \frac{1}{\gamma}\left(1 + \frac{1}{\gamma}\|\nabla^2 f(u)\|_{H\to H_{\mathrm{sub}}}\right), \tag{3.11}$$

where $u = P_\gamma(q)$.

Proof. We consider the Newton equation $\mathrm{D}G(q)v = r$ for a general right hand side $r \in H$. Since we have $c = \gamma$, the Newton operator has the structure

$$\mathrm{D}G(q) = \gamma\,\mathrm{Id} + \nabla^2 f(u)T.$$

The auxiliary step is determined by solving

$$T\mathrm{D}G(q)\tilde{v} = \left[\gamma T + T\nabla^2 f(u)T\right]\tilde{v} = Tr. \tag{3.12}$$

Again, this is equivalent to the minimization of the quadratic functional (3.9) or the variational formulation $(\cdot, \mathrm{D}G(q)\tilde{v})_T = (\cdot, r)_T$. Now, with respect to the Hilbert space H_T, we clearly have continuity and coercivity of the left hand side, i.e., it holds

$$\gamma\|w\|_T^2 \leq \gamma(w,w)_T + (Tw, \nabla^2 f(u)Tw) = (w, \mathrm{D}G(q)w)_T \quad \text{for all } w \in H,$$

which is a consequence of the convexity of f. Furthermore the right-hand side is continuous, i.e., we have

$$(r,w)_T \leq \|r\|_T\|w\|_T \leq \|r\|\|w\|_T \quad \text{for all } w \in H.$$

Therefore, by the Riesz representation theorem, equation (3.12) admits a unique solution $\tilde{v} \in H_T$ with the estimate $\|\tilde{v}\|_T \leq \|r\|/\gamma$. To obtain the full solution of $\mathrm{D}G(q)v = r$, we set

$$v = \frac{1}{\gamma}\left(r - \nabla^2 f(u)T\,\tilde{v}\right). \tag{3.13}$$

Here we have used that $T\tilde{v} \in H$ with Proposition 3.13. By reordering equation (3.12) we see directly that $T\tilde{v} = 1/\gamma\,(Tr - T\nabla^2 f(u)T\tilde{v})$. Comparing this with (3.13) immediately shows $Tv = T\tilde{v}$. Therefore v solves $\mathrm{D}G(q)v = r$. Supposing additionally that $r \in H_{\mathrm{sub}}$, we directly obtain $v \in H_{\mathrm{sub}}$ from (3.13). The corresponding estimate (3.11) is obvious. □

Remark 3.5. i) If we suppose that f is quadratic, the Hessian $\nabla^2 f$ is independent of the point u. Consequently, the bound (3.11) on the inverse of $\mathrm{D}G(q)$ is independent of q.

ii) If we omit in Lemma 3.14 the requirement $c = \gamma$ and the smoothing property, we can still show invertibility of $\mathrm{D}G(q)$ in the sense of an operator on H. However, such a result will not be sufficient for the analysis of the semismooth Newton method, in general.

The nonconvex case

For general functionals f we cannot expect $\nabla^2 f(u)$ to be a positive semidefinite operator, which we used in Lemma 3.14 in a central way. In this case, we impose an a priori assumption on the coercivity of $\nabla^2 f_\gamma(u) = \gamma + \nabla^2 f(u)$. We will formulate an assumption that allows us to directly carry over the result of Lemma 3.14.

Lemma 3.15. *Suppose that $c = \gamma > 0$ and that Assumption 3.2 holds. Furthermore assume that for a specific $q \in H$ and the corresponding $u = P_\gamma(q)$ there exists a constant $\nu > 0$ such that*

$$(v, DG(q)v)_T = \gamma(v,v)_T + (v, \nabla^2 f(u)Tv)_T \geq \nu(v,v)_T \quad \textit{for all } v \in H. \tag{3.14}$$

Then the Newton operator $\mathrm{D}G(q)$ as given in (3.7) is boundedly invertible. Furthermore, we have $\mathrm{D}G(q)^{-1}(H_{\mathrm{sub}}) \subset H_{\mathrm{sub}}$ with the estimate

$$\|\mathrm{D}G(q)^{-1}\|_{H_{\mathrm{sub}} \to H_{\mathrm{sub}}} \leq \frac{1}{\gamma}\left(1 + \frac{1}{\nu}\|\nabla^2 f(u)\|_{H \to H_{\mathrm{sub}}}\right). \tag{3.15}$$

To ensure that this result is applicable in the context of Theorem 3.7, the constant ν in Lemma 3.15 needs to be bounded independently of the point q. To guarantee this for a neighborhood of a stationary point, we can, for instance, require a stronger second order condition in this point.

Proposition 3.16. *Suppose that $c = \gamma > 0$ and that Assumption 3.2 holds. Assume that for a $\bar{q} \in H$ and the corresponding $\bar{u} = P_\gamma(\bar{q})$ there exists a constant $0 < \bar{\nu} \leq \gamma$ such that*

$$(v, \nabla^2 f_\gamma(\bar{u})v) \geq \bar{\nu}(v, v) \quad \text{for all } v \in H.$$

Then there exists a neighborhood $\mathcal{N}(\bar{q})$ in H such that for all $q \in \mathcal{N}(\bar{q})$ the property (3.14) *holds with $\nu = \bar{\nu}/2$. Consequently, the result of Lemma 3.15 holds with*

$$\|\mathrm{D}G(q)^{-1}\|_{H_{\mathrm{sub}} \to H_{\mathrm{sub}}} \leq \frac{1}{\gamma}\left(1 + \frac{2}{\hat{\nu}} \sup_{\tilde{q} \in \mathcal{N}(\bar{q})} \|\nabla^2 f(P_\gamma(\tilde{q}))\|_{H \to H_{\mathrm{sub}}}\right)$$

for all $q \in \mathcal{N}(\bar{q})$.

Proof. We need to verify (3.14). By continuity of $\nabla^2 f_\gamma(\cdot)$, we find an open ball $\mathcal{N}(\bar{u})$ in H around $\bar{u}$, such that for any $u \in \mathcal{N}(\bar{u})$ it holds

$$(v, \nabla^2 f_\gamma(u)v) = (v, \nabla^2 f(u)v) + \gamma(v, v) \geq \frac{\bar{\nu}}{2}(v, v) \quad \text{for all } v \in H.$$

We define the neighborhood of $\bar{q}$ as $\mathcal{N}(\bar{q}) = \{\, \bar{q} + (u - \bar{u}) \mid u \in \mathcal{N}(\bar{u}) \,\}$. Using the Lipschitz continuity (with constant one) of P_γ, it is easy to verify that $P_\gamma(\mathcal{N}(\bar{q})) \subset \mathcal{N}(\bar{u})$. For any $q \in \mathcal{N}(\bar{q})$, we compute

$$\begin{aligned}(v, DG(q)v)_T &= \gamma(v, v)_T + (v, \nabla^2 f(u)Tv)_T = \gamma(v, v)_T + (Tv, \nabla^2 f(u)Tv) \\ &\geq \gamma(v, v)_T - \gamma(Tv, Tv) + \frac{\bar{\nu}}{2}(Tv, Tv) = \frac{\bar{\nu}}{2}(v, v)_T + \left(\gamma - \frac{\bar{\nu}}{2}\right)((v, v)_T - (Tv, Tv))\end{aligned}$$

for all $v \in H$, using the uniform coercivity of $\nabla^2 f_\gamma(\cdot)$ from before. Since T is a symmetric, positive semidefinite operator with norm bound one we have $(Tv, Tv) \leq (v, v)_T$ for all $v \in H$. Together with $\gamma > \bar{\nu}/2$, the last term in the previous estimate is positive and we conclude the proof. □

Remark 3.6. The sufficient condition from Proposition 3.16 is *not necessary* for the result of Lemma 3.15 (for instance, coercivity of $\nabla^2 f_\gamma(\bar{u})$ is also required to hold on the kernel of T). It would we desirable to obtain second order sufficient conditions in the optimal solutions which are as close as possible to verifiable second order necessary conditions (see, e.g., [CHW12b] for a sparse control problem) and to derive the coercivity conditions (3.14) from them. For a control constrained problem, such an analysis can be found in [Ulb11, Section 4.3].

3.3. Superposition operators

In many examples (specifically, in all of the examples considered in this thesis), the variable u can be understood as a vector valued function $u\colon \Omega \to \hat{H}$ for a bounded domain Ω and a separable Hilbert space $\hat{H}$, and the convex functional ψ can be written in the form

$$\psi(u) = \int_\Omega \hat{\psi}(u(x))\,\mathrm{d}x \tag{3.16}$$

with a convex, proper and lower semicontinuous functional $\hat{\psi}\colon \hat{H} \to \mathbb{R} \cup \{\, \infty \,\}$. Consequently, the Hilbert space H is chosen as $H = L^2(\Omega, \hat{H})$ and we have

$$(u, v) = \int_\Omega (u(x), v(x))_{\hat{H}}\,\mathrm{d}x, \qquad \|u\|^2 = \int_\Omega \|u(x)\|_{\hat{H}}^2\,\mathrm{d}x$$

for all $v, u \in H$. The integration in (3.16) is to be understood with respect to the Lebesgue-Bochner integral for the vector-valued function u. In fact, ψ as in (3.16) is well defined, since $|\Omega|$ is finite (see [BC11, Proposition 9.32]). In this case, the computation of the proximal map of ψ can be reduced to the computation of the proximal map of $\hat{\psi}$ in $\hat{H}$.

Proposition 3.17. *The proximal map of the functional ψ is given by the pointwise superposition operator*

$$P_c(q)(x) = \hat{P}_c(q(x)) \quad \text{almost everywhere,} \tag{3.17}$$

where $\hat{P}_c \colon \hat{H} \to \hat{H}$ is the proximal map of $\hat{\psi}$ in $\hat{H}$.

Proof. According to the definition, we have to minimize the functional

$$u \mapsto \int_\Omega \frac{c}{2}\|u(x) - q(x)\|_{\hat{H}}^2 + \hat{\psi}(u(x))\,\mathrm{d}x$$

to find $u = P_c(q)$. It is clear that this is equivalent to minimizing the expression under the integral in a pointwise fashion, which leads to (3.17), since

$$\operatorname*{argmin}_{\hat{u}\in\hat{H}} \left[\frac{c}{2}\|\hat{u} - q(x)\|_{\hat{H}}^2 + \hat{\psi}(\hat{u})\right] = \hat{P}_c(q(x)). \qquad \square$$

More generally, we can also consider the case where $\hat{\psi}$ additionally depends on $x \in \Omega$ where ψ is given as

$$\psi(u) = \int_\Omega \hat{\psi}(x, u(x))\,\mathrm{d}x$$

However, in the extended real valued setting, the question whether for a given $\hat{\psi} \colon \Omega \times \hat{H} \to \mathbb{R} \cup \{+\infty\}$ the functional ψ as above is well-defined, convex, proper, and lower semicontinuous is more delicate. On this issue, we refer for instance to Rockafellar [Roc68; Roc71] and the references therein. Since for most of the problems under consideration here $\hat{\psi}$ is independent of x, we do not go into further detail. We only mention that for the case of box-constraints (with measurable constraints) or for a weighted L^1 norm as in section 2.2.3 the questions above can be answered directly and we can derive an analogous result as in Proposition 3.17.

3.3.1. Semismoothness of superposition operators

In this section, we will describe how semismoothness of proximal maps represented by pointwise superposition operators can be reduced to semismoothness of the underlying pointwise proximal map $\hat{P}_c$. Here, we can apply known general results; see Ulbrich [Ulb11, Section 3.3.3] or Schiela [Sch08]. For completeness, we are going to reproduce the subset of the theory given in [Sch08] that we need for the discussion of the concrete examples.

Let us first consider a general problem setting for a superposition operator F, induced by the pointwise operator $\hat{F}$. Later, we are going to apply the results with $\hat{F} = \hat{P}_c$ and $F = P_c$. However, for the next results, the special construction of $\hat{P}_c$ as a proximal mapping on a Hilbert space will not be of particular importance. Assume that for two given separable Banach spaces V_1, V_2, where V_1 is continuously embedded into V_2, the operator $\hat{F}$ is given as an operator $\hat{F} \colon \Omega \times V_1 \to V_2$, which is measurable in the first argument and continuous in the second (i.e., $\hat{F}$ is a Carathéodory function). Now, we define the superposition operator (or Nemyckii-operator) F by

$$F(q)(x) = \hat{F}(x, q(x)).$$

Due to the Carathéodory property, F maps measurable functions from Ω to V_1 to measurable functions from Ω to V_2. Furthermore, we assume that we have a (pointwise) generalized differential $\mathrm{D}\hat{F}\colon \Omega \times V_1 \to \mathcal{B}(V_1, V_2)$. To construct a generalized differential for F, we will consider the superposition

$$\mathrm{D}F(q)(x) = \mathrm{D}\hat{F}(x, q(x)).$$

We have to explain why $\mathrm{D}F$ maps measurable functions to measurable functions. Since we cannot assume $\mathrm{D}\hat{F}$ to be continuous in the second argument (as will become obvious in section 3.3.2), we follow [Sch08] and require $\mathrm{D}\hat{F}$ to be a Baire-Carathéodory function (cf. [AZ08]), i.e., a function which can be represented as a pointwise limit of Carathéodory functions. For this, we recall that $\mathrm{D}\hat{F}$ can be alternatively considered as a function of three arguments

$$\mathrm{D}\hat{F}\colon \Omega \times V_1 \times V_1 \to V_2, \qquad (x, \hat{q}, \delta\hat{q}) \mapsto \mathrm{D}\hat{F}(x, \hat{q})\delta\hat{q},$$

which is linear in the third argument. As motivated before, we now require that $\mathrm{D}\hat{F}(x, \hat{q}, \delta\hat{q}) = \lim_{k\to\infty} \mathrm{D}\hat{F}^k(x, \hat{q}, \delta\hat{q})$ for all $\hat{q}, \delta\hat{q} \in V_1$ and almost all $x \in \Omega$, where the $\mathrm{D}\hat{F}^k$ are measurable in the first and continuous in the second and third argument. Since pointwise limits of measurable functions are measurable, this property guarantees that the superposition of $\mathrm{D}F$ maps measurable to measurable functions. Now, we fix the general assumptions on $\mathrm{D}\hat{F}$ for the rest of this section.

Assumption 3.4. Assume that $\hat{F}$ is Carathéodory and uniformly Lipschitz continuous in the second argument. Assume further that there exists a $\mathrm{D}\hat{F}\colon \Omega \times V_1 \to \mathcal{B}(V_1, V_2)$ which is Baire-Carathéodory with the following properties:

- The values $\mathrm{D}\hat{F}(x, \hat{q})$ are uniformly bounded in $\mathcal{B}(V_1, V_2)$ for all $\hat{q} \in V_1$ and $x \in \Omega$.
- $\hat{F}(x, \cdot)\colon V_1 \to V_2$ is semismooth w.r.t. $\mathrm{D}\hat{F}(x, \cdot)$ for all $x \in \Omega$.

Due to Lipschitz continuity of $\hat{F}$ in the second argument, the superposition operator

$$F\colon L^p(\Omega, V_1) \to L^r(\Omega, V_2), \qquad F(q)(x) = \hat{F}(x, q(x))$$

is well defined for $1 \le r \le p \le \infty$. For this, we verify that $\hat{F}$ fullfills the growth bound $\|\hat{F}(x, \hat{q})\|_{V_2} \le c_1 + c_2\|\hat{q}\|_{V_1}$ for $x \in \Omega$. The goal is to show semismoothness of F w.r.t. the generalized differential given by the superposition operator

$$\mathrm{D}F\colon L^p(\Omega, V_1) \to \mathcal{B}(L^p(\Omega, V_1), L^r(\Omega, V_2)), \quad \mathrm{D}F(q)(x) = \mathrm{D}\hat{F}(x, q(x))$$

for any $r < p$. Note that $\mathrm{D}F$ is well defined for every $r \le p$ due to $\|\mathrm{D}\hat{F}(x, \hat{q})\delta\hat{q}\|_{V_2} \le c_3\|\delta\hat{q}\|_{V_1}$ for $x \in \Omega$. We start by defining the following function.

Definition 3.5. Fix a $q^* \in L^p(\Omega, V_1)$. We define the "pointwise Newton residual" at q^* as

$$\hat{R}^*(x, \hat{q}) = \begin{cases} \frac{1}{\|\hat{q} - q^*(x)\|_{V_1}} \left[\hat{F}(\hat{q}) - \hat{F}(q^*(x)) - \mathrm{D}\hat{F}(\hat{q})(\hat{q} - q^*(x))\right] & \text{for } \hat{q} \neq q^*(x), \\ 0 & \text{else} \end{cases}$$

for $\hat{q} \in V_1$ and $x \in \Omega$ almost everywhere.

We can verify the following properties of $\hat{R}^*$.

Proposition 3.18. *Suppose that Assumption 3.4 holds and let $q^* \in L^p(\Omega, V_1)$ arbitrary.*

(i) $\hat{R}^*(x,\cdot)$ *is continuous at* $\hat{q} = q^*(x)$ *for almost all* $x \in \Omega$.

(ii) The superposition operator induced by $\hat{R}^*$*, given by*

$$R^*(q)(x) = \hat{R}^*(x, q(x)), \qquad L^p(\Omega, V_1) \to L^s(\Omega, V_2)$$

is well defined for any $1 \leq s \leq \infty$.

Proof. Property (i) follows directly from the semismoothness of $\hat{F}$ w.r.t. $\mathrm{D}\hat{F}$ as in Assumption 3.4. By the Carathéodory and Baire-Carathéodory assumptions on $\hat{F}$ and $\mathrm{D}\hat{F}$, measurability of $R^*(q)$ for measurable q is clear (since quotients of measurable functions are measurable). For the mapping property (ii), we further combine the uniform Lipschitz continuity of $\hat{F}$ and the boundedness of $\mathrm{D}\hat{F}$ from Assumption 3.4 to obtain

$$\|\hat{F}(q(x)) - \hat{F}(q^*(x))\|_{V_2} + \|\mathrm{D}\hat{F}(q(x))(q(x) - q^*(x))\|_{V_2} \leq (c_2 + c_3)\|q(x) - q^*(x)\|_{V_1}$$

for any $q \in L^p(\Omega, V_1)$ and $x \in \Omega$ (almost everywhere). Thus, we have verified (ii) for $s = \infty$. The case $s < \infty$ is a direct consequence of Hölder's inequality. □

The semismoothness of F w.r.t. $\mathrm{D}F$ will be derived with the help of a norm-continuity result of R^*. Below, we give the relevant special case of Lemma 3.1 in [Sch08].

Lemma 3.19 ([Sch08, Lemma 3.1])**.** *Under Assumption 3.4, the superposition operator* R^*

$$R^* \colon L^p(\Omega, V_1) \to L^s(\Omega, V_2)$$

is norm-continuous at the point $q = q^*$ *for any* $s < \infty$.

Proof. We give the proof, since it is elementary in this specific situation. Recall that $R^*(q^*) = 0$ by definition. Assume that $q_n \to q^*$ in $L^p(\Omega, V_1)$ for $n \to \infty$. We define the functions $r_n \in L^\infty(\Omega)$ as

$$r_n(x) = \|R^*(q_n)(x)\|_{V_2}^s.$$

By Proposition 3.18.(i), r_n converges to zero pointwise almost everywhere in Ω. Furthermore, it is positive and bounded by $(c_2 + c_3)^s$ as in the proof of Proposition 3.18. Now, we apply Lebesgue's dominated convergence theorem to see that $\|R^*(q_n)(x)\|_{L^s(\Omega,V_2)} \to 0$. □

It is noteworthy, that the case $s = \infty$ is not included in Lemma 3.19. This is the main reason for the so called "norm gap" in the following theorem (see also [Ulb11, Theorem 3.49]).

Theorem 3.20 ([Sch08, Theorem 3.3])**.** *Under Assumption 3.4, for any* $q^* \in L^p(\Omega, V_1)$ *the operator* $F \colon L^p(\Omega, V_1) \to L^r(\Omega, V_2)$ *is semismooth at* q^* *w.r.t.* $\mathrm{D}F$ *for any* $1 \leq r < p \leq \infty$.

Proof. We include the proof for the sake of completeness. By construction, it holds that

$$F(q)(x) - F(q^*)(x) - \mathrm{D}F(q)(x)\,(q(x) - q^*(x)) = R^*(q)(x)\,\|q(x) - q^*(x)\|_{V_1}$$

for $x \in \Omega$ almost everywhere. Taking the $L^r(\Omega, V_2)$ norm, we obtain with Hölder's inequality on the right-hand side that

$$\begin{aligned}\|F(q) - F(q^*) - \mathrm{D}F(q)(q - q^*)\|_{L^r(\Omega,V_2)} &= \left(\int_\Omega \|R^*(q)(x)\|_{V_2}^r \|q(x) - q^*(x)\|_{V_1}^r \,\mathrm{d}x\right)^{1/r} \\ &\leq \|R^*(q)\|_{L^s(\Omega,V_2)} \|q - q^*\|_{L^p(\Omega,V_1)}\end{aligned}$$

for $1 < s < \infty$ given by $1/r = 1/s + 1/p$. Since $\|R^*(q)\|_{L^s(\Omega,V_2)} \to 0$ for $\|q - q^*\|_{L^p(\Omega,V_1)} \to 0$ with Lemma 3.19, we obtain with $\delta q = q - q^*$ that

$$\frac{1}{\|\delta q\|_{L^p(\Omega,V_1)}} \|F(q^* + \delta q) - F(q^*) - \mathrm{D}F(q^* + \delta q)\delta q\|_{L^r(\Omega,V_2)} \to 0,$$

for $\delta q \to 0$, which concludes the proof. □

In the following we will analyze several proximal maps with the help of this theorem and we will have to verify Assumption 3.4 in each concrete setting. Note, that global Lipschitz continuity is always fulfilled for proximal maps; cf. Proposition 3.4. Uniform boundedness of the values of $\mathrm{D}\hat{P}_c(\cdot)$ is also natural in this setting; cf. Assumption 3.3.(i). The only nontrivial part will be to verify the pointwise semismoothness, which then enables us to show semismoothness of

$$P_c \colon H_{\mathrm{sub}} = L^p(\Omega, \hat{H}_{\mathrm{sub}}) \to H = L^2(\Omega, \hat{H})$$

for $p > 2$ with respect to an algorithmically useful $\mathrm{D}P_c$.

Let us further point out that the assumptions on $\mathrm{D}P_c$ from section 3.2.2, which were needed for the invertibility of the Newton system, follow from the respective assumptions on the pointwise proximal map.

Proposition 3.21. *Suppose that for a Baire-Carathéodory function* $\mathrm{D}\hat{P} \colon \Omega \times \hat{H} \to \mathcal{B}(\hat{H}, \hat{H})$ *the pointwise operator* $\mathrm{D}\hat{P}(x, \hat{q}) \colon \hat{H} \to \hat{H}$ *fulfills Assumption 3.3 for every* $x \in \Omega$ *and* $\hat{q} \in \hat{H}$ *(i.e.,* $\mathrm{D}\hat{P}(x, \hat{q})$ *is symmetric, positive semidefinite, and* $\|\mathrm{D}\hat{P}(x, \hat{q})\|_{\hat{H} \to \hat{H}} \leq 1$*). Then the superposition* $\mathrm{D}P \colon H \to \mathcal{B}(H, H)$ *on* $H = L^2(\Omega, \hat{H})$ *fulfills the same assumption.*

Proof. It is straightforward to verify that the superposition operator is symmetric and positive definite. The norm bound is a consequence of Hölder's inequality, i.e.,

$$\|\mathrm{D}P(q)\delta q\|^2 = \int_\Omega \left(\mathrm{D}\hat{P}(x, q(x))\delta q(x)\right)^2 \mathrm{d}x \leq \sup_{x \in \Omega} \|\mathrm{D}\hat{P}(x, q(x))\|^2_{\hat{H} \to \hat{H}} \int_\Omega \delta q^2(x) \, \mathrm{d}x. \qquad \square$$

3.3.2. Concrete examples

Of course, the construction of the algorithm hinges upon the easy (or at least computationally efficient) practical realization of the proximal map. Unfortunately, the proximal map does not in general admit a closed form representation. However, for the concrete ψ considered here, we can always derive explicit formulas. Furthermore, is necessary that we find an appropriate generalized derivative, which is in general not an automatic process but requires some mathematical analysis. In the following we discuss the special cases that are considered in this thesis. We give the concrete formulas for the proximal maps, discuss possible generalized differentials and point out the concrete choices of the space H_{sub} in each case. Moreover, we give concrete interpretations of the spaces $H_{\mathrm{D}P_c(q)}$, which were introduced to solve the Newton system in section 3.2.2.

Box constraints

In the case of box constraints we have an admissible set given by

$$U_{\mathrm{ad}} = \{\, u \in L^2(\Omega) \mid u_a \le u \le u_b \ \text{ almost everywhere} \,\},$$

where $u_a, u_b \in L^\infty(\Omega)$ are given lower and upper bounds with $u_a \le u_b$. Historically, the theory of semismooth Newton methods in Banach spaces has been shaped by this example (see, e.g., Ulbrich [Ulb02], Hintermüller, Ito, and Kunisch [HIK03], or Ito and Kunisch [IK04]). To put it in the context of the given framework, we set

$$\psi(u) = \mathbb{I}_{U_{\mathrm{ad}}}(u) = \begin{cases} 0 & \text{if } u \in U_{\mathrm{ad}}, \\ \infty & \text{else.} \end{cases}$$

The Hilbert space is chosen as $H = L^2(\Omega)$. It is easy to see that the proximal map is given by the projection onto the admissible set

$$P_c(q) = P_{\mathrm{ad}}(q), \quad \text{where } P_{\mathrm{ad}}(q)(x) = \begin{cases} q(x) & \text{if } u_a(x) \le q(x) \le u_b(x), \\ u_a(x) & \text{if } q(x) < u_a(x), \\ u_b(x) & \text{if } q(x) > u_b(x), \end{cases}$$

for all $x \in \Omega$. In the following, we will mostly suppress the dependence on the spatial variable x where no ambiguity arises. In this notation, the directional derivatives are easily computed as

$$\mathrm{d}P_{\mathrm{ad}}(q, \delta q) = \begin{cases} \delta q & \text{where } u_a < q < u_b, \\ 0 & \text{where } q < u_a \text{ or } q > u_b, \\ \delta q^+ & \text{where } q = u_a, \\ -\delta q^- & \text{where } q = u_b. \end{cases}$$

Here, $(\cdot)^+ = \max(0, \cdot)$ and $(\cdot)^- = -\min(0, \cdot)$ denote the positive and negative part, respectively. We denote the active and strongly active set at $q \in H$ respectively by

$$\mathcal{A}(q) = \{\, x \in \Omega \mid q(x) \le u_a(x) \text{ or } q(x) \ge u_b(x) \,\},$$
$$\mathcal{A}^s(q) = \{\, x \in \Omega \mid q(x) < u_a(x) \text{ or } q(x) > u_b(x) \,\}.$$

We observe that the directional derivative is linear under the supposition that the "ambiguous" set $\mathcal{A}(q) \setminus \mathcal{A}^s(q) = \{\, q = u_a \text{ or } q = u_b \,\}$ has Lebesgue measure zero. To obtain a suitable generalized differential, we have to modify the directional derivative on this set. As candidates for the generalized derivative, usually a choice of

$$\mathrm{D}P_{\mathrm{ad}}(q)\delta q = \begin{cases} \delta q & \text{where } u_a < q < u_b, \\ 0 & \text{where } q < u_a \text{ or } q > u_b, \\ d\,\delta q & \text{where } q = u_a \text{ or } q = u_b, \end{cases}$$

for different d is considered. Note, that $\mathrm{D}P_c(q)$ is a pointwise multiplication operator. Therefore, with a slight abuse of notation, we can also specify it in the form

$$\mathrm{D}P_{\mathrm{ad}}(q) = \begin{cases} 1 & \text{where } u_a < q < u_b, \\ 0 & \text{where } q < u_a \text{ or } q > u_b, \\ d & \text{where } q = u_a \text{ or } q = u_b. \end{cases}$$

In [HIK03] a constant choice of $d \in \mathbb{R}$ arbitrary is considered. The construction in [Ulb11, Section 3.3.2] yields an arbitrary $d \in L^\infty(\mathcal{A}(q) \setminus \mathcal{A}^s(q))$ which fulfills $0 \leq d(x) \leq 1$ for almost all $x \in \mathcal{A}(q) \setminus \mathcal{A}^s(q)$. There is no difference in the superlinear convergence theory that can be developed based on either choice. However, from the point of view of Assumption 3.3 the restriction $0 \leq d(x) \leq 1$ is important. We will usually prefer the choice of $\mathrm{D}P_c$ as the indicator function of the inactive set $\mathcal{I}(q) = \Omega \setminus \mathcal{A}(q)$ for convenience of notation. We set

$$\mathrm{D}P_{\mathrm{ad}}(q) = \chi_{\mathcal{I}(q)} = \begin{cases} 1 & \text{where } u_a < q < u_b, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that this construction fulfills Assumption 3.4, which yields the well-known semismoothness of $P_{\mathrm{ad}} \colon H_{\mathrm{sub}} = L^r(\Omega) \to H = L^2(\Omega)$ for any $r > 2$ with respect to any $\mathrm{D}P_{\mathrm{ad}}$ as considered before; see [HIK03; Ulb11]. We will not go into further detail here, since this is well-established. Furthermore, the verification of Assumption 3.3 is trivial in this case (i.e., $\mathrm{D}P_{\mathrm{ad}} = \chi_{\mathcal{I}(q)}$ is a symmetric, positive semidefinite operator on $H = L^2(\Omega)$ with norm bound one).

Let us mention that here the space $H_{\mathrm{D}P_c(q)}$ is isometrically isomorphic to $L^2(\mathcal{I}(q))$, which is the canonical restriction of $L^2(\Omega)$ to the current inactive set $\mathcal{I}(q)$. This follows from

$$\operatorname{Ker} \mathrm{D}P_c(q) = \{\, u \in L^2(\Omega) \mid \chi_{\mathcal{I}(q)} u = 0 \,\} = L^2(\Omega \setminus \mathcal{I}(q))$$

and the identification

$$H_{\mathrm{D}P_c(q)} = {L^2(\Omega)}\big/{L^2(\Omega \setminus \mathcal{I}(q))} = L^2(\mathcal{I}(q)).$$

In this case the quotient space is closed w.r.t. the $\chi_{\mathcal{I}(q)}$-norm. Therefore, the proof of invertibility of the Newton operator as given in Lemma 3.14 has the following interpretation: on the inactive sets, the Newton system corresponds to a linear quadratic problem, which can be solved using the strong convexity. Then an expression for the update on the active sets can be derived with a pointwise formula. This interpretation corresponds to the usual strategy to prove invertibility of the in the context of semismooth Newton or active set methods.

Sparsity

A nonsmooth reformulation for sparse optimal control problems in conjunction with semismooth Newton methods was first discussed in Stadler [Sta09]. For sparse optimal control problems, we have

$$\psi(u) = \alpha \|u\|_{L^1(\Omega)} = \int_\Omega \alpha |u(x)| \, \mathrm{d}x.$$

Here, the Hilbert space is chosen as $H = L^2(\Omega)$. Let us compute the proximal map with Proposition 3.17. To find the pointwise proximal map $\hat{u} = \hat{P}_c(\hat{q})$, we have to minimize the one-dimensional functional

$$\hat{u} \mapsto \frac{c}{2}(\hat{u} - \hat{q})^2 + \alpha |\hat{u}|.$$

It is easy to see that the optimality condition, given by $c(\hat{q} - \hat{u}) \in \alpha \partial |\hat{u}|$ is equivalent to $\hat{u} = 0$ if $|\hat{q}| \leq \alpha/c$, and $\hat{u} = \hat{q} - \operatorname{sgn}(\hat{q})\, \alpha/c$ otherwise. The solution is therefore given by $\hat{P}_c(\hat{q}) = 1/c\,(c - \alpha/|\hat{q}|)^+ \,\hat{q} = (\hat{q} - \alpha/c)^+ - (\hat{q} + \alpha/c)^-$. Recall that $(\cdot)^+ = \max(0, \cdot)$ and $(\,\cdot\,)^- = -\min(0, \cdot)$ denote the positive and negative part, respectively. The operator

$$\operatorname{shrink}_{\alpha/c}(\hat{q}) = (\hat{q} - \alpha/c)^+ - (\hat{q} + \alpha/c)^-$$

is sometimes referred to as the "soft-shrinkage" operator for the parameter α/c. Thereby, the proximal mapping of ψ in H is given as the superposition

$$P_c(q) = \operatorname{shrink}_{\alpha/c}(\hat{q}) = \begin{cases} q - \alpha/c & \text{where } q \geq \alpha/c, \\ q + \alpha/c & \text{where } q \leq -\alpha/c, \\ 0 & \text{otherwise.} \end{cases}$$

Similar to the case of box constraints, a generalized derivative can be given by the indicator function of the inactive set

$$\mathrm{D}P_c(q) = \chi_{\mathcal{I}(q)} = \begin{cases} 1 & \text{where } |q| > \alpha/c, \\ 0 & \text{otherwise,} \end{cases}$$

which is the superposition of $\mathrm{D}\hat{P}_c(\hat{q}) = \chi_{\{\,|\hat{q}|>\alpha/c\,\}}$. The verification of the pointwise semismoothness and the conditions on the generalized derivative (Assumption 3.3) is the same as in the previous case of box-constraints; we choose $H_{\mathrm{sub}} = L^r(\Omega)$ for some $r > 2$.

Directional sparsity

The concept of "directional sparsity" and a semismooth Newton method for a corresponding optimal control formulation were first discussed in Herzog, Stadler, and Wachsmuth [HSW12]. In this case, we consider

$$\psi(u) = \alpha \|u\|_{L^1(\Omega, L^2(I))} = \int_\Omega \alpha \|u(x)\|_{L^2(I)} \,\mathrm{d}x,$$

where $I = (0, T)$ is a time interval and Ω is a bounded domain. Here, we choose the Hilbert space as $H = L^2(\Omega, L^2(I)) = L^2(I \times \Omega)$. In this setting, we think of the pointwise evaluation of a function $u \in H$ at the point x as the function $u(x) \in L^2(I) = \hat{H}$. Again, we can reduce the computation of P_c to the minimization of

$$\hat{u} \mapsto \frac{c}{2} \|\hat{u} - \hat{q}\|_{\hat{H}}^2 + \alpha \|\hat{u}\|_{\hat{H}},$$

for a given $\hat{q} \in \hat{H}$. It is clear that $\hat{u}$ must either be zero, or a positive scalar multiple of $\hat{q}$. More specifically, if $\hat{u}$ is not equal to zero, it follows from the first order conditions that

$$0 = c\,(\hat{u} - \hat{q}) + (\alpha\,\hat{u})/\|\hat{u}\|_{\hat{H}} = (c + \alpha/\|\hat{u}\|_{\hat{H}})\,\hat{u} - c\,\hat{q}.$$

Taking the $\hat{H}$ norm, we obtain $c\|\hat{u}\|_{\hat{H}} + \alpha = c\,\|\hat{q}\|_{\hat{H}}$. This directly yields the formula $\|\hat{u}\|_{\hat{H}} = \|\hat{q}\|_{\hat{H}} - \alpha/c$. Since $\|\hat{u}\|_{\hat{H}} < 0$ is not possible, we must have $\hat{u} = 0$ in the case $\|\hat{q}\|_{\hat{H}} \leq \alpha/c$, and we obtain $c\,\hat{P}_c(\hat{q}) = c\,\hat{u} = (c - \alpha/\|\hat{q}\|_{\hat{H}})^+\,\hat{q}$. By Proposition 3.17, P_c is therefore given by the "stripe-wise" soft-shrinkage operator

$$P_c(q)(x) = \frac{1}{c}\,(c - \alpha/\|q(x)\|_{\hat{H}})^+\,q(x) \tag{3.18}$$

for $x \in \Omega$ almost everywhere.

The directional derivative of the pointwise proximal map can be easily computed as

$$\mathrm{d}\hat{P}_c(\hat{q}; \delta\hat{q}) = \frac{1}{c} \begin{cases} (c - \alpha/\|\hat{q}\|_{\hat{H}})^+\,\delta\hat{q} + \frac{\alpha(\hat{q}, \delta\hat{q})_{\hat{H}}}{\|\hat{q}\|_{\hat{H}}^3}\,\hat{q} & \text{if } \|\hat{q}\|_{\hat{H}} > \alpha/c, \\ \frac{\alpha(\hat{q}, \delta\hat{q})_{\hat{H}}}{\|\hat{q}\|_{\hat{H}}^3}\,\hat{q} & \text{if } \|\hat{q}\|_{\hat{H}} = \alpha/c \text{ and } (\hat{q}, \delta\hat{q})_{\hat{H}} > 0, \\ 0 & \text{else.} \end{cases}$$

Again, it is nonlinear in $\delta\hat{q}$ only in the "ambiguous" case where $\|\hat{q}\|_{\hat{H}} = \alpha/c$. For the pointwise generalized derivative, a possible choice is given by

$$\mathrm{D}\hat{P}_c(\hat{q})\delta\hat{q} = \frac{1}{c}\begin{cases}(c - \alpha/\|\hat{q}\|_{\hat{H}})^+ \delta\hat{q} + \frac{\alpha(\hat{q},\delta\hat{q})_{\hat{H}}}{\|\hat{q}\|_{\hat{H}}^3}\,\hat{q} & \text{if } \|\hat{q}\|_{\hat{H}} > \alpha/c,\\ 0, & \text{else.}\end{cases}$$

Note that in this case, the generalized derivative cannot be solely understood as a pointwise multiplication. We can write the corresponding linear operator schematically as

$$\mathrm{D}\hat{P}_c(\hat{q}) = \frac{\chi_{\{\,\|\hat{q}\|_{\hat{H}} > \alpha/c\,\}}}{c}\left((c - \alpha/\|\hat{q}\|_{\hat{H}})^+ + \frac{\alpha}{\|\hat{q}\|_{\hat{H}}^3}\hat{q}\otimes\hat{q}\right), \tag{3.19}$$

where $(\hat{q}\otimes\hat{q})(\,\cdot\,) = (\hat{q},\,\cdot\,)_{\hat{H}}\,\hat{q}$ is the rank-one product of $\hat{q}$ with itself. We can show semismoothness of P_c with the help of the general framework. It is a slight generalization of the result given in [HSW12, Lemma 3.2] (where $r \geq 6$ is required).

Lemma 3.22. *The proximal mapping* (3.18)*, considered as an operator from $H_{\mathrm{sub}} = L^r(\Omega,\hat{H})$ to $H = L^2(\Omega,\hat{H})$ for some $r > 2$, is semismooth with respect to the generalized derivative given by*

$$\mathrm{D}P_c(q)\delta q = \frac{\chi_{\mathcal{I}(q)}}{c}\left((c - \alpha/\|q\|_{\hat{H}})^+\,\delta q + \frac{\alpha(q,\delta q)_{\hat{H}}}{\|q\|_{\hat{H}}^3}\,q\right)$$

for all $\delta q \in L^r(\Omega,\hat{H})$, where the stripe-wise inactive set is given by

$$\mathcal{I}(q) = \{\, x\in\Omega \;\mid\; \|q(x)\|_{\hat{H}} > \alpha/c\,\}\,.$$

Moreover, $\mathrm{D}P_c$ conforms to Assumption 3.3.

Proof. It is possible to show the Baire-Carathéodory property for (3.19) by approximating the characteristic function with a sequence of continuous functions. However, this was only needed to ensure that $\mathrm{D}P_c$ maps measurable functions to measurable functions, which can alternatively be seen directly. According to section 3.3.1, we now analyze the pointwise proximal map $\hat{P}$ with respect to the given $\mathrm{D}\hat{P}_c$. It is evident from representation (3.19) that $\mathrm{D}\hat{P}_c(\hat{q})$ is a symmetric, positive semidefinite operator on $\hat{H}$ (independently of $\hat{q}\in\hat{H}$). Furthermore, $\|\mathrm{D}\hat{P}_c(\hat{q})\|_{\hat{H}\to\hat{H}} \leq 1$ follows from a straightforward computation. With Proposition 3.21, these properties transfer to $\mathrm{D}P_c(\cdot)\colon H\to H$. To prove semismoothness, we will apply Theorem 3.20. It remains to show that for all $\hat{q}^*\in\hat{H}$ we have

$$\hat{R}^*(\hat{q}) = \frac{1}{\|\hat{q}^* - \hat{q}\|_{\hat{H}}}\left[\hat{P}_c(\hat{q}) - \hat{P}_c(\hat{q}^*) - \mathrm{D}\hat{P}_c(\hat{q})(\hat{q}^* - \hat{q})\right] \to 0 \quad \text{for } \|\hat{q}^* - \hat{q}\|_{\hat{H}} \to 0. \tag{3.20}$$

Define the function $F\colon \hat{H}\setminus\{\,0\,\}\to\hat{H}$ as $F\colon \hat{q}\to 1/c\,(c - \alpha/\|\hat{q}\|_{\hat{H}})\,\hat{q}$. F is twice continuously differentiable with gradient

$$\nabla F(\hat{q}) = \frac{1}{c}\left((c - \alpha/\|\hat{q}\|_{\hat{H}}) + \frac{\alpha}{\|\hat{q}\|_{\hat{H}}^3}\hat{q}\otimes\hat{q}\right).$$

Furthermore, we can write $\hat{P}_c(\cdot) = \chi_{\{\,\|\cdot\|_{\hat{H}} > \alpha/c\,\}}F(\cdot)$. We distinguish the cases $\|\hat{q}^*\|_{\hat{H}} < \alpha/c$ and $\|\hat{q}^*\|_{\hat{H}} \geq \alpha/c$. In the first case we have $\hat{P}_c(\hat{q}^*) = \hat{P}_c(\hat{q}) = \mathrm{D}P_c(\hat{q})(\hat{q} - \hat{q}^*) \equiv 0$ for all $\hat{q}$ from a neighborhood of $\hat{q}^*$ and (3.20) is trivially fulfilled. In the second case, we have

$$\hat{R}^*(\hat{q}) = \frac{1}{\|\hat{q}^* - \hat{q}\|_{\hat{H}}}\begin{cases}F(\hat{q}) - F(\hat{q}^*) - F'(\hat{q})(\hat{q} - \hat{q}^*) & \text{if } \|\hat{q}\|_{\hat{H}} > \alpha/c,\\ -F(\hat{q}^*) & \text{if } \|\hat{q}\|_{\hat{H}} \leq \alpha/c.\end{cases} \tag{3.21}$$

Since $\hat{q}^* \neq 0$, we have

$$F(\hat{q}) - F(\hat{q}^*) - F'(\hat{q})(\hat{q} - \hat{q}^*) \in o(\|\hat{q}^* - \hat{q}\|_{\hat{H}}),$$

due to Fréchet-differentiability of F at $\hat{q}^*$ and continuity of the derivative. Now, we further distinguish between $\|\hat{q}^*\|_{\hat{H}} = \alpha/c$ and $\|\hat{q}^*\|_{\hat{H}} > \alpha/c$. In the first case, we have $-F(\hat{q}^*) = 0$. In the second, we again observe that $\|\hat{q}\|_{\hat{H}} > \alpha/c$ for all $\hat{q}$ with $\|\hat{q}^* - \hat{q}\|_{\hat{H}} \leq \|\hat{q}^*\|_{\hat{H}} - \alpha/c$. Therefore, $\hat{R}^*(\hat{q}) \to 0$ for $\hat{q} \to \hat{q}^*$ is verified for all $\hat{q}^* \in \hat{H}$ and we can apply Theorem 3.20 to conclude the proof. □

Let us give an interpretation of the space $H_{\mathrm{D}P_c(q)}$ from section 3.2.3. Again, the kernel of $\mathrm{D}P_c(q)$ corresponds to the current inactive sets, i.e., we have that

$$\operatorname{Ker} \mathrm{D}P_c(q) = \{\, u \in L^2(\Omega, \hat{H}) \mid \chi_{\mathcal{I}} u = 0 \,\} = L^2(\Omega \setminus \mathcal{I}(q), \hat{H}).$$

Consequently, the quotient space ${}^{H}\!/_{\operatorname{Ker} \mathrm{D}P_c(q)}$ is isometrically isomorphic to the restriction of H to the inactive sets $L^2(\mathcal{I}(q), \hat{H})$. However, in this case the space $L^2(\mathcal{I}(q), \hat{H})$ is not closed if endowed with the inner product induced by $\mathrm{D}P_c(q)$, which is given by

$$(u, v)_{\mathrm{D}P_c(q)} = \frac{1}{c} \int_{\mathcal{I}(q)} \left((c - \alpha/\|q\|_{\hat{H}})^+ (u, v)_{\hat{H}} + \frac{\alpha}{\|q\|_{\hat{H}}^3} (q, u)_{\hat{H}} (q, v)_{\hat{H}} \right) \mathrm{d}x$$

for any $u, v \in H$. In fact, since $\hat{H}$ has more than one dimension, the space of functions $u \in H$ with $(q(x), u(x))_{\hat{H}} = 0$ for $x \in \Omega$ (almost everywhere) is not trivial. For all of these functions, the product given above corresponds to a *weighted* inner product with the weight $(c - \alpha/\|q\|_{\hat{H}})^+/c$. Since the weight is bounded by one, but not necessarily bounded away from zero, the closure of ${}^{H}\!/_{\operatorname{Ker} \mathrm{D}P_c(q)}$ with respect to the inner product given above is larger than $L^2(\mathcal{I}(q), \hat{H})$, in general. We obtain the identification

$$H_{\mathrm{D}P_c(q)} \cong \{\, u \colon \mathcal{I}(q) \to \hat{H} \;\mid\; \|u\|_{\mathrm{D}P_c(q)} < \infty \,\}$$

in the sense of the usual equivalence class construction for Lebesgue spaces.

Directional sparsity with positivity constraints

If we want to additionally enforce positivity of the controls, we consider

$$\psi(u) = \alpha \|u\|_{L^1(\Omega, L^2(I))} + \mathbb{I}_{\{\, u \geq 0 \text{ on } I \times \Omega \,\}}(u).$$

As in the previous section, we can reduce the computation of the proximal map P_c to the computation of the pointwise proximal map corresponding to

$$\hat{\psi}(\hat{u}) = \alpha \|\hat{u}\|_{L^2(I)} + \mathbb{I}_{\{\, \hat{u} \geq 0 \text{ on } I \,\}}(\hat{u}).$$

It is given by

$$\hat{P}_c(\hat{q}) = \frac{1}{c} \left(c - \alpha/\|\hat{q}^+\|_{L^2(I)} \right)^+ \hat{q}^+,$$

where $\hat{q}^+$ denotes the positive part of $\hat{q}$ in $L^2(I)$. In fact, for any $\hat{q} \in L^2(I)$ and the choice $\hat{u} = 1/c\,(c - \alpha/\|\hat{q}^+\|_{L^2(I)})^+ \hat{q}^+$ we compute that

$$c(\hat{q} - \hat{u}) = c\hat{q}^- + \left[(c - \alpha/\|\hat{q}^+\|_{L^2(I)})^+ - c \right] \hat{q}^+ \in \partial \mathbb{I}_{\{\, \hat{u} \geq 0 \text{ on } I \,\}}(\hat{u}) + \partial \|\hat{u}\|_{L^2(I)} \subseteq \partial \hat{\psi}(\hat{u})$$

using the sum rule for the convex subdifferential. With the equivalent characterization of the proximal map via the subdifferential (see Proposition 3.4.(i)) we obtain $\hat{u} = P_c(\hat{q})$. Therefore, the proximal map of ψ is given as the superposition

$$P_c(q) = \frac{1}{c}\left(c - \alpha/\|q^+\|_{L^2(I)}\right)^+ q^+. \tag{3.22}$$

In other words, the proximal map can be decomposed as

$$P_c = F_2 \circ F_1 \quad \text{where } F_2(q_2) = (c - \alpha/\|q_2\|_{L^2(I)})^+ q_2 \quad \text{and } F_1(q_1) = q_1^+, \tag{3.23}$$

where F_2 is the proximal map of $\alpha\|\cdot\|_{L^1(\Omega_c, L^2(I))}$ form the previous section and F_1 is the projection to the positive cone. With the chain rule, we obtain the following generalized differential.

Lemma 3.23. *The proximal mapping* (3.18)*, considered as an operator from* $H_{\text{sub}} = L^r(I \times \Omega) = L^r(\Omega, L^r(I))$ *to* $H = L^2(I \times \Omega) = L^2(\Omega, L^2(I))$ *for some* $r > 2$*, is semismooth with respect to the generalized derivative given by*

$$\mathrm{D}P_c(q)\delta q = \frac{\chi_{\mathcal{I}_\Omega(q^+)}}{c}\left(\left(c - \alpha/\|q^+\|_{L^2(I)}\right)^+ \chi_{\mathcal{I}_{I\times\Omega}(q)}\delta q + \frac{\alpha(q^+, \delta q)_{L^2(I)}}{\|q^+\|^3_{L^2(I)}}\, q^+\right)$$

for all $\delta q \in L^r(I \times \Omega)$*, where the stripe-wise and the space-time inactive set are given by*

$$\mathcal{I}_\Omega(q^+) = \{\, x \in \Omega \;\mid\; \|q^+(x)\|_{L^2(I)} > \alpha/c \,\},$$
$$\mathcal{I}_{I\times\Omega}(q) = \{\, (t,x) \in I \times \Omega \;\mid\; q(t,x) > 0 \,\}.$$

Moreover, $\mathrm{D}P_c$ *conforms to Assumption 3.3.*

Proof. We define the pointwise functions $\hat{F}_1 \colon L^r(I) \to L^2(I)$ and $\hat{F}_2 \colon L^2(I) \to L^2(I)$ according to (3.23). Semismoothness of $\hat{F}_1(\hat{q}) = (\hat{q})^+ = \max(\hat{q}, 0)$ with respect to $\mathrm{D}F_1(\hat{q}) = \chi_{\{\, t\in I \mid \hat{q}(t) > 0 \,\}}$ follows as in the case of box-constraints (with norm-gap). The semismoothness of $\hat{F}_2$ (without norm-gap) with respect to the generalized differential (3.19) has already been verified in Proposition 3.22. By the semismooth chain rule (see Proposition 3.9) it follows now that $\hat{P}_c = \hat{F}_2 \circ \hat{F}_1 \colon L^r(I) \to L^2(I)$ is semismooth with respect to the generalized differential

$$\begin{aligned}\mathrm{D}\hat{P}_c(\hat{q}) &= \frac{\chi_{\{\, \|\hat{q}^+\|_{L^2(I)} > \alpha/c \,\}}}{c}\left(\left(c - \alpha/\|\hat{q}^+\|_{L^2(I)}\right)^+ + \frac{\alpha}{\|\hat{q}^+\|^3_{L^2(I)}}\hat{q}^+ \otimes \hat{q}^+\right)\chi_{\{\, t\in I \mid \hat{q}(t) > 0 \,\}} \\ &= \frac{\chi_{\{\, \|\hat{q}^+\|_{L^2(I)} > \alpha/c \,\}}}{c}\left(\left(c - \alpha/\|\hat{q}^+\|_{L^2(I)}\right)^+ \chi_{\{\, t\in I \mid \hat{q}(t) > 0 \,\}} + \frac{\alpha}{\|\hat{q}^+\|^3_{L^2(I)}}\hat{q}^+ \otimes \hat{q}^+\right).\end{aligned}$$

It is possible to show that $\mathrm{D}\hat{P}_c$ is a Baire-Carathéodory function by approximating the characteristic functions with appropriate smooth functions. Alternatively, the fact that $\mathrm{D}P_c$ maps measurable to measurable functions can again be seen directly. Now, we verify that Assumption 3.3 holds for $\mathrm{D}\hat{P}_c(\hat{q}) \colon L^2(I) \to L^2(I)$. Symmetry and positivity are obvious and the norm bound $\|\mathrm{D}\hat{P}_c(\hat{q})\|_{L^2(I)\to L^2(I)} \leq 1$ for all $\hat{q} \in L^2(I)$ can be verified with a direct computation. Therefore, Assumption 3.3 also holds for the superposition operator $\mathrm{D}P_c(q)(x) = \mathrm{D}\hat{P}_c(q(x))$ on the space $L^2(\Omega, L^2(I))$; see Proposition 3.21.

Furthermore, since the embedding $L^r(I) \hookrightarrow L^2(I)$ is continuous, $\hat{P}_c \colon L^r(I) \to L^2(I)$ is globally Lipschitz continuous and $\mathrm{D}\hat{P}_c(\hat{q}) \colon L^r(I) \to L^2(I)$ is uniformly bounded for all $\hat{q} \in L^r(I)$. Now, we apply Theorem 3.20 to obtain semismoothness of the superposition operator $P_c(q)(x) = \hat{P}_c(q(x))$ from $L^r(\Omega, L^r(I))$ to $L^2(\Omega, L^2(I))$ with respect to the generalized differential defined by $\mathrm{D}P_c(q)(x) = \mathrm{D}\hat{P}_c(q(x))$. This directly leads to the form of $\mathrm{D}P_c$ as given above. □

3.4. Algorithmic aspects

In this section, we will discuss the practical aspects of the proposed optimization methods. In particular, we will describe an iterative approach to the solution of the Newton system and globalization strategies.

3.4.1. Iterative solution of the Newton system

In the following, we fix some $q \in H$, abbreviate $u = P_c(q)$, and let $T = \mathrm{D}P_c(q)$ be the corresponding generalized derivative of the proximal map. We will discuss a numerical solution strategy for the linear system $\mathrm{D}G(q)\delta q = -G(q)$. Note, that in most cases it is practically infeasible to compute a full representation of $\mathrm{D}G(q)$, since the functional f involves a control-to-state mapping associated to a PDE. Therefore, we decide to use an iterative solution procedure. We have already seen that the Newton operator $\mathrm{D}G(q)$ is symmetric with respect to the inner product $(\cdot,\cdot)_T$. Moreover, we can expect it to be positive definite for linear quadratic f with $\gamma > 0$ or under a second order condition. Furthermore, we have seen that a solution of the equation $\mathrm{D}G(q)\delta q = -G(q)$ can be reduced to the solution of the quadratic problem

$$\min_{v\in H} Q_q(v) = (G(q), v)_T + \frac{1}{2}(v, \mathrm{D}G(q)v)_T. \tag{3.24}$$

By Lemma 3.14 (or 3.15) we know that if (3.24) is coercive with respect to the space H_T, the solution to (3.24) can also be found in the original space H. In this section, we will assume that (3.24) is uniquely solvable (up to equivalence in $H/\operatorname{Ker} T$).

The numerical implementation follows closely the theoretical setup. To compute a specific solution $\delta\tilde{q} \in H$ of (3.24) we apply the method of conjugate gradients (cg-method), which can be regarded as an iterative minimization for $Q_q(h)$. In the method, we compute products of search directions $d \in H$ with the full (in general non-symmetric) system operator $\mathrm{D}G(q)$, and compute inner products with $(\cdot,\cdot)_T$. Define the Krylov-space

$$\mathcal{K}_m = \{\, [\mathrm{D}G(q)]^k\, G(q) \;\mid\; k = 0, 1, \ldots, m-1 \,\} \subset H.$$

Then, performing m steps of conjugate gradients will compute the minimum $\delta\tilde{q}_m \in H$ of

$$\delta\tilde{q}_m = \operatorname*{argmin}_{v\in\mathcal{K}_m} Q_q(v).$$

For each $m > 0$ the minimum $\delta\tilde{q}_m \in H$ is unique. It is possible that there is a $n < m$ such that $\delta\tilde{q}_n$ is also a minimizer for all $\tilde{n} > n$. In this case the conjugate gradient iteration stops with

$$T\mathrm{D}G(q)\delta\tilde{q}_n = -TG(q).$$

In general, we only obtain an approximate solution $\delta\tilde{q}_m \in \mathcal{K}_m$ of (3.24) for some $m > 0$, based on an appropriate stopping criterion.

Remark 3.7. i) Since the Hessian $\nabla^2 f(\cdot)$ is typically a compact operator, the Newton operator $\mathrm{D}G(\cdot) = \gamma\,\mathrm{Id} + \nabla^2 f(\cdot)T$ (for $c = \gamma > 0$) is a compact perturbation of the identity, and we can expect superlinear convergence for the cg-method; see, e.g., [Dan67; Win80].

ii) Usually, a cg-method for the operator $\mathrm{D}G(q)$ w.r.t. the inner product induced by T can be interpreted as a preconditioned cg-method for the symmetric "iteration matrix" $\tilde{A} = T\mathrm{D}G(q)$ with the symmetric preconditioner $\tilde{T} = T^{-1}$. However, since T is typically not invertible, this is not directly possible here. Moreover, we have seen that even if we factor out the kernel of T, invertibility is not automatically fulfilled (cf. the discussion of H_T for the case of directional sparsity in section 3.3.2). Consequently, the finite dimensional approximations to T which appear in practical computations can be arbitrarily ill-conditioned, even if the kernel is eliminated.

Algorithm 1 Conjugate gradients with final step

$r_0 = b = -G(q)$
$d_0 = r_0$
$\delta q_0 = 0$
for $k = 0, 1, \ldots$ **do**
  compute $Ad_k = \mathrm{D}G(q)d_k \in H$
  $\beta_k = \frac{\|r_k\|_T^2}{(d_k, A\,d_k)_T}$
  $\delta q_{k+1} = \delta q_k + \beta_k d_k$
  $r_{k+1} = r_k - \beta_k A d_k$
  if <tolerance reached> **then**
    $h_{k+2} = h_{k+1} + 1/c\; r_{k+1}$ {final step (3.25)}
    return h_{k+2} {"converged"}
  end if
  $d_{k+1} = r_{k+1} + \frac{\|r_{k+1}\|_T^2}{\|r_k\|_T^2}\, d_k$
end for

After having achieved a desired tolerance, we perform the additional final step (3.10) to obtain the full solution. Assume therefore that we have a $\delta\tilde{q} \in H$ solving (3.24) exactly and denote by δq the full solution of $\mathrm{D}G(q)\delta q = -G(q)$. We compute for the residual $R(\delta\tilde{q})$ of the full Newton system (3.2) that

$$R(\delta\tilde{q}) = -G(q) - \mathrm{D}G(q)\delta\tilde{q} = \mathrm{D}G(q)(\delta q - \delta\tilde{q}) = c(\delta q - \delta\tilde{q}),$$

since $T(\delta q - \delta\tilde{q}) = 0$. With this identity the full solution δq can be computed from any solution $\delta\tilde{q}$ of (3.24) with the formula

$$\delta q = \delta\tilde{q} + \frac{1}{c}\, R(\delta\tilde{q}). \tag{3.25}$$

Note, that the residual $R(\delta\tilde{q})$ is a byproduct of the cg-method, which does not require an additional evaluation of $\mathrm{D}G(q)$. The complete procedure is given in Algorithm 1.

Remark 3.8. In practice, it can be desirable to compute an approximate solution of (3.24) only up to a very large tolerance. Therefore, as an alternative strategy, we can obtain the final update (3.25) by minimizing the norm of the residual in direction $r = R(\delta\tilde{q})$ by setting

$$\delta q = \delta\tilde{q} + \theta r \;\text{ as the minimizer of }\; \min_{\theta\in\mathbb{R}} \|\mathrm{D}G(q)(\delta\tilde{q} + \theta r) + G(q)\|. \tag{3.26}$$

This might give a more robust computation of the final step in cases where the cg-method fails to solve (3.24) to a sufficient accuracy (at the expense of the additional evaluation of $\mathrm{D}G(q)r$).

3.5. Globalization approaches

In the following, we are going to discuss globalization approaches for the described semismooth Newton method. In many cases, in the context of globalization of semismooth Newton methods, the local Newton method is complemented by another (possibly completely different) first order optimization method; see [Ulb11; Mil15]. In each step of the method, a Newton step is computed. Based on a convergence indicator (e.g., descent in the cost functional), the step is either accepted or rejected. In the case of rejection, a step of the first order method is performed. Another approach is a dampening of the Newton steps based on descent in the squared residual; see [IK09; IK08].

Here, we will focus on a trust-region approach, which tries to achieve a more gradual transition between a cheap first order optimization step and the more expensive semismooth Newton step. However, at the moment, we are unable to give a full global convergence analysis, in contrast to, e.g. [Ulb11]. On the theoretical side, we derive some connections of the reduced cost functional and the normal map and prove global convergence of a related first order optimization method. It will turn out, that in the context of a reformulation based on the normal map, the negative of the current residual $G(q)$ provides a suitable descent direction (which coincidentally is the first search direction in the cg-method, see Algorithm 1). This result serves to give a partial theoretical justification of the following trust-region approach. Under some conditions, also a damped Newton direction is suitable for globalization based on the reduced objective. Let us point out that this stands in contrast to the standard approach to semismooth Newton; see section 3.6.

3.5.1. Theoretical aspects

We base a globalization strategy on the descent in the reduced objective functional

$$q \mapsto j_\gamma(P_c(q)).$$

Recall that $j_\gamma(u) = f_\gamma(u) + \psi(u) = f(u) + \psi(u) + \gamma/2\,\|u\|^2$. To this purpose, it is necessary to understand the influence of a perturbation of q on the reduced objective. First, we consider the convex part and derive a lemma which is related to the continuous differentiability of the Moreau envelope of a convex function (cf. [BC11, Proposition 12.29]).

Lemma 3.24. *Let $q, \tilde{q} \in H$ and denote $u = P_c(q), \tilde{u} = P_c(\tilde{q})$. Then we have*

$$\begin{aligned} &\psi(\tilde{u}) \geq \psi(u) + c(q-u, \tilde{u}-u), \\ \textit{and} \qquad &\psi(\tilde{u}) \leq \psi(u) + c(q-u, \tilde{u}-u) + r(\tilde{q}, q), \end{aligned}$$

where the remainder is given by $r(\tilde{q}, q) = c(\tilde{q}-q, \tilde{u}-u) - c\|\tilde{u}-u\|^2$.

Proof. The first inequality is a direct consequence of $c(q-u) \in \partial\psi(u)$; see Proposition 3.4.(i). Conversely, the second inequality follows from $c(\tilde{q}-\tilde{u}) \in \partial\psi(\tilde{u})$, which results in

$$\begin{aligned} \psi(\tilde{u}) &\leq \psi(u) - c(\tilde{q}-\tilde{u}, u-\tilde{u}) \\ &= \psi(u) + c(\tilde{q}-\tilde{u}, \tilde{u}-u) \\ &= \psi(u) + c(q-u, \tilde{u}-u) + c(\tilde{q}-q-(\tilde{u}-u), \tilde{u}-u), \end{aligned}$$

which directly yields the form of the remainder $r(\tilde{q}, q)$ given above. □

Next, we consider the differentiable part, where we use the following standard estimate based on Lipschitz continuity of the gradient of f_γ.

Lemma 3.25. *Let $u, \tilde{u} \in N \subset H$ for some convex subset of N of H, and denote by L_f the Lipschitz constant of ∇f on N, i.e., we set*

$$L_f = \sup_{u,\tilde{u}\in N} \frac{\|\nabla f(u) - \nabla f(\tilde{u})\|}{\|u - \tilde{u}\|}.$$

Then we have

$$f_\gamma(\tilde{u}) \leq f_\gamma(u) + (\nabla f_\gamma(u), \tilde{u} - u) + \frac{1}{2}(L_f + \gamma)\|\tilde{u} - u\|^2.$$

Now, we show that $-G(q)$ can serve as a "canonical" descent direction for $j_\gamma \circ P_c$.

Lemma 3.26. *Assume that ∇f is Lipschitz continuous on U_{ad}. Let $q \in H$ be arbitrary and set $u = P_c(q)$. Furthermore, define*

$$q_\theta = q - \theta G(q)$$

for $\theta > 0$ and set $u_\theta = P_c(q_\theta)$. Then we have

$$j_\gamma(u_\theta) \leq j_\gamma(u) - \frac{1}{2\theta}\|u_\theta - u\|^2$$

for all $\theta \leq \min\{\, 1/c,\, 1/(L_f + \gamma)\,\}$.

Proof. First we apply the Lemmas 3.24 and 3.25 to obtain

$$\begin{aligned} j_\gamma(u_\theta) &= f_\gamma(u_\theta) + \psi(u_\theta) \\ &\leq j_\gamma(u) + (\nabla f_\gamma(u) + c(q - u), u_\theta - u) + r(q_\theta, q) + \frac{1}{2}(L_f + \gamma)\|u_\theta - u\|^2 \\ &= j_\gamma(u) + (G(q), u_\theta - u) + r(q_\theta, q) + \frac{1}{2}(L_f + \gamma)\|u_\theta - u\|^2, \end{aligned}$$

using the definition of G. Furthermore, by the choice of q_θ, we have $G(q) = -(q_\theta - q)/\theta$. It follows

$$\begin{aligned} j_\gamma(u_\theta) &\leq j_\gamma(u) - \frac{1}{\theta}(q_\theta - q, u_\theta - u) + r(q_\theta, q) + \frac{1}{2}(L_f + \gamma)\|u_\theta - u\|^2 \\ &= j_\gamma(u) + \left(c - \frac{1}{\theta}\right)(q_\theta - q, u_\theta - u) + \left(\frac{1}{2}(L_f + \gamma) - c\right)\|u_\theta - u\|^2, \end{aligned}$$

taking into account that $r(q_\theta, q) = c(q_\theta - q, u_\theta - u) - c\|u_\theta - u\|^2$. For $\theta < 1/c$ the coefficient in the second term is negative and we have

$$\left(c - \frac{1}{\theta}\right)(q_\theta - q, u_\theta - u) \leq \left(c - \frac{1}{\theta}\right)\|u_\theta - u\|^2$$

by the firm nonexpansiveness of the proximal map (see Proposition 3.4.(ii)). This results in

$$j_\gamma(u_\theta) \leq j_\gamma(u) + \left(\frac{1}{2}(L_f + \gamma) - \frac{1}{\theta}\right)\|u_\theta - u\|^2.$$

For $\theta \leq 1/(L_f + \gamma)$, the result follows. □

The previous result shows that an optimization step in direction $-G(q)$ with a step-size $\theta \leq \min\{1/c, 1/(L_f + \gamma)\}$ will lead to a guaranteed descent in the objective. Furthermore, the size of the reduction in the objective value can be related to an expression of the change in the control $u = P_c(q)$. Based on this, we can prove global convergence of the corresponding first order algorithm.

Theorem 3.27. *Assume that ∇f is Lipschitz continuous on U_{ad} and that j_γ is bounded from below. Take any $q_0 \in H$ and define*

$$q_{n+1} = q_n - \theta_n G(q_n) \quad \text{for } n \in \mathbb{N}_0, \tag{3.27}$$

where $\theta_n > 0$ with $\inf_{n\in\mathbb{N}_0} \theta_n > 0$ and $\sup_{n\in\mathbb{N}_0} \theta_n \leq \min\{1/c, 1/(L_f + \gamma)\}$. Then we have $G(q_n) \to 0$ for $n \to \infty$, i.e., the first order optimality measure converges to zero.

Proof. For convenience of notation, abbreviate $u_n = P_c(q_n)$ for $n \in \mathbb{N}_0$. According to Lemma 3.26, we have

$$j_\gamma(u_{n+1}) \leq j_\gamma(u_n) - \frac{1}{2\theta_n}\|u_{n+1} - u_n\|^2. \tag{3.28}$$

As a consequence, the functional values $j_\gamma(u_n)$ are monotonously decreasing and consequently convergent (j_γ is bounded from below). By reordering (3.28) we derive

$$\|u_{n+1} - u_n\|^2 \leq 2\theta_n \left(j_\gamma(u_n) - j_\gamma(u_{n+1})\right) \to 0 \quad \text{for } n \to \infty,$$

since the θ_n are uniformly bounded. Now, we consider the development of the residual along the iterations. According to the definition of G and q_n it holds

$$\begin{aligned} G(q_{n+1}) &= c(q_{n+1} - u_{n+1}) + \nabla f_\gamma(u_{n+1}) \\ &= c(q_n - \theta_n G(q_n) - u_{n+1}) + \nabla f_\gamma(u_{n+1}) \\ &= (1 - c\theta_n)\, G(q_n) - c(u_{n+1} - u_n) + \nabla f_\gamma(u_{n+1}) - \nabla f_\gamma(u_n) \end{aligned}$$

for all $n \geq 0$. Applying the norm, using the triangle inequality and once again the Lipschitz-continuity of ∇f, we obtain

$$\|G(q_{n+1})\| \leq (1 - c\theta_n)\, \|G(q_n)\| + (|\gamma - c| + L_f)\, \|u_{n+1} - u_n\|.$$

Furthermore, we have $0 \leq (1 - c\theta_n) \leq (1 - c \inf_{n\in\mathbb{N}} \theta_n) = \sigma < 1$. Define the sequence $g_n = \|G(q_n)\| \geq 0$ for $n \in \mathbb{N}_0$. It fulfills the estimate $g_{n+1} \leq \sigma g_n + \varepsilon_n$ with $\sigma < 1$ for perturbations $0 \leq \varepsilon_n \to 0$. Therefore, g_n must converge to zero for $n \to \infty$ (see Proposition A.5 in the Appendix). □

Note, that there are two important special cases of algorithm (3.27). The first is valid for a choice of $c \geq L_f + \gamma$ in the definition of the normal map. Then we can choose a constant step-size $\theta_n = \theta = 1/c$ for all $k \in \mathbb{N}_0$, and we obtain

$$u_{n+1} = P_c\left(q_n - \theta\, G(q_n)\right) = P_c\left(u_n - \theta\, \nabla f_\gamma(u_n)\right),$$

such that the auxiliary variable q_n can be eliminated. This is the well-known proximal gradient method (a generalization of the projected gradient method). The convergence properties of this method are well understood; see, e.g. [CW05; NN13] and the references therein for the convex case or [Hin+09] for the projected gradient method in a non-convex, Banach space setting. For

a globalization strategy in the context of the semismooth Newton method from section 3.2, we are especially interested in the case $\gamma > 0$ with an associated choice of $c = \gamma$. In this case, we can also interpret (3.27) as the damped fixed point iteration

$$q_{n+1} = q_n - \theta_n G(q_n) = (1 - \tau_n)q_n - \tau_n \frac{1}{\gamma}\nabla f(P_c(q_n)),$$

with damping parameter $\tau_n = \gamma\theta_n \in (0, 1]$ for the optimality condition $\bar{q} = -1/\gamma\nabla f(P_c(\bar{q}))$.

Based on the globally convergent method from Theorem 3.27, we can now apply the general trust-region approach from [Ulb11] to globalize the semismooth Newton method from section 3.2, by alternating between (scaled) Newton and gradient steps in a suitable way. We do not develop this further here, but refer to [Ulb11, Chapter 7]. In section 3.5.2, we will describe a different (partly heuristic) trust-region method, that will also fall back to a step in direction $-G(q)$ in the small-radius case. This method will additionally try to use as much second order information as possible, by using a modification of the truncated cg-method approach due to Steihaug [Ste83]. Theorem 3.27 provides a first step towards a theoretical justification of this approach.

Another approach to globalization of Newton's method is a damping of the Newton steps. In the context of a semismooth reformulation based on the normal map, this appears feasible as well.

Proposition 3.28. *Suppose that P_c is directionally differentiable at $q \in H$ in direction $\delta q \in H$. Then, the directional derivative of the reduced objective $j_\gamma \circ P_c$ at the point $q \in H$ in direction $\delta q \in H$ is given by*

$$\mathrm{d}[j_\gamma \circ P_c](q, \delta q) = \frac{\mathrm{d}}{\mathrm{d}\tau} j_\gamma(P_c(q + \tau\delta q)) = (G(q), \mathrm{d}P_c(q, \delta q)),$$

where $\mathrm{d}P_c$ is the directional derivative of P_c.

Proof. For convenience of notation, define $u = P_c(q)$, $u_\tau = P_c(q_\tau)$, where $q_\tau = q + \tau\delta q$. By Lemmas 3.24 and 3.25 we obtain similarly as in the proof of Lemma 3.26 that

$$j_\gamma(u_\tau) = j_\gamma(u) + (G(q), u_\tau - u) + \tilde{r}(q_\tau, q),$$

where $|\tilde{r}(q_\tau, q)| \leq ((L_f + \gamma)/2 + c)\|q_\tau - q\|^2 \leq C\tau^2\|\delta q\|^2$, using the Lipschitz continuity of P_c and $(\mathrm{Id} - P_c)$. Dividing by τ, we obtain

$$\frac{1}{\tau}\left(j_\gamma(u_\tau) - j_\gamma(u)\right) = (G(q), (u_\tau - u)/\tau) + O(\tau).$$

For $\tau \to 0$, we obtain $(u_\tau - u)/\tau = (P_c(q_\tau) - P_c(q))/\tau \to \mathrm{d}P_c(q, \delta q)$, which implies the result. □

As a consequence, we can see that the Newton direction is a descent direction, under some conditions. Let us mention that the following corollary is only a weak result, since it does not guarantee a sufficiently large decrease (cf. Lemma 3.26).

Corollary 3.29. *Suppose that P_c is directionally differentiable at the point $q \in H$. Suppose that the conditions of Lemma 3.15 are fulfilled and let $\delta q = -\mathrm{D}G(q)^{-1}G(q)$. If we have $\mathrm{d}P_c(q; \delta q) = \mathrm{D}P_c(q)\delta q$, then it follows*

$$\mathrm{d}[j_\gamma \circ P_c](q; \delta q) = (G(q), \mathrm{D}P_c(q)\delta q) \leq -\nu(\delta q, \mathrm{D}P_c(q)\,\delta q),$$

where $\nu > 0$ is the coercivity constant from Lemma 3.15.

Proof. This is a direct consequence of Proposition 3.28, the equality $G(q) = -\mathrm{D}G(q)\delta q$, the assumption on $\mathrm{d}P_c(q;\delta q)$, and the coercivity of $\mathrm{D}G(q)$ in the inner product induced by the generalized derivative $\mathrm{D}P_c(q)$ as in (3.14). □

We see that a globalization approach based on a damped Newton direction could offer a promising alternative; at least in steps, where the conditions of Corollary 3.29 are fulfilled. However, a proper treatment of the nonsmooth aspect of the problem (to guarantee sufficient descent) and alternative strategies in the case that the prerequisites of Corollary 3.29 are violated, are still missing.

3.5.2. A trust region method

In the following, we will describe a heuristic trust region approach for the globalization of the semismooth Newton method from section 3.2. The algorithm is inspired by the truncated conjugate gradients approach due to Steihaug [Ste83]. In fact, in the smooth setting for $\psi \equiv 0$ and $P_c = \mathrm{Id}$, we will recover the original algorithm (more precisely, the Hilbert space adaptation thereof).

First, as a mathematical concept, we define the trust region subproblem at the iterate q_n with the $T_n = \mathrm{D}P_c(q_n)$ as

$$\min_{v\in H} Q_{q_n}(v) \quad \text{subject to } \|v\|_{T_n} \leq \sigma_n. \tag{3.29}$$

Note, that the solution of this problem is not unique if T_n has a nontrivial kernel. However, an approximate minimizer of (3.29) can be obtained by the Steihaug cg-method as described in Algorithm 2. The globalization strategy will be based on the descent in the objective functional values $j_\gamma(u_n)$. We propose to update the trust region radius σ_n by comparing the functional decrease

$$\rho_n^{\mathrm{act}} = j_\gamma(u_{n+1}) - j_\gamma(u_n) = j_\gamma(P_c(q_n + \delta q_n)) - j_\gamma(P_c(q_n))$$

to the model decrease predicted by the quadratic model

$$\rho_n^{\mathrm{pred}} = Q_{q_n}(\delta q_n) = (G(q_n), \delta q_n)_{T_n} + \frac{1}{2}(\delta q_n, \mathrm{D}G(q_n)\delta q_n)_{T_n}.$$

The ratio of these two quantities is defined as

$$\eta_n = \rho_n^{\mathrm{act}} / \rho_n^{\mathrm{pred}}.$$

By comparing the parameter η_n to one, we have some information about the error from both approximating f by a quadratic function and from approximating $P_c(q_n + \delta q_n) - P_c(q_n)$ by $\mathrm{D}P_c(q_n)\delta q_n$. For algorithmic purposes, we define two constants $0 < \eta^{(1)} < \eta^{(2)} < 1$. If $\eta_n > \eta^{(1)}$ we are satisfied with the objective function decrease and will accept the step, otherwise we will reject it by decreasing the trust region radius. If $\eta_n > \eta^{(2)}$, we will accept the step and additionally increase the trust region radius. A pseudo-code description is given in Algorithm 3.

At present, we are unable to give a satisfactory convergence analysis of Algorithm 3, even though it seems to perform well in practice. Concerning the global convergence behavior, it would be desirable to show that it performs no worse that the first order method (3.27) with an

Algorithm 2 Steihaug cg-method with final step

$r_0 = b = -G(q)$
$d_0 = r_0$
$\delta q_0 = 0$
for $k = 0, 1, \ldots$ **do**
  compute $Ad_k = \mathrm{D}G(q)d_k \in H$
  if $(d_k, Ad_k)_T \leq 0$ **then**
    $\delta q_{k+1} = \delta q_k + \zeta d_k$ with $\zeta > 0$ such that $\|\delta q_{k+1}\|_T = \sigma$
    return δq_{k+1} {"negative curvature"}
  end if
  $\beta_k = \frac{\|r_k\|_T^2}{(d_k, Ad_k)_T}$
  if $\|\delta q_k + \beta_k d_k\|_T < \sigma$ **then**
    $\delta q_{k+1} = \delta q_k + \beta_k d_k$
  else
    $\delta q_{k+1} = \delta q_k + \zeta d_k$ with $\zeta > 0$ so that $\|\delta q_{k+1}\|_T = \sigma$
    return δq_{k+1} {"trust region left"}
  end if
  $r_{k+1} = r_k - \beta_k Ad_k$
  if <tolerance reached> **then**
    compute final step with (3.25) (or (3.26))
    return $\delta q_{k+2} = \delta q_{k+1} + \theta r_{k+1}$ {"converged"}
  end if
  $d_{k+1} = r_{k+1} + \frac{\|r_{k+1}\|_T^2}{\|r_k\|_T^2} d_k$
end for

Algorithm 3 Trust region method

initial $q_0 \in H$
initial $\sigma_0 > 0$
for $n = 0, 1, 2, \ldots$ **do**
  $T_n = \mathrm{D}P_c(q_n)$
  compute δq_n from (3.29) with Algorithm 2
  if $\eta_n > \eta^{(1)}$ **then**
    if $\eta_n > \eta^{(2)}$ **then**
      <increased σ_{n+1}>
    end if
  else
    $q_{n+1} = q_n$
    <decreased σ_{n+1}>
  end if
end for

adaptive (e.g., Armijo-type) step-size. However, in contrast to the classical trust-region method (in the smooth setting) this appears difficult, since the error between the update $u_{n+1} - u_n$ and the linearized approximation $T_n(q_{n+1} - q_n) = T_n \delta q_n$ cannot be controlled in a systematic way.

From this point of view a model seems preferable, which is, e.g., of the form

$$\tilde{\rho}_n^{\mathrm{pred}} = \tilde{Q}_{q_n}(\delta q_n) = (G(q_n), u_{n+1} - u_n) + \frac{1}{2}(\delta q_n, \mathrm{D}G(q_n)\delta q_n)_{T_n},$$

where $u_{n+1} = P_c(q_n + \delta q_n)$ and $u_n = P(q_n)$. Such a model would provide a guarantee for the behavior of η_n in the small radius case (cf. the proof of Proposition 3.28), which could ensure that sufficiently large steps can be guaranteed. However, the discrepancy between this model and the natural quadratic model (for the Newton-method) would have to be taken into account in Algorithm 2. It would also be desirable to prove an eventual transition to fast local convergence (i.e., full steps are taken).

Remark 3.9. i) The trust-region method as given in Algorithm 3 fails in a corner-case which appears frequently for initial guesses q_0 far from the optimum. For instance, it appears for $\psi(\cdot) = \|\cdot\|_{L^1}$ when we initialize $q_0 = 0$, which results in $\mathrm{D}P_c(q_0) = T_0 = 0$ (all points are "fixed"). From the point of view of the first-order method with a constant step-size this is not problematic; cf. Theorem 3.27. However, for $T = 0$ the radius constraint in (3.29) becomes meaningless, and the given pseudo-code fails. One practically motivated solution approach is to provide a special case for $T = 0$ in the implementation (e.g., an Armijo-line search). Another approach is to change the radius computations to work with the full norm $\|\delta q\|$ instead of $\|\delta q\|_T$. Then, we also modify the final step (3.25) to take the radius constraint into account. Note however, that we lose the monotonicity of the size of the update δq_k in the cg-method. The quantity $\|\delta q_k\|_T$ is monotonously increasing in each step k; see [Ste83].

ii) Let us comment on the final step (3.25), which is the only modification of Algorithm 2 w.r.t. the original version in [Ste83]. This step is performed only in the case of convergence up to a sufficient tolerance. In the present theory, where the quadratic problem can be solved exactly on the smaller space H_T, this step is completely invisible from the point of view of the quadratic model, since it lies in the kernel of T. In practice, where a specific step $\delta q_m \in \mathcal{K}_m \subset H$ is computed by the cg-method (which also contains contributions in the kernel of T), the influence of this step is hard to judge: even though the cg-method guarantees only that $\|R(\delta q_m)\|_T$ is smaller than a prescribed tolerance, computational experience suggests that the full residual norm $\|R(\delta q_m)\|$ is generally of the same order of magnitude. Therefore, after computing δq_m up to a sufficiently high tolerance, the final step is usually negligible. A mathematical analysis (for the case $c = \gamma$) suggest that this is related to the clustering of the eigenvalues of $\mathrm{D}G(q) = \gamma\,\mathrm{Id} + \nabla^2 f(u)T\colon H_T \to H_T$ around γ, which results from the typical compactness of the operator $\nabla^2 f(u)$. This effect, which supports the outlined solution approach, can only be expected if the Krylov space $\mathcal{K}_m$ is sufficiently large and it is possible to construct counterexamples. However, the known counterexamples rely on finite termination of the cg-method, and also support the proposed implementation.

The problems considered in the following chapters are convex (even linear-quadratic with exception of the cost term ψ). Practical experience shows that, in combination with a continuation strategy in γ, the local semismooth Newton method generally exhibits global convergence in practice for these problems. Therefore, a globalization strategy is not needed for the following numerical experiments. For computational results obtained with the outlined trust-region method we instead refer to Kunisch, Pieper, and Rund [KPR14] where a time optimal control problem for the monodomain equations (an instationary reaction-diffusion system arising in cardiac electrophysiology) is solved with this algorithm. Additional computational results obtained with this algorithm can be found in Springer [Spr15].

3.6. Other reformulations

In the semismooth Newton literature, methods for problems of the structure $(\mathcal{P}_\gamma)$ are usually based on a different reformulation of the optimality condition. In this section we compare the semismooth Newton algorithm based on the normal map to other, more common formulations.

3.6.1. A reformulation based on the "natural residual"

In this case, the optimization method operates directly on the control u. For the nonsmooth reformulation, we define

$$F(u) = c\left[u - P_c\left(u - \frac{1}{c}\nabla f_\gamma(u)\right)\right] \tag{3.30}$$

for an arbitrary constant $c > 0$, which is sometimes referred to as the "natural residual". It is easy to see that the stationarity condition for $(\mathcal{P}_\gamma)$ is equivalent to $F(u) = 0$.

Remark 3.10. Usually, we would leave out the additional scaling factor c. We add it here for easier comparison with the definition of G as in (3.2). Clearly, the scaling factor does not matter for the purposes of a Newton-type method due to the affine invariance property.

In each step of the semismooth Newton method, we need to solve the Newton equation

$$\mathrm{D}F(u)\,\delta u = -F(u) \tag{3.31}$$

for the update δu, which will be applied to the variable u directly. The Newton operator of (3.30) is given as

$$\mathrm{D}F(u) = c\,(\mathrm{Id} - T_+) + T_+\,\nabla^2 f_\gamma(u). \tag{3.32}$$

In this case, the operator T_+ is given by

$$T_+ := \mathrm{D}P_c\left(u - \frac{1}{c}\nabla f_\gamma(u)\right).$$

Note, that in this formulation we make a quadratic approximation for f at the current iterate for the control u, whereas the generalized differential of the proximal map is evaluated at the shifted point $u^+ := u - 1/c\,\nabla f_\gamma(u)$. For this reason, it is not possible to directly relate $\mathrm{D}F(u)$ to a quadratic approximation of j_γ at the point u, which we could do for the Newton operator $\mathrm{D}G(q)$. However, in the optimum, both Newton matrices are transposes of each other.

Proposition 3.30. *Suppose that $\bar{q} \in H$, such that $G(\bar{q}) = 0$ and set $\bar{u} = P_c(\bar{q})$. Then we have*

$$\mathrm{D}F(\bar{u}) = \mathrm{D}G(\bar{q})^*.$$

Proof. In a stationary point we have the identity $c\bar{q} = c\bar{u} + \nabla f(\bar{u})$. From this follows that $T_+ = \mathrm{D}P_c(\bar{u}^+) = \mathrm{D}P_c(\bar{q}) = T$. Clearly, we have $[c(\mathrm{Id} - T) + T\,\nabla^2 f_\gamma(\bar{u})]^* = c(\mathrm{Id} - T) + \nabla^2 f_\gamma(\bar{u})T$, which implies the claim. □

Proposition 3.31. *Under the conditions of Lemma 3.15, for any $u \in H$, the Newton operator $\mathrm{D}F(u)\colon H \to H$ is boundedly invertible.*

Proof. We have the identity

$$\mathrm{D}F(u)^* = c\,(\mathrm{Id} - T_+) + \nabla^2 f_\gamma(u) T_+.$$

By the same argumentation as in the proof of Lemma 3.15, using the structural properties of T_+ instead of T, the operator on the right hand side is boundedly invertible on H. This implies the claim with

$$\|\mathrm{D}F(u)^{-1}\|_{H\to H} = \|\mathrm{D}F(u)^{-*}\|_{H\to H} = \|[c(\mathrm{Id} - T_+) + \nabla^2 f_\gamma(u) T_+]^{-1}\|_{H\to H}. \qquad \square$$

For the solution of this system we can choose between two approaches. On the one hand, we can iteratively solve $\mathrm{D}F(u)\delta u = -F(u)$ by using a Krylov subspace method which can work on non-symmetric matrices. The most obvious choice would probably be the GMRES method, which has both a well understood convergence theory and easily available stable implementations. On the other hand, we can again reduce the system to a symmetric form. First, we multiply (3.31) from the left by $Q\colon H \to \operatorname{Ker} T_+ \subset H$, defined as the orthogonal projection to $\operatorname{Ker} T_+$. We obtain the explicit relation

$$c\,Q\,\delta\tilde{u} = -QF(u),$$

where we have used the identity $QT_+ = T_+Q = 0$ (T_+ and Q are self-adjoint). Now, we split $\delta u = \delta u_1 + \delta u_2$, where $\delta u_2 = Q\delta u = -1/c\,QF(u) \in \operatorname{Ker} T_+$ and $\delta u_1 \in (\operatorname{Ker} T_+)^\perp$ and obtain

$$\mathrm{D}F(u)\delta u_1 = -F(u) - \mathrm{D}F(u)\delta u_2.$$

Finally, by (formally) parametrizing δu_1 as $\delta u_1 = T_+\delta\tilde{u}$, we obtain the symmetric system

$$\mathrm{D}F(u)T_+\delta\tilde{u} = \left(cT_+ + T_+(\nabla^2 f_\gamma(u) - c)T_+\right)\delta\tilde{u} = -F(u) - \mathrm{D}F(u)\delta u_2$$

in terms of the unknown $\delta\tilde{u}$. We leave out a detailed rigorous justification of the last step at this point. It can be done as in the proof of Lemma 3.15. Note, that this system can again be solved with conjugate gradients as discussed in section 3.4.1.

The important special case

As in the case of the reformulation with the normal map, for the Banach space analysis, we have to suppose that $\gamma > 0$. Choosing $c = \gamma$ gives

$$F(u) = \gamma u - \gamma P_\gamma\left(-\frac{1}{\gamma}\nabla f(u)\right),$$

which again leads to a Newton method that allows for a function space analysis. Here, we consider F as an operator from H to H. In this case, the Newton operator simplifies to

$$\mathrm{D}F(u) = \gamma + T_+\,\nabla^2 f(u).$$

This is the most popular formulation in the (infinite dimensional) semismooth Newton literature; cf., e.g., [HIK03; Sta09; Ulb11; HSW12; HV12].

The linear quadratic case

In general, an approach based on the normal map (3.2) will lead to a different Newton method than an approach based on the natural residual (3.30). However, for quadratic f, this gap can be closed by a correct interpretation. In this case, the Hessian of f is constant on H, i.e.,

$$\nabla^2 f(u) = \nabla^2 f \qquad \text{for all } u \in H.$$

The discrepancy between both methods vanishes if we relate the optimization variable q to the shifted point $u_+ = u - 1/c\,\nabla f_\gamma(u)$. We will only consider the case $c = \gamma > 0$ in the following. Then we have $u_+ = -1/\gamma\,\nabla f(u)$ and we obtain the following result.

Proposition 3.32. *Suppose that $c = \gamma > 0$ and that f is quadratic. Furthermore, take initial iterates $q_0, u_0 \in H$, such that*

$$\gamma q_0 = -\nabla f(u_0).$$

Define the Newton iterates according to the normal map and the natural residual inductively as $q_{n+1} = q_n - \mathrm{D}G(q_n)^{-1}G(q_n)$ and $u_{n+1} = u_n - \mathrm{D}F(u_n)^{-1}F(u_n)$ for all $n \in \mathbb{N}_0$. Then we have

$$\gamma q_n = -\nabla f(u_n)$$

for all $n \in \mathbb{N}_0$. In this sense, both methods are equivalent.

Proof. Note, that the Newton iterates are well-defined, since the Newton matrices are invertible according to Lemma 3.14 and Proposition 3.31. The first step $u_1 - u_0$ fulfills the Newton equation $\mathrm{D}F(u_0)(u_1 - u_0) = -F(u_0)$, which reads as

$$\left(\gamma + T\,\nabla^2 f\right)(u_1 - u_0) = -\gamma u_0 + \gamma P_\gamma(-1/\gamma \nabla f(u_0)),$$

where $T = \mathrm{D}P_\gamma(-1/\gamma\nabla f(u_0)) = \mathrm{D}P_\gamma(q_0)$. Multiplying by $-1/\gamma\,\nabla^2 f$ from the left and inserting the relation $q_0 = -1/\gamma\,\nabla f(u_0)$ we obtain

$$\left(\gamma + \nabla^2 f\,T\right)\left(-1/\gamma\,\nabla^2 f(u_1 - u_0)\right) = \nabla^2 f\,(u_0 - P_\gamma(q_0))\,.$$

Since f is quadratic, the gradient of f is affine linear, i.e., we have $\nabla f(u) - \nabla f(\tilde{u}) = \nabla^2 f(u - \tilde{u})$ for all $u, \tilde{u} \in H$. Inserting this expression on the left- and right-hand side, we obtain

$$\left(\gamma + \nabla^2 f\,T\right)(-1/\gamma\,\nabla f(u_1) - q_0) = \nabla f(u_0) - \nabla f(P_\gamma(q_0)) = \gamma q_0 - \nabla f(P_\gamma(q_0)).$$

It follows that $\tilde{q} = -1/\gamma\,\nabla f(u_1)$ solves the Newton system $\mathrm{D}G(q_0)(\tilde{q} - q_0) = -G(q_0)$. Since this system has a unique solution, it holds $q_1 = -1/\gamma\,\nabla f(u_1)$. The full result follows now by induction over $n \in \mathbb{N}_0$. □

As a corollary, in the case of a linear quadratic problem with box-constraints, we also obtain the (essential) equivalence of both approaches to the well-known primal-dual active set strategy; see [HIK03].

3.6.2. The "control reduced" approach

Another elegant approach to semismooth Newton for optimal control problems is not based on the reduced cost functional, but on a reformulation of the KKT system; see Schiela [Sch08]. This approach is particularly useful in the context of the so-called "variational discretization" concept due to Hinze [Hin05]. We will see that this approach leads to a very similar algorithm, when compared to the approach on the reduced cost functional with the normal map. To fix ideas, we consider an abstract nonlinear control problem with a control appearing linearly on the right-hand side:

$$\begin{aligned} \min_{u\in H, y\in Y} \quad & J(y) + \psi(u) + \frac{\gamma}{2}\|u\|^2, \\ \text{subject to} \quad & A(y) = Bu \quad \text{in } W^*. \end{aligned} \tag{3.33}$$

We suppose that $\gamma > 0$. We use the same notation as in section 2.5.4. Here the equation is given as

$$e(u, y) = A(y) - Bu = 0 \quad \text{in } W^*,$$

for $y \in Y$ and $u \in H$. The spaces Y and W are reflexive Banach spaces, $B\colon H \to W^*$ is a bounded linear operator and we assume that $J\colon Y \to \mathbb{R}$ and $A\colon Y \to W^*$ are C^2. Furthermore, we suppose that $A(y) = Bu$ has a unique solution for every $u \in H$ and that $A'(y)\colon Y \to W^*$ is an isomorphism for all $y \in Y$. As before, we define the Lagrange function as

$$\mathcal{L}(u, y, p) = J(y) - \langle A(y) - Bu, p\rangle \quad \text{for } (u, y, p) \in H \times Y \times W.$$

The solution operator of the state equation denoted by $S\colon u \mapsto y$ is C^2 as a consequence of the implicit function theorem (cf. section 2.5.4). We obtain the state and adjoint equations respectively as

$$Bu - A(y) = 0 \quad \text{in } W^*, \tag{3.34}$$

$$J'(y) - A'(y)^* p = 0 \quad \text{in } Y^*. \tag{3.35}$$

Define the reduced tracking functional as $f(u) = J(S(u))$. As in section 2.5.4, we obtain a representation for the gradient and Hessian of f at a point u as

$$\begin{aligned} \nabla f(u) &= B^* A'(y)^{-*} J'(y) = B^* p, \\ \nabla^2 f(u) &= B^* A'(y)^{-*}\, \mathcal{L}''_{yy}\, A'(y)^{-1} B. \end{aligned}$$

Therein $y = S(u)$ and $p = A'(y)^{-*}J'(y)$ are the corresponding state and adjoint solutions, and $\mathcal{L}''_{yy} = J''(y) - \langle A''(y)(\cdot,\cdot), p\rangle$ is the second derivative of the Lagrange function w.r.t. the state.

Now, we sketch the idea behind the control reduced approach: the stationarity condition for (3.33) is equivalent to the conditions (3.34), (3.35) and the control projection formula

$$u = P_\gamma\left(-\tfrac{1}{\gamma} B^* p\right).$$

Here, the right-hand side of the control projection formula depends solely on the adjoint state. Now, we insert this expression into the state equation to obtain the following optimality system, formulated in terms of the state and adjoint variable as

$$F(y, p) = \begin{pmatrix} J'(y) - A'(y)^* p \\ B P_\gamma\left(-1/\gamma\, B^* p\right) - A(y) \end{pmatrix} = 0.$$

To apply a Newton method to the equation using the semismoothness concept, we derive the (generalized) derivative as

$$\mathrm{D}F(y,p) = \begin{pmatrix} \mathcal{L}''_{yy} & -A'(y)^* \\ -A'(y) & -1/\gamma\, BTB^* \end{pmatrix},$$

where $T = \mathrm{D}P_\gamma\,(-1/\gamma B^* p)$ and $\mathcal{L}''_{yy}$ as before.

Proposition 3.33. *Take $p \in W$ and set $u = P_\gamma(-1/\gamma\, B^* p)$ and $y = S(u)$. Suppose that $(\delta y, \delta p) \in Y \times W$ solves the Newton system*

$$\mathrm{D}F(y,p)(\delta y, \delta p) = -F(y,p).$$

Then, $\delta q = -1/\gamma\, B^* \delta p$ solves the Newton equation

$$\mathrm{D}G(q)\delta q = -G(q),$$

for $q = -1/\gamma\, B^* p$, where $G(q) = \gamma q + \nabla f(P_\gamma(q))$, as before.

Proof. Writing out the system, we obtain

$$\begin{aligned} \mathcal{L}''_{yy}\delta y - A'(y)^*\delta p &= -J'(y) + A'(y)^* p, \\ -A'(y)\delta y - 1/\gamma\, BTB^*\delta p &= 0. \end{aligned}$$

Note, that the right-hand side of the second equation is zero, since $y = S(u)$. Now, we perform a Schur complement reduction by inserting $\delta y = -A'(y)^{-1}(1/\gamma\, BTB^*\delta p)$ into the first equation. We obtain

$$-A'(y)^*\delta p - 1/\gamma\, \mathcal{L}''_{yy}A'(y)^{-1}BTB^*\,\delta p = -J'(y) + A'(y)^* p.$$

Applying $B^*A'(y)^{-*}$ from the left and introducing the auxiliary variables $q = -1/\gamma\, B^* p$ and $\delta q = -1/\gamma\, B^*\delta p$, we end up with

$$\gamma\,\delta q + B^*A'(y)^{-*}\mathcal{L}''_{yy}A'(y)^{-1}B\,T\delta q = -\gamma q - B^*A'(y)^{-*}J'(y).$$

By the representation formulas for the first and second derivative, this is the same as $\gamma\delta q + \nabla^2 f(u)T\delta q = -\gamma q - \nabla f(u)$. □

As a consequence of this, we obtain the following result: combining the control reduced Newton method with a projection onto the state manifold $y = S(u)$ in each step, we obtain an equivalent algorithm to the approach from section 3.2.

3.6.3. Comparison

Let us compare the three presented approaches: We have seen that in the case of a quadratic f (i.e., linear quadratic control problems), the Newton methods resulting from all approaches are essentially equivalent; see Propositions 3.31 and 3.33. Therefore, there seems to be no clear advantage of any method over the other in this situation. However, differences arise in the context of discrete approximations to the infinite-dimensional control problem, for instance in the context of the "variational discretization" concept as in [Hin05]. Here the control variable is only discretized implicitly via $q_\sigma = P_\gamma(-1/\gamma\, B^* p_\sigma) = P_\gamma(q_\sigma)$, where $q_\sigma = -1/\gamma\, B^* p_\sigma$ is searched in a finite dimensional space, but the proximal map is still evaluated analytically.

Note, that in this case the control variable is not a discrete quantity and it requires additional implementation effort to evaluate and store it directly; see [HV12]. The control reduced approach offers an advantage here, since the control is only evaluated implicitly (e.g., with specialized quadrature formulas or adaptive quadrature; cf. [WGS08], where the control reduced approach is used in conjunction with an interior point reformulation). The same is true for the approach with the normal map, where only the auxiliary variable q is stored; cf. also [Spr15] for a more detailed discussion.

In the case of nonquadratic f (i.e., for control problems with nonlinear state equations), an appropriate globalization is a crucial issue. Here, the normal map approach seems to offer some advantages over the natural residual formulation, as discussed in section 3.6.1. We have seen that the Newton system can be directly related to a quadratic model for the objective function in the current iterate $u = P_c(q)$ and that a globalization strategy can be based on the cost functional in a, more or less, straightforward manner. This is not the case in the context of the natural residual: There, the globalization is usually based on the squared residual $\|F(u)\|^2$; see [IK08; IK09]. This is mainly due to the fact that the semismooth Newton method as in section 3.6.1 produces iterates which are in general inadmissible w.r.t. the constraint $u \in U_{\text{ad}}$. Certainly, a globalization based on the residual is also an option, but it appears to be not the canonical choice in the context of functional minimization. Note however, that it is also possible to use the reduced cost functional in this setting if one introduces an additional projection to the constraint set U_{ad} in each step; see [Ulb11]. Still, the normal map seems to offer a more direct approach, since feasibility is automatically fulfilled. Moreover, when using the cg-method, there appears to be a natural transition between a first order and the full semismooth Newton step.

For the control reduced approach to semismooth Newton, there appears to be no globalization approach in the literature, to the best of the authors knowledge. However, in this context, we can mention the approach by Gräser and Kornhuber [Grä08; GK09] (cf. also [HV12]), where a dual functional is constructed and descent is required w.r.t. this functional. The corresponding method can also be interpreted as operating on the adjoint variable [GK09, Section 5.1], and the theory covers a line-search based on a Newton-like direction. However, this approach seems to be limited to problems with linear state equations.

Let us also mention other related methods for a similar classes of optimization problems. On the one hand there are related versions of the SQP-method where a semismooth Newton method is used as an inner solver for the quadratic problem with box-constraints; see, e.g., [HH02; HH06] and the references therein. On the other hand there are nonlinear variants of the primal-dual active set method; see, e.g., [BIK99; KR02; IK04]. In both of these approaches, the optimization is split into an inner and an outer loop. Roughly speaking, in the former case, the outer iteration linearizes only the smooth parts of the Lagrange function and keeps the nonsmooth part (the box-constraints) intact. In the latter case, the constraints are linearized by introducing appropriate active sets and a resulting nonquadratic but smooth optimization problem has to be solved in the inner loop.

4. A priori error analysis for an elliptic problem

In this chapter we will consider a priori error estimates for the elliptic model problem given by

$$\min_{u\in\mathcal{M}(\Omega_c),\, y\in Y} \quad \frac{1}{2}\|y-y_d\|^2_{L^2(\Omega_o)} + \alpha\|u\|_{\mathcal{M}(\Omega_c)}, \tag{4.1a}$$

$$\text{subject to} \quad \begin{cases} -\Delta y = \chi_{\Omega_c} u & \text{in } \Omega, \\ y = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1b}$$

Here, $\Omega \subset \mathbb{R}^d$ for $d \in \{2, 3\}$ is a convex bounded domain with a $\mathcal{C}^{2,\beta}$-boundary $\partial\Omega$. The control variable u is searched for in the space of regular Borel measures $\mathcal{M}(\Omega_c)$, where the control set $\Omega_c \subset \Omega$ is relatively closed in Ω, i.e.,

$$\Omega_c = \bar{\Omega}_c \setminus \partial\Omega.$$

We will make additional assumption on the form of Ω_c below (such as $\partial\Omega_c \setminus \partial\Omega$ polygonal). The state variable y is the solution of the Poisson equation (4.1b). We consider a standard linear quadratic tracking term on the observation domain $\Omega_o \subset \Omega$ with desired state y_d given in $L^2(\Omega_o)$. For the purpose of optimal regularity and error estimates we will make further assumptions, such as $y_d \in L^p(\Omega_o)$ or $y_d \in L^\infty(\Omega_o)$; see below. The parameter α is assumed to be positive.

To discretize the problem (4.1), we consider linear finite elements in space and a discretization for the control by nodal Dirac delta functions as proposed by Casas, Clason, and Kunisch [CCK12]. We derive estimates for the objective functional of order $\mathcal{O}(h^{4-d}|\ln h|^\kappa)$ and for the error of the state on the observation domain of order $\mathcal{O}(h^{2-d/2}|\ln h|^{\kappa/2})$, which improves on the previous analysis by essentially doubling the rates. The results have already appeared in similar form in Pieper and Vexler [PV13]. We achieve this by a careful study of the regularity, L^p estimates for $p \neq 2$, and by employing uniform finite element error estimates due to Rannacher and Frehse [FR76] and Rannacher [Ran76]. Due to the analogy of problem (4.1) with a state constrained optimal control problem (cf. section 2.4), we also refer to Deckelnick and Hinze [DH07], where similar L^p techniques using uniform finite element estimates have been introduced for the analysis of a state constrained problem (cf. also [Mey08]).

Furthermore, in the case where $\Omega = \Omega_o = \Omega_c$, we prove a regularity result for the optimal solutions. Under the assumption that the desired state is bounded, we show that the optimal control $\bar{u}$ is an element of $H^{-1}(\Omega)$, which rules out Dirac delta functions. In this case, we can derive an improved convergence rate for the optimal states in the three dimensional case. This result has also already appeared in [PV13]. Let us point out that the mentioned regularity result is also new if transferred to a state constrained problem (cf. section 2.4). In fact, an analogous, more general result has subsequently been derived in this context by Casas, Mateos, and Vexler [CMV14].

Furthermore, we also consider a discretization of the regularized version of problem (4.1). Motivated by the discretization of the original problem, we consider a discretization of the

control based on linear finite elements and mass lumping, which is different from a similar discretization concept proposed in [CHW12a]. For a fixed regularization parameter, we are able to provide an estimate for a post-processing of the optimal control of optimal order $\mathcal{O}(h^2)$ (under an established structural assumption on the optimal solution). Since the regularized problem is similar to a control constrained problem, this essentially transfers the post-processing results due to Meyer and Rösch [MR04] for a piece-wise constant discretization to a piece-wise linear setting.

The chapter is structured as follows. In section 4.1 we discuss regularity of the state and adjoint state and derive some structural consequences of the optimality conditions. The discretization of the state and control is introduced in section 4.2. In section 4.3 we derive the error estimates for the optimal solutions which are valid in the general setting. Section 4.4 contains the estimates which are valid under additional conditions on the control and observation sets: we give the additional regularity result and the improved estimate, as mentioned above. The regularized problem is discussed in section 4.5. We give an asymptotic estimate for the regularization error based on a technique introduced by Hintermüller, Schiela, and Wollner [HSW14]. We also provide the finite element error analysis for the regularized problem. Numerical results are given in section 4.6.

Throughout this chapter, we denote by $(\cdot,\cdot)$ the $L^2(\Omega)$ inner product and by $\langle\cdot,\cdot\rangle$ the duality product between $\mathcal{M}(\Omega)$ and $\mathcal{C}_0(\Omega)$.

4.1. Precise regularity and optimality conditions

In the following, we derive precise regularity estimates for the state solution, summarize the existence and optimality theory for the optimization problem and derive optimality conditions. As the first step we recall the weak formulation of the state equation (4.1b). For a given $u \in \mathcal{M}(\Omega)$ the solution y is determined by

$$y \in W^{1,s}(\Omega) : \ (\nabla y, \nabla \varphi) = \langle \chi_{\Omega_c} u, \varphi \rangle \quad \text{for all } \ \varphi \in W_0^{1,s'}(\Omega).$$

By the general theory in section 2.2.2, the above formulation possesses a unique solution for all $1 \leq s < d/(d-1)$. Moreover, the behavior of the constant in the a priori estimate for $s \to d/(d-1)$ can be estimated in the following way.

Lemma 4.1. *For $\varepsilon > 0$ small enough, let s_ε be given as*

$$s_\varepsilon = \frac{d}{d-1} - \varepsilon.$$

There exists a constant C independent of ε, such that for all $u \in \mathcal{M}(\Omega)$ and the corresponding solution y of (4.1b) *the following estimate holds:*

$$\|y\|_{W_0^{1,s_\varepsilon}(\Omega)} \leq \frac{C}{\varepsilon} \|u\|_{\mathcal{M}(\Omega)}.$$

Proof. To obtain the precise dependence of ε we use the continuous embedding of $W_0^{1,s'_\varepsilon}(\Omega)$ into $\mathcal{C}_0(\Omega)$, where s'_ε is the conjugate index of s_ε with

$$\frac{1}{s'_\varepsilon} + \frac{1}{s_\varepsilon} = 1, \quad s'_\varepsilon > d.$$

With [Alt11, Theorem 8.10] we obtain

$$\|v\|_{\mathcal{C}_0(\Omega)} \leq \frac{C}{\varepsilon} \|v\|_{W_0^{1,s'_\varepsilon}(\Omega)} \quad \text{for all } v \in W_0^{1,s'_\varepsilon}(\Omega)$$

with a constant C independent of ε. Using the fact that the Dirichlet Laplacian is an isomorphism between $W_0^{1,s'_\varepsilon}(\Omega)$ and $W^{-1,s'_\varepsilon}(\Omega)$ for ε small enough as in Theorem 2.12 (see also [Mey63; AK92]), we estimate

$$\|\nabla y\|_{L^{s_\varepsilon}(\Omega)} \leq \sup_{f \in L^{s'_\varepsilon}(\Omega,\mathbb{R}^d)} \frac{(\nabla y, f)}{\|f\|_{L^{s'_\varepsilon}(\Omega)}} \leq C \sup_{v \in W_0^{1,s'_\varepsilon}(\Omega)} \frac{(\nabla y, \nabla v)}{\|\nabla v\|_{L^{s'_\varepsilon}(\Omega)}},$$

where C is independent of ε. Since y is the solution of the state equation corresponding to u we obtain

$$\|\nabla y\|_{L^{s_\varepsilon}(\Omega)} \leq C \sup_{v \in W_0^{1,s'_\varepsilon}(\Omega)} \frac{\langle u, v\rangle}{\|\nabla v\|_{L^{s'_\varepsilon}(\Omega)}} \leq \frac{C}{\varepsilon} \|u\|_{\mathcal{M}(\Omega)},$$

which completes the proof. □

In the following we use the same notation as in section 2.2. We restrict the parameter s to the interval $[2d/(d+2), d/(d-1))$, unless explicitly stated otherwise. We also denote the state solution y corresponding to $u \in \mathcal{M}(\Omega_c)$ by $y = S(u)$, where $S\colon \mathcal{M}(\Omega_c) \to W_0^{1,s}(\Omega)$ is the corresponding solution operator. With the continuous embedding of $W_0^{1,s}(\Omega)$ into $L^2(\Omega)$, a reduced cost functional can be defined defined for (4.1). We define for $y \in W_0^{1,s}(\Omega)$ the tracking functional by

$$J(y) = \frac{1}{2}\|\chi_{\Omega_o} y - y_d\|^2_{L^2(\Omega_o)},$$

where $\chi_{\Omega_o}\colon W_0^{1,s}(\Omega) \to L^2(\Omega_o)$ denotes the canonical embedding and restriction. Note that, in the following, by an abuse of notation, χ_{Ω_o} will also denote the characteristic function of Ω_o. For $u \in \mathcal{M}(\Omega_c)$ we define the reduced cost functional by

$$j(u) = J(S(u)) + \alpha\|u\|_{\mathcal{M}(\Omega_c)}.$$

As in Theorem 2.3 we obtain optimal solutions of (4.1).

Proposition 4.2. *The problem* (4.1) *possesses at least one optimal solution* $(\bar u, \bar y) = (\bar u, S(\bar u)) \in \mathcal{M}(\Omega_c) \times W_0^{1,s}(\Omega)$.

Note that uniqueness of the optimal solutions can only be expected under further conditions on Ω_c and Ω_o, which we will investigate in section 4.4. Nevertheless, the observation of the optimal state is unique.

Proposition 4.3. *For two optimal solutions of the problem* (4.1) *the values* $\chi_{\Omega_o}\bar y$ *coincide.*

Proof. The functional $1/2\,\|\cdot - y_d\|^2_{L^2(\Omega_o)}$ is strictly convex on $L^2(\Omega_o)$. Take two optimal solutions $(\bar u, \bar y)$ and $(\tilde u, \tilde y)$. Define the mean value as $u_{1/2} = (\bar u + \tilde u)/2$ and $y_{1/2} = (\bar y + \tilde y)/2 = S(u_{1/2})$. We have

$$j(u_{1/2}) = \frac{1}{2}\|\chi_{\Omega_o} y_{1/2} - y_d\|^2_{L^2(\Omega_o)} + \alpha\|u_{1/2}\|_{\mathcal{M}(\Omega_c)} \leq \frac{1}{2}\left(j(\bar u) + j(\tilde u)\right) = j(\bar u) = j(\tilde u).$$

by convexity of the reduced cost functional j. If we suppose now that $\chi_{\Omega_o}\bar y \neq \chi_{\Omega_o}\tilde y$, we obtain even strict inequality, which contradicts the optimality of $(\bar u, \bar y)$ and $(\tilde u, \tilde y)$. □

Moreover, the following optimality system can be obtained; see section 2.2.4.

Theorem 4.4. *There exists a unique adjoint state $\bar{p} \in W_0^{1,q}(\Omega)$ (with $q > d$) corresponding to any optimal solution $(\bar{u}, \bar{y}) = (\bar{u}, S(\bar{u}))$ of (4.1). It satisfies*

$$\begin{cases} -\Delta \bar{p} = \chi_{\Omega_o}(\bar{y} - y_d) & \text{in } \Omega, \\ \bar{p} = 0 & \text{on } \partial\Omega \end{cases} \tag{4.2}$$

in the sense of the standard weak formulation and

$$-\langle \chi_{\Omega_c}(u - \bar{u}), \bar{p} \rangle + \alpha \|\bar{u}\|_{\mathcal{M}(\Omega_c)} \leq \alpha \|u\|_{\mathcal{M}(\Omega_c)} \quad \text{for all } u \in \mathcal{M}(\Omega_c). \tag{4.3}$$

Furthermore, the variational inequality (4.3) is equivalent to the two conditions

$$\|\chi_{\Omega_c} \bar{p}\|_{\mathcal{C}_0(\Omega_c)} \leq \alpha, \quad \text{and } \langle \chi_{\Omega_c} \bar{u}, \bar{p} \rangle = \alpha \|\bar{u}\|_{\mathcal{M}(\Omega_c)}. \tag{4.4}$$

This implies that the support of $\bar{u}$ is contained in the set $\{ x \in \Omega_c \mid |\bar{p}(x)| = \alpha \}$, and for the Jordan-decomposition $\bar{u} = \bar{u}^+ - \bar{u}^-$ we have

$$\operatorname{supp} \bar{u}^+ \subset \{ x \in \Omega_c \mid \bar{p}(x) = -\alpha \} \quad \text{and} \quad \operatorname{supp} \bar{u}^- \subset \{ x \in \Omega_c \mid \bar{p}(x) = \alpha \}. \tag{4.5}$$

Remark 4.1. The optimality condition (4.3) can be equivalently reformulated as

$$(S(u) - \bar{y}, \chi_{\Omega_o}(\bar{y} - y_d)) + \alpha \|u\|_{\mathcal{M}(\Omega_c)} - \alpha \|\bar{u}\|_{\mathcal{M}(\Omega_c)} \geq 0 \quad \text{for all } u \in \mathcal{M}(\Omega_c). \tag{4.6}$$

The statement of the above theorem directly implies the following corollary on the structure of the optimal control $\bar{u}$.

Corollary 4.5. *There exist a constant $\eta > 0$, depending on the data of the problem, such that*

$$\operatorname{supp} \bar{u} \subset \Omega_\eta = \{ x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) > \eta \}, \tag{4.7}$$

and additionally

$$\operatorname{dist}(\operatorname{supp} \bar{u}^+, \operatorname{supp} \bar{u}^-) > \eta. \tag{4.8}$$

The first property implies that the support is compact.

Proof. The adjoint state $\bar{p}$ belongs to $W^{1,q}(\Omega)$ with $q > d$ and $W^{1,q}(\Omega) \hookrightarrow \mathcal{C}^\beta(\bar{\Omega})$ with some $\beta > 0$. This implies (due to the homogeneous Dirichlet boundary conditions) the existence of $\eta > 0$, such that

$$|\bar{p}(x)| < \frac{\alpha}{2} \quad \text{for } x \in \Omega \setminus \Omega_\eta.$$

We complete the first part of the proof using the statement on the support of $\bar{u}$ from Theorem 4.4. With a similar argument we derive the second statement, since due to (4.5), the adjoint state attains the values $\pm\alpha$ respectively on the support of $\bar{u}^-$ and $\bar{u}^+$. □

For the derivation of the error estimate, we also need a standard regularity result for the Poisson equation; see, e.g., [Gri85, Theorem 2.4.2.5].

Theorem 4.6. *(Elliptic regularity) Let $f \in L^q(\Omega)$ for some $q \in [2, \infty)$. The unique solution $w \in H_0^1(\Omega)$ to the elliptic problem*

$$\begin{cases} -\Delta w = f & \text{in } \Omega \\ w = 0 & \text{on } \partial\Omega, \end{cases}$$

has the regularity $w \in W^{2,q}(\Omega)$ with the corresponding a priori estimate

$$\|\nabla^2 w\|_{L^q(\Omega)} \leq C_q \|w\|_{L^q(\Omega)}.$$

4.2. Discretization

For the discretization of the state equation we use linear finite elements on a family of shape regular, quasi-uniform triangulations $\{\mathcal{T}_h\}_h$; see, e.g., [BS08]. The discretization parameter h_K denotes the diameter of the cells $K \in \mathcal{T}_h$ and $h = \max_{K\in\mathcal{T}_h} h_K$ the maximum thereof. We set

$$\bar{\Omega}_h = \bigcup_{K\in\mathcal{T}_h} \bar{K}$$

and make the usual assumption on the approximation of the boundary:

$$\operatorname{dist}(\partial\Omega, \partial\Omega_h) \leq Ch^2.$$

The space of linear finite elements associated with $\mathcal{T}_h$ is defined as usual by

$$V_h = \left\{ v_h \in \mathcal{C}_0(\Omega) \quad \middle| \quad v_h|_K \in \mathcal{P}_1(K) \quad \text{for all } K \in \mathcal{T}_h \text{ and } v_h = 0 \text{ on } \Omega \setminus \Omega_h \right\}.$$

For a given $u \in \mathcal{M}(\Omega_c)$ the discrete state solution $y_h = S_h(u) \in V_h$ is determined by

$$(\nabla y_h, \nabla \varphi_h) = \langle \chi_{\Omega_c} u, \varphi_h \rangle \quad \text{for all } \varphi_h \in V_h. \tag{4.9}$$

We denote the corresponding reduced cost functional for $u \in \mathcal{M}(\Omega_c)$ by

$$j_h(u) = J(S_h(u)) + \alpha \|u\|_{\mathcal{M}(\Omega_c)}.$$

To define the approximation of the optimal control problem (4.1) we follow the approach from [CCK12] and do not discretize the control space (cf. also the variational control discretization approach by Hinze [Hin05]). The discrete optimal control problem is then given as

$$\min_{u\in\mathcal{M}(\Omega_c)} \quad j_h(u). \tag{4.10}$$

The existence of optimal solutions can be shown as on the continuous level. Since S_h maps an infinite dimensional space to a finite dimensional one, it can not be injective. Therefore, and due to missing strict convexity of $\|\cdot\|_{\mathcal{M}(\Omega_c)}$, the solutions of (4.10) are highly non-unique. However, following [CCK12], we can identify a class of solutions which is numerically accessible.

By $\{ x_n \mid n = 1, \dots, N_h \}$ we denote the interior nodes of Ω_h and by $\{ e_n \mid n = 1, \dots, N_h \} \subset V_h$ the corresponding nodal basis functions. The standard nodal interpolation operator $i_h \colon \mathcal{C}_0(\Omega) \to V_h \subset \mathcal{C}_0(\Omega)$ is given as

$$i_h v = \sum_{n=1}^{N_h} v(x_n) e_n.$$

Furthermore, we introduce the space $\mathcal{M}_h$ consisting of linear combination of Dirac delta functions associated with the nodes x_n as

$$\mathcal{M}_h = \left\{ u_h = \sum_{n=1}^{N_h} u_n \, \delta_{x_n} \quad \middle| \quad u_n \in \mathbb{R},\ n = 1, \dots, N_h \right\} \subset \mathcal{M}(\Omega).$$

Now, we additionally discretize the control space $\mathcal{M}(\Omega_c)$ as $\mathcal{M}_h \cap \mathcal{M}(\Omega_c)$. We obtain the following finite-dimensional optimization problem

$$\min_{u_h\in\mathcal{M}_h\cap\mathcal{M}(\Omega_c)} \quad j_h(u_h). \tag{4.11}$$

We will show that, under a compatibility condition on the mesh and the control set, the problems (4.10) and (4.11) are essentially equivalent. To this purpose, we define the operator $\Lambda_h \colon \mathcal{M}(\Omega) \to \mathcal{M}_h$ by

$$\Lambda_h u = \sum_{n=1}^{N_h} \langle u, e_n \rangle \, \delta_{x_n}. \tag{4.12}$$

Note that $\Lambda_h \colon \mathcal{M}(\Omega) \to \mathcal{M}_h \subset \mathcal{M}(\Omega)$ can be derived as the transpose of the nodal interpolation i_h. It has the following properties; see [CCK12, Theorem 3.1].

Theorem 4.7. *For any $u \in \mathcal{M}(\Omega)$ it holds:*

$$\begin{aligned}
&(i) && \langle \Lambda_h u, v \rangle = \langle u, i_h v \rangle \quad \text{for all } v \in \mathcal{C}_0(\Omega), \\
&(ii) && \|\Lambda_h u\|_{\mathcal{M}(\Omega)} \le \|u\|_{\mathcal{M}(\Omega)}, \\
&(iii) && \Lambda_h u \rightharpoonup^* u \quad \text{in } \mathcal{M}(\Omega) \quad \text{and } \|\Lambda_h u\|_{\mathcal{M}(\Omega)} \to \|u\|_{\mathcal{M}(\Omega)} \quad \text{for } h \to 0.
\end{aligned}$$

In the following, we suppose that the nodes of the mesh are ordered such that for some $N_c \le N_h$ it holds $\{ x_n \mid n = 1, \dots, N_h,\ x_n \in \Omega_c \} = \{ x_n \mid n = 1, \dots, N_c \}$. For the derivation of the error estimates, we have to ensure that the operator Λ_h defined in (4.12) maps the control space $\mathcal{M}(\Omega_c)$ into $\mathcal{M}(\Omega_c) \cap \mathcal{M}_h$. Therefore, we require that for each h we have

$$\Omega_c \cap \overline{\Omega}_h = \bigcup_{K \in \mathcal{T}_h^c} \overline{K}, \tag{4.13}$$

where $\mathcal{T}_h^c \subset \mathcal{T}_h$ is the collection of all the cells of the triangulation which make up the control region. Then, we can verify that

$$\Lambda_h(u) \in \mathcal{M}(\Omega_c) \cap \mathcal{M}_h \quad \text{for all } u \in \mathcal{M}(\Omega_c).$$

Based on the previous identities, it is easy to see that any optimal solution $\tilde{u}$ of (4.10) can be replaced by a discrete optimal solution $\bar{u}_h = \Lambda_h(\tilde{u}) \in \mathcal{M}(\Omega_c) \cap \mathcal{M}_h$ with the same objective value.

Proposition 4.8. *Any solution of the fully discrete problem* (4.11) *also solves the semidiscrete problem* (4.10)*. Moreover, for any other solution of* (4.10) $\tilde{u} \in \mathcal{M}(\Omega_c)$ *we have that $\Lambda_h \tilde{u} = \bar{u}_h \in \mathcal{M}_h \cap \mathcal{M}(\Omega_c)$ is an optimal solution of both problems. If the solution to* (4.1) *is unique, there holds*

$$\bar{u}_h \rightharpoonup^* \bar{u} \quad \text{in } \mathcal{M}(\Omega_c) \quad \text{and} \quad \|\bar{u}_h\|_{\mathcal{M}(\Omega_c)} \to \|\bar{u}\|_{\mathcal{M}(\Omega_c)} \quad \text{for } h \to 0.$$

Proof. We take any solution $\tilde{u} \in \mathcal{M}(\Omega_c)$ of (4.10) and set $\bar{u}_h = \Lambda_h \tilde{u}$. Due to the properties of Λ_h it holds that $S_h(\bar{u}_h) = S_h(\tilde{u})$ and $j(\bar{u}_h) \le j(\tilde{u})$. Due to the minimality of $\tilde{u}$, it follows $j_h(\bar{u}_h) = j_h(\tilde{u})$ and $\bar{u}_h$ is an optimal solution of (4.10). Since $\|\bar{u}_h\|_{\mathcal{M}(\Omega_c)}$ is uniformly bounded (by minimality of $\bar{u}_h$) we can find a $\hat{u} \in \mathcal{M}(\Omega_c)$ and select a sequence $\bar{u}_h \rightharpoonup^* \hat{u}$ in $\mathcal{M}(\Omega_c)$ for $h \to 0$. By the weak lower semicontinuity of j it follows now that

$$j(\hat{u}) \le \liminf_{h \to 0} j(\bar{u}_h) \le \lim_{h \to 0} j_h(\bar{u}_h) + \lim_{h \to 0} |j_h(\bar{u}_h) - j(\bar{u}_h)| = j(\bar{u}),$$

which implies that $\hat{u}$ is an optimal solution of (4.1). The convergence of the functional values $j_h(\bar{u}_h) \to j(\bar{u})$ and the convergence $|j_h(\bar{u}_h) - j(\bar{u}_h)| \to 0$ for $h \to 0$ will be shown in Theorem 4.12 (the additional assumptions made there can be dropped, if we are not interested in optimal rates of convergence). If $\bar{u}$ is unique, all thusly constructed subsequences have the same limit point, and the whole sequence converges. □

For future reference, let us state again the fully discrete problem. It is given by

$$\begin{aligned} \min_{u_h \in \mathcal{M}_h,\, y_h \in V_h} \quad & \frac{1}{2}\|y_h - y_d\|^2_{L^2(\Omega_o)} + \alpha \|u_h\|_{\mathcal{M}(\Omega_c)} \\ \text{subject to} \quad & (\nabla y_h, \nabla \varphi_h) = \langle \chi_{\Omega_c} u_h, \varphi_h \rangle \quad \text{for all } \varphi_h \in V_h. \end{aligned} \tag{4.14}$$

Let us point out that for any $u_h = \sum_n u_n \delta_{x_n} \in \mathcal{M}_h$ the total variation norm is simply given by the ℓ^1 norm of the underlying nodal vector:

$$\|u_h\|_{\mathcal{M}(\Omega_c)} = \sum_{n=1}^{N_c} |u_n|.$$

Furthermore, for any $u_h = \sum_n u_n \delta_{x_n} \in \mathcal{M}_h$ and $v_h = \sum_n v_n e_n \in V_h$ the duality product is given simply as the Euclidean inner product of the nodal vectors:

$$\langle \chi_{\Omega_c} u_h, v_h \rangle = \sum_{n=1}^{N_c} u_n v_n.$$

This means that a finite dimensional equivalent of (4.14) can be derived in a straightforward way, by introducing appropriate mass and stiffness matrices. For the optimal solutions the following discrete version of the optimality conditions holds, which can be derived as in the continuous case.

Theorem 4.9. *There exists a unique discrete adjoint state $\bar{p}_h \in V_h$, which, for any discrete solution $(\bar{u}_h, \bar{y}_h) \in \mathcal{M}_h \cap \mathcal{M}(\Omega_c) \times V_h$, fulfills the discrete adjoint equation*

$$(\nabla \varphi_h, \nabla \bar{p}_h) = (\bar{y}_h - y_d, \chi_{\Omega_o} \varphi_h) \quad \textit{for all } \varphi_h \in V_h, \tag{4.15}$$

and the optimality condition

$$-\langle \chi_{\Omega_c}(u - \bar{u}_h), \bar{p}_h \rangle + \alpha \|\bar{u}_h\|_{\mathcal{M}(\Omega_c)} \leq \alpha \|u\|_{\mathcal{M}(\Omega_c)} \quad \textit{for all } \ u \in \mathcal{M}(\Omega_c). \tag{4.16}$$

As in Remark 4.1, the last condition can be equivalently rewritten as

$$(y_h(u) - \bar{y}_h, \chi_{\Omega_o}(\bar{y}_h - y_d)) + \alpha \|u\|_{\mathcal{M}(\Omega_c)} - \alpha \|\bar{u}_h\|_{\mathcal{M}(\Omega_c)} \geq 0 \quad \textit{for all } \ u \in \mathcal{M}(\Omega_c). \tag{4.17}$$

Note that in the previous result the optimal solution is found in the discrete space $\mathcal{M}_h \cap \mathcal{M}(\Omega_c)$, whereas the admissible test functions for the variational inequalities (4.16) and (4.17) can be chosen from the continuous space $\mathcal{M}(\Omega_c)$. This construction facilitates the derivation of error estimates.

4.3. General error estimates

In this section we derive error estimates which are valid in the generic case. Under additional assumptions on the location of the Ω_c and Ω_o with respect to each other better estimates are possible, which we will consider in the following section.

4.3.1. Estimates for the state solution

In order to prove our main result, we first provide some estimates of the state error for a fixed control $u \in \mathcal{M}(\Omega)$.

Lemma 4.10. *Let $u \in \mathcal{M}(\Omega)$ with associated continuous and discrete states $y = S(u)$ and $y_h = S_h(u)$ be given. Then there holds:*

$$(i) \quad \|y - y_h\|_{L^q(\Omega)} \le C_q h^{2-d/q'} \|u\|_{\mathcal{M}(\Omega)}, \quad q \in \left(1, \frac{d}{d-2}\right), \quad \frac{1}{q} + \frac{1}{q'} = 1$$

$$(ii) \quad \|y - y_h\|_{L^1(\Omega)} \le C h^2 |\ln h|^r \|u\|_{\mathcal{M}(\Omega)}$$

with $r = 2$ for $d = 2$ and $r = 11/4$ for $d = 3$.

Proof. For the first estimate in case $q = 2$ we refer to, e.g., [Cas85]. For the general case, $q \in (1, d/(d-2))$, we set $e = y - y_h$ and

$$g_q(x) = |e(x)|^{q-1} \operatorname{sgn}(e(x)).$$

By a direct calculation it follows $g_q \in L^{q'}(\Omega)$ and

$$\|g_q\|_{L^{q'}(\Omega)} = \|e\|^{q-1}_{L^q(\Omega)}.$$

We consider a dual problem for the auxiliary variable $w \in H_0^1(\Omega)$, which is given by

$$(\nabla w, \nabla \varphi) = (g_q, \varphi) \quad \text{for all } \varphi \in H_0^1(\Omega).$$

We define the corresponding Ritz projection $w_h \in V_h$ as

$$(\nabla w_h, \nabla \varphi_h) = (g_q, \varphi_h) \quad \text{for all } \varphi_h \in V_h.$$

With the help of this we can write

$$\begin{aligned} \|e\|^q_{L^q(\Omega)} &= (e, g_q) = (\nabla e, \nabla w) \\ &= (\nabla e, \nabla(w - w_h)) = (\nabla y, \nabla(w - w_h)) \\ &= \langle u, w - w_h \rangle \le \|u\|_{\mathcal{M}(\Omega)} \|w - w_h\|_{\mathcal{C}_0(\Omega)}, \end{aligned}$$

using Galerkin orthogonality for both errors $y - y_h$ and $w - w_h$. By elliptic regularity theory we obtain $w \in W^{2,q'}(\Omega)$ with

$$\|\nabla^2 w\|_{L^{q'}(\Omega)} \le C \|g_q\|_{L^{q'}(\Omega)}$$

and since $q' > 2/d$, a corresponding L^∞-estimate can be obtained. With an inverse estimate we get

$$\|w - w_h\|_{\mathcal{C}_0(\Omega)} \le \|w - i_h w\|_{\mathcal{C}_0(\Omega)} + C h^{-d/q'} \|i_h w - w_h\|_{L^{q'}(\Omega)},$$

where i_h is the nodal interpolation. With well-known interpolation estimates for the nodal interpolant in L^∞ and $L^{q'}$ and a further application of the triangle inequality, we arrive at

$$\|w - w_h\|_{\mathcal{C}_0(\Omega)} \le C h^{2-d/q'} \|\nabla^2 w\|_{L^{q'}(\Omega)} + C h^{-d/q'} \|w - w_h\|_{L^{q'}(\Omega)}.$$

The optimal estimate

$$\|w - w_h\|_{L^{q'}(\Omega)} \le C_q h^2 \|\nabla^2 w\|_{L^{q'}(\Omega)}$$

was first given in [RS82], albeit only for $d = 2$. However, the stability of the Ritz projection in $W^{1,q'}$, which is the central ingredient of the proof, is also known to hold for $d = 3$ (see [BS08, Theorem 8.5.3]), so the proof can be repeated one for one. Combining all the estimates, we obtain for the error $\|e\|_{L^q(\Omega)}$ the estimate

$$\begin{aligned}\|e\|_{L^q(\Omega)}^q &\le \|u\|_{\mathcal{M}(\Omega)} \|w - w_h\|_{\mathcal{C}_0(\Omega)} \\ &\le C_q h^{2-d/q'} \|u\|_{\mathcal{M}(\Omega)} \|e\|_{L^q(\Omega)}^{q-1},\end{aligned}$$

which gives the desired result.

To obtain the second estimate, we set $g_1 = \operatorname{sgn}(e) \in L^\infty(\Omega)$. There holds

$$\|e\|_{L^1(\Omega)} = (e, g_1).$$

As before, we consider the dual variable $w \in H_0^1(\Omega)$ and its Ritz projection $w_h \in V_h$ as the solutions of

$$\begin{aligned}(\nabla w, \nabla \varphi) &= (g_1, \varphi) \quad \text{for all } \varphi \in H_0^1(\Omega), \\ (\nabla w_h, \nabla \varphi_h) &= (g_1, \varphi_h) \quad \text{for all } \varphi_h \in V_h.\end{aligned}$$

Then we obtain using the Galerkin orthogonality for both errors $y - y_h$ and $w - w_h$, that

$$\begin{aligned}\|e\|_{L^1(\Omega)} &= (e, g_1) = (\nabla e, \nabla w) \\ &= (\nabla e, \nabla(w - w_h)) = (\nabla y, \nabla(w - w_h)) \\ &= \langle u, w - w_h \rangle \le \|u\|_{\mathcal{M}(\Omega)} \|w - w_h\|_{\mathcal{C}_0(\Omega)}.\end{aligned}$$

For the pointwise error in w we use the result from Frehse and Rannacher [FR76] for $d = 2$ and Rannacher [Ran76] for $d = 3$ and obtain

$$\|w - w_h\|_{\mathcal{C}_0(\Omega)} \le Ch^2 |\ln h|^r \|g_1\|_{L^\infty(\Omega)}.$$

This completes the proof. □

Via the Sobolev embedding theorem we can easily derive an estimate of the following form

$$\|y\|_{L^t(\Omega)} \le C_t \|u\|_{\mathcal{M}(\Omega)} \quad \text{for all } t < \frac{d}{d-2}.$$

for the continuous solutions. For the discrete solutions we can also give a result in the limiting case for t.

Lemma 4.11. *Let $u \in \mathcal{M}(\Omega)$ with the discrete solution $y_h = y_h(u)$ as above. Then we have*

$$\begin{aligned}\|y_h\|_{L^\infty(\Omega)} &\le C|\ln h|^{3/2} \|u\|_{\mathcal{M}(\Omega)} \quad \textit{for } d = 2, \\ \|y_h\|_{L^3(\Omega)} &\le C|\ln h| \|u\|_{\mathcal{M}(\Omega)} \qquad \textit{for } d = 3.\end{aligned}$$

Proof. In the first step we estimate

$$\|y_h\|_{L^\infty(\Omega)} \le C|\ln h|^{1/2} \|\nabla u_h\|_{L^2(\Omega)} \quad \text{for } d = 2$$

by the discrete Sobolev inequality (see [BS08, Lemma 4.9.2]) and

$$\|y_h\|_{L^3(\Omega)} \le C\|\nabla y_h\|_{L^{3/2}(\Omega)} \quad \text{for } d = 3,$$

by the Sobolev embedding. Defining $\sigma = d/(d-1)$ (i.e., $\sigma = 2$ and $\sigma = 3/2$ for $d = 2$ and $d = 3$ respectively), we proceed in a common way with an inverse estimate and the stability of the Ritz projection with respect to the $W^{1,s}$ seminorm (see [BS08, Theorem 8.5.3])

$$\begin{aligned}\|\nabla y_h\|_{L^\sigma(\Omega)} &\leq C\, h^{d/\sigma - d/s}\|\nabla y_h\|_{L^s(\Omega)} \\ &\leq Ch^{d/\sigma - d/s}\|\nabla y\|_{L^s(\Omega)},\end{aligned}$$

for any $1 < s < \sigma$, where the constant C is independent of s. Then we choose $s = s_\varepsilon = \sigma - \varepsilon$ for $0 < \varepsilon < \sigma - 1$, which implies that

$$\frac{d}{\sigma} - \frac{d}{s_\varepsilon} = -\frac{\varepsilon\, d}{\sigma(\sigma - \varepsilon)} > -\varepsilon\, d\sigma^{-1} = -\varepsilon(d-1).$$

We obtain by Lemma 4.1 that

$$\|\nabla y_h\|_{L^\sigma(\Omega)} \leq \frac{C}{\varepsilon}\, h^{-\varepsilon(d-1)}\|u\|_{\mathcal{M}(\Omega)}.$$

Choosing now $\varepsilon = 1/|\ln h|$ we obtain

$$\|\nabla y_h\|_{L^\sigma(\Omega)} \leq C|\ln h|\|u\|_{\mathcal{M}(\Omega)},$$

which, together with the first estimate, completes the proof. □

4.3.2. Estimates for the optimal solutions

In the next theorem we provide an error estimate for the error with respect to the cost functional. To state this theorem we need an assumption on the desired state y_d.

Assumption 4.1. We assume

$$y_d \in \begin{cases} L^\infty(\Omega), & \text{for } d = 2 \\ L^3(\Omega), & \text{for } d = 3. \end{cases}$$

First, we derive an error estimate for the optimal value of the objective functional.

Theorem 4.12. *Let Assumption 4.1 be fulfilled. Let moreover $\bar{u} \in \mathcal{M}(\Omega_c)$ be a solution to* (4.1) *and $\bar{u}_h \in \mathcal{M}_h$ be a solution to the discrete problem* (4.14)*. Then there holds*

$$|j(\bar{u}) - j_h(\bar{u}_h)| \leq C\, h^{4-d}|\ln h|^\kappa$$

with $\kappa = 7/2$ for $d = 2$ and $\kappa = 1$ for $d = 3$.

Proof. By optimality we obtain

$$j(\bar{u}) \leq j(\bar{u}_h) \quad \text{and} \quad j_h(\bar{u}_h) \leq j_h(\bar{u}).$$

Consequently we have

$$j(\bar{u}) - j_h(\bar{u}) \leq j(\bar{u}) - j_h(\bar{u}_h) \leq j(\bar{u}_h) - j_h(\bar{u}_h)$$

Therefore, it remains to estimate the error with respect to the cost functional for a fixed $u \in \mathcal{M}(\Omega_c)$, i.e., to estimate the term

$$|j(u) - j_h(u)| = \left| \frac{1}{2}\|S(u) - y_d\|^2_{L^2(\Omega_o)} - \frac{1}{2}\|S_h(u) - y_d\|^2_{L^2(\Omega_o)} \right|$$

and then to apply this estimate for both $u = \bar{u}$ and $u = \bar{u}_h$. For fixed $u \in \mathcal{M}(\Omega_c)$ we now use the notation $y = S(u)$ and $y_h = S_h(u)$. There holds:

$$\begin{aligned} j(u) - j_h(u) &= \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_o)} - \frac{1}{2}\|y_h - y_d\|^2_{L^2(\Omega_o)} \\ &= \frac{1}{2}(y - y_h, \chi_{\Omega_o}(y + y_h - 2y_d)) \\ &= -(y - y_h, \chi_{\Omega_o} y_d) + \frac{1}{2}\|y - y_h\|^2_{L^2(\Omega_o)} + (y - y_h, \chi_{\Omega_o} y_h). \end{aligned} \tag{4.18}$$

For the second term in (4.18) we obtain with Lemma 4.10.(i) for $q = 2$ that

$$\|y - y_h\|^2_{L^2(\Omega_o)} \leq Ch^{4-d}\|u\|^2_{\mathcal{M}(\Omega_c)}.$$

The other terms are estimated separately in two and three spatial dimensions. For $d = 2$ the first and last terms in (4.18) are estimated using Lemma 4.10.(ii):

$$\begin{aligned} (y - y_h, \chi_{\Omega_o} y_d) &\leq \|y - y_h\|_{L^1(\Omega_o)} \|y_d\|_{L^\infty(\Omega_o)} \leq Ch^2 |\ln h|^2 \|u\|_{\mathcal{M}(\Omega_c)}, \\ (y - y_h, \chi_{\Omega_o} y_h) &\leq \|y - y_h\|_{L^1(\Omega_o)} \|y_h\|_{L^\infty(\Omega_o)} \leq Ch^2 |\ln h|^2 \|u\|_{\mathcal{M}(\Omega_c)} \|y_h\|_{L^\infty(\Omega_o)}. \end{aligned}$$

Additionally, by Lemma 4.11 we have $\|y_h\|_{L^\infty(\Omega)} \leq |\ln h|^{3/2}\|u\|_{\mathcal{M}(\Omega_c)}$. For $d = 3$ we use Lemma 4.10.(i) with $q = 3/2$ for the remaining terms in (4.18) to obtain

$$\begin{aligned} (y - y_h, \chi_{\Omega_o} y_d) &\leq \|y - y_h\|_{L^{3/2}(\Omega_o)} \|y_d\|_{L^3(\Omega_o)} \leq Ch \|u\|_{\mathcal{M}(\Omega_c)}, \\ (y - y_h, \chi_{\Omega_o} y_h) &\leq \|y - y_h\|_{L^{3/2}(\Omega_o)} \|y_h\|_{L^3(\Omega_o)} \leq Ch \|u\|_{\mathcal{M}(\Omega_c)} \|y_h\|_{L^3(\Omega_o)}. \end{aligned}$$

We apply again Lemma 4.11 and complete the proof. □

Remark 4.2. i) Assumption 4.1 is only slightly stronger than the corresponding assumption in [CCK12], where $y_d \in L^4(\Omega)$ in two dimensions and $y_d \in L^{8/3}(\Omega)$ in three dimensions is assumed. By using a different exponent in the Hölder inequality at the end of the proof of Theorem 4.12, we can also derive a weaker estimate under such assumptions on y_d.

ii) Assumption 4.1 excludes the case where the desired state y_d is given as a Green's function. However, for construction of irregular examples with known exact solutions (see section 4.6), it is desirable to choose y_d to be the solution of

$$\begin{aligned} -\Delta y_d &= \delta_{x_0} \quad && \text{in } \Omega, \\ y_d &= 0 && \text{on } \partial\Omega \end{aligned}$$

for some $x_0 \in \Omega$. For this choice of y_d there holds:

$$\begin{aligned} y_d \in L^q(\Omega) \quad &\text{for all } q \in (1, \infty) \quad \text{for } d = 2, \\ \text{and} \quad y_d \in L^{3-\varepsilon}(\Omega) \quad &\text{for all } \varepsilon \in (0, 1) \quad \text{for } d = 3. \end{aligned}$$

The result of Theorem 4.12 can be directly extended to to this situation. In this case an additional logarithmic term $|\ln h|$ will appear.

In the next theorem we prove the main estimate for the error in the state variable.

Theorem 4.13. *Let the conditions of Theorem 4.12 be fulfilled. Then there holds*

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega_o)} \leq Ch^{2-d/2}|\ln h|^{\kappa/2}.$$

Proof. We use the optimality condition (4.6), choose $u = \bar{u}_h$ and obtain

$$(S(\bar{u}_h) - \bar{y}, \chi_{\Omega_o}(\bar{y} - y_d)) + \alpha\|\bar{u}_h\|_{\mathcal{M}(\Omega_c)} - \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c)} \geq 0.$$

For the corresponding discrete optimality condition (4.17) we choose $u = \bar{u}$ resulting in

$$(S_h(\bar{u}) - \bar{y}_h, \chi_{\Omega_o}(\bar{y}_h - y_d)) + \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c)} - \alpha\|\bar{u}_h\|_{\mathcal{M}(\Omega_c)} \geq 0.$$

Adding these two inequalities we arrive at

$$(S(\bar{u}_h) - \bar{y}, \chi_{\Omega_o}(\bar{y} - y_d)) + (S_h(\bar{u}) - \bar{y}_h, \chi_{\Omega_o}(\bar{y}_h - y_d)) \geq 0.$$

Rearranging the terms we obtain

$$\begin{aligned}(\bar{y}_h - \bar{y}, \chi_{\Omega_o}(\bar{y} - y_d)) + (S(\bar{u}_h) - \bar{y}_h, \chi_{\Omega_o}(\bar{y} - y_d)) \\ + (\bar{y} - \bar{y}_h, \chi_{\Omega_o}(\bar{y}_h - y_d)) + (S_h(\bar{u}) - \bar{y}, \chi_{\Omega_o}(\bar{y}_h - y_d)) \geq 0,\end{aligned}$$

resulting in

$$\begin{aligned}&\|\bar{y} - \bar{y}_h\|^2_{L^2(\Omega_o)} \leq (S(\bar{u}_h) - \bar{y}_h, \chi_{\Omega_o}(\bar{y} - y_d)) + (S_h(\bar{u}) - \bar{y}, \chi_{\Omega_o}(\bar{y}_h - y_d)) \\ &= (S(\bar{u}_h) - \bar{y}_h, \chi_{\Omega_o}(\bar{y} - S_h(\bar{u}))) + (S(\bar{u}_h) - \bar{y}_h, \chi_{\Omega_o}(S_h(\bar{u}) - y_d)) + (S_h(\bar{u}) - \bar{y}, \chi_{\Omega_o}(\bar{y}_h - y_d)).\end{aligned} \tag{4.19}$$

For the first term in (4.19) we obtain with Lemma 4.10.(i) for $p = 2$ that

$$\begin{aligned}(S(\bar{u}_h) - \bar{y}_h, \chi_{\Omega_o}(\bar{y} - S_h(\bar{u}))) &\leq \|S(\bar{u}_h) - \bar{y}_h\|_{L^2(\Omega_o)} \|\bar{y} - S_h(\bar{u})\|_{L^2(\Omega_o)} \\ &\leq Ch^{4-d}\|\bar{u}\|_{\mathcal{M}(\Omega_c)} \|\bar{u}_h\|_{\mathcal{M}(\Omega_c)}.\end{aligned}$$

The second and the third term in (4.19) are estimated by the same procedure as in the proof of Theorem 4.12 resulting in

$$\|\bar{y} - \bar{y}_h\|^2_{L^2(\Omega_o)} \leq Ch^{4-d}|\ln h|^{\kappa}.$$

This completes the proof. □

4.4. Improved estimates

In the following we derive improved error estimates, which are only valid under additional assumptions on the positions of Ω_c and Ω_c with respect to each other.

4.4.1. Global observation

At first, let us consider the case where the control set is contained in the observation domain, i.e., where

$$\Omega_c \subset \Omega_o.$$

In this situation, we can provide an error estimate as in Theorem 4.13 for the state and the control on the *whole* domain. We start with the observation that the optimal solutions of the continuous problem are unique.

Proposition 4.14. *Assume $\Omega_c \subset \Omega_o$. Then, problem* (4.1) *possesses a unique optimal solution.*

Proof. The observation $\chi_{\Omega_o}\bar{y}$ of the optimal state is unique; see Proposition 4.3. Additionally, the optimal control $\bar{u}$ is uniquely determined by $\chi_{\Omega_o}\bar{y}$ since the control to observation operator $\chi_{\Omega_o} \circ S$ is injective: Take $u \in \mathcal{M}(\Omega_c)$ and suppose that $\chi_{\Omega_o} y = 0$ for $y = S(u)$. This implies $y = 0$ with the maximum principle (cf., e.g., Lemma 4.20 below) and thus $u = 0$. Therefore, the optimal solution $(\bar{u}, \bar{y})$ is unique. □

As a preparatory result for the error estimate we need the following lemma.

Lemma 4.15. *Assume $\Omega_c \subset \Omega_o$. Then, it holds*

$$\operatorname{supp} \bar{u}_h \,,\, \operatorname{supp} \bar{u} \subset \{\, x \in \Omega_o \mid \operatorname{dist}(x, \partial\Omega_o) > \eta \,\}$$

for some $\eta > 0$ depending only on the problem data.

Proof. We use that the adjoint state $\bar{p}$ is Hölder continuous as in Corollary 4.5. For the discrete adjoint states we can obtain the uniform bound

$$\|\bar{p}_h\|_{C^{0,\beta}(\Omega)} \le C\|\bar{p}_h\|_{W^{1,q}(\Omega)} \le C\|p(\bar{y}_h)\|_{W^{1,q}(\Omega)} \le C\|\bar{y}_h - y_d\|_{L^2(\Omega_o)} \le C\|y_d\|_{L^2(\Omega_o)}$$

using the stability of the Ritz projection in $W^{1,q}$ for $q > d$, where $p(\bar{y}_h)$ solves the continuous adjoint equation (4.15) with the discrete $\bar{y}_h$ instead of $\bar{y}$ on the right hand side. Together with the Dirichlet boundary conditions and the conditions on the support of the optimal controls we therefore get

$$\operatorname{supp} \bar{u}_h \,,\, \operatorname{supp} \bar{u} \subset \{\, x \in \Omega_c \mid \operatorname{dist}(x, \partial\Omega) \ge \eta_1 \,\} = A_{\eta_1}$$

for some $\eta_1 > 0$ depending on the constant in the estimate before. The set A_{η_1} is compact, since Ω_c is relatively closed. With $A_{\eta_1} \subset \Omega_o$ and Ω_o open, we find a suitable $\eta \le \eta_1$ by considering that $\operatorname{dist}(\cdot, \partial\Omega_o) > 0$ must assume a minimum on A_{η_1}. □

Now, we will extend the error estimate for the state from Theorem 4.13 to an estimate on the whole domain and provide a complementary estimate for the optimal controls.

Theorem 4.16. *Assume $\Omega_c \subset \Omega_o$ and that the conditions of Theorem 4.12 are fulfilled. Then we have the estimate*

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} + \|\bar{u} - \bar{u}_h\|_{H^{-2}(\Omega)} \le C_\eta\, h^{2-d/2} |\ln h|^{\kappa/2}$$

with κ as in Theorem 4.12 and η from Lemma 4.15.

Proof. With the elliptic regularity and Lemma 4.10.(i) we obtain

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \leq \|S(\bar{u} - \bar{u}_h)\|_{L^2(\Omega)} + \|S(\bar{u}_h) - \bar{y}_h\|_{L^2(\Omega)} \leq \|\bar{u} - \bar{u}_h\|_{H^{-2}(\Omega)} + C\, h^{2-d/2}.$$

For the estimate of the control we choose a smooth function $\omega_\eta \in \mathcal{C}_0^\infty(\Omega)$ which is zero on $\Omega \setminus \Omega_o$ and equal to one on $\{\, x \in \Omega_o \mid \operatorname{dist}(x, \partial\Omega_o) > \eta \,\} \subseteq \Omega_c$. This is possible due to Lemma 4.15. Then we have for any $\psi \in H^2(\Omega)$ that

$$\langle \bar{u} - \bar{u}_h, \psi \rangle = \langle \bar{u} - \bar{u}_h, \omega_\eta \psi \rangle = (\nabla S(\bar{u} - \bar{u}_h), \nabla(\omega_\eta \psi)) = -(\bar{y} - S(\bar{u}_h), \Delta(\omega_\eta \psi)).$$

For the expression in the last term we obtain

$$\Delta(\omega_\eta \psi) = \Delta\omega_\eta \psi + 2\nabla\omega_\eta \nabla\psi + \omega_\eta \Delta\psi,$$

and since the derivatives of ω_η are bounded and depend only on η, we can estimate

$$\|\Delta(\omega_\eta \psi)\| \leq C_\eta \|\psi\|_{H^2(\Omega)}.$$

Moreover $\Delta(\omega_\eta \psi) = 0$ on $\Omega \setminus \Omega_o$ and thus we have

$$\langle \bar{u} - \bar{u}_h, \psi \rangle \leq C_\eta \|\psi\|_{H^2(\Omega)} \|\bar{y} - S(\bar{u}_h)\|_{L^2(\Omega_o)}.$$

Dividing by $\|\psi\|_{H^2(\Omega)}$ and taking the supremum, we obtain

$$\begin{aligned} \|\bar{u} - \bar{u}_h\|_{H^{-2}(\Omega)} &\leq C_\eta \|\bar{y} - S(\bar{u}_h)\|_{L^2(\Omega_o)} \\ &\leq C_\eta \left(\|\bar{y} - \bar{y}_h\|_{L^2(\Omega_o)} + \|\bar{y}_h - S(\bar{u}_h)\|_{L^2(\Omega_o)} \right) \leq C_\eta\, h^{2-d/2} |\ln h|^{\kappa/2}, \end{aligned}$$

where we applied Theorem 4.12 and Lemma 4.10.(i) for $q = 2$. This concludes the proof. □

4.4.2. Global control and observation

Now, we consider the case of control and observation on the whole domain, i.e., where

$$\Omega_c = \Omega_o = \Omega.$$

As we have seen in section 4.4.1, the optimal solution of the problem is unique. Furthermore, we can infer higher regularity of the optimal solution from a regularity assumption on the desired state. Specifically, we show that if the desired state y_d is bounded, the same holds for the optimal state $\bar{y}$. For instance, this immediately rules out Dirac measures for the optimal controls $\bar{u}$. The main result is given next.

Theorem 4.17. *Assume $\Omega_o = \Omega_c = \Omega$ and that the desired state y_d is in $L^\infty(\Omega)$. Then the optimal state $\bar{y}$ is also in $L^\infty(\Omega)$ and there holds*

$$\|\bar{y}\|_{L^\infty(\Omega)} \leq \|y_d\|_{L^\infty(\Omega)}.$$

For the proof of the result, we need some additional preparation. Let us give first a direct consequence of this theorem, which is an additional regularity for the optimal control $\bar{u}$ and for the optimal state $\bar{y}$.

Corollary 4.18. *Assume $\Omega_o = \Omega_c = \Omega$ and that the desired state y_d is in $L^\infty(\Omega)$. Then the optimal state $\bar{y}$ lies in $H_0^1(\Omega) \cap L^\infty(\Omega)$ and the optimal control $\bar{u}$ lies in $H^{-1}(\Omega)$. There holds*

$$\|\nabla \bar{y}\|_{L^2(\Omega)}^2 \le \|\bar{u}\|_{\mathcal{M}(\Omega)} \|y_d\|_{L^\infty(\Omega)} \quad \text{and} \quad \|\bar{u}\|_{H^{-1}(\Omega)} = \|\nabla \bar{y}\|_{L^2(\Omega)}.$$

In order to prove Theorem 4.17 and Corollary 4.18 we use some results from potential theory: First, introduce the Green's function $G_\Omega \colon \Omega \times \Omega \to \mathbb{R}^+ \cup \{+\infty\}$ as in, e.g., [AG01; Lan72]. Then, for a positive measure $\mu \in \mathcal{M}(\Omega)$, $\mu \ge 0$ we define the numeric function $v^* \colon \Omega \to \mathbb{R}^+ \cup \{+\infty\}$ by

$$v^* = S_G(\mu) := \int_\Omega G_\Omega(\cdot, x) \, \mathrm{d}\mu(x), \tag{4.20}$$

which is subharmonic and thus lower semicontinuous (see again [AG01]). If we normalize the Green's function G_Ω by the right constant, we obtain the following simple result.

Proposition 4.19. *For a compactly supported $\mu \in \mathcal{M}(\Omega)$, $\mu \ge 0$ the weak solution $v \in W_0^{1,s}(\Omega)$ with $1 \le s < d/(d-1)$ to the problem*

$$\begin{aligned} -\Delta v &= \mu \quad \text{in } \Omega, \\ v &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.21}$$

is equal to $v^* = S_G(\mu)$ (Lebesgue-)almost everywhere.

Proof. With [AG01, Theorem 4.3.8] the function v^* is a distributional solution of (4.21), and by a density argument, it is also a weak solution. □

With the help of the above representation, we obtain a pointwise representative of the optimal solution $y^* \colon \Omega \to \mathbb{R} \cup \{-\infty, +\infty\}$, defined as

$$y^* := S_G(\bar{u}^+) - S_G(\bar{u}^-) = S_G(\bar{u}).$$

Due to (4.7) the measures $\bar{u}^+$ and $\bar{u}^-$ are compactly supported, and with (4.8) y^* is well defined with values in $\mathbb{R} \cup \{-\infty, \infty\}$. With Proposition 4.19 we easily derive that $y^* = \bar{y}$ almost everywhere.

The next lemma states (roughly speaking), that if the optimal state is bounded on $\operatorname{supp} \bar{u}$, then it is bounded everywhere on Ω by the same constant. For positive measures of bounded variation this statement can be directly obtained from [Lan72, Theorem 1.6′] in the two-dimensional case. For $d = 3$, the analogous result [Lan72, Theorem 1.10] is stated only for $\Omega = \mathbb{R}^d$. Therefore, we provide a direct proof.

Lemma 4.20. *Let $\bar{u} \in \mathcal{M}(\Omega)$ be the optimal solution of (4.1). If $y^* = S_G(\bar{u})$ is bounded from above by some constant $C^+ \ge 0$ on $\operatorname{supp} \bar{u}^+$, then it is bounded everywhere by C^+. Analogously, if y^* is bounded from below by some $C^- \le 0$ on $\operatorname{supp} \bar{u}^-$, then y^* is bounded from below everywhere by C^-.*

Proof. Without restriction it suffices to show the first part of the result. Suppose that $y^* \le C^+$ on $\operatorname{supp} \bar{u}^+$. With (4.8) we estimate

$$S_G(\bar{u}^+) = y^* + S_G(\bar{u}^-) \le C^+ + C_\eta \|\bar{u}^-\|_{\mathcal{M}(\Omega)} \quad \text{on} \quad \operatorname{supp} \bar{u}^+ ,$$

where $C_\eta = C\log(\operatorname{diam}\Omega/\eta)$ for $d=2$ and $C_\eta = C/\eta$ for $d=3$ due to the growth properties of the Green's function. Thus, $S_G(\bar{u}^+)$ is bounded on $\operatorname{supp}\bar{u}^+$ as well. With [AG01, Corollary 4.5.2] we can now construct a sequence of compact sets K_i for $i \in \mathbb{N}$ with

$$\int_{\operatorname{supp}\bar{u}^+\setminus K_i} \mathrm{d}\bar{u}^+ = \bar{u}^+(\operatorname{supp}\bar{u}^+ \setminus K_i) \to 0 \quad \text{for } i \to \infty, \tag{4.22}$$

such that the functions $S_G(\bar{u}^+|_{K_i})$ are continuous. Now, we consider the solutions

$$y_i = S_G(\bar{u}^+|_{K_i}) - S_G(\bar{u}^-) \le y^*.$$

Recalling that $-S_G(\bar{u}^-)$ is upper semicontinuous, we obtain that each y_i is upper semicontinuous as well. For each x_0 on the boundary of $\Omega \setminus \operatorname{supp}\bar{u}^+$, which is a subset of $\operatorname{supp}\bar{u}^+ \cup \partial\Omega$, we have $y_i(x_0) \le y^*(x_0) \le C^+$ and with upper semicontinuity

$$\limsup_{x\to x_0} y_i(x) \le C^+. \tag{4.23}$$

Using the fact that y_i is subharmonic on $\Omega \setminus \operatorname{supp} q^+$ and the condition (4.23) we apply the maximum principle for subharmonic functions [AG01, Theorem 3.1.5], and obtain that y_i is bounded by C^+ everywhere on Ω for all $i \in \mathbb{N}$.

To complete the proof, it remains to show the convergence $y_i(x) \to y^*(x)$ for all $x \in \Omega \setminus \operatorname{supp}\bar{u}^+$. Let $x \in \Omega \setminus \operatorname{supp}\bar{u}^+$ be fixed. We denote by $\delta = \operatorname{dist}(x, \operatorname{supp}\bar{u}^+) > 0$. There holds

$$|y_i(x) - y^*(x)| = |S_G(\bar{u}^+|_{K_i})(x) - S_G(\bar{u}^+)(x)| \le C_\delta \int_{\operatorname{supp}\bar{u}^+\setminus K_i} \mathrm{d}\bar{u}^+ \to 0$$

for $i \to \infty$, where we have again used growth properties of the Green's function and (4.22). □

With these preparations we can give proofs of the claimed results.

Proof of Theorem 4.17. Assume the contrary, i.e., that we have constants $M, \varepsilon > 0$, such that $|y_d| \le M$ almost everywhere in Ω, but $|\bar{y}| > M + \varepsilon$ on some set of positive Lebesgue measure, i.e.,

$$|\{\, x \in \Omega \mid \bar{y}(x) > M + \varepsilon \,\}| > 0.$$

Due to Lemma 4.20 we can find a point $\hat{x} \in \operatorname{supp} u^+$ where $y^* = S_G(\bar{u})$ is larger than $M + \varepsilon$. Considering a ball $B_\eta(x)$ of radius η around this point, we have with Corollary 4.5 that $\bar{u}^-|_{B_\eta(\hat{x})} = 0$ and therefore that $S_G(\bar{u}|_{B_\eta(\hat{x})})$ is lower semicontinuous. We decompose

$$y^* = S_G(\bar{u}|_{B_\eta(\hat{x})}) + S_G(\bar{u}|_{\Omega\setminus B_\eta(\hat{x})})$$

and obtain that $S_G(\bar{u}|_{\Omega\setminus B_\eta(\hat{x})})$ is harmonic and consequently continuous on $B_\eta(\hat{x})$. This implies the lower semicontinuity of y^* on $B_\eta(\hat{x})$. Thus, the set

$$\{\, x \in B_\eta(x) \mid y^*(x) > M + \varepsilon \,\}$$

is open, and we can find a radius $r > 0$ such that $\bar{y} \ge M + \varepsilon$ almost everywhere in the ball $B_r(\hat{x})$.

Note that $\hat{x} \in \operatorname{supp}\bar{u}^+$ implies $\bar{p}(\hat{x}) = -\alpha$ with Theorem 4.4. We define w to be the solution to

$$\begin{aligned} -\Delta w &= \varepsilon \quad \text{in } B_r(\hat{x}), \\ w &= 0 \quad \text{on } \partial B_r(\hat{x}), \end{aligned}$$

which is clearly strictly positive at $\hat{x}$. Considering the minimum principle for $\tilde{p} = \bar{p} - w$ which fulfills

$$\begin{aligned} -\Delta\tilde{p} = \bar{y} - y_d - \varepsilon \geq 0 \quad & \text{in } B_r(\hat{x}), \\ \tilde{p} = \bar{p} \quad & \text{on } \partial B_r(\hat{x}), \end{aligned}$$

we see that the minimum value $p_{\min} = \inf_{x \in B_r(\hat{x})} \tilde{p}(\hat{x})$ must be attained for some $x' \in \partial B_r(\hat{x})$. Comparing with the center $\hat{x}$ we find

$$\bar{p}(x') = \tilde{p}(x') = (\bar{p} - w)(x') \leq (\bar{p} - w)(\hat{x}) < \bar{p}(\hat{x}) = -\alpha,$$

which is a violation of the bounds on the adjoint state (4.4) and thus a contradiction. □

Proof of Corollary 4.18. The result can be derived by considering a sequence of smooth approximations to $\bar{u}$, testing the corresponding state equation with the smooth solution and a subsequential weak limit argument.

However, the statement directly follows from a well-known classical result: Since y^* is Borel-measurable (as the difference of two lower semicontinuous functions) we can pair y^* with $\bar{u}$, and since, by the previous theorem, y^* is bounded, we obtain

$$\|\bar{u}\|_{\mathcal{M}(\Omega)} \|y^*\|_{L^\infty(\Omega)} \geq \langle \bar{u}, y^* \rangle = \int_\Omega y^*(x) \,\mathrm{d}\bar{u}(x) = \int_\Omega \int_\Omega G_\Omega(x, \tilde{x}) \,\mathrm{d}\bar{u}(x) \,\mathrm{d}\bar{u}(\tilde{x}).$$

With [Lan72, Theorem 1.20], this implies $\nabla y^* \in L^2(\Omega)$ and

$$\int_\Omega \int_\Omega G_\Omega(x, \tilde{x}) \,\mathrm{d}\bar{u}(x) \,\mathrm{d}\bar{u}(\tilde{x}) = \|\nabla y^*\|^2_{L^2(\Omega)},$$

which implies the first part of the claim. The second assertion is evident. □

In the following we use the additional regularity derived above to provide an improved estimate under the assumption that y_d is bounded.

Theorem 4.21. *Suppose $\Omega_o = \Omega_c = \Omega$. For both $d = 2$ and $d = 3$, let $(\bar{u}, \bar{y})$ be the solution to (4.1) and $(\bar{u}_h, \bar{y}_h) \in \mathcal{M}_h \times V_h$ be the discrete solution. Let moreover $y_d \in L^\infty(\Omega)$, which implies $\bar{y} \in H^1_0(\Omega) \cap L^\infty(\Omega)$ and $\bar{u} \in H^{-1}(\Omega)$ with Theorem 4.17 and Corollary 4.18. Then there holds*

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \leq C\, h |\ln h|^{r/2},$$

with the constant r as in Lemma 4.10.

Proof. First, we obtain an $L^2(\Omega)$ estimate for $\bar{y}_h - \bar{y}$ in terms of an $L^\infty(\Omega)$-error for the adjoint state. For that, we use the optimality condition (4.3), choosing $u = \bar{u}_h$

$$-\langle \bar{u}_h - \bar{u}, \bar{p} \rangle + \alpha \|\bar{u}\|_{\mathcal{M}(\Omega)} \leq \alpha \|\bar{u}_h\|_{\mathcal{M}(\Omega)},$$

and the optimality condition (4.16) choosing $u = \bar{u}$

$$-\langle \bar{u} - \bar{u}_h, \bar{p}_h \rangle + \alpha \|\bar{u}_h\|_{\mathcal{M}(\Omega)} \leq \alpha \|\bar{u}\|_{\mathcal{M}(\Omega)}.$$

Adding these two inequalities results in

$$\langle \bar{u}_h - \bar{u}, \bar{p} - \bar{p}_h \rangle \geq 0.$$

We introduce a discrete dual state $\hat{p}_h \in V_h$ for the continuous optimal solution defined by

$$(\nabla\varphi_h, \nabla\hat{p}_h) = (\bar{y} - y_d, \varphi_h) \quad \text{for all } \varphi_h \in V_h.$$

and $\hat{y}_h = y_h(\bar{u})$, the discrete solution for the continuous optimal control. There holds

$$\begin{aligned} 0 &\leq \langle \bar{u}_h - \bar{u}, \bar{p} - \bar{p}_h \rangle = \langle \bar{u}_h - \bar{u}, \bar{p} - \hat{p}_h \rangle + \langle \bar{u}_h - \bar{u}, \hat{p}_h - \bar{p}_h \rangle \\ &= \langle \bar{u}_h - \bar{u}, \bar{p} - \hat{p}_h \rangle + (\nabla(\bar{y}_h - \hat{y}_h), \nabla(\hat{p}_h - \bar{p}_h)) \\ &= \langle \bar{u}_h - \bar{u}, \bar{p} - \hat{p}_h \rangle + (\bar{y}_h - \hat{y}_h, \bar{y} - \bar{y}_h) \\ &= \langle \bar{u}_h - \bar{u}, \bar{p} - \hat{p}_h \rangle + (\bar{y} - \hat{u}_h, \bar{y} - \bar{y}_h) - \|\bar{y} - \bar{y}_h\|^2_{L^2(\Omega)}. \end{aligned}$$

Rearranging terms and using Young's inequality we obtain

$$\|\bar{y} - \bar{y}_h\|^2_{L^2(\Omega)} \leq \|\bar{u}_h - \bar{u}\|_{\mathcal{M}(\Omega)} \|\bar{p} - \hat{p}_h\|_{L^\infty(\Omega)} + \frac{1}{2}\|\bar{y} - \hat{y}_h\|^2_{L^2(\Omega)} + \frac{1}{2}\|\bar{y} - \bar{y}_h\|^2_{L^2(\Omega)},$$

which results in

$$\|\bar{y} - \bar{y}_h\|^2_{L^2(\Omega)} \leq C\|\bar{p} - \hat{p}_h\|_{L^\infty(\Omega)} + \|\bar{y} - \hat{y}_h\|^2_{L^2(\Omega)}, \tag{4.24}$$

since $\|\bar{u}\|_{\mathcal{M}(\Omega)}$ and $\|\bar{u}_h\|_{\mathcal{M}(\Omega)}$ are bounded. For the first term we obtain with an L^∞-estimate as in the proof of Lemma 4.10.(ii) that

$$\|\bar{p} - \hat{p}_h\|_{L^\infty(\Omega)} \leq Ch^2 |\ln h|^r \|\bar{y} - y_d\|_{L^\infty(\Omega)}.$$

The square root of the second term in (4.24) can be estimated by

$$\|\bar{y} - \hat{y}_h\|_{L^2(\Omega)} \leq Ch\|\bar{u}\|_{H^{-1}(\Omega)},$$

which can be obtained from standard estimates with a simple duality argument. Together with the improved regularity for $\bar{y}$ and $\bar{u}$ this completes the proof. □

4.5. Regularized problem

As discussed in chapter 2, for the numerical computation of optimal controls we are going to consider a regularized version of the optimal control problem. Here, we consider only the case where the control set can be written as the relative closure of an open set: we assume that

$$\Omega_c = \overline{\operatorname{int} \Omega_c} \cap \Omega.$$

This is needed to explain the space $L^2(\Omega_c)$. We remark that we could also discuss the case where Ω_c is a sufficiently smooth lower-dimensional manifold. However, for the regularity and error analysis we would have to use different techniques in this case (since we are not in the setting of a spatially distributed control). The regularized problem is given in the continuous setting by

$$\begin{aligned} \min_{u \in L^2(\Omega), y \in H^1_0(\Omega)} \quad & \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_o)} + \alpha\|u\|_{L^1(\Omega_c)} + \frac{\gamma}{2}\|u\|^2_{L^2(\Omega_c)} \\ \text{subject to} \quad & (\nabla y, \nabla\varphi) = (\chi_{\Omega_c} u, \varphi) \quad \text{for all } \varphi \in H^1_0(\Omega), \end{aligned} \tag{4.25}$$

where $\gamma > 0$ is the regularization parameter. For analysis of the problem (4.25) for a fixed nonzero γ we refer also to [Sta09] and [CHW12b; CHW12a] (in combination with a semilinear elliptic equation). By strong convexity of the reduced objective function corresponding to (4.25) we obtain the existence of a unique optimal solution.

Proposition 4.22. *The regularized problem* (4.25) *has a unique optimal solution* $(\bar{u}_\gamma, \bar{y}_\gamma) \in L^2(\Omega_c) \times (H_0^1(\Omega) \cap H^2(\Omega))$.

To derive optimality conditions for (4.25) we recall the proximal map of the L^1 norm from section 3.3.2, which is given for any $q \in L^2(\Omega_c)$ by the Nemyckii operator (the soft-shrinkage operator)

$$P_\gamma(q) = \operatorname{shrink}_{\alpha/\gamma}(q) = (q - \alpha/\gamma)^+ - (q + \alpha/\gamma)^-.$$

Thereby, the optimality condition for (4.25) can be obtained as follows.

Theorem 4.23. *Corresponding to the unique optimal solution* $(\bar{u}_\gamma, \bar{y}_\gamma)$ *of* (4.25) *there exists an adjoint state* $\bar{p}_\gamma$ *solving the adjoint equation* (4.2)*, which fulfills the variational inequality*

$$-(\bar{p}_\gamma + \gamma\bar{u}_\gamma, \chi_{\Omega_c}(u - \bar{u}_\gamma)) + \alpha\|\bar{u}_\gamma\|_{L^1(\Omega_c)} \leq \alpha\|u\|_{L^1(\Omega_c)} \quad \text{for all } u \in L^2(\Omega_c). \tag{4.26}$$

The variational inequality (4.26) *can be alternatively expressed by the pointwise proximal formula*

$$\bar{u}_\gamma = P_\gamma\left(-\frac{1}{\gamma}\chi_{\Omega_c}\bar{p}_\gamma\right) = \frac{1}{\gamma}\left((-\chi_{\Omega_c}\bar{p}_\gamma - \alpha)^+ - (-\chi_{\Omega_c}\bar{p}_\gamma + \alpha)^-\right). \tag{4.27}$$

Remark 4.3. The variational inequality (4.26) also has an equivalent pointwise representation. For every $x \in \Omega_c$ it holds that

$$-(\bar{p}_\gamma(x) + \gamma\bar{u}_\gamma(x))(\tilde{u} - \bar{u}_\gamma(x)) + \alpha|\bar{u}_\gamma(x)| \leq \alpha|\tilde{u}| \quad \text{for all } \tilde{u} \in \mathbb{R}.$$

Note that $\bar{u}_\gamma = 1/\gamma \operatorname{shrink}_\alpha(-\chi_{\Omega_c}\bar{p}_\gamma)$ is continuous. The pointwise variational inequality is simply the (sufficient) optimality condition for the pointwise proximal representation (4.27).

4.5.1. Regularization error analysis

In the case $\Omega = \Omega_c = \Omega_o$, convergence of $\bar{u}_\gamma \rightharpoonup^* \bar{u}$ in $\mathcal{M}(\Omega_c)$ has been shown in [CK11a]. In the general case, we apply Theorem 2.28 from section 2.5. Furthermore, we can derive an asymptotic a priori estimate for the error in the functional based on the techniques introduced by Hintermüller, Schiela, and Wollner [HSW14]. For this we need the additional assumption that Ω_c fulfills the cone condition. It is fulfilled, e.g., if Ω_c has a Lipschitz boundary (see [AF03, Chapter 4]). Following their notation, for any domain $D \subset \Omega$, we extend the definition of the space of Hölder continuous functions $\mathcal{C}^\beta(D)$ for exponents $1 < \beta \leq 2$. We define it as the space of continuously differentiable functions with $\beta - 1$ order Hölder continuous derivatives (usually denoted by $\mathcal{C}^{1,\beta-1}(D)$). Additionally we define the corresponding spaces with zero boundary conditions as $\mathcal{C}_0^\beta(D) = \{\, v \in \mathcal{C}^\beta(D) \mid v|_{\partial D} = 0 \,\}$. Now, we restate the following result from [HSW14].

Proposition 4.24. *For any* $0 < \beta \leq 1$ *and* $v \in \mathcal{C}^\beta(\Omega_c)$*, we have the interpolation estimate*

$$\|v\|_{L^\infty(\Omega_c)} \leq C\|v\|_{\mathcal{C}^\beta(\Omega_c)}^{1-\theta}\|v\|_{L^1(\Omega_c)}^{\theta} \quad \text{for } \theta = \frac{\beta}{\beta + d}.$$

Furthermore, for any domain $D \subset \Omega$ *and a positive function* $v \in \mathcal{C}_0^\beta(D)$ *with* $0 < \beta \leq 2$*, the estimate*

$$\|v\|_{L^\infty(D)} \leq C\|v\|_{\mathcal{C}_0^\beta(D)}^{1-\theta}\|v\|_{L^1(D)}^{\theta} \quad \text{for } \theta = \frac{\beta}{\beta + d}$$

is valid for all $0 < \beta \leq 2$.

Proof. The second part is proved in [HSW14, Proposition 2.4]. The zero boundary conditions are needed to ensure that the maximum of v is attained in the interior of D. As indicated in [HSW14, Remark 2.5], without zero boundary conditions the estimate for $\beta \in (0,1]$ remains valid on domains fulfilling the cone condition. □

Additionally, we need some a priori estimates for the optimal solutions which are independent of the regularization parameter.

Proposition 4.25. *For any $s < d/(d-1)$, $q < d/(d-2)$ and $\beta < 4-d$, there exists a constant $C > 0$, such that for all $\gamma > 0$ the following estimates are valid for the optimal triple $(\bar{u}_\gamma, \bar{u}_\gamma, \bar{p}_\gamma)$ of (4.25):*

$$\|\bar{u}_\gamma\|_{L^1(\Omega_c)} + \frac{\gamma}{2}\|\bar{u}_\gamma\|^2_{L^2(\Omega_c)} \leq C, \tag{4.28}$$

$$\|\bar{y}_\gamma\|_{L^q(\Omega)} + \|\bar{y}_\gamma\|_{W^{1,s}(\Omega)} \leq C, \tag{4.29}$$

$$\|\bar{p}_\gamma\|_{\mathcal{C}^\beta(\Omega)} + \|\bar{p}_\gamma\|_{W^{2,q}(\Omega)} \leq C. \tag{4.30}$$

Proof. The estimate (4.28) follows by straightforward arguments using the minimality of $\bar{u}_\gamma$; cf. Theorem 2.28. For the state, we apply now Proposition 2.8 and the Sobolev embedding with $1/q = 1/s - 1/d$. For the adjoint solution, we use the previous result to obtain an estimate for $\chi_{\Omega_o}(\bar{y}_\gamma - y_d)$ in $L^q(\Omega)$ and then apply Theorem 4.6. The estimate for $\bar{p}_\gamma$ in the Hölder norm is again a consequence of the Sobolev embedding with $\beta = 2 - d/q = 3 - d/s$. □

Thereby, we can obtain the following asymptotic estimate for the regularization error (cf. [HSW14, Corollary 2.6]).

Proposition 4.26. *The error in the objective functional due to regularization is bounded by*

$$0 \leq j_\gamma(\bar{u}_\gamma) - j(\bar{u}) \leq C\,\gamma^s,$$

where $s = 1/3$ in two dimensions and $s = 1/4 - \varepsilon$ (for $\varepsilon > 0$ arbitrary) in three dimensions. If we suppose that $\Omega_c = \Omega$, we obtain the improved rate $s = 1/2 - \varepsilon$ (for $\varepsilon > 0$ arbitrary) in two dimensions.

Proof. We define the positive and negative support sets of the control by

$$\Omega^+ = \{\, x \in \Omega_c \mid \bar{p}_\gamma(x) > \alpha \,\},$$
$$\Omega^- = \{\, x \in \Omega_c \mid \bar{p}_\gamma(x) < -\alpha \,\}.$$

With Hölder's inequality, the optimality conditions, and (4.28) we derive

$$\begin{aligned}\|\bar{u}_\gamma\|^2_{L^2(\Omega_c)} &\leq \|\bar{u}_\gamma\|_{L^1(\Omega_c)}\|\bar{u}_\gamma\|_{L^\infty(\Omega_c)} \leq C\|1/\gamma\ \mathrm{shrink}_\alpha(\bar{p}_\gamma)\|_{L^\infty(\Omega_c)} \\ &= \frac{C}{\gamma}\|\mathrm{shrink}_\alpha(\bar{p}_\gamma)\|_{L^\infty(\Omega_c)} = \frac{C}{\gamma}\max\{\,\|(\bar{p}_\gamma - \alpha)^+\|_{L^\infty(\Omega^+)}, \|(\bar{p}_\gamma + \alpha)^-\|_{L^\infty(\Omega^-)}\,\}.\end{aligned}$$

Now, we apply Proposition 4.24 separately in the case $\Omega_c = \Omega$ and the general case. In the general case, we set $\beta = \min\{\,1, 4 - d - \varepsilon\,\}$ for some $\varepsilon > 0$ according to Proposition 4.25 and note that $\mathrm{shrink}_\alpha(\bar{p}_\gamma) \in \mathcal{C}^\beta(\Omega_c)$, since the max-operator preserves Hölder continuity. We estimate for $\theta = \beta/(\beta + d)$ with the first part of Proposition 4.24 that

$$\|\mathrm{shrink}_\alpha(\bar{p}_\gamma)\|_{L^\infty(\Omega_c)} \leq \|\bar{p}_\gamma\|^{1-\theta}_{C^\beta(\Omega_c)}\|\mathrm{shrink}_\alpha(\bar{p}_\gamma)\|^\theta_{L^1(\Omega_c)}.$$

In the case $\Omega_c = \Omega$ and $d = 2$ an improvement is possible: due to the zero Dirichlet boundary conditions of $\bar{p}_\gamma$ we can guarantee that $(\bar{p}_\gamma - \alpha)^+|_{\partial\Omega^+} = 0$ and $(\bar{p}_\gamma - \alpha)^-|_{\partial\Omega^-} = 0$. Therefore, we can apply the second part of Proposition 4.24 with $\beta = 4 - d - \varepsilon$ and $\theta = \beta/(d + \beta)$ to obtain

$$\|(\bar{p}_\gamma - \alpha)^+\|_{L^\infty(\Omega^+)} \leq \|\bar{p}_\gamma\|^{1-\theta}_{C^\beta(\Omega)} \|(\bar{p}_\gamma - \alpha)^+\|^\theta_{L^1(\Omega^+)},$$

and similarly for the negative part. We proceed in a common way for the choices of β and θ as indicated above and obtain

$$\|\bar{u}_\gamma\|^2_{L^2(\Omega_c)} \leq \frac{C}{\gamma} \|\bar{p}_\gamma\|^{1-\theta}_{\mathcal{C}^\beta(\Omega)} \|\mathrm{shrink}_\alpha(\bar{p}_\gamma)\|^\theta_{L^1(\Omega_c)} = C\gamma^{\theta-1} \|\bar{p}_\gamma\|^{1-\theta}_{\mathcal{C}^\beta(\Omega)} \|\bar{u}_\gamma\|^\theta_{L^1(\Omega_c)} \leq C\gamma^{\theta-1},$$

using again the optimality conditions and the estimates (4.28) and (4.30). Finally, we apply Corollary 2.29 to compute

$$0 \leq j_\gamma(\bar{u}_\gamma) - j(\bar{u}) = \int_0^\gamma \frac{1}{2} \|\bar{u}_\sigma\|^2_{L^2(\Omega_c)} \,\mathrm{d}\sigma \leq C\gamma^\theta = C\gamma^{\beta/(\beta+d)}.$$

This results in $s = \theta = \beta/(\beta + d)$ for the choices of β indicated above. □

4.5.2. Optimization aspects

We can solve the regularized optimization problem with the methods from chapter 3. In the notation used there, we define the smooth and the convex part of the reduced cost functional as

$$f(u) = J(S(u)) \quad \text{and} \quad \psi(u) = \alpha\|u\|_{L^1(\Omega_c)} \quad \text{for } u \in L^2(\Omega_c),$$

respectively. As usual the reduced cost functional is defined as

$$j_\gamma(u) = f(u) + \psi(u) + \frac{\gamma}{2}\|u\|_{L^2(\Omega_c)} \quad \text{for } u \in L^2(\Omega_c),$$

Here, $S\colon L^2(\Omega_c) \to H^1_0(\Omega) \cap H^2(\Omega)$ is the solution operator of the state equation. The gradient and Hessian of f are easily derived as

$$\nabla f(u) = S^*(\chi_{\Omega_o}(S(u) - y_d)) \quad \text{and} \quad \nabla^2 f = S^* \chi_{\Omega_o} S,$$

where $S^* = \chi_{\Omega_c} S_{\mathrm{dual}}$ and S_{dual} is the solution operator of the dual equation $(\nabla\cdot, \nabla p) = (f, \cdot)$ which can be considered as an operator

$$S_{\mathrm{dual}}\colon L^2(\Omega) \to H^1_0(\Omega) \cap H^2(\Omega), \quad f \mapsto p.$$

The proximal map of ψ in $L^2(\Omega_c)$ evaluates to

$$P_\gamma(q) = \mathrm{shrink}_{\alpha/\gamma}(q)$$

for any given $q \in L^2(\Omega_c)$; see section 3.3.2. Since the adjoint solution operator S_{dual} maps $L^2(\Omega_o)$ continuously into $\mathcal{C}^\beta(\Omega)$ for some $\beta > 0$, we have a more than sufficient norm-gap to apply the general theory (see also [Sta09]).

Proposition 4.27. *Let $\bar{q}_\gamma = -1/\gamma\, \chi_{\Omega_c} \bar{p}_\gamma \in L^2(\Omega_c)$ be the optimal auxiliary variable with $\bar{u}_\gamma = P_\gamma(\bar{q}_\gamma)$. Suppose that for a given $q_0 \in L^r(\Omega_c)$ with $r > 2$, the distance $\|q_0 - \bar{q}_\gamma\|_{L^r(\Omega_c)}$ is sufficiently small. Then the semismooth Newton iterates, defined inductively as $q_{k+1} = q_k - \mathrm{D}G(q_k)^{-1} G(q_k)$ for $k \in \mathbb{N}$ converge superlinearly in $L^r(\Omega_c)$ towards $\bar{q}_\gamma$ (cf. section 3.2). The same holds for $u_k = P_\gamma(q_k)$ with limit $\bar{u}_\gamma$.*

Proof. Since f is convex and $\gamma > 0$, invertibility of the Newton operator is guaranteed; see Lemma 3.14. As discussed in section 3.3.2, $P_\gamma \colon L^r(\Omega_c) \to L^2(\Omega_c)$ is semismooth w.r.t. the generalized differential given there. Therefore, the iterates are well-defined. Now, we combine Theorem 3.7 and Proposition 3.11 with the choice $H = L^2(\Omega_c)$ and $H_{\text{sub}} = L^r(\Omega_c)$. □

In practice, if we solve (4.25) for a large initial parameter γ, we observe global convergence of the semismooth Newton method. In fact, global convergence for a control constrained problem was shown by Ito and Kunisch [IK04, Theorem 3.1] (the result is valid for a sufficiently large regularization parameter γ). Moreover, we observe the same behavior if we reduce the regularization parameter by a fixed constant and use the optimal solution for the previous parameter as an initial guess. With this algorithm, a globalization strategy of the semismooth Newton method is usually not needed.

4.5.3. Discretization of the regularized problem

The regularized problem (4.25) is very similar to a standard problem with control constraints. In fact, it is well known that (4.25) can be rewritten as a control constrained problem (with two controls) by splitting it into the positive and negative part $u = u^+ - u^-$. The L^1 norm then turns into a linear term in u^+ and u^-, respectively. For control constrained problems, we can find many finite element discretization concepts with a rigorous error analysis in the literature. For the sake of comparison, we briefly present the three most popular of those (in combination with linear or bilinear finite elements for the state). We will briefly explain the piece-wise constant, piece-wise linear, and variational discretization concept and summarize the known error estimates in each case; cf. also the overview given in Tröltzsch [Trö10a]. For all standard discretization concepts, the state equation is discretized in the standard conforming way as

$$(\nabla y_h, \nabla \varphi_h) = (\chi_{\Omega_c} u_h, \varphi_h) \quad \text{for all } \varphi_h \in V_h, \tag{4.31}$$

where u_h is searched for in different spaces $U_h \subset L^2(\Omega_c)$.

Both the piece-wise linear and piece-wise constant discretization do not yield the unregularized discrete problem (4.11) in the limiting case for $\gamma \to 0$. The variational concept has nice theoretical features, but is more difficult to realize in practice. Therefore, we propose a new variant based on mass lumping, which has a nice connection with the discretization of the unregularized problem (4.11) and combines an easy implementation with a rigorous error analysis. Under an established structural assumption on the optimal control of the regularized problem, we derive an estimate for the discretization error with the optimal order $\mathcal{O}(h^2)$. The result (which is analogous to known post-processing results for control constrained problems for piecewise constant control discretization due to Meyer and Rösch [MR04]) is based on a linear discretization of the control and seems to be new also in the context of control constrained optimization.

In the following, we will use the same notations for the discretization $\mathcal{T}_h$ as in section 4.2 and make the same assumptions. Specifically, we make the same compatibility assumption on the control set and the mesh as before.

Piecewise constant controls

Here, the control is discretized with piecewise constant, discontinuous finite elements. It is searched for in the space

$$U_h^0 = \left\{ u_h \in L^2(\Omega_c) \;\middle|\; u_h|_K \in \mathcal{P}_0(K) \quad \text{for all } K \in \mathcal{T}_h^c \right\}.$$

The control cost term is discretized in the standard conforming way and can be implemented in a straightforward manner. We obtain an optimality condition giving the optimal control as $\bar{u}_{h,\gamma} = 1/\gamma \operatorname{shrink}_\alpha(-\chi_{\Omega_c} P_h^0(\bar{p}_{h,\gamma}))$ (cell-wise), where $P_h^0 \colon V_h \to U_h^0$ is the L^2 projection. This control discretization is considered in [CHW12b; WW11], where an a priori estimate in L^2 for the optimal controls of the order $\mathcal{O}(h)$ is derived. This is in agreement with the approximation properties of piece-wise constants and the usual estimates for control constrained problems; see, e.g., [ACT02; CT03] and the references therein.

However, higher order reconstructions can be obtained from this discretization with the *post-processing* approach. The idea behind the reconstruction is to evaluate the continuous optimality condition with the discrete optimal adjoint state, i.e., we set

$$\bar{u}_{\sigma,\gamma} = \hat{P}_\gamma \left(-\frac{1}{\gamma} \chi_{\Omega_c} \bar{p}_{h,\gamma} \right) \quad \text{pointwise in } \Omega_c.$$

Under an additional assumption on the structure of the active sets, which is often fulfilled, one can show a convergence order of $\mathcal{O}(h^2)$ for the post-processed controls in L^2; see, e.g., [MR04; RV06] for the control constrained case. This discretization offers a good trade-off between ease of implementation and accuracy. In the context of the regularization of problem (4.14) it is inconvenient that there is no direct connection between the solutions of the discrete and the discrete regularized problem, since the controls are discretized with a nodal basis in the former and with a cell-wise basis in the latter case.

Piecewise linear controls

Here, the control is discretized with piecewise linear, continuous finite elements. It is searched for in the space

$$U_h^1 = \left\{ u_h = \chi_{\Omega_c} v_h \;\middle|\; v_h \in V_h \right\}$$

In this case, the practical implementation is not completely straightforward. For an arbitrary $u_h \in U_h^1$, the control cost term

$$\|u_h\|_{L^1(\Omega_c)} = \int_{\Omega_c} |u_h(x)| \, \mathrm{d}x$$

can not be accurately evaluated with a standard quadrature formula, or even by a matrix vector product on the level of the nodal vector. Note that the integrand $|u_h(\cdot)|$ is not twice differentiable on a cell $K \in \mathcal{T}_h$ if u_h changes sign there and that the L^1 norm is not a linear functional. However, this difficulty can be overcome by splitting the control into a positive and negative part on the level of coefficient vectors. For $u_h = \chi_{\Omega_c} \sum_n u_n e_n$ we can set

$$u_h^{\text{plus}} = \chi_{\Omega_c} \sum_{n=1}^{N_c} (u_n)^+ e_n \quad \text{and} \quad u_h^{\text{minus}} = \chi_{\Omega_c} \sum_{n=1}^{N_c} (u_n)^- e_n.$$

We have $u_h = u_h^{\text{plus}} - u_h^{\text{minus}}$ with $u_h^{\text{plus}}, u_h^{\text{minus}} \geq 0$ (pointwise in Ω_c) and the $L^1(\Omega_c)$ norm can be implemented as the sum of the two linear functionals

$$\|u_h\|_{L^1(\Omega_c),h} = \int_{\Omega_c} u_h^{\text{plus}}(x)\,\mathrm{d}x + \int_{\Omega_c} u_h^{\text{minus}}(x)\,\mathrm{d}x = \sum_{n=1}^{N_c} |u_n| \int_{\Omega_c} e_n(x)\,\mathrm{d}x.$$

Note that $\|u_h\|_{L^1(\Omega_c),h} \geq \|u_h\|_{L^1(\Omega_c)}$ for all $u_h \in U_h^1$. Consequently, we arrive at the discrete problem (cf. also [WW11, Section 4.4])

$$\min_{u_h \in U_h^1} J(S_h u_h) + \alpha \|u_h\|_{L^1(\Omega_c),h} + \frac{\gamma}{2}\|u_h\|_{L^2(\Omega_c)}^2.$$

Here, the optimality condition can not be given by a pointwise formula in terms of the adjoint state. This is due to the fact that the discrete version of the proximal map

$$P_\gamma^h(q_h) = \operatorname*{argmin}_{u_h \in V_h} \frac{\gamma}{2}\|u_h - q_h\|_{L^2(\Omega_c)}^2 + \alpha \|u_h\|_{L^1(\Omega_c),h}$$

does not possess a pointwise solution formula. However, we can obtain the following coordinate-wise optimality condition, linking the optimal adjoint state $\bar{p}_{h,\gamma} = \sum_n p_n e_n$, and the optimal control $\bar{u}_{h,\gamma} = \sum_n u_n e_n$. We obtain the variational inequality for the discrete problem as

$$(\bar{p}_{h,\gamma} + \gamma \bar{u}_{h,\gamma}, \chi_{\Omega_c}(u_h - \bar{u}_{h,\gamma})) + \alpha \|\bar{u}_{h,\gamma}\|_{L^1(\Omega_c),h} \leq \alpha \|u_h\|_{L^1(\Omega_c),h} \quad \text{for all } u_h \in U_h^1.$$

By choosing $u_h - \bar{u}_{h,\gamma} = \pm e_n$ for $n \in \{1, \dots, N_c\}$ we can obtain a corresponding discrete optimality system.

Remark 4.4. In practice, we observe that the optimal solutions are sparse; for $n \in \{1, \dots, N_c\}$ the inequality $|(\bar{p}_{h,\gamma}, \chi_{\Omega_c} e_n)| \leq \alpha \int_{\Omega_c} e_n$ implies $\bar{u}_{h,\gamma}(x_n) = 0$. Furthermore, it seems that the optimal solutions do not change sign on any given cell for h small enough, which justifies this implementation of the control cost term. A rigorous explanation of this effect is complicated by the non-local nature of $P_\gamma^h \colon U_h^1 \to U_h^1$ as defined above.

However, even though the implementation of the problem with linear controls is slightly more involved, the error analysis only yields a rate of $\mathcal{O}(h)$ for the controls in $L^2(\Omega_c)$; see, e.g., [Cas07; WW11; CHW12a]. A more refined analysis under a similar structural assumption on the optimal solutions as in the post-processing approach yields an improved rate of $\mathcal{O}(h^{3/2})$; see, e.g., [Rös06] for a control constrained problem where only the control is discretized or [BV07] for a convection-diffusion problem using stabilized finite elements. To the best of the authors knowledge, there are no post-processing results for this discretization to guarantee $\mathcal{O}(h^2)$ convergence.

Variational control discretization

In the variational approach due to Hinze [Hin05], the control is initially not discretized at all, i.e., we search $u_h \in L^2(\Omega_c)$. The practical realization of this approach is based on the optimality condition for the resulting semidiscrete problem, which is given by

$$\bar{u}_{h,\gamma} = P_\gamma\left(-\frac{1}{\gamma}\chi_{\Omega_c}\bar{p}_{h,\gamma}\right) \quad \text{pointwise in } \Omega_c,$$

where $\bar{p}_{h,\gamma}$ is the corresponding discrete optimal adjoint state. Note that $\bar{u}_{h,\gamma}$ is not an element of U_h^1, in general. However, it can be treated in practice by storing the discrete auxiliary variable

$q_h = -1/\gamma \chi_{\Omega_c} p_h$, or by other techniques; see [HV12]. In the numerical implementation, integrals of the form $(P_\gamma(q_h), \varphi_h)$ for $\varphi_h \in U_h^1$ have to be evaluated, which can be done by specialized quadrature formulas or by adaptive quadrature. However, this entails a high implementation effort.

The advantage of this method lies on the theoretical side: Since (in theory) the control is not discretized at all, there is no corresponding discretization error. Therefore, for a problem with control constraints, an L^2 error estimate for the controls of order $\mathcal{O}(h^2)$ can be derived directly; see [Hin05]. This can be carried over to the problem at hand with minor modifications; see [WW11, Corollary 4.5]. However, since we will not use this discretization concept (due to the high implementation effort), we will not go into further detail.

Piecewise linear controls with mass lumping

Since the controls are discretized as nodal Dirac measures in the unregularized problem, we are interested in a control discretization that yield an equivalent problem to (4.14) in the limiting case for $\gamma \to 0$. From the three approaches presented before, this is only the case for the variational discretization.

Remark 4.5. More precisely, we observe that the optimality conditions for the discretized problems for $\gamma = 0$ yields an optimality condition which implies the inequality

$$|(\bar{p}_{h,0}, \chi_{\Omega_c} \varphi_h)| \leq \alpha \int_{\Omega_c} \varphi_h \,\mathrm{d}x \quad \text{for all } \varphi_h \in U_h,$$

where $\bar{p}_{h,0}$ is the optimal adjoint state and U_h is the respective discrete control space. Note that only for $U_h = L^2(\Omega_c)$ this condition implies the inequality $|\bar{p}_{h,0}| \leq \alpha$ pointwise in Ω_c (as in the original discrete problem (4.14)). In the piece-wise constant case we only obtain $|P_h^0 \bar{p}_{h,0}| \leq \alpha$ and in the piece-wise linear case we obtain a variational inequality involving the Galerkin mass-matrix (which does not allow for a pointwise resolution, in general).

In the following, we consider an approach based on mass lumping, which yields an equivalent problem to (4.14) for $\gamma = 0$. We propose the following discretization, which is also derived directly from the regularized problem (4.14). It is given by

$$\begin{aligned} \min_{u_h \in U_h^1,\, y_h \in V_h} \quad & \frac{1}{2}\|y_h - y_d\|^2_{L^2(\Omega_o)} + \alpha \|u_h\|_{L^1(\Omega_c),h} + \frac{\gamma}{2}\|u_h\|^2_{L^2(\Omega_c),h} \\ \text{subject to} \quad & (\nabla y_h, \nabla \varphi_h) = (\chi_{\Omega_c} u_h, \varphi_h)_h \quad \text{for all } \varphi_h \in V_h. \end{aligned} \tag{4.32}$$

Here, the terms with the subscript h are computed by using the trapezoidal rule for the evaluation of the integrals from each cell (cf., e.g., [AKV92; Ran08]). For a given function $f \in \mathcal{C}_0(\Omega_c)$ we define

$$\left[\int_{\Omega_c} f(x) \,\mathrm{d}x\right]_h = \sum_{K \in \mathcal{T}_h^c} Q_{\mathrm{Trap},K}(f),$$

where $Q_{\mathrm{Trap},K}$ is the trapezoidal rule on the cell K. It is easy to see that with the nodal interpolation i_h we have

$$\left[\int_{\Omega_c} f(x) \,\mathrm{d}x\right]_h = \left[\int_{\Omega_c} (i_h f)(x) \,\mathrm{d}x\right]_h = \int_{\Omega_c} (i_h f)(x) \,\mathrm{d}x.$$

The first equality holds since the trapezoidal rule only evaluates the function in the grid points x_n, where we have $(i_h f)(x_n) = f(x_n)$, and the second holds since the trapezoidal rule is exact

for cell-wise linear functions. With this, we can directly derive that for $u_h \in U_h^1$ and $\varphi_h \in V_h$ we have

$$\begin{aligned} \|u_h\|_{L^1(\Omega_c),h} &= \left[\int_{\Omega_c} |u_h(x)| \,\mathrm{d}x\right]_h = \sum_{n=1}^{N_c} d_n \, |u_h(x_n)|, \\ \|u_h\|^2_{L^2(\Omega_c),h} &= \left[\int_{\Omega_c} u_h(x)^2 \,\mathrm{d}x\right]_h = \sum_{n=1}^{N_c} d_n \, u_h(x_n)^2, \\ \text{and} \quad (\chi_{\Omega_c} u_h, \varphi_h)_h &= \left[\int_{\Omega_c} u_h(x)\varphi(x) \,\mathrm{d}x\right]_h = \sum_{n=1}^{N_c} d_n \, u_h(x_n)\varphi_h(x_n), \end{aligned} \tag{4.33}$$

where $(d_n)_n$ for $n = 1, \dots, N_c$ is the diagonal of the *lumped mass matrix*. It is given by

$$d_n = \int_{\Omega_c} e_n(x) \,\mathrm{d}x \quad \text{for } n \in \{\, 1, \dots, N_c \,\}.$$

As a consequence of the discrete form of the lumped terms we easily see that the discrete problem for $\gamma = 0$ corresponds to (4.14).

Proposition 4.28. *Let $\bar{u}_h = \sum_n u_n \delta_{x_n} \in \mathcal{M}_h$ be an optimal solution of* (4.14)*. Then the finite element function given by*

$$\bar{u}_{h,0} = \chi_{\Omega_c} \sum_{n=1}^{N_c} \frac{u_n}{d_n} \, e_n \in U_h^1$$

solves the regularized problem (4.32) *for $\gamma = 0$.*

Proof. This is a direct consequence of the coordinate-wise expression for the lumped terms and the algebraic identities

$$\|\bar{u}_{h,0}\|_{L^1(\Omega_c),h} = \sum_{n=1}^{N_c} |u_n| = \|\bar{u}_h\|_{\mathcal{M}(\Omega_c)}$$

and

$$(\chi_{\Omega_c} \bar{u}_{h,0}, \varphi_h)_h = \sum_{n=1}^{N_c} u_n \varphi(x_n) = \langle \chi_{\Omega_c} \bar{u}_h, \varphi_h \rangle,$$

which holds for any $\varphi_h \in V_h$. □

It is well known that the lumped L^2 norm is equivalent to the L^2 norm on the discrete space, i.e., we have the inequalities

$$c_d \, \|u_h\|^2_{L^2(\Omega_c),h} \leq \|u_h\|^2_{L^2(\Omega_c)} \leq \|u_h\|^2_{L^2(\Omega_c),h} \quad \text{for all } u_h \in U_h^1.$$

for a constant c_d depending only on the dimension d. In fact, we have $c_2 = 1/4$ and $c_3 = 1/5$; cf. [CHW12a, Remark 3.1]. Therefore, it is justified to endow the space U_h^1 with the lumped inner product. To see that (4.32) has a unique solution, it suffices to see that the corresponding reduced cost functional is strongly convex.

Proposition 4.29. *The discrete regularized problem* (4.32) *has a unique optimal solution $(\bar{u}_{h,\gamma}, \bar{y}_{h,\gamma}) \in U_h^1 \times V_h$.*

By standard arguments, we obtain the following optimality conditions.

Proposition 4.30. *Let $(\bar{u}_{h,\gamma}, \bar{y}_{h,\gamma})$ be the optimal solution of (4.32). There exists a corresponding adjoint state $\bar{p}_{h,\gamma}$ solving the adjoint equation*

$$(\nabla\varphi_h, \nabla\bar{p}_{h,\gamma}) = (\chi_{\Omega_o}(\bar{y}_{h,\gamma} - y_d), \varphi_h) \quad \textit{for all } \varphi_h \in V_h. \tag{4.34}$$

It fulfills the following variational inequality

$$-(\bar{p}_{h,\gamma} + \gamma\bar{u}_{h,\gamma}, \chi_{\Omega_c}(u_h - \bar{u}_{h,\gamma}))_h + \alpha\|\bar{u}_{h,\gamma}\|_{L^1(\Omega_c),h} \le \alpha\|u_h\|_{L^1(\Omega_c),h} \quad \textit{for all } u_h \in U_h^1, \tag{4.35}$$

which can be equivalently expressed with the coordinate-wise formula

$$\bar{u}_{h,\gamma}(x_n) = \hat{P}_\gamma\left(-\frac{1}{\gamma}\,\bar{p}_{h,\gamma}(x_n)\right) = -\frac{1}{\gamma}\,\mathrm{shrink}_\alpha\left(\bar{p}_{h,\gamma}(x_n)\right) \quad \textit{for } n = 1, \dots, N_c. \tag{4.36}$$

Proof. The derivation of the adjoint equation and the variational inequality is standard. Let us justify the coordinate-wise formula: By (4.33) the variational inequality is equivalent to

$$\sum_{n=1}^{N_c} d_n\big[-(\bar{p}_{h,\gamma}(x_n) + \gamma\bar{u}_{h,\gamma}(x_n))(u_h(x_n) - \bar{u}_{h,\gamma}(x_n)) + \alpha|u_{h,\gamma}(x_n)|\big] \le \sum_{n=1}^{N_c} d_n\,\alpha|u_h(x_n)|$$

for all $u_h \in U_h^1$. Let now $i \in \{1, \dots, N_c\}$ be fixed. Choosing u_h with $u_h(x_i) = \tilde{u} \in \mathbb{R}$ and $u_h(x_n) = \bar{u}_{h,\gamma}(x_n)$ for $n \neq i$, we derive

$$-(\bar{p}_{h,\gamma}(x_i) + \gamma\bar{u}_{h,\gamma}(x_i))(\tilde{u} - \bar{u}_{h,\gamma}(x_i)) + \alpha|u_{h,\gamma}(x_i)| \le \alpha|\tilde{u}| \quad \text{for all } \tilde{u} \in \mathbb{R}.$$

This is the (sufficient) optimality condition for the problem

$$\bar{u}_{h,\gamma}(x_i) = \operatorname*{argmin}_{u\in\mathbb{R}}\left[\frac{\gamma}{2}u^2 + \bar{p}_{h,\gamma}(x_i)u + \alpha|u|\right],$$

which has the solution as given in (4.36). □

Remark 4.6. A similar mass lumping for discretization of L^1 control costs is also employed by Casas, Herzog, and Wachsmuth [CHW12a]. However, they consider the standard state equation (4.31) (without lumping) while the control cost term is evaluated as in (4.32) (with lumping). As a consequence they also obtain a coordinate-wise formula as in (4.36), where, however, the adjoint state is replaced with its Carstensen quasi interpolant (see [Car99]). They prove an L^2 error estimate for the controls of order $\mathcal{O}(h)$, which is confirmed by the numerical experiments. This is the same rate as for the $\mathcal{P}_0$ discretization and worse than the typical $\mathcal{O}(h^{3/2})$ rate for $\mathcal{P}_1$ and the $\mathcal{O}(h^2)$ rate for the variational or post-processing approach. In the following section, we provide an analysis for problem (4.32), where we obtain $\mathcal{O}(h^2)$ under an established structural assumption.

4.5.4. Finite element error analysis

In the following, for convenience of notation, we will abbreviate the optimal triples of (4.25) and (4.32) as

$$(\bar{u}, \bar{y}, \bar{p}) = (\bar{u}_\gamma, \bar{y}_\gamma, \bar{p}_\gamma) \quad \text{and} \quad (\bar{u}_h, \bar{y}_h, \bar{p}_h) = (\bar{u}_{h,\gamma}, \bar{y}_{h,\gamma}, \bar{p}_{h,\gamma}),$$

i.e., we drop the subscript γ. To obtain an error estimate for (4.32), we have to work with the discrete (lumped) L^2 norm as defined above. We first obtain an estimate for the discrete

error $\|\bar{u}_h - i_h\bar{u}\|_{L^2(\Omega_c),h}$. Note that this discrete error can be significantly smaller than the error $\|\bar{u}_h - \bar{u}\|_{L^2(\Omega_c)}$, which is typically restricted to $\mathcal{O}(h^{3/2})$ by the low regularity of the optimal control. This technique is also known in the context of discretization of optimal control problems with ordinary differential equations; see, e.g., [DHV01].

In the following we denote by $S_h \colon U_h^1 \to V_h$ the solution operator $S_h(u_h) = y_h$, where y_h is the solution of the discrete state equation

$$(\nabla y_h, \nabla \varphi_h) = (\chi_{\Omega_c} u_h, \varphi_h)_h \quad \text{for all } \varphi_h \in V_h.$$

Note that we could also extend the definition of S_h for arbitrary continuous controls $u \in \mathcal{C}_0(\Omega_c)$. However, this is equivalent to setting $y_h = S_h(i_h u)$ in these cases. We also denote by $S_{\mathrm{dual},h} \colon L^2(\Omega) \to V_h$ the dual solution operator $S_{\mathrm{dual},h}(f) = p_h$ corresponding to the dual equation

$$(\nabla \varphi_h, \nabla p_h) = (f, \varphi_h) \quad \text{for all } \varphi_h \in V_h.$$

We define the reduced discrete cost functional j_h, the corresponding differentiable part f_h, and the convex part ψ_h for $u_h \in U_h^1$ as

$$j_{h,\gamma}(u_h) = f_{h,\gamma}(u_u) + \psi_h(u_h) = J(S_h(u_h)) + \frac{\gamma}{2}\|u_h\|^2_{L^2(\Omega_c),h} + \|u_h\|_{L^1(\Omega_c),h}.$$

By a straightforward computation, we obtain the expressions for the first and second derivatives of $f_{h,\gamma}$ as

$$\begin{aligned} f'_{h,\gamma}(u_h)(\delta u_h) &= (\chi_{\Omega_o}(S_h(u_h) - y_d), S_h(\delta u_h)) + \gamma(u_h, \delta u_h)_h \\ \text{and} \quad f''_{h,\gamma}(u_h)(\delta u_h, \tau u_h) &= (\chi_{\Omega_o} S_h(\tau u_h), S_h(\delta u_h)) + \gamma(\tau u_h, \delta u_h)_h \end{aligned}$$

for any $\delta u_h, \tau u_h \in U_h^1$. Using the adjoint solution operator, we can also introduce corresponding expressions for the gradient. In the discrete setting it is natural to perform the identification of the gradient with the lumped mass matrix (to avoid a spurious inverse mass matrix). Since we have equipped U_h^1 with the lumped norm we require

$$(\nabla f_{h,\gamma}(u_h), \varphi_h)_h = f'_{h,\gamma}(u_h)(\varphi_h) \quad \text{for all } \varphi_h \in U_h^1.$$

This directly leads to the formula

$$\nabla f_{h,\gamma}(u_h) = \chi_{\Omega_c} S_{\mathrm{dual},h}(\chi_{\Omega_o}(S_h(u_h) - y_d)) + \gamma u_h$$

for the reduced gradient. A similar representation holds for the reduced Hessian.

Estimates for the state

To analyze the error of the state equation for a fixed control, it is necessary to consider the quadrature error. We have the following standard result.

Lemma 4.31. *Let u_h and φ_h be elements of U_h^1. For the mass lumping of the inner product we have the a priori estimate*

$$|(u_h, \varphi_h) - (u_h, \varphi_h)_h| \leq Ch^2 \|\nabla u_h\|_{L^2(\Omega_c)} \|\nabla \varphi_h\|_{L^2(\Omega_c)}.$$

This standard result can be found, e.g., in [AKV92; Ran08] (for the case $d = 2$). The proof is based on the usual transformation and localization argument and given in Appendix A.3 (also for the case $d = 3$).

Thereby, we can provide an estimate for the state with a fixed control. Since we evaluate the right-hand side with mass lumping, we are canonically restricted to continuous controls.

Theorem 4.32. *Let $u \in \mathcal{C}_0(\Omega_c)$ and denote by $y = S(u)$ and $y_h = S_h(i_h u)$ the state and discrete state solution, respectively. We have the a priori estimate*

$$\|y - y_h\|_{L^2(\Omega)} \leq Ch^2 \left(\|u\|_{L^2(\Omega_c)} + \|\nabla i_h u\|_{L^2(\Omega_c)} \right) + \|i_h u - u\|_{L^1(\Omega_c)}.$$

Proof. We introduce the Ritz projection $\hat{y}_h \in V_h$ of y as the solution of the conforming state equation (4.31) for $u_h = u$ and split the error as

$$\|y - y_h\|_{L^2(\Omega)} \leq \|y - \hat{y}_h\|_{L^2(\Omega)} + \|\hat{y}_h - y_h\|_{L^2(\Omega)}.$$

For the first term we apply the standard finite element error estimate

$$\|y - \hat{y}_h\|_{L^2(\Omega)} \leq Ch^2 \|u\|_{L^2(\Omega_c)},$$

using the classical Aubin-Nitsche duality argument. For the second term we use a (discrete) duality argument. We define the dual variable $w \in H_0^1(\Omega)$ and its Ritz projection $w_h \in V_h$ as the solutions of

$$(\nabla \varphi, \nabla w) = (\hat{y}_h - y_h, \varphi) \quad \text{for all } \varphi \in H_0^1(\Omega), \tag{4.37}$$

$$(\nabla \varphi_h, \nabla w_h) = (\hat{y}_h - y_h, \varphi_h) \quad \text{for all } \varphi_h \in V_h. \tag{4.38}$$

With this, we can write

$$\begin{aligned} \|\hat{y}_h - y_h\|_{L^2(\Omega)}^2 &= (\hat{y}_h - y_h, \hat{y}_h - y_h) = (\nabla(\hat{y}_h - y_h), \nabla w_h) = (u, w_h) - (i_h u, w_h)_h \\ &= (u - i_h u, w_h) + (i_h u, w_h) - (i_h u, w_h)_h \\ &\leq \|u - i_h u\|_{L^1(\Omega_c)} \|w_h\|_{L^\infty(\Omega)} + Ch^2 \|\nabla i_h u\|_{L^2(\Omega_c)} \|\nabla w_h\|_{L^2(\Omega)}, \end{aligned}$$

using the fact that $\hat{y}_h$ and y_h solve the respective state equations, Hölder's inequality and Lemma 4.31. Inserting w_h as an admissible test function into (4.38) and using the Cauchy-Schwarz inequality, we immediately obtain the stability estimate

$$\|\nabla w_h\|_{L^2(\Omega)} \leq C \|\hat{y}_h - y_h\|_{L^2(\Omega)}.$$

Furthermore we use the stability of the nodal interpolation and an inverse estimate to derive

$$\|w_h\|_{L^\infty(\Omega)} \leq \|i_h w\|_{L^\infty(\Omega)} + \|i_h w - w_h\|_{L^\infty(\Omega)} \leq \|w\|_{L^\infty(\Omega)} + Ch^{-d/2} \|i_h w - w_h\|_{L^2(\Omega)}.$$

For the first term we apply the Sobolev embedding and elliptic regularity theory to obtain

$$\|w\|_{L^\infty(\Omega)} \leq C \|w\|_{H^2(\Omega)} \leq C \|\hat{y}_h - y_h\|_{L^2(\Omega)}.$$

For the second term we use standard a priori estimates for the nodal interpolation and the Ritz projection to obtain

$$h^{-d/2} \|i_h w - w_h\|_{L^2(\Omega)} \leq h^{-d/2} \left(\|i_h w - w\|_{L^2(\Omega)} + \|w - w_h\|_{L^2(\Omega)} \right) \leq Ch^{2-d/2} \|\hat{y}_h - y_h\|_{L^2(\Omega)}.$$

Combining the estimates, it follows that

$$\|\hat{y}_h - y_h\|_{L^2(\Omega)}^2 \leq C \left(\|u - i_h u\|_{L^1(\Omega_c)} + h^2 \|\nabla i_h u\|_{L^2(\Omega_c)} \right) \|\hat{y}_h - y_h\|_{L^2(\Omega)}.$$

Dividing by $\|\hat{y}_h - y_h\|_{L^2(\Omega)}$ and combining with the first estimate, we conclude the proof. □

Estimates for the optimal solutions

In this section we will derive a priori estimates for the discretization of the regularized problem. Note that we will take care to make explicit the dependency of the estimates on the regularization parameter γ. All the generic constants which appear in the following are independent of both h and γ. As a first step we obtain the following stability estimate.

Lemma 4.33. *Let $(\bar{u}, \bar{y}, \bar{p})$ and $(\bar{u}_h, \bar{y}_h, \bar{p}_h)$ be the optimal triples of (4.5) and (4.32), respectively. It holds*

$$\gamma \| i_h \bar{u} - \bar{u}_h \|_{L^2(\Omega_c),h} \leq \| i_h \bar{p} - p_h(i_h \bar{u}) \|_{L^2(\Omega_c),h},$$

where $p_h(i_h\bar{u}) = S_{\mathrm{dual},h}(\chi_{\Omega_o}(S_h(i_h\bar{u}) - y_d))$ is the discrete adjoint state corresponding to $i_h\bar{u}$.

Proof. For arbitrary $u_h \in U_h^1$ we have

$$\begin{aligned}\gamma \| i_h \bar{u} - \bar{u}_h \|^2_{L^2(\Omega_c),h} &\leq f_h''(u_h)(i_h\bar{u} - \bar{u}_h, i_h\bar{u} - \bar{u}_h) = f_h'(i_h\bar{u})(i_h\bar{u} - \bar{u}_h) - f_h'(\bar{u}_h)(i_h\bar{u} - \bar{u}_h) \\ &= (\chi_{\Omega_c}(p_h(i_h\bar{u}) + \gamma i_h\bar{u}), i_h\bar{u} - \bar{u}_h)_h - (\chi_{\Omega_c}(\bar{p}_h + \gamma\bar{u}_h), i_h\bar{u} - \bar{u}_h)_h \quad (4.39)\end{aligned}$$

For the second term we use the discrete optimality condition for $\bar{u}_h$ from Proposition 4.30. The variational inequality (4.35) implies that

$$-(\chi_{\Omega_c}(\bar{p}_h + \gamma\bar{u}_h), i_h\bar{u} - \bar{u}_h)_h \leq \psi_h(i_h\bar{u}) - \psi_h(\bar{u}_h).$$

Consequently, we use the continuous optimality condition for $\bar{u}$ from Theorem 4.23. By a coordinate-wise interpretation, we obtain

$$\begin{aligned}\psi_h(i_h\bar{u}) - \psi_h(\bar{u}_h) &= \sum_{n=1}^{N_c} d_n \left(\alpha|\bar{u}(x_n)| - \alpha|\bar{u}_h(x_n)|\right) \\ &\leq -\sum_{n=1}^{N_c} d_n \left(\bar{p}(x_n) + \gamma\bar{u}(x_n)\right)\left(\bar{u}(x_n) - \bar{u}_h(x_n)\right) = -(\chi_{\Omega_c}(i_h\bar{p} + \gamma i_h\bar{u}), i_h\bar{u} - \bar{u}_h)_h,\end{aligned}$$

where the second inequality follows from the pointwise interpretation of the variational inequality (4.26); see Remark 4.3. We arrive at

$$\gamma \| i_h \bar{u} - \bar{u}_h \|^2_{L^2(\Omega_c),h} = (\chi_{\Omega_c}(p_h(i_h\bar{u}) - i_h\bar{p}), i_h\bar{u} - \bar{u}_h)_h.$$

An application of the Cauchy-Schwarz inequality and dividing by $\| i_h \bar{u} - \bar{u}_h \|_{L^2(\Omega_c),h}$ yields the desired result. □

Now, we come to the main error estimate for the regularized problem.

Theorem 4.34. *Let $(\bar{u}, \bar{y}, \bar{p})$ and $(\bar{u}_h, \bar{y}_h, \bar{p}_h)$ be the optimal triples of (4.5) and (4.32), respectively. We have the a priori estimate*

$$\| i_h \bar{u} - \bar{u}_h \|_{L^2(\Omega_c)} \leq C\gamma^{-1} \left(h^2 \left(1 + \gamma^{-1}\right) \| \bar{y} - y_d \|_{L^2(\Omega_o)} + \| \bar{u} - i_h\bar{u} \|_{L^1(\Omega_c)} \right). \quad (4.40)$$

Proof. By the equivalence of the lumped norm and Lemma 4.33 we have that

$$\| i_h \bar{u} - \bar{u}_h \|_{L^2(\Omega_c)} \leq \| i_h \bar{u} - \bar{u}_h \|_{L^2(\Omega_c),h} \leq C\gamma^{-1} \| i_h \bar{p} - p_h(i_h\bar{u}) \|_{L^2(\Omega_c)}$$

We introduce the Ritz projection $\hat{p}_h = S_{\mathrm{dual},h}(\chi_{\Omega_o}(\bar{y} - y_d))$ of the optimal adjoint state $\bar{p}$ and split

$$\|i_h\bar{p} - p_h(i_h\bar{u})\|_{L^2(\Omega_c)} \leq \|i_h\bar{p} - \bar{p}\|_{L^2(\Omega_c)} + \|\bar{p} - \hat{p}_h\|_{L^2(\Omega_c)} + \|\hat{p}_h - p_h(i_h\bar{u})\|_{L^2(\Omega_c)}.$$

The first and second term can be estimated by

$$\|i_h\bar{p} - \bar{p}\|_{L^2(\Omega)} + \|\bar{p} - \hat{p}_h\|_{L^2(\Omega)} \leq Ch^2\|\bar{p}\|_{H^2(\Omega)} \leq Ch^2\|\bar{y} - y_d\|_{L^2(\Omega_o)}$$

by a standard interpolation estimate for the nodal interpolant, the properties of the Ritz-projection (the standard Aubin-Nitsche duality argument) and elliptic regularity. Furthermore, we have

$$\|\hat{p}_h - p_h(i_h\bar{u})\|_{L^2(\Omega_c)} \leq \|S_{\mathrm{dual},h}(\chi_{\Omega_o}(\bar{y} - S(i_h\bar{u})))\|_{L^2(\Omega)} \leq C\|\bar{y} - S_h(i_h\bar{u})\|_{L^2(\Omega_o)}$$

with a standard stability estimate for the adjoint solution operator on $L^2(\Omega)$. Now, we apply Theorem 4.32 to obtain

$$\|\bar{y} - S_h(i_h\bar{u})\|_{L^2(\Omega_o)} \leq C\left(h^2\left(\|\bar{u}\|_{L^2(\Omega_c)} + \|\nabla i_h\bar{u}\|_{L^2(\Omega_c)}\right) + \|i_h\bar{u} - \bar{u}\|_{L^1(\Omega_c)}\right).$$

The gradient of $i_h\bar{u}$ can be estimated further by using the pointwise optimality condition $\bar{u}(x) = \hat{P}_\gamma(-1/\gamma\,\bar{p}(x))$ for $x \in \Omega_c$. We have $\bar{p} \in H^2(\Omega) \cap H_0^1(\Omega)$, which is continuously embedded into $W_0^{1,q}(\Omega)$ for all $q \leq 6$. Since $\hat{P}_\gamma\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with constant one, we also have $\bar{u} \in W^{1,q}(\Omega_c)$; see, e.g., [Zie89, Theorem 2.1.11]. Therefore, we obtain

$$\|\nabla i_h\bar{u}\|_{L^2(\Omega_c)} \leq C\|\nabla i_h\bar{u}\|_{L^q(\Omega_c)} \leq C\|\nabla\bar{u}\|_{L^q(\Omega_c)} \leq C\gamma^{-1}\|\bar{p}\|_{W_0^{1,q}(\Omega)}$$

with the stability of the nodal interpolation in $W^{1,q}(\Omega)$ for $q > d$. With elliptic regularity we have

$$\|\bar{p}\|_{W_0^{1,q}(\Omega)} \leq C\|\bar{p}\|_{H^2(\Omega)} \leq C\|\bar{y} - y_d\|_{L^2(\Omega_o)}$$

and conclude the proof. □

In Theorem 4.34 there still appears the interpolation error $\bar{u} - i_h\bar{u}$, which depends on the unknown solution. A generic estimate using $\bar{u} \in W^{1,q}(\Omega_c)$ for $q > d$ as in the proof of Theorem 4.34 yields only an estimate of the order $\mathcal{O}(h)$. To obtain an optimal $\mathcal{O}(h^2)$ bound for this interpolation error we require an additional assumption. We will see that, to obtain optimal estimates, it is important that this error has to be controlled only in the L^1 norm. For the same error in L^p with $1 \leq p \leq \infty$ we can only expect an estimate of the order $\mathcal{O}(h^{1+1/p})$. In the following, we adapt an established structural assumption on the optimal solution from the literature; cf., e.g., [MR04; Rös06; BV07]. It states, roughly speaking, that the interface between active and inactive sets of the optimal control has to be sufficiently regular. For instance, it is fulfilled if the "kink" set $\{\, x \in \Omega_c \mid \bar{p}(x) = \pm\alpha \,\}$ is a smooth $d-1$ dimensional submanifold of the control set.

Assumption 4.2. Let $(\bar{u}, \bar{y}, \bar{p}) = (\bar{u}_\gamma, \bar{y}_\gamma, \bar{p}_\gamma)$ be the optimal triple of (4.25). Define the set

$$\Omega_{\mathrm{kink}} = \bigcup \{\, K \in \mathcal{T}_h^c \mid \{\, x \in K \mid \bar{p}_\gamma(x) = \pm\alpha \,\} \neq \emptyset \,\},$$

composed of the cells where the optimal control $\bar{u}$ has a kink. Suppose that there exists a constant C_{kink} independent of h, such that the size of this set is bounded by

$$|\Omega_{\mathrm{kink}}| \leq C_{\mathrm{kink}}h.$$

Note that the constant in Assumption 4.2 can not be expected to be independent of γ in the general case. Furthermore, Assumption 4.1 on the desired state needs to be strengthened. We need $y_d \in L^p(\Omega)$ for some $p > d$, to obtain Lipschitz continuity of the adjoint state. These assumptions imply an estimate for the interpolation error.

Proposition 4.35. *Suppose that $y_d \in L^p(\Omega)$ for $p > d$ and that Assumption 4.2 holds. Then we have*

$$\|\bar{u} - i_h\bar{u}\|_{L^1(\Omega_c)} \leq C\gamma^{-1}h^2 \left(C_{\text{kink}}\|\bar{p}\|_{W^{1,\infty}(\Omega_{\text{kink}})} + \|\bar{p}\|_{H^2(\Omega)}\right).$$

Proof. It remains to estimate $\|\bar{u} - i_h\bar{u}\|_{L^1(\Omega_c)} = \|\bar{u} - i_h\bar{u}\|_{L^1(\Omega_{\text{kink}})} + \|\bar{u} - i_h\bar{u}\|_{L^1(\Omega_c\setminus\Omega_{\text{kink}})}$. On Ω_{kink} we use the estimate

$$\|\bar{u} - i_h\bar{u}\|_{L^1(\Omega_{\text{kink}})} \leq |\Omega_{\text{kink}}|\|\bar{u} - i_h\bar{u}\|_{L^\infty(\Omega_{\text{kink}})} \leq C_{\text{kink}}h\, Ch\|\bar{u}\|_{W^{1,\infty}(\Omega_{\text{kink}})}.$$

Using the equality $\bar{u} = P_\gamma(-1/\gamma\,\chi_{\Omega_c}\bar{p}) = -1/\gamma\,\text{shrink}_\alpha(\chi_{\Omega_c}\bar{p})$, we obtain the first part of the estimate. Note that the adjoint state is Lipschitz continuous due to $W^{2,q}(\Omega) \hookrightarrow W^{1,\infty}(\Omega)$, elliptic regularity, and the fact that $\chi_{\Omega_o}(\bar{y} - y_d) \in L^q(\Omega)$ for $q > d$. On the other cells we estimate

$$\begin{aligned}\|\bar{u} - i_h\bar{u}\|_{L^1(\Omega_c\setminus\Omega_{\text{kink}})} &\leq C\|\bar{u} - i_h\bar{u}\|_{L^2(\Omega_c\setminus\Omega_{\text{kink}})} \\ &= C\sum_{K\in\mathcal{T}_h^c,\,|\bar{p}|_K|\geq\alpha} \gamma^{-1}\|\bar{p} - i_h\bar{p}\|_{L^2(K)} \leq C\gamma^{-1}h^2\|\nabla^2\bar{p}\|_{L^2(\Omega)},\end{aligned}$$

where we have used that either $\bar{u}|_K = 0$ or $\bar{u}|_K = -1/\gamma\,(\bar{p}|_K \pm \alpha)$ for each cell not contained in Ω_{kink}. □

Before we summarize the given a priori estimates, we note that the solution dependent terms $\|\bar{p}_\gamma\|_{H^2(\Omega)}$ and $\|\bar{y}_\gamma - y_d\|_{L^2(\Omega_o)}$ can be bounded independently of γ. By elliptic regularity and minimality of $\bar{u}$ we have

$$\|\bar{p}_\gamma\|^2_{H^2(\Omega)} \leq C\|\bar{y}_\gamma - y_d\|^2_{L^2(\Omega_o)} \leq C\,j_\gamma(\bar{u}_\gamma) \leq C\,j_\gamma(0) = C\|y_d\|^2_{L^2(\Omega_o)},$$

which is obviously independent of γ. However, the term $\|\bar{p}_\gamma\|_{W^{1,\infty}(\Omega)}$ can only be estimated with a γ dependent constant, since a bound for $\bar{y}_\gamma - y_d$ in $L^q(\Omega_o)$ for $q > d$ is needed. We remark that for $d = 2$ such a bound could be obtained as in Proposition 4.24 (using Assumption 4.1) since $\|\bar{u}_\gamma\|_{L^1(\Omega_c)}$ is bounded independently of γ.

Theorem 4.36. *Let $(\bar{u}, \bar{y}, \bar{p}) = (\bar{u}_\gamma, \bar{y}_\gamma, \bar{p}_\gamma)$ and $(\bar{u}_h, \bar{y}_h, \bar{p}_h) = (\bar{u}_{h,\gamma}, \bar{y}_{h,\gamma}, \bar{p}_{h,\gamma})$ be the optimal triples of (4.5) and (4.32), respectively. Suppose that $y_d \in L^p(\Omega)$ for $p > d$ and that Assumption 4.2 holds. In the setting of Theorem 4.34 we obtain the a priori estimate*

$$\|i_h\bar{u} - \bar{u}_h\|_{L^2(\Omega_c)} + \|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} + \|\bar{p} - \bar{p}_h\|_{L^2(\Omega)} \leq C(\gamma)\,h^2,$$

where $C(\gamma) = C\gamma^{-1}(1 + \gamma^{-1} + \gamma^{-1}C_{\text{kink}}\|\bar{p}\|_{W^{1,\infty}(\Omega)})$. Furthermore, we define

$$\bar{u}_\sigma = P_\gamma(-1/\gamma\,\bar{p}_h) \in W^{1,\infty}(\Omega_c) \tag{4.41}$$

to be the post-processed discrete optimal control. For $\|\bar{u} - \bar{u}_\sigma\|_{L^2(\Omega_c)}$ we have the same estimate as above. If additionally $y_d \in L^\infty(\Omega)$, we even have the a priori estimate

$$\|\bar{u} - \bar{u}_\sigma\|_{L^\infty(\Omega_c)} \leq \tilde{C}(\gamma)\,h^2|\ln h|^r$$

with r as in Lemma 4.10, where $\tilde{C}(\gamma)$ depends continuously on γ.

Proof. We combine Proposition 4.35 with Theorem 4.34 to obtain the estimate for the controls. The estimates for the state and adjoint state follow from this with standard approximation and stability estimates for the Ritz projection. For the estimate of the post-processed controls, we derive first an L^∞ estimate for the optimal adjoint state using

$$\|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} \leq \|\bar{p} - \hat{p}_h\|_{L^\infty(\Omega)} + \|\hat{p}_h - \bar{p}_h\|_{L^\infty(\Omega)}.$$

The first term can be estimated with L^∞ estimates as in Lemma 4.10 by

$$\|\bar{p} - \hat{p}_h\|_{L^\infty(\Omega)} \leq Ch^2 |\ln h|^r \|\bar{y} - y_d\|_{L^\infty(\Omega)}$$

and the second one with a stability estimate as in Theorem 4.32. The result follows now from $\bar{u}_\sigma - \bar{u} = P_\gamma(-1/\gamma\,\bar{p}_h) - P_\gamma(-1/\gamma\,\bar{p})$ with the Lipschitz continuous superposition operator $P_\gamma(q)(x) = \hat{P}_\gamma(q(x))$ for $x \in \Omega_c$. □

Remark 4.7. The results for this section are written for a sparse control problem, where the nonsmooth term is given by

$$\psi(u) = \int_{\Omega_c} \hat{\psi}(u(x))\,\mathrm{d}x \quad \text{with } \hat{\psi}(\hat{u}) = \alpha|\hat{u}|.$$

However, the construction using mass lumping and the a priori estimates can be directly generalized to a problem with an arbitrary convex, proper, and lower semicontinuous functional $\hat{\psi}\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. In fact, everything up to and including Theorem 4.34 can be adapted line-by-line to this general setting by using the corresponding proximal map $\hat{P}_\gamma$. For instance, if we choose

$$\hat{\psi}(\hat{u}) = \mathbb{I}_{[u_a,u_b]}(\hat{u}) = \begin{cases} 0 & \text{for } u_a \leq u \leq u_b, \\ +\infty & \text{otherwise,} \end{cases}$$

we obtain a discretization concept and corresponding error estimate for a standard linear quadratic elliptic problem with (constant) box-constraints. In this case, Assumption 4.2 has to be modified in the obvious way.

4.6. Numerical results

We present some examples to verify the rates of convergence established in sections 4.3 and 4.4.

Example for spatial dimension two

We take $\Omega = B_1(0)$ as the unit ball and construct a radially symmetric example with the optimal state given as

$$\bar{y}(x) = -\frac{1}{2\pi} \ln(\max\{\rho, |x|\}),$$

with a kink in the radial direction at $\rho \in [0, 1)$. See Figure 4.1 for the representative cases $\rho = 1/2$ and $\rho = 0$. For $\rho = 0$ the state $\bar{y}$ is simply a Green's function, and the optimal control is then given by $\bar{u} = \delta_0$. For $\rho > 0$ we obtain the surface measure (given in terms of the 1-dimensional Hausdorff measure $\mathcal{H}^1$)

$$\bar{u} = \frac{1}{2\pi\rho} \mathcal{H}^1|_{\partial B_\rho(0)}$$

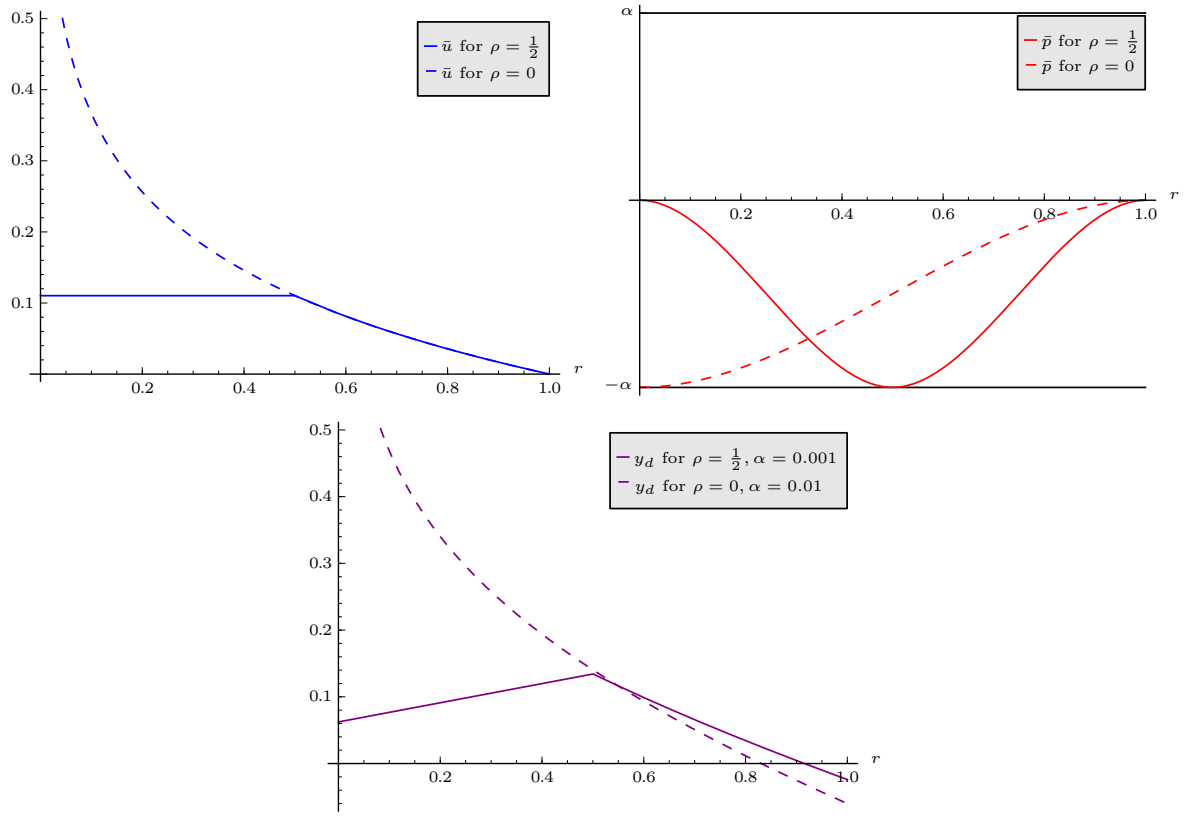


Figure 4.1.: Radially symmetric example for the unit circle in $\mathbb{R}^2$ in radial direction r

which, due to the choice of scaling, has a norm of $\|\bar{u}\|_{\mathcal{M}(\Omega)} = 1$. The optimal dual state can then be chosen as any element in $H^2(\Omega) \cap H_0^1(\Omega)$, such that $|\bar{p}| \leq \alpha$ and $\bar{p}|_{\partial B_\rho(0)} = -\alpha$. We make the specific choice

$$\bar{p}(x) = h(|x|),$$

where $h \in \mathcal{C}^1([0,1])$ is a piecewise cubic polynomial interpolating $h(0) = h(1) = 0$, $h(\rho) = -\alpha$ with the choices $h'(\rho) = h'(0) = h'(1) = 0$ (for $\rho = 0$, the conditions $h(0) = h'(0) = 0$ are dropped). This yields $\bar{p} \in \mathcal{C}^1(\Omega)$, which is piecewise twice continuously differentiable with bounded second derivatives, and a matching desired state $y_d \in L^\infty(\Omega)$ can be computed in strong formulation as

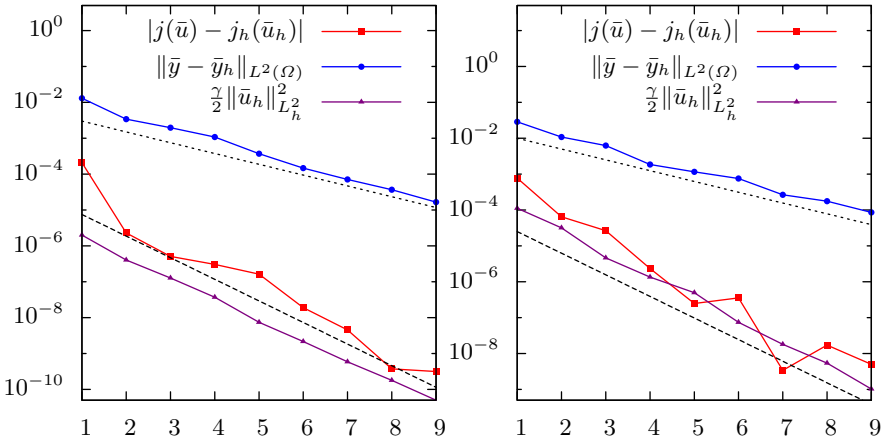
$$y_d = \Delta \bar{p} + \bar{y},$$

as depicted in Figure 4.1 for $\rho \in \{0, 1/2\}$. For the convenience of the reader, the exact formula for y_d is given by

$$y_d(r) = \begin{cases} \alpha \frac{6(3r-2\rho)}{\rho^3} - \frac{1}{2\pi}\ln(\rho) & \text{for } r < \rho \\ \alpha \frac{6\left(3r^2-2r\rho-2r+\rho\right)}{(\rho-1)^3 r} - \frac{1}{2\pi}\ln(r) & \text{for } r \geq \rho, \end{cases}$$

where $r = |x|$.

The convergence rates for a choice of $\rho = 1/2$ and $\rho = 0$ are given in Figure 4.2. The inital grid (refinement level 0) consists of five cells, a small square in the middle and four additional trapezoids at each edge, glued together at the corners. For both examples we plot the error in



(a) Errors for $\rho = 1/2$, $\alpha = 0.001$ (b) Errors for $\rho = 0$, $\alpha = 0.01$

Figure 4.2.: Convergence rates for $d = 2$ at different refinement levels.

the cost functional $J(\bar{u}, \bar{y}) - J(\bar{u}_h, \bar{y}_h)$ and the L^2-error in the state variable. The dashed lines indicate the orders of convergence $\mathcal{O}(h^2)$ and $\mathcal{O}(h)$, which are what theory predicts for the respective quantities (up to logarithmic contributions). Since the regularization is present in the numerical computations, we also report the size of the term $\gamma/2\,\|\bar{u}_h\|^2_{L^2(\Omega),h}$. As a parameter choice rule, at each refinement level the regularization parameter γ is decreased until

$$\frac{\gamma}{2}\|\bar{u}_h\|^2_{L^2(\Omega),h} \leq C_{\text{reg}} h^2$$

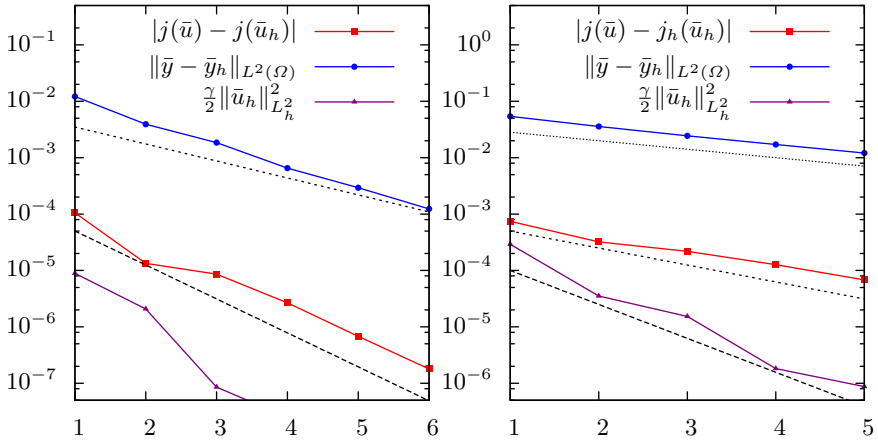
is fulfilled, where $C_{\text{reg}} > 0$ is a constant chosen heuristically in advance. This is done to ensure that at least the asymptotic best case convergence behaviour of the functional $\mathcal{O}(|\ln h|^\kappa h^2)$ should not be altered by the regularization. For instance, in Figure 4.2a we observe that the regularization term is an order of a magnitude smaller than the exact functional error, such that the reported error in the functional should be at least accurate in the first significant digit.

We see that the observed rates agree with the rates predicted by theory. In Figure 4.2a the rates seem to be even slightly better, however, this is far from conclusive. In Figure 4.2b, even though the rate for the functional is somewhat wiggly, we observe the expected rates. The wiggles could be caused by the fact that the initial mesh was perturbed slightly, and thus the approximation quality depends for a large part on the smallest distance of a grid-point to the origin, where the optimal control $\bar{u} = \delta_0$ is located. If we choose a mesh which has a point at the origin, the exact control is representable at each level, and the wiggles disappear. In the Dirac case, due to the low regularity of y_d, it is also clear that the rate of almost $\mathcal{O}(h)$ for the state error is the best theoretically possible.

Example for spatial dimension three

The construction of an example in three dimensions is completely analogous, except for the different Green's function

$$\bar{y}(x) = \frac{1}{4\pi}\left(\frac{1}{\max\{\rho, |x|\}} - 1\right),$$



(a) Errors for $\rho = 1/2$, $\alpha = 0.001$ (b) Errors for $\rho = 0$, $\alpha = 0.01$

Figure 4.3.: Convergence rates for $d = 3$ at different refinement levels.

thus we omit a detailed description. The final formula for y_d in this case is given by

$$y_d(r) = \begin{cases} \alpha \frac{6\,(4\,r - 3\,\rho)}{\rho^3} + \frac{1}{4\pi}\left(\frac{1}{\rho} - 1\right) & \text{for } r < \rho \\ \alpha \frac{6\left(4\,r^2 - 3\,r\rho - 3\,r + 2\,\rho\right)}{(\rho-1)^3 r} + \frac{1}{4\pi}\left(\frac{1}{r} - 1\right) & \text{for } r \geq \rho, \end{cases}$$

where $r = |x|$. The computational results can be seen in Figure 4.3. Note that the parameter choice rule for γ is simply the same as before. In this case, the general theory predicts an order of convergence close to $\mathcal{O}(h)$ for the functional and close to $\mathcal{O}(h^{1/2})$ for the L^2-error of the state. This is clearly observed in the case $\rho = 0$, where the optimal control $\bar{u}$ is a single Dirac delta function; see Figure 4.3b. In this case the rate for the state error is again the theoretically best possible. However, in the case $\rho = 1/2$, depicted in 4.3a, where y_d is bounded and the optimal control is a surface measure, the rates are clearly better. For visual comparison we plot the rates $\mathcal{O}(h)$ for the state in accordance with Theorem 4.21, and $\mathcal{O}(h^2)$ for the functional, which seems to be the closest match. Here, the order of convergence is the same as in the case $d = 2$.

5. A priori error analysis for a parabolic problem

In this chapter we will derive a priori finite element error estimates for the parabolic model problem given by

$$\min_{u\in\mathcal{M}(\Omega_c,L^2(I)),\, y\in Y} \quad \frac{1}{2}\|y-y_d\|^2_{L^2(I\times\Omega_o)} + \alpha\|u\|_{\mathcal{M}(\Omega_c,L^2(I))}, \tag{5.1a}$$

$$\text{subject to} \quad \begin{cases} \partial_t y - \Delta y = \chi_{\Omega_c} u & \text{on } I\times\Omega, \\ y = 0 & \text{on } I\times\partial\Omega, \\ y(0) = y_0 & \text{in } \Omega. \end{cases} \tag{5.1b}$$

Here, $\Omega \subset \mathbb{R}^2$ is a convex domain with a polygonal boundary $\partial\Omega$. The control variable u is searched for in the space of vector measures $\mathcal{M}(\Omega_c, L^2(I))$, where the control set $\Omega_c \subset \Omega$ is a (relatively closed) in Ω, i.e., we require

$$\Omega_c = \bar{\Omega}_c \setminus \partial\Omega.$$

We will make additional assumption on the form of Ω_c below (such as $\partial\Omega_c$ polygonal). The state variable y is the solution of the heat equation (5.1b) with zero Dirichlet boundary conditions and initial value $y_0 \in L^2(\Omega)$. We consider a standard linear quadratic tracking term on the observation domain $\Omega_o \subset \Omega$ with desired state $y_d \in L^2(I\times\Omega_o)$. For the purpose of optimal regularity and error estimates we will make the further assumption $y_d \in L^2(I, L^\infty(\Omega_o))$ and $y_0 \in H^1_0(\Omega)$; see below.

The problem setting (5.1) can be considered as a generalization of a pointwise parabolic control problem; a problem with a state equation of the form (5.1b) for the special case

$$u = \sum_{n=1}^{N} u_n(t)\delta_{x_n} \quad \text{for } u_n \in L^2(I), x_n \in \Omega_c,$$

where the positions $x_n \in \Omega_c$ are fixed and only the coefficients $u_n \in L^2(I)$ are subject to optimization. Pointwise parabolic control problems have been investigated by many authors; see, e.g., [Chr81; Lio92; DR00; MRVM00]. Finite element error estimates for parabolic pointwise control problems have been obtained by Gong, Hinze, and Zhou [GHZ14] and Leykekhman and Vexler [LV13]. In particular, we want to point out the latter paper, since the finite element error analysis of (5.1) given below will heavily rely on the estimates obtained there. A related, but different, sparse control problem is analyzed in Casas, Clason, and Kunisch [CCK13]. In Casas and Zuazua [CZ13] a control problem for the heat equation with measures on a subset of the parabolic cylinder is discussed. In the one-dimensional situation the authors are able to show that the minimizer is given by a finite sum of point sources. In Casas, Vexler, and Zuazua [CVZ14], where the control acts as an initial condition, the convergence of a corresponding finite element discretization is shown. With respect to finite element

discretization of optimal control problems governed by parabolic equations we also refer to Meidner and Vexler [MV08] for a standard linear quadratic problem with controls in $L^2(I \times \Omega)$.

We provide a numerical analysis for an appropriate finite element discretization of the problem (5.1). Following [Tho06; MV08; LV13], we employ a dG(0)cG(1) discretization with linear finite elements in space and piece-wise constants in time (which results in a variant of the implicit Euler scheme). Additionally, the control is discretized by nodal Dirac delta functions in space and piece-wise constants in time. We derive an a priori error estimate for the error between the objective functional values of order $\mathcal{O}(k+h^2)$ (up to a logarithmic factor) and between the optimal states on the observation domain of the continuous and discretized problems of order $\mathcal{O}(k^{1/2}+h)$ (up to a logarithmic factor), where k and h are temporal and spatial discretization parameters. This estimate seems to be optimal at least with respect to h; see the discussion in section 5.5. In comparison, the a priori estimates obtained in [CCK13] are of the order $\mathcal{O}(k^{1/2}+h)$ for the functional and $\mathcal{O}(k^{1/4}+h^{1/2})$ for the states. However, due to the more complicated structure of the controls considered there, the analysis given below is not directly extensible to their problem formulation; cf. the discussion in section 2.3.5. Most of the results of this chapter have already appeared in similar form in Kunisch, Pieper, and Vexler [KPV14].

This chapter is structured as follows. In section 5.1 we provide some necessary regularity results and derive consequences of the optimality conditions. Section 5.2 contains the discretization concept and analysis of the discrete problem. The error estimates mentioned above are contained in section 5.3; at first we derive estimates for the state with a fixed control, then we turn to estimates for the optimal solutions. The regularized problem is introduced and analyzed in section 5.4 and an estimate for the regularization error is provided. In section 5.5 we discuss a numerical example to verify the convergence rates in practice. Finally, section 5.6 describes the application to an inverse source location problem to demonstrate the practical applicability of the problem formulation with vector measures.

5.1. Optimality conditions

In this section we state some regularity results for the heat equation, which are improved w.r.t. the general analysis in section 2.3.2. They are mostly well-known but needed for an optimal error analysis. We also derive some consequences of the optimality system, i.e., a condition on the support and a higher regularity result for the optimal controls.

With respect to the general setting in section 2.3.2 we consider A to be the negative Laplacian with zero Dirichlet boundary conditions

$$A = -\Delta \colon W_0^{1,s}(\Omega) \to W^{-1,s}(\Omega)$$

on a the two dimensional polygonal and convex domain $\Omega \subset \mathbb{R}^2$. We remark that most of the following regularity results can be generalized in a suitable way to the general case and to three dimensions. Some of the finite element estimates employed in section 5.3.1 are, however, only available for $d = 2$ and the following techniques are in some cases restricted to two dimensions.

For a right-hand side $u \in \mathcal{M}(\Omega_c, L^2(I))$, the state $y = S(y_0, u)$ has the regularity

$$y \in L^2(I, W_0^{1,s}(\Omega)) \cap H^1(I, W^{-1,s}(\Omega))$$

for any $s < 2$; see Proposition 2.18. As before, we denote by S the corresponding solution operator with $y = S(u) = S(y_0, u)$ and abbreviate the reduced cost functional of (5.1a) by

$$j(u) = f(u) + \psi(u) = J(S(u)) + \alpha\|u\|_{\mathcal{M}(\Omega_c, L^2(I))},$$

where $J(y) = 1/2\,\|y - y_d\|^2_{L^2(I\times\Omega_o)}$ is the quadratic tracking term. We obtain the following additional estimate for the state solution.

Proposition 5.1. *The state solution $y = S(y_0, u)$ lies in the space $L^2(I, L^q(\Omega))$ for any $q \in [1, \infty)$ with the a priori estimate*

$$\|y\|_{L^2(I,L^q(\Omega))} \le C\, q\, (\|u\|_{\mathcal{M}(\Omega_c,L^2(I))} + \|y_0\|_{L^2(\Omega)}) \tag{5.2}$$

with a constant C independent of q.

Proof. We use the Sobolev embedding theorem and argue as in [LV13, Proposition 2.1] to obtain the dependence of the constant on q in (5.2). □

Proposition 5.2. *The state y is continuous in time in the sense that*

$$y \in \mathcal{C}(\bar{I}, (W_0^{1,s}(\Omega), W^{-1,s}(\Omega))_{1/2,2}) \hookrightarrow \mathcal{C}(\bar{I}, W^{-\varepsilon,s}(\Omega)) \tag{5.3}$$

for any $s < 2$ and $\varepsilon > 0$, where $(W_0^{1,s}(\Omega), W^{-1,s}(\Omega))_{1/2,2}$ is a real interpolation space.

Proof. The result follows by an application of the trace theorem [Ama95, Theorem III 4.10.2] and we refer to [Tri78, Theorem 4.6.1] for the embedding of the interpolation space. □

Remark 5.1. With methods as in Droniou and Raymond [DR00, Theorem 2.4], where a single point source is considered, it is possible to show that

$$y \in L^\infty(I, L^s(\Omega)) \quad \text{for any } s < 2.$$

Furthermore the mapping $t \mapsto y(t) \in L^s(\Omega)$ is continuous with respect to the weak topology in $L^s(\Omega)$.

As in section 2.3 we see that (5.1) is well-posed and obtain the following optimality system.

Theorem 5.3. *There exists a unique adjoint state $\bar{p} \in L^2(I, W_0^{1,s'}(\Omega))$ (with $s' > d$) corresponding to any optimal solution $(\bar{u}, \bar{y}) = (\bar{u}, S(\bar{u}))$ of (5.1). It satisfies*

$$\begin{cases} -\partial_t \bar{p} - \Delta \bar{p} = \chi_{\Omega_o}(\bar{y} - y_d) & \text{on } I \times \Omega, \\ \bar{p} = 0 & \text{on } I \times \partial\Omega, \\ \bar{p}(T) = 0 & \text{in } \Omega \end{cases} \tag{5.4}$$

in the sense of the standard weak formulation and

$$-\langle \chi_{\Omega_c}(u - \bar{u}), \bar{p}\rangle + \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))} \le \alpha\|u\|_{\mathcal{M}(\Omega_c,L^2(I))} \quad \text{for all } u \in \mathcal{M}(\Omega_c, L^2(I)). \tag{5.5}$$

Furthermore, the variational inequality (5.5) is equivalent to the two conditions

$$\|\chi_{\Omega_c}\bar{p}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} \le \alpha, \quad \text{and } \langle \chi_{\Omega_c}\bar{u}, \bar{p}\rangle = \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c),L^2(I))}. \tag{5.6}$$

This implies that the support of $\bar{u}$ is contained in the set $\{\, x \in \Omega_c \mid |\bar{p}(x)|_{L^2(I)} = \alpha \,\}$, and for the polar decomposition $\mathrm{d}\bar{u} = \bar{u}'\,\mathrm{d}|\bar{u}|$ we have

$$\bar{u}'(x) = \frac{1}{\alpha}\bar{p}(x) \quad \text{for } x \in \Omega_c \quad |\bar{u}|\text{-almost everywhere.} \tag{5.7}$$

For the adjoint state we can obtain improved regularity.

Lemma 5.4. *Let $f \in L^2(I, L^2(\Omega))$. The solution to the dual equation*

$$-\partial_t p - \Delta p = f, \quad p(T) = 0 \tag{5.8}$$

lies in the spaces $L^2(I, H^2(\Omega)) \cap H^1(I, L^2(\Omega))$ and $\mathcal{C}(\bar{I}, H_0^1(\Omega))$ with the corresponding estimate

$$\|\partial_t p\|_{L^2(I,L^2(\Omega))} + \|p\|_{L^2(I,H^2(\Omega))} + \|p\|_{\mathcal{C}(\bar{I},H_0^1(\Omega))} \leq C \|f\|_{L^2(I\times\Omega)}.$$

Proof. This can be proved by combining well-known techniques for parabolic equations (see, e.g., [Eva10]), with an elliptic regularity result for convex polygonal domains; see [Gri85]. □

We denote the solution operator of the dual equation by $p = S_{\text{dual}}(f)$. With the previous result and the Sobolev embedding, the adjoint state from Theorem 5.3 is an element of the space $L^2(I, \mathcal{C}^\delta(\bar{\Omega}))$ for any $\delta < 1$. For our purposes it will be convenient to exchange the order in which I and $\bar{\Omega}$ appear.

Proposition 5.5. *For any $0 < \delta \leq 1$, we have the continuous embedding*

$$L^2(I, \mathcal{C}^\delta(\bar{\Omega})) \hookrightarrow \mathcal{C}^\delta(\bar{\Omega}, L^2(I)).$$

Proof. Take any $v \in L^2(I, \mathcal{C}^\delta(\bar{\Omega}))$. For any x and $x + h \in \bar{\Omega}$ it holds

$$\begin{aligned}\|\bar{p}(x) - \bar{p}(x+h)\|_{L^2(I)}^2 &= \int_I |\bar{p}(t,x) - \bar{p}(t,x+h)|^2 \,\mathrm{d}t \\ &\leq \int_I \left(\sup_{\xi\in\bar{\Omega}} |\bar{p}(t,\xi) - \bar{p}(t,\xi+h)|^2 \right) \mathrm{d}t \leq \|\bar{p}\|_{L^2(I,\mathcal{C}^\delta(\bar{\Omega}))}^2 |h|^{2\delta}.\end{aligned}$$

Taking the square root implies the claim. □

By the regularity of the optimal adjoint state, we obtain now that the support of $\bar{u}$ must be compactly supported in Ω.

Proposition 5.6. *Define for $\eta > 0$ the domain Ω_η as*

$$\Omega_\eta = \{\, x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) > \eta \,\}.$$

There exists $\eta > 0$ such that $\operatorname{supp} \bar{u} \subset \Omega_c \cap \Omega_\eta$, where the constant η depends only on the data of the problem (5.1)*.*

Proof. With Lemma 5.4, the Sobolev embedding, and Proposition 5.5 we have $\bar{p} \in \mathcal{C}^\delta(\bar{\Omega}, L^2(I))$ for any $0 < \delta < 1$. The result now follows from the sparsity property of the support (2.26) and the zero Dirichlet boundary conditions. We have $\bar{p}(x) = 0$ for all $x \in \partial\Omega$ and can therefore choose $\eta < (\alpha/(2\|\bar{p}\|_{\mathcal{C}^\delta(\bar{\Omega},L^2(I))}))^{1/\delta}$ to finish the proof. □

Now, we make the following additional assumptions on the data.

Assumption 5.1. We require that the desired state fulfills

$$y_d \in L^2(I, L^\infty(\Omega_o)), \tag{5.9}$$

and that the initial condition fulfills $y_0 \in H_0^1(\Omega)$.

The regularity assumption on y_d is only slightly stronger than the natural regularity from Proposition 5.1. The assumption on y_0 will only be needed in section 5.3.1 since we employ optimal order estimates for the state equation in the $L^2(I \times \Omega)$ norm (see Meidner and Vexler [MV08]). With Assumption 5.1 and Proposition 5.1, the right-hand side of the adjoint equation (2.23) is even in $L^2(I, L^q(\Omega))$ for any $q < \infty$. For $p = S_{\mathrm{dual}}(f)$ with an arbitrary $f \in L^2(I, L^q(\Omega))$ we obtain that

$$\partial_t p,\, \Delta p \ \in L^2(I, L^q(\Omega)), \tag{5.10}$$

using maximal parabolic regularity; see, e.g., [GKR01]. However, from this we can not in general infer $L^2(I, W^{2,q}(\Omega))$ regularity without further assumptions on $\partial\Omega$. Nevertheless, we can obtain this regularity locally in the interior of the domain. To this purpose, we take the constant η from Proposition 5.6 and define a domain Ω^η that fulfills

$$\Omega_\eta \subset \Omega^\eta \subset \Omega_{\eta/2} \subset \Omega \quad \text{with } \partial\Omega^\eta \text{ of class } C^\infty.$$

It is clear, that such an Ω^η exists. For this domain, we can formulate the following result.

Lemma 5.7. *Let Ω^η as defined above with $\eta > 0$ from Proposition 5.6. We obtain for any solution of (5.8) with $f \in L^2(I, L^q(\Omega))$ for $q \in [1, \infty)$ that*

$$p|_{I\times\Omega^\eta} \in L^2(I, W^{2,q}(\Omega^\eta)) \cap H^1(I, L^q(\Omega^\eta)), \tag{5.11}$$

with the a priori estimate

$$\|p\|_{L^2(I,W^{2,q}(\Omega^\eta))} + \|\partial_t p\|_{L^2(I,L^q(\Omega^\eta))} \le C\, q \left(\|f\|_{L^2(I,L^q(\Omega_{\eta/3}))} + \eta^{-1} \|f\|_{L^2(I,L^2(\Omega))} \right).$$

Proof. See for instance [LV13, Lemma 2.2], where this is shown for any ball $B \subset \Omega_{\eta/2}$. The result follows since $\Omega^\eta \subset \Omega_{\eta/2}$ can be covered by finitely many balls $B \subset \Omega_{\eta/2}$. □

By applying this to the optimal adjoint state $\bar{p}$ and interpolating between both spaces from Lemma 5.7 with $\theta = 1 - \varepsilon$ for $\varepsilon > 0$ (see [Ama00, Theorem 5.2]) we obtain

$$\bar{p} \in \mathcal{C}^{1/2-\varepsilon}(\bar{I}, \mathcal{C}(\Omega^\eta)) \quad \text{for any } \varepsilon > 0. \tag{5.12}$$

Here, we have used the embedding $(W^{2,q}(\Omega^\eta), L^q(\Omega^\eta))_{1-\varepsilon,2} \hookrightarrow \mathcal{C}^\beta(\Omega^\eta)$ for $\beta = 2\varepsilon - d/q > 0$ and the compact embedding $\mathcal{C}^\beta(\Omega^\eta) \hookrightarrow \mathcal{C}(\Omega^\eta)$. With the help of the optimality conditions, we can now derive additional regularity for the optimal controls. We can show that $\bar{u}(t) \in \mathcal{M}(\Omega_c)$ is continuous in time.

Theorem 5.8. *With Assumption 5.1, we obtain the additional regularity*

$$\bar{u} \in \mathcal{C}^{1/2-\varepsilon}(I, \mathcal{M}(\Omega_c)) \quad \text{for any } \varepsilon > 0.$$

Proof. Using $\mathrm{d}\bar{u} = u'\,\mathrm{d}|u| = -1/\alpha\, \chi_{\Omega_c}\bar{p}\,\mathrm{d}|\bar{u}|$ we have for t_1 and t_2 in I that

$$\begin{aligned}\|\bar{u}(t_1) - \bar{u}(t_2)\|_{\mathcal{M}(\Omega_c)} &= \sup_{\|\varphi\|_{\mathcal{C}_0(\Omega)}=1} \langle \chi_{\Omega_c}(\bar{u}(t_1) - \bar{u}(t_2)), \varphi\rangle \\ &= \sup_{\|\varphi\|_{\mathcal{C}_0(\Omega)}=1} \int_{\Omega_c} \frac{1}{\alpha}(\bar{p}(t_2) - \bar{p}(t_1))\,\varphi\,\mathrm{d}|\bar{u}| \le \frac{1}{\alpha}\|\bar{p}(t_2) - \bar{p}(t_1)\|_{\mathcal{C}_0(\Omega_c\cap\Omega^\eta)}\,|\bar{u}|(\Omega_c),\end{aligned}$$

due to Proposition 5.6. Therefore, with the regularity (5.12) for $\bar{p}$ we have

$$\|\bar{u}(t_1) - \bar{u}(t_2)\|_{\mathcal{M}(\Omega_c)} \le \frac{1}{\alpha}\|\bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))}\|\bar{p}(t_2) - \bar{p}(t_1)\|_{\mathcal{C}_0(\Omega_c\cap\Omega^\eta)} \le |t_2 - t_1|^{1/2-\varepsilon}$$

for any t_1 and t_2 in I, which implies the claim. □

5.2. Discretization and numerical analysis

We discretize the state variable y with (linear) finite elements in space and discontinuous finite elements (of order $r \geq 0$) in time

$$y_{kh} \in X_k^r(I, V_h) \subset L^2(I, H_0^1(\Omega)).$$

Here, as in chapter 4, $V_h \subset H_0^1(\Omega)$ denotes the space of linear finite elements on a family of shape regular quasi-uniform triangulations $\{ \mathcal{T}_h \}_h$; see, e.g., [BS08]. The finite element space associated with $\mathcal{T}_h$ is defined as before by

$$V_h = \{ v_h \in \mathcal{C}_0(\Omega) \mid v_h|_K \in \mathcal{P}_1(K) \text{ for } K \in \mathcal{T}_h \}$$

The discretization parameter h denotes the maximal diameter of cells $K \in \mathcal{T}_h$. Furthermore, we suppose that Ω_c can be written as the union of a collection of cells or faces of $\mathcal{T}_h$ for all h; see section 4.2. For the time discretization we define for any Banach space V the semidiscrete space

$$X_k^r(I, V) = \{ v_k \in L^2(I, V) \mid v_k|_{I_m} \in \mathcal{P}_r(I_m, V),\ m = 1, 2, \ldots, M \}$$

as discontinuous, Banach space valued, piecewise polynomial functions on the disjoint partition of the temporal interval

$$\bar{I} = \{ 0 \} \cup I_1 \cup I_2 \cup \ldots \cup I_M,$$

where $I_m = (t_{m-1}, t_m]$ and $0 = t_0 < t_1 < \cdots < t_M = T$. By $k_m = t_m - t_{m-1}$ we denote the step length and by $k = \max_m k_m$ the maximum thereof. We employ the notation

$$w_m^- = \lim_{\varepsilon \to 0^+} w(t_m - \varepsilon), \quad w_m^+ = \lim_{\varepsilon \to 0^+} w(t_m + \varepsilon), \quad [w]_m = w_m^+ - w_m^-$$

for the left and right sided limits and the jump term (for any w where these limits are defined).

The discrete state equation is then given with the bilinear form

$$B(y, \varphi) = \sum_{m=1}^{M} \langle \partial_t y, \varphi \rangle_{I_m} + (\nabla y, \nabla \varphi)_I + \sum_{m=1}^{M-1} ([y]_m, \varphi_m^+) + (y_0^+, \varphi_0^+), \tag{5.13}$$

defined for $y, \varphi \in X_k^r(I, V_h)$. The distributional derivative of a discrete function $\partial_t y_{kh}|_{I_m}$ is given by the classical derivative of the polynomial (and vanishes for $r = 0$). The duality pairing $\langle \cdot\,, \cdot \rangle_{I_m}$ denotes the pairing of $L^2(I_m, W^{-1,s}(\Omega))$ with its dual. Therefore this definition can be extended to $y \in X_k^r(I, V_h) + Y^s$ and $\varphi \in X_k^r(I, V_h) + X^{s'}$. Furthermore, by applying integration by parts to (5.13) we obtain the equivalent dual formulation

$$B(y, \varphi) = -\sum_{m=1}^{M} \langle y, \partial_t \varphi \rangle_{I_m} + (\nabla y, \nabla \varphi)_I + \sum_{m=1}^{M-1} (-y_m^-, [\varphi]_m) + (y_M^-, \varphi_M^-). \tag{5.14}$$

Then, for any right hand side $u \in \mathcal{M}(\Omega_c, L^2(I))$ the discrete dG(r)cG(1) formulation of the state equation for the discretized state $y_{kh} \in X_k^r(I, V_h)$ is given as

$$B(y_{kh}, \varphi_{kh}) = \langle \chi_{\Omega_c} u, \varphi_{kh} \rangle + (y_0, \varphi_{kh,0}^+). \tag{5.15}$$

for all $\varphi_{kh} \in X_k^r(I, V_h)$. Since the right hand side is a linear functional on the discrete solution space, existence of a unique solution can be derived with standard arguments (see Thomée [Tho06]). Therefore, we can define a discrete solution operator with $y_{kh} = S_{kh}(u) =$

$S_{kh}(y_0, u)$. This operator and the bilinear form B are compatible with the continuous state solution $y = S(u)$ in the sense that

$$B(y, \varphi) = \langle \chi_{\Omega_c} u, \varphi \rangle + (y_0, \varphi_0^+), \tag{5.16}$$

for any $\varphi \in L^2(I, W_0^{1,s'}(\Omega))$ with $s' > 2$ such that the limits φ_m^+ for $m = 0, \dots, M-1$ are well defined in $W^{\varepsilon,s'}(\Omega)$ for some $\varepsilon > 0$. This follows from the state equation (2.20) since the jump terms in (5.13) vanish due to Proposition 5.2. With this we can verify the Galerkin orthogonality

$$B(y - y_{kh}, \varphi_{kh}) = 0, \tag{5.17}$$

for all $\varphi_{kh} \in X_k^r(I, V_h)$ and therefore y_{kh} is also referred to as the Galerkin projection of y. We can now formulate a semidiscrete version of (5.1) by replacing the continuous state equation (2.20) with the discrete equation (5.15). We formulate the semidiscrete problem as

$$\min_{u \in \mathcal{M}(\Omega_c, L^2(I))} j_{kh}(u) = J(S_{kh}(u)) + \alpha \|u\|_{\mathcal{M}(\Omega_c, L^2(I))}. \tag{5.18}$$

For the derivation of the optimality system, we define the discrete Lagrange function as

$$\mathcal{L}_{kh}(u, y, p) = J(y) - B(y, \varphi) + \langle \chi_{\Omega_c} u, p \rangle + (y_0, p_0^+)$$

for any $u \in \mathcal{M}(\Omega_c, L^2(I))$ and y and p as before. With the same methods as in the continuous case (cf. section 2.3.4) we can prove the following results.

Proposition 5.9. *The problem* (5.18) *possesses a globally optimal solution* $\tilde{u} \in \mathcal{M}(\Omega_c, L^2(I))$.

Proposition 5.10. *Let* $\tilde{u}$ *be an optimal solution of* (5.18) *and* $\bar{y}_{kh} = S_{kh}(\tilde{u})$ *the corresponding optimal state. There exists a unique discrete adjoint state* $\bar{p}_{kh} \in X_k^r(I, V_h)$ *solving the adjoint equation*

$$B(\varphi_{kh}, \bar{p}_{kh}) = (\chi_{\Omega_o}(\bar{y}_{kh} - y_d), \varphi_{kh}), \tag{5.19}$$

for all $\varphi_{kh} \in X_k^r(I, V_h)$ *and fulfilling the subgradient condition*

$$-\langle \chi_{\Omega_c}(u - \tilde{u}), \bar{p}_{kh} \rangle + \alpha \|\tilde{u}\|_{\mathcal{M}(\Omega_c, L^2(I))} \leq \alpha \|u\|_{\mathcal{M}(\Omega_c, L^2(I))} \tag{5.20}$$

for all $u \in \mathcal{M}(\Omega_c, L^2(I))$. *We alternatively express the first condition* (5.19) *with a solution operator by* $\bar{p}_{kh} = S_{\mathrm{dual},kh}(\chi_{\Omega_o}(\bar{y}_{kh} - y_d))$.

Since S_{kh} has a infinite-dimensional kernel, the solutions to (5.18) can not be expected to be unique. Therefore, as in the elliptic case, we now construct an appropriate subspace of $\mathcal{M}(\Omega_c, L^2(I))$ with the same approximation properties. By $\{x_n\}_n$ for $n = 1, 2 \dots, N_c$ we denote the nodes of the triangulation $\mathcal{T}_h$ contained in Ω_c and by $\{e_n\} \subset V_h$ the corresponding Lagrangian nodal basis functions. We introduce the space $\mathcal{M}_h$ consisting of linear combination of Dirac delta functional associated with the nodes x_n as in section 4.2. A suitable interpolation operator is now defined by duality as

$$\begin{aligned} \Lambda_{kh} \colon \mathcal{M}(\Omega_c, L^2(I)) &\to X_k^r(I, \mathcal{M}_h), \\ \langle \Lambda_{kh} u, \varphi \rangle &= \langle \chi_{\Omega_c} u,\, \pi_k i_h \varphi \rangle \quad \text{for all } \varphi \in \mathcal{C}_0(\Omega_c, L^2(I)) \end{aligned} \tag{5.21}$$

where $i_h \colon \mathcal{C}(\bar{\Omega}, L^2(I)) \to L^2(I, V_h)$ is the nodal interpolation operator and π_k is the L^2 projection on $X_k^r(I, L^2(\Omega)) \subset L^2(I \times \Omega)$. The interpolation operator i_h is given by

$$(i_h w)(x) = \sum_{n=1}^{N_c} w(x_n)\, e_n(x) \quad \text{for } x \in \Omega_c. \tag{5.22}$$

We can verify that for any $w \in \mathcal{C}_0(\Omega_c, L^2(I))$ the projection π_k has the pointwise formula

$$(\pi_k w)(x) = \sum_{m=1}^{M(1+r)} \frac{(\psi_m, w(x))_{L^2(I)}}{\|\psi_m\|^2_{L^2(I)}} \psi_m = \tilde{\pi}_k(w(x)) \quad \text{for } x \in \Omega_c, \tag{5.23}$$

where $\{\psi_m\}_m$ for $m = 1, \dots, M(1+r)$ is an orthogonal basis of $X_k^r(I, \mathbb{R})$ with respect to the inner product in $L^2(I)$ and $\tilde{\pi}_k$ is the L^2 projection in $L^2(I)$ onto $X_k^r(I, \mathbb{R})$. Therefore π_k and i_h commute and we have for $w \in \mathcal{C}_0(\Omega_c, L^2(I))$

$$i_h(\pi_k(w)) = \pi_k(i_h(w)) = \sum_{n=1}^{N_c} \sum_{m=1}^{M(1+r)} \frac{(\psi_m, w(x_n))_{L^2(I)}}{\|\psi_m\|^2_{L^2(I)}} \psi_m \, e_n,$$

which implies

$$\Lambda_{kh} u = \sum_{n=1}^{N_c} \sum_{m=1}^{M(1+r)} \frac{\langle u, \psi_m \, e_n \rangle}{\|\psi_m\|^2_{L^2(I)}} \psi_m \, \delta_{x_n}.$$

Remark 5.2. In the case $r = 0$ we take the piecewise constant functions $\psi_m = \chi_{I_m}$ as a suitable orthogonal basis for $X_k^r(I, \mathbb{R})$. In this case the operator Λ_{kh} can be written as

$$\Lambda_{kh} u = \sum_{n=1}^{N_c} \sum_{m=1}^{M} \frac{1}{k_m} \int_{I_m} \langle u(t), e_n \rangle \, \mathrm{d}t \, \chi_{I_m} \, \delta_n,$$

which is the same as given in [CCK13, Theorem 4.2].

Lemma 5.11. *For any $u \in \mathcal{M}(\Omega_c, L^2(I))$ we have*

$$\langle \Lambda_{kh} u, \varphi_{kh} \rangle = \langle u, \varphi_{kh} \rangle \quad \textit{for all } \varphi_{kh} \in X_k^r(I, V_h),$$
$$\textit{and} \quad \|\Lambda_{kh} u\|_{\mathcal{M}(\Omega_c, L^2(I))} \leq \|u\|_{\mathcal{M}(\Omega_c, L^2(I))}.$$

Proof. For the first property is immediately clear from the definition since $\chi_{\Omega_c}(\pi_k i_h \varphi_{kh}) = \chi_{\Omega_c}(i_h \varphi_{kh}) = \chi_{\Omega_c} \varphi_{kh}$ due to the assumptions on Ω_c. Furthermore we have

$$(\pi_k(i_h \varphi))(x) = \tilde{\pi}_k((i_h \varphi)(x)) \quad \text{for all } x \in \Omega_c$$

with (5.23) and since $\tilde{\pi}_k$ is an orthogonal projection we obtain with (5.22) that

$$\|\tilde{\pi}_k((i_h \varphi)(x))\|_{L^2(I)} \leq \|(i_h \varphi)(x)\|_{L^2(I)} \leq \|\varphi\|_{\mathcal{C}_0(\Omega_c, L^2(I))} \quad \text{for all } x \in \Omega_c.$$

With this the estimate $\|\pi_k(i_h \varphi)\|_{\mathcal{C}_0(\Omega_c, L^2(I))} \leq \|\varphi\|_{\mathcal{C}_0(\Omega_c, L^2(I))}$ is evident and by

$$\|\Lambda_{kh} u\|_{\mathcal{M}(\Omega_c, L^2(I))} = \sup_{\varphi \in \mathcal{C}_0(\Omega_c, L^2(I))} \frac{\langle \Lambda_{kh} u, \varphi \rangle}{\|\varphi\|_{\mathcal{C}_0(\Omega_c, L^2(I))}}$$

and the definition of Λ_{kh} as in (5.21) we obtain the second property. □

By familiar arguments (cf. section 4.8) it immediately follows that we can restrict the space for the optimal controls to $X_k^r(I, \mathcal{M}_h)$.

Proposition 5.12. *The semi-discrete solution operator* $S_{kh}\colon \mathcal{M}(\Omega_c, L^2(I)) \to X_k^r(I, V_h)$ *fulfills* $S_{kh} = S_{kh} \circ \Lambda_{kh}$ *and for each optimal solution* $\tilde{u} \in \mathcal{M}(\Omega_c, L^2(I))$ *of* (5.18) *the discrete control* $\bar{u}_{kh} = \Lambda_{kh}\tilde{u} \in X_k^r(I, \mathcal{M}_h)$ *fulfills*

$$j_{kh}(\tilde{u}) = j_{kh}(\bar{u}_{kh})$$

Thus, $\bar{u}_{kh} = \Lambda_{kh}\tilde{u}$ *is also an optimal solution of* (5.18).

Therefore, in the following, it suffices to consider the fully discrete problem

$$\min_{u_{kh} \in X_k^r(I,\mathcal{M}_h),\, y_{kh} \in X_k^r(I,V_h)} \quad \frac{1}{2}\|y_{kh} - y_d\|^2_{L^2(I\times\Omega_o)} + \alpha\|u_{kh}\|_{\mathcal{M}(\Omega_c,L^2(I))}$$
$$\text{subject to } \begin{cases} B(y_{kh}, \varphi_{kh}) = \langle \chi_{\Omega_c} u_{kh}, \varphi_{kh}\rangle + (y_0, \varphi_0^+) \\ \qquad\qquad\qquad\qquad \text{for all } \varphi_{kh} \in X_k^r(I, V_h). \end{cases} \tag{5.24}$$

which can be solved in practice. We point out that for any $u_{kh} = \sum_n u_n \delta_{x_n} \in X_k^r(I, \mathcal{M}_h)$ with $u_n \in X_k^r(I, \mathbb{R})$ the total variation norm is simply given by a weighted $\ell^1(\ell^2)$ norm of the underlying nodal vector:

$$\|u_{kh}\|_{\mathcal{M}(\Omega_c,L^2(I))} = \sum_{n=1}^{N_c} \|u_n\|_{L^2(I)} = \sum_{n=1}^{N_c} \left(\sum_{m=1}^{M(1+r)} \|\psi_m\|^2_{L^2(I)} u_{n,m}^2 \right)^{1/2}.$$

Furthermore, for any $u_{kh} = \sum_n u_n \delta_{x_n} \in \mathcal{M}_h$ and $v_{kh} = \sum_n v_n e_n \in V_h$ the duality product is given simply as the corresponding $L^2(I)$ inner product of the nodal vectors:

$$\langle \chi_{\Omega_c} u_h, v_h\rangle = \sum_{n=1}^{N_c} (u_n, v_n)_{L^2(I)} = \sum_{n=1}^{N_c} \sum_{m=1}^{M(1+r)} \|\psi_m\|^2_{L^2(I)} u_{n,m} v_{n,m}.$$

This means that a finite dimensional equivalent of (5.24) can be derived in a straightforward way, by introducing appropriate mass and stiffness matrices. Furthermore, the state equation (5.19) can be reformulated as a time-stepping scheme.

Note that for this problem the same optimality system holds as in Proposition 5.10, where we are allowed to insert any control from $\mathcal{M}(\Omega_c, L^2(I))$ in the subgradient condition (5.20), instead of only discrete controls. This is a direct consequence of Proposition 5.12 and will be important for the following error analysis.

5.3. Error estimates

For the error analysis we restrict attention to dG(0), which is a variant of the implicit Euler method. This restriction arises since we employ optimal estimates for the dG(r)cG(1) method in the $L^\infty(\Omega, L^2(I))$ norm, which are not considered in the standard finite element literature. These estimates were obtained recently for two dimensions by Leykekhman and Vexler [LV13] in the case $r = 0$. First, we will provide estimates for the state for a fixed control; then, we turn to estimates for the optimization problem.

5.3.1. Error analysis for the state

Define $i_k \colon \mathcal{C}(\bar{I}, V) \to X_k^0(I, V)$ to be the pointwise interpolation at the right time point in each interval

$$i_k w = \sum_{m=1}^{M} w(t_m)\chi_{I_m}, \tag{5.25}$$

where χ_{I_m} is the characteristic function of the interval I_m. We can obtain the following interpolation estimates for i_k.

Lemma 5.13. *For any $w = S_{\mathrm{dual}}(f)$ with $f \in L^2(I, L^2(\Omega))$ we have*

$$\|w - i_k w\|_{L^2(I,L^2(\Omega))} \leq C\, k \|f\|_{L^2(I,L^2(\Omega))}, \tag{5.26}$$

$$\|w - i_k w\|_{L^2(I,H_0^1(\Omega))} \leq C\, k^{1/2} \|f\|_{L^2(I,L^2(\Omega))}. \tag{5.27}$$

Proof. First, we note that $i_k w$ in $L^2(I, H_0^1(\Omega))$ since $w \in \mathcal{C}(\bar{I}, H_0^1(\Omega))$ with Lemma 5.4. The interpolation estimates can be obtained with standard techniques. Since (5.27) is not standard, we will give a proof in Appendix A.3. □

In the following estimates we are going to apply the best approximation properties obtained in [LV13].

Theorem 5.14 ([LV13, Theorem 3.1, Theorem 3.5])**.** *Let $w = S_{\mathrm{dual}}(f)$ be an adjoint solution and $w_{kh} = S_{\mathrm{dual},kh}(f)$ its Galerkin projection for some $f \in L^2(I, L^2(\Omega))$ and $1 \leq q \leq \infty$. Then we have for every $x \in \Omega$ that*

$$\begin{aligned}\|w(x) - w_{kh}(x)\|_{L^2(I)}^2 \\ \leq C|\ln h|^2 \inf_{\chi \in X_k^0(I,V_h)} \int_I \|w(t) - \chi(t)\|_{L^\infty(\Omega)}^2 + h^{-4/q} \|i_k w(t) - \chi(t)\|_{L^q(\Omega)}^2 \,\mathrm{d}t.\end{aligned}$$

Furthermore, for $x \in \Omega_\eta$ with $\eta > 4h > 0$ we have the local estimate

$$\begin{aligned}\|w(x) - w_{kh}(x)\|_{L^2(I)}^2 \\ \leq C|\ln h|^3 \inf_{\chi \in X_k^0(I,V_h)} \int_I \|w(t) - \chi(t)\|_{L^\infty(B_\eta(x))}^2 + h^{-4/q} \|i_k w(t) - \chi(t)\|_{L^q(B_\eta(x))}^2 \,\mathrm{d}t \\ + C\,\eta^{-2} |\ln h| \int_I \|w(t) - w_{kh}(t)\|_{L^2(\Omega)}^2 \,\mathrm{d}t.\end{aligned}$$

With this we can prove the following a priori error estimates.

Theorem 5.15. *Let $y = S(y_0, u)$ and its Galerkin projection $y_{kh} = S_{kh}(y_0, u)$ for arbitrary $u \in \mathcal{M}(\Omega_c, L^2(I))$ and $y_0 \in H_0^1(\Omega)$. Then we have the a priori estimate*

$$\|y - y_{kh}\|_{L^2(I\times\Omega)} \leq C|\ln h|^2 (k^{1/2} + h) \left(\|u\|_{\mathcal{M}(\Omega_c,L^2(I))} + \|y_0\|_{H^1(\Omega)} \right). \tag{5.28}$$

If additionally the measure is supported in the interior of the domain, i.e., $\operatorname{supp} u \subset \Omega^\eta$ for some $\eta > 0$, we obtain the improved estimate in a weaker norm

$$\|y - y_{kh}\|_{L^2(I,L^1(\Omega))} \leq C\,\eta^{-1} |\ln h|^{5/2} (k + h^2) \left(\|u\|_{\mathcal{M}(\Omega_c,L^2(I))} + \|y_0\|_{H^1(\Omega)} \right). \tag{5.29}$$

Proof. Consider that $y = S(y_0, 0) + S(0, u)$ and $y_{kh} = S_{kh}(y_0, 0) + S_{kh}(0, u)$. We have $S(y_0, 0) \in L^2(I, H^2(\Omega)) \cap H^1(I, L^2(\Omega))$ and the corresponding error estimate

$$\|S(y_0, 0) - S_{kh}(y_0, 0)\|_{L^2(I,L^2(\Omega))} \leq c\,(k + h^2)\|y_0\|_{H^1(\Omega)}$$

can be found, e.g., in [MV08]. Without restriction, we suppose $y_0 = 0$ in the following and employ a duality argument. Define the error $e = y - y_{kh}$ and introduce

$$\begin{aligned} g_2 &= e \in L^2(I, L^2(\Omega)), \\ g_1 &= \|e(t)\|_{L^1(\Omega)} \operatorname{sgn} e(t, x) \in L^2(I, L^\infty(\Omega)), \end{aligned}$$

for the first and second estimate respectively. For $l \in \{1, 2\}$ we define the auxiliary dual variable $w = S_{\mathrm{dual}}(g_l)$ and its Galerkin projection $w_{kh} = S_{\mathrm{dual},kh}(g_l)$. We can verify that $B(\varphi, w) = (\varphi, g_l)_I$ holds for any $\varphi \in L^2(I, W^{1,s}(\Omega))$ with $\varphi_m^- \in H^{-1}(\Omega)$ for $m = 1, \dots, M$, since the jump terms in the dual description of the bilinear form (5.13) vanish due to Lemma 5.4. We rewrite the error using this identity for w, Galerkin orthogonality for y (see (5.17)), Galerkin orthogonality for w and (5.16) to obtain

$$\begin{aligned} \|y - y_{kh}\|^2_{L^2(I,L^l(\Omega))} &= (y - y_{kh}, g_l)_I = B(y - y_{kh}, w) \\ &= B(y - y_{kh}, w - w_{kh}) = B(y, w - w_{kh}) \\ &= \langle u, \chi_{\Omega_c}(w - w_{kh})\rangle \leq \|u\|_{\mathcal{M}(\Omega_c,L^2(I))}\|w - w_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))}. \end{aligned} \tag{5.30}$$

In the following, we estimate the last term.

For the first estimate, where $l = 2$, we apply the global best approximation property from Theorem 5.14 with the choice $\chi = \pi_h i_k w$, where i_k is the pointwise interpolation defined in (5.25) and $\pi_h \colon L^1(\Omega) \to V_h$ is the Clément interpolation; see, e.g., [BG98]. This results in

$$\begin{aligned} &\|w - w_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} \\ &\qquad \leq C|\ln h| \left(\|w - \pi_h i_k w\|_{L^2(I,L^\infty(\Omega))} + h^{-2/q}\|i_k(w - \pi_h w)\|_{L^2(I,L^q(\Omega))}\right), \end{aligned} \tag{5.31}$$

where we choose any $q < \infty$. The first term is further estimated by

$$\begin{aligned} \|w - \pi_h i_k w\|_{L^2(I,L^\infty(\Omega))} &\leq \|w - \pi_h w\|_{L^2(I,L^\infty(\Omega))} + \|\pi_h(w - i_k w)\|_{L^2(I,L^\infty(\Omega))} \\ &\leq C\,h\|w\|_{L^2(I,H^2(\Omega))} + C\,h^{-2/q}\|\pi_h(w - i_k w)\|_{L^2(I,L^q(\Omega))} \end{aligned}$$

with an interpolation estimate for the Clément interpolation and an inverse estimate with the same q as above. With the stability of the Clément interpolation in $L^q(\Omega)$ and the Sobolev embedding we obtain

$$\|\pi_h(w - i_k w)\|_{L^2(I,L^q(\Omega))} \leq C\,\|w - i_k w\|_{L^2(I,L^q(\Omega))} \leq C\,q\,\|w - i_k w\|_{L^2(I,H^1_0(\Omega))},$$

see, e.g., [Alt11, Theorem 8.8] for the dependence of the embedding constant on $q < \infty$. With Lemma 5.13 we then get the estimate

$$\|\pi_h(w - i_k w)\|_{L^2(I,L^q(\Omega))} \leq C\,q\,k^{1/2}\|g_2\|_{L^2(I,L^2(\Omega))}.$$

The second term in (5.31) is estimated by the triangle inequality

$$\begin{aligned} &\|i_k(w - \pi_h w)\|_{L^2(I,L^q(\Omega))} \\ &\qquad \leq \|i_k w - w\|_{L^2(I,L^q(\Omega))} + \|w - \pi_h w\|_{L^2(I,L^q(\Omega))} + \|\pi_h(w - i_k w)\|_{L^2(I,L^q(\Omega))}, \end{aligned}$$

The single terms are treated as before and we arrive at

$$\begin{aligned}&\|w - w_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))}\\ &\qquad\leq c|\ln h|\left(h\|w\|_{L^2(I,H^2(\Omega))} + h^{-2/q}\left(q\,k^{1/2}\|g_2\|_{L^2(I,L^2(\Omega))} + h^{1+2/q}\|w\|_{L^2(I,H^2(\Omega))}\right)\right).\end{aligned}$$

Finally, with the choice $q = |\ln h|$ and Lemma 5.4 this implies

$$\begin{aligned}\|w - w_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} &\leq c|\ln h|\left(h + q\,h^{-2/q}k^{1/2}\right)\|g_2\|_{L^2(I,L^2(\Omega))}\\ &\leq c|\ln h|^2\left(h + k^{1/2}\right)\|y - y_{kh}\|_{L^2(I,L^2(\Omega))}.\end{aligned} \tag{5.32}$$

Combining (5.30) and (5.32) we obtain the result (5.28).

The second estimate, where $l = 1$, can be obtained in a similar fashion using the local estimate from Theorem 5.14 and choosing again $\chi = \pi_h i_k w$. Then we can use the approximation properties of the Clément interpolation, Lemma 5.13 and the regularity estimate from Lemma 5.7 for the first two terms and an L^2 estimate from [MV08] for the term $\|w - w_{kh}\|_{L^2(I\times\Omega)}$. We obtain

$$\begin{aligned}\|w - w_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} &\leq c\,\eta^{-1}|\ln h|^{1/2}\left(1 + qh^{-2/q}\right)\left(k + h^2\right)\|g_1\|_{L^2(I,L^q(\Omega))}\\ &\leq c\,\eta^{-1}|\ln h|^{3/2}\left(k + h^2\right)\|y - y_{kh}\|_{L^2(I,L^1(\Omega))}\end{aligned} \tag{5.33}$$

with $q = |\ln h|$ as above and we obtain (5.29). We omit a more detailed argument since it is analogous to the one in [LV13, Theorem 4.1], where an estimate for the special case $u(t) = \hat{u}(t)\delta_{x_0}$ for some $x_0 \in \Omega$ and $\hat{u} \in L^2(I)$ is proved. □

Remark 5.3. It is possible to derive a sharpened version of (5.28) without any $|\ln h|$ term if we require a coupling of k and h of the form

$$k = ch^2$$

for a constant independent of k and h, see [CCK13, Theorem 4.6]. Whether we can improve (5.28) without such a coupling is an open question to the best of the authors knowledge. However, such an improvement alone would yield no improvement for the estimates in section 5.3.2.

For the error analysis in the following section we need an additional stability property of the space-time discretization.

Lemma 5.16. *We have for every $y_0 \in L^2(\Omega)$ and $u \in \mathcal{M}(\Omega_c, L^2(I))$ that*

$$\|y_{kh}\|_{L^2(I,L^\infty(\Omega))} \leq C|\ln h|\left(\|y_0\|_{L^2(\Omega)} + \|u\|_{\mathcal{M}(\Omega_c,L^2(I))}\right).$$

Proof. We start by applying the discrete Sobolev inequality (see [BS08, Lemma 4.9.1])

$$\|y_{kh}\|^2_{L^2(I,L^\infty(\Omega))} = \int_I \|y_{kh}(t)\|^2_{L^\infty(\Omega)}\,\mathrm{d}t \leq C|\ln h|\|\nabla y_{kh}\|^2_{L^2(I\times\Omega)}. \tag{5.34}$$

Now, we can add the primal and dual representation of the bilinear form as in (5.13) and (5.14) with $y = \varphi = y_{kh}$ and divide by two, which yields

$$B(y_{kh}, y_{kh}) = (\nabla y_{kh}, \nabla y_{kh})_I + \frac{1}{2}\sum_{m=1}^{M}\|[y_{kh}]_m\|^2_{L^2(\Omega)} + \frac{1}{2}\|y^+_{kh,0}\|^2_{L^2(\Omega)} + \frac{1}{2}\|y^-_{kh,M}\|^2_{L^2(\Omega)}.$$

This allows us to estimate the $L^2(I, H_0^1(\Omega))$ seminorm in terms of the bilinear form and with the definition of the discrete state equation (5.15) we obtain

$$\begin{aligned}\|\nabla y_{kh}\|^2_{L^2(I\times\Omega)} &\le B(y_{kh}, y_{kh}) - \frac{1}{2}\|y^+_{kh,0}\|^2_{L^2(\Omega)} \\ &= \langle u, \chi_{\Omega_c} y_{kh}\rangle + (y_0, y^+_{kh,0}) - \frac{1}{2}\|y^+_{kh,0}\|^2_{L^2(\Omega)} \\ &= \langle u, \chi_{\Omega_c} y_{kh}\rangle - \frac{1}{2}\|y^+_{kh,0} - y_0\|^2_{L^2(\Omega)} + \frac{1}{2}\|y_0\|^2_{L^2(\Omega)} \\ &\le \|u\|_{\mathcal{M}(\Omega_c,L^2(I))}\|y_{kh}\|_{L^2(I,L^\infty(\Omega))} + \frac{1}{2}\|y_0\|^2_{L^2(\Omega)}.\end{aligned} \tag{5.35}$$

Finally, we apply (5.34) and use Young's inequality to derive

$$\begin{aligned}\|\nabla y_{kh}\|^2_{L^2(I\times\Omega)} &\le C\left(|\ln h|^{1/2}\|u\|_{\mathcal{M}(\Omega_c,L^2(I))}\|\nabla y_{kh}\|_{L^2(I\times\Omega)} + \frac{1}{2}\|y_0\|^2_{L^2(\Omega)}\right) \\ &\le \frac{1}{2}\|\nabla y_{kh}\|^2_{L^2(I\times\Omega)} + C|\ln h|\left(\|u\|^2_{\mathcal{M}(\Omega_c,L^2(I))} + \|y_0\|^2_{L^2(\Omega)}\right).\end{aligned}$$

We take the one term to the left-hand side take the square root, and combine the estimate again with (5.34) on the left-hand side to finish the proof. □

5.3.2. Error analysis for the optimal control problem

First we will consider convergence of the functional values.

Lemma 5.17. *For every optimal control $\bar u$ or $\bar u_{kh}$ we have*

$$\max\left\{\|\bar u\|_{\mathcal{M}(\Omega_c,L^2(I))}, \|\bar u_{kh}\|_{\mathcal{M}(\Omega_c,L^2(I))}\right\} \le C\left(\|y_0\|_{L^2(\Omega)} + \|y_d\|_{L^2(I\times\Omega_o)}\right). \tag{5.36}$$

Proof. For $\bar u_{kh}$ this is a consequence of the minimality, since

$$\|\bar u_{kh}\|_{\mathcal{M}(\Omega_c,L^2(I))} \le \frac{1}{\alpha}J(S_{kh}(y_0,0)) \le \frac{1}{2\alpha}\left(\|S_{kh}(y_0,0)\|_{L^2(I\times\Omega_o)} + \|y_d\|_{L^2(I\times\Omega_o)}\right).$$

The result follows by the stability estimate $\|S_{kh}(y_0,0)\|_{L^2(I\times\Omega)} \le C\|y_0\|_{L^2(\Omega)}$ for the dG(r)cG(1) method; see (5.35) for $u = 0$. The proof for $\bar u$ is similar. □

Theorem 5.18. *Let $\bar u \in \mathcal{M}(\Omega_c, L^2(I))$ be an optimal solution to* (5.1) *and $\bar u_{kh} \in X_k^0(I, \mathcal{M}_h)$ be a discrete optimal solution to* (5.24)*. We have for the associated optimal functional values*

$$|j(\bar u) - j_{kh}(\bar u_{kh})| \le C\,\eta^{-1}|\ln h|^4\left(k + h^2\right), \tag{5.37}$$

with a constant C independent of k and h, where η is the constant from Proposition 5.6.

Proof. Since we have

$$j(\bar u) - j_{kh}(\bar u) \le j(\bar u) - j_{kh}(\bar u_{kh}) \le j(\bar u_{kh}) - j_{kh}(\bar u_{kh})$$

by minimality of $\bar u$ and $\bar u_{kh}$, and Proposition 5.12 we obtain

$$|j(\bar u) - j_{kh}(\bar u_{kh})| \le \max\{|j(\bar u) - j_{kh}(\bar u)|,\, |j(\bar u_{kh}) - j_{kh}(\bar u_{kh})|\}.$$

Therefore we estimate the functional error $j(u) - j_{kh}(u) = J(S(u)) - J(S_{kh}(u))$ for a fixed $u \in \mathcal{M}(\Omega_c, L^2(I))$. We define $y = S(u)$ and $y_{kh} = S_{kh}(u)$ and by reordering terms and applying Hölders inequality we get

$$\begin{aligned} |j(u) - j_{kh}(u)| &= \frac{1}{2}|(\chi_{\Omega_o}(y - y_{kh}), y - y_{kh} + 2y_{kh} - 2y_d)_I| \\ &\leq \frac{1}{2}\|y - y_{kh}\|^2_{L^2(I\times\Omega)} + \|y - y_{kh}\|_{L^2(I,L^1(\Omega))}\|y_{kh} - y_d\|_{L^2(I,L^\infty(\Omega))}. \end{aligned} \tag{5.38}$$

The terms which contain $y - y_{kh}$ are treated with estimates (5.28) and (5.29) from Theorem 5.15 respectively. Furthermore we have

$$\|y_{kh} - y_d\|_{L^2(I,L^\infty(\Omega))} \leq \|y_d\|_{L^2(I,L^\infty(\Omega))} + C|\ln h|\left(\|y_0\|_{L^2(\Omega)} + \|u\|_{\mathcal{M}(\Omega_c,L^2(I))}\right)$$

with Lemma 5.16 and (5.9). Together with Lemma 5.17 we have shown (5.37). □

We also provide an error estimate for the optimal state solutions on the observation domain.

Theorem 5.19. *Let $\bar{u} \in \mathcal{M}(\Omega_c, L^2(I))$ be an optimal solution to (5.1) with associate state $\bar{y} = S(y_0, \bar{u})$ and $\bar{u}_{kh} \in X_k^0(I, \mathcal{M}_h)$ be a discrete optimal solution to (5.24) with $\bar{y}_{kh} = S_{kh}(y_0, \bar{u}_{kh})$. With assumption (5.9) we have the estimate*

$$\|\bar{y} - \bar{y}_{kh}\|_{L^2(I\times\Omega_o)} \leq C\,\eta^{-1/2}|\ln h|^2\left(k^{1/2} + h\right)$$

where $\eta > 0$ is the constant from Proposition 5.6.

Proof. We test the continuous subgradient condition (5.5) with the discrete solution, and the discrete one (5.20) with the continuous solution (which is possible due to Proposition 5.12) to obtain

$$\begin{aligned} -\langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p}\rangle + \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))} &\leq \alpha\|\bar{u}_{kh}\|_{\mathcal{M}(\Omega_c,L^2(I))}, \\ -\langle\chi_{\Omega_c}(\bar{u} - \bar{u}_{kh}), \bar{p}_{kh}\rangle + \alpha\|\bar{u}_{kh}\|_{\mathcal{M}(\Omega_c,L^2(I))} &\leq \alpha\|\bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))}. \end{aligned}$$

Adding both implies

$$-\langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p} - \bar{p}_{kh}\rangle \leq 0.$$

We introduce as auxiliary variables the Galerkin projections of $\bar{y}$ and $\bar{p}$ as $\hat{y}_{kh} = S_{kh}(\bar{u})$ and $\hat{p}_{kh} = S_{\mathrm{dual},kh}(\chi_{\Omega_o}(\bar{y} - y_d))$. With this, we can reformulate the inequality above to

$$\begin{aligned} 0 &\leq \langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p} - \bar{p}_{kh}\rangle \\ &= \langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p} - \hat{p}_{kh}\rangle + \langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \hat{p}_{kh} - \bar{p}_{kh}\rangle \\ &= \langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p} - \hat{p}_{kh}\rangle + (\bar{y}_{kh} - \hat{y}_{kh}, \chi_{\Omega_o}(\bar{y} - \bar{y}_{kh})) \\ &= \langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p} - \hat{p}_{kh}\rangle + (\bar{y} - \hat{y}_{kh}, \chi_{\Omega_o}(\bar{y} - \bar{y}_{kh})) - \|\bar{y} - \bar{y}_{kh}\|^2_{L^2(I\times\Omega_o)} \end{aligned}$$

We bring the last term above on the other side and treat the second with Young's inequality to obtain

$$\begin{aligned} \frac{1}{2}\|\bar{y} - \bar{y}_{kh}\|^2_{L^2(I\times\Omega_o)} &\leq \langle\chi_{\Omega_c}(\bar{u}_{kh} - \bar{u}), \bar{p} - \hat{p}_{kh}\rangle + \frac{1}{2}\|\bar{y} - \hat{y}_{kh}\|^2_{L^2(I\times\Omega_o)} \\ &\leq \|\bar{u}_{kh} - \bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))}\|\bar{p} - \hat{p}_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} + \frac{1}{2}\|\bar{y} - \hat{y}_{kh}\|^2_{L^2(I\times\Omega_o)}. \end{aligned}$$

Since $\|\bar{u}_{kh} - \bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))}$ can be bounded independently of k and h with Lemma 5.17 and the triangle inequality we obtain an estimate of the optimal state in terms of two Galerkin projection errors

$$\|\bar{y} - \bar{y}_{kh}\|^2_{L^2(I\times\Omega_o)} \leq C \left(\|\bar{p} - \hat{p}_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} + \|\bar{y} - \hat{y}_{kh}\|^2_{L^2(I\times\Omega_o)} \right).$$

For the second term on the right hand side we apply Theorem 5.15 to obtain

$$\|\bar{y} - \hat{y}_{kh}\|^2_{L^2(I\times\Omega_o)} \leq C\,|\ln h|^4 \left(k + h^2\right) \left(\|\bar{u}\|^2_{\mathcal{M}(\Omega_c,L^2(I))} + \|y_0\|^2_{H^1(\Omega)} \right).$$

For the first term we argue as in Theorem 5.15 for estimate (5.33) to obtain

$$\|\bar{p} - \hat{p}_{kh}\|_{\mathcal{C}_0(\Omega_c,L^2(I))} \leq C\,\eta^{-1}|\ln h|^{1/2} \left(1 + q\,h^{2/q}\right) \left(k + h^2\right) \|\bar{y} - y_d\|_{L^2(I,L^q(\Omega_o))}.$$

Then we use the regularity assumption on the desired state (5.9) and estimate (5.2) from Proposition 5.1 for

$$\|\bar{y} - y_d\|_{L^2(I,L^q(\Omega_o))} \leq \|y_d\|_{L^2(I,L^\infty(\Omega_o))} + C\,q \left(\|\bar{u}\|_{\mathcal{M}(\Omega_c,L^2(I))} + \|y_0\|_{L^2(\Omega)} \right).$$

Setting $q = |\ln h|$ and combining the above estimates, we complete the proof. □

5.4. Regularized problem

For the numerical realization, as discussed in in section 2.5, we consider a regularized version of (5.1). Since we have restricted attention to the case of a two dimensional Ω at the start of this chapter, we will only discuss this case. However, all of the following results can be generalized to the three dimensional case in a straightforward way. The regularized problem is given as

$$\begin{aligned} \min_{u\in L^2(I\times\Omega),\, y\in Y^2} \quad & \frac{1}{2}\|y - y_d\|^2_{L^2(I\times\Omega_o)} + \alpha\|u\|_{L^1(\Omega_c,L^2(I))} + \frac{\gamma}{2}\|u\|^2_{L^2(I\times\Omega_c)} \\ \text{subject to } & \begin{cases} (\partial_t y, \varphi) + (\nabla y, \nabla\varphi) = (\chi_{\Omega_c} u, \varphi) & \text{for all } \varphi \in L^2(I, H^1_0(\Omega)) \\ y(0) = y_0 \end{cases} \end{aligned} \tag{5.39}$$

As in the elliptic case, for the case of simplicity, we exclude Ω_c with complicated topology and only consider Ω_c which are the relative closure of an open set. In this case $L^q(\Omega_c)$ for $q \in \{1, 2\}$ is to be understood with respect to the Lebesgue measure. It is clear that the canonical embedding $L^1(\Omega_c, L^2(I)) \hookrightarrow \mathcal{M}(\Omega_c, L^2(I))$ is isometric and therefore

$$\|u\|_{\mathcal{M}(\Omega_c,L^2(I))} = \|u\|_{L^1(\Omega_c,L^2(I))} = \int_{\Omega_c} \|u(x)\|_{L^2(I)}\,\mathrm{d}x$$

for $u \in L^1(\Omega_c, L^2(I))$. We abbreviate the inner product in $L^2(I \times \Omega_c)$ by $(\cdot,\cdot)$. For an independent analysis of the problem (5.39) we refer to Herzog, Stadler, and Wachsmuth [HSW12].

As discussed in section 3.3.2, the proximal map corresponding to $\psi(\cdot) = \|\cdot\|_{L^1(\Omega_c,L^2(I))}$ for the parameter $\gamma > 0$ is given by

$$P_\gamma(q)(x) = \frac{1}{\gamma}\left(\gamma - \alpha/\|q(x)\|_{L^2(I)}\right)^+ q(x) \quad \text{for } q \in L^2(\Omega_c, L^2(I)).$$

As a consequence of Proposition 2.25 and this formula, we obtain the following result; cf. also [HSW12].

Proposition 5.20. *Let $\gamma > 0$. Problem (5.39) possesses a unique optimal solution $\bar{u}_\gamma \in L^2(I \times \Omega_c)$ with corresponding state $\bar{y}_\gamma = S(y_0, \bar{u}_\gamma)$ and adjoint state $\bar{p}_\gamma = S_{\mathrm{dual}}(\chi_{\Omega_o}(\bar{y}_\gamma - y_d))$. The optimality is characterized by the subgradient condition*

$$-(\chi_{\Omega_c}(u - \bar{u}_\gamma),\, \gamma\bar{u}_\gamma + \bar{p}_\gamma) + \alpha\|\bar{u}_\gamma\|_{L^1(\Omega_c,L^2(I))} \le \alpha\|u\|_{L^1(\Omega_c,L^2(I))} \tag{5.40}$$

for all $u \in L^1(\Omega_c, L^2(I))$, which is equivalent to the "stripe-wise" projection formula

$$\bar{u}_\gamma(t,x) = -\frac{1}{\gamma}\left(1 - \alpha/\|\bar{p}_\gamma(x)\|_{L^2(I)}\right)^+ \bar{p}_\gamma(t,x) \tag{5.41}$$

for almost all $(t,x) \in I \times \Omega_c$. This implies that $\operatorname{supp}|\bar{u}_\gamma|$ is contained in the closure of $\{\, x \in \Omega_c \mid \|\bar{p}_\gamma(x)\|_{L^2(I)} > \alpha \,\}$.

The regularized problem (5.39) can be solved efficiently with a semismooth Newton method which admits a Banach space analysis; see [HSW12, Theorem 3.7, Example 1.2]. As before, the gradient and Hessian of the smooth part of the reduced cost functional $f(u) = J(S(u))$ are given by

$$\begin{aligned} \nabla f(u) &= \chi_{\Omega_c} S_{\mathrm{dual}}(\chi_{\Omega_o}(S(u,y_0) - y_d)) && \text{for } u \in L^2(I \times \Omega_c) \\ \text{and}\quad \nabla^2 f(u)\delta u &= \chi_{\Omega_c} S_{\mathrm{dual}}(\chi_{\Omega_o}(S(\delta u, 0)) && \text{for } u \text{ and } \delta u \in L^2(I \times \Omega_c). \end{aligned}$$

Since the dual solution operator S_{dual} maps $L^2(I \times \Omega)$ continuously into $L^2(I, H^2(\Omega))$, which is embedded into $\mathcal{C}^\delta(\Omega_c, L^2(I))$ for all $0 < \delta < 1$ (cf. Proposition 5.5), we have a more than sufficient norm-gap. In terms of the general framework given in chapter 3, we obtain the following result.

Proposition 5.21. *Let $\bar{q}_\gamma = -1/\gamma\, \chi_{\Omega_c}\bar{p}_\gamma$ be the optimal auxiliary variable with $\bar{u}_\gamma = P_\gamma(\bar{q}_\gamma)$. Suppose that for a given $q_0 \in L^r(\Omega_c, L^2(I))$ with $r > 2$, the distance $\|q_0 - \bar{q}_\gamma\|_{L^r(\Omega_c,L^2(I))}$ is sufficiently small. Then the semismooth Newton iterates, defined inductively as $q_{k+1} = q_k - \mathrm{D}G(q_k)^{-1}G(q_k)$ for $k \in \mathbb{N}$ converge superlinearly in $L^r(\Omega_c, L^2(I))$ towards $\bar{q}_\gamma$. The same holds for $u_k = P_\gamma(q_k)$ with limit $\bar{u}_\gamma$.*

Proof. We combine Lemma 3.14, Proposition 3.11, and Theorem 3.7 with the choice $H = L^2(\Omega_c, L^2(I))$ and $H_{\mathrm{sub}} = L^r(\Omega_c, L^2(I))$ as in Lemma 3.22; cf. also Proposition 4.27. □

Moreover, we obtain the original problem (5.1) in the limiting case for $\gamma \to 0$; see Theorem 2.28.

Theorem 5.22. *For $\gamma \to 0$ we have $j(\bar{u}) \le j_\gamma(u_\gamma) \to j(\bar{u})$, where $\bar{u}$ is an (arbitrary) optimal solution of (5.1). Moreover, the sequence of solutions of (5.39) contains an accumulation point in the sense of weak-$*$ convergence and any such accumulation point is an optimal solution of (5.1).*

As before, we use the following procedure to compute $\bar{u}$ in practice. In an inner loop, we use the semismooth Newton method to compute the minimizer $\bar{u}_\gamma$ for a small value of γ. Then we decrease γ and use the previous solution as an initial guess for the new iteration. In the numerical experiments for this problem, the Newton method exhibited convergence in each iteration and a globalization strategy was not needed.

5.4.1. Regularization error

We can obtain the similar estimate for the regularization error as in section 4.5.1 for the elliptic problem. As before, we need some estimates for the optimal solutions which are independent of the regularization parameter. We employ the same notation for $\mathcal{C}^\beta$ for $\beta \in (0, 2]$ as in section 4.5.1.

Proposition 5.23. *Let $\eta > 0$ be the constant from Proposition 5.6. For any $s < 2$, $q < \infty$ and $\beta < 2$, there exists a constant $C > 0$, such that for all $\gamma > 0$ the following estimates are valid for the optimal triple $(\bar{u}_\gamma, \bar{u}_\gamma, \bar{p}_\gamma)$ of (5.39):*

$$\|\bar{u}_\gamma\|_{L^1(\Omega_c,L^2(I))} + \frac{\gamma}{2}\|\bar{u}_\gamma\|^2_{L^2(I\times\Omega_c)} \le C, \tag{5.42}$$

$$\|\bar{y}_\gamma\|_{L^2(I,L^q(\Omega))} + \|\bar{y}_\gamma\|_{L^2(I,W^{1,s}(\Omega))} \le C, \tag{5.43}$$

$$\|\bar{p}_\gamma\|_{L^2(I,\mathcal{C}^\beta(\Omega^\eta))} + \|\bar{p}_\gamma\|_{L^2(I,W^{2,q}(\Omega^\eta))} \le C. \tag{5.44}$$

Proof. The estimate (5.42) follows by straightforward arguments using the minimality of $\bar{u}_\gamma$; cf. Theorem 2.28. For the state, we apply now Proposition 5.1. For the adjoint solution, we then apply Lemma 5.7. The estimate for $\bar{p}_\gamma$ in the Hölder norm is again a consequence of the Sobolev embedding with $\beta = 2 - 2/q = 3 - 2/s$. □

Thereby, using the technique from Hintermüller, Schiela, and Wollner [HSW14], we can obtain an asymptotic estimate for the regularization error.

Proposition 5.24. *The error in the objective functional due to regularization is bounded by*

$$0 \le j_\gamma(\bar{u}_\gamma) - j(\bar{u}) \le C\,\gamma^s, \quad \text{where } s = 1/3.$$

Proof. We can adapt the proof of Proposition 4.26 with some modifications. We start again with the estimate

$$\begin{aligned}\|\bar{u}_\gamma\|^2_{L^2(I\times\Omega_c)} \le \|\bar{u}_\gamma\|_{L^1(\Omega_c,L^2(I))}\|\bar{u}_\gamma\|_{L^\infty(\Omega_c,L^2(I))} &\le C\Big\|\|\bar{u}_\gamma(\cdot)\|_{L^2(I)}\Big\|_{L^\infty(\Omega_c)} \\ &= \frac{C}{\gamma}\Big\|(\alpha - \|\bar{p}_\gamma(\cdot)\|_{L^2(I)})^+\Big\|_{L^\infty(\Omega_c)}\end{aligned}$$

for any $\gamma > 0$, using (5.42) and the optimality conditions. With estimate (5.44) and Proposition 5.5, we obtain that $\bar{p}_\gamma \in \mathcal{C}^1(\Omega^\eta, L^2(I))$ (recall that $\mathcal{C}^1$ denotes Lipschitz continuity). We define the positive function

$$v_\gamma(x) = (\alpha - \|\bar{p}_\gamma(x)\|_{L^2(I)})^+ \quad \text{for } x \in \Omega_c.$$

With similar arguments as in Proposition 5.6 we obtain now that the support of v is contained in the interior of the domain; i.e., we have $\operatorname{supp} v_\gamma \subset \{\, x \in \Omega_c \mid \|\bar{p}_\gamma(x)\| \ge \alpha \,\} \subset \Omega^\eta$. With the regularity of $\bar{p}_\gamma$ in the interior of the domain, this implies that v_γ is Lipschitz continuous with

$$\|v_\gamma\|_{\mathcal{C}^1(\Omega_c)} \le \|\bar{p}_\gamma\|_{\mathcal{C}^1(\Omega^\eta,L^2(I))} \le \|\bar{p}_\gamma\|_{L^2(I,\mathcal{C}^1(\Omega^\eta))} \le C.$$

With Proposition 4.24 we obtain now for $\theta = 2/3$ that

$$\begin{aligned}\|\bar{u}_\gamma\|^2_{L^2(I\times\Omega_c)} &\le \frac{C}{\gamma}\|v_\gamma\|_{L^\infty(\Omega_c)} \le \frac{C}{\gamma}\|v_\gamma\|^{1-\theta}_{\mathcal{C}^1(\Omega_c)}\|v_\gamma\|^\theta_{L^1(\Omega_c)} \\ &\le C\gamma^{\theta-1}\Big\|1/\gamma\,(\alpha - \|\bar{p}_\gamma(x)\|_{L^2(I)})^+\Big\|^\theta_{L^1(\Omega_c)} = C\gamma^{\theta-1}\|\bar{u}_\gamma\|^\theta_{L^1(\Omega_c,L^2(I))} \le C\gamma^{\theta-1}\end{aligned}$$

by the definition of v_γ, the optimality condition and (5.42). Combining this with Corollary 2.29 yields

$$0 \leq j_\gamma(\bar{u}_\gamma) - j(\bar{u}) = \int_0^\gamma \frac{1}{2}\|\bar{u}_\sigma\|^2_{L^2(I\times\Omega_c)}\,\mathrm{d}\sigma \leq C\gamma^\theta = C\gamma^{1/3},$$

as claimed above. □

5.5. Numerical results

In this section we construct a numerical example which is geared towards verification of the convergence results in section 5.3.2. A practically motivated example will be given in section 5.6. We design an example with an explicit solution on the interval $I = (0, T)$ and the two dimensional domain $\Omega = \Omega_c = \Omega_o = (-1, 1) \times (-1, 1)$. For the construction of the example the optimal control is chosen as

$$\bar{u}(t) = T^{-2}\,(T - t)\,\delta_0,$$

with a Dirac delta function in the origin. We can give the analytical solution $\bar{y}$ of $\partial_t y - \Delta y = \bar{u}$ with zero Dirichlet boundary conditions; see Figure 5.1. It can be represented by the series

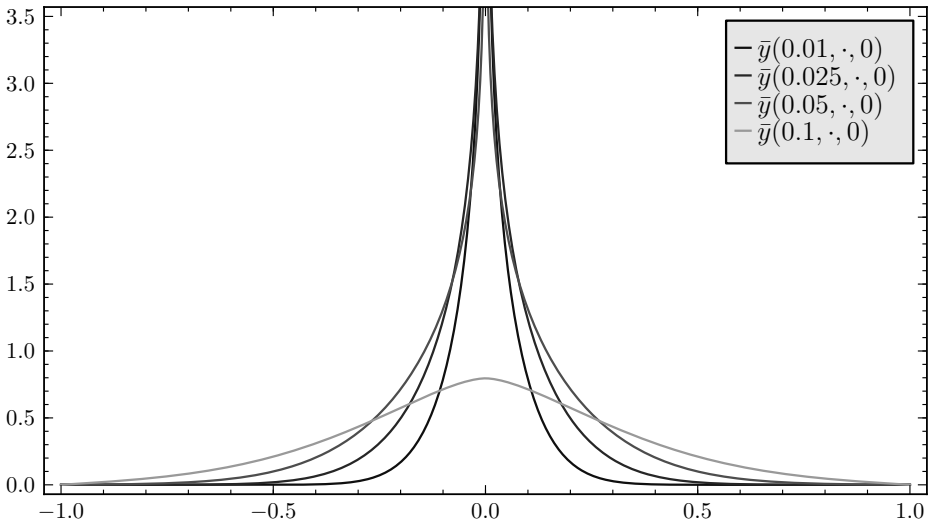


Figure 5.1.: Snapshots of the exact state solution $\bar{y}$ at $x_2 = 0$ (for $T = 0.1$).

$$\bar{y}(t, x) = \sum_{k\in\mathbb{Z}, l\in\mathbb{Z}} (-1)^{k+l}\, G(t,\, x_1 + 2k,\, x_2 + 2l), \tag{5.45}$$

where $x = (x_1, x_2)^t$ and G is the free space solution given by

$$G(t, x_1, x_2) = \frac{1}{4\pi\, T^2}\left((r^2/4 - T + t)\mathrm{Ei}\left(-r^2/(4t)\right) + te^{-r^2/(4t)}\right)$$

and $r^2 = x_1^2 + x_2^2$ is the squared distance to the origin. The function $\mathrm{Ei}(s) = \int_{-s}^{\infty} e^{-h}/h\;\mathrm{d}h$ is the exponential integral. The polar decomposition for $\bar{u} = \bar{u}'|\bar{u}|$ is given by

$$\bar{u}'(t) = \sqrt{3}\,T^{-3/2}\,(T - t)\,, \quad |\bar{u}| = \frac{1}{\sqrt{3}} T^{-1/2}\delta_0,$$

and a matching adjoint state $\bar{p}$ which fulfills the optimality conditions from Theorem 5.3 (and $\bar{p}(T) = 0$) can be chosen as

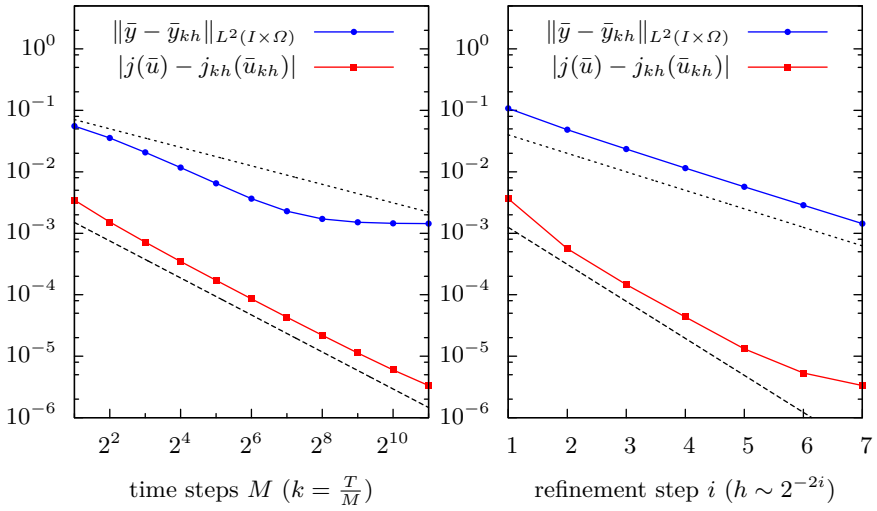
$$\bar{p}(t,x) = -\alpha\,\sqrt{3}\,T^{-3/2}\,(T-t)\,\cos\left(\pi/2\,x_1\right)\,\cos\left(\pi/2\,x_2\right).$$

The reader may verify that this $\bar{p}$ fulfills (5.6) and (5.7), which, in turn, implies the variational inequality (5.5). By inspection of the adjoint equation (5.4) we determine the desired state y_d to be

$$y_d = \bar{y} + \partial_t \bar{p} + \Delta \bar{p},$$

for which we can now derive an explicit formula by differentiating $\bar{p}$.

We choose the final time as $T = 0.1$ and $\alpha = 0.01$. For the practical verification of the convergence results we compute the optimal solutions $(\bar{u}_{kh}, \bar{y}_{kh})$ on an equidistant time grid with M steps and with a uniform triangulation of the square of different refinement levels. The series in (5.45) is approximated by the first nine terms, which yields a pointwise accuracy of about 10^{-12}. We use an adapted iterated quadrature formula in space to evaluate the integrals containing the singularity near $x = 0$ with sufficient accuracy. For the temporal integration, we use the box-rule. The convergence plots are given in Figure 5.2. We also plot the corresponding rates of convergence as predicted in Theorems 5.18 and 5.19 without the logarithmic factor. As



(a) Time refinement on grid level 7. (b) Space refinement with 2048 time steps.

Figure 5.2.: Error plots of the optimal solutions

we can see, the rates for the functional match the predicted order of almost $\mathcal{O}(k)$ and $\mathcal{O}(h^2)$, which are plotted for visual comparison. For the state error we make this observation only in the case of refinement in space: Figure 5.2b clearly shows a rate of $\mathcal{O}(h)$ in this case. For the case of time refinement, we seem to observe in Figure 5.2a a slightly better rate than the predicted $\mathcal{O}(\sqrt{k})$ (until the spatial error starts to dominate from 128 time steps on). For this reason we give the experimental orders of convergence in Table 5.1, which seem to indicate a possible rate close to $\mathcal{O}(k^{0.8})$.

Time steps	$\lvert j(\bar{u}) - j_{kh}(\bar{u}_{kh})\rvert$	Rate	$\lVert \bar{y} - \bar{y}_{kh}\rVert_{L^2(I\times\Omega)}$	Rate
2	$3.458 \cdot 10^{-3}$	–	$5.543 \cdot 10^{-2}$	–
4	$1.527 \cdot 10^{-3}$	1.17924	$3.553 \cdot 10^{-2}$	0.641629
8	$7.160 \cdot 10^{-3}$	1.09267	$2.072 \cdot 10^{-2}$	0.778014
16	$3.470 \cdot 10^{-4}$	1.04502	$1.172 \cdot 10^{-2}$	0.822051
32	$1.714 \cdot 10^{-4}$	1.01757	$6.509 \cdot 10^{-3}$	0.848465
64	$8.569 \cdot 10^{-5}$	1.00013	$3.658 \cdot 10^{-3}$	0.831381
128	$4.316 \cdot 10^{-5}$	0.98937	$2.291 \cdot 10^{-3}$	0.675078
256	$2.193 \cdot 10^{-5}$	0.97669	$1.716 \cdot 10^{-3}$	0.416928
512	$1.131 \cdot 10^{-5}$	0.95550	$1.512 \cdot 10^{-3}$	0.182591

Table 5.1.: Time refinement on grid level 7 (as in Figure 5.2a).

5.6. Point source identification

In this section we discuss a practical application of the abstract problem formulation to an inverse source problem. The state equation for the example is a simplified model for the transport and diffusion of a pollutant y in a lake (cf. [MRVM00]), given as

$$\left.\begin{aligned} \partial_t y - \nu\Delta y + b\cdot\nabla y &= u \quad \text{in } I\times\Omega,\\ \nu\partial_n y &= 0 \quad \text{on } I\times\partial\Omega\setminus\Gamma_{\text{in}},\\ \nu\partial_n y - n\cdot b\, y &= 0 \quad \text{on } I\times\Gamma_{\text{in}}, \end{aligned}\right\} \tag{5.46}$$

with initial condition $y(0) = 0$. The domain Ω describes the surface of the lake, the inflow boundary Γ_{in} is a subset of $\partial\Omega$, $\nu > 0$ is a diffusion parameter and b is assumed to be a static, smooth and divergence-free vector field (i.e., we assume the influence of y on the flow b to be negligible). We additionally define a outflow boundary Γ_{out}, such that b has the property

$$n\cdot b \quad \begin{cases} \leq 0 & \text{on } \Gamma_{\text{in}}\\ \geq 0 & \text{on } \Gamma_{\text{out}}\\ = 0 & \text{on } \partial\Omega\setminus(\Gamma_{\text{in}}\cup\Gamma_{\text{out}}), \end{cases}$$

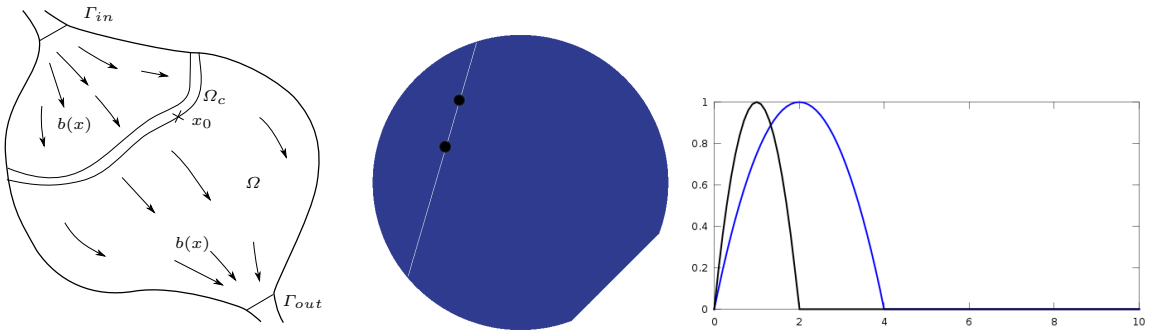
where $n\colon \partial\Omega \to \mathbb{R}^d$ is the outer normal. The source term u is assumed to consist of a finite number of pointwise inflows

$$\hat{u} = \sum_{i=1}^{N} \hat{u}_i(t)\delta_{\hat{x}_i} \tag{5.47}$$

where $\hat{x}_i \in \Omega_c$ are unknown locations and $\hat{u}_i(t)$ describes the unknown amount of substance leaking into the lake at $\hat{x}_i$ and time t. Furthermore we assume it is known that $\hat{x}_i \in \Omega_c$, where Ω_c is a line (e.g., a pipeline) intersecting Ω.

A schematic depiction of the setup and exemplary exact data is given in Figure 5.3. Furthermore, the diffusion coefficient is chosen as $\nu = 0.002$ and the vector field b is given by the negative gradient of a potential Φ on Ω. We set $b = -\nu\nabla\Phi$ and require

$$\begin{aligned} -\nabla\cdot\nu\nabla\Phi &= 0 \quad \text{in } \Omega,\\ \nu\partial_n\Phi &= 0 \quad \text{on } \partial\Omega\setminus(\Gamma_{\text{in}}\cup\Gamma_{\text{out}}),\\ \nu\partial_n\Phi &= \rho_{\text{in}} \geq 0 \quad \text{on } \Gamma_{\text{in}},\\ \nu\partial_n\Phi + \sigma\Phi &= \rho_{\text{out}} \leq 0 \quad \text{on } \Gamma_{\text{in}}. \end{aligned}$$



(a) Setup for (5.46). (b) Ω, Ω_c, $\hat{x}_1$ and $\hat{x}_2$. (c) Coefficients $\hat{u}_1$ and $\hat{u}_2$.

Figure 5.3.: Inverse problem setup.

The boundary conditions determine the shape of the velocity field in accordance with the conditions imposed on $\nabla \cdot b$ and $-n \cdot b$. The penalty factor $\sigma > 0$ is introduced to ensure unique solvability of the equation for Φ.

We solve the state equation for the data given in Figure 5.3. Corresponding snapshots for some $t \in I = (0, 10)$ of the state solution corresponding to the exact data are given in Figure 5.4.

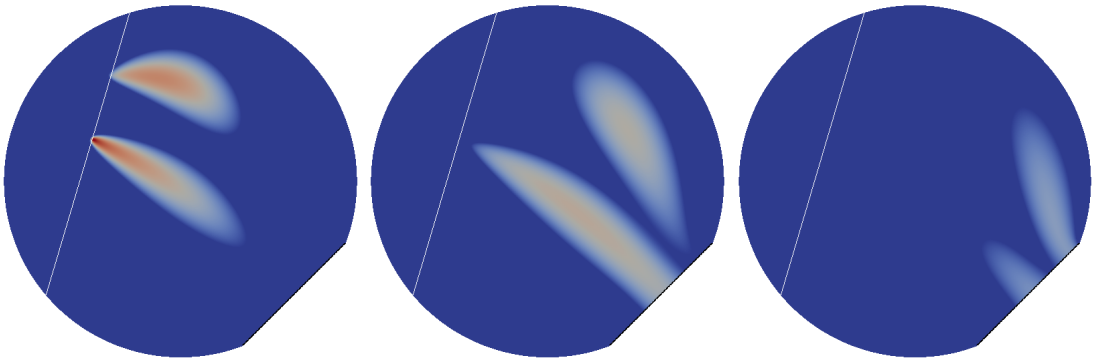


Figure 5.4.: Snapshots of the exact state $\hat{y}$ at $t = 2, 4, 6$

For the inverse problem we have available only the concentration of y on the outflow boundary in the form $y_{\text{obs}} = \hat{y}|_{I \times \Gamma_{\text{out}}} + \delta$. Here, $\hat{y}$ is the solution of (5.46) corresponding to the true source (5.47) and the noise term $\delta \in L^2(I \times \Gamma_{\text{out}})$ stands for an additional measurement error (which we will set to a deterministic function in our numerical experiments). For the concrete example from Figure 5.4 the corresponding observations are depicted in Figure 5.5.

To give a reconstruction of the source $\hat{u}$, we propose to solve the deterministic inverse problem

$$\min_{u \in \mathcal{M}(\Omega_c, L^2(I))} \frac{1}{2} \|S(u) - y_{\text{obs}}\|^2_{L^2(I \times \Gamma_{\text{out}})} + \alpha \|u\|_{\mathcal{M}(\Omega_c, L^2(I))}$$

where $S(u)$ is the solution of (5.46) corresponding to u. This inverse problem formulation is similar to the approach described in [BP13], if we would somehow replace the Hilbert space $L^2(I)$ with $\mathbb{R}^M$ for some $M \in \mathbb{N}$. For the concrete example with the depicted data we empirically determine $\alpha = 0.5$ to be an appropriate regularization parameter. The optimal state solution

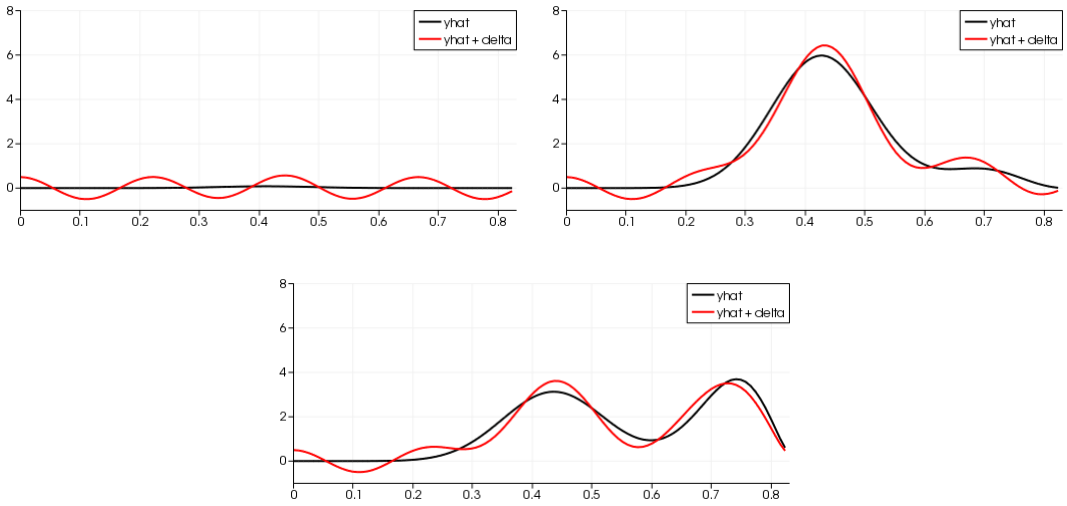


Figure 5.5.: Snapshots of the observation y_{obs} on the observation boundary Γ_{out} (with and without noise) at $t = 2, 4, 6$.

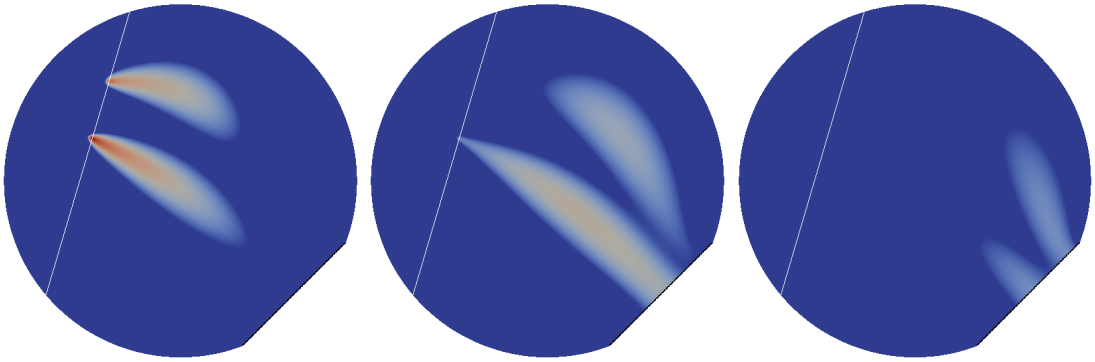
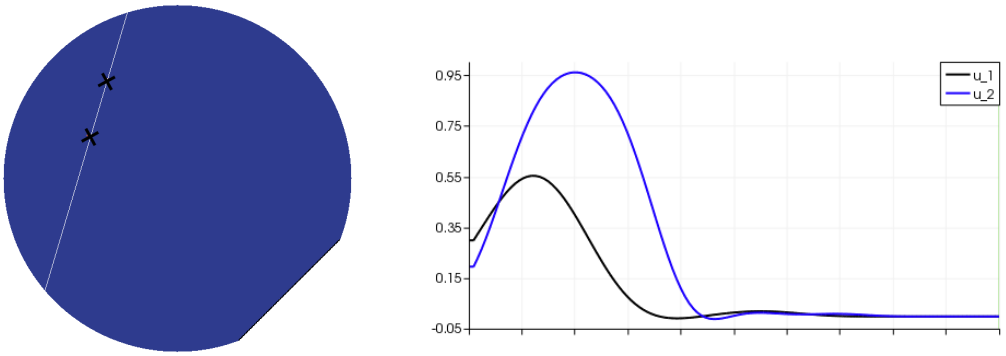


Figure 5.6.: Snapshots of the reconstructed $\bar{y}$ at $t = 2, 4, 6$ for $\alpha = 0.5$

$\bar{y}$ is visualized in Figure 5.6. For the numerical realization we added a small L^2 regularization term as in described in section 5.4, with a value of $\gamma = 10^{-6}$ in the depicted simulation. Due to discretization and the additional L^2 regularization, the discrete $\bar{u}_{kh}$ does have not the structure as in (5.47) (for $N = 2$) since it is the linear combination of more than two Dirac delta functions. As a postprocessing strategy, to obtain the visualization in Figure 5.7, we group all the connected components of the grid points in the support of $\bar{u}_{kh}$ and identify each of them with a central point $\tilde{x}_i$ of the component. In the concrete case we have exactly two components. Then we identify the spatial part of the of $\bar{u}_{kh}$ with $|\bar{u}_{kh}| \approx \sum_{i=1,2} c_i \delta_{\tilde{x}_i}$, where the c_i is the sum over all coefficients of $|\bar{u}_{kh}|$ in each component. From the optimality condition (2.27) we derive a reconstruction of the coefficients of the form $\tilde{u}_i(t) = -\frac{c_i}{\alpha}\bar{p}_{kh}(t, \tilde{x}_i)$; cf. Corollary 2.23.

We see that the outlined reconstruction procedure gives the main structural features of the exact source $\hat{u}$, such as the number and location of the points x_i, and a quantitatively adequate estimate of the coefficients u_i (which is in contrast to the results we would obtain with a regularization approach based on the L^2-norm). Certainly, there is a qualitative error between



(a) supp $\bar{u}_{kh}$ and reconstructed $\tilde{x}_1$ and $\tilde{x}_2$.

(b) Reconstructed $\tilde{u}_1$ and $\tilde{u}_2$.

Figure 5.7.: Postprocessing: visualization of the reconstructed $\bar{u}$.

$\hat{u}$ and $\bar{u}$ which stems from the noise δ and the nonzero regularization parameter α. However, a detailed study of the reconstruction error for a systematic choice of α depending on the magnitude of δ (as in [BP13]) is beyond the scope of this work.

6. A posteriori error analysis and adaptivity

In this chapter, we will consider a solution method based on a adaptive mesh refinement for an elliptic control problem of the following form:

$$\begin{aligned} \min_{u\in U_{\mathrm{ad}},\, y\in W_0^{1,s}(\Omega\cup\Gamma)} \quad & J(y)+\psi(u), \\ \text{subject to} \quad & e(y,u)(\varphi)=0 \quad \text{for all } \varphi\in W. \end{aligned} \tag{6.1}$$

We suppose that $\psi\colon \mathcal{M}(\Omega_c)\to\mathbb{R}$ is the weighted total variation norm

$$\psi(u)=\alpha\int_{\Omega_c}\mathrm{d}|u|(x),$$

and U_{ad} are (optional) positivity constraints (either $U_{\mathrm{ad}}=\mathcal{M}(\Omega_c)$ or $U_{\mathrm{ad}}=\mathcal{M}^+(\Omega_c)$). The analysis will be done for the elliptic model problem from section 2.2: the state equation is given by

$$e(y,u)(\cdot)=a(y,\cdot)-\langle\chi_{\Omega_c}u,\cdot\rangle,$$

where a is an elliptic bilinear form (see section 2.2). We assume that the domain Ω is polygonal. As before, Ω_c denotes the control set, and J is a quadratic tracking term on the observation region Ω_o (either distributed or boundary observation),

$$J(y)=\frac{1}{2}\|y-y_d\|^2_{L^2(\Omega_o)}.$$

The adaptive algorithm is based on the regularized problem

$$\begin{aligned} \min_{u\in L^2(\Omega_c)\cap U_{\mathrm{ad}},\, y\in H_0^1(\Omega\cup\Gamma)} \quad & J(y)+\psi(u)+\frac{\gamma}{2}\|u\|^2, \\ \text{subject to} \quad & e(u,y)(\varphi)=0 \quad \text{for all } \varphi\in H_0^1(\Omega\cup\Gamma). \end{aligned} \tag{6.2}$$

We consider a discretization of (6.2) based on isoparametric bilinear (or trilinear, in the three dimensional case) finite elements; see section 6.3.

We derive a posteriori error indicators for the solution of (6.1) on adaptive meshes. The refinement strategy is based on error indicators for the cost functional, obtained with the dual-weighted-residual (DWR) approach by Becker, Kapp, and Rannacher [BKR00; BR01]. The error between the cost functional of the continuous and discrete problem are expressed as weighted residuals of the state equation, adjoint equation, and the optimality condition for the control, which are then localized to the individual cells of the mesh. Thereby, we hope to identify an optimal mesh to achieve a specified accuracy for the optimal objective functional. We do not discuss adaptivity with respect to different "quantities of interest"; cf., e.g., [VW08]. The discretization error estimator (denoted by η_h) and the corresponding indicators are derived for the regularized problem (6.2). Since the problem (6.2) is similar to a control constrained optimization problem, we can adapt many of the ideas in Vexler and Wollner [VW08]. The main idea of our modifications, to cope with the missing differentiability of ψ, is to replace the

cost term by a linear expression with the subgradient (using the homogeneity of degree one of ψ and the optimality conditions from Proposition 2.5).

To ensure that the error due to regularization is sufficiently small, the parameter γ has to be chosen appropriately small. However, a large regularization parameter facilitates the numerical solution of (6.2) (it leads to more efficient optimization algorithms) and improves the quality of the error indicators for the discretization error (since they are derived under the assumption $\gamma > 0$). Therefore, we are interested in an accurate estimate of the regularization error in the optimal functional value. The computational estimate to be derived below, which we denote by η_γ, is based on an asymptotic model motivated by the a priori analysis of the regularization error; cf. [IK92; HK06a; HK06b]. In each step, the adaptive strategy will solve the discrete regularized problem, evaluate the indicators η_γ and η_h such that

$$j(\bar{u}) - j_{\gamma,h}(\bar{u}_{\gamma,h}) \approx \eta_\gamma + \eta_h,$$

and either refine the mesh according to the localization of η_h, or decrease the regularization parameter. The aim is to balance the relatives size of η_γ and η_h.

Similar methods are used for state constrained problems; see Wollner [Wol10; RVW12] for a goal-oriented adaptive algorithm in combination with an interior point reformulation. A related but different approach (which is based on convergence of an interior point method in function space) has been proposed by Günther and Schiela [SG11]. Other approaches for the derivation of goal-oriented error estimates for state constrained problems directly work with the unregularized problem formulation; see Benedix and Vexler [BV09] or Hintermüller and Hoppe [HH10]. Parameter updates for the regularization parameter in Moreau-Yosida regularization methods based on a priori estimates have been considered by Hintermüller and Hinze [HH09].

This chapter is structured as follows. In section 6.1 we explain the error estimation strategy for the regularization error. Section 6.3 contains the derivation of the discretization error estimator: First, we compute an equivalent representation for the error in terms of weighted residuals and a complementarity term. Then, we describe the practical evaluation and localization strategy. Some further specializations and justifications for the estimators for the case of piece-wise constant control discretization and piece-wise bilinear control discretization with and without mass lumping are provided. In section 6.4 we sketch the idea behind the algorithmic strategy. Section 6.5 contains numerical results, which demonstrate the efficiency of the estimators for two model problems. Finally, in section 6.6 we compare the adaptive algorithm to the nodal Dirac discretization from chapter 4 and give an outlook on possibilities for further improvements.

6.1. Problem setup

For the adaptive algorithm we work with the regularized problem for a decreasing sequence of parameters γ. First, we briefly recapitulate the necessary notation from the previous chapters. We denote the inner product and norm on $L^2(\Omega)$ by $(\cdot,\cdot) = (\cdot,\cdot)_{L^2(\Omega)}$. Furthermore, we abbreviate $U = L^2(\Omega_c)$ and denote the corresponding inner product by $(\cdot, \chi_{\Omega_c}\cdot)$. In cases where no confusion arises, we omit the characteristic function and we also denote the norm in U by $\|\cdot\| = \|\cdot\|_{L^2(\Omega_c)}$. Furthermore, we abbreviate $V = H_0^1(\Omega \cup \Gamma)$. The state equation, given by

$$e(y,u)(\varphi) = a(y,\varphi) - (u,\varphi) = 0 \quad \text{for all } \varphi \in V \tag{6.3}$$

admits a unique solution $y = S(u) \in V$ for any given $u \in U$. We define the reduced cost functionals f and f_γ for the smooth part, and j and j_γ for the full functional as before by the relations

$$j_\gamma(u) = j(u) + \frac{\gamma}{2}\|u\|^2 = f(u) + \psi(u) + \frac{\gamma}{2}\|u\|^2 = J(S(u)) + \psi(u) + \frac{\gamma}{2}\|u\|^2.$$

For convenience of notation, we only consider the unconstrained setting $U_{\mathrm{ad}} = U$. The modifications for the general case with constraints are obvious, in each case. As discussed section 2.5, an optimality condition for (6.2) is given by (6.3), the adjoint equation,

$$e'_y(y,u)(\varphi,p) = a(\varphi,p) = J'(y)(\varphi) \quad \text{for all } \varphi \in V, \tag{6.4}$$

and the optimality condition

$$e'_u(y,u)(\tilde{u},p) + \gamma(u,\tilde{u}) + \psi(u) = (p + \gamma u, \tilde{u}) \leq \psi(\tilde{u}) - \psi(u) \quad \text{for all } \tilde{u} \in U, \tag{6.5}$$

where $(u,y,p) = (\bar{u}_\gamma, \bar{y}_\gamma, \bar{p}_\gamma)$ are the (unique) optimal control, optimal state and optimal adjoint state, respectively. Furthermore, as in section 3.1, the subdifferential inclusion above can be rewritten with the proximal map P_γ as

$$\bar{u}_\gamma = P_\gamma(\bar{q}_\gamma),$$

where $\bar{q}_\gamma \in U$ is the gradient of $-1/\gamma\, f$ in the optimum. It is defined by

$$(\bar{q}_\gamma, \varphi) = -\frac{1}{\gamma} e'_u(y,u)(\varphi, \bar{p}_\gamma) = -\frac{1}{\gamma}(\chi_{\Omega_c}\bar{p}_\gamma, \varphi) \quad \text{for all } \varphi \in U.$$

The proximal map for the parameter $c > 0$ is given by

$$P_c(q) = \mathrm{shrink}_{\alpha/c}(q) = (q - \alpha/c)^+ - (q + \alpha/c)^-.$$

For an improved estimate of the regularization error, we will require the perturbation of the solution $\bar{u}_\gamma$ with respect to the regularization parameter. To this purpose, we first recall the definition of the generalized derivative of the proximal map. It is given by

$$\mathrm{D}P_c(q)\delta q = \chi_{\mathcal{I}(q)}\delta q,$$

where $\chi_{\mathcal{I}(q)}$ is the characteristic function of the set $\mathcal{I}(q) = \{\, x \in \Omega_c \mid -\alpha/c < q(x) < \alpha/c \,\}$. Now, we define the "sensitivity" $\dot{u}_\gamma$ as the unique solution of the auxiliary optimization problem

$$\begin{aligned} \min_{\dot{u} \in L^2(\mathcal{I}(\bar{q}_\gamma)),\, \dot{y} \in V} \quad & \frac{1}{2}\|\dot{y}\|^2_{L^2(\Omega_o)} + \frac{\gamma}{2}\|\dot{u}\|^2 + (\bar{u}_\gamma, \dot{u}), \\ \text{subject to} \quad & a(\dot{y}, \varphi) = (\dot{u}, \varphi) \quad \text{for all } \varphi \in V. \end{aligned} \tag{6.6}$$

With this definition of $\dot{u}_\gamma$, we can express the second derivative of the value function in an almost everywhere sense, using the results from Wachsmuth and Wachsmuth [WW11]; see Proposition 6.1. Furthermore, a straightforward computation reveals that $\dot{u}_\gamma$ can alternatively given by $\dot{u}_\gamma = \chi_{\mathcal{I}(\bar{q}_\gamma)}\dot{q}_\gamma$, where the auxiliary variable $\dot{q}_\gamma$ is defined as the solution to the linear equation

$$\gamma\dot{q}_\gamma + \nabla^2 f(\bar{u}_\gamma)\mathrm{D}P_c(q)\dot{q}_\gamma = -\bar{u}_\gamma. \tag{6.7}$$

Thereby, the solution (6.6) corresponds to a computation of one Newton step for the original problem (see section 3.2.2). Note, that with this definition of $\dot{u}_\gamma$ we can only obtain the

directional derivative of the map $\gamma \mapsto \bar{u}_\gamma$ in the case where the generalized derivative $\mathrm{D}P_c(q)$ coincides with the directional derivative $\mathrm{d}P_c(q,\cdot)$. This is exactly the case under the assumption that the set $\{\, x \in \Omega_c \mid -\alpha/c = \bar{q}_\gamma(x) \;\text{ or }\; \bar{q}_\gamma(x) = \alpha/c \,\}$ has Lebesgue measure zero. To obtain the derivative in the general case, we would have to replace the generalized derivative in (6.7) with the directional one, and work with the concept of Bouligand differentiability; see, e.g., [GV07; GGW08].

6.2. The regularization error

An asymptotic a priori error estimate for the elliptic linear quadratic model problem of the form

$$0 \leq j(\bar{u}) - j_\gamma(\bar{u}_\gamma) \leq C\,\gamma^s$$

for $s \in (0,1)$ has been derived in section 4.5.1. In this section, we will describe two heuristic strategies to evaluate this error a posteriori, based only on knowledge of the optimal triple $(\bar{u}_\gamma, \bar{y}_\gamma, \bar{p}_\gamma)$. As in section 2.26, we denote the optimal value function by

$$v(\gamma) = j_\gamma(\bar{u}_\gamma) = j(\bar{u}_\gamma) + \frac{\gamma}{2}\|\bar{u}_\gamma\|^2 \quad \text{for } \gamma > 0.$$

We recall that the value function is concave (see Proposition 2.26) and that the first derivative is given by

$$v'(\gamma) = \frac{1}{2}\|\bar{u}_\gamma\|^2 \quad \text{for } \gamma > 0.$$

Furthermore, we recall that in the present convex case, the derivative of v is Lipschitz continuous (see Proposition 2.32). Furthermore, the second derivative of the value function can be computed; cf. also [HK06a; HK06b; WW11].

Proposition 6.1. *With $\dot{u}_\gamma$ defined as the solution of* (6.6)*, the second derivative of v can be expressed as*

$$v''(\gamma) = (\bar{u}_\gamma, \dot{u}_\gamma) \quad \textit{for } \gamma > 0 \quad \textit{almost everywhere.}$$

Proof. For the sparse control problem under consideration this result can be found in [WW11, Lemma 3.5]. □

Motivated by this result, we will silently assume in the following that for the current value of $\gamma > 0$, the directional derivative in (6.7) is linear (and consequently the value function v is twice differentiable). As mentioned before, this is equivalent to the assumption that the set $\{\, x \in \Omega_c \mid \bar{q}_\gamma = -\alpha/\gamma \text{ or } \bar{q}_\gamma = \alpha/\gamma \,\}$ has Lebesgue measure zero, which is also referred to as *strict complementarity*.

Taylor expansion

A very simple strategy to estimate the error is given by a first order Taylor expansion at the point γ: we approximate

$$v(0) - v(\gamma) = -v'(\gamma)\gamma + R(\gamma) \approx \eta_\gamma^{\mathrm{triv}} = -v'(\gamma)\gamma = -\frac{\gamma}{2}\|\bar{u}_\gamma\|^2,$$

where $R(\gamma)$ is the corresponding remainder, which we neglect. In other words, we declare that the error between $j(\bar{u})$ and $j_\gamma(\bar{u}_\gamma)$ should consist exactly of the regularization term

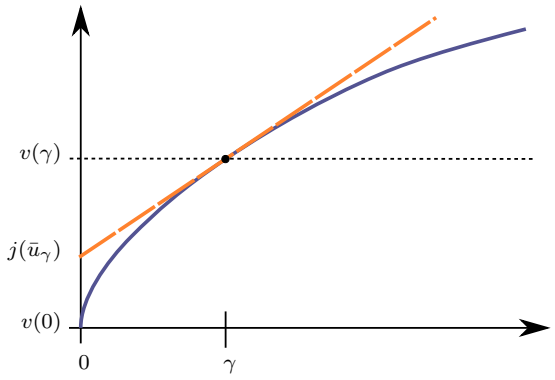


Figure 6.1.: $v(\gamma) \approx v(0) + c\gamma^s$

$\eta_\gamma^{\text{triv}} = -\gamma/2\,\|\bar{u}_\gamma\|^2$. Of course, this is cheap and trivial to implement, but unfortunately not asymptotically accurate. Typically, the corresponding effectivity index

$$I_{\text{eff}}^\gamma = \frac{j(\bar{u}) - j_\gamma(\bar{u}_\gamma)}{\eta_\gamma^{\text{triv}}}$$

does not converge to one for $\gamma \to 0$. To understand this, we make for the value function the ansatz $v(\gamma) \approx m(\gamma) = m_0 + c\gamma^s$ for some $m_0 \in \mathbb{R}$, $c > 0$, and $s \in (0, 1]$, which is motivated by the a priori analysis. For the model function m, we obtain with a simple computation the identity

$$-m'(\gamma)\gamma = s\,(m(0) - m(\gamma)).$$

Using $\eta_\gamma^{\text{triv}}$ to estimate the error $m(0) - m(\gamma)$, the resulting effectivity index is given by $I_{\text{eff}}^\gamma = (m(0) - m(\gamma))/(-m'(\gamma)\gamma) = s^{-1} \geq 1$. Unless s is equal to one, we underestimate the error by a constant factor; this is depicted in Figure 6.1. However, this estimator can be surprisingly useful in practice. This is due to the fact that the error is underestimated by the constant factor s^{-1} (which was bounded by 3 in the two dimensional case and $4 + \varepsilon$ in the three dimensional case; see Proposition 4.26), and a precise estimate is often not necessary; see section 6.5.

A model function

Motivated by the a priori analysis and the above discussion, we develop a second estimation approach. As substitute for the value function, we introduce the model function m, given by

$$v(\gamma) \approx m(\gamma) = m_0 + c\gamma^s$$

for some (m_0, c, s); as before. Then, for fixed $\gamma > 0$ we choose the parameters (m_0, c, s) based on current data: we require

$$\begin{aligned} m(\gamma) = m_0 + c\gamma^s &= v(\gamma) = j_\gamma(\bar{u}_\gamma), \\ m'(\gamma) = cs\gamma^{s-1} &= v'(\gamma) = \frac{1}{2}\|\bar{u}_\gamma\|^2, \\ m''(\gamma) = cs(s-1)\gamma^{s-2} &= v''(\gamma) = (\bar{u}_\gamma, \dot{u}_\gamma), \end{aligned}$$

We have seen that the sensitivity $\dot{u}_\gamma$ can be computed by solving a linear quadratic optimization problem. Under the assumption of strict complementarity this has a similar computational effort to the computation of one Newton step for the original problem. Based on this, we define the estimate

$$v(0) - v(\gamma) \approx \eta_\gamma^{\mathrm{mod}} = m_0 - m(\gamma) = -c_{\mathrm{est}}\gamma^{s_{\mathrm{est}}}.$$

A quick computation reveals that the relations above determine (c, s) to be given by $(c_{\mathrm{est}}, s_{\mathrm{est}})$ with

$$s_{\mathrm{est}} = 1 + \frac{\gamma v''(\gamma)}{v'(\gamma)} \quad \text{and} \quad c_{\mathrm{est}} = \frac{v'(\gamma)}{s_{\mathrm{est}}\gamma^{s_{\mathrm{est}}-1}}. \tag{6.8}$$

This implies that $\eta_\gamma^{\mathrm{mod}}$ has the closed form representation

$$\eta_\gamma^{\mathrm{mod}} = -c_{\mathrm{est}}\gamma^{s_{\mathrm{est}}} = -\frac{\gamma v'(\gamma)}{1 + \gamma v''(\gamma)/v'(\gamma)} = \frac{\eta_\gamma^{\mathrm{triv}}}{s_{\mathrm{est}}}. \tag{6.9}$$

Therefore, we can give another interpretation of this modified estimator: with the help of the estimated rate of convergence s_{est} we try to compensate the systematic error that results from a simple Taylor approximation as for $\eta_\gamma^{\mathrm{triv}}$.

6.3. The discretization error

In this section, we discretize the regularized problem (6.2). For the state, we use bilinear (trilinear) isoparametric finite elements and for the control we consider different spaces (cf. section 4.5.3). We denote for the discretization parameter h the triangulation by $\mathcal{T}_h = \{ K \}$, which is a collection of disjoint quadrilaterals (hexahedrals, in three dimensions) with

$$\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} \bar{K}.$$

For each $K \in \mathcal{T}_h$, the map $T_K \colon \hat{K} \to K$ denotes a bilinear (trilinear) transformation from the reference cell $\hat{K} = (0,1)^d$. We assume that for all h the boundary Γ can be exactly represented by the corresponding faces of the adjacent cell. The space of isoparametric bilinear finite elements (see, e.g., [BS08, Chapter 10.4]) associated with $\mathcal{T}_h$ is defined as usual by

$$V_h = \left\{ v_h \in \mathcal{C}(\bar{\Omega}) \;\middle|\; T_K^{-1} \circ v_h|_K \in \mathcal{Q}_1(\hat{K}) \quad \text{for all } K \in \mathcal{T}_h \right\} \cap H_0^1(\Omega \cup \Gamma).$$

The space $\mathcal{Q}_1(\hat{K})$ denotes the space of bilinear functions on the unit square (trilinear functions on the unit cube).

To enable local mesh refinement, we relax the usual regularity assumption on the triangulation by allowing so-called *hanging nodes*; cf., e.g., [BR01; BE03; Mei08]. For each $K \in \mathcal{T}_h$ and a corresponding face $F \subset \partial K$, we require that F is either a subset of $\partial\Omega$, identical to a face of a neighboring cell, or identical to the disjoint union of two (four, in three dimensions) neighboring cells. Note that this still results in a conforming finite element space, as defined above, if the degrees of freedom on the hanging nodes are fixed to the appropriate average of the values of the neighboring vertices. For details on the practical implementation of this approach we refer to [CO84]. For the evaluation of the error estimator we additionally assume that $\mathcal{T}_h$ has *patch structure*, i.e., it results from a uniform refinement of a coarse triangulation denoted by $\mathcal{T}_{2h}$.

The space of piecewise constant functions is defined as

$$V_h^0 = \left\{ u_h \in L^2(\Omega) \;\middle|\; u_h|_K \in \mathcal{P}_0(K) \quad \text{for all } K \in \mathcal{T}_h \right\}.$$

As discussed in section 4.5.3, we consider a discretization of the control with piecewise constants discontinuous or piecewise bilinear (trilinear) continuous finite elements. In the former case we define $U_h = U_h^0 = V_h^0 \cap L^2(\Omega_c)$ and in the latter case we define $U_h = U_h^1 = V_h \cap L^2(\Omega_c)$. We consider the controls on Ω_c as restrictions of finite element functions on Ω, which is always possible and leads to a simple notation. We also use the same mesh for the control and state variable in the practical implementation. However, from a practical standpoint, it would certainly be beneficial to consider different meshes for control and state. Since the error indicators derived below can be separated into errors stemming from control and state discretization, we will comment on possible extensions into this direction.

The discrete regularized problem is given by

$$\begin{aligned} \min_{u \in U_{\text{ad}} \cap U_h,\, y \in V_h} \quad & J_h(y_h) + \psi_h(u_h) + \frac{\gamma}{2}\|u_h\|_h^2, \\ \text{subject to} \quad & e_h(y_h, u_h)(\varphi_h) = 0 \quad \text{for all } \varphi_h \in V_h. \end{aligned} \tag{6.10}$$

The subscript h for each of the expressions e, J, ψ, and $\|\cdot\|$ indicates a possible evaluation of the corresponding terms by numerical quadrature. The discrete state equation is given as

$$e_h(y_h, u_h)(\varphi_h) = a_h(y_h, \varphi_h) - (u_h, \varphi_h)_h \quad \text{for } u_h \in U_h,\ y_h \in V_h,\ \text{and } \varphi_h \in V_h.$$

Since we work with isoparametric finite elements, we cannot expect to exactly evaluate the cell-wise contributions for $a(y_h, \varphi_h)$; therefore we also add an h here. In the case where all transformations T_K are linear and the coefficients in the bilinear form are constant in space, a quadrature formula of sufficiently high order on each cell (e.g., the tensor-product rule resulting from the two-point Gaussian rule) guarantees an exact evaluation. For the right-hand side an exact evaluation can be guaranteed if the control set is compatible with the mesh. Additionally, as motivated by the analysis in section 4.5.3, we will also consider an evaluation of the control term $(u_h, \varphi_h)_h$ with a lower order quadrature rule, i.e., the tensor-product trapezoidal rule, which is referred to as *mass lumping*. We will go into more detail below. The discrete tracking term is given by

$$J_h(y_h) = \frac{1}{2}\|y_h - y_d\|_{L^2(\Omega_o),h}^2 = \frac{1}{2}\left[\int_{\Omega_o} (y_h - y_d)^2 \,\mathrm{d}x\right]_h \quad \text{for } y_h \in V_h.$$

Again, if we use a quadrature formula of sufficiently high order, assume that Ω_o is approximated well by the corresponding cells of the mesh, and assume that y_d is smooth, we can expect an accurate evaluation of this term. The control cost term is defined as

$$\psi_h(u_h) + \frac{\gamma}{2}\|u_h\|_h^2 = \alpha\left[\int_{\Omega_c} |u_h| \,\mathrm{d}x\right]_h + \frac{\gamma}{2}\left[\int_{\Omega_c} u_h^2 \,\mathrm{d}x\right]_h \quad \text{for } u_h \in U_h.$$

We assume that Ω_c is approximated sufficiently well by the triangulation. The L^1 norm is always evaluated with mass lumping, which is motivated by the splitting into positive and negative part. This is exact in the case that u_h is either only positive or only negative on each single cell; cf. the discussion in section 4.5.3. For the regularization term we choose the

quadrature formula in accordance with the choice made for the control above. In all cases we endow the control space U_h with the discrete inner product

$$(u_h, \varphi_h)_h = \left[\int_{\Omega_c} u_h \varphi_h \,\mathrm{d}x\right]_h \quad \text{for } u_h \in U_h, \varphi_h \in U_h,$$

where the quadrature rule is chosen as for the regularization term and the control term. Again, we give a more detailed description below; see section 6.3.2.

We denote the solution operator of the discrete state equation by $S_h(u_h) = y_h$. In the following, we use the approach based on the dual-weighted-residual method (see [BR01]) to assess the discretization error in the discrete reduced cost functional

$$j_{\gamma,h}(u_h) = J_h(S_h(u_h)) + \psi_h(u_h) + \frac{\gamma}{2}\|u_h\|_h^2$$

in the optimal solution. Corresponding to the continuous Lagrange functional defined as

$$\mathcal{L}(u, y, p) = J(y) - e(y, u)(p)$$

for $u \in U$, $y \in V$, and $\varphi \in V$, we define the discrete Lagrange functional as

$$\mathcal{L}_h(u_h, y_h, p_h) = J_h(y_h) - e_h(y_h, u_h)(p_h)$$

for $u_h \in U_h$ $y_h \in V_h$, and $\varphi_h \in V_h$. As before, the unique optimal solution $(u_h, y_h, p_h) = (\bar{u}_{\gamma,h}, \bar{y}_{\gamma,h}, \bar{p}_{\gamma,h})$ of (6.10) is characterized by the relations

$$\begin{aligned}
e_h(y_h, u_h)(\varphi_h) = a_h(y_h, \varphi_h) - (u_h, \varphi_h)_h = 0 &\qquad \text{for all } \varphi_h \in V_h,\\
e'_{h,y}(y_h, u_h)(\varphi_h, p_h) = a_h(\varphi_h, p_h) = J'_h(y_h)(\varphi_h) &\qquad \text{for all } \varphi_h \in V_h,\\
e'_{h,u}(y, u)(\tilde{u}_h, p_h) + \gamma(u_h, \tilde{u}_h)_h = (p_h + \gamma u_h, \tilde{u}_h)_h \le \psi_h(\tilde{u}_h) - \psi_h(u_h) &\qquad \text{for all } \tilde{u}_h \in U_h.
\end{aligned}$$

As in the continuous case, we define the auxiliary variable $\bar{q}_{\gamma,h} \in U_h$ by

$$(\bar{q}_{\gamma,h}, \varphi_h)_h = -\frac{1}{\gamma} e'_{h,u}(\bar{y}_h, \bar{u}_h)(\varphi_h, \bar{p}_{\gamma,h}) = -\frac{1}{\gamma}(\bar{p}_{\gamma,h}, \varphi_h)_h \quad \text{for all } \varphi_h \in U_h.$$

In other words, we define $\bar{q}_{\gamma,h}$ as the (discrete) L^2-projection of $-1/\gamma\, \chi_{\Omega_c} \bar{p}_{\gamma,h}$ on the control space U_h. In the case of $U_h = U_h^1$ (piecewise linear controls) and if Ω_c is compatible with the triangulation, we have $\bar{q}_{\gamma,h} = -1/\gamma\, \chi_{\Omega_c} \bar{p}_{\gamma,h}$. With this variable, we can express the optimality condition alternatively with a proximal map as

$$\bar{u}_{\gamma,h} = P_{\gamma,h}(\bar{q}_{\gamma,h}) = \operatorname*{argmin}_{u_h \in U_h} \left[\frac{\gamma}{2}\|\bar{q}_{\gamma,h} - u_h\|_h^2 + \psi_h(u_h)\right].$$

Note that $P_{\gamma,h} \colon U_h \to U_h$ will generally not be the same operator as its continuous counterpart P_γ. In particular, $P_{\gamma,h}$ does not generally have a closed form representation; we will give a more detailed description below.

6.3.1. Finite element error for the regularized problem

We give a reformulation of the error in the functional with the DWR method. Since ψ is not smooth, we define the optimal subgradients $\bar{\lambda}_\gamma \in U$ and $\lambda_{\gamma,h} \in U_h$ in the subdifferential of ψ and ψ_h respectively as

$$\begin{aligned}
\bar{\lambda}_\gamma &= \gamma\,(\bar{q}_\gamma - \bar{u}_\gamma)\,,\\
\bar{\lambda}_{\gamma,h} &= \gamma\,(\bar{q}_{\gamma,h} - \bar{u}_{\gamma,h})\,.
\end{aligned}$$

In fact, with Proposition 3.4.(i), we have $\bar{\lambda}_\gamma \in \partial\psi(\bar{u}_\gamma)$ and $\bar{\lambda}_{\gamma,h} \in \partial_h\psi_h(\bar{u}_\gamma)$. The subdifferential $\partial_h\psi_h$ is defined w.r.t. the inner product in U_h; to be precise, we state again the full form

$$(\bar{\lambda}_{\gamma,h}, \tilde{u}_h)_h = \gamma(\bar{q}_{\gamma,h} - \bar{u}_{\gamma,h}, \tilde{u}_h)_h \leq \psi_h(\tilde{u}_h) - \psi_h(\bar{u}_{\gamma,h}) \quad \text{for all } \tilde{u}_h \in U_h.$$

Since ψ and ψ_h are positively homogeneous (of degree one), we can represent the functional value with the help of the subgradient (see Proposition 2.5) as

$$\begin{aligned} (\bar{\lambda}_\gamma, \bar{u}_\gamma) &= \psi(\bar{u}_\gamma), \\ (\bar{\lambda}_{\gamma,h}, \bar{u}_{\gamma,h})_h &= \psi_h(\bar{u}_{\gamma,h}). \end{aligned} \tag{6.11}$$

For convenience of notation, we abbreviate the optimal variables in the following without subscript γ by

$$\begin{aligned} \bar{\chi} &= (\bar{q}, \bar{\lambda}, \bar{y}, \bar{p}) = (\bar{q}_\gamma, \bar{\lambda}_\gamma, \bar{y}_\gamma, \bar{p}_\gamma), \\ \bar{\chi}_h &= (\bar{q}_h, \bar{\lambda}_h, \bar{y}_h, \bar{p}_h) = (\bar{q}_{\gamma,h}, \bar{\lambda}_{\gamma,h}, \bar{y}_{\gamma,h}, \bar{p}_{\gamma,h}). \end{aligned} \tag{6.12}$$

Furthermore, we introduce the modified Lagrange functionals

$$\begin{aligned} \tilde{\mathcal{L}}(\chi) = \tilde{\mathcal{L}}(u, \lambda, y, p) &= \mathcal{L}(u, y, p) + (\lambda, u) + \frac{\gamma}{2}\|u\|^2, \\ \tilde{\mathcal{L}}_h(\chi_h) = \tilde{\mathcal{L}}_h(u_h, \lambda_h, y_h, p_h) &= \mathcal{L}_h(u_h, y_h, p_h) + (\lambda_h, u_h)_h + \frac{\gamma}{2}\|u_h\|_h^2. \end{aligned}$$

The Lagrange functions $\tilde{\mathcal{L}}$ and $\tilde{\mathcal{L}}_h$ are smooth, so we can proceed to give an estimate for the error; cf. [BR01, Proposition 2.1], [BKR00, Section 4.3], and [VW08, Section 4.2].

Proposition 6.2. *For the optimal variables* (6.12) *it holds*

$$\begin{aligned} j_\gamma(\bar{u}) - j_{\gamma,h}(\bar{u}_h) &= \tilde{\mathcal{L}}(\bar{\chi}_h) - \tilde{\mathcal{L}}_h(\bar{\chi}_h) \\ &\quad + \frac{1}{2}\left[\rho_y(\bar{\chi}_h)(\bar{p} - \bar{p}_h) + \rho_p(\bar{\chi}_h)(\bar{y} - \bar{y}_h) + \rho_u(\bar{\chi}_h)(\bar{u} - \bar{u}_h) + (\bar{u} + \bar{u}_h, \bar{\lambda} - \bar{\lambda}_h)\right], \end{aligned} \tag{6.13}$$

where the residuals ρ_y, ρ_p, and ρ_u are defined as

$$\rho_y(\chi)(\cdot) = \tilde{\mathcal{L}}_p'(\chi) = -e(u, y)(\cdot), \tag{6.14}$$

$$\rho_p(\chi)(\cdot) = \tilde{\mathcal{L}}_y'(\chi) = J'(y)(\cdot) - e_y'(u, y)(\cdot, p) \tag{6.15}$$

$$\rho_u(\chi)(\cdot) = \tilde{\mathcal{L}}_u'(\chi) = (\gamma u + \lambda, \cdot) - e_u'(u, y)(\cdot, p), \tag{6.16}$$

for any $\chi = (u, \lambda, y, p) \in U \times U \times V \times V$.

Proof. By construction (see (6.11)), we have

$$\begin{aligned} j_\gamma(\bar{u}) - j_{\gamma,h}(\bar{u}_h) &= f_\gamma(\bar{u}) + \frac{\gamma}{2}\|\bar{u}\|^2 + \psi(\bar{u}) - f_{\gamma,h}(\bar{u}_h) - \frac{\gamma}{2}\|\bar{u}_h\|_h^2 - \psi_h(\bar{u}_h) \\ &= \mathcal{L}(\bar{u}, \bar{\lambda}, \bar{y}, \bar{p}) - \mathcal{L}_h(\bar{u}_h, \bar{\lambda}_h, \bar{y}_h, \bar{p}_h) = \mathcal{L}(\bar{\chi}) - \mathcal{L}_h(\bar{\chi}_h) = \mathcal{L}(\bar{\chi}_h) - \mathcal{L}_h(\bar{\chi}_h) + \mathcal{L}(\bar{\chi}) - \mathcal{L}(\bar{\chi}_h) \end{aligned}$$

Since $\mathcal{L}$ is smooth, we can apply the usual trick and rewrite the last term as an integral over the derivative, which is then evaluated with the trapezoidal rule to obtain

$$\tilde{\mathcal{L}}(\bar{\chi}) - \tilde{\mathcal{L}}(\bar{\chi}_h) = \int_0^1 \tilde{\mathcal{L}}'(t\bar{\chi}_h + (1-t)\bar{\chi})(\bar{\chi} - \bar{\chi}_h)\,\mathrm{d}t = \frac{1}{2}\left[\tilde{\mathcal{L}}'(\bar{\chi})(\bar{\chi} - \bar{\chi}_h) + \tilde{\mathcal{L}}'(\bar{\chi}_h)(\bar{\chi} - \bar{\chi}_h)\right].$$

Note that, since $\tilde{\mathcal{L}}'(\cdot)(\chi - \chi_h)$ is linear, the evaluation with the trapezoidal rule is exact. The result now follows by computing the partial derivatives of $\tilde{\mathcal{L}}$. It holds

$$\tilde{\mathcal{L}}'(\bar{\chi})(\bar{\chi} - \bar{\chi}_h) = \rho_y(\bar{\chi})(\bar{p} - \bar{p}_h) + \rho_p(\bar{\chi})(\bar{y} - \bar{y}_h) + \rho_u(\bar{\chi})(\bar{u} - \bar{u}_h) + (\bar{u}, \bar{\lambda} - \bar{\lambda}_h) = (\bar{u}, \bar{\lambda} - \bar{\lambda}_h),$$

since the first three terms vanish for the optimal solution $\bar{\chi}$. Similarly, we have

$$\tilde{\mathcal{L}}'(\bar{\chi}_h)(\bar{\chi} - \bar{\chi}_h) = \rho_y(\bar{\chi}_h)(\bar{p} - \bar{p}_h) + \rho_p(\bar{\chi}_h)(\bar{y} - \bar{y}_h) + \rho_u(\bar{\chi}_h)(\bar{u} - \bar{u}_h) + (\bar{u}_h, \bar{\lambda} - \bar{\lambda}_h).$$

This yields the result. □

Let us give an interpretation to these terms. The term

$$\eta_h^{\mathrm{quad}} = \tilde{\mathcal{L}}(\bar{\chi}_h) - \tilde{\mathcal{L}}_h(\bar{\chi}_h)$$

represents a quadrature error. It depends only on computable, discrete quantities, and can be assessed in practice by comparing with a higher order quadrature formula. In the case where we use high order quadrature formulas for the evaluation of the respective quantity, it is neglected in the numerical experiments. In the case where we use mass lumping, we give more details below. The terms

$$\rho_y(\chi_h)(\bar{p} - \bar{p}_h) = (\bar{u}_h, \bar{p} - \bar{p}_h) - a(\bar{y}_h, \bar{p} - \bar{p}_h) \tag{6.17}$$

$$\rho_p(\chi_h)(\bar{y} - \bar{y}_h) = (\chi_{\Omega_o}(\bar{y} - y_d), \bar{y} - \bar{y}_h) - a(\bar{y} - \bar{y}_h, \bar{p}_h) \tag{6.18}$$

are the residual of the state equation, weighted by the error of the adjoint variable, and the adjoint residual, weighted by the error of the state variable. To evaluate this error in practice, we will replace the exact variables $(\bar{y}, \bar{p})$ by (locally) higher order reconstructions; see section 6.3.2. The term

$$\rho_u(\bar{\chi})(\bar{u} - \bar{u}_h) = (\gamma \bar{u}_h + \bar{\lambda}_h, \bar{u} - \bar{u}_h) + (\bar{p}_h, \bar{u} - \bar{u}_h) = (\gamma \bar{q}_h + \bar{p}_h, \bar{u} - \bar{u}_h) \tag{6.19}$$

is an L^2-projection error between $\gamma \bar{q}_h \in U_h$ and $-\chi_{\Omega_c} \bar{p}_h \in \{\, \chi_{\Omega_c} v_h \mid v_h \in V_h \,\}$, weighted by the error of the control variable. If Ω_c is represented exactly by the triangulation, this error is zero in the case of bilinear controls (i.e., $U_h = U_h^1$). In the other cases, we will replace the exact variable $\bar{u}$ again by a (locally) higher order reconstruction; see section 6.3.2. Finally, the term

$$(\bar{u} + \bar{u}_h, \bar{\lambda} - \bar{\lambda}_h) \tag{6.20}$$

can be interpreted as a complementarity error. For instance, it is zero when the subgradients $\bar{\lambda}$ and $\bar{\lambda}_h$ coincide (to $\pm\alpha$) on the combined support of $\bar{u}$ and $\bar{u}_h$. We will localize this term with an approach based on a reconstruction the exact adjoint state and the pointwise formula for the proximal map P_γ.

6.3.2. Error representation

To obtain a computable error, we have to replace the continuous variables $(\bar{u}, \bar{y}, \bar{p})$ appearing in (6.14)-(6.16) with computable quantities. For $\bar{y}$ and $\bar{p}$, we use an established (heuristic) approach and replace them with locally higher order reconstructions of the computed solutions $\bar{y}_h$ and $\bar{p}_h$; see [BR01, Section 5]. To this purpose, we define the patch-wise interpolation operator

$$i_{2h} \colon V_h \to V_{2h}^2,$$

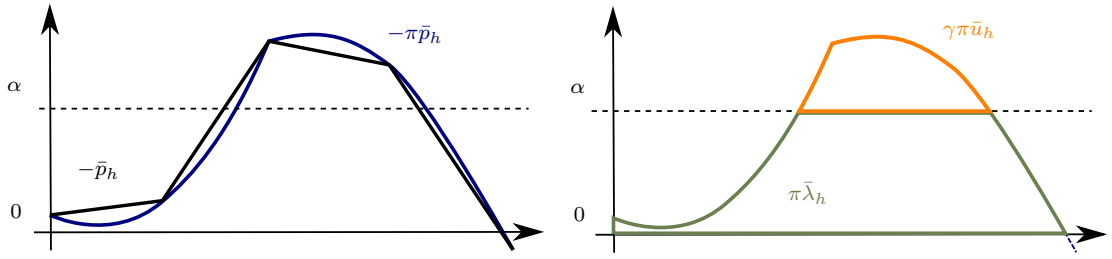


Figure 6.2.: Higher order reconstructions $\pi\bar{p}_h$, $\pi\bar{u}_h$ and $\pi\bar{\lambda}_h$

which interpolates the function v_h on each patch (a cell of the globally coarsened mesh $\mathcal{T}_{2h}$) by a biquadratic (triquadratic) function $i_{2h}v_h \in V_{2h}^2$. The space V_{2h}^2 is defined similarly to V_h on the coarse triangulation $\mathcal{T}_{2h}$ with the biquadratic (triquadratic) functions $\mathcal{Q}_2(\hat{K})$ on the reference cell. Then we replace $(\bar{y}, \bar{p})$ with $(\pi\bar{y}_h, \pi\bar{p}_h) = (i_{2h}\bar{y}_h, i_{2h}\bar{p}_h)$ for the evaluation of the weights in (6.14) and (6.15); see Becker and Rannacher [BR96] for more details on this approach. Consequently, we define

$$\eta_h^y = \rho_y(\bar{\chi}_h)(i_{2h}\bar{p}_h - \bar{p}_h),$$
$$\eta_h^p = \rho_p(\bar{\chi}_h)(i_{2h}\bar{y}_h - \bar{y}_h).$$

To obtain a locally higher order reconstruction for the optimal control $\bar{u}$ and the subgradient $\hat{\lambda}$, we follow Vexler and Wollner [VW08] and evaluate the continuous proximal map P_γ for the semidiscrete variable $\pi\bar{q}_h = -1/\gamma\, \chi_{\Omega_c}\pi\bar{p}_h \in L^2(\Omega_c)$. We set

$$\pi u_h = P_\gamma(\pi\bar{q}_h), \quad \text{and } \pi\bar{\lambda}_h = \gamma\left(\pi\bar{q}_h - \pi\bar{u}_h\right).$$

As before, we replace $(\bar{u}, \bar{\lambda})$ by $(\pi\bar{u}_h, \pi\bar{\lambda}_h)$ in the weight for (6.15) and the complementarity term. We define the corresponding estimators as

$$\eta_h^u = \rho_u(\bar{\chi}_h)(\pi\bar{u}_h - \bar{u}_h),$$
$$\eta_h^\lambda = (\pi\bar{u}_h + \bar{u}_h, \pi\bar{\lambda}_h - \bar{\lambda}_h).$$

Note, that these variables are generally not finite element functions and their numerical treatment requires some care. We will use summatory subdivided quadrature formulas, which only require a pointwise evaluation of these quantities. This is possible, since P_γ possesses a pointwise representation. A one-dimensional visualization of this approach is given in Figure 6.2.

Then, the error contributions from Proposition 6.2 are localized to the cells of the triangulation $\mathcal{T}_h$. The standard approach for ρ_u and ρ_z is a cell-wise integration by parts; see, e.g., [BR01]. We will use the filtering approach by Braack and Ern [BE03] in the numerical experiments. The error contributions for ρ_u and the complementarity term are localized directly, since they do not involve partial derivatives. To justify the localization procedure, it is usually argued that the discrete solutions appearing in the weights of (6.14)–(6.16) can be replaced by an interpolation of the continuous solution. Then, Galerkin orthogonality is applied in the residuals, which allows to replace the discrete optimal solution in the weight by an arbitrary discrete variable. Due to the nonsmoothness of ψ, we can not do this in general; cf. [VW08]. We will further elaborate in the concrete settings for the control space U_h; cf. also section 4.5.3.

In the following, for clarity of presentation, we neglect the errors from quadrature in a and J, and the errors due to the approximation of Ω_c by the mesh. Therefore, we make the following assumption.

Assumption 6.1. We assume that $a_h \equiv a$, $J_h \equiv J$, and make a compatibility assumption on Ω_c and the mesh as in chapter 4 and chapter 5.

The otherwise resulting errors are not directly related to the discussion. Furthermore, in the numerical experiments these assumptions are always fulfilled.

Piecewise constant controls

We discuss the case of piecewise constant controls $U_h = U_h^0$. We give the specialization of Proposition 6.2.

Proposition 6.3. *With Assumption 6.1, we obtain for the optimal variables* (6.12) *that*

$$\begin{aligned} &j_\gamma(\bar{u}) - j_{\gamma,h}(\bar{u}_h) \\ &\qquad = \frac{1}{2}\left[\rho_y(\bar{\chi}_h)(\bar{p} - p_h) + \rho_p(\bar{\chi}_h)(\bar{y} - y_h) + \rho_u(\bar{\chi}_h)(\bar{u} - u_h) + (\bar{u}_h + \bar{u}, \bar{\lambda} - \bar{\lambda}_h)\right], \end{aligned}$$

where $p_h \in V_h$, $y_h \in V_h$ and $u_h \in U_h^0$ are arbitrary. The residuals ρ_y, ρ_p, and ρ_u are defined as in Proposition 6.2.

Proof. With Assumption 6.1, all integrals in the Lagrange function can be evaluated exactly and we have $\tilde{\mathcal{L}}(\bar{\chi}_h) - \tilde{\mathcal{L}}_h(\bar{\chi}_h) = 0$. With Galerkin orthogonality for the state and adjoint equation, we have $\rho_y(\bar{\chi}_h)(\varphi_h) = \rho_p(\bar{\chi}_h)(\varphi_h) = 0$ for any $\varphi_h \in V_h$. The equality

$$\rho_u(\chi_h)(\varphi_h) = (\gamma\bar{u}_h + \bar{\lambda}_h, \varphi_h) + (\bar{p}_h, \varphi_h) = (\gamma\bar{q}_h + \bar{p}_h, \varphi_h) = 0.$$

for all $\varphi_h \in U_h^0$ follows by definition of $\bar{\lambda}_h$ and $\bar{q}_h$ and the fact that $(\cdot,\cdot)_h \equiv (\cdot,\cdot)$ due to Assumption 6.1. □

As a corollary, we can insert $(u_h, y_h, p_h) = (P_h^0\bar{u}, i_h\bar{y}, i_h\bar{p})$ in Proposition 6.3 (where i_h is the nodal interpolation and P_h^0 is the L^2-projection onto U_h^0), and give a partial justification for the localization of the first three terms to cell-wise contributions. For the last term, a rigorous justification is still missing. In the numerical experiments we usually observe that this term small (of higher order) when compared to the estimate of the weighted L^2-projection error

$$\rho_u(\chi_h)(\bar{u} - i_h\bar{u}) = (\bar{p}_h - \gamma\bar{q}_h, \bar{u} - P_h^0\bar{u}).$$

Piecewise linear controls with consistent mass

Now, we turn to the case of piecewise linear controls $U_h = U_h^1$. First, we discuss the case without mass lumping, i.e., where it holds $(\cdot,\cdot)_h \equiv (\cdot,\cdot)$ with Assumption 6.1. As mentioned before, we have $\bar{q}_h = -1/\gamma\,\chi_{\Omega_c}\bar{p}_h \in U_h^1$ under the compatibility assumption on Ω_c and the triangulation. However, due to the nondiagonal mass matrix of $(\cdot,\cdot)$ on the space U_h^1, the optimal solution $\bar{u}_h = P_{\gamma,h}(\bar{q}_h)$ does not have a closed form solution; cf. section 4.5.3. A similar remark applies for the subgradient $\bar{\lambda}_h = \gamma(\bar{q}_h - \bar{u}_h)$. Nevertheless, we obtain the following result.

Proposition 6.4. *With Assumption 6.1, we obtain for the optimal variables* (6.12) *that*

$$j_\gamma(\bar{u}) - j_{\gamma,h}(\bar{u}_h) = \frac{1}{2}\left[\rho_y(\bar{\chi}_h)(\bar{p} - p_h) + \rho_p(\bar{\chi}_h)(\bar{y} - y_h) + (\bar{u}_h + \bar{u}, \bar{\lambda} - \bar{\lambda}_h)\right],$$

where $p_h \in V_h$ and $y_h \in V_h$ are arbitrary. The residuals ρ_y, ρ_p, and ρ_u are defined as in Proposition 6.2.

Proof. As before, with Assumption 6.1, all integrals in the Lagrange function are evaluated exactly and we have $\tilde{\mathcal{L}}(\bar{\chi}_h) - \tilde{\mathcal{L}}_h(\bar{\chi}_h) = 0$. With Galerkin orthogonality for the state and adjoint equation, we have $\rho_y(\bar{\chi}_h)(\varphi_h) = \rho_p(\bar{\chi}_h)(\varphi_h) = 0$ for any $\varphi_h \in V_h$. Therefore it follows that

$$\rho_u(\chi_h)(\varphi) = (\gamma\bar{u}_h + \bar{\lambda}_h, \varphi) + (\bar{p}_h, \varphi) = (\gamma\bar{q}_h + \bar{p}_h, \varphi) = 0$$

for all $\varphi \in U$, and the residual vanishes. □

As before, we can insert $(y_h, p_h) = (i_h\bar{y}, i_h\bar{p}_h)$ in Proposition 6.4, and give a partial justification for the localization of the first two terms to cell-wise contributions. For the last term, we are unable to give a corresponding justification. In the numerical experiments we observe good effectivity values with the local reconstruction procedure with $(\pi\bar{u}_h, \pi\bar{\lambda}_h)$ also in the cases where the complementarity term gives a significant contribution; see below.

Piecewise linear controls with mass lumping

Now, we discuss the case $U_h = U_h^1$ with mass lumping, i.e., where the terms $(\cdot,\cdot)_h$, ψ_h, and $\|\cdot\|_h^2$ are evaluated with the trapezoidal rule on each cell; cf. section 4.5.3.

Proposition 6.5. *With Assumption 6.1 and mass lumping, we obtain for the optimal variables* (6.12) *that*

$$\begin{aligned} j_\gamma(\bar{u}) - j_{\gamma,h}(\bar{u}_h) = \frac{\gamma}{2}\left[\|\bar{u}_h\|_h^2 - \|\bar{u}_h\|^2\right] \\ + \frac{1}{2}\left[\rho_y(\bar{\chi}_h)(\bar{p} - i_h\bar{p}) + \rho_p(\bar{\chi}_h)(\bar{y} - y_h) + (\bar{u}_h + \bar{u}, \bar{\lambda} - \bar{\lambda}_h)\right] + \mathcal{R}, \end{aligned}$$

where $y_h \in V_h$ is arbitrary and $i_h\bar{p} \in V_h$ is the nodal interpolation of $\bar{p}$. The residuals ρ_y, ρ_p, and ρ_u are defined as in Proposition 6.2, and the remainder $\mathcal{R}$ is given by the quadrature error

$$\mathcal{R} = (\bar{u}_h, i_h\bar{p} - \bar{p}_h) - (\bar{u}_h, i_h\bar{p} - \bar{p}_h)_h.$$

Proof. With Assumption 6.1, all integrals in the Lagrange function with exception of the control terms are evaluated exactly. As mentioned before, it holds

$$\bar{\lambda}_h = \gamma(\bar{q}_h - \bar{u}_h) = -\chi_{\Omega_c}\bar{p}_h - \gamma\bar{u}_h \in U_h^1,$$

due to the compatibility of Ω_c and the triangulation. Therefore, we obtain

$$\tilde{\mathcal{L}}(\bar{\chi}_h) - \tilde{\mathcal{L}}_h(\bar{\chi}_h) = (\bar{\lambda}_h + \bar{p}_h, \bar{u}_h) + \frac{\gamma}{2}\|u_h\|^2 - (\bar{\lambda}_h + \bar{p}_h, \bar{u}_h)_h - \frac{\gamma}{2}\|u_h\|_h^2 = \frac{\gamma}{2}\|\bar{u}_h\|_h^2 - \frac{\gamma}{2}\|\bar{u}_h\|^2$$

with $\bar{\lambda}_h + \chi_{\Omega_c}\bar{p}_h = -\gamma\bar{u}_h$. With Galerkin orthogonality for the adjoint equation, we have $\rho_p(\bar{\chi}_h)(\varphi_h) = 0$ for any $\varphi_h \in V_h$. For the state residual, we can not apply Galerkin orthogonality. Here, the discrete state equation is given by

$$a(\bar{y}_h, \varphi_h) - (\bar{u}_h, \varphi_h)_h \quad \text{for all } \varphi_h \in V_h.$$

Therefore, we compute

$$\rho_y(\bar{\chi}_h)(\varphi_h) = a(\bar{y}_h, \varphi_h) - (\bar{u}_h, \varphi_h) = (\bar{u}_h, \varphi_h)_h - (\bar{u}_h, \varphi_h) \quad \text{for all } \varphi_h \in V_h.$$

Inserting $\varphi_h = i_h\bar{p} - \bar{p}_h$ gives the form above. The control residual vanishes due to $\bar{q}_h = -1/\gamma\, \chi_{\Omega_c}\bar{p}_h \in U_h^1$, as in Proposition (6.4). □

Motivated by Lemma 4.31, where an estimate for the quadrature error due to mass lumping of the form

$$|(\bar{u}_h, i_h\bar{p} - \bar{p}_h) - (\bar{u}_h, i_h\bar{p} - \bar{p}_h)_h| \leq Ch^2 \|\nabla \bar{u}_h\|_{L^2(\Omega_c)} \|\nabla(i_h\bar{p} - \bar{p}_h)\|_{L^2(\Omega_c)}$$

was derived (in the case of linear finite elements on a shape-regular triangulation), we decide to neglect the remainder term $|\mathcal{R}|$. By the a priori analysis in section 4.5.4, we can expect the error $\|i_h\bar{p} - \bar{p}_h\|_{H^1(\Omega)}$ to be of the order $\mathcal{O}(h)$, at least on quasi-uniform meshes. Under such an assumption, the remainder term is of the order $\mathcal{O}(h^3)$ (for fixed $\gamma > 0$) and therefore negligible.

6.4. Adaptive strategy

We base the refinement strategy upon the following representation for the combined discretization and regularization error:

$$j(\bar{u}) - j_{\gamma,h}(\bar{u}_{\gamma,h}) = j(\bar{u}) - j_\gamma(\bar{u}_\gamma) + j_\gamma(\bar{u}_\gamma) - j_{\gamma,h}(\bar{u}_{\gamma,h}) \approx \eta_\gamma + \eta_h.$$

The algorithm will solve the optimization problem (6.10) for given initial γ on an initial mesh, evaluate the indicators given by

$$\begin{aligned} \eta_\gamma &= \eta_\gamma^{\text{triv}} / s_{\text{est}}, \\ \eta_h &= \eta_h^{\text{quad}} + \eta_h^y + \eta_h^p + \eta_h^u + \eta_h^\lambda, \end{aligned}$$

as defined above. The adaptive strategy is based on an equilibration of both error terms: we try to keep a balance, such that

$$|\eta_\gamma| \approx c_{\text{equi}}\, |\eta_h|$$

is fulfilled throughout the iterations. Here, $c_{\text{equi}} \geq 1$ is a chosen equilibration factor. The introduction of this factor is motivated on the one hand by the corresponding theory, where we estimate the regularization error on the continuous level, and on the other hand by numerical experience; see below. In the numerical examples we use the following simple strategy: if $|\eta_\gamma| > 2\, c_{\text{equi}}\, |\eta_h|$, we decrease γ (by a fixed factor), if $c_{\text{equi}}\, |\eta_h| > 2\, \eta_\gamma$, we refine the discretization. Otherwise, if none of these conditions is fulfilled, we do both steps. The refinement of the discretization is based upon the localization of the error indicators as discussed above. For the selection of the relevant cells, different strategies are possible such as the *fixed-fraction* or the *fixed-error* strategy. We will use the optimization approach by Braack [Bra98, Section 4.4.2].

6.5. Numerical results

In this section we give some numerical evidence for the effectivity of the outlined estimation strategy. We set up two model configurations; one with global control and observation and one with control and observation on disjoint subdomains. We compute the effectivity indices of the discretization error estimator for a sequence of regularization parameters and compare the relative size of the error contributions due to control and state discretization. We asses the quality of the regularization error estimate on a sequence of adaptively generated meshes. The improved practical performance due to adaptivity is demonstrated by comparing with uniform mesh refinement. We mention that the corresponding algorithm is implemented in the PDE-optimization library RoDoBo [RoD], using the underlying finite element toolbox Gascoigne [Gas].

The test problem

We consider the linear quadratic optimal control problem from chapter 4. The weak form e is in this case given by

$$e(u,y)(p) = a(y,p) - (u,p) = (\nabla y, \nabla p) - (u,p),$$

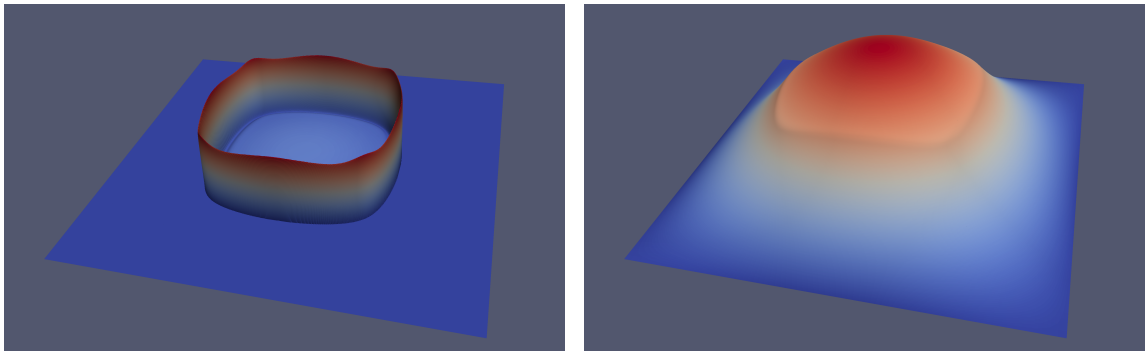
and we have $V = H_0^1(\Omega)$. The domain for this test example is chosen as the unit square $\Omega = (0,1)^2 \subset \mathbb{R}^2$. As discussed above, the objective functional consists of $\psi(u) = \alpha\|u\|_{\mathcal{M}(\Omega_c)}$ and the quadratic tracking J given by

$$J(y) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega_o)}$$

for a desired state $y_d \in L^2(\Omega_o)$. We consider two configurations:

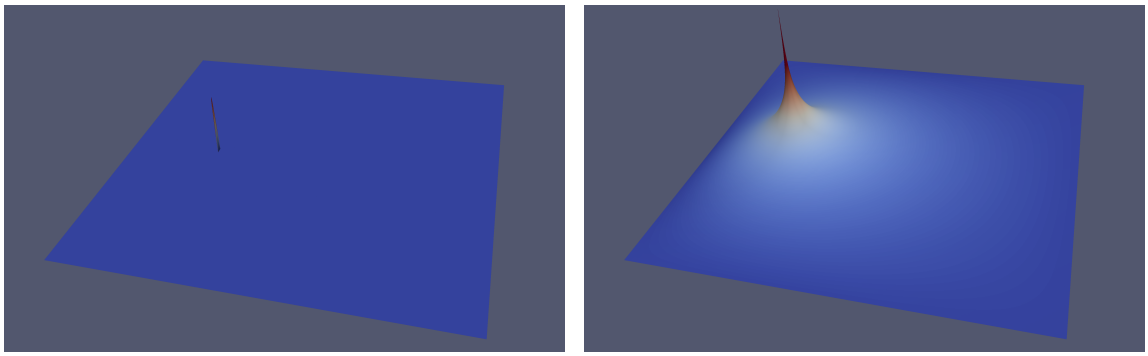
1. *Global control and observation:* We consider $\Omega_c = \Omega_o = \Omega = (0,1)^2$ with the desired state
$$y_d(x) = \frac{1}{\sigma}\exp\left(-\frac{|x - x_c|^2}{2\sigma^2}\right)$$
where $x_c = (1/2, 1/2)$ is the center and $\sigma = 0.3$. The cost parameter is set to $\alpha = 0.01$.
2. *Disjoint control and observation:* We consider a control domain in the left quarter of Ω and an observation domain in the right quarter, i.e., we set
$$\Omega_c = \{\, x = (x_1, x_2) \in \Omega \mid x_1 < 1/4 \,\},$$
$$\Omega_o = \{\, x = (x_1, x_2) \in \Omega \mid x_1 > 3/4 \,\}.$$
The desired state is chosen as $y_d(x) = \sin(\pi x_1)\sin(\pi x_2)^3$, and the cost parameter is set to $\alpha = 0.0001$.

By the improved regularity result from section 4.4.2, we know that the optimal solution to the first problem is an element of $H^{-1}(\Omega)$. In fact, by the numerical results we observe that the solution appears to be a line-measure on a smooth, closed curve with an even more regular, distributed L^2-part in the interior; see Figure 6.3. With the second example, we want to investigate also the case where the optimal control is a point source (which is possible in the former configuration only for desired states with singularities; see section 4.6). For this problem, the numerical results indicate that the optimal control is given by a Dirac delta function in



(a) control $\bar{u}_{\gamma,h}$ ($\gamma = 10^{-9}$) (b) state $\bar{y}_{\gamma,h}$ ($\gamma = 10^{-9}$)

Figure 6.3.: Numerical results for example 1.



(a) control $\bar{u}_{\gamma,h}$ ($\gamma \approx 10^{-15}$) (b) state $\bar{y}_{\gamma,h}$ ($\gamma \approx 10^{-18}$)

Figure 6.4.: Numerical results for example 2.

the point $x = (1/4, 1/2)$; see Figure 6.4. The underlying initial mesh is chosen as the twice globally refined unit cube (with 16 initial cells). For both of these examples, the bilinear form a can always be evaluated exactly and Ω_c and Ω_o are compatible with the mesh (on all possible refinements). The tracking term is evaluated with a Gaussian quadrature rule; therefore the corresponding (neglected) error term is of fourth order, since y_d is smooth.

Discretization error

Now, we give some numerical evidence for the accuracy of the proposed estimates for the discretization error. We compute the effectivity indices

$$I_{\text{eff}}^h = \frac{j(\bar{u}_\gamma) - j_{\gamma,h}(\bar{u}_{\gamma,h})}{\eta_h}$$

for fixed $\gamma \in \{ 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-11} \}$. Since no exact value is available, we compare with a reference value on a fine mesh. Since we noticed in the numerical experiments that the estimators η_h^y and η_h^p dominate the total error for example 1, we generally give the results for example 2 in the following, since this allows for a better assessment of the quality of the error estimators for the control discretization.

The values for the approach based on mass lumping are given in Table 6.1. We observe that the effectivity indices converge to one for fixed γ with increasing fineness of the discretization. The indices deteriorate for decreasing γ on a fixed mesh, which is to be expected. The indices for

#ref	$\gamma = 10^{-3}$	$\gamma = 10^{-5}$	$\gamma = 10^{-7}$	$\gamma = 10^{-9}$	$\gamma = 10^{-11}$
2	1.01	1.21	1.65	-0.49	-0.05
4	1.07	1.12	1.11	1.74	-0.29
6	1.01	1.07	1.05	1.25	-11.79
8	1.00	1.02	1.03	1.38	1.16
10	–	1.01	1.01	1.12	1.18
12	–	1.01	1.01	1.03	1.10

Table 6.1.: Effectivity indices for example 2 with mass lumping for fixed $\gamma > 0$.

the consistent discretization with piece-wise bilinears are given in Table 6.2. For completeness,

#ref	$\gamma = 10^{-3}$	$\gamma = 10^{-5}$	$\gamma = 10^{-7}$	$\gamma = 10^{-9}$	$\gamma = 10^{-11}$
2	3.23	0.69	0.88	0.65	0.02
4	1.00	0.91	0.90	0.95	0.47
6	1.18	1.04	1.06	1.01	1.02
8	2.00	1.03	1.04	1.07	1.01
10	1.12	1.02	0.95	0.98	1.01
12	1.02	1.01	0.97	1.00	1.01

Table 6.2.: Effectivity indices for example 2 with consistent mass for fixed $\gamma > 0$.

we also give the indices for piecewise constant discretization in Table 6.3.

#ref	$\gamma = 10^{-3}$	$\gamma = 10^{-5}$	$\gamma = 10^{-7}$	$\gamma = 10^{-9}$	$\gamma = 10^{-11}$
2	3.78	0.81	0.94	0.40	0.01
4	1.05	0.99	0.98	0.98	0.22
6	0.99	1.00	1.01	0.99	0.94
8	0.99	1.00	1.00	0.99	1.02
10	1.01	1.00	1.00	1.00	1.01
12	1.02	1.00	1.00	1.00	1.01

Table 6.3.: Effectivity indices for example 2 with piecewise constants for fixed $\gamma > 0$.

Regularization error and adaptive results

To assess the quality of the regularization error estimate, we run the adaptive algorithm for example 1 with an equilibration factor $c_{\text{equi}} = 10$. Thereby, we try to keep the discretization error one order of magnitude smaller than the regularization error to ensure that the numerical solution is sufficiently close to the continuous one (as in the theory). The results for the effectivity indices defined as

$$\widetilde{I_{\text{eff}}^{\gamma}} = \frac{j(\bar{u}) - j_{\gamma,h}(\bar{u}_{\gamma,h})}{\eta_\gamma} \quad \text{and} \quad I_{\text{eff}} = \frac{j(\bar{u}) - j_{\gamma,h}(\bar{u}_{\gamma,h})}{\eta_\gamma + \eta_h}$$

are given in table 6.4 (for a selection of generated refinement levels). The given values are computed with piece-wise bilinear discretization with mass lumping. The results for the other discretization concepts are similar. We observe that the estimated rate of convergence s_{est} tends

#ref	N_{dof}	γ	s_{est}	η_γ	η_h	$\widetilde{I_{\text{eff}}}^{\gamma}$	I_{eff}
4	447	$3.2{\cdot}10^{-6}$	0.97	$-2.1{\cdot}10^{-3}$	$3.6{\cdot}10^{-4}$	0.83	1.00
6	2345	$3.2{\cdot}10^{-7}$	0.90	$-2.8{\cdot}10^{-4}$	$6.7{\cdot}10^{-5}$	0.80	1.04
8	7337	10^{-7}	0.88	$-1.0{\cdot}10^{-4}$	$1.9{\cdot}10^{-5}$	0.85	1.04
10	17649	10^{-8}	0.84	$-1.5{\cdot}10^{-5}$	$5.5{\cdot}10^{-6}$	0.52	0.81
12	57821	$3.2{\cdot}10^{-9}$	0.82	$-6.1{\cdot}10^{-6}$	$1.9{\cdot}10^{-6}$	0.68	1.00
14	194537	10^{-9}	0.81	$-2.4{\cdot}10^{-6}$	$-1.9{\cdot}10^{-6}$	0.79	0.44
16	513963	10^{-9}	0.81	$-2.4{\cdot}10^{-6}$	$1.8{\cdot}10^{-7}$	0.92	1.00

Table 6.4.: Numerical results for example 1 ($c_{\text{equi}} = 10$).

to a value of approximately 0.8 for increasing refinement level. The experimental effectivity $\widetilde{I_{\text{eff}}}^{\gamma}$ is in all cases in the interval $(0.5, 1]$; i.e., we overestimate the error slightly. However, in most cases this is explained by the discretization error, as the values of I_{eff} suggest.

Now, we give the results of the adaptive computation for example 1 and example 2. We set an equilibration factor of $c_{\text{equi}} = 2$; the corresponding results are given in Table 6.5. We

#ref	γ	s_{est}	η_γ	η_h	I_{eff}
2	10^{-6}	0.97	$-6.5{\cdot}10^{-4}$	$2.0{\cdot}10^{-3}$	0.53
4	$3.2{\cdot}10^{-7}$	0.93	$-2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$	7.00
6	10^{-7}	0.85	$-0.1{\cdot}10^{-4}$	$5.3{\cdot}10^{-5}$	1.05
8	10^{-8}	0.83	$-1.5{\cdot}10^{-5}$	$8.5{\cdot}10^{-6}$	0.95
10	$3.2{\cdot}10^{-9}$	0.81	$-6.0{\cdot}10^{-6}$	$2.2{\cdot}10^{-6}$	1.00
12	$3.2{\cdot}10^{-10}$	0.80	$-9.7{\cdot}10^{-7}$	$1.3{\cdot}10^{-6}$	3.09
13	10^{-10}	0.80	$-9.7{\cdot}10^{-7}$	$9.8{\cdot}10^{-7}$	1.54

(a) Results for example 1.

#ref	γ	s_{est}	η_γ	η_h	I_{eff}
4	10^{-10}	0.72	$-1.5{\cdot}10^{-5}$	$-1.9{\cdot}10^{-5}$	0.53
6	10^{-11}	0.67	$-8.7{\cdot}10^{-6}$	$-8.9{\cdot}10^{-7}$	0.95
8	10^{-13}	0.65	$-1.0{\cdot}10^{-6}$	$-1.1{\cdot}10^{-6}$	0.50
10	10^{-14}	0.65	$-5.6{\cdot}10^{-7}$	$-1.0{\cdot}10^{-7}$	0.95
12	10^{-16}	0.67	$-6.4{\cdot}10^{-8}$	$-7.5{\cdot}10^{-8}$	0.56
14	10^{-17}	0.66	$-3.5{\cdot}10^{-8}$	$-1.4{\cdot}10^{-9}$	0.99
17	10^{-19}	0.67	$-6.4{\cdot}10^{-9}$	$1.1{\cdot}10^{-9}$	1.42

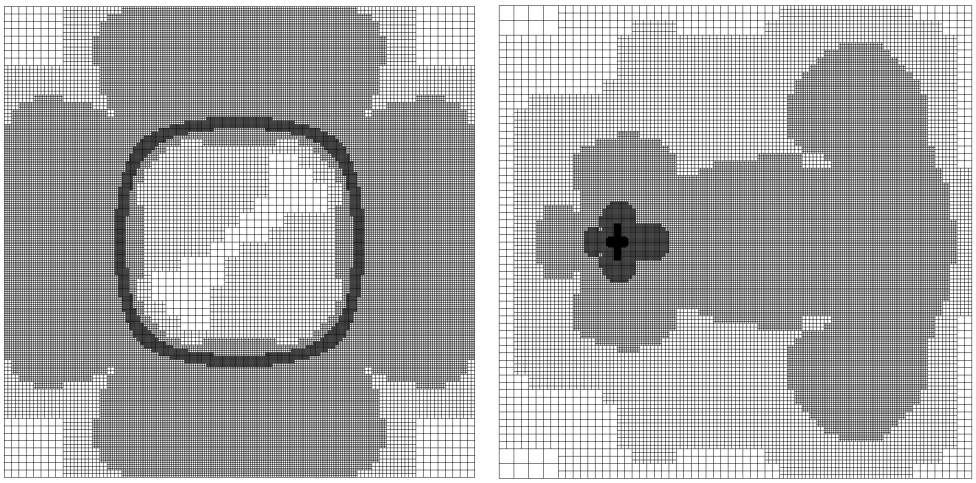
(b) Results for example 2.

Table 6.5.: Results of the adaptive algorithm ($c_{\text{equi}} = 2$).

observe that the efficiency indices are reasonably close to one for example 2. In the case of example 1, there are some notable outliers. However, in these cases the estimators η_h and η_γ are of the same magnitude and have opposing sign, which suggests a cancellation effect. If we set c_{equi} to a larger value, the efficiency values improve (as in Table 6.4). Note, that for example 1 the estimated rate of convergence is $s_{\text{est}} \approx 0.8$, and for example 2 it is $s_{\text{est}} \approx 0.67$ (on average throughout the iterations). Therefore we can conclude that also the simple estimator

$\eta_\gamma^{\mathrm{triv}} = -\gamma/2 \, \|\bar{u}_\gamma\|^2$ from section 6.2 would have yielded good results for these two problems: the quality of the effectivity indices would be decreased but the outcome of the numerical computation would be similar to a run with $\eta_\gamma^{\mathrm{mod}} = \eta_\gamma^{\mathrm{triv}}/s_{\mathrm{est}}$ and the scaled equilibration constant $c_{\mathrm{equi}}/s_{\mathrm{est}}$.

In Figure 6.5 we depict the generated meshes on a moderate refinement level. As expected, we observe local refinement especially in the areas where the optimal solution has singularities. This effect is particularly prominent for example 2, where the optimal control is a point source. Finally, we compare the accuracy that can be achieved with uniform and with local refinement,



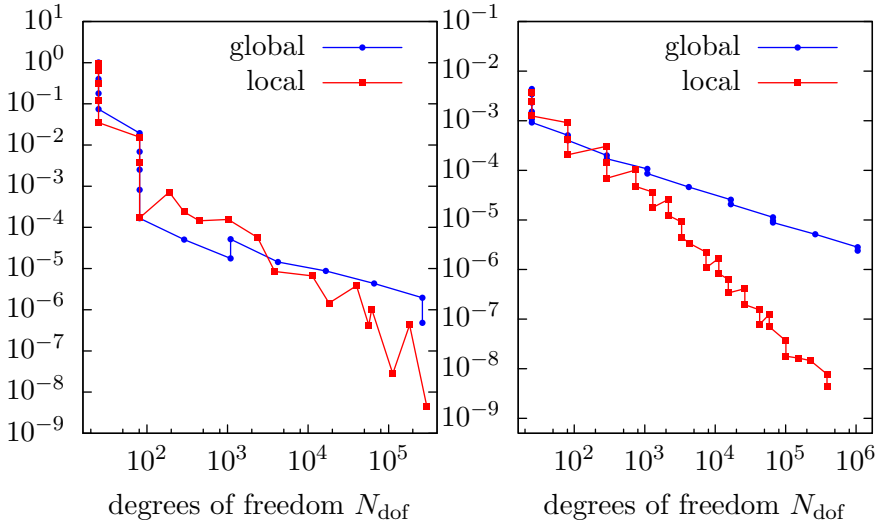
(a) Mesh for example 1. (b) Mesh for example 2.

Figure 6.5.: Generated meshes on refinement level 11.

by contrasting the results from Table 6.5 with the analogous results for uniform refinement. To this purpose, we modify the algorithm from section 6.4 and replace the local refinement step with a global refinement step. We use the consistent piece-wise bilinear discretization for the control (which generally seems to result in the smallest discretization error for the control) and a value of $c_{\mathrm{equi}} = 1$. An approximate description can be given as follows: we stop decreasing the regularization parameter γ as soon as the estimated regularization error is below the estimated discretization error and perform a global mesh refinement. We plot the achieved accuracy in the functional $|j(\bar{u}) - j_{\gamma,h}(\bar{u}_{\gamma,h})|$ against the number of degrees of freedom N_{dof} in Figure 6.6. We observe that the adaptive algorithm achieves a significantly higher accuracy with an equal number of discretization points.

6.6. Comparison with a nodal Dirac discretization and outlook

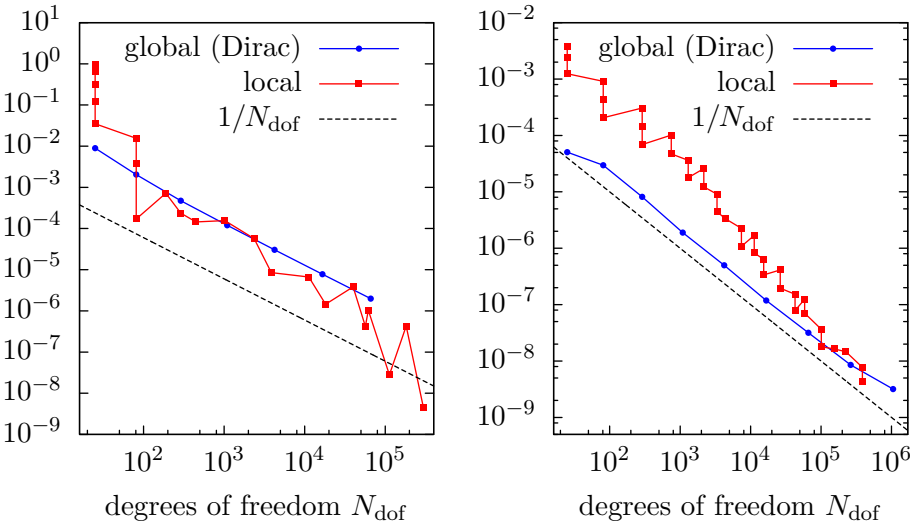
The comparison between global and local mesh refinement seemed to clearly favor the local mesh refinement. However, from the a priori analysis from section 4, we know that a (variational) discretization of the optimal measure without the introduction of a regularization parameter is able to achieve the optimal convergence rate $\mathcal{O}(h^2)$ (up to a logarithmic factor). Therefore, the comparison of local and global mesh refinement for the point source example depicted in Figure 6.6b is slightly surprising, since the "global" strategy does not achieve this convergence rate. Therefore, we also compare the results achieved with the adaptive algorithm



(a) Comparison for example 1. (b) Comparison for example 2.

Figure 6.6.: Accuracy of the objective functional for uniform and adaptive refinement.

to the discretization concept from section 4.2 on a uniform mesh. The results are given in Figure 6.7. For visual comparison, we also plot the reciprocal of N_{dof}, which in two dimensions is asymptotically equal to a constant times h^2 on a uniform mesh. We see the expected $\mathcal{O}(h^2)$ rate for the uniform discretization, as predicted by the theory in chapter 4. Furthermore, we observe that the local mesh refinement strategy is able to reproduce a similar accuracy after some iterations with a slightly lower amount of degrees of freedom. An explanation for the bad



(a) Comparison for example 1. (b) Comparison for example 2.

Figure 6.7.: Accuracy of the objective functional for uniform refinement with the nodal Dirac discretization and adaptive refinement.

performance of the algorithm based on error equilibration and global mesh refinement (as in

Figure 6.6b) can be given by analyzing the structure of the discretization error. By inspection of the numerical results, we observe that the estimates $\eta_h^\lambda / \eta_h^{\text{quad}} / \eta_h^u$ associated with the control discretization heavily dominate the overall error in the later steps for the "global" run from Figure 6.6b. In contrast, if the nodal Dirac discretization without regularization is employed, there is no control discretization error, as we have seen in section 4.2. This discussion points to a weakness of the presented adaptive strategy: since the control is interpreted as a L^2 function and discretized with corresponding finite elements, the corresponding discretization error has to be resolved by the mesh, and we have to invest additional degrees of freedom. As we can see by comparing Figure 6.6b and Figure 6.7b, this additional effort can be significant. Furthermore, we also observe a very drastic local refinement around the location of the point source in Figure 6.5b, which is not only caused by the singularity of the optimal state, but also to a significant degree by the control discretization error. For such cases, it appears desirable to also derive an adaptive algorithm for the nodal Dirac discretization concept, to possibly avoid some of this additional effort. This is left as a subject for future research.

However, an adaptive strategy that produces a mesh which can represent the optimal control as an L^2 function in each step seems also to be valuable. For instance, this is the case if we are interested in the optimal solution of (6.2) only for a moderately small value of γ or when the optimal solution possess higher regularity, such as in example 1. Note also, that the $\mathcal{O}(h^2)$ rate for the objective functional can not be expected anymore for a global discretization in three spatial dimensions. Therefore, we can already expect the local algorithm to perform significantly better than an approach based on global refinement, even if nodal Dirac delta functions are used for the global discretization and no regularization is employed there.

A. Appendix

In the appendix we provide some necessary auxiliary results, which are mostly well-known but not directly available in the literature and mainly of technical nature.

A.1. Approximation of measures by smooth functions

We verify that Assumption 2.1 from section 2.5 holds for the functionals ψ discussed in section 2.2.3 and section 2.3. As usual, we assume that Ω is open and $\Gamma \subset \partial\Omega$ a relatively closed part of the boundary. We make the additional assumption that $\Omega_c \subset \Omega \cup \Gamma$ is the relative closure in $\Omega \cup \Gamma$ of an open set, i.e.,

$$\Omega_c = \overline{\operatorname{int}(\Omega_c)} \cap (\Omega \cup \Gamma).$$

We verify a fundamental lemma that is formulated for elements of the space of vector measures $\mathcal{M}(\Omega_c, \hat{H})$ and elements of the Hilbert space $L^2(\Omega_c, \hat{H}) = L^2(\operatorname{int}(\Omega_c), \hat{H})$, where $\hat{H}$ is a separable Hilbert space (cf. section 2.3.1).

Proposition A.1. *Let $\bar{u} \in \mathcal{M}(\Omega_c, \hat{H})$. Define ψ as*

$$\psi(u) = \|u\|_{\mathcal{M}(\Omega_c, \hat{H})} = \int_{\Omega_c} \mathrm{d}|u|(x).$$

There exists a sequence of functions $\{ u_n \}_{n\in\mathbb{N}} \subset L^2(\Omega_c, \hat{H})$ with $\psi(u_n) \to \psi(\bar{u})$ and $u_n \rightharpoonup^* \bar{u}$ in $\mathcal{M}(\Omega_c, \hat{H})$ for $n \to \infty$. If $\bar{u}$ is positive, the functions u_n can also be chosen positively.

Proof. For $\hat{H} = \mathbb{R}$ we refer to [Bre11, Problem 24]. For the general case, we give a different, constructive proof based on convolution and duality. We denote by $B_\eta(x)$ the ball of radius $\eta > 0$ around $x \in \Omega_c$ and introduce the characteristic function $\omega_\eta(x, y) = \chi_{B_\eta(x)}(y)$ and the weight $w_\eta(x) = \int_{\Omega_c} \omega_\eta(x, y)\,\mathrm{d}y$ for $x \in \Omega_c$ and $y \in \Omega_c$. It is clear that for each $\eta > 0$ the weight $w_\eta(x)$ is bounded. Since Ω_c is the closure of an open set it also holds $\inf_{x\in\Omega_c} w_\eta(x) > 0$. We define $B_\eta \colon \mathcal{C}(\Omega_c, \hat{H}) \to \mathcal{C}(\Omega_c, \hat{H})$ for $\varphi \in \mathcal{C}(\Omega_c, \hat{H})$ as the average

$$(B_\eta\varphi)(x) = \frac{1}{w_\eta(x)} \int_{\Omega_c} \omega_\eta(x, y)\varphi(y)\,\mathrm{d}y \quad \text{for } x \in \Omega_c.$$

By elementary computations we verify that B_η is well-defined, its operator norm is bounded by one, and that it holds

$$\|B_\eta\varphi - \varphi\|_{\mathcal{C}(\Omega_c, \hat{H})} \to 0 \quad \text{for } \eta \to 0 \quad \text{for all } \varphi \in \mathcal{C}(\Omega_c, \hat{H}).$$

Now, we construct a sequence $u_n \in \mathcal{M}(\Omega_c, \hat{H})$ for $n \in \mathbb{N}$ to approximate $\bar{u}$ by duality as

$$\langle u_n, \varphi \rangle = \langle \bar{u}, B_{1/n}\varphi \rangle \quad \text{for } \varphi \in \mathcal{C}_0(\Omega_c, \hat{H}).$$

By construction, we directly obtain

$$|\langle u_n - \bar{u}, \varphi\rangle| \leq \|\bar{u}\|_{\mathcal{M}(\Omega_c,\hat{H})} \|B_{1/n}\varphi - \varphi\|_{\mathcal{C}(\Omega_c,\hat{H})} \to 0 \quad \text{for } n \to \infty$$

for all $\varphi \in \mathcal{C}_0(\Omega_c, \hat{H})$, which implies $u_n \rightharpoonup^* \bar{u}$ in $\mathcal{M}(\Omega_c, \hat{H})$. Furthermore, it holds

$$\|u_n\|_{\mathcal{M}(\Omega_c,\hat{H})} = \sup_{\|\varphi\|_{\mathcal{C}_0(\Omega_c,\hat{H})}\leq 1} \langle u_n, \varphi\rangle = \sup_{\|\varphi\|_{\mathcal{C}_0(\Omega_c,\hat{H})}\leq 1} \langle \bar{u}, B_{1/n}\varphi\rangle \leq \|\bar{u}\|_{\mathcal{M}(\Omega_c,\hat{H})}$$

Due to the weak lower semicontinuity of the norm, it follows $\psi(u_n) \to \psi(\bar{u})$ for $n \to \infty$. It remains to see that the u_n are actually elements of $L^2(\Omega_c, L^2(I))$. In fact, we have $u_n \in L^\infty(\Omega_c, \hat{H})$ with

$$u_n(y) = \int_{\Omega_c} \frac{\omega_{1/n}(x,y)}{w_{1/n}(x)} \,\mathrm{d}\bar{u}(x) = \int_{\Omega_c} \frac{\omega_{1/n}(x,y)}{w_{1/n}(x)} \bar{u}'(x) \,\mathrm{d}|\bar{u}|(x) \quad \text{for } x \in \Omega_c,$$

which can be verified by applying Fubini's theorem to the expression $\int_{\Omega_c} (u_n(y), \varphi(y))_{\hat{H}} \,\mathrm{d}y$ with the above definition. Since the weight $w_{1/n}$ is bounded from below for fixed $n \in \mathbb{N}$, each u_n is uniformly bounded. It is obvious that u_n is positive if $\bar{u}$ is positive, and we conclude the proof. □

Remark A.1. To extend the previous result also for a weighted integral with continuous weight $\hat{\alpha}$ as in section 2.2.3, we can simply replace the measure u by its weighted version $\tilde{u}$ defined as $\mathrm{d}\tilde{u} = \hat{\alpha}\,\mathrm{d}u$. Then we apply Proposition A.1 to obtain a sequence $\tilde{u}_n$ and set $u_n = \tilde{u}_n/\hat{\alpha}$.

A.2. Auxiliary results

We give the proofs of some elementary results that were needed in Chapter 3.

We define the space H_T as the Hilbert space induced by the inner product derived from a symmetric, positive operator $T\colon H \to H$ defined on a Hilbert space H with $\|T\|_{H\to H} \leq 1$.

Definition A.1. Define the symmetric and positive semi-definite form $(\cdot,\cdot)_T = (\cdot, T\,\cdot)$ and the associated seminorm $\|\cdot\|_T = \sqrt{(\cdot,\cdot)_T}$. The space H_T is given as

$$H_T = \overline{\left({}^{H}\!/_{\operatorname{Ker} T}\right)}^{\|\cdot\|_T},$$

which is the closure of the quotient space ${}^{H}\!/_{\operatorname{Ker} T}$ w.r.t. the T-norm.

Proposition A.2. *The bilinear form $(\cdot,\cdot)_T$, extended in the canonical way to the quotient space ${}^{H}\!/_{\operatorname{Ker} T}$, is symmetric and positive definite. Therefore, $\|\cdot\|_T$ is a norm on ${}^{H}\!/_{\operatorname{Ker} T}$.*

Proof. Symmetry and positivity of $(\cdot,\cdot)_T$ follow from Assumption 3.3. Defining a consistent extension of $(\cdot,\cdot)_T$ to the quotient space ${}^{H}\!/_{\operatorname{Ker} T}$ is straightforward. (Assume that $q_1 = q_2 + k$, where $k \in \operatorname{Ker} T$. It follows $(\cdot, Tq_1) = (\cdot, Tq_2)$.)

Suppose now that $v \in H$ with $\|v\|_T = 0$. With the spectral calculus for self-adjoint operators, we can introduce $T^{1/2}\colon H \to H$, since T is positive semi-definite. Since $(\cdot,\cdot)_T = (T^{1/2}\cdot, T^{1/2}\cdot)$ we derive $T^{1/2}v = 0$, which implies $Tv = T^{1/2}T^{1/2}v = 0$. In other words, $v = 0 + \operatorname{Ker} T$. It follows that $(\cdot,\cdot)_T$ is positive definite on ${}^{H}\!/_{\operatorname{Ker} T}$. □

Corollary A.3. *H_T, endowed with the inner product $(\cdot,\cdot)_T$, is a Hilbert space.*

Proposition A.4. *The operator $T\colon H \to H$ extends in a natural way to an operator $T\colon H_T \to H$ (denoted with the same symbol), such that*

$$\|T\|_{H_T \to H} \leq 1 \quad \textit{and} \quad T(v + \operatorname{Ker} T) = Tv \quad \textit{for all } v \in H.$$

Proof. The extension to the quotient space ${}^{H}\!/_{\operatorname{Ker} T}$ with the above properties is clear. Consider now an element v in the closure of ${}^{H}\!/_{\operatorname{Ker} T}$, i.e., we have $\|v_n - v\|_T \to 0$ for $n \to \infty$ for some sequence $\{ v_n \} \subset H$. Therefore, we have $T^{1/2}v_n \to T^{1/2}v$ in H and we set $Tv = \lim_{n\to\infty} Tv_n$. Furthermore we have $\|Tv\| = \lim_{n\to\infty}\|Tv_n\| \leq \|T^{1/2}\|_{H\to H} \lim_{n\to\infty}\|T^{1/2}v_n\| \leq \lim_{n\to\infty}\|v_n\|_T = \|h\|_T$ since $\|T^{1/2}\|_{H\to H} \leq 1$ due to Assumption 3.3.(i). It is now easy to verify that this extension of T yields a linear operator with the desired properties. □

We also needed the following elementary result.

Proposition A.5. *Consider a positive real sequence $g_n \geq 0$ for $n \in \mathbb{N}_0$, which fulfills the estimate $g_{n+1} \leq \sigma g_n + \varepsilon_n$ with $\sigma < 1$ for perturbations $0 \leq \varepsilon_n \to 0$ for $n \to \infty$. Then g_n converges to zero for $n \to \infty$.*

Proof. For convenience of notation, we can without restriction assume that $g_0 = 0$. By induction we have for all $n \in \mathbb{N}$ and $1 \leq m \leq n$ that

$$g_n \leq \sum_{k=0}^{n-1} \sigma^{n-1-k}\varepsilon_k = \sigma^{n-m} \sum_{k=0}^{m-1} \sigma^{m-1-k}\varepsilon_k + \sum_{k=m}^{n-1} \sigma^{n-1-k}\varepsilon_k \leq \frac{\sigma^{n-m}}{1-\sigma} \sup_{k\geq 0} \varepsilon_k + \frac{1}{1-\sigma} \sup_{k\geq m} \varepsilon_k$$

by the geometric series. By choosing m sufficiently large, the second term becomes arbitrarily small, since $\varepsilon_n \to 0$ for $n \to \infty$. The first term can be controlled by choosing $n > m$ sufficiently large, which shows $g_n \to 0$ for $n \to \infty$. □

A.3. Interpolation error estimates

Mass lumping

We prove an estimate for the error due to mass lumping stated in Lemma 4.31, which was needed in section 4.5.4. We use the same notation and make the same assumptions as there. For $d = 2$ a proof of this standard result (which is very similar to the one given below) can be found, e.g., in [AKV92; Ran08].

Lemma A.6. *Let u_h and φ_h be elements of U_h^1. For the mass lumping of the inner product we have the a priori estimate*

$$|(u_h, \varphi_h) - (u_h, \varphi_h)_h| \leq Ch^2 \|\nabla u_h\|_{L^2(\Omega_c)} \|\nabla \varphi_h\|_{L^2(\Omega_c)}.$$

Proof. The result is proved with the usual transformation and localization argument based on the Bramble-Hilbert lemma. We introduce the reference triangle $\hat{K}$ and define the function

$$\hat{f}\colon \hat{K} \to \mathbb{R}, \quad \hat{f}(x) = \hat{u}_h(x)\hat{\varphi}_h(x)$$

for given linear functions $\hat{u}_h$ and $\hat{\varphi}_h \in \mathcal{P}_1(\hat{K})$. On the reference cell, the error due to quadrature is given by

$$e_{\hat{K}} = \left| \int_{\hat{K}} \hat{f}(x)\,\mathrm{d}x - Q_{\mathrm{Trap},\hat{K}}(\hat{f}) \right| = \left| \int_{\hat{K}} \left[\hat{f}(x) - (i_h \hat{f})(x) \right] \mathrm{d}x \right| \leq \|\hat{f} - i_h \hat{f}\|_{L^1(\hat{K})}.$$

This term is estimated with the help of an interpolation estimate for the nodal interpolation. We apply [BS08, Theorem 4.4.4] to obtain

$$\|\hat{f} - i_h \hat{f}\|_{L^1(\hat{K})} \leq C \|\hat{f} - i_h \hat{f}\|_{L^q(\hat{K})} \leq C \|\nabla^2 \hat{f}\|_{L^q(\hat{K})},$$

for $q = 1$ in the case of $d = 2$ and $q > 3/2$ in the case of $d = 3$ with a constant depending only on q and $\hat{K}$ (recall that $W^{2,q}(\hat{K})$ embeds into the continuous functions for this choice of q). We compute

$$\nabla^2 \hat{f} = \nabla^2 (\hat{u}_h \hat{\varphi}_h) = \nabla^2 \hat{u}_h\, \hat{\varphi}_h + 2 \nabla \hat{u}_h \nabla \hat{\varphi}_h + \hat{u}_h\, \nabla^2 \hat{\varphi}_h \quad \text{on } \hat{K}$$

with the chain rule. Since $\hat{u}_h$ and $\hat{\varphi}_h$ are linear, the first and the third term vanish. Therefore we obtain

$$e_{\hat{K}} \leq C \|\nabla \hat{u}_h \nabla \hat{\varphi}_h\|_{L^p(\hat{K})} \leq C \|\nabla \hat{u}_h\|_{L^{2q}(\hat{K})} \|\nabla \hat{\varphi}_h\|_{L^{2q}(\hat{K})} \leq C \|\nabla \hat{u}_h\|_{L^2(\hat{K})} \|\nabla \hat{\varphi}_h\|_{L^2(\hat{K})}$$

with Hölder's inequality and the equivalence of all $L^q(\hat{K})$ norms on finite dimensional subspaces of $L^\infty(\hat{K})$ ($\nabla \hat{u}_h$ and $\nabla \hat{\varphi}_h$ are constant). Again, the constant depends only on $\hat{K}$ and the (arbitrary) choice of q above.

By a standard transformation argument, this implies for any given cell $K \in \mathcal{T}_h$ the estimate

$$\left| \int_K u_h(x) \varphi_h(x)\,\mathrm{d}x - Q_{\mathrm{Trap},K}(u_h \varphi_h) \right| \leq C h_K^2 \, \|\nabla u_h\|_{L^2(K)} \|\nabla \varphi_h\|_{L^2(K)},$$

where the constant C additionally depends on the "shape-regularity" of the mesh (see, e.g., [BS08, Section 4.4]). By summing over the contributions from each cell, we finally obtain

$$\begin{aligned}
|(u_h, \varphi_h) - (u_h, \varphi_h)_h| &= \left| \int_{\Omega_c \cap \Omega_h} u_h(x) \varphi_h(x)\,\mathrm{d}x - \sum_{K \in \mathcal{T}_h} Q_{\mathrm{Trap},K}(u_h \varphi_h) \right| \\
&\leq C h^2 \sum_{K \in \mathcal{T}_h} \|\nabla u_h\|_{L^2(K)} \|\nabla \varphi_h\|_{L^2(K)} \\
&\leq C h^2 \left(\sum_{K \in \mathcal{T}_h} \|\nabla u_h\|_{L^2(K)}^2 \right)^{1/2} \left(\sum_{K \in \mathcal{T}_h} \|\nabla \varphi_h\|_{L^2(K)}^2 \right)^{1/2} \\
&= C h^2 \, \|\nabla u_h\|_{L^2(\Omega_c)} \|\nabla \varphi_h\|_{L^2(\Omega_c)}
\end{aligned}$$

with the discrete version of Hölder's inequality, which concludes the proof. □

Nodal interpolation in time

We also give the proof of Lemma 5.13, which was needed for the analysis of the parabolic control problem in section 5.3. We use the same notation as there and state again the result for convenience.

Lemma A.7. *For any $w \in L^2(I, H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(I, L^2(\Omega))$ we have*

$$\|w - i_k w\|_{L^2(I,L^2(\Omega))} \leq C\, k \|\partial_t w\|_{L^2(I\times\Omega)}, \tag{A.1}$$

$$\|w - i_k w\|_{L^2(I,H_0^1(\Omega))} \leq C\, k^{1/2} \left(\|\partial_t w\|_{L^2(I\times\Omega)} + \|\Delta w\|_{L^2(I\times\Omega)} \right). \tag{A.2}$$

The estimate (A.1) is standard. To prove (A.2), we need an auxiliary result.

Proposition A.8. *For any $w \in L^2(I, H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(I, L^2(\Omega))$ we have the estimate*

$$\sup_{t\in I} \|\nabla(w(t) - w(T))\|_{L^2(\Omega)}^2 \leq C\, \|\partial_t w\|_{L^2(I,L^2(\Omega))} \|\Delta w\|_{L^2(I,L^2(\Omega))},$$

where the constant C is independent of T.

Proof. Since $w \in \mathcal{C}(\bar{I}, H_0^1(\Omega))$ with the trace theorem [Ama95, Theorem III 4.10.2] we have a unique, continuous representation $[0,T] \ni t \mapsto w(t) \in H_0^1(\Omega)$ and hence for $\Delta w(t) \in H^{-1}(\Omega)$. Since $\|\Delta w(\cdot)\|_{L^2(\Omega)}$ is square integrable, it is finite almost everywhere and we can choose a point $t_0 \in [0,T]$, such that

$$\|\Delta w(t_0)\|_{L^2(\Omega)}^2 \leq \frac{1}{T} \int_0^T \|\Delta w(t)\|_{L^2(\Omega)}^2 \,\mathrm{d}t.$$

We can estimate with the triangle inequality that

$$\sup_{t\in I} \|\nabla(w(t) - w(T))\|_{L^2(\Omega)} \leq 2 \sup_{t\in I} \|\nabla(w(t) - w(t_0))\|_{L^2(\Omega)}. \tag{A.3}$$

To estimate the term on the right we define the function $v = w - w(t_0)$, which is an element of $L^2(I, H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(I, L^2(\Omega))$. By construction, v fulfills $v(t_0) = 0$ and $\partial_t v = \partial_t w$ and we can estimate

$$\|\Delta v\|_{L^2(I,L^2(\Omega))} \leq 2 \|\Delta w\|_{L^2(I,L^2(\Omega))}$$

by the choice of t_0. Now, we can apply a well-known identity, integration by parts and Hölder's inequality to obtain for any $t \in I$ that

$$\begin{aligned}\|\nabla v(t)\|_{L^2(\Omega))}^2 &= \int_{t_0}^t \frac{\mathrm{d}}{\mathrm{d}s} \|\nabla v(s)\|_{L^2(\Omega))}^2 \,\mathrm{d}s = \int_{t_0}^t 2\,(\partial_t v(s), -\Delta v(s))\,\mathrm{d}s \\ &\leq 2\, \|\partial_t v\|_{L^2(I,L^2(\Omega))} \|\Delta v\|_{L^2(I,L^2(\Omega))} \leq 4\, \|\partial_t w\|_{L^2(I,L^2(\Omega))} \|\Delta w\|_{L^2(I,L^2(\Omega))},\end{aligned}$$

and we finish the proof by combining this with (A.3). □

Proof of Lemma A.7. To show estimate (A.2), we first prove it on the reference interval $I' = (0,1)$ for an arbitrary $\hat{w} \in L^2(I', H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(I', L^2(\Omega))$. With Proposition A.8 it holds on the reference interval that

$$\begin{aligned}\|\nabla(\hat{w} - \hat{w}(1))\|_{L^2(I',L^2(\Omega))}^2 &\leq \sup_{t\in I'} \|\nabla(\hat{w}(t) - \hat{w}(1))\|_{L^2(\Omega)}^2 \\ &\leq C\, \|\partial_t \hat{w}\|_{L^2(I',L^2(\Omega))} \|\Delta \hat{w}\|_{L^2(I',L^2(\Omega))}.\end{aligned}$$

By linear transformation this implies for w, restricted to an arbitrary time interval I_m, that

$$\begin{aligned}\|w - w(t_m)\|_{L^2(I_m,H_0^1(\Omega))}^2 &\leq C\, k_m \|\partial_t w\|_{L^2(I_m,L^2(\Omega))} \|\Delta w\|_{L^2(I_m,L^2(\Omega))} \\ &\leq C\, k_m \left(\|\partial_t w\|_{L^2(I_m,L^2(\Omega))}^2 + \|\Delta w\|_{L^2(I_m,L^2(\Omega))}^2 \right).\end{aligned}$$

The final result is obtained by summing these estimates over all intervals I_m for $m = 1 \dots M$ and taking the square root. □

Bibliography

[ACG15] Jean-Marc Azaïs, Yohann de Castro, and Fabrice Gamboa. "Spike detection from inaccurate samplings." In: *Appl. Comput. Harmon. Anal.* 38.2 (2015), pp. 177–195.

[ACT02] Nadir Arada, Eduardo Casas, and Fredi Tröltzsch. "Error Estimates for the Numerical Approximation of a Semilinear Elliptic Control Problem." In: *Comput. Optim. Appl.* 23.2 (2002), pp. 201–229.

[AF03] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*. 2nd ed. Vol. 140. Pure and Applied Mathematics. Elsevier/Academic Press, Amsterdam, 2003.

[AG01] David H. Armitage and Stephen J. Gardiner. *Classical Potential Theory*. Springer, London, 2001.

[AK92] Yu. A. Alkhutov and Vladimir A. Kondrat'ev. "Solvability of the Dirichlet problem for second-order elliptic equations in a convex domain." In: *Differ. Uravn.* 28.5 (1992), pp. 806–818, 917.

[AKV92] Andrey B. Andreev, V. A. Kascieva, and Michèle Vanmaele. "Some results in lumped mass finite-element approximation of eigenvalue problems using numerical quadrature formulas." In: *J. Comput. Appl. Math.* 43.3 (1992), pp. 291–311.

[AZ08] Jürgen Appell and Petr P. Zabrejko. *Nonlinear Superposition Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 2008.

[Alt11] Hans Wilhelm Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung.* 6th ed. Springer, Berlin Heidelberg, 2011.

[Ama00] Herbert Amann. "Compact embeddings of vector valued Sobolev and Besov spaces." In: *Glas. Mat. Ser. III* 35.1 (2000), pp. 161–177.

[Ama05] Herbert Amann. "Nonautonomous Parabolic Equations Involving Measures." In: *J. Math. Sci. (N. Y.)* 130.4 (2005), pp. 4780–4802.

[Ama95] Herbert Amann. *Linear and Quasilinear Parabolic Problems. Volume I: Abstract Linear Theory*. Vol. 89. Monographs in Mathematics. Birkhäuser, Basel, 1995.

[BC11] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

[BCS13] Alberto Bressan, Guiseppe Maria Coclite, and Wen Shen. "A Multidimensional Optimal-Harvesting Problem with Measure-Valued Solutions." In: *SIAM J. Control Optim.* 51.2 (2013), pp. 1186–1202.

[BE03] Malte Braack and Alexandre Ern. "A Posteriori Control of Modeling Errors and Discretization Errors." In: *Multiscale Model. Simul.* 1.2 (2003), pp. 221–238.

[BG98] C. Bernardi and V. Girault. "A Local Regularization Operator for Triangular and Quadrilateral Finite Elements." In: *SIAM J. Numer. Anal.* 35.5 (1998), pp. 1893–1916.

[BIK99] Maïtine Bergounioux, Kazufumi Ito, and Karl Kunisch. "Primal-Dual Strategy for Constrained Optimal Control Problems." In: *SIAM J. Control Optim.* 37.4 (1999), pp. 1176–1194.

[BKR00] Roland Becker, Hartmut Kapp, and Rolf Rannacher. "Adaptive Finite Element Methods for Optimal Control of Partial Differential Equations: Basic Concept." In: *SIAM J. Control Optim.* 39.1 (2000), pp. 113–132.

[BMR91] Alfredo Bermúdez, Aurea Martínez, and Carmen Rodríguez. "Un problème de contrôle ponctuel lié à l'emplacement optimal d'émissaires d'évacuation sous-marins." In: *C. R. Math. Acad. Sci. Paris* 313.8 (1991), pp. 515–518.

[BMV07] Roland Becker, Dominik Meidner, and Boris Vexler. "Efficient numerical solution of parabolic optimization problems by finite element methods." In: *Optim. Methods Softw.* 22.5 (2007), pp. 813–833.

[BP13] Kristian Bredies and Hanna Katriina Pikkarainen. "Inverse problems in spaces of measures." In: *ESAIM Control Optim. Calc. Var.* 19 (01 2013), pp. 190–218.

[BR01] Roland Becker and Rolf Rannacher. "An optimal control approach to a posteriori error estimation in finite element methods." In: *Acta Numer.* 10 (2001), pp. 1–102.

[BR96] Roland Becker and Rolf Rannacher. "A Feed-Back Approach to Error Control in Finite Element Methods: Basic Analysis and Examples." In: *East-West J. Numer. Math.* 4 (1996), pp. 237–264.

[BS00] J. Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Series in Operations Research. Springer, New York, 2000.

[BS08] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*. 3rd ed. Vol. 15. Texts in Applied Mathematics. Springer, New York, 2008.

[BV07] Roland Becker and Boris Vexler. "Optimal control of the convection-diffusion equation using stabilized finite element methods." In: *Numer. Math.* 106.3 (2007), pp. 349–367.

[BV09] Olaf Benedix and Boris Vexler. "A posteriori error estimation and adaptivity for elliptic optimal control problems with state constraints." In: *Comput. Optim. Appl.* 44.1 (2009), pp. 3–25.

[Bra98] Malte Braack. "An Adaptive Finite Element Method for Reactive Flow Problems." PhD Dissertation. Ruprecht-Karls-Universität Heidelberg, 1998.

[Bre11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[Bru+12] Patricia Brunner, Christian Clason, Manuel Freiberger, and Hermann Scharfetter. "A deterministic approach to the adapted optode placement for illumination of highly scattering tissue." In: *Biomed. Opt. Express* 3 (2012), pp. 1732–1743.

[CCK12] Eduardo Casas, Christian Clason, and Karl Kunisch. "Approximation of elliptic control problems in measure spaces with sparse solutions." In: *SIAM J. Control Optim.* 50.4 (2012), pp. 1735–1752.

[CCK13] Eduardo Casas, Christian Clason, and Karl Kunisch. "Parabolic Control Problems in Measure Spaces with Sparse Solutions." In: *SIAM J. Control Optim.* 51.1 (2013), pp. 28–63.

[CFG13] Emmanuel J. Candès and Carlos Fernandez-Granda. "Super-Resolution from Noisy Data." In: *J. Fourier Anal. Appl.* 19.6 (2013), pp. 1229–1254.

[CFG14] Emmanuel J. Candès and Carlos Fernandez-Granda. "Towards a mathematical theory of super-resolution." In: *Comm. Pure Appl. Math.* 67.6 (2014), pp. 906–956.

[CHW12a] Eduardo Casas, Roland Herzog, and Gerd Wachsmuth. "Approximation of sparse controls in semilinear equations by piecewise linear functions." In: *Numer. Math.* 122.4 (2012), pp. 645–669.

[CHW12b] Eduardo Casas, Roland Herzog, and Gerd Wachsmuth. "Optimality Conditions and Error Analysis of Semilinear Elliptic Control Problems with L^1 Cost Functional." In: *SIAM J. Optim.* 22.3 (2012), pp. 795–820.

[CK11a] Christian Clason and Karl Kunisch. "A duality-based approach to elliptic control problems in non-reflexive Banach spaces." In: *ESAIM Control Optim. Calc. Var.* 17.1 (2011), pp. 243–266.

[CK11b] Christian Clason and Karl Kunisch. "A measure space approach to optimal source placement." In: *Comput. Optim. Appl.* 53.1 (2011), pp. 155–171.

[CK14] Eduardo Casas and Karl Kunisch. "Optimal Control of Semilinear Elliptic Equations in Measure Spaces." In: *SIAM J. Control Optim.* 52.1 (2014), pp. 339–364.

[CMV14] Eduardo Casas, Mariano Mateos, and Boris Vexler. "New regularity results and improved error estimates for optimal control problems with state constraints." In: *ESAIM Control Optim. Calc. Var.* 20.3 (2014), pp. 803–822.

[CO84] Graham F. Carey and J. Tinsley Oden. *Finite elements. Vol. III.* The Texas Finite Element Series, III. Computational aspects. Prentice Hall Inc., Englewood Cliffs, NJ, 1984.

[CT03] Eduardo Casas and Fredi Tröltzsch. "Error Estimates for Linear-Quadratic Elliptic Control Problems." In: *Analysis and Optimization of Differential Systems.* Vol. 121. IFIP – The International Federation for Information Processing. Kluwer Acad. Publ., Boston, MA, 2003, pp. 89–100.

[CVZ14] Eduardo Casas, Boris Vexler, and Enrique Zuazua. "Sparse initial data identification for parabolic PDE and it's finite element approximations." In: *Math. Control Relat. Fields* (2014). accepted.

[CW05] Patrick L. Combettes and Valérie Wajs. "Signal Recovery by Proximal Forward-Backward Splitting." In: *Multiscale Model. Simul.* 4.4 (2005), pp. 1168–1200.

[CZ13] Eduardo Casas and Enrique Zuazua. "Spike controls for elliptic and parabolic PDEs." In: *Systems Control Lett.* 62.4 (2013), pp. 311–318.

[Car99] Carsten Carstensen. "Quasi-Interpolation and A Posteriori Error Analysis in Finite Element Methods." In: *ESAIM Math. Model. Numer. Anal.* 33.6 (1999), pp. 1187–1202.

[Cas07] Eduardo Casas. "Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems." In: *Adv. Comput. Math.* 26.1-3 (2007), pp. 137–153.

[Cas85] Eduardo Casas. "L^2 estimates for the finite element method for the Dirichlet problem with singular data." In: *Numer. Math.* 47.4 (1985), pp. 627–632.

[Cas86] Eduardo Casas. “Control of an elliptic problem with pointwise state constraints.” In: *SIAM J. Control Optim.* 24.6 (1986), pp. 1309–1318.

[Cas97] Eduardo Casas. “Pontryagin’s principle for state-constrained boundary control problems of semilinear parabolic equations.” In: *SIAM J. Control Optim.* 35.4 (1997), pp. 1297–1327.

[Chr81] Ion Chryssoverghi. “Approximate methods for optimal pointwise control of parabolic systems.” In: *Systems Control Lett.* 1.3 (1981/82), pp. 216–219.

[Cla13] Francis Clarke. *Functional analysis, calculus of variations and optimal control*. Vol. 264. Graduate Texts in Mathematics. Springer, London, 2013.

[DF95] Steven P. Dirkse and Michael C. Ferris. “The PATH Solver: A Non-Monotone Stabilization Scheme for Mixed Complementarity Problems.” In: *Optim. Methods Softw.* 5 (1995), pp. 123–156.

[DH07] Klaus Deckelnick and Michael Hinze. “Convergence of a Finite Element Approximation to a State-Constrained Elliptic Control Problem.” In: *SIAM J. Numer. Anal.* 45.5 (2007), pp. 1937–1953.

[DHV01] Asen L. Dontchev, William W. Hager, and Vladimir M. Veliov. “Second-Order Runge-Kutta Approximations in Control Constrained Optimal Control.” In: *SIAM J. Numer. Anal.* 38.1 (2001), pp. 202–226.

[DR00] Jérôme Droniou and Jean-Pierre Raymond. “Optimal pointwise control of semilinear parabolic equations.” In: *Nonlinear Anal.* 39.2 (2000), pp. 135–156.

[DU77] Joseph Diestel and John J. Uhl. *Vector measures*. Mathematical surveys. American Mathematical Society, 1977.

[Dan67] James W. Daniel. “The Conjugate Gradient Method for Linear and Nonlinear Operator Equations.” In: *SIAM J. Numer. Anal.* 4.1 (1967), pp. 10–26.

[Die69] Jean Dieudonné. *Foundations of modern analysis*. Vol. 10-I. Pure and Applied Mathematics. Enlarged and corrected printing. Academic Press, New York-London, 1969.

[Dro00] Jérôme Droniou. “Solving convection-diffusion equations with mixed, Neumann and Fourier boundary conditions and measures as data, by a duality method.” In: *Adv. Differential Equations* 5.10-12 (2000), pp. 1341–1396.

[EG92] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[ET99] Ivar Ekeland and Roger Témam. *Convex analysis and variational problems*. Vol. 28. Classics in Applied Mathematics. Translated from the French. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[Els11] Jürgen Elstrodt. *Maß- und Integrationstheorie*. 7th ed. Springer-Lehrbuch. Springer, Berlin Heidelberg, 2011.

[Eva10] Lawrence C. Evans. *Partial Differential Equations*. 2nd ed. American Mathematical Society, 2010.

[FR08] Massimo Fornasier and Holger Rauhut. “Recovery algorithms for vector-valued data with joint sparsity constraints.” In: *SIAM J. Numer. Anal.* 46.2 (2008), pp. 577–613.

[FR76] Jens Frehse and Rolf Rannacher. "Eine L^1-Fehlerabschätzung für diskrete Grundlösungen in der Methode der finiten Elemente." In: *Finite Elemente. Tagungsband des Sonderforschungsbereichs 72.* Vol. 89. Bonner Mathematische Schriften. Bonn, 1976, pp. 92–114.

[FS14] Massimo Fornasier and Francesco Solombrino. "Mean-field optimal control." In: *ESAIM Control Optim. Calc. Var.* 20.4 (2014), pp. 1123–1152.

[GGW08] Roland Griesse, Thomas Grund, and Daniel Wachsmuth. "Update strategies for perturbed nonsmooth equations." In: *Optim. Methods Softw.* 23.3 (2008), pp. 321–343.

[GHZ14] Wei Gong, Michael Hinze, and Zhaojie Zhou. "A Priori Error Analysis for Finite Element Approximation of Parabolic Optimal Control Problems with Pointwise Control." In: *SIAM J. Control Optim.* 52.1 (2014), pp. 97–119.

[GK09] Carsten Gräser and Ralf Kornhuber. "Nonsmooth Newton Methods for Set-Valued Saddle Point Problems." In: *SIAM J. Numer. Anal.* 47.2 (2009), pp. 1251–1273.

[GKR01] Jens A. Griepentrog, Hans-Christoph Kaiser, and Joachim Rehberg. "Heat Kernel and Resolvent Properties for Second Order Elliptic Differential Operators with General Boundary Conditions on L^p." In: *Adv. Math. Sci. Appl.* 11.1 (2001), pp. 87–112.

[GR01] Jens A. Griepentrog and Lutz Recke. "Linear Elliptic Boundary Value Problems with Non-Smooth Data: Normal Solvability on Sobolev-Campanato Spaces." In: *Math. Nachr.* 225.1 (2001), pp. 39–74.

[GV07] Roland Griesse and Boris Vexler. "Numerical Sensitivity Analysis for the Quantity of Interest in PDE-Constrained Optimization." In: *SIAM J. Sci. Comput.* 29.1 (2007), pp. 22–48.

[Gri85] Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains.* Vol. 24. Monographs and Studies in Mathematics. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[Grä08] Carsten Gräser. "Globalization of Nonsmooth Newton Methods for Optimal Control Problems." In: *Numerical Mathematics and Advanced Applications.* Springer Berlin Heidelberg, 2008, pp. 605–612.

[Grö89] Konrad Gröger. "A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations." In: *Math. Ann.* 283.4 (1989), pp. 679–687.

[HD+09] Robert Haller-Dintelmann, Christian Meyer, Joachim Rehberg, and Anton Schiela. "Hölder continuity and optimal control for nonsmooth elliptic problems." In: *Appl. Math. Optim.* 60.3 (2009), pp. 397–428.

[HDR09] Robert Haller-Dintelmann and Joachim Rehberg. "Maximal parabolic regularity for divergence operators including mixed boundary conditions." In: *J. Differential Equations* 247.5 (2009), pp. 1354–1396.

[HH02] Michael Hintermüller and Michael Hinze. "Globalization of SQP-Methods in Control of the Instationary Navier-Stokes Equations." In: *ESAIM Math. Model. Numer. Anal.* 36 (04 2002), pp. 725–746.

[HH06] Michael Hintermüller and Michael Hinze. "A SQP-Semismooth Newton-type Algorithm applied to Control of the instationary Navier–Stokes System Subject to Control Constraints." In: *SIAM J. Optim.* 16.4 (2006), pp. 1177–1200.

[HH09] Michael Hintermüller and Michael Hinze. "Moreau-Yosida regularization in state constrained elliptic control problems: error estimates and parameter adjustment." In: *SIAM J. Numer. Anal.* 47.3 (2009), pp. 1666–1683.

[HH10] Michael Hintermüller and Ronald H. W. Hoppe. "Goal-Oriented Adaptivity in Pointwise State Constrained Optimal Control of Partial Differential Equations." In: *SIAM J. Control Optim.* 48.8 (2010), pp. 5468–5487.

[HIK03] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. "The primal-dual active set strategy as a semismooth Newton method." In: *SIAM J. Optim.* 13.3 (2003), pp. 865–888.

[HK06a] Michael Hintermüller and Karl Kunisch. "Feasible and non-interior path-following in constrained minimization with low multiplier regularity." In: *SIAM J. Control Optim.* 45.4 (2006), pp. 1198–1221.

[HK06b] Michael Hintermüller and Karl Kunisch. "Path-following methods for a class of constrained minimization problems in function space." In: *SIAM J. Optim.* 17.1 (2006), pp. 159–187.

[HSW12] Roland Herzog, Georg Stadler, and Gerd Wachsmuth. "Directional Sparsity in Optimal Control of Partial Differential Equations." In: *SIAM J. Control Optim.* 50.2 (2012), pp. 943–963.

[HSW14] Michael Hintermüller, Anton Schiela, and Winnifred Wollner. "The Length of the Primal-Dual Path in Moreau–Yosida-Based Path-Following Methods for State Constrained Optimal Control." In: *SIAM J. Optim.* 24.1 (2014), pp. 108–126.

[HUSN84] Jean-Baptiste Hiriart-Urruty, Jean-Jacques Strodiot, and V. Hien Nguyen. "Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data." In: *Appl. Math. Optim.* 11.1 (1984), pp. 43–56.

[HV12] Michael Hinze and Morten Vierling. "The semi-smooth Newton method for variationally discretized control constrained elliptic optimal control problems; implementation, convergence and globalization." In: *Optim. Methods Softw.* 27.6 (2012), pp. 933–950.

[Hen15] Felix Henneke. "Sparse Time-Frequency Control of Bilinear Quantum Systems." (in preparation). PhD Dissertation. 2015.

[Hen96] Wolfgang Hensgen. "A simple proof of Singer's representation theorem." In: *Proc. Amer. Math. Soc.* 124.10 (1996), pp. 3211–3212.

[Hin+09] Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*. Vol. 23. Mathematical Modelling: Theory and Applications. Springer, New York, 2009.

[Hin05] Michael Hinze. "A Variational Discretization Concept in Control Constrained Optimization: The Linear-Quadratic Case." In: *Comput. Optim. Appl.* 30.1 (2005), pp. 45–61.

[IK04] Kazufumi Ito and Karl Kunisch. "The Primal-Dual Active Set Method for Nonlinear Optimal Control Problems with Bilateral Constraints." In: *SIAM J. Control Optim.* 43.1 (2004), pp. 357–376.

[IK08] Kazufumi Ito and Karl Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*. Vol. 15. Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[IK09] Kazufumi Ito and Karl Kunisch. "On a semi-smooth Newton method and its globalization." In: *Math. Program.* 118.2 (2009), pp. 347–370.

[IK92] Kazufumi Ito and Karl Kunisch. "On the Choice of the Regularization Parameter in Nonlinear Inverse Problems." In: *SIAM J. Optim.* 2.3 (1992), pp. 376–404.

[JK95] David Jerison and Carlos E. Kenig. "The Inhomogeneous Dirichlet Problem in Lipschitz Domains." In: *J. Funct. Anal.* 130.1 (1995), pp. 161–219.

[JLS09] Bangti Jin, Dirk A. Lorenz, and Stefan Schiffler. "Elastic-net regularization: error estimates and active set methods." In: *Inverse Problems* 25.11 (2009), Article ID 115022, 26 pp.

[KKV11] Barbara Kaltenbacher, Alana Kirchner, and Boris Vexler. "Adaptive discretizations for the choice of a Tikhonov regularization parameter in nonlinear inverse problems." In: *Inverse Problems* 27.12 (2011), Article ID 125008, 28 pp.

[KPR14] Karl Kunisch, Konstantin Pieper, and Armin Rund. "Time optimal control for a reaction diffusion system arising in cardiac electrophysiology – a monolithic approach." In: *ESAIM Math. Model. Numer. Anal.* (2014). accepted.

[KPV14] Karl Kunisch, Konstantin Pieper, and Boris Vexler. "Measure Valued Directional Sparsity for Parabolic Optimal Control Problems." In: *SIAM J. Control Optim.* 52.5 (2014), pp. 3078–3108.

[KR02] Karl Kunisch and Arnd Rösch. "Primal-Dual Active Set Strategy for a General Class of Constrained Optimal Control Problems." In: *SIAM J. Optim.* 13.2 (2002), pp. 321–334.

[KTV14] Karl Kunisch, Philip Trautmann, and Boris Vexler. "Optimal control of the undamped linear wave equation with measure valued controls." submitted. 2014.

[LR10] Dirk A. Lorenz and Arnd Rösch. "Error estimates for joint Tikhonov and Lavrentiev regularization of constrained control problems." In: *Appl. Anal.* 89.11 (2010), pp. 1679–1691.

[LV13] Dmitriy Leykekhman and Boris Vexler. "Optimal A Priori Error Estimates of Parabolic Optimal Control Problems with Pointwise Control." In: *SIAM J. Numer. Anal.* 51.5 (2013), pp. 2797–2821.

[Lan72] Naum S. Landkof. *Foundations of Modern Potential Theory*. Vol. 180. Die Grundlehren der mathematischen Wissenschaften. Translated from the Russian by A. P. Doohovskoy. Springer, New York-Heidelberg, 1972.

[Lan83] Serge Lang. *Real Analysis*. 2nd ed. Advanced Book Program. Addison-Wesley Publishing Company, Reading, MA, 1983.

[Lan93] Serge Lang. *Real and functional analysis*. 3rd ed. Vol. 142. Graduate Texts in Mathematics. Springer-Verlag, New York, 1993.

[Lio69] Jacques-Louis Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Collection études mathématiques. Dunod; Gauthier-Villars, Paris, 1969.

[Lio71] Jacques-Louis Lions. *Optimal control of systems governed by partial differential equations.* Vol. 170. Die Grundlehren der mathematischen Wissenschaften. Translated from the French by S. K. Mitter. Springer-Verlag, New York-Berlin, 1971.

[Lio92] Jacques-Louis Lions. "Pointwise Control for Distributed Systems." In: *Control and Estimation in Distributed Parameter Systems.* 1992. Chap. 1, pp. 1–39.

[MPS11] Christian Meyer, Lucia Panizzi, and Anton Schiela. "Uniqueness criteria for the adjoint equation in state-constrained elliptic optimal control." In: *Numer. Funct. Anal. Optim.* 32.9 (2011), pp. 983–1007.

[MR04] Christian Meyer and Arnd Rösch. "Superconvergence Properties of Optimal Control Problems." In: *SIAM J. Control Optim.* 43.3 (2004), pp. 970–985.

[MRVM00] Aurea Martínez, Carmen Rodríguez, and M. Ernesto Vázquez-Méndez. "Theoretical and Numerical Analysis of an Optimal Control Problem Related to Wastewater Treatment." In: *SIAM J. Control Optim.* 38.5 (2000), pp. 1534–1553.

[MV08] Dominik Meidner and Boris Vexler. "A Priori Error Estimates for Space-Time Finite Element Discretization of Parabolic Optimal Control Problems Part I: Problems Without Control Constraints." In: *SIAM J. Control Optim.* 47.3 (2008), pp. 1150–1177.

[Mei08] Dominik Meidner. "Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems." PhD Dissertation. Ruprecht-Karls-Universität Heidelberg, 2008.

[Mey08] Christian Meyer. "Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints." In: *Control Cybernet.* 37.1 (2008), pp. 51–83.

[Mey63] Norman G. Meyers. "An L^p-estimate for the gradient of solutions of second order elliptic divergence equations." In: *Ann. Scuola Norm. Sup. Pisa (3)* 17 (1963), pp. 189–206.

[Mez09] Lakhdar Meziani. "On the dual space $C_0^*(S, X)$." In: *Acta Math. Univ. Comenian. (N.S.)* 78.1 (2009), pp. 153–160.

[Mil15] Andre Milzarek. "Numerical methods for a class of nonsmooth optimization problems and generalized variational inequalties." (in preparation). PhD Dissertation. 2015.

[Mor65] Jean-Jaques Moreau. "Proximité et dualité dans un espace hilbertien." In: *Bull. Soc. Math. France* 93 (1965), pp. 273–299.

[NN13] Yurii Nesterov and Arkadi Nemirovski. "On first-order algorithms for ℓ_1/nuclear norm minimization." In: *Acta Numer.* 22 (2013), pp. 509–575.

[PV13] Konstantin Pieper and Boris Vexler. "A Priori Error Analysis for Discretization of Sparse Elliptic Optimal Control Problems in Measure Space." In: *SIAM J. Control Optim.* 51.4 (2013), pp. 2788–2808.

[Pri95] Alain Prignet. "Remarks on existence and uniqueness of solutions of elliptic problems with right-hand side measures." In: *Rend. Mat. Appl. (7)* 15.3 (1995), pp. 321–337.

[QS99] Liqun Qi and Defeng Sun. "A survey of some nonsmooth equations and smoothing Newton methods." In: *Progress in Optimization, volume 30 of Applied Optimization.* Kluwer Academic Publishers, 1999, pp. 121–146.

[RS82] Rolf Rannacher and Ridgway Scott. "Some optimal error estimates for piecewise linear finite element approximations." In: *Math. Comp.* 38.158 (1982), pp. 437–445.

[RV06] Arnd Rösch and Boris Vexler. "Optimal Control of the Stokes Equations: A Priori Error Analysis for Finite Element Discretization with Postprocessing." In: *SIAM J. Numer. Anal.* 44.5 (2006), pp. 1903–1920.

[RVW12] Rolf Rannacher, Boris Vexler, and Winnifried Wollner. "A posteriori error estimation in PDE-constrained optimization with pointwise inequality constraints." In: *Constrained optimization and optimal control for partial differential equations*. Vol. 160. Internat. Ser. Numer. Math. Birkhäuser, Basel, 2012, pp. 349–373.

[RZ98] Jean-Pierre Raymond and Hasnaa Zidani. "Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls." In: *SIAM J. Control Optim.* 36.6 (1998), pp. 1853–1879.

[Ral94] Daniel Ralph. "Global Convergence of Damped Newton's Method for Nonsmooth Equations via the Path Search." In: *Math. Oper. Res.* 19.2 (1994), pp. 352–389.

[Ran08] Rolf Rannacher. *Numerische Mathematik 2*. Lecture Notes, Universität Heidelberg: `http://numerik.iwr.uni-heidelberg.de/~lehre/notes/`. 2008.

[Ran76] Rolf Rannacher. "Zur L^∞-Konvergenz linearer finiter Elemente beim Dirichlet-Problem." In: *Math. Z.* 149.1 (1976), pp. 69–77.

[Rob92] Stephen M. Robinson. "Normal Maps Induced by Linear Transformations." In: *Math. Oper. Res.* 17.3 (1992), pp. 691–714.

[Rob94] Stephen M. Robinson. "Newton's method for a class of nonsmooth functions." In: *Set-Valued Var. Anal.* 2.1-2 (1994), pp. 291–305.

[Roc68] R. Tyrrell Rockafellar. "Integrals which are convex functionals." In: *Pacific J. Math.* 24.3 (1968), pp. 525–539.

[Roc71] R. Tyrrell Rockafellar. "Integrals which are convex functionals. II." In: *Pacific J. Math.* 39.2 (1971), pp. 439–469.

[Rud87] Walter Rudin. *Real and complex analysis*. 3rd ed. McGraw-Hill Book Co., New York, 1987.

[Rös06] Arnd Rösch. "Error estimates for linear-quadratic control problems with control constraints." In: *Optim. Methods Softw.* 21.1 (2006), pp. 121–134.

[SG11] Anton Schiela and Andreas Günther. "An interior point algorithm with inexact step computation in function space for state constrained optimal control." In: *Numer. Math.* 119.2 (2011), pp. 373–407.

[SW09] Otmar Scherzer and Birgit Walch. "Sparsity Regularization for Radon Measures." In: *Scale Space and Variational Methods in Computer Vision*. Vol. 5567. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, 2009, pp. 452–463.

[Sch08] Anton Schiela. "A Simplified Approach to Semismooth Newton Methods in Function Space." In: *SIAM J. Optim.* 19.3 (2008), pp. 1417–1432.

[Sch09] Anton Schiela. "Barrier Methods for Optimal Control Problems with State Constraints." In: *SIAM J. Optim.* 20.2 (2009), pp. 1002–1031.

[Sch13] Anton Schiela. "Analytic and algorithmic aspects of path-following in optimal control." Lecture Notes, Technische Universität München. 2013.

[Spr15] Andreas Springer. "Efficient higher order discontinuous Galerkin time discretization for parabolic optimal control problems." PhD Dissertation. Technische Universität München, 2015.

[Sta09] Georg Stadler. "Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices." In: *Comput. Optim. Appl.* 44.2 (2009), pp. 159–181.

[Sta65] Guido Stampacchia. "Le probléme de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus." In: *Ann. Inst. Fourier (Grenoble)* 15.1 (1965), pp. 189–257.

[Ste83] Trond Steihaug. "The Conjugate Gradient Method and Trust Regions in Large Scale Optimization." In: *SIAM J. Numer. Anal.* 20.3 (1983), pp. 626–637.

[Tho06] Vidar Thomée. *Galerkin finite element methods for parabolic problems*. 2nd ed. Vol. 25. Springer Series in Computational Mathematics. Springer, Berlin, 2006.

[Tri78] Hans Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. Vol. 18. North-Holland mathematical library. North-Holland Publ., Amsterdam, 1978.

[Tro87] Giovanni Maria Troianiello. *Elliptic differential equations and obstacle problems*. The University Series in Mathematics. Plenum Press, New York, 1987.

[Trö10a] Fredi Tröltzsch. "On Finite Element Error Estimates for Optimal Control Problems with Elliptic PDEs." In: *Large-Scale Scientific Computing*. Vol. 5910. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 40–53.

[Trö10b] Fredi Tröltzsch. *Optimal Control of Partial Differential Equations*. Vol. 112. Graduate Studies in Mathematics. Providence, Rhode Island: American Mathematical Society, 2010.

[Ulb02] Michael Ulbrich. "Semismooth Newton Methods for Operator Equations in Function Spaces." In: *SIAM J. Optim.* 13.3 (2002), pp. 805–841.

[Ulb11] Michael Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM Series on Optimization. SIAM, 2011.

[VM06] Georg Vossen and Helmut Maurer. "On L^1-minimization in optimal control and applications to robotics." In: *Optimal Control Appl. Methods* 27.6 (2006), pp. 301–321.

[VW08] Boris Vexler and Winnifred Wollner. "Adaptive Finite Elements for Elliptic Optimization Problems with Control Constraints." In: *SIAM J. Control Optim.* 47.1 (2008), pp. 509–534.

[WGS08] Martin Weiser, Tobias Gänzler, and Anton Schiela. "A control reduced primal interior point method for a class of control constrained optimal control problems." In: *Comput. Optim. Appl.* 41.1 (2008), pp. 127–145.

[WW11] Gerd Wachsmuth and Daniel Wachsmuth. "Convergence and regularization results for optimal control problems with sparsity functional." In: *ESAIM Control Optim. Calc. Var.* 17.3 (2011), pp. 858–886.

[Win80] Ragnar Winther. "Some Superlinear Convergence Results for the Conjugate Gradient Method." In: *SIAM J. Numer. Anal.* 17.1 (1980), pp. 14–17.

[Wol10] Winnifred Wollner. "A posteriori error estimates for a finite element discretization of interior point methods for an elliptic optimization problem with state constraints." In: *Comput. Optim. Appl.* 47.1 (2010), pp. 133–159.

[Zie89] William P. Ziemer. *Weakly differentiable functions. Sobolev spaces and functions of bounded variation.* Vol. 120. Graduate Texts in Mathematics. Springer, New York, 1989.

[Gas] Gascoigne. *The finite element toolkit.* `http://gascoigne.uni-hd.de`.

[RoD] RoDoBo. *A C++ library for optimization with stationary and nonstationary PDEs with interface to Gascoigne.* `http://www.rodobo.org`.