UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Optimal Control Problems and Algebraic Flux Correction Schemes

Dissertation zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

Universität Duisburg-Essen
Fakultät Mathematik

genehmigte Dissertation

von

Herrn Jens Baumgartner, M.Sc.
geboren in Dinslaken

Datum der Einreichung: 29.11.2021
Datum der Disputation: 15.02.2022

1. Gutachter: Herr Professor Dr. Arnd Rösch
2. Gutachter: Herr Professor Dr. Dmitri Kuzmin

DOI: 10.17185/duepublico/75439
URN: urn:nbn:de:hbz:465-20220222-105256-2

Dissertation

Optimal Control Problems and Algebraic Flux Correction Schemes

Fakultät Mathematik
Universität Duisburg-Essen

Datum der Einreichung: 29.11.2021
Datum der Disputation: 15.02.2022

1. Gutachter: Herr Professor Dr. Arnd Rösch
Universität Duisburg-Essen
Lehrstuhl Nichtlineare Optimierung

2. Gutachter: Herr Professor Dr. Dmitri Kuzmin
TU Dortmund
Lehrstuhl Angewandte Mathematik und Numerik

Zusammenfassung

In der vorliegenden Arbeit beschäftigen wir uns mit der Diskretisierung von Optimalsteuerproblemen, deren Zustandsgleichung eine Konvektion-Diffusion Reaktionsgleichung ist. Insbesondere in dem sogenannten konvektionsdominanten Fall können Lösungen solcher Gleichungen Grenzschichten enthalten, d.h. schmale Regionen mit steilen Gradienten. Zur Diskretisierung partieller Differentialgleichungen werden im Allgemeinen standardisierte Finite Elemente Methoden angewendet. Diese führen uns jedoch in dem konvektionsdominanten Fall zu Lösungen, welche nicht-physikalische Oszillationen enthalten. Dies motiviert den Einsatz von Stabilisierungstechniken, wie beispielsweise den Einsatz von algebraischen Korrekturschemata, den sogenannten Algebraic Flux Correction (AFC) Schemes. Die Hauptmotivation für die Konstruktion solcher AFC Schemata besteht in der Erfüllung des diskreten Maximumprinzips (DMP), sodass künstlich auftretende Oszillationen in den diskreten Lösungen verhindert werden. In dieser Arbeit diskretisieren wir Optimalsteuerprobleme mit Hilfe eines AFC Schemas. Im Allgemeinen werden in der Theorie der Optimalen Steuerung zur Diskretisierung der Optimierungsprobleme die Ansätze *optimize-then-discretize* und *discretize-then-optimize* herangezogen. Aufgrund der Nichtlinearität bzw. im Allgemeinen auch aufgrund der Nichtdifferenzierbarkeit der AFC Methode verwenden wir den *optimize-then-discretize*-Ansatz, d.h. wir diskretisieren die Optimalitätssysteme mit Hilfe eines AFC Schemas. Dadurch erhalten wir stabilisierte und gekoppelte Systeme. In dieser Arbeit beantworten wir die Frage nach der Lösbarkeit solcher Systeme. Zudem leiten wir L^2-Fehlerabschätzungen von den AFC Lösungen zu den optimalen Lösungen der jeweiligen kontinuierlichen Optimierungsproblemen her. Abschließend werden die theoretischen Resultate durch numerische Tests unterstützt.

Abstract

Solutions of convection-diffusion-reaction equations may possess layers, i.e. narrow regions where the solution has a large gradient (in particular for convection-dominated equations). Standard Finite Element Methods lead to discrete solutions which are polluted by spurious oscillations. The main motivation for the construction of the so-called Algebraic Flux Correction (AFC) schemes is the satisfaction of the discrete maximum principle (DMP) to avoid spurious oscillations in the discrete solutions. In this thesis, we apply the AFC method on several optimal control problems governed by a convection-diffusion-reaction equation. Due to the fact that the AFC schemes are nonlinear and usually non-differentiable, the approaches *optimize-then-discretize* and *discretize-then-optimize* do not commute. We use the *optimize-then-discretize*-approach, i.e. we discretize the state equation and the adjoint equation with the help of the AFC method. This leads us to coupled and discretized systems. We verify the existence of corresponding discrete solutions and derive L^2-error estimates for the control and the state. The stabilizing effect of the AFC method on the discrete solutions and L^2-errors are illustrated by numerical tests.

Acknowledgments

First of all, I would like to thank Arnd Rösch for giving me the opportunity to work as a PhD student in his research group. In the past I did not expect that this could be possible for me. Thanks Arnd, for supporting me during my PhD time with participations on conferences, mathematical discussions and not to forget the nice talks about sports apart from mathematics. I really enjoyed this. Of course, big thanks go to the other members of the research group: Nicole, Sven, Danilo, Marita and Aysel. I wish you and your families all the best for the future. Furthermore, I would like to thank Dmitri Kuzmin, Christoph Lohmann and Johannes Pfefferer for their time, open ears and willingness answering my questions.
Finally, an infinite amount of hugs, kisses and thanks go to my family: My parents Heinz-Georg and Iris, my brother Kai, my sister Kirsten with her boyfriend Lars, my grandparents Werner and Kläre Gaczenski, Lieselotte and Georg Baumgartner. You gave me much strength, power and comfort since I was born. There are no words which can transport my love for you and my thankfulness being a member of this family.

Without courage, wisdom bears
no fruit

Baltasar Gracián y Morales

Contents

1 Introduction

1.1 Motivation

Convection-diffusion reaction processes arise in many chemical, physical or biological applications. For instance, we have the air conditioning-process, the process of water resource recovery or the spreading-process of oil in the ocean when an oil-tanker has a lack. All processes are influenced by the directional motion (convection), the random motion (diffusion) and the reaction of particles. Apart from the huge meaning for the engineering the modelling and investigation of convection-diffusion reaction processes are also important for the mathematics. In the last decades many papers concerning the analysis and the numerical treatment of such convection-diffusion reaction equations have been published. The above-mentioned processes can be described by the following convection-diffusion reaction equation

$$-\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \tag{1.1.1}$$

where $\Omega \subseteq \mathbb{R}^2$ is an open and bounded domain. The constant diffusion coefficient is given by $\varepsilon > 0$; $c \in L^\infty(\Omega)$ is a nonnegative reaction where $c_0 := \operatorname{ess\,inf} c > 0$ holds. We further assume that the convection field $\mathbf{b} \in W^{1,\infty}(\Omega)^2$ satisfies

$$\operatorname{div}(\mathbf{b}) = 0. \tag{1.1.2}$$

Especially convection-dominated equations, i.e. equations where the convective transport dominates the diffusion, have often been analyzed. The reason for this widespread interest is that solutions of convection-dominated equations may possess layers, i.e. small regions where the solution has a large gradient. The computation of discrete solutions by standard Finite Element Methods leads to solutions which contain spurious oscillations. To reduce those oscillations, many stabilization methods have been invented. One of the first stabilization methods was the streamline-upwind Petrov-Galerkin method (SUPG) introduced in [BroHug82]. The discrete solution corresponding to the SUPG-stabilization possesses the layer at the correct position, but still oscillations appear near the layers. During the last years, further methods tied to the SUPG method were developed like the so-called SOLD methods [JoKno07, JoKno08], new stabilization techniques like edge-based stabilization techniques [BBK17, BurHan04] or the latest, the Algebraic Flux Correction (AFC) schemes.
The initial idea of the construction of AFC schemes goes back to the paper of Zalesak [Zal79] published in 1979. In the research field of computational mathematics, the AFC schemes have gradually gained importance from 2004 onwards until today. Especially consider the works of Kuzmin [Kuz06, Kuz10, Kuz12]. However, the analysis of AFC schemes has been established only since 2016 [BJK16, BBK17, BJK17, BJKR18]. The main motivation for the construction of an AFC scheme is the satisfaction of the discrete maximum principle (DMP) such that spurious oscillations in the discrete solutions are prevented. The treatment of convection-diffusion reaction equations is not only interesting for the Analysis and the Numerical mathematics.
Due to the fact that many physical or chemical optimization processes can be modelled by optimal control problems governed by a convection-diffusion reaction equation the investigation of these problems increases in the last years as well. In this work, we consider several optimal control problems governed by a convection-diffusion reaction equation. Apart from the analysis of continuous optimal control problems for reasons of implementation, many regularization resp. discretization methods have been applied on such problems. For detailed information we refer to [Cas86, Cas93, MyRöTr06, KruRö08, CheRö09, HtKu09, HzHt09]. For reasons of discretizing such optimal control problems, the *optimize-then-discretize*-approach and the *discretize-then-optimize*-approach have been usually investigated (for instance see [HzRö12])

where it is worth to mention that the *discretize-then-optimize*-approach has been applied more often. In the *optimize-then-discretize*-approach, the necessary optimality conditions are derived on the continuous level. Then, the optimality conditions are discretized by a Finite Element Method. In the *discretize-then-optimize*-approach, the state equation is discretized by a Finite Element Method such that the resulting optimization problem is finite dimensional. After that, the necessary optimality conditions are derived. Regarding the introduced stabilization methods, the application of a Finite Element Method in both approaches is implemented by adding a stabilization term to the state resp. in the *optimize-then-discretize*-approach also to the adjoint equation. In [Braa09] the author points out that for symmetric and bilinear stabilization terms both approaches coincide on the contrary to the case of non-symmetric stabilization terms. In contrast to the SUPG method or the edge based Galerkin method, the original AFC method contains in general solution-dependent, nonlinear and non-differentiable correction factors, the so-called flux limiters. Due to the non-differentiability of the basic limiters, Newton-like solvers cannot be applied on such AFC stabilized systems. Moreover, in the *discretize-then-optimize*-approach it is left unsaid how to define the corresponding adjoint operator for a nonlinear and non-differentiable operator. Additionally, sufficient and necessary optimality conditions of first order cannot be derived since the non-differentiability does not make it possible to compute Fréchet derivatives. It is worth mentioning that during the last years the flux limiters have undergone many modifications so that they become differentiable (see [BadBon17, Section 7] or [LohSP19, p. 127]).
However, in this work we focus on the original flux limiters, i.e. the implemented limiters are nonlinear and non-differentiable. Currently, the *discretize-then-optimize*-approach cannot be applied on our problem. Hence, in this work we connect the *optimize-then-discretize*-approach with the AFC method. Past publications concerned with the connection between optimal control problems and stabilization methods except the AFC method. For instance, in [ColHei02] the authors analyzed both approaches for an unconstrained optimal control problem. The authors investigate in both approaches the SUPG stabilization method. Due to the fact that the bilinear SUPG stabilization term is non-symmetric regarding [Braa09] both approaches do not coincide. The verification of the existence of discrete solutions and the derivation of a priori error estimates have been realized by the application of the theory of linear, continuous and invertible operators. In [HzYaZh09, YaZh09] the authors stabilize an unconstrained and control constrained optimal control problem by an edge based Galerkin method which has been established in [BurHan04]. The stabilization term of the edge based Galerkin discretization is bilinear and symmetric such that the *discretize-then-optimize* and the *optimize-then-discretize*-approach coincide. Another stabilization method which has been applied for discretizing an optimal control problem governed by a convection-diffusion reaction equation is the so-called LPS-method (see [Guer99]), i.e. a stabilization method based on local projections. In [BecVe07] the authors use the *discretize-then-optimize*-approach to discuss the discretization of an unconstrained and a control constrained optimal control problem by the LPS-method. As far as we know the papers mentioned above are the only papers which connect stabilization methods and optimal control problems. As already mentioned, we stabilize optimal control problems with the AFC method in this work. Due to the nonlinearity and the non-differentiability, the indroduced approaches *discretize-then-optimize* and *optimize-then-discretize* do not coincide in general. Currently, in the *discretize-then-optimize*-approach it is neither possible to construct an adjoint operator nor to compute Fréchet derivatives. Hence, in contrast to [BecVe07] and [HzYaZh09], proving the existence of discrete solutions and deriving sufficient and necessary optimality conditions of first order for the finite dimensional optimization problem cannot be realized by the *discretize-then-optimize*-approach. Additionally, in the *optimize-then-discretize*-approach the applied theory for linear, continuous and invertible operators introduced in [ColHei02] cannot be transferred to our problem. As a result, the discretization of an optimal control problem by an AFC

scheme requires new techniques to verify the existence of discrete solutions and for proving corresponding error estimates in the context of the *optimize-then-discretize*-approach.

1.2 Outline

In this work, we consider an open, bounded and convex polygonal domain $\Omega \subseteq \mathbb{R}^2$ with boundary Γ. We study the stabilization of linear-quadratic optimal control problems governed by a convection-diffusion reaction equation. In the first case we consider an optimization problem governed by a convection-diffusion reaction equation with Dirichlet boundary conditions

$$\left.\begin{array}{l} \min\limits_{u,y} \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b} \cdot \nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ u \in U_{ad} \\ y \in Y_{ad} \end{array}\right\} (P)$$

where $y_d \in L^2(\Omega)$ and $U_{ad}, Y_{ad} \subseteq L^2(\Omega)$ are closed and convex sets. Secondly, we investigate the stabilization of the following optimal control problem with Robin boundary control

$$\left.\begin{array}{l} \min\limits_{u,y} \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 \\ -\varepsilon\Delta y + \mathbf{b} \cdot \nabla y + cy = 0 \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n \cdot y}{2} = u \quad \text{on } \Gamma \\ u \in U_{ad}^{\Gamma} \end{array}\right\} (P_\Gamma)$$

where $y_d \in L^2(\Omega)$ and the set $U_{ad}^{\Gamma} \subseteq L^2(\Gamma)$ is closed and convex. This work is organized as follows. In Section 2, we introduce function spaces and basic results. According to (P) resp. (P_Γ), in Section 3 we analyze the convection-diffusion reaction equation with Dirichlet and Robin boundary conditions. In Section 4, we specify the introduced optimal control problems (P) resp. (P_Γ) and provide the corresponding analysis. Section 5 is dedicated to introduce the Finite Element Method and the general AFC methodology. In connection we show the general construction of an AFC scheme and discuss sufficient conditions such that discrete maximum principles hold. After that, in Section 6 we show the discretization of the optimal control problems by the AFC method in the context of the *optimize-then-discretize*-approach. In Section 7, the existence of discrete solutions and corresponding error estimates will be derived in an abstract framework. In Section 8, the abstract results of Section 7 will be applied on the several optimal control problems. In detail, we proof the existence of discrete solutions for the discretized systems provided in Section 6 and derive corresponding a priori error estimates. In this context we prove that for the introduced optimal control problems the following L^2-error estimates

$$\parallel u - u_h \parallel_{0,\Omega} + \parallel y - y_h \parallel_{0,\Omega} \leq Ch^{\frac{1}{2}} \tag{1.2.1}$$

and

$$\parallel u - u_h \parallel_{0,\Gamma} + \parallel y - y_h \parallel_{0,\Omega} \leq Ch^{\frac{1}{4}} \tag{1.2.2}$$

hold where $C > 0$ is a constant, y_h is the discrete state solution and u_h the control, both computed by the AFC method. In addition, we illustrate for every optimal control problem several numerical results, i.e. we show computed L^2-errors, L^2-convergence orders and plots of the AFC state resp. the AFC adjoint solutions. In Section 9, we will illustrate further optimal control problems where the abstract results of Section 7 are applicable. Finally, we show an optimal control problem which cannot be currently solved by the derived abstract theory.

1.3 List of symbols and notations

Table 1: Spaces, functions, notations

Symbol	Description
Ω	Domain
Γ	Boundary of Ω
$\{\Gamma_i\}_{i=1}^m$	Line segments of a polygonal boundary Γ
$\mathbb{N}$	Natural numbers (without zero)
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$
$\mathbb{R}$	Real numbers
$C^k(\Omega)$	k-times continuous and differentiable functions
$C_0^k(\Omega)$	Space of functions belonging to $C^k(\Omega)$ with compact support in Ω
$C^{0,\gamma}(\bar{\Omega})$	Hölder-continuous functions to exponents $\gamma \in (0,1]$
$\mathcal{M}(\bar{\Omega})$	Space of Radon measures
$L^p(\Omega)$	$\{f : \Omega \to \mathbb{R} \vert f$ Lebesgue-measurable, $\int_\Omega \vert f(x)\vert^p dx < \infty\}$
$L^\infty(\Omega)$	$\{f : \Omega \to \mathbb{R} \vert f$ Lebesgue-measurable, $\operatorname{ess\,sup}_{x\in\Omega} \vert f(x)\vert < \infty\}$
$W^{k,p}(\Omega)$	Sobolev space for integer k, i.e. space of functions whose weak derivatives of order up to k are in $L^p(\Omega)$
$W_0^{k,p}(\Omega)$	Closure of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$
$H_0^k(\Omega)$	$H_0^k(\Omega) := W_0^{k,2}(\Omega)$
$W^{-1,p'}(\Omega)$	Dual space of $W_0^{1,p}(\Omega)$
$H^{-1}(\Omega)$	$H^{-1}(\Omega) := W^{-1,2}(\Omega)$
$W^{s,p}(\Omega)$	Sobolev space for $s > 0$
$H^s(\Omega)$	Sobolev space for $s > 0$ and $p = 2$
$H^s(\Gamma)$	Sobolev space on Γ for $0 < s \leq 1$ and $p = 2$
$H^s(\Gamma_i)$	Sobolev space on line segments Γ_i for $0 < s \leq 1$ and $p = 2$
$\hookrightarrow$	Continuously embedded
$\overset{c}{\hookrightarrow}$	Compactly, continuously embedded
α	Multi-index
∇f	Gradient of f
n	Unit outward normal vector
$\partial_n f$	Normal derivative of f
Δf	Laplacian of f
$supp(f)$	Support of f
f^+	$\max(0, f)$
f^-	$\min(0, f)$
$\mathbb{P}_{[r_1,r_2]}(\cdot)$	Projection formula, $\mathbb{P}_{[r_1,r_2]}(\cdot) = \min\{r_2, \max\{r_1, \cdot\}\}$
$\Psi_k(\cdot)$	Truncation function, $\Psi_k(\cdot) = \mathbb{P}_{[-k,k]}(\cdot)$
$B_r(\bar{x})$	$B_r(\bar{x}) := \{x : \quad \Vert x - \bar{x} \Vert_2 \leq r\}$
$\lceil \cdot \rceil$	Ceiling function
δ	Moreau-Yosida regularization parameter
$\varepsilon > 0$	Diffusion coefficient
$\mathbf{b} \in W^{1,\infty}(\Omega)^2$	Convection field
$c \in L^\infty(\Omega)$	Reaction

Table 2: Norms

Symbol	Description
$\|\cdot\|_{C^k(\bar{\Omega})}$	Norm on the space $C^k(\bar{\Omega})$
$\|\cdot\|_{C^{0,\gamma}(\bar{\Omega})}$	Norm on the space $C^{0,\gamma}(\bar{\Omega})$
$\|\cdot\|_{0,p,\Omega}$	Norm on the space $L^p(\Omega)$ for $p \neq 2$
$\|\cdot\|_{0,\Omega}$	Norm on the space $L^2(\Omega)$
$\|\cdot\|_{k,p,\Omega}$	Norm on the space $W^{k,p}(\Omega)$ for $p \neq 2$
$\|\cdot\|_{s,\Omega}$	Norm on the space $H^s(\Omega)$ for $s > 0$
$\|\cdot\|_{s,\Gamma}$	Norm on the space $H^s(\Gamma)$ for $0 < s \leq 1$
$\|\cdot\|_{k,p,\Omega}$	Seminorm on the space $W^{k,p}(\Omega)$ for $p \neq 2$
$\|\cdot\|_{k,\Omega}$	Seminorm on the space $W^{k,2}(\Omega)$
$\|\cdot\|_2$	Euclidean norm on $\mathbb{R}^2$
$\|\cdot\|_\infty$	Maximum norm on $\mathbb{R}^2$

Table 3: Symbols referring to FEM and AFC methodology

Symbol	Description
$\mathcal{T}_h$	Triangulation
T	Mesh cell of $\mathcal{T}_h$
$\mathcal{E}_h$	Set of all edges of the triangulation
N	Dimension of the Finite Element space/ Total number of nodes
x_i	Nodal point with index $i \in \{1, \cdots, N\}$
S_i	Index set of neighbors for node x_i
Δ_i	Patches of node x_i
$diam(S)$	Diameter of $S \subset B_r$
h_T	Width of mesh cell T
h	Maximum mesh width
$\mathbb{P}_1$	Polynomials up to degree 1
V_h	Space of $\mathbb{P}_1$ Finite Elements
$V_{h,0}$	$V_{h,0} := V_h \cap H_0^1(\Omega)$
I_h	Lagrange interpolation operator
$\mathcal{A} = (a_{ij})_{i,j=1,\cdots,N}$	Finite Element stiffness matrix
$\mathcal{M}_{ass} = (m_{ij})_{i,j=1,\cdots,N}$	Finite Element mass matrix
$\mathcal{D} = (d_{ij})_{i,j=1,\cdots,N}$	Artificial diffusion matrix
$\tilde{\mathcal{D}} = (\tilde{d}_{ij})_{i,j=1,\cdots,N}$	Correction matrix
α_{ij}	Flux limiter
$d_h(\cdot;\cdot,\cdot)$	AFC stabilization term

2 Function spaces

In this section, we provide the fundamental function spaces which will be used on the following pages. We start with the introduction to the basic function spaces, i.e. continuous and Hölder-continuous function spaces. After that, the definitions of the L^p-spaces and the Sobolev spaces will be provided. Throughout this section we assume that $\Omega \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$ is an open, bounded domain with Lipschitz boundary Γ. According to [Gris85, Corollary 1.2.2.3], the following theory also holds for an open, bounded an convex domain. For the precise definition of Lipschitz boundaries we refer to [Gris85, Definition 1.2.1.1]. A detailed review of the upcoming functional analysis includes [Ada75, Brez11, Evans98, Gris85, Pfeff15].

2.1 Classical function spaces

Definition 2.1. *Let $m \in \mathbb{N}$. A multi-index α is a vector $\alpha = (\alpha_1, \cdots, \alpha_m)$ with $\alpha_j \in \mathbb{N}_0,\ j = 1, \cdots, m$ and*

$$|\alpha| := \sum_{j=1}^{m} \alpha_j.$$

Definition 2.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $|\alpha|$-times continuous differentiable function, then the α-partial derivative of f is defined by*

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}.$$

We denote the gradient of a continuous differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ by

$$\nabla f := \Big(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_d}\Big)^T \quad \text{with} \quad \nabla_j f = \frac{\partial f}{\partial x_j} \quad \forall j = 1, \cdots, d.$$

Definition 2.3. *Let $k \in \mathbb{N}_0$ and let $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2$ be a multi-index. The space $C^k(\Omega)$ is defined as the set of all functions f on Ω with continuous derivatives $D^\alpha f$ up to order k. Further, we set $C^\infty(\Omega) = \cap_{k=0}^\infty C^k(\Omega)$ and $C_0^k(\Omega) = \{f \in C^k(\Omega) : supp(f) \subset \Omega$ is compact$\}$. The space $C^k(\bar{\Omega})$ denotes the set of all functions f on Ω with bounded and uniformly continuous derivatives $D^\alpha f$ up to order k, i.e. the derivatives $D^\alpha f$ can continuously be extended to $\bar{\Omega}$ for $|\alpha| \leq k$. The norm in $C^k(\bar{\Omega})$ is defined by*

$$\| f \|_{C^k(\bar{\Omega})} := \sum_{|\alpha| \leq k} \sup_{x \in \Omega} |(D^\alpha f)(x)|.$$

For $k = 0$ we set $C(\Omega) = C^0(\Omega)$ and $C(\bar{\Omega}) = C^0(\bar{\Omega})$. Moreover, we define

$$C_0^\infty(\Omega) = \{f \in C^\infty(\Omega) : supp(f) \subset \Omega \text{ is compact}\}.$$

Definition 2.4. *Let $\gamma \in (0, 1]$. The space of all Hölder-continuous functions to the exponent γ is defined by*

$$C^{0,\gamma}(\bar{\Omega}) := \{f \in C(\bar{\Omega}) :\ \exists\, c > 0\ \forall x_1, x_2 \in \bar{\Omega} :\ |f(x_1) - f(x_2)| \leq c \| x_1 - x_2 \|_2^\gamma\}.$$

The associated norm is defined by

$$\| f \|_{C^{0,\gamma}(\bar{\Omega})} := \| f \|_{C(\bar{\Omega})} + \sup_{\substack{x_1, x_2 \in \bar{\Omega} \\ x_1 \neq x_2}} \frac{|f(x_1) - f(x_2)|}{\| x_1 - x_2 \|_2^\gamma}.$$

Definition 2.5. *The dual space of $C(\bar{\Omega})$, denoted by $\mathcal{M}(\bar{\Omega})$ is called the space of Radon measures on $\bar{\Omega}$. The space $\mathcal{M}(\bar{\Omega})$ is endowed with the norm*

$$\| \mu \|_{\mathcal{M}(\bar{\Omega})} = \sup_{z \in B_1} \int_{\Omega} z \, d\mu$$

where

$$B_1 := \{z \in C(\bar{\Omega}) : \quad \| z \|_{C(\bar{\Omega})} \leq 1\}.$$

Now, we introduce the L^p-spaces for a domain $\mathcal{G}$ which coincides with Ω, its boundary Γ or with a subset $\Gamma_s \subset \Gamma$. Note that in the case of a boundary integral, i.e. $\mathcal{G} \in \{\Gamma, \Gamma_s\}$, the variable of integration is s and in the case of $\mathcal{G} = \Omega$, the variable of integration is x.

Definition 2.6 (L^p-spaces)**.** *Let $\mathcal{G}$ be Ω, its boundary Γ or a subset of the boundary $\Gamma_s \subset \Gamma$ with $|\Gamma_s| > 0$. The Lebesgue space $L^p(\mathcal{G})$ with $1 \leq p < \infty$ is the space of all Lebesgue-measurable functions f such that*

$$\int_{\mathcal{G}} |f|^p < \infty.$$

For $1 \leq p < \infty$ the $L^p(\mathcal{G})$-norm is defined by

$$\| f \|_{0,p,\mathcal{G}} := \Big(\int_{\mathcal{G}} |f|^p \Big)^{\frac{1}{p}}.$$

We denote by $L^\infty(\mathcal{G})$ the Banach space of real valued functions f such that

$$|f(x)| \leq C \quad \text{a.e. in/on } \mathcal{G}$$

where $C > 0$ is constant. The $L^\infty(\mathcal{G})$-norm is defined by

$$\| f \|_{0,\infty,\mathcal{G}} := \operatorname*{ess\,sup}_{x \in \mathcal{G}} |f(x)| := \inf_{\substack{N \subset \mathcal{G} \\ |N|=0}} \sup_{x \in \mathcal{G} \setminus N} |f(x)|.$$

Remark 2.7. *The spaces $(L^p(\mathcal{G}), \| \cdot \|_{0,p,\mathcal{G}})$ and $(L^\infty(\mathcal{G}), \| \cdot \|_{0,\infty,\mathcal{G}})$ are Banach spaces. For $p = 2$, the space $L^2(\mathcal{G})$ is a Hilbert space with the inner product*

$$(f, g)_{\mathcal{G}} = \int_{\mathcal{G}} fg$$

and the induced norm

$$\| f \|_{0,2,\mathcal{G}} = (f, f)_{\mathcal{G}}^{\frac{1}{2}}.$$

In the following we set

$$\| f \|_{0,\mathcal{G}} := \| f \|_{0,2,\mathcal{G}} \, .$$

According to the L^p-spaces, we have the following Hölder inequality which can be found for instance in [Brez11, Theorem 4.6].

Lemma 2.8 (Hölder inequality)**.** *Let $1 \leq p \leq \infty$ and p' the conjugate exponent of p, i.e. $\frac{1}{p} + \frac{1}{p'} = 1$. For $f \in L^p(\mathcal{G})$ and $g \in L^{p'}(\mathcal{G})$ we have $fg \in L^1(\mathcal{G})$ and the inequality*

$$\Big| \int_{\mathcal{G}} fg \Big| \leq \| f \|_{0,p,\mathcal{G}} \| g \|_{0,p',\mathcal{G}} \, .$$

Now, we define the usual Sobolev spaces.

Definition 2.9 ($W^{k,p}(\Omega)$ – *spaces*)**.** *Let $1 \leq p \leq \infty$ and $k \in \mathbb{N}_0$. Furthermore, let $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2$ be a multi-index. The Sobolev space $W^{k,p}(\Omega)$ is the space of all functions $f \in L^p(\Omega)$ whose weak derivatives $D^\alpha f$ exist and belong to $L^p(\Omega)$ for $|\alpha| \leq k$. The space $W^{k,p}(\Omega)$ is equipped with the norm*

$$\| f \|_{k,p,\Omega} := \left(\sum_{|\alpha| \leq k} \| D^\alpha f \|_{0,p,\Omega}^p \right)^{\frac{1}{p}} \quad \text{if } 1 \leq p < \infty$$
$$\| f \|_{k,\infty,\Omega} := \sum_{|\alpha| \leq k} \| D^\alpha f \|_{0,\infty,\Omega} \, .$$

The seminorms are given by

$$|f|_{k,p,\Omega} := \left(\sum_{|\alpha| = k} \| D^\alpha f \|_{0,p,\Omega}^p \right)^{\frac{1}{p}} \quad \text{if } 1 \leq p < \infty$$
$$|f|_{k,\infty,\Omega} := \sum_{|\alpha| = k} \| D^\alpha f \|_{0,\infty,\Omega} \, .$$

For $p = 2$ we set

$$\| f \|_{k,\Omega} := \| f \|_{k,2,\Omega}$$
$$|f|_{k,\Omega} := |f|_{k,2,\Omega}.$$

Definition 2.10. *Let $1 \leq p \leq \infty$ and $k \in \mathbb{N}$. The space $W_0^{k,p}(\Omega)$ denotes the closure of $C_0^\infty(\Omega)$ with respect to the norm $\| \cdot \|_{k,p,\Omega}$, i.e.*

$$W_0^{k,p}(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{k,p,\Omega}}.$$

Lemma 2.11. *(Poincaré inequality) Let $1 \leq p < \infty$. Then, there exists a constant $C_p > 0$ (depending on Ω and p) such that*

$$\| f \|_{0,p,\Omega} \leq C_p \| \nabla f \|_{0,p,\Omega} \quad \forall f \in W_0^{1,p}(\Omega).$$

Proof. See [Brez11, Corollary 9.19]. □

Remark 2.12. *Let $1 \leq p < \infty$. Then, the norm*

$$|f|_{1,p} := \| \nabla f \|_{0,p,\Omega}$$

is equivalent to the norm $\| \cdot \|_{1,p,\Omega}$ on the space $W_0^{1,p}(\Omega)$.

In the following we extend Definition 2.9 to non-integers $s > 0$. This leads us to the so-called fractional Sobolev spaces which arise in many applications of partial differential equations and optimal control theory. For a general introduction to fractional Sobolev spaces we refer to [Gris85, Section 1.3] or [Trie78].

Definition 2.13 ($W^{s,p}(\Omega)$–*spaces*)**.** *Let $1 \leq p < \infty$ and $\alpha \in \mathbb{N}_0^2$ be a multi-index. Furthermore, let $s > 0$ be a non-integer where $s = k + \sigma$ with $k \in \mathbb{N}_0$ and $0 < \sigma < 1$. The space $W^{s,p}(\Omega)$ denotes the space of all functions which belong to $W^{k,p}(\Omega)$ and satisfy*

$$|f|_{s,p,\Omega} := \left(\sum_{|\alpha| = k} \int_\Omega \int_\Omega \frac{|(D^\alpha f)(x_1) - (D^\alpha f)(x_2)|^p}{|x_1 - x_2|^{2+\sigma p}} dx_1 \, dx_2 \right)^{\frac{1}{p}} < \infty.$$

The corresponding norm is given by

$$\| f \|_{s,p,\Omega} := \left(\| f \|_{k,p,\Omega}^p + |f|_{s,p,\Omega}^p \right)^{\frac{1}{p}} .$$

For $p = 2$ we set

$$\begin{aligned} \| f \|_{s,\Omega} &:= \| f \|_{s,2,\Omega} \\ |f|_{s,\Omega} &:= |f|_{s,2,\Omega}. \end{aligned}$$

Now, we define the Sobolev spaces $W^{s,p}(\Gamma)$ on the Lipschitz boundary Γ. Note that in this work, only the spaces $W^{s,p}(\Gamma)$ with $1 \leq p < \infty$ and $0 < s \leq 1$ will be used. However, the following definition also holds for an arbitrary $0 < s \leq k$ with $k \in \mathbb{N}$ when Ω is of class $C^{k-1,1}$. For detailed information on the general definition of Sobolev spaces on manifolds we refer to [Gris85, Section 1.3.3] or [Wlok92, Section 1§4].

Definition 2.14 ($W^{s,p}(\Gamma)$-spaces)**.** *Let $1 \leq p \leq \infty$ and $\alpha \in \mathbb{N}_0^2$ be a multi-index. For $k \in \{0, 1\}$ the space $W^{k,p}(\Gamma)$ is the space of all functions $f \in L^p(\Gamma)$ whose weak tangential derivatives $\partial_t^\alpha f$ exist and belong to $L^p(\Gamma)$ for $|\alpha| \leq k$. The corresponding norm is given by*

$$\begin{aligned} \| f \|_{k,p,\Gamma} &:= \left(\sum_{|\alpha| \leq k} \| \partial_t^\alpha f \|_{0,p,\Gamma}^p \right)^{\frac{1}{p}} \quad \text{if } 1 \leq p < \infty \\ \| f \|_{k,\infty,\Gamma} &:= \sum_{|\alpha| \leq k} \| \partial_t^\alpha f \|_{0,\infty,\Gamma} . \end{aligned}$$

The seminorm $|\cdot|_{k,p,\Gamma}$ is defined as in Defintion 2.9. Furthermore, we set $W^{0,p}(\Gamma) := L^p(\Gamma)$. For every non-integer $0 < s < k$ and $1 \leq p < \infty$ we set $s = m + \sigma$ with $m \in \mathbb{N}_0$ and $0 < \sigma < 1$. The space $W^{s,p}(\Gamma)$ denotes the space of all functions which belong to $W^{m,p}(\Gamma)$ and satisfy

$$|f|_{s,p,\Gamma} := \left(\sum_{|\alpha|=m} \int_\Gamma \int_\Gamma \frac{|(\partial_t^\alpha f)(x_1) - (\partial_t^\alpha f)(x_2)|^p}{|x_1 - x_2|^{1+\sigma p}} ds_{x_1} \, ds_{x_2} \right)^{\frac{1}{p}} < \infty.$$

The space $W^{s,p}(\Gamma)$ is equipped with the norm

$$\| f \|_{s,p,\Gamma} := \left(\| f \|_{m,p,\Gamma}^p + |f|_{s,p,\Gamma}^p \right)^{\frac{1}{p}} .$$

For $p = 2$ we set

$$\begin{aligned} \| f \|_{s,\Gamma} &:= \| f \|_{s,2,\Gamma} \\ |f|_{s,\Gamma} &:= |f|_{s,2,\Gamma}. \end{aligned}$$

Remark 2.15. *In this work, we also use Sobolev spaces on line segments of a polygonal boundary Γ. In detail, from Section 3 we consider an open, bounded, convex and polygonal domain $\Omega \subseteq \mathbb{R}^2$. Regarding [Gris85, Section 4, p. 182], a polygonal domain is a union of open line segments $\{\Gamma_i\}_{i=1}^m \subset \Gamma$ with $m \in \mathbb{N}$ and $\bigcup_{i=1}^m \bar{\Gamma}_i = \Gamma$. In Section 3.2.1 we will prove several regularity results for a solution of an elliptic partial differential equation with Robin boundary condition. For this, we will use for the data on the boundary the product space $\Pi_{i=1}^m W^{\frac{1}{2},2}(\Gamma_i)$. Note that the spaces $W^{\frac{1}{2},2}(\Gamma_i)$, $i = 1, \cdots, m$ can be defined analogously to Definition 2.14. In general, due to the lack of regularity in the corners of Γ the product space $\Pi_{i=1}^m W^{\frac{1}{2},2}(\Gamma_i)$ do not coincide with the space $W^{\frac{1}{2},2}(\Gamma)$.*

Definition 2.16. *For $p = 2$ and $s \geq 0$ we introduce the notation $(H^s(\Omega), \|\cdot\|_{s,\Omega})$, $(H^s_0(\Omega), |\cdot|_{s,\Omega})$, $(H^s(\Gamma), \|\cdot\|_{s,\Gamma})$ and $(H^s(\Gamma_i), \|\cdot\|_{s,\Gamma_i})$, $i = 1, \cdots, m$ where*

$$\begin{aligned} H^s(\Omega) &:= W^{s,2}(\Omega) \\ H^s_0(\Omega) &:= W^{s,2}_0(\Omega) \\ H^s(\Gamma) &:= W^{s,2}(\Gamma) \\ H^s(\Gamma_i) &:= W^{s,2}(\Gamma_i), \quad i = 1, \cdots, m. \end{aligned}$$

In the following, let us recall several embedding theorems which can be found for instance in [Alt06, Section 8]. For a detailed introduction to Sobolev spaces and embedding theorems we refer to [Gris85, Section 1.4], [Ada75] or [BeLöf76, Section 6].

Theorem 2.17. *Let $1 \leq p, q < \infty$ and let $s, t \geq 0$ be real numbers and k a non-negative integer. Then, the following assertions hold:*

1.) *Let $s - d/p = t - d/q$ and $s \geq t$. Then, the continuous embedding $W^{s,p}(\Omega) \hookrightarrow W^{t,q}(\Omega)$ is valid.*

2.) *Let $s - d/p > t - d/q$ and $s > t$. Then, the compact embedding $W^{s,p}(\Omega) \overset{c}{\hookrightarrow} W^{t,q}(\Omega)$ is valid.*

3.) *Let $s - d/p = k + \sigma$ and $0 < \sigma < 1$. Then, the continuous embedding $W^{s,p}(\Omega) \hookrightarrow C^{k,\sigma}(\bar{\Omega})$ is valid.*

4.) *Let $s - d/p > k + \sigma$ and $0 \leq \sigma \leq 1$. Then, the compact embedding $W^{s,p}(\Omega) \overset{c}{\hookrightarrow} C^{k,\sigma}(\bar{\Omega})$ is valid.*

For the investigation of boundary value problems it is often necessary that functions belonging to $W^{k,p}(\Omega)$ with $k \geq 1$ have boundary values in so far as:

Theorem 2.18. *Let $1 \leq p \leq \infty$. Then, there exists a linear and continuous mapping $\tau : W^{1,p}(\Omega) \to L^p(\Gamma)$ such that for all $f \in C(\bar{\Omega})$*

$$(\tau f)(x) = f(x) \quad \text{a.e. on } \Gamma$$

is valid. Moreover, there exists a constant $C_\tau > 0$ such that for all $f \in W^{1,p}(\Omega)$

$$\| \tau f \|_{0,p,\Gamma} \leq C_\tau \| f \|^{1-1/p}_{0,p,\Omega} \| f \|^{1/p}_{1,p,\Omega}$$

holds.

Proof. See [Alt06, Theorem A6.6] and [BreSco02, Theorem 1.6.6]. □

Definition 2.19. *The element τf is called trace of f on the boundary Γ and the mapping τ is called trace operator. In the following we use for the trace of a function $f \in W^{1,p}(\Omega)$ with $1 \leq p \leq \infty$ the notation $\tau f = f_{|\Gamma}$.*

Theorem 2.20. *Let $1 < p < \infty$. The trace operator is a bounded and linear operator from $\tau : W^{1,p}(\Omega) \to W^{1-\frac{1}{p},p}(\Gamma)$.*

Proof. See [Gris85, Theorem 1.5.1.2]. □

Lemma 2.21. *Let $\mathbf{f} \in W^{1,\infty}(\Omega)^2$. Then, the integration-by-parts formula yields*

$$\int_\Omega (\mathbf{f} \cdot \nabla z)\, \psi + (\mathbf{f} \cdot \nabla \psi)\, z + \operatorname{div}(\mathbf{f}) z \psi \; dx = \int_\Gamma \mathbf{f} \cdot n z \psi \; ds \quad \forall z, \psi \in H^1(\Omega).$$

Proof. See [DPE11, Section 2.1.5]. □

2.2 Elementary functions and results

Definition 2.22. *Let $i, j \in \mathbb{N}$. The Kronecker-Delta function δ_{ij} is defined by*

$$\delta_{ij} := \begin{cases} 1 & , i = j \\ 0 & , i \neq j. \end{cases}$$

Definition 2.23. *For $k > 0$ we define the truncation function $\Psi_k : \mathbb{R} \to \mathbb{R}$ by*

$$\Psi_k(z) := \begin{cases} k & , z > k \\ z & , |z| \leq k \\ -k & , z < -k. \end{cases} \tag{2.2.1}$$

Note that the function Ψ_k is Lipschitz continuous with a Lipschitz constant $L = 1$.

Definition 2.24. *For $r_1, r_2 \in \mathbb{R} \cup \{-\infty, \infty\}$ with $r_1 < r_2$ we define the pointwise projection formula $\mathbb{P}_{[r_1,r_2]} : \mathbb{R} \to \mathbb{R}$ by*

$$\mathbb{P}_{[r_1,r_2]}(z) = \min\{r_2, \max\{z, r_1\}\} \quad \forall z \in \mathbb{R}.$$

Remark 2.25. *For $k > 0$ we have*

$$\Psi_k(z) = \mathbb{P}_{[-k,k]}(z) \quad \forall z \in \mathbb{R}.$$

Lemma 2.26. *Let $r_1, r_2 \in \mathbb{R} \cup \{-\infty, \infty\}$ with $r_1 < r_2$. Then, we have for $z \in \mathbb{R}$*

$$\max\{0, z - r_2\} + \min\{0, z - r_1\} = z - \mathbb{P}_{[r_1,r_2]}(z). \tag{2.2.2}$$

Proof. First, we investigate the case $r_1 \leq z \leq r_2$. According to Definition 2.24, we have $\mathbb{P}_{[r_1,r_2]}(z) = z$ and

$$\begin{aligned} z - r_2 \leq 0 &\to \max\{0, z - r_2\} = 0 \\ z - r_1 \geq 0 &\to \min\{0, z - r_1\} = 0. \end{aligned}$$

Thus, we obtain

$$\max\{0, z - r_2\} + \min\{0, z - r_1\} = 0 = z - \mathbb{P}_{[r_1,r_2]}(z).$$

In the case $z \leq r_1 < r_2$ we have $\mathbb{P}_{[r_1,r_2]}(z) = r_1$ and

$$\begin{aligned} z - r_2 < 0 &\to \max\{0, z - r_2\} = 0 \\ z - r_1 \leq 0 &\to \min\{0, z - r_1\} = z - r_1 \end{aligned}$$

such that (2.2.2) holds. In the last case we investigate the case $r_1 < r_2 \leq z$. We have $\mathbb{P}_{[r_1,r_2]}(z) = r_2$ and

$$\begin{aligned} z - r_2 \geq 0 &\to \max\{0, z - r_2\} = z - r_2 \\ z - r_1 > 0 &\to \min\{0, z - r_1\} = 0 \end{aligned}$$

such that again (2.2.2) is valid. □

Lemma 2.27 (Young inequality)**.** *Let $a, b \geq 0$ and $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \tag{2.2.3}$$

holds. For $\gamma > 0$ we have for all $a, b \geq 0$

$$ab \leq \frac{\gamma a^2}{2} + \frac{b^2}{2\gamma}. \tag{2.2.4}$$

3 Elliptic boundary value problems

In this section, we analyze the convection-diffusion reaction equation (1.1.1) with two types of boundary conditions. Firstly, we investigate the convection-diffusion reaction equation with Dirichlet boundary condition which arises in (P) and secondly we provide the analysis of the convection-diffusion reaction equation with Robin boundary condition which arises in (P_Γ). From now on, let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, convex and polygonal domain with boundary Γ. Due to [Gris85, Theorem 1.2.2.3] the domain Ω is Lipschitz. As we have mentioned in Section 1.1 we consider the following data of the convection-diffusion reaction equation.

$(A1)$ $\varepsilon > 0$ is a constant diffusion coefficient.

$(A2)$ $\mathbf{b} \in W^{1,\infty}(\Omega)^2$ is a convection field which satisfies

$$\operatorname{div}(\mathbf{b}) = 0.$$

$(A3)$ $c \in L^\infty(\Omega)$ is a reaction coefficient which satisfies $c_0 := \operatorname{ess\,inf} c > 0$.

Before we start with the analysis of the elliptic boundary value problems, let us recall two fundamental results of functional analysis. Let $(V, \| \cdot \|_V)$ be a real Hilbert space, V^* its dual and $u \in V$. Then, a continuous linear functional F_u can be defined on V by

$$F_u(v) = (u, v)_V \,. \tag{3.0.1}$$

The following theorem of Riesz yields the reverse case, i.e. for every functional $F \in V^*$ there exists an element $u \in V$ such that (3.0.1) holds.

Theorem 3.1 (Riesz)**.** *Let F be a continuous linear functional on V. Then, F can be represented uniquely by*

$$F(v) = (u, v)_V$$

for some $u \in V$. Furthermore, we have

$$\| F \|_{V^*} = \| u \|_V \,.$$

Proof. See [BreSco02, Theorem 2.4.2]. □

The basic Lax-Milgram theorem is another relevant result which should be considered.

Lemma 3.2 (Lax-Milgram)**.** *Let V be a Hilbert space and let $a : V \times V \to \mathbb{R}$ be a continuous bilinear form. Assume that a is coercive, i.e. there exists a constant $C_1 > 0$ such that*

$$a(v, v) \geq C_1 \| v \|_V^2 \quad \forall v \in V.$$

Then, for every continuous linear form $F \in V^*$ there exists a unique $y \in V$ such that

$$a(y, v) = F(v) \quad \forall v \in V. \tag{3.0.2}$$

Furthermore, there exists a constant $C_2 > 0$, independent of F such that

$$\| y \|_V \leq C_2 \| F \|_{V^*} \,. \tag{3.0.3}$$

3.1 Dirichlet boundary condition

Now, let us investigate the convection-diffusion reaction equation with Dirichlet boundary condition. According to Theorem 3.1 resp. Lemma 3.2, we have $V = H_0^1(\Omega)$ and $V^* = H^{-1}(\Omega)$. For $u \in L^2(\Omega)$ we consider the equation

$$\begin{aligned} -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy &= u \quad \text{in } \Omega \\ y &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{3.1.1}$$

We introduce the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ given by

$$a(y,v) := \int_\Omega \varepsilon\nabla y\cdot\nabla v + (\mathbf{b}\cdot\nabla y)\, v + cyv \; dx \quad \forall y,v \in H_0^1(\Omega) \tag{3.1.2}$$

and define a linear continuous functional $F : L^2(\Omega) \to H^{-1}(\Omega)$ by

$$L^2(\Omega) \ni u \mapsto F_u(v) := \int_\Omega uv \; dx \quad \forall v \in H_0^1(\Omega). \tag{3.1.3}$$

Hence, the resulting weak formulation of (3.1.1) is

$$a(y,v) = F_u(v) \quad \forall v \in H_0^1(\Omega). \tag{3.1.4}$$

Definition 3.3. *An element $y \in H_0^1(\Omega)$ satisfying* (3.1.4) *is called weak solution of* (3.1.1) .

Lemma 3.4. *The bilinear form* (3.1.2) *is continuous and coercive, i.e. there exists a constant $C > 0$ such that for all $y \in H_0^1(\Omega)$*

$$a(y,y) \geq C \parallel y \parallel_{1,\Omega}^2 .$$

Proof. Obviously, the bilinear form (3.1.2) is continuous. Due to the fact that $\operatorname{div}(\mathbf{b}) = 0$ and $y = 0$ a.e. on Γ, integration-by-parts (Lemma 2.21) yields $\int_\Omega (\mathbf{b}\cdot\nabla y)\, y \; dx = 0$. Hence, we have

$$\begin{aligned} a(y,y) &= \int_\Omega \varepsilon\nabla y\cdot\nabla y + (\mathbf{b}\cdot\nabla y)\, y + cy^2 \; dx \\ &= \int_\Omega \varepsilon\nabla y\cdot\nabla y + cy^2 \; dx \\ &\geq \varepsilon|y|_{1,\Omega}^2 + c_0 \parallel y \parallel_{0,\Omega}^2 \\ &\geq \min\{\varepsilon, c_0\} \parallel y \parallel_{1,\Omega}^2 \end{aligned}$$

and obtain the coercivity of $a(\cdot,\cdot)$. □

Theorem 3.5. *For every $u \in L^2(\Omega)$ the equation* (3.1.1) *admits a unique weak solution $y \in H_0^1(\Omega)$. Moreover, there exists a constant $C > 0$, depending on the domain Ω and on the data of* (3.1.1) *such that*

$$\parallel y \parallel_{1,\Omega} \leq C \parallel u \parallel_{0,\Omega} . \tag{3.1.5}$$

Proof. First, the space $(H_0^1(\Omega), \|\cdot\|_{1,\Omega})$ is a Hilbert space. Lemma 3.4 yields the continuity and the coercivity of $a(\cdot,\cdot)$. The functional F, i.e. (3.1.3) is linear and satisfies

$$\begin{aligned} |F_u(v)| &\leq \parallel u \parallel_{0,\Omega} \parallel v \parallel_{0,\Omega} \\ &\leq C \parallel u \parallel_{0,\Omega} \parallel v \parallel_{1,\Omega} \quad \forall v \in H_0^1(\Omega). \end{aligned}$$

Hence, we obtain

$$\| F_u \|_{-1,\Omega} \leq C \| u \|_{0,\Omega} . \tag{3.1.6}$$

Lemma 3.2 (Lax-Milgram) yields the existence of a unique solution $y \in H_0^1(\Omega)$ for (3.1.4). Moreover, we obtain by virtue of (3.1.6) the a priori $H^1(\Omega)$-estimate

$$\begin{aligned} \min\{\varepsilon, c_0\} \| y \|_{1,\Omega} &\leq \| F_u \|_{-1,\Omega} \\ &\leq C \| u \|_{0,\Omega} . \end{aligned}$$

□

Next, we define a solution operator associated with the weak formulation (3.1.4). It is helpful to introduce a linear and continuous operator denoted by $G : L^2(\Omega) \to H^1(\Omega)$ that maps an arbitrary $u \in L^2(\Omega)$ to the unique solution $y \in H^1(\Omega)$. The embedding operator is denoted by $E_H : H^1(\Omega) \to L^2(\Omega)$. Then, we are able to define the solution operator $S : L^2(\Omega) \to L^2(\Omega)$ by

$$u \mapsto y, \; y = Su = E_H G u. \tag{3.1.7}$$

With respect to the solution operator S the corresponding adjoint operator S^* will be derived. For this, we consider for $w \in L^2(\Omega)$ the following convection-diffusion reaction equation

$$\begin{aligned} -\varepsilon \Delta p - \mathbf{b} \cdot \nabla p + cp &= w \quad \text{in } \Omega \\ p &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{3.1.8}$$

The associated bilinear form $\tilde{a} : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is given by

$$\tilde{a}(p, \psi) := \int_\Omega \varepsilon \nabla p \cdot \nabla \psi - (\mathbf{b} \cdot \nabla p)\, \psi + cp\psi \; dx.$$

Hence, Lemma 2.21 yields

$$\begin{aligned} \tilde{a}(p, \psi) &= \int_\Omega \varepsilon \nabla p \cdot \nabla \psi - (\mathbf{b} \cdot \nabla p)\, \psi + cp\psi \; dx \\ &= \int_\Omega \varepsilon \nabla p \cdot \nabla \psi + (\mathbf{b} \cdot \nabla \psi)\, p + cp\psi \; dx \\ &= a(\psi, p) \end{aligned}$$

such that the weak formulation is given by

$$a(\psi, p) = (w, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega). \tag{3.1.9}$$

Definition 3.6. *An element $p \in H_0^1(\Omega)$ satisfying* (3.1.9) *is called weak solution of* (3.1.8).

Lemma 3.7. *The adjoint operator $S^* : L^2(\Omega) \to L^2(\Omega)$ is given by*

$$S^* w := \hat{p}$$

where $\hat{p} \in H_0^1(\Omega)$ is the weak solution of (3.1.8) *with respect to $w \in L^2(\Omega)$.*

Proof. Let $f, w \in L^2(\Omega)$ be arbitrary functions. Moreover, let $\hat{p} \in H_0^1(\Omega)$ be the unique weak solution of (3.1.8), i.e. we have

$$a(z, \hat{p}) = (w, z)_\Omega \quad \forall z \in H_0^1(\Omega).$$

For $f \in L^2(\Omega)$ the solution operator S yields $Sf = y$ where $y \in L^2(\Omega)$ is the unique weak solution of

$$a(y, z) = (f, z)_\Omega \quad \forall z \in H_0^1(\Omega).$$

Now, we choose $z = y$ in the first and $z = \hat{p}$ in the second weak formulation so that we obtain

$$(w, y)_\Omega = (f, \hat{p})_\Omega.$$

Due to the fact that we have choosen $f \in L^2(\Omega)$ and $w \in L^2(\Omega)$ arbitrarily, the definition of the adjoint operator S^* implies

$$(f, S^* w)_\Omega = (Sf, w)_\Omega = (y, w)_\Omega = (f, \hat{p})_\Omega.$$

Hence, we obtain

$$S^* w = \hat{p}.$$

The operator

$$\begin{aligned} S^* : L^2(\Omega) &\to L^2(\Omega) \\ w &\mapsto S^* w := \hat{p} \end{aligned}$$

is linear. Following the lines of the proof of Theorem 3.5, we can verify the continuity of S^*. □

3.1.1 Higher regularity of solutions

Theorem 3.5 ensures the existence of a unique weak solution $y \in H_0^1(\Omega)$ of (3.1.4). In our numerical analysis we will need higher regularity of y.

Theorem 3.8. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, convex and polygonal domain. Then, for every $u \in L^2(\Omega)$ there exists a unique solution $y \in H^2(\Omega)$ of the elliptic partial differential equation*

$$\begin{aligned} -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + cy &= u \quad \text{in } \Omega \\ y &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{3.1.10}$$

Moreover, there exists a constant $C > 0$, depending only on the domain and the data of the partial differential equation (3.1.10) *such that*

$$\| y \|_{2,\Omega} \leq C \| u \|_{0,\Omega} \tag{3.1.11}$$

is satisfied.

Proof. See [Gris85, Theorem 3.2.1.2]. □

3.2 Robin boundary condition

In the following we analyze the convection-diffusion reaction equation with Robin boundary conditions. According to Theorem 3.1 and Lemma 3.2, we have $V = H^1(\Omega)$ and $V^* = H^1(\Omega)^*$. Now, we consider the equation

$$\begin{aligned} -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + cy &= 0 \quad \text{in } \Omega \\ \varepsilon \partial_n y - \frac{\mathbf{b} \cdot n \cdot y}{2} &= u \quad \text{on } \Gamma. \end{aligned} \tag{3.2.1}$$

For the derivation of the weak formulation which corresponds to (3.2.1), let us assume that y is sufficient regular. Then, Green's first identity yields for $v \in H^1(\Omega)$

$$\begin{aligned}
&\int_\Omega (-\varepsilon\Delta y)v + (\mathbf{b}\cdot\nabla y)v + cyv\ dx \\
&= \int_\Omega \varepsilon\nabla y\cdot\nabla v + (\mathbf{b}\cdot\nabla y)\,v + cyv\ dx - \int_\Gamma \varepsilon(\partial_n y)v\ ds \\
&= \int_\Omega \varepsilon\nabla y\cdot\nabla v + (\mathbf{b}\cdot\nabla y)\,v + cyv\ dx - \int_\Gamma \frac{\mathbf{b}\cdot n\cdot yv}{2} + uv\ ds.
\end{aligned}$$

We define the bilinear form $a_\Gamma : H^1(\Omega)\times H^1(\Omega)\to\mathbb{R}$ by

$$a_\Gamma(y,v) := \int_\Omega \varepsilon\nabla y\cdot\nabla v + (\mathbf{b}\cdot\nabla y)\,v + cyv\ dx - \int_\Gamma \frac{\mathbf{b}\cdot n\cdot yv}{2}\ ds \tag{3.2.2}$$

and a linear, continuous functional $F : L^2(\Gamma)\to H^1(\Omega)^*$ by

$$L^2(\Gamma)\ni u\mapsto F_u(v) := \int_\Gamma uv\ ds \quad \forall v\in H^1(\Omega). \tag{3.2.3}$$

The weak formulation of (3.2.1) is given by

$$a_\Gamma(y,v) = F_u(v) \quad \forall v\in H^1(\Omega). \tag{3.2.4}$$

Definition 3.9. *An element $y\in H^1(\Omega)$ satisfying* (3.2.4) *is called weak solution of* (3.2.1) *.*

Lemma 3.10. *The bilinear form* (3.2.2) *is continuous and coercive, i.e. there exists a constant $C>0$ such that for all $y\in H^1(\Omega)$*

$$a_\Gamma(y,y)\geq C\parallel y\parallel_{1,\Omega}^2 .$$

Proof. The partial integration formula (Lemma 2.21) yields with Assumption (A2)

$$\int_\Omega (\mathbf{b}\cdot\nabla y)\,y\ dx = \int_\Gamma \frac{\mathbf{b}\cdot n\cdot y^2}{2}\ ds.$$

Hence, the bilinear form (3.2.2) is coercive, i.e. we have

$$\begin{aligned}
a_\Gamma(y,y) &= \int_\Omega \varepsilon\nabla y\cdot\nabla y + (\mathbf{b}\cdot\nabla y)\,y + cy^2\ dx - \int_\Gamma \frac{\mathbf{b}\cdot n\cdot y^2}{2}\ ds \\
&= \int_\Omega \varepsilon\nabla y\cdot\nabla y + cy^2\ dx \\
&\geq \varepsilon|y|_{1,\Omega}^2 + c_0\parallel y\parallel_{0,\Omega}^2 \\
&\geq \min\{\varepsilon, c_0\}\parallel y\parallel_{1,\Omega}^2 .
\end{aligned}$$

□

Theorem 3.11. *For every $u\in L^2(\Gamma)$ the equation* (3.2.1) *admits a unique weak solution $y\in H^1(\Omega)$. Moreover, there exists a constant $C>0$, depending on the domain Ω and on the data of* (3.2.1) *such that*

$$\parallel y\parallel_{1,\Omega}\leq C\parallel u\parallel_{0,\Gamma} . \tag{3.2.5}$$

Proof. First, the space $(H^1(\Omega), \|\cdot\|_{1,\Omega})$ is a Hilbert space. The functional F, i.e. (3.2.3) is linear and satisfies for all $v \in H^1(\Omega)$

$$\begin{aligned} |F_u(v)| &\leq \| u \|_{0,\Gamma} \| v \|_{0,\Gamma} \\ &\leq C_\tau \| u \|_{0,\Gamma} \| v \|_{1,\Omega} \end{aligned}$$

where we have used the trace inequality (Lemma 2.18). Hence, we obtain

$$\| F_u \|_{H^1(\Omega)^*} \leq C_\tau \| u \|_{0,\Gamma} .$$

Combined with Lemma 3.10, Lemma 3.2 (Lax-Milgram) yields the existence of a unique solution $y \in H^1(\Omega)$ for (3.2.4). Moreover, we obtain the $H^1(\Omega)$-a priori estimate

$$\begin{aligned} \min\{\varepsilon, c_0\} \| y \|_{1,\Omega} &\leq \| F_u \|_{H^1(\Omega)^*} \\ &\leq C_\tau \| u \|_{0,\Gamma} . \end{aligned}$$

□

As in the case of Dirichlet boundary conditions, we define the solution operator associated with the weak formulation (3.2.4). We introduce a linear and continuous operator denoted by $G : L^2(\Gamma) \to H^1(\Omega)$ that maps an arbitrary $u \in L^2(\Gamma)$ to the unique solution $y \in H^1(\Omega)$. The embedding operator is denoted by $E_H : H^1(\Omega) \to L^2(\Omega)$. Thus, we are able to define the solution operator $S : L^2(\Gamma) \to L^2(\Omega)$ by

$$u \mapsto y, \; y = Su = E_H G u. \tag{3.2.6}$$

For the derivation of the adjoint operator corresponding to S we consider the following convection-diffusion reaction equation

$$\begin{aligned} -\varepsilon \Delta p - \mathbf{b} \cdot \nabla p + cp &= w \quad \text{in } \Omega \\ \varepsilon \partial_n p + \frac{\mathbf{b} \cdot n \cdot p}{2} &= 0 \quad \text{on } \Gamma \end{aligned} \tag{3.2.7}$$

where $w \in L^2(\Omega)$. The associated bilinear form $\tilde{a}_\Gamma : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ of (3.2.7) is given by

$$\tilde{a}_\Gamma(p, \psi) := \int_\Omega \varepsilon \nabla p \cdot \nabla \psi - (\mathbf{b} \cdot \nabla p)\, \psi + cp\psi \; dx + \int_\Gamma \frac{\mathbf{b} \cdot n \cdot p\psi}{2} \; ds.$$

Lemma 2.21 yields

$$\begin{aligned} &- \int_\Omega (\mathbf{b} \cdot \nabla p)\, \psi \; dx + \int_\Gamma \frac{\mathbf{b} \cdot n \cdot p\psi}{2} \; ds \\ &= - \int_\Omega (\mathbf{b} \cdot \nabla p)\, \psi \; dx + \int_\Gamma \mathbf{b} \cdot n \cdot p\psi \; ds - \int_\Gamma \frac{\mathbf{b} \cdot n \cdot p\psi}{2} \; ds \\ &= \int_\Omega (\mathbf{b} \cdot \nabla \psi)\, p \; dx - \int_\Gamma \frac{\mathbf{b} \cdot n \cdot p\psi}{2} \; ds \end{aligned}$$

such that

$$\begin{aligned} \tilde{a}_\Gamma(p, \psi) &= \int_\Omega \varepsilon \nabla p \cdot \nabla \psi - (\mathbf{b} \cdot \nabla p)\, \psi + cp\psi \; dx + \int_\Gamma \frac{\mathbf{b} \cdot n \cdot p\psi}{2} \; ds \\ &= \int_\Omega \varepsilon \nabla p \cdot \nabla \psi + (\mathbf{b} \cdot \nabla \psi)\, p + cp\psi \; dx - \int_\Gamma \frac{\mathbf{b} \cdot n \cdot p\psi}{2} \; ds \\ &= a_\Gamma(\psi, p). \end{aligned}$$

The weak formulation of (3.2.7) is given by

$$a_\Gamma(\psi, p) = (w, \psi)_\Omega \quad \forall \psi \in H^1(\Omega). \tag{3.2.8}$$

Following the lines of Theorem 3.11, one can easily show that for every $w \in L^2(\Omega)$ the equation (3.2.8) is uniquely solvable in $H^1(\Omega)$.

Lemma 3.12. *The adjoint operator $S^* : L^2(\Omega) \to L^2(\Gamma)$ is given by*

$$S^* w := \hat{p}_{|\Gamma}$$

where $\hat{p} \in H^1(\Omega)$ is the weak solution of (3.2.8) *with respect to $w \in L^2(\Omega)$.*

Proof. Let $f \in L^2(\Gamma)$ and $w \in L^2(\Omega)$ be arbitrary functions. Moreover, let $\hat{p} \in H^1(\Omega)$ be the unique weak solution of (3.2.8), i.e. we have

$$a_\Gamma(z, \hat{p}) = (w, z)_\Omega \quad \forall z \in H^1(\Omega).$$

For $f \in L^2(\Gamma)$ the solution operator S yields $Sf = y$ where $y \in L^2(\Omega)$ is the unique solution of

$$a_\Gamma(y, z) = (f, z)_\Gamma \quad \forall z \in H^1(\Omega).$$

If we choose $z = y$ in the first and $z = \hat{p}$ in the second weak formulation we obtain

$$\int_\Omega wy \, dx = \int_\Gamma f\hat{p} \, ds.$$

Due to the fact that we have choosen $f \in L^2(\Gamma)$ and $w \in L^2(\Omega)$ arbitrarily the definition of the adjoint operator S^* implies

$$(f, S^* w)_\Gamma = (Sf, w)_\Omega = (y, w)_\Omega = (f, \hat{p})_\Gamma.$$

Hence, we obtain

$$S^* w = \hat{p}_{|\Gamma}.$$

The operator

$$\begin{aligned} S^* : L^2(\Omega) &\to L^2(\Gamma) \\ w &\mapsto S^* w := \hat{p}_{|\Gamma} \end{aligned}$$

is linear. Thanks to the coercivity of $a_\Gamma(\cdot, \cdot)$ and the trace inequality, we obtain

$$\| \hat{p} \|_{0,\Gamma} \leq C_\tau \| \hat{p} \|_{1,\Omega} \leq C_\tau C \| w \|_{0,\Omega}$$

with constants $C_\tau, C > 0$ so that the operator S^* is continuous. □

3.2.1 Higher regularity of solutions

For the derivation and validity of some convergence results which arise especially in Section 7.2 and Section 8 we use a higher regularity of the solutions than Theorem 3.11 provided. In this context we refer to [JeKe81] where the authors verify $H^{\frac{3}{2}}(\Omega)$-regularity of the solution for the homogeneous Neumann boundary value problem

$$\begin{aligned} -\varepsilon \Delta y &= 0 \quad \text{in } \Omega \\ \varepsilon \partial_n y &= G \quad \text{on } \Gamma \end{aligned} \tag{3.2.9}$$

provided that $G \in L^2(\Gamma)$ and $\int_\Gamma G\ ds = 0$. Following [Dha12, Theorem 1.12], one can easily show that the inhomogeneous Neumann boundary value problem

$$\begin{aligned} -\varepsilon\Delta y &= F \quad \text{in } \Omega \\ \varepsilon\partial_n y &= G \quad \text{on } \Gamma \end{aligned} \tag{3.2.10}$$

possesses a solution $y \in H^{\frac{3}{2}}(\Omega)$ provided that $F \in H^{s-2}(\Omega)$ for some $s \in (\frac{3}{2}, 2]$, $G \in L^2(\Gamma)$ and $\int_\Omega F\ dx + \int_\Gamma G\ ds = 0$. Moreover, one can show that

$$\| y \|_{\frac{3}{2},\Omega} + \| y \|_{1,\Gamma} \leq C(\| F \|_{0,\Omega} + \| G \|_{0,\Gamma}) \tag{3.2.11}$$

holds. According to (3.2.1), the verification of the $H^{\frac{3}{2}}(\Omega)$-regularity and the validity of an a priori estimate can be done by defining $F := -cy - \mathbf{b} \cdot \nabla y$ and $G := \frac{\mathbf{b}\cdot n \cdot y}{2} + u$. Testing (3.2.4) with $v = 1$, one can prove that the solvability condition $\int_\Omega F\ dx + \int_\Gamma G\ ds = 0$ holds. Due to Theorem 3.11 we have $y \in H^1(\Omega)$ and consequently $F \in L^2(\Omega)$. By virtue of Theorem 2.20 and $u \in L^2(\Gamma)$ we obtain $G \in L^2(\Gamma)$. The combination of (3.2.11) and the $H^1(\Omega)$-a priori estimate of y (see Theorem 3.11) yields

$$\| y \|_{\frac{3}{2},\Omega} + \| y \|_{1,\Gamma} \leq C \| u \|_{0,\Gamma}\ .$$

Apart from the derived $H^{\frac{3}{2}}(\Omega)$-regularity we are able to prove that $H^2(\Omega)$-regularity holds for the solution of (3.2.1). For this, it is sufficient to assume that $G \in \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$ where $\{\Gamma_i\}_{i=1}^m \subset \Gamma$ are the line segments of the boundary Γ (see Remark 2.15). Next, in the case where Ω is a quadrant, we show that there exists a solution $y \in H^2(\Omega)$ for the partial differential equation

$$\begin{aligned} -\varepsilon\Delta y &= F \quad \text{in } \Omega \\ \varepsilon\partial_n y &= G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots m. \end{aligned} \tag{3.2.12}$$

However, by virtue of [Gris85, Section 1.5, pp.47], one can prove $H^2(\Omega)$-regularity also for a rectilinear polygon Ω. Moreover, the H^2-regularity still holds when a suitable C^2-coordinate transformation (see [Ada75, Section 3, Transformation of Coordinates]) changes the angle of a rectangular corner. For proving the existence of a solution $y \in H^2(\Omega)$ for (3.2.12) we start with the verification of the existence of a function $\hat{y} \in H^2(\Omega)$ such that

$$\varepsilon\partial_n \hat{y} = G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots, m.$$

Lemma 3.13. *Let $\Omega \subseteq \mathbb{R}^2$ be a quadrant. Moreover, let $G = (G_1, \cdots, G_m) \in \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$. Then, there exists a function $\hat{y} \in H^2(\Omega)$ such that*

$$\varepsilon\partial_n \hat{y} = G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots, m.$$

Proof. Let us consider the first quadrant $\mathbb{R}_+ \times \mathbb{R}_+$ of Ω. In the following, we prove the surjectivity of the operator $T : H^2(\mathbb{R}_+ \times \mathbb{R}_+) \to H^{\frac{3}{2}}(\mathbb{R}_+) \times H^{\frac{1}{2}}(\mathbb{R}_+) \times H^{\frac{3}{2}}(\mathbb{R}_+) \times H^{\frac{1}{2}}(\mathbb{R}_+)$ which is defined by

$$y \mapsto Ty := (G_{0,1}, G_1, G_{0,2}, G_2)$$

where

$$\begin{aligned} G_{0,1} &= y_{|x_2=0} \\ G_1 &= (D_{x_2} y)_{|x_2=0} \\ G_{0,2} &= y_{|x_1=0} \\ G_2 &= (D_{x_1} y)_{|x_1=0}\ . \end{aligned}$$

For the surjectivity of T we follow [Gris85, Theorem 1.5.2.4] where it is sufficient to verify that the conditions

$$G_{0,1}(0) = G_{0,2}(0) \tag{3.2.13}$$

and

$$\int_0^1 |G_1(t) - DG_{0,2}(t)|^2 \, \frac{dt}{t} < \infty \tag{3.2.14}$$

$$\int_0^1 |DG_{0,1}(t) - G_2(t)|^2 \, \frac{dt}{t} < \infty \tag{3.2.15}$$

hold. Now, the strategy is to define the functions $G_{0,1}, G_{0,2}$ such that (3.2.13)-(3.2.15) are satisfied. We start with condition (3.2.14). We define

$$z(t) := G_1(t) - DG_{0,2}(t) \quad \forall t \in [0,1]$$

and $G_{0,2} := \int G_1 \, dt + C_1$, i.e. $G_{0,2}$ is the antiderivative to G_1 with a constant C_1. Then, we have $z(t) = 0$ for all $t \in [0,1]$ and thus

$$\int_0^1 z(t)^2 \, \frac{dt}{t} = 0$$

such that (3.2.14) is fulfilled. For the verification of condition (3.2.15) we proceed in the same way. We define

$$z(t) := DG_{0,1}(t) - G_2(t) \quad \forall t \in [0,1]$$

and $G_{0,1} := \int G_2 \, dt + C_2$ with a constant C_2. Hence, condition (3.2.15) is satisfied. Note that due to the regularity $G_1, G_2 \in H^{\frac{1}{2}}(\mathbb{R}_+)$ we have $G_{0,1}, G_{0,2} \in H^{\frac{3}{2}}(\mathbb{R}_+)$ and thus with the Sobolev embedding theorem $G_{0,1}, G_{0,2} \in C(\mathbb{R}_+)$. Now, for the satisfaction of (3.2.13), we have to adjust the constants $C_1, C_2 \in \mathbb{R}$. Let us choose $C_1 = 0$ and

$$C_2 := \left(\int G_1 \, dt\right)(0) - \left(\int G_2 \, dt\right)(0).$$

Then, we obtain

$$G_{0,2}(0) = \left(\int G_1 \, dt\right)(0) = \left(\int G_2 \, dt\right)(0) + C_2 = G_{0,1}(0).$$

Hence, condition (3.2.13) is fulfilled. Altogether, with the definition

$$\begin{aligned} G_{0,1} &:= \int G_1 \, dt \in H^{\frac{3}{2}}(\mathbb{R}_+) \\ G_{0,2} &:= \int G_2 \, dt + \left(\int G_1 \, dt\right)(0) - \left(\int G_2 \, dt\right)(0) \in H^{\frac{3}{2}}(\mathbb{R}_+) \end{aligned}$$

the conditions (3.2.13)-(3.2.15) are valid. Thus, [Gris85, Theorem 1.5.2.4] yields the existence of a function $\hat{y} \in H^2(\Omega)$ which satisfies

$$\varepsilon \partial_n \hat{y} = G_i \quad \text{on } \Gamma_i, \quad i = 1, 2.$$

The proof for the other quadrants goes in the same way. □

In the following, we prove the existence of a solution $y \in H^2(\Omega)$ for (3.2.12).

Lemma 3.14. *Let $\Omega \subseteq \mathbb{R}^2$ be a quadrant. Then, for every $(F, G) \in L^2(\Omega) \times \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$ there exists a solution $y \in H^2(\Omega)$ for the partial differential equation*

$$\begin{aligned} -\varepsilon \Delta y &= F \quad \text{in } \Omega \\ \varepsilon \partial_n y &= G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots m. \end{aligned} \tag{3.2.16}$$

Proof. Let us consider the boundary data of (3.2.16), i.e. $(G_1, \cdots, G_m) \in \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$. In the case where Ω is a quadrant, Lemma 3.13 yields the existence of a function $\hat{y} \in H^2(\Omega)$ such that

$$\varepsilon \partial_n \hat{y} = G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots, m.$$

Let us define the function $\hat{F} := \varepsilon \Delta \hat{y} + F \in L^2(\Omega)$. Then, the homogeneous Neumann boundary problem

$$\begin{aligned} -\varepsilon \Delta Y &= \hat{F} \quad \text{in } \Omega \\ \varepsilon \partial_n Y &= 0 \quad \text{on } \Gamma \end{aligned} \tag{3.2.17}$$

possesses a unique solution $Y \in H^2(\Omega)$. Consequently , the function $y := Y + \hat{y} \in H^2(\Omega)$ solves

$$-\varepsilon \Delta y = -\varepsilon \Delta Y - \varepsilon \Delta \hat{y} = \hat{F} - \varepsilon \Delta \hat{y} = F \quad \text{in } \Omega$$

and

$$\varepsilon \partial_n y = \varepsilon \partial_n Y + \varepsilon \partial_n \hat{y} = \varepsilon \partial_n \hat{y} = G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots, m.$$

Finally, we can conclude that $y \in H^2(\Omega)$ solves the partial differential equation

$$\begin{aligned} -\varepsilon \Delta y &= F \quad \text{in } \Omega \\ \varepsilon \partial_n y &= G_i \quad \text{on } \Gamma_i \quad i = 1, \cdots, m. \end{aligned} \tag{3.2.18}$$

□

In the case of a convex and polygonal domain Ω, one can follow [Pfeff15, Definition 2.19] where the author partitions the domain by an intersection with circles centered at the corners which do not overlap. Hence, the domain Ω is subdivided in circular sectors around the corners and the remaining part. According to [Pfeff15, Section 3.1.1, p.27], the H^2-regularity of a solution is valid on the remaining part. Using a suitable C^2-coordinate transformation, the H^2-regularity for quadrants (see Lemma 3.14) can be transferred to each circular sector. Then, the local H^2-regularity can be extended globally to the convex polygonal domain Ω by a partition of unity. Thus, we can state the next result.

Theorem 3.15. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, convex and polygonal domain. Then, for every $(F, G) \in L^2(\Omega) \times \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$ there exists a solution $y \in H^2(\Omega)$ for the partial differential equation*

$$\begin{aligned} -\varepsilon \Delta y &= F \quad \text{in } \Omega \\ \varepsilon \partial_n y &= G_i \quad \text{on } \Gamma_i, \quad i = 1, \cdots m. \end{aligned} \tag{3.2.19}$$

Now, we apply Theorem 3.15 for proving the $H^2(\Omega)$-regularity of a solution y for (3.2.1).

Theorem 3.16. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, convex and polygonal domain. Then, for every $(f, u) \in L^2(\Omega) \times H^{\frac{1}{2}}(\Gamma)$ there exists a unique solution $y \in H^2(\Omega)$ for the elliptic partial differential equation*

$$\begin{aligned} -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy &= f \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} &= u \quad \text{on } \Gamma. \end{aligned} \tag{3.2.20}$$

Moreover, there exists a constant $C > 0$, depending only on the domain and the data of the partial differential equation (3.2.20) *such that*

$$\| y \|_{2,\Omega} \leq C \left(\| f \|_{0,\Omega} + \| u \|_{\frac{1}{2},\Gamma} \right) \tag{3.2.21}$$

is satisfied.

Proof. Following the lines of Theorem 3.11, one can easily prove that there exists a unique solution $y \in H^1(\Omega)$ for (3.2.20). For the verification of the $H^2(\Omega)$-regularity of y and the corresponding $H^2(\Omega)$-a priori estimate, we define the operator $\mathcal{A} : H^2(\Omega) \to L^2(\Omega) \times \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$ by

$$y \mapsto \mathcal{A}y := (f, u_1, \cdots, u_m)$$

where for $i = 1, \cdots, m$

$$\begin{aligned} -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy &= f \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} &= u_i \quad \text{on } \Gamma_i \end{aligned} \tag{3.2.22}$$

and $\{\Gamma_i\}_{i=1}^m$ are the line segments of the boundary Γ. Note that every function $u \in H^{\frac{1}{2}}(\Gamma)$ belongs to the product space $\Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$ and the solution y of (3.2.20) also solves (3.2.22). In the following, we prove that the operator $\mathcal{A}$ is linear, bijective and continuous. Obviously, the operator $\mathcal{A}$ is linear.

Surjectivity:

Let $(f, u_1, \cdots, u_m) \in L^2(\Omega) \times \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)$ be arbitrary. The application of Green's formula yields the following weak formulation of (3.2.22)

$$\int_\Omega \varepsilon\nabla y\cdot\nabla v + (\mathbf{b}\cdot\nabla y)\, v + cyv \; dx - \int_\Gamma \frac{\mathbf{b}\cdot n\cdot yv}{2} \; ds = \int_\Omega fv \; dx + \sum_{i=1}^m \int_{\Gamma_i} u_i v \; ds \quad \forall v \in H^1(\Omega).$$

Following the lines of Theorem 3.11, one can easily prove by the application of Lemma 3.2 (Lax-Milgram) that there exists a unique solution $y \in H^1(\Omega)$. According to the continuity of the trace operator $\tau : H^1(\Omega) \to H^{\frac{1}{2}}(\Gamma)$ (see Theorem 2.20), we obtain $y \in H^{\frac{1}{2}}(\Gamma)$. Due to $\mathbf{b} \in W^{1,\infty}(\Omega)^2$ we have $\mathbf{b}\cdot n\cdot y \in H^{\frac{1}{2}}(\Gamma_i)$ and consequently $\frac{\mathbf{b}\cdot n\cdot y}{2} + u_i \in H^{\frac{1}{2}}(\Gamma_i)$ for $i = 1, \cdots, m$. Now, the $H^2(\Omega)$-regularity of a unique solution y for (3.2.22) follows from the application of Theorem 3.15 applied on the partial differential equation with $F := -\mathbf{b}\cdot\nabla y - cy + f$ and boundary conditions $\varepsilon\partial_n y = \frac{\mathbf{b}\cdot n\cdot y}{2} + u_i =: G_i$ on Γ_i for $i = 1, \cdots, m$. Hence, the operator $\mathcal{A}$ is surjective. This implies the $H^2(\Omega)$-regularity of y.

Injectivity:

For the verification of the injectivity, let $y \neq \tilde{y}$ be two solutions. We have to check that $\mathcal{A}y \neq \mathcal{A}\tilde{y}$ holds. Let us assume that $\mathcal{A}y = \mathcal{A}\tilde{y}$. Then, the weak formulation of (3.2.22) implies

$$\int_{\Omega} \varepsilon \nabla(y-\tilde{y}) \cdot \nabla v + (\mathbf{b} \cdot \nabla y - \tilde{y})\, v + c(y-\tilde{y})v \; dx - \int_{\Gamma} \frac{\mathbf{b} \cdot n \cdot (y-\tilde{y})v}{2} \; ds = 0 \quad \forall v \in H^1(\Omega).$$

The choice $v = y - \tilde{y}$ yields

$$\min\{\varepsilon, c_0\} \parallel y - \tilde{y} \parallel_{1,\Omega} \leq 0$$

so that we obtain $y = \tilde{y}$. This is a contradiction proving the injectivity of the operator $\mathcal{A}$.

Continuity:

Since the operator $\mathcal{A}$ is linear it sufficies to prove the boundedness, i.e.

$$\begin{aligned} \parallel \mathcal{A}y \parallel_{L^2(\Omega)\times\Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i)} &= \parallel f \parallel_{0,\Omega} + \Sigma_{i=1}^m \parallel u_i \parallel_{\frac{1}{2},\Gamma_i} \\ &\leq C \parallel y \parallel_{2,\Omega} . \end{aligned}$$

The proof of the boundedness of the operator $\mathcal{A}$ is straight forward. For this, note that we have

$$\begin{aligned} \parallel f \parallel_{0,\Omega} &= \parallel -\varepsilon\Delta y + \mathbf{b} \cdot \nabla y + cy \parallel_{0,\Omega} \\ &\leq C_1 \parallel y \parallel_{2,\Omega} \end{aligned}$$

where $C_1 > 0$ is a constant. Moreover, using the facts that the unit outward normal n is constant on each Γ_i, $i = 1, \cdots, m$ and $y \in H^2(\Omega)$, one can easily derive with the application of the trace inequality (Theorem 2.20)

$$\begin{aligned} \parallel u_i \parallel_{\frac{1}{2},\Gamma_i} &= \parallel \varepsilon\partial_n y - \frac{\mathbf{b} \cdot n \cdot y}{2} \parallel_{\frac{1}{2},\Gamma_i} \\ &\leq C \parallel y \parallel_{2,\Omega} \end{aligned}$$

where $C > 0$ is a constant which depends on the data of the problem. Hence, the operator $\mathcal{A}$ is bounded. Finally, the inverse mapping theorem [Brez11, Corollary 2.7] yields that the operator $\mathcal{A}$ has a continuous inverse $\mathcal{A}^{-1} : L^2(\Omega) \times \Pi_{i=1}^m H^{\frac{1}{2}}(\Gamma_i) \to H^2(\Omega)$ which is given by

$$\mathcal{A}^{-1}(f, u_1, \cdots, u_m) = y$$

with respect to (3.2.22). Through the continuity of $\mathcal{A}^{-1}$ we obtain the desired $H^2(\Omega)$-a priori estimate

$$\begin{aligned} \parallel y \parallel_{2,\Omega} &= \parallel \mathcal{A}^{-1}(f, u_1, \cdots, u_m) \parallel_{2,\Omega} \\ &\leq C_{\mathcal{A}^{-1}} \left(\parallel f \parallel_{0,\Omega} + \sum_{i=1}^{m} \parallel u_i \parallel_{\frac{1}{2},\Gamma_i} \right) \end{aligned} \tag{3.2.23}$$

where $C_{\mathcal{A}^{-1}} > 0$ is a constant, independent of y. As we have mentioned a solution y of (3.2.20) also solves (3.2.22). The unique solution of (3.2.20) thus belongs to $H^2(\Omega)$ and the corresponding $H^2(\Omega)$-a priori estimate holds by virtue of (3.2.23). $\square$

4 The optimal control problems

In this thesis, we focus on the following optimal control problems. First, we consider optimal control problems with distributed control and Dirichlet boundary condition. In detail, we investigate an unconstrained optimal control problem

$$\left.\begin{array}{l} \min\ J^f(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \end{array}\right\} (P_f)$$

Then, we analyze a control constrained optimal control problem

$$\left.\begin{array}{l} \min\ J^b(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ u_a \le u \le u_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_b)$$

where u_a, u_b are the control constraints which will be specified below. Another interesting case which will be dealt with here is the following state constrained optimal control without constraints on the control

$$\left.\begin{array}{l} \min\ J^s(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ y_a \le y \le y_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_s)$$

The bounds y_a, y_b will be specified below. In Section 4.3.3 we will derive an error estimate from a solution corresponding to (P_s) to the solution of an appropriate regularization of (P_s). Apart from the investigation of the above-mentioned optimal control problems with distributed control we study the following control constrained optimal control problem with Robin boundary control

$$\left.\begin{array}{l} \min\ J^\Gamma(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = 0 \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} = u \quad \text{on } \Gamma \\ u_a^\Gamma \le u \le u_b^\Gamma \quad \text{a.e. on } \Gamma \end{array}\right\} (P_\Gamma)$$

with control constraints u_a^Γ, u_b^Γ. ∂_n denotes the normal derivative with respect to the unit outward normal n. Assumptions on the optimal control problems are formulated as follows.

$(B1)$ The desired state y_d is a given function in $L^2(\Omega)$.

$(B2)$ The functions y_a, y_b of the pointwise state constraints belong to $C^{0,1}(\bar{\Omega})$ and satisfy

$$y_a \le 0 \le y_b \quad \text{a.e. on } \Gamma.$$

$(B3)$ The control constraints $u_a, u_b \in L^\infty(\Omega)$ satisfy

$$u_a < u_b \quad \text{a.e. on } \Omega.$$

$(B4)$ The control constraints $u_a^\Gamma, u_b^\Gamma \in H^1(\Gamma) \cap L^\infty(\Gamma)$ satisfy

$$u_a^\Gamma < u_b^\Gamma \quad \text{a.e. on } \Gamma.$$

$(B5)$ The Tikhonov regularization parameter $\lambda > 0$ is a fixed real number.

In the following sections, we introduce the analysis of $(P_f), (P_b), (P_s)$ and (P_Γ). In detail we verify the existence and uniqueness of solutions and provide first order sufficient and necessary optimality conditions which correspond to the above-mentioned optimal control problems. The analysis of problems with distributed control or boundary control can be found for instance in [Troel] or [Lions71]. For the analysis of the state constrained optimal control problem we refer to [Cas86], [KruRö08], [HtKu17] and [HzHt09]. Due to the fact that the verification of the existence and uniqueness of optimal solutions for the optimization problems with distributed control follows standard arguments, we keep these sections briefly. However, the analysis of (P_s) has been extended such that the corresponding sections will be illustrated more in detail.

4.1 Unconstrained case

We start with the unconstrained optimal control problem

$$\left.\begin{array}{l} \min\ J^f(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \end{array}\right\} (P_f)$$

4.1.1 Existence and uniqueness

For the verification of the existence and uniqueness of optimal solutions we use the following basic result of the infinite-dimensional optimization which can be found for instance in [Troel, Theorem 2.14].

Theorem 4.1. *Let $(U, \| \cdot \|_U)$ and $(H, \| \cdot \|_H)$ be two Hilbert spaces. Let $U_{ad} \subseteq U$ be a nonempty, convex, bounded and closed set. Furthermore, let $y_d \in H$, $\lambda \in \mathbb{R}_{>0}$ and $S : U \to H$ is a linear and continuous operator. Then, the optimization problem*

$$\min_{u \in U_{ad}} g(u) := \frac{1}{2} \parallel Su - y_d \parallel_H^2 + \frac{\lambda}{2} \parallel u \parallel_U^2 \tag{4.1.1}$$

admits a unique optimal solution $\bar{u}$.

Proof. Since $g(u) \geq 0$ there exists

$$j := \inf_{u \in U_{ad}} g(u).$$

Consequently, there exists a sequence $\{u_n\}_{n=1}^\infty \subseteq U_{ad}$ such that $g(u_n) \to j$ for $n \to \infty$. Due to the boundedness, closedness and the convexity of U_{ad} and the fact that U is a Hilbert space, there exists a weakly converging subsequence $\{u_{n_k}\}_{k=1}^\infty$ such that $u_{n_k} \rightharpoonup \bar{u}$ with $\bar{u} \in U_{ad}$. The continuity of the operator S implies the continuity of g. In combination with the convexity of g we obtain that the objective functional g is weakly lower semicontinuous. Hence, we obtain

$$j = \liminf_{k \to \infty} g(u_{n_k}) \geq g(\bar{u}).$$

Due to the fact that $\bar{u} \in U_{ad}$ we have $g(\bar{u}) = j$ and by assumption $\lambda > 0$ that the optimal solution $\bar{u}$ is unique. □

With the help of the control-to-state operator S, we rewrite the optimization problem (P_f) only with respect to the control u. According to Theorem 4.1, we define the so-called reduced form of (P_f) which is given by

$$\min_{u \in L^2(\Omega)} g(u) := J^f(Su, u) = \frac{1}{2} \parallel Su - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 . \tag{4.1.2}$$

Theorem 4.2. *There exists a unique optimal solution $\bar{u}$ for the optimization problem* (4.1.2).

Proof. Since $U_{ad} = L^2(\Omega)$ is not bounded, the original proof of Theorem 4.1 cannot be applied directly for proving the existence and uniqueness of an optimal solution for (4.1.2) . To apply Theorem 4.1, we restrict the set of admissible controls $L^2(\Omega)$ in the following way. Let $u_0 \in L^2(\Omega)$ be arbitrary. We define the set

$$U^1_{ad} := \{u \in L^2(\Omega) : \quad \| u \|^2_{0,\Omega} > 2\lambda^{-1} g(u_0)\}.$$

Then, we obtain

$$g(u) = J^f(Su, u) = \frac{1}{2} \| Su - y_d \|^2_{0,\Omega} + \frac{\lambda}{2} \| u \|^2_{0,\Omega} \geq \frac{\lambda}{2} \| u \|^2_{0,\Omega} > g(u_0) \quad \forall u \in U^1_{ad}. \tag{4.1.3}$$

Hence, we restrict the optimization problem (4.1.2) on the nonempty, convex, bounded and closed set

$$U^2_{ad} := \{u \in L^2(\Omega) : \quad \| u \|^2_{0,\Omega} \leq 2\lambda^{-1} g(u_0)\}$$

and consider

$$\min_{u \in U^2_{ad}} g(u). \tag{4.1.4}$$

Theorem 4.1 yields a unique optimal solution $\bar{u} \in U^2_{ad}$ for the optimization problem (4.1.4) such that

$$\min_{u \in U^2_{ad}} g(u) = g(\bar{u}).$$

Regarding (4.1.3) and the fact $u_0 \in U^2_{ad}$, we are able to derive

$$g(u) > g(u_0) \geq g(\bar{u}) \quad \forall u \in U^1_{ad}$$

and by (4.1.4)

$$g(u) > g(\bar{u}) \quad \forall u \in U^2_{ad} \setminus \{\bar{u}\}.$$

Hence, we can conclude that $\bar{u}$ is the unique optimal solution for the unconstrained optimal control problem (4.1.2), i.e.

$$\min_{u \in L^2(\Omega)} g(u) = g(\bar{u}).$$

□

The combination of Theorem 3.5 and Theorem 4.2 yields the following result.

Theorem 4.3. *There exists a unique optimal solution $(\bar{y}, \bar{u})$ for (P_f).*

The following necessary and sufficient optimality conditions of first order can be derived by the Lagrangian approach (see [Troel, Section 2.10] and [Troel, Theorem 2.25]). Note that in [Troel, Theorem 2.25] the considered differential operator contains no convection field. Thanks to Assumption (A2) (see Section 3) on the convection field **b**, the bilinear form (3.1.2) satisfies the assumptions in [Troel, Theorem 2.25]. Hence, the results in [Troel, Section 2.8.2, pp. 51] can be transferred to (P_f).

Theorem 4.4. *A pair $(\bar{y}, \bar{u}) \in H^1(\Omega) \times L^2(\Omega)$ is a solution of (P_f) if and only if there exists an adjoint solution $\bar{p} \in H^1(\Omega)$ such that the following optimality system is satisfied*

$$a(\bar{y}, v) = (\bar{u}, v)_\Omega \quad \forall v \in H_0^1(\Omega) \tag{4.1.5}$$

$$a(\psi, \bar{p}) = (\bar{y} - y_d, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \tag{4.1.6}$$

$$(\lambda \bar{u} + \bar{p}, u - \bar{u})_\Omega \geq 0 \quad \forall u \in L^2(\Omega). \tag{4.1.7}$$

According to [Troel, Theorem 2.28], a pointwise discussion of the variational inequality (4.1.7) yields in the case without any control constraints

$$\bar{p} + \lambda \bar{u} = 0 \quad \text{a.e. in } \Omega. \tag{4.1.8}$$

Remark 4.5 (Higher regularity)**.** *Due to the fact that $\bar{y}, \bar{p} \in H^1(\Omega)$ and $\bar{u} = -\frac{1}{\lambda}\bar{p}$, Theorem 3.8 yields $\bar{y} \in H^2(\Omega)$. Since $y_d \in L^2(\Omega)$ we have $\bar{p} \in H^2(\Omega)$.*

4.2 Control constrained case

In this section, we provide the analysis of the following control constrained optimal control problem

$$\left.\begin{array}{l} \min \ J^b(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ u_a \leq u \leq u_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_b)$$

4.2.1 Existence and uniqueness

For the verification of the existence and uniqueness of an optimal solution for (P_b), let us introduce the following nonempty, closed, convex, and bounded set of admissible controls

$$U_{ad} := \{u \in L^2(\Omega) : u_a(x) \leq u(x) \leq u_b(x) \quad \text{a.e. in } \Omega\}.$$

Similar to Section 4.1, we define the reduced functional

$$\min_{u \in U_{ad}} g(u) := J^b(Su, u) = \frac{1}{2} \parallel Su - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 . \tag{4.2.1}$$

By virtue of Theorem 4.1 we are able to state the following results.

Theorem 4.6. *There exists a unique optimal solution $\bar{u}$ for the optimization problem* (4.2.1).

Theorem 4.7. *There exists a unique optimal solution $(\bar{y}, \bar{u})$ for (P_b).*

Regarding [Troel, Theorem 2.25], necessary and sufficient optimality conditions for (P_b) can be written as follows.

Theorem 4.8. *A pair $(\bar{y}, \bar{u}) \in H^1(\Omega) \times L^2(\Omega)$ is an optimal solution of (P_b) if and only if there exists an adjoint solution $\bar{p} \in H^1(\Omega)$ such that the following optimality system is satisfied*

$$a(\bar{y}, v) = (\bar{u}, v)_\Omega \quad \forall v \in H_0^1(\Omega) \tag{4.2.2}$$

$$a(\psi, \bar{p}) = (\bar{y} - y_d, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \tag{4.2.3}$$

$$(\lambda \bar{u} + \bar{p}, u - \bar{u})_\Omega \geq 0 \quad \forall u \in U_{ad}. \tag{4.2.4}$$

Due to [Troel, Theorem 2.28] condition (4.2.4) is equivalent to

$$\bar{u} = \mathbb{P}_{[u_a,u_b]}\left(-\frac{1}{\lambda}\bar{p}\right) \quad \text{a.e. in } \Omega$$

where $\mathbb{P}_{[u_a,u_b]}(\cdot)$ is defined by Definition 2.24.

Remark 4.9 (Higher regularity)**.** *Due to the fact that* $\bar{u} = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda}\bar{p}) \in L^2(\Omega)$*, Theorem* 3.8 *yields* $\bar{y} \in H^2(\Omega)$*. Since* $y_d \in L^2(\Omega)$ *we have* $\bar{p} \in H^2(\Omega)$*.*

4.3 State constrained case

In this section, we investigate the state constrained optimal control problem (P_s). For this, recall the formulated optimal control problem

$$\left.\begin{array}{l} \min\ J^s(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ y_a \le y \le y_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_s)$$

4.3.1 Existence and uniqueness

Now, we prove the existence and uniqueness of a solution for (P_s). For this, we introduce the set of admissible controls by

$$U_{ad}^S := \{u \in L^2(\Omega) :\ y_a(x) \le Su(x) \le y_b(x) \quad \text{a.e. in } \Omega\}. \tag{4.3.1}$$

Similar to the proofs in the unconstrained case resp. the control constrained case which correspond to the existence of a solution, we have to assume that there exists a feasible control so that $U_{ad}^S \neq \emptyset$. However, we require a stronger condition, that will ensure the existence of Lagrange multipliers (see Remark 4.13).

Assumption 4.10 (Slater-condition)**.** *There exists a function* $\hat{u} \in L^2(\Omega)$ *such that*

$$y_a(x) < \hat{y}(x) < y_b(x) \quad \forall x \in \bar{\Omega}$$

where $\hat{y} = S\hat{u}$*.*

Note that the Slater-condition implies the existence of a feasible control $\hat{u} \in U_{ad}^S$ for (P_s). Furthermore, U_{ad}^S is convex, bounded and closed. Similar to the previous sections, we define the reduced form which corresponds to (P_s) in the following way

$$\min_{u \in U_{ad}^S} g(u) := J^s(Su,u) = \frac{1}{2} \parallel Su - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 . \tag{4.3.2}$$

By application of Theorem 4.1, one can easily verify the existence of a unique optimal solution for the optimization problem (4.3.2).

Theorem 4.11. *Suppose that Assumption 4.10 is fulfilled. Then, the optimization problem* (4.3.2) *admits a unique optimal solution* $\bar{u} \in U_{ad}^S$*.*

Another proof corresponding to the existence of a unique optimal solution for (P_s) can be found for instance in [Cas86, Theorem 1]. Now, we provide sufficient and necessary optimality conditions of first order. A detailed proof can be found in [Cas86, Theorem 2].

Theorem 4.12. *A pair $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ is the optimal solution of problem (P_s) if and only if there exist $\mu_a, \mu_b \in \mathcal{M}(\bar{\Omega})$, $\bar{p} \in L^2(\Omega)$ such that*

$$\begin{aligned} -\varepsilon\Delta\bar{y} + \mathbf{b}\cdot\nabla\bar{y} + c\bar{y} &= \bar{u} \quad \text{in } \Omega \\ \bar{y} &= 0 \quad \text{on } \Gamma \end{aligned}$$

$$\begin{aligned} -\varepsilon\Delta\bar{p} - \mathbf{b}\cdot\nabla\bar{p} + c\bar{p} &= \bar{y} - y_d + \mu_b - \mu_a \quad \text{in } \Omega \\ \bar{p} &= 0 \quad \text{on } \Gamma \end{aligned}$$

$$\begin{aligned} \bar{p} + \lambda\bar{u} &= 0 \quad \text{a.e. in } \Omega \\ \int_\Omega \varphi \, \mathrm{d}\mu_i &\geq 0 \quad \forall \varphi \in C(\bar{\Omega}),\ \varphi(x) \geq 0 \quad \forall x \in \bar{\Omega},\ i \in \{a, b\} \\ \bar{y} &\geq y_a \quad \text{a.e. in } \Omega,\ \langle \mu_a, y_a - \bar{y}\rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})} = 0 \\ \bar{y} &\leq y_b \quad \text{a.e. in } \Omega,\ \langle \mu_b, y_b - \bar{y}\rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})} = 0. \end{aligned}$$

Remark 4.13. *Note that the optimality conditions provided in Theorem 4.12 can be derived also by the Lagrangian approach where the Lagrangian can be defined by*

$$\mathcal{L}(\bar{u}, \mu_b, \mu_a) := \frac{1}{2} \parallel S\bar{u} - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel \bar{u} \parallel_{0,\Omega}^2 + \int_\Omega (S\bar{u} - y_b) \ d\mu_b + \int_\Omega (y_a - S\bar{u}) \ d\mu_a$$

with Lagrange multipliers μ_b, μ_a. A general introduction to the Lagrangian approach can be found for instance in [Troel, Section 2.10, Section 2.11].

Remark 4.14. *According to Theorem 4.12, the adjoint equation cannot be understood in the usual weak sense since on the right hand side occur measures $\mu_a, \mu_b \in \mathcal{M}(\bar{\Omega})$. Hence, the weak formulation of the adjoint equation can be read as follows:*

$$a(\psi, \bar{p}) = (\bar{y} - y_d, \psi)_\Omega + \langle \mu_b - \mu_a, \psi\rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})} \quad \forall \psi \in W^{1,s}(\Omega) \tag{4.3.3}$$

where $s > 2$. Due to the fact that in Section 4.3.3, equation (4.3.3) will be tested by functions belonging to $H^2(\Omega)$, we mention that the Sobolev embedding theorem (see Theorem 2.17) yields $H^2(\Omega) \hookrightarrow W^{1,s}(\Omega)$ for any $s > 2$.

As we can see above, the Lagrange multipliers μ_a, μ_b are in general no regular functions. In contrast to Theorem 3.8 we cannot expect that $\bar{p} \in H^2(\Omega)$ and consequently $\bar{p} \in C(\bar{\Omega})$. This lack of regularity causes difficulties in the numerical treatment of such problems. Due to this fact, many regularizations of (P_s) have been constructed in the last years. In this work, we regularize the state constrained optimal control problem (P_s) by the following Moreau-Yosida regularization which has been applied for instance in [HtKu17].

4.3.2 Moreau-Yosida regularization

The objective functional of the Moreau-Yosida regularization is defined by

$$\begin{aligned} &J^{MY}(y_\delta, u_\delta) := \\ &\qquad \frac{1}{2} \parallel y_\delta - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u_\delta \parallel_{0,\Omega}^2 + \frac{\delta}{2} \parallel \max\{0, y_\delta - y_b\} \parallel_{0,\Omega}^2 + \frac{\delta}{2} \parallel \min\{0, y_\delta - y_a\} \parallel_{0,\Omega}^2 \end{aligned}$$

where $\delta > 0$ is the Moreau-Yosida regularization parameter that is taken large. The Moreau-Yosida regularization of (P_s) can be stated as follows

$$\left.\begin{array}{l} \min \ J^{MY}(y_\delta, u_\delta) \\ -\varepsilon\Delta y_\delta + \mathbf{b}\cdot\nabla y_\delta + c y_\delta = u_\delta \quad \text{in } \Omega \\ y_\delta = 0 \quad \text{on } \Gamma \end{array}\right\} (P_s^{MY})$$

The proof of the existence and the uniqueness of an optimal solution for (P_s^{MY}) can be done in the same way as in Theorem 4.2 resp. Theorem 4.3.

Theorem 4.15. *There exists a unique optimal solution $(\bar{y}_\delta, \bar{u}_\delta)$ for (P_s^{MY}).*

The next result shows sufficient and necessary optimality conditions which correspond to (P_s^{MY}). The proof is straight forward by the application of the Lagrangian approach mentioned in Remark 4.13.

Theorem 4.16. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}). Then, a function $\mu_\delta \in L^2(\Omega)$ and an adjoint state $\bar{p}_\delta \in H^1(\Omega)$ exist such that the following optimality system is satisfied*

$$\begin{aligned} -\varepsilon\Delta\bar{y}_\delta + \mathbf{b}\cdot\nabla\bar{y}_\delta + c\bar{y}_\delta &= \bar{u}_\delta \quad \text{in } \Omega \\ \bar{y}_\delta &= 0 \quad \text{on } \Gamma \end{aligned}$$

$$\begin{aligned} -\varepsilon\Delta\bar{p}_\delta - \mathbf{b}\cdot\nabla\bar{p}_\delta + c\bar{p}_\delta &= \bar{y}_\delta - y_d + \mu_\delta \quad \text{in } \Omega \\ \bar{p}_\delta &= 0 \quad \text{on } \Gamma \end{aligned}$$

$$\begin{aligned} \bar{p}_\delta + \lambda\bar{u}_\delta &= 0 \quad \text{a.e. in } \Omega \\ \mu_\delta &= \delta\cdot(\max\{0, \bar{y}_\delta - y_b\} + \min\{0, \bar{y}_\delta - y_a\}) \in L^2(\Omega). \end{aligned}$$

In contrast to Theorem 4.12, the functions arising in Theorem 4.16 are sufficient regular such that the numerical treatment of (P_s^{MY}) is easier than (P_s).

4.3.3 Error analysis

According to Section 3.1, the optimal states $\bar{y}, \bar{y}_\delta$ corresponding to (P_s) and (P_s^{MY}) belong to $H^2(\Omega)$. By virtue of the Sobolev embedding theorem (see Theorem 2.17), we obtain that $\bar{y}, \bar{y}_\delta \in C^{0,\gamma}(\bar{\Omega})$ for $0 < \gamma < 1$. From now on, let us fix such a quantity $0 < \gamma < 1$. This section is dedicated to show that the optimal solutions $\{(\bar{y}_\delta, \bar{u}_\delta)\}_{\delta>0}$ of (P_s^{MY}) converge to the optimal solution $(\bar{y}, \bar{u})$ of (P_s). Moreover, we proof that a $L^2(\Omega)$-convergence rate of $\mathcal{O}(\frac{\gamma}{2\gamma+1})$ holds. We would like to mention that a similar result has been proven in [HzHt09]. However, in contrast to our error estimate, the provided $L^2(\Omega)$-error estimate is connected to the Finite Element Method, i.e. the error estimate depends on the Finite Element mesh size h. Now, let us consider the following result which ensures the convergence of the Moreau-Yosida regularization. A proof can be found in [HtKu17, Remark 2.1].

Theorem 4.17. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Then, we have for $\delta \to \infty$*

$$\begin{aligned} &\bar{y}_\delta \to \bar{y} \quad \text{in } H^2(\Omega) \cap H_0^1(\Omega) \\ &\quad \text{and} \\ &\bar{u}_\delta \to \bar{u} \quad \text{in } L^2(\Omega). \end{aligned}$$

In the following, we investigate the convergence behavior of $(\bar{y}_\delta, \bar{u}_\delta)$. For this, we first have to prove the uniform boundedness of the controls $\{\bar{u}_\delta\}_{\delta>0}$ in the $L^2(\Omega)$-norm.

Lemma 4.18. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Then, we have the uniform boundedness of $\{\bar{u}_\delta\}_{\delta>0}$ in the $L^2(\Omega)$-norm, i.e.*

$$\| \bar{u}_\delta \|_{0,\Omega} \leq C \quad \forall \delta > 0$$

where $C > 0$ is a constant which does not depend on δ.

Proof. Recalling the definitions of the functionals $J^s(y, u)$ and $J^{MY}(y, u)$ the optimality of $(\bar{y}, \bar{u})$ for (P_s) and the optimality of $(\bar{y}_\delta, \bar{u}_\delta)$ for (P_s^{MY}) yield for $\delta > 0$

$$J^{MY}(\bar{y}_\delta, \bar{u}_\delta) \leq J^{MY}(\bar{y}, \bar{u}) = J^s(\bar{y}, \bar{u}) := C.$$

Due to the definition of $J^{MY}(\cdot, \cdot)$, we obtain the uniform boundedness

$$\lambda \| \bar{u}_\delta \|_{0,\Omega} \leq C.$$

□

Now, we proof that for $\delta \to \infty$ the penalization terms $\max\{0, \bar{y}_\delta - y_b\}$ and $\min\{0, \bar{y}_\delta - y_a\}$ converge to 0 in the $L^2(\Omega)$-norm.

Lemma 4.19. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Then, we have the following decay rates*

$$\| \max\{0, \bar{y}_\delta - y_b\} \|_{0,\Omega} = \mathcal{O}\left(\frac{1}{\sqrt{\delta}}\right) \tag{4.3.4}$$

$$\| \min\{0, \bar{y}_\delta - y_a\} \|_{0,\Omega} = \mathcal{O}\left(\frac{1}{\sqrt{\delta}}\right). \tag{4.3.5}$$

Proof. The optimality of $(\bar{y}, \bar{u})$ for (P_s) and the optimality of $(\bar{y}_\delta, \bar{u}_\delta)$ for (P_s^{MY}) yield

$$J^{MY}(\bar{y}_\delta, \bar{u}_\delta) \leq J^{MY}(\bar{y}, \bar{u}) = J^s(\bar{y}, \bar{u}) := C.$$

Due to the definition of $J^{MY}(\cdot, \cdot)$, we obtain

$$\frac{\delta}{2} \| \max\{0, \bar{y}_\delta - y_b\} \|_{0,\Omega}^2 + \frac{\delta}{2} \| \min\{0, \bar{y}_\delta - y_a\} \|_{0,\Omega}^2 \leq C$$

and consequently the desired decay rates. □

Theorem 4.20. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Then, we have the following estimate*

$$\begin{aligned} &\lambda \| \bar{u} - \bar{u}_\delta \|_{0,\Omega}^2 + \| \bar{y} - \bar{y}_\delta \|_{0,\Omega}^2 \\ &\quad + \delta \| \max\{0, \bar{y}_\delta - y_b\} \|_{0,\Omega}^2 + \delta \| \min\{0, \bar{y}_\delta - y_a\} \|_{0,\Omega}^2 \leq -\langle \mu_b - \mu_a, \bar{y} - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})}. \end{aligned} \tag{4.3.6}$$

Proof. According to Theorem 4.12 and Theorem 4.16, the optimality conditions of (P_s) and (P_s^{MY}) yield

$$\lambda \| \bar{u} - \bar{u}_\delta \|_{0,\Omega}^2 = (\bar{u} - \bar{u}_\delta, \bar{p}_\delta - \bar{p})_\Omega = a(\bar{y} - \bar{y}_\delta, \bar{p}_\delta - \bar{p})$$

and

$$a(\bar{y} - \bar{y}_\delta, \bar{p}_\delta - \bar{p}) = (\bar{y}_\delta - \bar{y}, \bar{y} - \bar{y}_\delta)_\Omega + (\mu_\delta, \bar{y} - \bar{y}_\delta)_\Omega - \langle \mu_b - \mu_a, \bar{y} - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})}.$$

Hence, we acquire

$$\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 = (\mu_\delta, \bar{y} - \bar{y}_\delta)_\Omega - \langle \mu_b - \mu_a, \bar{y} - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})}. \tag{4.3.7}$$

Now, we estimate the first term on the right hand side of (4.3.7). For this, we define the sets

$$\Omega_{y_b} := \{x \in \Omega : \bar{y}_\delta(x) > y_b(x) \quad \text{a.e. in } \Omega\},$$

$$\Omega_{y_a, y_b} := \{x \in \Omega : y_a(x) \leq \bar{y}_\delta(x) \leq y_b(x) \quad \text{a.e. in } \Omega\}$$

and

$$\Omega_{y_a} := \{x \in \Omega : \bar{y}_\delta(x) < y_a(x) \quad \text{a.e. in } \Omega\}.$$

Note that

$$\Omega = \Omega_{y_a} \cup \Omega_{y_b} \cup \Omega_{y_a, y_b}$$

and

$$\mu_\delta = \delta \cdot (\max\{0, \bar{y}_\delta - y_b\} + \min\{0, \bar{y}_\delta - y_a\}) = 0 \quad \text{in } \Omega_{y_a, y_b}.$$

Consequently, we get

$$(\mu_\delta, \bar{y} - \bar{y}_\delta)_\Omega = \int\limits_{\Omega_{y_b}} \delta(\bar{y}_\delta - y_b)(\bar{y} - \bar{y}_\delta)\ dx + \int\limits_{\Omega_{y_a}} \delta(\bar{y}_\delta - y_a)(\bar{y} - \bar{y}_\delta)\ dx. \tag{4.3.8}$$

Inserting y_b in the first term and y_a in the second term of the right hand side of (4.3.8), we are able to derive

$$\begin{aligned}
\int\limits_{\Omega_{y_b}} \delta(\bar{y}_\delta - y_b)(\bar{y} - \bar{y}_\delta)\ dx &= \int\limits_{\Omega_{y_b}} \delta \overbrace{(\bar{y}_\delta - y_b)}^{>0} \overbrace{(\bar{y} - y_b)}^{\leq 0}\ dx + \int\limits_{\Omega_{y_b}} \delta(\bar{y}_\delta - y_b)(y_b - \bar{y}_\delta)\ dx \\
&\leq -\delta \parallel \bar{y}_\delta - y_b \parallel_{0,\Omega_{y_b}}^2 \\
&= -\delta \parallel \max\{0, \bar{y}_\delta - y_b\} \parallel_{0,\Omega}^2
\end{aligned}$$

and

$$\begin{aligned}
\int\limits_{\Omega_{y_a}} \delta(\bar{y}_\delta - y_a)(\bar{y} - \bar{y}_\delta)\ dx &= \int\limits_{\Omega_{y_a}} \delta \overbrace{(\bar{y}_\delta - y_a)}^{<0} \overbrace{(\bar{y} - y_a)}^{\geq 0}\ dx + \int\limits_{\Omega_{y_a}} \delta(\bar{y}_\delta - y_a)(y_a - \bar{y}_\delta)\ dx \\
&\leq -\delta \parallel \bar{y}_\delta - y_a \parallel_{0,\Omega_{y_a}}^2 \\
&= -\delta \parallel \min\{0, \bar{y}_\delta - y_a\} \parallel_{0,\Omega}^2 .
\end{aligned}$$

Using (4.3.7) we receive

$$\begin{aligned}
&\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \\
&\qquad + \delta \parallel \max\{0, \bar{y}_\delta - y_b\} \parallel_{0,\Omega}^2 + \delta \parallel \min\{0, \bar{y}_\delta - y_a\} \parallel_{0,\Omega}^2 \leq -\langle \mu_b - \mu_a, \bar{y} - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})}.
\end{aligned}$$

□

From now on, we use for a better clarity for $r \in \mathbb{R}$ also the notation

$$r^+ = \max\{0, r\}$$
$$r^- = \min\{0, r\}.$$

The next result shows a further estimation of the right hand side of (4.3.6) so that we are able to use the decay rates derived in Lemma 4.19.

Corollary 4.21. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Then, we have the following estimate*

$$\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \leq \langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} - \langle \mu_a, (\bar{y}_\delta - y_a)^- \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})}. \quad (4.3.9)$$

Proof. Recall the estimate of Theorem 4.20, i.e.

$$\begin{aligned} &\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \\ &\quad + \delta \parallel \max\{0, \bar{y}_\delta - y_b\} \parallel_{0,\Omega}^2 + \delta \parallel \min\{0, \bar{y}_\delta - y_a\} \parallel_{0,\Omega}^2 \leq -\langle \mu_b - \mu_a, \bar{y} - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})}. \end{aligned} \quad (4.3.10)$$

Inserting y_a, y_b on the right hand side of (4.3.10) yields

$$\begin{aligned} \lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 &\leq -\langle \mu_b - \mu_a, \bar{y} - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \\ &= \langle \mu_b - \mu_a, \bar{y}_\delta - \bar{y} \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \\ &= \langle \mu_b, \bar{y}_\delta - y_b \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} + \langle \mu_b, y_b - \bar{y} \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \\ &\quad - \langle \mu_a, \bar{y}_\delta - y_a \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} - \langle \mu_a, y_a - \bar{y} \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})}. \end{aligned} \quad (4.3.11)$$

The complementary slackness conditions of (P_s) (see Theorem 4.12), i.e.

$$\langle \mu_a, y_a - \bar{y} \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} = 0$$
$$\langle \mu_b, y_b - \bar{y} \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} = 0$$

yield for (4.3.11)

$$\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \leq \langle \mu_b, \bar{y}_\delta - y_b \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} - \langle \mu_a, \bar{y}_\delta - y_a \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})}.$$

Due to $\mu_b, \mu_a \geq 0$ and the fact that

$$\bar{y}_\delta - y_b = (\bar{y}_\delta - y_b)^+ + (\bar{y}_\delta - y_b)^-$$

resp.

$$y_a - \bar{y}_\delta = (y_a - \bar{y}_\delta)^+ + (y_a - \bar{y}_\delta)^-$$

we obtain

$$\begin{aligned} \lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 &\leq \langle \mu_b, \bar{y}_\delta - y_b \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} - \langle \mu_a, \bar{y}_\delta - y_a \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \\ &\leq \langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} + \langle \mu_a, y_a - \bar{y}_\delta \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \\ &\leq \langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} + \langle \mu_a, (y_a - \bar{y}_\delta)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \\ &= \langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} - \langle \mu_a, (\bar{y}_\delta - y_a)^- \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \end{aligned}$$

where in the last step we have used

$$(y_a - \bar{y}_\delta)^+ = -(\bar{y}_\delta - y_a)^-.$$

□

Now, we derive a $L^2(\Omega)$-estimate for the right hand side of (4.3.9). In general, we do not have $L^2(\Omega) \hookrightarrow L^\infty(\Omega)$. However, the following result yields an upper $L^2(\Omega)$-bound for functions $f \in C^{0,\gamma}(\bar{\Omega})$ with $0 < \gamma \leq 1$ in the $L^\infty(\Omega)$-norm.

Lemma 4.22. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, convex and polygonal domain. Moreover, let $f \in C^{0,\gamma}(\bar{\Omega})$ for some $0 < \gamma \leq 1$ with $\| f \|_{C^{0,\gamma}(\bar{\Omega})} \leq \varpi$. Then, there exists a constant $C > 0$ such that the estimate*

$$\| f \|_{0,\infty,\Omega} \leq C \varpi^{\frac{2}{2\gamma+2}} \| f \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}} \tag{4.3.12}$$

is satisfied.

Proof. A proof can be found in [KruRö08, Lemma 4] where the authors verify the result for an open and bounded domain $E \subseteq \mathbb{R}^d$ with $d = 2, 3$ satisfying the inner cone condition. Note that polygonal domains Ω satisfy the inner cone condition. □

The next result shows that the boundedness condition of Lemma 4.22 is satisfied by $(\bar{y}_\delta - y_b)^+$ and $(\bar{y}_\delta - y_a)^-$. For this, the uniform boundedness of the controls $\{\bar{u}_\delta\}_{\delta>0}$ in the $L^2(\Omega)$-norm will be used.

Lemma 4.23. *Let $0 < \gamma < 1$. Then, we have*

$$\| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})} \leq C_1$$
$$\| (\bar{y}_\delta - y_a)^- \|_{C^{0,\gamma}(\bar{\Omega})} \leq C_2$$

where $C_1, C_2 > 0$ are constants, independent of δ.

Proof. We start with the verification of the boundedness of $\| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})}$. The proof for $\| (\bar{y}_\delta - y_a)^- \|_{C^{0,\gamma}(\bar{\Omega})}$ goes along the same lines. First, recall the regularity assumption on the constraint $y_b \in C^{0,1}(\bar{\Omega})$ and recall the fact that $\bar{y}_\delta \in H^2(\Omega)$ holds. For $0 < \gamma < 1$ the Sobolev embedding $H^2(\Omega) \hookrightarrow C^{0,\gamma}(\bar{\Omega})$ implies

$$\begin{aligned} \| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})} &\leq \| \bar{y}_\delta - y_b \|_{C^{0,\gamma}(\bar{\Omega})} \\ &\leq \| \bar{y}_\delta \|_{C^{0,\gamma}(\bar{\Omega})} + \| y_b \|_{C^{0,\gamma}(\bar{\Omega})} \\ &\leq C \| \bar{y}_\delta \|_{2,\Omega} + \| y_b \|_{C^{0,\gamma}(\bar{\Omega})} . \end{aligned}$$

Theorem 3.8 and Lemma 4.18 imply

$$\| \bar{y}_\delta \|_{2,\Omega} \leq C \| \bar{u}_\delta \|_{0,\Omega} \leq \tilde{C}$$

with constants $C, \tilde{C} > 0$ such that

$$\| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})} \leq C_1.$$

□

Corollary 4.24. *Let $0 < \gamma < 1$. Then, we have*

$$\| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \leq C_1 \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}}$$
$$\| (\bar{y}_\delta - y_a)^- \|_{0,\infty,\Omega} \leq C_2 \| (\bar{y}_\delta - y_a)^- \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}}$$

where $C_1, C_2 > 0$ are constants, independent of δ.

Now, we combine the $L^2(\Omega)$-decay rates provided in Lemma 4.19 with Corollary 4.24.

Corollary 4.25. *Let $0 < \gamma < 1$ and $\delta > 0$. Then, we have*

$$\| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \leq C_1 \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+2}}$$
$$\| (\bar{y}_\delta - y_a)^- \|_{0,\infty,\Omega} \leq C_2 \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+2}}$$

where $C_1, C_2 > 0$ are constants, independent of δ.

The next result yields that $\{(\bar{y}_\delta, \bar{u}_\delta)\}_{\delta>0}$ converge in the $L^2(\Omega)$-norm to the unique optimal solution $(\bar{y}, \bar{u})$. Furthermore, we obtain a $L^2(\Omega)$-convergence rate of $\mathcal{O}\left(\frac{\gamma}{4\gamma+4}\right)$.

Theorem 4.26. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Moreover, let $0 < \gamma < 1$ and $\delta > 0$. Then, there exists a constant $C > 0$, independent of δ such that*

$$\lambda \| \bar{u} - \bar{u}_\delta \|_{0,\Omega}^2 + \| \bar{y} - \bar{y}_\delta \|_{0,\Omega}^2 \leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+2}}.$$

Proof. Corollary 4.25 implies

$$\begin{aligned}\langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} &\leq \| \mu_b \|_{\mathcal{M}(\bar{\Omega})} \| (\bar{y}_\delta - y_b)^+ \|_{C(\bar{\Omega})} \\ &\leq \| \mu_b \|_{\mathcal{M}(\bar{\Omega})} \| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \\ &\leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+2}}\end{aligned}$$

and in the same way

$$\begin{aligned}\langle \mu_a, (\bar{y}_\delta - y_a)^- \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} &\leq \| \mu_a \|_{\mathcal{M}(\bar{\Omega})} \| (\bar{y}_\delta - y_a)^- \|_{0,\infty,\Omega} \\ &\leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+2}}.\end{aligned}$$

Altogether, we obtain with Corollary 4.21

$$\lambda \| \bar{u} - \bar{u}_\delta \|_{0,\Omega}^2 + \| \bar{y} - \bar{y}_\delta \|_{0,\Omega}^2 \leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+2}}$$

where $C > 0$ is a constant, independent of δ. □

Corollary 4.27. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Moreover, let $0 < \gamma < 1$ and $\delta > 0$. Then, we have*

$$\| \bar{u} - \bar{u}_\delta \|_{0,\Omega} + \| \bar{y} - \bar{y}_\delta \|_{0,\Omega} \leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{\gamma}{2\gamma+2}} = C \left(\frac{1}{\delta}\right)^{\frac{\gamma}{4\gamma+4}}.$$

4.3.4 Improved $L^2(\Omega)$-order of convergence

In this section, we improve the order of convergence $\mathcal{O}\left(\frac{\gamma}{4\gamma+4}\right)$ to $\mathcal{O}\left(\frac{\gamma}{4\gamma+2}\right)$. Regarding Corollary 4.27, we will use the verified $L^2(\Omega)$-decay rate of $\bar{u} - \bar{u}_\delta$ so that we are able to improve the estimate of the right hand side of (4.3.9). The next result shows that the derived $L^2(\Omega)$-decay rate of $\bar{u} - \bar{u}_\delta$ can be used to bound the states:

Lemma 4.28. *Let $0 < \gamma < 1$. Then, we have*

$$\begin{aligned} \| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})} &\leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega} \\ \| (\bar{y}_\delta - y_a)^- \|_{C^{0,\gamma}(\bar{\Omega})} &\leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega} \end{aligned}$$

where C is a constant, independent of δ.

Proof. We start with the verification of the boundedness of $\| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})}$. The fact $\bar{y} - y_b \leq 0$ a.e. in Ω yields

$$\begin{aligned} (\bar{y}_\delta - y_b)^+ &= (\bar{y}_\delta - \bar{y} + \bar{y} - y_b)^+ \\ &\leq (\bar{y}_\delta - \bar{y})^+. \end{aligned}$$

Hence, the regularity assumption $y_b \in C^{0,1}(\bar{\Omega})$ and the fact $\bar{y}_\delta, \bar{y} \in H^2(\Omega)$ yield

$$\begin{aligned} \| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})} &\leq \| (\bar{y}_\delta - \bar{y})^+ \|_{C^{0,\gamma}(\bar{\Omega})} \\ &\leq \| \bar{y}_\delta - \bar{y} \|_{C^{0,\gamma}(\bar{\Omega})} \\ &\leq C \| \bar{y}_\delta - \bar{y} \|_{2,\Omega} . \end{aligned} \tag{4.3.13}$$

Recall that we have

$$a(\bar{y}_\delta - \bar{y}, v) = (\bar{u}_\delta - \bar{u}, v)_\Omega \quad \forall v \in H_0^1(\Omega). \tag{4.3.14}$$

Note that (4.3.14) is the weak formulation associated with the partial differential equation

$$\begin{aligned} -\varepsilon\Delta(\bar{y}_\delta - \bar{y}) + \mathbf{b} \cdot \nabla(\bar{y}_\delta - \bar{y}) + c(\bar{y}_\delta - \bar{y}) &= \bar{u}_\delta - \bar{u} \quad \text{in } \Omega \\ \bar{y}_\delta - \bar{y} &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Theorem 3.8 yields the $H^2(\Omega)$-a priori estimate

$$\| \bar{y}_\delta - \bar{y} \|_{2,\Omega} \leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega} .$$

Consequently, by virtue of the Sobolev embedding theorem, we obtain for (4.3.13)

$$\| (\bar{y}_\delta - y_b)^+ \|_{C^{0,\gamma}(\bar{\Omega})} \leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega} .$$

The proof for $\| (\bar{y}_\delta - y_a)^- \|_{C^{0,\gamma}(\bar{\Omega})}$ goes along the same lines. Note that we have $\bar{y} - y_a \geq 0$ a.e. in Ω and thus

$$\begin{aligned} (\bar{y}_\delta - y_a)^- &= (\bar{y}_\delta - \bar{y} + \bar{y} - y_a)^- \\ &\geq (\bar{y}_\delta - \bar{y})^-. \end{aligned}$$

Due to the fact that $|(r)^-| = -(r)^- \geq 0$ for all $r \in \mathbb{R}$ we have

$$\begin{aligned} \| (\bar{y}_\delta - y_a)^- \|_{C^{0,\gamma}(\bar{\Omega})} &\leq \| (\bar{y}_\delta - \bar{y})^- \|_{C^{0,\gamma}(\bar{\Omega})} \\ &\leq \| \bar{y}_\delta - \bar{y} \|_{C^{0,\gamma}(\bar{\Omega})} \\ &\leq C \| \bar{y}_\delta - \bar{y} \|_{2,\Omega} \\ &\leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega} . \end{aligned}$$

□

By using Lemma 4.28, a further application of Lemma 4.22 leads us to the next result.

Corollary 4.29. *Let $0 < \gamma < 1$. Then*

$$\| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^{\frac{2}{2\gamma+2}} \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}}$$
$$\| (\bar{y}_\delta - y_a)^- \|_{0,\infty,\Omega} \leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^{\frac{2}{2\gamma+2}} \| (\bar{y}_\delta - y_a)^- \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}}$$

hold where $C > 0$ is a constant, independent of δ.

Before we are able to improve the order of convergence, we use the following auxiliary estimates.

Lemma 4.30. *Let $0 < \gamma < 1$. Then*

$$\| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \leq \frac{\lambda \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^2}{(2\gamma+2) \| \mu_b \|_{\mathcal{M}(\bar{\Omega})}} + C \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}}$$
$$\| (\bar{y}_\delta - y_a)^- \|_{0,\infty,\Omega} \leq \frac{\lambda \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^2}{(2\gamma+2) \| \mu_a \|_{\mathcal{M}(\bar{\Omega})}} + C \| (\bar{y}_\delta - y_a)^- \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}}$$

hold where $C > 0$ is a constant, independent of δ.

Proof. We only verify the first inequality. The second goes along the same lines. Corollary 4.29 and a rearrangement of the terms yield

$$\begin{aligned} &\| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \\ &\quad \leq C \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^{\frac{2}{2\gamma+2}} \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}} \\ &\quad = \left(\frac{\lambda}{\| \mu_b \|_{\mathcal{M}(\bar{\Omega})}} \right)^{\frac{1}{2\gamma+2}} \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^{\frac{2}{2\gamma+2}} \, C \left(\frac{\| \mu_b \|_{\mathcal{M}(\bar{\Omega})}}{\lambda} \right)^{\frac{1}{2\gamma+2}} \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}} . \end{aligned} \tag{4.3.15}$$

We set $p = 2\gamma + 2 \in (1, \infty)$ and $q = \frac{2\gamma+2}{2\gamma+1} \in (1, \infty)$ so that the Young inequality (2.2.4) implies

$$\begin{aligned} &\left(\frac{\lambda}{\| \mu_b \|_{\mathcal{M}(\bar{\Omega})}} \right)^{\frac{1}{2\gamma+2}} \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^{\frac{2}{2\gamma+2}} \, C \left(\frac{\| \mu_b \|_{\mathcal{M}(\bar{\Omega})}}{\lambda} \right)^{\frac{1}{2\gamma+2}} \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+2}} \\ &\leq \frac{\left(\frac{\lambda}{\|\mu_b\|_{\mathcal{M}(\bar{\Omega})}} \right)^{\frac{2\gamma+2}{2\gamma+2}} \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^{\frac{2(2\gamma+2)}{2\gamma+2}}}{2\gamma+2} + \frac{C \left(\frac{\|\mu_b\|_{\mathcal{M}(\bar{\Omega})}}{\lambda} \right)^{\frac{2\gamma+2}{(2\gamma+2)(2\gamma+1)}} \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma(2\gamma+2)}{(2\gamma+2)(2\gamma+1)}}}{\frac{2\gamma+2}{2\gamma+1}} \\ &= \frac{\lambda \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^2}{\| \mu_b \|_{\mathcal{M}(\bar{\Omega})} (2\gamma+2)} + \frac{C \left(\frac{\|\mu_b\|_{\mathcal{M}(\bar{\Omega})}}{\lambda} \right)^{\frac{1}{2\gamma+1}} \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}}}{\frac{2\gamma+2}{2\gamma+1}} \\ &= \frac{\lambda \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^2}{\| \mu_b \|_{\mathcal{M}(\bar{\Omega})} (2\gamma+2)} + C \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}} . \end{aligned}$$

Hence, we get for (4.3.15)

$$\| (\bar{y}_\delta - y_b)^+ \|_{0,\infty,\Omega} \leq \frac{\lambda \| \bar{u}_\delta - \bar{u} \|_{0,\Omega}^2}{(2\gamma+2) \| \mu_b \|_{\mathcal{M}(\bar{\Omega})}} + C \| (\bar{y}_\delta - y_b)^+ \|_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}} .$$

□

The previous result implies that the $L^2(\Omega)$-decay rate of $\bar{u}_\delta - \bar{u}$ can be transferred to the right hand side of (4.3.9).

Theorem 4.31. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Moreover, let $0 < \gamma < 1$ and $\delta > 0$. Then, we have*

$$\lambda \left(\frac{2\gamma}{2\gamma+2}\right) \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+1}}.$$

Proof. Recall (4.3.9), i.e.

$$\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \leq \langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} - \langle \mu_a, (\bar{y}_\delta - y_a)^- \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})}.$$

Lemma 4.30 implies

$$\begin{aligned}
\langle \mu_b, (\bar{y}_\delta - y_b)^+ \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} &\leq \parallel \mu_b \parallel_{\mathcal{M}(\bar{\Omega})} \parallel (\bar{y}_\delta - y_b)^+ \parallel_{C(\bar{\Omega})} \\
&\leq \parallel \mu_b \parallel_{\mathcal{M}(\bar{\Omega})} \parallel (\bar{y}_\delta - y_b)^+ \parallel_{0,\infty,\Omega} \\
&\leq \parallel \mu_b \parallel_{\mathcal{M}(\bar{\Omega})} \frac{\lambda \parallel \bar{u}_\delta - \bar{u} \parallel_{0,\Omega}^2}{\parallel \mu_b \parallel_{\mathcal{M}(\bar{\Omega})} (2\gamma+2)} + C \parallel \mu_b \parallel_{\mathcal{M}(\bar{\Omega})} \parallel (\bar{y}_\delta - y_b)^+ \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}} \\
&= \frac{\lambda \parallel \bar{u}_\delta - \bar{u} \parallel_{0,\Omega}^2}{(2\gamma+2)} + C \parallel (\bar{y}_\delta - y_b)^+ \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}}
\end{aligned}$$

and in the same way

$$\langle \mu_a, (\bar{y}_\delta - y_a)^- \rangle_{\mathcal{M}(\bar{\Omega}),C(\bar{\Omega})} \leq \frac{\lambda \parallel \bar{u}_\delta - \bar{u} \parallel_{0,\Omega}^2}{(2\gamma+2)} + C \parallel (\bar{y}_\delta - y_a)^- \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}}.$$

Altogether, we have

$$\begin{aligned}
&\lambda \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 \\
&\qquad \leq \frac{2\lambda \parallel \bar{u}_\delta - \bar{u} \parallel_{0,\Omega}^2}{(2\gamma+2)} + C \parallel (\bar{y}_\delta - y_b)^+ \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}} + C \parallel (\bar{y}_\delta - y_a)^- \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}}
\end{aligned}$$

and with Lemma 4.19

$$\begin{aligned}
\lambda \left(\frac{2\gamma}{2\gamma+2}\right) \parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega}^2 + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega}^2 &\leq C \parallel (\bar{y}_\delta - y_b)^+ \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}} + C \parallel (\bar{y}_\delta - y_a)^- \parallel_{0,\Omega}^{\frac{2\gamma}{2\gamma+1}} \\
&\leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{2\gamma}{2\gamma+1}}
\end{aligned}$$

where $C > 0$ is a constant, independent of δ. □

Corollary 4.32. *Let $(\bar{y}_\delta, \bar{u}_\delta)$ be the optimal solution of (P_s^{MY}) and $(\bar{y}, \bar{u})$ be the optimal solution of (P_s). Moreover, let $0 < \gamma < 1$ and $\delta > 0$. Then, we have*

$$\parallel \bar{u} - \bar{u}_\delta \parallel_{0,\Omega} + \parallel \bar{y} - \bar{y}_\delta \parallel_{0,\Omega} \leq C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{\gamma}{2\gamma+1}} = C \left(\frac{1}{\delta}\right)^{\frac{\gamma}{4\gamma+2}}$$

where $C > 0$ is a constant, independent of δ.

4.4 Control constrained case with Robin boundary control

In the previous sections, we have investigated several optimal control problems with distributed control. Now, we provide the analysis of the following control constrained optimal control problem with Robin boundary control

$$\left.\begin{array}{l} \min\ J^{\Gamma}(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = 0 \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} = u \quad \text{on } \Gamma \\ u_a^{\Gamma} \le u \le u_b^{\Gamma} \quad \text{a.e. on } \Gamma \end{array}\right\} (P_{\Gamma})$$

4.4.1 Existence and uniqueness

The verification of the existence and the uniqueness of an optimal solution for (P_{Γ}) goes in a similar way as in the control constrained case with distributed control. Hence, we keep this section briefly. The set of admissible controls is defined by

$$U_{ad}^{\Gamma} := \{u \in L^2(\Gamma) : u_a^{\Gamma}(x) \le u(x) \le u_b^{\Gamma}(x) \quad \text{a.e. on } \Gamma\}.$$

Note that U_{ad}^{Γ} is closed, convex, and bounded. Due to Assumption (B4) the set U_{ad}^{Γ} is nonempty. Regarding Section 3.2, the control-to-state operator is given by $S : L^2(\Gamma) \to L^2(\Omega)$. Now, we consider the reduced form of (P_{Γ}), i.e.

$$\min_{u\in U_{ad}^{\Gamma}} g(u) := J^{\Gamma}(Su,u) = \frac{1}{2} \parallel Su - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 . \tag{4.4.1}$$

By application of Theorem 4.1, we are able to formulate the next results.

Theorem 4.33. *There exists a unique optimal solution $\bar{u}$ for the optimization problem* (4.4.1).

Theorem 4.34. *There exists a unique optimal solution $(\bar{y},\bar{u})$ for (P_{Γ}).*

Similar to [Troel, Section 2.8.4], sufficient and necessary optimality conditions of first order can be derived in so far as:

Lemma 4.35. *A pair $(\bar{y},\bar{u}) \in H^1(\Omega)\times L^2(\Gamma)$ is an optimal solution of (P_{Γ}) if and only if there exists an adjoint solution $\bar{p} \in H^1(\Omega)$ such that*

$$\begin{array}{ll} -\varepsilon\Delta\bar{y} + \mathbf{b}\cdot\nabla\bar{y} + c\bar{y} = 0 \quad \text{in } \Omega & -\varepsilon\Delta\bar{p} - \mathbf{b}\cdot\nabla\bar{p} + c\bar{p} = \bar{y} - y_d \quad \text{in } \Omega \\ \varepsilon\partial_n\bar{y} - \dfrac{\mathbf{b}\cdot n\cdot\bar{y}}{2} = \bar{u} \quad \text{on } \Gamma & \varepsilon\partial_n\bar{p} + \dfrac{\mathbf{b}\cdot n\cdot\bar{p}}{2} = 0 \quad \text{on } \Gamma \end{array}$$

$$(\lambda\bar{u} + \bar{p}, u - \bar{u})_{\Gamma} \ge 0 \quad \forall u \in U_{ad}^{\Gamma}$$

is satisfied.

The weak formulation of the above optimality system is given by

$$\begin{aligned} a_{\Gamma}(\bar{y},v) &= (\bar{u},v)_{\Gamma} \quad \forall v \in H^1(\Omega) \\ a_{\Gamma}(\psi,\bar{p}) &= (\bar{y} - y_d,\psi)_{\Omega} \quad \forall \psi \in H^1(\Omega) \\ (\lambda\bar{u} + \bar{p}, u - \bar{u})_{\Gamma} &\ge 0 \quad \forall u \in U_{ad}^{\Gamma}. \end{aligned}$$

The pointwise discussion of the variational inequality leads us to the following representation of the control $\bar{u}$ with the help of the projection formula

$$\bar{u} = \mathbb{P}_{[u_a^{\Gamma},u_b^{\Gamma}]}\left(-\frac{1}{\lambda}\bar{p}_{|\Gamma}\right).$$

4.4.2 Higher regularity of solutions

As we have mentioned in Section 3.2, for the derivation of the $H^2(\Omega)$-regularity of $\bar{y}$, we have to ensure a higher regularity of the control $\bar{u} \in L^2(\Gamma)$. Regarding Theorem 3.16, for proving $\bar{y} \in H^2(\Omega)$, it is sufficient to prove $\bar{u} \in H^{\frac{1}{2}}(\Gamma)$. However, we are even able to ensure $\bar{u} \in H^1(\Gamma)$.

Lemma 4.36. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times L^2(\Gamma)$ be an optimal solution of (P_Γ) with a corresponding adjoint solution $\bar{p} \in H^1(\Omega)$. Then, we have $\bar{u} \in H^1(\Gamma)$.*

Proof. Since we have $\bar{y} \in H^1(\Omega)$, Theorem 3.16 implies $\bar{p} \in H^2(\Omega)$. Then, [Nec12, Theorem 4.11] yields $\bar{p} \in H^1(\Gamma)$. The assumption $u_a^\Gamma, u_b^\Gamma \in H^1(\Gamma)$ and the application of [KinSta80, Theorem A.1, p. 50] lead us to

$$\bar{u} = \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}\left(-\frac{1}{\lambda}\bar{p}_{|\Gamma}\right) \in H^1(\Gamma).$$

□

As a direct consequence of Theorem 3.16 and Lemma 4.36, we obtain $\bar{y}, \bar{p} \in H^2(\Omega)$.

Corollary 4.37. *Let $(\bar{y}, \bar{u}) \in H^1(\Omega) \times H^1(\Gamma)$ be the optimal solution of (P_Γ) with a corresponding adjoint solution $\bar{p} \in H^1(\Omega)$. Then, we have $\bar{y}, \bar{p} \in H^2(\Omega)$.*

5 Discretization

In this section, we introduce the Finite Element Method and discuss the AFC discretization method. First, according to [Ciar02] and [BreSco02], we start with an introduction to the Finite Element discretization. After that, based on the results in [BJK16], [BBK17], [BJK17], [BJKR18] and [Kuz12] we exemplify the construction of an AFC scheme for a general linear boundary value problem with Dirichlet boundary condition. However, the case of a Robin boundary value problem is also covered when some adjustments on the indices hold. The modifications will be demonstrated at the end of this section. Furthermore, based on the results in the above-mentioned AFC literature, we discuss the solvability of the derived AFC scheme and conditions such that a discrete solution satisfies discrete maximum principles. After this general introduction we consider the convection-diffusion reaction equations (3.1.1) and (3.2.1). We prove error estimates in a mesh-dependent norm which also imply error estimates in the $L^2(\Omega)$-norm. These results will be used especially in Section 8. Finally, we concern with the design of some limiters and show that the sufficient conditions for the solvability of the corresponding AFC scheme and sufficient conditions for the DMP hold. We mention that the range of AFC literature increases continuously. However, an in-depth discussion is beyond the scope of this work and we will only concentrate on the basic AFC results which are used in this work.

5.1 Finite Element discretization

Let $\{\mathcal{T}_h\}_{h>0}$ be a family of triangulations of $\bar{\Omega}$. The mesh $\mathcal{T}_h$ consists of open and pairwise disjoint triangles $T \in \mathcal{T}_h$ such that

$$\bar{\Omega} = \bigcup_{T \in \mathcal{T}_h} \bar{T}.$$

Furthermore, no vertex of any triangle lies in the interior of an edge of another cell. For each element $T \in \mathcal{T}_h$ we define the parameters $\rho_T = \sup\{diam(S) : S \subset \mathbb{R}^2 \text{ is a ball in } T\}$ and

$h_T = diam(T)$. The maximum mesh size is given by $h := \max_{T \in \mathcal{T}_h} h_T$. Moreover, we assume that the triangulations satisfy the following regularity assumptions:

- Shape-regularity: There exists a constant $\sigma > 0$, independent of h such that for all $T \in \mathcal{T}_h$

$$\frac{h_T}{\rho_T} \leq \sigma.$$

- Size-regularity: There exists a constant $\hat{\sigma} > 0$, independent of h such that

$$h \leq \hat{\sigma} \min_{T \in \mathcal{T}_h} h_T.$$

$\mathcal{E}_h$ denotes the set of all edges. The Finite Element spaces are given by

$$V_h := \{z_h \in C(\bar{\Omega}) : z_h|_T \in \mathbb{P}_1(T) \ \forall \ T \in \mathcal{T}_h\}$$

and

$$V_{h,0} := V_h \cap H_0^1(\Omega)$$

where $\mathbb{P}_1(T)$ denotes the space of polynomials of degree less than or equal to 1 on T. The set of nodes of $\mathcal{T}_h$ is denoted by $\{x_1, \cdots, x_N\}$. In detail, $\{x_1, \cdots, x_M\}$ is the set of inner nodes and $\{x_{M+1}, \cdots, x_N\} \subset \Gamma$ is the set of boundary nodes. For a node x_i, $i = 1, \cdots, N$ we define the patch $\Delta_i := \{T \in \mathcal{T}_h : x_i \in T\}$ and the index set of its neighbors

$$S_i := \{j \in \{1, \cdots N\} \setminus \{i\} : \ x_i \text{ and } x_j \text{ are the endpoints of the same edge } E \in \mathcal{E}_h\}.$$

The standard nodal basis functions of V_h are denoted by $\{\varphi_1, \cdots, \varphi_N\}$ and satisfy the conditions

$$\varphi_i(x_j) = \delta_{ij}, \quad i, j = 1, \cdots, N \tag{5.1.1}$$

where δ_{ij} are the Kronecker-Delta functions (see Definition 2.22). Since (5.1.1) we have

$$supp(\varphi_i) = \Delta_i, \quad i = 1, \cdots, N. \tag{5.1.2}$$

The Lagrange interpolation operator $I_h : C(\bar{\Omega}) \to V_h$ is defined by

$$I_h z(x) := \sum_{i=1}^{N} z(x_i)\varphi_i(x).$$

5.1.1 Basic estimates in FEM-theory

In this section, we introduce basic results corresponding to the Finite Element Method (FEM) which can be found for instance in [BreSco02] and [Ciar02, Section 3]. First, we provide an interpolation error estimate for functions $z \in C(\bar{\Omega}) \cap H^2(\Omega)$ on convex and polygonal domains. For detailed information on the proof we refer to [Ciar02, Theorem 3.1.6, Theorem 3.2.1].

Lemma 5.1 (Interpolation error estimate)**.** *There holds for all $z \in C(\bar{\Omega}) \cap H^2(\Omega)$*

$$\| I_h z - z \|_{0,\Omega} + h|I_h z - z|_{1,\Omega} \leq Ch^2|z|_{2,\Omega}$$

where $C > 0$ is a constant, independent of h.

Lemma 5.2 (Inverse inequality)**.** *Let $z_h \in V_h$ and $1 \leq p \leq \infty$. Then, we have*

$$\| z_h \|_{1,p,\Omega} \leq Ch^{-1} \| z_h \|_{0,p,\Omega}$$

where $C > 0$ is a constant, independent of h.

Proof. See [BreSco02, Lemma 4.5.3]. □

The next result provides an $L^\infty(\Omega)$-error estimate for functions $z_h \in V_h$. A detailed proof can be found for instance in [BreSco02, Lemma 4.9.1].

Lemma 5.3 (Discrete Sobolev inequality)**.** *There holds for all $z_h \in V_h$*

$$\| z_h \|_{0,\infty,\Omega} \leq C \left(1 + |\ln h|\right)^{\frac{1}{2}} \| z_h \|_{1,\Omega}$$

where $C > 0$ is a constant, independent of h.

5.2 AFC method for linear Dirichlet boundary value problems

In this section, we will introduce a general construction of an AFC scheme, generally based on the results in [BJK16]. For this, let us consider a linear boundary value problem with Dirichlet boundary condition. Furthermore, the case of a Robin boundary value problem is also covered. The appropriate modifications in the case of Robin boundary conditions will be illustrated later in this section. Let us return to the case of Dirichlet boundary conditions. We can discretize this problem using a conforming Finite Element Method. Then, the discrete solution can be represented by a vector $z \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective Finite Element space. The components $\{M+1, \cdots, N\}$ of z correspond to the nodes where Dirichlet boundary conditions are prescribed. The components $\{1, \cdots, M\}$ of z correspond to the inner nodes where these components are computed using the Finite Element discretization of the underlying PDE. Then, the discrete solution $z = (z_1, \cdots, z_N) \in \mathbb{R}^N$ satisfies a system of linear equations of the form

$$\begin{aligned} \sum_{j=1}^{N} a_{ij} z_j &= g_i, \quad i = 1, \cdots, M \\ z_i &= g_i, \quad i = M+1, \cdots, N \end{aligned} \tag{5.2.1}$$

where $G = (g_1, \cdots, g_M, g_{M+1}, \cdots, g_N) \in \mathbb{R}^N$ is a given right hand side vector.

Remark 5.4. *Due to the compact support property of the standard nodal basis functions, only nodes belonging to S_i produce nonvanishing entries in the i-th row of the mass matrix $\mathcal{M}_{mass} = (m_{ij})_{i,j=1}^N$ which is defined by*

$$m_{ij} = \int_\Omega \varphi_i \varphi_j \, dx.$$

We remark that subsequently all discrete operators have the same compact sparsity pattern as the matrix $\mathcal{M}_{mass} = (m_{ij})_{i,j=1}^N$.

Now, let us return to the system (5.2.1). We assume that the matrix $(a_{ij})_{i,j=1}^M$ is positive definite, i.e. there exists a constant $C > 0$ such that

$$\sum_{i,j=1}^{M} z_i a_{ij} z_j \geq C \sum_{i=1}^{M} z_i^2 > 0 \quad \forall \, (z_1, \cdots, z_M) \in \mathbb{R}^M \setminus \{0\}. \tag{5.2.2}$$

The positive definiteness of the matrix $(a_{ij})_{i,j=1}^{M}$ implies that

$$a_{ii} > 0, \quad i = 1, \cdots, M. \tag{5.2.3}$$

Regarding Remark 5.4, the entries of the matrix $(a_{ij})_{i,j=1}^{N}$ satisfy the condition

$$a_{ij} = 0 \quad \forall\, i, j \in \{1, \cdots, N\},\ j \notin S_i.$$

The motivation for the construction of an AFC scheme is that the discrete solution $z \in \mathbb{R}^N$ should be free of spurious oscillations. This property can be expressed by the satisfaction of the local discrete maximum principle (local DMP).

Definition 5.5. *We say that a solution $z \in \mathbb{R}^N$ of* (5.2.1) *satisfies the local DMP if for $i \in \{1, \cdots, M\}$*

$$g_i \leq 0 \Rightarrow z_i \leq \max_{j \in S_i} z_j^+ \tag{5.2.4}$$

$$g_i \geq 0 \Rightarrow z_i \geq \min_{j \in S_i} z_j^- \tag{5.2.5}$$

hold.

In [BJK16, Lemma 21] the authors show that the local DMP holds if the conditions

$$a_{ij} \leq 0 \quad \forall\, i \neq j,\ i = 1, \cdots, M,\ j = 1, \cdots, N \tag{5.2.6}$$

and

$$\sum_{j=1}^{N} a_{ij} \geq 0 \quad \forall\, i \in \{1, \cdots, M\} \tag{5.2.7}$$

are satisfied.

Lemma 5.6. *Let the conditions* (5.2.6) *and* (5.2.7) *be satisfied. Then, for a solution $z \in \mathbb{R}^N$ of* (5.2.1) *the local DMP* (5.2.4) *resp.* (5.2.5) *hold.*

Proof. We start with the verification of claim (5.2.4). For this, let $i \in \{1, \cdots, M\}$, $g_i \leq 0$ and $\bar{C} := \max_{j \in S_i} z_j^+$. Due to condition (5.2.7), we have

$$a_{ii} \geq \sum_{j \neq i} (-a_{ij}) \geq 0.$$

Using (5.2.1) and the compact sparsity pattern-property, we are able to derive the following inequalities

$$a_{ii} z_i \leq \sum_{j \neq i} (-a_{ij}) z_j = \sum_{j \in S_i} (-a_{ij}) \underbrace{(z_j - \bar{C})}_{\leq 0} + \sum_{j \in S_i} (-a_{ij}) \bar{C} \leq \left(\sum_{j \in S_i} (-a_{ij}) \right) \bar{C} \leq a_{ii}\ \bar{C}.$$

The positivity of a_{ii} yields the desired result. For the verification of claim (5.2.5) let $\underline{C} := \min_{j \in S_i} z_j^-$ and $g_i \geq 0$. As above, one can easily derive

$$a_{ii} z_i \geq \sum_{j \neq i} (-a_{ij}) z_j = \sum_{j \in S_i} (-a_{ij}) \underbrace{(z_j - \underline{C})}_{\geq 0} + \sum_{j \in S_i} (-a_{ij}) \underline{C} \geq \left(\sum_{j \in S_i} (-a_{ij}) \right) \underline{C} \geq a_{ii}\ \underline{C}$$

obtaining the desired result. □

5.2.1 Construction of an AFC scheme

The starting point of the AFC algorithm is the system matrix $\mathcal{A} = (a_{ij})_{i,j=1}^N$ which will be correspond in the following sections to the stiffness matrices of the bilinear forms $a(\cdot,\cdot)$ resp. $a_\Gamma(\cdot,\cdot)$. As we have mentioned above, condition (5.2.6) is sufficient for the validity of the discrete maximum principle. However, this condition is only satisfied in a few simple situations. For instance, consider the introduced convection-diffusion reaction equation with homogeneous Dirichlet boundary condition (3.1.1). The entries a_{ij} of the stiffness matrix are given by

$$a_{ij} := a(\varphi_j, \varphi_i) = \int_\Omega \varepsilon \nabla\varphi_j \cdot \nabla\varphi_i + (\mathbf{b}\cdot\nabla\varphi_j)\,\varphi_i + c\varphi_j\varphi_i \; dx, \quad i,j = 1,\cdots,N. \tag{5.2.8}$$

Integration by parts yields for $i,j = 1,\cdots,N$

$$\begin{aligned} a(\varphi_j,\varphi_i) :&= \int_\Omega \varepsilon\nabla\varphi_j\cdot\nabla\varphi_i + (\mathbf{b}\cdot\nabla\varphi_j)\,\varphi_i + c\varphi_j\varphi_i \; dx \\ &= \int_\Omega \varepsilon\nabla\varphi_j\cdot\nabla\varphi_i - (\mathbf{b}\cdot\nabla\varphi_i)\,\varphi_j + c\varphi_i\varphi_j \; dx \end{aligned}$$

such that the matrix entries of $(a_{ij})_{i,j=1}^M$, which correspond to the inner nodes, are not symmetric. Consequently, condition (5.2.6) is not satisfied. This non-symmetry holds also in the case of a convection-diffusion reaction equation with Robin boundary condition (3.2.1). To achieve that condition (5.2.6) is satisfied, we modify the discrete problem (5.2.1). For this, we introduce a symmetric artificial diffusion matrix $\mathcal{D} = \mathcal{D}(\mathcal{A}) = (d_{ij})_{i,j=1}^N$ which is defined by

$$d_{ij} := \begin{cases} -\max\{a_{ij}, 0, a_{ji}\} & , \quad i \neq j \\ -\sum\limits_{l\neq i} d_{il} & , \quad i = j. \end{cases} \tag{5.2.9}$$

Remark 5.7. *In [Kuz12, Section 5.1] the author verifies that given an arbitrary matrix $\mathcal{A} \in \mathbb{R}^{N\times N}$ the entries of $\mathcal{A}+\mathcal{D} = (a_{ij}+d_{ij})_{i,j=1}^N$ satisfy the condition*

$$a_{ij} + d_{ij} \leq 0 \quad \forall\, i \neq j, \quad i,j = 1,\cdots,N. \tag{5.2.10}$$

Moreover, row and column sums of $\mathcal{A}+\mathcal{D} = (a_{ij}+d_{ij})_{i,j=1}^N$ coincide with those of $\mathcal{A}$. The artificial diffusion matrix $\mathcal{D} \in \mathbb{R}^{N\times N}$ is positive semidefinite.

The combination of condition (5.2.2) and Remark 5.7 yields that the matrix $\mathcal{A}+\mathcal{D} = (a_{ij}+d_{ij})_{i,j=1}^N$ has positive diagonal entries

$$a_{ii} + d_{ii} > 0, \quad i = 1,\cdots,M$$

and non-positive off-diagonal entries. When condition (5.2.7) is valid, then the matrix $\mathcal{A}+\mathcal{D}$ satisfies sufficient conditions for the local DMP. After introducing the matrix $\mathcal{D}$, let us start with the modification of (5.2.1). With the help of the matrix $\mathcal{D}$, we consider for (5.2.1) an equivalent system

$$\begin{aligned} [(\mathcal{A}+\mathcal{D})z]_i &= g_i + (\mathcal{D}z)_i, \quad i = 1,\cdots,M \\ z_i &= g_i, \quad i = M+1,\cdots,N. \end{aligned} \tag{5.2.11}$$

Since the row sums of the matrix $\mathcal{D}$ vanish, it follows that

$$(\mathcal{D}z)_i = \sum_{j=1}^N d_{ij}(z_j - z_i), \quad i = 1,\cdots,N$$

where $d_{ij}(z_j - z_i)$ are the so-called fluxes. Now, the idea of the AFC method is to limit those fluxes that would otherwise leads to spurious oscillations in the discrete solution. By limiting the fluxes we obtain the system

$$\begin{aligned} \sum_{j=1}^{N} a_{ij} z_j + \sum_{j=1}^{N} (1 - \alpha_{ij}(z)) d_{ij}(z_j - z_i) &= g_i, \quad i = 1, \cdots, M \\ z_i &= g_i, \quad i = M+1, \cdots, N \end{aligned} \tag{5.2.12}$$

where $\alpha_{ij} \in [0,1]$ for $i, j \in \{1, \cdots, N\}$, $i \neq j$ are solution-dependent limiters. Note that for $\alpha_{ij} = 1$ we obtain the original discrete system (5.2.1). Moreover, we mention that the coefficients α_{ij} should be as close to 1 as possible to limit the modification of (5.2.1).

5.2.2 Solvability of an AFC scheme

In this section, we prove the solvability of (5.2.12). From now on, we require that the limiters are symmetric, i.e. we have

$$\alpha_{ij} = \alpha_{ji}, \quad i \neq j. \tag{5.2.13}$$

First, in [BJK15] the authors show that when the symmetry condition (5.2.13) does not hold, it is not ensured that there exists a solution for (5.2.12). To prove the existence of a discrete solution for (5.2.12), we use the following fixed-point theorem which can be found in [GirRav86, Lemma 1.4].

Lemma 5.8. *Let X be a finite-dimensional Hilbert space with inner product $(\cdot,\cdot)_X$ and norm $\|\cdot\|_X$. Let $T : X \to X$ be a continuous mapping and $K > 0$ a real number such that $(Tz, z)_X > 0$ for any $z \in X$ with $\| z \|_X = K$ Then, there exists $z \in X$ such that $\| z \|_X < K$ and $Tz = 0$.*

Apart from the symmetry assumption (5.2.13) for the verification of the solvability of (5.2.12), we have to require that the limiters are continuous in so far as:

Assumption 5.9. *Consider for $i, j \in \{1, \cdots, N\}$ with $j \neq i$ the map $\alpha_{ij} : \mathbb{R}^N \to [0,1]$. We assume that*

$$\mathbb{R}^N \ni z \mapsto \alpha_{ij}(z)(z_j - z_i) \in \mathbb{R}$$

is a continuous function of $z_1, \cdots, z_N$.

Now, with the help of the following auxiliary result the existence of a discrete solution can be proved.

Lemma 5.10. *Let $N \in \mathbb{N}_{\geq 1}$ be arbitrary. Consider any $\beta_{ij} = \beta_{ji} \leq 0$, $i, j = 1, \cdots, N$. Then*

$$\sum_{i,j=1}^{N} z_i \beta_{ij}(z_j - z_i) = -\sum_{\substack{i,j=1 \\ i<j}}^{N} \beta_{ij}(z_i - z_j)^2 \geq 0 \quad \forall z_1, \cdots, z_N \in \mathbb{R}.$$

Proof. See [BJK16, Lemma 1]. □

The next result shows that there exists a discrete solution for (5.2.12) when the limiters satisfy the continuity assumption, i.e. Assumption 5.9. A proof can be found for instance in [BJK16, Theorem 3]. However, due to the meaning of the result we illustrate the proof.

Theorem 5.11. *Let Assumption 5.9 be satisfied. Then, the AFC system* (5.2.12) *possesses at least one solution* $z \in \mathbb{R}^N$.

Proof. Let us denote by $\tilde{z} = (z_1, \cdots, z_M)$ the elements of the space $\mathbb{R}^M$. For $i \in \{M+1, \cdots, N\}$ we assume that $z_i = g_i$. To any $\tilde{z} \in \mathbb{R}^M$ we assign the vector $z := (z_1, \cdots, z_N)$. Let us define the operator $T : \mathbb{R}^M \to \mathbb{R}^M$ by

$$(T\tilde{z})_i := \sum_{j=1}^{N} a_{ij} z_j + \sum_{j=1}^{N} (1 - \alpha_{ij}(z)) d_{ij}(z_j - z_i) - g_i, \quad i = 1, \cdots, M. \tag{5.2.14}$$

Then, $z \in \mathbb{R}^N$ is a solution of (5.2.12) if and only if $T\tilde{z} = 0$. First, the operator T is continuous. Due to the positive definiteness (5.2.2), we can write

$$\sum_{i,j=1}^{M} z_i a_{ij} z_j \geq C_M \parallel \tilde{z} \parallel_2^2 \tag{5.2.15}$$

where $C_M > 0$ is a constant and $\| \cdot \|_2$ denotes the euclidean norm on $\mathbb{R}^M$. We have

$$\begin{aligned}(T\tilde{z}, \tilde{z})_2 = &\sum_{i,j=1}^{M} z_i a_{ij} z_j + \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i a_{ij} g_j + \sum_{i=1}^{M} \sum_{j=1}^{M} z_i (1 - \alpha_{ij}(z)) d_{ij}(z_j - z_i) \\ &- \sum_{i=1}^{M} g_i z_i + \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i (1 - \alpha_{ij}(z)) d_{ij}(g_j - z_i).\end{aligned} \tag{5.2.16}$$

By virtue of Lemma 5.10 we obtain

$$\sum_{i=1}^{M} \sum_{j=1}^{M} z_i (1 - \alpha_{ij}(z)) d_{ij}(z_j - z_i) \geq 0.$$

Furthermore, we have

$$\sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i (1 - \alpha_{ij}(z)) d_{ij}(g_j - z_i) \geq \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i (1 - \alpha_{ij}(z)) d_{ij} g_j.$$

(5.2.16) can be further estimated by

$$\begin{aligned}(T\tilde{z}, \tilde{z})_2 = &\sum_{i,j=1}^{M} z_i a_{ij} z_j + \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i a_{ij} g_j + \sum_{i=1}^{M} \sum_{j=1}^{M} z_i (1 - \alpha_{ij}(z)) d_{ij}(z_j - z_i) \\ &- \sum_{i=1}^{M} g_i z_i + \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i (1 - \alpha_{ij}(z)) d_{ij}(g_j - z_i) \\ \geq &\sum_{i,j=1}^{M} z_i a_{ij} z_j + \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i a_{ij} g_j - \sum_{i=1}^{M} g_i z_i + \sum_{i=1}^{M} \sum_{j=M+1}^{N} z_i (1 - \alpha_{ij}(z)) d_{ij} g_j.\end{aligned}$$

The positive definiteness (5.2.15) yields

$$\sum_{i,j=1}^{M} z_i a_{ij} z_j \geq C_M \parallel \tilde{z} \parallel_2^2 .$$

The application of Hölder's and Young's inequality leads us to

$$\begin{aligned}(T\tilde{z},\tilde{z})_2 &= \sum_{i,j=1}^{M} z_i a_{ij} z_j + \sum_{i=1}^{M}\sum_{j=M+1}^{N} z_i a_{ij} g_j - \sum_{i=1}^{M} g_i z_i + \sum_{i=1}^{M}\sum_{j=M+1}^{N} z_i(1-\alpha_{ij}(z))d_{ij} g_j \\ &\geq C_M \parallel \tilde{z} \parallel_2^2 - C_1 \parallel \tilde{z} \parallel_2 \\ &\geq \frac{C_M}{2} \parallel \tilde{z} \parallel_2^2 - C_2\end{aligned}$$

where C_1, C_2 are positive constants, independent of $\tilde{z}$. Then, for any $\parallel \tilde{z} \parallel_2 > \sqrt{\frac{2C_2}{C_M}}$, one has $(T\tilde{z},\tilde{z})_2 > 0$. Through the application of Lemma 5.8 we obtain the existence of a vector $\tilde{z} \in \mathbb{R}^M$ such that $T\tilde{z} = 0$. □

Remark 5.12. *Due to the fact that the limiters are in general nonlinear, the uniqueness of a solution $z \in \mathbb{R}^N$ for (5.2.12) is not guaranteed. In [BJKR18, Section 3.3] or [Loh19, Section 3] one can see that under certain Lipschitz-continuity assumptions on the limiters the uniqueness of a discrete solution $z \in \mathbb{R}^N$ can be ensured.*

5.2.3 Discrete maximum principle

First, there are many publications concerned with conditions such that discrete maximum principles hold. For instance, in [BJK16, Section 5] the authors verify that under certain assumptions on the system matrix $\mathcal{A}$ the local DMP holds for a solution $z \in \mathbb{R}^N$ of (5.2.12). In [LohSP19, Lemma 4.12, Lemma 4.16] sufficient conditions on the system matrix have been derived in an abstract framework. In the following, we show a general result for the discrete maximum principle which is inspired by [Knob19, Theorem 1] and [BJK16, Corollary 11]. First, note that the coefficients of the solution-dependent matrix $\tilde{\mathcal{D}} = (\tilde{d}_{ij})_{i,j=1}^N$ which correct the artificial diffusion operator $\mathcal{D}$ are defined by

$$\tilde{d}_{ij} := \begin{cases} \alpha_{ij} d_{ij} & , \quad i \neq j \\ -\sum\limits_{l \neq i} \alpha_{il} d_{il} & , \quad i = j. \end{cases} \tag{5.2.17}$$

Using (5.2.17) the AFC system, i.e. (5.2.12) can be written as

$$\begin{aligned}(\tilde{\mathcal{A}}z)_i := [(\mathcal{A}+\mathcal{D}-\tilde{\mathcal{D}})z]_i &= g_i, \quad i = 1, \cdots, M \\ z_i &= g_i, \quad i = M+1, \cdots, N.\end{aligned} \tag{5.2.18}$$

Let us consider $\tilde{\mathcal{A}} = (\tilde{a}_{ij})_{i,j=1}^N$ where

$$\tilde{a}_{ij} := \begin{cases} a_{ij} + (1-\alpha_{ij}(z))d_{ij} & , \quad i \neq j \\ a_{ii} - \sum\limits_{l \neq i}(1-\alpha_{il}(z))d_{il} & , \quad i = j. \end{cases}$$

Recall that the compact sparsity pattern-property (see Remark 5.4) also holds for the matrix $\tilde{\mathcal{A}}$. Now, inspired by [Knob19, Theorem 1] we consider a sufficient condition on the system matrix $\tilde{\mathcal{A}}$ such that a discrete solution of (5.2.12) satisfies the discrete maximum principle (Definition 5.5).

Assumption 5.13. *Consider any $z = (z_1, \cdots, z_N) \in \mathbb{R}^N$ and any $i \in \{1, \cdots, M\}$. If z_i is a strict local extremum of z with respect to the vertex x_i, i.e.*

$$z_i > z_j \quad \forall\, j \in S_i \quad \text{or} \quad z_i < z_j \quad \forall\, j \in S_i,$$

then

$$\tilde{a}_{ij} = a_{ij} + (1-\alpha_{ij}(z))d_{ij} \leq 0 \quad \forall\, j \in S_i.$$

The satisfaction of the DMP referring to [Knob19, Theorem 1] is stated as follows.

Theorem 5.14. *Let $z = (z_1, \cdots, z_N) \in \mathbb{R}^N$ be a solution of* (5.2.12) *with limiters satisfying Assumption 5.13. Moreover, let condition* (5.2.7) *be satisfied. Consider any $i \in \{1, \cdots, M\}$. Then, we have*

$$g_i \leq 0 \Rightarrow z_i \leq \max_{j \in S_i} z_j^+ \tag{5.2.19}$$

$$g_i \geq 0 \Rightarrow z_i \geq \min_{j \in S_i} z_j^- . \tag{5.2.20}$$

Proof. Let $i \in \{1, \cdots, M\}$ and $g_i \leq 0$. Recall that $z \in \mathbb{R}^N$ solves

$$\sum_{j=1}^{N} \tilde{a}_{ij} z_j = g_i. \tag{5.2.21}$$

Let us assume that assertion (5.2.19) does not hold, i.e. we have

$$z_i > \max_{j \in S_i} z_j^+ =: \bar{C}.$$

By condition (5.2.7) and the definition of d_{ij}, we obtain

$$\sum_{j=1}^{N} \tilde{a}_{ij} = a_{ii} - \sum_{l \neq i} (1 - \alpha_{il}(z)) d_{il} + \sum_{l \neq i} (a_{il} + (1 - \alpha_{il}(z)) d_{il}) = \sum_{j=1}^{N} a_{ij} \geq 0. \tag{5.2.22}$$

On the other hand, we have also by condition (5.2.3)

$$\tilde{a}_{ii} = a_{ii} - \sum_{l \neq i} (1 - \alpha_{il}(z)) d_{il} \geq a_{ii} > 0. \tag{5.2.23}$$

Similar to Lemma 5.6, we are able to derive the following inequalities

$$\tilde{a}_{ii} z_i \leq \sum_{j \neq i} (-\tilde{a}_{ij}) z_j = \sum_{j \in S_i} (-\tilde{a}_{ij}) \underbrace{(z_j - \bar{C})}_{\leq 0} + \sum_{j \in S_i} (-\tilde{a}_{ij}) \bar{C} \leq \left(\sum_{j \in S_i} (-\tilde{a}_{ij}) \right) \bar{C} \leq \tilde{a}_{ii} \, \bar{C}$$

where we have used $\tilde{a}_{ij} \leq 0$, i.e. Assumption 5.13. The last inequality is a consequence of (5.2.22), i.e.

$$\sum_{j \in S_i} (-\tilde{a}_{ij}) \leq \sum_{j \neq i} (-\tilde{a}_{ij}) \leq \tilde{a}_{ii}.$$

The positivity of $\tilde{a}_{ii}$ yields the contradiction proving the assertion (5.2.19). The proof of (5.2.20) goes in the same way. □

5.2.4 Stabilization term

According to the mentioned AFC literature, the correction term in (5.2.12) is often denoted as stabilization term. For a general definition of the stabilization term, let us define the evaluation of $z \in C(\bar{\Omega})$ at nodes x_i, $i = 1, \cdots, N$ by $z_i := z(x_i)$. Then, for $z, v, w \in C(\bar{\Omega})$, we introduce the nonlinear form $d_h(\cdot; \cdot, \cdot)$ by

$$d_h(z; v, w) := \sum_{i,j=1}^{N} (1 - \alpha_{ij}(z)) d_{ij} (v(x_j) - v(x_i)) w(x_i) \tag{5.2.24}$$

where $\alpha_{ij} = \alpha_{ji} \in [0,1]$ for $i,j \in \{1,\cdots,N\}$, $i \neq j$ are solution-dependent correction factors. The diffusion matrix $\mathcal{D} = (d_{ij})_{i,j=1}^N$ is defined by (5.2.9). Now, let us derive several basic properties from $d_h(\cdot;\cdot,\cdot)$ which will be used on the following pages. First, for any $z \in C(\bar{\Omega})$ we obtain

$$\begin{aligned}
d_h(z;v,w) :&= \sum_{i,j=1}^N (1-\alpha_{ij}(z))d_{ij}(v(x_j)-v(x_i))w(x_i) \\
&= \sum_{j<i}^N (1-\alpha_{ij}(z))d_{ij}(v(x_j)-v(x_i))w(x_i) + \sum_{j>i}^N (1-\alpha_{ij}(z))d_{ij}(v(x_j)-v(x_i))w(x_i) \\
&= \sum_{j<i}^N (1-\alpha_{ij}(z))d_{ij}(v(x_j)-v(x_i))w(x_i) + \sum_{j<i}^N (1-\alpha_{ji}(z))d_{ji}(v(x_i)-v(x_j))w(x_j) \\
&= \sum_{j<i}^N (1-\alpha_{ij}(z))d_{ij}(v(x_j)-v(x_i))(w(x_i)-w(x_j))
\end{aligned}$$

where we have used the symmetry of the limiters $\alpha_{ij} = \alpha_{ji}$ and the symmetry of the diffusion coefficients $d_{ij} = d_{ji}$. Note that $d_h(z;v,w) = d_h(z;w,v)$ such that the mapping $d_h(z;\cdot,\cdot)$ is a symmetric bilinear form. Moreover, it holds $d_h(z;1,v) = 0 = d_h(z;w,1)$ and $d_h(z;w,w) \geq 0$ for all $z,w,v \in C(\bar{\Omega})$.

5.2.5 Modifications for Robin boundary conditions

In the case of a pure Robin boundary value problem (3.2.1), the system of linear equations has the form

$$\sum_{j=1}^N a_{ij} z_j = g_i, \quad i = 1,\cdots,N. \tag{5.2.25}$$

In contrast to the Dirichlet boundary value problem, the set $\{x_{M+1},\cdots,x_N\}$ coincides with the set of nodes where Robin boundary conditions are prescribed. The construction of an AFC scheme for (5.2.25) is similar to (5.2.1). Note that due to the lack of Dirichlet boundary conditions, we have to adjust the indices of condition (5.2.2), i.e. the matrix $(a_{ij})_{i,j=1}^N$ is postive definite in so far as:

$$\sum_{i,j=1}^N z_i a_{ij} z_j \geq C \sum_{i=1}^N z_i^2 > 0 \quad \forall\, (z_1,\cdots,z_N) \in \mathbb{R}^N \setminus \{0\}$$

where $C > 0$ is a constant. The AFC system corresponding to (5.2.25) is given by

$$\sum_{j=1}^N a_{ij} z_j + \sum_{j=1}^N (1-\alpha_{ij}(z))d_{ij}(z_j - z_i) = g_i, \quad i = 1,\cdots,N \tag{5.2.26}$$

where $\alpha_{ij} \in [0,1]$ for $i,j \in \{1,\cdots,N\}$, $i \neq j$ are the solution-dependent limiters. The proof of the existence of a discrete solution $z \in \mathbb{R}^N$ for (5.2.26) goes along the same lines as in Theorem 5.11. Moreover, the results corresponding to the local discrete maximum principle have to be slightly modified. For detailed information we refer to [LohSP19, Section 4.2.2].

5.3 Application on the state equations

The goal of this section is to apply the AFC method on the state equation corresponding to $(P_f), (P_b), (P_s^{MY})$ and (P_Γ). This section is generally based on the results in [BJKR18]. We start with the application of the AFC method on the elliptic equations (3.1.4) and (3.2.4), i.e. we consider for an arbitrary function $g \in L^2(\Omega)$

$$a(y, v) = (g, v)_\Omega \quad \forall v \in H_0^1(\Omega) \tag{5.3.1}$$

and for an arbitrary function $g \in H^{\frac{1}{2}}(\Gamma)$

$$a_\Gamma(y, v) = (g, v)_\Gamma \quad \forall v \in H^1(\Omega) \tag{5.3.2}$$

where $a(\cdot, \cdot)$ is defined by (3.1.2) and $a_\Gamma(\cdot, \cdot)$ by (3.2.2). Recall that (5.3.1) and (5.3.2) are the weak formulations associated with the state equations of $(P_f), (P_b), (P_s^{MY})$ resp. (P_Γ). In the following, the superscript s emphasizes that we consider the state equation. Moreover, the superscript Γ indicates that we consider the state equation of the Robin boundary control problem (P_Γ). The reasons for the superscripts will become apparent in the following sections.

5.3.1 Dirichlet boundary condition

Next, we derive an AFC scheme for (5.3.1). The Galerkin method associated with (5.3.1) is given by: Find $y_h \in V_{h,0}$ such that

$$a(y_h, v_h) = (g, v_h)_\Omega \quad \forall v_h \in V_{h,0}. \tag{5.3.3}$$

We introduce the matrix $\mathcal{A} = (a_{ij})_{i,j=1}^N$ where $a_{ij} := a(\varphi_j, \varphi_i)$ and according to (5.2.9), the corresponding artificial diffusion matrix $\mathcal{D} = (d_{ij})_{i,j=1}^N$. Let $y_h \in V_{h,0}$ be a solution of (5.3.3). Then, we have $y_h = \sum_{i=1}^N y_i \varphi_i$ where $y_i = y_h(x_i)$ for all $i \in \{1, \cdots, N\}$. Consequently, the following system of linear equations is satisfied

$$\begin{aligned} \sum_{j=1}^N a_{ij} y_j = g_i, &\quad i = 1, \cdots, M \\ y_i = 0, &\quad i = M+1, \cdots, N \end{aligned} \tag{5.3.4}$$

with $g_i = (g, \varphi_i)_\Omega$ for $i = 1, \cdots, M$. According to (5.2.12), the AFC system of (5.3.4) is given by

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (g, v_h)_\Omega \quad \forall v_h \in V_{h,0} \tag{5.3.5}$$

where $d_h^s(\cdot; \cdot, \cdot)$ is defined by (5.2.24). Following the lines in Theorem 5.11 or [BJKR18, Theorem 1], we can state the next result.

Theorem 5.15. *Let Assumption 5.9 on the flux limiters α_{ij} be satisfied. Then, the equation* (5.3.5) *admits a solution $y_h \in V_{h,0}$.*

5.3.1.1 Discrete maximum principle

According to Section 5.2.3, we will verify that under the Assumption 5.13 a discrete solution $y_h \in V_{h,0}$ of (5.3.5) satisfies the local DMP, defined on patches Δ_i, $i = 1, \cdots, M$. The following proof can be found for instance in [BJKR18, Theorem 2].

Theorem 5.16 (Local DMP). *Let $y_h \in V_{h,0}$ be a solution of* (5.3.5) *with α_{ij} satisfying Assumption 5.13. Consider any $i \in \{1, \cdots .M\}$. Then*

$$g \leq 0 \text{ in } \Delta_i \;\Rightarrow\; \max_{\Delta_i} y_h \leq \max_{\partial\Delta_i} y_h^+ \tag{5.3.6}$$

$$g \geq 0 \text{ in } \Delta_i \;\Rightarrow\; \min_{\Delta_i} y_h \geq \min_{\partial\Delta_i} y_h^-. \tag{5.3.7}$$

Proof. We will only verify claim (5.3.6). The proof of claim (5.3.7) goes in the same way. Similar to Theorem 5.14, we start by proving the fulfillment of the conditions (5.2.22) and (5.2.23). For a solution $y_h = \sum_{i=1}^{N} y_j\varphi_j$ of (5.3.5) with $y_i = y_h(x_i)$, $i = 1, \cdots, N$, one has

$$\sum_{j=1}^{N} \tilde{a}_{ij} y_j = g_i, \quad i = 1, \cdots, M \tag{5.3.8}$$

where

$$\begin{aligned} \tilde{a}_{ij} &= a(\varphi_j, \varphi_i) + d_h^s(y_h; \varphi_j, \varphi_i), \quad i = 1, \cdots, M,\ j = 1, \cdots, N, \\ g_i &= (g, \varphi_i)_{\Delta_i}, \quad i = 1, \cdots, M. \end{aligned}$$

For $i = 1, \cdots, M$ we have

$$\begin{aligned} \tilde{a}_{ii} &\geq a(\varphi_i, \varphi_i) \geq \varepsilon |\varphi_i|_{1,\Omega}^2 > 0, \\ \sum_{j=1}^{N} \tilde{a}_{ij} &= a(1, \varphi_i) + d_h^s(y_h; 1, \varphi_i) = (c, \varphi_i)_{\Delta_i} > 0 \end{aligned} \tag{5.3.9}$$

so that the conditions (5.2.22) and (5.2.23) hold. We remark that these properties follow from the facts that $(\mathbf{b} \cdot \nabla v, v)_\Omega = 0$ for any $v \in H_0^1(\Omega)$, $d_h^s(y_h; \varphi_i, \varphi_i) \geq 0$ for $i = 1, \cdots, M$ and $\sum_{j=1}^{N} \varphi_j = 1$ in Ω. Now, for any $i \in \{1, \cdots, M\}$, let $g \leq 0$ in Δ_i so that $g_i \leq 0$. Let us assume that (5.3.6) does not hold, i.e. we have

$$\max_{\Delta_i} y_h > \max_{\partial\Delta_i} y_h^+ =: \bar{C}. \tag{5.3.10}$$

Then, it follows $\max_{\Delta_i} y_h > \max_{\partial\Delta_i} y_h$ and $y_i = y_h(x_i) > 0$ is a strict local maximum of $(y_1, \cdots, y_N)$ in S_i, i.e. we have

$$y_i > y_j \quad \forall j \in S_i.$$

Since $\tilde{a}_{ij} = 0$ for any $j \notin S_i \cup \{i\}$, we obtain by virtue of (5.3.8)

$$\tilde{a}_{ii} y_i = g_i + \sum_{j \in S_i} (-\tilde{a}_{ij}) y_j \leq \sum_{j \in S_i} (-\tilde{a}_{ij})(y_j - \bar{C}) + \sum_{j \in S_i} (-\tilde{a}_{ij}) \bar{C} \leq \sum_{j \in S_i} (-\tilde{a}_{ij}) \bar{C} \leq \tilde{a}_{ii} \bar{C} \tag{5.3.11}$$

where we have used $\tilde{a}_{ij} \leq 0$ for all $j \in S_i$ (Assumption 5.13) and

$$\tilde{a}_{ii} \geq \sum_{j \in S_i} (-\tilde{a}_{ij}).$$

The positivity of $\tilde{a}_{ii} > 0$ leads us to

$$y_i \leq \bar{C}$$

which is a contradiction to (5.3.10), proving the assertion. □

Remark 5.17 (Global DMP). *In [BJKR18, Theorem 3] the authors prove that under Assumption 5.9 and Assumption 5.13 the following global discrete maximum principle holds:*

$$g \leq 0 \text{ in } \Omega \;\Rightarrow\; \max_{\bar{\Omega}} y_h \leq \max_{\partial\Omega} y_h^+$$

$$g \geq 0 \text{ in } \Omega \;\Rightarrow\; \min_{\bar{\Omega}} y_h \geq \min_{\partial\Omega} y_h^-.$$

5.3.1.2 Error analysis

In this section, we derive an error estimate for $y - y_h$ where $y \in H^2(\Omega) \cap H_0^1(\Omega)$ is the unique solution of (5.3.1) and $y_h \in V_{h,0}$ is a solution of (5.3.5). For this, we define the mesh-dependent norm

$$\| v \|_h^s := \left(\varepsilon |v|_{1,\Omega}^2 + c_0 \| v \|_{0,\Omega}^2 + d_h^s(y_h; v, v)\right)^{\frac{1}{2}}.$$

The following theorem provides an error estimate for $\| y - y_h \|_h^s$. A proof can be found in [BJKR18, Theorem 4]. However, due to the fact that this result will be often used in this work, we illustrate the proof.

Theorem 5.18. *Let $y \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of* (5.3.1) *and let $y_h \in V_{h,0}$ be a solution of* (5.3.5)*. Then, there exists a constant $C > 0$, independent of h and the data of* (5.3.1) *such that*

$$\| y - y_h \|_h^s \leq C \left(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|y|_{2,\Omega} + d_h^s(y_h; I_h y, I_h y)^{\frac{1}{2}}.$$

Proof. We define

$$y - y_h = (y - I_h y) + (I_h y - y_h) =: w_h + e_h.$$

Note that $d_h^s(y_h; w_h, w_h) = 0$. Hence, the standard FE-estimate (Lemma 5.1) yields

$$\| w_h \|_h^s \leq C(\varepsilon + \| c \|_{0,\infty,\Omega} h) h|y|_{2,\Omega}.$$

Furthermore, we have

$$\begin{aligned}
(\| e_h \|_h^s)^2 &\leq a(e_h, e_h) + d_h^s(y_h; e_h, e_h) \\
&= a(I_h y, e_h) + d_h^s(y_h; I_h y, e_h) - \overbrace{(a(y_h, e_h) + d_h^s(y_h; y_h, e_h))}^{=(g,e_h)_\Omega} \\
&= a(I_h y, e_h) + d_h^s(y_h; I_h y, e_h) - a(y, e_h) \\
&= \underbrace{a(I_h y - y, e_h)}_{(1)} + \underbrace{d_h^s(y_h; I_h y, e_h)}_{(2)}.
\end{aligned}$$

The first term (1) can be estimated by the standard FE-estimate so that we obtain

$$a(I_h y - y, e_h) \leq C \left(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|y|_{2,\Omega} \| e_h \|_h^s .$$

Due to the fact that $d_h^s(y_h; \cdot, \cdot)$ is a symmetric positive semidefinite bilinear form, the Cauchy-Schwarz inequality yields

$$\begin{aligned}
d_h^s(y_h; I_h y, e_h) &\leq d_h^s(y_h; I_h y, I_h y)^{\frac{1}{2}} d_h^s(y_h; e_h, e_h)^{\frac{1}{2}} \\
&\leq d_h^s(y_h; I_h y, I_h y)^{\frac{1}{2}} \| e_h \|_h^s .
\end{aligned}$$

Consequently, we obtain the desired result. □

In the AFC literature $d_h^s(y_h; I_h y, I_h y)$ is referred to as consistency error. The following lemma yields a corresponding estimate which can also be found in [BJKR18, Lemma 2].

Lemma 5.19. *For all $z_h \in V_h, z \in C(\bar{\Omega})$ we have*

$$d_h^s(z_h; I_h z, I_h z) \leq C \left(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} h + \| c \|_{0,\infty,\Omega} h^2\right) |I_h z|_{1,\Omega}^2.$$

Proof. Note that

$$
\begin{aligned}
|a_{ij}| &= \Big| \int_\Omega \varepsilon \nabla\varphi_i \cdot \nabla\varphi_j + (\mathbf{b}\cdot\nabla\varphi_j)\,\varphi_i + c\varphi_i\varphi_j \; dx \Big| \\
&\le \varepsilon |\varphi_i|_{1,T}|\varphi_j|_{1,T} + \| \mathbf{b} \|_{0,\infty,\Omega} \, |\varphi_i|_{1,T} \, \| \varphi_j \|_{0,T} + \| c \|_{0,\infty,\Omega} \| \varphi_i \|_{0,T} \| \varphi_j \|_{0,T} \, .
\end{aligned}
$$

Hence, the coefficients of the diffusion matrix $\mathcal{D}$ can be estimated by

$$
\begin{aligned}
|d_{ij}| \le \sum_{\substack{T\in\mathcal{T}_h \\ x_i,x_j\in T}} \Big(& \varepsilon |\varphi_i|_{1,T}|\varphi_j|_{1,T} + \| c \|_{0,\infty,T} \| \varphi_i \|_{0,T} \| \varphi_j \|_{0,T} \\
& + \| \mathbf{b} \|_{0,\infty,T} \left\{ |\varphi_i|_{1,T} \, \| \varphi_j \|_{0,T} + |\varphi_j|_{1,T} \, \| \varphi_i \|_{0,T} \right\} \Big) \\
& \le C \sum_{\substack{T\in\mathcal{T}_h \\ x_i,x_j\in T}} \left(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} \, h_T + \| c \|_{0,\infty,\Omega} \, h_T^2 \right) \\
& \le \tilde{C} \left(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} \, h + \| c \|_{0,\infty,\Omega} \, h^2 \right) .
\end{aligned}
$$

Then, using Lemma 5.10 and $(1-\alpha_{ij}(z_h)) \le 1$, we obtain

$$
\begin{aligned}
d_h^s(z_h; I_h z, I_h z) &= \sum_{\substack{i,j=1 \\ i<j}}^{N} (1-\alpha_{ij}(z_h)) |d_{ij}| |z(x_i) - z(x_j)|^2 \\
&\le \sum_{\substack{T\in\mathcal{T}_h \\ x_i,x_j\in T}} |d_{ij}| |z(x_i) - z(x_j)|^2 \\
&\le \tilde{C} \left(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} \, h + \| c \|_{0,\infty,\Omega} \, h^2 \right) \sum_{\substack{T\in\mathcal{T}_h \\ x_i,x_j\in T}} |z(x_i) - z(x_j)|^2 .
\end{aligned}
$$

Since

$$
\sum_{x_i,x_j\in T} |z(x_i) - z(x_j)|^2 \le C |I_h z|_{1,\Omega}^2
$$

we get the desired result. □

As we have mentioned in Section 1.1, oscillations may occur in the Galerkin solutions when the convection dominates the diffusion, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega}$. On the one hand, an appropriate stabilization method should compute solutions free of spurious oscillations. On the other hand, the discrete solutions should also converge to the continuous solution for $h \to 0$. The combination of Theorem 5.18 and Lemma 5.19 yields that the solutions computed by the AFC method converge with an order of $\mathcal{O}(\frac{1}{2})$ to the continuous solution of (5.3.1) when the diffusion coefficient is sufficiently small.

Corollary 5.20. *In the convection-dominated case, i.e.* $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} \, h$ *we obtain*

$$
\| y - y_h \|_h^s \le C h^{\frac{1}{2}} \, \| y \|_{2,\Omega} \, .
$$

5.3.2 Robin boundary condition

Now, we derive an AFC scheme for (5.3.2). It follows the same lines as in Section 5.3.1. We start again with the Galerkin formulation of (5.3.2) which is given by: Find $y_h \in V_h$ such that

$$
a_\Gamma(y_h, v_h) = (g, v_h)_\Gamma \quad \forall v_h \in V_h. \tag{5.3.12}
$$

We introduce the matrix $\mathcal{A} = (a_{ij})_{i,j=1}^N$ where $a_{ij} := a_\Gamma(\varphi_j, \varphi_i)$ and the corresponding artificial diffusion matrix $\mathcal{D} = (d_{ij})_{i,j=1}^N$. Let $y_h \in V_h$ be a solution of (5.3.12). Then, we have $y_h = \sum_{i=1}^N y_i\varphi_i$ with $y_i = y_h(x_i)$ for all $i \in \{1, \cdots, N\}$. Hence, the following system of linear equations is satisfied

$$\sum_{j=1}^N a_{ij}y_j = g_i, \quad i = 1, \cdots, N \tag{5.3.13}$$

with

$$g_i = \begin{cases} 0 & , \quad i \in \{1, \cdots, M\} \\ (g, \varphi_i)_\Gamma & , \quad i \in \{M+1, \cdots, N\}. \end{cases}$$

According to (5.2.12), the AFC system associated with (5.3.12) is given by

$$a_\Gamma(y_h, v_h) + d_h^{s,\Gamma}(y_h; y_h, v_h) = (g, v_h)_\Gamma \quad \forall v_h \in V_h \tag{5.3.14}$$

where $d_h^{s,\Gamma}(\cdot;\cdot,\cdot)$ is defined by (5.2.24). Similar to Theorem 5.15, we can state the next result.

Theorem 5.21. *Let the Assumption 5.9 on the flux limiters α_{ij} be satisfied. Then, the equation* (5.3.14) *admits a solution $y_h \in V_h$.*

5.3.2.1 Discrete maximum principle

The next result shows the validity of the local DMP in the case of the pure Robin boundary value problem. Note that the local DMP only holds on patches Δ_i with $i = 1, \cdots, M$, i.e. on patches for the inner nodes of the triangulation. We will discuss this point in Remark 5.23.

Theorem 5.22 (Local DMP)**.** *Let $y_h \in V_h$ be a solution of* (5.3.14) *with α_{ij} satisfying Assumption 5.13. Consider any $i \in \{1, \cdots, M\}$. Then*

$$\max_{\Delta_i} y_h \le \max_{\partial\Delta_i} y_h^+ \tag{5.3.15}$$

$$\min_{\Delta_i} y_h \ge \min_{\partial\Delta_i} y_h^-. \tag{5.3.16}$$

Proof. We only prove claim (5.3.15). First, similar to Theorem 5.16, we prove sufficient conditions for the DMP. For this, let $y_h \in V_h$ be a solution of (5.3.14) and let us denote $y_i = y_h(x_i)$, $i = 1, \cdots, N$. Then, $y_h = \sum_{i=1}^N y_j\varphi_j$ and one has

$$\sum_{j=1}^N \tilde{a}_{ij}y_j = g_i, \quad i = 1, \cdots, N \tag{5.3.17}$$

where

$$\tilde{a}_{ij} = a_\Gamma(\varphi_j, \varphi_i) + d_h^{s,\Gamma}(y_h; \varphi_j, \varphi_i), \quad i = 1, \cdots, M, \ j = 1, \cdots, N$$

and

$$g_i = \begin{cases} 0 & , \quad i \in \{1, \cdots, M\} \\ (g, \varphi_i)_\Gamma & , \quad i \in \{M+1, \cdots, N\}. \end{cases}$$

For nodes $i \in \{1, \cdots, M\}$ we have

$$\int_\Gamma \frac{\mathbf{b} \cdot n \cdot \varphi_i}{2} \, ds = 0.$$

Hence, we can derive for $i \in \{1, \cdots, M\}$

$$\tilde{a}_{ii} \geq a_\Gamma(\varphi_i, \varphi_i) \geq \varepsilon |\varphi_i|_{1,\Omega}^2 > 0$$

and

$$\sum_{j=1}^N \tilde{a}_{ij} = a_\Gamma(1, \varphi_i) + d_h^{s,\Gamma}(y_h; 1, \varphi_i) = (c, \varphi_i)_{\Delta_i} - \int_\Gamma \frac{\mathbf{b} \cdot n \cdot \varphi_i}{2} \, ds = (c, \varphi_i)_{\Delta_i} > 0$$

such that the sufficient conditions for the DMP are fulfilled. Note that these properties are again results of the facts that $d_h^{s,\Gamma}(y_h; \varphi_i, \varphi_i) \geq 0$ for $i = 1, \cdots, N$ and $\sum_{j=1}^N \varphi_j = 1$ in Ω. Now, for $i \in \{1, \cdots, M\}$ the equation (5.3.17) reduces to

$$\sum_{j=1}^N \tilde{a}_{ij} y_j = 0. \tag{5.3.18}$$

Hence, for the verification of (5.3.15) we can follow the same lines of Theorem 5.16 where the local DMP poperty has been verified for a Dirichlet boundary value problem. □

Remark 5.23. *For indices corresponding to the Robin boundary nodes, i.e. x_i with $i = M+1, \cdots, N$ we do not know whether the boundary integral*

$$\int_\Gamma \frac{\mathbf{b} \cdot n \cdot \varphi_i}{2} \, ds$$

is positive or not. Since the non-negativity of the row sums (see Lemma 5.6 or [BJK16, Corollary 11]*) is necessary for verifying the local DMP the assertions* (5.3.15) *resp.* (5.3.16) *can only be verified for the set of inner nodes $\{1, \cdots, M\}$.*

Remark 5.24 (Global DMP)**.** *Similar to Remark 5.17, a solution $y_h \in V_h$ of* (5.3.14) *satisfies the following global discrete maximum principle when Assumption 5.9 and Assumption 5.13 hold:*

$$\max_{\bar{\Omega}} y_h \leq \max_{\partial\Omega} y_h^+$$
$$\min_{\bar{\Omega}} y_h \geq \min_{\partial\Omega} y_h^-.$$

5.3.2.2 Error analysis

In this section, we derive an error estimate for $y - y_h$ where $y \in H^2(\Omega)$ is the unique solution of (5.3.2) and $y_h \in V_h$ is a solution of (5.3.14). The mesh-dependent norm is defined by

$$\| v \|_h^{s,\Gamma} := \left(\varepsilon |v|_{1,\Omega}^2 + c_0 \| v \|_{0,\Omega}^2 + d_h^{s,\Gamma}(y_h; v, v) \right)^{\frac{1}{2}}.$$

The next result is inspired by Theorem 5.18 resp. [BJKR18, Theorem 4].

Theorem 5.25. *Let $y \in H^2(\Omega)$ be the solution of* (5.3.2) *and let $y_h \in V_h$ be a solution of* (5.3.14)*. Then, there exists a constant $C > 0$, independent of h and the data of* (5.3.2) *such that*

$$\| y - y_h \|_h^{s,\Gamma} \leq C \left(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|y|_{2,\Omega} + Ch|y|_{2,\Omega} + d_h^{s,\Gamma}(y_h; I_h y, I_h y)^{\frac{1}{2}}.$$

Proof. The first lines of the proof are the same as in Theorem 5.18. We define

$$y - y_h = (y - I_h y) + (I_h y - y_h) =: w_h + e_h.$$

The standard FE-estimate yields

$$\| w_h \|_h^{s,\Gamma} \leq C(\varepsilon + \| c \|_{0,\infty,\Omega} h) h|y|_{2,\Omega}. \tag{5.3.19}$$

Moreover, we are able to derive

$$(\| e_h \|_h^{s,\Gamma})^2 \leq a_\Gamma(I_h y - y, e_h) + d_h^{s,\Gamma}(y_h; I_h y, e_h). \tag{5.3.20}$$

The first term on the right hand side of (5.3.20) can be written as

$$\begin{aligned} a_\Gamma(I_h y - y, e_h) = & \underbrace{\varepsilon\,(\nabla(I_h y - y), \nabla e_h)_\Omega + (\mathbf{b} \cdot \nabla(I_h y - y), e_h)_\Omega + (c(I_h y - y), e_h)_\Omega}_{(1)} \\ & - \underbrace{\int_\Gamma \frac{\mathbf{b} \cdot n \cdot (I_h y - y) e_h}{2}\, ds}_{(2)}. \end{aligned} \tag{5.3.21}$$

Part (1) can be estimated by

$$(1) \leq C \left(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|y|_{2,\Omega} \| e_h \|_h^{s,\Gamma}.$$

The Hölder inequality yields for the second part (2)

$$\int_\Gamma \frac{\mathbf{b} \cdot n \cdot (I_h y - y) e_h}{2}\, ds \leq C \| \mathbf{b} \|_{0,\infty,\Gamma} \| I_h y - y \|_{0,\Gamma} \| e_h \|_{0,\Gamma}.$$

The application of the trace inequality (Theorem 2.18) and the interpolation error estimate (Lemma 5.1) result in

$$\begin{aligned} \| I_h y - y \|_{0,\Gamma} \leq & \| I_h y - y \|_{0,\Omega}^{1/2} \| I_h y - y \|_{1,\Omega}^{1/2} \\ \leq & C \left(h|y|_{2,\Omega}^{1/2}\right) \left(h^{1/2}|y|_{2,\Omega}^{1/2}\right) \\ \leq & Ch^{3/2}|y|_{2,\Omega} \end{aligned}$$

and

$$\| e_h \|_{0,\Gamma} \leq \| e_h \|_{0,\Omega}^{1/2} \| e_h \|_{1,\Omega}^{1/2}. \tag{5.3.22}$$

The inverse inequality (Lemma 5.2) yields

$$\| e_h \|_{1,\Omega}^{1/2} \leq Ch^{-1/2} \| e_h \|_{0,\Omega}^{1/2}$$

such that we get for (5.3.22)

$$\begin{aligned} \| e_h \|_{0,\Gamma} & \leq Ch^{-1/2} \| e_h \|_{0,\Omega} \\ & \leq Ch^{-1/2} \| e_h \|_h^{s,\Gamma}. \end{aligned} \tag{5.3.23}$$

Note that in the last inequality of (5.3.23) we have used

$$c_0 \parallel e_h \parallel_{0,\Omega}^2 \leq (\parallel e_h \parallel_h^{s,\Gamma})^2.$$

Consequently, term (2) can be estimated by

$$\int_\Gamma \frac{\mathbf{b} \cdot n \cdot (I_h y - y) e_h}{2} \, ds \leq Ch|y|_{2,\Omega} \parallel e_h \parallel_h^{s,\Gamma} .$$

Altogether, the estimates of (1) and (2) imply for (5.3.21)

$$a_\Gamma(I_h y - y, e_h) \leq C \left(\varepsilon + c_0^{-1}\{\parallel \mathbf{b} \parallel_{0,\infty,\Omega}^2 + \parallel c \parallel_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|y|_{2,\Omega} \parallel e_h \parallel_h^{s,\Gamma} + Ch|y|_{2,\Omega} \parallel e_h \parallel_h^{s,\Gamma} .$$

Due to the fact that $d_h^{s,\Gamma}(y_h; \cdot, \cdot)$ is a symmetric, positive semidefinite bilinear form, the Cauchy-Schwarz inequality yields

$$d_h^{s,\Gamma}(y_h; I_h y, e_h) \leq d_h^{s,\Gamma}(y_h; I_h y, I_h y)^{\frac{1}{2}} \parallel e_h \parallel_h^{s,\Gamma} .$$

Collecting the estimates of (5.3.19) and (5.3.20) leads us to the desired result. □

The following lemma is similar to Lemma 5.19 and yields an estimate for the consistency error $d_h^{s,\Gamma}(y_h; I_h y, I_h y)$. A similar proof can be found for instance in [LohSP19, Theorem 4.72].

Lemma 5.26. *For all $z_h \in V_h, z \in C(\bar{\Omega})$ we have*

$$d_h^{s,\Gamma}(z_h; I_h z, I_h z) \leq C(\varepsilon + \parallel \mathbf{b} \parallel_{0,\infty,\Omega} h + \parallel c \parallel_{0,\infty,\Omega} h^2)|I_h z|_{1,\Omega}^2.$$

Proof. Note that

$$\begin{aligned}
|a_{ij}^\Gamma| &= \Big| \int_\Omega \varepsilon \nabla\varphi_i \cdot \nabla\varphi_j + (\mathbf{b} \cdot \nabla\varphi_j)\, \varphi_i + c\varphi_i\varphi_j \, dx - \int_\Gamma \frac{\mathbf{b} \cdot n \cdot \varphi_i\varphi_j}{2} \, ds \Big| \\
&\leq \varepsilon|\varphi_i|_{1,T}|\varphi_j|_{1,T} + \parallel \mathbf{b} \parallel_{0,\infty,\Omega} |\varphi_i|_{1,T} \parallel \varphi_j \parallel_{0,T} + \parallel c \parallel_{0,\infty,\Omega} \parallel \varphi_i \parallel_{0,T} \parallel \varphi_j \parallel_{0,T} \\
&+ \parallel \mathbf{b} \parallel_{0,\infty,\Gamma} \parallel \varphi_i \parallel_{0,\Delta_i\cap\Delta_j\cap\Gamma} \parallel \varphi_j \parallel_{0,\Delta_i\cap\Delta_j\cap\Gamma} .
\end{aligned}$$

Now, we estimate the diffusion coefficient d_{ij}^Γ with the help of the trace inequality (Theorem 2.18).

$$\begin{aligned}
|d_{ij}^\Gamma| \leq \sum_{T\in\mathcal{T}_h, x_i, x_j \in T} \Big(&\varepsilon|\varphi_i|_{1,T}|\varphi_j|_{1,T} + \parallel c \parallel_{0,\infty,\Omega} \parallel \varphi_i \parallel_{0,T} \parallel \varphi_j \parallel_{0,T} \\
&+ \parallel \mathbf{b} \parallel_{0,\infty,\Omega} \{|\varphi_i|_{1,T} \parallel \varphi_j \parallel_{0,T} + |\varphi_j|_{1,T} \parallel \varphi_i \parallel_{0,T}\} \\
&+ C_\tau \parallel \mathbf{b} \parallel_{0,\infty,\Omega} \{\parallel \varphi_i \parallel_{0,\Delta_i\cap\Delta_j\cap\Gamma} \parallel \varphi_j \parallel_{0,\Delta_i\cap\Delta_j\cap\Gamma}\}\Big) \\
&\leq C\left(\varepsilon + \parallel \mathbf{b} \parallel_{0,\infty,\Omega} h + \parallel c \parallel_{0,\infty,\Omega} h^2\right).
\end{aligned}$$

The rest of the proof goes along the same lines as in Lemma 5.19. □

The combinition of Theorem 5.25 and Lemma 5.26 leads us to the next result.

Corollary 5.27. *In the convection-dominated case, i.e. $\varepsilon \ll \parallel \mathbf{b} \parallel_{0,\infty,\Omega} h$ we obtain*

$$\parallel y - y_h \parallel_h^{s,\Gamma} \leq Ch^{\frac{1}{2}} \parallel y \parallel_{2,\Omega} .$$

5.4 Application on the adjoint equations

The application of the AFC method to the adjoint equations corresponding to $(P_f), (P_b), (P_s^{MY})$ and (P_Γ) goes in the same way as in Section 5.3. Now, we consider the adjoint equation of $(P_f), (P_b), (P_s^{MY})$ which has for an arbitary right hand side $g \in L^2(\Omega)$ the form

$$a(\psi, p) = (g, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \tag{5.4.1}$$

and the adjoint equation of (P_Γ), i.e.

$$a_\Gamma(\psi, p) = (g, \psi)_\Omega \quad \forall \psi \in H^1(\Omega). \tag{5.4.2}$$

In the following, the superscript *ad* emphasizes that we should consider the adjoint equation. Moreover, the superscript Γ indicates the necessity to consider the adjoint equation of the Robin boundary control problem (P_Γ). The Galerkin formulations of (5.4.1) resp. (5.4.2) are given by

$$a(\psi_h, p_h) = (g, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \tag{5.4.3}$$

and

$$a_\Gamma(\psi_h, p_h) = (g, \psi_h)_\Omega \quad \forall \psi_h \in V_h. \tag{5.4.4}$$

The entries of the stiffness matrix corresponding to (5.4.3) are given by $a_{ij} = a(\varphi_j, \varphi_i)$ resp. the entries of the stiffness matrix corresponding to (5.4.4) are given by $a_{ij} = a_\Gamma(\varphi_j, \varphi_i)$. The artificial diffusion matrices are defined by (5.2.9).

Remark 5.28 (Existence of discrete solutions / Discrete maximum principle)**.** *According to* (5.4.3) *resp.* (5.4.4)*, the corresponding variational formulations of the stabilized equations are similar to* (5.3.5) *resp.* (5.3.14)*. Hence, regarding Theorem 5.15, the claim of the existence of a discrete stabilized solution holds when the continuity assumption on the flux limiter (Assumption 5.9) is satisfied. For the proof, we refer again to Theorem 5.11. Moreover, the fulfillment of the local DMP property by a discrete stabilized solution p_h can be verified in the same way as in Theorem 5.16 (Dirichlet boundary condition) resp. Theorem 5.22 (Robin boundary condition) when Assumption 5.13 is satisfied. Global discrete maximum principles (see Remark 5.17 and Remark 5.24) are also valid when Assumption 5.9 and Assumption 5.13 hold. Due to the fact that Section 5.3 can be transferred analogously to the adjoint equations* (5.4.3) *and* (5.4.4)*, we keep the following sections briefly.*

5.4.1 Dirichlet boundary condition

The variational formulation of the AFC stabilization corresponding to the adjoint equation (5.4.1) is given by: Find $p_h \in V_{h,0}$ such that

$$a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (g, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \tag{5.4.5}$$

where $d_h^{ad}(\cdot; \cdot, \cdot)$ is defined by (5.2.24).

Theorem 5.29. *Let Assumption 5.9 on the flux limiters α_{ij} be satisfied. Then, the equation* (5.4.5) *admits a solution $p_h \in V_{h,0}$.*

Theorem 5.30. *Let $p \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of* (5.4.1) *and let $p_h \in V_{h,0}$ be a solution of* (5.4.5)*. Then, there exists a constant $C > 0$, independent of h and the data of* (5.4.1) *such that*

$$\| p - p_h \|_h^{ad} \leq C \left(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|p|_{2,\Omega} + d_h^{ad}(p_h; I_h p, I_h p)$$

where the mesh-dependent norm $\| \cdot \|_h^{ad}$ is defined by

$$\| \psi \|_h^{ad} := \left(\varepsilon|\psi|_{1,\Omega}^2 + c_0 \| \psi \|_{0,\Omega}^2 + d_h^{ad}(p_h; \psi, \psi)\right)^{\frac{1}{2}}.$$

Due to the fact that the entries of the diffusion matrix $\mathcal{D}$ are symmetric, i.e. $d_{ij} = d_{ji}$ the consistency error estimate for the state equation (see Lemma 5.19) also hold for the consistency error $d_h^{ad}(p_h; I_h p, I_h p)$.

Lemma 5.31. *For all $z_h \in V_h, z \in C(\bar{\Omega})$ we have*

$$d_h^{ad}(z_h; I_h z, I_h z) \leq C \left(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} h + \| c \|_{0,\infty,\Omega} h^2\right) |I_h z|_{1,\Omega}^2.$$

Corollary 5.32. *In the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$ we obtain*

$$\| p - p_h \|_h^{ad} \leq C h^{\frac{1}{2}} \| p \|_{2,\Omega} .$$

5.4.2 Robin boundary condition

For the adjoint equation (5.4.2), the stabilization method is given by: Find $p_h \in V_h$ such that

$$a_\Gamma(\psi_h, p_h) + d_h^{ad,\Gamma}(p_h; p_h, \psi_h) = (g, \psi_h)_\Omega \quad \forall \psi_h \in V_h \tag{5.4.6}$$

where $d_h^{ad,\Gamma}(\cdot; \cdot, \cdot)$ is defined by (5.2.24).

Theorem 5.33. *Let Assumption 5.9 on the flux limiters α_{ij} be satisfied. Then, the equation* (5.4.6) *admits a solution $p_h \in V_h$.*

Theorem 5.34. *Let $p \in H^2(\Omega)$ be the solution of* (5.4.2) *and let $p_h \in V_h$ be a solution of* (5.4.6)*. Then, there exists a constant $C > 0$, independent of h and the data of* (5.4.6) *such that*

$$\| p - p_h \|_h^{ad,\Gamma} \leq C \left(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\}\right)^{\frac{1}{2}} h|p|_{2,\Omega} + d_h^{ad,\Gamma}(p_h; I_h p, I_h p)^{\frac{1}{2}}$$

where the mesh-dependent norm is defined by

$$\| \psi \|_h^{ad,\Gamma} := \left(\varepsilon|\psi|_{1,\Omega}^2 + c_0 \| \psi \|_{0,\Omega}^2 + d_h^{ad,\Gamma}(p_h; \psi, \psi)\right)^{\frac{1}{2}}.$$

For the same reasons as in Lemma 5.31 the proof of Lemma 5.26 can be applied to verify the following consistency error estimate.

Lemma 5.35. *For all $z_h \in V_h$ and $z \in C(\bar{\Omega})$ we have*

$$d_h^{ad,\Gamma}(z_h; I_h z, I_h z) \leq C \left(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} h + \| c \|_{0,\infty,\Omega} h^2\right) |I_h z|_{1,\Omega}^2.$$

Corollary 5.36. *In the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$ we obtain*

$$\| p - p_h \|_h^{ad,\Gamma} \leq C h^{\frac{1}{2}} \| p \|_{2,\Omega} .$$

5.5 AFC limiters

In this section, we introduce the construction of the Kuzmin limiter and the BJK limiter. The construction corresponds to a linear boundary value problem with Dirichlet boundary condition. However, by adjusting the indices the following definitions and results can be transferred to a linear boundary value problem with Robin boundary condition (see Section 5.2.5). For detailed information on the general design principle of the limiters we refer to [BJK16, Lemma 6]. We remark that for reasons of a better convergence behavior and implementation, many modified variations of the originally Kuzmin limiter resp. BJK limiter have been developed. For instance,

in [BadBon17] differentiable resp. regularized limiter have been constructed such that Newton-like solution strategies are applicable. In [Knob21] the author modifies the Kuzmin limiter such that the DMP is guaranteed on arbitrary meshes. In [Kuz18] a gradient based limiter is derived such that a higher order of convergence is obtained. Further information on other limiters and their modifications can be found for instance in [Kuz10], [BJKR18], [Loh19], [LohSP19] and [JhaTh20]. In the following, we show the design for both types of limiters (Kuzmin/BJK) and verify that Assumption 5.9 and Assumption 5.13 are satisfied such that the local resp. the global DMP property holds. Furthermore, we discuss the meaning of the linearity-preserving property of limiters. In this context, we mention the numerical results in [BBK17] and [BJK17] which suggest a better convergence behavior when the limiter possesses the linearity-preserving property (see also [Kuz12, Section 7]).

Assumption 5.37. *The limiters α_{ij} possess the linearity-preserving property, if*

$$\alpha_{ij}(z) = 1 \quad \text{for } z \in \mathbb{P}_1(\mathbb{R}^2), \quad i = 1, \cdots, M, \; j = 1, \cdots, N.$$

The construction of the following limiters is generally based on [BJK16], [BJK17], [BJKR18], [Kuz12] and [Kuz12/2]. We start with an introduction to the Kuzmin limiter.

5.5.1 Kuzmin limiter

The design of the Kuzmin limiter starts with the definition of the following values. For $i = 1, \cdots, N$ we compute

$$P_i^+ := \sum_{\substack{j \in S_i \\ a_{ji} \le a_{ij}}} f_{ij}^+ \,, \qquad Q_i^+ := -\sum_{j \in S_i} f_{ij}^-$$

$$P_i^- := \sum_{\substack{j \in S_i \\ a_{ji} \le a_{ij}}} f_{ij}^- \,, \qquad Q_i^- := -\sum_{j \in S_i} f_{ij}^+$$

where $f_{ij} := d_{ij}(z_j - z_i)$ are the fluxes, $f_{ij}^+ = \max\{0, f_{ij}\}$, and $f_{ij}^- = \min\{0, f_{ij}\}$. Then, one defines for $i = 1, \cdots, N$

$$R_i^+ := \min\Big\{1, \frac{Q_i^+}{P_i^+}\Big\}, \quad R_i^- := \min\Big\{1, \frac{Q_i^-}{P_i^-}\Big\}. \tag{5.5.1}$$

If $P_i^+ = 0$ or $P_i^- = 0$, we set $R_i^+ := 1$ resp. $R_i^- := 1$. Furthermore, at Dirichlet nodes, we also set

$$R_i^+ := 1, \quad R_i^- := 1, \quad i = M+1, \cdots, N. \tag{5.5.2}$$

Then, for any $i, j \in \{1, \cdots, N\}$, $i \neq j$ with $a_{ji} \le a_{ij}$ the Kuzmin limiter is defined by

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } \; f_{ij} > 0, \\ 1 & \text{if } \; f_{ij} = 0, \\ R_i^- & \text{if } \; f_{ij} < 0 \end{cases} \tag{5.5.3}$$

with $\alpha_{ji} := \alpha_{ij}$. In [BJK16, Theorem 3] and [BJKR18, Lemma 4] it is shown that the Kuzmin limiter satisfies the continuity assumption (Assumption 5.9). Note that a proof corresponding to the satisfaction of Assumption 5.9 will be performed in Lemma 5.38 for the BJK limiter. However, the proof can be transferred to the Kuzmin limiter as well. Moreover, when the entries of the system matrix satisfy

$$\min\{a_{ij}, a_{ji}\} \le 0, \quad i = 1, \cdots, M, \quad j = 1, \cdots, N, \quad i \neq j \tag{5.5.4}$$

then, [BJKR18, Lemma 5] yields that the Kuzmin limiter satisfies Assumption 5.13 so that the local and global DMP hold. However, it is worth to mention that (5.5.4) does not hold on arbitrary meshes. In detail, Barrenechea et al. discuss in [BJK16, Remark 14] that (5.5.4) is not guaranteed for so-called non-Delaunay meshes, i.e. for meshes which contain obtuse triangles with angles bigger than $\frac{\pi}{2}$. In [Knob21] the author provides a modification of (5.5.3) so that the local DMP property holds on arbitrary meshes. As we have mentioned in the introduction of this section, the satisfaction of the linearity-preserving property (Assumption 5.37) may leads us to a better convergence behavior of the AFC method. Hence, in the next section we will introduce a linearity-preserving limiter.

5.5.2 BJK limiter

Similar to the Kuzmin limiter, the design of the BJK limiter is initiated by the definition of the following values. For $i = 1, \cdots, N$ we compute

$$P_i^+ := \sum_{j \in S_i} f_{ij}^+ \,, \qquad Q_i^+ := q_i(z_i - z_i^{\max}) \tag{5.5.5}$$

$$P_i^- := \sum_{j \in S_i} f_{ij}^- \,, \qquad Q_i^- := q_i(z_i - z_i^{\min}) \tag{5.5.6}$$

where again $f_{ij} := d_{ij}(z_j - z_i)$ and

$$z_i^{\max} := \max_{j \in S_i \cup \{i\}} z_j, \quad z_i^{\min} := \min_{j \in S_i \cup \{i\}} z_j, \quad q_i := \gamma_i \cdot \sum_{j \in S_i} d_{ij}$$

with fixed constants $\gamma_i > 0$. The quantities R_i^+ and R_i^- are defined by (5.5.1) and (5.5.2). Then, for $i, j \in \{1, \cdots, N\}$, $i \neq j$ we set

$$\tilde{\alpha}_{ij} := \begin{cases} R_i^+ & \text{if } \; f_{ij} > 0, \\ 1 & \text{if } \; f_{ij} = 0, \\ R_i^- & \text{if } \; f_{ij} < 0 \end{cases} \tag{5.5.7}$$

and finally

$$\alpha_{ij} := \min\{\tilde{\alpha}_{ij}, \tilde{\alpha}_{ji}\}. \tag{5.5.8}$$

The next result shows that the BJK limiter is continuous and fulfills the sufficient condition for the local DMP property. The following proof can also be found in [BJKR18, Lemma 6].

Lemma 5.38. *The BJK limiter satisfies Assumption 5.9 (continuity) and Assumption 5.13 (local DMP).*

Proof. We start with the verification of Assumption 5.9. It sufficies to show that $\tilde{\alpha}_{ij}(z_h)(z_j - z_i)$ is a continuous function of $z_h \in V_h$ with $z_i := z_h(x_i)$, $i = 1, \cdots, N$. For this, let us consider $f_{ij}(z_h) > 0$ and $d_{ij} < 0$ since the case $d_{ij} = 0$ leads to $\alpha_{ij} = 1$. Due to $d_{ij} < 0$, we have $z_i > z_j$. We thus get $f_{ij}(\tilde{z}_h) > 0$ for functions $\tilde{z}_h$ in a neighborhood of z_h. Using definition (5.5.7) we obtain

$$\tilde{\alpha}_{ij}(z_h) = R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\} = \frac{\min\left\{P_i^+, Q_i^+\right\}}{f_{ij} + \tilde{P}_i^+} \quad \text{with} \quad \tilde{P}_i^+ = \sum_{\substack{k \in S_i \\ k \neq j}} f_{ik}^+.$$

The numerator and the denominator are continuous functions and by virtue of $f_{ij} > 0$ the denominator is positive in a neighborhood of z_h. Consequently, $\tilde{\alpha}_{ij}$ is a continuous function at z_h. In the case $f_{ij}(z_h) < 0$ we obtain with $z_i < z_j$

$$\tilde{\alpha}_{ij}(z_h) = R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\} = \frac{\min\left\{-P_i^-, -Q_i^-\right\}}{|f_{ij}| - \tilde{P}_i^-} \quad \text{with} \quad \tilde{P}_i^- = \sum_{\substack{k \in S_i \\ k \neq j}} f_{ik}^-.$$

Hence, we can conclude in the same way as for $f_{ij}(z_h) > 0$. The last case is $f_{ij}(z_h) = 0$ which leads with the help of (5.5.7) and $d_{ij} < 0$ to $\tilde{\alpha}_{ij}(z_h)(z_j - z_i) = (z_j - z_i) = 0$. The boundedness of $\alpha_{ij} \in [0, 1]$ yields

$$\tilde{\alpha}_{ij}(\tilde{z}_h)(\tilde{z}_j - \tilde{z}_i) \to 0$$

as $\tilde{z}_j \to \tilde{z}_i$. Finally, the BJK limiter satisfies Assumption 5.9.
Now, we prove that the BJK limiter satisfies Assumption 5.13. For this, we assume that z_i, $i \in \{1, \cdots, M\}$ is a strict local extremum of $z = (z_1, \cdots, z_N)$. Then, we have to prove

$$a_{ij} + (1 - \min\{\tilde{\alpha}_{ij}(z_h), \tilde{\alpha}_{ji}(z_h)\})d_{ij} \leq 0 \quad \forall j \in S_i. \tag{5.5.9}$$

If $d_{ij} = 0$, then $a_{ij} \leq 0$ and hence (5.5.9) holds. Now, let us assume that $d_{ij} < 0$ and $z_i > z_j$ for any $j \in S_i$. Then, $f_{ij} > 0$ and $z_i^{\max} = z_i$ so that $P_i^+ > 0$, $Q_i^+ = 0$. Hence, $\tilde{\alpha}_{ij} = R_i^+ = 0$. Combined with the fact that $a_{ij} + d_{ij} \leq 0$ we obtain (5.5.9). The proof of the case $z_i < z_j$ for any $j \in S_i$ goes in the same way. Thus, the BJK limiter satisfies Assumption 5.13. □

5.5.3 Linearity-preserving property

As we have mentioned the linearity-preserving property suggests a better convergence behavior of the AFC method. Now, we see that for the BJK limiter the linearity-preserving property holds. In [BJK17, Section 6] the authors show that the validity of Assumption 5.37 is ensured when

$$Q_i^+ \geq P_i^+ \quad \text{for} \quad f_{ij} > 0, \quad Q_i^- \leq P_i^- \quad \text{for} \quad f_{ij} < 0 \tag{5.5.10}$$

hold. Moreover, the authors illustrate that if $z \in \mathbb{P}_1(\mathbb{R}^2)$ satisfies

$$z_i - z_i^{\min} \leq \gamma_i(z_i^{\max} - z_i), \qquad z_i^{\max} - z_i \leq \gamma_i(z_i - z_i^{\min}) \tag{5.5.11}$$

for appropriate constants $\gamma_i > 0$, then (5.5.10) is valid. Hence, the constants γ_i will be adjusted in such a way that the linearity-preserving property holds. Note that by changing the sign of the function z the first inequality of (5.5.11) implies the second. Thus, it sufficies to use constants such that

$$z_i - z_i^{\min} \leq \gamma_i(z_i^{\max} - z_i) \quad \forall z \in \mathbb{P}_1(\mathbb{R}^2). \tag{5.5.12}$$

In the case where the patches Δ_i are symmetric with respect to the vertex x_i
[BJK17, Lemma 6.1] yields that (5.5.12) holds with $\gamma_i = 1$. In the case of general patches Δ_i
[BJK17, Theorem 6.1] yields that inequality (5.5.12) holds, when

$$\gamma_i = \frac{\max\limits_{x_j \in \partial\Delta_i} |x_i - x_j|}{\operatorname{dist}(x_i, \partial\Delta_i^{\text{conv}})}, \quad i = 1, \cdots, M$$

where Δ_i^{conv} is the convex hull of Δ_i and $\partial\Delta_i^{\text{conv}}$ its boundary. The next result shows that the BJK limiter is linearity-preserving under the assumption that the constants γ_i fulfill condition (5.5.12). A proof can also be found in [BJKR18, Lemma 7].

Lemma 5.39. *Let γ_i for $i = 1, \cdots, M$ fulfill* (5.5.12)*. Then, the BJK limiter* (5.5.8) *satisfies* Assumption 5.37.

Proof. Let $i \in \{1, \cdots, M\}$. It sufficies to verify that for any $z_h \in \mathbb{P}_1(\mathbb{R}^2)$, one has $R_i^+(z_h) = 1 = R_i^-(z_h)$. Using condition (5.5.12), we obtain

$$P_i^+ = \sum_{\substack{j \in S_i \\ z_j < z_i}} d_{ij}(z_j - z_i) \leq \sum_{j \in S_i} d_{ij}(z_i^{\min} - z_i) \leq \sum_{j \in S_i} d_{ij}\gamma_i(z_i - z_i^{\max}) = Q_i^+$$

and hence $R_i^+ = 1$. Similary, one obtains $R_i^- = 1$. □

6 Coupled formulation of optimality conditions and AFC discretization

In this section, we provide a discretization concept corresponding to $(P_f), (P_b), (P_s^{MY})$ and (P_Γ). In the context of the *optimize-then-discretize*-approach we will obtain for each problem a coupled and discretized system. For this, recall that in the *optimize-then-discretize*-approach, first the optimality conditions are derived on the continuous level. After that, the discretization will be organized on the derived optimality systems. In this work, the optimality systems of $(P_f), (P_b), (P_s^{MY})$ and (P_Γ) will be discretized by the AFC method. For detailed information on the analysis of $(P_f), (P_b), (P_s^{MY})$ and (P_Γ), we refer to Section 4. Furthermore, we remark that in Section 7 resp. Section 8 the following coupled and discretized systems will be analyzed.

6.1 Unconstrained case

Now, let us start with the discretization of the unconstrained optimal control problem

$$\left.\begin{array}{l} \min\ J^f(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b} \cdot \nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \end{array}\right\} (P_f)$$

For this, regarding Section 4.1, the unique solution $(\bar{y}, \bar{u})$ of (P_f) satisfies the following optimality system

$$\begin{array}{l} a(\bar{y}, v) = (\bar{u}, v)_\Omega \quad \forall v \in H_0^1(\Omega) \\ a(\psi, \bar{p}) = (\bar{y} - y_d, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \\ \bar{p} + \lambda\bar{u} = 0 \quad \text{a.e. in } \Omega \end{array}$$

where $\bar{p}$ is the corresponding adjoint solution.

6.1.1 Discretized system

Following the *optimize-then-discretize*-approach, we obtain a coupled and discretized system

$$\left.\begin{array}{l} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \\ a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \\ \lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \end{array}\right\} (P_h^f)$$

Note that $d_h^s(y_h; y_h, v_h)$ resp. $d_h^{ad}(p_h; p_h, \psi_h)$ are the stabilization terms introduced in (5.3.5) resp. (5.4.5).

Definition 6.1. *A pair* $(y_h, u_h) \in V_{h,0} \times V_{h,0}$ *is called solution for* (P_h^f) *if there exists a discrete adjoint solution* $p_h \in V_{h,0}$ *such that*

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \tag{6.1.1}$$
$$a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \tag{6.1.2}$$
$$\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \tag{6.1.3}$$

is satisfied.

6.2 Control constrained case

Now, we apply the *optimize-then-discretize*-approach on the control constrained optimal control problem

$$\left.\begin{array}{l} \min \ J^b(y,u) := \ \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ u_a \leq u \leq u_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_b)$$

Regarding Section 4.2, the solution $(\bar{y}, \bar{u})$ of (P_b) satisfies the following optimality system

$$\begin{aligned} a(\bar{y}, v) &= (\bar{u}, v)_\Omega \quad \forall v \in H_0^1(\Omega) \\ a(\psi, \bar{p}) &= (\bar{y} - y_d, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \\ \bar{u} &= \mathbb{P}_{[u_a, u_b]}\left(-\frac{1}{\lambda}\bar{p}\right) \quad \text{a.e. in } \Omega. \end{aligned}$$

6.2.1 Discretized system

The coupled and discretized system is given by

$$\left.\begin{array}{l} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \\ a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \\ u_h = \mathbb{P}_{[u_a, u_b]}\left(-\frac{1}{\lambda} p_h\right) \quad \text{a.e. in } \Omega \end{array}\right\} (P_h^b)$$

Definition 6.2. *A pair* $(y_h, u_h) \in V_{h,0} \times L^2(\Omega)$ *is called solution for* (P_h^b) *if there exists a discrete adjoint solution* $p_h \in V_{h,0}$ *such that*

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \tag{6.2.1}$$
$$a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \tag{6.2.2}$$
$$u_h = \mathbb{P}_{[u_a, u_b]}\left(-\frac{1}{\lambda} p_h\right) \quad \text{a.e. in } \Omega \tag{6.2.3}$$

is satisfied.

6.3 State constrained case - Moreau-Yosida regularization

As we have mentioned in Section 4.3, the lack of regularity of the measures μ_a, μ_b complicates the numerical treatment of (P_s). Thus, a direct application of the *optimize-then-discretize*-approach on (P_s) is not advisible and currently not realizable. The remedy is to be found in

the discretization of an appropriate regularization of (P_s) (see Section 4.3.2). In our work, we use the Moreau-Yosida regularization (P_s^{MY}). Now, the strategy is to construct a solution $(y_h, u_h) \in V_{h,0} \times V_{h,0}$ with a corresponding adjoint solution $p_h \in V_{h,0}$ for the Moreau-Yosida regularization which is also an approximation for the solution $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times H_0^1(\Omega)$ with $\bar{p} \in H_0^1(\Omega)$ corresponding to (P_s). For this, recall the weak formulation of the optimality conditions of (P_s^{MY}) (see Theorem 4.16):

$$\begin{aligned} a(\bar{y}_\delta, v) &= (\bar{u}_\delta, v)_\Omega \quad \forall v \in H_0^1(\Omega) \\ a(\psi, \bar{p}_\delta) &= (\bar{y}_\delta - y_d, \psi)_\Omega + (\mu_\delta, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \\ \lambda \bar{u}_\delta + \bar{p}_\delta &= 0 \quad \text{a.e. in } \Omega \\ \mu_\delta &= \delta \cdot (\max\{0, \bar{y}_\delta - y_b\} + \min\{0, \bar{y}_\delta - y_a\}) \quad \text{a.e. in } \Omega. \end{aligned} \tag{6.3.1}$$

In the following, we use a different representation of μ_δ. In detail, we use the pointwise projection formula $\mathbb{P}_{[y_a,y_b]}(\cdot)$ given by

$$\mathbb{P}_{[y_a,y_b]}(r) = \min\{y_b, \max\{r, y_a\}\}.$$

According to Lemma 2.26, we are able to rewrite the weak formulation (6.3.1) in so far as:

$$\begin{aligned} a(\bar{y}_\delta, v) &= (\bar{u}_\delta, v)_\Omega \quad \forall v \in H_0^1(\Omega) \\ a(\psi, \bar{p}_\delta) &= (\bar{y}_\delta - y_d, \psi)_\Omega + (\mu_\delta, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega) \\ \lambda \bar{u}_\delta + \bar{p}_\delta &= 0 \quad \text{a.e. in } \Omega \\ \mu_\delta &= \delta \cdot \big(\bar{y}_\delta - \mathbb{P}_{[y_a,y_b]}(\bar{y}_\delta)\big) \quad \text{a.e. in } \Omega. \end{aligned}$$

6.3.1 Discretized system

The *optimize-then-discretize*-approach yields the following coupled and discretized system

$$\left.\begin{aligned} &a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \\ &a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega + \big(\mu_h^\delta, \psi_h\big)_\Omega \quad \forall \psi_h \in V_{h,0} \\ &\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \\ &\mu_h^\delta = \delta \cdot (y_h - \mathbb{P}_{[y_a,y_b]}(y_h)) \quad \text{a.e. in } \Omega \end{aligned}\right\} (P_{s,h}^{MY})$$

Definition 6.3. *A pair $(y_h, u_h) \in V_{h,0} \times V_{h,0}$ is called solution for $(P_{s,h}^{MY})$ if there exists a discrete adjoint solution $p_h \in V_{h,0}$ such that*

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \tag{6.3.2}$$

$$a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega + \big(\mu_h^\delta, \psi_h\big)_\Omega \quad \forall \psi_h \in V_{h,0} \tag{6.3.3}$$

$$\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \tag{6.3.4}$$

$$\mu_h^\delta = \delta \cdot (y_h - \mathbb{P}_{[y_a,y_b]}(y_h)) \quad \text{a.e. in } \Omega \tag{6.3.5}$$

is satisfied.

6.4 Control constrained case with Robin boundary control

Apart from the optimal control problems with distributed control, we apply the *optimize-then-discretize*-approach on the following control constrained optimal control problem with Robin

boundary control.

$$\left.\begin{array}{l}\min\ J^{\Gamma}(y,u) := \frac{1}{2}\parallel y-y_d\parallel_{0,\Omega}^2+\frac{\lambda}{2}\parallel u\parallel_{0,\Gamma}^2\\ -\varepsilon\Delta y+\mathbf{b}\cdot\nabla y+cy=0 \quad \text{in } \Omega\\ \varepsilon\partial_n y-\frac{\mathbf{b}\cdot n\cdot y}{2}=u \quad \text{on } \Gamma\\ u_a^{\Gamma}\le u\le u_b^{\Gamma} \quad \text{a.e. on } \Gamma\end{array}\right\}(P_{\Gamma})$$

Section 4.4 provides that the solution $(\bar{y},\bar{u})$ of (P_{Γ}) satisfies

$$\begin{aligned} a_{\Gamma}(\bar{y},v) &= (\bar{u},v)_{\Gamma} \quad \forall v\in H^1(\Omega)\\ a_{\Gamma}(\psi,\bar{p}) &= (\bar{y}-y_d,\psi)_{\Omega} \quad \forall \psi\in H^1(\Omega)\\ \bar{u} &= \mathbb{P}_{[u_a^{\Gamma},u_b^{\Gamma}]}\left(-\frac{1}{\lambda}\bar{p}\right) \quad \text{a.e. on } \Gamma\end{aligned}$$

where $\bar{p}$ is the corresponding adjoint solution.

6.4.1 Discretized system

The *optimize-then-discretize*-approach yields the following coupled and discretized system

$$\left.\begin{array}{l} a_{\Gamma}(y_h,v_h)+d_h^{s,\Gamma}(y_h;y_h,v_h)=(u_h,v_h)_{\Gamma} \quad \forall v_h\in V_h\\ a_{\Gamma}(\psi_h,p_h)+d_h^{ad,\Gamma}(p_h;p_h,\psi_h)=(y_h-y_d,\psi_h)_{\Omega} \quad \forall \psi_h\in V_h\\ u_h=\mathbb{P}_{[u_a^{\Gamma},u_b^{\Gamma}]}\left(-\frac{1}{\lambda}p_h\right) \quad \text{a.e. on } \Gamma\end{array}\right\}(P_h^{\Gamma})$$

Note that the stabilization terms $d_h^{s,\Gamma}(y_h;y_h,v_h)$ resp. $d_h^{ad,\Gamma}(p_h;p_h,\psi_h)$ have been introduced in (5.3.14) resp. in (5.4.6).

Definition 6.4. *A pair $(y_h,u_h)\in V_h\times L^2(\Gamma)$ is called solution for (P_h^{Γ}) if there exists a discrete adjoint solution $p_h\in V_h$ such that*

$$a(y_h,v_h)+d_h^{s,\Gamma}(y_h;y_h,v_h)=(u_h,v_h)_{\Gamma} \quad \forall v_h\in V_h \tag{6.4.1}$$

$$a(\psi_h,p_h)+d_h^{ad,\Gamma}(p_h;p_h,\psi_h)=(y_h-y_d,\psi_h)_{\Omega} \quad \forall \psi_h\in V_h \tag{6.4.2}$$

$$u_h=\mathbb{P}_{[u_a^{\Gamma},u_b^{\Gamma}]}\left(-\frac{1}{\lambda}p_h\right) \quad \text{a.e. on } \Gamma \tag{6.4.3}$$

is satisfied.

7 Abstract formulation

An accurate observation of the discretized systems (P_h^f), (P_h^b), $(P_{s,h}^{MY})$, and (P_h^{Γ}) yields that the structure of all systems are quite similiar. In detail, we are able to rewrite the discrete systems by an abstract operator equation

$$K_h\mathbf{x}_h=Q\mathbf{x}_h+G \quad \text{in } \mathcal{X}_h^* \tag{7.0.1}$$

where $\mathcal{X}_h$ is a finite dimensional Hilbert space, $\mathcal{X}_h^*$ its dual, and $G\in\mathcal{X}_h^*$. The structure of (7.0.1) guides us to the supposition that the discrete problems can be analyzed in a unique way. Hence, in the first part of this section we prove for a general framework of (7.0.1) the existence of a discrete solution $\mathbf{x}_h$. After that, we derive for general coupled systems a L^2-error estimate for the control and the state. We remark that in Section 8, the following abstract results will be applied for the analysis of (P_h^f), (P_h^b), $(P_{s,h}^{MY})$, and (P_h^{Γ}).

7.1 Existence of discrete solutions

In the following let $\mathcal{X}$ be a Hilbert space with inner product $(\cdot,\cdot)_{\mathcal{X}}$ and norm $\|\cdot\|_{\mathcal{X}}$. Moreover, let $(\mathcal{X}_h, \|\cdot\|_{\mathcal{X}_h})$ with $\mathcal{X}_h \subset \mathcal{X}$ be a finite dimensional Hilbert space with inner product $(\cdot,\cdot)_{\mathcal{X}_h}$ and norm $\|\cdot\|_{\mathcal{X}_h}$. We consider a functional $G \in \mathcal{X}_h^*$ and continuous operators $K_h : \mathcal{X}_h \to \mathcal{X}_h^*$ and $Q : \mathcal{X}_h \to \mathcal{X}_h^*$.

Assumption 7.1. *We assume that the operator $K_h : \mathcal{X}_h \to \mathcal{X}_h^*$ is continuous and satisfies*

$$\langle K_h \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \geq C_K \| \mathbf{x}_h \|_{\mathcal{X}_h}^2 \quad \forall\, \mathbf{x}_h \in \mathcal{X}_h \tag{7.1.1}$$

where $C_K > 0$ is a constant.

Assumption 7.2. *We assume that the operator $Q : \mathcal{X}_h \to \mathcal{X}_h^*$ is continuous and satisfies*

$$\langle Q \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \leq C_Q \| \mathbf{x}_h \|_{\mathcal{X}_h} \quad \forall\, \mathbf{x}_h \in \mathcal{X}_h \tag{7.1.2}$$

where $C_Q > 0$ is a constant.

The following lemma yields the existence of a discrete solution $\mathbf{x}_h \in \mathcal{X}_h$ for the operator equation

$$K_h \mathbf{x}_h = Q \mathbf{x}_h + G \quad \text{in } \mathcal{X}_h^*.$$

Lemma 7.3. *Under the Assumptions 7.1 and 7.2 there exists a solution $\mathbf{x}_h \in \mathcal{X}_h$ such that*

$$K_h \mathbf{x}_h = Q \mathbf{x}_h + G \quad \text{in } \mathcal{X}_h^* \tag{7.1.3}$$

is satisfied.

Proof. We define the operator $T : \mathcal{X}_h \to \mathcal{X}_h^*$ by

$$T \mathbf{x}_h := K_h \mathbf{x}_h - Q \mathbf{x}_h - G.$$

Assumption 7.1 and Assumption 7.2 imply

$$\begin{aligned} \langle T \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} &= \langle K_h \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} - \langle Q \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} - \langle G, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \\ &\geq C_K \| \mathbf{x}_h \|_{\mathcal{X}_h}^2 - C_Q \| \mathbf{x}_h \|_{\mathcal{X}_h} - \| G \|_{\mathcal{X}_h^*} \| \mathbf{x}_h \|_{\mathcal{X}_h} . \end{aligned}$$

The application of Young's inequality on the last two terms on the right hand side results in

$$\begin{aligned} \langle T \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} &\geq C_K \| \mathbf{x}_h \|_{\mathcal{X}_h}^2 - \frac{C_K}{2} \| \mathbf{x}_h \|_{\mathcal{X}_h}^2 - \frac{C_Q^2}{C_K} - \frac{\| G \|_{\mathcal{X}_h^*}^2}{C_K} \\ &= \frac{C_K}{2} \| \mathbf{x}_h \|_{\mathcal{X}_h}^2 - \frac{C_Q^2}{C_K} - \frac{\| G \|_{\mathcal{X}_h^*}^2}{C_K}. \end{aligned} \tag{7.1.4}$$

By virtue of Riesz's respresentation theorem (see Theorem 3.1) we get

$$(T \mathbf{x}_h, \mathbf{x}_h)_{\mathcal{X}_h} = \langle T \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h}. \tag{7.1.5}$$

Hence, for all $\mathbf{x}_h \in \mathcal{X}_h$ with

$$\| \mathbf{x}_h \|_{\mathcal{X}_h} > \sqrt{\frac{2\left(C_Q^2 + \| G \|_{\mathcal{X}_h^*}^2\right)}{C_K^2}}$$

it holds

$$(T\mathbf{x}_h, \mathbf{x}_h)_{\mathcal{X}_h} > 0$$

with respect to (7.1.4) and (7.1.5). Lemma 5.8 yields the existence of a solution $\bar{\mathbf{x}}_h \in \mathcal{X}_h$ such that

$$T\bar{\mathbf{x}}_h = 0 \quad \text{in } \mathcal{X}_h.$$

Again, the application of Theorem 3.1 implies

$$T\bar{\mathbf{x}}_h = 0 \quad \text{in } \mathcal{X}_h^*$$

and consequently the desired result. □

7.2 General error estimates for coupled systems

Let $\mathcal{G}$ be Ω or the boundary Γ. In the case where $\mathcal{G} = \Omega$, let $X = H_0^1(\Omega)$ and $X_h = V_{h,0}$, the Finite Element space of $\mathbb{P}_1$ Finite Elements, introduced in Section 5.1. In the case $\mathcal{G} = \Gamma$, let $X = H^1(\Omega)$ and $X_h = V_h$. For the several cases of $\mathcal{G}$ we use the corresponding product spaces $\mathcal{X} = X \times X$ and $\mathcal{X}_h = X_h \times X_h$. Now, we introduce operators, sets and spaces which will be used in this section. Initially, let $U \subseteq L^2(\mathcal{G})$ be a nonempty, closed and convex set. We consider an operator $K^s \in \mathcal{L}(X, X^*)$ with the corresponding adjoint operator $K^{ad} := (K^s)^* \in \mathcal{L}(X, X^*)$ where $\mathcal{L}(X, X^*)$ denotes the set of all linear and continuous functionals from X to X^*. Moreover, we consider $G \in X^*$, operators $D^s_{(\cdot)} : X_h \to \mathcal{L}(X_h, X_h^*)$, $D^{ad}_{(\cdot)} : X_h \to \mathcal{L}(X_h, X_h^*)$, continuous operators

$$Z : C(\bar{\Omega}) \to U, \quad R : C(\bar{\Omega}) \to C(\bar{\Omega})$$

and constants $\delta, \hat{\delta} \in \mathbb{R}_{\geq 0}$ such that

$$\delta - \hat{\delta} > 0.$$

We assume that the following continuous problem possesses a solution $(y, u) \in (X \cap C(\bar{\Omega})) \times L^2(\mathcal{G})$ with a corresponding adjoint solution $p \in X \cap C(\bar{\Omega})$ such that for all $(v, \psi) \in \mathcal{X}$ the system

$$\left.\begin{aligned}
&\langle K^s y, v\rangle_{X^*,X} = (u, v)_{\mathcal{G}} \\
&\langle K^{ad} p, \psi\rangle_{X^*,X} = (\delta \cdot y - \hat{\delta} \cdot Ry, \psi)_\Omega + \langle G, \psi\rangle_{X^*,X} \\
&u = Zp \quad \text{a.e. in } \mathcal{G} \\
&(\lambda u + p, w - u)_{\mathcal{G}} \geq 0 \quad \forall w \in U
\end{aligned}\right\} (P)$$

is satisfied. In addition to the continuity assumption, the operator R satisfies the following condition.

Assumption 7.4. *The operator $R : C(\bar{\Omega}) \to C(\bar{\Omega})$ is Lipschitz continuous in so far as: For $z, \tilde{z} \in C(\bar{\Omega})$ we have*

$$(Rz - R\tilde{z}, v)_\Omega \leq \| z - \tilde{z} \|_{0,\Omega} \| v \|_{0,\Omega} \quad \forall v \in C(\bar{\Omega}).$$

Furthermore, we assume that for $z \in C(\bar{\Omega})$

$$\| Rz \|_{0,\Omega} \leq C_R$$

where $C_R > 0$ is a constant, independent of z.

Remark 7.5. *For varying $\delta, \hat{\delta}, \mathcal{G}, Z, R$, and U, the continuous problem (P) corresponds to the optimality systems of $(P_f), (P_b), (P_s^{MY})$, and (P_Γ). We will return to this discussion in Section 8.*

Now, let us consider the discrete counterpart to (P). We assume that there exists a solution $(y_h, u_h) \in X_h \times L^2(\mathcal{G})$ with a corresponding discrete adjoint solution $p_h \in X_h$ such that for all $(v_h, \psi_h) \in \mathcal{X}_h$ the system

$$\left.\begin{aligned}
&\langle K^s y_h, v_h\rangle_{X^*,X} + \langle D^s_{y_h} y_h, v_h\rangle_{X_h^*,X_h} = (u_h, v_h)_{\mathcal{G}} \\
&\langle K^{ad} p_h, \psi_h\rangle_{X^*,X} + \langle D^{ad}_{p_h} p_h, \psi_h\rangle_{X_h^*,X_h} = (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi_h)_\Omega + \langle G, \psi_h\rangle_{X^*,X} \\
&u_h = Z p_h \quad \text{a.e. in } \mathcal{G} \\
&(\lambda u_h + p_h, w - u_h)_{\mathcal{G}} \geq 0 \quad \forall w \in U
\end{aligned}\right\} (P_h)$$

is satisfied. Now, according to (P) and (P_h), we derive a general L^2-error estimate for the differences $y - y_h$ and $u - u_h$ with respect to the following auxiliary problem. For this, let us assume that $(y_h, u_h) \in X_h \times L^2(\mathcal{G})$ with $p_h \in X_h$ solves (P_h). Moreover, we assume that there exists a solution $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (C(\bar{\Omega}) \times C(\bar{\Omega}))$ such that for all $(v, \psi) \in \mathcal{X}$ the system

$$\left.\begin{aligned}
&\langle K^s \tilde{y}, v\rangle_{X^*,X} = (u_h, v)_{\mathcal{G}} \\
&\langle K^{ad} \tilde{p}, \psi\rangle_{X^*,X} = (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi)_\Omega + \langle G, \psi\rangle_{X^*,X}
\end{aligned}\right\} (P_{aux})$$

is satisfied.

Lemma 7.6. *Let Assumption 7.4 on the operator R be fulfilled. Furthermore, let $(y, u) \in (X \cap C(\bar{\Omega})) \times L^2(\mathcal{G})$ be a solution of (P) with a corresponding adjoint solution $p \in X \cap C(\bar{\Omega})$. Moreover, let $(y_h, u_h) \in X_h \times L^2(\mathcal{G})$ be a solution of (P_h) with a corresponding adjoint solution $p_h \in X_h$. The pair $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (C(\bar{\Omega}) \times C(\bar{\Omega}))$ solves (P_{aux}). Then, we have*

$$\begin{aligned}
\frac{\lambda}{2} \parallel u_h - u \parallel^2_{0,\mathcal{G}} + \frac{(\delta - \hat{\delta})}{2} \parallel y_h - y \parallel^2_{0,\Omega} \leq C \parallel p_h - \tilde{p} \parallel^2_{0,\mathcal{G}} + \hat{\delta} C_R \parallel y_h - \tilde{y} \parallel_{0,\Omega} \\
+ \frac{\delta^2}{2(\delta - \hat{\delta})} \parallel y_h - \tilde{y} \parallel^2_{0,\Omega}
\end{aligned}$$

where $C, C_R > 0$ are constants.

Proof. First, we consider the variational inequalities

$$(\lambda u + p, w - u)_{\mathcal{G}} \geq 0 \quad \forall w \in U \tag{7.2.1}$$

and

$$(\lambda u_h + p_h, w - u_h)_{\mathcal{G}} \geq 0 \quad \forall w \in U. \tag{7.2.2}$$

Note that $u = Zp \in U$ and $u_h = Zp_h \in U$. Hence, by virtue of (7.2.1) and (7.2.2), we are able to derive the inequality

$$\begin{aligned}
\lambda \parallel u_h - u \parallel^2_{0,\mathcal{G}} &\leq (p_h - p, u - u_h)_{\mathcal{G}} \\
&= \underbrace{(p_h - \tilde{p}, u - u_h)_{\mathcal{G}}}_{(1)} + \underbrace{(\tilde{p} - p, u - u_h)_{\mathcal{G}}}_{(2)}.
\end{aligned} \tag{7.2.3}$$

For (1), Young's inequality yields

$$\begin{aligned}
(p_h - \tilde{p}, u - u_h)_{\mathcal{G}} &\leq \parallel p_h - \tilde{p} \parallel_{0,\mathcal{G}} \parallel u_h - u \parallel_{0,\mathcal{G}} \\
&\leq C \parallel p_h - \tilde{p} \parallel^2_{0,\mathcal{G}} + \frac{\lambda}{2} \parallel u_h - u \parallel^2_{0,\mathcal{G}}.
\end{aligned}$$

For the second term (2), we get by using the equations of (P) resp. (P_{aux})

$$\begin{aligned}(\tilde{p}-p, u-u_h)_{\mathcal{G}} &= \langle K^s y - K^s \tilde{y}, \tilde{p}-p\rangle_{X^*,X} \\ &= \langle y-\tilde{y}, (K^s)^*\tilde{p} - (K^s)^* p\rangle_{X,X^*} \\ &= \langle y-\tilde{y}, K^{ad}\tilde{p} - K^{ad} p\rangle_{X,X^*} \\ &= (\delta\cdot(y_h-y) - \hat{\delta}\cdot(Ry_h - Ry), y-\tilde{y})_\Omega \\ &= \underbrace{\delta\cdot(y_h-y, y-\tilde{y})_\Omega}_{(3)} - \underbrace{\hat{\delta}\cdot(Ry_h-Ry, y-\tilde{y})_\Omega}_{(4)}.\end{aligned} \tag{7.2.4}$$

For (3), we obtain by inserting the discrete solution y_h

$$\begin{aligned}\delta\cdot(y_h-y, y-\tilde{y})_\Omega &= \delta\cdot(y_h-y, y-y_h)_\Omega + \delta\cdot(y_h-y, y_h-\tilde{y})_\Omega \\ &\leq -\delta \parallel y_h - y \parallel_{0,\Omega}^2 + \delta \parallel y_h - y \parallel_{0,\Omega} \parallel y_h - \tilde{y} \parallel_{0,\Omega}.\end{aligned}$$

For (4), we get

$$\begin{aligned}\hat{\delta}\cdot(Ry-Ry_h, y-\tilde{y})_\Omega &= \hat{\delta}\cdot(Ry-Ry_h, y_h-\tilde{y})_\Omega + \hat{\delta}\cdot(Ry-Ry_h, y-y_h)_\Omega \\ &\leq \hat{\delta} C_R \parallel y_h - \tilde{y} \parallel_{0,\Omega} + \hat{\delta} \parallel y - y_h \parallel_{0,\Omega}^2\end{aligned}$$

where we have used for the first term the uniform boundedness of R and for the second term the Lipschitz continuity of R (see Assumption 7.4). Hence, we can derive for (7.2.4)

$$(\tilde{p}-p, u-u_h)_{\mathcal{G}} \leq (\hat{\delta}-\delta)\cdot \parallel y_h - y \parallel_{0,\Omega}^2 + \delta \parallel y_h - y \parallel_{0,\Omega} \parallel y_h - \tilde{y} \parallel_{0,\Omega} + \hat{\delta} C_R \parallel y_h - \tilde{y} \parallel_{0,\Omega}.$$

Collecting and rearranging of the estimated terms $(1), (2)$ yields for (7.2.3)

$$\begin{aligned}\frac{\lambda}{2} \parallel u_h - u \parallel_{0,\mathcal{G}}^2 + (\delta-\hat{\delta}) \parallel y_h - y \parallel_{0,\Omega}^2 &\leq C \parallel p_h - \tilde{p} \parallel_{0,\mathcal{G}}^2 + \hat{\delta} C_R \parallel y_h - \tilde{y} \parallel_{0,\Omega} \\ &\quad + \delta \parallel y_h - y \parallel_{0,\Omega} \parallel y_h - \tilde{y} \parallel_{0,\Omega}.\end{aligned}$$

Consequently, Young's inequality (2.2.4) implies with $\gamma = \delta - \hat{\delta} > 0$

$$\begin{aligned}\frac{\lambda}{2} \parallel u_h - u \parallel_{0,\mathcal{G}}^2 + (\delta-\hat{\delta}) \parallel y_h - y \parallel_{0,\Omega}^2 &\leq C \parallel p_h - \tilde{p} \parallel_{0,\mathcal{G}}^2 + \hat{\delta} C_R \parallel y_h - \tilde{y} \parallel_{0,\Omega} \\ &\quad + \frac{(\delta-\hat{\delta})}{2} \parallel y_h - y \parallel_{0,\Omega}^2 + \frac{\delta^2}{2(\delta-\hat{\delta})} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2\end{aligned}$$

and finally

$$\begin{aligned}\frac{\lambda}{2} \parallel u_h - u \parallel_{0,\mathcal{G}}^2 + \frac{(\delta-\hat{\delta})}{2} \parallel y_h - y \parallel_{0,\Omega}^2 &\leq C \parallel p_h - \tilde{p} \parallel_{0,\mathcal{G}}^2 + \hat{\delta} C_R \parallel y_h - \tilde{y} \parallel_{0,\Omega} \\ &\quad + \frac{\delta^2}{2(\delta-\hat{\delta})} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2.\end{aligned}$$

□

We note that in the next section an estimate for $\parallel y_h - \tilde{y} \parallel_{0,\Omega}$ and $\parallel p_h - \tilde{p} \parallel_{0,\mathcal{G}}$ will be provided for a specific framework of (P_h) and (P_{aux}) such that the cases (P_h^f), (P_h^b), $(P_{s,h}^{MY})$, and (P_h^Γ) are covered. The application of Lemma 7.6 and the results of the following section will be demonstrated in Section 8.

7.2.1 Auxiliary error estimates

As we can see in Lemma 7.6, the estimate depends on $\| p_h - \tilde{p} \|_{0,\mathcal{G}}$ and $\| y_h - \tilde{y} \|_{0,\Omega}$ where $(\tilde{y}, \tilde{p})$ is the auxiliary solution of (P_{aux}). In the following, we derive for $\delta, \hat{\delta} \in \mathbb{R}_{\geq 0}$ with $\delta - \hat{\delta} > 0$ and $\mathcal{G} = \Omega$ resp. $\mathcal{G} = \Gamma$ an auxiliary error estimate for $p_h - \tilde{p}$ and $y_h - \tilde{y}$ in the L^2-norm. For this, let us recall the introduced auxiliary problem (see Section 7.2)

$$\left.\begin{aligned} &\langle K^s \tilde{y}, v\rangle_{X^*,X} = (u_h, v)_{\mathcal{G}} \\ &\langle K^{ad} \tilde{p}, \psi\rangle_{X^*,X} = (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi)_\Omega + \langle G, \psi\rangle_{X^*,X} \end{aligned}\right\} (P_{aux})$$

where $u_h = Z p_h$ a.e. in $\mathcal{G}$. Note that $(y_h, u_h) \in X_h \times L^2(\mathcal{G})$ with $p_h \in X_h$ solves for all $(v_h, \psi_h) \in \mathcal{X}_h$ the discrete system

$$\left.\begin{aligned} &\langle K^s y_h, v_h\rangle_{X^*,X} + \langle D^s_{y_h} y_h, v_h\rangle_{X_h^*,X_h} = (u_h, v_h)_{\mathcal{G}} \\ &\langle K^{ad} p_h, \psi_h\rangle_{X^*,X} + \langle D^{ad}_{p_h} p_h, \psi_h\rangle_{X_h^*,X_h} = (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi_h)_\Omega + \langle G, \psi_h\rangle_{X^*,X} \\ &u_h = Z p_h \quad \text{a.e. in } \mathcal{G} \\ &(\lambda u_h + p_h, w - u_h)_{\mathcal{G}} \geq 0 \quad \forall w \in U \end{aligned}\right\} (P_h)$$

As we have described in Section 7.2, the discretized systems (P_h^f), (P_h^b), $(P_{s,h}^{MY})$, and (P_h^Γ) can be expressed uniformly by (P_h) where the several problems are obtained by varying the definition of $\mathcal{G}$, $\delta, \hat{\delta}, Z, R$, and U. First, we derive an auxiliary $L^2(\Omega)$-error estimate for the differences $\tilde{y} - y_h$ and $\tilde{p} - p_h$ in the case $\mathcal{G} = \Omega$ such that the systems $[(P_h^f), (P_h^b), (P_{s,h}^{MY})]$ are covered. After that, we derive an auxiliary L^2-error estimate for the case $\mathcal{G} = \Gamma$ such that (P_h^Γ) is covered.

7.2.1.1 Auxiliary result for $\mathcal{G} = \Omega$

According to Section 7.1, for the investigation of (P_h^f), (P_h^b), $(P_{s,h}^{MY})$ we set $X_h = V_{h,0} \subseteq H_0^1(\Omega) = X$ and $\mathcal{X}_h = V_{h,0} \times V_{h,0} \subset \mathcal{X} = H_0^1(\Omega) \times H_0^1(\Omega)$ where $\mathcal{X}$ is equipped with the norm

$$\| (y, p) \|_{\mathcal{X}} = \sqrt{|y|_{1,\Omega}^2 + |p|_{1,\Omega}^2}.$$

The finite dimensional space $\mathcal{X}_h$ is equipped with the norm

$$\| \cdot \|_{\mathcal{X}_h} := \| \cdot \|_{\mathcal{X}} .$$

Throughout this section we consider the following operators:

- $K^s \in \mathcal{L}(X, X^*)$ defined by

$$\langle K^s y, v\rangle_{X^*,X} := a(y, v)$$

- $(K^s)^* = K^{ad} \in \mathcal{L}(X, X^*)$ defined by

$$\langle K^{ad} p, \psi\rangle_{X^*,X} := a(\psi, p)$$

- $G \in X^*$ defined by

$$\langle G, \psi\rangle_{X^*,X} := (-y_d, \psi)_\Omega$$

- $D^s_{(\cdot)} : X_h \to \mathcal{L}(X_h, X_h^*)$ defined by

$$w_h \mapsto D^s_{w_h} \in \mathcal{L}(X_h, X_h^*)$$

and

$$\langle D^s_{w_h} y_h, v_h \rangle_{X_h^*, X_h} := d_h^s(w_h; y_h, v_h)$$

- $D^{ad}_{(\cdot)} : X_h \to \mathcal{L}(X_h, X_h^*)$ defined by

$$w_h \mapsto D^{ad}_{w_h} \in \mathcal{L}(X_h, X_h^*)$$

and

$$\langle D^{ad}_{w_h} p_h, \psi_h \rangle_{X_h^*, X_h} := d_h^{ad}(w_h; p_h, \psi_h)$$

Note that $\delta, \hat{\delta} \in \mathbb{R}_{\geq 0}$ with $\delta - \hat{\delta} > 0$ are fixed constants. In the following, we consider the general Problem (P_h) which corresponds for specified $\delta, \hat{\delta} \in \mathbb{R}_{\geq 0}$ to (P_h^f), (P_h^b), and $(P_{s,h}^{MY})$. For the next results it is helpful to remark that the state and the adjoint equation of (P_h) can be written as

$$\begin{aligned} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) &= (u_h, v_h)_\Omega \quad \forall v_h \in V_{h,0} \\ a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) &= (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi_h)_\Omega - (y_d, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0}. \end{aligned} \tag{7.2.5}$$

(P_{aux}) can be expressed by

$$\begin{aligned} a(\tilde{y}, v) &= (u_h, v)_\Omega \quad \forall v \in H_0^1(\Omega) \\ a(\psi, \tilde{p}) &= (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi)_\Omega - (y_d, \psi)_\Omega \quad \forall \psi \in H_0^1(\Omega). \end{aligned} \tag{7.2.6}$$

According to Section 5.3 resp. Section 5.4, the state (adjoint) equation of (7.2.5) is the corresponding stabilized equation for the auxiliary state (adjoint) equation of (7.2.6). Before we start with the error analysis between (P_{aux}) and (P_h), we derive $H^2(\Omega)$-a priori estimates for the auxiliary solution $(\tilde{y}, \tilde{p})$. The meaning of these estimates will become apparent in Section 8.

Lemma 7.7. *Let $\mathcal{G} = \Omega$ and $(y_h, u_h) \in X_h \times L^2(\Omega)$ be a solution of (P_h) with a corresponding adjoint solution $p_h \in X_h$. Then, there exists a unique solution $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ for (P_{aux}). Moreover, the following $H^2(\Omega)$-a priori estimates hold:*

$$\| \tilde{y} \|_{2,\Omega} \leq C \| u_h \|_{0,\Omega} \tag{7.2.7}$$

$$\| \tilde{p} \|_{2,\Omega} \leq C \left(\delta \| y_h \|_{0,\Omega} + \hat{\delta} C_R + \| y_d \|_{0,\Omega} \right). \tag{7.2.8}$$

Proof. The operators K^s and K^{ad} are defined by the coercive and continuous bilinear form $a(\cdot, \cdot)$ on $\mathcal{X}$ (see Lemma 3.4). Considering (7.2.5) and (7.2.6), the right hand sides

$$X \ni v \mapsto (u_h, v)_\Omega \tag{7.2.9}$$

and

$$X \ni \psi \mapsto (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi)_\Omega - (y_d, \psi)_\Omega \tag{7.2.10}$$

define linear and continuous functionals on X such that Lemma 3.2 (Lax-Milgram) yields the existence of a unique solution $(\tilde{y}, \tilde{p}) \in \mathcal{X}$. Theorem 3.8 yields the $H^2(\Omega)$-regularity, i.e.

$(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$. By estimating (7.2.9), we obtain again by virtue of Theorem 3.8 the $H^2(\Omega)$-a priori estimate

$$\| \tilde{y} \|_{2,\Omega} \leq C \| u_h \|_{0,\Omega} .$$

By means of Assumption 7.4 on the operator R, we can derive the $H^2(\Omega)$-a priori estimate

$$\| \tilde{p} \|_{2,\Omega} \leq C \left(\delta \| y_h \|_{0,\Omega} + \hat{\delta} C_R + \| y_d \|_{0,\Omega} \right) .$$

□

The next result is inspired by Theorem 5.18 resp. Theorem 5.30 and provides auxiliary error estimates for the differences $\tilde{y} - y_h$ and $\tilde{p} - p_h$ in the mesh-dependent norm $\| \cdot \|_h^s$ resp. $\| \cdot \|_h^{ad}$.

Lemma 7.8. *Let $(y_h, u_h) \in X_h \times L^2(\Omega)$ be a solution of (P_h) with a corresponding adjoint solution $p_h \in X_h$. Moreover, let $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the unique solution of (P_{aux}). Then, we have*

$$\| \tilde{y} - y_h \|_h^s \leq C(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\})^{\frac{1}{2}} h |\tilde{y}|_{2,\Omega} + d_h^s(y_h; I_h\tilde{y}, I_h\tilde{y})^{\frac{1}{2}}$$

and

$$\| \tilde{p} - p_h \|_h^{ad} \leq C(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\})^{\frac{1}{2}} h |\tilde{p}|_{2,\Omega} + d_h^{ad}(p_h; I_h\tilde{p}, I_h\tilde{p})^{\frac{1}{2}}.$$

Proof. We illustrate the proof for $\tilde{y} - y_h$. The proof for $\tilde{p} - p_h$ goes along the same lines. We start by defining

$$\tilde{y} - y_h = (\tilde{y} - I_h\tilde{y}) + (I_h\tilde{y} - y_h) =: q_h + e_h.$$

Note that $d_h^s(y_h; q_h, q_h) = 0$. Hence, the standard FE-estimate yields

$$\| q_h \|_h^s \leq C(\varepsilon + \| c \|_{0,\infty,\Omega} h) h |\tilde{y}|_{2,\Omega}. \tag{7.2.11}$$

Similar to Section 5.3.1.2, we are able to derive

$$\begin{aligned}
(\| e_h \|_h^s)^2 &\leq a(e_h, e_h) + d_h^s(y_h; e_h, e_h) \\
&= a(I_h\tilde{y}, e_h) + d_h^s(y_h; I_h\tilde{y}, e_h) - \overbrace{(a(y_h, e_h) + d_h^s(y_h; y_h, e_h))}^{(u_h, e_h)_\Omega} \\
&= a(I_h\tilde{y}, e_h) + d_h^s(y_h; I_h\tilde{y}, e_h) - a(\tilde{y}, e_h) \\
&= a(I_h\tilde{y} - \tilde{y}, e_h) + d_h^s(y_h; I_h\tilde{y}, e_h)
\end{aligned} \tag{7.2.12}$$

where we have used the state equation of (7.2.5) and the auxiliary state equation of (7.2.6). The first term on the right hand side of (7.2.12) can be estimated by

$$a(I_h\tilde{y} - \tilde{y}, e_h) \leq C(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|_{0,\infty,\Omega}^2 + \| c \|_{0,\infty,\Omega}^2 h^2\})^{\frac{1}{2}} h |\tilde{y}|_{2,\Omega} \| e_h \|_h^s . \tag{7.2.13}$$

Due to the fact that $d_h^s(y_h; \cdot, \cdot)$ is a symmetric, positive semidefinite bilinear form, the Cauchy-Schwarz inequality yields

$$\begin{aligned}
d_h^s(y_h; I_h\tilde{y}, e_h) &\leq d_h^s(y_h; e_h, e_h)^{\frac{1}{2}} d_h^s(y_h; I_h\tilde{y}, I_h\tilde{y})^{\frac{1}{2}} \\
&\leq d_h^s(y_h; I_h\tilde{y}, I_h\tilde{y})^{\frac{1}{2}} \| e_h \|_h^s .
\end{aligned} \tag{7.2.14}$$

Collecting (7.2.11), (7.2.12), (7.2.13), and (7.2.14) we obtain the desired result. □

Corollary 7.9. *Let $(y_h, u_h) \in X_h \times L^2(\Omega)$ be a solution of (P_h) with a corresponding adjoint solution $p_h \in X_h$. Moreover, let $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the unique solution of (P_{aux}). Then, we have in the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$*

$$\| \tilde{y} - y_h \|_h^s \leq Ch \| \tilde{y} \|_{2,\Omega} + Ch^{\frac{1}{2}} |I_h \tilde{y}|_{1,\Omega}$$

and

$$\| \tilde{p} - p_h \|_h^{ad} \leq Ch \| \tilde{p} \|_{2,\Omega} + Ch^{\frac{1}{2}} |I_h \tilde{p}|_{1,\Omega}.$$

Proof. According to Lemma 5.19 resp. Lemma 5.31, the consistency errors $d_h^s(y_h; I_h\tilde{y}, I_h\tilde{y})$ resp. $d_h^{ad}(p_h; I_h\tilde{p}, I_h\tilde{p})$ can be estimated by

$$d_h^s(y_h; I_h\tilde{y}, I_h\tilde{y}) \leq C(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} h + \| c \|_{0,\infty,\Omega} h^2) |I_h\tilde{y}|_{1,\Omega}^2$$

resp.

$$d_h^{ad}(p_h; I_h\tilde{p}, I_h\tilde{p}) \leq C(\varepsilon + \| \mathbf{b} \|_{0,\infty,\Omega} h + \| c \|_{0,\infty,\Omega} h^2) |I_h\tilde{p}|_{1,\Omega}^2.$$

The combination of the above consistency error estimates with Lemma 7.8 leads us to the desired result. □

7.2.1.2 Auxiliary result for $\mathcal{G} = \Gamma$

In this section, we derive an auxiliary error estimate for the case $\mathcal{G} = \Gamma$. According to Section 7.2, we set $X_h = V_h \subseteq H^1(\Omega) = X$ and $\mathcal{X}_h = V_h \times V_h \subset \mathcal{X} = H^1(\Omega) \times H^1(\Omega)$ where $\mathcal{X}$ is equipped with the norm

$$\| (y, p) \|_{\mathcal{X}} = \sqrt{\| y \|_{1,\Omega}^2 + \| p \|_{1,\Omega}^2}.$$

The finite dimensional space $\mathcal{X}_h$ is equipped with the norm

$$\| \cdot \|_{\mathcal{X}_h} := \| \cdot \|_{\mathcal{X}} .$$

Throughout this section we consider the operators:

- $K_\Gamma^s \in \mathcal{L}(X, X^*)$ defined by

$$\langle K_\Gamma^s y, v \rangle_{X^*,X} := a_\Gamma(y, v)$$

- $(K_\Gamma^s)^* = K_\Gamma^{ad} \in \mathcal{L}(X, X^*)$ defined by

$$\langle K_\Gamma^{ad} p, \psi \rangle_{X^*,X} := a_\Gamma(\psi, p)$$

- $G \in X^*$ defined by

$$\langle G, \psi \rangle_{X^*,X} := (-y_d, \psi)_\Omega$$

- $D_{(\cdot)}^{s,\Gamma} : X_h \to \mathcal{L}(X_h, X_h^*)$ defined by

$$w_h \mapsto D_{w_h}^{s,\Gamma} \in \mathcal{L}(X_h, X_h^*)$$

  and

$$\langle D_{w_h}^{s,\Gamma} y_h, v_h \rangle_{X_h^*,X_h} := d_h^{s,\Gamma}(w_h; y_h, v_h)$$

- $D^{ad,\Gamma}_{(\cdot)} : X_h \to \mathcal{L}(X_h, X_h^*)$ defined by

$$w_h \mapsto D^{ad,\Gamma}_{w_h} \in \mathcal{L}(X_h, X_h^*)$$

and

$$\langle D^{ad,\Gamma}_{w_h} p_h, \psi_h \rangle_{X_h^*, X_h} := d_h^{ad,\Gamma}(w_h; p_h, \psi_h)$$

Now, we prove that (P_{aux}) possesses a solution $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$. Regarding Section 3.2.1 (Theorem 3.16), we assume $u_h \in H^{\frac{1}{2}}(\Gamma)$. Note that the higher regularity of u_h is sufficient to obtain the $H^2(\Omega)$-regularity of $\tilde{y}$. Thus, Theorem 2.17 yields $(\tilde{y}, \tilde{p}) \in C(\bar{\Omega}) \times C(\bar{\Omega})$.

Lemma 7.10. *Let $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ be a solution of (P_h) with a corresponding adjoint solution $p_h \in X_h$. Then, there exists a unique solution $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$ for (P_{aux}). Moreover, the following $H^2(\Omega)$-a priori estimates hold:*

$$\| \tilde{y} \|_{2,\Omega} \leq C \| u_h \|_{\frac{1}{2},\Gamma} \tag{7.2.15}$$

$$\| \tilde{p} \|_{2,\Omega} \leq C \left(\delta \| y_h \|_{0,\Omega} + \hat{\delta} C_R + \| y_d \|_{0,\Omega} \right). \tag{7.2.16}$$

Proof. The operators K^s_Γ and K^{ad}_Γ are defined by the coercive and continuous bilinear form $a_\Gamma(\cdot,\cdot)$ on $\mathcal{X}$ (see also Lemma 3.10). Moreover, the right hand sides of (P_{aux}), i.e.

$$X \ni v \mapsto (u_h, v)_\Gamma$$

resp.

$$X \ni \psi \mapsto (\delta \cdot y_h - \hat{\delta} \cdot R y_h, \psi)_\Omega - (y_d, \psi)_\Omega$$

define linear and continuous functionals on $H^1(\Omega)$ such that Lemma 3.2 (Lax-Milgram) yields the existence of a unique solution $(\tilde{y}, \tilde{p}) \in \mathcal{X}$. Theorem 3.16 yields the $H^2(\Omega)$-regularity, i.e. $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$. Moreover, Theorem 3.16 implies

$$\| \tilde{y} \|_{2,\Omega} \leq C \| u_h \|_{\frac{1}{2},\Gamma}$$

and by virtue of Assumption 7.4, we get

$$\| \tilde{p} \|_{2,\Omega} \leq C \left(\delta \| y_h \|_{0,\Omega} + \hat{\delta} C_R + \| y_d \|_{0,\Omega} \right).$$

□

Similar to Lemma 7.8, we can state the next result. The proof goes in the same way.

Lemma 7.11. *Let $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ be a solution of (P_h) with a corresponding discrete adjoint solution $p_h \in X_h$. Moreover, let $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$ be the solution of (P_{aux}). Then, we have*

$$\| \tilde{y} - y_h \|_h^{s,\Gamma} \leq C(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|^2_{0,\infty,\Omega} + \| c \|^2_{0,\infty,\Omega} h^2\})^{\frac{1}{2}} h |\tilde{y}|_{2,\Omega} + d_h^{s,\Gamma}(y_h; I_h\tilde{y}, I_h\tilde{y})^{\frac{1}{2}}$$

and

$$\| \tilde{p} - p_h \|_h^{ad,\Gamma} \leq C(\varepsilon + c_0^{-1}\{\| \mathbf{b} \|^2_{0,\infty,\Omega} + \| c \|^2_{0,\infty,\Omega} h^2\})^{\frac{1}{2}} h |\tilde{p}|_{2,\Omega} + d_h^{ad,\Gamma}(p_h; I_h\tilde{p}, I_h\tilde{p})^{\frac{1}{2}}.$$

Similar to Corollary 7.9, the combination of Lemma 5.26 resp. Lemma 5.35 and Lemma 7.11 yields the following auxiliary estimates for the convection-dominated case.

Corollary 7.12. *Let $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ be a solution of (P_h) with a corresponding discrete adjoint solution $p_h \in X_h$. Moreover, let $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$ be the solution of (P_{aux}). Then, we have in the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$*

$$\| \tilde{y} - y_h \|_h^{s,\Gamma} \leq Ch \| \tilde{y} \|_{2,\Omega} + Ch^{\frac{1}{2}} |I_h \tilde{y}|_{1,\Omega}$$

and

$$\| \tilde{p} - p_h \|_h^{ad,\Gamma} \leq Ch \| \tilde{p} \|_{2,\Omega} + Ch^{\frac{1}{2}} |I_h \tilde{p}|_{1,\Omega}.$$

Remark 7.13. *Regarding Lemma 7.8 resp. Lemma 7.11, we can see that the auxiliary error estimates depend on $\tilde{y}$ resp. $\tilde{p}$. Lemma 7.7 and Lemma 7.10 provide $H^2(\Omega)$-a priori estimates for $(\tilde{y}, \tilde{p})$ which depend on y_h resp. p_h. Hence, the auxiliary error estimates are currently not useful for the derivation of error estimates for $y - y_h$ and $u - u_h$ (see Lemma 7.6). However, in Section 8 we will see that for a specific setting of the problem (P) resp. (P_h) the discrete solutions y_h, p_h are uniformly bounded in the $L^2(\Omega)$-norm. In the case of Dirichlet boundary condition the control u_h is bounded in the $L^2(\Omega)$-norm. In the case of Robin boundary condition, we verify the $H^{\frac{1}{2}}(\Gamma)$-regularity of u_h and the uniform boundedness in the $H^{\frac{1}{2}}(\Gamma)$-norm. Therefore, the $H^2(\Omega)$-a priori estimates of $\tilde{y}, \tilde{p}$ become helpful and the auxiliary error estimates (Corollary 7.9/Corollary 7.12) as well.*

8 Application on optimal control problems

In this section, we apply the abstract results derived in Section 7 on the discrete systems (P_h^f), (P_h^b), $(P_{s,h}^{MY})$, and (P_h^Γ). First, we verify the existence of discrete solutions corresponding to (P_h^f), (P_h^b), and $(P_{s,h}^{MY})$. In the case of Dirichlet boundary conditions $[(P_h^f), (P_h^b), (P_{s,h}^{MY})]$, the Finite Element framework is different as in the case of Robin boundary condition (P_h^Γ). Hence, we discuss the discrete system (P_h^Γ) separately. According to Section 7.1, for the investigation of (P_h^f), (P_h^b), and $(P_{s,h}^{MY})$ we set $X_h = V_{h,0} \subseteq H_0^1(\Omega) = X$ and $\mathcal{X}_h = V_{h,0} \times V_{h,0} \subset \mathcal{X} = H_0^1(\Omega) \times H_0^1(\Omega)$ where $\mathcal{X}$ is equipped with the norm

$$\| (y, p) \|_{\mathcal{X}} = \sqrt{|y|_{1,\Omega}^2 + |p|_{1,\Omega}^2}.$$

The norm on the finite dimensional space $\mathcal{X}_h$ is given by

$$\| \cdot \|_{\mathcal{X}_h} := \| \cdot \|_{\mathcal{X}} .$$

8.1 Unconstrained case

Let us start with the coupled and discretized system, introduced in Section 6.1 which corresponds to the unconstrained optimal control problem (P_f). The system is given by

$$\left.\begin{array}{l} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \\ a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \\ \lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \end{array}\right\} (P_h^f)$$

8.1.1 Existence

By means of Lemma 7.3, we start with the verification of the existence of a discrete solution. For the application of Lemma 7.3, an intuitive strategy is to add the adjoint equation of (P_h^f) to the state equation. Then, after rearranging the terms, the operator K_h can be defined by the sum of the left hand sides of (P_h^f) and the operator $Q+G$ can be defined by the sum of the right hand sides. However, following this strategy the operator $Q+G$ does not satisfy condition (7.1.2) so that Lemma 7.3 is not applicable. Hence, we modify this strategy by considering an equivalent regularized system of (P_h^f) so that the sufficient conditions (7.1.1) and (7.1.2) for the application of Lemma 7.3 hold.

8.1.1.1 Regularized system

We regularize the coupled and discretized system (P_h^f) in the following way:

$$\left.\begin{array}{l} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \\ \frac{1}{\lambda} a(\psi_h, p_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, \psi_h) = \frac{1}{\lambda}(y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \\ \lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \end{array}\right\} (P_h^{f,rg})$$

Definition 8.1. *A pair $(y_h, u_h) \in \mathcal{X}_h$ is called solution for $(P_h^{f,rg})$ if there exists a discrete adjoint solution $p_h \in X_h$ such that*

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \tag{8.1.1}$$

$$\frac{1}{\lambda} a(\psi_h, p_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, \psi_h) = \frac{1}{\lambda}(y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \tag{8.1.2}$$

$$\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \tag{8.1.3}$$

is satisfied.

Due to the fact that (P_h^f) and $(P_h^{f,rg})$ are equivalent systems, a solution of $(P_h^{f,rg})$ is also a solution of (P_h^f) and vice versa. Hence, the strategy is to apply the theory of Section 7.1 on $(P_h^{f,rg})$ so that we are able to verify the existence of a discrete solution for (P_h^f). The regularized system can be transferred to the following equivalent operator equation: Find $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that

$$(K + D_h)\,\mathbf{x}_h = Q\mathbf{x}_h + G \quad \text{in } \mathcal{X}_h^* \tag{8.1.4}$$

where for $\mathbf{z}_h = (v_h, \psi_h) \in \mathcal{X}_h$

- $\langle K\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := \langle K^s y_h, v_h\rangle_{X^*,X} + \langle \frac{1}{\lambda} K^{ad} p_h, \psi_h\rangle_{X^*,X}$
- $\langle D_h\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := \langle D_{y_h}^s y_h, v_h\rangle_{X_h^*,X_h} + \langle \frac{1}{\lambda} D_{p_h}^{ad} p_h, \psi_h\rangle_{X_h^*,X_h}$
- $\langle Q\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := (-\frac{1}{\lambda} p_h, v_h)_\Omega + (\frac{1}{\lambda} y_h, \psi_h)_\Omega$
- $\langle G, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := -\frac{1}{\lambda}(y_d, \psi_h)_\Omega$.

Note that the operators have been introduced in Section 7.2.1.1.

Lemma 8.2. *The operator $K_h := (K + D_h) : \mathcal{X}_h \to \mathcal{X}_h^*$ satisfies Assumption 7.1 and the operator $\tilde{Q} := Q + G$ defined by*

$$\langle \tilde{Q}\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := \langle Q\mathbf{x}_h + G, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h}$$

satisfies Assumption 7.2.

Proof. First, the continuity of the limiters (see Assumption 5.9) implies the continuity of D_h. Regarding Section 7.2.1.1, the continuity of K and $\tilde{Q}$ is obvious. Moreover, we have to prove that $K_h := (K + D_h)$ satisfies

$$\langle K_h \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \geq C_K \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}^2 \quad \forall\, \mathbf{x}_h \in \mathcal{X}_h \tag{8.1.5}$$

and

$$\langle \tilde{Q} \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \leq C_Q \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h} \quad \forall\, \mathbf{x}_h \in \mathcal{X}_h \tag{8.1.6}$$

where $C_K, C_Q > 0$ are constants. We start with (8.1.5). For this, let $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ be arbitrary. The coercivity of $a(\cdot,\cdot)$ and $d_h^s(y_h; y_h, y_h) \geq 0$ resp. $d_h^{ad}(p_h; p_h, p_h) \geq 0$ imply

$$\begin{aligned}
\langle K_h \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} &= a(y_h, y_h) + \frac{1}{\lambda} a(p_h, p_h) + d_h^s(y_h; y_h, y_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, p_h) \\
&\geq \min\{\varepsilon, \frac{\varepsilon}{\lambda}\}(|y_h|_{1,\Omega}^2 + |p_h|_{1,\Omega}^2) \\
&= \min\{\varepsilon, \frac{\varepsilon}{\lambda}\} \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}^2 .
\end{aligned}$$

For the verification of the second condition we have

$$\begin{aligned}
\langle \tilde{Q} \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} &= \langle Q \mathbf{x}_h + G, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \\
&= (-\frac{1}{\lambda} p_h, y_h)_\Omega + (\frac{1}{\lambda} y_h, p_h)_\Omega - \frac{1}{\lambda}(y_d, p_h)_\Omega \\
&= -\frac{1}{\lambda}(y_d, p_h)_\Omega \\
&\leq \frac{\parallel y_d \parallel_{0,\Omega}}{\lambda} \parallel p_h \parallel_{0,\Omega} \\
&= \frac{\parallel y_d \parallel_{0,\Omega}}{\lambda} \sqrt{\parallel p_h \parallel_{0,\Omega}^2} \\
&\leq \frac{\parallel y_d \parallel_{0,\Omega}}{\lambda} \sqrt{\parallel y_h \parallel_{0,\Omega}^2 + \parallel p_h \parallel_{0,\Omega}^2} \\
&\leq C_p \frac{\parallel y_d \parallel_{0,\Omega}}{\lambda} \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}
\end{aligned}$$

where $C_p > 0$ is the Poincaré constant (see Lemma 2.11). □

Lemma 8.3. *There exists a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ for the operator equation* (8.1.4).

Proof. According to Lemma 8.2, the operators K_h and $Q+G$ satisfy the assumptions of Lemma 7.3. Consequently, we obtain a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that the operator equation (8.1.4) holds. □

Corollary 8.4. *There exists a solution $(y_h, u_h) \in \mathcal{X}_h$ with a corresponding adjoint solution $p_h \in X_h$ for the regularized system $(P_h^{f,rg})$.*

Proof. First, Lemma 8.3 yields the existence of a discrete solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ for the operator equation (8.1.4). Due to (8.1.3), we can define a discrete control by $u_h = -\frac{1}{\lambda} p_h$. Now, we consider the variational formulation of (8.1.4), i.e.

$$\langle (K + D_h)\, \mathbf{x}_h, \mathbf{z}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} = \langle Q \mathbf{x}_h + G, \mathbf{z}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \quad \forall\, \mathbf{z}_h \in \mathcal{X}_h. \tag{8.1.7}$$

Testing (8.1.7) with $\mathbf{z}_h^1 = (v_h, 0), \mathbf{z}_h^2 = (0, \psi_h) \in \mathcal{X}_h$ where $v_h, \psi_h \in X_h$ are arbitrary, we can conclude that according to Definition 8.1, the pair $(y_h, -\frac{1}{\lambda} p_h) \in \mathcal{X}_h$ is a solution of $(P_h^{f,rg})$ with a corresponding adjoint solution $p_h \in X_h$. □

As we have mentioned in Remark 7.13, we are also interested in $L^2(\Omega)$-norm estimates of the discrete solutions.

Lemma 8.5. *Let $(y_h, u_h) \in \mathcal{X}_h$ be a solution of $(P_h^{f,rg})$ with a corresponding adjoint solution $p_h \in X_h$. Then, we have the following $L^2(\Omega)$-norm estimates*

$$\| u_h \|_{0,\Omega} \leq C_c \tag{8.1.8}$$
$$\| y_h \|_{0,\Omega} \leq C_s \tag{8.1.9}$$
$$\| p_h \|_{0,\Omega} \leq C_{ad} \tag{8.1.10}$$

where $C_c, C_s, C_{ad} > 0$ are constants, independent of h.

Proof. We add the regularized state equation (8.1.1) to the adjoint equation (8.1.2) with respect to (8.1.3) and obtain for all $(v_h, \psi_h) \in \mathcal{X}_h$

$$a(y_h, v_h) + \frac{1}{\lambda} a(\psi_h, p_h) + d_h^s(y_h; y_h, v_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, \psi_h) = (-\frac{1}{\lambda} p_h, v_h)_\Omega + \frac{1}{\lambda}(y_h - y_d, \psi_h)_\Omega.$$

Now, we set $v_h = y_h$ and $\psi_h = p_h$ such that we obtain in combination with the coercivity of $a(\cdot,\cdot)$ and $d_h^s(y_h; y_h, y_h) \geq 0$ resp. $d_h^{ad}(p_h; p_h, p_h) \geq 0$

$$c_0 \| y_h \|_{0,\Omega}^2 + \frac{c_0}{\lambda} \| p_h \|_{0,\Omega}^2 \leq -\frac{1}{\lambda}(y_d, p_h)_\Omega. \tag{8.1.11}$$

Thus, we obtain

$$\| p_h \|_{0,\Omega} \leq \frac{1}{c_0} \| y_d \|_{0,\Omega} =: C_{ad}$$

and

$$\| u_h \|_{0,\Omega} = \| -\frac{1}{\lambda} p_h \|_{0,\Omega} \leq \frac{1}{\lambda c_0} \| y_d \|_{0,\Omega} =: C_c.$$

By means of (8.1.11), we get for a discrete state solution

$$\| y_h \|_{0,\Omega} \leq \sqrt{\frac{1}{\lambda c_0} \| y_d \|_{0,\Omega} C_{ad}} =: C_s.$$

□

The equivalence of the systems $(P_h^{f,rg})$ and (P_h^f) implies that $(y_h, u_h) = (y_h, -\frac{1}{\lambda} p_h) \in \mathcal{X}_h$ is also a solution of (P_h^f). Hence, we can state the next result.

Corollary 8.6. *A solution $(y_h, u_h) \in \mathcal{X}_h$ for $(P_h^{f,rg})$ is also a solution for (P_h^f) and the corresponding $L^2(\Omega)$-norm estimates* (8.1.8) - (8.1.10) *are valid.*

8.1.2 $L^2(\Omega)$-error estimates

In the following, we derive $L^2(\Omega)$-error estimates for a discrete solution $(y_h, u_h) \in \mathcal{X}_h$ of (P_h^f) to the solution $(\bar{y}, \bar{u}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ of (P_f). Regarding Section 7.2, we consider the following specific setting:

Assumption 8.7. *We assume that*

- $\mathcal{G} = \Omega$

- $U = L^2(\Omega)$
- $\delta = 1$ *and* $\hat{\delta} = 0$
- $Z : C(\bar{\Omega}) \to U$ *is defined by*

$$Zw := -\frac{1}{\lambda} w \quad a.e. \ in \ \Omega.$$

Due to Assumption 8.7, the discretized system (P_h^f) coincides with the general system (P_h) resp. (P_f) coincides with (P). Note that the variational inequality

$$(\lambda u_h + p_h, u - u_h)_\Omega \geq 0 \quad \forall u \in U$$

is satisfied by $\lambda u_h + p_h = 0$ a.e. in Ω resp.

$$(\lambda \bar{u} + \bar{p}, u - \bar{u})_\Omega \geq 0 \quad \forall u \in U$$

is satisfied by $\lambda \bar{u} + \bar{p} = 0$ a.e. in Ω. Note that we have in the unconstrained case $R = 0$ resp. $\hat{\delta} = 0$ such that Assumption 7.4 holds. Now, by virtue of Lemma 7.6 we can derive the following $L^2(\Omega)$-error estimate. For this, we remark that the Sobolev embedding theorem yields $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$.

Lemma 8.8. *Let $(\bar{y}, \bar{u}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_f) with the corresponding adjoint solution $\bar{p} \in X \cap H^2(\Omega)$. Moreover, we have the solution $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ of (P_{aux}) and a solution $(y_h, u_h) \in \mathcal{X}_h$ of (P_h^f) where $p_h \in X_h$ is the corresponding discrete adjoint solution. Then, we have*

$$\frac{\lambda}{2} \parallel u_h - \bar{u} \parallel_{0,\Omega}^2 + \frac{1}{2} \parallel y_h - \bar{y} \parallel_{0,\Omega}^2 \leq C \parallel p_h - \tilde{p} \parallel_{0,\Omega}^2 + \frac{1}{2} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2 .$$

Proof. Lemma 7.6 yields for general $\delta, \hat{\delta} \in \mathbb{R}_{\geq 0}$ with $\delta - \hat{\delta} > 0$ the estimate

$$\frac{\lambda}{2} \parallel u_h - \bar{u} \parallel_{0,\Omega}^2 + \frac{(\delta - \hat{\delta})}{2} \parallel y_h - \bar{y} \parallel_{0,\Omega}^2 \leq C \parallel p_h - \tilde{p} \parallel_{0,\Omega}^2 + \frac{\delta^2}{2(\delta - \hat{\delta})} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2 . \quad (8.1.12)$$

Setting $\delta = 1$ and $\hat{\delta} = 0$ in (8.1.12) leads us to the desired $L^2(\Omega)$-norm error estimate. □

According to Corollary 7.9 and Remark 7.13, the next result shows that the auxiliary solutions $\tilde{y}, \tilde{p}$ are uniformly bounded in the $H^2(\Omega)$-norm.

Lemma 8.9. *Let $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_{aux}). Moreover, let $(y_h, u_h) \in \mathcal{X}_h$ be a solution of (P_h^f) where $p_h \in X_h$ is the corresponding discrete adjoint solution. Then, we have*

$$\parallel \tilde{y} \parallel_{2,\Omega} \leq C_1 \quad (8.1.13)$$

$$\parallel \tilde{p} \parallel_{2,\Omega} \leq C_2 \quad (8.1.14)$$

where $C_1, C_2 > 0$ are constants, independent of h.

Proof. Lemma 7.7 yields the $H^2(\Omega)$-a priori estimates

$$\parallel \tilde{y} \parallel_{2,\Omega} \leq C \parallel u_h \parallel_{0,\Omega}$$

and

$$\parallel \tilde{p} \parallel_{2,\Omega} \leq C \left(\parallel y_h \parallel_{0,\Omega} + \parallel y_d \parallel_{0,\Omega} \right) .$$

By virtue of Lemma 8.5, we get the desired result. □

Due to our huge interest in the convection-dominated case, we combine Corollary 7.9, Lemma 8.8, and Lemma 8.9 to obtain the following $L^2(\Omega)$-error estimates corresponding to the control and the state solution.

Theorem 8.10. *Let $(\bar{y}, \bar{u}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_f) and $(y_h, u_h) \in \mathcal{X}_h$ be a solution of (P_h^f). Then, we have in the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$*

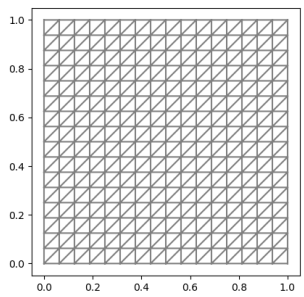
$$\| u_h - \bar{u} \|_{0,\Omega} + \| y_h - \bar{y} \|_{0,\Omega} \leq C h^{\frac{1}{2}}$$

where $C > 0$ is a constant, independent of h.

8.1.3 Numerical results

This section is dedicated to show the results of our numerical tests. On the one hand, for given data of a test problem, we compute the state resp. the adjoint solution by the Galerkin method. Here, we will see the occurrence of node-to-node oscillations. On the other hand, we have applied the AFC method for the computation of stabilized discrete solutions, i.e. solutions without oscillations. According to the derived $L^2(\Omega)$-error estimates, we show the computed $L^2(\Omega)$-convergence rates for the state solution and the adjoint solution. We remark that in the following sections, all numerical tests have been performed on a unit square mesh $\Omega = [0,1] \times [0,1]$. According to the FEM, in Figure 1 we can see the several refinement levels of the unit square mesh Ω. For the implementation of the AFC stabilization technique, we have used the continuous and linearity-preserving BJK limiter, introduced in Section 5.5.2. Due to the fact that the BJK limiter is nonlinear and non-differentiable, we use an iterative method for solving the coupled and discretized systems. In the last years, many regularized versions of the BJK limiter have been developed such that Newton methods can be applied on the stabilized equations. For detailed information, we refer for instance to [BadBon17] or [LohSP19]. In our numerical investigations, the coupled and discretized systems have been transferred to a relaxed preconditioned Richardson iteration (see [LohSP19, Section 6.2.3]). The method has been implemented in Python 3. Moreover, we have used the FEniCS package to build up the stiffness matrices of the bilinear forms of the state equation resp. to build up the stiffness matrices of the bilinear form of the adjoint equation. In all numerical tests we set

- $\varepsilon = 0.001$ diffusion coefficient
- $\mathbf{b} = (1,1)$ convection field
- $c = 1$ reaction coefficient
- $\lambda = 0.5$ Tikhonov parameter
- tolerance for the residual $= 10^{-10}$.



Grid level	mesh	mesh size
1	$(4,4)$	3.5355e-01
2	$(8,8)$	1.7678e-01
3	$(16,16)$	8.8388e-02
4	$(32,32)$	4.4194e-02
5	$(64,64)$	2.2097e-02
6	$(128,128)$	1.1049e-02
7	$(256,256)$	5.5243e-03
8	$(512,512)$	2.7621e-03

Figure 1: $(16,16)$-mesh (l.h.s.) and grid levels

8.1.3.1 Test problem

According to Section 6.1, we consider the following unconstrained optimal control problem

$$\left.\begin{array}{l} \min \ \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + cy = u + f \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \end{array}\right\} (P_f^{test})$$

where $f \in L^2(\Omega)$ is for the first time an arbitrary function which will be specified later in Section 8.1.3.2. The corresponding optimality system is given by

$$\begin{array}{ll} -\varepsilon \Delta \bar{y} + \mathbf{b} \cdot \nabla \bar{y} + c\bar{y} = \bar{u} + f \quad \text{in } \Omega & -\varepsilon \Delta \bar{p} - \mathbf{b} \cdot \nabla \bar{p} + c\bar{p} = \bar{y} - y_d \quad \text{in } \Omega \\ \bar{y} = 0 \quad \text{on } \Gamma & \bar{p} = 0 \quad \text{on } \Gamma \end{array}$$

$$\lambda \bar{u} + \bar{p} = 0 \quad \text{a.e. in } \Omega.$$

8.1.3.2 Analytical solutions

To review that the iterative method computes accurate AFC solutions, the right hand sides f, y_d have been adjusted such that the test problem (P_f^{test}) possesses the following analytical optimal state solution

$$\bar{y}(x_1, x_2) = \left(x_1 - \frac{e^{(-(1-x_1)/0.01)} - e^{(-1/0.01)}}{(1 - e^{(-1/0.01)})}\right) \cdot \left(x_2 - \frac{e^{(-(1-x_2)/0.01)} - e^{(-1/0.01)}}{(1 - e^{(-1/0.01)})}\right)$$

and the analytical optimal adjoint solution

$$\bar{p}(x_1, x_2) = \left((1 - x_1) - \frac{e^{(-x_1/0.01)} - e^{(-1/0.01)}}{(1 - e^{(-1/0.01)})}\right) \cdot \left((1 - x_2) - \frac{e^{(-x_2/0.01)} - e^{(-1/0.01)}}{(1 - e^{(-1/0.01)})}\right).$$

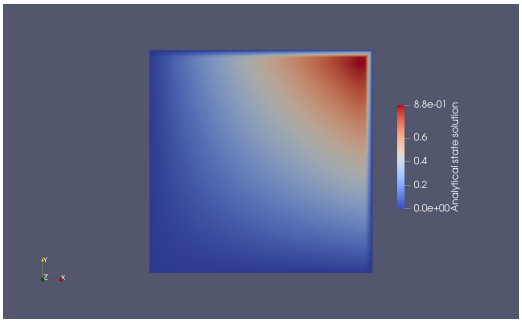


Figure 2: Analytical state solution $\bar{y}$

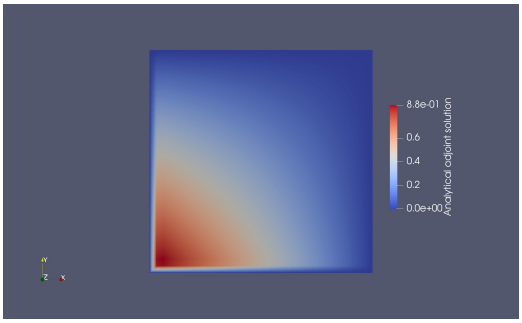


Figure 3: Analytical adjoint solution $\bar{p}$

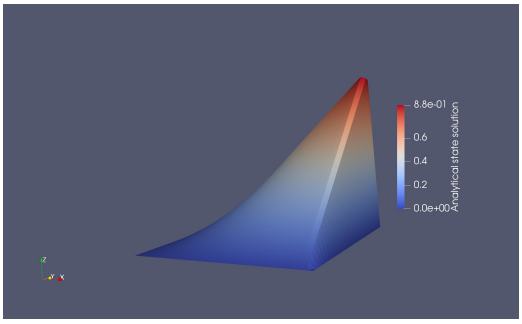


Figure 4: Analytical state solution $\bar{y}$

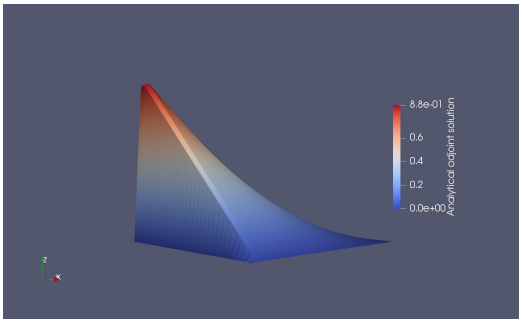


Figure 5: Analytical adjoint solution $\bar{p}$

8.1.3.3 Galerkin solutions

Below, we plot the Galerkin solutions computed on a $(32, 32)$-unit square mesh. In Figure 6/Figure 8 and Figure 7/Figure 9 we can see that the Galerkin solutions possess node-to-node oscillations. A reason for such a behavior is that sufficient conditions for the discrete maximum principle are not satisfied by the stiffness matrices corresponding to the state and the adjoint equation. Moreover, standard FEM cannot resolve the narrow region of the layers. As we have mentioned in the previous sections, we prevent the occurrance of oscillations by the satisfaction of the DMP property resp. by the application of the AFC method. Hence, in the next sections we show the results of our numerical test corresponding to the applied AFC method.

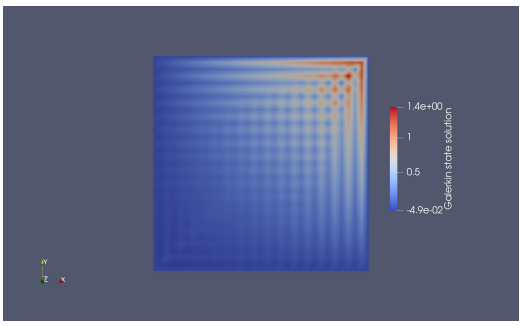


Figure 6: Galerkin state solution

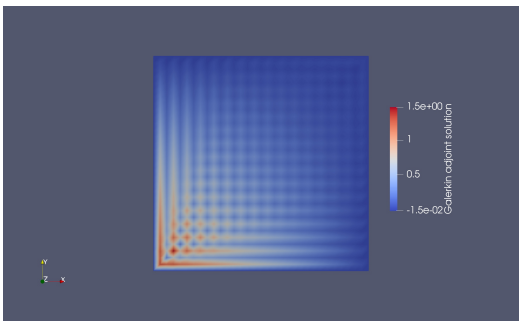


Figure 7: Galerkin adjoint solution

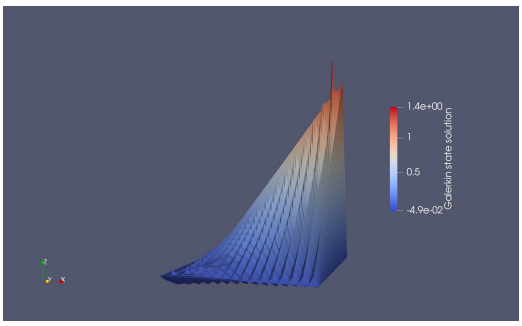


Figure 8: Galerkin state solution

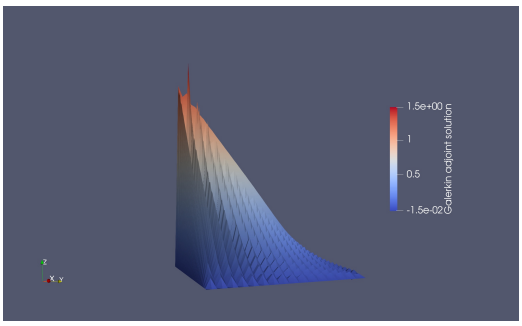


Figure 9: Galerkin adjoint solution

8.1.3.4 Experimental order of convergence

According to Theorem 8.10, we will review $L^2(\Omega)$-convergence rates for the state and the adjoint solution. For this, we compute the AFC state and the AFC adjoint solution with respect to the adjusted right hand sides f, y_d. For a given mesh size h we compute the $L^2(\Omega)$-errors of

$$e_h := \bar{y} - y_h$$
$$k_h := \bar{p} - p_h$$

where $\bar{y}, \bar{p}$ are the analytical solutions and y_h, p_h the corresponding stabilized and discrete solutions computed by the AFC method. The orders of convergence are calculated by the experimental order of convergence (EOC). For this, let $z \in X$ be a continuous solution and $z_h, z_{h'} \in X_h$ are the corresponding Finite Element solutions for mesh sizes $0 < h \neq h' < 1$. Then, for a given norm $\| \cdot \|$ the EOC is calculated by

$$\text{EOC} = \text{EOC}_{\|\cdot\|}(h, h') = \frac{\ln(\| z - z_h \|) - \ln(\| z - z_{h'} \|)}{\ln(h) - \ln(h')}.$$

In Table 4 we see $L^2(\Omega)$-convergence rates for $e_h = \bar{y} - y_h$ and $k_h = \bar{p} - p_h$. Since we have $\bar{u} = -\frac{1}{\lambda}\bar{p}$ and $u_h = -\frac{1}{\lambda}p_h$, for a comparison of the theoretical and the numerical results corresponding to the controls, it sufficies to compare the AFC adjoint solution p_h and the analytical adjoint solution $\bar{p}$. The iteration for the first grid level $l = 1$ is started with the zero-state and the zero-adjoint solution. For a grid level $l \in \{2, \cdots, 7\}$ the start solutions of the iteration are the interpolated AFC solutions of the previous grid level $l - 1$.

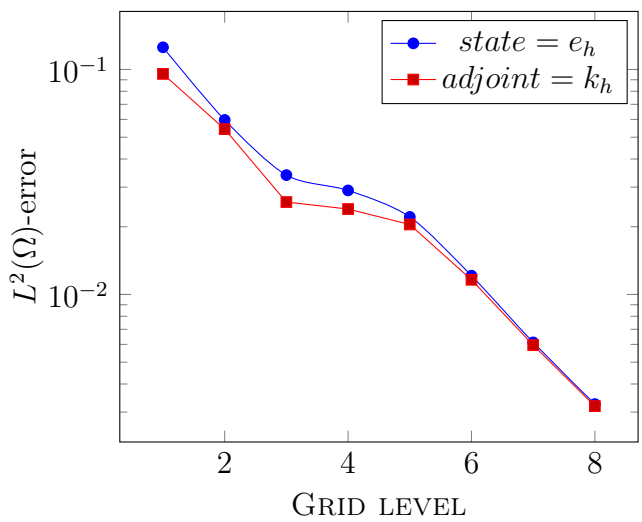
Table 4: $L^2(\Omega)$-errors / EOC (P_f^{test})

Grid level	$c_0 \parallel e_h \parallel_{0,\Omega}$	EOC $c_0 \parallel e_h \parallel_{0,\Omega}$	$c_0 \parallel k_h \parallel_{0,\Omega}$	EOC $c_0 \parallel k_h \parallel_{0,\Omega}$
1	1.2548e-01	-	9.5584e-02	-
2	5.9655e-02	1.07	5.4306e-02	0.82
3	3.3926e-02	0.81	2.5750e-02	1.08
4	2.9008e-02	0.23	2.3975e-02	0.1
5	2.2116e-02	0.39	2.0466e-02	0.23
6	1.2117e-02	0.87	1.1621e-02	0.82
7	6.1237e-03	0.98	5.9555e-03	0.96
8	3.2566e-03	0.91	3.1901e-03	0.90

Remark 8.11. *As we can see in Table 4, the EOC of the adjoint and the state solution are nearly $\mathcal{O}(1)$. This result does not confirm the theoretical order of $\mathcal{O}(\frac{1}{2})$ provided in Theorem 8.10. A reason for the higher order of convergence in our numerical test could be that in the estimation of the consistency errors (see Lemma 5.19/Lemma 5.31) the factors $(1 - \alpha_{ij})$ have been estimated too roughly by 1. Additionally, the limiters α_{ij} depend nonlinear on the structure of a solution. In Section 5.5.2, we have shortly discussed the question in which way the structure of a solution influences the order of convergence, in the context of the linearity-preserving property of the limiter. For this, recall the numerical tests in [BJK17] where it is shown that the linearity-preserving property leads to higher orders of convergence in the tests than the theoretical results predict.*

Figure 10 illustrates the behavior of the $L^2(\Omega)$-errors e_h and k_h. We can see that the $L^2(\Omega)$-errors decrease asymptotically while the grid level increases resp. the mesh size h decreases.

Figure 10: $L^2(\Omega)$-errors behavior (P_f^{test})



8.1.3.5 AFC solutions

In the following, we plot the AFC solutions computed on a $(32, 32)$-unit square mesh. We can see that the AFC solutions are free of oscillations. Moreover, a comparison of the layers between the AFC solutions (Figure 13/Figure 14) and the analytical solutions (Figure 4/Figure 5) leads us to the conclusion that the iterative method computes accurate discrete solutions. Finally, we can state that the AFC method is a remedy to prevent oscillations in the discrete solutions corresponding to (P_f^{test}).

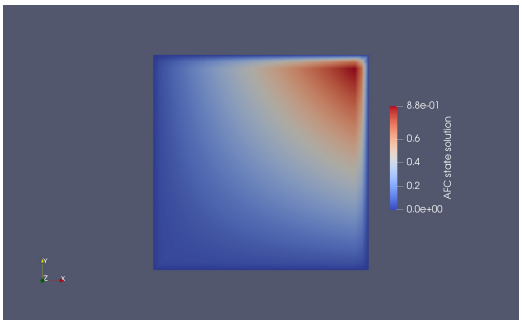


Figure 11: AFC state solution

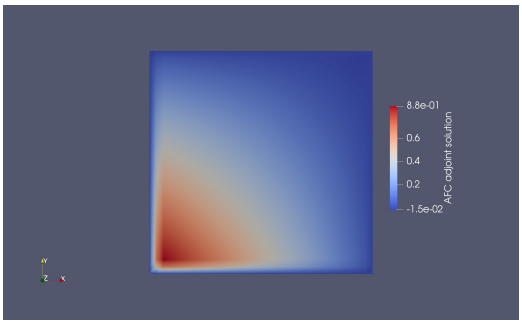


Figure 12: AFC adjoint solution

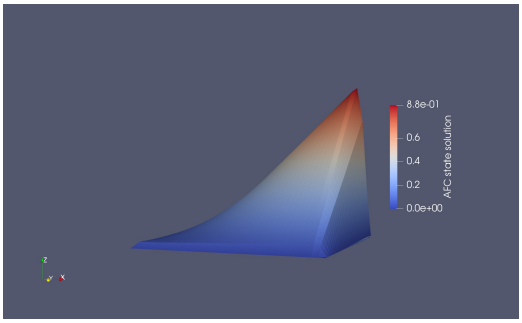


Figure 13: AFC state solution

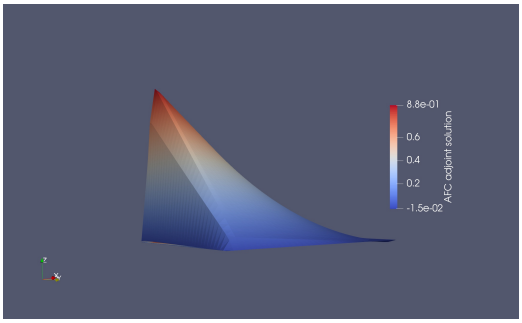


Figure 14: AFC adjoint solution

8.2 Control constrained case

In this section, we investigate the coupled and discretized system corresponding to the control constrained optimal control problem (P_b). Regarding Section 6.2, we consider the system

$$\left.\begin{array}{l} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \\ a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \\ u_h = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h) \quad \text{a.e. in } \Omega \end{array}\right\} (P_h^b)$$

8.2.1 Existence

As in the unconstrained case, the existence of a discrete solution for (P_h^b) cannot be verified directly since the uniform boundedness of y_h in the $L^2(\Omega)$-norm cannot be ensured from the above system. Following the strategy in the unconstrained case (see Section 8.1.1.1), the operator $Q + G : \mathcal{X}_h \to \mathcal{X}_h^*$ defined by

$$\langle (Q+G)\mathbf{x}_h, \mathbf{z}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} := (\mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h), v_h)_\Omega + (y_h - y_d, \psi_h)_\Omega$$

where $\mathbf{x}_h = (y_h, p_h)$ and $\mathbf{z}_h = (v_h, \psi_h)$ does not satisfy condition (7.1.2) in Assumption 7.2. Hence, as in Section 8.1.1.1 we regularize (P_h^b) so that we are able to apply Lemma 7.3 for the verification of the existence of a discrete solution.

8.2.1.1 Regularized system

Let $k \in \mathbb{N}$ be arbitrary. Instead of (P_h^b), we consider the following truncated, coupled and discretized system

$$\left.\begin{array}{l} a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \\ a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (\Psi_k(y_h) - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \\ u_h = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h) \quad \text{a.e. in } \Omega \end{array}\right\} (P_h^{b,k})$$

where Ψ_k is defined by (2.2.1).

Definition 8.12. *Let $k \in \mathbb{N}$. A pair $(y_h, u_h) \in X_h \times L^2(\Omega)$ is called solution for $(P_h^{b,k})$ if there exists a discrete adjoint solution $p_h \in X_h$ such that*

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \tag{8.2.1}$$

$$a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (\Psi_k(y_h) - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \tag{8.2.2}$$

$$u_h = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h) \quad \text{a.e. in } \Omega \tag{8.2.3}$$

is satisfied.

Now, the strategy is to solve the truncated system $(P_h^{b,k})$ for an arbitrary $k \in \mathbb{N}$. After that, we will verify that a discrete state solution $y_h \in V_{h,0}$ of $(P_h^{b,k})$ satisfies

$$\Psi_k(y_h) = y_h \quad \text{a.e. in } \Omega$$

for a $k \in \mathbb{N}$ which is large enough. First, the system (8.2.1)-(8.2.3) can be transferred to the following operator equation: Find $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that

$$(K + D_h)\,\mathbf{x}_h = Q_k \mathbf{x}_h + G \quad \text{in } \mathcal{X}_h^* \tag{8.2.4}$$

where for $\mathbf{z}_h = (v_h, \psi_h) \in \mathcal{X}_h$

- $\langle K\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := \langle K^s y_h, v_h\rangle_{X^*,X} + \langle K^{ad} p_h, \psi_h\rangle_{X^*,X}$
- $\langle D_h\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := \langle D_{y_h}^s y_h, v_h\rangle_{X_h^*,X_h} + \langle D_{p_h}^{ad} p_h, \psi_h\rangle_{X_h^*,X_h}$
- $\langle Q_k\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := (\mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h), v_h)_\Omega + (\Psi_k(y_h), \psi_h)_\Omega$
- $\langle G, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := (-y_d, \psi_h)_\Omega$.

Lemma 8.13. *Let $k \in \mathbb{N}$ be arbitrary. The operator $K_h := (K + D_h) : \mathcal{X}_h \to \mathcal{X}_h^*$ satisfies Assumption 7.1 and $\tilde{Q}_k := Q_k + G$ defined by*

$$\langle \tilde{Q}_k\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} = \langle Q_k\mathbf{x}_h + G, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h}$$

satisfies Assumption 7.2.

Proof. First, for the continuity of the operator K_h we refer to Lemma 8.2. Due to the Lipschitz continuity of the projection formula $\mathbb{P}_{[u_a,u_b]}(\cdot)$ and $\Psi_k(\cdot)$ we obtain the continuity of $\tilde{Q}_k := Q_k + G$. The verification of condition (7.1.1) in Assumption 7.1 goes in the same way as in Lemma 8.2. The proof of condition (7.1.2) in Assumption 7.2, i.e.

$$\langle \tilde{Q}_k \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \leq C_Q \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h} \quad \forall\, \mathbf{x}_h \in \mathcal{X}_h \tag{8.2.5}$$

where $C_Q > 0$ is slightly different. For this, let $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ be arbitrary. Then, we have

$$\begin{aligned}
\langle \tilde{Q}_k \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} &= \langle Q_k \mathbf{x}_h + G, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \\
&= (\mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h), y_h)_\Omega + (\Psi_k(y_h), p_h)_\Omega - (y_d, p_h)_\Omega \\
&\leq C \parallel y_h \parallel_{0,\Omega} + C(k) \parallel p_h \parallel_{0,\Omega} + \parallel y_d \parallel_{0,\Omega} \parallel p_h \parallel_{0,\Omega} \\
&\leq C_k \left(\sqrt{\parallel y_h \parallel_{0,\Omega}^2} + \sqrt{\parallel p_h \parallel_{0,\Omega}^2} \right) \\
&\leq \sqrt{2} C_k \sqrt{\parallel y_h \parallel_{0,\Omega}^2 + \parallel p_h \parallel_{0,\Omega}^2} \\
&\leq \sqrt{2} C_k C_p \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}
\end{aligned}$$

where $C_p > 0$ is the Poincaré constant and $C_k > 0$ a constant which depends on k. □

Lemma 8.14. *Let $k \in \mathbb{N}$ be arbitrary. Then, there exists a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ for the operator equation* (8.2.4).

Proof. According to Lemma 8.13, the sufficient conditions for the application of Lemma 7.3 are fulfilled by the operators K_h and $\tilde{Q}_k := Q_k + G$. Consequently, we obtain a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that the operator equation (8.2.4) holds. □

Following the lines of Corollary 8.4, one can easily prove the next result which is a direct consequence of Lemma 8.14.

Corollary 8.15. *Let $k \in \mathbb{N}$ be arbitrary. Then, there exists a solution $(y_h, u_h) \in X_h \times L^2(\Omega)$ with a corresponding adjoint solution $p_h \in X_h$ for the regularized and discretized system $(P_h^{b,k})$.*

Note that the control u_h is given by $u_h = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h)$. Now, we prove for a specific $k \in \mathbb{N}$ that a solution $(y_h, u_h) \in X_h \times L^2(\Omega)$ of $(P_h^{b,k})$ also solves (P_h^b). For the right choice of $k \in \mathbb{N}$, we consider the following a priori $L^\infty(\Omega)$-estimate corresponding to a state solution $y_h \in X_h$ of (8.2.1).

Lemma 8.16. *Let $y_h \in X_h$ be a solution of the state equation* (8.2.1). *Then, we have*

$$\parallel y_h \parallel_{0,\infty,\Omega} \leq C(1 + |\ln(h)|)^{\frac{1}{2}} \frac{C(u_a, u_b, \Omega)}{\min\{\varepsilon, c_0\}}.$$

Proof. Setting $v_h = y_h$ in (8.2.1), the coercivity of $a(\cdot,\cdot)$ and the positivity of $d_h^s(y_h; y_h, y_h)$ imply

$$\begin{aligned}
\min\{\varepsilon, c_0\} \parallel y_h \parallel_{1,\Omega}^2 &\leq (\mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h), y_h)_\Omega \\
&\leq C(u_a, u_b, \Omega) \parallel y_h \parallel_{0,\Omega} \\
&\leq C(u_a, u_b, \Omega) \parallel y_h \parallel_{1,\Omega}
\end{aligned}$$

and consequently

$$\parallel y_h \parallel_{1,\Omega} \leq \frac{C(u_a, u_b, \Omega)}{\min\{\varepsilon, c_0\}}.$$

The discrete Sobolev inequality (Lemma 5.3) yields the desired result. □

In the following, we set

$$k = k(h) := \lceil C(1 + |\ln(h)|)^{\frac{1}{2}} \frac{C(u_a, u_b, \Omega)}{\min\{\varepsilon, c_0\}} \rceil \tag{8.2.6}$$

where $\lceil r \rceil$ denotes the usual ceiling function.

Corollary 8.17. *Let $k \in \mathbb{N}$ be defined by (8.2.6). Then, a solution $(y_h, u_h) \in X_h \times L^2(\Omega)$ of $(P_h^{b,k})$ with a corresponding adjoint solution $p_h \in X_h$ is also a solution for (P_h^b).*

Proof. Lemma 8.16 yields that a discrete state solution $y_h \in X_h$ satisfies

$$\| y_h \|_{0,\infty,\Omega} \leq C(1 + |\ln(h)|)^{\frac{1}{2}} \frac{C(u_a, u_b, \Omega)}{\min\{\varepsilon, c_0\}} \leq k$$

such that $\Psi_k(y_h) = y_h$ a.e. in Ω. Consequently, the regularized system $(P_h^{b,k})$ corresponds to the discretized system (P_h^b). Hence, a solution $(y_h, u_h) \in X_h \times L^2(\Omega)$ of $(P_h^{b,k})$ with a discrete adjoint solution $p_h \in X_h$ solves (P_h^b). □

Remark 8.18. *In the special case that the control constraints are nonnegative, i.e. $u_a, u_b \geq 0$ a.e. in Ω resp. nonpositive, i.e. $u_a, u_b \leq 0$ a.e. in Ω, the global discrete maximum principle (Remark 5.17) implies that the discrete state solution y_h of (P_h^b) satisfies*

$$\begin{aligned} u_a, u_b \leq 0 \text{ in } \Omega \;&\Rightarrow\; \max_{\bar{\Omega}} y_h \leq \max_{\partial\Omega} y_h^+ \\ u_a, u_b \geq 0 \text{ in } \Omega \;&\Rightarrow\; \min_{\bar{\Omega}} y_h \geq \min_{\partial\Omega} y_h^-. \end{aligned}$$

Lemma 8.19. *Let $(y_h, u_h) \in X_h \times L^2(\Omega)$ be a solution of (P_h^b) with a corresponding adjoint solution $p_h \in X_h$. Then, we have the following $L^2(\Omega)$-norm estimates*

$$\| u_h \|_{0,\Omega} \leq C_c \tag{8.2.7}$$

$$\| y_h \|_{0,\Omega} \leq C_s \tag{8.2.8}$$

$$\| p_h \|_{0,\Omega} \leq C_{ad} \tag{8.2.9}$$

where $C_c, C_s, C_{ad} > 0$ are constants, independent of h.

Proof. The boundedness of the control u_h in the $L^2(\Omega)$-norm follows directly by definition, i.e.

$$\| u_h \|_{0,\Omega} = \| \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda} p_h) \|_{0,\Omega} =: C_c. \tag{8.2.10}$$

By virtue of (8.2.1), we are able to derive for $y_h \in X_h$

$$\begin{aligned} c_0 \| y_h \|_{0,\Omega}^2 &\leq a(y_h, y_h) + d_h^s(y_h; y_h, y_h) \\ &= (u_h, y_h)_\Omega \\ &\leq \| u_h \|_{0,\Omega} \| y_h \|_{0,\Omega} \\ &\leq C_c \| y_h \|_{0,\Omega} . \end{aligned}$$

Thus, we have

$$\| y_h \|_{0,\Omega} \leq \frac{C_c}{c_0} =: C_s. \tag{8.2.11}$$

By means of (8.2.2) and Corollary 8.17 we obtain for the discrete adjoint solution p_h

$$\begin{aligned} c_0 \parallel p_h \parallel_{0,\Omega}^2 &\leq a(p_h, p_h) + d_h^{ad}(p_h; p_h, p_h) \\ &= (y_h - y_d, p_h)_\Omega \\ &\leq (\parallel y_h \parallel_{0,\Omega} + \parallel y_d \parallel_{0,\Omega}) \parallel p_h \parallel_{0,\Omega} . \end{aligned}$$

Due to (8.2.11), we get

$$c_0 \parallel p_h \parallel_{0,\Omega} \leq C_s + \parallel y_d \parallel_{0,\Omega} \tag{8.2.12}$$

and consequently

$$\parallel p_h \parallel_{0,\Omega} \leq \frac{C_s + \parallel y_d \parallel_{0,\Omega}}{c_0} =: C_{ad}.$$

□

8.2.2 $L^2(\Omega)$-error estimates

In this section, we derive error estimates in the $L^2(\Omega)$-norm for a solution $(y_h, u_h) \in X_h \times L^2(\Omega)$ of (P_h^b) to the corresponding solution $(\bar{y}, \bar{u}) \in (X \cap H^2(\Omega)) \times L^2(\Omega)$ of (P_b). Regarding Section 7.2, we have in the control constrained case the following assumptions:

Assumption 8.20. *We assume that*

- $\mathcal{G} = \Omega$
- $U := U_{ad} = \{u \in L^2(\Omega) : u_a(x) \leq u(x) \leq u_b(x) \quad a.e. \text{ in } \Omega\}$.
- $\delta = 1$ *and* $\hat{\delta} = 0$
- $Z : C(\bar{\Omega}) \to U$ *is defined by*

$$Zw := \mathbb{P}_{[u_a, u_b]}\left(-\frac{1}{\lambda} w\right) \quad a.e. \text{ in } \Omega.$$

Due to Assumption 8.20, the discretized system (P_h^b) coincides with the general system (P_h) and (P_b) coincides with (P). The general L^2-error estimate (see Lemma 7.6) leads us to the next result. Note that the proof goes along the same lines as in Lemma 8.8.

Lemma 8.21. *Let $(\bar{y}, \bar{u}) \in (X \cap H^2(\Omega)) \times L^2(\Omega)$ be the solution of (P_b) with a corresponding adjoint solution $\bar{p} \in X \cap H^2(\Omega)$. Moreover, we have the solution $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ of (P_{aux}) and a solution $(y_h, u_h) \in X_h \times L^2(\Omega)$ of (P_h^b) where $p_h \in X_h$ is the corresponding discrete adjoint solution. Then, we have*

$$\frac{\lambda}{2} \parallel u_h - \bar{u} \parallel_{0,\Omega}^2 + \frac{1}{2} \parallel y_h - \bar{y} \parallel_{0,\Omega}^2 \leq C \parallel p_h - \tilde{p} \parallel_{0,\Omega}^2 + \frac{1}{2} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2 .$$

The combination of Lemma 7.7 and Lemma 8.19 yields the uniform boundedness of the auxiliary solutions $\tilde{y}, \tilde{p}$ in the $H^2(\Omega)$-norm.

Lemma 8.22. *Let $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_{aux}). Moreover, let $(y_h, u_h) \in X_h \times L^2(\Omega)$ be a solution of (P_h^b) where $p_h \in X_h$ is the corresponding discrete adjoint solution. Then, we have*

$$\parallel \tilde{y} \parallel_{2,\Omega} \leq C_1 \tag{8.2.13}$$

$$\parallel \tilde{p} \parallel_{2,\Omega} \leq C_2 \tag{8.2.14}$$

where $C_1, C_2 > 0$ are constants, independent of h.

The combination of Corollary 7.9, Lemma 8.21, and Lemma 8.22 leads us in the convection-dominated case to the next result.

Theorem 8.23. *Let $(\bar{y}, \bar{u}) \in (X \cap H^2(\Omega)) \times L^2(\Omega)$ be the solution of (P_b) and $(y_h, u_h) \in X_h \times L^2(\Omega)$ be a solution of (P_h^b). Then, we have in the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$*

$$\| u_h - \bar{u} \|_{0,\Omega} + \| y_h - \bar{y} \|_{0,\Omega} \leq C h^{\frac{1}{2}}$$

where $C > 0$ is a constant which depends on the data of the problem.

8.2.3 Numerical results

In this section, we see the results of our numerical tests concerned with the application of the AFC method for discretizing the control constrained optimal control problem (P_b). For this, the test problem has been constructed such that the analytical solutions $\bar{y}, \bar{p}$ introduced in Section 8.1.3.2 are the optimal solutions of the test problem. Moreover, we use the same iterative solver as in the unconstrained case. The following test problem has been solved on a $[0,1] \times [0,1]$ unit square mesh. For the stabilization of the discrete solutions we have used the BJK limiter, introduced in Section 5.5.2. For a general introduction to the numerics, i.e. data of the convection-diffusion reaction equation, Tikhonov parameter, grid levels etc. we refer to Section 8.1.3.

8.2.3.1 Test problem

According to Section 6.2, we consider the following optimal control problem

$$\left.\begin{array}{l} \min \ \frac{1}{2} \| y - y_d \|_{0,\Omega}^2 + \frac{\lambda}{2} \| u \|_{0,\Omega}^2 \\ -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + c y = u + f \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ u \in U_{ad} \end{array}\right\} (P_b^{test}) \tag{8.2.15}$$

where $f \in L^2(\Omega)$ will be defined later in this section. The set of admissible controls is given by

$$U_{ad} := \{ u \in L^2(\Omega) : \ u_a(x) \leq u(x) \leq u_b(x) \quad \text{a.e. in } \Omega \}.$$

In the numerical test, we use the following control constraints

- $u_a = -1$
- $u_b = 1$.

Recall that for (P_b^{test}) the optimality system is given by

$$\begin{array}{ll} -\varepsilon \Delta \bar{y} + \mathbf{b} \cdot \nabla \bar{y} + c \bar{y} = \bar{u} + f \quad \text{in } \Omega & -\varepsilon \Delta \bar{p} - \mathbf{b} \cdot \nabla \bar{p} + c \bar{p} = \bar{y} - y_d \quad \text{in } \Omega \\ \bar{y} = 0 \quad \text{on } \Gamma & \bar{p} = 0 \quad \text{on } \Gamma \end{array}$$

$$(\lambda \bar{u} + \bar{p}, u - \bar{u})_\Omega \geq 0 \quad \forall u \in U_{ad}.$$

The functions f, y_d have been adjusted in the same way as for (P_f^{test}). The analytical control corresponding to the variational inequality in the optimality system above can be expressed by $\bar{u} = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda}\bar{p})$. Since the projection formula is Lipschitz continuous, the $L^2(\Omega)$-convergence rates of the adjoint solution can be transferred to obtain $L^2(\Omega)$-convergence rates for the control $u_h = \mathbb{P}_{[u_a,u_b]}(-\frac{1}{\lambda}p_h)$. Moreover, since the discrete AFC adjoint solution has no oscillations, the AFC control has no oscillations as well. The computed AFC solutions are the same as illustrated in Figure 13 and Figure 14.

8.2.3.2 Experimental order of convergence

In the following, we see $L^2(\Omega)$-convergence rates for $e_h = \bar{y} - y_h$ and $k_h = \bar{p} - p_h$ where y_h, p_h are the computed AFC solutions for a given mesh size h. As we can see in Table 5, the orders of convergence are quite similar to the orders in the unconstrained case (see Section 8.1.3.4). This observation confirms the theoretical results derived in Theorem 8.10 and Theorem 8.23 which predict a similar convergence behavior of the discrete solutions in these two cases. For an explanation referring to the higher convergence rates in our numerical test than provided in the theoretical results we refer to Remark 8.11.

Table 5: $L^2(\Omega)$-errors / EOC (P_b^{test})

Grid level	$c_0 \parallel e_h \parallel_{0,\Omega}$	EOC $c_0 \parallel e_h \parallel_{0,\Omega}$	$c_0 \parallel k_h \parallel_{0,\Omega}$	EOC $c_0 \parallel k_h \parallel_{0,\Omega}$
1	1.1846e-01	-	9.4863e-02	-
2	5.6450e-02	1.07	5.4000e-02	0.81
3	3.0850e-02	0.87	2.5383e-02	1.09
4	2.7093e-02	0.19	2.4073e-02	0.08
5	2.1647e-02	0.32	2.0554e-02	0.23
6	1.2056e-02	0.84	1.1638e-02	0.82
7	6.1042e-03	0.98	5.9624e-03	0.96
8	3.2474e-03	0.91	3.1941e-03	0.90

8.3 State constrained case - Moreau-Yosida regularization

In this section, we analyze the discretized system ($P_{s,h}^{MY}$) corresponding to the Moreau-Yosida regularization (P_s^{MY}). As we have mentioned in Section 6.3 a direct application of the *optimize-then-discretize*-approach on the optimality system of (P_s) (see Theorem 4.12) is currently not possible. Hence, we have discretized the Moreau-Yosida regularization (P_s^{MY}) by the AFC method. Following the *optimize-then-discretize*-approach, we have got the coupled and discretized system

$$\left.\begin{aligned}
&a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \quad \forall v_h \in X_h \\
&a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega + \left(\mu_h^\delta, \psi_h\right)_\Omega \ \forall \psi_h \in X_h \\
&\mu_h^\delta = \delta \cdot (y_h - \mathbb{P}_{[y_a,y_b]}(y_h)) \quad \text{a.e. in } \Omega \\
&\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega
\end{aligned}\right\} (P_{s,h}^{MY})$$

8.3.1 Existence

Now, we verify the existence of a discrete solution for ($P_{s,h}^{MY}$). Note that the uniform boundedness of y_h in the $L^2(\Omega)$-norm cannot be derived directly from ($P_{s,h}^{MY}$). Hence, as in the unconstrained case resp. in the control constrained case, we regularize ($P_{s,h}^{MY}$). For this, we rewrite the right hand side of the adjoint equation in so far as:

$$(y_h - y_d, \psi_h)_\Omega + \left(\mu_h^\delta, \psi_h\right)_\Omega = (\delta + 1) \cdot (y_h, \psi_h)_\Omega - (y_d, \psi_h)_\Omega - \delta \cdot \left(\mathbb{P}_{[y_a,y_b]}(y_h), \psi_h\right)_\Omega .$$

Then, instead of ($P_{s,h}^{MY}$), we consider for all $(v_h, \psi_h) \in \mathcal{X}_h$ the system

$$a(y_h, v_h) + d_h^s(y_h; y_h, v_h) = (u_h, v_h)_\Omega \tag{8.3.1}$$

$$a(\psi_h, p_h) + d_h^{ad}(p_h; p_h, \psi_h) = (\delta + 1) \cdot (y_h, \psi_h)_\Omega - (y_d, \psi_h)_\Omega - \delta \cdot \left(\mathbb{P}_{[y_a,y_b]}(y_h), \psi_h\right)_\Omega \tag{8.3.2}$$

$$\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega. \tag{8.3.3}$$

8.3.1.1 Regularized system

For the verification of the existence of a discrete solution, we derive again an equivalent and regularized system to (8.3.1)-(8.3.3) such that we are able to apply Lemma 7.3. For this, the state equation (8.3.1) is multiplied with $\delta + 1$ and the adjoint equation (8.3.2) is divided by λ so that we obtain for all $(v_h, \psi_h) \in \mathcal{X}_h$ the following regularized system

$$\left.\begin{array}{l} (\delta+1)\cdot a(y_h, v_h) + (\delta+1)\cdot d_h^s(y_h; y_h, v_h) = (\delta+1)\cdot (u_h, v_h)_\Omega \\ \frac{1}{\lambda} a(\psi_h, p_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, \psi_h) = \frac{(\delta+1)}{\lambda}(y_h, \psi_h)_\Omega - \frac{1}{\lambda}(y_d, \psi_h)_\Omega - \frac{\delta}{\lambda}\left(\mathbb{P}_{[y_a,y_b]}(y_h), \psi_h\right)_\Omega \\ \lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \end{array}\right\} (P_{s,h}^{MY,rg})$$

Definition 8.24. *A pair $(y_h, u_h) \in \mathcal{X}_h$ is called solution for $(P_{s,h}^{MY,rg})$ if there exists a discrete adjoint solution $p_h \in X_h$ such that for all $(v_h, \psi_h) \in \mathcal{X}_h$*

$$(\delta+1)\cdot a(y_h, v_h) + (\delta+1)\cdot d_h^s(y_h; y_h, v_h) = (\delta+1)\cdot (u_h, v_h)_\Omega \tag{8.3.4}$$

$$\frac{1}{\lambda} a(\psi_h, p_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, \psi_h) = \frac{(\delta+1)}{\lambda}(y_h, \psi_h)_\Omega - \frac{1}{\lambda}(y_d, \psi_h)_\Omega - \frac{\delta}{\lambda}\left(\mathbb{P}_{[y_a,y_b]}(y_h), \psi_h\right)_\Omega \tag{8.3.5}$$

$$\lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega \tag{8.3.6}$$

is satisfied.

Now, the system $(P_{s,h}^{MY,rg})$ can be transferred to the following operator equation: Find $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that

$$(K + D_h)\,\mathbf{x}_h = Q\mathbf{x}_h + G \quad \text{in } \mathcal{X}_h^* \tag{8.3.7}$$

where for $\mathbf{z}_h = (v_h, \psi_h) \in \mathcal{X}_h$

- $\langle K\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := (\delta+1)\cdot\langle K^s y_h, v_h\rangle_{X^*,X} + \langle \frac{1}{\lambda} K^{ad} p_h, \psi_h\rangle_{X^*,X}$
- $\langle D_h\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := (\delta+1)\cdot\langle D_{y_h}^s y_h, v_h\rangle_{X_h^*,X_h} + \frac{1}{\lambda}\langle D_{p_h}^{ad} p_h, \psi_h\rangle_{X_h^*,X_h}$
- $\langle Q\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := -\frac{(\delta+1)}{\lambda}\cdot(p_h, v_h)_\Omega + \frac{(\delta+1)}{\lambda}\cdot(y_h, \psi_h)_\Omega - \frac{\delta}{\lambda}\cdot\left(\mathbb{P}_{[y_a,y_b]}(y_h), \psi_h\right)_\Omega$
- $\langle G, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := -\frac{1}{\lambda}(y_d, \psi_h)_\Omega$.

Note that the operators are defined in Section 7.2.1.1.

Lemma 8.25. *The operator $K_h := (K + D_h) : \mathcal{X}_h \to \mathcal{X}_h^*$ satisfies Assumption 7.1 and $\tilde{Q} := Q + G$ defined by*

$$\langle \tilde{Q}\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := \langle Q\mathbf{x}_h + G, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h}$$

satisfies Assumption 7.2.

Proof. First, for the continuity of the operators K_h and $\tilde{Q}$ we refer to Lemma 8.2. Now, we prove the satisfaction of the conditions (7.1.1) and (7.1.2). For this, let $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ be arbitrary. The coercivity of $a(\cdot,\cdot)$ and $d_h^s(y_h; y_h, y_h) \geq 0$ resp. $d_h^{ad}(p_h; p_h, p_h) \geq 0$ imply

$$\begin{aligned} \langle K_h\mathbf{x}_h, \mathbf{x}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} &= (\delta+1)\cdot a(y_h, y_h) + \frac{1}{\lambda} a(p_h, p_h) \\ &\quad + (\delta+1)\cdot d_h^s(y_h; y_h, y_h) + \frac{1}{\lambda} d_h^{ad}(p_h; p_h, p_h) \\ &\geq \min\left\{(\delta+1)\cdot\varepsilon, \frac{\varepsilon}{\lambda}\right\}\left(|y_h|_{1,\Omega}^2 + |p_h|_{1,\Omega}^2\right) \\ &= \min\left\{(\delta+1)\cdot\varepsilon, \frac{\varepsilon}{\lambda}\right\} \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}^2 . \end{aligned}$$

For proving the second condition, we are able to derive

$$\begin{aligned}
\langle \tilde{Q}\mathbf{x}_h, \mathbf{x}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} &= \langle Q\mathbf{x}_h + G, \mathbf{x}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} \\
&= -\frac{(\delta+1)}{\lambda}\,(p_h, y_h)_\Omega + \frac{(\delta+1)}{\lambda}\,(y_h, p_h)_\Omega \\
&- \frac{\delta}{\lambda}\cdot\left(\mathbb{P}_{[y_a,y_b]}(y_h), p_h\right)_\Omega - \frac{1}{\lambda}\,(y_d, p_h)_\Omega \\
&\leq \left(\frac{C\delta + \parallel y_d \parallel_{0,\Omega}}{\lambda}\right) \parallel p_h \parallel_{0,\Omega} \\
&= \left(\frac{C\delta + \parallel y_d \parallel_{0,\Omega}}{\lambda}\right) \sqrt{\parallel p_h \parallel_{0,\Omega}^2} \\
&\leq \left(\frac{C\delta + \parallel y_d \parallel_{0,\Omega}}{\lambda}\right) \sqrt{\parallel y_h \parallel_{0,\Omega}^2 + \parallel p_h \parallel_{0,\Omega}^2} \\
&\leq C_P \left(\frac{C\delta + \parallel y_d \parallel_{0,\Omega}}{\lambda}\right) \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}
\end{aligned}$$

where $C > 0$ is a constant and $C_P > 0$ the Poincaré constant. □

Lemma 8.26. *There exists a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ for the operator equation* (8.3.7).

Proof. According to Lemma 8.25, the operators K_h and $\tilde{Q} = Q + G$ satisfy the sufficient conditions for the application of Lemma 7.3. Consequently, we obtain a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that the operator equation (8.3.7) holds. □

Corollary 8.27. *There exists a solution $(y_h, u_h) \in \mathcal{X}_h$ with a corresponding adjoint solution $p_h \in X_h$ for the system $(P_{s,h}^{MY,rg})$.*

Proof. Lemma 8.26 yields the existence of a discrete solution $(y_h, p_h) \in \mathcal{X}_h$ for the operator equation (8.3.7). According to Definition 8.24, the discrete control can be defined by $u_h = -\frac{1}{\lambda}p_h$ such that $(y_h, u_h) = (y_h, -\frac{1}{\lambda}p_h) \in \mathcal{X}_h$ is a solution for the regularized system $(P_{s,h}^{MY,rg})$. □

Now, we prove the $L^2(\Omega)$-boundedness of a solution (y_h, u_h). As we have mentioned in the previous sections, these estimates are meaningful for the $H^2(\Omega)$-estimates of the auxiliary solutions $\tilde{y}, \tilde{p}$ provided in Lemma 7.7. We will return to this discussion in Section 8.3.2.

Lemma 8.28. *Let $(y_h, u_h) \in \mathcal{X}_h$ be a solution of $(P_{s,h}^{MY,rg})$ with a corresponding adjoint solution $p_h \in X_h$. Then, we have the following $L^2(\Omega)$-norm estimates*

$$\parallel y_h \parallel_{0,\Omega} \leq C_s \tag{8.3.8}$$

$$\parallel p_h \parallel_{0,\Omega} \leq C_{ad} \tag{8.3.9}$$

$$\parallel u_h \parallel_{0,\Omega} \leq C_c \tag{8.3.10}$$

where $C_s, C_{ad}, C_c > 0$ are constants, independent of h.

Proof. We add the regularized state equation (8.3.4) to the adjoint equation (8.3.5) and obtain for all $(v_h, \psi_h) \in \mathcal{X}_h$

$$\begin{aligned}
&(\delta+1)\cdot a(y_h, v_h) + \frac{1}{\lambda}a(\psi_h, p_h) + (\delta+1)\cdot d_h^s(y_h; y_h, v_h) + \frac{1}{\lambda}d_h^{ad}(p_h; p_h, \psi_h) \\
&\quad = -\frac{(\delta+1)}{\lambda}\cdot(p_h, v_h)_\Omega + \frac{(\delta+1)}{\lambda}\cdot(y_h, \psi_h)_\Omega - \frac{1}{\lambda}(y_d, \psi_h)_\Omega - \frac{\delta}{\lambda}\cdot\left(\mathbb{P}_{[y_a,y_b]}(y_h), \psi_h\right)_\Omega .
\end{aligned}$$

Setting $v_h = y_h$ and $\psi_h = p_h$, we get in combination with the coercivity of $a(\cdot,\cdot)$ and $d_h^s(y_h; y_h, y_h) \geq 0$ resp. $d_h^{ad}(p_h; p_h, p_h) \geq 0$

$$(\delta+1)c_0 \parallel y_h \parallel_{0,\Omega}^2 + \frac{c_0}{\lambda} \parallel p_h \parallel_{0,\Omega}^2 \leq -\frac{1}{\lambda}(y_d, p_h)_\Omega - \frac{\delta}{\lambda}\left(\mathbb{P}_{[y_a,y_b]}(y_h), p_h\right)_\Omega . \tag{8.3.11}$$

Thus, we obtain with the pointwise boundedness of $\mathbb{P}_{[y_a,y_b]}(\cdot)$

$$\parallel p_h \parallel_{0,\Omega} \leq \frac{1}{c_0} \parallel y_d \parallel_{0,\Omega} + \frac{\delta}{c_0} C(y_a, y_b, \Omega) =: C_{ad} \tag{8.3.12}$$

where $C(y_a, y_b, \Omega) > 0$ is a constant, independent of h. By means of (8.3.11) and (8.3.12), we are able to derive for a discrete state solution $y_h \in X_h$

$$\parallel y_h \parallel_{0,\Omega} \leq \sqrt{\frac{\left(\frac{1}{\lambda} \parallel y_d \parallel_{0,\Omega} + \frac{\delta}{\lambda} C(y_a, y_b, \Omega)\right) C_{ad}}{(\delta+1)c_0}} =: C_s.$$

The $L^2(\Omega)$-boundedness of $u_h = -\frac{1}{\lambda} p_h$ is a direct consequence of (8.3.12). □

Finally, by virtue of Corollary 8.27, we obtain the existence of a discrete solution for $(P_{s,h}^{MY})$ with the corresponding $L^2(\Omega)$-norm estimates.

Corollary 8.29. *A solution $(y_h, u_h) \in \mathcal{X}_h$ of $(P_{s,h}^{MY,rg})$ is also a solution for $(P_{s,h}^{MY})$. Furthermore, the $L^2(\Omega)$-estimates* (8.3.8)-(8.3.10) *are valid.*

8.3.2 $L^2(\Omega)$-error estimates

In this section, we derive $L^2(\Omega)$-error estimates for the differences $\bar{y}_\delta - y_h$ and $\bar{u}_\delta - u_h$ where $(y_h, u_h) \in \mathcal{X}_h$ is a discrete solution of $(P_{s,h}^{MY})$ and $(\bar{y}_\delta, \bar{u}_\delta) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ is the optimal solution of (P_s^{MY}), with the help of Lemma 7.6. Finally, we will return to the state constrained optimal control problem (P_s) and show that a solution $(y_h, u_h) \in \mathcal{X}_h$ of $(P_{s,h}^{MY})$ is also an approximation for the solution $(\bar{y}, \bar{u})$ of (P_s), with respect to the data of the problem and the regularization parameters.

8.3.2.1 Moreau-Yosida regularization

According to Section 7.2, the discretized system $(P_{s,h}^{MY})$ coincides to the general system (P_h) and (P_s^{MY}) coincides to (P) when the following assumptions hold:

Assumption 8.30. *We assume that*

- $\mathcal{G} = \Omega$
- $U = L^2(\Omega)$
- $\delta := \delta + 1$ *and* $\hat{\delta} := \delta$
- *The operator* $Z : C(\bar{\Omega}) \to U$ *is given by*

$$Zw := -\frac{1}{\lambda} w \quad a.e. \text{ in } \Omega$$

- *The operator* $R : C(\bar{\Omega}) \to C(\bar{\Omega})$ *is defined by*

$$R(z) := \mathbb{P}_{[y_a,y_b]}(z)$$

where $\mathbb{P}_{[y_a,y_b]}(\cdot)$ *is given by Definition 2.24.*

Note that the operator $R : C(\bar{\Omega}) \to C(\bar{\Omega})$ is Lipschitz continuous in the sense that for $z, \tilde{z} \in C(\bar{\Omega})$ the condition

$$(Rz - R\tilde{z}, v)_\Omega \leq \| z - \tilde{z} \|_{0,\Omega} \| v \|_{0,\Omega} \quad \forall v \in C(\bar{\Omega})$$

holds. Moreover, the operator R fulfills the condition

$$\| Rz \|_{0,\Omega} \leq C_R := C(y_a, y_b, \Omega).$$

Consequently, with respect to Assumption 8.30, the sufficient conditions for the application of Lemma 7.6 are satisfied so that the general $L^2(\Omega)$-error estimate (see Lemma 7.6) leads us to the next result.

Lemma 8.31. *Let $(\bar{y}_\delta, \bar{u}_\delta) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_s^{MY}) with the corresponding adjoint solution $\bar{p}_\delta \in X \cap H^2(\Omega)$. Moreover, we have the solution $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ of (P_{aux}) and a solution $(y_h, u_h) \in \mathcal{X}_h$ of $(P_{s,h}^{MY})$ with a corresponding discrete adjoint solution $p_h \in X_h$. Then, it holds*

$$\frac{\lambda}{2} \| u_h - \bar{u}_\delta \|_{0,\Omega}^2 + \frac{1}{2} \| y_h - \bar{y}_\delta \|_{0,\Omega}^2 \leq C \| p_h - \tilde{p} \|_{0,\Omega}^2 + \delta C_R \| y_h - \tilde{y} \|_{0,\Omega} + \frac{(\delta+1)^2}{2} \| y_h - \tilde{y} \|_{0,\Omega}^2$$

where $C_R = C(y_a, y_b, \Omega) > 0$ and $C > 0$ are constants, independent of h.

For the boundedness of the auxiliary solutions $\tilde{y}, \tilde{p}$ in the $H^2(\Omega)$-norm, we use Lemma 7.7 and Lemma 8.28 so that we can state the next result.

Lemma 8.32. *Let $(y_h, u_h) \in \mathcal{X}_h$ be a solution of $(P_{s,h}^{MY})$ where $p_h \in X_h$ is a corresponding discrete adjoint solution. Moreover, let $(\tilde{y}, \tilde{p}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_{aux}). Then, we have*

$$\| \tilde{y} \|_{2,\Omega} \leq C_1 \tag{8.3.13}$$

$$\| \tilde{p} \|_{2,\Omega} \leq C_2 \tag{8.3.14}$$

where $C_1, C_2 > 0$ are constants, independent of h.

The combination of Corollary 7.9, Lemma 8.31 and Lemma 8.32 yields in the convection-dominated case the following $L^2(\Omega)$-error estimate.

Theorem 8.33. *Let $(\bar{y}_\delta, \bar{u}_\delta) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_s^{MY}) and $(y_h, u_h) \in \mathcal{X}_h$ be a solution of $(P_{s,h}^{MY})$. Then, we have in the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$*

$$\| u_h - \bar{u}_\delta \|_{0,\Omega} + \| y_h - \bar{y}_\delta \|_{0,\Omega} \leq C_\delta C_R h^{\frac{1}{4}} + C_\delta h^{\frac{1}{2}}$$

where $C_\delta, C_R > 0$ are constants, independent of h.

It is worth to mention that the constant $C_\delta > 0$ depends linear on the regularization parameter $\delta > 0$ (see Lemma 8.31).

8.3.2.2 State constrained case

Now, we return to the state constrained optimal control problem (P_s). As we have mentioned, the Moreau-Yosida regularization of (P_s) has been only a tool to construct a discrete solution which approximates the unique solution $(\bar{y}, \bar{u})$ of (P_s). The next result shows that a discrete solution of the regularized problem $(P_{s,h}^{MY})$ is also an approximation for the solution of (P_s). However, as we can see in Theorem 8.33, the accuracy of the approximation in the $L^2(\Omega)$-norm depends on the choice of the regularization parameter $\delta > 0$. For this, recall also the error analysis in Section 4.3.3 where the $L^2(\Omega)$-errors of $\bar{y} - \bar{y}_\delta$ and $\bar{u} - \bar{u}_\delta$ have been analyzed. Now, the combination of Corollary 4.32 and Theorem 8.33 yields the following result.

Theorem 8.34. *Let $(\bar{y}, \bar{u}) \in \mathcal{X} \cap (H^2(\Omega) \times H^2(\Omega))$ be the solution of (P_s) and $(y_h, u_h) \in \mathcal{X}_h$ be a solution of $(P_{s,h}^{MY})$. Then, we have for $\delta > 0$ and $0 < \gamma < 1$ in the convection-dominated case, i.e. $\varepsilon \ll \| \mathbf{b} \|_{0,\infty,\Omega} h$*

$$\| u_h - \bar{u} \|_{0,\Omega} + \| y_h - \bar{y} \|_{0,\Omega} \leq C_\delta C_R h^{\frac{1}{4}} + C_\delta h^{\frac{1}{2}} + C \left(\frac{1}{\sqrt{\delta}} \right)^{\frac{\gamma}{2\gamma+1}}$$

where $C, C_\delta, C_R > 0$ are constants, independent of h.

Remark 8.35. *We remark that an optimal parameter adjustment between the mesh size h and the regularization parameter $\delta > 0$ is currently not possible. Moreover, since $\frac{1}{\varepsilon}$ resp. $\delta > 0$ arise in the constants C, C_δ, it is worth to mention that the error estimate provided in Theorem 8.34 should be regarded for fixed $\varepsilon > 0$ and $\delta > 0$.*

8.3.3 Numerical results

In this section, we show the results of our numerical tests referring to the application of the AFC method to the Moreau-Yosida regularization (P_s^{MY}). The iterative solver is the same as in the unconstrained case resp. in the control constrained case. The following test problems have been solved on a unit square mesh $\Omega = [0,1] \times [0,1]$ with the BJK limiter introduced in Section 5.5.2. For a general introduction to the numerics, i.e. data of the convection-diffusion reaction equation, Tikhonov parameter, grid levels etc., we refer to Section 8.1.3.

8.3.3.1 Test problem

The basis of our numerical investigations is the following state constrained optimal control problem

$$\left.\begin{array}{l} \min \ J^s(y,u) := \frac{1}{2} \| y - y_d \|_{0,\Omega}^2 + \frac{\lambda}{2} \| u \|_{0,\Omega}^2 \\ -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + cy = u + f \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ y_a \leq y \leq y_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_s^{test})$$

In all tests we have considered $y_d = 4$ and $f = 1$. Moreover, we have the following state constraints

$$\begin{aligned} y_a &= -0.5 \\ y_b &= 0.7. \end{aligned}$$

As we have seen in Section 4.3.2, the solution $(\bar{y}_\delta, \bar{u}_\delta)$ of the Moreau-Yosida regularization converge to the solution $(\bar{y}, \bar{u})$ of (P_s) when $\delta \to \infty$. The Moreau-Yosida regularization corresponding to (P_s^{test}) is given by

$$\left.\begin{array}{l} \min\ J^{MY}(y_\delta, u_\delta) \\ -\varepsilon\Delta y_\delta + \mathbf{b}\cdot\nabla y_\delta + c y_\delta = u_\delta + f \quad \text{in } \Omega \\ y_\delta = 0 \quad \text{on } \Gamma \end{array}\right\} (P_s^{test,MY}) \tag{8.3.15}$$

where the objective functional is defined by

$$J^{MY}(y_\delta, u_\delta) := J^s(y_\delta, u_\delta) + \frac{\delta}{2}\left(\| \max\{0, y_\delta - y_b\} \|_{0,\Omega}^2 + \| \min\{0, y_\delta - y_a\} \|_{0,\Omega}^2\right)$$

with a regularization parameter $\delta > 0$ that is taken large. Moreover, according to Section 4.3.2, we have derived that the solution $(\bar{y}_\delta, \bar{u}_\delta)$ of $(P_s^{test,MY})$ satisfies the following optimality system

$$\begin{array}{ll} -\varepsilon\Delta\bar{y}_\delta + \mathbf{b}\cdot\nabla\bar{y}_\delta + c\bar{y}_\delta = \bar{u}_\delta + f \quad \text{in } \Omega & -\varepsilon\Delta\bar{p}_\delta - \mathbf{b}\cdot\nabla\bar{p}_\delta + c\bar{p}_\delta = \bar{y}_\delta - y_d + \mu_\delta \quad \text{in } \Omega \\ \bar{y}_\delta = 0 \quad \text{on } \Gamma & \bar{p}_\delta = 0 \quad \text{on } \Gamma \end{array}$$

$$\begin{array}{r} \mu_\delta = \delta\cdot(\max\{0, \bar{y}_\delta - y_b\} + \min\{0, \bar{y}_\delta - y_a\}) = \delta\cdot\left(\bar{y}_\delta - \mathbb{P}_{[y_a,y_b]}(\bar{y}_\delta)\right) \quad \text{a.e. in } \Omega \\ \lambda\bar{u}_\delta + \bar{p}_\delta = 0 \quad \text{a.e. in } \Omega. \end{array}$$

Regarding Section 6.3.1, the discrete version of the optimality system corresponding to $(P_s^{test,MY})$ has the form

$$\begin{array}{l} a(y_h, v_h) = (u_h + f, v_h)_\Omega \quad \forall v_h \in V_{h,0} \\ a(\psi_h, p_h) = (y_h - y_d, \psi_h)_\Omega + (\mu_h^\delta, \psi_h)_\Omega \quad \forall \psi_h \in V_{h,0} \end{array}$$

where

$$\begin{array}{r} \mu_h^\delta = \delta\cdot(\max\{0, y_h - y_b\} + \min\{0, y_h - y_a\}) = \delta\cdot\left(y_h - \mathbb{P}_{[y_a,y_b]}(y_h)\right) \quad \text{a.e. in } \Omega \\ \lambda u_h + p_h = 0 \quad \text{a.e. in } \Omega. \end{array}$$

In our numerical tests, we have reviewed the stabilizing effect of the AFC method on discrete solutions for different choices of $\delta > 0$. Note that for given $\delta > 0$, the analytical solutions of $(P_s^{test,MY})$ are unknown. Hence, we use a specific formula to compute the experimental order of convergence in the $L^2(\Omega)$-norm.

8.3.3.2 Experimental order of convergence

Due to the fact that the computation of the $EOC_{\|\cdot\|}$ where $\|\cdot\|$ is an appropriate norm is not often reliable by using a computed reference solution, we have performed our numerical tests with the help of a specific formula for the $EOC_{\|\cdot\|}$. Let us start with the derivation of the formula. For a grid level $l \in \{1, \cdots, 8\}$, we consider the corresponding mesh size h_l and the computed discrete state solution y_{h_l} resp. the discrete adjoint solution p_{h_l}. Note that the coarest mesh size h_1 is refined by $h_l = \frac{1}{2^{l-1}}h_1$ for $l = 2, \cdots, 8$ resp. $h_{l+1} = \frac{1}{2}h_l$ for $l = 1, \cdots, 7$. Now, a numerical method of order q should satisfy

$$w = w_{h_l} + e(w)(h_l)^q \tag{8.3.16}$$

where $e(w)(h_l)^q$ is the error between the computed discrete solution w_{h_l} and the analytical solution w of the test problem. As we can see in (8.3.16), the discrete solution converges to the analytical solution when $h_l \to 0$. For the computation of the order q, we consider

$$\begin{aligned} w &= w_{h_l} + e(w)(h_l)^q \\ w &= w_{h_{l+1}} + e(w)(h_{l+1})^q \\ w &= w_{h_{l+2}} + e(w)(h_{l+2})^q. \end{aligned}$$

Now, we get

$$\begin{aligned} w_{h_l} - w_{h_{l+1}} &= e(w)(h_{l+1})^q - e(w)(h_l)^q \\ &= e(w)\left(\left(\frac{1}{2}\right)^q h_l^q - h_l^q\right) \\ &= e(w)\left(\left(\frac{1}{2}\right)^q - 1\right) h_l^q \end{aligned} \tag{8.3.17}$$

and

$$\begin{aligned} w_{h_{l+1}} - w_{h_{l+2}} &= e(w)(h_{l+2})^q - e(w)(h_{l+1})^q \\ &= e(w)\left(\left(\frac{1}{2}\right)^q h_{l+1}^q - h_{l+1}^q\right) \\ &= e(w)\left(\left(\frac{1}{2}\right)^q - 1\right) h_{l+1}^q \\ &= e(w)\left(\left(\frac{1}{2}\right)^q - 1\right)\left(\frac{1}{2}\right)^q h_l^q. \end{aligned} \tag{8.3.18}$$

Hence, we obtain by virtue of (8.3.17) and (8.3.18)

$$\ln\left(\frac{\| w_{h_{l+2}} - w_{h_{l+1}} \|}{\| w_{h_{l+1}} - w_{h_l} \|}\right) = q \cdot \ln\left(\frac{1}{2}\right)$$

and consequently the formula for the computation of the order q

$$\frac{\ln\left(\frac{\|w_{h_{l+2}} - w_{h_{l+1}}\|}{\|w_{h_{l+1}} - w_{h_l}\|}\right)}{\ln\left(\frac{1}{2}\right)} = q =: EOC_{\|\cdot\|}. \tag{8.3.19}$$

In the following, we show for regularization parameters $\delta = 1, 10, 100$ the computed $EOC_{\|\cdot\|_{0,\Omega}}$ for discrete state solutions y_{h_l} and discrete adjoint solutions p_{h_l} in the $L^2(\Omega)$-norm. For this, the computed $EOC_{\|\cdot\|_{0,\Omega}}$ of the state solution is denoted by q_l^s resp. the $EOC_{\|\cdot\|_{0,\Omega}}$ of the adjoint solution is denoted by q_l^{ad} for $l = 1, \cdots, 6$. Moreover, we introduce for $l = 1, \cdots, 7$ the differences

$$\begin{aligned} e_{h_l} &:= y_{h_{l+1}} - y_{h_l} \\ k_{h_l} &:= p_{h_{l+1}} - p_{h_l}. \end{aligned} \tag{8.3.20}$$

8.3.3.3 Numerical results $\delta = 1$

In Table 6, for the differences (8.3.20), the computed $L^2(\Omega)$-norms are illustrated for the regularization parameter $\delta = 1$. For the state and the adjoint solutions, the values of the $L^2(\Omega)$-norms decrease while the grid level increases. Moreover, by using the EOC-formula (8.3.19), the $L^2(\Omega)$-order of convergence is nearly $\mathcal{O}(1)$.

Table 6: $L^2(\Omega)$-errors / EOC ($P_s^{test,MY}$), $\delta = 1$

l	$c_0 \parallel e_{h_l} \parallel_{0,\Omega}$	q_l^s	$c_0 \parallel k_{h_l} \parallel_{0,\Omega}$	q_l^{ad}
1	1.8278e-01	-	2.2768e-01	-
2	1.2829e-01	0.51	1.7061e-01	0.42
3	9.4977e-02	0.43	1.2622e-01	0.43
4	6.7941e-02	0.48	9.0822e-02	0.47
5	4.7805e-02	0.51	6.4380e-02	0.50
6	2.8016e-02	0.77	3.7896e-02	0.76
7	1.3770e-02	1.02	1.8683e-02	1.02

The following plots of the AFC state solution and the AFC adjoint solution has been computed on a $(128, 128)$-unit square mesh for a regularization parameter $\delta = 1$. As we can see, there occur no spurious oscillations in the computed discrete solutions.

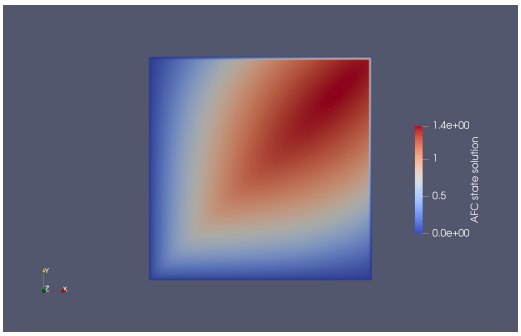


Figure 15: AFC state solution

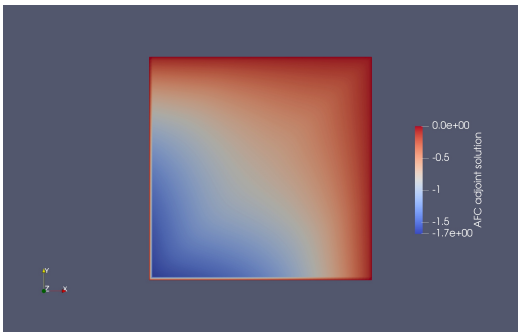


Figure 16: AFC adjoint solution

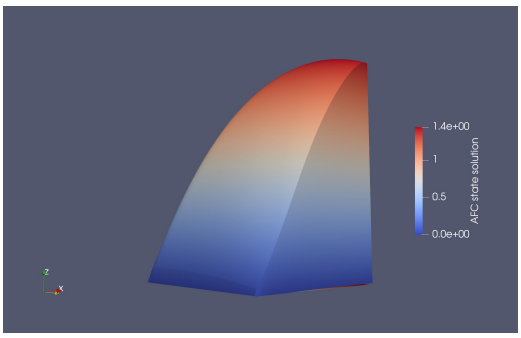


Figure 17: AFC state solution

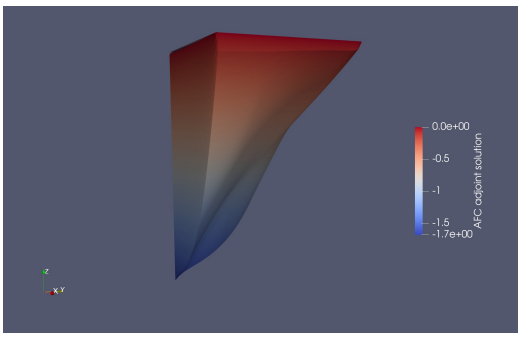


Figure 18: AFC adjoint solution

8.3.3.4 Numerical results $\delta = 10$

In Table 7, for the differences (8.3.20), the computed $L^2(\Omega)$-norms are illustrated for the regularization parameter $\delta = 10$. As in the case of $\delta = 1$, the values of the $L^2(\Omega)$-norms decrease while the grid level increases. Moreover, the EOC-formula (8.3.19) leads us to a $L^2(\Omega)$-order of convergence nearly $\mathcal{O}(1)$ for the state and the adjoint solutions.

Table 7: $L^2(\Omega)$-errors / EOC ($P_s^{test,MY}$), $\delta = 10$

l	$c_0 \parallel e_{h_l} \parallel_{0,\Omega}$	q_l^s	$c_0 \parallel k_{h_l} \parallel_{0,\Omega}$	q_l^{ad}
1	1.4887e-01	-	2.1034e-01	-
2	1.0148e-01	0.55	1.4968e-01	0.49
3	7.5458e-02	0.43	1.0985e-01	0.45
4	5.4358e-02	0.47	7.9610e-02	0.46
5	3.8342e-02	0.50	5.6657e-02	0.49
6	2.2548e-02	0.77	3.3389e-02	0.76
7	1.1088e-02	1.02	1.6471e-02	1.02

Now, we plot the AFC state resp. the AFC adjoint solution computed for a regularization parameter $\delta = 10$ on a $(128, 128)$-unit square mesh.

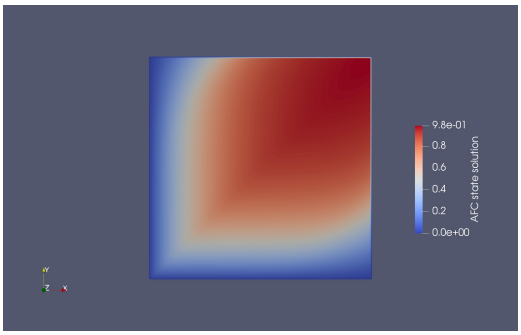


Figure 19: AFC state solution

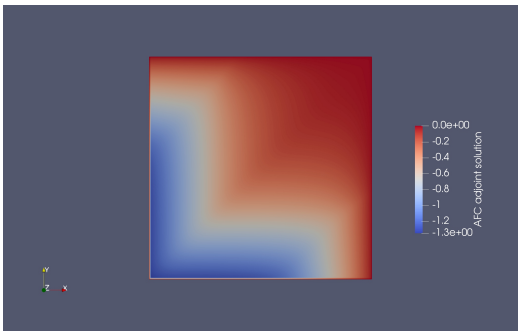


Figure 20: AFC adjoint solution

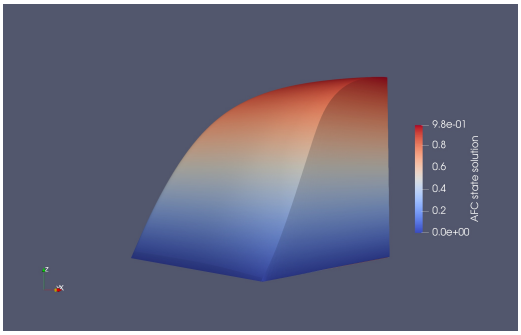


Figure 21: AFC state solution

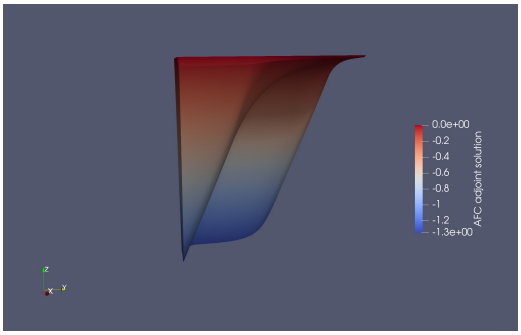


Figure 22: AFC adjoint solution

8.3.3.5 Numerical results $\delta = 100$

In Table 8, for the differences (8.3.20), the computed $L^2(\Omega)$-norms are illustrated for the regularization parameter $\delta = 100$. As in the cases $\delta = 1$ and $\delta = 10$, the values of the $L^2(\Omega)$-norms decrease while the grid level increases. Moreover, the EOC-formula (8.3.19) leads us to a $L^2(\Omega)$-order of convergence nearly $\mathcal{O}(1)$ for the state and the adjoint solutions.

Table 8: $L^2(\Omega)$-errors / EOC ($P_s^{test,MY}$), $\delta = 100$

l	$c_0 \parallel e_{h_l} \parallel_{0,\Omega}$	q_l^s	$c_0 \parallel k_{h_l} \parallel_{0,\Omega}$	q_l^{ad}
1	1.3171e-01	-	1.9526e-01	-
2	8.5484e-02	0.62	1.3936e-01	0.49
3	6.3388e-02	0.43	1.0213e-01	0.45
4	4.5958e-02	0.46	7.4082e-02	0.46
5	3.2425e-02	0.50	5.2792e-02	0.49
6	1.9155e-02	0.76	3.1403e-02	0.75
7	9.3454e-03	1.04	1.5368e-02	1.03

Now, we can see the computed AFC state resp. the AFC adjoint solution for a regularization parameter $\delta = 100$ on a $(128, 128)$-unit square mesh.

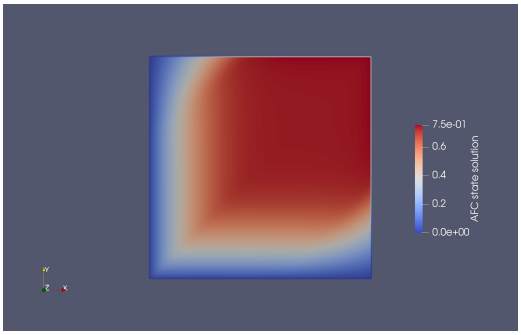


Figure 23: AFC state solution

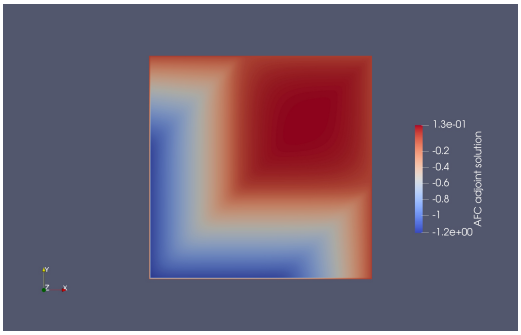


Figure 24: AFC adjoint solution

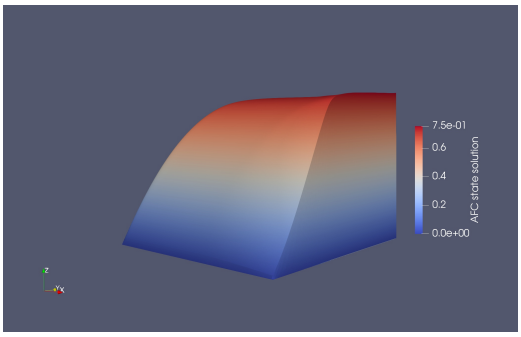


Figure 25: AFC state solution

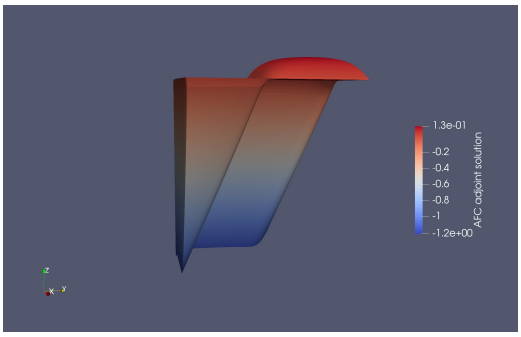


Figure 26: AFC adjoint solution

8.4 Control constrained case with Robin boundary control

In this section, we investigate the discrete system (P_h^Γ) corresponding to the control constrained optimal control problem with Robin boundary control. According to Section 7.1, we set $X_h = V_h \subseteq H^1(\Omega) = X$ and $\mathcal{X}_h = V_h \times V_h \subset \mathcal{X} = H^1(\Omega) \times H^1(\Omega)$ where $\mathcal{X}$ is equipped with the norm

$$\parallel (y,p) \parallel_{\mathcal{X}} = \sqrt{\parallel y \parallel_{1,\Omega}^2 + \parallel p \parallel_{1,\Omega}^2}.$$

The finite dimensional space $\mathcal{X}_h$ is endowed with the norm

$$\parallel \cdot \parallel_{\mathcal{X}_h} := \parallel \cdot \parallel_{\mathcal{X}} .$$

The starting point of the next sections is the following coupled and discretized system, provided in Section 6.4.

$$\left.\begin{aligned} &a_\Gamma(y_h, v_h) + d_h^{s,\Gamma}(y_h; y_h, v_h) = (u_h, v_h)_\Gamma \quad \forall v_h \in X_h \\ &a_\Gamma(\psi_h, p_h) + d_h^{ad,\Gamma}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \\ &u_h = \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}\left(-\tfrac{1}{\lambda} p_h\right) \quad \text{a.e. on } \Gamma. \end{aligned}\right\} (P_h^\Gamma)$$

8.4.1 Existence

First, we state that the systems (P_h^b) and (P_h^Γ) have a similar structure. Hence, for the verification of the existence of a discrete solution for (P_h^Γ), we follow the regularization strategy of (P_h^b), provided in Section 8.2.1.

8.4.1.1 Regularized system

Let $k \in \mathbb{N}$ be arbitrary. We consider the following truncated, coupled and discretized system

$$\left.\begin{aligned} &a_\Gamma(y_h, v_h) + d_h^{s,\Gamma}(y_h; y_h, v_h) = (u_h, v_h)_\Gamma \quad \forall v_h \in X_h \\ &a_\Gamma(\psi_h, p_h) + d_h^{ad,\Gamma}(p_h; p_h, \psi_h) = (\Psi_k(y_h) - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \\ &u_h = \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(-\tfrac{1}{\lambda} p_h) \quad \text{a.e. on } \Gamma. \end{aligned}\right\} (P_h^{\Gamma,k})$$

Definition 8.36. *Let $k \in \mathbb{N}$. A pair $(y_h, u_h) \in X_h \times L^2(\Gamma)$ is called solution for $(P_h^{\Gamma,k})$ if there exists a discrete adjoint solution $p_h \in X_h$ such that*

$$a_\Gamma(y_h, v_h) + d_h^{s,\Gamma}(y_h; y_h, v_h) = (u_h, v_h)_\Gamma \quad \forall v_h \in X_h \tag{8.4.1}$$

$$a_\Gamma(\psi_h, p_h) + d_h^{ad,\Gamma}(p_h; p_h, \psi_h) = (\Psi_k(y_h) - y_d, \psi_h)_\Omega \quad \forall \psi_h \in X_h \tag{8.4.2}$$

$$u_h = \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(-\frac{1}{\lambda} p_h) \quad \text{a.e. on } \Gamma \tag{8.4.3}$$

is satisfied.

Similar to the control constrained case (see Section 8.2.1), the strategy is to solve the system $(P_h^{\Gamma,k})$ for a $k \in \mathbb{N}$ which is large enough. After that, we verify that $\Psi_k(y_h) = y_h$ a.e. in Ω. First, the system (8.4.1)-(8.4.3) can be transferred to the following operator equation: Find $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that

$$(K + D_h)\,\mathbf{x}_h = Q_k \mathbf{x}_h + G \quad \text{in } \mathcal{X}_h^* \tag{8.4.4}$$

where for $\mathbf{z}_h = (v_h, \psi_h) \in \mathcal{X}_h$

- $\langle K\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := \langle K_\Gamma^s y_h, v_h\rangle_{X^*,X} + \langle K_\Gamma^{ad} p_h, \psi_h\rangle_{X^*,X}$
- $\langle D_h\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := \langle D_{y_h}^{s,\Gamma} y_h, v_h\rangle_{X_h^*,X_h} + \langle D_{p_h}^{ad,\Gamma} p_h, \psi_h\rangle_{X_h^*,X_h}$
- $\langle Q_k\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := (\mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(-\frac{1}{\lambda} p_h), v_h)_\Gamma + (\Psi_k(y_h), \psi_h)_\Omega$
- $\langle G, \mathbf{z}_h\rangle_{\mathcal{X}^*,\mathcal{X}} := (-y_d, \psi_h)_\Omega$.

Note that the above operators have been defined in Section 7.2.1.2.

Lemma 8.37. *Let $k \in \mathbb{N}$ be arbitrary. The operator $K_h := (K + D_h) : \mathcal{X}_h \to \mathcal{X}_h^*$ satisfies Assumption 7.1 and $\tilde{Q}_k := Q_k + G$ defined by*

$$\langle \tilde{Q}_k\mathbf{x}_h, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h} := \langle Q_k\mathbf{x}_h + G, \mathbf{z}_h\rangle_{\mathcal{X}_h^*,\mathcal{X}_h}$$

satisfies Assumption 7.2.

Proof. First, the continuity of the operator K_h is obvious. Due to the Lipschitz continuity of the projection formula $\mathbb{P}_{[u_a^\Gamma,u_b^\Gamma]}(\cdot)$ and $\Psi_k(\cdot)$, we get the continuity of $\tilde{Q}_k := Q_k + G$. The verification of condition (7.1.1) in Assumption 7.1 goes in the same way as in Lemma 8.2. For this, let $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ be arbitrary. The coercivity of $a_\Gamma(\cdot,\cdot)$ and $d_h^{s,\Gamma}(y_h; y_h, y_h) \geq 0$ resp. $d_h^{ad,\Gamma}(p_h; p_h, p_h) \geq 0$ imply

$$\langle K_h \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \geq \min\{\varepsilon, c_0\} \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}^2 .$$

For the satisfaction of condition (7.1.2) in Assumption 7.2, we are able to derive

$$\begin{aligned}
\langle \tilde{Q}_k \mathbf{x}_h, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} &= \langle Q_k \mathbf{x}_h + G, \mathbf{x}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} \\
&= (\mathbb{P}_{[u_a^\Gamma,u_b^\Gamma]}(-\frac{1}{\lambda} p_h), y_h)_\Gamma + (\Psi_k(y_h), p_h)_\Omega - (y_d, p_h)_\Omega \\
&\leq C \parallel y_h \parallel_{0,\Gamma} + C(k) \parallel p_h \parallel_{0,\Omega} + \parallel y_d \parallel_{0,\Omega} \parallel p_h \parallel_{0,\Omega} \\
&\leq C_\tau C \parallel y_h \parallel_{1,\Omega} + C(k) \parallel p_h \parallel_{0,\Omega} + \parallel y_d \parallel_{0,\Omega} \parallel p_h \parallel_{0,\Omega} \\
&\leq \tilde{C}_k \left(\sqrt{\parallel y_h \parallel_{1,\Omega}^2} + \sqrt{\parallel p_h \parallel_{1,\Omega}^2} \right) \\
&\leq \sqrt{2} \tilde{C}_k \sqrt{\parallel y_h \parallel_{1,\Omega}^2 + \parallel p_h \parallel_{1,\Omega}^2} \\
&= \hat{C}_k \parallel \mathbf{x}_h \parallel_{\mathcal{X}_h}
\end{aligned}$$

where $\hat{C}_k > 0$ is a constant which depends on k and on the constant $C_\tau > 0$ of the trace inequality (see Theorem 2.18). □

Lemma 8.38. *There exists a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ for the operator equation* (8.4.4).

Proof. According to Lemma 8.37, the operators K_h and $\tilde{Q}_k := Q_k + G$ satisfy the sufficient conditions for the application of Lemma 7.3. Consequently, we obtain a solution $\mathbf{x}_h = (y_h, p_h) \in \mathcal{X}_h$ such that the operator equation (8.4.4) holds. □

Now, we are able to define a solution for $(P_h^{\Gamma,k})$ with the help of $\mathbf{x}_h = (y_h, p_h)$. According to (8.4.3), the control can be defined by $u_h = \mathbb{P}_{[u_a^\Gamma,u_b^\Gamma]}(-\frac{1}{\lambda} p_h)$ a.e. on Γ. Hence, the following corollary is a direct consequence of Lemma 8.38.

Corollary 8.39. *There exists a solution $(y_h, u_h) \in X_h \times L^2(\Gamma)$ with a corresponding adjoint solution $p_h \in X_h$ for the regularized discretized system $(P_h^{\Gamma,k})$.*

In the same way as in Section 8.2.1, we consider for the right choice of $k \in \mathbb{N}$, the following $L^\infty(\Omega)$-a priori estimate of a discrete solution $y_h \in X_h$.

Lemma 8.40. *Let $y_h \in X_h$ be a solution of the state equation* (8.4.1). *Then, we have*

$$\parallel y_h \parallel_{0,\infty,\Omega} \leq C(1 + |\ln(h)|)^{\frac{1}{2}} \frac{\tilde{C}(u_a^\Gamma, u_b^\Gamma, \Omega)}{\min\{\varepsilon, c_0\}}.$$

Proof. Setting $v_h = y_h$ in (8.4.1), the coercivity of $a_\Gamma(\cdot,\cdot)$ and the positivity of $d_h^{s,\Gamma}(y_h; y_h, y_h)$ imply

$$\begin{aligned}
\min\{\varepsilon, c_0\} \parallel y_h \parallel_{1,\Omega}^2 &\leq (\mathbb{P}_{[u_a^\Gamma,u_b^\Gamma]}(-\frac{1}{\lambda} p_h), y_h)_\Gamma \\
&\leq C(u_a^\Gamma, u_b^\Gamma, \Omega) \parallel y_h \parallel_{0,\Gamma} \\
&\leq C_\tau C(u_a^\Gamma, u_b^\Gamma, \Omega) \parallel y_h \parallel_{1,\Omega}
\end{aligned}$$

where we have used the trace inequality. Hence, we get

$$\| y_h \|_{1,\Omega} \leq \frac{\tilde{C}(u_a^\Gamma, u_b^\Gamma, \Omega)}{\min\{\varepsilon, c_0\}}$$

with a constant $\tilde{C}(u_a^\Gamma, u_b^\Gamma, \Omega) > 0$. Finally, the discrete Sobolev inequality (Lemma 5.3) yields the desired result. □

In the following, we set

$$k = k(h) := \lceil C(1 + |\ln(h)|)^{\frac{1}{2}} \frac{\tilde{C}(u_a^\Gamma, u_b^\Gamma, \Omega)}{\min\{\varepsilon, c_0\}} \rceil \tag{8.4.5}$$

where $\lceil r \rceil$ denotes the usual ceiling function.

Corollary 8.41. *Let $k \in \mathbb{N}$ be defined by (8.4.5). Then, a solution $(y_h, u_h) \in X_h \times L^2(\Gamma)$ of $(P_h^{\Gamma,k})$ with a corresponding adjoint solution $p_h \in X_h$ is also a solution for (P_h^Γ).*

Proof. Lemma 8.40 yields that a discrete state solution y_h satisfies

$$\| y_h \|_{0,\infty,\Omega} \leq C(1 + |\ln(h)|)^{\frac{1}{2}} \frac{\tilde{C}(u_a^\Gamma, u_b^\Gamma, \Omega)}{\min\{\varepsilon, c_0\}} \leq k$$

such that $\Psi_k(y_h) = y_h$ a.e. in Ω. Consequently, the regularized system $(P_h^{\Gamma,k})$ coincides with the discretized system (P_h^Γ) such that $(y_h, u_h) \in X_h \times L^2(\Gamma)$ with $p_h \in X_h$ solves (P_h^Γ). □

Next, we verify higher regularity of u_h, i.e. $u_h = \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(-\frac{1}{\lambda} p_h) \in H^{\frac{1}{2}}(\Gamma)$. According to Lemma 7.10 and Lemma 7.11, the higher regularity of u_h is sufficient to obtain the $H^2(\Omega)$-regularity of the auxiliary solution $\tilde{y}$. The meaning of this result will become apparent in Section 8.4.2. In addition to the higher regularity of u_h, we verify that y_h, p_h are uniformly bounded in the $L^2(\Omega)$-norm resp. u_h is uniformly bounded in the $H^{\frac{1}{2}}(\Gamma)$-norm.

Lemma 8.42. *Let $(y_h, u_h) \in X_h \times L^2(\Gamma)$ be a solution of (P_h^Γ) with a corresponding adjoint solution $p_h \in X_h$. Then, $u_h \in H^{\frac{1}{2}}(\Gamma)$ and we have the following estimates*

$$\| u_h \|_{\frac{1}{2},\Gamma} \leq C_c \tag{8.4.6}$$

$$\| y_h \|_{0,\Omega} \leq C_s \tag{8.4.7}$$

$$\| p_h \|_{0,\Omega} \leq C_{ad} \tag{8.4.8}$$

where $C_c, C_s, C_{ad} > 0$ are constants, independent of h.

Proof. We start with the higher regularity of u_h. Since $p_h \in X_h \subset H^1(\Omega)$, we have $(p_h)_{|\Gamma} \in H^{\frac{1}{2}}(\Gamma)$. Hence, with the regularity assumption (B4), i.e. $u_a^\Gamma, u_b^\Gamma \in H^1(\Gamma)$, [KinSta80, Theorem A.1, p. 50] yields $u_h = \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(-\frac{1}{\lambda} p_h) \in H^{\frac{1}{2}}(\Gamma)$. Moreover, we obtain the uniform boundedness of the control in the $H^{\frac{1}{2}}(\Gamma)$-norm resp. in the $L^2(\Gamma)$-norm by

$$\| u_h \|_{0,\Gamma} \leq \| u_h \|_{\frac{1}{2},\Gamma} = \| \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(-\frac{1}{\lambda} p_h) \|_{\frac{1}{2},\Gamma} := C_c. \tag{8.4.9}$$

Now, we prove the boundedness of y_h and p_h in the $L^2(\Omega)$-norm. Considering the discretized state equation (8.4.1) with $v_h = y_h$, we obtain with (8.4.9) and the trace inequality (Theorem 2.18)

$$\begin{aligned}
\min\{\varepsilon, c_0\} \| y_h \|_{1,\Omega}^2 &\leq a_\Gamma(y_h, y_h) + d_h^{s,\Gamma}(y_h; y_h, y_h) \\
&= (u_h, y_h)_\Gamma \\
&\leq \| u_h \|_{0,\Gamma} \| y_h \|_{0,\Gamma} \\
&\leq C_c C_\tau \| y_h \|_{1,\Omega} .
\end{aligned}$$

Thus, we get

$$\| y_h \|_{1,\Omega} \leq \frac{C_c C_\tau}{\min\{\varepsilon, c_0\}} =: C_s. \tag{8.4.10}$$

For the verification of the $L^2(\Omega)$-boundedness of p_h, we consider the discretized adjont equation (8.4.2). Setting $\psi_h = p_h$ in (8.4.2), leads us to

$$\begin{aligned} c_0 \| p_h \|_{0,\Omega}^2 &\leq a_\Gamma(p_h, p_h) + d_h^{ad,\Gamma}(p_h; p_h, p_h) \\ &= (y_h - y_d, p_h)_\Omega \\ &\leq (\| y_h \|_{0,\Omega} + \| y_d \|_{0,\Omega}) \| p_h \|_{0,\Omega} . \end{aligned}$$

By means of (8.4.10), we get

$$c_0 \| p_h \|_{0,\Omega} \leq C_s + \| y_d \|_{0,\Omega} \tag{8.4.11}$$

and consequently

$$\| p_h \|_{0,\Omega} \leq \frac{C_s + \| y_d \|_{0,\Omega}}{c_0} =: C_{ad}.$$

□

8.4.2 L^2-error estimates

In this section, we derive a $L^2(\Omega)$-error estimate for $\bar{y} - y_h$ and a $L^2(\Gamma)$-error estimate for $\bar{u} - u_h$ where $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ is a discrete solution of (P_h^Γ) and $(\bar{y}, \bar{u}) \in H^2(\Omega) \times H^1(\Gamma)$ is the solution of (P_Γ) (see Corollary 4.37). Regarding Section 7.2, the discretized system (P_h^Γ) coincides with the general system (P_h) and (P_Γ) coincides with (P) when the following assumptions hold:

Assumption 8.43. *We assume that*

- $\mathcal{G} = \Gamma$
- $U := U_{ad}^\Gamma = \{u \in L^2(\Gamma) : u_a^\Gamma(x) \leq u(x) \leq u_b^\Gamma(x) \quad a.e. \text{ on } \Gamma\}$.
- $\delta = 1$ *and* $\hat{\delta} = 0$.
- $Z : C(\bar{\Omega}) \to U$ *is defined by*

$$Zw := \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}\left(-\frac{1}{\lambda} w\right) \quad a.e. \text{ on } \Gamma.$$

Now, we derive a L^2-error estimate for the state and the control.

Lemma 8.44. *Let $(\bar{y}, \bar{u}) \in H^2(\Omega) \times H^1(\Gamma)$ be the solution of (P_Γ) with the corresponding adjoint solution $\bar{p} \in H^2(\Omega)$. Moreover, we have the solution $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$ of (P_{aux}) and a solution $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ of (P_h^Γ) with a corresponding discrete adjoint solution $p_h \in X_h$. Then, we have*

$$\frac{\lambda}{2} \| u_h - \bar{u} \|_{0,\Gamma}^2 + \frac{1}{2} \| y_h - \bar{y} \|_{0,\Omega}^2 \leq C \| p_h - \tilde{p} \|_{0,\Gamma}^2 + \frac{1}{2} \| y_h - \tilde{y} \|_{0,\Omega}^2 .$$

Proof. Lemma 7.6 yields for $\mathcal{G} = \Gamma$ and general $\delta, \hat{\delta} \in \mathbb{R}_{\geq 0}$ with $\delta - \hat{\delta} > 0$ the estimate

$$\frac{\lambda}{2} \parallel u_h - \bar{u} \parallel_{0,\Gamma}^2 + \frac{(\delta - \hat{\delta})}{2} \parallel y_h - \bar{y} \parallel_{0,\Omega}^2 \leq C \parallel p_h - \tilde{p} \parallel_{0,\Gamma}^2 + \frac{\delta^2}{2(\delta - \hat{\delta})} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2 .$$

Thus, setting $\delta = 1$ and $\hat{\delta} = 0$ leads us to the desired result. □

Lemma 8.45. *Let $(\tilde{y}, \tilde{p}) \in H^2(\Omega) \times H^2(\Omega)$ be the solution of (P_{aux}). Moreover, let $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ be a solution of (P_h^Γ) where $p_h \in X_h$ is a corresponding discrete adjoint solution. Then, we have*

$$\parallel \tilde{y} \parallel_{2,\Omega} \leq C_1 \tag{8.4.12}$$

$$\parallel \tilde{p} \parallel_{2,\Omega} \leq C_2 \tag{8.4.13}$$

where $C_1, C_2 > 0$ are constants, independent of h.

Proof. Lemma 7.10 yields the a priori estimates

$$\parallel \tilde{y} \parallel_{2,\Omega} \leq C \parallel u_h \parallel_{\frac{1}{2},\Gamma}$$

$$\parallel \tilde{p} \parallel_{2,\Omega} \leq C \left(\parallel y_h \parallel_{0,\Omega} + \parallel y_d \parallel_{0,\Omega} \right).$$

By virtue of Lemma 8.42, we obtain the desired result. □

Theorem 8.46. *Let $(\bar{y}, \bar{u}) \in H^2(\Omega) \times H^1(\Gamma)$ be the solution of (P_Γ) and $(y_h, u_h) \in X_h \times H^{\frac{1}{2}}(\Gamma)$ be a solution of (P_h^Γ).Then, we have in the convection-dominated case, i.e.*
$\varepsilon \ll \parallel \mathbf{b} \parallel_{0,\infty,\Omega} h$

$$\parallel u_h - \bar{u} \parallel_{0,\Gamma} + \parallel y_h - \bar{y} \parallel_{0,\Omega} \leq C \frac{h^{\frac{1}{2}}}{\varepsilon^{\frac{1}{4}}} + C h^{\frac{1}{2}}.$$

Proof. Lemma 8.44 provides the estimate

$$\frac{\lambda}{2} \parallel u_h - \bar{u} \parallel_{0,\Gamma}^2 + \frac{1}{2} \parallel y_h - \bar{y} \parallel_{0,\Omega}^2 \leq C \parallel p_h - \tilde{p} \parallel_{0,\Gamma}^2 + \frac{1}{2} \parallel y_h - \tilde{y} \parallel_{0,\Omega}^2$$

where $p_h \in X_h$ is a corresponding discrete adjoint solution and $\tilde{y}, \tilde{p}$ the auxiliary solutions of (P_{aux}). The auxiliary error estimate (see Corollary 7.12) yields with (8.4.12)

$$\parallel y_h - \tilde{y} \parallel_{0,\Omega} \leq C h^{\frac{1}{2}}. \tag{8.4.14}$$

For the estimation of $\parallel p_h - \tilde{p} \parallel_{0,\Gamma}^2$, we use the trace inequality (Theorem 2.18) and get

$$\parallel p_h - \tilde{p} \parallel_{0,\Gamma} \leq C_\tau \parallel p_h - \tilde{p} \parallel_{0,\Omega}^{\frac{1}{2}} \parallel p_h - \tilde{p} \parallel_{1,\Omega}^{\frac{1}{2}} . \tag{8.4.15}$$

The $L^2(\Omega)$-norm can be estimated by Lemma 7.12 and (8.4.13) such that we obtain

$$\parallel p_h - \tilde{p} \parallel_{0,\Omega} \leq C \parallel p_h - \tilde{p} \parallel_h^{ad,\Gamma} \leq C h^{\frac{1}{2}}.$$

The $H^1(\Omega)$-norm can be estimated by

$$\begin{aligned}
\parallel p_h - \tilde{p} \parallel_{1,\Omega} &\leq \frac{\varepsilon^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}} \parallel p_h - \tilde{p} \parallel_{1,\Omega} \\
&= \frac{1}{\varepsilon^{\frac{1}{2}}} \left(\varepsilon \parallel p_h - \tilde{p} \parallel_{0,\Omega}^2 + \varepsilon |p_h - \tilde{p}|_{1,\Omega}^2 \right)^{\frac{1}{2}} \\
&\leq \frac{1}{\varepsilon^{\frac{1}{2}}} \left(C\varepsilon (\parallel p_h - \tilde{p} \parallel_h^{ad,\Gamma})^2 + (\parallel p_h - \tilde{p} \parallel_h^{ad,\Gamma})^2 \right)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{1 + C\varepsilon}{\varepsilon}} \parallel p_h - \tilde{p} \parallel_h^{ad,\Gamma} \\
&\leq \frac{C}{\varepsilon^{\frac{1}{2}}} h^{\frac{1}{2}}
\end{aligned}$$

where we have used in the last step again Lemma 7.12. Consequently, we get for (8.4.15)

$$\begin{aligned} \| p_h - \tilde{p} \|_{0,\Gamma} &\leq C h^{\frac{1}{4}} \cdot \frac{C}{\varepsilon^{\frac{1}{4}}} h^{\frac{1}{4}} \\ &\leq C \frac{h^{\frac{1}{2}}}{\varepsilon^{\frac{1}{4}}} \end{aligned}$$

and by virtue of (8.4.14) the desired result. □

8.4.3 Numerical results

In this section, we show the numerical results concerning the application of the AFC method to the control constrained optimal control problem with Robin boundary control (P_Γ). We use the same iterative solver as in the numerical tests of the previous sections. Moreover, the following test problem has been solved on a unit square mesh $\Omega = [0,1] \times [0,1]$ where we have used the BJK limiter, introduced in Section 5.5.2. For a general introduction to the numerics, i.e. data of the convection-diffusion reaction equation, Tikhonov parameter, grid levels etc., we refer to Section 8.1.3.

8.4.3.1 Test problem

For the numerical investigation, we consider the following test problem

$$\left.\begin{array}{l} \min \ \frac{1}{2} \| y - y_d \|_{0,\Omega}^2 + \frac{\lambda}{2} \| u - u_d \|_{0,\Gamma}^2 \\ -\varepsilon \Delta y + \mathbf{b} \cdot \nabla y + c y = f \quad \text{in } \Omega \\ \varepsilon \partial_n y - \frac{\mathbf{b} \cdot n \cdot y}{2} = u + g \quad \text{on } \Gamma \\ u_a^\Gamma \leq u \leq u_b^\Gamma \quad \text{a.e. on } \Gamma \end{array}\right\} (P_\Gamma^{test})$$

where $f \in L^2(\Omega)$ and $g, u_d \in L^2(\Gamma)$ will be specified later. The optimality system corresponding to (P_Γ^{test}) is given by

$$\begin{aligned} -\varepsilon \Delta \bar{y} + \mathbf{b} \cdot \nabla \bar{y} + c \bar{y} &= f \quad \text{in } \Omega & -\varepsilon \Delta \bar{p} - \mathbf{b} \cdot \nabla \bar{p} + c \bar{p} &= \bar{y} - y_d \quad \text{in } \Omega \\ \varepsilon \partial_n \bar{y} - \frac{\mathbf{b} \cdot n \cdot \bar{y}}{2} &= \bar{u} + g \quad \text{on } \Gamma & \varepsilon \partial_n \bar{p} + \frac{\mathbf{b} \cdot n \cdot \bar{p}}{2} &= 0 \quad \text{on } \Gamma \end{aligned}$$

$$(\lambda(\bar{u} - u_d) + \bar{p}, u - \bar{u})_\Gamma \geq 0 \quad \forall u \in U_{ad}^\Gamma$$

The set of admissible controls is defined by

$$U_{ad}^\Gamma := \{ u \in L^2(\Gamma) : u_a^\Gamma(x) \leq u(x) \leq u_b^\Gamma(x) \quad \text{a.e. on } \Gamma \}.$$

8.4.3.2 Analytical solutions

The functions f, y_d on the right hand sides of (P_Γ^{test}) resp. the function g on the boundary have been adjusted such that the optimal control problem possesses the following analytical optimal state solution

$$\bar{y}(x_1, x_2) = x_2 \cdot (1 - x_2) \cdot \Big(x_1 - \frac{e^{(-(1-x_1)/0.01)} - e^{(-1/0.01)}}{(1 - e^{(-1/0.01)})} \Big)$$

resp. the analytical optimal adjoint solution

$$\bar{p}(x_1, x_2) = x_2 \cdot (1 - x_2) \cdot \Big((1 - x_1) - \frac{e^{(-x_1/0.01)} - e^{(-1/0.01)}}{(1 - e^{(-1/0.01)})} \Big).$$

The function g on the boundary has been defined by

$$g := \varepsilon \partial_n \bar{y} - \frac{\mathbf{b} \cdot n \cdot \bar{y}}{2} - \bar{u}.$$

Moreover, the desired control u_d is given by

$$u_d := \left(\frac{1}{\lambda} \bar{p} + \tilde{u} \right) \quad \text{a.e. on } \Gamma$$

with

$$\tilde{u}(x_1, x_2) := \mathbb{P}_{[u_a^\Gamma, u_b^\Gamma]}(\sin(\pi x_1) \cdot \sin(\pi x_2)).$$

Thus, the variational inequality in the optimality system is fulfilled by the optimal control

$$\bar{u} := \tilde{u}|_\Gamma = 0.$$

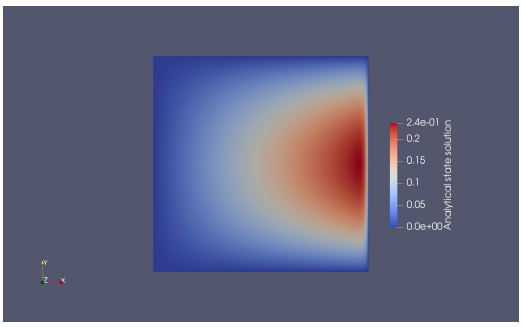


Figure 27: Analytical state solution $\bar{y}$

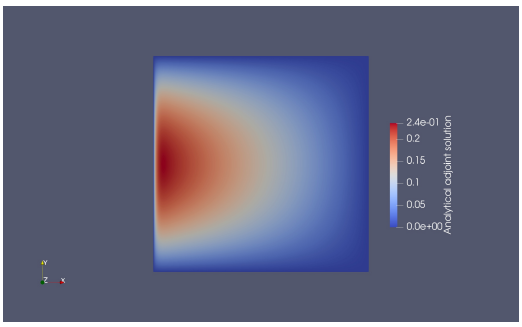


Figure 28: Analytical adjoint solution $\bar{p}$

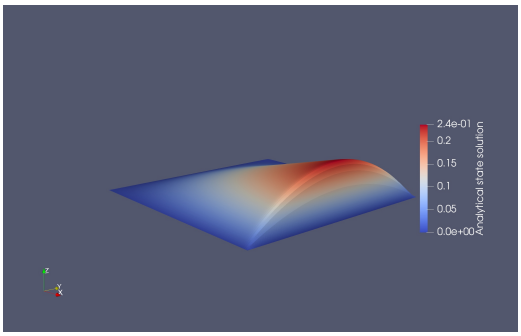


Figure 29: Analytical state solution $\bar{y}$

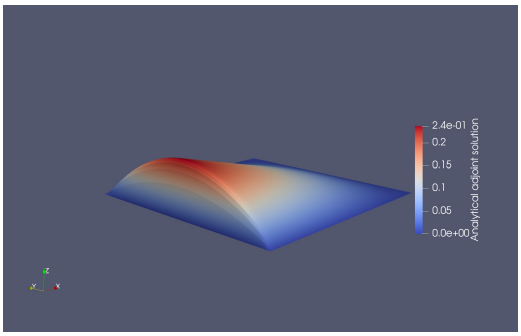


Figure 30: Analytical adjoint solution $\bar{p}$

8.4.3.3 Experimental order of convergence

In this section, we show the computed $L^2(\Omega)$-errors and $L^2(\Omega)$-convergence orders for the differences

$$e_h := \bar{y} - y_h$$
$$k_h := \bar{p} - p_h$$

where $\bar{y}, \bar{p}$ are the analytical optimal solutions and y_h, p_h are the computed AFC solutions corresponding to the state resp. the adjoint equation. Note that the focus of this test lies on the stabilizing effect of the AFC method for the state and the adjoint equation.

Table 9: $L^2(\Omega)$-errors / EOC (P_Γ^{test})

Grid level	$c_0 \parallel e_h \parallel_{0,\Omega}$	EOC $c_0 \parallel e_h \parallel_{0,\Omega}$	$c_0 \parallel k_h \parallel_{0,\Omega}$	EOC $c_0 \parallel k_h \parallel_{0,\Omega}$
1	2.2385e-01	-	2.3250e-01	-
2	6.6020e-02	1.76	6.5494e-02	1.83
3	1.7123e-02	1.95	1.6349e-02	2.00
4	8.7421e-03	0.97	8.4029e-03	0.96
5	5.6111e-03	0.64	5.4623e-03	0.62
6	2.9642e-03	0.92	2.9018e-03	0.91
7	1.4827e-03	1.00	1.4547e-03	1.00
8	7.8458e-04	0.92	7.7105e-04	0.92

8.4.3.4 AFC solutions

In the figures below, we plot the AFC solutions computed on a $(128, 128)$-unit square mesh. Firstly, the solutions possess no oscillations and secondly, a comparison with the analytical solutions Figure 29/Figure 30 leads us to the conclusion that the iterative solver computes accurate discrete solutions.

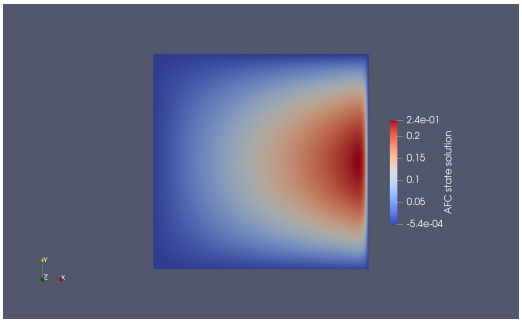


Figure 31: AFC state solution

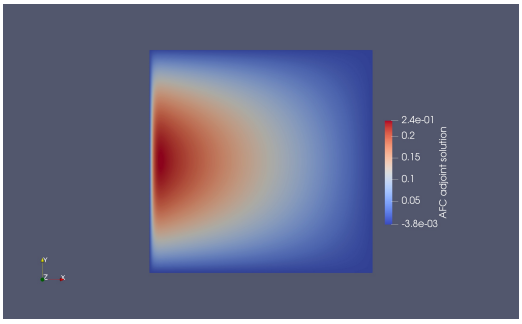


Figure 32: AFC adjoint solution

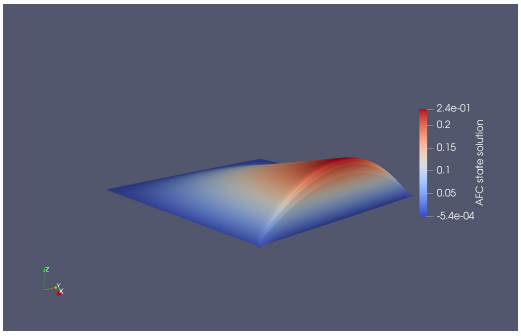


Figure 33: AFC state solution

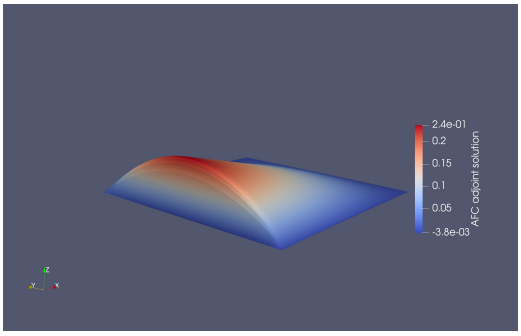


Figure 34: AFC adjoint solution

9 Further applications and an open problem

In the previous sections, we have investigated several optimal control problems where the AFC method has been applied in the context of the *optimize-then-discretize*-approach. Now, we see further optimal control problems where the existence of a discrete solution and corresponding

error estimates can be derived with the help of Section 7. We remark that the proofs for the existence of discrete solutions and the derivation of error estimates for the following optimal control problems are similar to the proofs provided in Section 8. Hence, we keep this section brief. In contrast to this we will show an unconstrained optimal control problem with Robin boundary control where the theory of Section 7 is not applicable.

9.1 State and control constrained case

First, let us introduce the following state and control constrained optimal control problem

$$\left.\begin{array}{l} \min\ J^{sc}(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = u \quad \text{in } \Omega \\ y = 0 \quad \text{on } \Gamma \\ u_a \le u \le u_b \quad \text{a.e. in } \Omega \\ y \le y_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_{sc})$$

Optimal control problems of type (P_{sc}) have been investigated for instance in [Cas86], [HtKu09], [HtKu17], or [KruRö08] where in [KruRö08], a state and control constrained Robin boundary control problem has been considered. However, the results in [KruRö08] can also be transferred to the case of distributed control, i.e. (P_{sc}). In contrast to Section 8.3, we consider control constraints, but we do not have any lower state constraints. For the application of the abstract results derived in Section 7, one can use as in the previous sections, the optimality system corresponding to a regularization of (P_{sc}). Following the strategy in [KruRö08], the authors regularize problem (P_{sc}) by using the so-called virtual control approach. The virtual control problem corresponding to (P_{sc}) reads as follows:

$$\left.\begin{array}{l} \min\ J^{VC}(y_\kappa,u_\kappa,v_\kappa) := \frac{1}{2} \parallel y_\kappa - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u_\kappa \parallel_{0,\Omega}^2 + \frac{\sigma(\kappa)}{2} \parallel v_\kappa \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y_\kappa + \mathbf{b}\cdot\nabla y_\kappa + cy_\kappa = u_\kappa \quad \text{in } \Omega \\ y_\kappa = 0 \quad \text{on } \Gamma \\ u_a \le u_\kappa \le u_b \quad \text{a.e. in } \Omega \\ y_\kappa \le y_b + \varsigma(\kappa)v_\kappa \quad \text{a.e. in } \Omega \end{array}\right\} (P_{sc}^{VC})$$

where $\kappa > 0$ is a regularization parameter and $\sigma(\kappa), \varsigma(\kappa)$ are positive and real valued functions. In an analogous way to [KruRö08, Corollary 2], one can prove the following $L^2(\Omega)$-error estimate

$$\lambda \parallel \bar{u}_\kappa - \bar{u} \parallel_{0,\Omega} + \parallel \bar{y}_\kappa - \bar{y} \parallel_{0,\Omega} \le C \left(\frac{\varsigma(\kappa)}{\sqrt{\sigma(\kappa)}} \right)^{\frac{1}{3}} \tag{9.1.1}$$

where $(\bar{y},\bar{u})$ is the optimal solution of (P_{sc}) and $(\bar{y}_\kappa,\bar{u}_\kappa,\bar{v}_\kappa)$ is the optimal solution of (P_{sc}^{VC}). Choosing $\sigma(\kappa), \varsigma(\kappa)$ so that

$$\lim_{\kappa\to\infty} \frac{\varsigma(\kappa)}{\sqrt{\sigma(\kappa)}} = 0$$

we get the following $L^2(\Omega)$-convergence

$$\begin{array}{l} \bar{y}_\kappa \to \bar{y} \quad \text{in } L^2(\Omega) \\ \bar{u}_\kappa \to \bar{u} \quad \text{in } L^2(\Omega). \end{array}$$

However, we are currently not able to apply the discretization technique provided in Section 6 or a modified technique to the virtual control problem (P_{sc}^{VC}). In Section 8.3, we have seen that the Moreau-Yosida regularization is an appropriate tool to construct a discrete solution for the state constrained optimal control problem (P_s). The Moreau-Yosida regularization of (P_{sc}) is given by

$$\left.\begin{array}{l} \min\ J_{sc}^{MY}(y_\delta, u_\delta) := \min \frac{1}{2} \parallel y_\delta - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u_\delta \parallel_{0,\Omega}^2 + \frac{\delta}{2} \parallel \max\{0, y_\delta - y_b\} \parallel_{0,\Omega}^2 \\ -\varepsilon\Delta y_\delta + \mathbf{b}\cdot\nabla y_\delta + c y_\delta = u_\delta \quad \text{in } \Omega \\ y_\delta = 0 \quad \text{on } \Gamma \\ u_a \le u_\delta \le u_b \quad \text{a.e. in } \Omega \end{array}\right\} (P_{sc}^{MY})$$

where $\delta > 0$ is a Moreau-Yosida regularization parameter that is taken large. Fortunately, under suitable assumptions on the regularization parameters $\delta, \sigma(\kappa), \varsigma(\kappa)$, the virtual control problem (P_{sc}^{VC}) coincides with the Moreau-Yosida regularization (P_{sc}^{MY}). In detail, setting $\delta = \delta(\kappa) := \frac{\sigma(\kappa)}{\varsigma^2(\kappa)}$, the optimal solution $(\bar{y}_\delta, \bar{u}_\delta)$ of (P_{sc}^{MY}) coincides with $(\bar{y}_\kappa, \bar{u}_\kappa)$ and by virtue of (9.1.1) we obtain

$$\lambda \parallel \bar{u}_\delta - \bar{u} \parallel_{0,\Omega} + \parallel \bar{y}_\delta - \bar{y} \parallel_{0,\Omega} \le C \left(\frac{1}{\sqrt{\delta}}\right)^{\frac{1}{3}}.$$

The combination of the discretization techniques provided in Section 8.2.1 and Section 8.3.1 leads us to a discretization technique for (P_{sc}^{MY}) and consequently to a discrete solution which approximates the optimal solution of (P_{sc}). In this context, we remark that similar to Section 6.3, the optimality system corresponding to (P_{sc}^{MY}) is the basis of the application of the *optimize-then-discretize*-approach. Corresponding $L^2(\Omega)$-error estimates for the control and the state can be derived as in the previous sections by the general L^2-error estimate provided in Lemma 7.6. Finally, one can prove that the following $L^2(\Omega)$-error estimate holds:

$$\parallel u_h - \bar{u}_\delta \parallel_{0,\Omega} + \parallel y_h - \bar{y}_\delta \parallel_{0,\Omega} \le C_\delta C_R h^{\frac{1}{2}} + C h^{\frac{1}{2}}$$

where $(y_h, u_h) \in V_{h,0} \times L^2(\Omega)$ is an AFC solution for $(\bar{y}_\delta, \bar{u}_\delta)$ and $C, C_\delta, C_R > 0$ are constants, independent of h.

9.2 Robin boundary control with boundary observation

Now, we consider two types of a pure Robin boundary control problem where the *optimize-then-discretize*-approach combined with the AFC method is applicable. First, we introduce the following unconstrained pure Robin boundary control problem

$$\left.\begin{array}{l} \min\ J^{\Gamma,2}(y,u) := \frac{1}{2} \parallel y - y_\Gamma \parallel_{0,\Gamma}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + c y = 0 \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} = u \quad \text{on } \Gamma \\ u \in U_{ad}^\Gamma \end{array}\right\} (P_\Gamma^2)$$

where $U_{ad}^\Gamma = L^2(\Gamma)$, $\lambda > 0$ and $y_\Gamma \in L^2(\Gamma)$. The starting point of the discretization of (P_Γ^2) in the context of the *optimize-then-discretize*-approach is again the optimality system. Then, similar to the case of (P_f), the theory of Section 7 will be applied to a regularized, discrete, and coupled system. We remark that one can regularize the discrete system in the same way as $(P_h^{f,rg})$, by dividing the adjoint equation with the Tikhonov parameter λ. Hence, the discretization concept, the results corresponding to the existence of a discrete solution, and the derivation of

L^2-error estimates for (P_h^f) can be transferred to the case of (P_Γ^2). Another boundary control problem where the theory of Section 7 is applicable is the following control constrained Robin boundary control problem

$$\left.\begin{array}{l} \min\ J^{\Gamma,3}(y,u) := \frac{1}{2} \parallel y - y_\Gamma \parallel_{0,\Gamma}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = 0 \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} = u \quad \text{on } \Gamma \\ u \in U_{ad}^\Gamma \end{array}\right\} (P_\Gamma^3)$$

where $\lambda > 0$, $y_\Gamma \in L^2(\Gamma)$ and

$$U_{ad}^\Gamma := \{u \in L^2(\Gamma) : u_a^\Gamma(x) \leq u(x) \leq u_b^\Gamma(x) \quad \text{a.e. on } \Gamma\}.$$

For the discretization of (P_Γ^3), one can use the technique of the control constrained optimal control problem (P_b) provided in Section 8.2.1. The structure of the weak formulation corresponding to the optimality systems of (P_b) resp. (P_Γ^3) are quite similar. The differences are the right hand sides of the state equation and the adjoint equation. In detail, the right hand sides of (P_b) are $L^2(\Omega)$-inner products and the right hand sides of (P_Γ^3) are $L^2(\Gamma)$-inner products. However, these differences do not influence the way of discretizing (P_Γ^3). Firstly, as in Section 8.2.1, one can follow the *optimize-then-discretize*-approach. Secondly, one can regularize the resulting coupled and discretized system by truncating the discrete state on the right hand side of the adjoint equation. A $L^\infty(\Gamma)$-a priori estimate of a discrete AFC state solution and an appropriate choice of the truncation parameter $k \in \mathbb{N}$ will lead us to a discrete solution for the initial coupled and discretized system. Finally, for (P_Γ^2) and (P_Γ^3), one can derive the following L^2-error estimates by the application of a modified version of the abstract L^2-estimate provided in Lemma 7.6

$$\parallel u_h - \bar{u} \parallel_{0,\Gamma} + \parallel y_h - \bar{y} \parallel_{0,\Gamma} \leq Ch^{\frac{1}{4}}$$

where $(\bar{y}, \bar{u})$ is the unique optimal solution of (P_Γ^2) resp. (P_Γ^3) and (y_h, u_h) is a AFC solution of the corresponding coupled and discretized systems. The constant C is independent of h. We mention that in contrast to the problems with distributed control, we obtain a convergence order of $\mathcal{O}(\frac{1}{4})$, since the application of the trace inequality and the inverse inequality reduces the order of such boundary control problems. For detailed information see also Theorem 8.46 in Section 8.4.

9.3 An open problem

In this section, we show an optimal control problem where the application of the theory in Section 7 is currently not possible. For this, let us consider the following unconstrained Robin boundary control problem

$$\left.\begin{array}{l} \min\ J^{\Gamma,4}(y,u) := \frac{1}{2} \parallel y - y_d \parallel_{0,\Omega}^2 + \frac{\lambda}{2} \parallel u \parallel_{0,\Gamma}^2 \\ -\varepsilon\Delta y + \mathbf{b}\cdot\nabla y + cy = 0 \quad \text{in } \Omega \\ \varepsilon\partial_n y - \frac{\mathbf{b}\cdot n\cdot y}{2} = u \quad \text{on } \Gamma \\ u \in U_{ad}^\Gamma \end{array}\right\} (P_\Gamma^4)$$

where $U_{ad}^\Gamma = L^2(\Gamma)$. Following the *optimize-then-discretize*-approach, leads us to the coupled and discretized system

$$\left.\begin{array}{l} a_\Gamma(y_h, v_h) + d_h^{s,\Gamma}(y_h; y_h, v_h) = (u_h, v_h)_\Gamma \quad \forall v_h \in V_h \\ a_\Gamma(\psi_h, p_h) + d_h^{ad,\Gamma}(p_h; p_h, \psi_h) = (y_h - y_d, \psi_h)_\Omega \quad \forall \psi_h \in V_h \\ u_h = -\frac{1}{\lambda} p_h \quad \text{a.e. on } \Gamma \end{array}\right\} (P_h^{\Gamma,4})$$

In contrast to the previous sections, we are not able to derive the uniform $L^2(\Omega)$-boundedness of y_h or p_h from $(P_h^{\Gamma,4})$. Hence, following the strategy provided in Section 8.4.1 the operator $Q : \mathcal{X}_h \to \mathcal{X}_h^*$ defined by

$$\langle Q(y_h, p_h), (v_h, \psi_h)\rangle_{\mathcal{X}_h^*, \mathcal{X}_h} := (-\frac{1}{\lambda} p_h, v_h)_\Gamma + (y_h - y_d, \psi_h)_\Omega$$

does not satisfy condition (7.1.2) in Assumption 7.2. Additionally, for proving the uniform $L^2(\Omega)$-boundedness of y_h or p_h, an appropriate regularization corresponding to $(P_h^{\Gamma,4})$ is currently not available. Hence, Lemma 7.3 cannot be applied to $(P_h^{\Gamma,4})$. Furthermore, the lack of the uniform $L^2(\Omega)$-boundedness of y_h and p_h has also an influence on the derivation of a L^2-error estimate for the state and the control. For instance, we have seen in Lemma 8.45 that the $H^2(\Omega)$-norm of the auxiliary solutions $\tilde{y}$ and $\tilde{p}$ is bounded by the L^2-norms of the corresponding right hand sides of the state and the adjoint equation. Due to the lack of the uniform L^2-boundedness of y_h and p_h, it is not ensured that the $H^2(\Omega)$-a priori error estimates of $\tilde{y}$ and $\tilde{p}$ do not depend on the mesh size h. In this case, the L^2-error estimate (see Theorem 8.46) is not useful. Thus, we cannot conclude that the discrete solution $(y_h, u_h) = (y_h, -\frac{1}{\lambda} p_h)$ approximate the optimal solution $(\bar{y}, \bar{u})$ of (P_Γ^4).

10 Summary

In this thesis, we have investigated the discretization of optimal control problems by applying the AFC method. We started this work with an introduction to function spaces and elementary results in Section 2, which have been used throughout this work. The focus of Section 3 was the analysis of an elliptic boundary value problem with two types of boundary conditions. Firstly, we have analyzed a convection-diffusion reaction equation with Dirichlet boundary conditions and secondly, we have provided the analysis of a convection-diffusion reaction equation with Robin boundary conditions. In detail, for both types of boundary conditions, the existence of a weak solution and the validity of $H^2(\Omega)$-regularity were proven. After that, in Section 4, we have provided the analysis of the continuous optimal control problems, which were the main objects of investigation in this work. Moreover, the Moreau-Yosida regularization was also investigated for the state constrained optimal control problem. We have seen that the optimal solution of the state constrained optimal control problem has been approximated by the solutions of the Moreau-Yosida regularization with a $L^2(\Omega)$-convergence order of $\mathcal{O}(\frac{1}{6})$. The Finite Element Method has been introduced in Section 5. In addition, we have also shown the construction and the analysis of an AFC scheme for a general linear boundary value problem. Recall that the goal of the construction of an AFC scheme is that the discrete solution should satisfy the discrete maximum prinicple such that spurious oscillations are prevented. For this, we have seen in Section 5 sufficient conditions for the discrete maximum principle. Moreover, according to Section 3, we have applied the AFC method on the introduced elliptic boundary value problems. Finally, we have seen the construction of the Kuzmin limiter and the BJK limiter. For both limiters, the corresponding AFC scheme possesses a discrete solution satisfying the discrete maximum principle. Moreover, for the BJK limiter it was shown that the linearity-preserving property holds. In Section 6, we have discretized the several optimal control problems with the AFC method. Due to the fact that the limiters are in general nonlinear and non-differentiable, we have used the *optimize-then-discretize*-approach. During this section, we have provided a discretization concept for optimal control problems with distributed control and for optimal control problems with Robin boundary control. Due to the lack of regularity of the measure arising in the optimality system of the state constrained optimal control problem, we have discretized the optimality system of the Moreau-Yosida regularization. Collecting the derived systems, we have seen that all systems have a similar structure. Hence, in Section 7, we have provided abstract results such that the existence of a discrete solution and corresponding L^2-error estimates can be proven for the derived coupled systems. The application of these results on the introduced optimal control problems has been demonstrated in Section 8. Here, we have investigated the existence of a discrete solution for the several coupled and discretized systems. Furthermore, we have verified L^2-error estimates between the AFC state solution and the optimal state solution respective between the computed AFC control and the optimal control corresponding to the continuous problem. In the case of distributed control, we have proved a $L^2(\Omega)$-convergence order of $\mathcal{O}(\frac{1}{2})$ where for the state constrained optimal control problem the convergence order of $\mathcal{O}(\frac{1}{2})$ only holds for the solution of the discretized Moreau-Yosida regularization. The derived $L^2(\Omega)$-error estimate between the solution of the state constrained optimal control and a solution of the discretized Moreau-Yosida regularization depends on the regularization parameter δ. Apart from the theoretical results, there were several numerical tests for proving the stabilizing effect of the AFC method and for reviewing the derived L^2-error estimates performed. In the numerical tests, we have solved the AFC systems by a preconditioned relaxed Richardson iteration on a $\Omega = [0,1] \times [0,1]$ unit square mesh. For the stabilization of the discrete solutions, we have used the BJK limiter. It was shown that every computed AFC state and AFC adjoint solutions are free of spurious oscillations. Moreover, a comparison to the analytical solutions led us to the conclusion that the iterative method

computed accurate AFC solutions with layers at the correct position. In addition, as we can see in the numerical tests of the mentioned AFC literature, the computed L^2-convergence rates are higher than the theoretical results predicted. In the last section, we have shown further optimal control problems which can be discretized and solved in a similar way as illustrated in the previous sections. For this, we have mentioned that the abstract results provided in Section 7 have to be modified at some points. However, the demonstrated strategy in the context of the *optimize-then-discretize*-approach still holds for investigating the optimal control problems provided in Section 9. In contrast to the problems where the derived theory is applicable, we have also seen an unconstrained optimal control problem with Robin boundary control. It was shortly demonstrated that the general results in Section 7 are not applicable since the uniform L^2-boundedness of the discrete solutions cannot be ensured. This leads us to the following conclusion with an outlook on possible further research work.

11 Conclusion and Outlook

The goal of this work was to apply the AFC methodology for the discretization of optimal control problems governed by a convection-diffusion reaction equation. As we have seen, the AFC method stabilizes the discrete solutions in the context of the *optimize-then-discretize*-approach. Especially in the unconstrained, the control constrained case and in the case of Robin boundary control we are convinced that the application of the AFC method is helpful to obtain stabilized and accurate solutions. The derived discretization concepts are quite easy to understand and the corresponding numerical calculation of the solutions does not require much effort. However, in the case of state constrained optimal control problems, we also see potential of improvement since an optimal parameter adjustment between the regularization parameter δ of the Moreau-Yosida regularization and the mesh size h is not derived yet. Finally, we have also the opinion that the non-differentiability of the method complicate the numerical treatment since many established solvers like Newton-solvers for optimal control problems cannot be used. As we have mentioned in the introduction of this thesis, the combination of AFC schemes with optimal control problem is currently not investigated by the community of optimal control problems. Since, the analysis of AFC schemes has been established recently in 2016, there are consequently many open problems for the optimal control and AFC community.

11.1 AFC

Due to the fact that the first papers concerning the theoretical analysis of the AFC methodology have been established in 2016, there are many problems which are currently of interest. As we have mentioned in the context of the numerical tests, we have seen higher convergence orders than our theorectical results predicted. A reason for this behavior could be that the limiters have been estimated too roughly by 1 (see Remark 8.11). Hence, it should be interesting to establish sharper estimates for the limiters. Another part of investigation is the development of further limiters respective modified limiters such that the DMP, the linearity-preserving, continuity or differentiability properties hold. Differentiable limiters (see [BadBon17]) are of huge interest so that efficient nonlinear solutions strategies can be applied. For detailed information, we refer also to [JhaJo19], [JhaJo20] or [Loh21]. During the last years, the works of Kuzmin [Kuz12, Kuz12/2, Kuz18] and Barrenechea et. al [BJK16, BBK17, BJK17] have provided basic design criteria of limiter such that discrete maximum prinicples and the continuity property are satisfied. However, the basic Kuzmin limiter introduced in Section 5.5.1 does not guarantee the DMP property on arbitrary meshes. The work of Knobloch [Knob21] provides a modified version of the basic Kuzmin limiter such that the DMP holds. Here, we can see that the

structure of the Finite Element meshes also influences the need of new resp. modified limiter. According to steady-state convection-diffusion reaction equations, an a priori error analysis has been provided in [BJKR18]. A counterpart to the a priori error analysis is the a posteriori error analysis where the quality of the computed discrete solutions can be evaluated. First results to a posteriori error estimators can be found for instance in [JhaTh20] or [Jha21]. Finally, we remark that according to the general Flux-correction methodology (see [Zal79]), there are not only stationary convection-diffusion reaction equations which are currently under investigation. For instance, in [JoKno21] the authors analyze an evolutionary convection-diffusion reaction equation: Find $w(x,t)$ such that

$$\frac{\partial w}{\partial t} - \varepsilon \Delta w + \mathbf{b} \cdot \nabla w + cw = g \quad \text{in } \Omega \times (0,T]$$

is satisfied where $(0,T]$ is the time interval and $\Omega \subseteq \mathbb{R}^d$ with $d = 2,3$ a bounded polygonal or polyhedral domain with Lipschitz-continuous boundary Γ. Note that in [JoKno21], Neumann and Dirichlet boundary conditions are prescribed. Furthermore, the limit case of the diffusion coefficient, i.e. $\varepsilon = 0$ is analyzed in the works of Kuzmin [Kuz12, Kuz18] where the author often investigates time-dependent transport equations: Find $w(x,t)$ such that

$$\frac{\partial w}{\partial t} + \nabla \cdot (\mathbf{v} w) = g \quad \text{in } \Omega \times (0,T]$$

where $\mathbf{v} = \mathbf{v}(x,t)$ is a velocity field. A compact overview on the application of limiting strategies can be found for instance in [LohSP19] where stationary and time-dependent advection-reaction equations have been considered. In addition, the author extends the application of limiting to an advection-reaction equation for symmetric tensor quantities.

11.2 Optimal control theory

In this work, we have seen that the *optimize-then-discretize*-approach is a possible strategy for discretizing several optimal control problems. Especially, for general nonlinear and non-differentiable discretization methods like the AFC method the *optimize-then-discretize*-approach is a useful choice. In the theory of optimal control, the problems are often discretized by the *discretize-then-optimize*-approach where the existence of discrete solutions is usually verified by standard techniques in contrast to the *optimize-then-discretize*-approach. In this work, all derived discrete systems have undergone some regularizations before the existence of discrete solutions could be proved. Due to the fact that there are many different types of optimal control problems (see Section 9), one has to develop further regularizations such that the *optimize-then-discretize*-approach is applicable. However, in the state constrained case, we have seen that a direct application of the *optimize-then-discretize*-approach is currently not possible. For the approximation of the optimal solution corresponding to the state constrained optimal control problem, we have used the coupled and discretized system corresponding to the Moreau-Yosida regularization. Currently, we cannot prove an optimal parameter adjustment for the $L^2(\Omega)$-error estimate provided in Theorem 8.34. Hence, one would like to derive other concepts for discretizing state constrained optimal control problems respective for discretizing the regularized counterparts. In addition to this, we remark that there exist many regularizations of state constrained optimal control problems. For instance, we have the so-called Lavrientiev regularization where the application on a state constrained optimal control problem has been demonstrated in [CheRö09]. For this, it will be interesting how to discretize this regularization in the framework of the *optimize-then-discretize*-approach. According to the current state of research, the application of the AFC method on parabolic optimal control problems or the derivation of a posteriori error estimates of AFC-discretized optimal control problems can be

investigated in future work. Due to the fact that the combination of optimal control problems with the AFC methodology is completely new, we are convinced that future work will arise many, currently unknown but nevertheless relevant problems.

References

[Ada75] Adams, R.A., Sobolev spaces. Academic Press, New York-London, Pure and Applied Mathematics, Vol. 65, 1975.

[Alt06] Alt, H.W., Lineare Funktionalanalysis. Springer, Berlin, 5. überarbeitete Auflage, 2006.

[BadBon17] Badia, S., Bonilla, J., Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. Comput. Methods Appl. Mech. Eng. 313, 133–158, 2017.

[BJKR18] Barrenechea, G.R., John, V., Knobloch, P., Rankin, R., A unified analysis of algebraic flux correction schemes for convection–diffusion equations. SeMA Journal 75, 655-685, 2018.

[BBK17] Barrenechea, G.R., Burman, E., Karakatsani, F., Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes. Numer. Math. 135(2), 521–545, 2017.

[BJK17] Barrenechea, G.R., John, V., Knobloch, P., An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. Math. Models Methods Appl. Sci. 27(3), 525–548, 2017.

[BJK16] Barrenechea, G.R., John, V., Knobloch, P., Analysis of algebraic flux correction schemes. SIAM J. Numer. Anal. Vol. 54, No. 4, 2427-2451, 2016.

[BJK15] Barrenechea, G.R., John, V., Knobloch, P., Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension. IMA J. Numer. Anal. 35(4), 1729–1756, 2015.

[BecVe07] Becker, R., Vexler, B., Optimal control of the convection-diffusion equation using stabilized Finite Element Methods. Numer. Math., 106(3), 349-367, 2007.

[Braa09] Braack, M., Optimal control in fluid mechanics by finite elements with symmetric stabilization. SIAM J. Control Optim. 48, no. 2, 672 - 687, 2009.

[BeLöf76] Bergh, J., Löfström, J., Interpolation Spaces, An Introduction. Springer, Berlin, 1976.

[BreSco02] Brenner, S., Scott, R., The Mathematical Theory of Finite Elements. 2nd ed., Springer, New York, 2002.

[Brez11] Brezis, H., Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer, New York, 2011.

[BroHug82] Brooks, A.N., Hughes, T.J.R., Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. Comput. Methods Appl. Mech. Eng. 32(1–3), 199–259 (1982). FENOMECH '81, Part I (Stuttgart, 1981).

[BurHan04] Burman, E., Hansbo, P., Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems. Comput. Methods Appl. Mech. Eng. 193(15–16), 1437–1453, 2004.

[Cas86] Casas, E., Control of an elliptic problem with pointwise state constraints. SIAM J. Cont. Optim. 24, 1309–1322, 1986.

[Cas93] Casas, E., Boundary control of semilinear elliptic equations with pointwise state constraints. SIAM J. Control Optim. 31, no. 4, 993–1006, 1993.

[CheRö09] Cherednichenko, S., Rösch, A., Error estimates for the discretization of elliptic control problems with pointwise control and state constraints. Comput. Optim. Appl., 44(1):27–55, 2009.

[Ciar02] Ciarlet, P.G., The finite element method for elliptic problems. SIAM, Philadelphia, PA., 40, 2002.

[ColHei02] Collis, S.S., Heinkenschloss, M., Analysis of the Streamline Upwind/Petrov Galerkin Method applied to the solution of optimal control problems. CAAM TR02- 01, 2002.

[Dha12] Dhamo, V., Optimal boundary control of quasilinear elliptic partial differential equations: theory and numerical analysis. PhD thesis. TU Berlin, 2012.

[DPE11] Di Pietro, D.A., Ern., A., Mathematical aspects of discontinuous Galerkin methods. Vol. 69. Springer Science and Business Media, 2011.

[Evans98] Evans, L.C., Partial Differential Equations. Volume 19. American Mathematical Society, Providence, Rhode Island, 1998.

[GirRav86] Girault, V., Raviart, P.A., Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms. Berlin-Heidelberg-New York-Tokyo, Springer Verlag, 1986.

[Gris85] Grisvard, P., Elliptic Problems in nonsmooth domains. Pitmann. Boston, 1985.

[Guer99] Guermond, J.L., Stabilization of Galerkin approximations of transport equations by subgrid modeling. Model. Math. Anal. Numer., 36, 1293–1316, 1999.

[HeiLy10] Heinkenschloss, M., Leykekhman, D., Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems. SIAM Journal on Numerical Analysis, Vol. 47, No. 6, 4607-4638, 2010.

[HtKu09] Hintermüller, M., Kunisch, K., Pde-constrained optimization subject to pointwise constraints on the control, the state and its derivative. SIAM J. Optim. 20(3), 1133–1156, 2009.

[HtKu17] Hintermüller, M., Kunisch, K., Feasible and non-interior path-following in constrained minimization with low multiplier regularity. SIAM J. Control Optim., 45(4): 1198–1221, 2017.

[HzHt09] Hinze, M., Hintermüller, M., Moreau–Yosida Regularization in State Constrained Elliptic Control Problems: Error Estimates and Parameter Adjustment. SIAM J. Numer. Anal., 47(3), 1666–1683, 2009.

[HzRö12] Hinze, M., Rösch, A., Discretization of optimal control problems. In: G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, and M. Ulbrich, and S. Ulbrich, editors, Constrained optimization and optimal control for partial differential equations, volume 160 of International Series of Numerical Mathematics, 391–430. Birkhäuser/Springer Basel AG, Basel, 2012.

[HzSchi11] Hinze, M., Schiela, A., Discretization of interior point methods for state constrained elliptic optimal control problems: optimal error estimates and parameter adjustment. Comput. Optim. Appl. 48(3): 581-600, 2011.

[HzYaZh09] Hinze, M., Yan, N., Zhou, Z., Variational discretization for optimal control governed by convection dominated diffusion equations. J. Comput. Math., 27, Nr. 2-3, 237-253, 2009.

[JeKe81] Jerison, D., Kenig, C., The Neumann problem on Lipschitz domains. Bull. Amer. Math.Soc., Vol. 4, Nr.2, 203–207, 1981.

[JhaJo19] Jha, A., John, V., A study of solvers for nonlinear AFC discretizations of convection-diffusion equations. Comput. Math. Appl. 78, no. 9, 3117–3138, 2019.

[JhaJo20] Jha, A., John, V., On basic iteration schemes for nonlinear AFC discretizations. Boundary and interior layers, computational and asymptotic methods - BAIL 2018, 113–128, Lect. Notes Comput. Sci. Eng., 135, Springer, Cham, 2020.

[JhaTh20] Jha, A., Numerical Algorithms for Algebraic Stabilizations of Scalar Convection-Dominated Problems. Thesis (D.Sc.)–Freie Universität Berlin (Germany). 183 pp. ISBN: 979-8698-55207-9, ProQuest LLC, 2020.

[Jha21] Jha, A., A residual based a posteriori error estimators for AFC schemes for convection-diffusion equations. Comput. Math. Appl. 97, 86–99, 2021.

[JoKno07] John, V., Knobloch, P., On spurious oscillations at layers diminishing (SOLD) methods for convection– diffusion equations. I. A review. Comput. Methods Appl. Mech. Eng. 196(17–20), 2197–2215, 2007.

[JoKno08] John, V., Knobloch, P., On spurious oscillations at layers diminishing (SOLD) methods for convection– diffusion equations. II. Analysis for P1 and Q1 finite elements. Comput. Methods Appl. Mech. Eng. 197(21–24), 1997–2014, 2008.

[JoKno21] John, V., Knobloch, P., Existence of solutions of a finite element flux-corrected-transport scheme. Appl. Math. Lett. 115, Paper No. 106932, 6 pp., 2021.

[KinSta80] Kinderlehrer, D., Stampacchia, G., An introduction to variational inequalities and their applications. Vol. 88. Pure and Applied Mathematics. Academic Press, New York, 1980.

[Knob19] Knobloch, P., A Linearity Preserving Algebraic Flux Correction Scheme of Upwind Type Satisfying the Discrete Maximum Principle on Arbitrary Meshes. Numerical Mathematics and Advanced Applications ENUMATH 2017. Springer International Publishing, 2019.

[Knob21] Knobloch, P., A new algebraically stabilized method for convection-diffusion-reaction equations. Numerical mathematics and advanced applications ENUMATH 2019, 605–613, Lect. Notes Comput. Sci. Eng., 139, Springer, Cham, 2021.

[Kuz06] Kuzmin, D., On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. J. Comput. Phys., 219(2):513–531, 2006.

[Kuz10] Kuzmin, D., A Guide to Numerical Methods for Transport Equations. url: http://www.mathematik.uni-dortmund.de/ kuzmin/Transport.pdf, 2010.

[Kuz12] Kuzmin, D., Algebraic Flux Correction I. Scalar Conservation Laws. In: Flux-Corrected Transport. Ed. by D. Kuzmin, R. Löhner, and S. Turek. Scientific Computation. Springer Netherlands, 145–192, 2012.

[Kuz12/2] Kuzmin, D., Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. In: Journal of Computational and Applied Mathematics 236.9, 2317–2337, 2012.

[Kuz18] Kuzmin, D., Gradient–based limiting and stabilization of continuous Galerkin methods. Tech. rep. Ergebnisberichte des Instituts für Angewandte Mathematik, Nummer 589. Fakultät für Mathematik, TU Dortmund, 07/2018.

[Kru08] Krumbiegel, K., Numerical concepts and error analysis for elliptic Neumann boundary control problems with pointwise state and control constraints. PhD thesis, 2008.

[KruRö08] Krumbiegel, K., Rösch, A., On the regularization error of state constrained Neumann control problems. Control Cybern. 37(2), 369–392, 2008.

[Lions71] Lions, J.L., Optimal Control of Systems Governed by Partial Differential Equations. Springer-Verlag, Berlin, 1971.

[LohSP19] Lohmann, C., Physics-compatible finite element methods for scalar and tensorial advection problems. Springer, 2019.

[Loh19] Lohmann, C., On the solvability and iterative solution of algebraic flux correction problems for convection-reaction equations. Ergebnisberichte des Instituts für Angewandte Mathematik Nummer 612, Fakultät für Mathematik, TU Dortmund, 612, 2019.

[Loh21] Lohmann, C., An algebraic flux correction scheme facilitating the use of Newton-like solution strategies. Comput. Math. Appl. 84, 56–76, 2021.

[MyRöTr06] Meyer, C., Rösch, A., Tröltzsch, F., Optimal control of PDEs with regularized pointwise state constraints. Comput. Optim. Appl. 33, 209–28, 2006.

[MyPrTr07] Meyer, C., Prüfert, U., Tröltzsch, F., On two numerical methods for state-constrained elliptic control problems. Optim. Methods Softw. 22(6), 871-899, 2007.

[Nec12] Nečas, J., Direct Methods in the Theory of Elliptic Equations. Springer, Heidelberg, 2012.

[NePfRö15] Neitzel, I., Pfefferer, J., Rösch, A., Finite element discretization of state-constrained elliptic optimal control problems with semilinear state equation. SIAM J. Control Optim., 53(2):874–904, 2015.

[Pfeff15] Pfefferer, J., Numerical analysis for elliptic Neumann boundary control problems on polygonal domains. PhD thesis, 2015.

[Trie78] Triebel, H., Theory of function spaces, differential operators. North-Holland Publ. Co., Amsterdam, 1978.

[Troel] Tröltzsch, F., Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen. Vieweg + Teubner, Wiesbaden, 2005.

[Wlok92] Wloka, J., Thomas, C. B., Thomas, M. J., Partial Differential Equations. Cambridge, England, Cambridge University Press, 1992.

[Wink20] Winkler, M., Error estimates for the finite element approximation of bilinear boundary control problems. Computational Optimization and Applications (COAP), 76(1), 155-199, 2020.

[YuBe14] Yücel, H., Benner, P., Adaptive discontinuous Galerkin methods for state constrained optimal control problems governed by convection diffusion equations. Computational Optimization and Applications. 62, 2014.

[YaZh09] Yan, N., Zhou, Z., A priori and a posteriori error analysis of edge stabilization Galerkin method for the optimal control problem governed by convection dominated diffusion equation. Journal of Computational and Applied Mathematics 223, 198–217, 2009.

[Zal79] Zalesak, S.T., Fully multidimensional flux-corrected transport algorithms for fluids. J. Comput. Phys. 31(3), 335–362, 1979.

[ZowKur79] Zowe, J., Kurcyusz, S., Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. 5(1), 49–62, 1979.