Dipl.-Ing. Richard Löscher, BSc

On unified frameworks for space-time optimal control and least squares problems

DOCTORAL THESIS

to achieve the university degree of

Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dr. Olaf Steinbach

Institut für Angewandte Mathematik

Graz, November 2023

Abstract

The aim of this thesis is twofold. Firstly, it focuses on the analysis of distributed optimal control problems constrained by partial differential equations (PDEs) and state and/or control constraints, with the goal to reach a given target up to some prescribed accuracy spending minimal costs. Under some natural assumptions on the involved differential operators, it is shown that this class of problems admits a certain structure and can be analyzed in an abstract framework. This structure carries over to the discrete setting, where optimal approximation results are gained, depending on the regularity of the desired target and the cost parameter. Various model problems, including a space-time optimal control problem for the wave equation, are shown to fit into this framework and numerical examples will be given that support the theory.

Secondly, an approach for the stable solution of (conditionally stable) problems of PDEs is discussed when using a least squares approach. Again, this will be done in an abstract framework. Under merely the same assumptions as in the case of optimal control problems, a full analysis of the continuous and the discrete setting are carried out, revealing that the approach provides a reliable error estimator under a standard saturation assumption. As a model problem, we will discuss the application of the framework to the wave equation and give numerical examples.

Zusammenfassung

Diese Arbeit verfolgt zwei Ziele. Zum Einen liegt der Fokus auf der Analyse von verteilten optimalen Steuerungsproblemen, deren Nebenbedingung durch partielle Differentialgleichungen (PDEs) und Zustands- und/oder Kontrollbeschränkungen gegeben sind. Das Ziel dabei ist es einen vorgegebenen Zustand unter Aufwendung minimaler Kosten so genau wie gewünscht zu erreichen. Unter einigen natürlichen Annahmen über die zugrundeliegenden Differentialoperatoren wird gezeigt, dass diese Klasse von Problemen eine bestimmte Struktur aufweist und in einem abstrakten Rahmen analysiert werden kann. Diese Struktur überträgt sich auf die Diskretisierung, wo optimale Approximationsresultate, abhängig von der Regularität des gewünschten Ziels und dem Kostenparameter erzielt werden. Für verschiedene Modellprobleme, einschließlich eines Raum-Zeit optimalen Steuerungsproblems für die Wellengleichung, wird gezeigt, dass sie in dieses Setting passen, und numerische Beispiele werden gegeben, welche die theoretischen Ergbnisse bestätigen.

Zweitens wird ein Zugang für die stabile Lösung (bedingt stabiler) Probleme von PDEs, mit der Methode der kleinsten Fehlerquadrate, in einem abstrakten Rahmen

diskutiert. Unter ähnlichen Annahmen wie im Fall der optimalen Steuerungsprobleme wird eine vollständige Analyse der kontinuierlichen und diskreten Formulierung durchgeführt. Unter einer Saturationsannahme liefert dieser Zugang einen effizienten und zuverlässigen Fehlerschätzer. Als Anwedungsbeispiel wird die Raum-Zeit Formulierung der Wellengleichung diskutiert und numerische Beispiele werden gegeben.

Contents

1 Introduction

The most common procedure to computationally solve time-dependent partial differential equations (PDEs) are either the *method of lines*, built upon first discretizing in space, e.g., by a finite element method (FEM), see [18, 27, 39, 105], and then applying a time stepping method, see [60, 61], as discussed in [33, 112] for Galerkin finite element methods, or by *Rothes method* [72, 101], i.e., first applying a suitable time stepping method and then discretizing in space. In this work though, we will focus on space time methods, which arise when treating time just as another (spatial) variable and then discretizing the space time domain at once. Pioneering work dates already back to Argyris and Scharpf [5], who treated the time in a variational sense, rather than using a time stepping method, and to Hughes and Hulbert [66, 67], using discontinuous Galerkin methods in time. A tremendous collection of recent advances in space time methods is given in [78], see also [102, 108]. The rising interest in this methodology can be explained by the advantages it bears, stemming from its full space time flexibility. Most prominent are adaptivity locally in space and time simultaneously, see, e.g., [8, 9, 35, 106], the handling of (low) space time regularity in a natural way, e.g., [62, 110, 111], parallelization without causality constraints, see, e.g., [46], and the treatment of moving domains, e.g., [93, 94]. Recent work shows, that even reduced basis methods apply in space and time simultaneously, see [13, 104]. When discretized, space time schemes usually lead to huge systems of (linear) equations, as spatial and temporal degrees of freedom (DoFs) are solved simultaneously. However, especially the computational advances of the last decade made it possible to tackle this objective and gave access to the solutions of such systems. This has lead to major progress in the development of efficient solvers, see [48, 82, 102] and references therein. Of special interest for this work is, that space time approaches make it easy to access the whole temporal evolution of a time dependent problem, or its adjoint/backward in time problem, at once. This becomes particularly handy for optimal control problems, subject to PDE constraints, where one usually needs the solution of the forward and backward problem [84, 113]. While applying a time stepping scheme would require to first step forward in time and then step backward in time, space time approaches lead to a larger system, where the problem is solved at once [10, 79, 81, 91]. Moreover, having the information on the whole evolution history, allows to construct residual based error estimators in space and time simultaneously, see, e.g., [37], which in most cases helps to reduce the number of DoFs drastically and leads to more efficient schemes. In this thesis we will address both, optimal control problems as well as the direct solution of PDEs and analyze the advantages of space time methods stated above for these applications. The remainder of this work is structured as follows.

In Chapter 2 we recall some well-established results concerning the unique solvability of variational formulations in a functional analytical framework and describe the (Hilbert) spaces, needed for our space time analysis later on. Further, we discuss the properties of suitable finite dimensional approximation spaces for a numerical treatment of the problems we investigate. In Chapter 3 we give an abstract framework for distributed optimal control problems, where we aim to reach a desired target y_d, up to a prescribed accuracy, by a state y_ϱ, describing the solution of a PDE $By_\varrho = u_\varrho \in X^*$, under acceptable costs $\|u_\varrho\|_{X^*}$. Motivated by the pioneering work of Neumüller and Steinbach [95], we first derive relations between the cost parameter $\varrho > 0$ and the regularity of the target y_d, revealing a connection between optimal control problems and inverse problems, see, e.g., [24, 92]. Then an abstract discretization, using conforming finite dimensional subspaces, is outlined and stability and error estimates are derived, linking the cost parameter to the best approximation error. In particular, for a finite element discretization, we can derive the optimal choice $\varrho = h^2$, between the cost parameter and the mesh size, which is of interest in the design of robust iterative solvers for optimal control problems, see, e.g., [75, 80, 81]. As in many applications the incorporation of state and/or control constraints is required, we phrase them in the abstract framework, in both cases charging us with the task to solve a variational inequality. The analysis on the continuous and discrete level will be discussed in detail. Chapter 4 will then show some applications of the proposed framework. Starting with an elliptic PDE, we will discuss different choices of measuring the control and compare the approaches with existing examples in the literature. While in the elliptic case the incorporation of constraints, the applicability of adaptive schemes and adaptive regularization parameters seems to be more commonly discussed, the full capacity of the framework will be revealed by considering a hyperbolic optimal control problem, in Section 4.2. A space time analysis of the problem, will allow to directly transfer all the ideas of the abstract setting to the hyperbolic case. We just mention, that for parabolic problems the same analysis applies, as outlined in [81], see also [79, 80]. As optimal control problems usually aim to minimize convex functionals, they bear a close relation to systems of saddle point structure, as discussed, e.g., in the tremendous overview by Bochev and Gunzberger [14, 15]. In Chapter 5 we will exploit this relation, to solve the direct formulation of space time PDEs. In many cases a direct discretization of space time variational formulations leads to the cumbersome task of finding stable pairs of trial spaces. Recent work by Führer and Karkulik [43] shows the capacity of least squares methods, as they overcome certain, unpleasant restrictions in the choice of approximation spaces and come with inbuilt error estimators. However, to be of practical use, their formulation is based on a first order reformulation of the problem in order to keep the error estimator in localizable norm. This, in turns, comes with the disadvantage of introducing additional DoFs and convergence can only be shown, when assuming higher regularity on the source term. Although, there are already some workarounds available, see [45], our approach will be different, in the sense that we can directly set up a practical least squares/saddle point problem, using the Riesz lift of the residuum, which does neither need any additional regularity assumptions, nor to introduce unpleasant DoFs. Using the framework developed for the abstract analysis of optimal control problems in Chapter 3, we will be able to give a full stability and error analysis and prove efficiency and reliability of the inbuilt error estimator. To show the capacity of the method, we consider its application to the wave equation, though many more applications are covered.

2 Preliminaries

2.1 Functional analytical background

For our analysis, we will consider a well-established setting of spaces, having a certain structure. The name orgininates from the „rigged Hilbert spaces“ introduced by Gel’fand and Vilenkin [51, Chapter 4.2]. Though, we will give a slightly more general definition.

Definition 2.1 (Gelfand triple [42, Definition 4]). *We call a triple of spaces $X \subset H \subset X^*$ a Gelfand triple if $(X, \|\cdot\|_X)$ is a separable Banach space, which is dense in some Hilbert space H, where X^* denotes the dual space of X with respect to H.*

The following theorems are well-established results, concerning the existence and uniqueness of solutions of variational formulations. In the case of elliptic differential equations, this dates back to Lax and Milgram [83], see also [99] for a constructive proof. To formulate the result, we give a precise definition of the involved operator.

Definition 2.2. *Let $X \subset H \subset X^*$ be a Gelfand triple and let $T : X \to X^*$ be a linear operator. We say that T is*

- self-adjoint*, if*
$$\forall x, x' \in X : \langle Tx, x'\rangle_{X^*,X} = \langle Tx', x\rangle_{X^*,X},$$
- bounded*, if*
$$\exists c_2^T > 0 \, \forall x \in X : \|Tx\|_{X^*} \leq c_2^T \|x\|_X,$$
- X-elliptic*, if*
$$\exists c_1^T > 0 \, \forall x \in X : \langle Tx, x\rangle_{X^*,X} \geq c_1^T \|x\|_X^2.$$

Theorem 2.3 (Lax–Milgram [39, c.f. Lemma 2.2]). *Let X be a Hilbert space and let $T : X \to X^*$ be a linear, self-adjoint, bounded and X-elliptic operator. Then, for every $f \in X^*$ the problem to find $x_f \in X$ such that*

$$\langle Tx_f, x\rangle_{X^*,X} = \langle f, x\rangle_{X^*,X}, \quad \textit{for all } x \in X, \tag{2.1}$$

admits a unique solution, satisfying the a-priori estimate

$$\|x_f\|_X \leq \frac{1}{c_1^T} \|f\|_{X^*}.$$

We will also need a generalization of the Lemma of Lax–Milgram to the non-elliptic case. Its foundations were laid by Nečas [96] and it was popularized by Babuška [6, 7], thus we call it BN Theorem.

THEOREM 2.4 (Babuška-Nečas c.f. [39, Theorem 2.6]). *Let X be a Banach space and Y be a reflexive Banach space. Further, let $T : X \to Y^*$ be a bounded linear operator, i.e., there exists $c_2^T > 0$ for all $x \in X$ such that*

$$\|Tx\|_{Y^*} \leq c_2^T \|x\|_X,$$

with the following properties:

(BN1)

$$\exists c_1^T > 0 : \inf_{0 \neq x \in X} \sup_{0 \neq y \in Y} \frac{\langle Tx, y \rangle_{Y^*,Y}}{\|x\|_X \|y\|_Y} \geq c_1^T,$$

(BN2)

$$\forall y \in Y \setminus \{0\} \, \exists x_y \in X : \langle Tx_y, y \rangle_{Y^*,Y} \neq 0.$$

Then, for every $f \in Y^*$ the problem to find $x_f \in X$ such that

$$\langle Tx_f, y \rangle_{Y^*,Y} = \langle f, y \rangle_{Y^*,Y} \quad \text{for all } y \in Y,$$

admits a unique solution, satisfying the a-priori estimate

$$\|x_f\|_X \leq \frac{1}{c_1^T} \|f\|_{Y^*}.$$

A particular case of the generalized form of mixed ansatz and test spaces X and Y arises when considering saddle point problems. The structure of these problems allows for more precise, but equivalent conditions to (BN1)-(BN2), guaranteeing unique solvability, as stated in the next theorem.

THEOREM 2.5 ([39, Theorem 2.34]). *Let X and Y be reflexive Banach spaces and $f \in X^*$ and $g \in Y^*$ be given. Let $A : X \to X^*$ and $B : Y \to X^*$ be bounded linear operators, i.e., there exist $c_2^A, c_2^B > 0$ for all $q \in X$ and $z \in Y$ such that*

$$\|Aq\|_{X^*} \leq c_2^A \|q\|_X \quad \text{and} \quad \|Bz\|_{X^*} \leq c_2^B \|z\|_Y,$$

and let the following properties hold true:

(A-BN1)

$$\exists c_1^A > 0 : \inf_{0 \neq p \in ker(B^*)} \sup_{0 \neq q \in ker(B^*)} \frac{\langle Ap, q \rangle_{X^*,X}}{\|p\|_X \|q\|_X} \geq c_1^A,$$

(A-BN2)

$$\forall q \in ker(B^*) \setminus \{0\}\, \exists p_q \in ker(B^*) : \langle Ap_q, q\rangle_{X^*,X} \neq 0,$$

(B-BN1)

$$\exists c_1^B > 0 : \inf_{0\neq z\in Y} \sup_{0\neq q\in X} \frac{\langle Bz, q\rangle_{X^*,X}}{\|q\|_X\|z\|_Y} \geq c_1^B,$$

where $ker(B^*) := \{q \in X : \langle Bz, q\rangle_{X^*,X} = 0 \text{ for all } z \in Y\}$*. Then, the variational formulation to find* $(p, y) \in X \times Y$ *such that*

$$\begin{aligned} \langle Ap, q\rangle_{X^*,X} + \langle By, q\rangle_{X^*,X} &= \langle f, q\rangle_{X^*,X} \quad \text{for all } q \in X, \\ \langle B^*p, z\rangle_{Y^*,Y} &= \langle g, z\rangle_{Y^*,Y} \quad \text{for all } z \in Y, \end{aligned} \tag{2.2}$$

admits a unique solution and the following a-priori estimates hold true:

$$\begin{aligned} \|p\|_X &\leq \frac{1}{c_1^A}\|f\|_{X^*} + \frac{1}{c_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right)\|g\|_{Y^*} \\ \|y\|_Y &\leq \frac{1}{c_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right)\|f\|_{X^*} + \frac{c_2^A}{[c_1^B]^2}\left(1 + \frac{c_2^A}{c_1^A}\right)\|g\|_{Y^*}. \end{aligned}$$

Remark 2.6. *If* $A : X \to X^*$ *self-adjoint and* $ker(B^*)$*-elliptic or even* X*-elliptic then (A-BN1)-(A-BN2) are fulfilled.*

So far, we only treated variational formulations with equalities. We also want to consider the existence and uniqueness of solutions for a class of variational inequalities, which, to the best of our knowledge, originated from Lions and Stampacchia [86]. A good overview can be found in Glowinski [52].

Theorem 2.7 ([86, c.f. Theorem 2.1]). *Let* X *be a Hilbert space and* $K \subset X$ *a closed and convex subset. Let* $T : X \to X^*$ *be a linear, bounded and* X*-elliptic operator. Then for every* $f \in X^*$ *the problem to find* $x_f \in K$ *such that*

$$\langle Tx_f, x - x_f\rangle_{X^*,X} \geq \langle f, x - x_f\rangle_{X^*,X} \quad \text{for all } x \in K, \tag{2.3}$$

admits a unique solution and the map $f \to x_f$ *(in general non-linear) is continuous.*

Remark 2.8.
- *In the case of* $K \equiv X$*, the variational inequality* (2.3) *reduces to* (2.1) *and existence and uniqueness of solutions follows by the Lemma of Lax–Milgram (Theorem 2.3).*
- *If* $T : X \to X^*$ *is self-adjoint, then* (2.3) *is equivalent to find* $x_f \in K$ *minimizing*

  $$\mathcal{J}(x) = \frac{1}{2}\langle Tx, x\rangle_{X^*,X} - \langle f, x\rangle_{X^*,X}$$

  for given $f \in X^*$.

As known for elliptic partial differential equations, the regularity of the solution depends on the regularity of the source and properties of the domain, see, e.g., Lions, Magenes [85], Evans [40] for smoothly bounded domains or Grisvard [59] or Dauge [30] for more general domains. The following theorem reveals that the regularity of the solution of the variational inequality admits merely the same properties. The results date back to Brezis and Stampacchia and can be found in [19, 53].

THEOREM 2.9 ([19, c.f. Theorémè I.1, Remarque I.4, Remarque I.5]). *Let the assumptions of Theorem 2.7 hold true and let $X \subset H \subset X^*$ be a Gelfand triple. Then for $f \in H$ the unique solution $x_f \in K$ of (2.3) fulfills*

$$Tx_f \in H \quad \text{and} \quad \|Tx_f\|_H \leq C < \infty,$$

where $C = C(K, f)$.

In the following we study the structure of elliptic operators in more detail. In particular, they can be used to define equivalent norms. This will later on be used to define interpolation spaces.

LEMMA 2.10. *Let X be a Hilbert space and let $T : X \to X^*$ be a linear, self-adjoint, bounded and X-elliptic operator. Then $\|x\|_T := \sqrt{\langle Tx, x\rangle_{X^*,X}}$ induces an equivalent norm on X. More precisely,*

$$\sqrt{c_1^T}\|x\|_X \leq \|x\|_T \leq \sqrt{c_2^T}\|x\|_X \quad \text{for all } x \in X. \tag{2.4}$$

Furthermore, for all $f \in X^*$, $\|f\|_{X^*,T} := \sqrt{\langle f, T^{-1}f\rangle_{X^*,X}}$ induces an equivalent norm to

$$\|f\|_{X^*} = \sup_{0 \neq x \in X} \frac{\langle f, x\rangle_{X^*,X}}{\|x\|_X},$$

i.e.,

$$\frac{1}{\sqrt{c_2^T}}\|f\|_{X^*} \leq \|f\|_{X^*,T} \leq \frac{1}{\sqrt{c_1^T}}\|f\|_{X^*} \quad \text{for all } f \in X^*. \tag{2.5}$$

Proof. The first estimate in (2.4) follows directly from the ellipticity condition. The second estimate follows from duality and the boundedness, i.e.,

$$\langle Tx, x\rangle_{X^*,X} \leq \|Tx\|_{X^*}\|x\|_X \leq c_2^T\|x\|_X^2,$$

and we conclude (2.4) by taking the square root. By the Lemma of Lax–Milgram (Theorem 2.3), we know that there exists a unique solution $x_f \in X$ of

$$Tx_f = f \quad \text{in } X^*,$$

and that $\|x_f\|_X \leq \frac{1}{c_1^T}\|f\|_{X^*}$. Thus, we get

$$\langle f, T^{-1}f\rangle_{X^*,X} = \langle f, x_f\rangle_{X^*,X} \leq \|f\|_{X^*}\|x_f\|_X \leq \frac{1}{c_1^T}\|f\|_{X^*}^2.$$

Moreover, using the definition of the dual norm and (2.4), we estimate

$$\begin{aligned}\|f\|_{X^*} &= \sup_{0\neq x\in X}\frac{\langle f, x\rangle_{X^*,X}}{\|x\|_X} = \sup_{0\neq x\in X}\frac{\langle Tx_f, x\rangle_{X^*,X}}{\|x\|_X} \\ &\leq \sup_{0\neq x\in X}\frac{\|x_f\|_T\|x\|_T}{\|x\|_X} \leq \sqrt{c_2^T}\|x_f\|_T.\end{aligned}$$

Now, we conclude (2.5) by

$$\langle f, T^{-1}f\rangle_{X^*,X} = \langle Tx_f, x_f\rangle_{X^*,X} = \|x_f\|_T^2 \geq \frac{1}{c_2^T}\|f\|_{X^*}^2,$$

and taking the square root. □

Using the spectral decomposition of elliptic operators, we are can define interpolation spaces.

Lemma 2.11 ([85, cf. Definition 2.1, Proposition 2.3]). *Let $X_1 \subset X_0$ be seperable Hilbert spaces, such that X_1 is dense in X_0 and continuously embedded, i.e.,*

$$\exists c_X > 0 : \|x\|_{X_0} \leq c_X\|x\|_{X_1} \quad \text{for all } x \in X_1.$$

Further, define the operator $S : dom(S) \subset X_0 \to X_0$ as

$$\langle Sx, y\rangle_{X_0} = \langle x, y\rangle_{X_1} \quad \text{for all } x, y \in X_1,$$

where $dom(S) := \{z \in X_0 : x \to \langle z, x\rangle_{X_1} \text{is continuous for all } x \in X_1\}$. Then, for $\theta \in [0, 1]$, the interpolation space defined as

$$X_\theta = [X_0, X_1]_\theta := dom(S^{\theta/2}),$$

is a Hilbert space when endowed with the graph norm

$$\|x\|_{X_\theta} := \sqrt{\|x\|_{X_0}^2 + \|S^{\theta/2}x\|_{X_0}}.$$

Morover, for all $x \in X_1$ there holds

$$\|x\|_{X_\theta} \leq c\|x\|_{X_1}^\theta\|x\|_{X_0}^{1-\theta}.$$

COROLLARY 2.12. *Let $X \subset H \subset X^*$ be a Gelfand triple and let $T : X \to X^*$ be a linear, self-adjoint, bounded and X-elliptic operator. Then the interpolation space X_θ for $\theta \in [0,1]$ defined by*

$$X_\theta = [H, X]_\theta := dom(T^{\theta/2}),$$

is a Hilbert space equipped with the graph norm

$$\|x\|_{X_\theta} := \sqrt{\|x\|_H^2 + \|T^{\theta/2}x\|_H^2}.$$

In particular, $H = X_0$ and $X = X_1$. Moreover, for all $x \in X$ there holds

$$\|x\|_{X_\theta} \leq c\|x\|_X^\theta \|x\|_H^{1-\theta}.$$

We will also need mapping properties of operators defined on the interpolation spaces. The next theorem shows under which assumptions we can conclude boundedness.

THEOREM 2.13 ([85, cf. Théorème 5.1]). *Let X_0, X_1, Y_0 and Y_1 be Hilbert spaces, such that $X_0 \subset X_1$ and $Y_0 \subset Y_1$ are dense with respect to continuous injection. Let $T : X_i \to Y_i$ be a bounded linear operator for $i = 0, 1$. Then*

$$T : [X_0, X_1]_\theta \to [Y_0, Y_1]_\theta$$

is a bounded linear operator, for all $\theta \in [0,1]$.

2.2 Sobolev spaces

We briefly introduce the function spaces, we will use for the analysis of the variational (in-)equalities, stated in the previous section. For further reading we refer to [1, 11, 89].

Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded Lipschitz-domain, i.e., its boundary can be locally parametrized by Lipschitz continuous functions, see, e.g., [105, Definition 2.1]. In order to give a self-contained introduction, we introduce the spaces by completion of infinitely differentiable functions in their corresponding norm. Let $\alpha \in \mathbb{N}_0^d$ be a multiindex, with $|\alpha| := \alpha_1 + \ldots + \alpha_d$ and let $D^\alpha v(x) = (\partial_{x_1})^{\alpha_1} \ldots (\partial_{x_d})^{\alpha_d} v(x)$ for $x \in \Omega$. Let $\mathcal{C}(\Omega)$ denote the space of continuous functions $v : \Omega \to \mathbb{R}$ and define

$$\begin{aligned}
\mathcal{C}^\infty(\Omega) &:= \{v \in \mathcal{C}(\Omega) : D^\alpha v \in \mathcal{C}(\Omega), \text{for all } \alpha \in \mathbb{N}_0^d\}, \\
\mathcal{C}_0^\infty(\Omega) &:= \{v \in \mathcal{C}^\infty(\Omega) : \operatorname{supp}(v) \subset\subset \Omega\},
\end{aligned}$$

where $\operatorname{supp}(v) := \overline{\{x \in \Omega : v(x) \neq 0\}}$. In particular, for $v \in \mathcal{C}_0^\infty(\Omega)$ it holds $v(x) = 0$ for all $x \in \partial\Omega$. Now, let us consider functions $v : \Omega \to \mathbb{R}$, which are square integrable, and define the norm

$$\|v\|_{L^2(\Omega)} := \left(\int_\Omega |v(x)|^2 \, dx \right)^{1/2} < \infty.$$

Then, we define

$$L^2(\Omega) := \overline{\mathcal{C}^\infty(\Omega)}^{\|\cdot\|_{L^2(\Omega)}} = \overline{\mathcal{C}_0^\infty(\Omega)}^{\|\cdot\|_{L^2(\Omega)}},$$

and note that $L^2(\Omega)$ is a Hilbert space endowed with the inner product

$$\langle u, v \rangle_{L^2(\Omega)} := \int_\Omega u(x)v(x) \, dx \quad \text{for all } u, v \in L^2(\Omega).$$

We further introduce the spaces of square integrable functions admitting generalized derivatives of order $m \in \mathbb{N}$. Therefore, consider the norm

$$\|v\|_{H^m(\Omega)} := \left(\sum_{|\alpha| \le m} \|D^\alpha v\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

Then we define the spaces

$$H^m(\Omega) := \overline{\mathcal{C}^\infty(\Omega)}^{\|\cdot\|_{H^m(\Omega)}} \quad \text{and} \quad H_0^m(\Omega) := \overline{\mathcal{C}_0^\infty(\Omega)}^{\|\cdot\|_{H^m(\Omega)}}.$$

Again $H^m(\Omega)$ and $H_0^m(\Omega)$ are Hilbert spaces, endowed with the inner product

$$\langle u, v \rangle_{H^m(\Omega)} := \sum_{|\alpha| \le m} \langle D^\alpha u, D^\alpha v \rangle_{L^2(\Omega)} \quad \text{for all } u, v \in H^m(\Omega).$$

Moreover, consider $0 < s = m + \sigma$, where $m \in \mathbb{N}_0$ and $\sigma \in (0,1)$ and define the Sobolev-Slobodeckii norm

$$\|v\|_{H^s(\Omega)} := \sqrt{\|v\|_{H^m(\Omega)}^2 + |v|_{H^s(\Omega)}^2},$$

where

$$|v|_{H^s(\Omega)} := \left(\sum_{|\alpha| = m} \int_\Omega \int_\Omega \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x-y|^{d+2\sigma}} \, dx \, dy \right)^{1/2}.$$

Then, the intermediate spaces are given as

$$H^s(\Omega) := \overline{\mathcal{C}^\infty(\Omega)}^{\|\cdot\|_{H^s(\Omega)}} \quad \text{and} \quad H_0^s(\Omega) := \overline{\mathcal{C}_0^\infty(\Omega)}^{\|\cdot\|_{H^s(\Omega)}}.$$

In particular, for $s = 1$, there holds the Poincaré inequality,

$$\|v\|_{L^2(\Omega)} \leq c_P \|\nabla v\|_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega), \tag{2.6}$$

and we can define an equivalent innner product for $u, v \in H_0^1(\Omega)$

$$\langle u, v\rangle_{H_0^1(\Omega)} := \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx.$$

For $s > 0$, the space $H^{-s}(\Omega)$ is defined, as the dual space of $H_0^s(\Omega)$, with norm

$$\|f\|_{H^{-s}(\Omega)} := \sup_{0 \neq v \in H_0^s(\Omega)} \frac{\langle f, v\rangle_\Omega}{\|v\|_{H^s(\Omega)}},$$

where $\langle\cdot,\cdot\rangle_\Omega$ denotes the duality paring, as extension of the L^2-inner product.

Properties: Due to the inclusion $\mathcal{C}_0^\infty(\Omega) \subset \mathcal{C}^\infty(\Omega)$ we have that

$$H_0^s(\Omega) \subset H^s(\Omega) \quad \text{for all } s \geq 0.$$

Further, due to the definition of the norm, we have that

$$H^s(\Omega) \subset H^r(\Omega) \subset L^2(\Omega), \quad \text{for all } s > r > 0.$$

If $2m > d$, there holds by the Sobolev embedding theorem, e.g., [1, Theorem 5.4],

$$H^{j+m}(\Omega) \subset \mathcal{C}^j(\overline{\Omega}), \quad \text{for all } j \in \mathbb{N}_0,$$

i.e., the functions are continuous up to the boundary of Ω and pointwise evaluation is well-defined. Moreover, the inclusion $H^s(\Omega) \hookrightarrow L^2(\Omega)$ is continuous for all $s > 0$, as

$$\|v\|_{L^2(\Omega)} \leq \|v\|_{H^s(\Omega)} \quad \text{for all } v \in H^s(\Omega).$$

For the dual spaces, we have by standard properties of duality, that

$$L^2(\Omega) \subset H^{-r}(\Omega) \subset H^{-s}(\Omega) \quad \text{for all } s > r > 0.$$

By definiton, we see that $H^s(\Omega)$ and $H_0^s(\Omega)$ are dense subspaces of $L^2(\Omega)$. In particular, for all $s > 0$,

$$H^{-s}(\Omega) \subset L^2(\Omega) \subset H_0^s(\Omega)$$

is a Gelfand triple. Thanks to the assumptions on the domain Ω, the intermediate spaces $H^s(\Omega)$ and $H_0^s(\Omega)$, $0 < s < m$, $m \in \mathbb{N}$, can be equivalently characterized by the interpolation spaces, see, e.g., [11, 89], with $\theta = \frac{s}{m} \in (0, 1)$,

$$H^s(\Omega) := [L^2(\Omega), H^m(\Omega)]_\theta \quad \text{and} \quad H_0^s(\Omega) := [L^2(\Omega), H_0^m(\Omega)]_\theta.$$

The abstract boundedness of operators of Theorem 2.13 can be transferred to Sobolev spaces as follows.

Theorem 2.14. *Let $0 \leq r_0 \leq r_1$ and $0 \leq s_0 \leq s_1$ be given. Assume that the operator $T : H^{r_i}(\Omega) \to H^{s_i}(\Omega)$ is bounded for $i = 0, 1$, i.e., there exists constants $M_i > 0$ such that*

$$\|Tu\|_{H^{s_i}(\Omega)} \leq M_i \|u\|_{H^{r_i}(\Omega)} \quad \textit{for all } u \in H^{r_i}(\Omega) \textit{ and } i = 0, 1.$$

Then $T : H^{(1-\theta)r_0+\theta r_1}(\Omega) \to H^{(1-\theta)s_0+\theta s_1}(\Omega)$ is bounded for all $\theta \in [0, 1]$ and it holds

$$\|Tu\|_{H^{(1-\theta)s_0+\theta s_1}(\Omega)} \leq c M_0^{1-\theta} M_1^{\theta} \|u\|_{H^{(1-\theta)r_0+\theta r_1}(\Omega)}.$$

Proof. The statement can be proven defining equivalent norms on Sobolev spaces using the K-method. This goes beyond the scope of this work. For details we refer to McLean [89, Theorem B.2]. □

2.3 Sobolev spaces in the space time domain

Let $0 < T < \infty$ and $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a Lipschitz domain. In order to define Sobolev spaces on the space time cylinder $Q := \Omega \times (0, T)$, we will first introduce the concept of Bochner spaces, see, e.g., [26, 85, 116]. We will briefly state the main results. Let X be a seperable, real Hilbert space. Then we define

$$L^2(0, T; X) := \left\{ v : (0, T) \to X \text{ measurable w.r.t } dt : \int_0^T \|v(t)\|_X^2 \, dt < \infty \right\}.$$

As in the case of spatial Sobolev spaces, we can ask for generalized derivatives in time, by defining

$$H^m(0, T; X) := \{ v \in L^2(0, T; X) : (\partial_t)^k v \in L^2(0, T; X), \forall 0 \leq k \leq m \}.$$

The intermediate spaces for $0 < s < m$, $m \in \mathbb{N}$, are defined by function space interpolation as

$$H^s(0, T; X) := [L^2(0, T; X), H^m(0, T; X)]_\theta, \quad \theta = \frac{s}{m} \in (0, 1).$$

Now, for $r, s \geq 0$, we can define the anisotropic Sobolev spaces

$$H^{r,s}(Q) := L^2(0, T; H^r(\Omega)) \cap H^s(0, T; L^2(\Omega)),$$

which are Hilbert spaces, endowed with the inner product

$$\langle u, v \rangle_{H^{r,s}(Q)} := \int_0^T \langle u(t, \cdot), v(t, \cdot) \rangle_{H^r(\Omega)} \, dt + \int_\Omega \langle u(\cdot, x), v(\cdot, x) \rangle_{H^s(0,T)} \, dx.$$

To incorporate boundary conditions on the spatial boundary, we consider the space

$$H^{r,s}_{0;}(Q) := L^2(0,T;H^r_0(\Omega)) \cap H^s(0,T;L^2(\Omega)).$$

Additionally, we will also need initial and terminal conditions. For $s > \frac{1}{2}$ recall, that $H^s(0,T) \subset \mathcal{C}([0,T])$. Thus, the spaces

$$H^s_{0,}(0,T) := \{v \in H^s(0,T) : v(0) = 0\} \text{ and } H^s_{,0}(0,T) := \{v \in H^s(0,T) : v(T) = 0\}$$

are well-defined for $s > \frac{1}{2}$. Similarly, $H^s(0,T;L^2(\Omega)) \subset \mathcal{C}([0,T];L^2(\Omega))$ for $s > \frac{1}{2}$, and the spaces

$$\begin{aligned} H^s_{0,}(0,T;L^2(\Omega)) &:= \{v \in H^s(0,T;L^2(\Omega)) : v(0) = 0 \text{ in } L^2(\Omega)\}, \\ H^s_{,0}(0,T;L^2(\Omega)) &:= \{v \in H^s(0,T;L^2(\Omega)) : v(T) = 0 \text{ in } L^2(\Omega)\}, \end{aligned}$$

are well-defined. With this reasoning, we define the spaces incorporating spatial and intitial/terminal conditions for $r \geq 0$ and $s > \frac{1}{2}$ by

$$\begin{aligned} H^{r,s}_{0;0,}(Q) &:= L^2(0,T;H^r_0(\Omega)) \cap H^s_{0,}(0,T;L^2(\Omega)) \\ H^{r,s}_{0;,0}(Q) &:= L^2(0,T;H^r_0(\Omega)) \cap H^s_{,0}(0,T;L^2(\Omega)). \end{aligned}$$

In particular, for $s = r = 1$, $H^{1,1}_{0;0,}(Q)$ and $H^{1,1}_{0;,0}(Q)$ are Hilbert spaces, which, due to the Poincarè inequality, can be endowed with the inner product

$$\langle u, v\rangle_{H^1(Q)} := \langle \partial_t u, \partial_t v\rangle_{L^2(Q)} + \langle \nabla_x u, \nabla_x v\rangle_{L^2(Q)}$$

and the induced norm

$$|v|_{H^1(Q)} = \sqrt{\langle v, v\rangle_{H^1(Q)}}.$$

The dual spaces, $[H^{1,1}_{0;0,}(Q)]^*$ and $[H^{1,1}_{0;,0}(Q)]^*$ are characterized by completion of $L^2(Q)$ with respect to the norms

$$\begin{aligned} \|g\|_{[H^{1,1}_{0;0,}(Q)]^*} &:= \sup_{0 \neq v \in H^{1,1}_{0;0,}(Q)} \frac{\langle g, v\rangle_Q}{|v|_{H^1(Q)}}, \\ \|f\|_{[H^{1,1}_{0;,0}(Q)]^*} &:= \sup_{0 \neq v \in H^{1,1}_{0;,0}(Q)} \frac{\langle f, v\rangle_Q}{|v|_{H^1(Q)}}, \end{aligned}$$

where $\langle \cdot, \cdot\rangle_Q$ denotes the duality pairing, as extension of the $L^2(Q)$-inner product.

2.4 Approximation error estimates

The following section shall give an overview of results concerning a finite dimensional approximation of the infinite dimensional problems discussed in the preceeding section. We start with the well known results for elliptic problems.

THEOREM 2.15 ([105, Theorem 8.1 (Cea's Lemma)]). *Let the assumptions of Theorem 2.3 hold true and let $X_h \subset X$ be a conforming subspace. For the unique solution $x_{fh} \in X_h$ of*

$$\langle Tx_{fh}, x_h\rangle_{X^*,X} = \langle f, x_h\rangle_{X^*,X} \quad \textit{for all } x_h \in X_h,$$

there holds the stability estimate

$$\|x_{fh}\|_X \leq \frac{1}{c_1^T}\|f\|_{X^*}.$$

and the error estimate

$$\|x_f - x_{fh}\|_X \leq \frac{c_2^T}{c_1^T}\inf_{x_h \in X_h}\|x_f - x_h\|_X,$$

where $x_f \in X$ denotes the unique solution of (2.1).

For the approximation of variational inequalities, we can give a similiar result, depending on the approximation of the space and the set of constraints.

THEOREM 2.16 ([41, c.f. Theorem 1]). *Let $X \subset H \subset X^*$ be a Gelfand triple, $T : X \to X^*$ be a linear, self-adjoint, bounded and X-elliptic operator and let $f \in H$. Further, let $X_h \subset X$ be a conforming subspace and $K_h \subset X_h$ a closed and convex set. Then for the unique solution $x_{fh} \in K_h$ of*

$$\langle Tx_{fh}, x_h - x_{fh}\rangle_{X^*,X} \geq \langle f, x_h - x_{fh}\rangle_{X^*,X} \quad \textit{for all } x_h \in K_h$$

there holds the error estiamte

$$\|x_f - x_{fh}\|_T \leq \left\{\|x_f - x_h\|_T^2 + 2\|f - Tx_f\|_H\left[\|x_f - x_h\|_H + \|x_{fh} - x\|_H\right]\right\}^{1/2}$$

for all $x_h \in K_h$ and $x \in K$, where $x_f \in K$ denotes the unique solution of (2.3).

2.5 Approximation spaces

The quasi-optimal error estimates derived in the last section depend on the best approximation of the finite dimensional space X_h in X. In the following we will give a brief overview on the construction of such finite dimesionsal approximation spaces, using the finite element method, and state their approximation properties. For further reading we refer to [18, 23, 39, 105].

2.5.1 1D approximation spaces

Let $0 < L < \infty$ and consider the interval $I = (0, L)$. In order to define a finite dimensional approximation space, let $M \in \mathbb{N}$ be any number and consider the nodes $0 = x_0 < x_1 < \ldots < x_M = L$. Then, we can divide the interval into subintervals $\tau_k = (x_{k-1}, x_k)$, $k = 1, \ldots, M$, for which

$$\overline{I} = \bigcup_{k=1}^{M} \overline{\tau}_k$$

holds. We say that τ_k is an *element*, and denote by $h_k = |\tau_k| = x_k - x_{k-1}$ for $k = 1, \ldots, M$ and by $h = \max_{k=1,\ldots,M} h_k$ the local and global mesh size respectively. To achieve approximation spaces of arbitrary order, we are going to use the theory of B-splines, following [69]. Let us start with the approximation space of piecewise constant, globally discontinuous functions

$$S_h^0(0, L) = \left\{ v \in L^2(0, L) : \, v\big|_{[x_{k-1}, x_k)} \in \mathbb{P}_0([x_{k-1}, x_k)) \right\},$$

where $\mathbb{P}_0([x_{k-1}, x_k))$ is the space of all constant functions (polynomials of degree $p = 0$) on $[x_{k-1}, x_k)$. Obviously, this space is spanned by piecewise constant functions, i.e.,

$$S_h^0(0, L) := \operatorname{span}\{\varphi_{h,k}^0\}_{k=0}^{M-1}, \quad \text{with} \quad \varphi_{h,k}^0(x) = \begin{cases} 1, & x_k \le x < x_{k+1} \\ 0, & \text{else}, \end{cases}$$

and each $v_h \in S_h^0(0, L)$ admits the representation

$$v_h(x) = \sum_{k=0}^{M-1} v_k \varphi_{h,k}^0(x), \quad x \in (0, L),$$

with coefficients $(v_0, \ldots, v_{M-1}) = \mathbf{v}_h \in \mathbb{R}^M$. This allows to uniquely identify each function with a vector $S_h^0(0, L) \ni v_h \leftrightarrow \mathbf{v}_h \in \mathbb{R}^M$ and is called the finite element isomorphism. Note, that the basis functions are defined on half-open intervals $[x_k, x_{k+1})$ rather than on the elements $\tau_{k+1} = (x_k, x_{k+1})$. This ensures right continuity in the left point of the interval and will be necessary later on, to define functions of higher order and higher regularity. For ease of presentation we might drop the dependency of the basis funtions on h in the following, writing $\varphi_{h,k}^0 = \varphi_k^0$. Since functions in $S_h^0(0, L)$ are in general discontinuous, by the Sobolev imbedding $H^s(0, L) \subset \mathcal{C}([0, L])$ for $2s > 1$ we see that $S_h^0(0, L) \subset H^s(0, L)$ only for $s < \frac{1}{2}$. In order to achieve conformity also in Sobolev spaces of higher order, we will define B-splines of higher order, see, e.g., [31, 69]. The space of B-splines of degree $p \in \mathbb{N}$ will be denoted by

$S_h^p(0,L)$ and is spanned by basis functions φ_k^p, which can be contructed recursively as

$$\varphi_k^p(x) := \begin{cases} \frac{x-x_{k-p}}{x_k-x_{k-p}}\varphi_{k-1}^{p-1}(x) + \frac{x_{k+1}-x}{x_{k+1}-x_{k+1-p}}\varphi_k^{p-1}(x), & k=0,\dots,M+p-1, \\ 0, & \text{else}, \end{cases}$$

starting from φ_k^0. Here we use the convention that $x_{-j} = x_0$ and $x_{M+j} = x_M$ for all $j \in \mathbb{N}$ and fractions with zero denominators are zero. In order to avoid asymmetry, we enforce continuity at the right endpoint, i.e., we define

$$\varphi_k^p(L) = \lim_{\substack{x\to L \\ x<L}} \varphi_k^p(x), \quad k = 0,\dots,M+p-1.$$

In particular, for the B-splines of degree $p=1$ the basis functions $\{\varphi_k^1\}_{k=0}^M$ are given as

$$\begin{aligned}
\varphi_0^1(x) &= \frac{x_1 - x}{h_1}\varphi_0^0(x), \\
\varphi_k^1(x) &= \frac{x - x_{k-1}}{h_k}\varphi_{k-1}^0(x) + \frac{x_{k+1}-x}{h_{k+1}}\varphi_k^0(x), \quad k = 1,\dots,M-1, \\
\varphi_M^1(x) &= \frac{x - x_{M-1}}{h_M}\varphi_{M-1}^0(x).
\end{aligned}$$

Whereas for $p=2$ we get the basis functions $\{\varphi_k^2\}_{k=0}^{M+1}$ defined as

$$\begin{aligned}
\varphi_0^2(x) &= \frac{x_1 - x}{h_1}\varphi_0^1(x), \\
\varphi_1^2(x) &= \frac{x - x_0}{h_1}\varphi_0^1(x) + \frac{x_2 - x}{h_1 + h_2}\varphi_1^1(x), \\
\varphi_k^2(x) &= \frac{x - x_{k-2}}{h_{k-1}+h_k}\varphi_{k-1}^1(x) + \frac{x_{k+1}-x}{h_k + h_{k+1}}\varphi_k^1(x), \quad k = 2,\dots,M-1, \\
\varphi_M^2(x) &= \frac{x - x_{M-2}}{h_{M-1}+h_M}\varphi_{M-1}^1(x) + \frac{x_M - x}{h_M}\varphi_M^1(x), \\
\varphi_{M+1}^2(x) &= \frac{x - x_{M-1}}{h_M}\varphi_M^1(x).
\end{aligned}$$

THEOREM 2.17 ([69, c.f. Theorem 4]). *It holds that $S_h^p(0,L) \subset \mathcal{C}^{p-1}([0,L])$, $p \in \mathbb{N}$. In particular,*

$$S_h^1(0,L) \subset H^1(0,L) \quad \textit{and} \quad S_h^2(0,L) \subset H^2(0,L).$$

Remark 2.18.

- *It holds, see [69, Theorem 6], that*

$$S_h^p(0,L) = \Big\{ v \in \mathcal{C}^{p-1}([0,L]) : v\big|_{[x_{k-1},x_k)} \in \mathbb{P}_p([x_{k-1},x_k)),\ k=1,\ldots,M-1,$$
$$v\big|_{[x_{M-1},x_M]} \in \mathbb{P}_p([x_{M-1},x_M]) \Big\}.$$

- *For $p=1$ the basis functions φ_k^1 correspond to the Lagrangian basis functions/ hat functions*

$$\varphi_k^1(x) = \begin{cases} 1, & x = x_k, \\ 0, & x = x_i,\ i \neq k, \\ \textit{linear}, & \textit{else}, \end{cases}$$

and thus $S_h^1(0,L)$ is the well-known space of piecewise linear, globally continuous functions.

- *The support of the basis functions grows with the degree of the B-splines, i.e., supp$(\varphi_k^p) = [x_{k-p}, x_{k+1}]$, with the convention $x_{-j} = x_0$ and $x_{M+j} = x_M$ for $j \in \mathbb{N}$.*

- *The B-splines are non-negative and positive inside their support. Further, for each $x \in [0,L]$ the B-splines satisfy the partition of unity*

$$\sum_{k=0}^{M+p-1} \varphi_k^p(x) = 1, \quad \textit{for all } p \in \mathbb{N}_0,$$

see [69, p. 6, p. 21].

Remark 2.19. *Though, we are only considering functions on an interval, approximation spaces for higher dimensions can be defined by a tensor product structure, i.e., for an orthotope $R = (0,L_1) \times \ldots \times (0,L_d) \subset \mathbb{R}^d$, one can define the space*

$$S_h^p(R) := S_h^p(0,L_1) \otimes \ldots \otimes S_h^p(0,L_d), \quad p \in \mathbb{N}_0.$$

2.5.2 Approximation properties

In order to show how well a given function $v : (0,L) \to \mathbb{R}$ of sufficient regularity can be approximated using splines, we will construct an approximation $v_h \in S_h^p(0,L)$ by introducing a suitable quasi-interpolation, following the presentation in [69]. Before we proceed, we need an auxiliary result, concerning the approximability of functions by polynomials.

Lemma 2.20 ([69, cf Theorem 15]). *Let $v \in H^{m+1}([a,b])$, $m \in \mathbb{N}_0$, for some interval $[a,b] \subset [0,L]$, and let*

$$g(x) = T_m v(x;a) = \sum_{i=0}^{m} \frac{(x-a)^i}{i!} \partial_x^i v(a), \quad x \in [a,b],$$

be the Taylor polynomial of degree m in a. Then

$$\|\partial_x^j (v-g)\|_{L^2([a,b])} \leq c(m,j)(b-a)^{m+1-j} \|\partial_x^{m+1} v\|_{L^2([a,b])}, \quad j = 0,\dots,m.$$

Proof. By the integral representation of the remainder we have that

$$v(x) - g(x) = v(x) - T_m v(x;a) = \frac{1}{m!} \int_a^x (x-y)^m \partial_y^{m+1} v(y)\, dy, \quad x \in [a,b].$$

Differentiating and a Cauchy–Schwarz inequality gives

$$\begin{aligned} |\partial_x^j (v-g)(x)| &= \Big| \frac{1}{(m-j)!} \int_a^x (x-y)^{m-j} \partial_y^{m+1} v(y)\, dy \Big| \\ &\leq \frac{1}{(m-j)!} \Big(\int_a^x (x-y)^{2(m-j)}\, dy \Big)^{1/2} \|\partial_x^{m+1} v\|_{L^2([a,b])} \\ &\leq \frac{(b-a)^{m-j+1/2}}{(m-j)!(2(m-j)+1)^{1/2}} \|\partial_x^{m+1} v\|_{L^2([a,b])}. \end{aligned}$$

Now, after squaring and integrating over $[a,b]$ we arrive at

$$\|\partial_x^j (v-g)\|_{L^2([a,b])}^2 \leq c(m,j)(b-a)^{2(m-j+1)} \|\partial_x^{m+1} v\|_{L^2([a,b])}^2,$$

from which the result immediately follows. □

Let us introduce some useful notation before we proceed. For $k = 0,\dots,M$ recall that $\tau_k = (x_{k-1}, x_k)$ and consider $I_{k,p} = (x_{k-p-1}, x_{k+p})$ with the convention that $x_{-j} = x_0$ and $x_{M+j} = x_M$ for $j \in \mathbb{N}$. Moreover, let $h_{k,\min} := \arg\min_{\tau_\ell \subset I_{k,p}} |\tau_\ell|$ and similarly $h_{k,\max} := \arg\max_{\tau_\ell \subset I_{k,p}} |\tau_\ell|$. The next lemma shows, under which conditions a quasi-interpolation will give a good approximation.

Lemma 2.21. *Let $p \in \mathbb{N}_0$ and $\Pi_h^p : L^2(0,L) \to S_h^p(0,L)$ be a projection, i.e.,*

$$\Pi_h^p v_h = v_h \quad \text{for all } v_h \in S_h^p(0,L),$$

with the following properties:

(QI1) *There exist $c_\Pi > 0$ for all $v \in L^2(I_{k,p})$, $k = 0,\dots,M$, such that*

$$\|\partial_x^j \Pi_h^p v\|_{L^2(\tau_k)} \leq c_\Pi h_{k,min}^{-j} \|v\|_{L^2(I_{k,p})}, \quad j = 0,\dots,p.$$

(QI2) Π_h^p *reproduces* $\mathbb{P}_p$. *More precisely, for each* $g \in \mathbb{P}_p(I_{k,p})$, $k = 0, \dots, M$, *it holds*

$$(\Pi_h^p g)(x) = g(x), \quad x \in I_{k,p}.$$

Assume that $\frac{h_{k,max}}{h_{k,min}} \leq c_L$. *Then for* $v \in H^{m+1}(I_{k,p})$, $m = 0, \dots p$, *there holds*

$$\|\partial_x^j (v - \Pi_h^p v)\|_{L^2(\tau_k)} \leq c(p, m, j, c_\Pi, c_L) h_k^{m+1-j} \|\partial_x^{m+1} v\|_{L^2(I_{k,p})}, \quad j = 0, \dots, m.$$

Moreover, if $v \in H^{m+1}(0, L)$, $m = 0, \dots, p$, *we get*

$$\|\partial_x^j (v - \Pi_h^p v)\|_{L^2(0,L)} \leq c(p, m, j, c_\Pi, c_L) h^{m+1-j} \|\partial_x^{m+1} v\|_{L^2(0,L)}, \quad j = 0, \dots, m.$$

Proof. For any $g \in \mathbb{P}_p(I_{k,p})$ we compute for $j = 0, \dots, p$, using (QI1) and (QI2)

$$\begin{aligned} \|\partial_x^j (v - \Pi_h^p v)\|_{L^2(\tau_k)} &\leq \|\partial_x^j (v - g)\|_{L^2(\tau_k)} + \|\partial_x^j (\Pi_h^p (v - g))\|_{L^2(\tau_k)} \\ &\leq \|\partial_x^j (v - g)\|_{L^2(\tau_k)} + c_\Pi h_{k,\min}^{-j} \|v - g\|_{L^2(I_{k,p})}. \end{aligned}$$

By Lemma 2.20, choosing $g = T_m v(x; x_{k-p-1}) \in \mathbb{P}_m(I_{k,p})$ as the Taylor polynomial of degree m in x_{k-p-1}, we can bound the first term as

$$\|\partial_x^j (v - g)\|_{L^2(\tau_k)} \leq c(m, j) h_k^{m+1-j} \|\partial_x^{m+1} v\|_{L^2(\tau_k)}$$

and the second term by

$$\begin{aligned} \|v - g\|_{L^2(I_{k,p})} &\leq c(m)(x_{k+p} - x_{k-p-1})^{m+1} \|\partial_x^{m+1} v\|_{L^2(I_{k,p})} \\ &\leq c(m)(2p+1)^{m+1} h_{k,\max}^{m+1} \|\partial_x^{m+1} v\|_{L^2(I_{k,p})}. \end{aligned}$$

The estimates together with the assumption that $h_{k,\max} \leq c_L h_{k,\min}$ give the desired local bound. The global bound follows by squaring and summation. □

An operator fulfilling (QI1)-(QI2) can be constructed. We follow [69, Section 1.5.3.1]. Therefore, we define

$$\Pi_h^p v(x) = \sum_{k=0}^{M+p-1} \lambda_{k,p}(v) \varphi_k^p(x), \tag{2.7}$$

with coefficients

$$\lambda_{k,p}(v) = \frac{1}{x_{k+1} - x_{k-p}} \int_{x_{k-p}}^{x_{k+1}} \Big(\sum_{i=0}^{p} a_{k,i} \Big(\frac{x - x_{k-p}}{x_{k+1} - x_{k-p}} \Big)^i \Big) v(x)\, dx,$$

where $a_{k,i}$, $i = 0, \dots, p$, are chosen such that we have for each $j = k-p, \dots, k+1$

$$\lambda_{k,p}(\varphi_j^p) = \begin{cases} 1, & j = k \\ 0, & \text{else.} \end{cases} \tag{2.8}$$

Note, that the representation

$$\Big(\frac{x - x_{k-p}}{x_{k+1} - x_{k-p}}\Big)^i = \sum_{j=k-p}^{k+1} c_{k,i,j}\varphi_j^p(x), \quad x \in [x_{k-p}, x_{k+1}),\, i = 0, \dots, p,$$

holds true, and thus (2.8) is equivalent to require

$$\lambda_{k,p}\Big(\Big(\frac{x - x_{k-p}}{x_{k+1} - x_{k-p}}\Big)^i\Big) = c_{k,i,k}.$$

The operator Π_h^p has the following properties.

LEMMA 2.22. *Let $p \in \mathbb{N}_0$. The operator (2.7) is a well-defined projection for functions in $L^2(0,L)$, i.e., $\Pi_h^p : L^2(0,L) \to S_h^p(0,L)$ satisfies*

$$(\Pi_h^p\varphi_j^p)(x) = \sum_{k=0}^{M+p-1} \lambda_{k,p}(\varphi_j^p)\varphi_k^p(x) = \varphi_j^p(x) \quad \textit{for all } x \in [0,L],\, j = 0, \dots, M+p-1.$$

Moreover, it fulfills the assumptions (QI1)-(QI2).

Proof. The proof is technical and exceeds the scope of this work. A detailed proof can be found in [69, Section 1.5.3.1] □

We summarize our findings in the main theorem of this section.

THEOREM 2.23. *Let $p \in \mathbb{N}_0$ and let $v \in H^{m+1}(0,L)$ for $m = 0, \dots, p$. Then*

$$\inf_{v_h \in S_h^p(0,L)} \|\partial_x^j(v - v_h)\|_{L^2(0,L)} \le c h^{m+1-j}\|\partial_x^{m+1}v\|_{L^2(0,L)}, \quad j = 0, \dots, m,$$

holds, where $c > 0$ is a constant independent of the mesh size h.

We also want to have best approximation error estimates in the broken Sobolev-spaces $H^s(0,L)$, $s > 0$, $s \notin \mathbb{N}_0$. For this purpose we will first introduce projection operators, which realize the best approximation. Combined with the given approximation error estimate and the interpolation result of Theorem 2.14, we will then be able to derive the desired estimates.

We denote by $Q_h^p : L^2(0,L) \to S_h^p(0,L)$, $p \in \mathbb{N}_0$, the L^2-projection onto $S_h^p(0,L)$ defined as

$$\langle Q_h^p u, v_h\rangle_{L^2(0,L)} = \langle u, v_h\rangle_{L^2(0,L)}, \quad \text{for all } v_h \in S_h^p(0,L).$$

It admits the well-known properties.

LEMMA 2.24. *Let $u \in L^2(0, L)$. Then, the following properties hold true:*

(i) $\|Q_h^p u\|_{L^2(0,L)} \leq \|u\|_{L^2(0,L)}$

(ii) $\|u - Q_h^p u\|_{L^2(0,L)} \leq \|u\|_{L^2(0,L)}$

(iii) $\inf_{v_h \in S_h^p(0,L)} \|u - v_h\|_{L^2(0,L)} = \|u - Q_h^p u\|_{L^2(0,L)}$

We further denote by $P_h^p : H^1(0, L) \to S_h^p(0, L)$, $p \in \mathbb{N}$ the H^1-projection onto $S_h^p(0, L)$ defined as

$$\langle P_h^p u, v_h \rangle_{L^2(0,L)} + \langle \partial_x (P_h^p u), \partial_x v_h \rangle_{L^2(0,L)} = \langle u, v_h \rangle_{L^2(0,L)} + \langle \partial_x u, \partial_x v_h \rangle_{L^2(0,L)},$$

for all $v_h \in S_h^p(0, L)$. Note, that by Theorem 2.17 $S_h^p(0, L) \subset H^1(0, L)$, $p \geq 1$, and thus the projection is well-defined. As the L^2-projection, it admits the following, well-known properties.

LEMMA 2.25. *Let $u \in H^1(0, L)$. Then, the following properties hold true:*

(i) $\|P_h^p u\|_{H^1(0,L)} \leq \|u\|_{H^1(0,L)}$

(ii) $\|u - P_h^p u\|_{H^1(0,L)} \leq \|u\|_{H^1(0,L)}$

(iii) $\inf_{v_h \in S_h^p(0,L)} \|u - v_h\|_{H^1(0,L)} = \|u - P_h^p u\|_{H^1(0,L)}$

The interpolation error estimates, for a discretization that is fine enough, i.e., $h < 1$, are stated in the next theorem.

THEOREM 2.26. *Let the assumptions of Theorem 2.23 hold true. Assume that $u \in H^s(0, L)$ with $0 \leq s \leq p + 1$. Then,*

$$\inf_{v_h \in S_h^p(0,L)} \|u - v_h\|_{L^2(0,L)} = \|u - Q_h^p u\|_{L^2(0,L)} \leq ch^s \|u\|_{H^s(0,L)}. \tag{2.9}$$

Moreover, for $p \geq 1$ and $1 \leq s \leq p + 1$, we get

$$\inf_{v_h \in S_h^p(0,L)} \|u - v_h\|_{H^1(0,L)} = \|u - P_h^p u\|_{H^1(0,L)} \leq ch^{s-1} \|u\|_{H^s(0,L)}. \tag{2.10}$$

Proof. For $u \in L^2(0, L)$ we get by Lemma 2.24 (ii) that

$$\|u - Q_h^p u\|_{L^2(0,L)} \leq \|u\|_{L^2(0,L)}.$$

Further, for $u \in H^{p+1}(0, L)$, we have, using Lemma 2.24 (iii) and Theorem 2.23 that

$$\|u - Q_h^p u\|_{L^2(0,L)} = \inf_{v_h \in S_h^p(0,L)} \|u - v_h\|_{L^2(0,L)} \leq ch^{p+1} \|u\|_{H^{p+1}(0,L)}.$$

Thus, the operator $I - Q_h^p : H^{r_i}(0, L) \to L^2(0, L)$, $i = 0, 1$, is bounded for $0 = r_0 < r_1 = p + 1$. Now, with Theorem 2.14 we get that $I - Q_h^p : H^r(0, L) \to L^2(0, L)$ is bounded for all $s = (1 - \theta)r_0 + \theta r_1 = \theta(p + 1) \in [0, p + 1]$, for $\theta \in [0, 1]$, and

$$\|u - Q_h^p u\|_{L^2(0,L)} \leq ch^{(p+1)\theta}\|u\|_{H^s(0,L)} = ch^s\|u\|_{H^s(0,L)}$$

which gives (2.9). The inequality (2.10) can be derived analogously, using Lemma 2.25. □

2.5.3 Space(-time) approximation spaces

In the last section we discussed approximation spaces on intervals, which can be extended by tensor products to higher dimensions. For several reasons we will now discuss other discrete trial spaces that fulfill the same approximation properties as discussed in the one dimensional case, but can be defined on a simplicial decomposition of a d-dimensional domain. Firstly, for a more complex geometry an approximation with orthotopes might not be accurate. Secondly, when we want to define adaptive schemes, we will need to have local refinements, which are easily realizable on simplicial meshes. For further reading we refer to [23, 39, 105].

In the following let $\mathcal{D} \subset \mathbb{R}^n$, $n \in \mathbb{N}$, denote a bounded Lipschitz domain, and assume that $\mathcal{D}$ is polygonally ($n = 2$), polyhedrally ($n = 3$) or polychorally ($n = 4$) bounded. Further, let $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$ denote a decomposition of $\mathcal{D}$ into non-overlapping, simplicial elements $\tau_\ell \subset \mathbb{R}^n$, $\ell = 1, \ldots, N$, i.e.,

$$\overline{\mathcal{D}} = \bigcup_{\ell=1}^N \overline{\tau_\ell}, \quad \text{with} \quad \tau_\ell \cap \tau_k = \emptyset, \text{ for all } k \neq \ell.$$

A simplex $\tau_\ell \subset \mathbb{R}^n$ is defined as the interior of the convex hull of $n + 1$ vertices $\{x_i\}_{i=1}^{n+1}$, for which the vectors $\{x_2 - x_1, \ldots, x_{n+1} - x_1\}$ are linearly independent. The collection of all vertices are the nodes of the decomposition $\mathcal{T}_h$, which are denoted by $\{x_k\}_{k=1}^M$. Further, we denote by $h_\ell = |\tau_\ell|^{1/n}$, $\ell = 1, \ldots, N$ the local mesh size of each element and by $h = h_{\max} = \max_{\ell=1,\ldots,N} h_\ell$ the maximal global mesh size and by $h_{\min} = \max_{\ell=1,\ldots,N} h_\ell$ the minimal global mesh size. We assume that the decomposition fulfills the following properties:

(D1) (Admissibility) Neighboring elements share either a node $n = 1, 2, 3, 4$, an edge $n = 2, 3, 4$, a face $n = 3, 4$ or a tetraeder $n = 4$, i.e., we avoid „hanging nodes“.

(D2) (Shape regularity) We define the diameter of each element as

$$d_\ell := \sup_{x,y \in \tau_\ell} |x - y|$$

and its radius as

$$r_\ell := \arg\max\{r > 0 : \ B_r(x) \subset \tau_\ell, \text{for any } x \in \tau_\ell\},$$

i.e., the radius of the largest ball that can be inscribed in τ_ℓ. Then we say that $\mathcal{T}_h$ is shape regular if

$$d_\ell \le c_F r_\ell, \quad \text{for all } \ell = 1, \ldots, N,$$

with a constant independent of the decomposition $\mathcal{T}_h$.

We say that $\mathcal{T}_h$ is *globally quasi-uniform* if

$$\frac{h_{\max}}{h_{\min}} \le c_G,$$

for a constant $c_G > 0$ independent of h and $\mathcal{T}_h$ is *locally quasi-uniform* if

$$\frac{h_\ell}{h_k} \le c_L,$$

for all neighboring elements $\overline{\tau_\ell} \cap \overline{\tau_k} \neq \emptyset$ and a constant $c_L > 0$ independent of h.

Let us first consider the space of globally discontinuous, piecewise constant functions defined as

$$S_h^0(\mathcal{T}_h) := \{v_h \in L^2(\mathcal{D}) : \ v_h\big|_{\tau_\ell} \in \mathbb{P}_0(\tau_\ell) \text{ for all } \ell = 1, \ldots, N\},$$

where $\mathbb{P}_0(\tau_\ell)$ is the space of all constant functions on τ_ℓ. The space is obviously spanned by the functions $\{\varphi_\ell^0\}_{\ell=1}^N$, which are given as

$$\varphi_\ell^0(x) = \begin{cases} 1, & x \in \tau_\ell, \\ 0, & \text{else.} \end{cases}$$

As in the one dimensional case, by the Sobolev embedding theorem we see that $S_h^0(\mathcal{T}_h) \not\subset H^1(\mathcal{D})$. So let us define the approximation space of globally continuous, piecewise linear functions defined as

$$S_h^1(\mathcal{T}_h) := \left\{v_h \in \mathcal{C}(\overline{\mathcal{D}}) : \ v_h\big|_{\tau_\ell} \in \mathbb{P}_1(\tau_\ell) \text{ for all } \ell = 1, \ldots, N\right\},$$

where $\mathbb{P}_1(\tau_\ell)$ denotes the space of all polynomials of degree one on τ_ℓ.

Remark 2.27.

- *The hat functions* $\{\varphi_k^1\}_{k=1}^M$*, defined as*

$$\varphi_k^1(x) = \begin{cases} 1, & x = x_k, \\ 0, & x = x_i,\ i \neq k, \\ \textit{linear}, & \textit{else}, \end{cases}$$

  are a basis for $S_h^1(\mathcal{T}_h)$*, see e.g. [39, Proposition 1.78].*

- *It holds that* $S_h^1(\mathcal{T}_h) \subset H^1(\Omega)$*, see [39, Proposition 1.74].*
- *For* $d = 1$ *we have* $S_h^1(\mathcal{T}_h) = S_h^1(\Omega)$*, where* $\Omega = (0, L)$*.*

We now want to derive best approximation results as in the one-dimensional case. Let us first consider the nodal interpolation operator $I_h : \mathcal{C}(\overline{\mathcal{D}}) \to S_h^1(\mathcal{T}_h)$, defined as

$$I_h v(x) = \sum_{k=1}^{M} v(x_k)\varphi_k^1(x) \quad \text{for all } x \in \mathcal{D}.$$

Since, for $\mathcal{D} \subset \mathbb{R}^n$, $n = 1, 2, 3$ we have by the Sobolev embedding theorem that $H^2(\mathcal{D}) \subset \mathcal{C}(\overline{\mathcal{D}})$ we can state the following result.

Theorem 2.28. *Let* $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$ *be a locally quasi-uniform decomposition of* $\mathcal{D} \subset \mathbb{R}^n$ *for* $n = 1, 2, 3$ *and let* $v \in H^2(\tau_\ell)$, $\ell = 1, \ldots, N$*. Then* $I_h : H^2(\tau_\ell) \to S_h^1(\mathcal{T}_h)$ *is well-defined and*

$$\|v - I_h v\|_{L^2(\tau_\ell)}^2 + h_\ell^2 \|\nabla(v - I_h v)\|_{L^2(\tau_\ell)}^2 \leq c h_\ell^4 \sum_{|\alpha|=2} \|D^\alpha v\|_{L^2(\tau_\ell)}^2.$$

Moreover, if $v \in H^2(\mathcal{D})$*,* $I_h : H^2(\mathcal{D}) \to S_h^1(\mathcal{T}_h)$ *is well-defined and*

$$\|v - I_h v\|_{L^2(\mathcal{D})}^2 + h^2 \|\nabla(v - I_h v)\|_{L^2(\mathcal{D})}^2 \leq c h^4 \sum_{|\alpha|=2} \|D^\alpha v\|_{L^2(\mathcal{D})}^2.$$

Proof. As $H^2(\omega) \subset \mathcal{C}(\overline{\omega})$ for all $\omega = \tau_\ell$ or $\omega = \mathcal{D}$ if $n = 1, 2, 3$, the interpolation of $I_h v \in S_h^1(\mathcal{T}_h)$ of $v \in H^2(\omega)$ is well-defined and the error estimates are well-known, see [39, Remark 1.105] or [105, Lemma 9.9]. □

The previous Lemma just covers the case for $n = 1, 2, 3$, as for $n = 4$ the nodal interpolation is not a well-defined operator for functions in $H^2(\mathcal{D})$. To obtain similar best approximation estimates, we will construct a quasi-interpolation operator. Before we proceed, let us introduce some useful notation. For an element τ_ℓ, $\ell = 1, \ldots, N$, define the element patch as

$$\overline{\omega}_{\tau_\ell} := \bigcup_{\{j=1,\ldots,N:\, \overline{\tau_\ell}\cap\overline{\tau_j}\neq\emptyset\}} \overline{\tau}_j, \quad \ell = 1, \ldots, N.$$

Moreover, for $v \in H^j(\mathcal{D})$, $j \in \mathbb{N}_0$, we denote the $H^j(\mathcal{D})$ semi-norm by

$$|v|_{H^j(\mathcal{D})} := \sqrt{\sum_{|\alpha|=j} \|D^\alpha v\|^2_{L^2(\mathcal{D})}}.$$

For sufficiently regular functions we have the following approximation property of polynomials.

Lemma 2.29. *Assume that $\mathcal{T}_h$ is locally quasi-uniform. Let $v \in H^{m+1}(\omega_\ell)$, $m \in \mathbb{N}_0$, for $\omega_\ell = \tau_\ell$ or $\omega_\ell = \omega_{\tau_\ell}$, $\ell = 1, \ldots, N$. Then*

$$\inf_{g\in\mathbb{P}_m(\omega_\ell)} |v - g|_{H^j(\omega_\ell)} \leq c(n, m, c_L, c_F) h_\ell^{m+1-j} |v|_{H^{m+1}(\omega_\ell)}, \quad j = 0, \ldots, m.$$

Proof. The result can be proved applying a generalized version of the Bramble-Hilbert Lemma given in [36, Theorem 7.1]. The details for the application to this specific case can be found in [103, p. 490]. □

Now, to get a grip on the approximation properties of a quasi-interpolation operator, we need an auxiliary result, following Ciarlet [23], which deals with the mapping of a reference setting. Therefore, let $\hat{\tau} \subset \mathbb{R}^n$, be an arbitrary but fixed simplex, called a *reference element*. Then, for each element τ_ℓ, $\ell = 1, \ldots, N$, there exists an invertible, affine-linear mapping $F_\ell : \hat{\tau} \to \tau_\ell$ given as

$$F_\ell(\hat{x}) = B_\ell \hat{x} + b_\ell, \quad B_\ell \in \mathbb{R}^{n\times n},\ b_\ell \in \mathbb{R}^n,$$

such that $F_\ell(\hat{\tau}) = \tau_\ell$.

Lemma 2.30. *If $v \in H^j(\tau_\ell)$ for some $j \geq 0$ and $\ell = 1, \ldots, N$, then $\hat{v} := v \circ F_\ell$ satisfies $\hat{v} \in H^j(\hat{\tau})$ and, in addition, there exists a constant $c = c(j, n, c_F, \hat{\tau})$ such that*

$$|\hat{v}|_{H^j(\hat{\tau})} \leq c h_\ell^j |\det(B_\ell)|^{-1/2} |v|_{H^j(\tau_\ell)}. \tag{2.11}$$

and

$$|v|_{H^j(\tau_\ell)} \leq c h_\ell^{-j} |\det(B_\ell)|^{1/2} |\hat{v}|_{H^j(\hat{\tau})}. \tag{2.12}$$

Proof. The mapping properties

$$|\hat{v}|_{H^j(\hat{\tau})} \leq c \|B_\ell\|_2^j |\det(B_\ell)|^{-1/2} |v|_{H^j(\tau_\ell)}.$$

and

$$|v|_{H^j(\tau_\ell)} \leq c\|B_\ell^{-1}\|_2^j |\det(B_\ell)|^{1/2} |\hat{v}|_{H^j(\hat{\tau})},$$

are well-known, see, e.g., Ciarlet [23, Theorem 3.1.2]. Moreover, it holds, see [23, Theorem 3.1.3], that

$$\|B_\ell\|_2 \leq \frac{d_\ell}{\hat{r}} \quad \text{and} \quad \|B_\ell^{-1}\|_2 \leq \frac{\hat{d}}{r_\ell},$$

where

$$d_\ell = \sup_{x,y\in\tau_\ell} |x-y| \quad \text{and} \quad \hat{d} = \sup_{\hat{x},\hat{y}\in\hat{\tau}} |\hat{x}-\hat{y}|,$$

and the radii are defined as

$$\begin{aligned} r_\ell &:= \arg\max\{r>0 :\ B_r(x) \subset \tau_\ell, \text{for any } x \in \tau_\ell\}, \\ \hat{r} &:= \arg\max\{r>0 :\ B_r(\hat{x}) \subset \hat{\tau}, \text{for any } \hat{x} \in \hat{\tau}\}. \end{aligned}$$

Now, note that a ball $B_r(x) \subset \mathbb{R}^n$ with radius $r > 0$, has the volume

$$|B_r(x)| = \begin{cases} 2r, & n=1, \\ \pi r^2, & n=2, \\ \frac{4}{3}\pi r^3, & n=3, \\ \frac{\pi^2 r^4}{2}, & n=4. \end{cases}$$

Recall, that the local mesh size of an element $\tau_\ell \subset \mathbb{R}^n$ is defined as $h_\ell = |\tau_\ell|^{1/n}$. When taking $(r_\ell, x_\ell) := \arg\max\{r>0,\ x\in\tau_\ell :\ B_r(x) \subset \tau_\ell\}$, we have that $|B_{r_\ell}(x_\ell)| \leq |\tau_\ell|$. By the assumption of shape regularity (D2), we also have the reverse direction, i.e., $|\tau_\ell| \leq d_\ell^n \leq c_F^n r_\ell^n \leq c(n, c_F)|B_{r_\ell}(x_\ell)|$ and thus the relation

$$c_1 h_\ell \leq r_\ell \leq c_2 h_\ell$$

for constants $c_i(n, c_F) > 0$, $i = 1, 2$. With this we can estimate

$$\|B_\ell\|_2 \leq \frac{d_\ell}{\hat{r}} \leq c_F c_2 \frac{h_\ell}{\hat{r}} \leq \hat{c} h_\ell \quad \text{and} \quad \|B_\ell^{-1}\|_2 \leq \frac{\hat{d}}{r_\ell} \leq c_1^{-1}\frac{\hat{d}}{h_\ell} \leq \hat{c} h_\ell^{-1},$$

which completes the proof. □

Now we can state the main result, for the construction of a quasi-interpolation operator.

LEMMA 2.31. *Assume that $\mathcal{T}_h$ is locally quasi-uniform and let $\omega_\ell = \tau_\ell$ or $\omega_\ell = \omega_{\tau_\ell}$, $\ell = 1, \ldots, N$. Let $p \in \mathbb{N}_0$ and $\Pi_h^p : L^2(\mathcal{D}) \to S_h^p(\mathcal{T}_h)$ be a projection, i.e.,*

$$\Pi_h^p v_h = v_h \quad \text{for all } v_h \in S_h^p(\mathcal{T}_h),$$

with the following properties:

(QI1-d) *For all $v \in L^2(\omega_\ell)$ there exist $c_\Pi > 0$, such that*

$$|\Pi_h^p v|_{H^j(\tau_\ell)} \le c_\Pi h_\ell^{-j} \|v\|_{L^2(\omega_\ell)}, \quad j = 0, \dots, p.$$

(QI2-d) Π_h^p *reproduces $\mathbb{P}_p$. More precisely, for each $g \in \mathbb{P}_p(\omega_\ell)$, it holds*

$$(\Pi_h^p g)(x) = g(x), \quad x \in \omega_\ell.$$

Then for $v \in H^{m+1}(\omega_\ell)$, $m = 0, \dots p$, there holds

$$|v - \Pi_h^p v|_{H^j(\tau_\ell)} \le c(n, m, c_\Pi, c_L, c_F) h_\ell^{m+1-j} |v|_{H^{m+1}(\omega_\ell)}, \quad j = 0, \dots, m.$$

Moreover, if $v \in H^{m+1}(\mathcal{D})$, $m = 0, \dots, p$, we get

$$|v - \Pi_h^p v|_{H^j(\mathcal{D})} \le c(n, m, c_\Pi, c_L, c_F) h^{m+1-j} |v|_{H^{m+1}(\mathcal{D})}, \quad j = 0, \dots, m.$$

Proof. Let $\ell = 1, \dots, N$ be arbitrary but fixed and let $v \in H^{m+1}(\omega_\ell)$. First, observe, that by combining (2.11)-(2.12), we get for all $j = 0, \dots, m$, that

$$|v|_{H^j(\tau_\ell)} \le c h_\ell^{-j} |\det(B_\ell)|^{1/2} |\hat{v}|_{H^j(\hat{\tau})} \le c h_\ell^{-j} |\det(B_\ell)|^{1/2} \sum_{i=0}^{m} |\hat{v}|_{H^i(\hat{\tau})} \le c \sum_{i=0}^{m} h_\ell^{i-j} |v|_{H^i(\tau_\ell)}.$$

Using the local quasi-uniformity, we can further deduce that

$$|v|_{H^j(\omega_\ell)} \le c(m) c_L \sum_{i=0}^{m} h_\ell^{i-j} |v|_{H^i(\omega_\ell)} \tag{2.13}$$

Now, for any $j = 0, \dots, m$ and any $g \in \mathbb{P}_m(\omega_\ell)$, using a triangle inequality together with (QI2-d), (QI1-d) and (2.13), we can estimate

$$\begin{aligned} |v - \Pi_h^p v|_{H^j(\tau_\ell)} &\le |v - g|_{H^j(\tau_\ell)} + |\Pi_h^p (v - g)|_{H^j(\tau_\ell)} \\ &\le |v - g|_{H^j(\omega_\ell)} + c_\Pi h_\ell^{-j} \|v - g\|_{L^2(\omega_\ell)} \\ &\le c(m, c_L, c_\Pi) \sum_{i=0}^{m} h_\ell^{i-j} |v - g|_{H^i(\omega_\ell)} \end{aligned}$$

and the assertion follows, when applying Lemma 2.29. □

In the following we construct quasi-interpolation operators, meeting the assumptions (QI1-d) and (QI2-d) for the approximation spaces $S_h^p(\mathcal{T}_h)$, $p = 0, 1$. For $S_h^0(\mathcal{T}_h)$ consider the L^2-projection $\Pi_h^0 := Q_h^0 : L^2(\mathcal{D}) \to S_h^0(\mathcal{T}_h)$, defined as

$$\langle Q_h^0 u, v_h \rangle_{L^2(\mathcal{D})} = \langle u, v_h \rangle_{L^2(\mathcal{D})}, \quad \text{for all } v_h \in S_h^0(\mathcal{T}_h). \tag{2.14}$$

The properties of the projection are summarized in the next lemma.

Lemma 2.32. *The operator* (2.14), *is well-defined. Furthermore, it is a projection onto* $S_h^0(\mathcal{T}_h)$ *and satisfies (QI1-d) and (QI2-d).*

Proof. Since the integral exist the operator is well-defined and the projection property can be shown easily. Moreover, since on each element τ_ℓ, $\ell = 1, \dots, N$, it holds that $(Q_h^0 u)(x) = c_\ell \varphi_\ell^0(x)$, $c_\ell = \frac{1}{|\tau_\ell|}\int_{\tau_\ell} u(x)\,dx$ we have that

$$\begin{aligned}
\|Q_h^0 u\|_{L^2(\tau_\ell)}^2 &= \langle Q_h^0 u, Q_h^0 u\rangle_{L^2(\tau_\ell)} = c_\ell \langle Q_h^0 u, \varphi_\ell^0\rangle_{L^2(\tau_\ell)} \\
&= c_\ell \langle Q_h^0 u, \varphi_\ell^0\rangle_{L^2(\mathcal{D})} = c_\ell \langle u, \varphi_\ell^0\rangle_{L^2(\mathcal{D})} \\
&= c_\ell \langle u, \varphi_\ell^0\rangle_{L^2(\tau_\ell)} = \langle u, Q_h^0 u\rangle_{L^2(\tau_\ell)} \\
&\le \|u\|_{L^2(\tau_\ell)} \|Q_h^0 u\|_{L^2(\tau_\ell)},
\end{aligned}$$

showing (QI1-d). Being a linear projection it holds $Q_h^0 \varphi_\ell^0 = \varphi_\ell^0$ for all $\ell = 1, \dots, N$ and (QI2-d) follows from the fact that each $g \in \mathbb{P}_0(\tau_\ell)$ can be written as $g(x) = c\varphi_\ell^0(x)$, $c \in \mathbb{R}$. □

For $S_h^1(\mathcal{T}_h)$ we will construct a quasi-interpolation operator, following Clément [25], see also [39, Section 1.6.1], [105, Section 9.4] and [3]. For a node x_k, $k = 1, \dots, M$, we call

$$\overline{\omega}_{x_k} := \bigcup_{\{\ell=1,\dots,M:\, x_k \in \overline{\tau}_\ell\}} \overline{\tau}_\ell, \quad k = 1, \dots, M,$$

the node patch and we define the local L^2-projections $Q_{h,k}^1 : L^2(\omega_{x_k}) \to S_h^1(\omega_{x_k})$ as

$$\langle Q_{h,k}^1 u, v_h\rangle_{L^2(\omega_{x_k})} = \langle u, v_h\rangle_{L^2(\omega_{x_k})} \quad \text{for all } v_h \in S_h^1(\omega_{x_k}).$$

Using the local projection operators, we now define the quasi-interpolation operator $\Pi_h^1 : L^2(\mathcal{D}) \to S_h^1(\mathcal{T}_h)$ as

$$(\Pi_h^1 u)(x) = \sum_{k=1}^{M} (Q_{h,k}^1 u)(x_k)\varphi_k^1(x), \quad x \in \mathcal{D}, \tag{2.15}$$

which has the following properties.

Lemma 2.33. *The operator* (2.15) *is well-defined. Moreover it is projection onto* $S_h^1(\mathcal{T}_h)$ *and satisfies (QI1-d) and (QI2-d).*

Proof. Since all the local projections $Q_{h,k}^1$ are well-defined for functions in $L^2(\mathcal{D})$, so is Π_h^1. The projection property $\Pi_h^1 v_h = v_h \in S_h^1(\mathcal{T}_h)$ is easy to check. Moreover, each $g \in \mathbb{P}_1(\omega_{x_k})$ can be written as $g(x) = \sum_{\{j:\, x_j \in \overline{\omega}_{x_k}\}} g(x_j)\varphi_j^1(x)$, $x \in \omega_{x_k}$ and thus $\Pi_h^1 g = g$, showing (QI2-d). To prove (QI1-d) first note, that by the inverse inequality in Lemma 2.37, for all $v_h \in S_h^1(\mathcal{T}_h)$ it holds that

$$\|v_h\|_{L^\infty(\tau_\ell)} \le c h_\ell^{-n/2} \|v_h\|_{L^2(\tau_\ell)}.$$

Therefore, using the boundedness of the local projection, we have for $x_k \in \tau_\ell$ that

$$|Q^1_{h,k}u(x_k)| \leq \|Q^1_{h,k}u\|_{L^\infty(\tau_\ell)} \leq ch_\ell^{-n/2}\|Q^1_{h,k}u\|_{L^2(\tau_\ell)} \leq ch_\ell^{-n/2}\|u\|_{L^2(\omega_{x_k})}, \qquad (2.16)$$

Moreover, since the shape functions satisfy $\varphi^1_k(x) \leq 1$ for all $x \in \mathcal{D}$, we can bound

$$\|\varphi^1_k\|_{L^2(\tau_\ell)} \leq |\tau_\ell|^{1/2} = h_\ell^{n/2}$$

and using an inverse inequality, see Lemma 2.37, we get

$$\|\nabla\varphi^1_k\|_{L^2(\tau_\ell)} \leq c_I h_\ell^{-1}\|\varphi^1_k\|_{L^2(\tau_\ell)} \leq c_I h_\ell^{n/2-1}$$

resulting overall in

$$|\varphi^1_k|_{H^j(\tau_\ell)} \leq ch_\ell^{n/2-j}, \quad j = 0, 1. \qquad (2.17)$$

Now, we can estimate, using (2.16) and (2.17),

$$\begin{aligned} |\Pi^1_h u|_{H^j(\tau_\ell)} &\leq \sum_{\{k=1,\ldots,M:\, x_k\in\overline{\tau}_\ell\}} |Q^1_{h,k}u(x_k)||\varphi^1_k|_{H^j(\tau_\ell)} \\ &\leq c \sum_{\{k=1,\ldots,M:\, x_k\in\overline{\tau}_\ell\}} h_\ell^{-n/2}h_\ell^{n/2-j}\|u\|_{L^2(\omega_{x_k})} \\ &\leq ch_\ell^{-j}\|u\|_{L^2(\omega_{\tau_\ell})}, \end{aligned}$$

for $j = 0, 1$, from which we can deduce the bound (QI1-d). This finishes the proof.□

REMARK 2.34. *The Clément quasi-interpolation* Π^1_h *defined in* (2.15) *is suitable for showing the approximation properties, but does not preserve boundary conditions. This problem can be avoided when using the Scott-Zhang interpolation operator given in [103], with the drawback that this operator is only well-defined for functions in* $H^1(\mathcal{D})$. *A quasi-interpolation of minimal regularity, preserving boundary conditions, was constructed by Bernardi [12].*

We can now summarize the approximation property of the spaces in higher dimensions $n \in \mathbb{N}$.

THEOREM 2.35. *Let* $\mathcal{T}_h$ *be a locally quasi-uniform decomposition of the domain* $\mathcal{D} \subset \mathbb{R}^n$, $n = 1, 2, 3, 4$. *Further, let* $p = 0, 1$ *and let* $v \in H^{m+1}(\mathcal{D})$ *for* $m = 0, \ldots, p$. *Then*

$$\inf_{v_h \in S^p_h(\mathcal{D})} |v - v_h|_{H^j(\mathcal{D})} \leq ch^{m+1-j}|v|_{H^{m+1}(\mathcal{D})}, \quad j = 0, \ldots, m,$$

holds, where $c > 0$ *is a constant independent of the mesh size* h.

In order to achieve estimates in Sobolev spaces of real order, we proceed as in the one dimensional case, by defining the global L^2-projection $Q_h^p : L^2(\mathcal{D}) \to S_h^p(\mathcal{T}_h)$, $p = 0, 1$ by

$$\langle Q_h^p u, v_h \rangle_{L^2(\mathcal{D})} = \langle u, v_h \rangle_{L^2(\mathcal{D})}, \quad \text{for all } v_h \in S_h^p(\mathcal{T}_h)$$

and the global H^1-projection $P_h^1 : H^1(\mathcal{D}) \to S_h^1(\mathcal{D})$ by

$$\langle P_h^1 u, v_h \rangle_{L^2(\mathcal{D})} + \langle \nabla(P_h^1 u), \nabla v_h \rangle_{L^2(\mathcal{D})} = \langle u, v_h \rangle_{L^2(\mathcal{D})} + \langle \nabla u, \nabla v_h \rangle_{L^2(\mathcal{D})},$$

for all $v_h \in S_h^1(\mathcal{T}_h)$. They admit the same properties as in the one dimensional case, see Lemma 2.24 and Lemma 2.25. Using a space interpolation argument and a discretization that is fine enough, i.e., $h < 1$, we can state the following theorem.

THEOREM 2.36. *Let $p = 0, 1$ and $u \in H^s(\mathcal{D})$ with $0 \le s \le p + 1$. Then*

$$\inf_{v_h \in S_h^p(\mathcal{T}_h)} \|u - v_h\|_{L^2(\mathcal{D})} = \|u - Q_h^p u\|_{L^2(\mathcal{D})} \le c h^s \|u\|_{H^s(\mathcal{D})}. \tag{2.18}$$

Moreover, let $u \in H^s(\mathcal{D})$ with $1 \le s \le 2$. Then

$$\inf_{v_h \in S_h^1(\mathcal{T}_h)} \|u - v_h\|_{H^1(\mathcal{D})} = \|u - P_h^1 u\|_{H^1(\mathcal{D})} \le c h^{s-1} \|u\|_{H^s(\mathcal{D})}. \tag{2.19}$$

Proof. The proof follows the lines of the one dimensional case in Theorem 2.26, using the approximation results of Theorem 2.35, the properties of the projections Q_h^1 and P_h^1 and the interpolation in Sobolev spaces (Theorem 2.14). □

2.5.4 Inverse inequalities

It is well-known that on finite dimensional spaces all norms are equivalent. The goal of on inverse inequality is to specify the equivalence constants for $S_h^1(\mathcal{T}_h)$ depending on the mesh size. This will be a powerful tool for the error analysis later on. We refer to [39, Section 1.7].

LEMMA 2.37. *Let $v_h \in S_h^1(\mathcal{T}_h)$. Then for each $\ell = 1, \ldots, N$ there hold the local inverse inequalities*

$$\|\nabla v_h\|_{L^2(\tau_\ell)} \le c_I h_\ell^{-1} \|v_h\|_{L^2(\tau_\ell)}.$$

and

$$\|v_h\|_{L^p(\tau_\ell)} \le c h_\ell^{n(\frac{1}{p} - \frac{1}{q})} \|v_h\|_{L^q(\tau_\ell)}, \quad p, q = 1, \ldots, \infty,$$

where

$$\|v_h\|_{L^p(\tau_\ell)} := \begin{cases} \left(\int_{\tau_\ell} |v_h(x)|^p \, dx \right)^{1/p}, & 1 \le p < \infty, \\ \sup_{x \in \tau_\ell} |v_h(x)|, & p = \infty. \end{cases}$$

Moreover, we can estimate

$$\|\nabla v_h\|^2_{L^2(\mathcal{D})} \leq c_I \sum_{\ell=1}^{N} h_\ell^{-2} \|v_h\|^2_{L^2(\tau_\ell)}.$$

If $\mathcal{T}_h$ is globally quasi-uniform there holds the global inverse inequality

$$\|\nabla v_h\|_{L^2(\mathcal{D})} \leq c_I h^{-1} \|v_h\|_{L^2(\mathcal{D})}.$$

Proof. The proof relies on the mapping properties to a reference setting, see Lemma 2.30, and the fact that in finite dimensions all norms are equivalent. Details can be found in [39, Lemma 1.138, Corollary 1.141]. See also [105, Lemma 9.6, Lemma 9.8] for sharp constants. □

3 A UNIFIED ANALYSIS FOR OPTIMAL CONTROL PROBLEMS WITH ENERGY REGULARIZATION

In this chapter we will give a unified framework to analyze optimal control problems, which admit a certain structure. To phrase the problem statement, let us consider the Gelfand triples

$$Y \subset H_Y \subset Y^* \quad \text{and} \quad X \subset H_X \subset X^*.$$

Then, for a given target $y_d \in H_Y$, a given cost parameter $\varrho > 0$ and an operator $B : Y \to X^*$, we want to minimze the cost functional

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{H_Y} + \frac{\varrho}{2}\|u_\varrho\|^2_{X^*}, \tag{3.1}$$

over all $y_\varrho \in Y$ and $u_\varrho \in X^*$ that fulfill

$$By_\varrho = u_\varrho \quad \text{in } X^*. \tag{3.2}$$

In general, we cannot compute the solution directly as $y_\varrho \in Y$ but $y_d \in H_Y \not\subset Y$. So, we aim to find a good approximation of the target y_d under acceptable costs for the control u_ϱ. Since $\operatorname{dom}(B) = Y \subset H_Y$, $B : H \to X^*$ is in general unbounded and the minimization of $\|y_d - y_\varrho\|^2_{H_Y}$ subject to (3.2) is ill-conditioned. This is well-known in inverse problems, see [38, 68]. A remedy is to add a regularization term, e.g. Tikhonov regularization, see [100], or a suitable norm of the state, see, e.g., [20]. We note, that the reconstructed state and control depend on the choice of the norm. For some applications it is favorable to have solutions of low regularity, which can be realized by using norms in Banach spaces and is known as directional sparsity, see, e.g., [63]. In our case, the cost term of the control $\frac{\varrho}{2}\|u_\varrho\|^2_{X^*}$ serves as a regularization and is crucial for the analysis of the problem. The cost parameter is then related to a regularization parameter, which should fulfill certain properties to have an optimal balance of the regularization and minimization term.

Let us assume that $B : Y \to X^*$ is an isomorphism. Then the norm equivalence

$$\|u_\varrho\|_{X^*} = \|By_\varrho\|_{X^*} \simeq \|y_\varrho\|_Y,$$

holds true. Thus, the operator induces a norm on the space of the control, the so called *energy norm*. Therefore, we call this setting the *energy regularization*.

Many problems can be casted into this framework, including energy regularization of elliptic distributed optimal control problems [17, 73, 80, 95], elliptic optimal control problems with boundary control [4, 54, 97] and parabolic distributed optimal control problems [79] as well as hyperbolic distributed optimal control problems [87, 91] in the context of space-time methods, to mention some of them. The rest of this chapter is structured as follows. In Section 3.1 we will analyze the continuous problem (3.1)-(3.2), stating the assumptions on the operator B in more detail. We will also discuss the relation of the regularity of the target y_d to the cost/regularization parameter $\varrho > 0$. Moreover, we will introduce a conforming discretization in finite dimensional subspaces $X_h \subset X$ and $Y_h \subset Y$ of the continuous setting, analyze unique solvability and derive quasi-optimal error estimates for the computable state $y_{\varrho h} \in Y_h$. The section ends with a discussion on a discrete reconstruction of the control $u_{\varrho H} \in U_H \subset X^*$ from a given, computed state $y_{\varrho h} \in Y_h$. In Section 3.2 we will discuss the handling of constraints in the abstract setting of optimal control problems with energy regularization. We will again start with the analysis of the continuous setting. In particular, we will derive analogous relations between the regularization parameter and the regularity of the target. We finally conclude, by analyzing the discrete setting.

3.1 The energy regularization

For ease of presentation, we will henceforth restrict ourselves to the case $H_X = H_Y = H$. As already mentioned in the introduction a crucial property of the energy regularization is to assume that $B : Y \to X^*$ is an isomorphism. In view of the BN-Theorem (Theorem 2.4), we will make the following assumptions:

Assumption 3.1. *Let $B : Y \to X^*$ fulfill the following properties:*

(B1) *(Boundedness)* $\exists c_2^B > 0 : \|By\|_{X^*} \leq c_2^B \|y\|_Y$ *for all* $y \in Y$,

(B2) *(Injectivity)* $\exists c_1^B > 0 : c_1^B \|y\|_Y \leq \sup_{0\neq q\in X} \frac{\langle By,q\rangle_H}{\|q\|_X}$ *for all* $y \in Y$,

(B3) *(Surjectivity)* $\forall q \in X \setminus \{0\}\, \exists y_q \in Y : \langle By_q, q\rangle_H \neq 0$.

Using the assumptions (B1) and (B2) we have for all $y \in Y$ that

$$c_1^B \|y\|_Y \leq \sup_{0\neq q\in Y} \frac{\langle By, q\rangle_H}{\|q\|_X} = \|By\|_{X^*} \leq c_2^B \|y\|_Y, \tag{3.3}$$

and thus $\|y\|_Y \simeq \|By\|_{X^*}$. As a next step we want to give a computable realization of the norm on X^*. Therefore, we introduce an operator $A : X \to X^*$, for which we make the following assumptions.

Assumption 3.2. *Let $A : X \to X^*$ fulfill the following properties:*

(A1) *(Boundedness)* $\exists c_2^A > 0 : \; \|Aq\|_{X^*} \leq c_2^A \|q\|_X \quad$ *for all* $q \in X$,

(A2) *(Self-adjointness)* $\langle Ap, q\rangle_H = \langle p, Aq\rangle_H \quad$ *for all* $p, q \in X$,

(A3) *(Ellipticity)* $\exists c_1^A > 0 : \; \langle Aq, q\rangle_H \geq c_1^A \|q\|_X^2 \quad$ *for all* $q \in X$.

By Lemma 2.10 we know that for all $q \in X$ and all $v \in X^*$

$$\|q\|_A := \sqrt{\langle Aq, q\rangle_H} \quad \text{and} \quad \|v\|_{A^{-1}} := \sqrt{\langle v, A^{-1}v\rangle_H},$$

define equivalent norms on X and X^* respectively, with norm equivalence constants

$$\sqrt{c_1^A}\|q\|_X \leq \|q\|_A \leq \sqrt{c_2^A}\|q\|_X \quad \text{and} \quad \frac{1}{\sqrt{c_2^A}}\|v\|_{X^*} \leq \|v\|_{A^{-1}} \leq \frac{1}{\sqrt{c_1^A}}\|v\|_{X^*}. \tag{3.4}$$

Now, using the operator constraint (3.2) and the fact, that $B : Y \to X^*$ is an isomorphism, we can replace $\|u_\varrho\|_{X^*} \simeq \|u_\varrho\|_{A^{-1}} = \|By_\varrho\|_{A^{-1}}$ and consider the reduced cost functional

$$\begin{aligned} \tilde{\mathcal{J}}(y_\varrho) &= \frac{1}{2}\|y_\varrho - y_d\|_H^2 + \frac{\varrho}{2}\|By_\varrho\|_{A^{-1}}^2 && (3.5)\\ &= \frac{1}{2}\langle y_\varrho - y_d, y_\varrho - y_d\rangle_H + \frac{\varrho}{2}\langle By_\varrho, A^{-1}By_\varrho\rangle_H. && (3.6) \end{aligned}$$

Then the problem is to find $y_\varrho \in Y$ such that

$$\tilde{\mathcal{J}}(y_\varrho) \leq \tilde{\mathcal{J}}(y) \quad \text{for all } y \in Y.$$

Thus, $y_\varrho \in Y$ must fulfill the gradient equation

$$\varrho B^* A^{-1} B y_\varrho + (y_\varrho - y_d) = 0 \quad \text{in } Y^*, \tag{3.7}$$

where $B^* : X \to Y^*$ is the formally adjoint operator of $B : Y \to X^*$.

Remark 3.3. *Since $B : Y \to X^*$ is an isomorphism, we can also consider the operator constraint* (3.2) *as $y_\varrho = B^{-1}u_\varrho$ and the reduced cost functional*

$$\begin{aligned} \hat{\mathcal{J}}(u_\varrho) &= \frac{1}{2}\|B^{-1}u_\varrho - y_d\|_H^2 + \frac{\varrho}{2}\|u_\varrho\|_{A^{-1}}^2 \\ &= \frac{1}{2}\langle B^{-1}u_\varrho - y_d, B^{-1}u_\varrho - y_d\rangle_H + \frac{\varrho}{2}\langle A^{-1}u_\varrho, u_\varrho\rangle_H. \end{aligned} \tag{3.8}$$

Then, we want to find $u_\varrho \in X^*$ fulfilling

$$\hat{\mathcal{J}}(u_\varrho) \leq \hat{\mathcal{J}}(v) \quad \textit{for all } v \in X^*,$$

which is characterized by the gradient equation

$$(B^*)^{-1}(B^{-1}u_\varrho - y_d) + \varrho A^{-1}u_\varrho = 0.$$

Defining $p_\varrho = (B^*)^{-1}(B^{-1}u_\varrho - y_d) \in X$ *we thus need to solve the optimality system*

$$By_\varrho = u_\varrho \text{ in } X^*, \quad B^*p_\varrho = y_\varrho - y_d \text{ in } Y^*, \quad p_\varrho + \varrho A^{-1}u_\varrho = 0 \text{ in } X, \tag{3.9}$$

consisting of the forward problem, the adjoint/backward problem and the gradient equation. Eliminating the control by $By_\varrho = u_\varrho$ *we derive the equivalent system, to find* $(p_\varrho, y_\varrho) \in X \times Y$ *such that*

$$\begin{array}{rrcll} \varrho^{-1}Ap_\varrho + & By_\varrho = & 0 & \text{in } X^*, \\ -B^*p_\varrho + & y_\varrho = & y_d & \text{in } Y^*, \end{array}$$

which reads in variational form

$$\begin{array}{rrcl} \varrho^{-1}\langle Ap_\varrho, q\rangle_H + & \langle By_\varrho, q\rangle_H = & 0 & \text{for all } q \in X, \\ -\langle B^*p_\varrho, z\rangle_H + & \langle y_\varrho, z\rangle_H = & \langle y_d, z\rangle_H, & \text{for all } z \in Y. \end{array} \tag{3.10}$$

Now, eliminating the adjoint state $p_\varrho = -\varrho A^{-1}By_\varrho$ *we arrive at the Schur complement system to find* $y_\varrho \in Y$ *such that*

$$\varrho B^*A^{-1}By_\varrho + y_\varrho = y_d \quad \text{in } Y^*,$$

which is again (3.7)*. Note, that this approach is of interest when considering control constraints.*

In order to show that a unique solution to (3.7) exists, let us consider the following auxiliary result.

LEMMA 3.4. *The operator* $S := B^*A^{-1}B : Y \to Y^*$ *is self-adjoint, bounded and* Y*-elliptic. In particular,* $\|y\|_S := \sqrt{\langle By, A^{-1}By\rangle_H}$ *defines an equivalent norm on* Y*, i.e.,*

$$\sqrt{c_1^S}\|y\|_Y \le \|y\|_S \le \sqrt{c_2^S}\|y\|_Y,$$

with constants $c_1^S = \frac{[c_1^B]^2}{c_2^A}$ *and* $c_2^S = \frac{[c_2^B]^2}{c_1^A}$.

Proof. As A is self-adjoint by (A2), so is S. Using (3.3) and (3.4) we get for all $y \in Y$ that

$$\langle Sy, y\rangle_H = \langle By, A^{-1}By\rangle_H = \|By\|^2_{A^{-1}} \ge \frac{1}{c_2^A}\|By\|^2_{X^*} \ge \frac{[c_1^B]^2}{c_2^A}\|y\|^2_Y,$$

which gives ellipticity. For boundedness, consider

$$\begin{aligned}
\|Sy\|_{Y^*} &= \sup_{0\neq z\in Y}\frac{\langle Sy,z\rangle_H}{\|z\|_Y} = \sup_{0\neq z\in Y}\frac{\langle By,A^{-1}Bz\rangle_H}{\|z\|_Y}\\
&\leq \sup_{0\neq z\in Y}\frac{\|By\|_{A^{-1}}\|Bz\|_{A^{-1}}}{\|z\|_Y}\\
&\leq c_2^B\|By\|_{A^{-1}}\sup_{0\neq z\in Y}\frac{\|Bz\|_{A^{-1}}}{\|Bz\|_{X^*}}\\
&\leq \frac{c_2^B}{c_1^A}\|By\|_{X^*}\sup_{0\neq z\in Y}\frac{\|Bz\|_{X^*}}{\|Bz\|_{X^*}}\\
&\leq \frac{[c_2^B]^2}{c_1^A}\|y\|_Y,
\end{aligned}$$

which concludes the proof. □

Using the operator $S := B^*A^{-1}B : Y \to Y^*$, we consider (3.7) as variational formulation to find $y_\varrho \in Y$ such that

$$\varrho\langle Sy_\varrho, z\rangle_H + \langle y_\varrho, z\rangle_H = \langle y_d, z\rangle_H \quad \text{for all } z \in Y. \tag{3.11}$$

LEMMA 3.5. *Let $y_d \in H$ be given. Then* (3.11) *admits a unique solution $y_\varrho \in Y$ and the stability estimates*

$$\|y_\varrho\|_H \leq \|y_d\|_H \quad \textit{and} \quad \sqrt{\varrho}\|y_\varrho\|_S \leq \|y_d\|_H \tag{3.12}$$

hold true.

Proof. The unique solvability of (3.11) follows directly from the Y-ellipticity of $T := \varrho S + I : Y \to Y^*$, see Lemma 3.4, and the Lemma of Lax–Milgram (Theorem 2.3). For the estimates, consider (3.11) with test function $z = y_\varrho \in Y$. Then

$$\varrho\|y_\varrho\|_S^2 + \|y_\varrho\|_H^2 = \langle y_d, y_\varrho\rangle_H \leq \|y_d\|_H\|y_\varrho\|_H,$$

from which we first get

$$\|y_\varrho\|_H \leq \|y_d\|_H,$$

and subsequently deduce that

$$\sqrt{\varrho}\|y_\varrho\|_S \leq \|y_d\|_H. \qquad \square$$

Note, that for $\varrho \to 0$ the bound $\|y_\varrho\|_Y \leq \frac{1}{\sqrt{\varrho}}\|y_d\|_H$ explodes and we cannot expect that $y_\varrho \in Y$ if $y_d \in H$ but $y_d \notin Y$. Hence, the stabilization, i.e., $\varrho > 0$, is crucial in this case.

3.1.1 Regularization error estimates

In the following, we will derive error estimates for $\|y_\varrho - y_d\|_H$ depending on the regularization parameter ϱ and the regularity of the target y_d, as was done in [95] for elliptic problems, in [81] for parabolic problems and in [87] for hyperbolic problems.

LEMMA 3.6. *Let $y_d \in H$ be given. For the unique solution $y_\varrho \in Y$ of* (3.11) *there holds*

$$\|y_\varrho - y_d\|_H \le \|y_d\|_H. \tag{3.13}$$

Further, if $y_d \in Y$, then

$$\|y_\varrho - y_d\|_H \le \sqrt{\varrho}\|y_d\|_S \quad \text{and} \quad \|y_\varrho - y_d\|_S \le \|y_d\|_S. \tag{3.14}$$

Moreover, it holds

$$\|y_\varrho\|_S \le \|y_d\|_S. \tag{3.15}$$

At last, if $y_d \in Y$ such that $Sy_d \in H$ it holds

$$\|y_\varrho - y_d\|_H \le \varrho\|Sy_d\|_H \quad \text{and} \quad \|y_\varrho - y_d\|_S \le \sqrt{\varrho}\|Sy_d\|_H, \tag{3.16}$$

and, in this case we also have

$$\|Sy_\varrho\|_H \le \|Sy_d\|_H. \tag{3.17}$$

Proof. Testing (3.11) with $z = y_\varrho \in Y$ gives

$$\varrho\|y_\varrho\|_S^2 = \varrho\langle Sy_\varrho, y_\varrho\rangle_H = \langle y_d - y_\varrho, y_\varrho\rangle_H = \langle y_d - y_\varrho, y_d\rangle_H - \langle y_d - y_\varrho, y_d - y_\varrho\rangle_H$$

and reordering terms and using the Cauchy-Schwarz inequality gives

$$\varrho\|y_\varrho\|_S^2 + \|y_d - y_\varrho\|_H^2 = \langle y_d, y_d - y_\varrho\rangle_H \le \|y_d\|_H\|y_d - y_\varrho\|_H,$$

from which we conclude (3.13).

If $y_d \in Y$, we get with $z = y_d$ in (3.11), that $\langle y_d, y_d - y_\varrho\rangle_H = \varrho\langle y_d, Sy_\varrho\rangle_H$ and thus

$$\varrho\|y_\varrho\|_S^2 + \|y_d - y_\varrho\|_H^2 = \langle y_d, y_d - y_\varrho\rangle_H = \varrho\langle y_d, Sy_\varrho\rangle_H \le \varrho\|y_d\|_S\|y_\varrho\|_S,$$

showing (3.15). Moreover, choosing $z = y_d - y_\varrho \in Y$ in (3.11) we can compute

$$\begin{aligned}\|y_d - y_\varrho\|_H^2 &= \langle y_d - y_\varrho, y_d - y_\varrho\rangle_H = \varrho\langle Sy_\varrho, y_d - y_\varrho\rangle_H \\ &= \varrho\langle Sy_d, y_d - y_\varrho\rangle_H + \varrho\langle S(y_d - y_\varrho), y_\varrho - y_d\rangle_H.\end{aligned}$$

Again reordering gives

$$\|y_d - y_\varrho\|_H^2 + \varrho\|y_d - y_\varrho\|_S^2 = \varrho\langle Sy_d, y_d - y_\varrho\rangle_H.$$

and estimating $\langle Sy_d, y_d - y_\varrho\rangle_H \le \|y_d\|_S \|y_d - y_\varrho\|_H$ we conclude (3.14), i.e.,

$$\|y_d - y_\varrho\|_S \le \|y_d\|_S \quad \text{and} \quad \|y_d - y_\varrho\|_H \le \sqrt{\varrho}\|y_d\|_S.$$

If additionally $Sy_\varrho \in H$ we can estimate $\langle Sy_d, y_d - y_\varrho\rangle_H \le \|Sy_d\|_H \|y_d - y_\varrho\|_H$ to get (3.16), i.e.,

$$\|y_d - y_\varrho\|_H \le \varrho\|Sy_d\|_H \quad \text{and} \quad \|y_d - y_\varrho\|_S \le \sqrt{\varrho}\|Sy_d\|_H.$$

From (3.7) we know that $\varrho Sy_\varrho = y_d - y_\varrho \in H$ if $y_d \in H$ and therefore we have

$$\varrho\|Sy_\varrho\|_H = \|y_d - y_\varrho\|_H. \tag{3.18}$$

Together with (3.16), this shows (3.17). □

Corollary 3.7. *If we reconsider* (3.18)*, used in the previous proof, together with* (3.13) *and* (3.14)*, we get*

$$\|Sy_\varrho\|_H \le \begin{cases} \frac{1}{\varrho}\|y_d\|_H, & \text{if } y_d \in H, \\ \frac{1}{\sqrt{\varrho}}\|y_d\|_S, & \text{if } y_d \in Y. \end{cases} \tag{3.19}$$

Using the above results we can bound the objective functional, depending on the regularity of the target.

Corollary 3.8. *With the norm representation* $\|u_\varrho\|_{A^{-1}} = \|y_\varrho\|_S$ *and the stability bounds* (3.13)-(3.14) *and* (3.12) *we derive that*

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|_H^2 + \frac{\varrho}{2}\|u_\varrho\|_{A^{-1}}^2 \le \begin{cases} \|y_d\|_H^2, & \text{if } y_d \in H, \\ \varrho\|y_d\|_S^2, & \text{if } y_d \in Y. \end{cases}$$

We can also bound the adjoint state, depending on the regularization parameter.

Lemma 3.9. *Let* $y_d \in H$ *be given and let* $(p_\varrho, y_\varrho) \in X \times Y$ *be the unique solution of* (3.10)*. Then*

$$\|p_\varrho\|_X \le \sqrt{\frac{\varrho}{c_1^A}}\|y_\varrho - y_d\|_H^{1/2}\|y_d\|_H^{1/2}. \tag{3.20}$$

Proof. Choosing $z = y_\varrho$ and $q = p_\varrho$ in (3.10) gives

$$\frac{c_1^A}{\varrho}\|p_\varrho\|_X^2 = \frac{1}{\varrho}\langle Ap_\varrho, p_\varrho\rangle_H = -\langle By_\varrho, p_\varrho\rangle_H = -\langle B^* p_\varrho, y_\varrho\rangle_H = \langle y_d - y_\varrho, y_\varrho\rangle_H$$

and with a Cauchy-Schwarz inequality and (3.12) we conclude

$$c_1^A\|p_\varrho\|_Y^2 \le \varrho\|y_\varrho - y_d\|_H\|y_\varrho\|_H \le \varrho\|y_\varrho - y_d\|_H\|y_\varrho\|_H.$$ □

3.1.2 Discretization

We will now proceed with a discretized setting, that will later on be the starting point for the numerical computation of solutions of the optimal control problem. Let therefore $Y_h \subset Y$ denote a finite dimensional subspace and let us consider the Galerkin variational formulation to find $y_{\varrho h} \in Y_h$ such that

$$\varrho\langle S y_{\varrho h}, z_h\rangle_H + \langle y_{\varrho h}, z_h\rangle_H = \langle y_d, z_h\rangle_H, \quad \text{for all } z_h \in Y_h. \tag{3.21}$$

LEMMA 3.10. *The discrete variational formulation* (3.21) *admits a unique solution* $y_{\varrho h} \in Y_h$. *Moreover, the quasi-optimal error estimate (Cea's Lemma)*

$$\varrho\|y_\varrho - y_{\varrho h}\|_S^2 + \|y_\varrho - y_{\varrho h}\|_H^2 \le \inf_{z_h \in Y_h} \left[\varrho\|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2\right]$$

holds true.

Proof. The operator $T := \varrho S + I : Y \to Y^*$ is self-adjoint, bounded and Y-elliptic. Thus, the statement is a direct consequence of Lemma 2.15. □

The main statement of this section is the following theorem.

THEOREM 3.11. *Let* $y_d \in H$. *For the unique solution* $y_{\varrho h} \in Y_h$ *of* (3.21) *there holds the error estimate*

$$\|y_{\varrho h} - y_d\|_H \le \|y_d\|_H. \tag{3.22}$$

If additionally, $y_d \in Y$, *there holds*

$$\|y_{\varrho h} - y_d\|_H \le c\Big(\sqrt{\varrho}\|y_d\|_S + \inf_{z_h \in Y_h} \left[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\right]^{1/2}\Big) \tag{3.23}$$

and

$$\sqrt{\varrho}\|y_{\varrho h} - y_d\|_S \le c\Big(\sqrt{\varrho}\|y_d\|_S + \inf_{z_h \in Y_h} \left[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\right]^{1/2}\Big). \tag{3.24}$$

Moreover, if $y_d \in Y$ *and* $S y_d \in H$ *we have the error estimates*

$$\|y_{\varrho h} - y_d\|_H \le c\Big(\varrho\|S y_d\|_H + \inf_{z_h \in Y_h} \left[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\right]^{1/2}\Big) \tag{3.25}$$

and

$$\sqrt{\varrho}\|y_{\varrho h} - y_d\|_S \le c\Big(\varrho\|S y_d\|_H + \inf_{z_h \in Y_h} \left[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\right]^{1/2}\Big). \tag{3.26}$$

REMARK 3.12. *Note, that the assumptions $y_d \in Y$ and $Sy_d \in H$ are regularity assumptions on the target y_d. Therefore, the estimates of Theorem 3.11 give a connection of the regularization parameter $\varrho > 0$, the approximation property of $Y_h \subset Y$ and the regularity of the target.*

Proof. Let $y_d \in H$. Choosing $z_h = y_{\varrho h}$ in (3.21) we get that

$$\varrho\|y_{\varrho h}\|_S^2 = \varrho\langle Sy_{\varrho h}, y_{\varrho h}\rangle_H = \langle y_d - y_{\varrho h}, y_{\varrho h}\rangle_H = -\|y_d - y_{\varrho h}\|_H^2 + \langle y_d - y_{\varrho h} y_{\varrho h}\rangle_H.$$

Reordering and using a Cauchy–Schwarz inequalit we then get

$$\varrho\|y_{\varrho h}\|_S^2 + \|y_d - y_{\varrho h}\|_H^2 \leq \|y_d\|_H \|y_d - y_{\varrho h}\|_H,$$

which gives (3.22).

Let $y_d \in Y$. Adding and subtracting y_ϱ, using a triangle inequality and Hölders inequality, i.e., $(a+b)^2 \leq 2(a^2+b^2)$ we can estimate

$$\begin{aligned}
&\varrho\|y_{\varrho h} - y_d\|_S^2 + \|y_{\varrho h} - y_d\|_H^2 \\
&\quad \leq \varrho\big(\|y_{\varrho h} - y_\varrho\|_S + \|y_\varrho - y_d\|_S\big)^2 + \big(\|y_{\varrho h} - y_\varrho\|_H + \|y_\varrho - y_d\|_H\big)^2 \\
&\qquad \leq 2\left\{\varrho\|y_{\varrho h} - y_\varrho\|_S^2 + \|y_{\varrho h} - y_\varrho\|_H^2 + \varrho\|y_\varrho - y_d\|_S^2 + \|y_\varrho - y_d\|_H^2\right\}
\end{aligned}$$

Now, with Lemma 3.10 and adding and subtracting y_d we can estimate the first term

$$\begin{aligned}
&\varrho\|y_{\varrho h} - y_\varrho\|_S^2 + \|y_{\varrho h} - y_\varrho\|_H^2 \\
&\quad \leq \inf_{z_h \in Y_h} \left[\varrho\|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2\right] \\
&\qquad \leq 2\big(\varrho\|y_\varrho - y_d\|_S^2 + \|y_\varrho - y_d\|_H^2 + \inf_{z_h \in Y_h} \left[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\right]\big).
\end{aligned}$$

With the regularization error estimates of Lemma 3.6 (3.14) we thus get

$$\varrho\|y_{\varrho h} - y_d\|_S^2 + \|y_{\varrho h} - y_d\|_H^2 \leq 12\varrho\|y_d\|_S^2 + 4 \inf_{z_h \in Y_h} [\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2],$$

from which we get (3.23) and (3.24). Whereas, for $y_d \in Y$ such that $Sy_d \in H$ we can use Lemma 3.6 (3.16), to estimate

$$\varrho\|y_{\varrho h} - y_d\|_S^2 + \|y_{\varrho h} - y_d\|_H^2 \leq 12\varrho^2\|Sy_d\|_H^2 + 4 \inf_{z_h \in Y_h} [\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2],$$

which gives (3.25) and (3.26). □

As last part of this section, we are going to replace the continuous operator $S = B^*A^{-1}B : Y \to Y^*$ by a computable approximation and derive error estimates for the solution of the resulting variational formulation. Therefore, we define $\tilde{S} : Y \to Y^*$ by $\tilde{S}y = B^* p_{yh}$, where $p_{yh} \in X_h \subset X$ is the solution of

$$\langle A p_{yh}, q_h \rangle_H = \langle By, q_h \rangle_H \quad \text{for all } q_h \in X_h.$$

Then we consider the perturbed variational formulation to find $\tilde{y}_{\varrho h} \in Y_h$ such that

$$\varrho \langle \tilde{S} \tilde{y}_{\varrho h}, z_h \rangle_H + \langle \tilde{y}_{\varrho h}, z_h \rangle_H = \langle y_d, z_h \rangle_H, \quad \text{for all } z_h \in Y_h. \tag{3.27}$$

REMARK 3.13. *The computable approximation $\tilde{S}$ is exactly the Schur complement operator, when discretizing the system* (3.10) *directly, i.e.,* (3.21) *is equivalent to find $(p_{\varrho h}, \tilde{y}_{\varrho h}) \in X_h \times Y_h$ such that*

$$\begin{aligned} \varrho^{-1} \langle A p_{\varrho h}, q_h \rangle_H \quad &+ \langle B \tilde{y}_{\varrho_h}, q_h \rangle_H &&= 0 && \text{for all } q_h \in X_h, \\ - \langle B^* p_{\varrho h}, z_h \rangle_H \quad &+ \langle \tilde{y}_{\varrho h}, z_h \rangle_H &&= \langle y_d, z_h \rangle_H && \text{for all } z_h \in Y_h. \end{aligned} \tag{3.28}$$

LEMMA 3.14. *The perturbed operator $\tilde{S} : Y \to Y^*$ is self-adjoint, bounded and positive semi-definite. Furthermore, the perturbed variational formulation* (3.27) *admits a unique solution $\tilde{y}_{\varrho h} \in Y_h \subset H$.*

Proof. For arbitrary $z \in Y$, let $p_{zh} \in X_h$ be defined as unique solution of

$$\langle A p_{zh}, q_h \rangle_H = \langle Bz, q_h \rangle_H, \quad \text{for all } q_h \in Y_h.$$

In order to show that $\tilde{S}$ is self adjoint, take $y, z \in Y$ arbitrary but fixed. Using definition of $\tilde{S}$ and the self-adjointness of $A : X \to X^*$ (A2) we compute

$$\begin{aligned} \langle \tilde{S}y, z \rangle_H &= \langle B^* p_{yh}, z \rangle_H = \langle Bz, p_{yh} \rangle_H \\ &= \langle A p_{zh}, p_{yh} \rangle_H = \langle A p_{yh}, p_{zh} \rangle_H \\ &= \langle By, p_{zh} \rangle_H = \langle B^* p_{zh}, y \rangle_H \\ &= \langle \tilde{S}z, y \rangle_H. \end{aligned}$$

To show boundedness, first note that for all $z \in Y$

$$c_1^A \|p_{zh}\|_X^2 \le \langle A p_{zh}, p_{zh} \rangle_H = \langle Bz, p_{zh} \rangle_H \le c_2^B \|z\|_Y \|p_{zh}\|_X$$

implies that

$$\|p_{zh}\|_X \le \frac{c_2^B}{c_1^A} \|z\|_Y.$$

Thus, we can estimate

$$\begin{aligned}\|\tilde{S}z\|_{Y^*} &= \|B^* p_{zh}\|_{Y^*} = \sup_{0\neq y\in Y} \frac{\langle B^* p_{zh}, y\rangle_H}{\|y\|_Y} \\ &= \sup_{0\neq y\in Y} \frac{\langle By, p_{zh}\rangle_H}{\|y\|_Y} \leq c_2^B \|p_{zh}\|_X \leq \frac{[c_2^B]^2}{c_1^A} \|z\|_Y.\end{aligned}$$

Moreover, for arbitrary but fixed $z \in Y$ we compute that

$$\langle \tilde{S}z, z\rangle_H = \langle B^* p_{zh}, z\rangle_H = \langle Bz, p_{zh}\rangle_H = \langle Ap_{zh}, p_{zh}\rangle_H = \|p_{zh}\|_A^2 \geq 0,$$

which shows that $\tilde{S} \geq 0$. Unique solvability now follows, since $\tilde{T} := \varrho\tilde{S} + I \geq I$, which is at least H-elliptic. □

Using a Strang Lemma argument, we can give error estimates for the solution of the perturbed system (3.27), which now additionally depends on the approximation of the operator.

THEOREM 3.15. *Let $y_d \in H$. Then the unique solution $\tilde{y}_{\varrho h} \in Y_h$ of (3.27) admits the estimate*

$$\|\tilde{y}_{\varrho h} - y_d\|_H \leq \|y_d\|_H. \tag{3.29}$$

Let $y_d \in Y$ and let $p_{y_d} \in X$ be the unique solution of

$$\langle Ap_{y_d}, q\rangle_H = \langle By_d, q\rangle_H \quad \textit{for all } q \in X.$$

Further, assume that for all $z_h \in Y_h$ the inverse inequality

$$\|z_h\|_Y \leq c_I h^{-1} \|z_h\|_H, \tag{3.30}$$

holds. Then we get

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_d\|_H \leq c\Big(&\big[h^{-1}\varrho + \sqrt{\varrho}\big]\|y_d\|_S + h^{-1}\varrho \inf_{q_h\in X_h} \|p_{y_d} - q_h\|_X \\ &+ \big[h^{-1}\sqrt{\varrho} + 1\big] \inf_{z_h\in Y_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big).\end{aligned} \tag{3.31}$$

Moreover, if $y_d \in Y$ and $Sy_d \in H$ we have the error estimate

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_d\|_H \leq c\Big(&\big[h^{-1}\varrho^{3/2} + \varrho\big]\|Sy_d\|_H + h^{-1}\varrho \inf_{q_h\in X_h} \|p_{y_d} - q_h\|_X \\ &+ \big[h^{-1}\sqrt{\varrho} + 1\big] \inf_{z_h\in Y_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big).\end{aligned} \tag{3.32}$$

REMARK 3.16. *Note, that the inverse inequality* (3.30) *is defined for the full norm and thus, thinking of a discretization using finite elements, requires a globally quasi-uniform mesh. When applying an adaptive refinement though, the global quasi-uniformity will in general degenerate, even when starting with a globally quasi-uniform mesh. However, if Y and H are Sobolev spaces of positive order we can localize the norm and a local quasi-uniform mesh is sufficient to derive the estimate. Moreover, we are able to show quasi-optimal error estimates for adaptive schemes, when considering a mesh dependent regularization parameter ϱ for elliptic problems in Section 4.1.1.*

Proof. Choosing $z_h = \tilde{y}_{\varrho h} \in Y_h$ as test function in (3.27), we get

$$\begin{aligned}\varrho\langle \tilde{S}\tilde{y}_{\varrho h}, \tilde{y}_{\varrho h}\rangle_H &= \langle y_d - \tilde{y}_{\varrho h}, \tilde{y}_{\varrho h}\rangle_h \\ &= \langle y_d - \tilde{y}_{\varrho h}, y_d\rangle_H - \langle y_d - \tilde{y}_{\varrho h}, y_d - \tilde{y}_{\varrho h}\rangle_H.\end{aligned}$$

Reordering terms and using a Cauchy-Schwarz inequality we thus get

$$\varrho\langle \tilde{S}\tilde{y}_{\varrho h}, \tilde{y}_{\varrho h}\rangle_H + \|y_d - \tilde{y}_{\varrho h}\|_H^2 = \langle y_d - \tilde{y}_{\varrho h}, \tilde{y}_{\varrho h}\rangle_H \le \|y_d - \tilde{y}_{\varrho h}\|_H \|y_d\|_H,$$

which, using the positive-semidefiniteness of $\tilde{S}$, gives (3.29).

Subtracting the perturbed variational formulation (3.27) from (3.21), gives

$$\varrho\,\langle S y_{\varrho h} - \tilde{S}\tilde{y}_{\varrho h}, z_h\rangle_H + \langle y_{\varrho h} - \tilde{y}_{\varrho h}, z_h\rangle_H = 0 \quad \text{for all } z_h \in Y_h,$$

i.e.,

$$\varrho\,\langle (S - \widetilde{S}) y_{\varrho h}, z_h\rangle_H + \langle y_{\varrho h} - \tilde{y}_{\varrho h}, z_h\rangle_H = \varrho\,\langle \widetilde{S}(\tilde{y}_{\varrho h} - y_{\varrho h}), z_h\rangle_Q \quad \text{for all } z_h \in Y_h.$$

In particular for $z_h = \tilde{y}_{\varrho h} - y_{\varrho h}$ we further conclude

$$\begin{aligned}0 &\le \varrho\,\langle \tilde{S}(\tilde{y}_{\varrho h} - y_{\varrho h}), \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H \\ &= \varrho\,\langle (S - \tilde{S}) y_{\varrho h}, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H + \langle y_{\varrho h} - \tilde{y}_{\varrho h}, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H,\end{aligned}$$

i.e., reordering and using the assumed inverse inequality (3.30) in X_h,

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H^2 &\le \varrho\,\langle (S - \tilde{S}) y_{\varrho h}, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H \\ &= \varrho\,\langle B^*(p_{y_{\varrho h}} - p_{y_{\varrho h} h}), \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H \\ &= \varrho\,\langle p_{y_{\varrho h}} - p_{y_{\varrho h} h}, B(\tilde{y}_{\varrho h} - y_{\varrho h})\rangle_H \\ &\le \varrho c_2^B\,\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_Y \|p_{y_{\varrho h}} - p_{y_{\varrho h} h}\|_X \\ &\le c_I c_2^B\,\varrho\, h^{-1}\,\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H \|p_{y_{\varrho h}} - p_{y_{\varrho h} h}\|_X.\end{aligned} \tag{3.33}$$

Hence, with a triangle inequality, this gives

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H &\le c\,\varrho h^{-1}\,\|p_{y_{\varrho h}} - p_{y_{\varrho h} h}\|_X \\ &\le c\,\varrho h^{-1}\left[\|p_{y_{\varrho h}} - p_{y_d}\|_X + \|p_{y_d} - p_{y_d h}\|_X + \|p_{y_d h} - p_{y_{\varrho h} h}\|_X\right].\end{aligned} \tag{3.34}$$

For the first term we further have

$$\begin{aligned} c_1^A \|p_{y_{\varrho h}} - p_{y_d}\|_X^2 &\leq \langle A(p_{y_{\varrho h}} - p_{y_d}), p_{y_{\varrho h}} - p_{y_d}\rangle_H \\ &= \langle B(y_{\varrho h} - y_d), p_{y_{\varrho h}} - p_{y_d}\rangle_H \\ &\leq \|B(y_{\varrho h} - y_d)\|_{X^*} \|p_{y_{\varrho h}} - p_{y_d}\|_X \\ &\leq c_2^B \, \|y_{\varrho h} - y_d\|_X \|p_{y_{\varrho h}} - p_{y_d}\|_Y, \end{aligned}$$

i.e.,

$$\|p_{y_{\varrho h}} - p_{y_d}\|_X \leq \frac{c_2^B}{c_1^A} \|y_{\varrho h} - y_d\|_X \leq \frac{c_2^B}{c_1^A \sqrt{c_1^S}} \|y_{\varrho h} - y_d\|_S = \frac{c_2^B}{c_1^B} \frac{\sqrt{c_2^A}}{c_1^A} \|y_{\varrho h} - y_d\|_S.$$

Thus, with Theorem 3.11 (3.24) we conclude that for $y_d \in Y$

$$\|p_{y_{\varrho h}} - p_{y_d}\|_X \leq \frac{c}{\sqrt{\varrho}} \Big(\sqrt{\varrho} \|y_d\|_S + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2 \big]^{1/2} \Big)$$

and if additionally $Sy_d \in H$ holds, we get, using (3.26), that

$$\|p_{y_{\varrho h}} - p_{y_d}\|_X \leq \frac{c}{\sqrt{\varrho}} \Big(\varrho \|Sy_d\|_H + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2 \big]^{1/2} \Big).$$

Analogously, we can estimate the third term by

$$\|p_{y_d h} - p_{y_{\varrho h} h}\|_X \leq \frac{c}{\sqrt{\varrho}} \Big(\sqrt{\varrho} \|y_d\|_S + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2 \big]^{1/2} \Big),$$

if $y_d \in Y$, and if additionally $Sy_d \in H$ holds

$$\|p_{y_d h} - p_{y_{\varrho h} h}\|_X \leq \frac{c}{\sqrt{\varrho}} \Big(\varrho \|Sy_d\|_H + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2 \big]^{1/2} \Big).$$

To estimate the remaining term, let us first recall that $p_{y_d} \in X$ solves

$$\langle Ap_{y_d}, q\rangle_H = \langle By_d, q\rangle_H \quad \text{for all } q \in X,$$

while $p_{y_d h} \in X_h$ solves

$$\langle Ap_{y_d h}, q_h\rangle_H = \langle By_d, q_h\rangle_H \quad \text{for all } q_h \in X_h.$$

Thus, we conclude the Galerkin orthogonality

$$\langle A(p_{y_d} - p_{y_d h}), q_h\rangle_H = 0 \quad \text{for all } q_h \in X_h,$$

and Cea's lemma,

$$\|p_{y_d} - p_{y_d h}\|_X \le \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X.$$

Altogether, we have for $y_d \in Y$, that

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H \le ch^{-1}\Big(&\varrho\|y_d\|_S + \varrho \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X \\ &+ \sqrt{\varrho} \inf_{z_h \in Y_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big)\end{aligned}$$

and for $y_d \in Y$ such that $Sy_d \in H$ we get

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H \le ch^{-1}\Big(&\varrho^{3/2}\|Sy_d\|_H + \varrho \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X \\ &+ \sqrt{\varrho} \inf_{z_h \in Y_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big).\end{aligned}$$

Using the triangle inequality

$$\|\tilde{y}_{\varrho h} - y_d\|_H \le \|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H + \|y_{\varrho h} - y_d\|_H$$

together with the estimates (3.23) and (3.25) of Theorem 3.11 the proof is concluded. □

Corollary 3.17. *Reconsidering* (3.34) *in the previous proof, replacing* p_{y_d} *and* $p_{y_d h}$ *by* p_{y_ϱ} *and* $p_{y_\varrho h}$ *respectively, and redoing the same steps, we can prove the estimate*

$$\|\tilde{y}_{\varrho h} - y_\varrho\|_H \le c\varrho h^{-1}\Big(\|y_{\varrho h} - y_\varrho\|_S + \inf_{q_h \in X_h} \|p_{y_\varrho} - q_h\|_X\Big). \tag{3.35}$$

For the adjoint state $p_{\varrho h}$ we have an analogon to the continuous estimate in Lemma 3.9.

Lemma 3.18. *Let* $y_d \in H$ *be given and let* $(\tilde{y}_{\varrho h}, p_{\varrho h}) \in Y_h \times X_h$ *be the unique solution of* (3.28)*. Then*

$$\|p_{\varrho h}\|_X \le \sqrt{\frac{\varrho}{c_1^A}}\|\tilde{y}_{\varrho h} - y_d\|_H^{1/2}\|y_d\|_H^{1/2}. \tag{3.36}$$

Proof. The proof follows the lines of Lemma 3.9 choosing $z_h = \tilde{y}_{\varrho h}$ and $q_h = p_{\varrho h}$ in (3.28). □

Corollary 3.19. *From Lemma 3.9 and Lemma 3.18 and a triangle inequality we easily conclude*

$$\|p_\varrho - p_{\varrho h}\|_X \le \sqrt{\frac{\varrho}{c_1^A}}(\|y_\varrho - y_d\|_H^{1/2} + \|\tilde{y}_{\varrho h} - y_d\|_H^{1/2})\|y_d\|_H^{1/2}. \tag{3.37}$$

3.1.3 Reconstruction of the control

As an application of this framework, we have optimal control problems in mind. Thus, one is usually interested in the control u_ϱ, which is given as

$$u_\varrho = By_\varrho \quad \text{in } X^*. \tag{3.38}$$

Since $B : Y \to X^*$ is an isomorphism, the control is uniquely determined. For a computable reconstruction, we consider yet another finite dimensional conforming space $U_H \subset X^*$ and want to find $u_{\varrho H} \in U_H$ as solution of

$$\langle u_{\varrho H}, q_h \rangle_H = \langle B\tilde{y}_{\varrho h}, q_h \rangle_H, \text{ for all } q_h \in X_h. \tag{3.39}$$

Using the finite element isomorphism $U_H \ni u_{\varrho H} \leftrightarrow \mathbf{u}_{\varrho H} \in \mathbb{R}^{\dim(U_H)}$ and $Y_h \ni y_{\varrho h} \leftrightarrow \mathbf{y}_{\varrho h} \in \mathbb{R}^{\dim(Y_h)}$, the system matrix is a discretization of the identity, i.e., $\langle u_{\varrho H}, q_h \rangle_H = (\hat{M}_h \mathbf{u}_{\varrho H}, \mathbf{y}_{\varrho h})_2$. Though, in general, there does not exist a unique solution if $\dim(U_H) \neq \dim(Y_h)$. Thus, we reformulate (3.38) by equivalently finding the control $u_\varrho \in X^*$ as the minimizer of

$$u_\varrho = \arg\min_{v \in X^*} \frac{1}{2} \|v - By_\varrho\|_{X^*}^2 = \arg\min_{v \in X^*} \frac{1}{2} \langle v - By_\varrho, A^{-1}(v - By_\varrho) \rangle_H,$$

where we used the equivalent representation of the dual norm (3.4) induced by A^{-1}. Then, $u_\varrho \in X^*$ satisfies the gradient equation

$$A^{-1}(u_\varrho - By_\varrho) = 0.$$

Introducing $\hat{p} = A^{-1}(By_\varrho - u_\varrho) \in X$, noting that $\hat{p} = 0$, this is equivalent to the solution $(\hat{p}, u_\varrho) \in X \times X^*$ of the system

$$\langle A\hat{p}, q \rangle_H + \langle u_\varrho, q \rangle_H = \langle By_\varrho, q \rangle_H, \quad \langle v, \hat{p} \rangle_H = 0, \tag{3.40}$$

for all $(q, v) \in X \times X^*$.

Thus, the discrete reconstruction $\tilde{u}_{\varrho H} \in U_H$ can be computed as solution of the discretized saddle point formulation of (3.40) replacing y_ϱ by $\tilde{y}_{\varrho h}$, i.e., we seek to find the solution $(\hat{p}_h, \tilde{u}_{\varrho H}) \in X_h \times U_H$ of

$$\langle A\hat{p}_h, q_h \rangle_H + \langle \tilde{u}_{\varrho H}, q_h \rangle_H = \langle B\tilde{y}_{\varrho h}, q_h \rangle_H, \quad \langle v_H, \hat{p}_h \rangle_H = 0, \tag{3.41}$$

for all $(q_h, v_H) \in X_h \times U_H$. The unique solvability and errror estimates are given in the next theorem.

THEOREM 3.20. *Let the discrete inf-sup stability condition*

$$c_S \|v_H\|_{X^*} \leq \sup_{0 \neq q_h \in X_h} \frac{\langle v_H, q_h \rangle_H}{\|q_h\|_X} \quad \textit{for all } v_H \in U_H \tag{3.42}$$

hold true, for some $c_S > 0$. *Then the discrete variational formulation* (3.41) *admits a unique solution* $(\hat{p}_h, \tilde{u}_{\varrho H}) \in X_h \times U_H$ *and the estimate*

$$\|u_\varrho - \tilde{u}_{\varrho H}\|_{X^*} \leq \left(1 + \frac{1}{c_S}\left[1 + \frac{c_2^A}{c_1^A}\right]\right) \inf_{v_H \in U_H} \|u_\varrho - v_H\|_{X^*} + \frac{c_2^B}{c_S}\left[1 + \frac{c_2^A}{c_1^A}\right] \|\tilde{y}_{\varrho h} - y_\varrho\|_Y \tag{3.43}$$

holds, where $(\hat{p}, u_\varrho) \in X \times X^*$ *denotes the unique solution of* (3.40).

Proof. Unique solvability follows from Theorem 2.5, as $A : X \to X^*$ is elliptic on $X_h \subset X$ and the discrete inf-sup stability (3.42) is assumed to hold true. For the error estimate, first note that by a triangle inequality we have for arbitrary but fixed $v_H \in U_H$

$$\|u_\varrho - \tilde{u}_{\varrho H}\|_{X^*} \leq \|u_\varrho - v_H\|_{X^*} + \|v_H - \tilde{u}_{\varrho H}\|_{X^*}.$$

Further, using (3.42) and (3.40), (3.41) we estimate

$$\begin{aligned}
c_S \|v_H - \tilde{u}_{\varrho H}\|_{X^*} &\leq \sup_{0 \neq q_h \in X_h} \frac{\langle v_H - \tilde{u}_{\varrho H}, q_h \rangle_H}{\|q_h\|_X} \\
&= \sup_{0 \neq q_h \in X_h} \frac{\langle v_H, q_h \rangle_H - \langle B\tilde{y}_{\varrho h}, q_h \rangle_H + \langle A\hat{p}_h, q_h \rangle_H}{\|q_h\|_X} \\
&= \sup_{0 \neq q_h \in Y_h} \frac{\langle v_H - u_\varrho, q_h \rangle_H - \langle B(\tilde{y}_{\varrho h} - y_\varrho), q_h \rangle_H + \langle A\hat{p}_h, q_h \rangle_H}{\|q_h\|_Y} \\
&\leq \|v_H - u_\varrho\|_{X^*} + c_2^B \|\tilde{y}_{\varrho h} - y_\varrho\|_Y + c_2^A \|\hat{p}_h\|_X.
\end{aligned}$$

Now using (3.41), for $\|\hat{p}_h\|_X$ we have

$$\begin{aligned}
c_1^A \|\hat{p}_h\|_X^2 &\leq \langle A\hat{p}_h, \hat{p}_h \rangle_H = \langle B\tilde{y}_{\varrho h}, \hat{p}_h \rangle_H - \langle \tilde{u}_{\varrho H}, \hat{p}_h \rangle_H \\
&= \langle B(\tilde{y}_{\varrho h} - y_\varrho), \hat{p}_h \rangle_H + \langle u_\varrho - v_H, \hat{p}_h \rangle_H - \langle \tilde{u}_{\varrho H} - v_H, \hat{p}_h \rangle_H \\
&= \langle B(\tilde{y}_{\varrho h} - y_\varrho), \hat{p}_h \rangle_H + \langle u_\varrho - v_H, \hat{p}_h \rangle_H \\
&\leq \big(c_2^B \|\tilde{y}_{\varrho h} - y_\varrho\|_Y + \|u_\varrho - v_H\|_{X^*}\big) \|\hat{p}_h\|_X,
\end{aligned}$$

which gives $\|\hat{p}_h\|_X \leq \frac{c_2^B}{c_1^A} \|\tilde{y}_{\varrho h} - y_\varrho\|_Y + \frac{1}{c_1^A} \|u_\varrho - v_H\|_{X^*}$ and concludes the proof. □

REMARK 3.21. *Recall, that by the definition of the norm of* X^* *we have*

$$\|v_H\|_{X^*} = \sup_{0 \neq q \in X} \frac{\langle v_H, q \rangle_H}{\|q\|_X}.$$

Thus, the discrete inf-sup condition (3.42) *is fulfilled, if the discrete trial space* $Y_h \subset Y$ *is chosen rich enough with respect to the choice of* $U_H \subset X^*$.

3.2 State and control constraints

In the former section, we found the optimal state $y_\varrho \in Y$ and control $u_\varrho \in X^*$ minimizing the functional

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_\varrho - y_d\|_H^2 + \frac{\varrho}{2}\|u_\varrho\|_{X^*}^2, \tag{3.44}$$

subject to

$$By_\varrho = u_\varrho \quad \text{in } X^*, \tag{3.45}$$

for a given target $y_d \in H$ and $\varrho > 0$. Though, in many applications, either the state or the control admits some constraints. In this section we will consider such constrained optimal control problems and discuss how to handle them within the abstract framework. In order to incorporate constraints, we need to adapt the abstract setting, in the sense that barrier functions will make sense. Therefore, let the Gelfand triples $X \subset H \subset X^*$ and $Y \subset H \subset Y^*$ be given as before, but now assume that the duality is with respect to $H = L^2(\mathcal{D})$, where $\mathcal{D} = \Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ is either some bounded domain, or $\mathcal{D} = \Omega \times (0, T)$ is a space-time domain with finite time horizon $T < \infty$. In case of boundary control, one might also consider $\mathcal{D} = \partial\Omega$ or $\mathcal{D} = \partial\Omega \times (0, T)$. Further, let X and Y be Sobolev spaces of positive index, such that point evaluation almost everywhere is well defined.

State Constraints

Recall, that the solution $y_\varrho \in Y$ of (3.44)-(3.45), is given as the minimizer of the reduced cost functional (3.5)

$$\tilde{\mathcal{J}}(y_\varrho) = \frac{1}{2}\|y_\varrho - y_d\|_H^2 + \frac{\varrho}{2}\|By_\varrho\|_{A^{-1}}^2 = \frac{1}{2}\langle y_\varrho - y_d, y_\varrho - y_d\rangle_H + \frac{\varrho}{2}\langle A^{-1}By_\varrho, By_\varrho\rangle_H.$$

To impose constraints on the state, we consider

$$y_\varrho \in K_s := \big\{z \in Y : \, g_-(x) \le z(x) \le g_+(x), \text{ for a.a. } x \in \mathcal{D}\big\}, \tag{3.46}$$

where $g_\pm \in Y$ are given barrier functions, for which we assume $g_-(x) \le 0 \le g_+(x)$ for all $x \in \mathcal{D}$. Then, we want to find $y_\varrho \in K_s$ such that

$$\tilde{\mathcal{J}}(y_\varrho) \le \tilde{\mathcal{J}}(z) \quad \text{for all } z \in K_s. \tag{3.47}$$

We easily check, that $K_s \subset Y$ is closed and convex. With Theorem 2.7 and Remark 2.8, choosing $T := \varrho S + I : Y \to Y^*$ and $f = y_d \in H \subset Y^*$, we conclude that $y_\varrho \in K_s$ is characterized as the unique solution of the variational inequality

$$\varrho\langle Sy_\varrho, z - y_\varrho\rangle_H + \langle y_\varrho, z - y_\varrho\rangle_H \ge \langle y_d, z - y_\varrho\rangle_H \quad \text{for all } z \in K_s, \tag{3.48}$$

where $y_d \in H$ is given and $S := B^*A^{-1}B : Y \to Y^*$ is self-adjoint, bounded and elliptic, see Lemma 3.4.

Control constraints

In order to include constraints on the control $u_\varrho \in X^*$, we will minimize the reduced cost functional (3.8)

$$\begin{aligned}\hat{\mathcal{J}}(u_\varrho) &= \frac{1}{2}\|B^{-1}u_\varrho - y_d\|_H + \frac{\varrho}{2}\|u_\varrho\|^2_{A^{-1}}\\ &= \frac{1}{2}\langle B^{-1}u_\varrho - y_d, B^{-1}u_\varrho - y_d\rangle_H + \frac{\varrho}{2}\langle u_\varrho, A^{-1}u_\varrho\rangle_H\end{aligned}$$

over all functions

$$u_\varrho \in U_c := \big\{v \in X^* : \langle h_-, q\rangle_H \le \langle v, q\rangle_H \le \langle h_+, q\rangle_H, \text{ for all } q \in X,\ q \ge 0\big\}, \quad (3.49)$$

where $h_\pm \in H$ are given barrier functions and we assume that $h_-(x) \le 0 \le h_+(x)$ for a.a. $x \in \mathcal{D}$. As in the case of state constraints, we can argue that there exists a unique minimizer $u_\varrho \in U_c$ fulfilling

$$\hat{\mathcal{J}}(u_\varrho) \le \hat{\mathcal{J}}(v), \quad \text{for all } v \in U_c, \quad (3.50)$$

and that (3.50) is equivalent to find $u_\varrho \in U_c$ as unique solution of the variational inequality

$$\langle (B^*)^{-1}(B^{-1}u_\varrho - y_d), v - u_\varrho\rangle_H + \varrho\langle u_\varrho, A^{-1}(v - u_\varrho)\rangle_H \ge 0 \quad \text{for all } v \in U_c. \quad (3.51)$$

Recall, that $B : Y \to X^*$ is an isomorphism and that $By_\varrho = u_\varrho$, i.e., $u_\varrho = B^{-1}y_\varrho$. If we introduce

$$K_c := B^{-1}(U_c) = \Big\{z \in Y : \langle h_-, q\rangle_H \le \langle Bz, q\rangle_H \le \langle h_+, q\rangle_H, \text{ for all } q \in X,\ q \ge 0\Big\},$$

we can equivalently phrase the variational inequality (3.51) as: find $y_\varrho \in K_c$ such that

$$\langle y_\varrho - y_d, z - y_\varrho\rangle_H + \varrho\langle Sy_\varrho, z - y_\varrho\rangle_H \ge 0 \quad \text{for all } z = B^{-1}v \in K_c. \quad (3.52)$$

This is exactly (3.48), but with a different set of constraints.

So both, state and control constraints, admit the same structure. Namely, for a convex and closed subset $K \subset Y$, where $0 \in K$, and given $y_d \in H$, find $y_\varrho \in K$ such that

$$\varrho\langle Sy_\varrho, z - y_\varrho\rangle_H + \langle y_\varrho, z - y_\varrho\rangle_H \ge \langle y_d, z - y_\varrho\rangle_H, \quad \text{for all } z \in K. \quad (3.53)$$

In the following we will thus give abstract stability and regularization error estimates for this variational formulation. Related estimates for elliptic optimal control problems were studied in [50].

LEMMA 3.22. *Let $y_d \in H$ be given. Then the unique solution $y_\varrho \in K$ of* (3.53) *fulfills*

$$\|y_\varrho\|_H \leq \|y_d\|_H \quad \textit{and} \quad \sqrt{\varrho}\|y_\varrho\|_S \leq \|y_d\|_H. \tag{3.54}$$

Moreover, if $y_d \in K$, it holds

$$\|y_\varrho\|_S \leq \|y_d\|_S. \tag{3.55}$$

Proof. By assumption, we have that $z = 0 \in K$ is a valid test function in (3.53) and gives

$$\varrho\|y_\varrho\|_S^2 + \|y_\varrho\|_H^2 \leq \langle y_d, y_\varrho\rangle_H \leq \|y_d\|_H\|y_\varrho\|_H,$$

which gives (3.54). Further, if $y_d \in K$, choosing $z = y_d$ in (3.53) we can estimate

$$\begin{aligned}
\varrho\|y_\varrho\|_S^2 &= \varrho\langle Sy_\varrho, y_\varrho\rangle_H \\
&= -\varrho\langle Sy_\varrho, y_d - y_\varrho\rangle_H + \varrho\langle Sy_\varrho, y_d\rangle_H \\
&\leq -\langle y_d - y_\varrho, y_d - y_\varrho\rangle_H + \varrho\langle Sy_\varrho, y_d\rangle_H.
\end{aligned}$$

Now, reordering and using a Cauchy-Schwarz inequality gives

$$\varrho\|y_\varrho\|_S^2 \leq \varrho\|y_\varrho\|_S^2 + \|y_\varrho - y_d\|_H^2 \leq \varrho\|y_d\|_S\|y_\varrho\|_S,$$

from which we conclude (3.55). □

As in the case without constraints, see Lemma 3.6, we can prove the following regularization estimates.

LEMMA 3.23 ([50, cf Lemma 2.1]). *Let $y_d \in H$ be given. For the unique solution $y_\varrho \in K$ of* (3.53) *there holds*

$$\|y_\varrho - y_d\|_H \leq \|y_d\|_H. \tag{3.56}$$

Further, if $y_d \in K$, then

$$\|y_\varrho - y_d\|_H \leq \sqrt{\varrho}\|y_d\|_S \quad \textit{and} \quad \|y_\varrho - y_d\|_S \leq \|y_d\|_S. \tag{3.57}$$

If in addition $Sy_d \in H$ it holds

$$\|y_\varrho - y_d\|_H \leq \varrho\|Sy_d\|_H \quad \textit{and} \quad \|y_\varrho - y_d\|_S \leq \sqrt{\varrho}\|Sy_d\|_H. \tag{3.58}$$

Proof. From (3.53) we get that

$$\varrho\langle Sy_\varrho, z - y_\varrho\rangle_H \geq \langle y_d - y_\varrho, z - y_\varrho\rangle_H$$

for all $v \in K$. In particular, choosing $z = 0$ this gives

$$\varrho\langle Sy_\varrho, y_\varrho\rangle_H \leq \langle y_d - y_\varrho, y_\varrho\rangle_H = -\|y_d - y_\varrho\|_H^2 + \langle y_d - y_\varrho, y_d\rangle_H.$$

Now, reordering and applying a Cauchy–Schwarz inequality gives

$$\varrho\|y_\varrho\|_S^2 + \|y_d - y_\varrho\|_H^2 \leq \|y_d - y_\varrho\|_H \|y_d\|_H,$$

from which (3.56) follows. If $y_d \in K$, we can choose $z = y_d \in K$ in (3.53) to immediately obtain

$$\|y_d - y_\varrho\|_H^2 \leq \varrho\langle Sy_\varrho, y_d - y_\varrho\rangle_H = \varrho\langle S(y_\varrho - y_d), y_d - y_\varrho\rangle_H + \varrho\langle Sy_d, y_d - y_\varrho\rangle_H. \quad (3.59)$$

Again, reordering and using a Cauchy-Schwarz inequality gives

$$\|y_d - y_\varrho\|_H^2 + \varrho\|y_d - y_\varrho\|_S^2 \leq \varrho\|y_d\|_S \|y_d - y_\varrho\|_S,$$

from which (3.57) follows. If in addition $Sy_d \in H$ we can estimate (3.59) differently to obtain

$$\|y_d - y_\varrho\|_H^2 + \varrho\|y_d - y_\varrho\|_S^2 \leq \varrho\|Sy_d\|_H \|y_d - y_\varrho\|_H,$$

which gives (3.58). □

3.2.1 Complementarity conditions

In order to derive meaningful complementarity conditions, we will make use of the following regularity result, which is a direct consequence of Theorem 2.9.

LEMMA 3.24. *Let $y_d \in H$. Then the unique solution $y_\varrho \in K$ of* (3.53) *fulfills*

$$\varrho Sy_\varrho + y_\varrho \in H \quad \text{and} \quad \|\varrho Sy_\varrho + y_\varrho\|_H \leq C < \infty,$$

where $C = C(K, y_d)$. In particular, $Sy_\varrho \in H$.

State constraints

Recall, that the set of constraints in this case is given as

$$K = K_s = \big\{z \in Y : g_-(x) \leq z(x) \leq g_+(x), \text{ for a.a. } x \in \mathcal{D}\big\},$$

where $g_\pm \in Y$, $g_-(x) \leq 0 \leq g_+(x)$ for a.a. $x \in \mathcal{D}$ and we now additionally assume that $Sg_\pm \in H$. This assumption is natural, as $Sy_\varrho \in H$ by Lemma 3.24. We introduce the auxiliary variable $\lambda := \varrho Sy_\varrho + y_\varrho - y_d \in Y^*$. By Lemma 3.24 and (3.48) we have that $\lambda \in H = L^2(\mathcal{D})$ and fulfills

$$\langle \lambda, z - y_\varrho\rangle_H \geq 0, \quad \text{for all } z \in K_s. \quad (3.60)$$

In order to derive complementarity conditions, we further introduce

$$\mathcal{D}_{s,\pm} := \{x \in \mathcal{D} :\ y_\varrho(x) = g_\pm(x)\} \tag{3.61}$$

and first consider $w_\pm \in Y$ satisfying

$$0 \le w_-(x) \le y_\varrho(x) - g_-(x) \quad \text{and} \quad 0 \le w_+(x) \le g_+(x) - y_\varrho(x) \quad \text{for a.a. } x \in \mathcal{D}.$$

Then $w_\pm(x) = 0$ for all $x \in \mathcal{D}_{s,\pm}$ and testing (3.60) with $z = y_\varrho + w_+ \in K_s$ gives

$$0 \le \langle \lambda, w_+ \rangle_H = \langle \lambda, w_+ \rangle_{L^2(\mathcal{D}\setminus\mathcal{D}_{s,+})},$$

while testing with $z = y_\varrho - w_- \in K_s$ gives

$$0 \ge \langle \lambda, w_- \rangle_H = \langle \lambda, w_- \rangle_{L^2(\mathcal{D}\setminus\mathcal{D}_-)}.$$

From this we conclude the complementarity conditions

$$\begin{aligned} \lambda &= 0, & g_- < y_\varrho &< g_+, && \text{on } \mathcal{D}\setminus\mathcal{D}_{s,\pm}, \\ \lambda &\ge 0, & y_\varrho &= g_-, && \text{on } \mathcal{D}_{s,-}, \\ \lambda &\le 0, & y_\varrho &= g_+, && \text{on } \mathcal{D}_{s,+}, \end{aligned} \tag{3.62}$$

which enable us to give an additional regularization error estimate.

LEMMA 3.25. *For the unique solution $y_\varrho \in K_s$ of* (3.48) *there holds*

$$\varrho\|Sy_\varrho\|_H^2 = \|y_d - y_\varrho\|^2_{L^2(\mathcal{D}\setminus\mathcal{D}_{s,\pm})} + \varrho^2\|Sg_-\|^2_{L^2(\mathcal{D}_{s,-})} + \varrho^2\|Sg_+\|^2_{L^2(\mathcal{D}_{s,+})}.$$

In particular, if $y_d \in K_s$ such that $Sy_d \in H$, we have that

$$\|Sy_\varrho\|_H \le \|Sy_d\|_H + \|Sg_+\|_H + \|Sg_-\|_H.$$

Proof. From the complementarity conditions (3.62) we get that $\lambda = \varrho S y_\varrho + y_\varrho - y_d = 0$ on $\mathcal{D}\setminus\mathcal{D}_{s,\pm}$. Subsequently, we have

$$\varrho S y_\varrho = y_d - y_\varrho \quad \text{on } \mathcal{D}\setminus\mathcal{D}_{s,\pm}$$

and therefore

$$\varrho\|Sy_\varrho\|_{L^2(\mathcal{D}\setminus\mathcal{D}_{s,\pm})} = \|y_d - y_\varrho\|_{L^2(\mathcal{D}\setminus\mathcal{D}_{s,\pm})}.$$

Further, we have that

$$Sy_\varrho(x) = Sg_\pm(x) \quad \text{for a.a. } x \in \mathcal{D}_{s,\pm}.$$

Thus,

$$
\begin{aligned}
\varrho^2\|Sy_\varrho\|_H^2 &= \varrho^2\|Sy_\varrho\|^2_{L^2(\mathcal{D}\setminus\mathcal{D}_{s,\pm})} + \varrho^2\|Sy_\varrho\|^2_{L^2(\mathcal{D}_{s,-})} + \varrho^2\|Sy_\varrho\|^2_{L^2(\mathcal{D}_{s,+})} \\
&= \|y_d - y_\varrho\|^2_{L^2(\mathcal{D}\setminus\mathcal{D}_{s,\pm})} + \varrho^2\|Sg_-\|^2_{L^2(\mathcal{D}_{s,-})} + \varrho^2\|Sg_+\|^2_{L^2(\mathcal{D}_{s,+})}.
\end{aligned}
$$

If $y_d \in K_s$ such that $Sy_d \in H$, we can use Lemma 3.23 (3.58) to obtain

$$
\begin{aligned}
\varrho\|Sy_\varrho\|_H &\le \big(\|y_d - y_\varrho\|_H^2 + \varrho^2\|Sg_-\|_H^2 + \varrho^2\|Sg_+\|_H^2\big)^{1/2} \\
&\le \big(\varrho^2\|Sy_d\|_H^2 + \varrho^2\|Sg_-\|_H^2 + \varrho^2\|Sg_+\|_H^2\big)^{1/2} \\
&\le \varrho\big(\|Sy_d\|_H + \|Sg_+\|_H + \|Sg_-\|_H\big),
\end{aligned}
$$

which concludes the proof. □

Control constraints

In this case, the set of constraints is given as

$$
K = K_c = \Big\{z \in Y : \langle h_-, q\rangle_H \le \langle Bz, q\rangle_H \le \langle h_+, q\rangle_H, \text{ for all } q \in X,\ q \ge 0\Big\},
$$

where $h_\pm \in H$ and $h_-(x) \le 0 \le h_+(x)$ for almost all (a.a.) $x \in \mathcal{D}$. To derive complementarity conditions we introduce the auxiliary variable $w_\lambda \in X$ fulfilling

$$
B^* w_\lambda = \lambda = \varrho S y_\varrho + y_\varrho - y_d \in X^*.
$$

By the regularity of $\lambda \in H$, see Lemma 3.24, we have that $B^* w_\lambda \in H$ and by (3.52), we get

$$
0 \le \langle \lambda, z - y_\varrho\rangle_H = \langle B^* w_\lambda, z - y_\varrho\rangle_H = \langle w_\lambda, B(z - y_\varrho)\rangle_H, \quad \text{for all } z \in K_c. \tag{3.63}
$$

In order to proceed our analysis, we make the following assumption.

Assumption 3.26. *If $Sy_\varrho = B^*A^{-1}By_\varrho \in H$, then $u_\varrho = By_\varrho \in H$.*

This guarantess that point evaluation of u_ϱ is almost everywhere (a.e.) well-defined and we can introduce

$$
\mathcal{D}_{c,\pm} := \big\{x \in \mathcal{D} :\ u_\varrho(x) = h_\pm(x)\big\}.
$$

Further, let $z_\pm \in Y$ be the unique solutions of

$$
Bz_+ = By_\varrho + \psi_+ \text{ in } X^* \quad \text{and} \quad Bz_- = By_\varrho - \psi_- \text{ in } X^*,
$$

where $\psi_\pm \in H$ fulfilling $\psi_\pm \geq 0$ and

$$\langle \psi_-, q\rangle_H \leq \langle By_\varrho - h_-, q\rangle_H \quad \text{and} \quad \langle \psi_+, q\rangle_H \leq \langle h_+ - By_\varrho, q\rangle_H,$$

for all $q \in X$ such that $q \geq 0$. This guarantees, that $z_\pm \in K_c$. Choosing $q \in \mathcal{C}_0^\infty(\mathcal{D}_{c,\pm})$, we see that $\psi_\pm(x) = 0$ for a.a. $x \in \mathcal{D}_{c,\pm}$. Testing (3.63) with z_+ now gives

$$0 \leq \langle w_\lambda, \psi_+\rangle_H = \langle w_\lambda, \psi_+\rangle_{L^2(\mathcal{D}\setminus\mathcal{D}_{c,+})},$$

whereas testing with z_- gives

$$0 \geq \langle w_\lambda, \psi_-\rangle_H = \langle w_\lambda, \psi_-\rangle_{L^2(\mathcal{D}\setminus\mathcal{D}_{c,-})}.$$

Altogether, we thus conclude the complementarity conditions

$$\begin{aligned} &w_\lambda = 0, \quad h_- < By_\varrho = u_\varrho < h_+, \quad \text{on } \mathcal{D}\setminus\mathcal{D}_{c,\pm},\\ &w_\lambda \geq 0, \qquad\quad By_\varrho = u_\varrho = h_-, \quad \text{on } \mathcal{D}_{c,-},\\ &w_\lambda \leq 0, \qquad\quad By_\varrho = u_\varrho = h_+, \quad \text{on } \mathcal{D}_{c,+}. \end{aligned} \tag{3.64}$$

As in the case of state constraints we want to give additional regularization error estimates. Therefore, we need to make the following assumption.

ASSUMPTION 3.27. *There exists $c_c > 0$ for $h_\pm \in H$ such that*

$$\|B^*A^{-1}h_\pm\|_H \leq c_c\|h_\pm\|_H.$$

REMARK 3.28.

- *In order to introduce the sets $\mathcal{D}_{c,\pm}$, we need to have pointwise a.e. evaluation of u_ϱ. This is guaranteed by Assumption 3.26. Note, that $u_\varrho \in H$ is always guaranteed if $X^* = H$ and $A = I : H \to H$ or if $X = Y$ and $A = B$ (as then $A = B = S$). Otherwise, this needs to be checked for the specific application.*
- *Assumption 3.27 is fulfilled with $c_c = 1$, whenever $X = Y$ and $A = B = S$.*

LEMMA 3.29. *Let the Assumptions 3.26 and 3.27 hold. Then for the unique solution $y_\varrho \in K_c$ of (3.52) there holds*

$$\varrho\|Sy_\varrho\|_H^2 = \|y_d - y_\varrho\|^2_{L^2(\mathcal{D}\setminus\mathcal{D}_{c,\pm})} + c_c^2\varrho^2\big[\|h_-\|_H^2 + \|h_+\|_H^2\big].$$

In particular, if $y_d \in K_c$ such that $Sy_d \in H$, we have that

$$\|Sy_\varrho\|_H \leq c\big(\|Sy_d\|_H + \|h_+\|_H + \|h_-\|_H\big).$$

Proof. From the complementarity conditions (3.64) we get that $w_\lambda(x) = 0$ for a.a. $x \in \mathcal{D} \setminus \mathcal{D}_{c,\pm}$ and conclude,

$$0 = B^* w_\lambda = \lambda = \varrho S y_\varrho + y_\varrho - y_d \quad \text{on } \mathcal{D} \setminus \mathcal{D}_{c,\pm}.$$

Thus

$$\varrho S y_\varrho = y_d - y_\varrho \quad \text{on } \mathcal{D} \setminus \mathcal{D}_{c,\pm}.$$

Subsequently,

$$\varrho^2 \|S y_\varrho\|_H^2 = \|y_d - y_\varrho\|^2_{L^2(\mathcal{D}\setminus\mathcal{D}_{c,\pm})} + \varrho^2 \|S y_\varrho\|^2_{L^2(\mathcal{D}_{c,-})} + \varrho^2 \|S y_\varrho\|^2_{L^2(\mathcal{D}_{c,+})}.$$

Now, using the complementarity conditions (3.64) and Assumption 3.27, we can bound

$$\begin{aligned} \|S y_\varrho\|_{L^2(\mathcal{D}_{c,\pm})} &= \|B^* A^{-1} B y_\varrho\|_{L^2(\mathcal{D}_{c,\pm})} = \|B^* A^{-1} h_\pm\|_{L^2(\mathcal{D}_{c,\pm})} \\ &\le \|B^* A^{-1} h_\pm\|_H \le c_c \|h_\pm\|_H. \end{aligned}$$

The second estimate can now be shown following the lines of the proof of Lemma 3.25 using the regularization estimate (3.58). □

3.2.2 Discretization

We saw that both, state and control constraints, lead to the same variational formulation (3.53), but with different sets of constraints $K \in \{K_s, K_c\}$. Thus, we only need to analyze the discrete variational formulation to find the solution $y_{\varrho h} \in K_h$ of

$$\varrho \langle S y_{\varrho h}, z_h - y_{\varrho h}\rangle_H + \langle y_{\varrho h}, z_h - y_{\varrho h}\rangle_H \ge \langle y_d, z_h - y_{\varrho h}\rangle_H \quad \text{for all } z_h \in K_h, \tag{3.65}$$

where $X_h \subset X$ is a finite dimensional subspace and $K_h \subset X_h$ is a convex and closed set. By Theorem 2.7 we know, that the discrete variational formulation admits a unique solution. Further, using Theorem 2.16, we can show the following result.

Lemma 3.30. *Let $y_\varrho \in K$ and $y_{\varrho h} \in K_h$ be the unique solutions of* (3.53) *and* (3.65) *respectively. Then*

$$\begin{aligned} &\varrho \|y_\varrho - y_{\varrho h}\|_S^2 + \|y_\varrho - y_{\varrho h}\|_H^2 \\ &\quad \le c\Big(\inf_{z_h \in K_h} \left[\varrho \|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2\right] \\ &\qquad + \|y_d - y_\varrho\|_H^2 + \varrho^2 \|S y_\varrho\|_H^2 \Big). \end{aligned}$$

Proof. Let $T := \varrho S + I : Y \to Y^*$, which induces the norm $\|z\|_T^2 = \varrho\|z\|_S^2 + \|z\|_H^2$. By Lemma 3.24 we have that $Ty_\varrho \in H$ and from Theorem 2.16 we get for all $z_h \in K_h$ and all $z \in K$ that

$$\begin{aligned}
&\varrho\|y_\varrho - y_{\varrho h}\|_S^2 + \|y_\varrho - y_{\varrho h}\|_H^2 \\
&\qquad \le \varrho\|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2 \\
&\qquad\qquad +2\|y_d - y_\varrho - \varrho S y_\varrho\|_H \big[\|y_\varrho - z_h\|_H + \|y_{\varrho h} - z\|_H\big].
\end{aligned}$$

Now, choosing $z = y_\varrho \in K$ and using a triangle inequality and the estimate $ab \le 2a^2 + \frac{b^2}{8}$ we obtain

$$\begin{aligned}
&\varrho\|y_\varrho - y_{\varrho h}\|_S^2 + \|y_\varrho - y_{\varrho h}\|_H^2 \\
&\quad \le \varrho\|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2 + 2\|y_d - y_\varrho\|_H \big[\|y_\varrho - z_h\|_H + \|y_{\varrho h} - y_\varrho\|_H\big] \\
&\qquad +2\varrho\|Sy_\varrho\|_H \big[\|y_\varrho - z_h\|_H + \|y_{\varrho h} - y_\varrho\|_H\big] \\
&\quad \le \varrho\|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2 + 4\|y_d - y_\varrho\|_H^2 + \frac{1}{4}\|y_\varrho - z_h\|_H^2 + \frac{1}{4}\|y_{\varrho h} - y_\varrho\|_H^2 \\
&\qquad +4\varrho^2\|Sy_\varrho\|_H^2 + \frac{1}{4}\|y_\varrho - z_h\|_H^2 + \frac{1}{4}\|y_{\varrho h} - y_\varrho\|_H^2 \\
&\quad = \varrho\|y_\varrho - z_h\|_S^2 + \frac{3}{2}\|y_\varrho - z_h\|_H^2 + 4\|y_d - y_\varrho\|_H^2 + 4\varrho^2\|Sy_\varrho\|_H^2 \\
&\qquad +\frac{1}{2}\|y_{\varrho h} - y_\varrho\|_H^2 \\
&\quad \le \frac{3}{2}\big(\varrho\|y_\varrho - z_h\|_S^2 + \|y_\varrho - z_h\|_H^2\big) + 4\|y_d - y_\varrho\|_H^2 + 4\varrho^2\|Sy_\varrho\|_H^2 \\
&\qquad +\frac{1}{2}\big(\varrho\|y_\varrho - y_{\varrho h}\|_S^2 + \|y_\varrho - y_{\varrho h}\|_H^2\big).
\end{aligned}$$

Subtracting $\frac{1}{2}\big(\varrho\|y_\varrho - y_{\varrho h}\|_S^2 + \|y_\varrho - y_{\varrho h}\|_H^2\big)$ from both sides, multiplying by $\frac{1}{2}$ and taking the infimum over all $z_h \in K_h$ concludes the proof. □

Now we are in the position to state the main theorem of this section, which is an analogon to the quasi-best approximation result in the unconstrained case in Theorem 3.11.

THEOREM 3.31. *Let $y_{\varrho h} \in K_h$ denote the unique solution of (3.65). If $y_d \in H$ and $0 \in K_h$ then*

$$\|y_d - y_{\varrho h}\|_H \le \|y_d\|_H. \tag{3.66}$$

In the case of state contraints, if in addition $y_d \in K = K_s$ such that $Sy_d \in H$ then there holds

$$\begin{aligned}
\|y_d - y_{\varrho h}\|_H \;\le\; & c\Big(\varrho\|Sy_d\|_H + \varrho\|Sg_\pm\|_H \\
& + \inf_{z_h \in K_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big)
\end{aligned} \tag{3.67}$$

and

$$\sqrt{\varrho}\|y_d - y_{\varrho h}\|_S \;\leq\; c\Big(\varrho\|Sy_d\|_H + \varrho\|Sg_\pm\|_H + \inf_{z_h\in K_h}\big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big). \tag{3.68}$$

In the case of control constraints, if the Assumptions 3.26 and 3.27 hold true and $y_d \in K = K_c$ *such that* $Sy_d \in H$ *we have*

$$\|y_d - y_{\varrho h}\|_H \;\leq\; c\Big(\varrho\|Sy_d\|_H + \varrho\|h_\pm\|_H + \inf_{z_h\in K_h}\big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big) \tag{3.69}$$

and

$$\sqrt{\varrho}\|y_d - y_{\varrho h}\|_S \;\leq\; c\Big(\varrho\|Sy_d\|_H + \varrho\|h_\pm\|_H + \inf_{z_h\in K_h}\big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big). \tag{3.70}$$

Proof. By assumption $z_h = 0 \in K_h$ is a valid test function for (3.65). Thus, the estimate (3.66) can be shown analogously to the continuous case in Lemma 3.23.

Using a triangle inequality, we can first estimate

$$\begin{aligned}
&\|y_d - y_{\varrho h}\|_H^2 + \varrho\|y_d - y_{\varrho h}\|_S^2 \\
&\quad\leq 2\|y_d - y_\varrho\|_H^2 + 2\varrho\|y_d - y_\varrho\|_S^2 && (3.71)\\
&\qquad +2\|y_\varrho - y_{\varrho h}\|_H^2 + 2\varrho\|y_\varrho - y_{\varrho h}\|_S^2. && (3.72)
\end{aligned}$$

With the regularization error estimates (3.58) of Lemma 3.23 we bound (3.71) by

$$\|y_d - y_\varrho\|_H^2 + \varrho\|y_d - y_\varrho\|_S^2 \leq \varrho^2\|Sy_d\|_H^2. \tag{3.73}$$

For (3.72) we use Lemma 3.30 to obtain the bound

$$\begin{aligned}
&\|y_\varrho - y_{\varrho h}\|_H^2 + \varrho\|y_\varrho - y_{\varrho h}\|_S^2 \\
&\quad\leq c\Big(\inf_{z_h\in K_h}\big[\|y_\varrho - z_h\|_H^2 + \varrho\|y_\varrho - z_h\|_S^2\big] && (3.74)\\
&\qquad +\|y_d - y_\varrho\|_H^2 + \varrho^2\|Sy_\varrho\|_H^2\Big). && (3.75)
\end{aligned}$$

Now, for (3.74) we add and subtract y_d, apply a triangle inequality and use (3.73) to get

$$\begin{aligned}
&\inf_{z_h\in K_h}\big[\|y_\varrho - z_h\|_H^2 + \varrho\|y_\varrho - z_h\|_S^2\big] \\
&\quad\leq 2\inf_{z_h\in K_h}\big[\|y_d - z_h\|_H^2 + \varrho\|y_d - z_h\|_S^2\big] + 2\|y_\varrho - y_d\|_H^2 + 2\varrho\|y_\varrho - y_d\|_S^2 \\
&\quad\leq 2\inf_{z_h\in K_h}\big[\|y_d - z_h\|_H^2 + \varrho\|y_d - z_h\|_S^2\big] + 2\varrho^2\|Sy_d\|_H^2.
\end{aligned}$$

As by (3.73) we have that $\|y_d - y_\varrho\|_H^2 \le \varrho^2 \|Sy_d\|_H^2$ it remains to bound $\varrho^2\|Sy_\varrho\|_H^2$ in (3.75). Therefore, we use Lemma 3.25 in the case of state constraints, i.e., $K = K_s$, to get

$$\varrho^2 \|Sy_\varrho\|_H^2 \le \varrho^2 \big(\|Sy_d\|_H + \|Sg_\pm\|_H\big)^2.$$

In the case of control constraints $K = K_c$ we use Lemma 3.29, to get

$$\varrho^2 \|Sy_\varrho\|_H^2 \le c\varrho^2 \big(\|Sy_d\|_H + \|h_\pm\|_H\big)^2.$$

This concludes the proof. $\square$

As in the unconstrained setting, we might not be able to realize the operator $S = B^* A^{-1} B : Y \to Y^*$ but only a computable counterpart $\tilde{S} : Y \to Y^*$, see Remark 3.13. Recall, the definition $\tilde{S}y := B^* p_{yh}$, where $p_{yh} \in X_h \subset X$ is the unique solution of

$$\langle Ap_{yh}, q_h \rangle_H = \langle By, q_h \rangle_H \quad \text{for all } q_h \in X_h. \tag{3.76}$$

We consider the perturbed variational formulation to find $\tilde{y}_{\varrho h} \in K_h$ such that

$$\varrho \langle \tilde{S}\tilde{y}_{\varrho h}, z_h - \tilde{y}_{\varrho h} \rangle_H + \langle \tilde{y}_{\varrho h}, z_h - \tilde{y}_{\varrho h} \rangle_H \ge \langle y_d, z_h - \tilde{y}_{\varrho h} \rangle_H \quad \text{for all } z_h \in K_h. \tag{3.77}$$

For each given $y_d \in H$ the perturbed variational inequality (3.77) admits a unique solution $\tilde{y}_{\varrho h} \in K_h \subset H$, due to the properties of $\tilde{S}$, see Lemma 3.14. Moreover, we can give the following error estimates, as analogon to the unconstrained case in Theorem 3.15.

THEOREM 3.32. *Let $y_d \in H$ and $0 \in K_h$. Then for the unique solution $\tilde{y}_{\varrho h} \in K_h$ of (3.77) there holds*

$$\|\tilde{y}_{\varrho h} - y_d\|_H \le \|y_d\|_H. \tag{3.78}$$

Moreover, let $y_d \in K$ such that $Sy_d \in H$ and let $p_{y_d} \in X$ be the unique solution of

$$\langle Ap_{y_d}, q \rangle_H = \langle By_d, q \rangle_H, \quad \textit{for all } q \in X.$$

and assume that for all $z_h \in K_h$ the inverse inequality

$$\|z_h\|_Y \le c_I h^{-1} \|z_h\|_H,$$

holds. Then, in the case of state constraints $K = K_s$, we get

$$\begin{aligned} \|\tilde{y}_{\varrho h} - y_d\|_H \le c\Big(&\big[h^{-1}\varrho^{3/2} + \varrho\big]\big(\|Sy_d\|_H + \|Sg_\pm\|_H\big) + h^{-1}\varrho \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X \\ &+ \big[h^{-1}\sqrt{\varrho} + 1\big] \inf_{z_h \in K_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big). \end{aligned} \tag{3.79}$$

Whereas, in the case of control constraints $K = K_c$, if the Assumptions 3.26 and 3.27 hold true, we get

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_d\|_H \leq c\Big(&\big[h^{-1}\varrho^{3/2} + \varrho\big]\big(\|Sy_d\|_H + \|h_\pm\|_H\big) + h^{-1}\varrho \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X \\ &+ \big[h^{-1}\sqrt{\varrho} + 1\big] \inf_{z_h \in K_h} \big[\varrho\|y_d - z_h\|_S^2 + \|y_d - z_h\|_H^2\big]^{1/2}\Big).\end{aligned} \tag{3.80}$$

Proof. Since $z_h = 0 \in K_h$ is a valid test function in (3.77), the estimate (3.78) follows the lines of the continuous setting of the proof of Lemma 3.23.

Let $y_{\varrho h} \in K_h$ denote the unique solution of the unperturbed variational inequality (3.65) and note that testing (3.65) with $\tilde{y}_{\varrho h} \in K_h$ gives

$$\langle \varrho S y_{\varrho h} + y_{\varrho h}, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H \leq \langle y_d, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H$$

whereas testing (3.77) with $y_{\varrho h} \in K_h$ gives

$$\langle \varrho \tilde{S} \tilde{y}_{\varrho h} + \tilde{y}_{\varrho h}, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H \leq \langle y_d, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H.$$

Now, since $\tilde{S} \geq 0$ is positive semi-definite, it holds

$$\begin{aligned}\|y_{\varrho h} - \tilde{y}_{\varrho h}\|_H^2 &\leq \|y_{\varrho h} - \tilde{y}_{\varrho h}\|_H^2 + \varrho\langle \tilde{S}(y_{\varrho h} - \tilde{y}_{\varrho h}), y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H \\ &= \langle y_{\varrho h} + \varrho\tilde{S} y_{\varrho h} - (\tilde{y}_{\varrho h} + \varrho \tilde{S}\tilde{y}_{\varrho h}), y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H \\ &= \varrho\langle (\tilde{S} - S) y_{\varrho h}, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H \\ &\quad + \langle y_{\varrho h} + \varrho S y_{\varrho h}, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H + \langle \tilde{y}_{\varrho h} + \varrho\tilde{S}\tilde{y}_{\varrho h}, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H \\ &\leq \varrho\langle (\tilde{S} - S) y_{\varrho h}, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H \\ &\quad + \langle y_d, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H + \langle y_d, \tilde{y}_{\varrho h} - y_{\varrho h}\rangle_H \\ &= \varrho\langle (\tilde{S} - S) y_{\varrho h}, y_{\varrho h} - \tilde{y}_{\varrho h}\rangle_H.\end{aligned}$$

This is exactly the estimate (3.33) as in the unconstrained case and we can proceed the estimate in the same way. We sketch the main steps. Using the assumed inverse inequality, we can now derive the bound

$$\|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H \leq c\varrho h^{-1}\|p_{y_{\varrho h}} - p_{y_{\varrho h} h}\|_X,$$

where $p_{y_{\varrho h}} \in X$ is the unique solution of

$$\langle A p_{y_{\varrho h}}, q\rangle_H = \langle B y_{\varrho h}, q\rangle_H \quad \text{for all } q \in X,$$

and $p_{y_{\varrho h} h} \in Y_h$ solves (3.76) for $y = y_{\varrho h}$. Let $p_{y_d h} \in Y_h$ denote the unique solution of (3.76) for $y = y_d$. Then, using a triangle inequality, we can estimate

$$\|p_{y_{\varrho h}} - p_{y_{\varrho h} h}\|_X \leq \|p_{y_{\varrho h}} - p_{y_d}\|_X + \|p_{y_d} - p_{y_d h}\|_X + \|p_{y_d h} - p_{y_{\varrho h} h}\|_X.$$

We can estimate the terms as in the proof of Theorem 3.15 by

$$\|p_{y_{\varrho h}} - p_{y_d}\|_X \leq c\|y_{\varrho h} - y_d\|_S \quad \text{and} \quad \|p_{y_d h} - p_{y_{\varrho h} h}\|_X \leq c\|y_{\varrho h} - y_d\|_S$$

and from the Galerkin orthogonality, we can bound

$$\|p_{y_d} - p_{y_d h}\|_X \leq \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X,$$

Using a triangle inequality we get

$$\begin{aligned} \|\tilde{y}_{\varrho h} - y_d\|_H &\leq \|\tilde{y}_{\varrho h} - y_{\varrho h}\|_H + \|y_{\varrho h} - y_d\|_H \\ &\leq c\varrho h^{-1}\big(\|y_{\varrho h} - y_d\|_S + \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_X\big) + \|y_{\varrho h} - y_d\|_H \end{aligned}$$

and with Theorem 3.31 (3.67)-(3.68) in the case of state constraints and (3.69)-(3.70) in the case of control constraints we conclude the proof. $\square$

4 Model problems and numerical illustration of the optimal control framework

In this chapter we discuss various examples, where the framework of Chapter 3 is applicable. Firstly, we consider a distributed optimal control problem subject to the Poisson equation, for which the analysis on the continuous level leads to well-known formulations and spaces. We discuss the energy regularization in the case where we measure the control in $H^{-1}(\Omega)$, which in some sense turns out to be the natural choice, and also in the case, where we want to measure the control in $L^2(\Omega)$, as considered, e.g., by Brenner in [17]. The regularization error estimates in both cases have already been derived in [95] and fit perfectly to the abstract framework. Moreover, we will discuss conforming discretizations and give a full stability and error analysis, out of which we deduce the optimal choice of the regularization parameter $\varrho > 0$ depending on the mesh size h and the regularity of the target. These dependencies are of special interest in the design of solvers with optimal complexity [75, 76, 80]. Furthermore, we describe an adaptive finite element method and analyze the choice of a mesh dependent regularization parameter, similar to [74], which is beneficial in terms of stability, but still leads to quasi best approximation error estimates. The discussion of elliptic optimal control problems will be concluded by taking state and/or control constraints into account. Numerical examples will support our theory.

While space time formulations for parabolic optimal control problems are well-studied [10, 47, 77, 79] and the energy regularization fits into the abstract framework, see [81], much less is known in the case of hyperbolic optimal control problems. To show the full capacity of the abstract theory developed, in the last section we will consider a distributed optimal control problem subject to the wave equation with energy regularization. Using a space time setting developed in [111, 116], we redo the analysis as in the elliptic case, including regularization error estimates, the optimal choice of $\varrho > 0$ and adaptive schemes, as well as state and control constraints.

4.1 An elliptic model problem

Let $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ be a bounded Lipschitz domain. Then we consider the optimal control problem to reach a given desired state $y_d \in L^2(\Omega)$ as good as possible under bearable costs, by functions that fulfill the operator equation

$$-\Delta y = u \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega. \tag{4.1}$$

This is modeled, by finding the minimizer $(y_\varrho, u_\varrho) \in Y \times U$ of

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_U \tag{4.2}$$

for some cost/regularization parameter $\varrho > 0$. In the following, different choices for the spaces Y and U are discussed. In Section 4.1.4 we will consider the variational formulation for the operator equation (4.1), leading to the state being in $Y = H^1_0(\Omega)$, which seems to be the natural setting. Despite of the quiet common approach to measure the control in $L^2(\Omega)$, to fit the abstract framework, we will see that we rather need to consider $U = H^{-1}(\Omega)$. This framework will turn out to be especially useful when considering less regular targets $y_d \in H^s_0(\Omega)$ for $s \in [0, 1)$. We will then derive a discretization and related error estimates, using the results from the abstract setting. With the a posteriori error estimates we can formulate an adaptive scheme and, observing that for $u_\varrho \in H^{-1}(\Omega)$ we are able to understand the regularization parameter as a diffusion coefficient in a reduced optimality system, it will be an easy task to consider $\varrho = \varrho(x)$ as a function. This will be used to derive an optimal choice of the parameter for adaptive schemes and give related error estimates. In Section 4.1.3 we discuss the incorporation of state and control constraints. Moreover, in Section 4.1.4 we will answer the question whether the energy regularization also covers controls in $U = L^2(\Omega)$. Indeed, it turns out that the setting is applicable also in this case, but leads to higher regularity assumptions on the state. Related discretization and error estimates follow out of the box, if we can guarantee conforming ansatz spaces.

4.1.1 The energy regularization in $H^{-1}(\Omega)$

Since the Laplace operator and its mapping properties are well-understood and more commonly known, this section aims to be a starting point in the justification of the abstract framework, derived in Chapter 3. We will redo some of the steps and derivations already given in the abstract framework, to make the reader familiar with the tools and approaches used. At some points we will also compare the concept of the energy regularization with the more common approach for the $L^2(\Omega)$ regularization, to point out the differences. To start with, let us consider the operator equation (4.1) in a variational sense, i.e., we want to find $y \in Y := H^1_0(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla z \, dx = \langle u, z\rangle_\Omega \quad \text{for all } z \in H^1_0(\Omega). \tag{4.3}$$

Note, that with this choice we already fix $Y = H^1_0(\Omega)$. The variational formulation (4.3) admits a unique solution for all $u \in H^{-1}(\Omega)$, as the next lemma shows.

Lemma 4.1. *The operator* $-\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)$ *is an isomorphism. More precisely, for each* $u \in H^{-1}(\Omega)$ *there exists exactly one* $y \in H_0^1(\Omega)$ *solving* (4.3) *and*

$$\|y\|_{H_0^1(\Omega)}^2 = \|\nabla y\|_{L^2(\Omega)}^2 = \langle -\Delta y, y\rangle_\Omega = \|u\|_{H^{-1}(\Omega)}^2. \tag{4.4}$$

Proof. The operator $B = -\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)$ defined as

$$\langle By, z\rangle_\Omega := \int_\Omega \nabla y \cdot \nabla z \, dx \quad \text{for all } y, z \in H_0^1(\Omega),$$

is self-adjoint by definition. Since $\|y\|_{H_0^1(\Omega)} = \|\nabla y\|_{L^2(\Omega)}$, we immediately get boundedness and $H_0^1(\Omega)$-ellipticity with $c_1^B = c_2^B = 1$, i.e., for all $y, z \in H_0^1(\Omega)$

$$\langle By, y\rangle_\Omega = \|y\|_{H_0^1(\Omega)}^2 \quad \text{and} \quad \langle By, z\rangle_\Omega \le \|y\|_{H_0^1(\Omega)} \|z\|_{H_0^1(\Omega)}.$$

By the Lemma of Lax–Milgram (Theorem 2.3), for $u \in H^{-1}(\Omega)$ there exists a unique solution $y \in H_0^1(\Omega)$ of (4.3). This shows the property of being an isomorphism. Further, we compute that

$$\begin{aligned} \|u\|_{H^{-1}(\Omega)} &= \sup_{0 \ne z \in H_0^1(\Omega)} \frac{\langle u, z\rangle_\Omega}{\|z\|_{H_0^1(\Omega)}} = \sup_{0 \ne z \in H_0^1(\Omega)} \frac{\langle By, z\rangle_\Omega}{\|z\|_{H_0^1(\Omega)}} \\ &= \sup_{0 \ne z \in H_0^1(\Omega)} \frac{\int_\Omega \nabla y \cdot \nabla z \, dx}{\|\nabla z\|_{L^2(\Omega)}} \le \|y\|_{H_0^1(\Omega)} \end{aligned}$$

and vice versa

$$\|y\|_{H_0^1(\Omega)}^2 = \langle By, y\rangle_\Omega = \langle u, y\rangle_\Omega \le \|u\|_{H^{-1}(\Omega)} \|y\|_{H_0^1(\Omega)}$$

showing $\|y\|_{H_0^1(\Omega)} = \|u\|_{H^{-1}\Omega}$, which gives the desired equality of norms. □

With $\|u_\varrho\|_{H^{-1}(\Omega)} = \|y_\varrho\|_{H_0^1(\Omega)} = \|\nabla y_\varrho\|_{L^2(\Omega)}$ the reduced cost functional now becomes

$$\tilde{\mathcal{J}}(y_\varrho) = \frac{1}{2}\|y_\varrho - y_d\|_{L^2(\Omega)}^2 + \frac{\varrho}{2}\|\nabla y_\varrho\|_{L^2(\Omega)}^2, \tag{4.5}$$

and the mimizer $y_\varrho \in H_0^1(\Omega)$ has to fulfill the gradient equation

$$\varrho\langle \nabla y_\varrho, \nabla z\rangle_{L^2(\Omega)} + \langle y_\varrho - y_d, z\rangle_{L^2(\Omega)} = 0 \quad \text{for all } z \in H_0^1(\Omega), \tag{4.6}$$

for given $y_d \in L^2(\Omega)$, which is the weak formulation of the Dirichlet boundary value problem

$$-\varrho\Delta y_\varrho + y_\varrho = y_d \text{ in } \Omega \quad \text{and} \quad y_\varrho = 0 \text{ on } \partial\Omega. \tag{4.7}$$

Before we proceed analyzing (4.6), we want to give a different derivation, which involves the dual problem and will be useful to compare the approach to the common L^2-regularization. Instead of the reduced cost functional (4.5), we can use that $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism and substitute $y_\varrho = B^{-1}u_\varrho$. Together with the norm representation (4.4) this gives

$$\begin{aligned} \hat{\mathcal{J}}(u_\varrho) &= \frac{1}{2}\|B^{-1}u_\varrho - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_{H^{-1}(\Omega)} && (4.8)\\ &= \frac{1}{2}\langle B^{-1}u_\varrho - y_d, B^{-1}u_\varrho - y_d\rangle_{L^2(\Omega)} + \frac{\varrho}{2}\langle u_\varrho, B^{-1}u_\varrho\rangle_\Omega, && (4.9)\end{aligned}$$

for which the minimizer $u_\varrho \in H^{-1}(\Omega)$ fulfills the gradient equation

$$(B^*)^{-1}(B^{-1}u_\varrho - y_d) + \varrho B^{-1}u_\varrho = 0 \quad \text{in } H_0^1(\Omega).$$

Introducing $p_\varrho \in H_0^1(\Omega)$ as solution of the dual problem

$$B^* p_\varrho = B^{-1}u_\varrho - y_d \quad \text{in } H^{-1}(\Omega),$$

which is by the self-adjointness $B = B^*$ nothing but

$$-\Delta p_\varrho = y_\varrho - y_d \text{ in } \Omega \quad \text{and} \quad p_\varrho = 0 \text{ on } \partial\Omega,$$

we get the gradient equation

$$p_\varrho + \varrho B^{-1}u_\varrho = 0 \quad \text{in } H_0^1(\Omega),$$

and we end up with the optimality system

$$\begin{array}{llll} -\Delta y_\varrho = u_\varrho, & -\Delta p_\varrho = y_\varrho - y_d, & p_\varrho + \varrho y_\varrho = 0 & \text{in } \Omega,\\ y_\varrho = 0, & p_\varrho = 0, & & \text{on } \partial\Omega, \end{array} \tag{4.10}$$

consisting of the forward equation, the adjoint/backward equation and the gradient equation. Using the forward equation we can eliminate the control $u_\varrho = -\Delta y_\varrho$ and derive the variational formulation, testing the backward and the gradient equation with $q, z \in H_0^1(\Omega)$, respectively, and applying integration by parts, which then reads: find $(y_\varrho, p_\varrho) \in H_0^1(\Omega) \times H_0^1(\Omega)$ such that

$$\begin{array}{llll} \varrho^{-1}\displaystyle\int_\Omega \nabla p_\varrho \cdot \nabla q\, dx & + \displaystyle\int_\Omega \nabla y_\varrho \cdot \nabla q\, dx & = 0, & \text{for all } q \in H_0^1(\Omega),\\ -\displaystyle\int_\Omega \nabla p_\varrho \cdot \nabla z\, dx & + \displaystyle\int_\Omega y_\varrho z\, dx & = \displaystyle\int_\Omega y_d z\, dx, & \text{for all } z \in H_0^1(\Omega). \end{array} \tag{4.11}$$

Moreover, we can eliminate the adjoint variable $p_\varrho = -\varrho y_\varrho$, to conclude

$$-\varrho\Delta y_\varrho = \Delta p_\varrho = y_d - y_\varrho,$$

which is (4.7) again.

REMARK 4.2 (L^2-regularization). *Instead of measuring the control in $H^{-1}(\Omega)$, it is common to consider the cost functional*

$$\mathcal{I}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_{L^2(\Omega)},$$

which implies that we assume $u_\varrho \in L^2(\Omega)$. Of course, the variational formulation (4.3) *admits a unique solution $y \in H^1_0(\Omega)$ also in this case, as $L^2(\Omega) \subset H^{-1}(\Omega)$. Thus, we can define the solution operator $\mathcal{S} : L^2(\Omega) \to H^1_0(\Omega)$ by $\mathcal{S}u = y \in H^1_0(\Omega)$ for $u \in L^2(\Omega)$ and, as before, we can consider the reduced cost functional*

$$\hat{\mathcal{I}}(u_\varrho) = \frac{1}{2}\|\mathcal{S}u_\varrho - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_{L^2(\Omega)},$$

for which the minimizer is characterized as solution of the gradient equation

$$\mathcal{S}^*(\mathcal{S}u_\varrho - y_d) + \varrho u_\varrho = 0 \quad \text{in } L^2(\Omega),$$

where $\mathcal{S}^*$ is the formally L^2-adjoint of $\mathcal{S}$, defined as

$$\langle \mathcal{S}^* y, u\rangle_{L^2(\Omega)} := \langle y, \mathcal{S}u\rangle_{L^2(\Omega)} \quad \text{for all } y \in H^1_0(\Omega),\ u \in L^2(\Omega).$$

Let us introduce $p_\varrho = \mathcal{S}^*(\mathcal{S}u_\varrho - y_d)$ and assume that $p_\varrho \in H^1_0(\Omega)$, then it is the solution of the adjoint problem

$$-\Delta p_\varrho = y_\varrho - y_d \text{ in } \Omega \quad p_\varrho = 0 \text{ on } \partial\Omega,$$

and we end up with the optimality system

$$\begin{aligned} -\Delta y_\varrho &= u_\varrho, & -\Delta p_\varrho &= y_\varrho - y_d, & p_\varrho + \varrho u_\varrho &= 0 & &\text{in } \Omega, \\ y_\varrho &= 0, & p_\varrho &= 0 & & & &\text{on } \partial\Omega. \end{aligned} \tag{4.12}$$

Eliminiating the control $u_\varrho = -\Delta y_\varrho$, we now get that $p_\varrho - \varrho\Delta y_\varrho = 0$ and we can phrase the variational formulation: find $(y_\varrho, p_\varrho) \in H^1_0(\Omega) \times H^1_0(\Omega)$ such that

$$\begin{aligned} \varrho^{-1}\int_\Omega p_\varrho q\,dx \quad &+ \int_\Omega \nabla y_\varrho \cdot \nabla q\,dx &&= 0 && \text{for all } q \in H^1_0(\Omega), \\ -\int_\Omega \nabla p_\varrho \cdot \nabla z\,dx \quad &+ \int_\Omega y_\varrho z\,dx &&= \int_\Omega y_d z\,dx, && \text{for all } z \in H^1_0(\Omega). \end{aligned} \tag{4.13}$$

Moreover, we can eliminate the adjoint variable $p_\varrho = -\varrho u_\varrho = \varrho\Delta y_\varrho$ to conclude

$$\varrho\Delta^2 y_\varrho = \Delta p_\varrho = y_d - y_\varrho$$

and therefore

$$\varrho\Delta^2 y_\varrho + y_\varrho = y_d \text{ in } \Omega, \qquad y_\varrho = \Delta y_\varrho = 0 \text{ on } \partial\Omega. \tag{4.14}$$

Compared to (4.7) *using the common L^2-regularization leads to a fourth order PDE, as was already considered in [95]. Moreover, considering the optimality system* (4.12)*, note that*

$$u_\varrho = -\varrho^{-1} p_\varrho \in H_0^1(\Omega),$$

imposing probably unpleasant, additional boundary conditions on the control, which might even affect the numerical solution of (4.13)*. In Section 4.1.4 we will show that using the abstract framework, we can give a formulation that guarantees $u_\varrho \in L^2(\Omega)$, without adding additional regularity or boundary conditions to the control.*

So far, we did the derivation of the optimality system and their variational formulation step by step to show the techniques and tools needed. Now, we cast (4.1)-(4.2) into the framework of Chapter 3, to show the applicability of the abstract framework and, subsequently, use the theoretical results concerning the regularization error estimates, the discretization and the reconstruction of the control. Since the operator constraint (4.1) is given by the Poisson equation and we have by Lemma 4.1 that $B := -\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism, it fulfills Assumption (B1)-(B3). Moreover, B is self-adjoint, bounded and $H_0^1(\Omega)$-elliptic and we can choose $A = B = -\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)$, which then satisfies the Assumptions (A1)-(A3). Thus, with the choice

$$X = Y = H_0^1(\Omega) \subset H = L^2(\Omega) \subset H^{-1}(\Omega) = X^* = Y^* \quad \text{and} \quad A = B : Y \to Y^*,$$

we can apply the theory from Chapter 3. We start, noting that $S = B^* A^{-1} B = A = -\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)$, by (3.11) the minimizer $y_\varrho \in H_0^1(\Omega)$ of (4.5) is uniquely determined as the solution of

$$\varrho \langle -\Delta y_\varrho, z\rangle_\Omega + \langle y_\varrho, z\rangle_{L^2(\Omega)} = \langle y_d, z\rangle \quad \text{for all } v \in H_0^1(\Omega),$$

which is exactly (4.6). Using that $\|v\|_S = \|\nabla v\|_{L^2(\Omega)} = \|v\|_{H_0^1(\Omega)}$, by Lemma 3.5, Lemma 3.6 and Coroallary 3.7 we get the stability and regularization error estimates.

LEMMA 4.3. *Let $y_d \in L^2(\Omega)$ be given. For the unique solution $y_\varrho \in H_0^1(\Omega)$ of* (4.6) *there holds*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}, \tag{4.15}$$

as well as

$$\|y_\varrho\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}, \quad \|y_\varrho\|_{H_0^1(\Omega)} \le \frac{1}{\sqrt{\varrho}} \|y_d\|_{L^2(\Omega)}, \quad \|\Delta y_\varrho\|_{L^2(\Omega)} \le \frac{1}{\varrho} \|y_d\|_{L^2(\Omega)}. \tag{4.16}$$

Further, if $y_d \in H_0^1(\Omega)$*, then*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \leq \sqrt{\varrho}\|y_d\|_{H_0^1(\Omega)} \quad \text{and} \quad \|y_\varrho - y_d\|_{H_0^1(\Omega)} \leq \|y_d\|_{H_0^1(\Omega)}. \tag{4.17}$$

Moreover, it holds

$$\|y_\varrho\|_{H_0^1(\Omega)} \leq \|y_d\|_{H_0^1(\Omega)} \quad \text{and} \quad \|\Delta y_\varrho\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{\varrho}}\|y_d\|_{H_0^1(\Omega)}. \tag{4.18}$$

At last, if $y_d \in H_0^1(\Omega)$ *such that* $\Delta y_d \in L^2(\Omega)$ *it holds*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \leq \varrho\|\Delta y_d\|_{L^2(\Omega)} \quad \text{and} \quad \|y_\varrho - y_d\|_{H_0^1(\Omega)} \leq \sqrt{\varrho}\|\Delta y_d\|_{L^2(\Omega)}, \tag{4.19}$$

and, in this case we also have

$$\|\Delta y_\varrho\|_{L^2(\Omega)} \leq \|\Delta y_d\|_{L^2(\Omega)}. \tag{4.20}$$

Using a space interpolation argument, we can now state the main stability and regularization error estimates, which coincide with [95, Theorem 3.2].

THEOREM 4.4. *Let* $y_d \in H_0^s(\Omega)$ *for* $s \in [0,1]$ *or* $y_d \in H_0^1(\Omega) \cap H^s(\Omega)$ *for* $s \in (1,2]$*. Then,*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \leq c\varrho^{s/2}\|y_d\|_{H^s(\Omega)} \quad \text{for } s \in [0,2], \tag{4.21}$$

and

$$\|y_\varrho - y_d\|_{H_0^1(\Omega)} \leq c\varrho^{(s-1)/2}\|y_d\|_{H^s(\Omega)} \quad \text{for } s \in (1,2]. \tag{4.22}$$

Moreover, if Ω *is convex or the boundary of* $\partial\Omega$ *is* $\mathcal{C}^2$*, it holds that*

$$\|y_\varrho\|_{H^s(\Omega)} \leq c\|y_d\|_{H^s(\Omega)} \quad \text{for } s \in [0,2]. \tag{4.23}$$

Proof. To show (4.21) consider the linear mapping $Ty_d := y_\varrho - y_d$. By (4.15) and (4.19) it holds that

$$\begin{aligned}
\|Ty_d\|_{L^2(\Omega)} &\leq \|y_d\|_{L^2(\Omega)}, \quad \text{for all } y_d \in L^2(\Omega),\\
\|Ty_d\|_{L^2(\Omega)} &\leq \sqrt{\varrho}\|y_d\|_{H_0^1(\Omega)} \leq \sqrt{\varrho}\|y_d\|_{H^1(\Omega)}, \quad \text{for all } y_d \in H_0^1(\Omega),\\
\|Ty_d\|_{L^2(\Omega)} &\leq \varrho\|\Delta y_d\|_{L^2(\Omega)} \leq \varrho\|y_d\|_{H^2(\Omega)}, \quad \text{for all } y_d \in H_0^1(\Omega) \cap H^2(\Omega),
\end{aligned}$$

i.e, $T : X \to L^2(\Omega)$ is bounded for $X \in \{L^2(\Omega), H_0^1(\Omega), H_0^1(\Omega) \cap H^2(\Omega)\}$. Using Theorem 2.14 we deduce that $T : H_0^s(\Omega) \to L^2(\Omega)$ is bounded for all $s \in [0,1]$ and $T : H_0^1(\Omega) \cap H^s(\Omega) \to L^2(\Omega)$ is bounded for all $s \in (1,2]$ and

$$\|y_\varrho - y_d\|_{L^2(\Omega)} = \|Ty_d\|_{L^2(\Omega)} \leq c\varrho^{s/2}\|y_d\|_{H^s(\Omega)}.$$

The estimate (4.22) follows analogously defining $T : X \to H_0^1(\Omega)$ and using (4.17) and (4.19). For (4.23) first note if Ω is convex or $\partial\Omega$ is $\mathcal{C}^2$, we have that

$$\sum_{|\alpha|=2} \|D^\alpha y\|_{L^2(\Omega)} \le c\|\Delta y\|_{L^2(\Omega)}, \tag{4.24}$$

for all $y \in H^1_\Delta(\Omega) := \{y \in H_0^1(\Omega) : \Delta y \in L^2(\Omega)\}$, see, e.g., [30, 40, 59]. Obviously, it holds that

$$\|\Delta y\|_{L^2(\Omega)} \le \|y\|_{H^2(\Omega)} \quad \text{for all } y \in H_0^1(\Omega) \cap H^2(\Omega).$$

Using (4.24) and the Poincare inequality we can estimate

$$\|y\|_{H^2(\Omega)} \le c\big(\|y\|_{H_0^1(\Omega)} + \|\Delta y\|_{L^2(\Omega)}\big).$$

Further, integrating by parts, gives for $y \in H^1_\Delta(\Omega)$, that

$$\|y\|^2_{H_0^1(\Omega)} = \int_\Omega \nabla y \cdot \nabla y \, dx = \langle -\Delta y, y\rangle_\Omega \le \|\Delta y\|_{L^2(\Omega)}\|y\|_{L^2(\Omega)} \le c_P\|\Delta y\|\|y\|_{H_0^1(\Omega)}$$

and therefore $\|y\|_{H_0^1(\Omega)} \le c\|\Delta y\|_{L^2(\Omega)}$. Altogether, we see that

$$\|\Delta y\|_{L^2(\Omega)} \le \|y\|_{H^2(\Omega)} \le c\|\Delta y\|_{L^2(\Omega)} \quad \text{and} \quad H^1_\Delta(\Omega) = H_0^1(\Omega) \cap H^2(\Omega).$$

Now consider the linear mapping $\hat{T}y_d = y_\varrho$. By (4.16), (4.18) and (4.20) we have that

$$\begin{aligned}
\|\hat{T}y_d\|_{L^2(\Omega)} &= \|y_\varrho\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)} \quad \text{for all } y_d \in L^2(\Omega),\\
\|\hat{T}y_d\|_{H^1(\Omega)} &= \|y_\varrho\|_{H^1(\Omega)} \le c\|y_\varrho\|_{H_0^1(\Omega)} \le c\|y_d\|_{H^1(\Omega)} \quad \text{for all } y_d \in H_0^1(\Omega),\\
\|\hat{T}y_d\|_{H^2(\Omega)} &= \|y_\varrho\|_{H^2(\Omega)} \le c\|\Delta y_\varrho\|_{L^2(\Omega)} \le c\|y_d\|_{H^2(\Omega)} \quad \text{for all } y_d \in H_0^1(\Omega) \cap H^2(\Omega).
\end{aligned}$$

Thus, $\hat{T} : X \to X$ is bounded for $X \in \{L^2(\Omega), H_0^1(\Omega), H_0^1(\Omega) \cap H^2(\Omega)\}$ and by Theorem 2.14 $\hat{T} : H_0^s(\Omega) \to H_0^s(\Omega)$ is bounded for all $s \in [0, 1]$ and $\hat{T} : H_0^1(\Omega) \cap H^s(\Omega) \to H_0^1(\Omega) \cap H^s(\Omega)$ is bounded for $s \in (1, 2]$ and

$$\|y_\varrho\|_{H^s(\Omega)} = \|\hat{T}y_d\|_{H^s(\Omega)} \le c\|y_d\|_{H^s(\Omega)}. \qquad \square$$

Discretization

In the following we discuss the discretization and give discretization error estimates. We assume that $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a bounded and convex Lipschitz domain, which is polygonally ($d = 2$) or polyhedrally ($d = 3$) bounded. For the discretization we consider the space $S_h^1(\mathcal{T}_h)$ of globally continuous, piecewise linear functions defined on a admissible and shape regular decomposition $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$. As a conforming ansatz space, we use $Y_h = S_h^1(\mathcal{T}_h) \cap H_0^1(\Omega)$. Then the discrete variational formulation of (4.6) is to find $y_{\varrho h} \in Y_h$ such that

$$\varrho \langle \nabla y_{\varrho h}, \nabla z_h \rangle_{L^2(\Omega)} + \langle y_{\varrho h}, z_h \rangle_{L^2(\Omega)} = \langle y_d, z_h \rangle_{L^2(\Omega)} \quad \text{for all } z_h \in Y_h. \tag{4.25}$$

This is exaclty (3.21), and we get unique solvability by Lemma 3.10 and best approximation error estimates by Theorem 3.11.

THEOREM 4.5. *Let $y_d \in L^2(\Omega)$. For the unique solution $y_{\varrho h} \in Y_h$ of (4.25) there holds*

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}. \tag{4.26}$$

If additionally, $y_d \in H_0^1(\Omega)$, we have

$$\|y_{\varrho h} - y_d\|_{L(\Omega)} \le c\Big(\sqrt{\varrho}\|y_d\|_{H_0^1(\Omega)} + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|^2_{H_0^1(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big) \tag{4.27}$$

and

$$\sqrt{\varrho}\|y_{\varrho h} - y_d\|_{H_0^1(\Omega)} \le c\Big(\sqrt{\varrho}\|y_d\|_{H_0^1(\Omega)} + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|^2_{H_0^1(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big). \tag{4.28}$$

Moreover, if $y_d \in H_0^1(\Omega)$ and $\Delta y_d \in L^2(\Omega)$ we have the error estimates

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le c\Big(\varrho\|\Delta y_d\|_{L^2(\Omega)} + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|^2_{H_0^1(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big) \tag{4.29}$$

and

$$\sqrt{\varrho}\|y_{\varrho h} - y_d\|_{H_0^1(\Omega)} \le c\Big(\varrho\|\Delta y_d\|_{L^2(\Omega)} + \inf_{z_h \in Y_h} \big[\varrho \|y_d - z_h\|^2_{H_0^1(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big). \tag{4.30}$$

Using these results and the best approximation properties we can now derive the optimal choice of the regularization parameter $\varrho > 0$ and error estimates in broken Sobolev spaces.

THEOREM 4.6. *Let $y_d \in H_0^s(\Omega)$ for $s \in [0,1]$ or $y_d \in H_0^1(\Omega) \cap H^s(\Omega)$ for $s \in (1,2]$. If $\varrho = h^2$, then*

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le c h^s \|y_d\|_{H^s(\Omega)} \quad \textit{for all } s \in [0,2] \tag{4.31}$$

and

$$\|y_{\varrho h} - y_d\|_{H_0^1(\Omega)} \le c h^{s-1} \|y_d\|_{H^s(\Omega)} \quad \textit{for all } s \in [1,2]. \tag{4.32}$$

Proof. Firstly, for $y_d \in L^2(\Omega)$ (4.26) gives

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}.$$

Using the best approximation estimates of Y_h, see Theorem 2.28, we get for $y_d \in H_0^1(\Omega)$ that

$$\inf_{z_h \in Y_h} \left[h^2 \|y_d - z_h\|^2_{H_0^1(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\right]^{1/2} \le ch\|y_d\|_{H^1(\Omega)},$$

while for $y_d \in H_0^1(\Omega) \cap H^2(\Omega)$ we get

$$\inf_{z_h \in Y_h} \left[h^2 \|y_d - z_h\|^2_{H_0^1(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\right]^{1/2} \le ch^2\|y_d\|_{H^2(\Omega)}.$$

With $\varrho = h^2$ the estimate (4.27) now becomes

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le c(\sqrt{\varrho} + h)\|y_d\|_{H^1(\Omega)} = ch\|y_d\|_{H^1(\Omega)},$$

and, analogously, (4.29) can be bounded by

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le c(\varrho + h^2)\|y_d\|_{H^1(\Omega)} = ch^2\|y_d\|_{H^2(\Omega)}.$$

Therefore, the operator $T : X \to L^2(\Omega)$ defined as $Ty_d = y_{\varrho h} - y_d$ is bounded for $X \in \{L^2(\Omega), H_0^1(\Omega), H_0^1(\Omega) \cap H^2(\Omega)\}$ and by Theorem 2.14, $T : H_0^s(\Omega) \to L^2(\Omega)$ is bounded for all $s \in [0, 1]$ and $T : H_0^1(\Omega) \cap H^s(\Omega) \to L^2(\Omega)$ is bounded for all $s \in (1, 2]$ with

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} = \|Ty_d\|_{L^2(\Omega)} \le ch^s\|y_d\|_{H^s(\Omega)},$$

which is (4.31). The estimate (4.32) follows analogously, using the estimates (4.28) and (4.30), $\varrho = h^2$ and the space interpolation argument. We skip the details. □

The next Lemma states the convergence rates for the cost functional, depending on the regularity of the target $y_d \in H_0^s(\Omega)$, $s \in [0, 1]$.

LEMMA 4.7. *Let $y_d \in H_0^s(\Omega)$ for $s \in [0, 1]$ and let $y_{\varrho h} \in Y_h$ be the unique solution of (4.25). Let $\mathcal{T}_h$ be locally quasi-uniform and choose $\varrho = h^2$, then*

$$\widetilde{\mathcal{J}}(y_{\varrho h}) = \frac{1}{2}\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|\nabla y_{\varrho h}\|^2_{L^2(\Omega)} \le ch^{2s}\|y_d\|^2_{H^s(\Omega)}. \tag{4.33}$$

Proof. Let $y_d \in L^2(\Omega)$. Then by (4.31) it holds

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}.$$

Testing (4.25) with $z_h = y_{\varrho h} \in Y_h$, we obtain

$$\varrho\|\nabla y_{\varrho h}\|^2_{L^2(\Omega)} + \|y_{\varrho h}\|^2_{L^2(\Omega)} = \langle y_d, y_{\varrho h}\rangle_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}\|y_{\varrho h}\|_{L^2(\Omega)},$$

from which we immediately conclude

$$\|y_{\varrho h}\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)} \quad \text{and} \quad \|\nabla y_{\varrho h}\|_{L^2(\Omega)} \le \frac{1}{\sqrt{\varrho}}\|y_d\|_{L^2(\Omega)}.$$

Thus, we get

$$\begin{aligned} \widetilde{\mathcal{J}}(y_{\varrho h}) &= \frac{1}{2}\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|\nabla y_{\varrho h}\|^2_{L^2(\Omega)} \\ &\le \frac{1}{2}\|y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\frac{1}{\varrho}\|y_d\|^2_{L^2(\Omega)} \\ &= \|y_d\|^2_{L^2(\Omega)}. \end{aligned} \tag{4.34}$$

Now, let $y_d \in H^1_0(\Omega)$. Then by (4.31) we have

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le ch\|y_d\|_{H^1(\Omega)}.$$

Further, let $Q_h : L^2(\Omega) \to Y_h$ denote the L^2-projection, defined as

$$\langle Q_h u, z_h\rangle_{L^2(\Omega)} = \langle u, z_h\rangle_{L^2(\Omega)} \quad \text{for all } z_h \in Y_h.$$

Then we compute by (4.25)

$$\begin{aligned} \varrho\|\nabla y_{\varrho h}\|^2_{L^2(\Omega)} &= \langle y_d - y_{\varrho h}, y_{\varrho h}\rangle_{L^2(\Omega)} = \langle Q_h(y_d - y_{\varrho h}), y_{\varrho h}\rangle_{L^2(\Omega)} \\ &= -\|Q_h(y_d - y_{\varrho h})\|^2_{L^2(\Omega)} + \langle Q_h(y_d - y_{\varrho h}), y_d\rangle_{L^2(\Omega)} \\ &\le \langle y_d - y_{\varrho h}, Q_h y_d\rangle_{L^2(\Omega)} \\ &= \varrho\langle \nabla y_{\varrho h}, \nabla Q_h y_d\rangle_{L^2(\Omega)} \\ &\le \varrho\|\nabla y_{\varrho h}\|_{L^2(\Omega)}\|Q_h y_d\|_{H^1(\Omega)}, \end{aligned}$$

where we used that Q_h is self-adjoint. By the local quasi-uniformity of the triangulation $\mathcal{T}_h$, we know that the L^2-projection is H^1-stable, see [16, 21], and we get $\|Q_h y_d\|_{H^1(\Omega)} \le c\, \|y_d\|_{H^1(\Omega)}$ from which we conclude

$$\|\nabla y_{\varrho h}\|_{L^2(\Omega)} \le c\|y_d\|_{H^1(\Omega)}.$$

Thus, for the cost functional we have, using $\varrho = h^2$, that

$$\begin{aligned} \widetilde{\mathcal{J}}(y_{\varrho h}) &= \frac{1}{2}\|y_d - y_{\varrho h}\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|\nabla y_{\varrho h}\|^2_{L^2(\Omega)} \\ &\le \frac{1}{2}ch^2\|y_d\|^2_{H^1(\Omega)} + \frac{\varrho}{2}c\|y_d\|^2_{H^1(\Omega)} \\ &\le ch^2\|y_d\|^2_{H^1(\Omega)}. \end{aligned} \tag{4.35}$$

Interpolating (4.34) and (4.35), see Theorem 2.14, we get

$$\widetilde{\mathcal{J}}(y_{\varrho h}) \le ch^{2s}\|y_d\|^2_{H^s(\Omega)}, \quad s \in [0, 1]. \qquad \square$$

REMARK 4.8. *We showed optimal orders of convergence in* (4.33) *for* $y_d \in H^s(\Omega)$, *only for* $s \in [0,1]$ *and for the choice* $\varrho = h^2$, *where the best rate possible is*

$$\widetilde{\mathcal{J}}(y_{\varrho h}) \leq h^2 \|y_d\|^2_{H^1(\Omega)} \quad \text{if } y_d \in H^1_0(\Omega). \tag{4.36}$$

Now, consider $y_d \in H^1_0(\Omega) \cap H^s(\Omega)$ *for* $s \in (1,2]$. *By Theorem 4.5* (4.29) *we have*

$$\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} \leq c(\varrho^2 + \varrho h^2 + h^4)\|y_d\|^2_{H^2(\Omega)}$$

and we can bound

$$\|\nabla y_{\varrho h}\|_{L^2(\Omega)} \leq c\|y_d\|_{H^1(\Omega)} \leq c\|y_d\|_{H^2(\Omega)}$$

as in the preceeding proof. For the cost functional we then get,

$$\widetilde{\mathcal{J}}(y_{\varrho h}) \leq c(\varrho^2 + \varrho h^2 + h^4 + \varrho)\|y_d\|^2_{H^2(\Omega)}.$$

This reveals that also in the case of targets of higher regularity, we will only see a quadratic rate of convergence when choosing $\varrho = h^2$. *Thus, the energy regularization in* $H^{-1}(\Omega)$ *is well suited for targets that are less regular, e.g. discontinuous targets, while for more regular targets one should consider the control in* L^2, *as will be discussed in Section 4.1.4.*

Numerical results

The finite elements space $Y_h = S^1_h(\mathcal{T}_h) \cap H^1_0(\Omega)$ is spanned by piecewise linear, globally continuous functions $\{\varphi^1_k\}^M_{k=1}$. Therefore, for all $z_h \in Y_h$ the representation

$$z_h(x) = \sum_{k=1}^{M} z_k \varphi^1_k(x), \quad \text{with } z_k = z_h(x_k)$$

holds true, which defines the finite element isomorphism $Y_h \ni z_h \leftrightarrow \mathbf{z}_h \in \mathbb{R}^M$, where $\mathbf{z}_h[k] = z_k$, $k = 1, \ldots, M$. The discrete variational formulation (4.25) is then equivalent to the linear system of equations

$$(\varrho K_h + M_h)\mathbf{y}_{\varrho h} = \mathbf{y}_{dh}, \tag{4.37}$$

where the stiffness matrix and the mass matrix are given as

$$K_h[i,j] = \int_\Omega \nabla\varphi^1_j(x) \cdot \nabla\varphi^1_i(x)\, dx \quad \text{and} \quad M_h[i,j] = \int_\Omega \varphi^1_j(x)\varphi^1_i(x)\, dx, \quad i,j = 1, \ldots, M$$

and the load vector has the entries

$$\mathbf{y}_{dh}[i] = \int_\Omega y_d(x)\varphi^1_i(x)\, dx, \quad i = 1, \ldots, M.$$

Remark 4.9. *The discretization of the* L^2*-regularization* (4.13) *is equivalent to the linear system of equations*

$$\begin{pmatrix} \varrho^{-1} M_h & K_h \\ -K_h^\top & M_h \end{pmatrix} \begin{pmatrix} \boldsymbol{p}_{\varrho h} \\ \boldsymbol{y}_{\varrho h} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0}_h \\ \boldsymbol{y}_{dh} \end{pmatrix}. \tag{4.38}$$

A thorough analysis including the derivation of regularization and finite element error estimates in this case is given in [76]. A different derivation is given in Section 4.1.4. We will not repeat the analysis at this point, but we will compare the current approach to this method. It is important to note, that for $y_d \in H_0^s(\Omega)$*,* $s \in [0,1]$ *or* $y_d \in H_0^1(\Omega) \cap H^s(\Omega)$*,* $s \in (1,2]$ *the error estimate*

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \leq c h^s \|y_d\|_{H^s(\Omega)}$$

holds, for the choice $\varrho = h^4$*.*

To show the sharpness of the theoretical results, we consider three targets of different regularity defined on the unit square in $\Omega = (0,1)^2 \subset \mathbb{R}^2$. First, we consider $y_{d,1} \in \mathcal{C}^2(\overline{\Omega}) \cap H_0^1(\Omega)$ defined as

$$y_{d,1}(x,y) = \begin{cases} \frac{1}{2}(6y-3x-2)^3(3x-6y)^3 \sin(\pi x), & x \leq 2y \text{ and } 6y-3x \leq 2, \\ 0, & \text{else.} \end{cases} \tag{4.39}$$

As a second example we consider a piecewise bilinear function $y_{d,2} \in H^{3/2-\varepsilon}(\Omega) \cap H_0^1(\Omega)$, $\varepsilon > 0$, defined as

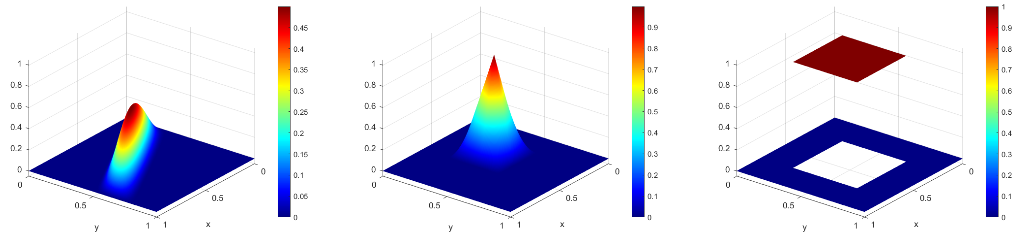
$$y_{d,2}(x,y) = \phi(x)\phi(y), \quad \phi(x) = \begin{cases} 1, & x = 0.45, \\ 0, & x \notin [0.2, 0.6], \\ \text{linear}, & \text{else.} \end{cases} \tag{4.40}$$

And finally, a discontinuous target $y_{d,3} \in H_0^{1/2-\varepsilon}(\Omega)$, $\varepsilon > 0$, defined as

$$y_{d,3}(x,y) = \begin{cases} 1, & (x,y) \in (0.25, 0.75)^2 \subset \Omega, \\ 0, & \text{else.} \end{cases} \tag{4.41}$$

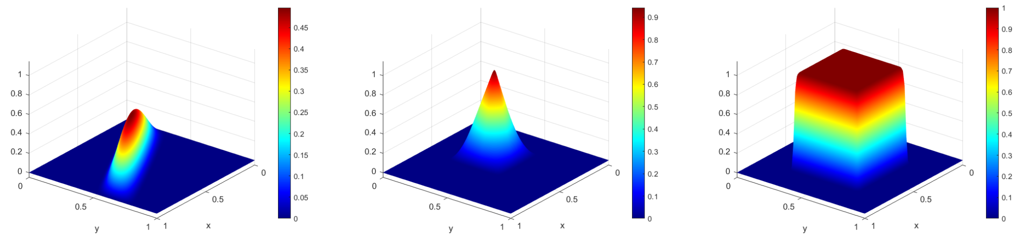
The targets are depicted in Figure 4.1.

The convergence rates for a uniform refinement are computed for an initial triangulation with $N = 128$ elements and $M = 49$ degrees of freedom (DoFs), see Figure 4.3, for all three targets for both, the energy regularization in $H^{-1}(\Omega)$ solving (4.37) and the common L^2-regularization solving (4.38). Firstly, for a fixed parameter $\varrho > 0$, we clearly see optimal convergence rates at first, which break down when $h = \varrho^{1/2}$ in the case of the energy regularization and $h = \varrho^{1/4}$ in the case of the common



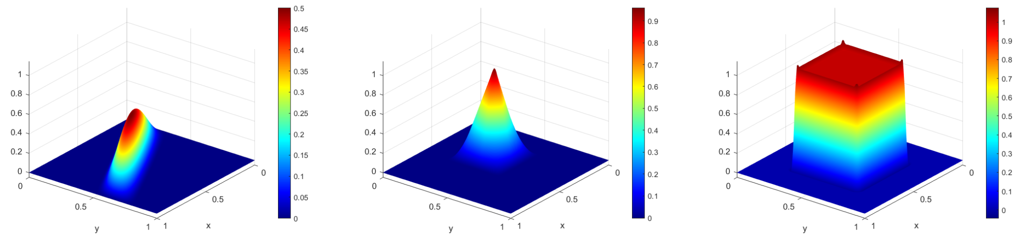
a) $y_{d,1}$ b) $y_{d,2}$ c) $y_{d,3}$

Figure 4.1: Target functions $y_{d,i}$, $i = 1, 2, 3$.



a) $y_{\varrho h,1}$ b) $y_{\varrho h,2}$ c) $y_{\varrho h,3}$

Energy regularization (4.37) with $\varrho = h^2$



d) $y_{\varrho h,1}$ e) $y_{\varrho h,2}$ f) $y_{\varrho h,3}$

Common L^2-regularization (4.38) with $\varrho = h^4$

Figure 4.2: Reconstructed target functions $y_{\varrho h,i}$, $i = 1, 2, 3$, on a mesh with $N = 32768$ elements and $M = 16129$ DoFs.

L^2-regularization, but independent of the regularity of the target. This is in total agreement with our theory, as the estimates in Theorem 4.5 show best approximation properties until the constant term dominates the error. Secondly, Figure 4.3 shows the convergence for the optimal choice $\varrho = h^2$ for the energy regularization and $\varrho = h^4$ for the common L^2-regularization. We see optimal convergence for all three targets, again supporting the theoretical error estimates in Theorem 4.6. Although, the convergence rates indicate a similar behavior of the reconstructed targets, Figure 4.2 reveals a qualitatively different behavior in the case of the discontinuous target. While for the common L^2-regularization one observes oscillations around the jump, the energy regularization gives sharp results. In Figure 4.4 the convergence of the cost functional is plotted for a fixed parameter $\varrho = 10^{-8}$ and the optimal choice $\varrho = h^2$. We clearly see, that for a fixed parameter the convergence is optimal up to the point where $h^4 \sim \varrho$, while for $\varrho = h^2$ we only see the optimal rate for the target $y_{d,3} \in H^{1/2-\varepsilon}(\Omega)$ and a quadratic rate for the other two targets, which supports the results in Remark 4.8.

Reconstruction of the control

In order to compute a discrete reconstruction of the control $u_\varrho \in H^{-1}(\Omega)$ we have two different options. Firstly, we can choose $U_h = Y_h \subset H^{-1}(\Omega)$ as a conforming trial space spanned by the piecewise linear, globally continuous functions $\{\varphi^1_{h,i}\}_{i=1}^{M_h}$. Then, for the unique solution $y_{\varrho h} \in Y_h$ of (4.25) we compute $u_{\varrho h} \in Y_h$ by solving

$$\langle u_{\varrho h}, v_h \rangle_{L^2(\Omega)} = \langle \nabla y_{\varrho h}, \nabla v_h \rangle_{L^2(\Omega)} \quad \text{for all } v_h \in Y_h. \tag{4.42}$$

Using the fe-isomorphism this is equivalent to the linear system of equations

$$M_h \mathbf{u}_{\varrho h} = K_h \mathbf{y}_{\varrho h}, \tag{4.43}$$

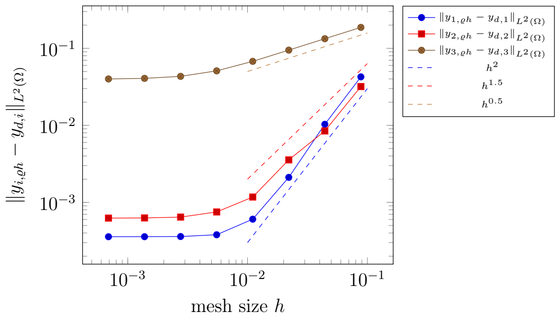
with mass and stiffness matrices as in (4.37). In this case $u_{\varrho h} \in Y_h$ is the discrete Riesz representant of $-\Delta y_{\varrho h} \in H^{-1}(\Omega)$. For a target $y_d \in H^s(\Omega) \cap H^1_0(\Omega)$, $s \in [1,2]$, using a Strang Lemma argument, we can derive the error estimate

$$\|u_\varrho - u_{\varrho h}\|_{H^{-1}(\Omega)} \leq c h^s \|y_d\|_{H^s(\Omega)}, \; s \in [1,2].$$

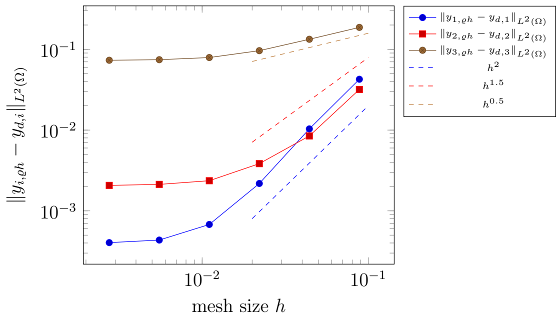
The results are depicted in Figure 4.5. However, $u_{\varrho h} \in Y_h$ enforces homogeneous Dirichlet boundary conditions and, for discontinuous controls $u_\varrho \in H^{-1}(\Omega)$, this approach seems not suitable, as $u_{\varrho h} \in Y_h \subset \mathcal{C}(\Omega)$ is continuous. Hence, we follow the ideas of Section 3.1.3 to give a more rigorous approach. Therefore, we choose $U_H = S^0_H(\mathcal{T}_H) \subset H^{-1}(\Omega)$ as conforming subspace, spanned by the piecewise constant functions $\{\varphi^0_{H,\ell}\}_{\ell=1}^{N_H}$, where we assume that the decompositions $\mathcal{T}_H$ and $\mathcal{T}_h$ are nested, i.e., $S^0_H(\mathcal{T}_H) \subset S^0_h(\mathcal{T}_h)$. Then we need to solve (3.41), which in this case reads: find $(\hat{p}_h, u_{\varrho H}) \in Y_h \times U_H$ such that

$$\langle \nabla \hat{p}_h, \nabla q_h \rangle_{L^2(\Omega)} + \langle u_{\varrho H}, q_h \rangle_{L^2(\Omega)} = \langle \nabla y_{\varrho h}, \nabla q_h \rangle_{L^2(\Omega)}, \quad \langle v_H, \hat{p}_h \rangle_{L^2(\Omega)} = 0, \tag{4.44}$$

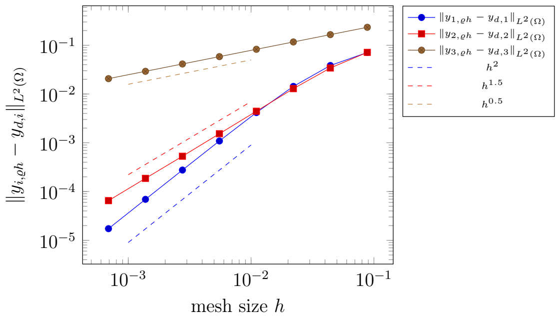
for all $(q_h, v_H) \in Y_h \times U_H$. Transferring Theorem 3.20 we get the following.



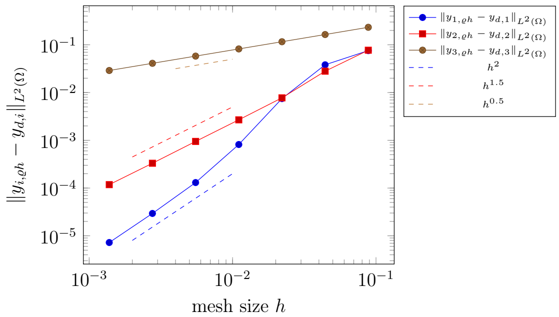
a) Energy regularization, $\varrho = 10^{-5}$



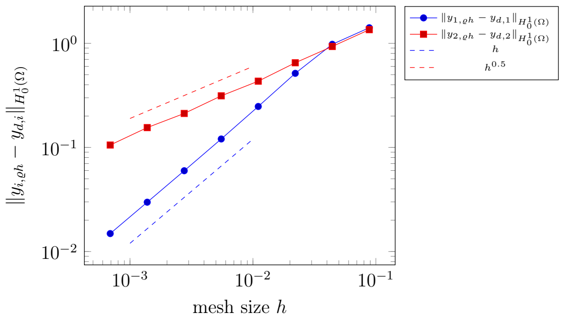
b) L^2-regularization, $\varrho = 10^{-8}$



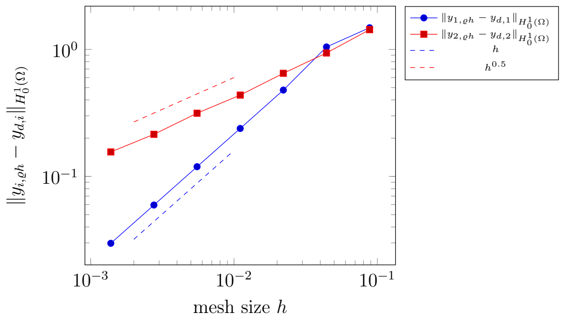
c) Energy regularization, $\varrho = h^2$



d) L^2-regularization, $\varrho = h^4$

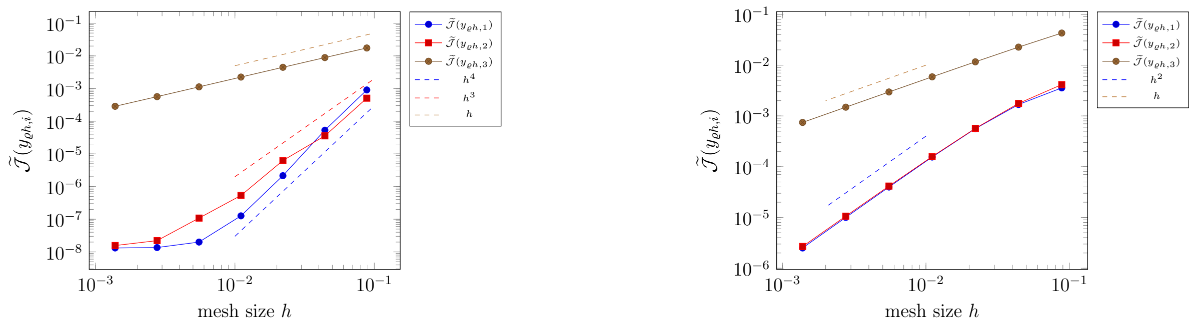


e) Energy regularization, $\varrho = h^2$



f) L^2-regularization, $\varrho = h^4$

Figure 4.3: Convergence for the three different target functions $y_{d,i}$, $i = 1, 2, 3$ for the energy regularization in $H^{-1}(\Omega)$ solving (4.37) and the common L^2 regularization solving (4.38) for different choices of $\varrho > 0$.



a) $\varrho = 10^{-8}$ b) $\varrho = h^2$

Figure 4.4: Convergence of the cost functional $\widetilde{\mathcal{J}}$.

THEOREM 4.10. *Let the discrete inf-sup stability condition*

$$c_S \|v_H\|_{H^{-1}(\Omega)} \leq \sup_{0 \neq q_h \in Y_h} \frac{\langle v_H, q_h \rangle_{L^2(\Omega)}}{\|\nabla q_h\|_{L^2(\Omega)}} \quad \text{for all } v_H \in U_H \tag{4.45}$$

hold true, for some $c_S > 0$. Then the discrete variational formulation (4.44) *admits a unique solution $(\hat{p}_h, u_{\varrho H}) \in Y_h \times U_H$. Further, denote by $u_\varrho = -\Delta y_\varrho \in H^{-1}(\Omega)$, where $y_\varrho \in H_0^1(\Omega)$ denotes the unique solution of* (4.6)*. If $y_d \in H_0^1(\Omega) \cap H^s(\Omega)$ for $s \in [1, 2]$ and $\varrho = h^2$, then*

$$\|u_\varrho - u_{\varrho H}\|_{H^{-1}(\Omega)} \leq c H^{s-1} \|y_d\|_{H^s(\Omega)}. \tag{4.46}$$

Proof. Unique solvability follows directly from Theorem 3.20, and we also get the best approximation error estimate

$$\|u_\varrho - u_{\varrho H}\|_{H^{-1}(\Omega)} \leq c\Big(\inf_{v_H \in U_H} \|u_\varrho - v_H\|_{H^{-1}(\Omega)} + \|\nabla(y_{\varrho h} - y_\varrho)\|_{L^2(\Omega)} \Big). \tag{4.47}$$

Firstly, note that subtracting (4.25) from (4.6) we conclude Galerkin orthogonality and we get Cea's Lemma, i.e.,

$$\begin{aligned} &\varrho \|\nabla(y_\varrho - y_{\varrho h})\|^2_{L^2(\Omega)} + \|y_\varrho - y_{\varrho h}\|^2_{L^2(\Omega)} \\ &\qquad \leq \inf_{z_h \in Y_h} \Big[\varrho \|\nabla(y_\varrho - z_h)\|^2_{L^2(\Omega)} + \|y_\varrho - z_h\|^2_{L^2(\Omega)} \Big]. \end{aligned}$$

For $y_d \in H_0^1(\Omega)$, we can therefore estimate, using (4.4) and (4.18),

$$\|u_\varrho - u_{\varrho H}\|_{H^{-1}(\Omega)} \leq c\big(\|u_\varrho\|_{H^{-1}(\Omega)} + \|\nabla y_\varrho\|_{L^2(\Omega)} \big) \leq c \|y_\varrho\|_{H_0^1(\Omega)} \leq c \|y_d\|_{H_0^1(\Omega)}. \tag{4.48}$$

Now, consider $y_d \in H_0^1(\Omega) \cap H^2(\Omega)$. Then by (4.20), we have that $u_\varrho = -\Delta y_\varrho \in L^2(\Omega)$. In order to estimate the first term in (4.47) denote by $Q_H^0 : L^2(\Omega) \to U_H$ the L^2-projection defined as

$$\langle Q_H^0 u, v_H \rangle_{L^2(\Omega)} = \langle u, v_H \rangle_{L^2(\Omega)} \quad \text{for all } v_H \in U_H.$$

Using the Galerkin orthogonality, i.e., $\langle u_\varrho - Q_H^0 u_\varrho, v_H\rangle_{L^2(\Omega)} = 0$ for all $v_H \in U_H$, and the best approximation property (Theorem 2.35) we can estimate

$$\begin{aligned}
\|u_\varrho - Q_H^0 u_\varrho\|_{H^{-1}(\Omega)} &= \sup_{0\neq z\in H_0^1(\Omega)} \frac{\langle u_\varrho - Q_H^0 u_\varrho, z\rangle_{L^2(\Omega)}}{\|\nabla z\|_{L^2(\Omega)}} \\
&= \sup_{0\neq z\in H_0^1(\Omega)} \frac{\langle u_\varrho - Q_H^0 u_\varrho, z - Q_H^0 z\rangle_{L^2(\Omega)}}{\|\nabla z\|_{L^2(\Omega)}} \\
&\leq \sup_{0\neq z\in H_0^1(\Omega)} \frac{\|u_\varrho - Q_H^0 u_\varrho\|_{L^2(\Omega)}\|z - Q_H^0 z\|_{L^2(\Omega)}}{\|\nabla z\|_{L^2(\Omega)}} \\
&\leq \sup_{0\neq z\in H_0^1(\Omega)} \frac{\|u_\varrho\|_{L^2(\Omega)} cH \|\nabla z\|_{L^2(\Omega)}}{\|\nabla z\|_{L^2(\Omega)}} \\
&= cH\|u_\varrho\|_{L^2(\Omega)} = cH\|\Delta y_\varrho\|_{L^2(\Omega)} \leq cH\|y_d\|_{H^2(\Omega)}.
\end{aligned}$$

The second term in (4.47) can be estimated, using $\varrho = h^2$ and the best approximation error estimates, by

$$\|\nabla(y_\varrho - y_{\varrho h})\|_{L^2(\Omega)} \leq ch\|y_\varrho\|_{H^2(\Omega)} \leq ch\|y_d\|_{H^2(\Omega)},$$

and overall (4.47) admits the estimate

$$\|u_\varrho - u_{\varrho H}\|_{H^{-1}(\Omega)} \leq cH\|y_d\|_{H^2(\Omega)}. \tag{4.49}$$

Thus, interpolating (4.48) and (4.49), we conclude

$$\|u_\varrho - u_{\varrho h}\|_{H^{-1}(\Omega)} \leq cH^{s-1}\|y_d\|_{H^s(\Omega)}$$

for all $s \in [1, 2]$. □

REMARK 4.11. *The convergence estiamte* (4.46) *only holds when considering the target $y_d \in H_0^1(\Omega) \cap H^s(\Omega)$ for $s \in [1, 2]$. In order to have convergence rates also for less regular targets, one can derive error estimates in $H^{-2}(\Omega)$, as was done in [50].*

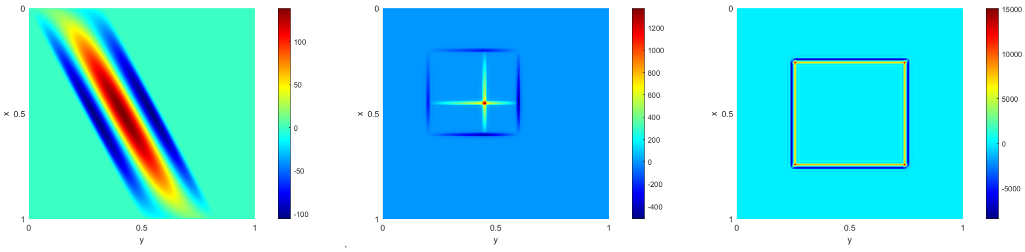
By the fe-isomorphism the variational formulation (4.44) is equivalent to the system of linear equations

$$\begin{pmatrix} K_h & \hat{M}_h \\ \hat{M}_h^\top & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{p}}_h \\ \mathbf{u}_{\varrho H} \end{pmatrix} = \begin{pmatrix} K_h \mathbf{y}_{\varrho h} \\ \mathbf{0}_H \end{pmatrix}, \tag{4.50}$$

with stiffness matrix K_h as above and mass matrix

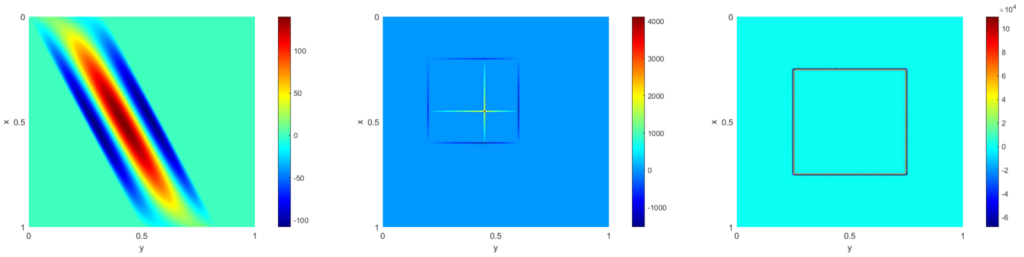
$$\hat{M}_h[i, \ell] = \langle \varphi_{H,\ell}^0, \varphi_{h,i}^1\rangle_{L^2(\Omega)}, \quad i = 1, \ldots, M_h,\ \ell = 1, \ldots, N_H.$$

In order to fulfill the discrete inf-sup stability (4.45) we choose $h = H/4$, i.e., $\mathcal{T}_h$ is twice uniformly refined with respect to $\mathcal{T}_H$. The reconstructed controls are depicted in Figure 4.5.



a) $u_{\varrho h,1}$ b) $u_{\varrho h,2}$ c) $u_{\varrho h,3}$

$u_{\varrho h,i} \in Y_h$ computed by solving (4.43)



d) $u_{\varrho H,1}$ e) $u_{\varrho H,2}$ f) $u_{\varrho H,3}$

$u_{\varrho H,i} \in U_H$ computed by solving (4.50)

Figure 4.5: Reconstructed controls on a mesh with $N = 32768$ elements and $M = 16129$ DoFs.

4.1.2 Adaptive refinement

Our main goal in this section is to reconstruct a given desired state $y_d \in L^2(\Omega)$ by the computed state $y_{\varrho h} \in Y_h$ as solution of the energy regularization (4.25) with only as much effort as needed. Noting, that when measuring the distance $\|y_d - y_{\varrho h}\|_{L^2(\Omega)}$, all parameters are known, we can easily compute the local error on each element, by

$$\eta_\ell = \|y_{\varrho h} - y_d\|_{L^2(\tau_\ell)}, \quad \ell = 1, \ldots, N.$$

The global error then fulfills

$$\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} = \sum_{\ell=1}^{N} \eta_\ell^2$$

and we can use η_ℓ, $\ell = 1, \ldots, N$ as error indicator. An adaptive refinement scheme will then refine all elements $\tau_\ell \in \mathcal{T}_h$ that are marked by

$$\eta_\ell > \theta \max_{i=1,\ldots,N} \eta_i, \quad \text{for some } \theta > 0.$$

This idea dates back to Dörfler [34] and we will refer to it as Dörfler marking. Recall, that the optimal choice for the regularization parameter for the energy regularization is $\varrho = h^2$ for a uniform mesh. In general, when applying an adaptive refinement scheme, the resulting meshes will get heavily non-uniform, leading to the question, which choice of the regularization parameter is appropriate in this situation. In Figure 4.3 we saw, when choosing a fixed parameter $\varrho > 0$ we have optimal orders of convergence, whenever $h > \varrho^{1/2}$. Thus, we might always choose $\varrho = h_{min}^2$. Although, if some elements are not refined at all, this choice is a vast overestimation. Especially, for discontinuous targets $y_d \in H_0^s(\Omega)$ with $0 \leq s < 1$, it is of highest interest to keep the regularization parameter as large as possible, as in this case the problem does not admit a solution $y_\varrho \in H_0^1(\Omega)$ without regularization. In the following, we will redo the error analysis of the discrete setting, under the assumption that $\varrho = \varrho(x)$ is a function that fulfills $0 < \underline{\varrho} \leq \varrho(x) \leq \overline{\varrho} < \infty$. We will show, that the choice $\varrho(x) = h_\ell^2$ for $x \in \tau_\ell$ is an appropriate choice, to regain optimal orders of convergence, as stated for the uniform refinement in the last section.

Instead of the variational formulation (4.6), for a constant parameter $\varrho > 0$ we will now consider: for given $y_d \in L^2(\Omega)$ find $y_\varrho \in H_0^1(\Omega)$ such that

$$\int_\Omega \varrho(x) \nabla y_\varrho(x) \cdot \nabla z(x)\, dx + \int_\Omega y_\varrho(x) z(x)\, dx = \int_\Omega y_d(x) z(x)\, dx \quad \text{for all } z \in H_0^1(\Omega). \tag{4.51}$$

Note, that this corresponds to the variational formulation of the diffusion equation

$$-\operatorname{div}(\varrho(x)\nabla y_\varrho(x)) + y_\varrho(x) = y_d(x),\ x \in \Omega \quad \text{and} \quad y_\varrho(x) = 0, \ \text{ for } x \in \partial\Omega, \tag{4.52}$$

with diffusion coefficient $\varrho(x)$ which is uniformly bounded from above and below. Hence, unique solvability follows by standard arguments, as for a constant coefficient. To derive error estimates, we will first give regularization error estimates.

LEMMA 4.12. *Let $y_d \in L^2(\Omega)$ and let $y_\varrho \in H_0^1(\Omega)$ be the unique solution of (4.51). Then*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \leq \|y_d\|_{L^2(\Omega)}. \tag{4.53}$$

Further, if $y_d \in H_0^1(\Omega)$, then

$$\|y_\varrho - y_d\|_{L^2(\Omega)}^2 \leq \int_\Omega \varrho(x) |\nabla y_d(x)|^2\, dx \tag{4.54}$$

and

$$\int_\Omega \varrho(x) |\nabla (y_\varrho(x) - y_d(x))|^2\, dx \leq \int_\Omega \varrho(x) |\nabla y_d(x)|^2\, dx. \tag{4.55}$$

Proof. Testing (4.51) with $z = y_\varrho \in H_0^1(\Omega)$, we obtain

$$\begin{aligned}\int_\Omega \varrho(x)|\nabla y_\varrho(x)|^2\,dx &= \int_\Omega (y_d(x) - y_\varrho(x))y_\varrho(x)\,dx \\ &= -\int_\Omega (y_d(x) - y_\varrho(x))(y_d(x) - y_\varrho(x))\,dx \\ &\quad + \int_\Omega y_d(x)(y_d(x) - y_\varrho(x))\,dx.\end{aligned}$$

Reordering gives

$$\|y_d - y_\varrho\|^2_{L^2(\Omega)} + \int_\Omega \varrho(x)|\nabla y_\varrho(x)|^2\,dx = \int_\Omega y_d(x)(y_d(x) - y_\varrho(x))\,dx$$

from which we deduce (4.53) when applying the Cauchy–Schwarz inequality. If $y_d \in H_0^1(\Omega)$, we can choose $z = y_d - y_\varrho$ in (4.51) and compute

$$\begin{aligned}\|y_d - y_\varrho\|^2_{L^2(\Omega)} &= \int_\Omega \varrho(x)\nabla y_\varrho(x)\cdot\nabla(y_d(x) - y_\varrho(x))\,dx \\ &= -\int_\Omega \varrho(x)|\nabla(y_d(x) - y_\varrho(x))|^2\,dx \\ &\quad + \int_\Omega \varrho(x)\nabla y_d(x)\cdot\nabla(y_d(x) - y_\varrho(x))\,dx,\end{aligned}$$

from which by reordering we conclude that

$$\begin{aligned}&\|y_d - y_\varrho\|^2_{L^2(\Omega)} + \int_\Omega \varrho(x)|\nabla(y_d(x) - y_\varrho(x))|^2\,dx \\ &\quad \le \sqrt{\int_\Omega \varrho(x)|\nabla y_d(x)|^2\,dx}\sqrt{\int_\Omega \varrho(x)|\nabla(y_d(x) - y_\varrho(x))|^2\,dx}.\end{aligned}$$

This gives (4.55) and (4.54) and concludes the proof. □

Let us now turn to the discretization and give discretization error estimates, as for the uniform case. For given $y_d \in L^2(\Omega)$, the discrete variational formulation is to find $y_{\varrho h} \in Y_h = S_h^1(\mathcal{T}_h) \cap H_0^1(\Omega)$ such that

$$\int_\Omega \varrho(x)\nabla y_{\varrho h}(x)\cdot\nabla z_h(x)\,dx + \int_\Omega y_{\varrho h}(x)z_h(x)\,dx = \int_\Omega y_d(x)z_h(x)\,dx \quad \text{for all } z_h \in Y_h. \tag{4.56}$$

Again, unique solvability follows as in the uniform case, as $\varrho(x)$ is bounded uniformly and the Lemma of Lax–Milgram applies. When subtracting (4.51) from (4.56) we see that the Galerkin orthogonality

$$\int_\Omega \varrho(x)\nabla(y_{\varrho h}(x) - y_\varrho(x))\cdot\nabla z_h(x)\,dx + \int_\Omega (y_{\varrho h}(x) - y_\varrho(x))z_h(x)\,dx = 0$$

holds for all $z_h \in Y_h$, from which we immediately conclude Cea's Lemma, i.e.,

$$\int_\Omega \varrho(x)|\nabla(y_{\varrho h}(x) - y_\varrho(x))|^2\,dx + \int_\Omega |y_{\varrho h}(x) - y_\varrho(x)|^2\,dx \\ \le \inf_{z_h \in Y_h} \left[\int_\Omega \varrho(x)|\nabla(z_h(x) - y_\varrho(x))|^2\,dx + \int_\Omega |z_h(x) - y_\varrho(x)|^2\,dx \right] \tag{4.57}$$

Using the estimates we derived up to this point, we can now state the convergence properties for the adaptive refinement.

THEOREM 4.13. *Let $y_d \in L^2(\Omega)$ and let $y_{\varrho h} \in Y_h$ be the unique solution of (4.56). Then*

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}. \tag{4.58}$$

Moreover, let $\mathcal{T}_h$ be a locally quasi-uniform triangulation and let $y_d \in H^1_0(\Omega)$. If $\varrho(x) = h_\ell^2$ for $\ell = 1, \ldots, N$ it holds

$$\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} \le c \sum_{\ell=1}^N h_\ell^2 \|\nabla y_d\|^2_{H^1(\tau_\ell)}. \tag{4.59}$$

Proof. The estimate (4.58) follows as in the continuous setting in Lemma 4.12, testing (4.56) with $z_h = y_{\varrho h}$. Now let $y_d \in H^1_0(\Omega)$ for all $\ell = 1, \ldots, N$. Using a triangle inequality and Hölders inequality, we compute that

$$\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} \le 2\big(\|y_{\varrho h} - y_\varrho\|^2_{L^2(\Omega)} + \|y_\varrho - y_d\|^2_{L^2(\Omega)}\big). \tag{4.60}$$

For the second term we use the estimate (4.54) to conclude

$$\|y_\varrho - y_d\|^2_{L^2(\Omega)} \le \int_\Omega \varrho(x)|\nabla y_d(x)|^2\,dx = \sum_{\ell=1}^N h_\ell^2 \int_{\tau_\ell} |\nabla y_d(x)|^2\,dx.$$

For the first term we apply (4.57) to get

$$\|y_{\varrho h} - y_\varrho\|^2_{L^2(\Omega)} \le \inf_{z_h \in Y_h} \left[\int_\Omega \varrho(x)|\nabla(z_h(x) - y_\varrho(x))|^2\,dx + \int_\Omega |z_h(x) - y_\varrho(x)|^2\,dx \right].$$

Choosing $z_h = \Pi_h^1 y_d$ as the Scott-Zhang quasi-interpolation operator, for which

$$\|y_d - \Pi_h y_d\|^2_{L^2(\tau_\ell)} \le c h_\ell^2 \int_{\omega_\ell} |\nabla y_d(x)|^2\,dx$$

and

$$\|\nabla(y_d - \Pi_h y_d)\|^2_{L^2(\tau_\ell)} \le c \int_{\omega_\ell} |\nabla y_d(x)|^2\,dx$$

holds, where $\overline{\omega}_\ell := \bigcup_{\{j=1,\dots,N:\, \overline{\tau_\ell}\cap\overline{\tau_j}\neq\emptyset\}} \overline{\tau}_j$, see Remark 2.34 and [103], we get, together with (4.54),

$$\begin{aligned}\int_{\tau_\ell} |\Pi_h y_d(x) - y_\varrho(x)|^2\,dx &\leq 2\|\Pi_h y_d - y_d\|^2_{L^2(\tau_\ell)} + 2\|y_d - y_\varrho\|^2_{L^2(\tau_\ell)}\\ &\leq 2ch_\ell^2 \int_{\omega_\ell} |\nabla y_d(x)|^2\,dx + 2\|y_d - y_\varrho\|^2_{L^2(\tau_\ell)}.\end{aligned}$$

In the same fashion, using $\varrho(x) = h_\ell^2$ on τ_ℓ, we compute

$$\begin{aligned}\int_{\tau_\ell} \varrho(x)|\nabla(\Pi_h y_d(x) - y_\varrho(x))|^2\,dx &\leq 2h_\ell^2\|\nabla(\Pi_h y_d - y_d)\|^2_{L^2(\tau_\ell)}\\ &\qquad +2\int_{\tau_\ell}\varrho(x)|\nabla(y_d(x) - y_\varrho(x))|^2\,dx\\ &\leq 2ch_\ell^2\int_{\omega_\ell}|\nabla y_d(x)|^2\,dx\\ &\qquad +2\int_{\tau_\ell}\varrho(x)|\nabla(y_d(x) - y_\varrho(x))|^2\,dx.\end{aligned}$$

Thus, using (4.54) and (4.55), we can bound

$$\begin{aligned}\|y_{\varrho h} - y_\varrho\|^2_{L^2(\Omega)} &\leq \inf_{z_h\in Y_h}\sum_{\ell=1}^N \Big[\int_{\tau_\ell}\varrho(x)|\nabla(z_h(x) - y_\varrho(x))|^2\,dx\\ &\qquad + \int_{\tau_\ell}|z_h(x) - y_\varrho(x)|^2\,dx\Big]\\ &\leq \sum_{\ell=1}^N\Big[\int_{\tau_\ell}\varrho(x)|\nabla(\Pi_h y_d(x) - y_\varrho(x))|^2\,dx + \int_{\tau_\ell}|\Pi_h y_d(x) - y_\varrho(x)|^2\,dx\Big]\\ &\leq c\Big(\sum_{\ell=1}^N\Big[h_\ell^2\int_{\omega_\ell}|\nabla y_d(x)|^2\,dx\Big] + \int_\Omega \varrho(x)|\nabla(y_\varrho(x) - y_d(x))|^2\,dx + \|y_\varrho - y_d\|^2_{L^2(\Omega)}\Big)\\ &\leq c\Big(\sum_{\ell=1}^N\Big[h_\ell^2\int_{\omega_\ell}|\nabla y_d(x)|^2\,dx\Big] + \int_\Omega\varrho(x)|\nabla y_d(x)|^2\,dx\Big)\\ &= c\Big(\sum_{\ell=1}^N\Big[h_\ell^2\int_{\omega_\ell}|\nabla y_d(x)|^2\,dx\Big] + \sum_{\ell=1}^N\Big[h_\ell^2\int_{\tau_\ell}|\nabla y_d|^2\,dx\Big]\Big)\\ &\leq c\sum_{\ell=1}^N h_\ell^2\|\nabla y_d\|^2_{L^2(\omega_\ell)}.\end{aligned}$$

Altogether, using the local quasi-uniformity, we can bound

$$\begin{aligned}
\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} &\leq 2\|y_{\varrho h} - y_\varrho\|^2_{L^2(\Omega)} + 2\|y_\varrho - y_d\|^2_{L^2(\Omega)} \\
&\leq c\Big(\sum_{\ell=1}^N h_\ell^2 \|\nabla y_d\|^2_{L^2(\omega_\ell)} + \sum_{\ell=1}^N h_\ell^2 \|\nabla y_d\|^2_{L^2(\tau_\ell)}\Big) \\
&\leq c\sum_{\ell=1}^N \Big(c_L \sum_{\tau_k \subset \omega_\ell} h_k^2 \int_{\tau_k} |\nabla y_d(x)|^2\,dx + \sum_{k=1}^N h_k^2 \int_{\tau_k} |\nabla y_d(x)|^2\,dx\Big) \\
&\leq c\sum_{\ell=1}^N h_\ell^2 \int_{\tau_\ell} |\nabla y_d(x)|^2\,dx,
\end{aligned}$$

where c_L denotes the local quasi uniformity constant, i.e., $h_\ell^2 \leq c_L h_k^2$ for all $\tau_k \in \omega_\ell$, $\ell = 1, \dots, N$. □

Remark 4.14. *Note, that in contrast to the case of constant $\varrho > 0$, we only derived error estimates, when the regularity of $y_d \in H_0^s(\Omega)$ for $s = 0, 1$. A higher order of convergence could be analyzed, when considering more regularity on $\varrho(x)$ and applying integration by parts. A rigorous analysis and the optimal choice for more regular $\varrho(x)$ depending on the discretization is still open.*

Remark 4.15. *In the case of constant $\varrho > 0$ we applied a space interpolation argument, to derive convergence rates for $y_d \in H_0^s(\Omega)$, $s \in [0, 1]$. This is not directly applicable here, as (4.59) involves a sum and not a constant to be interpolated. One option to resolve this issue, is to consider the bound*

$$\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} \leq c\sum_{\ell=1}^N h_\ell^2 \|\nabla y_d\|^2_{H^1(\tau_\ell)} \leq c h_{max}^2 \|y_d\|^2_{H_0^1(\Omega)}, \tag{4.61}$$

which we can interpolate with (4.58). But since h_{max} does not necessarly tend to zero in an adaptive refinement routine, this does not give a meaningful result. Another approach is based on the observation that

$$\|y\|_{H_0^1(\Omega),\varrho} := \int_\Omega \varrho(x)|\nabla y(x)|^2\,dx$$

defines an equivalent norm on $H_0^1(\Omega)$, as $0 < \underline{\varrho} \leq \varrho(x) \leq \overline{\varrho} < \infty$ for all $x \in \Omega$. Now, consider the eigenfunctions $\phi_k \in H_0^1(\Omega)$ of the boundary value problem

$$-\operatorname{div}(\varrho(x)\nabla\phi_k(x)) = \lambda_k(\varrho)\phi_k(x),\ x \in \Omega, \quad \phi_k(x) = 0,\ x \in \partial\Omega.$$

If they are normalized, i.e., $\|\phi_k\|_{L^2(\Omega)} = 1$ they form an orthonormal basis in $L^2(\Omega)$ and the eigenvalues fulfill $0 < \lambda_0(\varrho) \leq \lambda_1(\varrho) \leq \dots$ and $\lambda_k(\varrho) \to \infty$ for $k \to \infty$.

Thus, each $y \in L^2(\Omega)$ admits the representation

$$y(x) = \sum_{k=0}^{\infty} y_k \phi_k(x), \quad \text{where } y_k = \int_\Omega y(x)\phi_k(x)\,dx.$$

With this we easily compute for $y \in H_0^1(\Omega)$

$$\|y\|^2_{L^2(\Omega)} = \sum_{k=0}^{\infty} |y_k|^2 \quad \text{and} \quad \|y\|^2_{H_0^1(\Omega),\varrho} = \sum_{k=0}^{\infty} \lambda_k(\varrho)|y_k|^2$$

and we can define the interpolation norm

$$\|y\|^2_{H_0^s(\Omega),\varrho} = \sum_{k=0}^{\infty} \lambda_k(\varrho)^s |y_k|^2,$$

which defines an equivalent norm on $H_0^s(\Omega)$ for all $s \in [0,1]$. Then, interpolating (4.58) *and* (4.59) *gives*

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \leq \|y_d\|_{H_0^s(\Omega),\varrho} = \sqrt{\sum_{k=0}^{\infty} \lambda_k(\varrho)^s |y_{d,k}|^2}, \quad y_{d,k} = \int_\Omega y_d(x)\phi_k(x)\,dx.$$

The explicit dependence on ϱ now depends on the eigenvalues, which depend on the geometry of the domain Ω.

Numerical results

Using the fe-isomorphism $\mathbb{R}^M \ni \mathbf{y}_{\varrho h} \leftrightarrow y_{\varrho h} \in Y_h = S_h^1(\mathcal{T}_h) \cap H_0^1(\Omega)$, the discrete variational formulation (4.56) is equivalent to the linear system of equations

$$(K_{\varrho h} + M_h)\mathbf{y}_{\varrho h} = \mathbf{y}_{dh}, \tag{4.62}$$

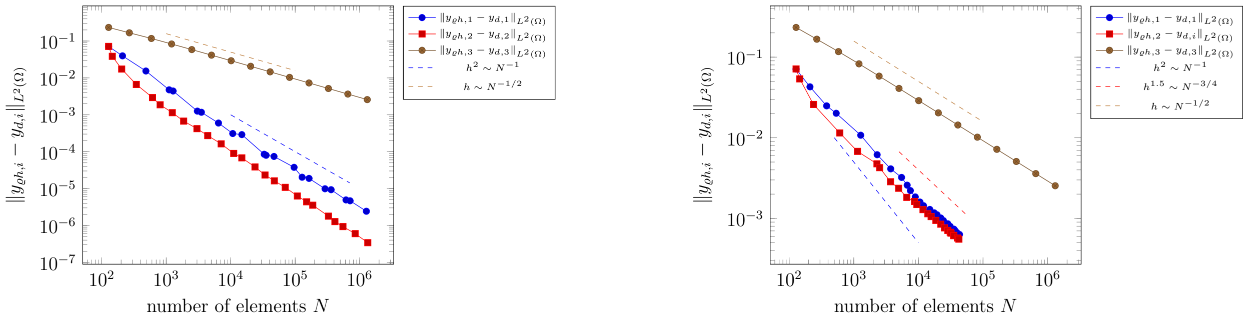
where the stiffness matrix and the mass matrix are given as

$$K_{\varrho h}[i,j] = \int_\Omega \varrho(x)\nabla\varphi_j^1(x)\cdot\nabla\varphi_i^1(x)\,dx \quad \text{and} \quad M_h[i,j] = \int_\Omega \varphi_j^1(x)\varphi_i^1(x)\,dx,\ i,j = 1,\ldots,M$$

and the load vector has the entries

$$\mathbf{y}_{dh}[i] = \int_\Omega y_d(x)\varphi_i^1(x)\,dx, \quad i = 1,\ldots,M.$$

Using a Dörfler marking strategy with $\theta = 0.5$, the results for an adaptive refinement scheme are depicted in Figure 4.6, first, for $\varrho(x) = h_{\min}^2$ and secondly, for $\varrho(x) = h_\ell^2$ for $x \in \tau_\ell$. For the continuous targets $y_{d,1}$ and $y_{d,2}$, we see that the choice of $\varrho = h_{\min}^2$



a) $\varrho \;= h_{\min}^2$ b) $\varrho(x) = h_\ell^2,\, x \in \tau_\ell$

Figure 4.6: Convergence rates of the diffusion regularization

gives an optimal order of quadratic convergence, i.e., the adaptive scheme resolves the singularity for $y_{d,2}$. For the discontinuous target, we only gain linear convergence. This can be explained as follows, using a counting argument. Let the initial mesh consist of $N \sim n^2$ elements, where n is approximately the number of elements in each column/row, as depicted in Figure 4.7. In a uniform refinement scheme, we refine each element, i.e., $N = \mathcal{O}(n^2)$ elements in total. As the discontinuity of $y_{d,3}$ is just along the boundary of $[0.25, 0.75]^2$, in order to regain the same order of convergence, it is sufficient to refine all elements that touch this boundary. Hence, in an adaptive scheme, in each step we will only refine $\mathcal{O}(n) = \mathcal{O}(\sqrt{N})$ elements, but keep the same order of convergence. Note, that this is optimal, as we can not refine less elements. Therefore, the adaptive scheme will produce approximately the same error with only $\mathcal{O}(\sqrt{N})$ of the elements compared with the uniform scheme. Using the same counting argument one can prove that the convergence behavior is dependent on the space dimension, as was observed in [74]. Furthermore, for the choice $\varrho(x) = h_\ell^2$, $\ell = 1, \ldots, N$, we observe the same, optimal, order of linear convergence, for the discontinuous target, while for the continuous targets, we see diminished orders. This is in agreement with the theory, as we cannot prove a higher order for a discontinuous function $\varrho(x)$, see Remark 4.14. In Table 4.1 the errors of the adaptive and uniform refinement schemes are compared.

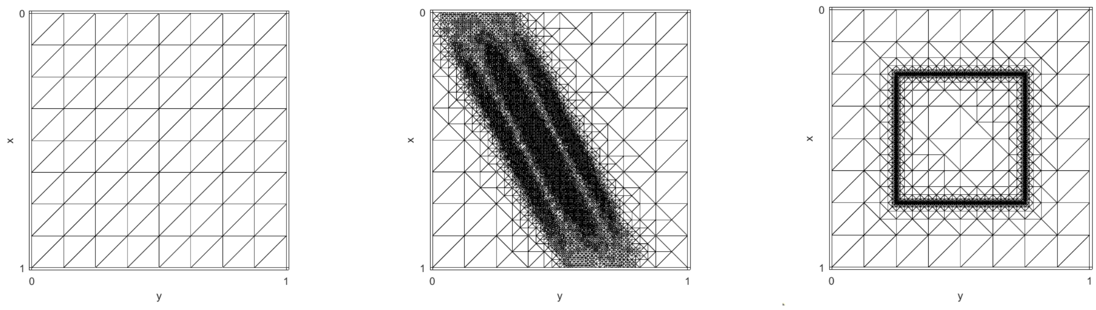
4.1.3 State and control constraints

In this section we will stick to the energy regularization in $H^{-1}(\Omega)$, i.e., in terms of the abstract theory we have

$$X = Y = H_0^1(\Omega) \quad \text{and} \quad A = B = S = -\Delta : H_0^1(\Omega) \to H^{-1}(\Omega).$$

In this setting we consider the optimal control problem to minimize

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_{H^{-1}(\Omega)} \tag{4.63}$$



a) Initial mesh $N = 128$ b) $y_{d,1}$, $N = 33103$ c) $y_{d,3}$, $N = 20340$

Figure 4.7: Initial mesh and adaptively refined meshes for the target functions $y_{d,1}$ and $y_{d,3}$.

Adaptive		Uniform	
# DoFs	$\|y_{\varrho h,3} - y_{d,3}\|_{L^2(\Omega)}$	# DoFs	$\|y_{\varrho h,3} - y_{d,3}\|_{L^2(\Omega)}$
49	2.33419e−1	49	2.33419e−1
119	1.66761e−1	225	1.65100e−1
275	1.17046e−1	961	1.16764e−1
587	8.25631e−2	3,969	8.25719e−2
1,219	5.81065e−2	16,129	5.83897e−2
2,491	4.09238e−2	65,025	4.12886e−2
5,043	2.88572e−2	261,121	2.91958e−2
10,155	2.03690e−2	1,046,529	2.06447e−2
20,387	1.43876e−2		
40,859	1.01671e−2		
81,811	7.18661e−3		
1630723	5.08065e−3		
327,555	3.59215e−3		

Table 4.1: Comparison of the errors of the adaptive refinement scheme solving (4.62) with $\varrho(x) = h_\ell^2$ for $x \in \tau_\ell$ and the uniform refinement scheme (4.37) for $\varrho = h^2$ for the discontinuous target $y_{d,3}$.

subject to

$$-\Delta y_\varrho = u_\varrho \text{ in } \Omega \quad \text{and} \quad y_\varrho = 0 \text{ on } \partial\Omega \tag{4.64}$$

and subject to either state constraints

$$g_-(x) \le y_\varrho(x) \le g_+(x) \quad \text{for } x \in \Omega,$$

where $g_\pm : \Omega \to \mathbb{R}$, which fulfill $g_-(x) \le 0 \le g_+(x)$, or subject to control constraints

$$\langle h_-, q\rangle_{L^2(\Omega)} \le \langle u_\varrho, q\rangle_\Omega \le \langle h_+, q\rangle_{L^2(\Omega)} \quad \text{for all } q \in H_0^1(\Omega),\ q(x) \ge 0,\ x \in \Omega,$$

for given $h_\pm : \Omega \to \mathbb{R}$ for which $h_-(x) \le 0 \le h_+(x)$, $x \in \Omega$ holds. The incorporation of state constraints in this setting was studied in [55], while state and control constraints are given in [50], where the analysis is trimmed to fit to the Poisson equation. In the following, we will cast this problem into the abstract framework of Section 3.2. Starting with state constraints, we derive regularization and discretization error estimates. Moreover, we will consider control constraints and, in the end, state and control constraints, and redo the same steps. The theory will be complemented by several numerical examples.

State constraints

In this case we want to find the minimizer

$$y_\varrho \in K_s := \{z \in H_0^1(\Omega) : \, g_-(x) \le z(x) \le g_+(x) \text{ for a.a. } x \in \Omega\},$$

of (4.63)-(4.64), for given functions $g_\pm \in H_0^1(\Omega)$ for which we assume that $g_-(x) \le 0 \le g_+(x)$ for a.a. $x \in \Omega$ and additionally $\Delta g_\pm \in L^2(\Omega)$. Using the relation $\|u_\varrho\|_{H^{-1}(\Omega)} = \|\nabla y_\varrho\|_{L^2(\Omega)}$ we can consider the reduced cost functional

$$\widetilde{\mathcal{J}}(y_\varrho) = \frac{1}{2}\|y_\varrho - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|\nabla y_\varrho\|^2_{L^2(\Omega)}$$

and the problem is equivalent to find $y_\varrho \in K_s$ such that

$$\widetilde{\mathcal{J}}(y_\varrho) \le \widetilde{\mathcal{J}}(z) \quad \text{for all } z \in K_s,$$

which is exactly (3.47). Thus, by (3.48) the minimizer is characterized as the unique solution $y_\varrho \in K_s$ of the variational inequality

$$\varrho\langle \nabla y_\varrho, \nabla(z - y_\varrho)\rangle_{L^2(\Omega)} + \langle y_\varrho, z - y_\varrho\rangle_{L^2(\Omega)} \ge \langle y_d, z - y_\varrho\rangle_{L^2(\Omega)} \quad \text{for all } z \in K_s. \tag{4.65}$$

By Lemma 3.23 we get the following regularization error estimates.

LEMMA 4.16 ([50, cf Lemma 2.1]). *Let $y_d \in L^2(\Omega)$ be given. For the unique solution $y_\varrho \in K_s$ of* (4.65) *there holds*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}. \tag{4.66}$$

Further, if $y_d \in K_s$, then

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \sqrt{\varrho}\|\nabla y_d\|_{L^2(\Omega)} \quad \textit{and} \quad \|\nabla(y_\varrho - y_d)\|_{L^2(\Omega)} \le \|\nabla y_d\|_{L^2(\Omega)}. \tag{4.67}$$

If in addition $\Delta y_d \in L^2(\Omega)$ it holds

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \varrho\|\Delta y_d\|_{L^2(\Omega)} \quad \textit{and} \quad \|\nabla(y_\varrho - y_d)\|_{L^2(\Omega)} \le \sqrt{\varrho}\|\Delta y_d\|_{L^2(\Omega)}. \tag{4.68}$$

To get grip on the constraints in the numerical treatment later on, we need an indicator that specifies whether the computed state fulfills the constraints or not. This indicator should be easy to realize such that the conditions can be checked efficiently. Therefore, let us introduce the auxiliary variable $\lambda := -\varrho\Delta y_\varrho + y_\varrho - y_d \in H^{-1}(\Omega)$, which by (4.65), satisfies

$$\langle \lambda, z - y_\varrho \rangle_{L^2(\Omega)} \geq 0, \quad \text{for all } z \in K_s.$$

Note, that by Lemma 3.24 the unique solution $y_\varrho \in K_s$ satisfies $\Delta y_\varrho \in L^2(\Omega)$, which implies $\lambda \in L^2(\Omega)$ and pointwise a.e. evaluation is well-defined. Let us introduce the sets, where the constraints are fulfilled exactly, i.e.,

$$\Omega_{s,\pm} := \{x \in \Omega : \, y_\varrho(x) = g_\pm(x)\}.$$

Then, the complementarity conditions of (3.62) transfer as follows:

$$\begin{aligned} \lambda = 0, &\quad g_- < y_\varrho < g_+, &&\text{on } \Omega \setminus \Omega_{s,\pm}, \\ \lambda \geq 0, &\quad y_\varrho = g_-, &&\text{on } \Omega_{s,-}, \\ \lambda \leq 0, &\quad y_\varrho = g_+, &&\text{on } \Omega_{s,+}. \end{aligned} \tag{4.69}$$

Discretization

As in the case without constraints, we assume that $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a bounded and convex Lipschitz domain, which is polygonally ($d = 2$) or polyhedrally ($d = 3$) bounded and we consider the trial space $Y_h = S_h^1(\mathcal{T}_h) \cap H_0^1(\Omega)$ of globally continuous, piecewise linear functions defined on an admissible and shape regular decomposition $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$. In order to get a good discrete approximation of the set of state constraints K_s, let us consider

$$K_{sh} := \big\{ z_h \in Y_h : \, I_h g_-(x) \leq z_h(x) \leq I_h g_+(x), \quad \text{for all } x \in \Omega \big\},$$

where $I_h : \mathcal{C}(\overline{\Omega}) \to Y_h$ denotes the nodal interpolation operator, defined as

$$I_h v(x) := \sum_{k=1}^{M} v(x_k) \varphi_k^1(x), \quad x \in \Omega,$$

and $\{x_k\}_{k=1}^M$ denote the vertices of $\mathcal{T}_h$. Note, since Ω is convex and we assumed $\Delta g_\pm \in L^2(\Omega)$, we actually have that $g_\pm \in H^2(\Omega) \subset \mathcal{C}(\overline{\Omega})$, see, e.g., [30, 59] and $I_h g_\pm$ is well-defined. The discrete variational formulation is then to find $y_{\varrho h} \in K_{sh}$ such that

$$\varrho \langle \nabla y_{\varrho h}, \nabla (z_h - y_{\varrho h}) \rangle_{L^2(\Omega)} + \langle y_{\varrho h}, z_h - y_{\varrho h} \rangle_{L^2(\Omega)} \geq \langle y_d, z_h - y_{\varrho h} \rangle_{L^2(\Omega)} \quad \text{for all } z_h \in K_{sh}, \tag{4.70}$$

which is uniquely solvable by Theorem 2.7 and the following error estimates follow out directly from Theorem 3.31 in the abstract setting.

THEOREM 4.17. *Let $y_{\varrho h} \in K_{sh}$ denote the unique solution of* (4.70)*. If $y_d \in L^2(\Omega)$ then*

$$\|y_d - y_{\varrho h}\|_{L^2(\Omega)} \leq \|y_d\|_{L^2(\Omega)}. \tag{4.71}$$

If in addition $y_d \in K_s$ such that $\Delta y_d \in L^2(\Omega)$ then there holds

$$\begin{aligned} \|y_d - y_{\varrho h}\|_{L^2(\Omega)} \;&\leq\; c\Big(\varrho\|\Delta y_d\|_{L^2(\Omega)} + \varrho\|\Delta g_\pm\|_{L^2(\Omega)} \\ &\quad + \inf_{z_h \in K_{sh}} \big[\varrho\|\nabla(y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big) \end{aligned} \tag{4.72}$$

and

$$\begin{aligned} \sqrt{\varrho}\|\nabla(y_d - y_{\varrho h})\|_{L^2(\Omega)} \;&\leq\; c\Big(\varrho\|\Delta y_d\|_{L^2(\Omega)} + \varrho\|\Delta g_\pm\|_{L^2(\Omega)} \\ &\quad + \inf_{z_h \in K_{sh}} \big[\varrho\|\nabla(y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big). \end{aligned} \tag{4.73}$$

We can now state the main theorem of this section, revealing the optimal choice $\varrho = h^2$, as in the unconstrained case.

THEOREM 4.18 ([50, cf Corollary 3.3]). *Let $y_{\varrho h} \in K_{sh}$ denote the unique solution of* (4.70) *and let $y_d \in K_s \cap H^r(\Omega)$ for $r \in (1,2]$ or $y_d \in H^r_0(\Omega)$ for $r \in [0,1]$, where we additionally assume $g_-(x) \leq y_d(x) \leq g_+(x)$ for a.a. $x \in \Omega$. If $\varrho = h^2$ then*

$$\|y_d - y_{\varrho h}\|_{L^2(\Omega)} \leq ch^r(\|y_d\|_{H^r(\Omega)} + \|g_\pm\|_{H^r(\Omega)}).$$

Proof. Let us first assume that $y_d \in K_s \cap H^2(\Omega) \subset \mathcal{C}(\overline{\Omega})$. Then the nodal interpolation $I_h y_d \in Y_h$ is well defined and obviously $I_h y_d \in K_{sh}$. With the interpolation error estimates of Theorem 2.28 it holds

$$\inf_{z_h \in K_{sh}} \|y_d - z_h\|_{L^2(\Omega)} \leq \|y_d - I_h y_d\|_{L^2(\Omega)} \leq ch^2\|y_d\|_{H^2(\Omega)}$$

and

$$\inf_{z_h \in K_{sh}} \|\nabla(y_d - z_h)\|_{L^2(\Omega)} \leq \|\nabla(y_d - I_h y_d)\|_{L^2(\Omega)} \leq ch\|y_d\|_{H^2(\Omega)}.$$

Thus, with $\varrho = h^2$ estimate (4.72) becomes

$$\begin{aligned} \|y_d - y_{\varrho h}\|_{L^2(\Omega)} \;&\leq\; c\Big(h^2\|\Delta y_d\|_{L^2(\Omega)} + h^2\|\Delta g_\pm\|_{L^2(\Omega)} + \big[h^2h^2 + h^4\big]^{1/2}\|y_d\|_{H^2(\Omega)}\Big) \\ &\leq\; h^2\Big(\|y_d\|_{H^2(\Omega)} + \|g_\pm\|_{H^2(\Omega)}\Big). \end{aligned}$$

Interpolating this estimate with (4.71) gives the desired result. □

Numerical results

Using the fe-isomorphism the variational formulation (4.70) is equivalent to find $K_{sh} \ni y_{\varrho h} \leftrightarrow \mathbf{y}_{\varrho h} \in \mathbb{R}^M$ such that

$$\varrho(K_h \mathbf{y}_{\varrho h}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 + (M_h \mathbf{y}_{\varrho h}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 \geq (\mathbf{y}_{dh}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2, \tag{4.74}$$

for all $K_{sh} \ni z_h \leftrightarrow \mathbf{z}_h \in \mathbb{R}^M$, where the stiffness and mass matrices are given as

$$K_h[i,j] = \int_\Omega \nabla\varphi_j^1(x) \cdot \nabla\varphi_i^1(x)\, dx \quad \text{and} \quad M_h[i,j] = \int_\Omega \varphi_j^1(x)\varphi_i^1(x)\, dx, \quad i,j = 1,\ldots,M$$

and the load vector admits the entries

$$\mathbf{y}_{dh}[i] = \int_\Omega y_d(x)\varphi_i^1(x)\, dx, \quad i = 1,\ldots,M.$$

To incorporate the constraints, we define the auxiliary variable

$$\boldsymbol{\lambda}_h := \varrho K_h \mathbf{y}_{\varrho h} + M_h \mathbf{y}_{\varrho h} - \mathbf{y}_{dh},$$

as in the continuous case. Further, let the set of active nodes be defined as

$$\mathcal{A}_{s,\pm} := \{k = 1,\ldots,M : \mathbf{y}_{\varrho h}[k] = g_\pm(x_k)\}. \tag{4.75}$$

Then we conclude, analogously to the continuous case, the discrete complementarity conditions

$$\begin{aligned}
\boldsymbol{\lambda}_h[k] = 0, &\quad g_-(x_k) < \mathbf{y}_{\varrho h}[k] < g_+(x_k), &\quad \text{for } k \notin \mathcal{A}_{s,\pm}, \\
\boldsymbol{\lambda}_h[k] \geq 0, &\quad \mathbf{y}_{\varrho h}[k] = g_-(x_k), &\quad \text{for } k \in \mathcal{A}_{s,-}, \\
\boldsymbol{\lambda}_h[k] \leq 0, &\quad \mathbf{y}_{\varrho h}[k] = g_+(x_k), &\quad \text{for } k \in \mathcal{A}_{s,+}.
\end{aligned} \tag{4.76}$$

These are equivalent to

$$\boldsymbol{\lambda}_h[k] = \min\{0, \boldsymbol{\lambda}_h[k] + \alpha(\mathbf{g}_{+h}[k] - \mathbf{y}_{\varrho h}[k])\} + \max\{0, \boldsymbol{\lambda}_h[k] + \alpha(\mathbf{g}_{-h}[k] - \mathbf{y}_{\varrho h}[k])\},$$

for some $\alpha > 0$ and $Y_h \ni I_h g_\pm \leftrightarrow \mathbf{g}_{\pm h} \in \mathbb{R}^M$. We introduce the functions

$$F_1(\mathbf{y}_{\varrho h}, \boldsymbol{\lambda}_h) = \varrho K_h \mathbf{y}_{\varrho h} + M_h \mathbf{y}_{\varrho h} - \boldsymbol{\lambda}_h - \mathbf{y}_d$$

and

$$\begin{aligned}
F_2(\mathbf{y}_{\varrho h}, \boldsymbol{\lambda}_h) \;=\; & \boldsymbol{\lambda}_h - \min\{0, \boldsymbol{\lambda}_h[k] + \alpha(\mathbf{g}_{+h}[k] - \mathbf{y}_{\varrho h}[k])\} \\
& - \max\{0, \boldsymbol{\lambda}_h[k] + \alpha(\mathbf{g}_{-h}[k] - \mathbf{y}_{\varrho h}[k])\}.
\end{aligned}$$

Now, in order to compute a solution of (4.74), we want to find the roots of those functions simultaneously, i.e., we have to solve the system of (non-)linear equations

$$\mathbf{F}(\mathbf{y}_{\varrho h}, \boldsymbol{\lambda}_h) := \begin{pmatrix} F_1(\mathbf{y}_{\varrho h}, \boldsymbol{\lambda}_h) \\ F_2(\mathbf{y}_{\varrho h}, \boldsymbol{\lambda}_h) \end{pmatrix} = \begin{pmatrix} \mathbf{0}_h \\ \mathbf{0}_h \end{pmatrix}. \tag{4.77}$$

This can be done applying a semi-smooth Newton algorithm, see, e.g., [64]. Therefore, we successively compute the iterates

$$\begin{pmatrix} \mathbf{y}_{\varrho h}^{m+1} \\ \boldsymbol{\lambda}_h^{m+1} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_{\varrho h}^{m} \\ \boldsymbol{\lambda}_h^{m} \end{pmatrix} - \left(D\mathbf{F}(\mathbf{y}_{\varrho h}^m, \boldsymbol{\lambda}_h^m)\right)^{-1} \mathbf{F}(\mathbf{y}_{\varrho h}^m, \boldsymbol{\lambda}_h^m), \tag{4.78}$$

where the Jacobian is given as

$$D\mathbf{F}(\mathbf{v}_h, \boldsymbol{\mu}_h) = \begin{pmatrix} \varrho K_h + M_h & -I \\ \alpha L'(\mathbf{g}_{\pm h}, \mathbf{v}_h, \boldsymbol{\mu}_h) & I - L'(\mathbf{g}_{\pm h}, \mathbf{v}_h, \boldsymbol{\mu}_h) \end{pmatrix}. \tag{4.79}$$

The diagonal entries of the diagonal matrices

$$L'(\mathbf{g}_{\pm h}, \mathbf{v}_h, \boldsymbol{\mu}_h) := L'_{\min}(\mathbf{g}_{+h}, \mathbf{v}_h, \boldsymbol{\mu}_h) + L'_{\max}(\mathbf{g}_{-h}, \mathbf{v}_h, \boldsymbol{\mu}_h),$$

are given as

$$\begin{aligned} L'_{min}(\mathbf{g}_{+h}, \mathbf{v}_h, \boldsymbol{\mu}_h) &= \operatorname{diag}\Big(\ell'_{min}\big(\boldsymbol{\mu}_h[k] + \alpha\,[\mathbf{g}_{+h}[k] - \mathbf{v}_h[k]]\big)\Big), \\ L'_{max}(\mathbf{g}_{-h}, \mathbf{v}_h, \boldsymbol{\mu}_h) &= \operatorname{diag}\Big(\ell'_{max}\big(\boldsymbol{\mu}_h[k] + \alpha\,[\mathbf{g}_{-h}[k] - \mathbf{v}_h[k]]\big)\Big), \end{aligned}$$

with the slant derivatives of the functions $\ell_{\min}(z) = \min\{0, z\}$ and $\ell_{\max}(z) = \max\{0, z\}$ defined by

$$\ell'_{\min}(z) = \begin{cases} 1, & z < 0, \\ 0, & z \geq 0, \end{cases} \quad \text{and} \quad \ell'_{\max}(z) = \begin{cases} 0, & z \leq 0, \\ 1, & z > 0. \end{cases}$$

Rewriting the system (4.78) gives

$$D\mathbf{F}(\mathbf{y}_{\varrho h}^m, \boldsymbol{\lambda}_h^m) \begin{pmatrix} \mathbf{y}_{\varrho h}^m - \mathbf{y}_{\varrho h}^{m+1} \\ \boldsymbol{\lambda}_h^m - \boldsymbol{\lambda}_h^{m+1} \end{pmatrix} = \mathbf{F}(\mathbf{y}_{\varrho h}^m, \boldsymbol{\mu}_h^m). \tag{4.80}$$

And with the definition of the Jacobian from the first line we get

$$(\varrho K_h + M_h)(\mathbf{y}_{\varrho h}^m - \mathbf{y}_{\varrho h}^{m+1}) - \boldsymbol{\lambda}_h^m + \boldsymbol{\lambda}_h^{m+1} = (\varrho K_h + M_h)\mathbf{y}_{\varrho h}^m - \boldsymbol{\lambda}_h^m - \mathbf{y}_{dh},$$

from which we conclude

$$(\varrho K_h + M_h)\mathbf{y}_{\varrho h}^{m+1} - \boldsymbol{\lambda}_h^{m+1} = \mathbf{y}_{dh}.$$

With

$$z^m_{+,k} := \boldsymbol{\lambda}_h[k]^m + \alpha\,[\mathbf{g}_{+h} - \mathbf{y}^m_{\varrho h}[k]] \quad \text{and} \quad z^m_{-,k} := \boldsymbol{\lambda}_h[k]^m + \alpha\,[\mathbf{g}_{-h} - \mathbf{y}^m_{\varrho h}[k]],$$

the second line reads, componentwise,

$$\begin{aligned}\alpha[\ell'_{\min}(z^m_{+,k}) + \ell'_{\max}(z^m_{-,k})](\mathbf{y}^m_{\varrho h}[k] - \mathbf{y}^{m+1}_{\varrho h}[k]) + \boldsymbol{\lambda}^m_h[k] - \boldsymbol{\lambda}^{m+1}_h[k]\\ -[\ell'_{\min}(z^m_{+,k})) + \ell'_{\max}(z^m_{-,k})](\boldsymbol{\lambda}^m_h[k] - \boldsymbol{\lambda}^{m+1}_h[k])\\ = \boldsymbol{\lambda}^m_h[k] - \min\{0, z^m_{+,k}\} - \max\{0, z^m_{-,k}\}.\end{aligned} \tag{4.81}$$

We distinguish the following three cases.

(i) $z^m_{+,k} \geq 0$ and $z^m_{-,k} \leq 0$: Then, $\ell'_{min}(z^m_{+,k}) = \ell'_{max}(z^m_{-,k}) = 0$, and from (4.81) we compute

$$\boldsymbol{\lambda}^{m+1}_h[k] = 0.$$

(ii) $z^m_{-,k} > 0$: From this we get $\boldsymbol{\lambda}^m_h[k] > \alpha\,[\mathbf{y}^m_{\varrho h}[k] - g_-(x_k)]$ and we compute

$$\begin{aligned}\boldsymbol{\lambda}^m_h[k] + \alpha\,[g_+(x_k) - \mathbf{y}^m_{\varrho h}[k]] &> \alpha\,[\mathbf{y}^m_{\varrho h}[k] - g_-(x_k) + g_+(x_k) - \mathbf{y}^m_{\varrho h}[k]]\\ &= \alpha\,[g_+(x_k) - g_-(x_k)] > 0,\end{aligned}$$

i.e., $z^m_{+,k} > 0$. Therefore, $\ell'_{\min}(z^m_{+,k}) = 0$ and $\ell'_{\max}(z^m_{-,k}) = 1$, and we get from (4.81) that

$$\mathbf{y}^{m+1}_{\varrho h}[k] = g_-(x_k).$$

(iii) $z^m_{+,k} < 0$: Then, as in the second case, we compute $z^m_{-,k} < 0$ to get $\ell'_{\min}(z^m_{+,k}) = 1$, $\ell'_{\max}(z^m_{-,k}) = 0$, and thus (4.81) becomes

$$\mathbf{y}^{m+1}_{\varrho h}[k] = g_+(x_k).$$

Therefore we see, that the iterates of the semi-smooth Newton method (4.78) fulfill the active set strategy as given in Algorithm 1.

To support our theoretical results, we will consider the domain $\Omega = (0,1)^2$ and the target functions $y_{cd,i} \in \mathcal{C}^\infty(\Omega) \cap H^1_0(\Omega)$, $i = 1, 2$, given as

$$y_{cd,1}(x, y) = \sin(\pi x)\sin(\pi y),$$

and

$$y_{cd,2}(x, y) = H_k(x)H_k(y), \text{ where } H_k(s) = \frac{1}{1 + e^{-k(s-0.25)}} - \frac{1}{1 + e^{-k(s-0.75)}},$$

Algorithm 1 Active set algorithm [64]

Require: Initial values $\mathbf{y}_h^0, \boldsymbol{\lambda}_h^0$
(a) $m = 0$
(b) Set

$$z_{+,k}^m = \boldsymbol{\lambda}_h^m[k] + \alpha\,[g_+(x_k) - \mathbf{y}_h^m[k]] \quad \text{and} \quad z_{-,k}^m = \boldsymbol{\lambda}_h^m[k] + \alpha\,[g_-(x_k) - \mathbf{y}_h^m[k]]$$

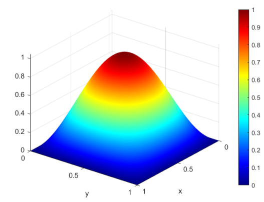
while stop criterion is not fulfilled **do**
(i) Set

$$\mathcal{I}^m = \{k:\, z_{+,k}^m \geq 0,\, z_{-,k}^m \leq 0\}, \quad \mathcal{A}_-^m = \{k:\, z_{-,k}^m > 0\}, \quad \mathcal{A}_+^m = \{k:\, z_{+,k}^m < 0\}$$

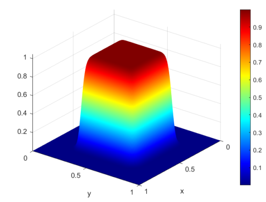
(ii) Solve

$$\begin{aligned}&(\varrho K_h + M_h)\mathbf{y}_h^{m+1} - \boldsymbol{\lambda}^{m+1} = \mathbf{y}_{dh},\\ &\mathbf{y}_h^{m+1}[k] = g_\pm(x_k),\ k \in \mathcal{A}_\pm^m,\\ &\boldsymbol{\lambda}_h^{m+1}[k] = 0,\ k \in \mathcal{I}^m.\end{aligned}$$

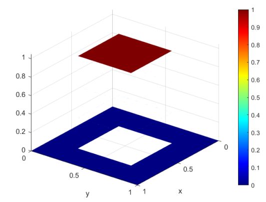
(iii) $m = m + 1$
end while



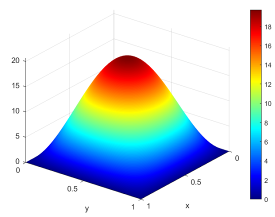
a) $y_{cd,1}$



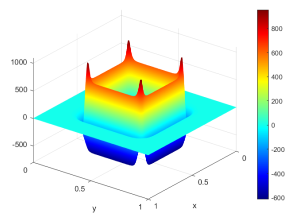
b) $y_{cd,2}$



c) $y_{cd,3}$



d) $u_1 = -\Delta y_{cd,1}$



e) $u_2 = -\Delta y_{cd,2}$

Figure 4.8: Target functions and controls $u_j = -\Delta y_{cd,j}$, $j = 1, 2$.

for $k = 40$, see Figure 4.8, for which we can compute $u_i = -\Delta y_{cd,i}$ analytically. We also consider the discontinuous target $y_{cd,3} := \lim_{k\to\infty} y_{cd,2} \in H^{1/2-\varepsilon}(\Omega)$, $\varepsilon > 0$, for which we *cannot* compute the control analytically, given as

$$y_{cd,3}(x,y) = \begin{cases} 1, & (x,y) \in (0.25, 0.75)^2, \\ 0, & \text{else.} \end{cases}$$

We consider the upper and lower constraints $g_\pm$ given by

$$g_-(x) \equiv 0, \; g_+(x) = 0.5 \cdot y_{cd,1}(x),$$

which are incorporated by solving the system (4.78) successively, with $\varrho = h^2$ and initial guess

$$\mathbf{y}^0_{\varrho h} = (h^2 K_h + M_h)^{-1}\mathbf{y}_{cd,h} \in \mathbb{R}^M \quad \text{and} \quad \boldsymbol{\lambda}^0_h = \mathbf{0}_h.$$

As a stopping criterion we choose the maximal absolute error in each node, i.e., we stop if

$$\mathrm{tol}_s := \max\{\mathrm{tol}_{s,+}, \mathrm{tol}_{s,-}\} < 10^{-5}, \tag{4.82}$$

where

$$\begin{aligned} \mathrm{tol}_{s,+} &:= \max_{\{k:\mathbf{y}_{\varrho h}[k] > g_+(x_k)\}} |\mathbf{y}_{\varrho h}[k] - g_+(x_k)|, \\ \mathrm{tol}_{s,-} &:= \max_{\{k:\mathbf{y}_{\varrho h}[k] < g_-(x_k)\}} |\mathbf{y}_{\varrho h}[k] - g_-(x_k)|. \end{aligned}$$

To reconstruct the control, we apply the method discussed in Section 4.1.1, i.e., we solve (4.50). The results are depicted in Figure 4.9.

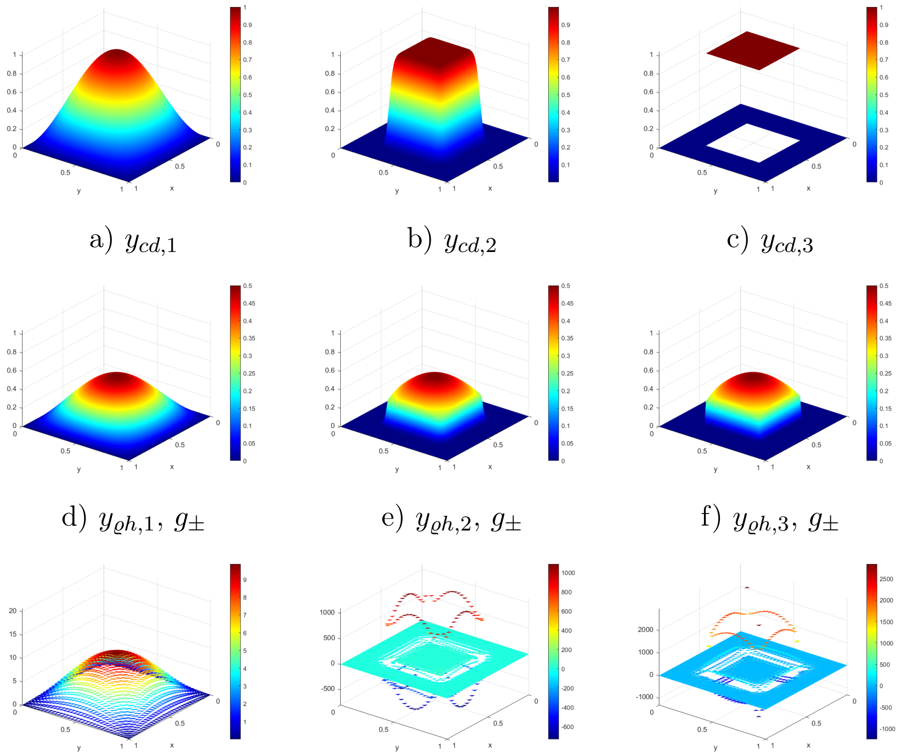
Control constraints

In this case we want to find the minimizer $y_\varrho \in H^1_0(\Omega)$ of (4.63)-(4.64) such that $u_\varrho = -\Delta y_\varrho \in U_c$, with

$$U_c := \big\{ v \in H^{-1}(\Omega) : \langle h_-, q\rangle_{L^2(\Omega)} \le \langle v, q\rangle_\Omega \le \langle h_+, q\rangle_{L^2(\Omega)} \, \forall q \in H^1_0(\Omega),\, q \ge 0 \big\},$$

where $h_\pm \in L^2(\Omega)$ are given, and we assume $h_-(x) \le 0 \le h_+(x)$ for a.a. $x \in \Omega$. Let us define $B := -\Delta : H^1_0(\Omega) \to H^{-1}(\Omega)$ and recall that B defines an isomorphism, see Lemma 4.1. Then, with $y_\varrho = B^{-1}u_\varrho$, we can introduce the reduced cost functional

$$\hat{\mathcal{J}}(u_\varrho) = \frac{1}{2}\|B^{-1}u_\varrho - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_{H^{-1}(\Omega)}$$



a) $y_{cd,1}$ b) $y_{cd,2}$ c) $y_{cd,3}$

d) $y_{\varrho h,1}$, $g_\pm$ e) $y_{\varrho h,2}$, $g_\pm$ f) $y_{\varrho h,3}$, $g_\pm$

g) $u_{\varrho H,1}$, $g_\pm$ h) $u_{\varrho H,2}$, $g_\pm$ i) $u_{\varrho H,3}$, $g_\pm$

Figure 4.9: Targets $y_{cd,i}$, computed constrained states $y_{\varrho h,i}$, $= 1, 2, 3$ on a mesh with $N = 32768$ elements and $M = 16129$ DoFs with constraints $g_\pm$ and reconstruction of the controls $u_{\varrho H,i}$ on a mesh with $N_H = 2048$ elements.

and the problem is to find the minimizer $u_\varrho \in U_c$ that fulfills

$$\hat{\mathcal{J}}(u_\varrho) \leq \hat{\mathcal{J}}(v) \quad \text{for all } v \in U_c.$$

This is exactly (3.50) and thus the minimizer is characterized as unique solution $u_\varrho \in U_c$ of the variational inequality

$$\langle (B^{-1})^*(B^{-1}u_\varrho - y_d), v - u_\varrho\rangle_\Omega + \varrho\langle B^{-1}u_\varrho, v - u_\varrho\rangle_\Omega \geq 0 \quad \text{for all } v \in U_c.$$

Using the property that B is an isomorphism again, we can as well consider the set $K_c = B^{-1}(U_c)$ and with the relations $y_\varrho = B^{-1}u_\varrho \in K_c$ and $z = B^{-1}v \in K_c$, the problem now becomes to find $y_\varrho \in K_c$ such that

$$\langle y_\varrho - y_d, z - y_\varrho\rangle_{L^2(\Omega)} + \varrho\langle \nabla y_\varrho, \nabla(z - y_\varrho)\rangle_{L^2(\Omega)} \geq 0 \quad \text{for all } z \in K_c. \tag{4.83}$$

Now we observe that (4.83) is exactly (4.65), but with a different set of constraints, i.e., K_c instead of K_s. So, the regularization error estimates of Lemma 4.16 remain valid, when replacing K_s by K_c at every occurence.

As in the case of state constraints, we need to derive complementarity conditions, to get grip on the constraints in the numerical treatment later on. Therefore, we introduce the auxiliary variable $w_\lambda \in H_0^1(\Omega)$ as unique solution of

$$-\Delta w_\lambda = \lambda = -\varrho\Delta y_\varrho + y_\varrho - y_d \text{ in } \Omega, \quad w_\lambda = 0 \text{ on } \partial\Omega.$$

In variational form this reads to find $w_\lambda \in H_0^1(\Omega)$ such that

$$\langle \nabla w_\lambda, \nabla z\rangle_{L^2(\Omega)} = \varrho\langle \nabla y_\varrho, \nabla z\rangle_{L^2(\Omega)} + \langle y_\varrho - y_d, z\rangle_{L^2(\Omega)} \quad \text{for all } z \in H_0^1(\Omega)$$

and in view of the variational inequality (4.83), we see that

$$0 \leq \langle \lambda, z - y_\varrho\rangle_\Omega = \langle -\Delta w_\lambda, z - y_\varrho\rangle_\Omega = \langle \nabla w_\lambda, \nabla(z - y_\varrho)\rangle_{L^2(\Omega)} \quad \text{for all } z \in K_c.$$

If $y_d \in L^2(\Omega)$, by Lemma 3.24 we have that $-\Delta w_\lambda = \lambda \in L^2(\Omega)$ and in particular $u_\varrho = -\Delta y_\varrho \in L^2(\Omega)$. Therefore, we can consider the sets

$$\Omega_{c,\pm} := \{x \in \Omega : \, u_\varrho(x) = h_\pm(x)\},$$

where the constraints are fulfilled with equality and by (3.64) we conclude complementarity conditions

$$\begin{aligned} w_\lambda &= 0, & h_- < u_\varrho &< h_+, & &\text{on } \Omega \setminus \Omega_{c,\pm}, \\ w_\lambda &\geq 0, & u_\varrho &= h_-, & &\text{on } \Omega_{c,-}, \\ w_\lambda &\leq 0, & u_\varrho &= h_+, & &\text{on } \Omega_{c,+}. \end{aligned} \tag{4.84}$$

Discretization

Since the continuous variational inequalities for state constraints (4.65) and control constraints (4.83) are equal, up to the set of constraints, the only difference in the discretization of control constraints is the definition of a suitable set $K_{ch} \subset Y_h = S_h^1(\mathcal{T}_h) \cap H_0^1(\Omega)$, that yields a good approximation of the set K_c. We therefore consider the set

$$K_{ch} := \big\{ z_h \in Y_h : \langle h_-, q_h \rangle_{L^2(\Omega)} \le \langle \nabla z_h, \nabla q_h \rangle_{L^2(\Omega)} \le \langle h_+, q_h \rangle_{L^2(\Omega)}, \, \forall q_h \in Y_h, \, q_h \ge 0 \big\}.$$

As in the case of state constraints, the discrete variational formulation is to find $y_{\varrho h} \in K_{ch}$ such that

$$\varrho \langle \nabla y_{\varrho h}, \nabla (z_h - y_{\varrho h}) \rangle_{L^2(\Omega)} + \langle y_{\varrho h}, z_h - y_{\varrho h} \rangle_{L^2(\Omega)} \ge \langle y_d, z_h - y_{\varrho h} \rangle_{L^2(\Omega)} \quad \text{for all } z_h \in K_{ch}, \tag{4.85}$$

which is uniquely solvable by Theorem 2.7. Error estimates follow from Theorem 3.31 and Remark 3.28.

THEOREM 4.19. *Let $y_{\varrho h} \in K_{ch}$ denote the unique solution of* (4.85). *If $y_d \in L^2(\Omega)$ then*

$$\|y_d - y_{\varrho h}\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}. \tag{4.86}$$

If in addition $y_d \in K_c$ such that $\Delta y_d \in L^2(\Omega)$ then there holds

$$\begin{aligned} \|y_d - y_{\varrho h}\|_{L^2(\Omega)} \;\le\; & c\Big(\varrho \|\Delta y_d\|_{L^2(\Omega)} + \varrho \|h_\pm\|_{L^2(\Omega)} \\ & + \inf_{z_h \in K_{ch}} \big[\varrho \|\nabla (y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)} \big]^{1/2} \Big) \end{aligned} \tag{4.87}$$

and

$$\begin{aligned} \sqrt{\varrho} \|\nabla (y_d - y_{\varrho h})\|_{L^2(\Omega)} \;\le\; & c\Big(\varrho \|\Delta y_d\|_{L^2(\Omega)} + \varrho \|h_\pm\|_{L^2(\Omega)} \\ & + \inf_{z_h \in K_{ch}} \big[\varrho \|\nabla (y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)} \big]^{1/2} \Big). \end{aligned} \tag{4.88}$$

The main theorem of this section is again an error estimate linking the regularity if the target and the regularization parameter, which gives the optimal choice $\varrho = h^2$.

THEOREM 4.20 ([50, cf Theorem 3.4]). *Let $y_{\varrho h} \in K_{ch}$ denote the unique solution of* (4.85) *and let $y_d \in K_c \cap H^2(\Omega)$. If $\varrho = h^2$ then*

$$\|y_d - y_{\varrho h}\|_{L^2(\Omega)} + h \|\nabla (y_d - y_{\varrho h})\|_{L^2(\Omega)} \le c h^2 (\|y_d\|_{H^2(\Omega)} + \|h_\pm\|_{L^2(\Omega)}).$$

Proof. Let $y_d \in K_c \cap H^2(\Omega)$ and consider the H^1-projection $P_h^1 : H_0^1(\Omega) \to Y_h$ defined as

$$\langle \nabla(P_h^1 y), \nabla q_h \rangle_{L^2(\Omega)} = \langle \nabla y, \nabla q_h \rangle_{L^2(\Omega)} \quad \text{for all } q_h \in Y_h.$$

In particular, considering $P_h^1 y_d \in Y_h$, we have that

$$\langle \nabla y_d, \nabla q_h \rangle_{L^2(\Omega)} = \langle \nabla(P_h^1 y_d), \nabla q_h \rangle_{L^2(\Omega)} \quad \text{for all } q_h \in Y_h,\ q_h \geq 0,$$

from which $P_h^1 y_d \in K_{ch}$ follows. Thus, by the approximation property, see Theorem 2.36, we first have that

$$\|\nabla(y_d - P_h^1 y_d)\|_{L^2(\Omega)} \leq ch\|y_d\|_{H^2(\Omega)}.$$

To get an error estimate for the $L^2(\Omega)$-norm, let us introduce the auxiliary problem to find $e \in H_0^1(\Omega)$ such that

$$\langle \nabla e, \nabla v \rangle_{L^2(\Omega)} = \langle y_d - P_h^1 y_d, v \rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

By the assumption that Ω is convex, using elliptic regularity, see, e.g., [30, 59], we have $e \in H^2(\Omega)$ and

$$\|e\|_{H^2(\Omega)} \leq c\|\Delta e\|_{L^2(\Omega)} = c\|y_d - P_h^1 y_d\|_{L^2(\Omega)}.$$

Using the Galerkin orthogonality $\langle \nabla(y_d - P_h^1 y_d), \nabla q_h \rangle_{L^2(\Omega)} = 0$ for all $q_h \in Y_h$, we can now estimate

$$\begin{aligned}
\|y_d - P_h^1 y_d\|_{L^2(\Omega)}^2 &= \langle y_d - P_h^1 y_d, y_d - P_h^1 y_d \rangle_{L^2(\Omega)} \\
&= \langle \nabla e, \nabla(y_d - P_h^1 y_d) \rangle_{L^2(\Omega)} \\
&= \langle \nabla(e - P_h^1 e), \nabla(y_d - P_h^1 y_d) \rangle_{L^2(\Omega)} \\
&\leq \|\nabla(e - P_h^1 e)\|_{L^2(\Omega)} \|\nabla(y_d - P_h^1 y_d)\|_{L^2(\Omega)} \\
&\leq ch\|e\|_{H^2(\Omega)} ch\|y_d\|_{H^2(\Omega)} \\
&\leq ch^2\|y_d - P_h^1 y_d\|_{L^2(\Omega)} \|y_d\|_{H^2(\Omega)},
\end{aligned}$$

i.e., $\|y_d - P_h^1 y_d\|_{L^2(\Omega)} \leq ch^2\|y_d\|_{H^2(\Omega)}$. Now, with $\varrho = h^2$ we get

$$\begin{aligned}
&\inf_{z_h \in K_{ch}} [h^2\|\nabla(y_d - z_h)\|_{L^2(\Omega)}^2 + \|y_d - z_h\|_{L^2(\Omega)}^2] \\
&\qquad \leq h^2\|\nabla(y_d - P_h^1 y_d)\|_{L^2(\Omega)}^2 + \|y_d - P_h^1 y_d\|_{L^2(\Omega)}^2 \\
&\qquad \leq ch^4\|y_d\|_{H^2(\Omega)}
\end{aligned}$$

and estimate (4.87) becomes

$$\begin{aligned}
\|y_d - y_{\varrho h}\|_{L^2(\Omega)} &\leq c\big(h^2\|\Delta y_d\|_{L^2(\Omega)} + h^2\|h_\pm\|_{L^2(\Omega)} + h^2\|y_d\|_{H^2(\Omega)}\big) \\
&\leq ch^2\big(\|y_d\|_{H^2(\Omega)} + \|h_\pm\|_{L^2(\Omega)}\big),
\end{aligned}$$

whereas in the same fashion the estimate (4.88) becomes

$$h\|\nabla(y_d - y_{\varrho h})\|_{L^2(\Omega)} \leq ch^2\big(\|y_d\|_{H^2(\Omega)} + \|h_\pm\|_{L^2(\Omega)}\big). \qquad \square$$

Numerical results

As in the case of state constraints, using the fe-isomorphism, the solution of (4.85) $K_{ch} \ni y_{\varrho h} \leftrightarrow \mathbf{y}_{\varrho h} \in \mathbb{R}^M$ has to fulfill

$$\varrho(K_h \mathbf{y}_{\varrho h}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 + (M_h \mathbf{y}_{\varrho h}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 \geq (\mathbf{y}_{dh}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2, \tag{4.89}$$

for all $K_{ch} \ni z_h \leftrightarrow \mathbf{z}_h \in \mathbb{R}^M$. To incorporate the constraints, we will consider the auxiliary variable $\mathbf{w}_{\lambda,h} \in \mathbb{R}^M$ solving

$$K_h \mathbf{w}_{\lambda,h} = \varrho K_h \mathbf{y}_{\varrho h} + M_h \mathbf{y}_{\varrho h} - \mathbf{y}_{dh}.$$

Introducing the set of active nodes as

$$\mathcal{A}_{c,\pm} := \{k = 1, \dots, M : (K_h \mathbf{y}_{\varrho h})[k] = \mathbf{h}_{\pm h}[k]\},$$

where the entries of $\mathbf{h}_{\pm h} \in \mathbb{R}^M$ are given as

$$\mathbf{h}_{\pm}[k] = \int_\Omega h_{\pm}(x) \varphi_k^1(x)\, dx,\ k = 1, \dots, M,$$

we can conclude the discrete complementarity conditions

$$\begin{aligned} &\mathbf{w}_{\lambda h}[k] = 0, \quad \mathbf{h}_{-h}[k] < (K_h \mathbf{y}_{\varrho h})[k] < \mathbf{h}_{+h}[k], \quad \text{for } k \notin \mathcal{A}_{c,\pm}, \\ &\mathbf{w}_{\lambda h}[k] \geq 0, \qquad\qquad\quad (K_h \mathbf{y}_{\varrho h})[k] = \mathbf{h}_{-h}[k], \quad \text{for } k \in \mathcal{A}_{c,-}, \\ &\mathbf{w}_{\lambda h}[k] \leq 0, \qquad\qquad\quad (K_h \mathbf{y}_{\varrho h})[k] = \mathbf{h}_{+h}[k], \quad \text{for } k \in \mathcal{A}_{c,+}. \end{aligned} \tag{4.90}$$

as in the continuous case. These are equivalent to

$$\mathbf{w}_{\lambda h}[k] = \min\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h+} - (K_h \mathbf{y}_{\varrho h})[k])\} + \max\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h-} - (K_h \mathbf{y}_{\varrho h})[k])\},$$

for some $\alpha > 0$. Thus, we want to find the roots of the functions

$$\tilde{\mathbf{F}}_1(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) = \varrho K_h \mathbf{y}_{\varrho h} + M_h \mathbf{y}_{\varrho h} - K_h \mathbf{w}_{\lambda h} - \mathbf{y}_{dh}$$

and

$$\begin{aligned} \tilde{\mathbf{F}}_2(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) = \mathbf{w}_{\lambda h} \quad &- \quad \min\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h+}[k] - (K_h \mathbf{y}_{\varrho h})[k])\} \\ &- \quad \max\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h-}[k] - (K_h \mathbf{y}_{\varrho h})[k])\} \end{aligned}$$

simultaneously. As in the case of state constraints, this can be achieved by applying a semi-smooth Newton method, i.e., we compute the iterates

$$\begin{pmatrix} \mathbf{y}_{\varrho h}^{m+1} \\ \mathbf{w}_{\lambda h}^{m+1} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_{\varrho h}^{m} \\ \mathbf{w}_{\lambda h}^{m} \end{pmatrix} - \left(D\tilde{\mathbf{F}}(\mathbf{y}_{\varrho h}^m, \mathbf{w}_{\lambda h}^m)\right)^{-1} \tilde{\mathbf{F}}(\mathbf{y}_{\varrho h}^m, \mathbf{w}_{\lambda h}^m),$$

where the Jacobian of

$$\tilde{\mathbf{F}}(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) = \begin{pmatrix} \tilde{\mathbf{F}}_1(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) \\ \tilde{\mathbf{F}}_2(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) \end{pmatrix}$$

is given as

$$D\tilde{\mathbf{F}}(\mathbf{v}_h, \mathbf{w}_h) = \begin{pmatrix} \varrho K_h + M_h & -K_h \\ \alpha L'(\mathbf{h}_{\pm h}, K_h \mathbf{v}_h, \mathbf{w}_h) K_h & I - L'(\mathbf{h}_{\pm h}, K_h \mathbf{v}_h, \mathbf{w}_h) \end{pmatrix}.$$

We consider the target functions $y_{cd,i}$, $i = 1, 2, 3$, again, but now the constraints on the control $h_{\pm}^{(j)}$ given by

$$h_-^{(1)}(x,y) \equiv 0, \quad \text{and} \quad h_+^{(1)}(x,y) = \min\{\max\{u_1(x,y), 0\}, 10\}$$

and

$$h_-^{(2)}(x,y) \equiv 0 \quad \text{and} \quad h_+^{(2)}(x,y) = 1000 \cdot y_{cd,2}(x,y).$$

We choose $\varrho = h^2$ and the initial guesses

$$\mathbf{y}_{\varrho h}^0 = (h^2 K_h + M_h)^{-1} \mathbf{y}_{cd,h} \in \mathbb{R}^M \quad \text{and} \quad \mathbf{w}_{\lambda h}^0 = \mathbf{0}_h.$$

As a stopping criterion we choose the maximal absolute error in each node, i.e., we stop if

$$\mathrm{tol}_c := \max\{\mathrm{tol}_{c,+}, \mathrm{tol}_{c,-}\} < 10^{-5}, \tag{4.91}$$

where

$$\begin{aligned} \mathrm{tol}_{c,+} &:= \max_{\{k:(K_h \mathbf{y}_{\varrho h})[k] > \mathbf{h}_{+h}[k]\}} |(K_h \mathbf{y}_{\varrho h})[k] - \mathbf{h}_{+h}[k]|, \\ \mathrm{tol}_{c,-} &:= \max_{\{k:(K_h \mathbf{y}_{\varrho h})[k] < \mathbf{h}_{-h}[k]\}} |(K_h \mathbf{y}_{\varrho h})[k] - \mathbf{h}_{-h}[k]|. \end{aligned}$$

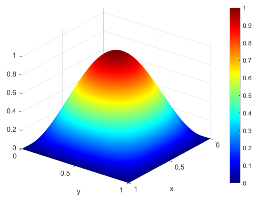
To reconstruct the control, we apply the method discussed in Section 4.1.1, i.e., we solve (4.50). The results are depicted in Figure 4.10.

State and control constraints

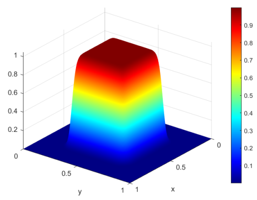
We saw that on the continuous level, the variational inequalities (4.65) and (4.83) to incorporate constraints, had the form to find $y_\varrho \in K$ such that

$$\varrho \langle \nabla y_\varrho, \nabla(z - y_\varrho) \rangle_{L^2(\Omega)} + \langle y_\varrho, z - y_\varrho \rangle_{L^2(\Omega)} \geq \langle y_d, z - y_\varrho \rangle \quad \text{for all } z \in K,$$

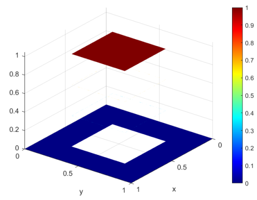
where $K \subset H_0^1(\Omega)$ is a convex set, i.e., $K = K_s$ in the case of state constraints and $K = K_c$ in the case of control constraints. Thus, at this point it is easy to



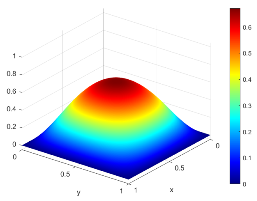
a) $y_{cd,1}$



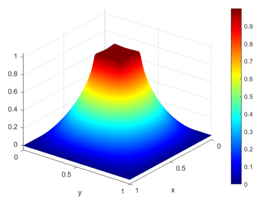
b) $y_{cd,2}$



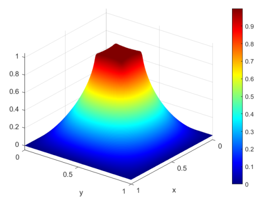
c) $y_{cd,3}$



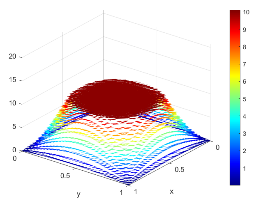
d) $y_{\varrho h,1}$, $h_\pm^{(1)}$



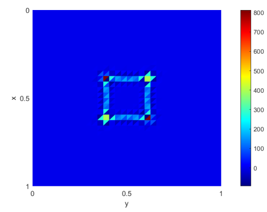
e) $y_{\varrho h,2}$, $h_\pm^{(2)}$



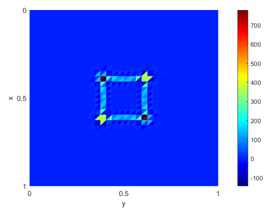
f) $y_{\varrho h,3}$, $h_\pm^{(2)}$



g) $u_{\varrho H,1}$, $h_\pm^{(1)}$



h) $u_{\varrho H,2}$, $h_\pm^{(2)}$



i) $u_{\varrho H,3}$, $h_\pm^{(2)}$

Figure 4.10: Targets $y_{cd,i}$, computed states $y_{\varrho h,i}$, $= 1, 2, 3$ on a mesh with $N = 32768$ elements and $M = 16129$ DoFs with constraints $h_\pm^{(j)}$ and reconstruction of the controls $u_{\varrho H,i}$ on a mesh with $N_H = 2048$ elements.

incorporate both constraints at once, by choosing $K = K_s \cap K_c$, which is again convex as an intersection of convex sets. Furthermore, the regularization error estimates of Lemma 4.16 stay valid, when replacing K_s by $K_s \cap K_c$. For a discretization, we will consider the discretized version, to find $y_{\varrho h} \in K_h := K_{sh} \cap K_{ch}$ such that

$$\varrho\langle \nabla y_{\varrho h}, \nabla(z_h - y_{\varrho h})\rangle_{L^2(\Omega)} + \langle y_{\varrho h}, z_h - y_{\varrho h}\rangle_{L^2(\Omega)} \geq \langle y_d, z_h - y_{\varrho h}\rangle \quad \text{for all } z_h \in K_h. \tag{4.92}$$

Note, that the discretization error estimates of Theorem 3.31 transfer to the following.

THEOREM 4.21. *Let $y_{\varrho h} \in K_h$ denote the unique solution of (4.92). If $y_d \in L^2(\Omega)$ then*

$$\|y_d - y_{\varrho h}\|_{L^2(\Omega)} \leq \|y_d\|_{L^2(\Omega)}. \tag{4.93}$$

If in addition $y_d \in K = K_s \cap K_c$ such that $\Delta y_d \in L^2(\Omega)$ then there holds

$$\begin{aligned} \|y_d - y_{\varrho h}\|_{L^2(\Omega)} \;\leq\; & c\Big(\varrho\|\Delta y_d\|_{L^2(\Omega)} + \varrho\|\Delta g_\pm\|_{L^2(\Omega)} + \varrho\|h_\pm\|_{L^2(\Omega)} \\ & + \inf_{z_h \in K_h} \left[\varrho\|\nabla(y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\right]^{1/2}\Big) \end{aligned} \tag{4.94}$$

and

$$\begin{aligned} \sqrt{\varrho}\|\nabla(y_d - y_{\varrho h})\|_{L^2(\Omega)} \;\leq\; & c\Big(\varrho\|\Delta y_d\|_{L^2(\Omega)} + \varrho\|\Delta g_\pm\|_{L^2(\Omega)} + \varrho\|h_\pm\|_{L^2(\Omega)} \\ & + \inf_{z_h \in K_h} \left[\varrho\|\nabla(y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\right]^{1/2}\Big). \end{aligned} \tag{4.95}$$

Using the fe-isomorphism, (4.92) is again equivalent to the system to find $K_h \ni y_{\varrho h} \leftrightarrow \mathbf{y}_{\varrho h} \in \mathbb{R}^M$ solving

$$\varrho(K_h\mathbf{y}_{\varrho h}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 + (M_h\mathbf{y}_{\varrho h}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 \geq (\mathbf{y}_{dh}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 \tag{4.96}$$

for all $K_h \ni z_h \leftrightarrow \mathbf{z}_h \in \mathbb{R}^M$. Firstly, to incorporate the state constraints, we additionally add the Lagrange multiplier $\boldsymbol{\lambda}_h \in \mathbb{R}^M$ solving

$$\boldsymbol{\lambda}_h = \varrho K_h\mathbf{y}_{\varrho h} + M_h\mathbf{y}_{\varrho h} - \mathbf{y}_{dh} \tag{4.97}$$

and fulfilling the complementarity conditions (4.76), i.e,

$$\boldsymbol{\lambda}_h[k] = \min\{0, \boldsymbol{\lambda}_h[k] + \alpha(g_+(x_k) - \mathbf{y}_{\varrho h}[k])\} + \max\{0, \boldsymbol{\lambda}_h[k] + \alpha(g_-(x_k) - \mathbf{y}_{\varrho h}[k])\}.$$

Secondly, to fulfill the control constraints, we add the Lagrange multiplier $\mathbf{w}_{\lambda h} \in \mathbb{R}^M$ that solves

$$K_h\mathbf{w}_{\lambda h} = \varrho K_h\mathbf{y}_{\varrho h} + M_h\mathbf{y}_{\varrho h} - \mathbf{y}_{dh} \tag{4.98}$$

and fulfills the complementarity conditions (4.90), which are equivalent to

$$\begin{aligned}\mathbf{w}_{\lambda h}[k] &= \min\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h+}[k] - (K_h\mathbf{y}_{\varrho h})[k])\} \\ &\quad + \max\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h-}[k] - (K_h\mathbf{y}_{\varrho h})[k])\}.\end{aligned}$$

Now, using that $K_h\mathbf{w}_{\lambda h} = \boldsymbol{\lambda}_h$, we want to find the roots of the function

$$\hat{\mathbf{F}}(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) = \begin{pmatrix}\hat{\mathbf{F}}_1(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) \\ \hat{\mathbf{F}}_2(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h})\end{pmatrix},$$

where

$$\hat{\mathbf{F}}_1(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) = \varrho K_h\mathbf{y}_{\varrho h} + M_h\mathbf{y}_{\varrho h} - K_h\mathbf{w}_{\lambda h} - \mathbf{y}_{dh},$$

and

$$\begin{aligned}\hat{\mathbf{F}}_2(\mathbf{y}_{\varrho h}, \mathbf{w}_{\lambda h}) &= (K_h + \beta I)\mathbf{w}_{\lambda h} \\ &\quad - \min\{0, (K_h\mathbf{w}_{\lambda h})[k] + \alpha(g_+(x_k) - \mathbf{y}_{\varrho h}[k])\} \\ &\quad\quad - \max\{0, (K_h\mathbf{w}_{\lambda h})[k] + \alpha(g_-(x_k) - \mathbf{y}_{\varrho h}[k])\} \\ &\quad\quad\quad -\beta\Big(\min\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h+} - (K_h\mathbf{y}_{\varrho h})[k])\} \\ &\quad\quad\quad\quad - \max\{0, \mathbf{w}_{\lambda h}[k] + \alpha(\mathbf{h}_{h-} - (K_h\mathbf{y}_{\varrho h})[k])\}\Big),\end{aligned}$$

where we introduce the parameter $\beta > 0$, to balance the relation between state and control constraints. This can again be achieved, applying a semi-smooth Newton method, with the Jacobian

$$D\hat{\mathbf{F}}(\mathbf{y}_h, \mathbf{w}_h) = \begin{pmatrix}\varrho K_h + M_h & -K_h \\ \alpha\big(L'(\mathbf{g}_{\pm h}) + \beta L'(\mathbf{h}_{\pm h})K_h\big) & \big(I - L'(\mathbf{g}_{\pm h})\big)K_h + \beta\big(I - L'(\mathbf{h}_{\pm h})\big)\end{pmatrix},$$

where, for the sake of presentation, we used the notation

$$L'(\mathbf{g}_{\pm h}) = L'(\mathbf{g}_{\pm h}, \mathbf{y}_h, \mathbf{w}_h) \quad \text{and} \quad L'(\mathbf{h}_{\pm h}) = L'(\mathbf{h}_{\pm h}, K_h\mathbf{y}_h, \mathbf{w}_h)$$

As a test example we consider $y_{cd,1}$ with state constraints

$$\hat{g}_-(x,y) \equiv 0 \quad \text{and} \quad \hat{g}_+(x,y) = \min\{y_{cd,1}(x,y), 0.5\}$$

and the control constraints

$$\hat{h}_-(x,y) \equiv 0 \quad \text{and} \quad \hat{h}_+(x,y) = \min\{u_1(x,y), 10\},$$

which is solved by applying the semi-smooth Newton method, where we choose $\beta \in \{1, 80\}$. The algorithm stopped for $\beta = 1$ after 13 iterations with a tolerance of $\mathrm{tol} = \max\{\mathrm{tol}_s, \mathrm{tol}_c\} = 6.103515\mathrm{e}{-04}$, , since after this point we do not see an

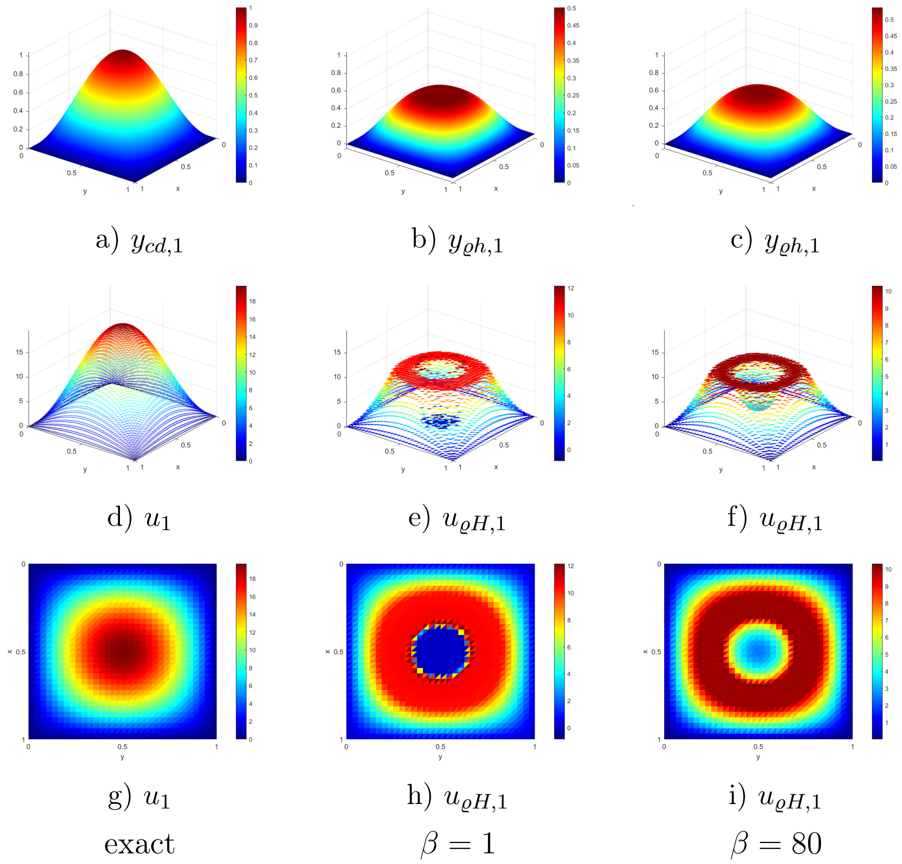


Figure 4.11: Target $y_{cd,1}$, computed constrainted states $y_{\varrho h,1}$ on a mesh with $N = 131072$ elements and $M = 65025$ DoFs and reconstructed constrained control $u_{\varrho H,1}$ on a mesh with $N_H = 8192$ elements for different values of β.

improvement of the solution. For $\beta = 80$ we stopped after 14 iterations with a tolerance of $\text{tol} = \max\{\text{tol}_s, \text{tol}_c\} = 3.646019\text{e}{-02}$. The results are depicted in Figure 4.11 and show that we meet both constraints pretty well. While the state constraints are met very well for both choices of β, the control constraints are more accurate for the a larger choice of β. This might also explain the larger tolerance.

To conclude this section, let us summarize that both state and control constraints can be casted in the setting of the energy regularization and efficient numerical methods can be derived for their incorporation. We saw that state and control constraints can be considered separately, as was done in [50], but also a simultaneous treatment is possible. The numerical computation of solutions involves the application of a semi-smooth Newton method. With this we inherit problems known from non-linear problems, as, e.g., finding the global minima of functions. In the numerical treatment we also observe that the behavior of solutions depends on the parameters involved in the Newton scheme, in particular, α and β. This needs to be further studied, but is far out of the scope of this work.

4.1.4 The energy regularization in $L^2(\Omega)$

Recall the model problem (4.1)-(4.2). In this section we will consider the space of the control $U = L^2(\Omega)$ for which the model problem now is to find $y_\varrho \in Y$ minimizing

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|u_\varrho\|^2_{L^2(\Omega)} \tag{4.99}$$

subject to

$$-\Delta y_\varrho = u_\varrho \text{ in } \Omega \quad \text{and} \quad y_\varrho = 0 \text{ on } \partial\Omega, \tag{4.100}$$

for given $y_d \in L^2(\Omega)$ and $\varrho > 0$. The goal of this section is to cast this problem in the abstract setting of the energy regularization and provide a rigorous finite element analysis. In order to apply the abstract theory, we first have to meet the Assumptions (B1)-(B3) and Assumptions (A1)-(A3) made on the operators $B : Y \to X^*$ and $A : X \to X^*$ and the underlying spaces. Note, that by the choice $U = L^2(\Omega) = X^*$ we already fixed $X = L^2(\Omega)$. Therefore, we can choose $A = I : L^2(\Omega) \to L^2(\Omega)$, for which it is obvious that A is self-adjoint, bounded and elliptic, i.e., Assumptions (A1)-(A3) are fulfilled. So, we need to find the appropriate space Y such that $B : Y \to L^2(\Omega)$ defines an isomorphism. Let us recall the space

$$H^1_\Delta(\Omega) := \{z \in H^1_0(\Omega) : \Delta z \in L^2(\Omega)\}$$

defined in the proof of Theorem 4.4. We stress that $H^1_\Delta(\Omega)$ is a Hilbert-space endowed with the inner product

$$\langle y, z\rangle_{H^1_\Delta(\Omega)} := \langle y, z\rangle_{H^1_0(\Omega)} + \langle \Delta y, \Delta z\rangle_{L^2(\Omega)}.$$

THEOREM 4.22. *The operator $B := -\Delta : H^1_\Delta(\Omega) \to L^2(\Omega)$ fulfills the Assumptions (B1)-(B3), i.e., it is an isomorphism. In particular,*

$$\|\Delta y\|_{L^2(\Omega)} \simeq \|y\|_{H^1_\Delta(\Omega)}$$

defines an equivalent norm for all $y \in H^1_\Delta(\Omega)$.

Proof. Let $y \in H^1_\Delta(\Omega)$ be arbitrary but fixed. For the boundedness (B1) we compute

$$\|By\|^2_{L^2(\Omega)} = \|\Delta y\|^2_{L^2(\Omega)} \le \|y\|^2_{H^1_0(\Omega)} + \|\Delta y\|^2_{L^2(\Omega)} = \|y\|^2_{H^1_\Delta(\Omega)}.$$

In order to show injectivity (B2), we first compute, using the Poincaré inequality,

$$\langle \nabla y, \nabla y\rangle_{L^2(\Omega)} \quad = \quad \langle -\Delta y, y\rangle_{L^2(\Omega)} \le \|\Delta y\|_{L^2(\Omega)}\|y\|_{L^2(\Omega)} \le c_P\|\Delta y\|_{L^2(\Omega)}\|\nabla y\|_{L^2(\Omega)},$$

i.e., $\|y\|_{H^1_0(\Omega)} = \|\nabla y\|_{L^2(\Omega)} \le c_P\|\Delta y\|_{L^2(\Omega)}$. With this we further have that

$$\|y\|^2_{H^1_\Delta(\Omega)} = \|y\|^2_{H^1_0(\Omega)} + \|\Delta y\|^2_{L^2(\Omega)} \le (c_P^2 + 1)\|\Delta y\|^2_{L^2(\Omega)},$$

and we conclude

$$\sup_{0 \ne q \in L^2(\Omega)} \frac{\langle -\Delta y, q\rangle_{L^2(\Omega)}}{\|q\|_{L^2(\Omega)}} = \|\Delta y\|_{L^2(\Omega)} \ge \frac{1}{\sqrt{c_P^2 + 1}}\|y\|_{H^1_\Delta(\Omega)}.$$

To show the surjectivity (B3), let $q \in L^2(\Omega) \setminus \{0\}$ be arbitrary but fixed. Let $y_q \in H^1_0(\Omega)$ denote the unique solution of

$$\langle \nabla y_q, \nabla \varphi\rangle_{L^2(\Omega)} = \langle q, \varphi\rangle_{L^2(\Omega)} \quad \text{for all } \varphi \in H^1_0(\Omega).$$

By density, we see that $-\Delta y_q = q \in L^2(\Omega)$ and thus $y_q \in H^1_\Delta(\Omega)$, which concludes the proof. □

With the preceding results, we can now apply the abstract theory, when choosing

$$H = X = L^2(\Omega) \quad \text{and} \quad Y = H^1_\Delta(\Omega)$$

and the operators

$$A = I : L^2(\Omega) \to L^2(\Omega) \quad \text{and} \quad B = -\Delta : H^1_\Delta(\Omega) \to L^2(\Omega).$$

In this case the Schur-complement operator is then given as $S := B^* A^{-1} B = \Delta^2 : H^1_\Delta(\Omega) \to [H^1_\Delta(\Omega)]^*$ and the minimizer $y_\varrho \in H^1_\Delta(\Omega)$ is characterized as the unique solution of the variational formulation (3.11), which reads

$$\varrho\langle \Delta y_\varrho, \Delta z\rangle_{L^2(\Omega)} + \langle y_\varrho, z\rangle_{L^2(\Omega)} = \langle y_d, z\rangle_{L^2(\Omega)} \quad \text{for all } z \in H^1_\Delta(\Omega). \tag{4.101}$$

The regularization error estimates can be transferred from Lemma 3.6.

Lemma 4.23. *Let $y_d \in L^2(\Omega)$ be given. For the unique solution $y_\varrho \in H^1_\Delta(\Omega)$ of* (4.101) *there holds*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}.$$

Further, if $y_d \in H^1_\Delta(\Omega)$, then

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \sqrt{\varrho}\|\Delta y_d\|_{L^2(\Omega)} \quad \text{and} \quad \|\Delta(y_\varrho - y_d)\|_{L^2(\Omega)} \le \|\Delta y_d\|_{L^2(\Omega)}. \tag{4.102}$$

Moreover, it holds

$$\|\Delta y_\varrho\|_{L^2(\Omega)} \le \|\Delta y_d\|_{L^2(\Omega)}. \tag{4.103}$$

At last, if $y_d \in H^1_\Delta(\Omega)$ such that $\Delta^2 y_d \in L^2(\Omega)$ it holds

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le \varrho\|\Delta^2 y_d\|_{L^2(\Omega)} \quad \text{and} \quad \|\Delta(y_\varrho - y_d)\|_{L^2(\Omega)} \le \sqrt{\varrho}\|\Delta^2 y_d\|_{L^2(\Omega)}, \tag{4.104}$$

and, in this case we also have

$$\|\Delta^2 y_\varrho\|_{L^2(\Omega)} \le \|\Delta^2 y_d\|_{L^2(\Omega)}. \tag{4.105}$$

Using a space interpolation argument, we can derive the following regularization error estimates.

Theorem 4.24 ([95, cf Theorem 4.1]). *Let $y_d \in H^s_0(\Omega)$ for $s \in [0,1]$ or $y_d \in H^1_0(\Omega) \cap H^s(\Omega)$ for $s \in (1,4]$. Then*

$$\|y_\varrho - y_d\|_{L^2(\Omega)} \le c\varrho^{s/4}\|y_d\|_{H^s(\Omega)} \text{ for } s \in [0,4] \tag{4.106}$$

and

$$\|\Delta(y_\varrho - y_d)\|_{L^2(\Omega)} \le c\varrho^{(s-2)/4}\|y_d\|_{H^s(\Omega)} \text{ for } s \in [2,4]. \tag{4.107}$$

Proof. Since we can estimate $\|\Delta y_d\|_{L^2(\Omega)} \le \|y_d\|_{H^2(\Omega)}$ and $\|\Delta^2 y_d\|_{L^2(\Omega)} \le \|y_d\|_{H^4(\Omega)}$, this is a direct consequence of the regularization error estimates in Lemma 4.23 and Theorem 2.14. We skip the details. □

Before we discuss the discretization, let us give some remarks.

Remark 4.25. *Instead of the variational formulation* (4.101) *we can also consider the equivalent system, as derived in* (3.10), *to find $(p_\varrho, y_\varrho) \in L^2(\Omega) \times H^1_\Delta(\Omega)$ such that*

$$\begin{aligned} \varrho^{-1}\langle p_\varrho, q\rangle_{L^2(\Omega)} - \langle \Delta y_\varrho, q\rangle_{L^2(\Omega)} &= 0 && \text{for all } q \in L^2(\Omega), \\ \langle p_\varrho, \Delta z\rangle_{L^2(\Omega)} + \langle y_\varrho, z\rangle_{L^2(\Omega)} &= \langle y_d, z\rangle_{L^2(\Omega)}, && \text{for all } z \in H^1_\Delta(\Omega). \end{aligned} \tag{4.108}$$

Assuming that $p_\varrho, q \in H^1(\Omega)$ *we can apply integration by parts to get*

$$\begin{aligned} \varrho^{-1}\langle p_\varrho, q\rangle_{L^2(\Omega)} + \langle \nabla y_\varrho, \nabla q\rangle_{L^2(\Omega)} &= \langle n\cdot\nabla y_\varrho, q\rangle_{\partial\Omega}, \\ -\langle \nabla p_\varrho, \nabla z\rangle_{L^2(\Omega)} + \langle y_\varrho, z\rangle_{L^2(\Omega)} &= \langle y_d, z\rangle_{L^2(\Omega)} - \langle p_\varrho, n\cdot\nabla z\rangle_{\partial\Omega}. \end{aligned}$$

Now choosing discrete trial spaces $X_h = Y_h = S_h^1(\mathcal{T}_h) \cap H_0^1(\Omega)$ *we can aim to find* $(p_{\varrho h}, y_{\varrho h}) \in X_h \times Y_h$ *such that*

$$\begin{aligned} \varrho^{-1}\langle p_{\varrho h}, q_h\rangle_{L^2(\Omega)} + \langle \nabla y_{\varrho h}, \nabla q_h\rangle_{L^2(\Omega)} &= 0 \quad \text{for all } q_h \in X_h, \\ -\langle \nabla p_{\varrho h}, \nabla z_h\rangle_{L^2(\Omega)} + \langle y_{\varrho h}, z_h\rangle_{L^2(\Omega)} &= \langle y_d, z_h\rangle_{L^2(\Omega)} \quad \text{for all } z_h \in Y_h, \end{aligned} \tag{4.109}$$

which is exactly the system (4.38) *we gained for the common* L^2*-regularization. However, the derivation shows that this formulation implicitly assumes higher regularity and Dirichlet boundary conditions on the adjoint state* $p_\varrho = \varrho\Delta y_\varrho = -u_\varrho \in H_0^1(\Omega)$*, and thus on the control.*

Discretization

For the discretization, we want to stay with conforming discretization spaces, to be able to apply the theory from the abstract setting. As mentioned in Remark 4.25 there is also mutual interest in the non-conforming setting, see, e.g., [75], where an optimal nested iteration is built based on mass lumping techniques. An overview of different discretization techniques is given in the survey by Brenner [17], where the incorporation of state constraints in this setting is discussed as well. For the rest of this subsection let us assume that $\Omega = [0, L_1] \times \ldots \times [0, L_d] \subset \mathbb{R}^d$, $d = 1,2,3$ is a rectangle ($d = 2$) or a quad ($d = 3$). Since the domain is convex, it holds that $H_\Delta^1(\Omega) = H_0^1(\Omega) \cap H^2(\Omega)$, see, e.g., [30, 59]. As a conforming ansatz space we can now use splines of second order, i.e., $S_h^2(\Omega) := S_h^2([0, L_1]) \otimes \ldots \otimes S_h^2([0, L_d])$ and $Y_h := S_h^2(\Omega) \cap H_0^1(\Omega) \subset H_\Delta^1(\Omega)$. The discrete variational formulation is then to find $y_{\varrho h} \in Y_h$ such that

$$\varrho\langle \Delta y_{\varrho h}, \Delta z_h\rangle_{L^2(\Omega)} + \langle y_{\varrho h}, z_h\rangle_{L^2(\Omega)} = \langle y_d, z_h\rangle_{L^2(\Omega)} \quad \text{for all } z_h \in Y_h. \tag{4.110}$$

This corresponds to the abstract discrete variational formulation (3.21) and transferring the results from Lemma 3.10 and best approximation error estimates by Theorem 3.11 we obtain unique solvability and the following finite element error estimates.

THEOREM 4.26. *Let* $y_d \in L^2(\Omega)$*. For the unique solution* $y_{\varrho h} \in Y_h$ *of* (4.110) *there holds the error estimate*

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}. \tag{4.111}$$

If additionally, $y_d \in H_\Delta^1(\Omega)$*, there holds*

$$\begin{aligned} \|y_{\varrho h} - y_d\|_{L^2(\Omega)} \le\ & c\big(\sqrt{\varrho}\|\Delta y_d\|_{L^2(\Omega)} \\ & + \inf_{z_h \in Y_h} \big[\varrho\|\Delta(y_d - z_h)\|^2_{L^2(\Omega)} + \|y_d - z_h\|^2_{L^2(\Omega)}\big]^{1/2}\big) \end{aligned} \tag{4.112}$$

and

$$\sqrt{\varrho}\|\Delta(y_{\varrho h}-y_d)\|_{L^2(\Omega)} \le c\Big(\sqrt{\varrho}\|\Delta y_d\|_{L^2(\Omega)} + \inf_{z_h\in Y_h}\big[\varrho\|\Delta(y_d-z_h)\|^2_{L^2(\Omega)}+\|y_d-z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big). \tag{4.113}$$

Moreover, if $y_d \in H^1_\Delta(\Omega)$ *and* $\Delta^2 y_d \in L^2(\Omega)$ *we have the error estimates*

$$\|y_{\varrho h}-y_d\|_{L^2(\Omega)} \le c\Big(\varrho\|\Delta^2 y_d\|_{L^2(\Omega)} + \inf_{z_h\in Y_h}\big[\varrho\|\Delta(y_d-z_h)\|^2_{L^2(\Omega)}+\|y_d-z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big) \tag{4.114}$$

and

$$\sqrt{\varrho}\|\Delta(y_{\varrho h}-y_d)\|_{L^2(\Omega)} \le c\Big(\varrho\|\Delta^2 y_d\|_{L^2(\Omega)} + \inf_{z_h\in Y_h}\big[\varrho\|\Delta(y_d-z_h)\|^2_{L^2(\Omega)}+\|y_d-z_h\|^2_{L^2(\Omega)}\big]^{1/2}\Big). \tag{4.115}$$

Similar to the case of the energy regularization in $H^{-1}(\Omega)$, we can now derive the optimal choice of the regularization parameter $\varrho > 0$ and error estimates in broken Sobolev spaces using the best approximation properties.

THEOREM 4.27. *Let* $y_d \in H^s_0(\Omega)$ *for* $s \in [0,1]$ *or* $y_d \in H^1_0(\Omega)\cap H^s(\Omega)$ *for* $s \in (1,3]$. *If* $\varrho = h^4$, *then*

$$\|y_{\varrho h}-y_d\|_{L^2(\Omega)} \le ch^s\|y_d\|_{H^s(\Omega)} \quad \text{for all } s\in[0,3] \tag{4.116}$$

and

$$\|\Delta(y_{\varrho h}-y_d)\|_{L^2(\Omega)} \le ch^{s-1}\|y_d\|_{H^s(\Omega)} \quad \text{for all } s\in[2,3]. \tag{4.117}$$

Proof. Our aim is to use a space interpolation argument, as in the proof of Theorem 4.6. Firstly, for $y_d \in L^2(\Omega)$ (4.111) gives

$$\|y_{\varrho h}-y_d\|_{L^2(\Omega)} \le \|y_d\|_{L^2(\Omega)}. \tag{4.118}$$

Secondly, consider $y_d \in H^1_0(\Omega)\cap H^3(\Omega)$. A triangle inequality first gives

$$\|y_{\varrho h}-y_d\|_{L^2(\Omega)} \le \|y_{\varrho h}-y_\varrho\|_{L^2(\Omega)} + \|y_\varrho-y_d\|_{L^2(\Omega)}.$$

We can estimate the second term by (4.106) as

$$\|y_\varrho-y_d\|_{L^2(\Omega)} \le \varrho^{3/4}\|y_d\|_{H^3(\Omega)}.$$

Using Cea's Lemma (Lemma 3.10) the first term is estimates as

$$\begin{aligned}\|y_{\varrho h}-y_\varrho\|^2_{L^2(\Omega)} &\le \inf_{z_h\in Y_h}\Big[\varrho\|\Delta(y_\varrho-z_h)\|^2_{L^2(\Omega)}+\|y_\varrho-z_h\|^2_{L^2(\Omega)}\Big]\\ &\le 2\inf_{z_h\in Y_h}\Big[\varrho\|\Delta(y_d-z_h)\|^2_{L^2(\Omega)}+\|y_d-z_h\|^2_{L^2(\Omega)}\Big]\\ &\quad +2\varrho\|\Delta(y_\varrho-y_d)\|^2_{L^2(\Omega)}+2\|y_\varrho-y_d\|^2_{L^2(\Omega)}\end{aligned}$$

and by the best approximation of Y_h, see Theorem 2.26, we have that

$$\inf_{z_h\in Y_h}\left[\varrho\|\Delta(y_d-z_h)\|^2_{L^2(\Omega)}+\|y_d-z_h\|^2_{L^2(\Omega)}\right]\le c(\varrho h^2+h^6)\|y_d\|^2_{H^3(\Omega)}.$$

Together with (4.106) and (4.107), we can thus estimate

$$\|y_{\varrho h}-y_\varrho\|^2_{L^2(\Omega)}\le c(\varrho h^2+h^6+\varrho\varrho^{1/2}+\varrho^{3/2})\|y_d\|^2_{H^3(\Omega)}.$$

For $\varrho=h^4$ we conclude the estimate

$$\|y_{\varrho h}-y_d\|_{L^2(\Omega)}\le ch^3\|y_d\|_{H^3(\Omega)}. \tag{4.119}$$

Interpolating the estimates (4.118)-(4.119) by using Theorem 2.14, we conclude the estimate (4.116). In a similar fashion we can derive

$$\|\Delta(y_{\varrho h}-y_d)\|_{L^2(\Omega)}\le h\|y_d\|_{H^3(\Omega)} \tag{4.120}$$

if $\varrho=h^4$ and the estimate (4.117) follows from interpolating (4.113) with (4.120). We skip the details. □

Before we give numerical examples, we state the convergence rates for the cost functional, depending on the regularity of the target.

LEMMA 4.28. *Let $y_d\in H^s_0(\Omega)$ for $s\in[0,1]$ or $y_d\in H^1_0(\Omega)\cap H^s(\Omega)$ for $s\in(1,2]$ and let $y_{\varrho h}\in Y_h$ be the unique solution of (4.110). Assume that the L^2-projection $Q^2_h:L^2(\Omega)\to Y_h$ fulfills*

$$\|\Delta Q^2_h z\|_{L^2(\Omega)}\le c\|\Delta z\|_{L^2(\Omega)}\quad \text{for all } z\in H^1_\Delta(\Omega)$$

and choose $\varrho=h^4$, then

$$\widetilde{\mathcal{J}}(y_{\varrho h})=\frac{1}{2}\|y_{\varrho h}-y_d\|^2_{L^2(\Omega)}+\frac{\varrho}{2}\|\Delta y_{\varrho h}\|^2_{L^2(\Omega)}\le ch^{2s}\|y_d\|^2_{H^s(\Omega)},\quad s\in[0,2].$$

Proof. Let $y_d\in L^2(\Omega)$. Then by (4.116) it holds

$$\|y_{\varrho h}-y_d\|_{L^2(\Omega)}\le\|y_d\|_{L^2(\Omega)}.$$

Testing (4.110) with $z_h=y_{\varrho h}\in Y_h$, we obtain

$$\varrho\|\Delta y_{\varrho h}\|^2_{L^2(\Omega)}+\|y_{\varrho h}\|^2_{L^2(\Omega)}=\langle y_d,y_{\varrho h}\rangle_{L^2(\Omega)}\le\|y_d\|_{L^2(\Omega)}\|y_{\varrho h}\|_{L^2(\Omega)}$$

from which we immediately conclude

$$\|y_{\varrho h}\|_{L^2(\Omega)}\le\|y_d\|_{L^2(\Omega)}\quad\text{and}\quad\|\Delta y_{\varrho h}\|_{L^2(\Omega)}\le\frac{1}{\sqrt{\varrho}}\|y_d\|_{L^2(\Omega)}.$$

Thus, we get

$$\begin{aligned}
\widetilde{\mathcal{J}}(y_{\varrho h}) &= \frac{1}{2}\|y_{\varrho h} - y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|\Delta y_{\varrho h}\|^2_{L^2(\Omega)} \\
&\leq \frac{1}{2}\|y_d\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\frac{1}{\varrho}\|y_d\|^2_{L^2(\Omega)} \\
&= \|y_d\|^2_{L^2(\Omega)}.
\end{aligned}$$

Now, let $y_d \in H^1_\Delta(\Omega)$. Then by (4.116) we have

$$\|y_{\varrho h} - y_d\|_{L^2(\Omega)} \leq ch^2\|y_d\|_{H^2(\Omega)}.$$

Further we compute, using (4.110), that

$$\begin{aligned}
\varrho\|\Delta y_{\varrho h}\|^2_{L^2(\Omega)} &= \langle y_d - y_{\varrho h}, y_{\varrho h}\rangle_{L^2(\Omega)} = \langle Q_h^2(y_d - y_{\varrho h}), y_{\varrho h}\rangle_{L^2(\Omega)} \\
&= -\|Q_h^2(y_d - y_{\varrho h})\|^2_{L^2(\Omega)} + \langle Q_h^2(y_d - y_{\varrho h}), y_d\rangle_{L^2(\Omega)} \\
&\leq \langle y_d - y_{\varrho h}, Q_h^2 y_d\rangle_{L^2(\Omega)} \\
&= \varrho\langle \Delta y_{\varrho h}, \Delta Q_h^2 y_d\rangle_{L^2(\Omega)} \\
&\leq \varrho\|\Delta y_{\varrho h}\|_{L^2(\Omega)}\|\Delta Q_h^2 y_d\|_{L^2(\Omega)} \\
&\leq c\varrho\|\Delta y_{\varrho h}\|_{L^2(\Omega)}\|\Delta y_d\|_{L^2(\Omega)},
\end{aligned}$$

where we used the self-adjointness and the assumed stability of Q_h^2. We conclude

$$\|\Delta y_{\varrho h}\|_{L^2(\Omega)} \leq c\|y_d\|_{H^2(\Omega)}.$$

Thus, for the cost functional we have, using $\varrho = h^4$

$$\begin{aligned}
\widetilde{\mathcal{J}}(y_{\varrho h}) &= \frac{1}{2}\|y_d - y_{\varrho h}\|^2_{L^2(\Omega)} + \frac{\varrho}{2}\|\Delta y_{\varrho h}\|^2_{L^2(\Omega)} \\
&\leq \frac{1}{2}ch^4\|y_d\|^2_{H^2(\Omega)} + \frac{\varrho}{2}c\|y_d\|^2_{H^2(\Omega)} \\
&\leq ch^4\|y_d\|^2_{H^2(\Omega)}.
\end{aligned}$$

Using a space interpolation argument we get

$$\widetilde{\mathcal{J}}(y_{\varrho h}) \leq ch^{2s}\|y_d\|^2_{H^s(\Omega)} \quad \text{for } s \in [0,2].$$ □

Numerical results

For $p \in \mathbb{N}_0$ the space $S_h^p(\Omega) = S_h^p([0, L_1]) \otimes \ldots \otimes S_h^p([0, L_d])$ is spanned by the functions

$$\varphi_k^p(x_1, \ldots, x_d) := \prod_{i=1}^{d} \varphi_{k_i}^p(x_i), \quad 0 \leq k_i \leq M_i + p - 1,\ i = 1, \ldots, d.$$

Reordering the DoFs, we can write

$$Y_h = S_h^2(\Omega) \cap H_0^1(\Omega) = \operatorname{span}\{\varphi_k^2\}_{k=1}^M.$$

Using the fe-isomorphism the discrete variational formulation (4.110) is then equivalent to the linear system of equations

$$(\varrho D_h + M_h)\mathbf{y}_{\varrho h} = \mathbf{y}_{dh}, \tag{4.121}$$

where the stiffness matrix and the mass matrix are given as

$$D_h[i,j] = \int_\Omega \Delta\varphi_j^2 \cdot \Delta\varphi_i^2 \, dx \quad \text{and} \quad M_h[i,j] = \int_\Omega \varphi_j^2 \varphi_i^2 \, dx, \quad i,j = 1,\dots,M$$

and the load vector has the entries

$$\mathbf{y}_{dh}[i] = \int_\Omega y_d \varphi_i^2 \, dx, \quad i = 1,\dots,M.$$

To show the sharpness of the theoretical results, we consider again three targets of different regularity defined on the unit square in $\Omega = (0,1)^2 \subset \mathbb{R}^2$. First, we consider $\hat{y}_{d,1} \in H^3(\Omega) \cap H_0^1(\Omega)$ defined as

$$\hat{y}_{d,1}(x,y) = \begin{cases} \sin(\pi x) 4^8 (y-0.25)^4 (0.75-y)^4, & 0.25 \le y \le 0.75, \\ 0, & \text{else.} \end{cases} \tag{4.122}$$

As a second example we consider a piecewise bilinear function $y_{d,2} \in H^{3/2-\varepsilon}(\Omega) \cap H_0^1(\Omega)$, $\varepsilon > 0$, defined as

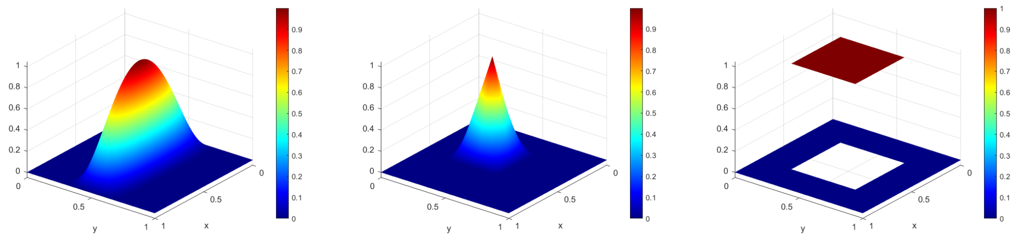
$$y_{d,2}(x,y) = \phi(x)\phi(y), \quad \phi(x) = \begin{cases} 1, & x = 0.45, \\ 0, & x \notin [0.2, 0.6], \\ \text{linear}, & \text{else.} \end{cases} \tag{4.123}$$

And finally, a discontinuous target $y_{d,3} \in H_0^{1/2-\varepsilon}(\Omega)$, $\varepsilon > 0$, defined as

$$y_{d,3}(x,y) = \begin{cases} 1, & (x,y) \in (0.25, 0.75)^2 \subset \Omega, \\ 0, & \text{else.} \end{cases} \tag{4.124}$$

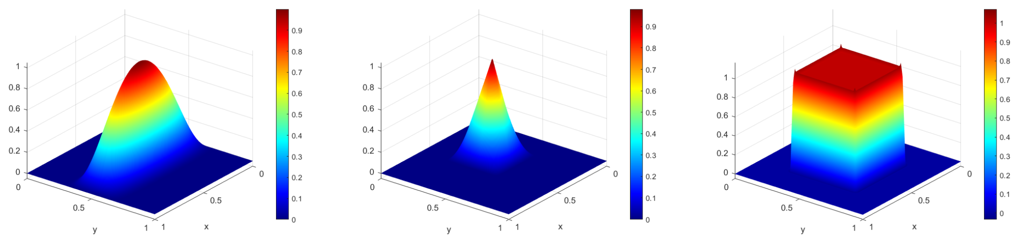
The targets are depicted in Figure 4.12.

The convergence rates for a uniform refinement are computed for an initial quadrilitteral mesh with $N = 36$ elements and $M = 36$ degrees of freedom (DoFs), see Figure 4.13. The initial mesh is chosen non-uniform to prevent superconvergence results for $\hat{y}_{d,1}$. For all three targets the results are depicted in Figure 4.14. Firstly, for a fixed parameter $\varrho = 10^{-8}$, we clearly see optimal convergence rates at first, which break down when $h = \varrho^{1/4}$ but independent of the regularity of the target. This is in total agreement with our theory, as the estimates in Theorem 4.26 reveal, i.e., we see the best approximation error up to the point where the constant terms are larger than the error. Secondly, Figure 4.14 shows that we get optimal orders of convergence for the



a) $\hat{y}_{d,1}$ b) $y_{d,2}$ c) $y_{d,3}$

Target functions $y_{d,i}$



d) $y_{\varrho h,1}$ e) $y_{\varrho h,2}$ f) $y_{\varrho h,3}$

Reconstruction of the target functions $y_{\varrho h,i}$

Figure 4.12: Target functions $y_{d,i}$, $i = 1, 2, 3$ and Reconstructed target functions $y_{\varrho h,i}$ on a mesh with $N = 36484$ elements and $M = 36481$ DoFs.

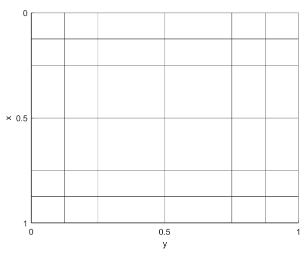
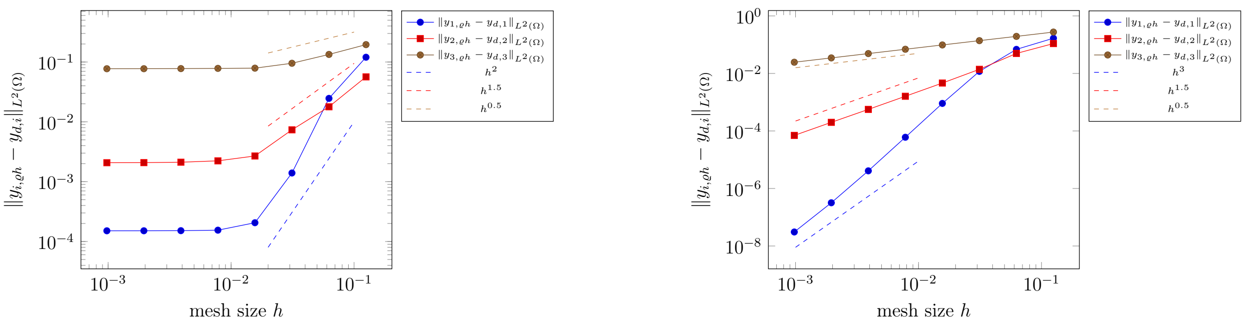


Figure 4.13: Initial mesh

choice $\varrho = h^4$, supporting the theoretical error estimates in Theorem 4.27. The reconstructed targets are depicted in Figure 4.12. We observe oscillations around the jump of the discontinuous target $y_{d,3}$. We saw the same behavior for the common L^2-regularization when using piecewise linear, globally continuous elements in Section 4.1.1, Figure 4.2. This indicates, that the behavior is a result of the higher regularity imposed by the method, i.e., by measuring the control in $L^2(\Omega)$ rather than $H^{-1}(\Omega)$ and does not stem from oscillations imposed by the higher order approximation. The convergence of the cost functional is plotted in Figure 4.14 for the choices $\varrho = 10^{-8}$ and $\varrho = h^4$. The results confirm the convergence rates proved in Lemma 4.28.

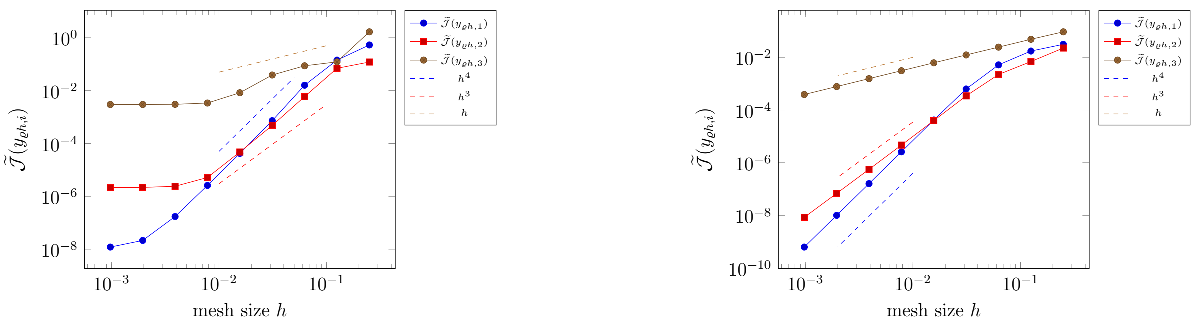
Reconstruction of the control

To compute a suitable reconstruction of the control $u_\varrho \in L^2(\Omega)$, we proceed as follows. As a conforming subspace for the discrete reconstruction, we choose $U_h = S_h^0(\Omega) = S_h^0([0, L_1]) \otimes \ldots \otimes S_h^0([0, L_d]) = \operatorname{span}\{\varphi_k^0\}_{k=1}^M$. Note, that $\dim(Y_h) = \dim(U_h)$ and we can compute the control on the same mesh as the state in this case. For the unique



a) $\varrho = 10^{-8}$ b) $\varrho = h^4$

Figure 4.14: Convergence for the three different target functions $\hat{y}_{d,1}$ and $y_{d,i}$, $i = 2, 3$ for the energy regularization in $L^2(\Omega)$ solving (4.121) for different choices of $\varrho > 0$.



a) $\varrho = 10^{-8}$ b) $\varrho = h^4$

Figure 4.15: Convergence of the cost functional $\widetilde{\mathcal{J}}$.

solution $y_{\varrho h} \in Y_h$ of (4.110), we can compute $u_{\varrho h} \in U_h$ as

$$\langle u_{\varrho h}, q_h \rangle_{L^2(\Omega)} = -\langle \Delta y_{\varrho h}, q_h \rangle_{L^2(\Omega)} \quad \text{for all } q_h \in U_h. \tag{4.125}$$

The following result holds true.

LEMMA 4.29. *The variational formulation* (4.125) *admits a unique solution* $u_{\varrho h} \in U_h$. *Let* $u_\varrho = -\Delta y_\varrho \in L^2(\Omega)$, *where* $y_\varrho \in H^1_\Delta(\Omega)$ *denotes the unique solution of* (4.101). *If* $y_\varrho \in H^{s+2}(\Omega)$, $s \in [0,1]$ *and* $\varrho = h^4$ *then it holds*

$$\|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)} \le c h^s \|y_\varrho\|_{H^{s+2}(\Omega)}.$$

Proof. Since the identity $I : L^2(\Omega) \to L^2(\Omega)$ is bounded, self-adjoint and elliptic on $L^2(\Omega)$, unique solvability follows by the Lemma of Lax–Milgram (Theorem 2.3). For the error estimate we use a Strang argument. Therefore, let $\tilde{u}_{\varrho h} \in U_h$ denote the unique solution of

$$\langle \tilde{u}_{\varrho h}, q_h \rangle_{L^2(\Omega)} = \langle -\Delta y_\varrho, q_h \rangle_{L^2(\Omega)} \quad \text{for all } q_h \in U_h. \tag{4.126}$$

To proceed, we use a triangle inequality to estimate

$$\|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)} \le \|u_\varrho - \tilde{u}_{\varrho h}\|_{L^2(\Omega)} + \|\tilde{u}_{\varrho h} - u_{\varrho h}\|_{L^2(\Omega)}. \tag{4.127}$$

Using Cea's Lemma (Theorem 2.15) and the best approximation results (Theorem 2.26) we immediately estimate the first term in (4.127) by

$$\begin{aligned} \|u_\varrho - \tilde{u}_{\varrho h}\|_{L^2(\Omega)} &\le \inf_{q_h \in U_h} \|u_\varrho - q_h\|_{L^2(\Omega)} \le c h^s \|u_\varrho\|_{H^s(\Omega)} \\ &= c h^s \|\Delta y_\varrho\|_{H^s(\Omega)} \le c h^s \|y_\varrho\|_{H^{s+2}(\Omega)}, \end{aligned}$$

for $s \in [0,1]$. Testing (4.126) and (4.125) with $q_h = \tilde{u}_{\varrho h} - u_{\varrho h} \in U_h$ we compute for the second term in (4.127)

$$\begin{aligned} \|\tilde{u}_{\varrho h} - u_{\varrho h}\|^2_{L^2(\Omega)} &= \langle \tilde{u}_{\varrho h} - u_{\varrho h}, \tilde{u}_{\varrho h} - u_{\varrho h} \rangle_{L^2(\Omega)} \\ &= \langle -\Delta(y_\varrho - y_{\varrho h}), \tilde{u}_{\varrho h} - u_{\varrho h} \rangle_{L^2(\Omega)} \\ &\le \|\Delta(y_\varrho - y_{\varrho h})\|_{L^2(\Omega)} \|\tilde{u}_{\varrho h} - u_{\varrho h}\|_{L^2(\Omega)}, \end{aligned}$$

i.e.,

$$\|\tilde{u}_{\varrho h} - u_{\varrho h}\|_{L^2(\Omega)} \le \|\Delta(y_\varrho - y_{\varrho h})\|_{L^2(\Omega)}.$$

Noting that for the solution $y_{\varrho h} \in Y_h$ of (4.110) Cea's Lemma holds true, we have

$$\begin{aligned} &\varrho \|\Delta(y_\varrho - y_{\varrho h})\|^2_{L^2(\Omega)} + \|y_\varrho - y_{\varrho h}\|^2_{L^2(\Omega)} \\ &\qquad \le \inf_{z_h \in Y_h} \left[\varrho \|\Delta(y_\varrho - z_h)\|^2_{L^2(\Omega)} + \|y_\varrho - z_h\|^2_{L^2(\Omega)} \right] \end{aligned}$$

With this, using $\varrho = h^4$ and the best approximation results (Theorem 2.23), we compute

$$\|\Delta(y_\varrho - y_{\varrho h})\|_{L^2(\Omega)} \leq \begin{cases} c\|y_\varrho\|_{H^2(\Omega)}, & \text{for } y_\varrho \in H_0^1(\Omega) \cap H^2(\Omega), \\ ch\|y_\varrho\|_{H^3(\Omega)}, & \text{for } y_\varrho \in H_0^1(\Omega) \cap H^3(\Omega), \end{cases}$$

and subsequently $\|\Delta(y_\varrho - y_{\varrho h})\|_{L^2(\Omega)} \leq ch^s \|y_\varrho\|_{H^{s+2}(\Omega)}$ for $s \in [0,1]$, which concludes the proof. $\square$

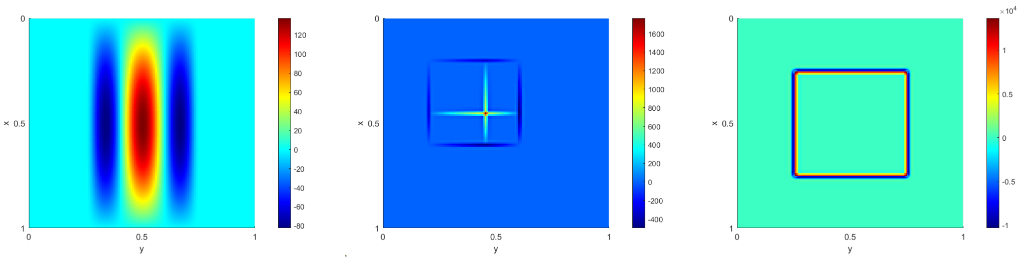
Using the fe-isomorphism, (4.125) is equivalent to the linear system of equations

$$\bar{M}_h \mathbf{u}_{\varrho h} = K_h \mathbf{y}_{\varrho h} \tag{4.128}$$

where the mass matrix and the stiffness matrix are given as

$$\bar{M}_h[\ell,k] = \langle \varphi_\ell^0, \varphi_k^0 \rangle_{L^2(\Omega)} \quad \text{and} \quad K_h[\ell,j] = \langle -\Delta\varphi_j^2, \varphi_\ell^0 \rangle_{L^2(\Omega)}, \quad \ell,k = 1,\ldots,M.$$

The results on for the three different targets are depicted in Figure 4.16.



a) $\hat{y}_{d,1}$ b) $y_{d,2}$ c) $y_{d,3}$

Figure 4.16: Reconstruction of the controls $u_{\varrho h,i}$ from the computed targets $y_{\varrho h,i}$ on a mesh with $N = 36864$ elements.

4.2 A hyperbolic model problem

So far we only considered the application of the abstract framework developed in Chapter 3 to elliptic problems. To show its full capacity, we will now consider the optimal control problem subject to the homogeneous wave equation as a model for a hyperbolic optimal control problem. Therefore, let $\Omega \subset \mathbb{R}^d$, $d = 1,2,3$, be a bounded convex domain with Lipschitz boundary $\Gamma = \partial\Omega$, for $d = 2,3$, and let $0 < T < \infty$ be a given finite time horizon. Then we introduce the space-time domain $Q := \Omega \times (0,T)$ and the lateral boundary $\Sigma := \Gamma \times (0,T)$. For a given target

$y_d \in L^2(Q)$ and a regularization parameter $\varrho > 0$, we consider the minimization of the cost functional

$$\mathcal{J}(y_\varrho, u_\varrho) := \frac{1}{2}\|y_\varrho - y_d\|^2_{L^2(Q)} + \frac{\varrho}{2}\|u_\varrho\|^2_{X^*} \tag{4.129}$$

subject to the initial boundary value problem for the wave equation with homogeneous Dirichlet boundary conditions,

$$\begin{aligned} \Box y_\varrho(x,t) := \partial_{tt} y_\varrho(x,t) - \Delta_x y_\varrho(x,t) &= u_\varrho(x,t) && \text{for } (x,t) \in Q, \\ y_\varrho(x,t) &= 0 && \text{for } (x,t) \in \Sigma, \\ y_\varrho(x,0) = \partial_t y_\varrho(x,t)_{|t=0} &= 0 && \text{for } x \in \Omega. \end{aligned} \tag{4.130}$$

4.2.1 The energy regularization in $[H^{1,1}_{0;,0}(Q)]^*$

To apply the abstract framework of Chapter 3, we will employ a space-time variational formulation of (4.129)-(4.130), following [87]. The first crucial property will be to find Hilbert spaces such that the wave operator $B := \Box : Y \to X^*$, satisfies the assumptions (B1)-(B3), i.e., it defines an isomorphism. Such spaces were constructed in [111, 116], from which we rephrase the main ideas. First, let us consider $u_\varrho \in X^* = L^2(Q)$. To derive a variational formulation for (4.130), we multiply by a test function $q(x,t) \in \mathcal{C}^\infty(Q)$, integrate over the space-time domain Q and apply integration by parts, to get

$$\begin{aligned} \int_0^T \int_\Omega u_\varrho(x,t) q(x,t)\, dx\, dt &= \int_0^T \int_\Omega \partial_{tt} y_\varrho(x,t) q(x,t) - \Delta_x y_\varrho(x,t) q(x,t)\, dx\, dt \\ &= \int_\Omega \left[\partial_t y_\varrho(x,t) q(x,t) \Big|_0^T - \int_0^T \partial_t y_\varrho(x,t) \partial_t q(x,t)\, dt \right] dx \\ &\quad + \int_0^T \left[\int_{\partial\Omega} n_x \cdot \nabla_x y_\varrho(x,t) q(x,t)\, ds_x + \int_\Omega \nabla_x y_\varrho(x,t) \cdot \nabla_x q(x,t)\, dx \right]. \end{aligned}$$

Now, since $\partial_t y_\varrho(x,t)\big|_{t=0} = 0$, choosing a test function that satisfies $q(x,T) = 0$ and $q\big|_\Sigma \equiv 0$, we get

$$b(y_\varrho, q) := -\langle \partial_t y_\varrho, \partial_t q \rangle_{L^2(Q)} + \langle \nabla_x y_\varrho, \nabla_x q \rangle_{L^2(Q)} = \langle u_\varrho, q \rangle_{L^2(Q)}. \tag{4.131}$$

Recall the spaces, incorporating the spatial boundary condition and the initial/terminal condition in a weak sense

$$\begin{aligned} H^{1,1}_{0;0,}(Q) &= H^1_{0,}(0,T; L^2(\Omega)) \cap L^2(0,T; H^1_0(\Omega)), \\ H^{1,1}_{0;,0}(Q) &= H^1_{,0}(0,T; L^2(\Omega)) \cap L^2(0,T; H^1_0(\Omega)), \end{aligned}$$

both equipped with the norm

$$|q|_{H^1(Q)} = \sqrt{\|\partial_t q\|^2_{L^2(Q)} + \|\nabla_x q\|^2_{L^2(Q)}}. \tag{4.132}$$

Then the bilinear form $b(\cdot,\cdot) : H^{1,1}_{0;0,}(Q) \times H^{1,1}_{0;,0}(Q) \to \mathbb{R}$, is well-defined and and bounded, i.e.,

$$|b(z,q)| \le |z|_{H^1(Q)} |q|_{H^1(Q)}$$

and for any $u_\varrho \in L^2(Q)$ the variational formulation (4.131) admits a unique solution $y_\varrho \in H^{1,1}_{0;0,}(Q)$, fulfilling the stability estimate, see e.g. [71, Theorem 5.1, p.169] and [110, Theorem 5.1],

$$\|y_\varrho\|_{H^{1,1}_{0;0,}(Q)} \le \frac{T}{\sqrt{2}} \|u_\varrho\|_{L^2(Q)}.$$

But, although we have unique solvability, the operator $B : H^{1,1}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ associated to the bilinear form, by

$$\langle By, q\rangle_Q := b(y,q) \quad \text{for all } y \in H^{1,1}_{0;0,}(Q),\ q \in H^{1,1}_{0;,0}(Q),$$

does not define an isomorphism.

THEOREM 4.30 ([111, Theorem 1.1]). *There does not exist a constant $c > 0$ such that each right-hand side $u_\varrho \in L^2(Q)$ and the corresponding solution $y_\varrho \in H^{1,1}_{0;0,}(Q)$ of* (4.131) *satisfy*

$$\|y_\varrho\|_{H^{1,1}_{0;0,}(Q)} \le c\, \|u_\varrho\|_{[H^{1,1}_{0;,0}(Q)]^*}.$$

In particular, Assumption (B2)

$$c_1^B \, \|y\|_{H^{1,1}_{0;0,}(Q)} \le \sup_{0 \ne q \in H^{1,1}_{0;,0}(Q)} \frac{\langle By, q\rangle_{L^2(Q)}}{\|q\|_{H^{1,1}_{0;,0}(Q)}} \quad \textit{for all}\ y \in H^{1,1}_{0;0,}(Q)$$

with a constant $c_1^B > 0$ does not *hold true.*

In the following, let us define suitable spaces, such that the wave operator is an isomorphism. The first issue to overcome is the establishment of Assumption (B2). It fails to hold, since the initial condition $\partial_t y_\varrho(x,t)\big|_{t=0}$ enters the variational formulation naturally, which is not appropriate in this case. In order to incorporate it in a meaningful sense, we will modify the ansatz space. Let $Q_- := \Omega \times (-T,T)$ denote the enlarged space-time domain and define the zero extension of a function $y \in L^2(Q)$ by

$$\widetilde{y}(x,t) := \begin{cases} y(x,t) & \text{for } (x,t) \in Q, \\ 0, & \text{else.} \end{cases}$$

Then, we consider the application of the wave operator in a distributional sense, i.e., for all $\varphi \in C_0^\infty(Q_-)$ we define

$$\langle \Box \widetilde{y}, \varphi \rangle_{Q_-} := \int_{Q_-} \widetilde{y}(x,t)\, \Box \varphi(x,t)\, dx\, dt = \int_Q y(x,t)\, \Box \varphi(x,t)\, dx\, dt.$$

Using this definition, we can introduce the space

$$\mathcal{H}(Q) := \Big\{ y = \widetilde{y}_{|Q} : \widetilde{y} \in L^2(Q_-),\, \widetilde{y}_{|\Omega\times(-T,0)} = 0,\, \Box \widetilde{y} \in [H_0^1(Q_-)]^* \Big\},$$

with the graph norm

$$\|y\|_{\mathcal{H}(Q)} := \sqrt{\|y\|^2_{L^2(Q)} + \|\Box \widetilde{y}\|^2_{[H_0^1(Q_-)]^*}}\,.$$

The normed vector space $(\mathcal{H}(Q), \|\cdot\|_{\mathcal{H}(Q)})$ is a Banach space, and it holds that, see [111, Lemma 3.5], $H^{1,1}_{0;0,}(Q) \subset \mathcal{H}(Q)$ i.e.,

$$\|\Box \widetilde{y}\|_{[H_0^1(Q_-)]^*} \le \|y\|_{H^{1,1}_{0;0,}(Q)} \quad \text{for all } y \in H^{1,1}_{0;0,}(Q). \tag{4.133}$$

Therefore, we can introduce the space

$$\mathcal{H}_{0;0,}(Q) := \overline{H^{1,1}_{0;0,}(Q)}^{\|\cdot\|_{\mathcal{H}(Q)}} \subset \mathcal{H}(Q),$$

which will serve as ansatz space. For $y \in \mathcal{H}_{0;0,}(Q)$, an equivalent norm is given as, see [111, Lemma 3.6],

$$\|y\|_{\mathcal{H}_{0;0,}(Q)} = \|\Box \widetilde{y}\|_{[H_0^1(Q_-)]^*}.$$

For given $u_\varrho \in [H^{1,1}_{0;,0}(Q)]^*$ we now study the variational formulation to find $y_\varrho \in \mathcal{H}_{0;0,}(Q)$ such that

$$\tilde{b}(y_\varrho, q) := \langle \Box \widetilde{y}_\varrho, \mathcal{E} q \rangle_{Q_-} = \langle u_\varrho, q \rangle_Q \quad \text{for all } q \in H^{1,1}_{0;,0}(Q), \tag{4.134}$$

where $\mathcal{E} : H^{1,1}_{0;,0}(Q) \to H_0^1(Q_-)$ is a suitable extension operator, e.g., reflection in time with respect to $t = 0$, satisfying

$$\|\mathcal{E} q\|_{H_0^1(Q_-)} \le 2\, \|q\|_{H^{1,1}_{0;,0}(Q)} \quad \text{for all } q \in H^{1,1}_{0;,0}(Q).$$

We conclude that the bilinear form $\tilde{b}(\cdot,\cdot) : \mathcal{H}_{0;0,}(Q) \times H^{1,1}_{0;,0}(Q) \to \mathbb{R}$ within the variational formulation (4.134) is bounded, as for all $y \in \mathcal{H}_{0;0,}(Q)$ and $q \in H^{1,1}_{0;,0}(Q)$ we have

$$|\tilde{b}(y_\varrho, q)| \le \|\Box \widetilde{y}\|_{[H_0^1(Q_-)]^*} \|\mathcal{E} q\|_{H_0^1(Q_-)} \le 2\, \|y\|_{\mathcal{H}_{0;0,}(Q)} \|q\|_{H^{1,1}_{0;,0}(Q)}. \tag{4.135}$$

Moreover, we have the following result.

THEOREM 4.31 ([111, Theorem 3.9]). *For each given $u_\varrho \in [H^{1,1}_{0;,0}(Q)]^*$, there exists a unique solution $y_\varrho \in \mathcal{H}_{0;0,}(Q)$ of the variational formulation (4.134) satisfying*

$$\|y_\varrho\|_{\mathcal{H}_{0;0,}(Q)} = \|\Box \widetilde{y}_\varrho\|_{[H^1_0(Q_-)]^*} = \|u_\varrho\|_{[H^{1,1}_{0;,0}(Q)]^*}.$$

In particular, there holds the inf-sup stability condition

$$\|y\|_{\mathcal{H}_{0;0,}(Q)} \leq \sup_{0 \neq q \in H^{1,1}_{0;,0}(Q)} \frac{\tilde{b}(y,q)}{\|q\|_{H^{1,1}_{0;,0}(Q)}} \quad \text{for all } y \in \mathcal{H}_{0;0,}(Q). \tag{4.136}$$

This result now gives raise to an isomorphic operator for the solution of the wave equation for controls $u_\varrho \in [H^{1,1}_{0;,0}(Q)]^*$. We summarize the findings in the next lemma.

LEMMA 4.32. *Let $\widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ be the linear operator associated with the bilinear form $\tilde{b}(\cdot,\cdot)$ defined as*

$$\langle \widetilde{B}y, q\rangle_Q := \tilde{b}(y,q) = \langle \Box \widetilde{y}, \mathcal{E}q\rangle_{Q_-} \quad \text{for all } y \in \mathcal{H}_{0;0,}(Q),\ q \in [H^{1,1}_{0;,0}(Q)]^*.$$

Then $\widetilde{B}$ satisfies the Assumptions (B1)-(B3) with constants $c_1^B = c_2^B = 1$. Moreover, the restriction fulfills

$$\widetilde{B}\big|_{H^{1,1}_{0;0,}(Q)} = B : H^{1,1}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*,$$

or, more precisely,

$$\tilde{b}(y,q) = b(y,q) \quad \text{for all } y \in H^{1,1}_{0;0,}(Q),\ q \in H^{1,1}_{0;,0}(Q). \tag{4.137}$$

Proof. Assumption (B1), i.e., boundedness of $\tilde{b}(\cdot,\cdot) : \mathcal{H}_{0;0,}(Q) \times H^{1,1}_{0;,0}(Q) \to \mathbb{R}$, was already established in (4.135). The improved boundedness constant $c_2^B = 1$ is given in [111, p. 22, (3.9)]. Assumption (B2) is exactly (4.136). To show (B3) let $\hat{q} \in H^{1,1}_{0;,0}(Q)$ be arbitrary but fixed and consider the linear functional $\langle u_{\hat{q}}, \cdot\rangle_Q := \langle \hat{q}, \cdot\rangle_{H^{1,1}_{0;,0}(Q)} : H^{1,1}_{0;,0}(Q) \to \mathbb{R}$. By Theorem 4.31, there exists a unique $y_{\hat{q}} \in \mathcal{H}_{0;0,}(Q)$, such that

$$\tilde{b}(y_{\hat{q}}, q) = \langle u_{\hat{q}}, q\rangle_Q \quad \text{for all } q \in H^{1,1}_{0;,0}(Q).$$

In particular, for $q = \hat{q}$, we have

$$\tilde{b}(y_{\hat{q}}, \hat{q}) = \langle u_{\hat{q}}, \hat{q}\rangle_Q = \|\hat{q}\|^2_{H^{1,1}_{0;,0}(Q)} \neq 0.$$

A proof of the property (4.137) is given in [111, Lemma 3.5]. $\square$

Remark 4.33. *If we consider conforming trial spaces* $Y_h \subset H^{1,1}_{0;0,}(Q) \subset \mathcal{H}_{0;0,}(Q)$ *and* $X_h \subset H^{1,1}_{0;,0}(Q)$, *by (4.137) it holds*

$$\tilde{b}(y_h, q_h) = -\langle \partial_t y_h, \partial_t q_h \rangle_{L^2(Q)} + \langle \nabla_x y_h, \nabla_x q_h \rangle_{L^2(Q)} \text{ for all } y_h \in Y_h,\ q_h \in X_h.$$

This is of particular interest when considering the discrete setting, as piecewise linear continuous functions are in $H^1(Q)$.

Moreover, we need to have an operator, realizing the norm in $[H^{1,1}_{0;,0}(Q)]^*$, which is guaranteed by the Assumptions (A1)-(A3).

Lemma 4.34. *The operator* $A : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ *defined as*

$$\langle Ap, q \rangle_Q := \langle \partial_t p, \partial_t q \rangle_{L^2(Q)} + \langle \nabla_x p, \nabla_x q \rangle_{L^2(Q)} \quad \text{for all } p, q \in H^{1,1}_{0;,0}(Q), \tag{4.138}$$

fulfills Assumptions (A1)-(A3) with constants $c_1^A = c_2^A = 1$ *and thus for all* $u \in [H^{1,1}_{0;,0}(Q)]^*$ *it holds that*

$$\|u\|_{[H^{1,1}_{0;,0}(Q)]^*} = \|u_\varrho\|_{A^{-1}} = \sqrt{\langle u, A^{-1}u \rangle_Q}.$$

Proof. By definition the operator is self-adjoint, which gives (A2). Moreover, we easily compute, using the Cauchy-Schwarz inequality twice, that

$$\langle Ap, q \rangle_Q \leq \|p\|_{H^{1,1}_{0;,0}(Q)} \|q\|_{H^{1,1}_{0;,0}(Q)} \quad \text{for all } p, q \in H^{1,1}_{0;,0}(Q),$$

i.e., (A1). (A3), i.e., ellipticity, follows directly as

$$\langle Aq, q \rangle_Q = |q|^2_{H^1(Q)} = \|q\|^2_{H^{1,1}_{0;,0}(Q)} \quad \text{for all } q \in H^{1,1}_{0;,0}(Q).$$

□

Thus, the optimal control problem (4.129)-(4.130) fits into the framework of Chapter 3, when choosing the spaces

$$H = L^2(Q), \quad X = H^{1,1}_{0;,0}(Q), \quad Y = \mathcal{H}_{0;0,}(Q),$$

and the operators

$$A : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^* \quad \text{and} \quad \widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$$

defined as in Lemma 4.34 and Lemma 4.32, and we can directly apply all the results derived in the abstract setting. Firstly, using that $\widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ is an isomorphism, we can eliminate the control from $\widetilde{B}y_\varrho = u_\varrho$ and consider the reduced cost functional

$$\begin{aligned}\tilde{\mathcal{J}}(y_\varrho) &= \frac{1}{2}\|y_\varrho - y_d\|^2_{L^2(Q)} + \frac{\varrho}{2}\|\widetilde{B}y_\varrho\|^2_{[H^{1,1}_{0;,0}(Q)]^*} \\ &= \frac{1}{2}\langle y_\varrho - y_d, y_\varrho - y_d\rangle_{L^2(Q)} + \frac{\varrho}{2}\langle \widetilde{B}y_\varrho, A^{-1}\widetilde{B}y_\varrho\rangle_Q,\end{aligned}$$

for which the minimizer is characterized as the unique solution $y_\varrho \in Y$ of the variational equation (3.11)

$$\varrho\langle Sy_\varrho, z\rangle_{L^2(Q)} + \langle y_\varrho, z\rangle_{L^2(Q)} = \langle y_d, z\rangle_{L^2(Q)} \quad \text{for all } z \in \mathcal{H}_{0;0,}(Q) \tag{4.139}$$

where $S := \widetilde{B}^* A^{-1}\widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [\mathcal{H}_{0;0,}(Q)]^*$ denotes the Schur complement operator, which is symmetric, self-adjoint and elliptic and defines an equivalent norm on $\mathcal{H}_{0;0,}(Q)$, see Lemma 3.4, i.e.,

$$\|y\|_{\mathcal{H}_{0;0,}(Q)} \le \|y\|_S := \sqrt{\langle \widetilde{B}y, A^{-1}\widetilde{B}y\rangle_{L^2(Q)}} \le \|y\|_{\mathcal{H}_{0;0,}(Q)}. \tag{4.140}$$

Now, by Lemma 3.6, we get the following regularization error estimates, depending on the regularization parameter $\varrho > 0$.

LEMMA 4.35 ([87, Theorem 3.2]). *Let $y_d \in L^2(Q)$ be given. For the unique solution $y_\varrho \in \mathcal{H}_{0;0,}(Q)$ of (4.139) there holds*

$$\|y_\varrho - y_d\|_{L^2(Q)} \le \|y_d\|_{L^2(Q)}. \tag{4.141}$$

Further, if $y_d \in \mathcal{H}_{0;0,}(Q)$, then

$$\|y_\varrho - y_d\|_{L^2(Q)} \le \sqrt{\varrho}\|y_d\|_S \quad \text{and} \quad \|y_\varrho - y_d\|_S \le \|y_d\|_S. \tag{4.142}$$

Moreover, it holds

$$\|y_\varrho\|_S \le \|y_d\|_S. \tag{4.143}$$

At last, if $y_d \in \mathcal{H}_{0;0,}(Q)$ such that $Sy_d \in L^2(Q)$ it holds

$$\|y_\varrho - y_d\|_{L^2(Q)} \le \varrho\|Sy_d\|_{L^2(Q)} \quad \text{and} \quad \|y_\varrho - y_d\|_S \le \sqrt{\varrho}\|Sy_d\|_{L^2(Q)}, \tag{4.144}$$

and, in this case we also have

$$\|Sy_\varrho\|_{L^2(Q)} \le \|Sy_d\|_{L^2(Q)}. \tag{4.145}$$

Note, that the operator $A := -\Delta_{(x,t)} : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ corresponds to the space-time Laplacian with mixed Dirichlet and Neumann boundary conditions. Therefore, the solution $p \in H^{1,1}_{0;,0}(Q)$ of $Ap = u$ in Q admits the regularity $p \in H^{r+1}(Q) \cap H^{1,1}_{0;,0}(Q)$ for given $u \in H^{r-1}(Q)$ and some $0 \le r \le 1$ depending on the geometry of the space-time domain, see e.g. [30, 59]. In particular, for $y_d \in H^2(Q) \cap H^{1;1}_{0;0,}(Q)$ it holds that $\widetilde{B}y_d \in L^2(Q)$. But, we can in general *not* guarantee that $A^{-1}\widetilde{B}y_d \in H^2(Q)$ and subsequently, $Sy_d = \widetilde{B}^* A^{-1}\widetilde{B}y_d \in L^2(Q)$ does *not* need to hold true.

THEOREM 4.36 ([87, c.f. Corollary 3.3]). *Let* $y_d \in H^{s,s}_{0;0,}(Q) = [L^2(Q), H^{1,1}_{0;0,}(Q)]_s$*, for* $s \in [0,1]$ *or* $y_d \in H^s(Q) \cap H^{1,1}_{0;0,}(Q)$ *such that* $Sy_d \in H^{s-2}(Q)$ *for* $s \in [1,2]$*. Then*

$$\|y_d - y_\varrho\|_{L^2(Q)} \leq c\varrho^{s/2}\|y_d\|_{H^s(Q)}, \quad s \in [0,1], \tag{4.146}$$

and

$$\|y_d - y_\varrho\|_{L^2(Q)} \leq c\varrho^{s/2}\|y_d\|_{H^s(Q)}, \quad s \in [1,2]. \tag{4.147}$$

Proof. For $y_d \in H^{1,1}_{0;0,}(Q)$ we get from the estimate (4.142), using the norm equivalence (4.140) and the property (4.133), that

$$\|y_d - y_\varrho\|_{L^2(Q)} \leq \varrho^{1/2}\|y_d\|_S \leq \varrho^{1/2}\|y_d\|_{\mathcal{H}_{0;0,}(Q)} \leq \varrho^{1/2}\|y_d\|_{H^{1,1}_{0;0,}(Q)}. \tag{4.148}$$

Now interpolating with (4.141), using Theorem 2.14, gives (4.146). Moreover, if $y_d \in H^2(Q) \cap H^{1,1}_{0;0,}(Q)$ such that $Sy_d \in L^2(Q)$, we get, using (4.144), that

$$\|y_d - y_\varrho\|_{L^2(Q)} \leq \varrho\|Sy_d\|_{L^2(Q)} \leq c\varrho\|y_d\|_{H^2(Q)}$$

and interpolating with (4.148) gives (4.147). □

REMARK 4.37 (L^2-regularization). *Recall, that* $B : H^{1,1}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ *does not define an isomorphism. Though, the problem to find* $y_\varrho \in H^{1,1}_{0;0,}(Q)$ *such that* $By_\varrho = u_\varrho$ *in* $[H^{1,1}_{0;,0}(Q)]^*$ *admits a unique solution for all* $u_\varrho \in L^2(Q)$*. This approach does not fit into the framework of the energy regularization, but we can derive an optimality system, analogously to the procedure for the Poisson equation in Remark 4.2. Namely, we can define the solution operator* $\mathcal{S} : L^2(Q) \to H^{1,1}_{0;0,}(Q)$ *by* $\mathcal{S}u_\varrho = y_\varrho$ *and consider the reduced cost functional*

$$\hat{\mathcal{I}}(u_\varrho) = \frac{1}{2}\|\mathcal{S}u_\varrho - y_d\|^2_{L^2(Q)} + \frac{\varrho}{2}\|u_\varrho\|^2_{L^2(Q)},$$

for which the minimizer satisfies the gradient equation

$$p_\varrho + \varrho u_\varrho = 0 \quad \text{in } L^2(Q), \tag{4.149}$$

where $p_\varrho \in H^{1,1}_{0;,0}(Q)$ *is the weak solution of the adjoint problem*

$$\begin{aligned} \Box p_\varrho(x,t) = \partial_{tt}p_\varrho(x,t) - \Delta_x p_\varrho(x,t) &= y_\varrho(x,t) - y_d(x,t) && \text{for } (x,t) \in Q, \\ p_\varrho(x,t) &= 0 && \text{for } (x,t) \in \Sigma, \\ p_\varrho(x,T) = \partial_t p_\varrho(x,t)_{|t=T} &= 0 && \text{for } x \in \Omega. \end{aligned} \tag{4.150}$$

We end up with the optimality system, including the forward equation (4.130)*, the adjoint equation* (4.150) *and the gradient equation* (4.149)*. Eliminating the control*

$u_\varrho = \Box y_\varrho$, from (4.130), *we get by the gradient equation that $p_\varrho + \varrho \Box y_\varrho = 0$ and we can phrase the variational formulation: find $(p_\varrho, y_\varrho) \in H^{1,1}_{0;,0}(Q) \times H^{1,1}_{0;0,}(Q)$ such that*

$$\begin{array}{rcccll} \varrho^{-1}\langle p_\varrho, q\rangle_{L^2(Q)} & + & b(y_\varrho, q) & = & 0 & \text{for all } q \in H^{1,1}_{0;,0}(Q) \\ -b(z, p_\varrho) & + & \langle y_\varrho, z\rangle_{L^2(Q)} & = & \langle y_d, z\rangle_{L^2(Q)}, & \text{for all } z \in H^{1,1}_{0;0,}(Q). \end{array} \tag{4.151}$$

Moreover, we can eliminate the adjoint variable $p_\varrho = -\varrho u_\varrho = -\varrho \Box y_\varrho$ in the adjoint equation (4.150) *to conclude*

$$\varrho \Box^2 y_\varrho = \Box p_\varrho = y_d - y_\varrho$$

and therefore

$$\begin{array}{rcrcll} & & \varrho \Box^2 y_\varrho(x,t) + y_\varrho(x,t) & = & y_d & \text{for } (x,t) \in Q, \\ y_\varrho(x,t) & = & \Box y_\varrho(x,t) & = & 0 & \text{for } (x,t) \in \Sigma, \\ y_\varrho(x,t) & = & \partial_t y_\varrho(x,t)_{|t=0} & = & 0 & \text{for } x \in \Omega, \\ \Box y_\varrho(x,T) & = & \partial_t \Box y_\varrho(x,t)_{|t=T} & = & 0 & \text{for } x \in \Omega, \end{array} \tag{4.152}$$

which leads to a kind of Bi-wave equation with boundary conditions inherited from the adjoint state p_ϱ. Although, we will not give a complete analysis for this approach, we can discretize the variational formulation (4.151) *and will numerically compare the solutions of the energy regularization to solutions of the L^2-regularization.*

Discretization

We consider a Galerkin finite element discretization of the variational formulation (4.139), with the conforming ansatz space $Y_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;0,}(Q) = \operatorname{span}\{\varphi_k^1\}_{k=1}^{M_Y} \subset \mathcal{H}_{0;0,}(Q)$ of piecewise linear, globally continuous functions, defined with respect to some admissible, globally quasi-uniform decomposition $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$ of the space-time domain Q into shape regular, simplicial finite elements τ_ℓ of mesh size h_ℓ. Then we have to find $y_{\varrho h} \in Y_h$ such that

$$\varrho \langle S y_{\varrho h}, z_h\rangle_Q + \langle y_{\varrho h}, z_h\rangle_Q = \langle y_d, z_h\rangle, \quad \text{for all } z_h \in Y_h. \tag{4.153}$$

Note, that at this point the action of the operator $S = \widetilde{B}^* A^{-1} \widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [\mathcal{H}_{0;0,}(Q)]^*$ can not be computed explicitly. To actually derive a numerical scheme, we replace the operator by a computable approximation $\tilde{S} : \mathcal{H}_{0;0,}(Q) \to [\mathcal{H}_{0;0,}(Q)]^*$, which action is defined by $\tilde{S} y = \widetilde{B}^* p_{yh}$, where $p_{yh} \in X_h := S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;,0}(Q) \subset H^{1,1}_{0;,0}(Q)$ is the unique solution of

$$\langle A p_{yh}, q_h\rangle_Q = \langle \widetilde{B} y, q_h\rangle_Q \quad \text{for all } q_h \in X_h. \tag{4.154}$$

Then we consider the perturbed system to find $\widetilde{y}_{\varrho h} \in Y_h$ such that

$$\varrho\langle\tilde{S}\tilde{y}_{\varrho h}, z_h\rangle_Q + \langle\tilde{y}_{\varrho h}, z_h\rangle_Q = \langle y_d, z_h\rangle, \quad \text{for all } z_h \in Y_h. \tag{4.155}$$

By Lemma 3.14 the variational formulation (4.155) admits a unique solution $\tilde{y}_{\varrho h} \in Y_h$, as $\tilde{S} \geq 0$ is positive semi-definite. Moreover, for any arbitrary but fixed $z_h \in Y_h \subset H^{1,1}_{0;0,}(Q)$, we have by (4.133) and Lemma 2.37 that there holds the inverse inequality

$$\|z_h\|_{\mathcal{H}_{0;0,}(Q)} \leq \|z_h\|_{H^{1,1}_{0;0,}(Q)} = \|\nabla_{(x,t)} z_h\|_{L^2(Q)} \leq c_I h^{-1}\|z_h\|_{L^2(Q)}. \tag{4.156}$$

Thus we can apply Theorem 3.15 to get the following finite element error estimates, depending on the regularization parameter $\varrho > 0$, the regularity of the target y_d and the approximation property of the trial spaces.

LEMMA 4.38. *Let $y_d \in L^2(Q)$. Then the unique solution $\tilde{y}_{\varrho h} \in Y_h$ of (4.155) admits the estimate*

$$\|\tilde{y}_{\varrho h} - y_d\|_{L^2(Q)} \leq \|y_d\|_{L^2(Q)}. \tag{4.157}$$

Let $y_d \in H^{1,1}_{0;0,}(Q)$ and let $p_{y_d} \in H^{1,1}_{0;,0}(Q)$ be the unique solution of

$$\langle Ap_{y_d}, q\rangle_Q = \langle By_d, q\rangle_Q \quad \text{for all } q \in H^{1,1}_{0;,0}(Q). \tag{4.158}$$

Then we get

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_d\|_{L^2(Q)} \leq c\Big(&\big[h^{-1}\varrho + \sqrt{\varrho}\big]\|y_d\|_{H^{1,1}_{0;0,}(Q)} + h^{-1}\varrho \inf_{q_h\in X_h}\|\nabla_{(x,t)}(p_{y_d} - q_h)\|_{L^2(Q)} \\ &+\big[h^{-1}\sqrt{\varrho}+1\big]\inf_{z_h\in Y_h}\big[\varrho\|\nabla_{(x,t)}(y_d - z_h)\|^2_{L^2(Q)} + \|y_d - z_h\|^2_{L^2(Q)}\big]^{1/2}\Big).\end{aligned} \tag{4.159}$$

Moreover, if $y_d \in H^2(Q) \cap H^{1,1}_{0;0,}(Q)$ and $Sy_d \in L^2(Q)$ we have the error estimate

$$\begin{aligned}\|\tilde{y}_{\varrho h} - y_d\|_{L^2(Q)} \leq c\Big(&\big[h^{-1}\varrho^{3/2} + \varrho\big]\|y_d\|_{H^2(Q)} + h^{-1}\varrho \inf_{q_h\in X_h}\|\nabla_{(x,t)}(p_{y_d} - q_h)\|_{L^2(Q)} \\ &+\big[h^{-1}\sqrt{\varrho}+1\big]\inf_{z_h\in Y_h}\big[\varrho\|\nabla_{(x,t)}(y_d - z_h)\|^2_{L^2(Q)} + \|y_d - z_h\|^2_{L^2(Q)}\big]^{1/2}\Big).\end{aligned} \tag{4.160}$$

Recall, that $A : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ corresponds to the space-time Laplacian with mixed boundary conditions and therefore the unique solution of (4.158) fulfills $p_{y_d} \in H^r(Q) \cap H^{1,1}_{0;,0}(Q)$ for some $0 \leq r \leq 1$ depending on the regularity of y_d and the geometry of the space-time domain. In the following let us restrict to a convex space-time domain Q and to functions $y_d \in H^s(Q) \cap H^{1,1}_{0;0,}(Q)$ such that $p_{y_d} \in H^s(Q) \cap H^{1,1}_{0;,0}(Q)$ for $s \in [1, 2]$. This can be achieved in practice if the initial and terminal conditions do not play a major role for y_d, e.g., when $y_d \equiv 0$ on the lateral boundary $\Sigma = \partial\Omega \times (0,T)$ and on $\Omega \times (0,\varepsilon)$ and $\Omega \times (T-\varepsilon, T)$ for some $\varepsilon > 0$. Then, for $s = 2$ the estimate

$$\|p_{y_d}\|_{H^2(Q)} \leq c\|Ap_{y_d}\|_{L^2(Q)} = c\|By_d\|_{L^2(Q)} \leq c\|y_d\|_{H^2(Q)} \tag{4.161}$$

holds true. The main statement of this section is then, the following error estimate.

THEOREM 4.39 ([87, Corollary 4.4]). *Let $\tilde{y}_{\varrho h} \in Y_h$ be the unique solution of the variational formulation* (4.155) *and let $y_d \in H^{s,s}_{0;0,}(Q) := [L^2(Q), H^{1,1}_{0;0,}(Q)]_s$ for $s \in [0,1]$ or $y_d \in H^s(Q) \cap H^{1,1}_{0;0,}(Q)$ such that $Sy_d \in H^{s-2}(Q)$ and $p_{y_d} \in H^s(Q) \cap H^{1,1}_{0;,0}(Q)$ for $s \in [1,2]$. If $\varrho = h^2$, then*

$$\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)} \leq ch^s\|y_d\|_{H^s(Q)} \quad s \in [0,2].$$

Proof. Firstly, for $y_d \in H^{1,1}_{0;0,}(Q)$, by the best approximation properties (Theorem 2.36) and using $\varrho = h^2$, we can estimate

$$\inf_{z_h \in Y_h} \left[h^2\|\nabla_{(x,t)}(y_d - z_h)\|^2_{L^2(Q)} + \|y_d - z_h\|^2_{L^2(Q)}\right]^{1/2} \leq ch\|y_d\|_{H^1(Q)},$$

and

$$\inf_{q_h \in X_h} \|\nabla_{(x,t)}(p_{y_d} - q_h)\|_{L^2(Q)} \leq \|\nabla_{(x,t)}p_{y_d}\|_{L^2(Q)}.$$

Moreover, we can compute that

$$\|\nabla_{(x,t)}p_{y_d}\|^2_{L^2(Q)} = \langle Ap_{y_d}, p_{y_d}\rangle_Q = \langle By_d, p_{y_d}\rangle_Q \leq \|\nabla_{(x,t)}y_d\|_{L^2(Q)}\|\nabla_{(x,t)}p_{y_d}\|_{L^2(Q)},$$

i.e., $\|\nabla_{(x,t)}p_{y_d}\|_{L^2(Q)} \leq \|\nabla_{(x,t)}y_d\|_{L^2(Q)}$. So estimate (4.159) becomes

$$\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)} \leq ch\|y_d\|_{H^1(Q)}, \tag{4.162}$$

and interpolating the estimate with (4.157), using Theorem 2.14, gives

$$\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)} \leq ch^s\|y_d\|_{H^s(Q)}, \quad s \in [0,1].$$

Now, let $y_d \in H^2(Q) \cap H^{1,1}_{0;0,}(Q)$ such that $Sy_d \in L^2(Q)$ and $p_{y_d} \in H^2(Q) \cap H^{1,1}_{0;,0}(Q)$. As above we can estimate

$$\inf_{z_h \in Y_h} \left[h^2\|\nabla_{(x,t)}(y_d - z_h)\|^2_{L^2(Q)} + \|y_d - z_h\|^2_{L^2(Q)}\right]^{1/2} \leq ch^2\|y_d\|_{H^2(Q)},$$

and, using (4.161),

$$\inf_{q_h \in X_h} \|\nabla_{(x,t)}(p_{y_d} - q_h)\|_{L^2(Q)} \leq ch\|p_{y_d}\|_{H^2(Q)} \leq ch\|y_d\|_{H^2(Q)}.$$

Thus, (4.160) can be bounded by

$$\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)} \leq ch^2\|y_d\|_{H^2(Q)}$$

and interpolating with (4.162) gives

$$\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)} \leq ch^s\|y_d\|_{H^s(Q)}, \quad s \in [1,2]. \qquad \square$$

Numerical results

Recall the finite element spaces used for the Galerkin variational formulation of (4.155) are

$$Y_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;0,}(Q) = \operatorname{span}\{\varphi_k^Y\}_{k=1}^{M_Y}$$

and

$$X_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;,0}(Q) = \operatorname{span}\{\varphi_k^X\}_{k=1}^{M_X}.$$

Furthermore, by Remark 3.13, we know that the solution of the perturbed variational formulation (4.155) is exactly the solution of the system

$$\begin{aligned}
\varrho^{-1}\langle A p_{\varrho h}, q_h\rangle_Q \quad & +\langle B\tilde{y}_{\varrho_h}, q_h\rangle_Q && =0 && \text{for all } q_h \in X_h \\
-\langle B^* p_{\varrho h}, z_h\rangle_Q \quad & +\langle \tilde{y}_{\varrho h}, z_h\rangle_{L^2(Q)} && =\langle y_d, z_h\rangle_{L^2(Q)} && \text{for all } z_h \in Y_h.
\end{aligned} \tag{4.163}$$

Thus, using the fe-isomorphism, we have to solve the equivalent system of linear equations

$$\begin{pmatrix} \varrho^{-1} A_h & B_h \\ -B_h^\top & M_h \end{pmatrix} \begin{pmatrix} \mathbf{p}_{\varrho h} \\ \tilde{\mathbf{y}}_{\varrho h} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_h \\ \mathbf{y}_{dh} \end{pmatrix}, \tag{4.164}$$

where the matrices are given as

$$\begin{aligned}
A_h[i,j] &= \langle \nabla_{(x,t)}\varphi_j^X, \nabla_{(x,t)}\varphi_i^X\rangle_{L^2(Q)}, \\
B_h[i,k] &= -\langle \partial_t\varphi_k^Y, \partial_t\varphi_i^X\rangle_{L^2(Q)} + \langle \nabla_x\varphi_k^Y, \nabla_x\varphi_i^X\rangle_{L^2(Q)}, \\
M_h[\ell,k] &= \langle \varphi_k^Y, \varphi_\ell^Y\rangle_{L^2(Q)},
\end{aligned}$$

for $i,j = 1,\ldots,M_X$ and $k,\ell = 1,\ldots,M_Y$ and the entries of the load vector are

$$\mathbf{y}_{dh}[k] = \langle y_d, \varphi_k^Y\rangle_{L^2(Q)}, \quad k = 1,\ldots,M_Y.$$

Remark 4.40. *Although we do not carry out a thorough analysis, we will compare our results for the energy regularization in* $[H^{1,1}_{0;,0}(Q)]^*$ *with the ones stemming from an* L^2*-regularization given by the variational formulation* (4.151). *The discretized system in this case is given by*

$$\begin{pmatrix} \varrho^{-1} \bar{M}_h & B_h \\ -B_h^\top & M_h \end{pmatrix} \begin{pmatrix} \boldsymbol{p}_{\varrho h} \\ \tilde{\boldsymbol{y}}_{\varrho h} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0}_h \\ \boldsymbol{y}_{dh}, \end{pmatrix} \tag{4.165}$$

where the only difference is the matrix

$$\bar{M}_h[i,j] = \langle \varphi_j^X, \varphi_i^X\rangle_{L^2(Q)}, \quad i,j = 1,\ldots,M_X,$$

realizing the norm in $L^2(Q)$. *As for the optimal control problem subject to the Poisson equation, see Remark 4.9 and Theorem 4.27, the optimal choice of the relaxation parameter in this case is* $\varrho = h^4$.

For our numerical examples, we consider the space time domain $Q = (0, L) \times (0, T) \subset \mathbb{R}^{1+1}$ for $L = T = 1$ and three targets of different regularity, see Figure 4.17. Namely, $y_{d,1} \in \mathcal{C}^2(\overline{Q}) \cap H_0^1(Q)$ defined as

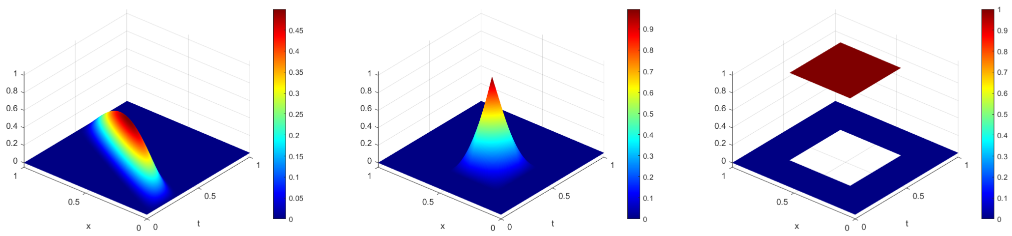
$$y_{d,1}(x,t) = \begin{cases} \frac{1}{2}(6t - 3x - 2)^3(3x - 6t)^3 \sin(\pi x), & x \le 2t \text{ and } 6t - 3x \le 2, \\ 0, & \text{else}, \end{cases} \tag{4.166}$$

the piecewise bilinear function $y_{d,2} \in H^{3/2-\varepsilon}(Q) \cap H_0^1(Q)$, $\varepsilon > 0$, defined as

$$y_{d,2}(x,t) = \phi(x)\phi(t), \quad \phi(s) = \begin{cases} 1, & s = 0.45, \\ 0, & s \notin [0.2, 0.6], \\ \text{linear}, & \text{else}, \end{cases} \tag{4.167}$$

and finally, the discontinuous target $y_{d,3} \in H_0^{1/2-\varepsilon}(Q)$, $\varepsilon > 0$, defined as

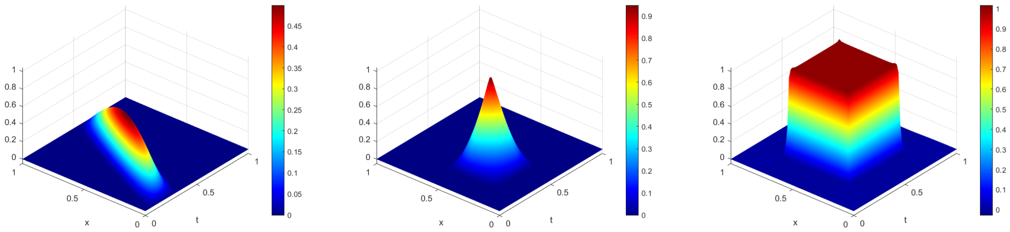
$$y_{d,3}(x,t) = \begin{cases} 1, & (x,t) \in (0.25, 0.75)^2 \subset Q, \\ 0, & \text{else}. \end{cases} \tag{4.168}$$



a) $y_{d,1}$ b) $y_{d,2}$ c) $y_{d,3}$

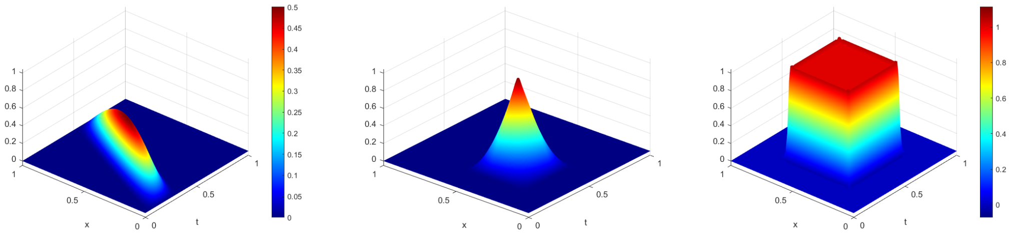
Figure 4.17: Target functions $y_{d,i}$, $i = 1, 2, 3$.

The convergence rates for a uniform refinement are depicted in Figure 4.19 for an initial triangulation with $N = 128$ elements and $M = 56$ degrees of freedom (DoFs) and for all three targets for both, the energy regularization in $[H^{1,1}_{0;,0}(Q)]^*$ solving (4.164) and the L^2-regularization solving (4.165). The behavior is similar to the one observed for the optimal control problem subject to the Poisson equation, see Figure 4.3. Firstly, for a fixed parameter $\varrho > 0$, we clearly see optimal convergence rates at first, which break down when $h = \varrho^{1/2}$ in the case of the energy regularization and $h = \varrho^{1/4}$ in the case of the L^2-regularization, but independent of the regularity of the target. This is agreement with the estimates in Lemma 4.38. Secondly, Figure 4.19 shows the convergence for the optimal choice $\varrho = h^2$ for the energy regularization and $\varrho = h^4$ for the L^2-regularization. We see optimal convergence for all three targets, as



a) $y_{\varrho h,1}$ b) $y_{\varrho h,2}$ c) $y_{\varrho h,3}$

Energy regularization (4.164) with $\varrho = h^2$



d) $y_{\varrho h,1}$ e) $y_{\varrho h,2}$ f) $y_{\varrho h,3}$

Common L^2-regularization (4.165) with $\varrho = h^4$

Figure 4.18: Reconstructed target functions $y_{\varrho h,i}$, $i = 1,2,3$, on a mesh with $N = 32768$ elements and $M = 16256$ DoFs.

predicted in Theorem 4.39. Moreover, the reconstructed targets in Figure 4.2 reveal a qualitative different behavior for the discontinuous target, as observed in the case of the Poisson equation with energy regularization in $H^{-1}(\Omega)$ in Section 4.1.1. While for the common L^2-regularization one observes oscialltive behavior around the jump, the energy regularization gives sharp results. Again this stems from imposing higher regularity on the state when measuring the control in $L^2(Q)$.

Furthermore, we want to check how the cost functional

$$\tilde{\mathcal{J}}(\tilde{y}_{\varrho h}) = \frac{1}{2}\|y_d - \tilde{y}_{\varrho h}\|^2_{L^2(Q)} + \frac{\varrho}{2}\|\widetilde{B}\tilde{y}_{\varrho h}\|^2_{[H^{1,1}_{0;,0}(Q)]^*}$$

behaves. Using the norm equivalence (4.140) and (4.133), we compute

$$\|\widetilde{B}\tilde{y}_{\varrho h}\|_{[H^{1,1}_{0;,0}(Q)]^*} = \|\tilde{y}_{\varrho h}\|_S \le \|\tilde{y}_{\varrho h}\|_{\mathcal{H}_{0;0,}(Q)} \le \|\tilde{y}_{\varrho h}\|_{H^{1,1}_{0;0,}(Q)} = \|\nabla_{(x,t)}\tilde{y}_{\varrho h}\|_{L^2(Q)}$$

and we get the upper estimate

$$\overline{\mathcal{J}}(\tilde{y}_{\varrho h}) = \frac{1}{2}\|y_d - \tilde{y}_{\varrho h}\|^2_{L^2(Q)} + \frac{\varrho}{2}\|\nabla_{(x,t)}\tilde{y}_{\varrho h}\|^2_{L^2(Q)}.$$

Figure 4.20 shows the convergence of the cost functional for a fixed parameter $\varrho = 10^{-8}$ and the optimal choice $\varrho = h^2$. We clearly see, that for a fixed parameter the convergence for all three targets is optimal up to the point where $h^4 \sim \varrho$, while for $\varrho = h^2$ we only see the optimal rate for the target $y_{d,3} \in H^{1/2-\varepsilon}(Q)$ and a quadratic rate for the other two. This is similar to the results for the Poisson equation, see Lemma 4.7 and Remark 4.8 and indicates that also in the case of the wave equation the energy regularization in $[H^{1,1}_{0;,0}(Q)]^*$ is especially well-suited for targets of low regularity, i.e., $y_d \in H^{s,s}_{0;0,}(Q)$ for $s \in [0,1]$.

As mentioned, since $A := -\Delta_{(x,t)} : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ corresponds to the space-time Laplacian with mixed boundary conditions, this has an effect on the regularity. Namely, even if $y_d \in H^2(Q) \cap H^{1,1}_{0;0,}(Q)$ we can not expect $Sy_d = \widetilde{B}A^{-1}\widetilde{B}y_d \in L^2(Q)$. This is observed when considering the smooth target,

$$y_{d,4}(x,t) = t\sin(t\pi)\sin(x\pi), \quad (x,t) \in Q.$$

As Table 4.2 shows we only see convergence of order $\mathcal{O}(h^{1.5})$ for the optimal choice $\varrho = h^2$. To remedy this behavior, one might either consider to choose $\varrho = h^3$, or change to measure the control in L^2, as there boundary conditions do not play a role. Moreover, an adaptive refinement scheme, discussed in the next section, helps to overcome this issue.

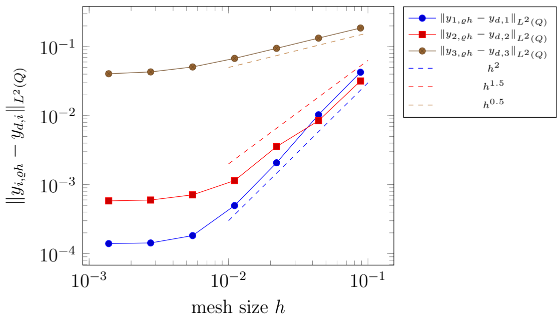
Level	DoFs	N	h	$\varrho\,(h^2)$	$\lVert\tilde{y}_{4,\varrho h} - y_{d,4}\rVert_{L^2(Q)}$	eoc
0	56	128	0.088	$7.81 \cdot 10^{-3}$	$1.31 \cdot 10^{-2}$	0.00
1	240	512	0.044	$1.95 \cdot 10^{-3}$	$4.52 \cdot 10^{-3}$	1.53
2	992	2,048	0.022	$4.88 \cdot 10^{-4}$	$1.62 \cdot 10^{-3}$	1.48
3	4,032	8,192	0.011	$1.22 \cdot 10^{-4}$	$5.82 \cdot 10^{-4}$	1.48
4	16,256	32,768	0.006	$3.05 \cdot 10^{-5}$	$2.08 \cdot 10^{-4}$	1.49
5	65,280	131,072	0.003	$7.63 \cdot 10^{-6}$	$7.39 \cdot 10^{-5}$	1.49
6	261,632	524,288	0.001	$1.91 \cdot 10^{-6}$	$2.62 \cdot 10^{-5}$	1.50
7	1,047,552	2,097,152	0.001	$4.77 \cdot 10^{-7}$	$9.28 \cdot 10^{-6}$	1.50

Table 4.2: Errors and orders of convergence for $y_{d,4}$ in the case of an uniform refinement strategy with $\varrho = h^2$.

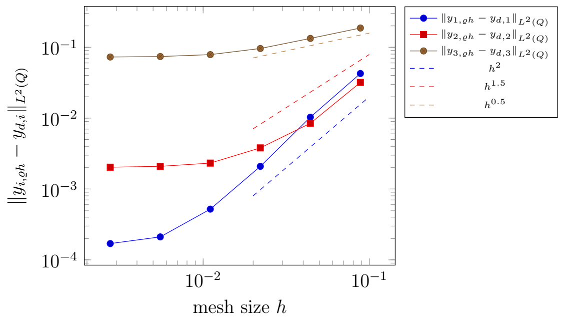
Reconstruction of the control

As in all the applications discussed before, we now want to reconstruct a discrete approximation of the control $\tilde{u}_{\varrho H} \in U_H \subset [H^{1,1}_{0;,0}(Q)]^*$ from the computed state $\tilde{y}_{\varrho h} \in Y_h$. Since, in general the control $u_\varrho \in [H^{1,1}_{0;,0}(Q)]^*$ is discontinuous, we consider

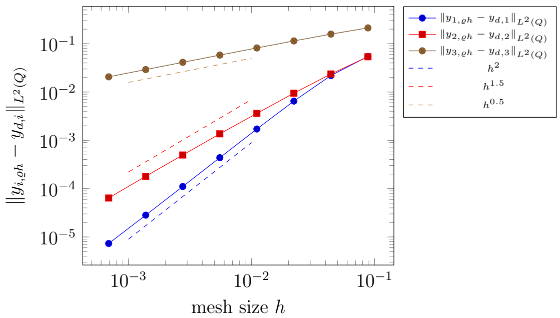
$$U_H = S^0_H(\mathcal{T}_H) = \operatorname{span}\{\varphi^0_\ell\}_{\ell=1}^{N_H},$$



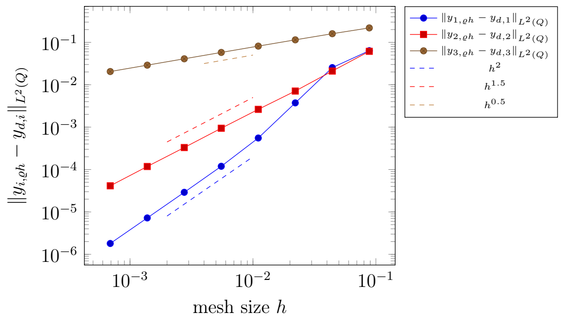
a) Energy regularization, $\varrho = 10^{-5}$



b) L^2-regularization, $\varrho = 10^{-8}$

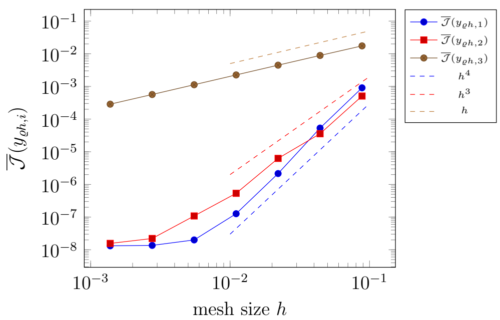


c) Energy regularization, $\varrho = h^2$

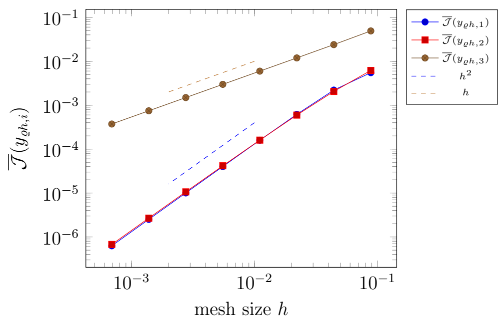


d) L^2-regularization, $\varrho = h^4$

Figure 4.19: Convergence for the three different target functions $y_{d,i}$, $i = 1, 2, 3$ for the energy regularization in $[H^{1,1}_{0;,0}(Q)]^*$ solving (4.164) and the L^2- regularization solving (4.165) for different choices of $\varrho > 0$.



a) $\varrho = 10^{-8}$



b) $\varrho = h^2$

Figure 4.20: Convergence of the cost functional $\overline{\mathcal{J}}$.

with respect to some coarser, nested decomposition $\mathcal{T}_H$, and consider the problem to find $\tilde{u}_{\varrho H} \in U_H$ as the minimizer of

$$\begin{aligned}\tilde{u}_{\varrho H} &= \arg\min_{v_H \in U_H} \frac{1}{2}\|v_H - B\tilde{y}_{\varrho h}\|^2_{[H^{1,1}_{0;,0}(Q)]^*} \\ &= \arg\min_{v_H \in U_H} \frac{1}{2}\langle v_H - B\tilde{y}_{\varrho h}, A^{-1}(v_H - B\tilde{y}_{\varrho h})\rangle_{[H^{1,1}_{0;,0}(Q)]^*}.\end{aligned}$$

This is equivalent to the saddle point formulation (3.41), i.e., to find $(\hat{p}_h, \tilde{u}_{\varrho H}) \in X_h \times U_H$ such that

$$\langle A\hat{p}_h, q_h\rangle_Q + \langle \tilde{u}_{\varrho H}, q_h\rangle_{L^2(Q)} = \langle B\tilde{y}_{\varrho h}, q_h\rangle_Q, \quad \langle v_H, \hat{p}_h\rangle_{L^2(Q)} = 0,$$

for all $(q_h, v_H) \in X_h \times U_H$. In terms of matrices this can be computed by solving the system

$$\begin{pmatrix} A_h & \hat{M}_h \\ \hat{M}_h^\top & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{p}}_h \\ \tilde{\mathbf{u}}_{\varrho H} \end{pmatrix} = \begin{pmatrix} B_h \tilde{\mathbf{y}}_{\varrho h} \\ \mathbf{0}_H \end{pmatrix}, \tag{4.169}$$

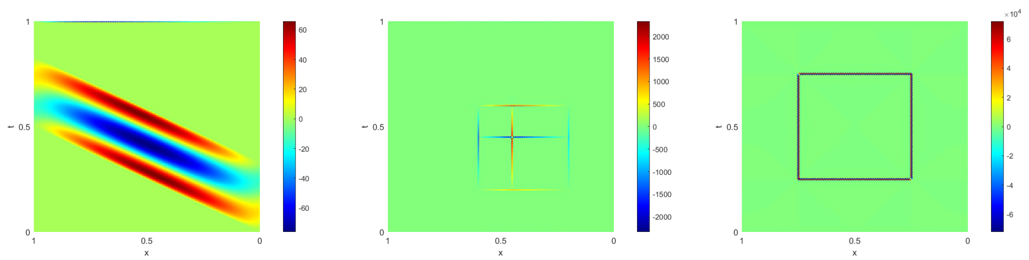
where A_h and B_h are as in (4.164) and

$$\hat{M}_h[i,\ell] = \langle \varphi^0_{H,\ell}, \varphi^1_{h,i}\rangle_{L^2(Q)}, \quad i = 1, \ldots, M_X, \ell = 1, \ldots, N_H.$$

Unique solvability and related error estimates follow from the abstract framework by Theorem 3.20, if the discrete inf-sup stability

$$c_S \|v_H\|_{[H^{1,1}_{0;,0}(Q)]^*} \leq \sup_{0 \neq q_h \in X_h} \frac{\langle v_H, q_h\rangle_{L^2(Q)}}{\|\nabla_{(x,t)} q_h\|_{L^2(Q)}} \quad \text{for all } v_H \in U_H,$$

is satisfied with some uniform constant $c_S > 0$. This can be achieved by the choice $h = H/4$, i.e., $\mathcal{T}_h$ is produced by twice uniformly refining $\mathcal{T}_H$. The reconstructed controls for the three different targets are depicted in Figure 4.21.



a) $\tilde{u}_{\varrho H,1}$ b) $\tilde{u}_{\varrho H,2}$ c) $\tilde{u}_{\varrho H,3}$

Figure 4.21: Reconstructed controls $\tilde{u}_{\varrho H,i} \in U_H$ computed by solving (4.169) on a mesh with $N = 32768$ elements and $M = 16256$ DoFs.

4.2.2 Adaptive refinement

The applicability of adaptive schemes is of interest for many reason. To mention some, note, that we observed the existence of smooth targets for which we do not see optimal orders of convergence, with the optimal choice $\varrho = h^2$, see Table 4.2. We will see that the adaptive scheme will regain optimal rates in this case and even raise the order of convergence for the less smooth targets $y_{d,2}$ and $y_{d,3}$. Moreover, it reduces the computational effort, as less elements are needed, to compute a solution of the same accuracy. To drive an adaptive scheme, note that for given $y_d \in L^2(Q)$, we can compute the error $\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)}$, where $\tilde{y}_{\varrho h} \in Y_h$ is the unique solution of (4.155). Thus, we can define the local error indicators on each element $\tau_\ell \in \mathcal{T}_h$ by

$$\tilde{\eta}_\ell := \|y_d - \tilde{y}_{\varrho h}\|_{L^2(\tau_\ell)}, \quad \ell = 1, \dots, N,$$

which satisfy

$$\|y_d - \tilde{y}_{\varrho h}\|^2_{L^2(Q)} = \sum_{\ell=1}^{N} \tilde{\eta}_\ell^2,$$

and we can drive an adaptive scheme using Dörfler marking [34], i.e., we refine all elements $\tau_\ell \in \mathcal{T}_h$ that satisfy

$$\tilde{\eta}_\ell > \theta \max_{i=1,\dots,N} \tilde{\eta}_i, \quad \text{for some } \theta > 0.$$

As in the case of adaptive refinement for the Poisson equation in Section 4.1.2, the question about the optimal choice of the regularization parameter ϱ arises, as adaptively refined meshes get heavily non-uniform. One obvious choice is $\varrho = h^2_{\min}$, but, especially for targets in $[H^{1,1}_{0;,0}(Q)]^*$ we want to keep the regularization parameter as large as possible, to have a suitable stabilization of the problem. Therefore, we also consider the choice of an adaptive parameter $\varrho(x,t) = h_\ell^2$ for $\ell = 1, \dots, N$. Without a rigorous analysis, we just give some numerical examples by solving

$$\begin{pmatrix} A_{\varrho^{-1}h} & B_h \\ -B_h^\top & M_h \end{pmatrix} \begin{pmatrix} \mathbf{p}_{\varrho h} \\ \tilde{\mathbf{y}}_{\varrho h} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_h \\ \mathbf{y}_{dh} \end{pmatrix}, \tag{4.170}$$

with B_h and M_h as in (4.164) and

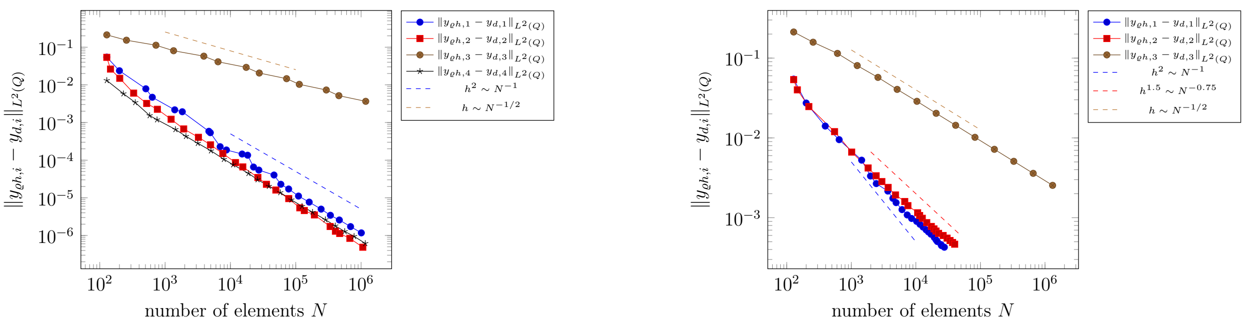
$$A_{\varrho^{-1}h}[i,j] = \int_0^T \int_\Omega \varrho(x,t)^{-1} \nabla_{(x,t)} \varphi_j^X(x,t) \cdot \nabla_{(x,t)} \varphi_i^X(x,t) \, dx \, dt.$$

The results for the different targets are depicted in Figure 4.22. We see a similar behavior as in the case of the adaptive refinement for the Poisson equation. Especially, for the continuous targets $y_{d,i}$, $i = 1, 2, 4$ we gain quadratic orders of convergence, when choosing $\varrho = h^2_{\min}$, while the choice $\varrho(x,t) = h_\ell^2$ seems again especially well-suited for the discontinuous target $y_{d,2}$. The resulting adaptive meshes are depicted in Figure 4.23.

Remark 4.41. *It is worth noting, that the solution of the distributed optimal control problem subject to the wave equation does not require a so-called Courant–Friedrichs–Lewy–condition (CFL-condition), dating back to [28], which is usually observed in the direct solution of the wave equation and will be adressed in Chapter 5 in more detail. For a uniform tensor product mesh of the space time domain $Q = (0,1)^d \times (0,T)$, the condition can explicitly be computed to be*

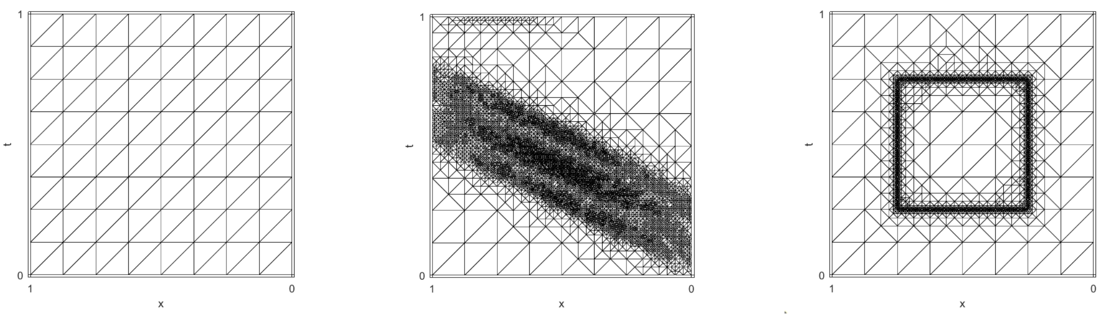
$$h_t \leq d^{-1/2} h_x,$$

where h_t and h_x denotes temporal and spatial mesh sizes, respectively, see [110, p. 49].



a) $\varrho = h_{\min}^2$ b) $\varrho(x) = h_\ell^2$, $x \in \tau_\ell$

Figure 4.22: Convergence rates for the adaptive scheme solving (4.170) with $\theta = 0.5$ for different choices of ϱ.



a) Initial mesh $N = 128$ b) $y_{d,1}$, $N = 15095$ c) $y_{d,3}$, $N = 5130$

Figure 4.23: Initial mesh and adaptively refined meshes for the target functions $y_{d,1}$ and $y_{d,3}$.

4.2.3 State and control constraints

In this section we will discuss the incorporation of state or control constraints into the optimal control problem subject to the wave equation. Recall, that the abstract theory applies to the unconstrained problem (4.129)-(4.130), when choosing

$$H = L^2(Q), \quad X = H^{1,1}_{0;,0}(Q), \quad Y = \mathcal{H}_{0;0,}(Q),$$

and the operators

$$A = -\Delta_{(x,t)} : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^* \quad \text{and} \quad \widetilde{B} = \mathcal{E}^* \widetilde{\Box(\cdot)} : \mathcal{H}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*.$$

State constraints

The problem admitting state constraints is then to minimize the cost functional

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(Q)} + \frac{\varrho}{2}\|u_\varrho\|^2_{[H^{1,1}_{0;,0}(Q)]^*}$$

subject to the wave equation

$$\begin{aligned} \Box y_\varrho(x,t) := \partial_{tt} y_\varrho(x,t) - \Delta_x y_\varrho(x,t) &= u_\varrho(x,t) && \text{for } (x,t) \in Q, \\ y_\varrho(x,t) &= 0 && \text{for } (x,t) \in \Sigma, \\ y_\varrho(x,0) = \partial_t y_\varrho(x,t)_{|t=0} &= 0 && \text{for } x \in \Omega, \end{aligned}$$

and to the state constraints

$$g_-(x,t) \le y_\varrho(x,t) \le g_+(x,t) \quad \text{for } (x,t) \in Q,$$

for a given target $y_d \in L^2(Q)$ and barrier functions $g_\pm : Q \to \mathbb{R}$, which fulfill $g_-(x,t) \le 0 \le g_+(x,t)$ for a.a. $(x,t) \in Q$ and $\widetilde{B}^* A^{-1} \widetilde{B} g_\pm \in L^2(Q)$. We already established that $\widetilde{B} : Y \to X^*$ is an isomorphism, and thus, considering the reduced cost functional

$$\tilde{\mathcal{J}}(y_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(Q)} + \frac{\varrho}{2}\|\widetilde{B} y_\varrho\|^2_{[H^{1,1}_{0;,0}(Q)]^*},$$

we want to find

$$y_\varrho \in K_s := \{z \in \mathcal{H}_{0;0,}(Q) : \, g_-(x,t) \le z(x,t) \le g_+(x,t), \text{ for a.a. } (x,t) \in Q\}$$

such that

$$\tilde{\mathcal{J}}(y_\varrho) \le \tilde{\mathcal{J}}(z), \quad \text{for all } z \in K_s.$$

This is exactly (3.47) and the minimizer is characterized as the unique solution of the variational inequality (3.48), i.e., we want to find $y_\varrho \in K_s$ such that

$$\varrho \langle S y_\varrho, z - y_\varrho \rangle_Q + \langle y_\varrho, z - y_\varrho \rangle_{L^2(Q)} \ge \langle y_d, z - y_\varrho \rangle_{L^2(Q)} \quad \text{for all } z \in K_s, \tag{4.171}$$

where $S := \widetilde{B}^* A^{-1} \widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [\mathcal{H}_{0;0,}(Q)]^*$. Thus, with Lemma 3.23 and the norm equivalence $\|y\|_S = \|y\|_{\mathcal{H}_{0;0,}(Q)}$, see (4.140), we immediately get the following regularization error estimates.

LEMMA 4.42. *Let $y_d \in L^2(Q)$ be given. For the unique solution $y_\varrho \in K_s$ of* (4.171) *there holds*

$$\|y_\varrho - y_d\|_{L^2(Q)} \le \|y_d\|_{L^2(Q)}. \tag{4.172}$$

Further, if $y_d \in K_s$, then

$$\|y_\varrho - y_d\|_{L^2(Q)} \le \sqrt{\varrho}\|y_d\|_{\mathcal{H}_{0;0,}(Q)} \quad \textit{and} \quad \|y_\varrho - y_d\|_{\mathcal{H}_{0;0,}(Q)} \le \|y_d\|_{\mathcal{H}_{0;0,}(Q)}. \tag{4.173}$$

If in addition $Sy_d \in L^2(Q)$ it holds

$$\|y_\varrho - y_d\|_{L^2(Q)} \le \varrho\|Sy_d\|_{L^2(Q)} \quad \textit{and} \quad \|y_\varrho - y_d\|_{\mathcal{H}_{0;0,}(Q)} \le \sqrt{\varrho}\|Sy_d\|_{L^2(Q)}. \tag{4.174}$$

Now let us derive complementarity conditions by introducing the auxiliary variable $\lambda := \varrho S y_\varrho + y_\varrho - y_d \in [\mathcal{H}_{0;0,}(Q)]^*$, which satisfies, see (4.171)

$$\langle \lambda, z - y_\varrho \rangle_{L^2(\Omega)} \ge 0, \quad \text{for all } z \in K_s.$$

By Lemma 3.24 the unique solution $y_\varrho \in K_s$ satisfies $Sy_\varrho \in L^2(Q)$, which implies $\lambda \in L^2(Q)$ and pointwise evaluation is a.e. well-defined. With the sets

$$Q_{s,\pm} := \{(x,t) \in Q : \, y_\varrho(x,t) = g_\pm(x,t)\}.$$

by (3.62), the following complementarity conditions hold true:

$$\begin{array}{lrl} \lambda = 0, & g_- < y_\varrho < g_+, & \text{on } Q \setminus Q_{s,\pm}, \\ \lambda \ge 0, & y_\varrho = g_-, & \text{on } Q_{s,-}, \\ \lambda \le 0, & y_\varrho = g_+, & \text{on } Q_{s,+}. \end{array} \tag{4.175}$$

Discretization

As in the case without constraints, we consider an admissible, globally quasi-uniform decomposition $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$ of the space-time domain Q into shape regular, simplicial finite elements τ_ℓ and the ansatz and test spaces

$$Y_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;0,}(Q) = \text{span}\{\varphi_k^Y\}_{k=1}^{M_Y}$$

and

$$X_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;,0}(Q) = \text{span}\{\varphi_k^X\}_{k=1}^{M_X}.$$

of globally continuous, piecewise linear functions. In order to get a good discrete approximation of the set of state constraints K_s, let us consider

$$K_{sh} := \big\{ z_h \in Y_h : \Pi_h g_-(x,t) \le z_h(x,t) \le \Pi_h g_+(x,t), \quad \text{for all } (x,t) \in Q \big\}.$$

where $\Pi_h : \mathrm{dom}(\Pi_h) \subset L^2(Q) \to Y_h$ denotes some suitable (quasi-)interpolation operator. Since again we are not able to compute the action of S, we replace it by a computable approximation $\tilde{S} : \mathcal{H}_{0;0,}(Q) \to [\mathcal{H}_{0;0,}(Q)]^*$ defined as $\tilde{S}y = \widetilde{B}^* p_{yh}$ for all $y \in \mathcal{H}_{0;0,}(Q)$, where $p_{yh} \in X_h$ solves (4.154). The discrete variational formulation is then to find $\tilde{y}_{\varrho h} \in K_{sh}$ such that

$$\varrho\langle \tilde{S}\tilde{y}_{\varrho h}, z_h - \tilde{y}_{\varrho h}\rangle_Q + \langle \tilde{y}_{\varrho h}, z_h - \tilde{y}_{\varrho h}\rangle_{L^2(Q)} \ge \langle y_d, z_h - \tilde{y}_{\varrho h}\rangle_{L^2(Q)} \quad \text{for all } z_h \in K_{sh}, \tag{4.176}$$

Recall, that the inverse inequality (4.156), i.e., $\|z_h\|_{\mathcal{H}_{0;0,}(Q)} \le c_I h^{-1} \|z_h\|_{L^2(Q)}$ holds, for all $z_h \in Y_h$. Therefore, Theorem 4.43 is applicable and gives unique solvability and the following error estimates for the variational inequality (4.176).

THEOREM 4.43. *Let $y_d \in L^2(Q)$ and $0 \in K_{sh}$. Then for the unique solution $\tilde{y}_{\varrho h} \in K_{sh}$ of (4.176) there holds*

$$\|\tilde{y}_{\varrho h} - y_d\|_{L^2(Q)} \le \|y_d\|_{L^2(Q)}. \tag{4.177}$$

Moreover, let $y_d \in K_s \cap H^{1,1}_{0;0,}(Q)$ such that $Sy_d \in L^2(Q)$ and $p_{y_d} \in H^{1,1}_{0;,0}(Q) \cap H^2(Q)$ be the unique solution of

$$\langle A p_{y_d}, q\rangle_Q = \langle B y_d, q\rangle_Q, \quad \text{for all } q \in H^{1,1}_{0;,0}(Q).$$

Then,

$$\begin{aligned} \|\tilde{y}_{\varrho h} - y_d\|_{L^2(Q)} \le c\Big(&\big[h^{-1}\varrho^{3/2} + \varrho\big]\big(\|Sy_d\|_{L^2(Q)} + \|Sg_\pm\|_{L^2(Q)}\big) \\ &+ h^{-1}\varrho \inf_{q_h \in X_h} \|p_{y_d} - q_h\|_{H^{1,1}_{0;,0}(Q)} \\ &+ \big[h^{-1}\sqrt{\varrho} + 1\big] \inf_{z_h \in K_h} \big[\varrho \|y_d - z_h\|^2_{\mathcal{H}_{0;0,}(Q)} + \|y_d - z_h\|^2_{L^2(Q)}\big]^{1/2}\Big). \end{aligned} \tag{4.178}$$

REMARK 4.44 ((Quasi-)interpolation operator). *In practice, we want K_{sh} to be a good approximation of the set K_h. Thus, we say that $\Pi_h : dom(\Pi_h) \subset L^2(Q) \to Y_h$ is a suitable (quasi-)interpolation operator if it satisfies*

(K1) $\Pi_h y \in K_{sh}$ *for all* $y \in K_s \cap dom(\Pi_h)$,

(K2) $\|y - \Pi_h y\|_{L^2(Q)} + h\|y - \Pi_h y\|_{\mathcal{H}_{0;0,}(Q)} \le c h^2 \|y\|_{H^2(Q)}$ *for all* $y \in H^2(Q) \cap H^{1,1}_{0;0,}(Q)$.

Whenever $Q \subset \mathbb{R}^n$, $n = 1, 2, 3$, we have $H^2(Q) \subset \mathcal{C}(\overline{Q})$ and one can choose $\Pi_h = I_h : \mathcal{C}(\overline{Q}) \to Y_h$ as the nodal interpolation, for which the above assumptions hold true. For four dimensional space-time domains and more complex geometries, the construction of a suitable projection is an open task, that exceeds the scope of this work.

The main statement of this section is again an interpolation argument, revealing the optimal choice $\varrho = h^2$, as in the unconstrained case.

THEOREM 4.45. *Let $\tilde{y}_{\varrho h} \in K_{sh}$ denote the unique solution of (4.176) and let $y_d \in K_s \cap H^r(Q)$ for $r \in (1,2]$ or $y_d \in H^{r;r}_{0;0,}(Q)$ for $r \in [0,1]$, where we additionally assume $g_-(x,t) \le y_d(x,t) \le g_+(x,t)$ for a.a. $(x,t) \in Q$. Moreover, we assume that $Sy_d \in H^{r-2}$, $g_\pm \in H^2(Q)$ and $p_{y_d} \in H^{1,1}_{0;,0}(Q) \cap H^r(Q)$ for all $r \in [0,2]$ and (4.161) holds. If $\varrho = h^2$, then*

$$\|y_d - y_{\varrho h}\|_{L^2(Q)} \le ch^r\big(\|y_d\|_{H^r(\Omega)} + \|g_\pm\|_{H^r(\Omega)}\big).$$

Proof. Let us first assume that $y_d \in K_s \cap H^2(\Omega)$. Then, by assumption (K1) we have $\Pi_h y_d \in K_{sh}$ and, using (K2), it holds

$$\inf_{z_h \in K_{sh}} \|y_d - z_h\|_{L^2(Q)} \le \|y_d - \Pi_h y_d\|_{L^2(Q)} \le ch^2 \|y_d\|_{H^2(Q)}$$

and

$$\inf_{z_h \in K_{sh}} \|y_d - z_h\|_{\mathcal{H}_{0;0,}(Q)} \le \|y_d - \Pi_h y_d\|_{\mathcal{H}_{0;0,}(Q)} \le ch\|y_d\|_{H^2(Q)}.$$

Moreover, since $p_{y_d} \in H^2(Q)$, we get by best approximation, see Theorem 2.36, and (4.161) that

$$\inf_{q_h \in X_h} \|p_{y_d} - q_h\|_{H^{1,1}_{0;,0}(Q)} \le ch\|p_{y_d}\|_{H^2(Q)} \le ch\|y_d\|_{H^2(Q)}.$$

Thus, with $\varrho = h^2$ estiamte (4.178) becomes

$$\|y_d - \tilde{y}_{\varrho h}\|_{L^2(Q)} \le ch^2\big(\|y_d\|_{H^2(Q)} + \|g_\pm\|_{H^2(Q)}\big).$$

Interpolating this estimate with (4.177) gives the desired result. $\square$

Numerical results

Using the fe-isomorphism the variational formulation (4.176) is equivalent to find $K_{sh} \ni \tilde{y}_{\varrho h} \leftrightarrow \mathbf{y}_{\varrho h} \in \mathbb{R}^{M_Y}$ such that

$$\varrho(S_h \tilde{\mathbf{y}}_{\varrho h}, \mathbf{z}_h - \tilde{\mathbf{y}}_{\varrho h})_2 + (M_h \tilde{\mathbf{y}}_{\varrho h}, \mathbf{z}_h - \tilde{\mathbf{y}}_{\varrho h})_2 \ge (\mathbf{y}_{dh}, \mathbf{z}_h - \tilde{\mathbf{y}}_{\varrho h})_2 \tag{4.179}$$

for all $K_{sh} \ni z_h \leftrightarrow \mathbf{z}_h \in \mathbb{R}^{M_Y}$, where $S_h = B_h^\top A_h^{-1} B_h$ with matrices and load vector defined as in (4.164). To incorporate the constraints, we define the auxiliary variable

$$\tilde{\boldsymbol{\lambda}}_h := \varrho S_h \tilde{\mathbf{y}}_{\varrho h} + M_h \tilde{\mathbf{y}}_{\varrho h} - \mathbf{y}_{dh},$$

and the set of active nodes be defined as

$$\mathcal{A}_{s,\pm} := \{k = 1, \ldots, M_Y : \tilde{\mathbf{y}}_{\varrho h}[k] = (\Pi_h g)_\pm(x_k)\}. \tag{4.180}$$

Then we obtain the discrete complementarity conditions

$$\begin{aligned}
\tilde{\boldsymbol{\lambda}}_h[k] &= 0, \quad (\Pi_h g_-)(x_k) < \tilde{\mathbf{y}}_{\varrho h}[k] < (\Pi_h g_+)(x_k), && \text{for } k \notin \mathcal{A}_{s,\pm}, \\
\tilde{\boldsymbol{\lambda}}_h[k] &\geq 0, \qquad \tilde{\mathbf{y}}_{\varrho h}[k] = (\Pi_h g_-)(x_k), && \text{for } k \in \mathcal{A}_{s,-}, \\
\tilde{\boldsymbol{\lambda}}_h[k] &\leq 0, \qquad \tilde{\mathbf{y}}_{\varrho h}[k] = (\Pi_h g_+)(x_k), && \text{for } k \in \mathcal{A}_{s,+}.
\end{aligned} \tag{4.181}$$

These are equivalent to

$$\tilde{\boldsymbol{\lambda}}_h[k] = \min\{0, \tilde{\boldsymbol{\lambda}}_h[k] + \alpha(\mathbf{g}_{+h}[k] - \tilde{\mathbf{y}}_{\varrho h}[k])\} + \max\{0, \tilde{\boldsymbol{\lambda}}_h[k] + \alpha(\mathbf{g}_{-h}[k] - \tilde{\mathbf{y}}_{\varrho h}[k])\},$$

for some $\alpha > 0$ and $Y_h \ni \Pi_h g_\pm \leftrightarrow \mathbf{g}_{\pm h} \in \mathbb{R}^{M_Y}$. Now in order to solve (4.179), we compute the roots of the functions

$$F_1(\tilde{\mathbf{y}}_{\varrho h}, \tilde{\boldsymbol{\lambda}}_h) = \varrho S_h \tilde{\mathbf{y}}_{\varrho h} + M_h \tilde{\mathbf{y}}_{\varrho h} - \tilde{\boldsymbol{\lambda}}_h - \mathbf{y}_d$$

and

$$\begin{aligned}
F_2(\tilde{\mathbf{y}}_{\varrho h}, \tilde{\boldsymbol{\lambda}}_h) = \tilde{\boldsymbol{\lambda}}_h \quad &- \quad \min\{0, \tilde{\boldsymbol{\lambda}}_h[k] + \alpha(\mathbf{g}_{+h}[k] - \tilde{\mathbf{y}}_{\varrho h}[k])\} \\
&- \quad \max\{0, \tilde{\boldsymbol{\lambda}}_h[k] + \alpha(\mathbf{g}_{-h}[k] - \tilde{\mathbf{y}}_{\varrho h}[k])\}.
\end{aligned}$$

simultaneously. This is achieved applying a semi-smooth Newton algorithm, which is equivalent to and active set strategy, see Algorithm 1, as in the case of constraints for the Poisson equation, see Section 4.1.3. Therefore, we successively compute the iterates

$$\begin{pmatrix} \tilde{\mathbf{y}}_{\varrho h}^{m+1} \\ \tilde{\boldsymbol{\lambda}}_h^{m+1} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{y}}_{\varrho h}^{m} \\ \tilde{\boldsymbol{\lambda}}_h^{m} \end{pmatrix} - \left(D\mathbf{F}(\tilde{\mathbf{y}}_{\varrho h}^m, \tilde{\boldsymbol{\lambda}}_h^m)\right)^{-1} \mathbf{F}(\tilde{\mathbf{y}}_{\varrho h}^m, \tilde{\boldsymbol{\lambda}}_h^m), \tag{4.182}$$

where the Jacobian is given as

$$D\mathbf{F}(\mathbf{z}_h, \boldsymbol{\mu}_h) = \begin{pmatrix} \varrho S_h + M_h & -I \\ \alpha L'(\mathbf{g}_{\pm h}, \mathbf{z}_h, \boldsymbol{\mu}_h) & I - L'(\mathbf{g}_{\pm h}, \mathbf{z}_h, \boldsymbol{\mu}_h) \end{pmatrix},$$

where the entries of L' are as in (4.79). To support our theoretical results, we will consider the domain $Q = (0, L) \times (0, T)$, $T = L = 1$, and the target functions, see Figure 4.17, $y_{d,1} \in \mathcal{C}^2(\overline{Q})$, given as

$$y_{d,1}(x,t) = \begin{cases} \frac{1}{2}(6t - 3x - 2)^3(3x - 6t)^3 \sin(\pi x), & x \leq 2t \text{ and } 6t - 3x \leq 2, \\ 0, & \text{else,} \end{cases}$$

and the discontinuous target $y_{d,3} \in H_0^{1/2-\varepsilon}(Q)$, $\varepsilon > 0$, defined as

$$y_{d,3}(x,t) = \begin{cases} 1, & (x,t) \in (0.25, 0.75)^2 \subset Q, \\ 0, & \text{else.} \end{cases}$$

The constraints on the state are imposed by the barrier functions $g_-^{(j)} \equiv 0$, $j = 1, 2$, and

$$g_+^{(1)}(x,t) = \min\{0.3, y_{d,1}(x,t)\} \quad \text{for } (x,t) \in Q$$

and

$$g_+^{(2)}(x,t) = 0.5 y_{d,1}(x,t) \quad \text{for } (x,t) \in Q.$$

We solve (4.182) successively, with $\varrho = h^2$ and initial guess

$$\tilde{\mathbf{y}}_{\varrho h}^0 = (h^2 S_h + M_h)^{-1} \tilde{\mathbf{y}}_{d,h} \in \mathbb{R}^{M_Y} \quad \text{and} \quad \tilde{\boldsymbol{\lambda}}_h^0 = \mathbf{0}_h \in \mathbb{R}^{M_Y}.$$

As a stopping criterion we choose the maximal absolute error in each node, i.e., we stop if

$$\text{tol}_s := \max\{\text{tol}_{s,+}, \text{tol}_{s,-}\} < 10^{-5}, \tag{4.183}$$

where

$$\begin{aligned} \text{tol}_{s,+} &:= \max_{\{k : \tilde{\mathbf{y}}_{\varrho h}[k] > g_+(x_k)\}} |\tilde{\mathbf{y}}_{\varrho h}[k] - g_+(x_k)|, \\ \text{tol}_{s,-} &:= \max_{\{k : \tilde{\mathbf{y}}_{\varrho h}[k] < g_-(x_k)\}} |\tilde{\mathbf{y}}_{\varrho h}[k] - g_-(x_k)|. \end{aligned}$$

To reconstruct the control, we solve (4.169), as in the unconstrained case. The results are depicted in Figure 4.24.

Control constraints

In this case we want to minimize

$$\mathcal{J}(y_\varrho, u_\varrho) = \frac{1}{2}\|y_d - y_\varrho\|^2_{L^2(Q)} + \frac{\varrho}{2}\|u_\varrho\|^2_{[H^{1,1}_{0;,0}(Q)]^*},$$

subject to the wave equation

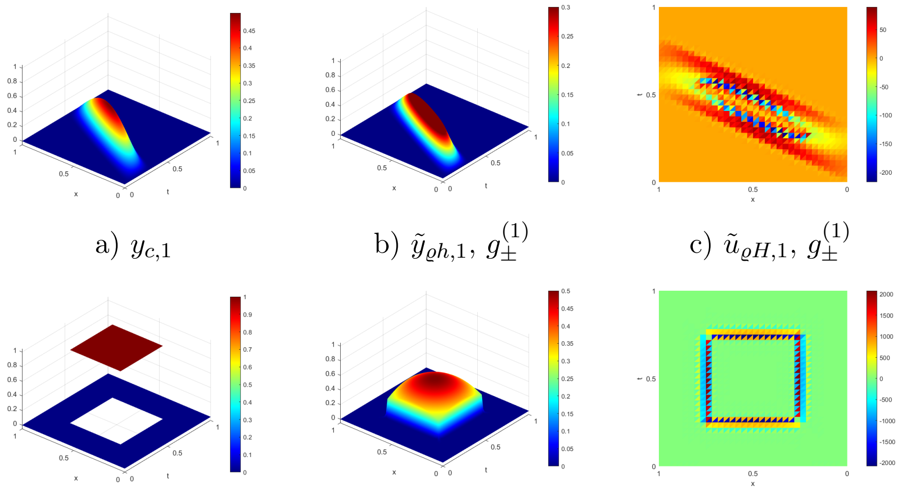
$$\begin{aligned} \Box y_\varrho(x,t) := \partial_{tt} y_\varrho(x,t) - \Delta_x y_\varrho(x,t) &= u_\varrho(x,t) && \text{for } (x,t) \in Q, \\ y_\varrho(x,t) &= 0 && \text{for } (x,t) \in \Sigma, \\ y_\varrho(x,0) = \partial_t y_\varrho(x,t)_{|t=0} &= 0 && \text{for } x \in \Omega, \end{aligned}$$

and to the control constraints

$$h_-(x,t) \le u_\varrho(x,t) \le h_+(x,t) \quad \text{for } (x,t) \in Q,$$

for a given target $y_d \in L^2(Q)$ and barrier functions $h_\pm \in L^2(Q)$, which fulfill $h_-(x,t) \le 0 \le h_+(x,t)$ for a.a. $(x,t) \in Q$. Following the abstract framework in Section 3.2, this is equivalent to find $y_\varrho \in K_c$, such that

$$\varrho \langle S y_\varrho, z - y_\varrho \rangle_Q + \langle y_\varrho, z - y_\varrho \rangle_{L^2(Q)} \ge \langle y_d, z - y_\varrho \rangle_{L^2(Q)} \quad \text{for all } z \in K_c, \tag{4.184}$$



a) $y_{c,1}$ b) $\tilde{y}_{\varrho h,1}$, $g_\pm^{(1)}$ c) $\tilde{u}_{\varrho H,1}$, $g_\pm^{(1)}$

d) $y_{d,3}$ e) $\tilde{y}_{\varrho h,3}$, $g_\pm^{(2)}$ f) $\tilde{u}_{\varrho H,3}$, $g_\pm^{(2)}$

Figure 4.24: Targets $y_{d,i}$, computed constrained states $\tilde{y}_{\varrho h,i}$, $= 1,3$ on a mesh with $N = 32768$ elements and $M = M16256$ DoFs with constraints $g_\pm^{(j)}$, $j = 1,2$ and reconstruction of the controls $\tilde{u}_{\varrho H,i}$ on a mesh with $N_H = 2048$ elements.

where the set of constraints is defined as

$$K_c := \{z \in \mathcal{H}_{0;0,}(Q) : \langle h_-, q\rangle_{L^2(Q)} \le \langle \widetilde{B}z, q\rangle_Q \le \langle h_+, q\rangle_{L^2(Q)}\, q \in H^{1,1}_{0;,0}(Q),\, q \ge 0\}.$$

As observed, this problem admits the same structure as state constraints and all the estimates from Lemma 4.42 carry over verbatim, when replacing K_s by K_c. The complementarity conditions can be derived considering the auxiliary variable $w_\lambda \in H^{1,1}_{0;,0}(Q)$ as unique solution of

$$\langle \widetilde{B}^* w_\lambda, z\rangle_Q = \langle \varrho S + y_\varrho - y_d, z\rangle_Q \quad \text{for all } z \in \mathcal{H}_{0;0,}(Q),$$

for which we get, by (3.64),

$$\begin{aligned} w_\lambda &= 0, & h_- < u_\varrho &< h_+, && \text{on } Q \setminus Q_{c,\pm}, \\ w_\lambda &\ge 0, & u_\varrho &= h_-, && \text{on } Q_{c,-}, \\ w_\lambda &\le 0, & u_\varrho &= h_+, && \text{on } Q_{c,+}, \end{aligned} \tag{4.185}$$

where the sets $Q_{c,\pm}$ are defined, if Assumption 3.26 is satisfied, as

$$Q_{c,\pm} := \{(x,t) \in Q : u_\varrho(x,t) = h_\pm(x,t)\}.$$

For the discretization, we introduce the set

$$K_{ch} := \{z_h \in Y_h : \langle h_-, q_h\rangle_{L^2(Q)} \le \langle B z_h, q_h\rangle_Q \le \langle h_+, q_h\rangle,\, q_h \in Y_h,\, q_h \ge 0\}$$

and consider the discrete variational problem to find $\tilde{y}_{\varrho h} \in K_{ch}$ such that

$$\varrho\langle \tilde{S}\tilde{y}_{\varrho h}, z_h - \tilde{y}_{\varrho h}\rangle_Q + \langle \tilde{y}_{\varrho h}, z_h - \tilde{y}_{\varrho h}\rangle_{L^2(Q)} \geq \langle y_d, z_h - \tilde{y}_{\varrho h}\rangle_{L^2(Q)} \quad \text{for all } z_h \in K_{ch}, \tag{4.186}$$

for which related error estimates as in Theorem 4.43 (replacing K_s by K_c, K_{sh} by K_{ch} and $\|Sg_\pm\|_{L^2(Q)}$ by $\|h_\pm\|_{L^2(Q)}$) follow, if we restrict to functions satisfying Assumption 3.26 and Assumption 3.27. As in the case of state constraints, using the fe-isomorphism, the solution of (4.186) $K_{ch} \ni \tilde{y}_{\varrho h} \leftrightarrow \tilde{\mathbf{y}}_{\varrho h} \in \mathbb{R}^{M_Y}$ has to fulfill

$$\varrho(S_h\tilde{\mathbf{y}}_{\varrho h}, \mathbf{z}_h - \tilde{\mathbf{y}}_{\varrho h})_2 + (M_h\tilde{\mathbf{y}}_{\varrho h}, \mathbf{z}_h - \tilde{\mathbf{y}}_{\varrho h})_2 \geq (\mathbf{y}_{dh}, \mathbf{z}_h - \mathbf{y}_{\varrho h})_2 \tag{4.187}$$

for all $K_{ch} \ni z_h \leftrightarrow \mathbf{z}_h \in \mathbb{R}^{M_Y}$. To incorporate the constraints, we will consider the auxiliary variable $\tilde{\mathbf{w}}_{\lambda,h} \in \mathbb{R}^{M_X}$ solving

$$B_h^\top \tilde{\mathbf{w}}_{\lambda,h} = \varrho S_h\tilde{\mathbf{y}}_{\varrho h} + M_h\tilde{\mathbf{y}}_{\varrho h} - \mathbf{y}_{dh}.$$

Introducing the set of active nodes as

$$\mathcal{A}_{c,\pm} := \{k = 1, \ldots, M_Y : (B_h\tilde{\mathbf{y}}_{\varrho h})[k] = \mathbf{h}_{\pm h}[k]\},$$

where the entries of $\mathbf{h}_{\pm h} \in \mathbb{R}^{M_X}$ are given as

$$\mathbf{h}_\pm[k] = \langle h_\pm, \varphi_k^X\rangle_{L^2(Q)}\, k = 1, \ldots, M_X,$$

we can conclude the discrete complementartity conditions

$$\begin{aligned}
\tilde{\mathbf{w}}_{\lambda h}[k] &= 0, \quad \mathbf{h}_{-h}[k] < (B_h\tilde{\mathbf{y}}_{\varrho h})[k] < \mathbf{h}_{+h}[k], && \text{for } k \notin \mathcal{A}_{c,\pm},\\
\tilde{\mathbf{w}}_{\lambda h}[k] &\geq 0, \quad (B_h\tilde{\mathbf{y}}_{\varrho h})[k] = \mathbf{h}_{-h}[k], && \text{for } k \in \mathcal{A}_{c,-},\\
\tilde{\mathbf{w}}_{\lambda h}[k] &\leq 0, \quad (B_h\tilde{\mathbf{y}}_{\varrho h})[k] = \mathbf{h}_{+h}[k], && \text{for } k \in \mathcal{A}_{c,+},
\end{aligned} \tag{4.188}$$

as in the continuous case. These are equivalent to

$$\tilde{\mathbf{w}}_{\lambda h}[k] = \min\{0, \tilde{\mathbf{w}}_{\lambda h}[k] + \alpha(\mathbf{h}_{h+} - (B_h\tilde{\mathbf{y}}_{\varrho h})[k])\} + \max\{0, \tilde{\mathbf{w}}_{\lambda h}[k] + \alpha(\mathbf{h}_{h-} - (B_h\tilde{\mathbf{y}}_{\varrho h})[k])\},$$

for some $\alpha > 0$. Thus, we want to find the roots of the functions

$$\tilde{\mathbf{F}}_1(\tilde{\mathbf{y}}_{\varrho h}, \tilde{\mathbf{w}}_{\lambda h}) = \varrho S_h\tilde{\mathbf{y}}_{\varrho h} + M_h\tilde{\mathbf{y}}_{\varrho h} - B_h^\top\tilde{\mathbf{w}}_{\lambda h} - \mathbf{y}_{dh},$$

and

$$\begin{aligned}
\tilde{\mathbf{F}}_2(\tilde{\mathbf{y}}_{\varrho h}, \tilde{\mathbf{w}}_{\lambda h}) = \tilde{\mathbf{w}}_{\lambda h} \quad &- \quad \min\{0, \tilde{\mathbf{w}}_{\lambda h}[k] + \alpha(\mathbf{h}_{h+}[k] - (B_h\tilde{\mathbf{y}}_{\varrho h})[k])\}\\
&- \quad \max\{0, \tilde{\mathbf{w}}_{\lambda h}[k] + \alpha(\mathbf{h}_{h-}[k] - (B_h\tilde{\mathbf{y}}_{\varrho h})[k])\},
\end{aligned}$$

simultaneously, which can be achieved by applying a semi-smooth Newton method, i.e., we compute the iterates

$$\begin{pmatrix} \tilde{\mathbf{y}}_{\varrho h}^{m+1} \\ \tilde{\mathbf{w}}_{\lambda h}^{m+1} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{y}}_{\varrho h}^{m} \\ \tilde{\mathbf{w}}_{\lambda h}^{m} \end{pmatrix} - \left(D\tilde{\mathbf{F}}(\tilde{\mathbf{y}}_{\varrho h}^m, \tilde{\mathbf{w}}_{\lambda h}^m)\right)^{-1}\tilde{\mathbf{F}}(\tilde{\mathbf{y}}_{\varrho h}^m, \tilde{\mathbf{w}}_{\lambda h}^m),$$

where the Jacobian is given as

$$D\tilde{\mathbf{F}}(\mathbf{z}_h, \mathbf{w}_h) = \begin{pmatrix} \varrho S_h + M_h & -B_h^\top \\ \alpha L'(\mathbf{h}_{\pm h}, B_h\mathbf{z}_h, \mathbf{w}_h)B_h & I - L'(\mathbf{h}_{\pm h}, B_h\mathbf{z}_h, \mathbf{w}_h) \end{pmatrix}.$$

REMARK 4.46. *To avoid setting up the matrix $S_h = B_h^\top A_h^{-1} B_h$, which is dense due to the inverse of A_h, we can introduce $\boldsymbol{p}_{\varrho h} = -\varrho A_h^{-1} B_h \tilde{\boldsymbol{y}}_{\varrho h}$ and yet another function*

$$\tilde{\boldsymbol{F}}_0(\boldsymbol{p}_{\varrho h}, \tilde{\boldsymbol{y}}_{\varrho h}, \tilde{\boldsymbol{w}}_{\lambda h}) = \varrho^{-1} A_h \boldsymbol{p}_{\varrho h} + B_h \tilde{\boldsymbol{y}}_{\varrho h}$$

and find the roots of

$$\overline{\boldsymbol{F}}(\boldsymbol{p}_{\varrho h}, \tilde{\boldsymbol{y}}_{\varrho h}, \tilde{\boldsymbol{w}}_{\lambda h}) = \begin{pmatrix} \tilde{\boldsymbol{F}}_0 \\ \tilde{\boldsymbol{F}}_1 \\ \tilde{\boldsymbol{F}}_2 \end{pmatrix} (\boldsymbol{p}_{\varrho h}, \tilde{\boldsymbol{y}}_{\varrho h}, \tilde{\boldsymbol{w}}_{\lambda h}),$$

iteratively solving a semi smooth Newton method with Jacobian

$$D\overline{\boldsymbol{F}}(\boldsymbol{p}_{\varrho h}, \boldsymbol{z}_h, \boldsymbol{w}_h) = \begin{pmatrix} \varrho^{-1} A_h & B_h & 0 \\ B_h^\top & I & -B_h^\top \\ 0 & \alpha L'(\boldsymbol{h}_{\pm h}, B_h \boldsymbol{z}_h, \boldsymbol{w}_h) B_h & I - L'(\boldsymbol{h}_{\pm h}, B_h \boldsymbol{z}_h, \boldsymbol{w}_h) \end{pmatrix}.$$

Although, we gain additional degrees of freedom, the Jacobian is sparse.

As a test example, we consider the target

$$y_{cd}(x,t) = \begin{cases} 1, & t \geq 1 - x, \\ 0, & \text{else}, \end{cases}$$

on the space time domain $Q = (0,1)^2$ and the barrier functions

$$h_-(x,t) \equiv 0, \quad \text{and} \quad h_+(x,t) = 1000.$$

We choose $\varrho = h^2$ and the initial guesses

$$\tilde{\mathbf{y}}^0_{\varrho h} = (h^2 S_h + M_h)^{-1} \mathbf{y}_{d,h} \in \mathbb{R}^{M_Y}, \quad \mathbf{p}^0_{\varrho h} = \varrho A_h^{-1} B_h \tilde{\mathbf{y}}^0_{\varrho h} \in \mathbb{R}^{M_X}, \quad \tilde{\mathbf{w}}^0_{\lambda h} = \mathbf{0}_h \in \mathbb{R}^{M_X}.$$

As a stopping criterion we choose the maximal absolute error in each node, i.e., we stop if

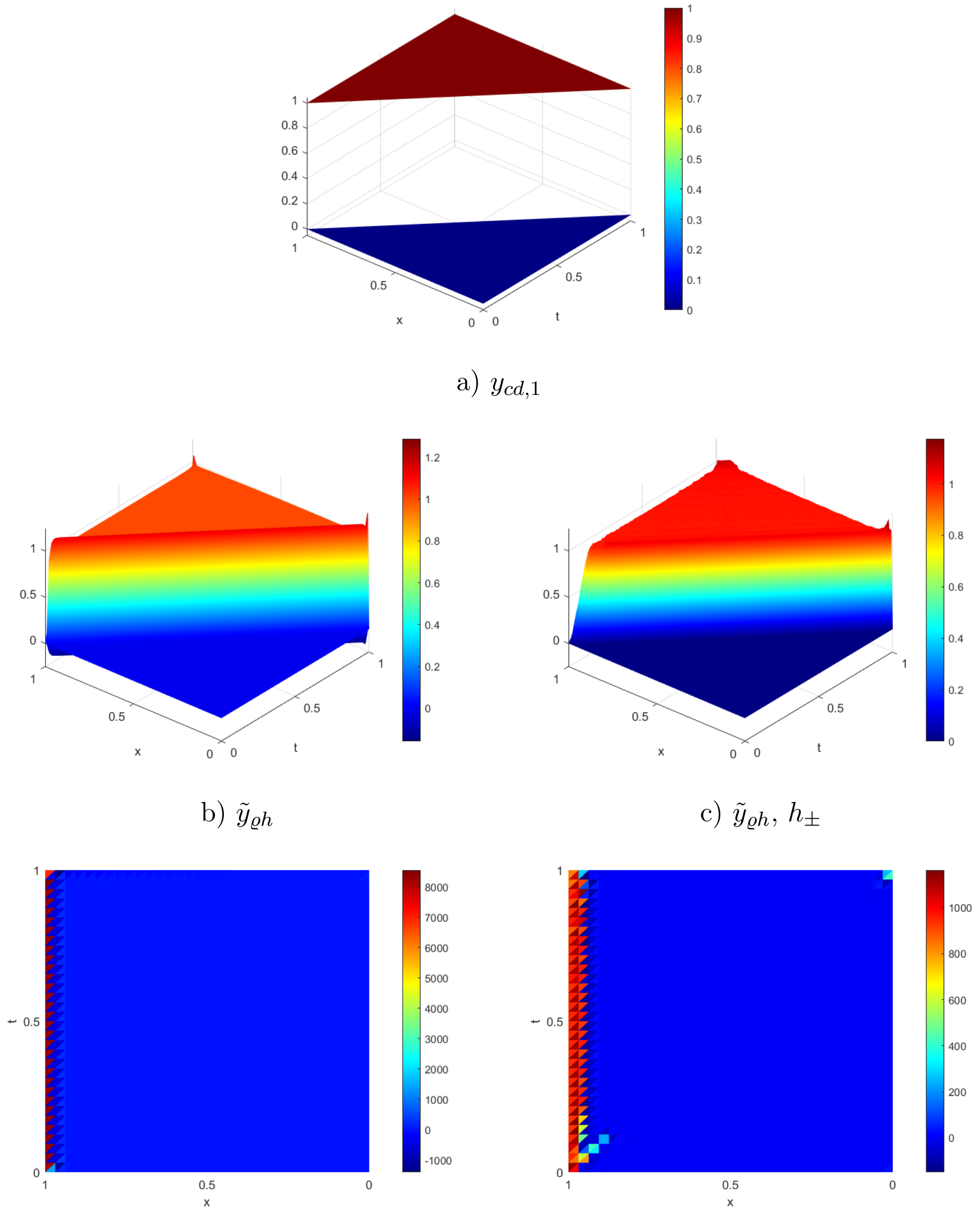
$$\mathrm{tol}_c := \max\{\mathrm{tol}_{c,+}, \mathrm{tol}_{c,-}\} < 10^{-5}, \tag{4.189}$$

where

$$\begin{aligned} \mathrm{tol}_{c,+} &:= \max_{\{k:(B_h \tilde{\mathbf{y}}_{\varrho h})[k] > \mathbf{h}_{+h}[k]\}} |(B_h \tilde{\mathbf{y}}_{\varrho h})[k] - \mathbf{h}_{+h}[k]|, \\ \mathrm{tol}_{c,-} &:= \max_{\{k:(B_h \tilde{\mathbf{y}}_{\varrho h})[k] < \mathbf{h}_{-h}[k]\}} |(B_h \tilde{\mathbf{y}}_{\varrho h})[k] - \mathbf{h}_{-h}[k]|. \end{aligned}$$

To reconstruct the control, we solve (4.169), as in the unconstrained case. The results are depicted in Figure 4.25.

To conclude, both state and control constraints fit into the abstract framework and can be handled as in the elliptic case in Section 4.1.3 and the final discussion therein. We point out, that all the numerical examples were carried out using MATLAB, where we adapted refinement routines from [44].



a) $y_{cd,1}$

b) $\tilde{y}_{\varrho h}$

c) $\tilde{y}_{\varrho h}$, $h_{\pm}$

d) $\tilde{u}_{\varrho H}$ unconstrained

e) $\tilde{u}_{\varrho H}$, $h_{\pm}$ constrained

Figure 4.25: Target y_{cd} and computed state $\tilde{y}_{\varrho h}$ on a mesh with $N = 32768$ elements and $M = 16129$ DoFs and reconstruction of the control $\tilde{u}_{\varrho H}$ on a mesh with $N_H = 2048$ elements with and without constraints $h_{\pm}$.

5 An adaptive least squares space(-time) framework

In this section we will analyze numerical methods for the direct solution of PDEs, based on a least squares framework, following ideas of [70]. In particular, we will see, that a majority of the concepts that have been used in the case of optimal control problems, can be reused and trimmed to this application. To outline the framework, let $X \subset H_X \subset X^*$ and $Y \subset H_Y \subset Y^*$ be two Gelfand triples of Hilbert spaces. Then for given $f \in X^*$, we want to find $y \in Y$ as solution of the operator equation

$$By = f \quad \text{in } X^*. \tag{5.1}$$

This equation admits a unique solution $y \in Y$ for any $f \in X^*$, if and only if $B : Y \to X^*$ is an isomorphism. The main interest now lies in the discretization of (5.1). A well-known approach is to equivalently consider the variational formulation, i.e, to find $y \in Y$ such that

$$\langle By, q\rangle_{H_X} = \langle f, q\rangle_{H_X} \quad \text{for all } q \in X, \tag{5.2}$$

for given $f \in X^*$, where $\langle \cdot, \cdot\rangle_{H_X}$ denotes the duality pairing as extension of the inner product in H_X, and then to find suitable, finite dimensional, trial spaces X_h and Y_h and determine the solution $y_h \in Y_h$ of

$$\langle By_h, q_h\rangle_{H_X} = \langle f, q_h\rangle_{H_X} \quad \text{for all } q_h \in Y_h. \tag{5.3}$$

At this point, it is important to note, that the property of $B : Y \to X^*$ being an isomorphism, does not transfer to the discrete setting, even when choosing conforming spaces $X_h \subset X$ and $Y_h \subset Y$. Thus, the crucial part of this procedure is to construct X_h and Y_h such that (5.3) admits a unique solution, preferably for any right hand side $f \in X^*$. In particular, in the finite dimensional setting this requires $\dim(X_h) = \dim(Y_h)$, which is a major restriction in many applications, as it narrows the choice of trial spaces fiercely. This, in turns, might restrict the applicability of adaptive schemes or parallel algorithms for the solution of the discretized problem. In addition, in practice, we prefer to have discrete spaces that are easy to implement in a numerical scheme, at best, spaces that can be used for a wide class of problems, e.g., standard finite element spaces. Before we state the alternative approach for the solution of the operator equation (5.1), we will consider a motivational example, illustrating the obstacles just stated.

5.1 Motivation

Let us consider the simple initial boundary value problem of a spatially one-dimensional wave equation with homogeneous Dirichlet boundary conditions on the space-time domain $Q = (0, L) \times (0, T)$, $0 < T, L < \infty$, given as

$$\begin{aligned} \partial_{tt} y(x,t) - \partial_{xx} y(x,t) &= f(x,t) && \text{for } (x,t) \in Q, \\ y(x,t) &= 0 && \text{for } (x,t) \in \Sigma = \{0, L\} \times (0, T), \\ y(x,0) = \partial_t y(x,t)_{|t=0} &= 0 && \text{for } x \in (0, L). \end{aligned} \tag{5.4}$$

With the theory outlined in Section 4.2, we can derive the variational formulation to find $y \in \mathcal{H}_{0;0,}(Q)$ such that

$$\langle \widetilde{B} y, q \rangle_Q = \langle f, q \rangle_Q \quad \text{for all } q \in H^{1,1}_{0;,0}(Q), \tag{5.5}$$

where the operator $\widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ defines an isomorphism and thus (5.5) admits a unique solution $y \in \mathcal{H}_{0;0,}(Q)$ for all $f \in [H^{1,1}_{0;,0}(Q)]^*$. For a numerical scheme, we consider an admissible, globally quasi-uniform, decomposition $\mathcal{T}_h = \{\tau_\ell\}_{\ell=1}^N$ of the space time domain Q into shape regular triangles τ_ℓ, of mesh size h_ℓ, $\ell = 1, \ldots, N$ and the trial spaces $X_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;0,}(Q)$ and $Y_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;,0}(Q)$. Using (4.137), the variational formulation is then to find $y_h \in Y_h$ such that

$$\int_0^T \int_0^L -\partial_t y_h(x,t) \partial_t q_h(x,t) + \partial_x y_h(x,t) \partial_x q_h(x,t) \, dt \, dx = \langle f, q_h \rangle_Q, \tag{5.6}$$

for all $q_h \in X_h$.

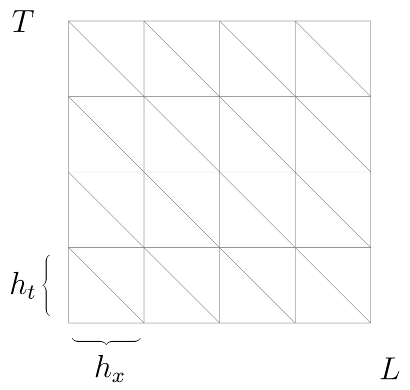


Figure 5.1: Structured space-time mesh

As a test example, let us consider a structured space time triangulation, see Figure 5.1, where we can steer the ratio between h_t and h_x and the solution of the wave equation

$$y_1(x,t) = t^2 \sin(\pi t) \sin(\pi x / L), \tag{5.7}$$

for $(x,t) \in Q$. In Table 5.1 we observe optimal orders of convergence, if $h_t \leq h_x$, whereas the convergence breaks down if $h_t > h_x$. This is known as the CFL-condition, see Remark 4.41, an can be derived explicitly for tensor product meshes, see, e.g., [110]. Since a small violation against this conditions already leads to an unstable numerical scheme, the discrete variational formulation (5.6) is obviously not suitable for adaptive schemes, as, in general we can not guarantee that an

N	h_t	h_x	$\lvert y_1 - y_h\rvert_{H^1(Q)}$	eoc	h_x	$\lvert y_1 - y_h\rvert_{H^1(Q)}$	eoc
8	0.500	0.500	$7.98 \cdot 10^{-1}$	0.000	0.450	$7.79 \cdot 10^{-1}$	0.000
32	0.250	0.250	$4.75 \cdot 10^{-1}$	0.749	0.225	$4.69 \cdot 10^{-1}$	0.734
128	0.125	0.125	$2.46 \cdot 10^{-1}$	0.948	0.113	$2.42 \cdot 10^{-1}$	0.956
512	0.063	0.063	$1.24 \cdot 10^{-1}$	0.987	0.056	$9.30 \cdot 10^{-1}$	−1.945
2,048	0.031	0.031	$6.22 \cdot 10^{-2}$	0.997	0.028	$1.30 \cdot 10^{5}$	−17.097
8,192	0.016	0.016	$3.11 \cdot 10^{-2}$	0.999	0.014	$4.70 \cdot 10^{16}$	−38.389
32,768	0.008	0.008	$1.56 \cdot 10^{-2}$	1.000	0.007	$1.62 \cdot 10^{41}$	−81.513

a) $h_t \leq h_x$ b) $h_t \not\leq h_x$

Table 5.1: Computation for the test example (5.7) on a mesh satisfying and violating the CFL-condition

adaptively refined mesh will meet this condition. Possible workarounds are the use of a suitable transformation operator such as the modified Hilbert transformation, see [88], which leads to an unconditionally stable Galerkin Bubnov scheme or a simpler stabilization approach on a tensor product space-time mesh, see [109], which was extended to arbitrary higher order schemes in [115]. But, as the latter approaches, require a tensor product structure (so far), they are not well-suited for fully space time adaptive schemes. To overcome the tensor product structure one might consider the idea of tent pitching, [56, 57], which guarantees that all elements satisfy the CFL-condition, or a suitable discontinuous Galerkin method, e.g., [35, 58], where for the latter the analysis relies on a density argument first stated in [32], which so far is only provable on hypercubes. Yet another approach, that has the advantage of an inbuilt error estimator, is to replace the direct problem by a least squares/minimal residual approach. This has been studied by Führer and Karkulik in the case of first order least squares systems (FOSLS) [43] or in the case of minimal residual Petrov-Galerkin discretizations in [2, 90], see also [29] for conditionally stable PDEs. Following [70], we will take the point of view of the latter approaches, ending up to with a mixed finite element scheme, where the second variable is the Riesz lift of the residual and serves as an error estimator. An abstract framework, describing the method, will be outlined in the next section, using the tools already developed in Chapter 3.

5.2 Abstract framework for adaptive least squares problems

In this section we will outline an abstract framework for the solution of the operator equation (5.1), using a least squares approach. For ease of presentation, let us again restrict to the case, where $H_X = H_Y = H$. As already mentioned, in order to have a unique solution $y \in Y$ of (5.1) for every $f \in X^*$, it is mandatory for $B : Y \to X^*$ to define an isomorphism. Recall, that this is guaranteed by the assumptions:

ASSUMPTION 5.1.

(B1) *(Boundedness)* $\exists c_2^B > 0 : \|By\|_{X^*} \leq c_2^B \|y\|_Y$ *for all* $y \in Y$,

(B2) *(Injectivity)* $\exists c_1^B > 0 : c_1^B \|y\|_Y \leq \sup_{0 \neq q \in X} \frac{\langle By, q\rangle_H}{\|q\|_X}$ *for all* $y \in Y$,

(B3) *(Surjectivity)* $\forall q \in X \setminus \{0\} \, \exists y_q \in Y : \langle By_q, q\rangle_H \neq 0$.

Now, note that the solution $y \in Y$ of the operator equation $By = f$ in X^* is equivalently characterized as the minimizer of

$$y = \frac{1}{2} \arg\min_{z \in Y} \|Bz - f\|_{X^*}^2. \tag{5.8}$$

As in the case of optimal control problems, we now want to give a computable realization of the dual norm X^*, to make the approach practical for applications. This is achieved by the operator $A : X \to X^*$, if it satisfies the assumptions:

ASSUMPTION 5.2.

(A1) *(Boundedness)* $\exists c_2^A > 0 : \|Aq\|_{X^*} \leq c_2^A \|q\|_X$ *for all* $q \in X$,

(A2) *(Self-adjointness)* $\langle Ap, q\rangle_H = \langle p, Aq\rangle_H$ *for all* $p, q \in X$,

(A3) *(Ellipticity)* $\exists c_1^A > 0 : \langle Aq, q\rangle_H \geq c_1^A \|q\|_X^2$ *for all* $q \in X$.

Then, by Lemma 2.10, we have for all $q \in X$ and all $v \in X^*$ that

$$\|q\|_A := \sqrt{\langle Aq, q\rangle_H} \quad \text{and} \quad \|v\|_{A^{-1}} := \sqrt{\langle v, A^{-1}v\rangle_H}$$

define equivalent norms on X and X^* respectively, with norm equivalence constants

$$\sqrt{c_1^A}\|q\|_X \leq \|q\|_A \leq \sqrt{c_2^A}\|q\|_X \quad \text{and} \quad \frac{1}{\sqrt{c_2^A}}\|v\|_{X^*} \leq \|v\|_{A^{-1}} \leq \frac{1}{\sqrt{c_1^A}}\|v\|_{X^*}. \tag{5.9}$$

In particular, the minimization problem (5.8) now becomes

$$y = \frac{1}{2} \arg\min_{z \in Y} \|Bz - f\|_{A^{-1}}^2 = \frac{1}{2} \arg\min_{z \in Y} \langle Bz - f, A^{-1}(Bz - f)\rangle_H, \tag{5.10}$$

for which the minimizer has to fulfill the gradient equation

$$B^* A^{-1}(By - f) = 0 \quad \text{in } Y^*. \tag{5.11}$$

Introducing the operator $S := B^*A^{-1}B : Y \to Y^*$, we thus need to find the solution $y \in Y$ of the variational equation

$$\langle Sy, z\rangle_H = \langle B^*A^{-1}f, z\rangle_H \quad \text{for all } z \in Y,$$

which admits a unique solution by the Lemma of Lax–Milgram (Theorem 2.3), as $S : Y \to Y^*$ is bounded, self-adjoint and elliptic, see Lemma 3.4. Now let us introduce the auxiliary variable $p = A^{-1}(f - By) \in X$. Noting that $p \equiv 0$, we can equivalenty phrase (5.11) as saddle point formulation to find $(p, y) \in X \times Y$ such that

$$\begin{array}{rcrcll} \langle Ap, q\rangle_H & + & \langle By, q\rangle_H & = & \langle f, q\rangle_H & \text{for all } q \in X, \\ -\langle B^*p, z\rangle_H & & & = & 0 & \text{for all } z \in Y. \end{array} \tag{5.12}$$

5.2.1 Discretization

Let us consider the conforming finite dimensional trial spaces $Y_H \subset Y$ and $X_h \subset X$. Then the discrete variational formulation of (5.12) is to find $(p_h, y_H) \in X_h \times Y_H$ such that

$$\begin{array}{rcrcll} \langle Ap_h, q_h\rangle_H & + & \langle By_H, q_h\rangle_H & = & \langle f, q_h\rangle_H & \text{for all } q_h \in X_h, \\ -\langle B^*p_h, z_H\rangle_H & & & = & 0 & \text{for all } z_H \in Y_H. \end{array} \tag{5.13}$$

Remark 5.3.

- *Let us stress that, due to the saddle point structure, in this case it is* not *required that dim*(X_h) = *dim*(Y_H) *holds to expect solvability of the system* (5.13). *On the contrary, this even increases the flexibility when looking for stable pairs of trial spaces, and it will be mandatory to obtain an error estimator, as we will see later on.*
- *Although, on the continuous level* $p \equiv 0$ *by construction, in general* $p_h \neq 0$, *which enables us to use* p_h *as an error estimator for* $\|y - y_H\|_Y$. *We will prove, its efficiency and reliability.*

Assuming a discrete inf-sup stability, we can prove unique solvability and error estimates for the discrete variational formulation (5.13).

Theorem 5.4 ([70, cf Lemma 2.8]). *Let* $f \in X^*$ *and assume the discrete inf-sup stability*

$$\exists \tilde{c}_1^B > 0 : \tilde{c}_1^B \|y_H\|_Y \leq \sup_{0 \neq q_h \in X_h} \frac{\langle By_H, q_h\rangle_H}{\|q_h\|_X}, \quad \textit{for all } y_H \in Y_H. \tag{5.14}$$

Then, the variational formulation (5.13) *admits a unique solution* $(p_h, y_H) \in X_h \times Y_H$. *Moreover, the error estimate*

$$\|y - y_H\|_Y \leq \left[1 + \frac{c_2^B}{\tilde{c}_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right)\right] \inf_{z_H \in Y_H} \|y - z_H\|_Y$$

holds, where $y \in Y$ *denotes the unique solution of* (5.1).

Proof. Since $A : X \to X^*$ is self-adjoint, bounded and elliptic, unique solvability follows from the assumed inf-sup stability (5.14) and Theorem 2.5. To show the error estimate, let $z_H \in Y_H$ be arbitrary but fixed. Then, by (5.14) and (5.13), using that $By = f$, we have

$$\begin{aligned}
\tilde{c}_1^B \|z_H - y_H\|_Y &\leq \sup_{0 \neq q_h \in X_h} \frac{\langle B(z_H - y_H), q_h \rangle_H}{\|q_h\|_X} \\
&= \sup_{0 \neq q_h \in X_h} \frac{\langle B(z_H - y), q_h \rangle_H + \langle A p_h, q_h \rangle_H}{\|q_h\|_X} \\
&\leq c_2^B \|z_H - y\|_Y + c_2^A \|p_h\|_X.
\end{aligned}$$

and further, using that $\langle B z_H, p_h \rangle_H = 0$ for all $z_H \in Y_H$, we compute

$$\begin{aligned}
c_1^A \|p_h\|_X^2 &\leq \langle A p_h, p_h \rangle_H = \langle B(y - y_H), p_h \rangle_H \\
&= \langle B(y - z_H), p_h \rangle_H \leq c_2^B \|y - z_H\|_Y \|p_h\|_X^2,
\end{aligned}$$

i.e.,

$$\|z_H - y_H\|_Y \leq \frac{c_2^B}{\tilde{c}_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right) \|y - z_H\|_Y.$$

By a triangle inequality

$$\|y - y_H\|_Y \leq \|y - z_H\|_Y + \|z_H - y_H\|_Y,$$

we conclude the estimate. □

REMARK 5.5. *In some applications, e.g., [106], it might be useful to define a discretization dependent norm on the ansatz space* Y *to achieve the discrete inf-sup stability* (5.14). *More precisely, let* $\|\cdot\|_{Y,h} : Y \to \mathbb{R}$ *define a norm on* Y_H *for which* $\|z_H\|_{Y,h} \leq \|z_H\|_Y$ *holds for all* $z_H \in Y_H$ *and assume that*

$$\exists \hat{c}_1^B > 0 : \hat{c}_1^B \|y_H\|_{Y,h} \leq \sup_{0 \neq q_h \in X_h} \frac{\langle B y_H, q_h \rangle_H}{\|q_h\|_X}, \quad \textit{for all } y_H \in Y_H.$$

Then the statements of Theorem 5.4 remain valid, but the error estimate becomes

$$\|y - y_H\|_{Y,h} \leq \left[1 + \frac{c_2^B}{\hat{c}_1^B}\left(1 + \frac{c_2^A}{c_1^A}\right)\right] \inf_{z_H \in Y_H} \|y - z_H\|_Y,$$

i.e., we only get the quasi-optimal estimates in the discretization dependent norm.

Recall, that by Assumption (B2) we have that

$$\exists c_1^B > 0 : \ c_1^B \|y_H\|_Y \le \sup_{0\neq q \in X} \frac{\langle B y_h, q\rangle_H}{\|q\|_X} \quad \text{for all } y_H \in Y_H \subset Y.$$

Thus, if the test space X_h is chosen rich enough, we can guarantee that the discrete inf-sup stability (5.14) holds true. The next lemma will shed light on how rich we need to choose X_h.

LEMMA 5.6 ([70, Theorem 2.7]). *For a given finite dimensional ansatz space $Y_H \subset Y$ let $X_h \subset X$ be such that*

$$\inf_{q_h \in X_h} \|p_{z_H} - q_h\|_A \le \delta \|p_{z_H}\|_A, \quad \textit{for all } z_H \in Y_H, \tag{5.15}$$

where $p_{z_H} = A^{-1} B z_H \in Y$. Then there holds the discrete inf-sup stability condition

$$c_1^B \sqrt{\frac{c_1^A}{c_2^A}} (1-\delta) \|z_H\|_Y \le \sup_{0 \neq q_h \in X_h} \frac{\langle B z_H, q_h\rangle_H}{\|q_h\|_X} \quad \textit{for all } z_H \in Y_H.$$

Proof. Firstly, let us recall the norm equivalences for the norms induced by the elliptic, bounded and self-adjoint operators $A : X \to X^*$ and $S := B^* A^{-1} B : Y \to Y^*$, i.e.,

$$\sqrt{c_1^A} \|q\|_X \le \|q\|_A := \sqrt{\langle A q, q\rangle_H} \le \sqrt{c_2^A} \|q\|_X, \quad \text{for all } q \in X,$$

and

$$\frac{c_1^B}{\sqrt{c_2^A}} \|z\|_Y \le \|z\|_S := \sqrt{\langle S z, z\rangle_H} \le \frac{c_2^B}{\sqrt{c_1^A}} \|z\|_Y, \quad \text{for all } z \in Y,$$

see (5.9) and Lemma 3.4. For $z_H \in Y_H$ arbitrary but fixed, note, that we can compute

$$\|p_{z_H}\|_A^2 = \langle A p_{z_H}, p_{z_H}\rangle_H = \langle B z_H, A^{-1} B z_H\rangle_H = \|z_H\|_S^2.$$

Further, let $p_{z_H h} \in X_h$ be the unique solution of

$$\langle A p_{z_H h}, q_h\rangle_H = \langle B z_H, q_h\rangle_H \quad \text{for all } q_h \in X_h,$$

for which we have the stability estimate

$$\|p_{z_H h}\|_A \le \|p_{z_H}\|_A$$

and Cea's Lemma, i.e., the quasi-best approximation estimate

$$\|p_{z_H} - p_{z_H h}\|_A \le \inf_{q_h \in X_h} \|p_{z_H} - q_h\|_A.$$

Together with assumption (5.15), we now compute that

$$\begin{aligned}\langle Bz_H, p_{z_H h}\rangle_H &= \langle Ap_{z_H}, p_{z_H}\rangle_H - \langle Ap_{z_H}, p_{z_H} - p_{z_H h}\rangle_H \\ &\geq \|p_{z_H}\|_A^2 - \|p_{z_H}\|_A \|p_{z_H} - p_{z_H h}\|_A \\ &\geq \|p_{z_H}\|_A^2 - \delta \|p_{z_H}\|_A^2 \\ &\geq (1-\delta)\|p_{z_H}\|_A \|p_{z_H h}\|_A \\ &\geq c_1^B \sqrt{\frac{c_1^A}{c_2^A}}(1-\delta)\|z_H\|_Y \|p_{z_H h}\|_X,\end{aligned}$$

which gives

$$c_1^B \sqrt{\frac{c_1^A}{c_2^A}}(1-\delta)\|z_H\|_Y \leq \frac{\langle Bz_H, p_{z_H h}\rangle_H}{\|p_{z_H h}\|_X} \leq \sup_{0\neq q_h \in X_h} \frac{\langle Bz_H, q_h\rangle_H}{\|q_h\|_X},$$

and concludes the proof. □

We stress again, that using the least squares framework comes with the advantage of *not* requiring $\dim(Y_H) = \dim(X_h)$, which gives us more flexibility for the choice of spaces that meet the condition (5.15). Although, it also comes with the drawback of introducing an additional unknown $p_h \in X_h$, which in terms of the numerical treatment has the disadvantage of more degrees of freedom in our system of equations. Though, in the following we will show that we can actually use p_h as an error estimator, to drive an adaptive scheme.

LEMMA 5.7 ([70, Lemma 2.5]). *Let $(p_h, y_H) \in X_h \times Y_H$ be the unique solution of* (5.13)*. Then,*

$$\|p_h\|_X \leq \frac{c_2^B}{c_1^A}\|y - y_H\|_Y.$$

Proof. Testing the first line in (5.13) with $q_h = p_h$ and using that $f = By$, we get

$$c_1^A \|p_h\|_X^2 \leq \langle Ap_h, p_h\rangle_H = \langle B(y-y_H), p_h\rangle \leq c_2^B \|y - y_H\|_Y \|p_h\|_X,$$

which already gives the desired result. □

While this shows the efficiency of the error estimator, the reliability is more cumbersome and needs an additional setup. Therefore, let us consider yet another finite dimensional space, that fulfills $Y_H \subset \overline{Y}_H \subset Y$ and assume that the discrete inf-sup stability

$$\exists \overline{c}_1^B > 0 : \overline{c}_1^B \|\overline{z}_H\|_Y \leq \sup_{0\neq q_h \in X_h} \frac{\langle B\overline{z}_H, q_h\rangle_H}{\|q_h\|_X}, \quad \text{for all } \overline{z}_H \in \overline{Y}_H \tag{5.16}$$

holds true. Then there exists a unique solution $(\overline{p}_h, \overline{y}_H) \in X_h \times \overline{Y}_H$ of

$$\begin{array}{rclcll} \langle A\overline{p}_h, q_h\rangle_H & + & \langle B\overline{y}_H, q_h\rangle_H & = & \langle f, q_h\rangle_H & \text{for all } q_h \in X_h, \\ -\langle B^*\overline{p}_h, \overline{z}_H\rangle_H & & & = & 0 & \text{for all } \overline{z}_H \in \overline{Y}_H. \end{array} \tag{5.17}$$

Now, we are in the position to the prove that p_h is a reliable error estimator, when assuming a saturation assumption.

Lemma 5.8 ([70, Lemma 2.6]). *Let $(p_h, y_H) \in X_h \times Y_H$ and $(\overline{p}_h, \overline{y}_H) \in X_h \times \overline{Y}_H$ be the unique solutions of* (5.13) *and* (5.17)*, respectively. Assume that the saturation assumption*

$$\|y - \overline{y}_H\|_Y \le \eta \|y - y_H\|_Y, \quad \textit{for some } \eta \in (0,1) \tag{5.18}$$

holds, where $y \in Y$ denotes the unique solution of (5.1)*. Then,*

$$\|y - y_H\|_Y \le \frac{[c_2^A]^2}{[\overline{c}_1^B]^2} \frac{c_2^B}{c_1^A} \frac{1}{1-\eta} \|p_h\|_X.$$

Remark 5.9.

- *Note, that since $Y_H \subset \overline{Y}_H$, we expect that $\overline{y}_H \in \overline{Y}_H$ is a better approximation to $y \in Y$ than $y_H \in Y_H$ is. Thus it makes sense to assume the saturation assumption* (5.18)*. Though, even for explicit examples, it is hard to justify that such an assumption holds, see [22].*
- *The solution $\overline{y}_H \in \overline{Y}_H$ is only needed for the theoretical treatment. Thus, the space $\overline{Y}_H$ can be arbitrarily non-constructive.*
- *Also note, that* (5.16) *implies* (5.14).

Proof. When subtracting (5.13) from (5.17) we obtain the Galerkin orthogonality

$$\langle B(\overline{y}_H - y_H), q_h\rangle_H = \langle A(p_h - \overline{p}_h), q_h\rangle_H, \quad \text{for all } q_h \in X_h.$$

Thus, for the difference $\overline{y}_H - y_H \in \overline{Y}_H$, we get, using the discrete inf-sup stability (5.16) that

$$\overline{c}_1^B \|\overline{y}_H - y_H\|_Y \le \sup_{0 \ne q_h \in X_h} \frac{\langle B(\overline{y}_H - y_H), q_h\rangle_H}{\|q_h\|_X} = \sup_{0 \ne q_h \in X_h} \frac{\langle A(p_h - \overline{p}_h), q_h\rangle_H}{\|q_h\|_Y} \le c_2^A \|p_h - \overline{p}_h\|_X.$$

Moreover, by the second line of (5.17) we have that $\langle B(\overline{z}_H - y_H), \overline{p}_h\rangle_H = 0$ for all $\overline{z}_H \in \overline{Y}_H$ and we can estimate

$$\begin{aligned} c_1^A \|p_h - \overline{p}_h\|_X^2 &\le \langle A(p_h - \overline{p}_h), p_h - \overline{p}_h\rangle_H = \langle B(\overline{y}_H - y_H), p_h - \overline{p}_h\rangle_H \\ &= \langle B(\overline{y}_H - y_H), p_h\rangle_H \le c_2^B \|\overline{y}_H - y_H\|_Y \|p_h\|_X. \end{aligned}$$

Altogether, we obtain

$$\|\overline{y}_H - y_H\|_Y \le \frac{[c_2^A]^2}{[\overline{c}_1^B]^2}\frac{c_2^B}{c_1^A}\|p_h\|_X$$

and, using a triangle inequality and the saturation assumption (5.18), gives

$$\|y - y_H\|_Y \le \|y - \overline{y}_H\|_Y + \|\overline{y}_H - y_H\|_Y \le \eta\|y - y_H\|_Y + \frac{[c_2^A]^2}{[\overline{c}_1^B]^2}\frac{c_2^B}{c_1^A}\|p_h\|_X,$$

which concludes the proof. □

5.3 A hyperbolic model problem

To forge a bridge to the motivational example, we will apply the abstract framework just developed to the one-dimensional wave equation in this section. Nevertheless, we like to point put, that its capacity ranges way further, paving the way for finite element methods for elliptic and parabolic problems, as discussed in [70], and even for boundary element methods for elliptic problems [107] and for hyperbolic problems [65].

Now, recall that we already derived the functional analytic setting for the wave equation, which reads to find $y \in \mathcal{H}_{0;0,}(Q)$ such that

$$\widetilde{B}y = f \quad \text{in } [H^{1,1}_{0;,0}(Q)]^*,$$

for given $f \in [H^{1,1}_{0;,0}(Q)]^*$, which admits a unique solution, as $\widetilde{B} : \mathcal{H}_{0;0,}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ satisfies the Assumptions (B1)-(B3). Moreover, in Section 4.2.1, we already saw that the operator $A : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ defined as

$$\langle Ap, q\rangle_Q = \langle \partial_t p, \partial_t q\rangle_{L^2(Q)} + \langle \nabla_x p, \nabla_x q\rangle_{L^2(Q)},$$

satisfies Assumptions (A1)-(A3) with constants $c_1^A = c_2^A = 1$, see Lemma 4.34, i.e.

$$\|q\|_A = \|\nabla_{(x,t)} q\|_{L^2(Q)} = \|q\|_{H^{1,1}_{0;,0}(Q)}.$$

Thus, the minimization problem (5.8) is to find

$$y = \arg\min_{z \in \mathcal{H}_{0;0,}(Q)} \frac{1}{2}\|\widetilde{B}y - f\|^2_{A^{-1}},$$

and we can derive the equivalent saddle point formulation to find $(p, y) \in H^{1,1}_{0;,0}(Q) \times \mathcal{H}_{0;0,}(Q)$ such that

$$\begin{array}{rcrcll} \langle Ap, q\rangle_Q & + & \langle \widetilde{B}y, q\rangle_Q & = & \langle f, q\rangle_Q & \text{for all } q \in H^{1,1}_{0;,0}(Q), \\ -\langle \widetilde{B}^* p, z\rangle_Q & & & = & 0 & \text{for all } z \in \mathcal{H}_{0;0,}(Q). \end{array} \tag{5.19}$$

For the discretization, we choose

$$Y_H = S_H^1(\mathcal{T}_H) \cap H^{1,1}_{0;0,}(Q) = \operatorname{span}\{\varphi_i^Y\}_{k=1}^{M_Y} \subset \mathcal{H}_{0;0,}(Q),$$

and

$$X_h = S_h^1(\mathcal{T}_h) \cap H^{1,1}_{0;,0}(Q) = \operatorname{span}\{\varphi_i^X\}_{k=1}^{M_X},$$

as spaces of piecewise linear, globally continuous functions defined on two admissible and shape regular triangulations $\mathcal{T}_h = \{\tau_{h,\ell}\}_{\ell=1}^{N_h}$ and $\mathcal{T}_H = \{\tau_{H,\ell}\}_{\ell=1}^{N_H}$, which we assume to be nested, i.e. $\tau_{H,\ell} = \bigcup_{j=1}^{4^L} \tau_{h,j}$, for some $L \geq 0$. The discrete variational formulation is then to find $(p_h, y_H) \in X_h \times Y_H$ such that

$$\begin{array}{rcrcll} \langle Ap_h, q_h\rangle_Q & + & \langle \widetilde{B}y_H, q_h\rangle_Q & = & \langle f, q_h\rangle_Q & \text{for all } q_h \in X_h, \\ -\langle \widetilde{B}^* p_h, z_H\rangle_Q & & & = & 0 & \text{for all } z_H \in Y_H. \end{array} \tag{5.20}$$

In order to admit a unique solution, we have to make sure that the discrete inf-sup stability condition holds true. Although, we will not give a rigorous analysis, let us recall, that the operator $A : H^{1,1}_{0;,0}(Q) \to [H^{1,1}_{0;,0}(Q)]^*$ corresponds to the space time Laplacian with mixed Dirichlet and Neumann boundary conditions. Now fix $H > 0$ and take an arbitrary but fixed $z_H \in Y_H$. It is well-known that $Y_H \subset H^{3/2-\varepsilon}(Q)$, $\varepsilon > 0$. Therefore, we get $\widetilde{B}z_H \in H^{-1/2}(Q)$ and for the solution $p_{z_H} \in H^{1,1}_{0;,0}(Q)$ of $Ap_{z_H} = \widetilde{B}z_H$ we expect that $p_{z_H} \in H^{1+r}(Q) \cap H^{1,1}_{0;,0}(Q)$, $r > 0$. Hence, using the best approximation error estimate (Theorem 2.36), we have that

$$\inf_{q_h \in X_h} \|p_{z_H} - q_h\|_A \leq ch^r \|p_{z_H}\|_{H^{1+r}(Q)} \leq \delta \|p_{z_H}\|_A, \quad \delta \in (0,1),$$

if h is sufficiently small with respect to H, i.e., by Lemma 5.6, the discrete inf-sup stability can be achieved if $h = H/2^L$ for some sufficiently large $L > 0$. Then, using that $\|y\|_{\mathcal{H}_{0;0,}(Q)} \leq \|y\|_{H^{1,1}_{0;0,}(Q)}$ for all $y \in H^{1,1}_{0;0,}(Q)$, see (4.133), by Theorem 5.4 and the best approximation properties (Theorem 2.36) we immediately get the following result.

THEOREM 5.10 ([70, Theorem 5.1]). *For any $f \in [H^{1,1}_{0;,0}(Q)]^*$ (5.20) admits a unique solution $(p_h, y_H) \in X_h \times Y_H$, if $h = H/2^L$ and $L > 0$ is sufficiently large. Moreover, if the unique solution of (5.1) satisfies $y \in H^s(Q) \cap H^{1,1}_{0;0,}(Q)$, then*

$$\|y - y_H\|_{\mathcal{H}_{0;0,}(Q)} \leq c \inf_{z_H \in Y_H} \|y - z_H\|_{\mathcal{H}_{0;0,}(Q)} \leq c \inf_{z_H \in Y_H} \|y - z_H\|_{H^{1,1}_{0;0,}(Q)} \leq cH^s \|y\|_{H^s(Q)}.$$

5.3.1 Numerical results

Using the fe-isomorphism, we need to solve the equivalent system of linear equations

$$\begin{pmatrix} A_h & B_h \\ -B_h^\top & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p}_h \\ \mathbf{y}_H \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ \mathbf{0}_H \end{pmatrix}, \tag{5.21}$$

with matrices

$$\begin{aligned} A_h[i,j] &= \langle \nabla_{(x,t)}\varphi_j^X, \nabla_{(x,t)}\varphi_i^X \rangle_{L^2(Q)}, \\ B_h[i,k] &= -\langle \partial_t \varphi_k^Y, \partial_t \varphi_i^X \rangle_{L^2(Q)} + \langle \nabla_x \varphi_k^Y, \nabla_x \varphi_i^X \rangle_{L^2(Q)}, \end{aligned}$$

for $i, j = 1, \ldots, M_X$ and $k = 1, \ldots, M_Y$ and load vector

$$\mathbf{f}[i] = \langle f, \varphi_i^X \rangle_{L^2(Q)}, \quad i = 1, \ldots, M_X.$$

REMARK 5.11. *Note, that in general the matrix B_h in* (5.21) *is* not *square. On the contrary, if $h = H$, then B_h is square and if the discrete inf-sup stability holds true, then $p_h = 0$ and we end up with the discrete variational formulation of the direct formulation, i.e., to find $y_h \in Y_h$ s.t.*

$$\langle \widetilde{B} y_h, q_h \rangle_Q = \langle f, q_h \rangle_Q \quad \textit{for all } q_h \in X_h.$$

Hence, in order to obtain an error estimator we need to have $h < H$.

As a first test example, let us again consider (5.7) on a structured triangulation of the space-time domain $Q = (0, L) \times (0, T)$. To gain stability, we choose $h = H/2$. The results depicted in Table 5.2 clearly indicate, that we overcome the CFL-condition $H_t \leq H_x$, that was present in the solution of the direct formulation. Moreover, we see optimal convergence rates in the energy norm, as predicted by the theory.

Secondly, to show the capacity of the error estimator, we consider the test example, see Figure 5.2,

$$y_2(x,t) := \begin{cases} \frac{1}{2}(t - x - 2)^3 (x - t)^3 \sin \frac{\pi}{3} x & \text{for } x \leq t \text{ and } t - x \leq 2, \\ 0 & \text{else,} \end{cases} \tag{5.22}$$

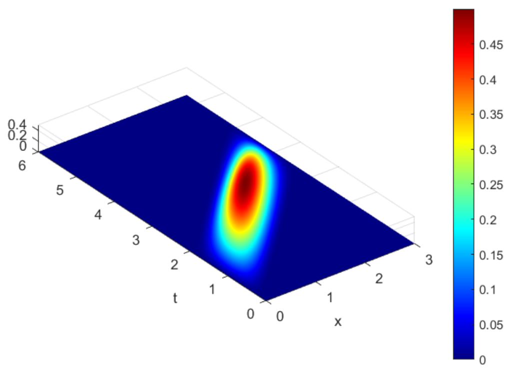
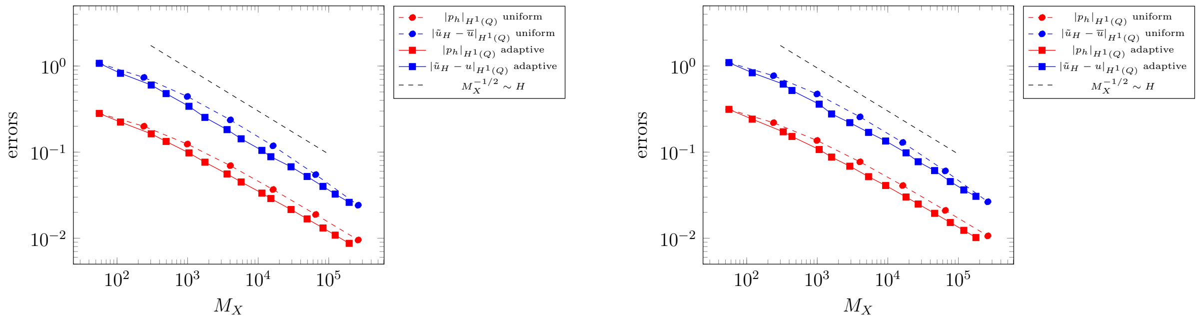


Figure 5.2: Test example y_2 (5.22)

N_H	H_t	H_x	$\lvert y_1 - y_H \rvert_{H^1(Q)}$	eoc
8	0.500	0.250	$7.05 \cdot 10^{-1}$	0.000
32	0.250	0.125	$5.63 \cdot 10^{-1}$	0.323
128	0.125	0.063	$3.43 \cdot 10^{-1}$	0.715
512	0.063	0.031	$1.68 \cdot 10^{-1}$	1.028
2,048	0.031	0.016	$7.63 \cdot 10^{-2}$	1.141
8,192	0.016	0.008	$3.44 \cdot 10^{-2}$	1.150
32,768	0.008	0.004	$1.61 \cdot 10^{-2}$	1.097

Table 5.2: Computation for the test example (5.7) solving (5.21), with $h = H/2$, on a mesh violating the CFL-condition $H_t \leq H_x$.



a) $h = H/2$ b) $h = H/4$

Figure 5.3: Errors $|u - \widetilde{u}_H|_{H^1(Q)}$ and estimators $|p_h|_{H^1(Q)}$ for adaptive and uniform refinements for different choices of h and H for the smooth solution (5.22).

on the space-time domain $Q = (0,3) \times (0,6)$. We apply a Dörfler marking scheme, where we define the error estimator for the local error by

$$\eta_\ell^2 = \|\nabla_{(x,t)} p_h\|^2_{L^2(\tau_{H,\ell})},$$

which fulfills

$$\sum_{\ell=1}^{N_H} \eta_\ell^2 = \sum_{\ell=1}^{N_H} \|\nabla_{(x,t)} p_h\|^2_{L^2(\tau_{H,\ell})} = \|\nabla_{(x,t)} p_h\|^2_{L^2(Q)},$$

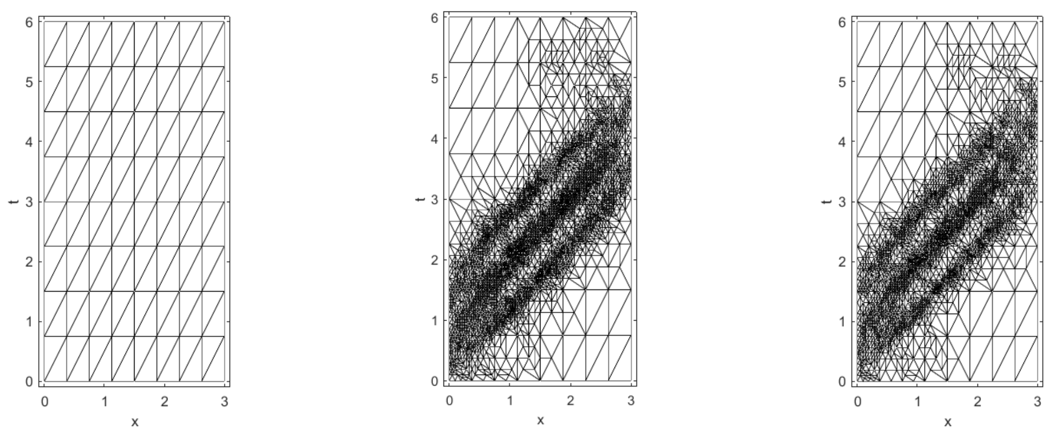
and refine all elements $\tau_{H,\ell}$, that satisfy

$$\eta_\ell > \theta \max_{i=1,\ldots,N_H} \eta_i,$$

where we choose $\theta = 0.5$. In view of Lemma 5.7, we have that η_ℓ is efficient. To get an reliable error estimator, note that the pairing X_h and Y_H was stable for the choice $h = H/2$. Hence, if we choose $h = H/4$, then $\overline{Y}_H = Y_{H/2}$ satisfies the discrete inf-sup stability condition (5.16) and $Y_H \subset Y_{H/2} \subset \mathcal{H}_{0;0,}(Q)$. If in addition the saturation assumption (5.18) holds, which we would expect, since we compute the solution $\overline{y}_H \in Y_{H/2}$ on a finer mesh, the estimator is reliable. Figure 5.3, shows optimal orders of convergence for a uniform refinement and improved results when driving an adaptive scheme, with the error estimator η_ℓ. Moreover, we see that the curves of the error estimator and the actual error are parallel, supporting, that the indicator works well. Surprisingly, the convergence also reveals that we already get an error estimator for the choice $h = H/2$. Thus, there seems to be a space $\overline{Y}_H$, that meets the assumptions of Lemma 5.8, e.g., one might think of adding the bubble functions

$$\varphi_\ell^B(x,t) = \begin{cases} \prod_{\{k=1,\ldots,M_Y : x_k \in \tau_{H,\ell}\}} \varphi_k^Y(x,t), & (x,t) \in \tau_{H,\ell}, \\ 0, & \text{else}, \end{cases}$$

for $\ell = 1, \ldots, N_H$, to Y_H to obtain $\overline{Y}_H$, but as already mentioned, this space can be arbitrarily non constructive. The adaptively refined meshes for the test example (5.22) are shown in Figure 5.4. We point out, that all the numerical examples were carried out using MATLAB, where we adapted refinement routines from [44].



a) initial, $M_X = 56$ b) $h = H/2$, $M_X = 3624$ c) $h = H/4$, $M_X = 2893$

Figure 5.4: Meshes from the adaptive refinement for the solution (5.22).

6 Conclusions and Outlook

In this thesis, we tackled two different types of problems. Firstly, we formulated and analyzed an abstract framework for distributed optimal control problems subject to the operator constraint $By_\varrho = u_\varrho$. Assuming that $B : Y \to X^*$ defines an isomorphism, we were able to derive regularization error estimates on an abstract level, as was first done for the Poisson equation [95]. Moreover, we investigated the discrete setting, showing stability for conforming trial spaces and quasi-best approximation error estimates. The incorporation of state or control constraints was observed to be equivalent to abstract variational inequalities, for which the analysis for the continuous and the discretized setting were carried out. To show the capacity of the abstract setting, we considered an elliptic and a hyperbolic distributed optimal control problem, where in both cases we compared different spaces for the control. For the elliptic problem, we also gave an alternative point of view for the L^2-regularization, fitting into the energy regularization approach and shed light on the differences between the energy approach and the common approach. All discussions were complemented by numerical examples, supporting the theoretical findings. Moreover, we proposed adaptive finite element schemes and an adaptive choice of the cost parameter and presented its performance for target functions of different regularity, including discontinuous targets. The findings show, that the proposed methods work equally well, for the elliptic and for the hyperbolic case, where for the latter we used a space time analysis and totally unstructured space time finite elements, leading to full space time adaptivity.

Secondly, we considered a least squares framework for the direct solution of the operator equation $By = f$. Again, considering that $B : Y \to X^*$ defines an isomorphism, we gave an abstract analysis for the continuous and discrete setting, including stability and related quasi best approximation error estimates. Assuming a saturation assumption, we were able to show efficiency and reliability of the inbuilt error estimator provided by the least squares formulation. As a model problem we considered the space time formulation of the wave equation. Numerical examples were carried out supporting the theoretical findings. In particular, we observed unconditional stability and were able to drive a space time adaptive scheme on fully unstructured simplicial space time finite element meshes.

As most applications require to consider two or three spatial dimensions, i.e., the space time domain is in $n = d + 1 = 3$ or $n = 4$ dimensions, there is a particular need for fast solvers of the arising systems. For optimal control problems we derived the optimal choice $\varrho = h^{\mathrm{ord}(S)}$ for the cost/regularization parameter, where $\mathrm{ord}(S)$ denotes the order of the differential operator $S = B^*A^{-1}B : Y \to Y^*$, in the case of the energy regularization. This not only leads to optimal orders of convergence of the method, but also implies that the Schur complement matrix of the optimality system is spectrally equivalent to the mass matrix. In particular, for iterative solvers, we can use the lumped mass matrix as preconditioner, leading to robust solvers of optimal complexity, see [75, 80, 81]. Although, in the least squares framework the Schur complement is also symmetric and positive definite, the design of robust preconditioning seems more involved and is at this time still an open task. Moreover, the extension of the proposed methods to non-linear problems is of practical interest, to integrate space time methods in applications, e.g., electrical machines, see [49].

References

[1] R. A. Adams, *Sobolev spaces*, Pure and Applied Mathematics, Vol. 65, Academic Press, New York-London, 1975.

[2] R. Andreev, *Stability of sparse space-time finite element discretizations of linear parabolic evolution equations*, IMA J. Numer. Anal., 33 (2013), pp. 242–260.

[3] T. Apel and J. Melenk, *Interpolation and quasi-interpolation in h- and hp-version finite element spaces*, in Encyclopedia of Computational Mechanics, second edition, E. Stein, R. de Borst, T. Hughes (eds.), 2017, pp. 1–33.

[4] T. Apel, O. Steinbach, and M. Winkler, *Error estimates for Neumann boundary control problems with energy regularization*, J. Numer. Math., 24 (2016), pp. 207–233.

[5] J. Argyris and D. Scharpf, *Finite elements in time and space*, Nuclear Engineering and Design, 10 (1969), pp. 456–464.

[6] I. Babuška, *Error-bounds for finite element method*, Numer. Math., 16 (1970/71), pp. 322–333.

[7] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972), Academic Press, New York, 1972, pp. 1–359.

[8] W. Bangerth, M. Geiger, and R. Rannacher, *Adaptive Galerkin finite element methods for the wave equation*, Comput. Methods Appl. Math., 10 (2010), pp. 3–48.

[9] W. Bangerth and R. Rannacher, *Adaptive finite element techniques for the acoustic wave equation*, J. Comput. Acoust., 9 (2001), pp. 575–591.

[10] N. Beranek, M. Reinhold, and K. Urban, *A space–time variational method for optimal control problems: well-posedness, stability and numerical solution*, Comput. Optim. Appl., 86 (2023), pp. 767—-794.

[11] J. Bergh and J. Löfström, *Interpolation spaces. An introduction*, Grundlehren der Mathematischen Wissenschaften, No. 223, Springer-Verlag, Berlin-New York, 1976.

[12] C. Bernardi, *Optimal finite-element interpolation on curved domains*, SIAM Journal on Numerical Analysis, 26 (1989), pp. 1212–1240.

[13] M. Bernreuther and S. Volkwein, *An adaptive certified space-time reduced basis method for nonsmooth parabolic partial differential equations*, 2022. arXiv:2212.13744.

[14] P. Bochev and M. Gunzburger, *Least-squares methods for hyperbolic problems*, in Handbook of numerical methods for hyperbolic problems, vol. 17 of Handb. Numer. Anal., Elsevier/North-Holland, Amsterdam, 2016, pp. 289–317.

[15] P. B. Bochev and M. D. Gunzburger, *Least-squares finite element methods*, vol. 166 of Applied Mathematical Sciences, Springer, New York, 2009.

[16] J. H. Bramble, J. E. Pasciak, and O. Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp., 71 (2002), pp. 147–156.

[17] S. C. Brenner, *Finite element methods for elliptic distributed optimal control problems with pointwise state constraints (survey)*, Advances in Mathematical Sciences, 21 (2020), pp. 3–16.

[18] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, 1994.

[19] H. R. Brezis and G. Stampacchia, *Sur la régularité de la solution d'inéquations elliptiques*, Bull. Soc. Math. France, 96 (1968), pp. 153–180.

[20] E. Burman, P. Hansbo, and M. G. Larson, *Solving ill-posed control problems by stabilized finite element methods: an alternative to Tikhonov regularization*, Inverse Problems, 34 (2018).

[21] C. Carstensen, *Merging the Bramble-Pasciak-Steinbach and the Crouzeix-Thomée criterion for H^1-stability of the L^2-projection onto finite element spaces*, Math. Comp., 71 (2002), pp. 157–163.

[22] C. Carstensen, D. Gallistl, and J. Gedicke, *Justification of the saturation assumption*, Numer. Math., 134 (2016), pp. 1–25.

[23] P. G. Ciarlet, *The finite element method for elliptic problems*, vol. Vol. 4 of Studies in Mathematics and its Applications, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.

[24] C. Clason and B. Kaltenbacher, *Optimal control and inverse problems*, Inverse Problems, 36 (2020).

[25] P. Clément, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér., 9 (1975).

[26] M. Costabel, *Boundary integral operators for the heat equation*, Integral Equations Operator Theory, 13 (1990), pp. 498–552.

[27] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[28] R. Courant, K. Friedrichs, and H. Lewy, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Mathematische Annalen, 100 (1928), pp. 32–74.

[29] W. Dahmen, H. Monsuur, and R. Stevenson, *Least squares solvers for ill-posed PDEs that are conditionally stable*, ESAIM: Math. Model. Numer. Anal., 57 (2023), pp. 2227–2255.

[30] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, vol. 1341 of Lecture Notes in Mathematics, Springer Berlin, Heidelberg, 1988.

[31] C. de Boor, *A practical guide to splines*, vol. 27 of Applied Mathematical Sciences, Springer-Verlag, New York-Berlin, 1978.

[32] L. Demkowicz, J. Gopalakrishnan, S. Nagaraj, and P. Sepúlveda, *A Spacetime DPG Method for the Schrödinger Equation*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 1740–1759.

[33] P. Deuflhard and F. Bornemann, *Numerische Mathematik. II*, de Gruyter Lehrbuch, Walter de Gruyter & Co., Berlin, 1994. Integration gewöhnlicher Differentialgleichungen.

[34] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[35] W. Dörfler, C. Wieners, and D. Ziegler, *Space-time discontinuous Galerkin methods for linear hyperbolic systems and the application to the forward problem in seismic imaging*, in Finite volumes for complex applications IX—methods, theoretical aspects, examples—FVCA 9, Bergen, Norway, June 2020, vol. 323 of Springer Proc. Math. Stat., Springer, Cham, 2020, pp. 477–485.

[36] T. Dupont and R. Scott, *Polynomial Approximation of Functions in Sobolev Spaces*, Mathematics of Computation, 34 (1980), pp. 441–463.

[37] B. Endtmayer, U. Langer, and A. Schafelner, *Goal-oriented Adaptive Space-Time Finite Element Methods for Regularized Parabolic p-Laplace Problems*, 2023. arXiv:2306.07167.

[38] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, vol. 375 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1996.

[39] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, 2004.

[40] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[41] R. S. Falk, *Error estimates for the approximation of a class of variational inequalities*, Math. Comput., 28 (1974), pp. 963–971.

[42] H. Feichtinger, *Banach Gelfand Triples for Applications in Physics and Engineering*, in AIP Conference Proceedings, vol. 1146 (1), 2009, p. 189–228.

[43] T. Führer and M. Karkulik, *Space-time least-squares finite elements for parabolic equations*, Comput. Math. Appl., 92 (2021), pp. 27–36.

[44] S. Funken, D. Praetorius, and P. Wissgott, *Efficient implementation of adaptive P1-FEM in Matlab*, Comput. Methods Appl. Math., 11 (2011), pp. 460–490.

[45] T. Führer, *On a mixed FEM and a FOSLS with H^{-1} loads*, 2023. arXiv:2210.14063.

[46] M. J. Gander, *50 years of time parallel integration*, in Multiple Shooting and Time Domain Decomposition, Springer Verlag, Heidelberg, Berlin, 2015, pp. 69–114.

[47] M. J. Gander and L.-D. Lu, *New time domain decomposition methods for parabolic control problems i: Dirichlet-neumann and neumann-dirichlet algorithms*, 2023.

[48] M. J. Gander and M. Neumüller, *Analysis of a new space-time parallel multigrid algorithm for parabolic problems*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2173–A2208.

[49] P. Gangl, M. Gobrial, and O. Steinbach, *A space-time finite element method for the eddy current approximation of rotating electric machines*, 2023. arXiv:2307.00278.

[50] P. Gangl, R. Löscher, and O. Steinbach, *Regularization and finite element error estimates for elliptic distributed optimal control problems with energy regularization and state or control constraints*, 2023. arXiv:2306.15316.

[51] I. M. Gelfand and N. Y. Vilenkin, *Generalized functions. Vol. 4*, AMS Chelsea Publishing, Providence, RI, 2016.

[52] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer Series in Computational Physics, Springer-Verlag, New York, 1984.

[53] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical analysis of variational inequalities.*, North-Holland Publishing Co., Amsterdam-New York,, 1981.

[54] W. Gong, M. Mateos, J. Singler, and Y. Zhang, *Analysis and approximations of Dirichlet boundary control of Stokes flows in the energy space*, SIAM J. Numer. Anal., 60 (2022), pp. 450–474.

[55] W. Gong and Z. Tan, *A new finite element method for elliptic optimal control problems with pointwise state constraints in energy spaces*, 2023. arXiv:2306.03246v1.

[56] J. Gopalakrishnan, M. Hochsteger, J. Schöberl, and C. Wintersteiger, *An Explicit Mapped Tent Pitching Scheme for Maxwell Equations*, in Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2018, Springer International Publishing, 2020, pp. 359–369.

[57] J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger, *Mapped Tent Pitching Schemes for Hyperbolic Systems*, SIAM Journal on Scientific Computing, 39 (2017), pp. B1043–B1063.

[58] J. Gopalakrishnan and P. Sepúlveda, *A space-time DPG method for the wave equation in multiple dimensions*, in Space-Time Methods, U. Langer and O. Steinbach, eds., De Gruyter, Berlin, Boston, 2019, pp. 117–140.

[59] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 69 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[60] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, vol. 8 of Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2008.

[61] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations I: Stiff and Differential-Algebraic Problems*, vol. 14 of Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2010.

[62] Henning, Julian, Palitta, Davide, Simoncini, Valeria, and Urban, Karsten, *An ultraweak space-time variational formulation for the wave equation: Analysis and efficient numerical solution*, ESAIM: M2AN, 56 (2022), pp. 1173–1198.

[63] R. Herzog, G. Stadler, and G. Wachsmuth, *Directional sparsity in optimal control of partial differential equations*, SIAM Journal on Control and Optimization, 50 (2012), pp. 943–963.

[64] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semi-smooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888.

[65] D. Hoonhout, R. Löscher, O. Steinbach, and C. Urzúa-Torres, *Stable least-squares space-time boundary element methods for the wave equation.* to be submitted, 2023.

[66] T. J. R. Hughes and G. M. Hulbert, *Space-time finite element methods for elastodynamics: formulations and error estimates*, Comput. Methods Appl. Mech. Engrg., 66 (1988), pp. 339–363.

[67] ——, *Space-time finite element methods for second-order hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 84 (1990), pp. 327–348.

[68] V. Isakov, *Inverse problems for partial differential equations*, vol. 127 of Applied Mathematical Sciences, Springer, Cham, third ed., 2017.

[69] A. Kunoth, T. Lyche, G. Sangalli, and S. Serra-Capizzano, *Splines and PDEs: from approximation theory to numerical linear algebra*, vol. 2219 of Lecture Notes in Mathematics, Springer, Cham; Fondazione C.I.M.E., Florence, 2018.

[70] C. Köthe, R. Löscher, and O. Steinbach, *Adaptive least-squares space-time finite element methods*, 2023. arX:2309.14300.

[71] O. A. Ladyzhenskaya, *The boundary value problems of mathematical physics*, vol. 49 of Applied Mathematical Sciences, Springer-Verlag, New York, 1985.

[72] J. Lang, *Adaptive multilevel solution of nonlinear parabolic PDE systems*, vol. 16 of Lecture Notes in Computational Science and Engineering, Springer-Verlag, Berlin, 2001. Theory, algorithm, and applications.

[73] U. Langer, R. Löscher, O. Steinbach, and H. Yang, *Robust iterative solvers for algebraic systems arising from elliptic optimal control problems.* TU Graz, Berichte aus dem Institut für Angewandte Mathematik, Bericht 2023/2, accepted.

[74] ——, *An adaptive finite element method for distributed elliptic optimal control problems with variable energy regularization*, 2023. arXiv:2209.08811.

[75] ——, *Mass-lumping discretization and solvers for distributed elliptic optimal control problems*, 2023. arXiv:2304.14664.

[76] U. Langer, R. Löscher, O. Steinbach, and H. Yang, *Robust Finite Element Discretization and Solvers for Distributed Elliptic Optimal Control Problems*, Computational Methods in Applied Mathematics, 23 (2023), pp. 989–1005.

[77] U. Langer and A. Schafelner, *Adaptive space-time finite element methods for parabolic optimal control problems*, J. Numer. Math., 30 (2022), pp. 247–266.

[78] U. Langer and O. Steinbach, eds., *Space-Time Methods*, De Gruyter, Berlin, Boston, 2019.

[79] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang, *Space-time finite element discretization of parabolic optimal control problems with energy regularization*, SIAM J. Numer. Anal., 59 (2021), pp. 675–695.

[80] U. Langer, O. Steinbach, and H. Yang, *Robust discretization and solvers for elliptic optimal control problems with energy regularization*, Comput. Meth. Appl. Math., 22 (2022), pp. 97–111.

[81] U. Langer, O. Steinbach, and H. Yang, *Robust space-time finite element error estimates for parabolic distributed optimal control problems with energy regularization*, 2022. arXiv:2206.06455.

[82] U. Langer and M. Zank, *Efficient direct space-time finite element solvers for parabolic initial-boundary value problems in anisotropic sobolev spaces*, SIAM Journal on Scientific Computing, 43 (2021), pp. A2714–A2736.

[83] P. D. Lax and A. N. Milgram, *Parabolic equations*, in Contributions to the theory of partial differential equations, Annals of Mathematics Studies, no. 33, Princeton University Press, Princeton, N.J., 1954, pp. 167–190.

[84] J.-L. Lions, *Optimal control of systems governed by partial differential equations*, Die Grundlehren der mathematischen Wissenschaften, Band 170, Springer-Verlag, New York-Berlin, 1971.

[85] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications. Vol. 1*, Travaux et Recherches Mathématiques, No. 17, Dunod, Paris, 1968.

[86] J. L. Lions and G. Stampacchia, *Variational inequalities*, Communications on Pure and Applied Mathematics, 20 (1967), pp. 493–519.

[87] R. Löscher and O. Steinbach, *Space-time finite element methods for distributed optimal control of the wave equation.* arXiv:2211.02562v1, 2023, accepted.

[88] R. Löscher, O. Steinbach, and M. Zank, *Numerical results for an unconditionally stable space-time finite element method for the wave equation*, in Domain Decomposition Methods in Science and Engineering XXVI, S. C. Brenner, E. Chung, A. Klawonn, F. Kwok, J. Xu, and J. Zou, eds., vol. 145 of Lecture Notes in Computational Science and Engineering, Cham, 2022, Springer, pp. 625–632.

[89] W. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, Cambridge, 2000.

[90] H. Monsuur, R. Stevenson, and J. Storn, *Minimal residual methods in negative or fractional Sobolev norms*, 2023. arXiv:2301.10484.

[91] S. Montaner and A. Münch, *Approximation of controls for linear wave equations: a first order mixed formulation*, Math. Control Relat. Fields, 9 (2019), pp. 729–758.

[92] F. Natterer, *Error bounds for Tikhonov regularization in Hilbert scales*, Applicable Anal., 18 (1984), pp. 29–37.

[93] M. Neumüller, *Space-Time Methods. Fast Solvers and Applications*, vol. 20, Verlag der Technischen Universität Graz, 2013.

[94] M. Neumüller and O. Steinbach, *Refinement of flexible space–time finite element meshes and discontinuous Galerkin methods*, Comput. Visual. Sci, 14 (2011), pp. 189–205.

[95] M. Neumüller and O. Steinbach, *Regularization error estimates for distributed control problems in energy spaces*, Math. Methods Appl. Sci., 44 (2021), pp. 4176–4191.

[96] J. Nečas, *Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3), 16 (1962), pp. 305–326.

[97] G. Of, T. X. Phan, and O. Steinbach, *An energy space finite element approach for elliptic Dirichlet boundary control problems*, Numer. Math., 129 (2015), pp. 723–748.

[98] I. Perugia, C. Schwab, and M. Zank, *Exponential convergence of hp-time-stepping in space-time discretizations of parabolic PDEs*, ESAIM Math. Model. Numer. Anal., 57 (2023), pp. 29–67.

[99] W. V. Petryshyn, *Constructional proof of Lax-Milgram lemma and its application to non-K-p.d. abstract and differential operator equations*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 404–420.

[100] F. PÖRNER, *Regularization Methods for Ill-Posed Optimal Control Problems*, PhD thesis, Julius–Maximilians–Universität Würzburg, 2018.

[101] E. ROTHE, *Über die Wärmeleitungsgleichung mit nichtkonstanten Koeffizienten im räumlichen Falle*, Math. Ann., 104 (1931), pp. 340–354.

[102] A. SCHAFELNER, *Space-time Finite Element Methods*, PhD thesis, Johannes Kepler Universität Linz, 2021.

[103] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[104] K. STEIH AND K. URBAN, *Space-Time Reduced Basis Methods for Time-Periodic Partial Differential Equations*, IFAC Proceedings Volumes, 45 (2012), pp. 710–715.

[105] O. STEINBACH, *Numerical approximation methods for elliptic boundary value problems*, Springer, New York, 2008.

[106] ——, *Space-time finite element methods for parabolic problems*, Comput. Methods Appl. Math., 15 (2015), pp. 551–566.

[107] O. STEINBACH, *An adaptive least squares boundary element method for elliptic boundary value problems*, 2023. Berichte aus dem Institut für Angewandte Mathematik, Bericht 2023/1, TU Graz.

[108] O. STEINBACH AND H. YANG, *Space-time finite element methods for parabolic evolution equations: discretization, a posteriori error estimation, adaptivity and solution*, in Space-time methods—applications to partial differential equations, vol. 25 of Radon Ser. Comput. Appl. Math., De Gruyter, Berlin, [2019] ©2019, pp. 207–248.

[109] O. STEINBACH AND M. ZANK, *A stabilized space–time finite element method for the wave equation*, in Advanced Finite Element Methods with Applications: Selected Papers from the 30th Chemnitz Finite Element Symposium 2017, T. Apel, U. Langer, A. Meyer, and O. Steinbach, eds., Springer International Publishing, Cham, 2019, pp. 341–370.

[110] ——, *Coercive space-time finite element methods for initial boundary value problems*, Electron. Trans. Numer. Anal., 52 (2020), pp. 154–194.

[111] ——, *A generalized inf-sup stable variational formulation for the wave equation*, J. Math. Anal. Appl., 505 (2022). Paper No. 125457, 24 pp.

[112] V. THOMÉE, *Galerkin finite element methods for parabolic problems*, vol. 25 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2006.

[113] F. Tröltzsch, *Optimal control of partial differential equations*, vol. 112 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2010.

[114] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

[115] M. Zank, *Higher-Order Space-Time Continuous Galerkin Methods for the Wave Equation*, in WCCM-ECCOMAS2020, vol. 700 of Numerical Methods and Algorithms in Science and Engineering, 2020.

[116] M. Zank, *Inf-Sup stable Space-Time Methods for Time-Dependent Partial Differential Equations*, vol. 38 of Computation in Engeneering and Science, Verlag der Technischen Universität Graz, 2020.