UNIVERSITÄT KOBLENZ · LANDAU

Mathematisches Institut

Ph. D. Thesis

Sparse discretization of sparse control problems with measures

von

Evelyn Christin Herberg
aus Mölln

Angenommene Dissertation zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
Fachbereich 3: Mathematik/Naturwissenschaften
Universität Koblenz-Landau

Gutachter:
Prof. Dr. Michael Hinze
Prof. Dr. Christian Clason

Prüfungskommission:
Prof. Dr.-Ing. Dietrich Paulus
Prof. Dr. Michael Hinze
Prof. Dr. Thomas Götz

Tag der mündlichen Prüfung: 11. Juni, 2021

Whether you think you can or think you can’t - you are right.

Abstract

We consider variational discretization of three different optimal control problems.

The first being a parabolic optimal control problem governed by space-time measure controls. This problem has a nice sparsity structure, which motivates our aim to achieve maximal sparsity on the discrete level. Due to the measures on the right hand side of the partial differential equation, we consider a very weak solution theory for the state equation and need an embedding into the continuous functions for the pairings to make sense. Furthermore, we employ Fenchel duality to formulate the predual problem and give results on solution theory of both the predual and the primal problem. Later on, the duality is also helpful for the derivation of algorithms, since the predual problem can be differentiated twice so that we can apply a semismooth Newton method. We then retrieve the optimal control by duality relations.

For the state discretization we use a Petrov-Galerkin method employing piecewise constant states and piecewise linear and continuous test functions in time. For the space discretization we choose piecewise linear and continuous functions. As a result the controls are composed of Dirac measures in space-time, centered at points on the discrete space-time grid. We prove that the optimal discrete states and controls converge strongly in L^q and weakly-$*$ in $\mathcal{M}$, respectively, to their smooth counterparts, where $q \in (1, \min\{2, 1 + 2/d\}]$ is the spatial dimension. The variational discrete version of the state equation with the above choice of spaces yields a Crank-Nicolson time stepping scheme with half a Rannacher smoothing step.

Furthermore, we compare our approach to a full discretization of the corresponding control problem, precisely a discontinuous Galerkin method for the state discretization, where the discrete controls are piecewise constant in time and Dirac measures in space. Numerical experiments highlight the sparsity features of our discrete approach and verify the convergence results.

The second problem we analyze is a parabolic optimal control problem governed by bounded initial measure controls. Here, the cost functional consists of a tracking term corresponding to the observation of the state at final time. Instead of a regularization term for the control in the cost functional, we consider a bound on the measure norm of the initial control. As in the first problem we observe a sparsity structure, but here the control resides only in space at initial time, so we focus on the space discretization to achieve maximal sparsity of the control. Again, due to the initial measure in the partial differential equation, we rely on a very weak solution theory of the state equation.

We employ a dG(0) approximation of the state equation, i.e. we choose piecewise linear and continuous functions in space, which are piecewise constant in time for our ansatz and test space. Then, the variational discretization of the problem together with the optimality conditions induce maximal discrete sparsity of the initial control, i.e. Dirac measures in space. We present numerical experiments to illustrate our approach and investigate the sparsity structure

As third problem we choose an elliptic optimal control governed by functions of bounded variation (BV) in one space dimension. The cost functional consists of a tracking term for the state and a BV-seminorm in terms of the derivative of the control. We derive a sparsity structure for the derivative of the BV control. Additionally, we utilize the mixed formulation for the state equation.

A variational discretization approach with piecewise constant discretization of the state and piecewise linear and continuous discretization of the adjoint state yields that the derivative of the control is a sum of Dirac measures. Consequently the control is a piecewise constant function. Under a structural assumption we even get that the number of jumps of the control is finite. We prove error estimates for the variational discretization approach in combination with the mixed formulation of the state equation and confirm our findings in numerical experiments that display the convergence rate.

In summary we confirm the use of variational discretization for optimal control problems with measures that inherit a sparsity. We are able to preserve the sparsity on the discrete level without discretizing the control variable.

Zusammenfassung

Wir betrachten variationelle Diskretisierung angewandt auf drei verschiedene Optimalsteuerungsprobleme.

Das erste Problem ist ein parabolisches Optimalsteuerungsproblem gesteuert durch Raum-Zeit Maßkontrollen. Dieses Problem hat eine Dünnbesetztheitseigenschaft, die unser Ziel motiviert, maximale Dünnbesetztheit auf der diskreten Ebene zu erhalten. Da auf der rechten Seite der partiellen Differentialgleichung Maße auftauchen, betrachten wir eine sehr schwache Lösungstheorie für die Zustandsgleichung und benötigen eine Einbettung in die stetigen Funktionen, damit die Paarungen sinnvoll sind. Des Weiteren nutzen wir Fenchel Dualität, um das präduale Problem zu formulieren und Resultate zur Lösungstheorie des prädualen und des primalen Problems anzugeben. Später ist die Dualität auch hilfreich um Algorithmen herzuleiten, da das präduale Problem zweimal differenzierbar ist, wodurch die semiglatte Newton Methode angewendet werden kann. Wir erhalten die optimale Kontrolle dann durch die Dualitäts-Relationen. Für die Diskretisierung des Zustands nutzen wir eine Petrov-Galerkin Methode mit stückweise konstanten Zuständen und stückweise linearen und stetigen Testfunktionen in der Zeit. Für die räumliche Diskretisierung wählen wir stückweise lineare und stetige Funktionen. Daraus resultierend bestehen die Kontrollen aus Diracmaßen in der Raum-Zeit, welche in den Gitterpunkten des diskreten Raum-Zeit Gitters zentriert sind. Wir beweisen, dass die optimalen diskreten Zustände und Kontrollen jeweils stark in L^q und schwach-$*$ in $\mathcal{M}$ zu ihren stetigen Gegenstücken konvergieren, wobei $q \in (1, \min\{2, 1 + 2/d\}]$ die räumliche Dimension ist. Die variationell diskrete Version der Zustandsgleichung mit der obigen Wahl der Räume ergibt ein Crank-Nicolson Zeitschrittschema mit einem halben Rannacher Glättungsschritt. Außerdem vergleichen wir unseren Ansatz mit einer vollen Diskretisierung des entsprechenden Kontrollproblems, genauer einer diskontinuierlichen Galerkin Methode für die Diskretisierung des Zustands, bei der die diskreten Kontrollen stückweise konstant in der Zeit und Diracmaße im Raum sind. Numerische Experimente verdeutlichen die Dünnbesetztheitseigenschaften unseres diskreten Ansatzes und verifizieren die Konvergenzresultate.

Das zweite Problem, welches wir analysieren, ist ein parabolisches Optimalsteuerungsproblem gesteuert durch beschränkte Anfangswert-Maßkontrollen. Hier besteht das Kostenfunktional aus einem Term zur Überwachung des Zustands zum finalen Zeitpunkt. Anstelle eines Regularisierungsterms für die Kontrolle im Kostenfunktional betrachten wir eine Beschränkung der Maßnorm der Anfangswertkontrolle. Wie im ersten Problem beobachten wir eine Dünnbesetztheitseigenschaft, doch hier existiert die Kontrolle nur im Raum zur Anfangszeit, sodass wir uns auf die Diskretisierung des Raumes fokussieren, um maximale Dünnbesetztheit der Kontrolle zu erreichen. Wieder betrachten wir eine sehr schwache Lösungstheorie der Zustandsgleichung wegen des Maßes als Anfangswert in der partiellen Differentialgleichung. Wir nutzen eine dG(0) Approximation der Zustandsgleichung, d.h. wir wählen stückweise lineare und stetige Funktion im Raum, die stückweise konstant in der Zeit sind für unsere Ansatz- und Testfunktionen. Die variationalle Diskretisierung des Problems zusammen mit den Optimalitätsbedingungen induziert maximale diskrete Dünnbesetztheit der Anfangswertkontrollen, d.h. Diracmaße im Raum. Wir präsentieren numerische Experimente um unseren Ansatz zu illustrieren und die Dünnbesetztheitsstruktur zu untersuchen.

Als drittes Problem wählen wir ein elliptisches Optimalsteuerungsproblem gesteuert durch Funktionen mit beschränkter Variation (BV) in einer Raumdimension. Das Kostenfunktional besteht aus einem Zustands-Überwachungsterm und einer BV-Seminorm für die Ableitung der Kontrolle. Wir leiten eine Dünnbesetztheitsstruktur für die Ableitung der BV Kontrolle her. Zusätzlich nutzen wir die gemischte Formulierung der Zustandsgleichung. Ein variationeller Diskretisierungsansatz mit stückweise konstanter Diskretisierung des Zustands und stückweise linearer und stetiger Diskretisierung des adjungierten Zustands liefert, dass die Ableitung der Kontrolle eine Summe von Diracmaßen ist. Infolgedessen ist die Kontrolle eine stückweise konstante Funktion. Unter strukturellen Annahmen erhalten wir sogar, dass die Anzahl der Sprünge der Kontrolle endlich ist. Wir beweisen Fehlerschätzer für die variationelle Diskretisierung in Kombination mit der gemischten Formulierung der Zustandsgleichung und bestätigen unsere Erkenntnisse in numerischen Experimenten, die die Konvergenzrate zeigen.

Zusammenfassend verifizieren wir den Nutzen der variationellen Diskretisierung für Optimalsteuerungsprobleme mit Maßen, welche eine Dünnbesetztheitseigenschaft aufweisen. Wir sind in der Lage die Dünnbesetztheit auf der diskreten Ebene zu erhalten, ohne die Kontrollvariable zu diskretisieren.

Publications

Some of the results of this thesis have already been published or submitted.

- **Section 3** is an extended version of

  [47] E. Herberg, M. Hinze, and H. Schumacher. "Maximal discrete sparsity in parabolic optimal control with measures". In: *Mathematical Control and Related Fields* 10.4 (Dec. 2020), pp. 735-759.

  Some of the results in the article [47] are based on the authors master thesis [45] with the title "Variational discretization of parabolic control problemsin space-time measure spaces". We hereafter clarify the improvements made in [47] as enhancement of [45]:

  Throughout the whole work the definition of the control space was corrected, so that the solvability of the state equation can be guaranteed. More details on said solvability of the state equation are added in [47, Section 2.1.]. Furthermore, the Fenchel duality was only discussed for the discretized problems in [45], while in [47, section 2.2.] this has been generalized to the continuous setting and then applied to discretized problems. Also, in [47, Section 2.2.] a linear operator that embeds into the space of continuous functions has been introduced, to make sense of the dual pairing with the controls living in a measure space. This operator - in the respective discrete setting - has also been added in [47, Section 3, Section 4]. Lastly, the computational results of [45] have been completely redone, new examples were examined and a convergence analysis added in [47, Section 5].

- **Section 4** is an extended version of

  [46] E. Herberg and M. Hinze. "Variational discretization approach applied to an optimal control problem with bounded measure controls", *arXiv preprint arXiv:2003.14380* (2020), which has been accepted for publication in the Radon book series.

These collaborations are an essential part of the research that led to this thesis.

Danksagung

Diese Arbeit wäre nicht möglich gewesen ohne die vielfältige Unterstützung, die ich während ihrer Entstehung bekommen habe. Ich möchte diese Möglichkeit nutzen, mich bei einigen Menschen explizit zu bedanken.

Zu aller erst bedanke ich mich bei meinem Betreuer Herrn Prof. Dr. Michael Hinze, der sich immer Zeit für mich genommen hat und mir stets gesagt hat, ich solle "die Ohren steifhalten", wenn es mal nicht so lief, wie ich mir das vorgestellt habe. Sowohl deine wissenschaftliche als auch deine menschliche Betreuung, vor allem nach dem Umzug nach Koblenz, hat mir sehr geholfen. Des Weiteren möchte ich mich dafür bedanken, dass du mir Verantwortung übertragen und so fortwährend die Möglichkeit gegeben hast, mich weiterzuentwickeln.

Außerdem habe ich viele tolle Kollegen in Hamburg und Koblenz kennengelernt, die meine Promotion spannender, witziger und lehrreicher gemacht haben.
Danke Henrik, für deine Unterstützung und die Zusammenarbeit.
Danke Carmen, dass du mir alles erklärt hast. Während der gemeinsamen Planung der Sommerschule habe ich so viel von dir gelernt - vor allem Ruhe zu bewahren.
Danke Christina, dass du immer als Mentorin, Vorbild und Freundin für mich da bist.
Danke Denis, dass wir zusammen die Challenge Umzug gemeistert haben und für viele tiefgründige Gespräche.
Danke Christian, für deine Expertise, dein offenes Ohr und deine Geduld.
Danke auch an alle anderen, die ich hier nicht namentlich erwähnt habe, aber die wissen, dass ich mich über Gespräche in den Kaffeepausen oder beim Mittagessen jedes Mal sehr gefreut habe.

Und schlussendlich danke ich meiner Familie, die immer an mich glauben und mir stets den Rücken freihalten.
Danke für einen Ort, an den ich jederzeit heimkommen kann, für guten Rat, für leckeres Essen und Spaziergänge.
Danke Mama und Papa, für all die Antworten auf meine unzähligen Fragen - schon mein Leben lang.
Danke Caro, dass du in so vielen Momenten in den letzten Jahren einfach da warst.
Ich könnte noch hunderte weitere Dinge aufzählen, aber ich fasse es wie folgt zusammen:
Danke Mama, Papa und Caro für alles.

Contents

Nomenclature **vii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Structure . . . 1
1.3 Literature overview and novelty of this work . . . 2

2 Mathematical background **5**
2.1 Functional analysis . . . 5
2.2 Optimization . . . 10
2.2.1 Optimal control . . . 12

3 Parabolic optimal control governed by space-time measure controls **15**
3.1 Problem Formulation . . . 15
3.2 Continuous optimality system . . . 17
3.2.1 State equation . . . 17
3.2.2 Fenchel duality . . . 18
3.2.3 Sparsity structure . . . 21
3.3 Variational discretization . . . 25
3.4 Discontinuous Galerkin discretization . . . 39
3.5 Computational results . . . 40

4 Parabolic optimal control governed by bounded initial measure controls **47**
4.1 Problem formulation . . . 47
4.2 Continuous optimality system . . . 48
4.3 Variational discretization . . . 54
4.4 Computational results . . . 59
4.4.1 Positive sources (problem $(P^{+}_{\alpha,\sigma})$) . . . 59
4.4.2 The general case (problem $(P_{\alpha,\sigma})$) . . . 65

5 Elliptic optimal control governed by functions of bounded variation **75**
5.1 Problem formulation . . . 75
5.2 Continuous optimality system . . . 75
5.3 Variational discretization . . . 80
5.3.1 Error estimates . . . 84
5.4 Computational results . . . 89
5.4.1 Semismooth Newton method . . . 89
5.4.2 Optimization algorithm . . . 92
5.4.3 Numerical Examples . . . 93

6 Conclusion **97**

A Appendix **99**
A.1 Density argument 1 99
A.2 Density argument 2 100
A.3 Fourier modes 101

Nomenclature

$\mathbb{R}$	field of real numbers – p. 5
$\mathbb{R}^n$	space of real n-dimensional vectors – p. 5
$\bar{\mathbb{R}}$	field of real numbers joined with $\{\infty\}$ – p. 18
$\mathbb{N}_0$	set of natural numbers including 0 – p. 9
$\mathbb{N}$	set of natural numbers – p. 32
$\emptyset$	empty set – p. 6
$\|\cdot\|_X$	norm of the normed space X – p. 5
$\bar{\Omega}$	closure of a set Ω in a metric space – p. 5
$(\cdot,\cdot)_H$	inner product of vector space H – p. 6
$\|\cdot\|_{X,Y}$	operator norm for an operator that maps from X to Y – p. 6
$dF(x;h)$	directional derivative of F in direction h at x – p. 6
X^*	dual space of X – p. 7
$\langle\cdot,\cdot\rangle_{X^*,X}$	dual pairing of X^* and X – p. 7
$\partial\Omega$	boundary of a set Ω in a metric space – p. 9
$B(x;r)$	open ball around $x \in \Omega$ with radius r – p. 9
$D^\alpha u$	α-th weak partial derivative of u – p. 9
$\partial^{cl}G(x)$	Clarke's generalized Jacobian of G – p. 11
conv	convex hull – p. 11
supp	support – p. 16
sgn	signum – p. 20
diam	diameter – p. 25
span	span – p. 25
div	divergence – p. 48
Lip	Lipschitz constant – p. 99
$\mathcal{B}(\Omega)$	Borel field of Ω – p. 8
$BV(\Omega)$	set of all functions of bounded variation on Ω – p. 10
$C(\Omega)$	space of continuous functions on Ω – p. 5
$C_c(\Omega)$	space of continuous functions on Ω whose support is compact – p. 8
$C_0(\Omega)$	space of continuous functions on Ω which vanish at infinity – p. 8
$C^k(\mathbb{R}^{n-1})$	space of continuous functions on $\mathbb{R}^{n-1}$ which are k times continuously differentiable – p. 9

$\mathcal{L}(X, Y)$	space of linear operators – p. 6
$L^p(\Omega)$	Lebesgue space of equivalence classes of functions from Ω to $\mathbb{R}$ for which the p-th power of the absolute value is Lebesgue integrable – p. 7
$L^\infty(\Omega)$	space of equivalence classes of functions from Ω to $\mathbb{R}$ which are essentially bounded – p. 7
$\mathcal{L}^p(\Omega)$	Lebesgue space of functions from Ω to $\mathbb{R}$ for which the p-th power of the absolute value is Lebesgue integrable – p. 7
$\mathcal{L}^p_{\text{loc}}(\Omega)$	space of functions which are locally in $\mathcal{L}^p(\Omega)$ – p. 7
$L^p_{\text{loc}}(\Omega)$	space of functions which are locally in $L^p(\Omega)$ – p. 7
$\mathcal{M}(\Omega)$	space of real, regular Borel measures on Ω – p. 8
$\mathcal{M}^+(\Omega)$	space of positive, real, regular Borel measures on Ω – p. 8
$W^{k,p}(\Omega)$	Sobolev space of integer order k with weak derivatives in $L^p(\Omega)$ – p. 9
$H^k(\Omega)$	simplified notation for $W^{k,2}(\Omega)$ – p. 9
$W^{k,p}_0(\Omega)$	space of functions in $W^{k,p}(\Omega)$ for each of which there is a defining sequence vanishing on the boundary $\partial\Omega$ – p. 9
$H^k_0(\Omega)$	simplified notation for $W^{k,2}_0(\Omega)$ – p. 10
$W^{k,1}_r(\Omega' \times I)$	anisotropic Sobolev space with weak temporal derivative in $L^r(\Omega' \times I)$ and the first k weak spacial derivatives in $L^r(\Omega' \times I)$ – p. 17
$H(\text{div}; \Omega)$	set of all functions in $L^2(\Omega)$ with divergence in $L^2(\Omega)$ – p. 76
Ω	usually a subset of $\mathbb{R}^n$ – p. 5
$\mathscr{L}$	Lagrangian – p. 11
δ_{x_0}	dirac delta function in x_0 – p. 24
e_{x_j}	piecewise linear function with $e_{x_j}(x_i) = 1$ for $i = j$ and $e_{x_j}(x_i) = 0$ for $i \neq j$ – p. 25
χ_k	indicator function of the time interval I_k – p. 25
M_h	mass matrix – p. 34
A_h	stiffness matrix – p. 34
$\mathcal{M}_\sigma$	space-time mass matrix – p. 35
$\mathbb{I}_N$	identity matrix of size $N \times N$ – p. 38
$\mathbb{1}_{N_h}$	vector of ones of length N_h – p. 61
$1_{(x,1)}$	characteristic function of the interval $(x, 1)$ – p. 79

Chapter 1

Introduction

In this chapter we will motivate our work, describe its structure, give an overview of the related literature to put it into context and explain its novelty.

1.1 Motivation

Applications like source identification and actuator placement motivate the study of sparse control problems. In those applications it is of interest to identify the precise location in space and (if applicable) the exact time instance of the controls support, which gives rise to the idea that the control is sparse and might be supported in one or multiple space(-time) points. There exist several ways to formulate such sparse control problems. The two main approaches to achieve a sparsity structure are to either introduce a L^1-norm regularization in the target functional or to consider measure-valued controls. We will investigate two different cases of the latter approach. Additionally, we consider a case where the control is a function of bounded variation (BV). This last case is strongly related to measure-valued controls. We specify the setup of the cases further in Section 1.2.

We discretize each of the chosen sparse optimal control problems. In order to retain the sparsity structure of the continuous problem on the discrete level, we choose a sparse discretization, i.e. we propose a discrete concept which delivers discrete controls with a maximal sparsity structure. This is achieved by the use of variational discretization from [49]. The key feature of variational discretization is to not discretize the control space. Instead, via the discretization of the test space and the optimality conditions, an implicit discretization of the control is achieved. This is how we can *control* the discrete structure of the controls through the choice of Petrov-Galerkin ansatz and test spaces in the discretization of the state equation. It is in fact the relation between the optimal adjoint state and the control that shows that the discrete structure of the test space affects the structure of the optimal controls. For problems with sparsity structure and measure control we aim at choosing the ansatz and test spaces in such a way that the induced structure of the controls is a sum of measures - without explicitly discretizing the control space. Similarly, in the case of BV controls we will choose the ansatz and test spaces, such that the first derivative of the control is a sum of measures and therefore the control is piecewise constant - again without discretizing the control.

1.2 Structure

We organize this thesis as follows: After introducing some useful basic concepts in Chapter 2 we move on to the main part of the work, which consists of Chapter 3, Chapter 4, and Chapter 5 - each of them dealing with one specific sparse optimal control problem, i.e.

- Chapter 3: a parabolic optimal control problem with a total variation norm of the space-time measure control in the target functional,

- Chapter 4: a parabolic optimal control problem with a bound on the total variation norm of the initial measure control,
- Chapter 5: an elliptic optimal control problem with a BV-seminorm of the BV control in the target functional.

Every one of these chapters has the same basic structure:

- First the problem formulation is introduced,
- then the continuous optimality system is analyzed,
- afterwards the variational discretization of the problem is established, and
- finally computational results are presented.

In Chapter 3 we additionally discuss the discontinuous Galerkin discretization of the given optimal control problem in comparison to the variational discretization approach. Lastly, we summarize the work in Chapter 6.

1.3 Literature overview and novelty of this work

The idea of studying optimal control with a sparsity structure has first been addressed in [69], where an $L^1(\Omega)$ control for linear elliptic equations was investigated. Motivating examples for the analysis of sparse control problems are for example given in [34] and [57], where the goals are finding the pollution source in a river and determining a heat source, respectively. Further approaches with an L^1-norm regularization in the target functional have been taken in [15, 23, 24, 48]. Spatio-temporally sparse optimal control problems of semilinear parabolic equations have been studied in [15], where the following three different sparsity promoting terms have been considered in the objective functional: $L^1(\Omega \times I)$, $L^2(I; L^1(\Omega))$ or $L^1(\Omega; L^2(I))$. Error estimates of fully discrete finite element approximations for the choice $L^1(\Omega; L^2(I))$ have been proven in [23] and improved in [24]. Another work in this direction is [48], where the directional sparsity control for parabolic equation is considered. Here, the controls are sparse in space but not necessarily in time and the sparsity pattern does not change over time.

Another way to achieve a sparsity structure is to consider measure controls. Control of elliptic partial differential equations with measures has been studied in [13, 19, 29, 30, 62], while control of parabolic partial differential equations with measures is content of [14, 20, 21, 25, 40, 41, 54, 55, 56, 71].

For the elliptic case we find a more general approach, where the control can be a measure or a function of bounded variation, in [29]. The optimal control problem with measure cost is treated with a duality approach and numerical results are presented. A similar problem, where the control u resides in $\mathcal{M}(\Omega)$ and a control cost is part of the target functional, is presented in [13]. Here, error estimates are proven and computational results are displayed. These a priori error estimates are then improved in [62]. Furthermore, the case with a semilinear elliptic partial differential equation governing an optimal control problem with measure control is discussed in [19]. Also, in [30] a similar optimal control problem but without control cost is analyzed, Fenchel duality is applied and numerical results are shown.

In the parabolic case there are a few different possible choices for the control space. An optimal control problem with separate measure data in time and space, control in $\mathcal{M}(\bar{\Omega}_T)$ and control cost in the L^2-norm is considered in [40]. Based on these results, in [41] pointwise control in $L^2(0, T; \mathbb{R}^m)$ is considered and the problem is variationally discretized. Error estimates and numerical results are presented in both works. Another approach is taken in [14], where the control resides in $L^2(I, \mathcal{M}(\Omega))$ and the control cost is taken in the respective norm. Here, convergence rates are proven. The choice $\mathcal{M}(\Omega_c, L^2(I))$ and control cost in the respective norm in [54] delivers a directional sparsity, since the spatial support is independent of time in this setting.

Optimal control of the linear second order wave equation with measure valued controls in $\mathcal{M}(\Omega; L^2(I))$ and control cost term is considered in [55]. In [71] measure-valued optimal control problems for 1D wave equation

with control space of either measure-valued functions $L^2_{w^*}(I;\mathcal{M}(\Omega))$ or vector measures $\mathcal{M}(\Omega, L^2(I))$ are treated, error estimates for the optimal state variable and the error measured in the cost functional are derived.

Further, in [20, 21, 25, 56] initial controls are examined. In [25] the parabolic optimal control problem with initial control in $\mathcal{M}(\Omega)$ is understood as an inverse source identification problem. The desired state only needs to be attained in an approximate sense and the target functional contains solely the control cost in its measure norm. A convergence result is presented. Then, in [20] initial control is combined with space-time measures as forcing functions, so that we have the set of controls $(u, u_0) \in \mathcal{M}(Q_c) \times \mathcal{M}(\Omega)$ and consequently two control cost terms.

In [56] initial control in $\mathcal{M}(\Omega)$ is considered with a cost functional consisting of a final time tracking term and control cost. Error estimates and numerical results for a full discretization are presented. In contrast, no control cost is considered in [21], but a bound on the total variation of the measure control is enforced. The target functional consists of a final time tracking term.

The cases we study in Chapter 3 and Chapter 4 are positioned in the category of parabolic optimal control with measures. The novelty of our work in both cases being the application of the variational discretization to a given optimal control problem.

In Chapter 3 the variational discrete approach is applied to a parabolic optimal control problem governed by space-time measure controls from [20]. Of particular interest for solving our optimal control problem in Chapter 3 are the techniques proposed in [29], where Fenchel duality is used to set up a predual problem for the elliptic case - we adapt this to the parabolic case. In this context we also consulted the lecture notes by Christian Clason [27], where Fenchel duality is discussed in detail. Furthermore, the time discrete scheme in a variational discrete setting for parabolic optimal control has been analyzed in [31, 39]. We choose a Petrov-Galerkin method employing piecewise constant states and piecewise linear and continuous test functions in time for the time discretization of the state equation. This induces an optimal control, that by construction of the state discretization and the variational discretization concept, is the sum of Dirac measures.

In Chapter 4 the initial optimal control problem that has been discussed in [21] is variationally discretized. Here, the control set is constrained, instead of incorporating a penalty term for the control in the target functional, which then only consists of a final time tracking term in the state variable. By chosing piecewise linear functions as ansatz and test space in the discretization of the state equation we see that the optimal control, which is not discretized, has the induced structure of being a sum of Dirac measures in space as in the previous problem.

Closely related to optimal control with measures is optimal control governed by BV functions, since the distributional derivative of a BV function resides in a measure space. Before optimal control with BV functions was first discussed, there already existed articles on optimization with BV functions. An early result in optimization with BV functions and regularization by BV-seminorms is [22]. This work was motivated by the application of denoising blocky images with very high noise. Error estimates and numerical analysis for inverse problems involving BV functions can be found in [6, 7]. An inverse problem with BV control and total variation seminorm goverened by an elliptic partial differential equation is applied to Quantitative Susceptibility Mapping in [9] with high quality results. There exist studies of elliptic optimal control with total variation regularization and control in $L^\infty(\Omega)$, see [28, 50]. Controls from the space $BV(\Omega) \cap L^\infty(\Omega)$ are considered in [16] as weights in the weighted p-Laplace problem with box-type constraints.

Optimal control governed by a semilinear parabolic equation and control cost in a total bounded variation seminorm is discussed in [17], a convergence result is shown and numerical experiments are presented. A similar problem is analyzed in [18], but with semilinear elliptic equation. In [51] the BV source in an elliptic system is supposed to be recovered. To this end, total variation regularization is employed. For 1D elliptic optimal control problems with BV control error estimates and numerical results for two discretization techniques are presented in [43]. The techniques being variational discretization and piecewise constant control discretization.

In Chapter 5 we also discuss 1D elliptic optimal control governed by BV functions. Here, we consider the mixed formulation of the elliptic partial differential equation and regard the variational discrete formulation of the resulting problem. We prove error estimates and provide computational experiments to confirm our findings.

Chapter 2

Mathematical background

This chapter gives a brief overview of established concepts and results that will be useful throughout this thesis. It consists of excerpts of functional analysis and optimization taken from [4, 52, 53, 66, 72, 73], but neither topic will be completely covered. Furthermore, the notation from given references will be adapted to the notation within the thesis to guarantee comprehensibility. We remark that there exist several other references for the presented concepts and results.

2.1 Functional analysis

The goal of this section is to generate a common understanding of the spaces, which will be used throughout the thesis. We begin with the definition of Banach and Hilbert spaces.

Definition 2.1. *(Norm, Banach space, [52, Definition 1.1.])*
Let X be a real vector space.

*i) A mapping $\|\cdot\|_X \colon X \mapsto [0,\infty)$ is a **norm** on X, if*

1) $\|u\|_X = 0 \Leftrightarrow u = 0$,

2) $\|\lambda u\|_X = |\lambda| \|u\|_X \quad \forall u \in X,\ \lambda \in \mathbb{R}$,

3) $\|u+v\|_X \leq \|u\|_X + \|v\|_X \quad \forall u, v \in X$.

*ii) A normed real vector space X is called (real) **Banach space** if it is complete, i.e., if any Cauchy sequence $(u_n)_n$ has a limit $u \in X$, more precisely, if $\lim_{m,n\to\infty} \|u_m - u_n\|_X = 0$ then there exists $u \in X$ with $\lim_{n\to\infty} \|u_n - u\|_X = 0$.*

A special Banach space, we will be interested in, is the following:

Definition 2.2. *(Space of continuous functions, [52, Example 1.1.])*
For $\Omega \subset \mathbb{R}^n$ consider the function space

$$C(\Omega) = \{u : \Omega \to \mathbb{R} : u \text{ continuous}\}.$$

If Ω is bounded then $C(\bar{\Omega})$ is a Banach space with the sup-norm

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

Later on, we will look at subspaces of the space of continuous functions with certain properties and identify measure spaces in a duality sense with one of these subspaces.

Definition 2.3. *(Inner product, Hilbert space, [52, Definition 1.2.]) Let H be a real vector space. A mapping $(\cdot,\cdot)_H \colon H \times H \mapsto \mathbb{R}$ is an **inner product** on H, if*

i) $(u,v)_H = (v,u)_H \quad \forall\, u, v \in H$,

ii) for every $v \in H$ the mapping $u \in H \mapsto (u,v)_H$ is linear,

iii) $(u,u)_H \geq 0 \quad \forall\, u \in H$ *and* $(u,u)_H = 0 \quad \Leftrightarrow \quad u = 0$.

A vector space H with inner product $(\cdot,\cdot)_H$ and associated norm

$$\|u\|_H := \sqrt{(u,u)_H}$$

*is called **Pre-Hilbert space**.*
*A Pre-Hilbert space $(H, (\cdot,\cdot)_H)$ is called **Hilbert space** if it is complete under its norm $\|\cdot\|_H$.*

Definition 2.4. *(space of linear operators, operator norm, [52, Definition 1.4.ii)])*
*Let X, Y be normed real vector spaces with norms $\|\cdot\|_X, \|\cdot\|_Y$. By $\mathcal{L}(X,Y)$ we denote the **space of all linear operators** $A : X \to Y$ that are bounded in the sense that*

$$\|A\|_{X,Y} := \sup_{\|u\|_X = 1} \|Au\|_Y < \infty.$$

*$\mathcal{L}(X,Y)$ is a normed space with the **operator norm** $\|\cdot\|_{X,Y}$.*

Theorem 2.5. *([52, Theorem 1.2.])*
If Y is a Banach space then $\mathcal{L}(X,Y)$ is a Banach space.

We extend the concept of differentiability to operators between Banach spaces.

Definition 2.6. *(directionally / Gâteaux / Fréchet differentiable, [52, Definition 1.29])*
Let $F : U \subset X \to Y$ be an operator with Banach spaces X, Y and open $U \neq \emptyset$.

*i) F is called **directionally differentiable** at $x \in U$ if the limit*

$$dF(x;h) = \lim_{t \searrow 0} \frac{F(x+th) - F(x)}{t} \in Y$$

exists for all $h \in X$. In this case, $dF(x;h)$ is called directional derivative of F in the direction h.

*ii) F is called **Gâteaux differentiable** at $x \in U$ if F is directionally differentiable at x and the directional derivative*

$$F'(x) : X \to Y, \qquad h \mapsto dF(x;h)$$

is bounded and linear, i.e. $F'(x) \in \mathcal{L}(X,Y)$.

*iii) F is called **Fréchet differentiable** at $x \in U$ if F is Gâteaux differentiable at x and if the following approximation condition holds:*

$$\|F(x+h) - F(x) - F'(x)h\|_Y = o(\|h\|_X) \qquad \text{for } \|h\|_X \to 0.$$

iv) If F is directionally / Gâteaux / Fréchet differentiable at every $x \in V, V \subset U$ open, then F is called directionally / Gâteaux / Fréchet differentiable on V.

A very important concept we will be using is duality.

Definition 2.7. *(dual space, dual pairing, [52, Definition 1.5.])*
*The space $X^* := \mathcal{L}(X, \mathbb{R})$ of linear functionals on X is called **dual space** of X and is a Banach space with the operator norm*

$$\|u^*\|_{X^*} := \sup_{\|u\|_X = 1} |u^*(u)|.$$

We use the notation

$$\langle u^*, u \rangle_{X^*,X} := u^*(u),$$

*and call $\langle \cdot, \cdot \rangle_{X^*,X}$ the **dual pairing** of X^* and X.*

Definition 2.8. *(pre-dual space)*
*In the given setting we call X the **pre-dual space** of X^*.*

We give the following result for Hilbert spaces here. The application to the space of continuous functions will be discussed later.

Theorem 2.9. *(Riesz representation theorem, [52, Theorem 1.4.])*
The dual space H^* of a Hilbert space H is isometric to H itself. More precisely, for every $v \in H$ the linear functional u^* defined by

$$\langle u^*, u \rangle_{H^*,H} := (v, u)_H \quad \forall u \in H$$

is in H^* with norm $\|u^*\|_{H^*} = \|v\|_H$. Vice versa, for any $u^* \in H^*$ there exists a unique $v \in H$ such that

$$\langle u^*, u \rangle_{H^*,H} := (v, u)_H \quad \forall u \in H,$$

and $\|u^*\|_{H^*} = \|v\|_H$.

We move on to defining the following spaces with their respective norms:

Definition 2.10. *($L^p(\Omega)$, [52, Definition 1.11.])*
Let $\Omega \subset \mathbb{R}^n$ be Lebesgue-measurable. We define for $p \in [1, \infty)$ the seminorm

$$\|u\|_{L^p(\Omega)} := \left(\int_{\mathbb{R}^n} |u(x)|^p \right)^{\frac{1}{p}},$$

and

$$\|u\|_{L^\infty(\Omega)} := \operatorname{ess\,sup}_{x \in \Omega} |u(x)|.$$

Now, for $1 \leq p \leq \infty$ we define the spaces

$$\mathcal{L}^p(\Omega) := \{u : \Omega \to \mathbb{R} \text{ Lebesgue measurable } : \|u\|_{L^p(\Omega)} < \infty\}.$$

These are not normed spaces since there exist measurable functions $u : \Omega \to \mathbb{R}, u \neq 0$, with $\|u\|_{L^p(\Omega)} = 0$.
We use the equivalence relation

$$u \sim v \in L^p(\Omega) \quad :\Leftrightarrow \quad \|u - v\|_{L^p(\Omega)} = 0 \quad \Leftrightarrow \quad u = v \quad a.e.$$

to define $L^p(\Omega) = \mathcal{L}^p(\Omega)/\sim$ as the space of equivalence classes of a.e. identical functions, equipped with the norm $\|\cdot\|_{L^p(\Omega)}$.
Finally we define

$$\mathcal{L}^p_{\text{loc}}(\Omega) := \{u : \Omega \to \mathbb{R} \text{ Lebesgue measurable } : u \in \mathcal{L}^p(K) \text{ for all } K \subset \Omega \text{ compact}\}$$

and set $L^p_{\text{loc}}(\Omega) := \mathcal{L}^p_{\text{loc}}/\sim$.

Theorem 2.11. *(Fischer-Riesz theorem, [52, Theorem 1.6.])*
The spaces $L^p(\Omega)$, $p \in [1, \infty]$, are Banach spaces. The space $L^2(\Omega)$ is a Hilbert space with inner product

$$(u, v)_{L^2(\Omega)} := \int_\Omega uv \, dx.$$

Definition 2.12. *(reflexive, [52, Definition 1.17.])*
*A Banach space X is called **reflexive** if the mapping $x \in X \mapsto \langle \cdot, x \rangle_{X^*,X} \in (X^*)^*$ is surjective, i.e., if for any $x^{**} \in (X^*)^*$ there exists $x \in X$ with*

$$\langle x^{**}, x^* \rangle_{(X^*)^*,X^*} = \langle x^*, x \rangle_{X^*,X} \quad \forall x^* \in X^*.$$

Remark 2.13. *([52, Remark 1.8.])*
L^p is reflexive for $1 < p < \infty$, since we have the isometric isomorphisms $(L^p)^* = L^q, \frac{1}{p} + \frac{1}{q} = 1$, and thus $((L^p)^*)^* = (L^q)^* = L^p$. Moreover, any Hilbert space is reflexive by the Riesz representation theorem.

In this thesis we are working with measures, which are defined as follows:

Definition 2.14. *($\mathcal{M}(\Omega)$,$\mathcal{M}^+(\Omega)$, [4, Section 4.2.1.])*
Let $\Omega \subset \mathbb{R}^n$ open and $\mathcal{B}(\Omega)$ its Borel field. We denote the set of all real-valued Borel measures by $\mathcal{M}(\Omega)$. It is the vectorial space of all the set functions $\mu : \mathcal{B}(\Omega) \to \mathbb{R}$ satisfying $\mu(\emptyset) = 0$ and σ-additivity. The subset of nonnegative elements is denoted by $\mathcal{M}^+(\Omega)$.

The **total variation** of a measure $\mu \in \mathcal{M}(\Omega)$ is the real-valued set function $|\mu|$, defined for all Borel sets B of Ω by

$$|\mu|(B) := \sup \left\{ \sum_{i=0}^{\infty} |\mu(B_i)| : \bigcup_{i=0}^{\infty} B_i = B \right\},$$

where the supremum is taken over all the partitions of B in $\mathcal{B}(\Omega)$ (compare [4, p. 125]). It holds $|\mu| \in \mathcal{M}^+(\Omega)$.

We now define the following two subsets of continuous functions with additional properties.

Definition 2.15. *($C_c(\Omega)$, [66, 2.9 Definition])*
The collection of all continuous functions on $\Omega \subset \mathbb{R}^n$ whose support is compact is denoted by $C_c(\Omega)$.

Definition 2.16. *($C_0(\Omega)$, [66, 3.16 Definition])*
A complex function f on a locally compact Hausdorff space Ω is said to vanish at infinity if to every $\epsilon > 0$ there exists a compact set $K \subset \Omega$, such that $|f(x)| < \epsilon$ for all x not in K.
The class of all continuous f on Ω which vanish at infinity is called $C_0(\Omega)$.
It is clear that $C_c(\Omega) \subset C_0(\Omega)$, and that the two classes coincide if Ω is compact. In that case we write $C(\Omega)$ for either of them.

Now we give the Riesz representation theorem - which we saw for Hilbert spaces earlier - in the context of spaces of continuous functions.

Theorem 2.17. *(Riesz representation theorem, [66, 6.19 Theorem])*
If Ω is a locally compact Hausdorff space, then every bounded linear functional ϕ on $C_0(\Omega)$ is represented by a unique regular complex Borel measure u, in the sense that

$$\phi(f) = \int_\Omega f \, du$$

for every $f \in C_0(\Omega)$.

So, by Riesz representation theorem, we may identify the space $\mathcal{M}(\Omega)$ with the dual space of $C_0(\Omega)$ and moreover, the total variation norm for measures coincide with the dual norm:

$$\|u\|_{\mathcal{M}(\Omega)} = \sup_{\|f\|_{C_0(\Omega)} \leq 1} \int_\Omega f \, du.$$

We stress that $C_0(\Omega)$ is not reflexive.

We define the following boundary condition, which will provide regularity needed for solvability of state equations in the work.

Definition 2.18. *($C^{k,\beta}$-boundary, Lipschitz-boundary [52, Definition 1.13.])*
*Let $\Omega \subset \mathbb{R}^n$ be open and bounded. We say that Ω has a $C^{k,\beta}$-**boundary**, $k \in \mathbb{N}_0 \cup \{\infty\}, 0 \leq \beta \leq 1$, if for any $x \in \partial\Omega$ there exists $r > 0, k \in \{1, \ldots, n\}, \sigma \in \{-1, +1\}$, and a function $\gamma \in C^k(\mathbb{R}^{n-1})$ such that*

$$\Omega \cap B(x; r) = \{y \in B(x; r) : \sigma y_k < \gamma(y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_n)\},$$

*where $B(x; r)$ denotes the open ball around x with radius r. Instead of $C^{0,1}$-boundary we say also **Lipschitz-boundary***

Next, we introduce subspaces $W^{k,p}(\Omega)$ of $L^p(\Omega)$. To this end we introduce the concept of weak derivatives.

Definition 2.19. *(weak partial derivative, [35, Section 5.2.1.])*
*Suppose $u, v \in L^1_{\text{loc}}(\Omega)$, and α is a multiindex. We say that v is the α^{th}-**weak partial derivative** of u, written $D^\alpha u = v$, provided*

$$\int_\Omega u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega v\phi \, dx$$

for all test functions $\phi \in C_c^\infty(\Omega)$.

Definition 2.20. *(Sobolev space, [52, Definition 1.14.])*
*Let $\Omega \subset \mathbb{R}^n$ be open. For $k \in \mathbb{N}_0, p \in [1, \infty]$, we define the **Sobolev space** $W^{k,p}(\Omega)$ by*

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) : u \text{ has weak derivatives } D^\alpha u \in L^p(\Omega) \text{ for all } |\alpha| \leq k\}$$

equipped with the norm

$$\|u\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}, \quad p \in [1, \infty),$$
$$\|u\|_{W^{k,\infty}(\Omega)} := \sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^\infty(\Omega)}.$$

Theorem 2.21. *([52, Theorem 1.11.])*
Let $\Omega \subset \mathbb{R}^n$ be open, $k \in \mathbb{N}_0$, and $p \in [1, \infty]$. Then $W^{k,p}(\Omega)$ is a Banach space.
Moreover, the space $H^k(\Omega) := W^{k,2}(\Omega)$ is a Hilbert space with inner product

$$(u, v)_{H^k(\Omega)} = \sum_{|\alpha| \leq k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}.$$

Definition 2.22. *($W_0^{k,p}(\Omega)$, [53, Definition 8.1.1.])*
For $p \in [1, \infty)$, by $W_0^{1,p}(\Omega)$ we mean the subset of $W^{1,p}(\Omega)$ consisting of all functions for each of which there is a defining sequence vanishing on the boundary $\partial\Omega$. For $k \in \{2, 3, \ldots\}$, let

$$W_0^{k,p}(\Omega) = W_0^{1,p}(\Omega) \cap W^{k,p}(\Omega).$$

The spaces for $k \in \{1, 2, \ldots\}$ are equipped with the same norm as $W^{k,p}(\Omega)$ and are Banach spaces. The space $H_0^k(\Omega) = W_0^{k,2}(\Omega)$ is a Hilbert space.

Another type of functions we are interested in are functions of bounded variation (BV) for the special case $\Omega \subset \mathbb{R}$.

Definition 2.23. *(BV(Ω),[4, Definition 10.1.1.])*
*We say that a function $u : \Omega \to \mathbb{R}$ is a **function of bounded variation** if and only if*

i) it belongs to $L^1(\Omega)$, and

ii) its distributional derivative u' belongs to $\mathcal{M}(\Omega)$.

We denote the set of all functions of bounded variation by $BV(\Omega)$.

We can also write

$$BV(\Omega) = \left\{ u \in L^1(\Omega) : \|u'\|_{\mathcal{M}(\Omega)} < \infty \right\}.$$

The space $BV(\Omega)$ is equipped with the following norm (see e.g. [4, p. 372]), which extends the classical norm in $W^{1,1}(\Omega)$:

$$\|u\|_{BV(\Omega)} := \|u\|_{L^1(\Omega)} + \|u'\|_{\mathcal{M}(\Omega)}.$$

Theorem 2.24. *([4, Theorem 10.1.1.])*
Equipped with its norm $\|\cdot\|_{BV(\Omega)}$, $BV(\Omega)$ is a Banach space.

Theorem 2.25. *([4, Theorem 10.1.3. and Theorem 10.1.4])*
Let Ω be a 1-regular open bounded subset of $\mathbb{R}$. The embedding

$$BV(\Omega) \hookrightarrow L^p(\Omega)$$

i) is continuous for all $1 \leq p \leq \infty$, and

ii) is compact for all $1 \leq p < \infty$.

2.2 Optimization

Let us begin with the helpful property of convexity.

Definition 2.26. *(convex set, [73, Definition 6.1])*
*The set $\Omega \subset \mathbb{R}^n$ is called **convex**, if for all $x, y \in \Omega$ and all $\lambda \in [0, 1]$ it holds*

$$\lambda x + (1 - \lambda) y \in \Omega.$$

Definition 2.27. *((strictly) convex function, [73, Definition 6.2])*
The function $f : \Omega \to \mathbb{R}$ defined on a convex set Ω is called

*i) **convex**, if for all $x, y \in \Omega$ and all $\lambda \in [0, 1]$ it holds*

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

*ii) **strictly convex**, if for all $x, y \in \Omega$, with $x \neq y$ and all $\lambda \in (0, 1)$ it holds*

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y).$$

Now, we introduce a few basic concepts of finite-dimensional optimization that will be needed as tools to solve the discretized problems. To this end, we consider the general nonlinear optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{such that} \quad g(x) \le 0,\ h(x) = 0, \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p$ are continuously differentiable functions.

Definition 2.28. *(Lagrangian, [73, Definition 16.17])*
The function $\mathscr{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$,

$$\mathscr{L}(x, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$$

is called the ***Lagrangian*** *of problem* (2.1).

Definition 2.29. *(Slater condition, [73, p. 110])*
The ***Slater condition*** *is satisfied for problem* (2.1)*, if there exists* $y \in \mathbb{R}^n$*, such that*

$$g_i(y) < 0 \quad \forall\, i = 1, \dots, m \qquad \text{and} \qquad h_j(y) = 0 \quad \forall\, j = 1, \dots, p.$$

The Slater condition is a constraint qualification for every admissible point of problem (2.1).

Theorem 2.30. *(Karush-Kuhn-Tucker conditions, [73, Satz 16.14])*
Let $\bar{x} \in \mathbb{R}^n$ *be a local solution to* (2.1)*, which fulfills a constraint qualification. The Karush-Kuhn-Tucker-conditions (KKT-conditions) hold:*
There exist Lagrange-multipliers $\bar{\lambda} \in \mathbb{R}^m$ *and* $\bar{\mu} \in \mathbb{R}^p$*, such that*

i) $\nabla_x \mathscr{L}(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0$,

ii) $h(\bar{x}) = 0$,

iii) $\bar{\lambda} \ge 0, \quad g(\bar{x}) \le 0, \quad \bar{\lambda}^\top g(\bar{x}) = 0$ *(complementarity condition).*

For semismooth problems we will make use of the following concepts.

Definition 2.31. *(Clarke's generalized Jacobian, [52, Example 2.4.])*
For locally Lipschitz-continuous functions $G : \mathbb{R}^n \to \mathbb{R}^m$*, we define* ***Clarke's generalized Jacobian*** *by*

$$\partial^{cl} G(x) = \operatorname{conv}\left\{ M : x^k \overset{k\to\infty}{\longrightarrow} x, G'(x^k) \longrightarrow M, G \text{ differentiable at } x^k \right\}.$$

(This definition is justified since G' *exists almost everywhere on* $\mathbb{R}^n$ *by Rademacher's theorem.)*

The following algorithm, a generalization of Newton's method, converges locally towards $\bar{x}$ satisfying $G(\bar{x}) = 0$, where G is a locally Lipschitz-continuous function.

Algorithm 2.32: Semismooth Newton method, [52, Algorithm 2.11.]

input: $x^0 \in X$ (sufficiently close to the solution $\bar{x}$)
for $k = 0, 1, \dots$ **do**
  Choose $M_k \in \partial^{cl} G(x^k)$.
  Obtain s^k by solving $M_k s^k = -G(x^k)$.
  Set $x^{k+1} = x^k + s^k$.

2.2.1 Optimal control

We will deal with optimization problems with partial differential equation constraints in this work, so we give a brief introduction (see e.g. [52, 58, 65, 72] for more details on this topic).

We consider the optimal control problem of the general form

$$\min_{(y,u)\in Y\times U} J(y,u) \quad \text{s.t.} \quad e(y,u)=0 \quad \text{and} \quad u\in U_{\text{ad}},$$

where $J : Y \times U \to \mathbb{R}$ is the objective functional, Y and U are the state and control space respectively. $y \in Y$ describes the state of the considered system, which is described by $e(y,u) = 0$ (state equation) and will be a partial differential equation in this work. The control $u \in U$ is supposed to be adapted in an optimal way. The set of admissible controls is denoted by U_{ad}.
Now, assuming that $y = y(u)$ the unique solution to the state equation $e(y,u) = 0$ exists, we can formulate the reduced problem

$$\min_{u\in U_{\text{ad}}} \hat{J}(u) := J(y(u),u). \tag{2.2}$$

Definition 2.33. *(optimal control, [72, Section 1.4.2])*
*A control $\bar{u} \in U_{\text{ad}}$ is called **optimal control** (or global solution) to* (2.2)*, if it holds*

$$\hat{J}(\bar{u}) \le \hat{J}(u) \qquad \forall u \in U_{\text{ad}}.$$

Then, $\bar{y} = y(\bar{u})$ is the associated optimal state.

When proving existence of solutions to the optimal control problem, the concepts of weak convergence, weak sequential compactness and lower semi continuity are essential.

Definition 2.34. *(weak convergence, [52, Definition 1.16])*
*Let X be a Banach space. We say that a sequence $(x_k)_k \subset X$ **converges weakly** to $x \in X$, written*

$$x_k \rightharpoonup x,$$

if

$$\langle x^*, x_k\rangle_{X^*,X} \to \langle x^*, x\rangle_{X^*,X} \qquad \text{as } k\to\infty \qquad \forall x^*\in X^*.$$

It is easy to check that strong convergence $x_k \to x$ implies weak convergence $x_k \rightharpoonup x$. For BV-functions we have the following:

Definition 2.35. *(weak convergence in $BV(\Omega)$, [4, Definition 10.1.2.])*
A sequence $(u_n)_n \in BV(\Omega)$ converges weakly to some $u \in BV(\Omega)$, and we write $u_n \rightharpoonup u$, if and only if the following convergences hold:

i) $u_n \to u \in L^1(\Omega)$*, and*

ii) $u_n' \rightharpoonup u' \in \mathcal{M}(\Omega)$*.*

Theorem 2.36. *(Weak sequential compactness, [52, Theorem 1.17])*
Let X be a reflexive Banach space. Then the following holds

i) Every bounded sequence $(x_k)_k \subset X$ contains a weakly convergent subsequence, i.e. there are $(x_{k'})_{k'} \subset (x_k)_k$ and $x \in X$ with $x_{k'} \rightharpoonup x$.

*ii) Every bounded, closed and convex subset $C \subset X$ is **weakly sequentially compact**, i.e. every sequence $(x_k)_k \subset C$ contains a weakly convergent subsequence $(x_{k'})_{k'} \subset (x_k)_k$ with $x_{k'} \rightharpoonup x$, where $x \in C$.*

Theorem 2.37. *(Lower semi continuity, [52, Theorem 1.18])*
*Let X be a Banach space. Then any continuous, convex functional $F : X \to \mathbb{R}$ is **weakly lower semi continuous**, i.e.*

$$x_k \rightharpoonup x \quad \Rightarrow \quad F(x) \leq \liminf_{k\to\infty} F(x_k).$$

We also introduce an optimality condition for problem (2.2) under the following assumptions:

Assumption 2.38.

i) $U_{\text{ad}} \subset U$ *is nonempty, convex and closed.*

ii) $J : Y \times U \to \mathbb{R}$ *and* $e : Y \times U \to Z$ *are continuously Fréchet differentiable and* U, Y, Z *are Banach spaces.*

iii) For all $u \in V$ *in a neighborhood* $V \subset U$ *of* U_{ad}*, the state equation* $e(y, u) = 0$ *has a unique solution* $y = y(u) \in Y$*.*

iv) $\frac{\partial}{\partial y} e(y(u), u) \in \mathcal{L}(Y, Z)$ *has a bounded inverse for all* $u \in V \supset U_{\text{ad}}$*.*

Under these assumptions the mapping $u \mapsto y(u)$ is continuously Fréchet differentiable by the implicit function theorem.

Theorem 2.39. *([52, Theorem 1.48])*
Let Assumption 2.38 hold. If $\bar{u}$ is a local solution of (2.2)*, then $\bar{u}$ satisfies the following optimality condition*

$$\bar{u} \in U_{\text{ad}}, \qquad \langle \hat{J}'(\bar{u}), u - \bar{u} \rangle_{U^*,U} \geq 0 \quad \forall u \in U_{\text{ad}}. \tag{2.3}$$

Remark 2.40. *([52, Remark 1.19])*
A condition of the form (2.3) *is called **variational inequality**.*

Chapter 3

Parabolic optimal control governed by space-time measure controls

This chapter is based on the article [47] with the title "Maximal discrete sparsity in parabolic optimal control with measures". Furthermore, some of the results in the article are based on the authors master thesis [45] with the title "Variational discretization of parabolic control problems in space-time measure spaces".

We organize the chapter as follows: We state the optimal control problem and the main convergence result in Section 3.1. In Section 3.2 we analyze the continuous problem (P) and its sparsity structure. We then set up the predual problem, show that it has a unique solution and apply the Fenchel duality theorem. The predual problem is discretized with two different strategies. The first one is by variational discretization. We discuss it in Section 3.3, where we derive also a semismooth Newton method to solve the variational discrete problem (P_σ). The second strategy is a discontinuous Galerkin discretization (see Section 3.4). The emerging fully discrete problem (P_{DG}) is solved analogously to (P_σ). Computational results of both approaches are compared in Section 3.5.

3.1 Problem Formulation

We consider the continuous minimization problem from [20]

$$\min_{(u_0,u)\in\mathcal{M}(\bar{\Omega}_c)\times\mathcal{M}(\bar{Q}_c)} J(u_0,u) \coloneqq \frac{1}{q}\|y-y_{\mathrm{d}}\|^q_{L^q(Q)} + \alpha\,\|u\|_{\mathcal{M}(\bar{Q}_c)} + \beta\,\|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}, \tag{P}$$

where the *state* $y \in L^q(Q)$ solves the following parabolic *state equation*

$$\begin{cases} \partial_t y - \Delta y &= u & \text{in } Q = \Omega\times(0,T),\\ y(x,0) &= u_0 & \text{in } \Omega,\\ y(x,t) &= 0 & \text{on } \Sigma = \Gamma\times(0,T), \end{cases} \tag{3.1}$$

with real, regular Borel measures $u \in \mathcal{M}(\bar{Q}_c)$ and $u_0 \in \mathcal{M}(\bar{\Omega}_c)$. Here $\Omega \subset \mathbb{R}^n$ is an open, bounded domain with boundary $\Gamma \coloneqq \partial\Omega$ of regularity to be discussed later. We fix an open, relatively compact interval $I_c \subset\subset I \coloneqq (0,T)$ and a relatively compact subdomain $\Omega_c \subset\subset \Omega$ and define the space-time control domain

$$Q_c \;\coloneqq\; \Omega_c \times I_c.$$

By the Riesz representation theorem (see Theorem 2.9), we may identify the space of regular $\mathcal{M}(X)$ Borel measures on a subset $X \subset \mathbb{R}^{d+1}$ with the dual space of $C_0(X)$, the closure of the space of continuous, compactly supported

functions in the supremum norm. In particular, we have

$$\mathcal{M}(\bar{\Omega}_c) = C(\bar{\Omega}_c)^*, \qquad \mathcal{M}(\bar{Q}_c) = C(\bar{Q}_c)^*, \qquad \mathcal{M}(\Omega) := C_0(\Omega)^*, \qquad \mathcal{M}(Q) := C_0(Q)^*.$$

Moreover, the total variation norm for measures coincides with the dual norms:

$$\|u_0\|_{\mathcal{M}(\bar{\Omega}_c)} = \sup_{\|f\|_{C(\bar{\Omega}_c)} \le 1} \int_{\bar{\Omega}_c} f \, du_0,$$
$$\|u\|_{\mathcal{M}(\bar{Q}_c)} = \sup_{\|f\|_{C(\bar{Q}_c)} \le 1} \int_{\bar{Q}_c} f \, du.$$

Where appropriate, we identify $\mathcal{M}(\bar{Q}_c)$ with the space $\{u \in \mathcal{M}(Q) : \operatorname{supp}(u) \subseteq \bar{Q}_c\}$, and accordingly $\mathcal{M}(\bar{\Omega}_c)$ with $\{u_0 \in \mathcal{M}(\Omega) : \operatorname{supp}(u_0) \subseteq \bar{\Omega}_c\}$. Furthermore, $\alpha > 0, \beta > 0$ are given penalty parameters.

The state y is supposed to solve (3.1) in the following very weak sense, equivalent to [20, Definition 2.1.] :

Definition 3.1. *A function $y \in L^q(Q)$ is a solution to* (3.1)*, if the identity*

$$\int_Q -(\partial_t w + \Delta w)\, y \, dx\, dt = \int_{\bar{Q}_c} w \, du + \int_{\bar{\Omega}_c} w(0) \, du_0 \tag{3.2}$$

holds for all $w \in C^\infty(\bar{Q})$ with $w(T) = 0$ and $w|_\Sigma = 0$.

The solvability of (3.2) and of problem (P) have already been established in [20, Theorem 2.2.] and [20, Theorem 2.7.]. We will also discuss this matter in greater detail later in Section 3.2. For the moment, we just state that both (3.2) and (P) are well-posed with unique solutions provided that

(i) Ω is sufficiently regular (e.g. Ω is of class $C^{1,1}$), and

(ii) $q \in (1, \min\{2, 1 + 2/n\}]$.

For the practical implementation, we propose a discrete concept which delivers discrete controls with a maximal sparsity structure, i.e., variational discretization from [49], which allows to *control* the discrete structure of the controls through the choice of Petrov-Galerkin ansatz and test spaces in the discretization of the state equation. The problem can also be discretized by a full discretization approach as is proposed in, e.g., [20], where piecewise constant controls in time and Dirac measures in space are used. This limits the maximal possible sparsity in this setting to controls which are constant on time intervals.

For the variational discretization we obtain analogous convergence results as reported in [20, Theorem 4.3.]. More precisely, we will prove Theorem 3.2, where (P_σ) denotes the variational discrete version of the problem (P). We denote the implicitly discrete control space $U_h \times \mathcal{U}_{\mathrm{vd}}$, in which ($P_\sigma$) has a unique solution (see Theorem 3.12), and the discretization parameter $\sigma := (\tau, h)$, where τ indicates time and h indicates space (see Section 3.3 for more details on the notation). Our main result reads as follows:

Theorem 3.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded open domain of class $C^{1,1}$ and let $q \in (1, 2]$ satisfy $q < 1 + 2/n$. For fixed σ, let $(\bar{u}_{0,h}, \bar{u}_\sigma)$ be the unique solution of problem (P_σ) that belongs to $U_h \times \mathcal{U}_\sigma$, and denote the associated state by $\bar{y}_\sigma$.*
Then for each sequence of discretizations with $|\sigma| \to 0$, we have the following convergence properties:

$$\bar{y}_\sigma \to \bar{y} \quad \text{in } L^q(Q), \tag{3.3}$$
$$\bar{u}_\sigma \overset{*}{\rightharpoonup} \bar{u} \quad \text{in } \mathcal{M}(\bar{Q}_c) \quad \text{and} \quad \bar{u}_{0,h} \overset{*}{\rightharpoonup} \bar{u}_0 \quad \text{in } \mathcal{M}(\bar{\Omega}_c), \tag{3.4}$$
$$\|\bar{u}_\sigma\|_{\mathcal{M}(\bar{Q}_c)} \to \|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} \quad \text{and} \quad \|\bar{u}_{0,h}\|_{\mathcal{M}(\bar{\Omega}_c)} \to \|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)}, \tag{3.5}$$

where $(\bar{u}_0, \bar{u})$ is the unique solution of (P) and $\bar{y}$ its associated state.

The proof is given on page 30.

3.2 Continuous optimality system

In this section, we take a closer look at the solution structure of (P).

3.2.1 State equation

First, we have to discuss solvability of the state equation (3.1), to be interpreted in the form (3.2). To this end, for an arbitrary open domain $\Omega' \subset \mathbb{R}^n$, we introduce the following anisotropic Sobolev spaces

$$W_r^{k,1}(\Omega' \times I) := \left\{ w \in L^r(\Omega' \times I) : \partial_t w,\ \partial_x^1 w, \dots, \partial_x^k w \in L^r(\Omega' \times I) \right\},$$

$k \in \mathbb{N}_0$ and $r \in (1, \infty)$, and define the space

$$W := \left\{ w \in W_2^{1,1}(Q) : w|_\Sigma = 0,\ w(T) = 0 \text{ and } -(\partial_t + \Delta)w \in L^p(Q) \right\},$$

where $p = \left(1 - \frac{1}{q}\right)^{-1} \in (2, \infty)$ is the Hölder conjugate of q. Because of $L^p(\Omega) \subset L^2(\Omega)$, the existence and uniqueness theory for weak solutions of parabolic partial differential equations (see e.g., [35, Chapter 7]) implies that the operator

$$L := -(\partial_t + \Delta) \colon W \to L^p(Q)$$

is an isomorphism of vector spaces. Equipped with the norm

$$\|w\|_W := \|L\, w\|_{L^p(Q)},$$

W is a reflexive Banach space and L is an isomorphism of Banach spaces. The heat operator $\partial_t - \Delta$ is the adjoint of L in the sense that $\langle Lw, y\rangle_{L^p,L^q} = \langle L^* y, w\rangle_{W^*,W}$ holds for all $w \in W$ and all $y \in L^q(Q)$:

$$L^* := \partial_t - \Delta : L^q(Q) \to W^*.$$

Since L is continuously invertible, so is its adjoint L^*.

Next we show that $\mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c)$ embeds continuously into W^*. This will justify putting measures on the right hand side of (P).

We choose two further subdomains $\Omega', \Omega'' \subset \Omega$ with smooth boundaries such that $\Omega_c \subset\subset \Omega' \subset\subset \Omega'' \subset\subset \Omega$. By interior regularity estimates (see, e.g., [53, Theorem 4.3.7]), there exists a $C \geq 0$ such that

$$\|w\|_{L^p(\Omega' \times I)} + \|\partial_t w\|_{L^p(\Omega' \times I)} + \|\partial_x^2 w\|_{L^p(\Omega' \times I)} \leq C\, \|L\, w\|_{L^p(\Omega'' \times I)} \leq C\, \|w\|_W.$$

Notice that the norms on the left hand side topologize $W_p^{2,1}(\Omega' \times I)$. Thus, the restriction operator defined as

$$r \colon W \to X_1 := \{ v \in W_p^{2,1}(\Omega' \times I) : v(T) = 0 \}, \qquad r(w) := w|_{Q_c},$$

is continuous. For $p \in (1 + n/2, \infty)$ (which corresponds to $q \in (1, 1 + 2/n)$), one has a continuous embedding of $W_p^{2,1}(\Omega' \times I)$ into $C(\bar{\Omega}' \times \bar{I})$ (see [8, Theorem 10.4]). Thus, we have a continuous embedding

$$j \colon X_1 \to X_2, \qquad X_2 := \{ f \in C(\bar{\Omega}' \times \bar{I}) : f(T) = 0 \}.$$

Utilizing the restriction operator

$$s : X_2 \to C(\bar{\Omega}_c) \times C(\bar{Q}_c), \qquad s(f) := (f(0)|_{\bar{\Omega}_c}, f|_{\bar{Q}_c}),$$

we define the continuous linear operator

$$\Phi := s \circ j \circ r \colon W \to C(\bar{\Omega}_c) \times C(\bar{Q}_c).$$

This allows us to rewrite (3.1) and (3.2) into the following very compact form:

$$L^* y = \Phi^*(u_0, u). \tag{3.6}$$

As we have already established that L^* is continuously invertible, the well-posedness of (3.1) and (3.2) is now evident. It will soon be essential that Φ^* is injective. Because of

$$\langle \Phi^*(u_0, u), w \rangle_{W^*,W} = \int_{\bar{\Omega}_c} w(0)|_{\bar{\Omega}_c} \, du_0 + \int_{\bar{Q}_c} w|_{\bar{Q}_c} \, du,$$

it suffices to show that $\Phi \colon W \to C(\bar{\Omega}_c) \times C(\bar{Q}_c)$ has dense image. By virtue of Tietze's extension theorem (see [59, Theorem 35.1.]), the restriction operator s is surjective. So it suffices to show that $j \circ r$ has dense image. Next we observe that Ω' is an extension domain and that each element of X_2 can be arbitrarily well approximated by the restriction of a function $f \in C^\infty(\mathbb{R}^n \times \bar{I})$ with $f(T) = 0$. We pick a mollifier $\varphi \in C^\infty(\mathbb{R}^n)$ with $\operatorname{supp}(\varphi) \subset \Omega$ and $\varphi|_{\bar{\Omega}'} = 1$ and put

$$w(x,t) := \varphi(x) \, f(x,t) \qquad \forall (x,t) \in Q.$$

By construction, we have $w \in W$ and $j \circ r(w) = f$, showing that Φ has indeed a dense image. Thus

$$\Phi^* \colon \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c) \to W^*$$

is injective.

3.2.2 Fenchel duality

As we want to solve our problem numerically by utilizing Newton-based methods, we have to cope with the fact that, because of $q < 2$, the contribution $y \mapsto \frac{1}{q} \|y - y_\mathrm{d}\|_{L^q}^q$ in the objective function J is not twice differentiable. Even worse, $u \mapsto \|u\|_{\mathcal{M}(\bar{Q}_c)}$ and $u_0 \mapsto \|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}$ are not differentiable at all. Fortunately, as demonstrated in [29, Chapter 2.1.], Fenchel duality can help here: It allows us to transform problem (P) into an optimization problem (P^*) that enjoys sufficient differentiability to make it amenable to the semismooth Newton method. As a convenient side effect, this will also allow us to show that (P) has a unique solution.

We define the problem (P^*) as follows; as it will turn out in Theorem 3.5, this problem is indeed the Fenchel *pre*dual of (P):

$$\min_{w \in W} \; K(w) := F(w) + G(\Phi w), \tag{P^*}$$

where $F \colon W \to \mathbb{R}$ and $G \colon C(\bar{\Omega}_c) \times C(\bar{Q}_c) \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ are given by

$$F(w) := \frac{1}{p} \|Lw\|_{L^p(Q)}^p + \langle Lw, y_\mathrm{d} \rangle_{L^p(Q), L^q(Q)},$$

$$G(f_0, f) := \begin{cases} 0, & \text{if } \|f_0\|_{C(\bar{\Omega}_c)} \le \beta \text{ and } \|f\|_{C(\bar{Q}_c)} \le \alpha, \\ \infty, & \text{else.} \end{cases}$$

Theorem 3.3. *Let $q \in (1, 2]$ satisfy $q < 1 + 2/n$. Then problem (P^*) has a unique solution $\bar{w} \in W$.*

Proof. Let $\{w_k\}_k \subset W$ be a minimizing sequence so that

$$\lim_{k \to \infty} K(w_k) = \inf_{w \in W} K(w) =: \underline{K}.$$

From $K(0) = 0$, we know that $\underline{K} \leq 0$, which allows us to assume without loss of generality that

$$K(w_k) \leq 0 < \infty \qquad \forall k,$$

and hence $G(\Phi w_k) = 0$ for all k. With Hölder's inequality and Young's inequality in the form

$$-a\,b \geq -\frac{1}{p}\,a^p - \frac{1}{q}\,b^q,$$

we see

$$\begin{aligned} K(w_k) = F(w_k) &= \frac{1}{p}\|Lw_k\|^p_{L^p(Q)} + \langle Lw_k, y_\mathrm{d}\rangle_{L^p(Q),L^q(Q)} \\ &\geq \frac{1}{p}\|Lw_k\|^p_{L^p(Q)} - \|Lw_k\|_{L^p(Q)}\|y_\mathrm{d}\|_{L^q(Q)} \\ &\geq -\frac{1}{q}\|y_\mathrm{d}\|^q_{L^q(Q)}, \end{aligned}$$

and hence $\underline{K} > -\infty$. For all k we have $F(w_k) \leq K(w_k) \leq 0$, which implies that $\{w_k\}_k$ is a bounded sequence in W. Recall that W is reflexive. Hence the bounded sequence $\{w_k\}_k$ admits a weakly convergent subsequence: $w_{k'} \rightharpoonup \bar{w}$ in W as $k' \to \infty$. Likewise, we have $\Phi w_{k'} \rightharpoonup \Phi \bar{w}$ in $C(\bar{\Omega}_c) \times C(\bar{Q}_c)$. As the indicator function of a closed, convex set, G is convex and weakly lower semi continuous Thus, we are lead to $G(\Phi \bar{w}) = 0$. Also F is weakly lower semi continuous, so we obtain:

$$\underline{K} \leq K(\bar{w}) = F(\bar{w}) \leq \liminf_{k'\to\infty} F(w_{k'}) = \lim_{k'\to\infty} K(w_{k'}) = \lim_{k\to\infty} K(w_k) = \underline{K}.$$

This shows that $\bar{w}$ is a minimizer of (P^*). Moreover, G is convex and F is strictly convex (L is injective and $2 \leq p < \infty$), hence $K = F + G \circ \Phi$ is also strictly convex. Thus, there cannot be more than one solution of (P^*). □

Now we recall the Fenchel duality theorem in a similar notation as in [27] and [29, Chapter 1.1.3.]. We also refer to [33, Chapter III.4.]. For a convex, lower semi continuous functional $F\colon R \to \bar{\mathbb{R}}$ on a normed space R with $\inf_{r\in R} F(r) < \infty$, we define its *Fenchel conjugate* by

$$F^*\colon R^* \to \bar{\mathbb{R}}, \quad F^*(\varrho) := \sup_{r\in R} \langle \varrho, r\rangle_{R^*,R} - F(r).$$

Theorem 3.4 (Fenchel Duality)**.** *Let R and S be normed spaces with topological duals R^* and S^* and let $\Lambda : R \to S$ be a continuous linear operator. Let $F\colon R \to \bar{\mathbb{R}}$ and $G\colon S \to \bar{\mathbb{R}}$ be convex, lower semi continuous functionals and suppose that F and G are not identically equal to ∞. Consider the* primal problem

$$\inf_{r\in R} F(r) + G(\Lambda r), \tag{3.7}$$

and the dual problem

$$\sup_{\sigma\in S^*} -F^*(\Lambda^*\sigma) - G^*(-\sigma). \tag{3.8}$$

Suppose the following two conditions are fulfilled:

- *The primal problem* (3.7) *has at least one solution.*
- *The* regular point conditions *is fulfilled, i.e., there exists an $r_0 \in R$, such that $F(r_0) < \infty$ and $G(\Lambda r) < \infty$ for all r in a sufficiently small neighborhood of r_0.*

Then also the dual problem *has at least one solution and one has the identity*

$$\min_{r\in R} F(r) + G(\Lambda r) = \max_{\sigma\in S^*} -F^*(\Lambda^*\sigma) - G^*(-\sigma). \tag{3.9}$$

Furthermore, for $r \in R$ and $\sigma \in S^*$, the following three statements are equivalent:

1. *r is a solution of the primal problem* (3.7) *and σ is a solution of the dual problem* (3.8).
2. $F(r) + G(\Lambda r) = -F^*(\Lambda^*\sigma) - G^*(-\sigma)$.
3. $\Lambda^*\sigma \in \partial F(r)$ *and* $-\sigma \in \partial G(\Lambda r)$.

We are going to apply the Fenchel duality theorem to $R = W$, $S := C(\bar{\Omega}_c) \times C(\bar{Q}_c)$, and $\Lambda = \Phi$. To this end, we show first that (P^*) and (P) are dual to each other.

Theorem 3.5. *The Fenchel dual problem of* (P^*) *coincides with* (P).

Proof. It suffices to show that $J(u_0, u) = F^*(\Phi^*(u_0, u)) + G^*(-u_0, -u)$.
In order to calculate F^*, we use the equivalence of the following statements on $w \in W$ and $\xi \in W^*$:

$$F^*(\xi) = \langle \xi, w \rangle_{W^*,W} - F(w) \quad \text{if and only if} \quad \xi \in \partial F(w). \tag{3.10}$$

Here $\partial F(w)$ denotes the subdifferential of the convex functional F. Since F is Fréchet differentiable, we have $\partial F(w) = \{DF(w)\}$. Hence $\xi \in \partial F(w)$ is given by $\xi = L^*(|Lw|^{p-2}Lw + y_\mathrm{d})$, where $\frac{1}{p} + \frac{1}{q} = 1$. Solving for w leads to

$$w = L^{-1}(\operatorname{sgn}(L^{-*}\xi - y_\mathrm{d})\ |L^{-*}\xi - y_\mathrm{d}|^{\frac{1}{(p-1)}}). \tag{3.11}$$

Substituting this into (3.10) and utilizing $\langle \xi, L^{-1}z \rangle_{W^*,W} = \langle z, L^{-*}\xi \rangle_{L^p(Q),L^q(Q)}$ for all $z \in L^p(Q)$, we derive

$$\begin{aligned}
F^*(\xi) &= \langle \xi, w \rangle_{W^*,W} - \frac{1}{p}\|Lw\|^p_{L^p(Q)} - \langle Lw, y_\mathrm{d} \rangle_{L^p(Q),L^q(Q)} \\
&= \langle \operatorname{sgn}(L^{-*}\xi - y_\mathrm{d})\,|L^{-*}\xi - y_\mathrm{d}|^{\frac{1}{(p-1)}}, L^{-*}\xi - y_\mathrm{d} \rangle_{L^p(Q),L^q(Q)} - \frac{1}{p}\|L^{-*}\xi - y_\mathrm{d}\|^q_{L^q(Q)} \\
&= \left(1 - \frac{1}{p}\right) \int_Q |L^{-*}\xi - y_\mathrm{d}|^{\frac{p}{(p-1)}}\, dx\, dt \\
&= \frac{1}{q}\|L^{-*}\xi - y_\mathrm{d}\|^q_{L^q(Q)}.
\end{aligned}$$

In order to derive $G^*(u_0, u)$ we can interpret $G(f_0, f)$, as consisting of two summands, which represent an indicator function with only one constraint respectively, where we want to use the following notation:

$$\ell_\alpha(0, f) + \ell_\beta(f_0, 0) := G(f_0, f).$$

Here, we make use of $(f_0, f) \in C(\bar{\Omega}_c) \times C(\bar{Q}_c)$ and

$$\left(C(\bar{\Omega}_c) \times C(\bar{Q}_c)\right)^* = \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c).$$

From [67, Theorem 2.2.8] we know that for $\tilde{u} = (u_0, 0) + (0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c)$ it holds that :

$$G^*(\tilde{u}) = (\ell_\alpha + \ell_\beta)^*(u_0, u) = \ell^*_\alpha(0, u) + \ell^*_\beta(u_0, 0).$$

Looking at both conjugates separately, we can use (3.10) and derive:

$$\begin{aligned}
\ell^*_\alpha(0, u) &= \sup_{(0,f) \in C(\bar{\Omega}_c) \times C(\bar{Q}_c)} \langle (0, u), (0, f) \rangle - \ell_\alpha(0, f) \\
&= \sup_{f \in C(\bar{Q}_c),\, \|f\|_{C(\bar{Q}_c)} \le \alpha} \int_{\bar{Q}_c} f\, du \\
&= \alpha\, \|u\|_{\mathcal{M}(\bar{Q}_c)},
\end{aligned} \tag{3.12}$$

where

$$\langle (u_0, u), (f_0, f) \rangle := \int_{\bar{\Omega}_c} f_0 \, du_0 + \int_{\bar{Q}_c} f \, du.$$

Analogously, we observe that $\ell_\beta^*(u_0, 0) = \beta \|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}$. Assembling this information, we obtain

$$G^*(\tilde{u}) = \alpha \|u\|_{\mathcal{M}(\bar{Q}_c)} + \beta \|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}.$$

□

Theorem 3.6. *Problem (P) has a unique solution $(\bar{u}_0, \bar{u})$, which is characterized by*

$$\Phi^*(\bar{u}_0, \bar{u}) = DF(\bar{w}) = L^*(|L\bar{w}|^{p-2} L\bar{w} + y_\mathrm{d}) \quad \textit{and} \quad -(\bar{u}_0, \bar{u}) \in \partial G(\Phi(\bar{w})), \tag{3.13}$$

where $\bar{w}$ is the unique solution of (P^*). The optimal state $\bar{y}$ can be retrieved from $\bar{w}$ via

$$\bar{y} = L^{-*} \Phi^*(\bar{u}_0, \bar{u}) = |L\bar{w}|^{p-2} L\bar{w} + y_\mathrm{d}.$$

Proof. We have seen already in the proof of Theorem 3.3 that both F and G are convex and lower semi continuous and that F is even strictly convex. Furthermore we set $\Lambda = \Phi$. By Theorem 3.3, the problem (P^*) has a at least one minimizer. For $w_0 = 0$, we have $\Phi w_0 = 0$ and $Lw_0 = 0$, showing that $F(w_0) < \infty$ and $G(\Phi w_0) < \infty$. Due to $\alpha, \beta > 0$, G is continuous at Φw_0 so that also the regular point condition is fulfilled. Thus, we may apply the Fenchel duality theorem, which implies that (P) has at least one solution. Since $L^{-*}\Phi^*$ is injective and $q > 1$, J is strictly convex, hence there is only one solution $(\bar{u}_0, \bar{u})$. By the the third condition from Theorem 3.4, each solution has to satisfy $\Phi^*(\bar{u}_0, \bar{u}) \in \partial F(\bar{w}) = \{DF(\bar{w})\}$ and $-(\bar{u}_0, \bar{u}) \in \partial G(\Phi(\bar{w}))$, where $\bar{w}$ is a solution of (P^*). □

3.2.3 Sparsity structure

We recall the optimality conditions for the solution $(\bar{u}_0, \bar{u})$ of (P) from [20, Theorem 3.1.] and the resulting sparsity structure [20, Corollary 3.2.] of the optimal controls $(\bar{u}_0, \bar{u})$. These results can be directly transferred to our setting since the latter is a special case of the formulation in [20].

Lemma 3.7. *Let $(\bar{u}_0, \bar{u})$ denote a solution to (P) with associated state $\bar{y}$. Denote by $\bar{w} \in W$ the unique solution of*

$$L\bar{w} = |\bar{y} - y_\mathrm{d}|^{q-2}(\bar{y} - y_\mathrm{d}) \in L^p(Q),$$

which follows from the first part of (3.13) by solving for $L\bar{w}$. This $\bar{w}$ satisfies

$$\int_{\bar{\Omega}_c} \bar{w}(0) \, d\bar{u}_0 + \beta \|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)} = 0, \tag{3.14}$$

$$\int_{\bar{Q}_c} \bar{w} \, d\bar{u} + \alpha \|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} = 0, \tag{3.15}$$

and

$$\begin{aligned}
|\bar{w}(x,t)| &= \alpha \quad \textit{for all } (x,t) \in \bar{Q}_c \cap \operatorname{supp}(\bar{u}),\\
|\bar{w}(x,t)| &\le \alpha \quad \textit{for all } (x,t) \in \bar{Q}_c \setminus \operatorname{supp}(\bar{u}),\\
|\bar{w}(x,0)| &= \beta \quad \textit{for all } x \in \bar{\Omega}_c \cap \operatorname{supp}(\bar{u}_0),\\
|\bar{w}(x,0)| &\le \beta \quad \textit{for all } x \in \bar{\Omega}_c \setminus \operatorname{supp}(\bar{u}_0).
\end{aligned}$$

Proof. From the first part of (3.13) with $\frac{1}{q} + \frac{1}{p} = 1$ we deduce

$$\begin{aligned} \bar{y} = L^{-*}\Phi^*(\bar{u}_0, \bar{u}) &= |L\,\bar{w}|^{p-2} L\,\bar{w} + y_\mathrm{d} \\ \Rightarrow \qquad \bar{y} - y_\mathrm{d} &= |L\,\bar{w}|^{p-2} L\,\bar{w} \\ \Rightarrow \ |\bar{y} - y_\mathrm{d}|^{q-2}(\bar{y} - y_\mathrm{d}) &= L\,\bar{w}. \end{aligned}$$

By optimality of $(\bar{u}_0, \bar{u})$ we have

$$J(\bar{u}_0, \bar{u}) \le J(u_0, u) \qquad \forall\, (u_0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c).$$

Also, J is convex, so for $\lambda \in [0, 1]$ it holds

$$\begin{aligned} & J(\bar{u}_0, \bar{u}) \le J(\bar{u}_0 + \lambda(u_0 - \bar{u}_0), \bar{u} + \lambda(u - \bar{u})) && \forall\, (u_0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c) \\ \Rightarrow \qquad & 0 \le \frac{J(\bar{u}_0 + \lambda(u_0 - \bar{u}_0), \bar{u} + \lambda(u - \bar{u})) - J(\bar{u}_0, \bar{u})}{\lambda} && \forall\, (u_0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c). \end{aligned}$$

Now let $\lambda \searrow 0$ and with $J(u_0, u) = F^*(\Phi^*(u_0, u)) + G^*(-u_0, -u)$ for all $(u_0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c)$ we get

$$\begin{aligned} 0 &\le dF^*(\Phi^*(\bar{u}_0, \bar{u}); (u_0 - \bar{u}_0, u - \bar{u})) + dG^*((-\bar{u}_0, -\bar{u}); (u_0 - \bar{u}_0, u - \bar{u})) \\ &= dF^*(\Phi^*(\bar{u}_0, \bar{u}); (u_0 - \bar{u}_0, u - \bar{u})) + dG^*((-\bar{u}_0, 0); (u_0 - \bar{u}_0, 0)) + dG^*((0, -\bar{u}); (0, u - \bar{u})) \\ &\le dF^*(\Phi^*(\bar{u}_0, \bar{u}); (u_0 - \bar{u}_0, u - \bar{u})) + \beta\left(\|u_0\|_{\mathcal{M}(\bar{\Omega}_c)} - \|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)}\right) + \alpha\left(\|u\|_{\mathcal{M}(\bar{Q}_c)} - \|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)}\right) \\ &= \int_{\bar{\Omega}_c} \bar{w}(0)\, d(u_0 - \bar{u}_0) + \int_{\bar{Q}_c} \bar{w}\, d(u - \bar{u}) + \beta\left(\|u_0\|_{\mathcal{M}(\bar{\Omega}_c)} - \|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)}\right) + \alpha\left(\|u\|_{\mathcal{M}(\bar{Q}_c)} - \|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)}\right). \end{aligned} \tag{3.16}$$

First, we set $u_0 = \bar{u}_0$ in (3.16) to get

$$-\int_{\bar{Q}_c} \bar{w}\, d(u - \bar{u}) + \alpha\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} \le \alpha\|u\|_{\mathcal{M}(\bar{Q}_c)} \qquad \forall u \in \mathcal{M}(\bar{Q}_c). \tag{3.17}$$

By setting $u = 0$ and $u = 2\bar{u}$ in this inequality we get

$$\int_{\bar{Q}_c} \bar{w}\, d\bar{u} + \alpha\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} \le 0 \qquad \text{and} \qquad -\int_{\bar{Q}_c} \bar{w}\, d\bar{u} - \alpha\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} \le 0,$$

which immediately shows (3.15). Analogously, we can choose $u = \bar{u}$ in (3.16) and then combine the information from setting $u_0 = 0$ and $u_0 = 2\bar{u}_0$ to see (3.14). We remark, that these conditions can also be followed from the second part of (3.13), where we have that $-(\bar{u}_0, \bar{u}) \in \partial G(\Phi(\bar{w})) \Leftrightarrow \Phi(\bar{w}) \in \partial G^*(-(\bar{u}_0, \bar{u}))$.
We insert (3.15) into (3.17) and see

$$-\int_{\bar{Q}_c} \bar{w}\, du \le \alpha\|u\|_{\mathcal{M}(\bar{Q}_c)} \qquad \forall u \in \mathcal{M}(\bar{Q}_c).$$

Let $u = \delta_x \otimes \delta_t$ for arbitrary $(x, t) \in \bar{Q}_c$, then

$$\begin{aligned} & -\int_{\bar{Q}_c} \bar{w}\, d\,(\delta_x \otimes \delta_t) \le \alpha\|\delta_x \otimes \delta_t\|_{\mathcal{M}(\bar{Q}_c)} \\ \Rightarrow \qquad & -\bar{w}(x, t) \le \alpha \sup_{\|f\|_{C(\bar{Q}_c)} \le 1} \int_{\bar{Q}_c} f\, d\,(\delta_x \otimes \delta_t) \\ \Rightarrow \qquad & -\bar{w}(x, t) \le \alpha. \end{aligned}$$

In the same way for $u = -(\delta_x \otimes \delta_t)$ we get $\bar{w}(x, t) \le \alpha$. Since $(x, t) \in \bar{Q}_c$ was chosen arbitrary, altogether, it holds

$$|\bar{w}(x, t)| \le \alpha \qquad \forall (x, t) \in \bar{Q}_c.$$

Furthermore, from (3.15) we can derive

$$\int_{\bar{Q}_c} \bar{w}\, d\bar{u} = -\alpha \|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} = \sup_{\|f\|_{C(\bar{Q}_c)} \leq 1} \int_{\bar{Q}_c} -\alpha f\, d\bar{u} = \sup_{\|f\|_{C(\bar{Q}_c)} = \alpha} \int_{\bar{Q}_c} f\, d\bar{u}.$$

For the case that $\bar{u} \neq 0$ the above equality shows

$$|\bar{w}(x,t)| = \alpha \qquad \forall (x,t) \in \bar{Q}_c \cap \operatorname{supp}(\bar{u}).$$

Analogously, we can deduce $|\bar{w}(x,0)| \leq \beta$ for all $x \in \bar{\Omega}_c$ and $|\bar{w}(x,0)| = \beta$ for all $x \in \bar{\Omega}_c \cap \operatorname{supp}(\bar{u}_0)$. □

We introduce the following helpful Lemma, which is based on [14, Lemma 3.4]. We change the sign in the equations and therefore find the signs also changed in the resulting sets. We give the proof in detail to demonstrate that the change of sign does indeed work throughout the whole argument.

Lemma 3.8. *Let $u_0 \in \mathcal{M}(\bar{\Omega}_c), u \in \mathcal{M}(\bar{Q}_c), f_0 \in C(\bar{\Omega}_c), f \in C(\bar{Q}_c)$, all of them not zero, be such that*

$$\begin{aligned} \langle u, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} &= -\|u\|_{\mathcal{M}(\bar{Q}_c)} \|f\|_{C(\bar{Q}_c)}, \\ \langle u_0, f_0 \rangle_{\mathcal{M}(\bar{\Omega}_c), C(\bar{\Omega}_c)} &= -\|u_0\|_{\mathcal{M}(\bar{\Omega}_c)} \|f_0\|_{C(\bar{\Omega}_c)}, \end{aligned} \tag{3.18}$$

and let $u_0 = u_0^+ - u_0^-, u = u^+ - u^-$ be the Jordan decompositions. Then we have

$$\operatorname{supp}(u^+) \subset \{(x,t) \in \bar{Q}_c : f(x,t) = -\|f\|_{C(\bar{Q}_c)}\}, \tag{3.19}$$

$$\operatorname{supp}(u^-) \subset \{(x,t) \in \bar{Q}_c : f(x,t) = +\|f\|_{C(\bar{Q}_c)}\}, \tag{3.20}$$

$$\operatorname{supp}(u_0^+) \subset \{x \in \bar{\Omega}_c : f_0(x) = -\|f_0\|_{C(\bar{\Omega}_c)}\}, \tag{3.21}$$

$$\operatorname{supp}(u_0^-) \subset \{x \in \bar{\Omega}_c : f_0(x) = +\|f_0\|_{C(\bar{\Omega}_c)}\}. \tag{3.22}$$

Proof. We will only prove (3.19), since (3.20)-(3.22) can be proven analogously.
Let $\nu \in \mathcal{M}(\bar{Q}_c)$, such that $\|\nu\|_{\mathcal{M}(\bar{Q}_c)} \leq \|u\|_{\mathcal{M}(\bar{Q}_c)}$, then

$$\langle u, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} = -\|u\|_{\mathcal{M}(\bar{Q}_c)} \|f\|_{C(\bar{Q}_c)} \leq -\|\nu\|_{\mathcal{M}(\bar{Q}_c)} \|f\|_{C(\bar{Q}_c)} \leq \langle \nu, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)}. \tag{3.23}$$

We have also that

$$\begin{aligned} \langle u, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} &= \langle u^+, f^+ \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} + \langle u^-, f^- \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} - \langle u^+, f^- \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} - \langle u^-, f^+ \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} \\ &\geq -\langle u^+, f^- \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} - \langle u^-, f^+ \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} \end{aligned}$$

Moreover, the inequality is strict, unless $\langle u^+, f^+ \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} = \langle u^-, f^- \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} = 0$, which happens if

$$\begin{aligned} \operatorname{supp}(u^+) \subset A_- &:= \left\{(x,t) \in \bar{Q}_c : f(x,t) \leq 0\right\}, \\ \operatorname{supp}(u^-) \subset A_+ &:= \left\{(x,t) \in \bar{Q}_c : f(x,t) \geq 0\right\}. \end{aligned}$$

Now, we define $\nu = \nu^+ - \nu^-$, as $\nu^+ = u^+{}_{|A_-}, \nu^- = u^-{}_{|A_+}$. So obviously it holds $\|\nu\|_{\mathcal{M}(\bar{Q}_c)} \leq \|u\|_{\mathcal{M}(\bar{Q}_c)}$. Furthermore, it is easy to see that

$$\langle u, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} = \langle \nu, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} + \langle u^+{}_{|A_+}, f^+ \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} + \langle u^-{}_{|A_-}, f^- \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)}.$$

Let us assume that $\operatorname{supp}(u^+) \not\subset A_-$ or $\operatorname{supp}(u^-) \not\subset A_+$, then $\langle u, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)} > \langle \nu, f \rangle_{\mathcal{M}(\bar{Q}_c), C(\bar{Q}_c)}$, but this contradicts (3.23), so we have that

$$\operatorname{supp}(u^+) \subset A_- \quad \text{and} \quad \operatorname{supp}(u^-) \subset A_+.$$

Now, we distinguish two cases:

1^{st} *case:* $\min_{(x,t)\in\bar{Q}_c} f(x,t) > -\|f\|_{C(\bar{Q}_c)}$
We will prove that in this case $u^+ = 0$, which gives

$$\operatorname{supp}(u^+) = \emptyset = \{(x,t) \in \bar{Q}_c : f(x,t) = -\|f\|_{C(\bar{Q}_c)}\}.$$

Let $(x_0,t_0) \in \bar{Q}_c$, such that $f(x_0,t_0) = \|f\|_{C(\bar{Q}_c)}$ and define $\nu = -u^+(\bar{Q}_c)(\delta_{x_0}\otimes\delta_{t_0}) - u^-$. For this choice it is obvious that $\|\nu\|_{\mathcal{M}(\bar{Q}_c)} = \|u\|_{\mathcal{M}(\bar{Q}_c)}$. Assume $u^+ \neq 0$, then

$$\begin{aligned}\langle \nu, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)} &= -\|f\|_{C(\bar{Q}_c)}u^+(\bar{Q}_c) - \langle u^-, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)}\\ &< \min_{(x,t)\in\bar{Q}_c} f(x,t)\, u^+(\bar{Q}_c) - \langle u^-, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)}\\ &\leq \langle u, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)}\end{aligned}$$

This is in contradiction with (3.23), so $u^+ = 0$ must hold and this shows (3.19).

2^{nd} *case:* $\min_{(x,t)\in\bar{Q}_c} f(x,t) = -\|f\|_{C(\bar{Q}_c)}$
Let $(x_0,t_0) \in \bar{Q}_c$, such that $f(x_0,t_0) = -\|f\|_{C(\bar{Q}_c)}$. We will show the claim in this case by contradiction. To this end let

$$S \coloneqq \{(x,t) \in \bar{Q}_c : 0 \geq f(x,t) > -\|f\|_{C(\bar{Q}_c)}\}.$$

Now, assume $u^+(S) > 0$. Here, we define $\nu = u^+(\bar{Q}_c)(\delta_{x_0}\otimes\delta_{t_0}) - u^-$ and again, it is obvious that $\|\nu\|_{\mathcal{M}(\bar{Q}_c)} = \|u\|_{\mathcal{M}(\bar{Q}_c)}$. Due to $u^+(S) > 0$, we see

$$\begin{aligned}\langle \nu, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)} &= -\|f\|_{C(\bar{Q}_c)}u^+(\bar{Q}_c) - \langle u^-, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)}\\ &< \langle u^+, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)} - \langle u^-, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)}\\ &= \langle u, f\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)}\end{aligned}$$

So again, this contradicts (3.23), and $u^+(S) = 0$ must hold. Since we already have $\operatorname{supp}(u^+) \subset A_-$ and $A_- \setminus S = \{(x,t) \in \bar{Q}_c : f(x,t) = -\|f\|_{C(\bar{Q}_c)}\}$, this shows (3.19). □

Remark 3.9. *Under the assumptions of Lemma 3.7 we have the following sparsity structure:*

$$\begin{aligned}\operatorname{supp}(\bar{u}^+) &\subset \{(x,t) \in \bar{Q}_c : \bar{w}(x,t) = -\alpha\},\\ \operatorname{supp}(\bar{u}^-) &\subset \{(x,t) \in \bar{Q}_c : \bar{w}(x,t) = +\alpha\},\\ \operatorname{supp}(\bar{u}_0^+) &\subset \{x \in \bar{\Omega}_c : \bar{w}(x,0) = -\beta\},\\ \operatorname{supp}(\bar{u}_0^-) &\subset \{x \in \bar{\Omega}_c : \bar{w}(x,0) = +\beta\},\end{aligned}$$

where $\bar{u} = \bar{u}^+ - \bar{u}^-$ and $\bar{u}_0 = \bar{u}_0^+ - \bar{u}_0^-$ are the Jordan decompositions. Let us note that $\bar{w}$ is the adjoint variable in (P) in the sense of (3.2).

Proof. For $\bar{u} = 0$ and $\bar{u}_0 = 0$ the claim is obvious. Let $\bar{u} \neq 0$ and $\bar{u}_0 \neq 0$, then from (3.14) and (3.15) in combination with $|\bar{w}(x,t)| = \alpha$ and $|\bar{w}(x,0)| = \beta$ we get

$$\begin{aligned}\langle \bar{u}, \Phi(\bar{w})\rangle_{\mathcal{M}(\bar{Q}_c),C(\bar{Q}_c)} &= \int_{\bar{Q}_c} \bar{w}\, d\bar{u} \qquad = -\alpha\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} = -|\bar{w}(x,t)|\,\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} = -\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)}\|\Phi(\bar{w})\|_{C(\bar{Q}_c)},\\ \langle \bar{u}_0, \Phi(\bar{w}(0))\rangle_{\mathcal{M}(\bar{\Omega}_c),C(\bar{\Omega}_c)} &= \int_{\bar{\Omega}_c} \bar{w}(0)\, d\bar{u}_0 = -\beta\|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)} = -|\bar{w}(x,0)|\,\|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)} = -\|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)}\|\Phi(\bar{w}(0))\|_{C(\bar{\Omega}_c)}.\end{aligned}$$

Then, by Lemma 3.8 and inserting properties of the embedding the claim follows. □

If one considers it as the generic case that the function $\bar{w}$ is not constant on sets of measure greater than zero, the controls have support sets of measure zero. This is our motivation to propose a discretization strategy which reflects this behavior on the discrete level in space and time.

3.3 Variational discretization

Here we want to achieve the desired maximal discrete sparsity, i.e., Dirac-measures in space-time, by choosing the Petrov-Galerkin ansatz and test space that will induce this structure. The variational discretization concept was introduced in [49] and its key feature is to not discretize the control space. Instead, via the discretization of the test space and the optimality conditions, an implicit discretization of the control is achieved. This is how we *control* the discrete structure of the controls. Looking at the relation (3.11) between the optimal adjoint state $\bar{w}$ and $\bar{\xi} = \Phi^*(u_0, u)$, it is obvious, that the discrete structure of the test space affects the structure of the optimal controls $(u_0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c)$.

This motivates the following choice of discrete spaces: We define the state space $\mathcal{Y}_\sigma$ consisting of continuous and piecewise linear functions in space and piecewise constant functions in time, whereas we define the test space $\mathcal{W}_\sigma$ consisting of continuous and piecewise linear functions in space and time.

In the following we discretize (P) and analyze the structure of the controls. We will see that the above choice of discrete state and test spaces in combination with the optimality system of the discrete problem induces Dirac measures in space and time for the controls. Afterwards, we discuss the existence and uniqueness of solutions to the variational discrete problem. We then prove the convergence properties stated in Theorem 3.2. Afterwards we discretize (P^*), reformulate the problem equivalently, and derive an optimality system by a Lagrange approach. If the necessary conditions are fulfilled, we can apply a semismooth Newton method to solve the optimality system. Utilizing Fenchel duality, we can finally calculate the optimal solution of (P).

As a first step to characterizing the discrete spaces, we have to set up the space-time grid. Define the partition $0 = t_0 < t_1 < \ldots < t_{N_\tau} = T$. The time interval I is decomposed in subintervals $I_k := (t_{k-1}, t_k]$ for $k = 1, \ldots, N_{\tau-1}$ and $I_{N_\tau} := (t_{N_\tau - 1}, t_{N_\tau})$. We point out that we defined the intervals I_k such that they cover the full time interval I, which is crucial because we deal with measures that can be supported on isolated points. The temporal grid size is denoted by $\tau = \max_{1 \le k \le N_\tau} \tau_k$, where $\tau_k := t_k - t_{k-1}$. Let $\mathcal{K}_h$ be a finite triangulation of Ω with grid size $h = \max_{K \in \mathcal{K}_h} \operatorname{diam}(K)$. We set $\bar{\Omega}_h := \bigcup_{K \in \mathcal{K}_h} K$ and denote by Ω_h the interior, by Γ_h the boundary of $\bar{\Omega}_h$, and by $Q_h := \Omega_h \times I$ the discrete space-time domain. We assume that vertices on Γ_h are points on Γ. The interior vertices of $\mathcal{K}_h$ are denoted by $\{x_j\}_{j=1}^{N_h}$. We combine the two discretization parameters τ and h into the vector $\sigma = (\tau, h)$ and define the following discrete state and test spaces:

$$\begin{aligned} \mathcal{Y}_\sigma &:= \operatorname{span}\{e_{x_j} \otimes \chi_k : 1 \le j \le N_h \text{ and } 1 \le k \le N_\tau\}, \\ \mathcal{W}_\sigma &:= \operatorname{span}\{e_{x_j} \otimes e_{t_k} : 1 \le j \le N_h \text{ and } 0 \le k \le N_\tau - 1\}, \end{aligned} \tag{3.24}$$

such that $w_\sigma(T) = 0$ for $w_\sigma \in \mathcal{W}_\sigma$ is ensured. Here, $(e_{x_j})_{j=1}^{N_h}$ and $(e_{t_k})_{k=0}^{N_\tau - 1}$ denote the nodal basis formed by continuous, piecewise linear functions on Ω_h and $\bar{I}$, respectively. Moreover, χ_k denotes the indicator function of the time interval I_k. We also define the space

$$Y_h := \operatorname{span}\{e_{x_j} : 1 \le j \le N_h\}.$$

In order to set up the variational discrete state equation, we start by deriving a very weak formulation of (3.1), which will be discretized afterwards. By multiplication with $w \in W$, integration over the domain Q, and utilizing $w(x, T) = 0$, we arrive at

$$\int_Q (-y\,\partial_t w + \nabla y \nabla w)\,dx\,dt = \int_{\bar{\Omega}_c} w(0)\,du_0 + \int_{\bar{Q}_c} w\,du. \tag{3.25}$$

Inserting $y_\sigma \in \mathcal{Y}_\sigma$ and testing against all $w_\sigma \in \mathcal{W}_\sigma$ yields the following variational discrete representation of the state equation: Find $y_\sigma \in \mathcal{Y}_\sigma$, such that

$$\int_Q (-y_\sigma \partial_t w_\sigma + \nabla y_\sigma \nabla w_\sigma)\, dx\, dt = \int_{\bar{\Omega}_c} w_\sigma(0)\, du_0 + \int_{\bar{Q}_c} w_\sigma\, du \tag{3.26}$$

holds for all $w_\sigma \in \mathcal{W}_\sigma$. This allows us to formulate the variational discrete problem

$$\min_{(u_0,u)\in\mathcal{M}(\bar{\Omega}_c)\times\mathcal{M}(\bar{Q}_c)} J_\sigma(u_0,u) \coloneqq \frac{1}{q}\|y_\sigma(u_0,u) - y_\mathrm{d}\|^q_{L^q(Q_h)} + \alpha\|u\|_{\mathcal{M}(\bar{Q}_c)} + \beta\|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}, \tag{P_σ}$$

where $y_\sigma(u_0,u)$ denotes the unique solution of (3.26).

We refrain from giving a detailed derivation for an optimality system and the sparsity structure for this problem as this would closely follow the procedure in the continuous setting (see Lemma 3.7 and Remark 3.9). Instead, we focus on analyzing how the controls (u_0,u) are implicitly discretized. First we define the following sets of indices:

$$\mathcal{I}_\sigma \coloneqq \{(j,k) : (x_j,t_k) \in \bar{Q}_c\} \qquad \text{and} \qquad \mathcal{I}_h \coloneqq \{j : x_j \in \bar{\Omega}_c\}.$$

We may suppose that the space-time discretization is sufficiently fine so that $\mathcal{I}_\sigma$ and $\mathcal{I}_h$ are nonempty. Utilizing these sets, we define the following discrete spaces:

$$V_h \coloneqq \operatorname{span}\{e_{x_j}|_{\bar{\Omega}_c} : j \in \mathcal{I}_h\} \quad \text{and} \quad \mathcal{V}_\sigma \coloneqq \operatorname{span}\{(e_{x_j}\otimes e_{t_k})|_{\bar{Q}_c} : (j,k)\in\mathcal{I}_\sigma\}.$$

Notice that both spaces are spaces of continuous functions, i.e., we have $V_h \subset C(\bar{\Omega}_c)$ and $\mathcal{V}_\sigma \subset C(\bar{Q}_c)$. The discrete version of the mapping Φ, decomposed into two parts, reads as follows:

$$\Phi_h : \mathcal{W}_\sigma \to V_h, \quad \Phi_h(w_\sigma) \coloneqq w_\sigma(0)|_{\bar{\Omega}_c}, \tag{3.27}$$

$$\Phi_\sigma : \mathcal{W}_\sigma \to \mathcal{V}_\sigma, \quad \Phi_\sigma(w_\sigma) \coloneqq w_\sigma|_{\bar{Q}_c}. \tag{3.28}$$

We suppose that $\bar{\Omega}_c$ is polygonal and that $\mathcal{K}_h$ is an exact triangulations of $\bar{\Omega}_c$, i.e., $\bar{\Omega}_c = \bigcup_{K\in\mathcal{K}_h : K\subset\bar{\Omega}_c} K$ and similarly for $\bar{Q}_c$.
From the discrete optimality system (whose precise derivation has been omitted), we obtain:

$$\max_{j\in\mathcal{I}_h} |\bar{w}_{j,0}| = \|\bar{w}_\sigma(0)\|_{\infty,\bar{\Omega}_c} = \|\Phi_h(\bar{w}_\sigma)\|_{\infty,\bar{\Omega}_c} \le \beta, \tag{3.29}$$

$$\max_{(j,k)\in\mathcal{I}_\sigma} |\bar{w}_{j,k}| = \|\bar{w}_\sigma\|_{\infty,\bar{Q}_c} \quad = \|\Phi_\sigma(\bar{w}_\sigma)\|_{\infty,\bar{Q}_c} \le \alpha, \tag{3.30}$$

where the $\bar{w}_{j,k}$ denote the coefficients of the optimal adjoint variable $\bar{w}_\sigma \in \mathcal{W}_\sigma$, i.e.,

$$\bar{w}_\sigma = \sum_{j=1}^{N_h}\sum_{k=0}^{N_\tau-1} \bar{w}_{j,k}\,(e_{x_j}\otimes e_{t_k}).$$

Here, we mention the mapping Φ explicitly for clarity. As discrete sparsity structure (analogous to Remark 3.9), we obtain the following

$$\begin{aligned}
&\operatorname{supp}(\bar{u}_0^+) \subset \{x\in\bar{\Omega}_c : \bar{w}_\sigma(x,0) = -\beta\},\\
&\operatorname{supp}(\bar{u}_0^-) \subset \{x\in\bar{\Omega}_c : \bar{w}_\sigma(x,0) = +\beta\},\\
&\operatorname{supp}(\bar{u}^+) \subset \{(x,t)\in\bar{Q}_c : \bar{w}_\sigma(x,t) = -\alpha\},\\
&\operatorname{supp}(\bar{u}^-) \subset \{(x,t)\in\bar{Q}_c : \bar{w}_\sigma(x,t) = +\alpha\}.
\end{aligned} \tag{3.31}$$

By construction, the discrete adjoint state $\bar{w}_\sigma$ is piecewise linear, both in space and time. Thus, $\Phi_\sigma(\bar{w}_\sigma)$ and $\Phi_h(\bar{w}_\sigma)$ attain their extremal values $\pm\alpha$ and $\pm\beta$ in the grid points contained in $\bar{\Omega}_c$ and $\bar{Q}_c$, respectively. Generically, $\Phi_\sigma(\bar{w}_\sigma)$ and $\Phi_h(\bar{w}_\sigma)$ attain their extrema *only* in these grid points, in which case we have

$$\operatorname{supp}(\bar{u}) \subset \{(x_j, t_k) : (j,k) \in \mathcal{I}_\sigma\} \qquad \text{and} \qquad \operatorname{supp}(\bar{u}_0) \subset \{x_j : j \in \mathcal{I}_h\}.$$

This leads us in a natural way to the discrete control spaces

$$U_h = \operatorname{span}\{\delta_{x_j} : j \in \mathcal{I}_h\} \qquad \subset \mathcal{M}(\bar{\Omega}_c), \tag{3.32}$$

$$\mathcal{U}_\sigma = \operatorname{span}\{\delta_{x_j} \otimes \delta_{t_k} : (j,k) \in \mathcal{I}_\sigma\} \subset \mathcal{M}(\bar{Q}_c). \tag{3.33}$$

Notice also that the natural pairings $\mathcal{M}(\bar{\Omega}_c) \times C(\bar{\Omega}_c) \to \mathbb{R}$ and $\mathcal{M}(\bar{Q}_c) \times C(\bar{Q}_c) \to \mathbb{R}$ induce the dualities $V_h^* \cong U_h$ and $\mathcal{V}_\sigma^* \cong \mathcal{U}_\sigma$ in the discrete setting. Here we see the effect of the variational discretization concept:
The choice for the discretization of the test space induces a natural discretization for the controls.
The following operators will be useful for the discussion of solutions to (P_σ):

Lemma 3.10. *Let the linear operators Υ_h and Π_h be defined as below:*

$$\Upsilon_h : \mathcal{M}(\bar{\Omega}_c) \to U_h \subset \mathcal{M}(\bar{\Omega}_c), \qquad \Upsilon_h u_0 := \sum_{j \in \mathcal{I}_h} \delta_{x_j} \int_{\bar{\Omega}_c} e_{x_j} du_0,$$

$$\Pi_h : C(\bar{\Omega}_c) \to V_h \subset C(\bar{\Omega}_c), \qquad \Pi_h f_0 := \sum_{j \in \mathcal{I}_h} f_0(x_j)\, e_{x_j}.$$

Then for every $u_0 \in \mathcal{M}(\bar{\Omega}_c)$, $f_0 \in C(\bar{\Omega}_c)$ and $v_h \in V_h$ the following properties hold.

$$\langle u_0, v_h\rangle = \langle \Upsilon_h u_0, v_h\rangle, \tag{3.34}$$

$$\langle u_0, \Pi_h f_0\rangle = \langle \Upsilon_h u_0, f_0\rangle, \tag{3.35}$$

$$\|\Upsilon_h u_0\|_{\mathcal{M}(\bar{\Omega}_c)} \le \|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}, \tag{3.36}$$

$$\Upsilon_h u_0 \overset{*}{\rightharpoonup} u_0 \in \mathcal{M}(\bar{\Omega}_c) \quad \text{and} \quad \|\Upsilon_h u_0\|_{\mathcal{M}(\bar{\Omega}_c)} \overset{h\to 0}{\longrightarrow} \|u_0\|_{\mathcal{M}(\bar{\Omega}_c)}. \tag{3.37}$$

These results follow directly from restricting [20, Proposition 4.1.] to $\bar{\Omega}_c \subset \Omega$. We give the proof for completeness:

Proof. We see (3.34) by the following calculations:

$$\begin{aligned}
\langle u_0, v_h\rangle &= \int_{\bar{\Omega}_c} \sum_{j \in \mathcal{I}_h} v_j\, e_{x_j}\, du_0 \\
&= \sum_{j \in \mathcal{I}_h} v_h(x_j) \int_{\bar{\Omega}_c} e_{x_j}\, du_0 \\
&= \sum_{j \in \mathcal{I}_h} \langle \delta_{x_j}, v_h\rangle \int_{\bar{\Omega}_c} e_{x_j}\, du_0 \\
&= \langle \Upsilon_h u_0, v_h\rangle.
\end{aligned}$$

And (3.35) by:

$$\begin{aligned}
\langle u_0, \Pi_h f_0\rangle &= \int_{\bar{\Omega}_c} \sum_{j \in \mathcal{I}_h} f_0(x_j)\, e_{x_j}\, du_0 \\
&= \sum_{j \in \mathcal{I}_h} \langle \delta_{x_j}, f_0\rangle \int_{\bar{\Omega}_c} e_{x_j}\, du_0 \\
&= \langle \Upsilon_h u_0, f_0\rangle.
\end{aligned}$$

Next, to show (3.36) we estimate:

$$\begin{aligned}\|\Upsilon_h u_0\|_{\mathcal{M}(\bar{Q}_c)} &= \Big\| \sum_{j\in\mathcal{I}_h} \delta_{x_j} \int_{\bar{Q}_c} e_{x_j}\, du_0 \Big\|_{\mathcal{M}(\bar{Q}_c)} \\ &\le \sum_{j\in\mathcal{I}_h} \underbrace{\|\delta_{x_j}\|_{\mathcal{M}(\bar{Q}_c)}}_{=1} \int_{\bar{Q}_c} e_{x_j}\, d|u_0| \\ &\le \int_{\bar{Q}_c} d|u_0| \;= \|u_0\|_{\mathcal{M}(\bar{Q}_c)}\,.\end{aligned}$$

Consequently, there exists a subsequence, denoted in the same way, such that

$$\Upsilon_h u_0 \overset{*}{\rightharpoonup} \bar{u}_0 \in \mathcal{M}(\bar{Q}_c) \text{ as } |h| \to 0.$$

For any $f_0 \in C(\bar{Q}_c)$, it holds that $\Pi_h f_0 \to f_0 \in C(\bar{Q}_c)$, as $|h| \to 0$. We derive

$$\langle \bar{u}_0, f_0\rangle = \lim_{|h|\to 0} \langle \Upsilon_h u_0, f_0\rangle \overset{(3.35)}{=} \lim_{|h|\to 0} \langle u_0, \Pi_h f_0\rangle = \langle u_0, f_0\rangle.$$

This shows $\bar{u}_0 = u_0$ and $\Upsilon_h u_0 \overset{*}{\rightharpoonup} u_0$ for the whole sequence and hence

$$\|u_0\|_{\mathcal{M}(\bar{Q}_c)} \le \liminf_{|h|\to 0} \|\Upsilon_h u_0\|_{\mathcal{M}(\bar{Q}_c)} \overset{(3.36)}{\le} \|u_0\|_{\mathcal{M}(\bar{Q}_c)}.$$

□

We also derive an analogous result for the space-time discrete spaces $\mathcal{U}_\sigma$ and $\mathcal{V}_\sigma$, similar to [20, Proposition 4.2.], but adjusted to our choice of spaces. The structure of the proof remains, only technical calculations are different.

Lemma 3.11. *Let the linear operators Υ_σ and Π_σ be defined as below:*

$$\begin{aligned}&\Upsilon_\sigma : \mathcal{M}(\bar{Q}_c) \to \mathcal{U}_\sigma \subset \mathcal{M}(\bar{Q}_c), \qquad \Upsilon_\sigma u \coloneqq \sum_{(j,k)\in\mathcal{I}_\sigma} \delta_{x_j}\otimes\delta_{t_k} \int_{\bar{Q}_c} e_{x_j}\otimes e_{t_k}\, du,\\ &\Pi_\sigma : C(\bar{Q}_c) \to \mathcal{V}_\sigma \subset C(\bar{Q}_c), \qquad \Pi_\sigma f \coloneqq \sum_{(j,k)\in\mathcal{I}_\sigma} f(x_j,t_k)\,(e_{x_j}\otimes e_{t_k}).\end{aligned}$$

Then for every $u \in \mathcal{M}(\bar{Q}_c)$, $f \in C(\bar{Q}_c)$ and $v_\sigma \in \mathcal{V}_\sigma$ the following properties hold.

$$\langle u, v_\sigma\rangle = \langle \Upsilon_\sigma u, v_\sigma\rangle, \tag{3.38}$$

$$\langle u, \Pi_\sigma f\rangle = \langle \Upsilon_\sigma u, f\rangle, \tag{3.39}$$

$$\|\Upsilon_\sigma u\|_{\mathcal{M}(\bar{Q}_c)} \le \|u\|_{\mathcal{M}(\bar{Q}_c)}, \tag{3.40}$$

$$\Upsilon_\sigma u \overset{*}{\rightharpoonup} u \;\in\; \mathcal{M}(\bar{Q}_c) \quad \textit{and} \quad \|\Upsilon_\sigma u\|_{\mathcal{M}(\bar{Q}_c)} \overset{|\sigma|\to 0}{\longrightarrow} \|u\|_{\mathcal{M}(\bar{Q}_c)}. \tag{3.41}$$

Proof. We verify (3.38) by the following calculation:

$$\begin{aligned}\langle u, v_\sigma\rangle &= \int_{\bar{Q}_c} \sum_{(j,k)\in\mathcal{I}_\sigma} v_{j,k}\,(e_{x_j}\otimes e_{t_k})\, du \\ &= \sum_{(j,k)\in\mathcal{I}_\sigma} v_\sigma(x_j,t_k) \int_{\bar{Q}_c} e_{x_j}\otimes e_{t_k}\, du \\ &= \sum_{(j,k)\in\mathcal{I}_\sigma} \langle \delta_{x_j}\otimes\delta_{t_k}, v_\sigma\rangle \int_{\bar{Q}_c} e_{x_j}\otimes e_{t_k}\, du \\ &= \langle \Upsilon_\sigma u, v_\sigma\rangle\,.\end{aligned}$$

Equation (3.39) can be seen by another short calculation:

$$\begin{aligned}\langle u, \Pi_\sigma f\rangle &= \int_{\bar{Q}_c} \sum_{(j,k)\in\mathcal{I}_\sigma} f(x_j,t_k)\,(e_{x_j}\otimes e_{t_k})\,du \\ &= \sum_{(j,k)\in\mathcal{I}_\sigma} \langle \delta_{x_j}\otimes\delta_{t_k}, f\rangle \int_{\bar{Q}_c} e_{x_j}\otimes e_{t_k}\,du \\ &= \langle \Upsilon_\sigma u, f\rangle\,.\end{aligned}$$

Inequality (3.40) is obtained as follows:

$$\begin{aligned}\|\Upsilon_\sigma u\|_{\mathcal{M}(\bar{Q}_c)} &= \Big\| \sum_{(j,k)\in\mathcal{I}_\sigma} \delta_{x_j}\otimes\delta_{t_k} \int_{\bar{Q}_c} e_{x_j}\otimes e_{t_k}\,du \Big\|_{\mathcal{M}(\bar{Q}_c)} \\ &\le \sum_{(j,k)\in\mathcal{I}_\sigma} \underbrace{\|\delta_{x_j}\otimes\delta_{t_k}\|_{\mathcal{M}(\bar{Q}_c)}}_{=1} \int_{\bar{Q}_c} e_{x_j}\otimes e_{t_k}\,d|u| \\ &\le \int_{\bar{Q}_c} d|u| \;= \|u\|_{\mathcal{M}(\bar{Q}_c)}\,.\end{aligned}$$

We deduce that there exists a subsequence, denoted in the same way, such that

$$\Upsilon_\sigma u \overset{*}{\rightharpoonup} \bar{u} \;\in \mathcal{M}(\bar{Q}_c) \text{ as } |\sigma|\to 0.$$

For any $f\in C(\bar{Q}_c)$, it holds that $\Pi_\sigma f\to f\in C(\bar{Q}_c)$ as $|\sigma|\to 0$. We derive

$$\langle \bar{u}, f\rangle = \lim_{|\sigma|\to 0}\langle \Upsilon_\sigma u, f\rangle \overset{(3.39)}{=} \lim_{|\sigma|\to 0}\langle u, \Pi_\sigma f\rangle = \langle u, f\rangle\,.$$

This shows $\bar{u}=u$ and $\Upsilon_\sigma u \overset{*}{\rightharpoonup} u$ for the whole sequence and hence

$$\|u\|_{\mathcal{M}(\bar{Q}_c)} \le \liminf_{|\sigma|\to 0}\|\Upsilon_\sigma u\|_{\mathcal{M}(\bar{Q}_c)} \overset{(3.40)}{\le} \|u\|_{\mathcal{M}(\bar{Q}_c)}\,.$$

□

Next, we observe that J_σ is convex, but not strictly convex. In the continuous setting, the strict convexity of J was caused by the norm $\|\cdot\|_{L^q(Q)}$ for $q>1$ and the injectivity of $L^{-*}\Phi^*$. Here we have the discrete operator $L_\sigma^*:\mathcal{Y}_\sigma\to\mathcal{W}_\sigma^*$, defined as

$$\langle L_\sigma^* y_\sigma, w_\sigma\rangle := \int_Q (-y_\sigma\,\partial_t w_\sigma + \nabla y_\sigma \nabla w_\sigma)\,dx\,dt,$$

with $L_\sigma:\mathcal{W}_\sigma\to\mathcal{Y}_\sigma^*$. We can rewrite the discrete state equation (3.26) as

$$y_\sigma(u_0,u) = L_\sigma^{-*}(\Phi_h^*\Upsilon_h u_0 + \Phi_\sigma^*\Upsilon_\sigma u). \tag{3.42}$$

In general, the mapping $\mathcal{M}(\bar{\Omega}_c)\times\mathcal{M}(\bar{Q}_c)\ni(u_0,u)\mapsto y_\sigma(u_0,u)$ is not injective, hence the uniqueness of the solution to (P_σ) cannot be concluded. In the implicitly discrete setting however, we can prove uniqueness similarly as done in [14, Section 4.3.].

Theorem 3.12. *The problem (P_σ) has at least one solution in $\mathcal{M}(\bar{\Omega}_c)\times\mathcal{M}(\bar{Q}_c)$ and there exists a unique solution $(\bar{u}_{0,h},\bar{u}_\sigma)\in U_h\times\mathcal{U}_\sigma$. Furthermore we know for every solution $(\hat{u}_0,\hat{u})\in\mathcal{M}(\bar{\Omega}_c)\times\mathcal{M}(\bar{Q}_c)$ of (P_σ) that*

$$(\Upsilon_h\hat{u}_0, \Upsilon_\sigma\hat{u}) = (\bar{u}_{0,h},\bar{u}_\sigma). \tag{3.43}$$

Proof. The existence of solutions can be derived as in the proof of Theorem 3.6 because the control domain is still continuous. Let $(\hat{u}_0, \hat{u}) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c)$ be a solution of problem (P_σ) and let

$$(\bar{u}_{0,h}, \bar{u}_\sigma) := (\Upsilon_h \hat{u}_0, \Upsilon_\sigma \hat{u}) \in U_h \times \mathcal{U}_\sigma.$$

We deduce from (3.34) and (3.38) that

$$y_\sigma(u_0, u) = y_\sigma(\Upsilon_h u_0, \Upsilon_\sigma u) \quad \text{for all } (u_0, u) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c). \tag{3.44}$$

Additionally, (3.36) and (3.40) deliver

$$\begin{aligned} \|\bar{u}_{0,h}\|_{\mathcal{M}(\bar{\Omega}_c)} &\le \|\hat{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)}, \\ \|\bar{u}_\sigma\|_{\mathcal{M}(\bar{Q}_c)} &\le \|\hat{u}\|_{\mathcal{M}(\bar{Q}_c)}. \end{aligned}$$

Combining these properties, we deduce $J_\sigma(\bar{u}_{0,h}, \bar{u}_\sigma) \le J_\sigma(\hat{u}_0, \hat{u})$. This validates the existence of solutions in the discrete space $U_h \times \mathcal{U}_{\mathrm{vd}}$.

The operators Υ_h and Υ_σ act as identities on U_h and $\mathcal{U}_\sigma$. Furthermore, the operator

$$(\Phi_h, \Phi_\sigma)^* : U_h \times \mathcal{U}_\sigma \to \mathcal{W}_\sigma^*$$

is injective and we also know that $\dim(\mathcal{Y}_\sigma) = \dim(\mathcal{W}_\sigma^*)$. Hence we deduce the injectivity of $(u_0, u) \mapsto y_\sigma(u_0, u)$ for discrete controls $(u_0, u) \in U_h \times \mathcal{U}_\sigma$. Now strict convexity of J_σ on $U_h \times \mathcal{U}_{\mathrm{vd}}$ follows from $q > 1$. Consequently, problem (P_σ) has a unique discrete solution $(\bar{u}_{0,h}, \bar{u}_\sigma) \in U_h \times \mathcal{U}_\sigma$.

For every solution $(\hat{u}_0, \hat{u})$ of (P_σ), the projection $(\Upsilon_h \hat{u}_0, \Upsilon_\sigma \hat{u})$ is a discrete solution. Moreover, there exists only one discrete solution. So we deduce that all projections must coincide, showing (3.43). □

Since all projections of solutions of (P_σ) yield the unique discrete solution $(\bar{u}_{0,h}, \bar{u}_\sigma)$, it suffices to analyze the convergence properties of $(\bar{u}_{0,h}, \bar{u}_\sigma)$ for $|\sigma| \to 0$. Furthermore we may find solutions of (P_σ) numerically, by restricting the control space to $U_h \times \mathcal{U}_\sigma$.

We now prove the convergence result formulated in Theorem 3.2 along the lines of the proof of [20, Theorem 4.3.].

Proof. Observe that

$$J_\sigma(\bar{u}_{0,h}, \bar{u}_\sigma) \le J_\sigma(0, 0) = \frac{1}{q} \|y_{\mathrm{d}}\|^q_{L^q(Q_h)}.$$

This implies that the norms $\|\bar{y}_\sigma\|_{L^q(Q_h)}$, $\|\bar{u}_{0,h}\|_{\mathcal{M}(\bar{\Omega}_c)}$, and $\|\bar{u}_\sigma\|_{\mathcal{M}(\bar{Q}_c)}$ are uniformly bounded for all σ. Now, let $\{\sigma_n\}_n$ be a sequence with $|\sigma_n| \to 0$. Boundedness in norm implies that there exists a subsequence $\{\sigma_{n_k}\}_k$, such that the following holds true for $k \to \infty$:

$$(\bar{u}_{0,h_{n_k}}, \bar{u}_{\sigma_{n_k}}) \overset{*}{\rightharpoonup} (\tilde{u}_0, \tilde{u}) \text{ in } \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c) \quad \text{and} \quad \bar{y}_{\sigma_{n_k}} \rightharpoonup \tilde{y} \text{ in } L^q(Q). \tag{3.45}$$

We will split the proof into three steps.

Step I - $\tilde{y}$ is the solution of (3.1) *corresponding to* $(\tilde{u}_0, \tilde{u})$*, i.e.,* $L^*\tilde{y} = \Phi^*(\tilde{u}_0, \tilde{u})$.
By constructing a suitable Friedrichs smoothing operator and by using standard results in approximation theory (e.g., cubic spline interpolation, see [2, Theorem 1]), one realizes that

$$\{\psi \in C^2(\bar{I}; \mathbb{R}) : \psi(T) = 0\} \otimes (C^2(\Omega) \cap W_0^{1,p}(\Omega))$$

is dense in W (for more details see Appendix A.1). Consequently, it is sufficient to test the smooth state equation (3.2) against $w = \varphi \otimes \psi$ with $\psi \in C^2(\bar{I}; \mathbb{R})$ satisfying $\psi(T) = 0$ and $\varphi \in C^2(\Omega) \cap W_0^{1,p}(\Omega)$.

Let φ be approximated by $\varphi_h \in Y_h$, such that

$$\int_\Omega \nabla\varphi_h \nabla z_h \, dx = \int_\Omega \nabla\varphi \nabla z_h \, dx \quad \forall w_h \in Y_h \quad \text{and} \quad \|\varphi - \varphi_h\|_{C(\bar\Omega)} \overset{h\to 0}{\longrightarrow} 0. \tag{3.46}$$

Indeed, one has the error estimate

$$\|\varphi - \varphi_h\|_{L^\infty(\Omega)} \le C\, h^2 \, \log\left(\frac{1}{h}\right) \|\varphi\|_{W^{2,\infty}(\Omega)}$$

for the Ritz projection. For details see Corollary 2 and Remark 4 in [60].
Moreover, let $\psi_\tau = \sum_k \psi(t_k)\, e_{t_k}$ be the continuous, piecewise linear interpolation of ψ on the time grid and put $w_\sigma = \varphi_h \otimes \psi_\tau \in \mathcal{W}_\sigma$. By construction, we have $w_\sigma \to w \in C(\bar Q)$ for $|\sigma| \to 0$. Furthermore, $\partial_t w_\sigma \to \partial_t w \in L^p(Q)$, where $\frac{1}{q} + \frac{1}{p} = 1$, for $|\sigma| \to 0$ since

$$\begin{aligned}
\|\partial_t w - \partial_t w_\sigma\|_{L^p(Q)} &\le \|(\varphi - \varphi_h)\otimes\psi'\|_{L^p(Q)} + \|\varphi_h \otimes (\psi' - \psi'_\tau)\|_{L^p(Q)} \\
&= \|\varphi - \varphi_h\|_{L^p(\Omega)}\, \|\psi'\|_{L^p(I)} + \|\varphi_h\|_{L^p(\Omega)}\, \|\psi' - \psi'_\tau\|_{L^p(I)} \\
&\le |Q|^{1/p} \left(\|\varphi - \varphi_h\|_{C(\bar\Omega)}\, \|\psi'\|_{L^p(I)} + \|\varphi_h\|_{C(\bar\Omega)}\, \|\psi' - \psi'_\tau\|_{L^p(I)}\right).
\end{aligned}$$

We have $\|\varphi - \varphi_h\|_{C(\bar\Omega)} \overset{h\to 0}{\longrightarrow} 0$ from (3.46) and $\|\psi' - \psi'_\tau\|_{L^p(I)} \le C\, h \|\psi''\|_{C(\bar I)} \overset{\tau\to 0}{\longrightarrow} 0$ can be confirmed by splitting $\bar I$ into its subintervals I_k, integrating and using $\psi(t_k) = \psi_\tau(t_k)$ for all $k \in \{1, \dots, N_\tau\}$. Testing (3.26) against this w_σ, we obtain

$$\langle L^*_\sigma \bar y_\sigma, w_\sigma\rangle = \int_{\bar\Omega_c} w_\sigma(0)\, d\bar u_{0,h} + \int_{\bar Q_c} w_\sigma \, d\bar u_\sigma. \tag{3.47}$$

On the right hand side, we can perform the limit directly:

$$\int_{\bar\Omega_c} w_\sigma(0)\, d\bar u_{0,h} + \int_{\bar Q_c} w_\sigma \, d\bar u_\sigma \overset{|\sigma|\to 0}{\longrightarrow} \int_{\bar\Omega_c} w(0)\, d\tilde u_0 + \int_{\bar Q_c} w \, d\tilde u.$$

The left hand side of (3.47) can be expanded to

$$\langle L^*_\sigma \bar y_\sigma, w_\sigma\rangle = -\int_Q \bar y_\sigma (\varphi_h \otimes \psi'_\tau)\, dx\, dt + \int_Q \nabla \bar y_\sigma \nabla(\varphi_h \otimes \psi_\tau)\, dx\, dt. \tag{3.48}$$

Applying the very definition of φ_h and integration by parts, we observe that

$$\int_Q \nabla \bar y_\sigma\, \nabla(\varphi_h \otimes \psi_\tau)\, dx\, dt = -\int_Q \bar y_\sigma\, (\Delta\varphi \otimes \psi_\tau)\, dx\, dt \overset{|\sigma|\to 0}{\longrightarrow} -\int_Q \tilde y\, \Delta w\, dx\, dt.$$

Along with $(\varphi_h \otimes \psi'_\tau) = \partial_t w_\sigma$ and $-\int_Q \bar y_\sigma\, \partial_t w_\sigma\, dx\, dt \to -\int_Q \tilde y\, \partial_t w\, dx\, dt$, this implies that

$$\langle L^*_\sigma \bar y_\sigma, w_\sigma\rangle \to \langle L^* \tilde y, w\rangle$$

for all tensor product functions $w = \varphi \otimes \psi$. Thus, we deduce $L^*\tilde y = \Phi^*(\tilde u_0, \tilde u)$ from (3.47).

Step II - $(\tilde u_0, \tilde u)$ coincides with the unique solution $(\bar u_0, \bar u)$ of (P_σ) that lies in $U_h \times \mathcal{U}_\sigma$
In order to prove this, it suffices to show

$$J_\sigma(\tilde u_0, \tilde u) \le J_\sigma(\bar u_0, \bar u).$$

Recall that we identified $\mathcal{M}(\bar\Omega_c)$ and $\mathcal{M}(\bar Q_c)$ with $\{u_0 \in \mathcal{M}(\Omega)\colon\ \mathrm{supp}(u_0) \subset \bar\Omega_c\}$ and $\{u \in \mathcal{M}(Q)\colon\ \mathrm{supp}(u) \subset \bar Q_c\}$, respectively. In this sense, the sets

$$\{f_0 \in C^\infty(\Omega)\colon\ \mathrm{supp}(f_0) \subset \bar\Omega_c\} \quad \text{and} \quad \{f \in C^\infty(Q)\colon\ \mathrm{supp}(f) \subset \bar Q_c\}$$

are dense in $\mathcal{M}(\bar{\Omega}_c)$ and $\mathcal{M}(\bar{Q}_c)$ with respect to the sequential weak-* topology. This can be seen by utilizing that $\bar{\Omega}_c$ and $\bar{Q}_c$ have to satisfy certain uniform cone conditions (because they are Lipschitz domains, see [1, Paragraph 4.8]) and by convolution against suitable Friedrichs mollifiers that are compactly supported in the interior of finite, convex cones (for further details see Appendix A.2). Notice also that such convolutions do not increase the $\mathcal{M}$-norms.
Consequently, we may pick a specific minimizing sequence

$$(u_{0,m}, u_m) = (f_{0,m}\, dx, f_m\, dx \otimes dt),$$

where $f_{0,m} \in C^\infty(\Omega)$ and $f_m \in C^\infty(Q)$ satisfy $\operatorname{supp}(f_{0,m}) \subset \Omega_c$ and $\operatorname{supp}(f_m) \subset Q_c$. Then the states

$$y_m := L^{-*}\Phi^*(u_{0,m}, u_m)$$

are solutions of the heat equations

$$\begin{cases} \partial_t y_m - \Delta y_m = f_m, & \text{in } Q, \\ y_m(x,0) = f_{0,m}, & \text{in } \Omega, \\ y_m(x,t) = 0, & \text{on } \Sigma. \end{cases}$$

It follows now from maximal regularity (recall that Ω is now assumed to be of class $C^{1,1}$), that $y_m \in W_r^{2,1}(Q)$ for all $2 \le r < \infty$. Thus the finite element discretizations of the states converge to y_m in $L^2(Q)$ and thus also in $L^q(Q)$. More precisely, we have for each fixed m that

$$\lim_{|\sigma|\to 0} \|L_\sigma^{-*}(\Phi_h^* \oplus \Phi_\sigma^*)(u_{0,m}, u_m) - y_m\|_{L^q(Q)} = 0.$$

We choose a suitable subsequence of $\{\sigma_{n_k}\}_{k\in\mathbb{N}}$ as follows: We put $k_1 := 1$ and pick $k_m \ge k_{m-1}$ recursively such that

$$\|L_{\sigma_{n_{k_m}}}^{-*}(\Phi_{h_{n_{k_m}}}^* \oplus \Phi_{\sigma_{n_{k_m}}}^*)(u_{0,m}, u_m) - y_m\|_{L^q(Q)}^q \le \frac{1}{m} \quad \text{for each } m \ge 2.$$

Now using the projection properties (3.36),(3.40) and (3.44) in combination with the above, we obtain

$$\begin{aligned} J_\sigma(\bar{u}_{0,h_{n_{k_m}}}, \bar{u}_{\sigma_{n_{k_m}}}) &\le J_\sigma(\Upsilon_{h_{n_{k_m}}} u_{0,m}, \Upsilon_{\sigma_{n_{k_m}}} u_m) \\ &\le \frac{1}{q}\|L_{\sigma_{n_{k_m}}}^{-*}(\Phi_{h_{n_{k_m}}}^* \oplus \Phi_{\sigma_{n_{k_m}}}^*)(u_{0,m}, u_m) - y_m + y_m - y_\mathrm{d}\|_{L^q(Q)}^q + \alpha\|u_m\|_{\mathcal{M}(\bar{Q}_c)} + \beta\|u_{0,m}\|_{\mathcal{M}(\bar{\Omega}_c)} \\ &\le \frac{1}{q}\Big(\frac{1}{m} + \|y_m - y_\mathrm{d}\|_{L^q(Q)}^q\Big) + \alpha\|u_m\|_{\mathcal{M}(\bar{Q}_c)} + \beta\|u_{0,m}\|_{\mathcal{M}(\bar{\Omega}_c)}. \end{aligned}$$

Next, we use the weakly lower semi continuity of J in combination with (3.45), then apply $\liminf_{m\to\infty}$ to both sides of the above inequality, and finally use the facts that $(u_{0,m}, u_m)$ is a minimizing sequence and that $(\bar{u}_0, \bar{u})$ solves problem (P):

$$\begin{aligned} J(\tilde{u}_0, \tilde{u}) &\le \liminf_{m\to\infty} J_\sigma(\bar{u}_{0,h_{n_{k_m}}}, \bar{u}_{\sigma_{n_{k_m}}}) \\ &\le \liminf_{m\to\infty} \frac{1}{q}\left(\left(\frac{1}{m} + \|y_m - y_\mathrm{d}\|_{L^q(Q)}^q\right) + \alpha\|u_m\|_{\mathcal{M}(\bar{Q}_c)} + \beta\|u_{0,m}\|_{\mathcal{M}(\bar{\Omega}_c)}\right) \\ &\le \limsup_{m\to\infty} J_\sigma(u_{0,m}, u_m) \\ &\le \frac{1}{q}\|\bar{y} - y_\mathrm{d}\|_{L^q(Q)}^q + \alpha\|\bar{u}\|_{\mathcal{M}(\bar{Q}_c)} + \beta\|\bar{u}_0\|_{\mathcal{M}(\bar{\Omega}_c)} \\ &= J(\bar{u}_0, \bar{u}). \end{aligned}$$

Step III - proof of (3.3)*,* (3.4) *and* (3.5)
Altogether, we know that every sequence $\{\sigma_n\}_n$ with $|\sigma_n| \to 0$ has a subsequence $\{\sigma_{n_{k_m}}\}_m$, such that for $m \to \infty$

$$(\bar{u}_{0,h_{n_{k_m}}}, \bar{u}_{\sigma_{n_{k_m}}}) \overset{*}{\rightharpoonup} (\bar{u}_0, \bar{u}) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c) \quad \text{and} \quad \bar{y}_{\sigma_{n_{k_m}}} \rightharpoonup \bar{y} \in L^q(Q).$$

Since the limits are always the same and $(\bar{u}_0, \bar{u})$ and $\bar{y}$ are unique, this implies already that

$$(\bar{u}_{0,h}, \bar{u}_\sigma) \overset{*}{\rightharpoonup} (\bar{u}_0, \bar{u}) \in \mathcal{M}(\bar{\Omega}_c) \times \mathcal{M}(\bar{Q}_c) \quad \text{and} \quad \bar{y}_\sigma \rightharpoonup \bar{y} \in L^q(Q) \quad \text{for } |\sigma| \to 0.$$

This shows (3.4). Next, we can calculate

$$\begin{aligned}
\frac{1}{q}\|\bar{y} - y_\mathrm{d}\|^q_{L^q(Q)} &\leq \liminf_{|\sigma|\to 0} \frac{1}{q}\|\bar{y}_\sigma - y_\mathrm{d}\|^q_{L^q(Q)} \leq \limsup_{|\sigma|\to 0} \frac{1}{q}\|\bar{y}_\sigma - y_\mathrm{d}\|^q_{L^q(Q)} \\
&\leq \limsup_{|\sigma|\to 0} \left(J_\sigma(\bar{u}_{0,h}, \bar{u}_\sigma) - \alpha\,\|\bar{u}_\sigma\|_{\mathcal{M}(Q)} - \beta\,\|\bar{u}_{0,h}\|_{\mathcal{M}(\Omega)}\right) \\
&\leq \limsup_{|\sigma|\to 0} J_\sigma(\bar{u}_{0,h}, \bar{u}_\sigma) - \liminf_{|\sigma|\to 0} \left(\alpha\,\|\bar{u}_\sigma\|_{\mathcal{M}(Q)} + \beta\,\|\bar{u}_{0,h}\|_{\mathcal{M}(\Omega)}\right) \\
&= J(\bar{u}_0, \bar{u}) - \liminf_{|\sigma|\to 0} \left(\alpha\|\bar{u}_\sigma\|_{\mathcal{M}(Q)} + \beta\|\bar{u}_{0,h}\|_{\mathcal{M}(\Omega)}\right) \\
&\overset{(3.45)}{\leq} J(\bar{u}_0, \bar{u}) - \left(\alpha\|\bar{u}\|_{\mathcal{M}(Q)} + \beta\|\bar{u}\|_{\mathcal{M}(\Omega)}\right) \\
&\leq \frac{1}{q}\|\bar{y} - y_\mathrm{d}\|^q_{L^q(Q)}.
\end{aligned}$$

Because of $1 < q < \infty$, the space $L^q(Q)$ is an uniformly convex Banach space; thus weak convergence together with the convergence of the norms implies strong convergence (see [10, Proposition 3.32]). This shows (3.3). In a similar way we can prove the first part of (3.5)

$$\begin{aligned}
\alpha\,\|\bar{u}\|_{\mathcal{M}(Q)} &\overset{(3.45)}{\leq} \liminf_{|\sigma|\to 0} \alpha\,\|\bar{u}_\sigma\|_{\mathcal{M}(Q)} \leq \limsup_{|\sigma|\to 0} \alpha\,\|\bar{u}_\sigma\|_{\mathcal{M}(Q)} \\
&\leq \limsup_{|\sigma|\to 0} \left(J_\sigma(\bar{u}_{0,h}, \bar{u}_\sigma) - \frac{1}{q}\|\bar{y}_\sigma - y_\mathrm{d}\|^q_{L^q(Q)} - \beta\,\|\bar{u}_{0,h}\|_{\mathcal{M}(\Omega)}\right) \\
&\overset{(3.3)}{\leq} J(\bar{u}_0, \bar{u}) - \frac{1}{q}\|\bar{y} - y_\mathrm{d}\|^q_{L^q(Q)} - \liminf_{|\sigma|\to 0} \beta\,\|\bar{u}_{0,h}\|_{\mathcal{M}(\Omega)} \\
&\overset{(3.45)}{\leq} J(\bar{u}_0, \bar{u}) - \frac{1}{q}\|\bar{y} - y_\mathrm{d}\|^q_{L^q(Q)} - \beta\|\bar{u}_0\|_{\mathcal{M}(\Omega)} \\
&= \alpha\,\|\bar{u}\|_{\mathcal{M}(Q)}.
\end{aligned}$$

Finally, the second part of (3.5) follows directly from $\lim_{|\sigma|\to 0} J_\sigma(\bar{u}_{0,h}, \bar{u}_\sigma) = J(\bar{u}_0, \bar{u})$ and the fact that we already showed the convergence of the other two terms. □

Now we discretize (P^*) with $w_\sigma \in \mathcal{W}_\sigma$ and equivalently reformulate the problem in the following way:

$$\min_{w_\sigma \in \mathcal{W}_\sigma} K_\sigma(w_\sigma) \coloneqq \frac{1}{p}\|L_\sigma w_\sigma\|^p_{L^p(Q_h)} + \langle L_\sigma w_\sigma, y_\mathrm{d}\rangle_{L^p(Q_h), L^q(Q_h)} \tag{P^*_σ}$$

$$\text{s.t. } \|\Phi_h(w_\sigma)\|_{\infty,\bar{Q}_c} \leq \beta \text{ and } \|\Phi_\sigma(w_\sigma)\|_{\infty,\bar{Q}_c} \leq \alpha \tag{3.49}$$

Similar to the continuous setting, it can be shown that (P^*_σ) is the Fenchel predual of the problem (P_σ) restricted to $(u_0, u) \in U_h \times \mathcal{U}_\sigma$. In order to solve (P^*_σ), we want to represent $L_\sigma : \mathcal{W}_\sigma \to \mathcal{Y}^*_\sigma$ by a matrix, as done in [14]. From [31, Section 4] and [39] we know that the matrix representation of $L^*_\sigma : \mathcal{Y}_\sigma \to \mathcal{W}^*_\sigma$ yields a Crank-Nicolson scheme with a smoothing step. We will derive this first.

Let $M_h := (\langle e_{x_j}, e_{x_k}\rangle)_{j,k=1}^{N_h}$ be the mass matrix and $A_h := (\int_\Omega \nabla e_{x_j} \nabla e_{x_k}\, dx)_{j,k=1}^{N_h}$ the stiffness matrix corresponding to Y_h. We define

$$\begin{aligned} y_{k,h} &:= y_\sigma|_{I_k} \in Y_h && \text{for } k \in \{1,\dots,N_\tau\},\\ w_{k,h} &:= w_\sigma(\cdot, t_k) \in Y_h && \text{for } k \in \{0,\dots,N_\tau - 1\},\\ w_k &:= w_{k,h} \otimes e_{t_k} \in \mathcal{W}_\sigma && \text{for } k \in \{0,\dots,N_\tau - 1\}. \end{aligned}$$

We then obtain the following :

$$\begin{aligned} \langle L_\sigma^* y_\sigma, w_k\rangle &= \int_Q -y_\sigma\, \partial_t w_k + \nabla y_\sigma \nabla w_k \, dx\, dt\\ &= \int_{I_k}\int_\Omega -y_{k,h} w_{k,h} \underbrace{\partial_t e_{t_k}}_{=\frac{1}{\tau_k}} + (\nabla y_{k,h}\nabla w_{k,h})\, e_{t_k}\, dx\, dt + \int_{I_{k+1}}\int_\Omega -y_{k+1,h} w_{k,h} \underbrace{\partial_t e_{t_k}}_{=-\frac{1}{\tau_{k+1}}} + (\nabla y_{k+1,h}\nabla w_{k,h})\, e_{t_k}\, dx\, dt\\ &= \int_\Omega -y_{k,h} w_{k,h}\, dx + \frac{\tau_k}{2}\int_\Omega \nabla y_{k,h}\nabla w_{k,h}\, dx + \int_\Omega y_{k+1,h} w_{k,h}\, dx + \frac{\tau_{k+1}}{2}\int_\Omega \nabla y_{k+1,h}\nabla w_{k,h}\, dx\\ &= (y_{k+1,h} - y_{k,h})^\top M_h\, w_{k,h} + (\frac{\tau_k}{2}\, y_{k,h} + \frac{\tau_{k+1}}{2}\, y_{k+1,h})^\top A_h\, w_{k,h}, \end{aligned}$$

for all $k \in \{1,\dots,N_\tau - 1\}$. So, for $k = 0$, we have

$$\begin{aligned} \langle L_\sigma^* y_\sigma, w_0\rangle &= \int_Q -y_\sigma\, \partial_t w_0 + \nabla y_\sigma \nabla w_0 \, dx\, dt\\ &= \int_{I_1}\int_\Omega -y_{1,h} w_{0,h} \underbrace{\partial_t e_{t_0}}_{=-\frac{1}{\tau_1}} + (\nabla y_{1,h}\nabla w_{0,h})\, e_{t_0}\, dx\, dt\\ &= \int_\Omega y_{1,h} w_{0,h}\, dx + \frac{\tau_1}{2}\int_\Omega \nabla y_{1,h}\nabla w_{0,h}\, dx\\ &= y_{1,h}^\top M_h\, w_{0,h} + \frac{\tau_1}{2} y_{1,h}^\top A_h\, w_{0,h}. \end{aligned}$$

In order to represent the discrete state equation (3.26) by a system of equations, we also need to calculate

$$r(w_\sigma) := \int_{\bar\Omega_c} w_\sigma(0)\, du_0 + \int_{\bar Q_c} w_\sigma\, du.$$

Due to the implicit discrete structure of the controls (u_0, u), we can define $\tilde u \in V_h^* \times \mathcal{V}_\sigma^*$ with $[\tilde u]_{j,k} = u_{j,k}$ and $[u_{0,h}]_j = u_{j,0}$ for $j \in \mathcal{I}_h$ and $[u_\sigma]_{j,k} = u_{j,k}$ for $(j,k) \in \mathcal{I}_\sigma$. Then we have

$$(\Phi_h + \Phi_\sigma)^*(u_{0,h}, u_\sigma) = \sum_{j=1}^{N_h}\sum_{k=0}^{N_\tau - 1} u_{j,k}\delta_{x_j} \otimes \delta_{t_k} \in \mathcal{W}_\sigma^*,$$

with $u_{j,k} = 0$ for $(j,0)$, $j \notin \mathcal{I}_h$ and $(j,k) \notin \mathcal{I}_\sigma$. Since $w_\sigma(0)|_{\bar\Omega_c} \in V_h = U_h^*$ and $w_\sigma|_{\bar Q_c} \in \mathcal{V}_\sigma = \mathcal{U}_\sigma^*$, we get

$$\begin{aligned} r(w_\sigma) &= \sum_{j\in\mathcal{I}_h} u_{j,0} w_{j,0} + \sum_{(j,k)\in\mathcal{I}_\sigma} u_{j,k} w_{j,k}\\ &= \langle (\Phi_h + \Phi_\sigma)^*(u_{0,h}, u_\sigma), w_\sigma\rangle_{\mathcal{W}_\sigma^*, \mathcal{W}_\sigma}. \end{aligned}$$

For the remainder of this section, we identify elements from $\mathcal{Y}_\sigma$ and $\mathcal{W}_\sigma$ with vectors in $\mathbb{R}^{N_\sigma}$, $N_\sigma := N_h \cdot N_\tau$ and elements from U_h and $\mathcal{U}_\sigma$ with vectors in $\mathbb{R}^{|\mathcal{I}_h|}$ and $\mathbb{R}^{|\mathcal{I}_\sigma|}$, respectively. The discrete elements can be expressed via their respective expansion coefficients. To simplify the notation, we define $y_k := (y_{1,k}, \dots, y_{N_h,k})^\top \in \mathbb{R}^{N_h}$ and write $y_\sigma = (y_1^\top, \dots, y_{N_\tau}^\top)^\top \in \mathbb{R}^{N_\sigma}$. Analogously, we define w_k for $k = 0, \dots, N_\tau - 1$.

We represent the discrete state equation by the following $(N_\sigma \times N_\sigma)$-matrix:

$$\mathcal{L}^\top \coloneqq \begin{pmatrix} (M_h + \frac{\tau_1}{2}A_h) & 0 & \dots & \dots & 0 \\ (-M_h + \frac{\tau_1}{2}A_h) & (M_h + \frac{\tau_2}{2}A_h) & & & \vdots \\ 0 & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & (-M_h + \frac{\tau_{N_\tau-1}}{2}A_h) & (M_h + \frac{\tau_{N_\tau}}{2}A_h) \end{pmatrix} \tag{3.50}$$

We point out that $\mathcal{L}^\top$ is in fact the operator L_σ^* concatenated with the space-time mass matrix $\mathcal{M}_\sigma$ that maps from $L^p(Q_h)$ to $(L^q(Q_h))^*$, which is an important detail for the implementation. A representation of L_σ, also concatenated with a space-time mass matrix, is the matrix $\mathcal{L} = (\mathcal{L}^\top)^\top$. If we now actually want the representative of $L_\sigma w_\sigma$, it is necessary to multiply with the inverse of the space-time mass matrix $\mathcal{M}_\sigma^{-1}$.
Furthermore the embedding $(\Phi_h \oplus \Phi_\sigma) : \mathcal{W}_\sigma \to (V_h \times \mathcal{V}_\sigma)$, defined in (3.27) and (3.28), can be represented by a restriction matrix:

$$(\mathcal{R}_h + \mathcal{R}_\sigma) : \mathbb{R}^{N_\sigma} \to \mathbb{R}^{|\mathcal{I}_h|+|\mathcal{I}_\sigma|}, \qquad w_\sigma \mapsto ((w_{0,j})_{j\in\mathcal{I}_h}^\top, (w_{j,k})_{(j,k)\in\mathcal{I}_\sigma}^\top)^\top.$$

From duality we conclude that $(\Phi_h \oplus \Phi_\sigma)^*$ can be represented by $(\mathcal{R}_h + \mathcal{R}_\sigma)^\top$, such that $\mathcal{L}^\top y_\sigma = (\mathcal{R}_h + \mathcal{R}_\sigma)^\top (u_{0,h}^\top, u_\sigma^\top)^\top$ is the matrix vector formulation of (3.26).
We equivalently reformulate the constraints (3.49) in P_σ^* using (3.29) and (3.30):

$$\begin{aligned} & \max_{j\in\mathcal{I}_h} |w_{j,0}| \le \beta \quad \text{and} \quad \max_{(j,k)\in\mathcal{I}_\sigma} |w_{j,k}| \le \alpha, \\ \Leftrightarrow \quad & \max_{j\in\mathcal{I}_h} \{w_{j,0}, -w_{j,0}\} - \beta \le 0 \quad \text{and} \quad \max_{(j,k)\in\mathcal{I}_\sigma} \{w_{j,k}, -w_{j,k}\} - \alpha \le 0. \end{aligned}$$

We now formulate linear inequality constraints that are equivalent to (3.49). All inequalities are strictly fulfilled for $w_\sigma = 0$, thus $w_\sigma = 0$ is an interior point of the feasible set and thus the Slater condition is satisfied (see Definition 2.29). We discretize the desired state by sampling it on the dual time grid:

$$y_{\mathrm{d},\sigma} = \sum_{j=1}^{N_h} \sum_{k=1}^{N_\tau} y_\mathrm{d}(x_j, (t_{k-1}+t_k)/2)\, e_{x_j} \otimes \chi_k \in \mathcal{Y}_\sigma. \tag{3.51}$$

By doing so, we assume that y_d has a certain minimum continuity. However, discretization by local averaging is also possible. If we make sure that $y_{\mathrm{d},\sigma} \to y_\mathrm{d}$ for $|\sigma| \to 0$, then using $y_{\mathrm{d},\sigma}$ in (P_σ) instead of y_d does not interfere with the convergence result Theorem 3.2. We proceed by setting up the corresponding Lagrangian $\mathscr{L}$ with multipliers $\lambda^{(1)}, \lambda^{(2)} \in \mathbb{R}^{|\mathcal{I}_\sigma|}$ and $\lambda^{(3)}, \lambda^{(4)} \in \mathbb{R}^{|\mathcal{I}_h|}$:

$$\begin{aligned} \mathscr{L}(w_\sigma, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}) = \frac{1}{p}\|\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma\|_{L^p(Q_h)}^p + \langle \mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma, y_{\mathrm{d},\sigma}\rangle_{L^p(Q_h),L^q(Q_h)} \\ + \sum_{(j,k)\in\mathcal{I}_\sigma} \lambda_{j,k}^{(1)}(w_{j,k} - \alpha) + \sum_{(j,k)\in\mathcal{I}_\sigma} \lambda_{j,k}^{(2)}(-w_{j,k} - \alpha) \\ + \sum_{j\in\mathcal{I}_h} \lambda_j^{(3)}(w_{j,0} - \beta) + \sum_{j\in\mathcal{I}_h} \lambda_j^{(4)}(-w_{j,0} - \beta). \end{aligned}$$

For the sake of simplified numerics, we use a lumped mass matrix approach for computing $K_\sigma(w_\sigma)$, where w_σ is a vector, i.e., with a slight abuse of notation, we employ

$$\begin{aligned} \|\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma\|_{L^p(Q_h)}^p &\coloneqq \sum_{j=1}^{N_h}\sum_{k=1}^{N_\tau} |(\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma)_{j,k}|^p\, \omega_j\, \tau_k, \\ \langle \mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma, y_{\mathrm{d},\sigma}\rangle_{L^p(Q_h),L^q(Q_h)} &\coloneqq \sum_{j=1}^{N_h}\sum_{k=1}^{N_\tau} (\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma)_{j,k} y_\mathrm{d}(x_j,t_k)\, \omega_j\, \tau_k, \end{aligned}$$

where $\omega_j := \int_\Omega e_{x_j} dx$. We also use a lumped mass matrix approach for $\mathcal{M}_\sigma^{-1}$.

We can now form the optimality system using the Karush-Kuhn-Tucker conditions (see Theorem 2.30). Since the Slater condition is satisfied, the Karush-Kuhn-Tucker conditions state that at the minimum w_σ, there must be $\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}$ the partial differential $\frac{\partial \mathscr{L}}{\partial w_\sigma}$ of $\mathscr{L}$ at the point $(w_\sigma, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)})$ has to vanish and that the following complementary conditions have to be fulfilled:

$$\lambda_{j,k}^{(i)}\Big(\frac{w_{j,k}}{(-1)^{(i-1)}} - \alpha\Big) = 0, \;\; \lambda_{j,k}^{(i)} \geq 0 \text{ and } \Big(\frac{w_{j,k}}{(-1)^{(i-1)}} - \alpha\Big) \leq 0 \quad \forall (j,k) \in \mathcal{I}_\sigma, i \in \{1,2\}, \tag{3.52}$$

$$\lambda_{j}^{(i)}\Big(\frac{w_{j,0}}{(-1)^{(i-1)}} - \beta\Big) = 0, \;\; \lambda_{j}^{(i)} \geq 0 \text{ and } \Big(\frac{w_{j,0}}{(-1)^{(i-1)}} - \beta\Big) \leq 0 \quad \forall j \in \mathcal{I}_h, i \in \{3,4\}. \tag{3.53}$$

Lemma 3.13. *The following conditions are equivalent for $\lambda, g \in \mathbb{R}$ and an arbitrary $\kappa > 0$:*

$$\max\{0,\, \lambda + \kappa g\} - \lambda = 0 \tag{3.54}$$

$$\Leftrightarrow \qquad \lambda g = 0 \quad \wedge \quad \lambda \geq 0 \quad \wedge \quad g \leq 0 \tag{3.55}$$

Proof.
"⇒":
Let (3.54) hold. This directly leads to $\lambda = \max\{0,\, \lambda + \kappa g\} \geq 0$. Next we will look at the two possible cases.

1^{st} *case:* $\max\{0,\, \lambda + \kappa g\} = 0$

$$\lambda = \max\{0,\, \lambda + \kappa g\} = 0 \quad \Rightarrow \quad \lambda g = 0$$
$$\max\{0,\, \lambda + \kappa g\} = 0 \quad \Rightarrow \quad \kappa g \leq 0 \quad \Rightarrow \quad g \leq 0$$

2^{nd} *case:* $\max\{0,\, \lambda + \kappa g\} = \lambda + \kappa g$

$$\lambda = \max\{0,\, \lambda + \kappa g\} = \lambda + \kappa g \quad \Rightarrow \quad \kappa g = 0 \quad \Rightarrow \quad g = 0 \quad \wedge \quad \lambda g = 0$$

"⇐":
Let (3.55) hold. We will also look at two cases here.

1^{st} *case:* $g = 0$

$$\max\{0,\, \lambda + \kappa g\} \overset{(\lambda \geq 0)}{=} \lambda - \lambda = 0$$

2^{nd} *case:* $g < 0$
From $\lambda g = 0$ we can immediately conclude that $\lambda = 0$. And thus:

$$\max\{0,\, \lambda + \kappa g\} - \lambda = \max\{0,\, \kappa g\} = 0.$$

□

Using this Lemma we can express (3.52) - (3.53) equivalently with an arbitrary $\kappa > 0$ for all $(j,k) \in \mathcal{I}_\sigma$ and $j \in \mathcal{I}_h$ by the following equations:

$$\begin{aligned}
N_{j,k}^{(1)} &:= \max\Big\{0,\, \lambda_{j,k}^{(1)} + \kappa(w_{j,k} - \alpha)\Big\} \;\; - \lambda_{j,k}^{(1)} = 0, \\
N_{j,k}^{(2)} &:= \max\Big\{0,\, \lambda_{j,k}^{(2)} + \kappa(-w_{j,k} - \alpha)\Big\} - \lambda_{j,k}^{(2)} = 0, \\
N_{j}^{(3)} &:= \max\Big\{0,\, \lambda_{j}^{(3)} + \kappa(w_{j,0} - \beta)\Big\} \;\; - \lambda_{j}^{(3)} = 0, \\
N_{j}^{(4)} &:= \max\Big\{0,\, \lambda_{j}^{(4)} + \kappa(-w_{j,0} - \beta)\Big\} - \lambda_{j}^{(4)} = 0.
\end{aligned}$$

We define

$$\mathscr{F}(w_\sigma, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}) := \begin{pmatrix} \frac{\partial \mathscr{L}}{\partial w_\sigma} & N^{(1)} & N^{(2)} & N^{(3)} & N^{(4)} \end{pmatrix}^\top \in \mathbb{R}^{N_\sigma + 2(|\mathcal{I}_\sigma| + |\mathcal{I}_h|)}$$

containing the left sides of our optimality system and solve the equation $\mathscr{F}(w_\sigma, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \lambda^{(4)}) = 0$ by a semismooth Newton method (Algorithm 1). In this algorithm we need to choose the matrix to be used in the semismooth Newton equation from the set of matrices denoted by Clarke's generalized Jacobian (Definition 2.31), which we repeat here:

$$\partial^{cl}\mathscr{F}(x) := \operatorname{conv}\left\{M : x^k \overset{k\to\infty}{\longrightarrow} x,\ \mathscr{F}'(x^k) \longrightarrow M,\ \mathscr{F} \text{ differentiable at } x^k\right\}.$$

Here, a choice has to be made for the numerics since the generalized Jacobians of $N^{(i)}$, $i \in \{1, 2, 3, 4\}$ need not be singletons. This is due to the max-functions and because we have

$$\partial_x(\max\{0, g(x)\}) = \begin{cases} 0, & \text{if } g(x) < 0, \\ \operatorname{conv}\{0, \partial_x g(x)\}, & \text{if } g(x) = 0, \\ \partial_x g(x), & \text{if } g(x) > 0, \end{cases}$$

for every differentiable scalar function $g(x)$.
Here, we make the decision to always choose $\partial_x(\max\{0, g(x)\}) = \partial_x g(x)$, if $g(x) = 0$. Using this, we define:

$$D\mathscr{F} := \begin{pmatrix} \frac{\partial^2 \mathscr{L}}{\partial^2 w_\sigma^2} & \frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(1)}} & \frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(2)}} & \frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(3)}} & \frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(4)}} \\ \frac{\partial N^{(1)}}{\partial w_\sigma} & \frac{\partial N^{(1)}}{\partial \lambda^{(1)}} & 0 & 0 & 0 \\ \frac{\partial N^{(2)}}{\partial w_\sigma} & 0 & \frac{\partial N^{(2)}}{\partial \lambda^{(2)}} & 0 & 0 \\ \frac{\partial N^{(3)}}{\partial w_\sigma} & 0 & 0 & \frac{\partial N^{(3)}}{\partial \lambda^{(3)}} & 0 \\ \frac{\partial N^{(4)}}{\partial w_\sigma} & 0 & 0 & 0 & \frac{\partial N^{(4)}}{\partial \lambda^{(4)}} \end{pmatrix} \in \partial^{cl}\mathscr{F}, \tag{3.56}$$

for the semismooth Newton method. To specify the non-zero entries of $D\mathscr{F}$, we first look at the derivative of $\mathscr{L}$ by w_σ in direction ϕ:

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial w_\sigma}(\phi) &= \frac{\partial}{\partial w_\sigma}(\phi)\left(\frac{1}{p}\|\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma\|^p_{L^p(Q_h)} + \left\langle \mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma, y_{\mathrm{d},\sigma}\right\rangle_{L^p(Q_h),L^q(Q_h)}\right) + \begin{pmatrix}\lambda^{(3)}\\ \lambda^{(1)}\end{pmatrix} - \begin{pmatrix}\lambda^{(4)}\\ \lambda^{(2)}\end{pmatrix} \\ &= \int_{Q_h} |\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma|^{(p-2)}(\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma)(\mathcal{M}_\sigma^{-1}\mathcal{L}\phi)\,dx\,dt + \left\langle \mathcal{M}_\sigma^{-1}\mathcal{L}\phi, y_{\mathrm{d},\sigma}\right\rangle_{L^p(Q_h),L^q(Q_h)} + \begin{pmatrix}\lambda^{(3)}\\ \lambda^{(1)}\end{pmatrix} - \begin{pmatrix}\lambda^{(4)}\\ \lambda^{(2)}\end{pmatrix}. \end{aligned}$$

We then can apply product rule to obtain the second derivative of $\mathscr{L}$ by w_σ in direction (ϕ, ξ):

$$\begin{aligned} \frac{\partial^2 \mathscr{L}}{\partial^2 w_\sigma^2}(\phi, \xi) &= \int_{Q_h} \frac{\partial}{\partial w_\sigma}(\xi)\left(|\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma|^{(p-2)}\right)(\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma)(\mathcal{M}_\sigma^{-1}\mathcal{L}\phi)\,dx\,dt \\ &\quad + \int_{Q_h} |\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma|^{(p-2)}\frac{\partial}{\partial w_\sigma}(\xi)(\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma)(\mathcal{M}_\sigma^{-1}\mathcal{L}\phi)\,dx\,dt \\ &= (p-1)\int_{Q_h} |\mathcal{M}_\sigma^{-1}\mathcal{L}w_\sigma|^{(p-2)}(\mathcal{M}_\sigma^{-1}\mathcal{L}\xi)(\mathcal{M}_\sigma^{-1}\mathcal{L}\phi)\,dx\,dt. \end{aligned}$$

The remaining four blocks in the first row are

$$\begin{aligned} \frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(1)}} &= \begin{pmatrix} 0 \\ \mathbb{I}_{|\mathcal{I}_\sigma|} \end{pmatrix} \quad \in \mathbb{R}^{(|\mathcal{I}_h| + |\mathcal{I}_\sigma|)\times|\mathcal{I}_\sigma|}, \\ \frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(2)}} &= \begin{pmatrix} 0 \\ -\mathbb{I}_{|\mathcal{I}_\sigma|} \end{pmatrix} \quad \in \mathbb{R}^{(|\mathcal{I}_h| + |\mathcal{I}_\sigma|)\times|\mathcal{I}_\sigma|}, \end{aligned}$$

$$\frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(3)}} = \begin{pmatrix} \mathbb{I}_{|\mathcal{I}_h|} \\ 0 \end{pmatrix} \quad \in \mathbb{R}^{(|\mathcal{I}_h|+|\mathcal{I}_\sigma|)\times|\mathcal{I}_h|},$$
$$\frac{\partial^2 \mathscr{L}}{\partial w_\sigma \partial \lambda^{(4)}} = \begin{pmatrix} -\mathbb{I}_{|\mathcal{I}_h|} \\ 0 \end{pmatrix} \quad \in \mathbb{R}^{(|\mathcal{I}_h|+|\mathcal{I}_\sigma|)\times|\mathcal{I}_h|},$$

where by $\mathbb{I}_N$ the identity matrix of size $N \times N$ is denoted. Next we move on to the other blocks in the first column. we observe the sparsity structures

$$\frac{\partial N^{(1)}_{j,k}}{\partial w_{l,m}} = \frac{\partial N^{(2)}_{j,k}}{\partial w_{l,m}} = 0 \quad \text{, if } j \neq l \text{ or } k \neq m,$$
$$\frac{\partial N^{(3)}_{j}}{\partial w_{l,m}} = \frac{\partial N^{(4)}_{j}}{\partial w_{l,m}} = 0 \quad \text{, if } j \neq l \text{ or } 0 \neq m.$$

The non-zero elements are the following

$$\frac{\partial N^{(1)}_{j,k}}{\partial w_{j,k}} = \begin{cases} 0, & \text{if } \lambda^{(1)}_{j,k} + \kappa(w_{j,k} - \alpha) < 0, \\ \kappa, & \text{else.} \end{cases}$$
$$\frac{\partial N^{(2)}_{j,k}}{\partial w_{j,k}} = \begin{cases} 0, & \text{if } \lambda^{(2)}_{j,k} + \kappa(-w_{j,k} - \alpha) < 0, \\ -\kappa, & \text{else.} \end{cases}$$
$$\frac{\partial N^{(3)}_{j}}{\partial w_{j,0}} = \begin{cases} 0, & \text{if } \lambda^{(3)}_{j} + \kappa(w_{j,0} - \beta) < 0, \\ \kappa, & \text{else.} \end{cases}$$
$$\frac{\partial N^{(4)}_{j}}{\partial w_{j,0}} = \begin{cases} 0, & \text{if } \lambda^{(4)}_{j} + \kappa(-w_{j,0} - \beta) < 0, \\ -\kappa, & \text{else.} \end{cases}$$

Finally, we specify the remaining blocks on the diagonal $\frac{\partial N^{(i)}}{\partial \lambda^{(i)}}$ for $i \in \{1, 2, 3, 4\}$. Again we observe a sparsity structure, i.e. only the diagonal entries of the matrices will be non-zero. So it suffices to characterize those:

$$\frac{\partial N^{(1)}_{j,k}}{\partial \lambda^{(1)}_{j,k}} = \begin{cases} -1, & \text{if } \lambda^{(1)}_{j,k} + \kappa(w_{j,k} - \alpha) < 0, \\ 0, & \text{else.} \end{cases}$$
$$\frac{\partial N^{(2)}_{j,k}}{\partial \lambda^{(2)}_{j,k}} = \begin{cases} -1, & \text{if } \lambda^{(2)}_{j,k} + \kappa(-w_{j,k} - \alpha) < 0, \\ 0, & \text{else.} \end{cases}$$
$$\frac{\partial N^{(3)}_{j}}{\partial \lambda^{(3)}_{j}} = \begin{cases} -1, & \text{if } \lambda^{(3)}_{j} + \kappa(w_{j,0} - \beta) < 0, \\ 0, & \text{else.} \end{cases}$$
$$\frac{\partial N^{(4)}_{j}}{\partial \lambda^{(4)}_{j}} = \begin{cases} -1, & \text{if } \lambda^{(4)}_{j} + \kappa(-w_{j,0} - \beta) < 0, \\ 0, & \text{else.} \end{cases}$$

An interesting observation is that for $\kappa = 1$ we have a symmetric matrix on the active sets.
We want to remark that we could follow [14] and employ Fenchel duality to recover a problem in the variable $\tilde{u}$. However, this would require to add the representation of the adjoint from (3.11) to our optimality system. As $p > 2$ the exponent $\frac{1}{p-1}$ is strictly smaller than 1, which is problematic for derivative based methods and was our main motivation to use the Fenchel duality approach in the first place. Instead, we solve for the optimal adjoint $\bar{w}_\sigma$ and recover the optimal control $(\bar{u}_{0,h}, \bar{u}_\sigma)$ through the discrete version of (3.13):

$$(\mathcal{R}_h + \mathcal{R}_\sigma)^\top (\bar{u}_{0,h}^\top, \bar{u}_\sigma^\top)^\top = \mathcal{L}^\top (|\mathcal{M}_\sigma^{-1} \mathcal{L} \bar{w}_\sigma|^{p-2} \mathcal{M}_\sigma^{-1} \mathcal{L} \bar{w}_\sigma + y_{\mathrm{d},\sigma}). \tag{3.57}$$

One could come to the conclusion that this is problematic since $(\mathcal{R}_h + \mathcal{R}_\sigma)^\top$ is in general only injective, not surjective. But the expansion coefficients $\bar{u}_{j,0}$ and $\bar{u}_{j,k}$ of the discrete solution $(\bar{u}_{0,h}, \bar{u}_\sigma)$ have to vanish anyways for $j \notin \mathcal{I}_h$ and $(j,k) \notin \mathcal{I}_\sigma$ (see (3.31)). So the remaining coefficients have merely to be read off.

3.4 Discontinuous Galerkin discretization

This section deals with a full discretization concept of (P), namely discontinuous Galerkin discretization, which is suggested in [20]. The discretization strategy will be adapted to our setting and notation. Also a convergence result for the fully discrete problem (P_{DG}) analogous to Theorem 3.2 is proven in [20].

We use the discrete spaces $\mathcal{Y}_\sigma, Y_h, U_h$ as introduced before in (3.24), (3.25), (3.32), respectively and the space-time discrete control space:

$$\mathcal{U}_{\mathrm{DG}} \coloneqq \operatorname{span}\{\delta_{x_j} \otimes \chi_k : j \in \mathcal{I}_h, k \in \mathcal{I}_\tau\}, \qquad \mathcal{I}_\tau \coloneqq \{k : I_k \subset \bar{I}_c\}.$$

In [20], an implicit Euler time stepping scheme is used for the discrete state equation. For $y_\sigma \in \mathcal{Y}_\sigma$ and for every $k \in \{1, \dots, N_\tau\}$, we define $y_{k,h} \coloneqq y_\sigma|_{I_k} \in Y_h$. Let $(u_{0,h}, u_\sigma) \in U_h \times \mathcal{U}_{\mathrm{DG}}$ be given and $z_h \in Y_h$ arbitrary. Then the following equations form the discrete state equation:

$$\begin{cases} \langle y_{k,h} - y_{k-1,h}, z_h \rangle_{L^2} + \tau_k \int_\Omega \nabla y_{k,h} \nabla z_h \, dx = \int_{\bar{Q}_c} (z_h \otimes \chi_k) \, du_\sigma, \quad k \in \{1, \dots, N_\tau\}, \\ y_{0,h} = y_{0h}, \end{cases} \tag{3.58}$$

where $y_{0h} \in Y_h$ is the unique element satisfying:

$$\langle y_{0h}, z_h \rangle_{L^2} = \int_{\bar{\Omega}_c} z_h \, du_{0,h} \qquad \forall\, z_h \in Y_h. \tag{3.59}$$

Here $\langle \cdot, \cdot \rangle_{L^2}$ denotes the scalar product in $L^2(\Omega)$. We denote the solution of the discrete state equation (3.58) by $y_\sigma(u_{0h}, u_\sigma)$, and define the discrete objective function

$$J_{\mathrm{DG}}(u_{0h}, u_\sigma) \coloneqq \frac{1}{q} \|y_\sigma(u_{0h}, u_\sigma) - y_{\mathrm{d}}\|^q_{L^q(Q_h)} + \alpha \|u_\sigma\|_{\mathcal{M}(\bar{Q}_c)} + \beta \|u_{0h}\|_{\mathcal{M}(\bar{\Omega}_c)}.$$

This allows us to formulate the following discrete optimization problem

$$\min_{(u_{0h}, u_\sigma) \in U_h \times \mathcal{U}_{\mathrm{DG}}} J_{\mathrm{DG}}(u_{0h}, u_\sigma). \tag{P_{DG}}$$

Similar to [14], we set up the system matrix for the discrete state equation. One difference that we need to consider is $u_{0,h} \neq 0$. This leads to N_h further degrees of freedom for the state and also to N_h additional columns and N_h additional rows in the system matrix; see [70, Chapter 12] for further details. With the mass matrix $M_h = (\langle e_{x_j}, e_{x_k} \rangle)_{j,k=1}^{N_h}$ and the stiffness matrix $A_h = (\int_\Omega \nabla e_{x_j} \nabla e_{x_k} \, dx)_{j,k=1}^{N_h}$, the left hand sides of (3.58) and (3.59) can be encoded into the following matrix of size $(N_\sigma + N_h) \times (N_\sigma + N_h)$:

$$\mathcal{L}_{\mathrm{DG}}^\top \coloneqq \begin{pmatrix} M_h & 0 & \dots & \dots & 0 \\ -M_h & M_h + \tau_1 A_h & & & \vdots \\ 0 & -M_h & M_h + \tau_2 A_h & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -M_h & M_h + \tau_{N_\tau} A_h \end{pmatrix}.$$

As in Section 3.3, this matrix represents the state-to-control-operator concatenated with the space-time mass matrix $\mathcal{M}_\sigma$. The representation of the adjoint state equation is $\mathcal{L}_{\mathrm{DG}} = (\mathcal{L}_{\mathrm{DG}}^\top)^\top$. With the discrete representations

$$u_\sigma = \sum_{j\in\mathcal{I}_h}\sum_{i\in\mathcal{I}_\tau} u_{j,i}\,\delta_{x_j}\otimes\chi_i \in \mathcal{U}_{\mathrm{DG}} \quad \text{and} \quad z_h = \sum_{l=1}^{N_h} z_l\, e_{x_l} \in Y_h$$

and using that the "mass matrix" $(\langle\delta_{x_j}, e_{x_l}\rangle)_{j,l=1}^{N_h}$ is the identity in $\mathbb{R}^{N_h\times N_h}$, we obtain the following in (3.58):

$$\int_{\bar{Q}_c} (z_h\otimes\chi_k)\,du_\sigma = \begin{cases}\tau_k \sum\limits_{j\in\mathcal{I}_h} u_{j,k}\,z_j, & \text{if } k\in\mathcal{I}_\tau,\\ 0, & \text{else.}\end{cases}$$

Analogously, with $u_{0,h} = \sum\limits_{j\in\mathcal{I}_h} u_j\,\delta_{x_j} \in U_h$, the right hand side from (3.59) turns into

$$\int_{\bar{\Omega}_c} z_h\,du_{0,h} = \sum_{j\in\mathcal{I}_h} u_j\,z_j.$$

Restriction matrices can be derived similar as in Section 3.3 to write the discrete state equation in matrix form. This can be used to discretize (P^*), leading to an optimization problem in the variable

$$w_{\mathrm{DG}} \;=\; (w_{j,k})_{j=1,k=0}^{N_h,N_\tau} \;\in\; \mathbb{R}^{N_\sigma+N_h}.$$

The setup for the fully discrete problem and the derivation of the optimality system are almost identical to the procedures from Section 3.3; one only has to replace $\mathcal{L}$ by $\mathcal{L}_{\mathrm{DG}}$ and to keep in mind that the number of degrees of freedom changes from N_σ to $N_\sigma + N_h$.

3.5 Computational results

We numerically solve (P_σ^*) by a semismooth Newton method as derived in Section 3.3. To simplify, we fix $u_0 = 0$. Therefore the condition $\|\Phi_h(w_\sigma)\|_{\infty,\bar{Q}_c} \le \beta$ in problem (P_σ^*) disappears and so do $\lambda^{(3)}$ and $\lambda^{(4)}$ in the Lagrangian $\mathscr{L}$. The dimension of the optimality system is reduced accordingly since $N^{(3)}$ and $N^{(4)}$ do not have to be considered. For the discontinuous Galerkin discretization from Section 3.4 of problem (P^*), we proceed similarly. Furthermore, the first row and column of $\mathcal{L}_{\mathrm{DG}}^\top$ can be eliminated here.

In this section all variables are specified as their discrete representatives, hence we omit the indices. As our domain for both examples, we choose $\Omega = (0,1)\subset\mathbb{R}$, $I = (0,\frac{3}{2})$, and the relatively compact Lipschitz domain

$$Q_c \coloneqq \left(\frac{1}{4},\frac{3}{4}\right)\times\left(\frac{1}{4},\frac{5}{4}\right) \subset\subset Q = \Omega\times I.$$

We assume an equidistant mesh, consequently every cell is of size $\tau\cdot h$. We set $\kappa = 1$ and $q = \frac{4}{3}$, so that $p = 4$.

Let us remark that $p > 2$ can lead the the matrix $D\mathscr{F}(w_\sigma,\lambda)$ being singular. The cause of this trouble is the second derivative of $z\mapsto\frac{1}{p}\|z\|_{L^p(Q)}^p$; it appears as central building block of $\frac{\partial^2\mathscr{L}}{\partial^2 w^2}$ and is nearly singular whenever z is not pointwise bounded away from 0. We circumvent this problem by adding a suitable multiple of the residual $\|\mathscr{F}\|\,\mathcal{M}_\sigma$ to the diagonal entries d_{ii} of the second derivative of $z\mapsto\frac{1}{p}\|z\|_{L^p(Q)}^p$; as it is well-known (see, e.g., [63]), such a regularization does not deteriorate the convergence rate of Newton's method.

In particular, we find the following setup to be suitable: $p = 0.1, r = \min(1, 10 * \text{residual})$ and

$$d_{ii} = \frac{d_{ii}^2 + d_{ii}\cdot p + r\cdot p}{d_{ii} + p}.$$

The choice for r delivers a smaller regularization for decreasing residual in the Newton's method.

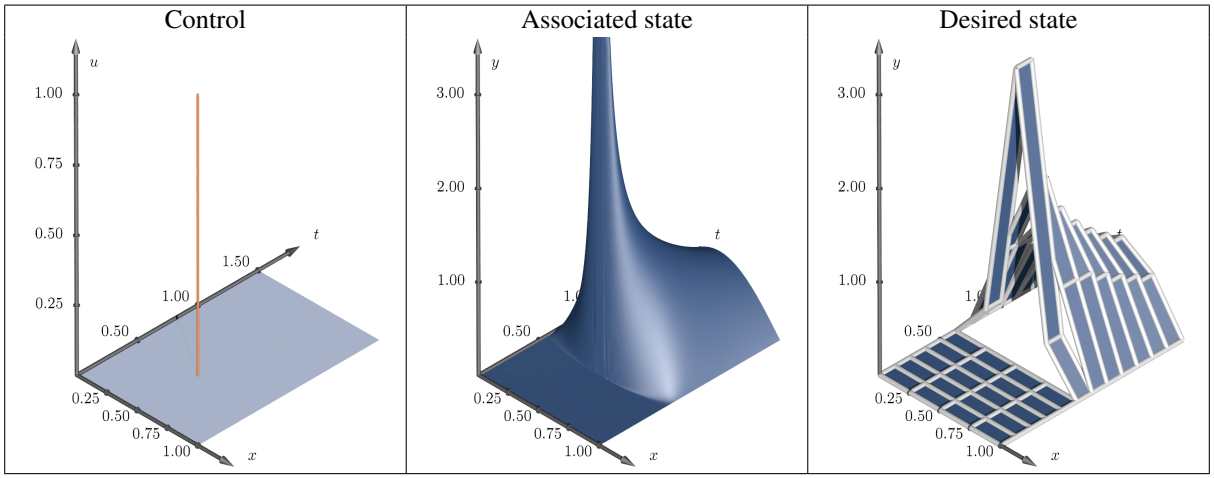

Figure 3.1: Numerical setup on 4×12 space-time grid with $q = \frac{4}{3}$. From left to right: control $u = \delta_{(1/2,1/2)}$, associated state $y(u)$ (sampled from the analytic solution with spacial Fourier modes - for more details see Appendix A.3), and discrete desired state y_d.

The purpose of our first numerical example is to illustrate the differences between variational discretization and discontinuous Galerkin discretization. Therefore, we use a relatively coarse space-time grid with $h = \frac{1}{4}$ and $\tau = \frac{h}{2} = \frac{1}{8}$. We generate a discrete desired state y_d by setting $y_\mathrm{d} := y(u) = L^{-*}\Phi^*(u)$ for the measure control $u = \delta_{(1/2,1/2)}$; afterwards, we discretize y_d according to (3.51), where we utilize a truncated Fourier expansion in order to evaluate y_d on the points $(x_j, (t_{k-1} + t_k)/2)$. Consequently, this problem is a source identification example that inherits sparsity.

If the penalty parameter α equals zero, the only admissible point for the predual problem is $w \equiv 0$. Hence, (3.57) shows that in this case the optimal discrete controls $u_{\sigma,0}$ and $u_{\mathrm{DG},0}$ for the two discretization approaches can be calculated by applying the discrete heat operator to y_d. This is meaningful since, for $\alpha = 0$, the only term remaining in the objective functional is the tracking term $\frac{1}{q}\|y - y_\mathrm{d}\|^q_{L^q(Q)}$. In this sense the controls $u_{\sigma,0}$ and $u_{\mathrm{DG},0}$ are the solutions to (P_σ) and (P_DG) for $\alpha = 0$ with the chosen y_d. Due to the different discretization approaches the calculated controls differ, which can be observed in Figure 3.2. As a consequence of the discretization error in y_d, we are not able to reproduce u exactly in either case.

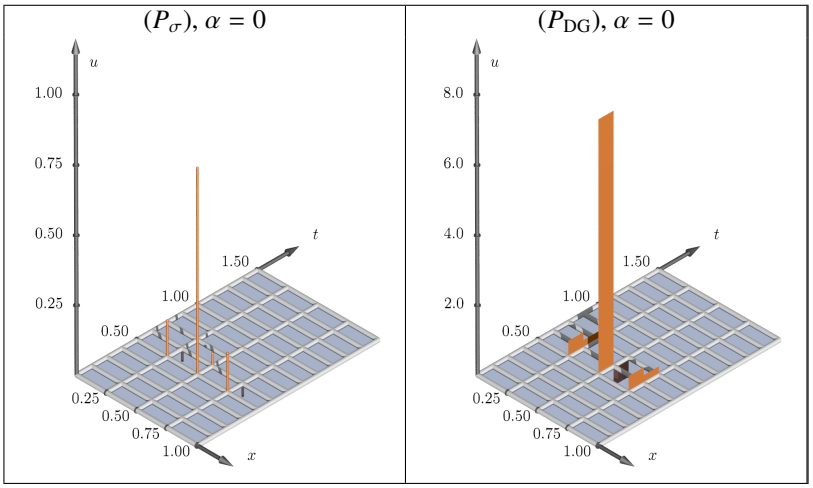

Figure 3.2: Numerical setup on 4×12 space-time grid with $q = \frac{4}{3}$. From left to right: calculated controls $u_{\sigma,0}$ and $u_{\mathrm{DG},0}$ for $\alpha = 0$. Here the controls are represented by their coefficients. In the case of piecewise constant controls in time, which are used in the discontinuous Galerkin setting, this leads to coefficient values, which are scaled with $\frac{1}{\tau} = 8$.

To identify the source location, we raise the penalty parameter α because this will lead to a decrease in the norm of the control and we expect a smaller support. The influence of α can be observed by plotting the norm of $u_{\sigma,\alpha}$ and $u_{\mathrm{DG},\alpha}$ respectively for a range of α. For each $i \in \{\sigma, \mathrm{DG}\}$, there exists a value $\bar{\alpha}_i$, such that for all $\alpha_i \geq \bar{\alpha}_i$ the optimal control corresponding to $y_{\mathrm{d},\sigma}$ is $u_{i,\alpha_i} \equiv 0$. Additionally it is interesting to look at the values of $\|y_{i,\alpha} - y_{\mathrm{d}}\|_{L^{4/3}}$ for $i \in \{\sigma, \mathrm{DG}\}$ and various values of α; we plotted the dependences in Figure 3.3 and Figure 3.4.

According to our expectations, the control norms are monotonically decreasing in α and eventually go to zero, while the errors in the tracking terms $\|y_{i,\alpha} - y_{\mathrm{d}}\|_{L^{4/3}}$ grow. The graphs for both strategies look very similar. This makes perfect sense, as we discretize the same problem and both discretization strategies converge towards the true solution.

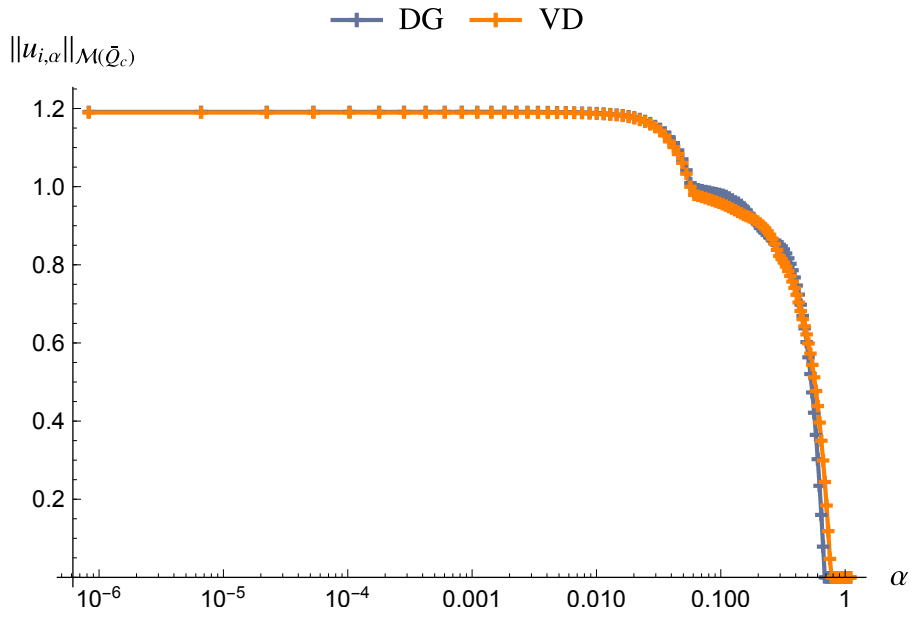

Figure 3.3: The dependence on the penalty parameter α of the measure norm of $u_{\sigma,\alpha}$ and $u_{\mathrm{DG},\alpha}$.

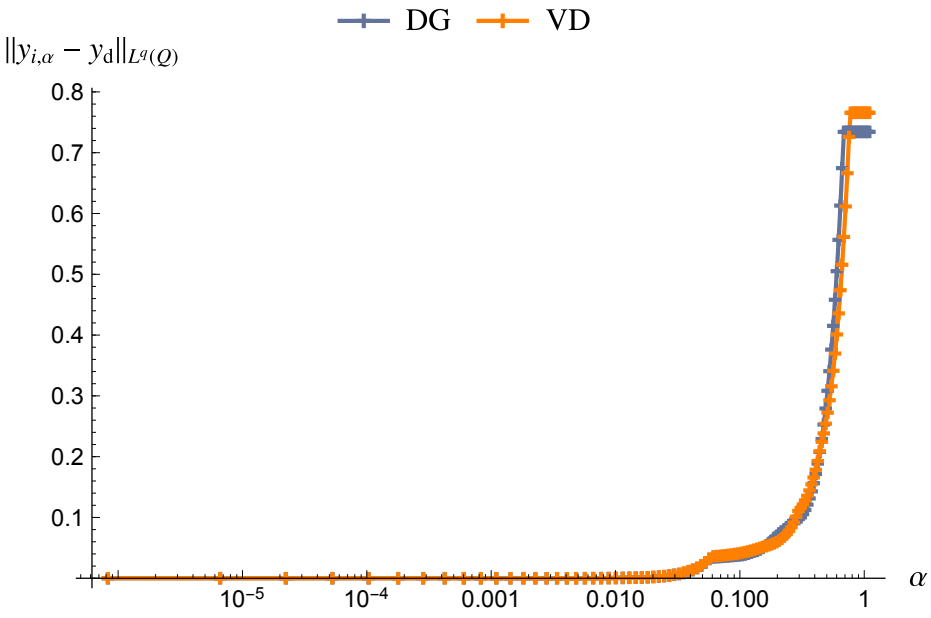

Figure 3.4: The dependence on the penalty parameter α of the errors $y_{\sigma,\alpha} - y_{\mathrm{d}}$ and $y_{\mathrm{DG},\alpha} - y_{\mathrm{d}}$ in the $L^{4/3}$ norm.

For further comparison of the two discretization strategies, we choose a value of α that leads to a norm of the controls, that is neither zero nor maximal. For $\alpha = 0.456$, the reconstructed controls and states are displayed in Figure 3.5. Here we see that the measure norm values $\|u_{\sigma,\alpha}\|_{\mathcal{M}(\bar{Q}_c)} = 0.6610$ and $\|u_{\mathrm{DG},\alpha}\|_{\mathcal{M}(\bar{Q}_c)} = 0.6692$ both differ from the true value $\|u\|_{\mathcal{M}(\bar{Q}_c)} = 1$. Furthermore we observe that $\mathrm{supp}(u_{\sigma,\alpha}) = \{(\frac{1}{1}, \frac{1}{2})\}$ and $\mathrm{supp}(u_{\mathrm{DG},\alpha}) = \{\frac{1}{2}\}\times(\frac{1}{2}, \frac{5}{8}]$, where the first coincides with the support of $u = \delta_{(1/2,1/2)}$. The control in the discontinuous Galerkin discrete setting can also be represented by a Dirac measure. In order to do so, it has to be multiplied by τ and the location of the measure in the time interval has to be chosen.

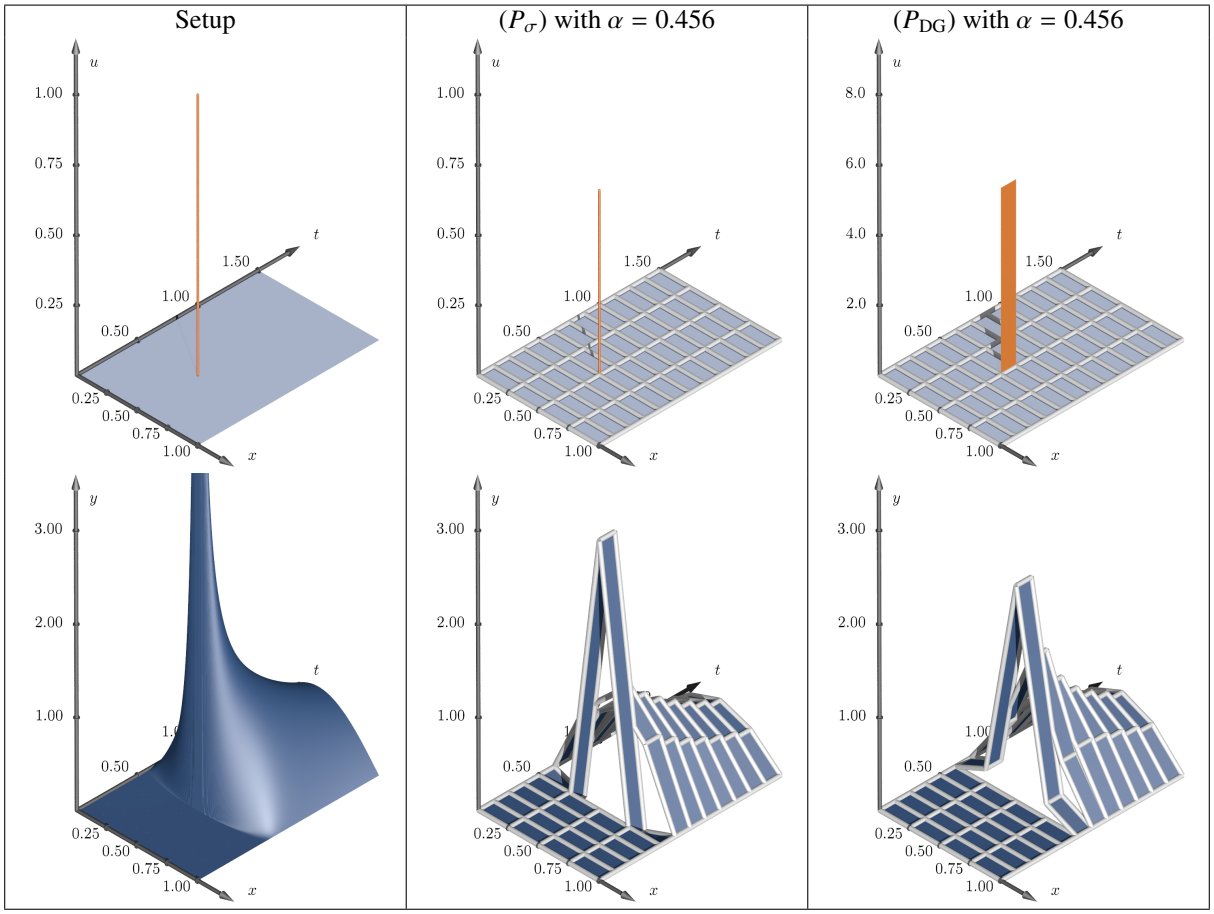

Figure 3.5: Top row: The measure control and the optimal controls $u_{\sigma,0.456}$ and $u_{\mathrm{DG},0.456}$. Bottom row: The associated state $y(u)$ (sampled from the analytic solution with spacial Fourier modes - for more details see Appendix A.3) and the associated states $y_{\sigma,0.456}$ and $y_{\mathrm{DG},0.456}$.

If the control u is not located on our space-time grid, it will be impossible to reproduce its support exactly. In the variational discretization approach a remedy might be choosing a test space $\mathcal{W}_\sigma$ consisting of piecewise quadratic – or even higher order – functions in time. Thereby the maximal values of the test functions $\pm\alpha$ could be attained not only at grid points, but also inside the time intervals. Determining the location of these maximal values would mean to determine the exact position in time of the potential support of the control. This will be part of further research.

Our second numerical example aims at visualizing the convergence properties from 3.2 and [20, Theorem 4.3.] for $|\sigma| \to 0$. Utilizing Fenchel duality, we can generate a discrete desired state y_d from a chosen true solution u_{true} for chosen $\bar{\alpha} > 0$. From (3.13), we deduce:

$$y_\mathrm{d} = L^{-*}\Phi^* u_{\mathrm{true}} - |Lw_{\bar{\alpha}}|^{p-2} Lw_{\bar{\alpha}}.$$

We set $u_{\text{true}} = \delta_{(1/2,1/2)}$ and the associated state $y_{\text{true}} := y(u_{\text{true}})$ sampled from the analytic solution with spacial Fourier modes (for more details see Appendix A.3). Furthermore, we take $w_{\bar{\alpha}}$, such that $w_{\bar{\alpha}}(1/2, 1/2) = -\bar{\alpha}$ and for all other values $(x, t) \in \bar{Q}_c$ it holds $|w_{\bar{\alpha}}(x, t)| < \bar{\alpha}$. For example, with $\bar{\alpha} = \frac{1}{4}$, this is fulfilled by

$$w_{\bar{\alpha}}(x, t) := -\frac{1}{4}\Big(((t - 0.5)^2 - 1)^2\,((2x - 1)^2 - 1)^2\Big).$$

In the following we fix $\alpha = \bar{\alpha}$. The setup is visualized exemplary on an equidistant 4×48-grid in Figure 3.6.

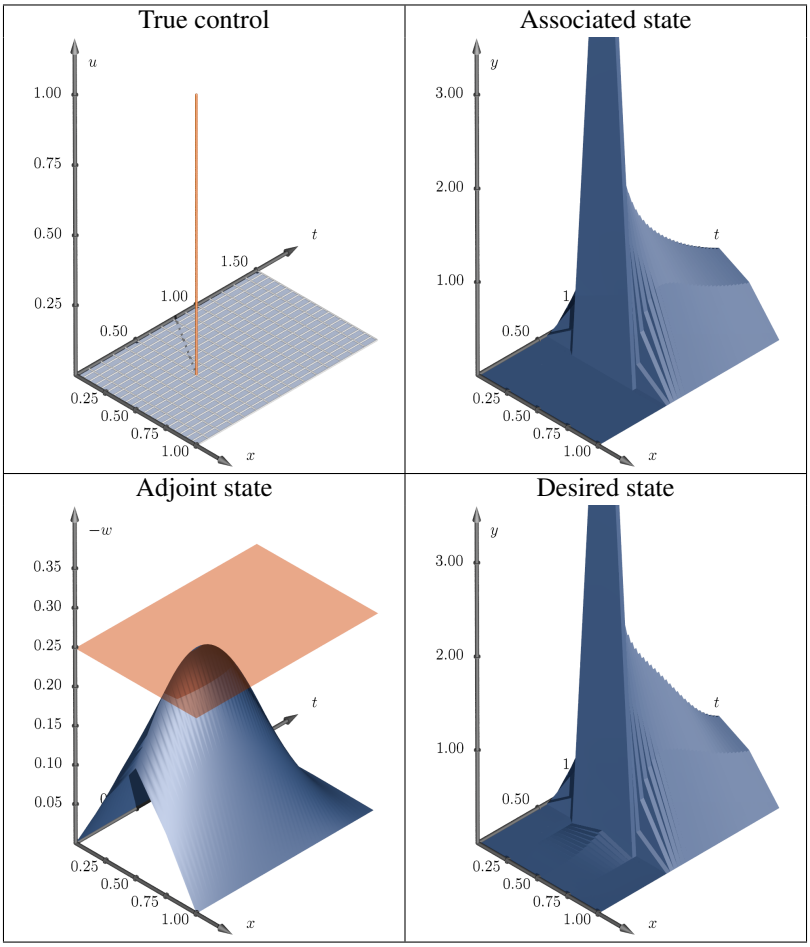

Figure 3.6: Numerical setup on a 4×48 space-time grid with $q = \frac{4}{3}$. From top left to bottom right: true control $u_{\text{true}} = \delta_{(1/2,1/2)}$, interpolation of the associated state $y(u_{\text{true}})$, adjoint state $w_{\bar{\alpha}}$ multiplied by -1 for easier visualization and desired state y_{d} calculated using Fenchel duality.

For both discretization strategies, i.e. $i \in \{\sigma, \text{DG}\}$, we have the following convergence properties:

$$\lim_{|\sigma|\to 0} \|u_i\|_{\mathcal{M}(\bar{Q}_c)} = \|u_{\text{true}}\|_{\mathcal{M}(\bar{Q}_c)} \qquad \text{and} \qquad \lim_{|\sigma|\to 0} \|y_i - y_{\text{true}}\|_{L^q(Q)} = 0.$$

In 3.7 we log-log-plot the errors $|\,\|u_i\|_{\mathcal{M}(\bar{Q}_c)} - \|u_{\text{true}}\|_{\mathcal{M}(\bar{Q}_c)}|$ and $\|y_i - y_{\text{true}}\|_{L^q(Q)}$, $i \in \{\sigma, \text{DG}\}$ versus the gridsize h. The proven convergence properties can be observed, as the errors go to zero for $h \to 0$, which is equivalent to $|\sigma| \to 0$ since τ is always linked to h. It is interesting to mention that in the variational discrete setting, oscillations in the state y_σ occur for small gridsizes, where $\tau = \frac{h}{2}$. This is caused by the Crank-Nicolson-like scheme. Nevertheless, we see convergence also in this case.

We calculate the convergence order h^α for the refinement from some gridsize h_1 to some other gridsize h_2, see Table 3.1 and Table 3.2. Here, we write $\|\cdot\|_{\mathcal{M}}$ as short notation for $\|\cdot\|_{\mathcal{M}(\bar{Q}_c)}$. Furthermore, in the table, we round the values of h_1 and h_2 to 4 digits after the comma. In computations we used negative potencies of 2 for the gridsize h.

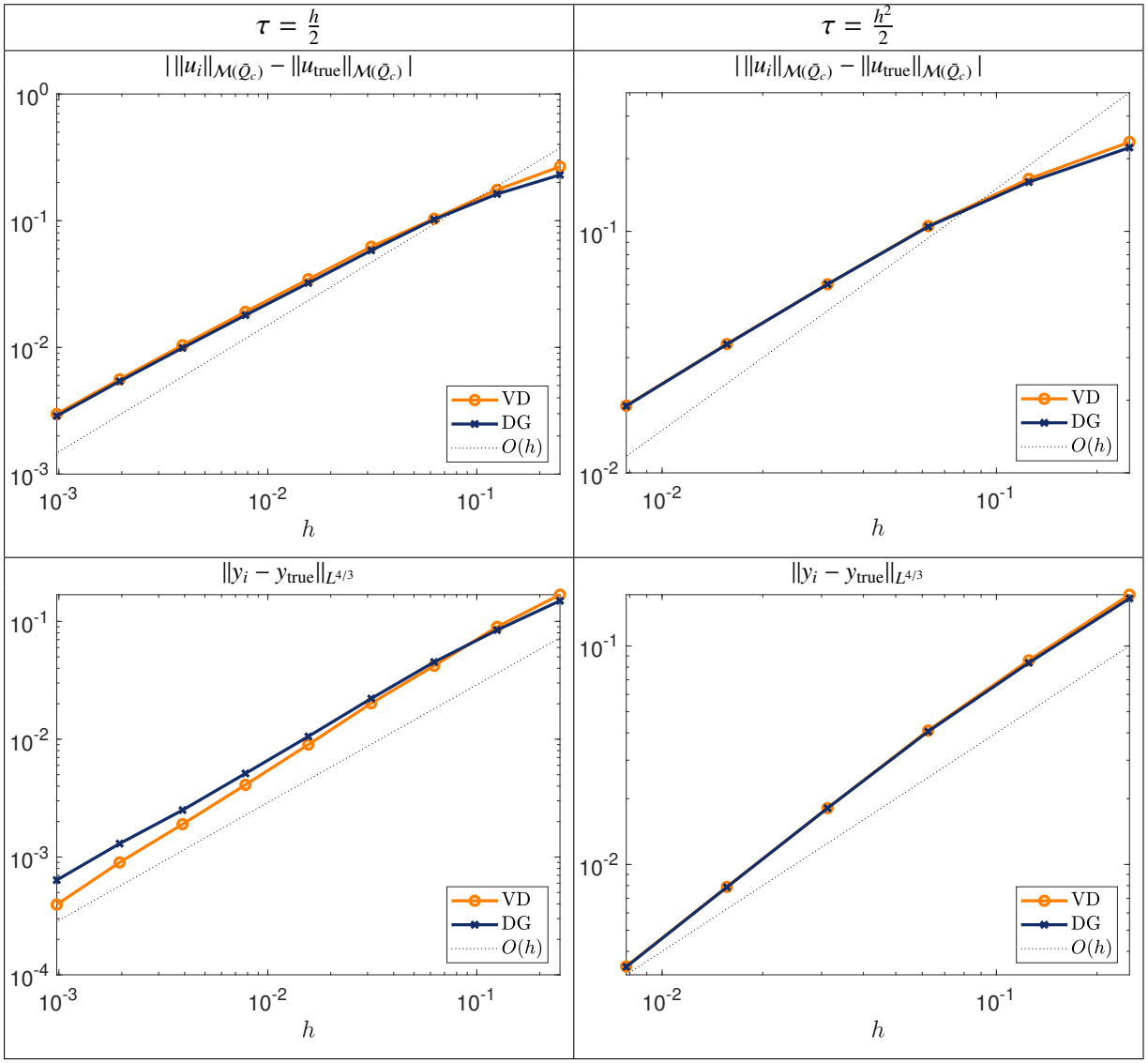

Figure 3.7: Top row: The difference of the measure norms of the true control u_{true} and calculated optimal control u_i, $i \in \{\sigma, \text{DG}\}$ for $\tau = \frac{h}{2}$ (left) and $\tau = \frac{h^2}{2}$ (right). Bottom row: The $L^{4/3}$-norm of the difference of the associated states $y_i - y_{\text{true}}$, $i \in \{\sigma, \text{DG}\}$ for $\tau = \frac{h}{2}$ (left) and $\tau = \frac{h^2}{2}$ (right). We remark that for $\tau = \frac{h}{2}$ we have more data points available, so we display a wider range of h values on the x-axis.

We observe many similarities in the derivation of the algorithms to solve the discrete problems. The implementation and the level of difficulty in programming is comparable for both approaches and we also observe similar iteration counts. The main advantage of the variational discretization compared to the discontinuous Galerkin discretization is the maximal discrete sparsity of the control achieved by choosing suitable ansatz and test spaces in the Petrov-Galerkin scheme.

h_1	h_2	$\lvert \lVert u_{\mathrm{VD}} \rVert_{\mathcal{M}} - \lVert u_{\mathrm{true}} \rVert_{\mathcal{M}} \rvert$	$\lvert \lVert u_{\mathrm{DG}} \rVert_{\mathcal{M}} - \lVert u_{\mathrm{true}} \rVert_{\mathcal{M}} \rvert$	$\lVert y_{\mathrm{VD}} - y_{\mathrm{true}} \rVert_{L^{4/3}}$	$\lVert y_{\mathrm{DG}} - y_{\mathrm{true}} \rVert_{L^{4/3}}$
0.25	0.125	0.6083	0.5020	0.9067	0.8240
0.125	0.0625	0.7568	0.6660	1.0974	0.9014
0.0625	0.0313	0.7338	0.8123	1.0577	1.0289
0.0313	0.0156	0.8539	0.8548	1.1645	1.0702
0.0156	0.0078	0.8531	0.8426	1.1351	1.0435
0.0078	0.0039	0.8764	0.8596	1.1100	1.0345
0.0039	0.0020	0.8931	0.8745	1.0780	0.9434
0.0020	0.0010	0.9083	0.9092	1.1892	1.0299
	mean	0.8105	0.7901	1.0923	0.9845
	slope of best fit	0.8196	0.8075	1.1020	1.0006

Table 3.1: convergence order (potency of gridsize h) of the respective errors when the grid is refined from gridsize h_1 to gridsize h_2 in the case $\tau = \frac{h}{2}$

h_1	h_2	$\lvert \lVert u_{\mathrm{VD}} \rVert_{\mathcal{M}} - \lVert u_{\mathrm{true}} \rVert_{\mathcal{M}} \rvert$	$\lvert \lVert u_{\mathrm{DG}} \rVert_{\mathcal{M}} - \lVert u_{\mathrm{true}} \rVert_{\mathcal{M}} \rvert$	$\lVert y_{\mathrm{VD}} - y_{\mathrm{true}} \rVert_{L^{4/3}}$	$\lVert y_{\mathrm{DG}} - y_{\mathrm{true}} \rVert_{L^{4/3}}$
0.25	0.125	0.5068	0.4744	0.9983	0.9792
0.125	0.0625	0.6485	0.6132	1.0670	1.0437
0.0625	0.0313	0.8005	0.7923	1.1796	1.1655
0.0313	0.0156	0.8223	0.8260	1.1973	1.2015
0.0156	0.0078	0.8478	0.8491	1.2084	1.2114
	mean	0.7252	0.7110	1.1301	1.1203
	slope of best fit	0.7355	0.7218	1.1361	1.1258

Table 3.2: convergence order (potency of gridsize h) of the respective errors when the grid is refined from gridsize h_1 to gridsize h_2 in the case $\tau = \frac{h^2}{2}$

Chapter 4

Parabolic optimal control governed by bounded initial measure controls

This chapter is based on article [46] with the title "Variational discretization approach applied to an optimal control problem with bounded measure controls".

The plan of this chapter is as follows: We state the optimal control problem in Section 4.1 , analyze the continuous problem, its sparsity structure and the special case of positive controls in Section 4.2. Thereafter we apply variational discretization to the optimal control problem in Section 4.3. Finally in Section 4.4 we apply the semismooth Newton method to the optimal control problem with positive controls (Subsection 4.4.1) and to the original optimal control problem (Subsection 4.4.2). For the latter we add a penalty term before applying the semismooth Newton method. For both cases we provide numerical examples.

4.1 Problem formulation

We consider the following optimal control problem which was analyzed in [21]:

$$\min_{u\in U_\alpha} J(u) = \frac{1}{2}\|y_u(T) - y_d\|^2_{L^2(\Omega)}. \tag{P_α}$$

Here let $y_d \in L^2(\Omega)$, and $U_\alpha := \{u \in \mathcal{M}(\bar{\Omega}) : \|u\|_{\mathcal{M}(\bar{\Omega})} \leq \alpha\}$, where $\mathcal{M}(\bar{\Omega})$ denotes the space of regular Borel measures on $\bar{\Omega}$ equipped with the norm

$$\|u\|_{\mathcal{M}(\bar{\Omega})} := \sup_{\|\phi\|_{C(\bar{\Omega})}\leq 1} \int_{\bar{\Omega}} \phi(x)\, du(x) = |u|(\bar{\Omega}).$$

The state y_u solves the parabolic equation

$$\begin{cases} \partial_t y_u + Ay_u &= f, \qquad \text{in } Q = \Omega \times (0,T), \\ y_u(x,0) &= u, \qquad \text{in } \bar{\Omega}, \\ \partial_n y_u(x,t) &= 0, \qquad \text{on } \Sigma = \Gamma \times (0,T), \end{cases} \tag{4.1}$$

where $f \in L^1(0,T;L^2(\Omega))$ is given, $\Omega \subset \mathbb{R}^n (n = 1,2,3)$ denotes an open, connected and bounded set with Lipschitz boundary $\Gamma = \partial\Omega$, and for the remainder of this chapter, let A be the elliptic operator defined by

$$Ay_u := -a\Delta y_u + b(x,t)\cdot\nabla y_u + c(x,t)y_u, \tag{4.2}$$

with a constant $a > 0$ and functions $b \in L^\infty(Q)^n$ and $c \in L^\infty(Q)$.

The state is supposed to solve (4.1) in the following very weak sense, see e.g. [21, Definition 2.1]:

Definition 4.1. *We say that a function $y \in L^1(Q)$ is a solution of* (4.1) *if the following identity holds:*

$$\int_Q (-\partial_t \phi + A^* \phi) y \, dxdt = \int_Q f\phi \, dxdt + \int_{\bar{\Omega}} \phi(0) \, du \quad \forall \, \phi \in \Phi, \tag{4.3}$$

where

$$\Phi := \{\phi \in L^2(0,T; H^1(\Omega)) : -\partial_t \phi + A^* \phi \in L^\infty(Q), \partial_n \phi = 0 \text{ on } \Sigma, \phi(T) = 0 \in \Omega\}$$

and $A^* \bar{\varphi} := -a\Delta\bar{\varphi} - \operatorname{div}[b(x,t)\bar{\varphi}] + c\bar{\varphi}$ denotes the adjoint operator of A.

The existence and uniqueness of solutions, in the sense of Definition 4.1, to the state equation (4.1), which is a deep result from the theory of diffusion-convection equations, has been proven in [21, Theorem 2.2].

4.2 Continuous optimality system

In this section we summarize properties of (P_α), which have been established in [21]. For completeness we also give the corresponding proofs. To begin, we repeat the following result from [21, Theorem 2.4]:

Theorem 4.2. *Problem (P_α) has a unique solution $\bar{u} \in U_\alpha$.*

To prove existence we will need a convergence result, given in [21, Theorem 2.3]:

Lemma 4.3. *Let $\{u_k\}_k \subset \mathcal{M}(\bar{\Omega})$, so that $u_k \overset{*}{\rightharpoonup} u \in \mathcal{M}(\bar{\Omega})$ with $\{y_k\}_k$ and y the associated states, respectively. Then, $y_k \rightharpoonup y \in L^r(0,T; W^{1,p}(\Omega))$ (for $p, r \in [1,2)$ with $\frac{2}{r} + \frac{n}{p} > n+1$) and*

$$\lim_{k\to\infty} \|y_k - y\|_{C([t_0,T];L^2(\Omega))} = 0 \quad \forall t_0 \in (0,T).$$

In particular, the convergence $y_k(T) \to y(T) \in L^2(\Omega)$ holds.

Proof. (of Theorem 4.2)
First, we prove existence of a solution. Let $\{u_k\}_k \subset U_\alpha$ be a minimizing sequence, such that

$$\lim_{k\to\infty} J(u_k) = \inf_{u\in U_\alpha} J(u) =: \underline{J}.$$

Due to $J \geq 0$ we have $\underline{J} \geq 0 > -\infty$. We observe that U_α is bounded in $\mathcal{M}(\bar{\Omega})$. Furthermore, for any weakly* convergent sequence $\{v_k\}_k \subset U_\alpha$ with $v_k \overset{*}{\rightharpoonup} v \in \mathcal{M}(\bar{\Omega})$ we have

$$\|v\|_{\mathcal{M}(\bar{\Omega})} = \sup_{\|f\|_{C(\bar{\Omega})}\leq 1} \int_{\bar{\Omega}} f \, dv = \lim_{k\to\infty} \sup_{\|f\|_{C(\bar{\Omega})}\leq 1} \int_{\bar{\Omega}} f \, dv_k = \lim_{k\to\infty} \|v_k\|_{\mathcal{M}(\bar{\Omega})} \leq \alpha,$$

and hence U_α is also weakly*-closed in $\mathcal{M}(\bar{\Omega})$. Then, from Banach-Alaoglu-Bourbaki Theorem (see e.g. [10, Theorem 3.16]) we have that U_α is weakly*-compact. It follows that any minimizing sequence is bounded in $\mathcal{M}(\bar{\Omega})$ and any weak* limit resides in U_α. From Lemma 4.3 we get

$$\begin{aligned} & u_k \overset{*}{\rightharpoonup} \bar{u} && \in \mathcal{M}(\bar{\Omega}) \\ \Rightarrow \quad & y_{u_k}(T) \to y_{\bar{u}}(T) && \in L^2(\Omega). \end{aligned}$$

So, by taking a suitable subsequence $\{u_{k'}\}_{k'}$ and from weakly lower semi continuity of J we have

$$\underline{J} \leq J(\bar{u}) \leq \liminf_{k'\to\infty} J(u_{k'}) = \lim_{k'\to\infty} J(u_{k'}) = \lim_{k\to\infty} J(u_k) = \underline{J},$$

so $\bar{u}$ solves (P_α).

Let us remark that indeed it is not necessary to argue with weakly lower semi continuity of J, since we have that $\{y_{u_k}(T)\}_k$ converges strongly in $L^2(\Omega)$.

Secondly, we prove uniqueness of the solution. To this end we assume there exist two solutions to (P_α), namely $\bar{u}_1$ and $\bar{u}_2 \in U_\alpha$ with associated states $\bar{y}_1$ and $\bar{y}_2$. Assume $\bar{y}_1(T) \neq \bar{y}_2(T)$, then for $\lambda \in (0,1)$:

$$\begin{aligned} J(\lambda\,\bar{u}_1 + (1-\lambda)\,\bar{u}_2) &= \frac{1}{2}\|\lambda\,\bar{y}_1(T) + (1-\lambda)\,\bar{y}_2(T) - y_d\|^2_{L^2(\Omega)} \\ &< \frac{1}{2}\lambda\,\|\bar{y}_1(T) - y_d\|^2_{L^2(\Omega)} + \frac{1}{2}(1-\lambda)\,\|\bar{y}_2(T) - y_d\|^2_{L^2(\Omega)} \\ &= \lambda\,J(\bar{u}_1) + (1-\lambda)\,J(\bar{u}_2) \\ &= J(\bar{u}_1), \end{aligned}$$

where we use the convexity of U_α, the strict convexity of J and the fact that $\bar{u}_1$ and $\bar{u}_2$ are solutions. The above inequality then contradicts $\bar{u}_1$ being a solution, hence we deduce that in fact $\bar{y}_1(T) = \bar{y}_2(T)$ must hold.

Now, set $\bar{y} = \bar{y}_2 - \bar{y}_1$ and let $t_0 \in (0,T)$ arbitrary, then one can see by elementary calculations

$$\bar{y}(x,t) = \frac{1}{t-t_0} z(x,t) \qquad \forall t \in (t_0, T),$$

where z satisfies

$$\begin{cases} \partial_t z + Az &= \bar{y}, \qquad \text{in } \Omega \times (t_0,T), \\ z(x,t_0) &= 0, \qquad \text{in } \bar{\Omega}, \\ \partial_n z(x,t) &= 0, \qquad \text{on } \Gamma \times (t_0,T). \end{cases}$$

From [21, Theorem 2.2] we know that in this setting $\bar{y} \in L^2(t_0,T;L^2(\Omega))$. Therefore, $z \in L^2(t_0,T;H^1(\Omega))$ and

$$\begin{cases} \partial_t z - a\Delta z &= \bar{y} + g, \qquad \text{in } \Omega \times (t_0,T), \\ z(x,t_0) &= 0, \qquad \text{in } \bar{\Omega}, \\ \partial_n z(x,t) &= 0, \qquad \text{on } \Gamma \times (t_0,T), \end{cases}$$

with $g = -b \cdot \nabla z - cz \in L^2(t_0,T;L^2(\Omega))$. We define

$$D(\Delta) := \left\{ \phi \in H^1(\Omega) : \Delta\phi \in L^2(\Omega) \text{ and } \partial_n \phi = 0 \text{ on } \Gamma \right\}.$$

From standard results on evolution equations (see e.g. [68, pp. 113-114]) we infer that

$$z \in C([t_0,T]; H^1(\Omega)) \cap L^2(t_0,T;D(\Delta)).$$

Since we have $z(x,t) = (t-t_0)\bar{y}(x,t)$, we get

$$\bar{y} \in C([\delta,T]; H^1(\Omega)) \cap L^2(\delta,T;D(\Delta)) \qquad \forall t_0 < \delta < T.$$

We have chosen $t_0 \in (0,T)$ arbitrary, therefore the above regularity of $\bar{y}$ also holds for arbitrary $\delta \in (0,T)$. Furthermore, since $\bar{y}$ satisfies the state equation (4.1), for almost every $t \in (\delta,T)$, we have

$$\|\partial_t \bar{y}(t) - a\Delta\bar{y}(t)\|_{L^2(\Omega)} = \|b(t)\cdot\nabla\bar{y}(t) + c(t)\bar{y}(t)\|_{L^2(\Omega)} \leq C\,\|\bar{y}(t)\|_{H^1(\Omega)},$$

for a suitable constant $C \in \mathbb{R}$. By definition of $\bar{y}$ we have

$$\bar{y}(T) = \bar{y}_2(T) - \bar{y}_1(T) = 0.$$

So, from backward uniqueness of the parabolic equation [37, Theorem 1.1] we conclude that $\bar{y}(t) = 0$ for all $t \in [\delta, T]$. Since we can take $\delta > 0$ arbitrarily small, we even get

$$\bar{y}(t) = 0 \qquad \forall t \in (0, T].$$

Take $r = 1$ in Lemma 4.3 to get

$$\bar{y} \in L^1(0, T; W^{1,p}(\Omega)) \qquad \text{for } 2 + \frac{n}{p} > n + 1.$$

The condition on p and n can equivalently be formulated as $p < \frac{n}{n-1}$. We also have

$$\partial_t \bar{y} = f - A\bar{y} \in L^1(0, T; W^{1,q}(\Omega)^*),$$

where $\frac{1}{p} + \frac{1}{q} = 1$. Together, we infer that $\bar{y} : [0, T] \to W^{1,q}(\Omega)^*$ is continuous.

Finally, this gives

$$\bar{u}_2 - \bar{u}_1 = \bar{y}_2(0) - \bar{y}_1(0) = \bar{y}(0) = \lim_{t \to 0} \bar{y}(t) = 0.$$

So, $\bar{u}_1 = \bar{u}_2$ and we can conclude uniqueness of the solution. □

Let $\bar{u}$ be the unique solution of (P_α) with associated state $\bar{y}$. We then say that $\bar{\varphi} \in L^2(0, T; H^1(\Omega)) \cap C\left(\bar{\Omega} \times [0, T]\right)$ is the associated adjoint state of $\bar{u}$, if it solves

$$\begin{cases} -\partial_t \bar{\varphi} + A^* \bar{\varphi} &= 0, & \text{in } Q, \\ \bar{\varphi}(x, T) &= \bar{y}(x, T) - y_d, & \text{in } \Omega, \\ \partial_n \bar{\varphi}(x, t) &= 0, & \text{on } \Sigma. \end{cases} \tag{4.4}$$

We recall the optimality conditions for (P_α) from [21, Theorem 2.5]:

Theorem 4.4. *Let $\bar{u}$ be the solution of (P_α) with $\bar{y}$ and $\bar{\varphi}$ the associated state and adjoint state, respectively. Then, the following properties hold*

1. *If $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$, then $\bar{y}(T) = y_d$ and $\bar{\varphi} = 0 \in Q$.*
2. *If $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$, then*

$$\operatorname{supp}(\bar{u}^+) \subset \{x \in \bar{\Omega} : \bar{\varphi}(x, 0) = -\|\bar{\varphi}(0)\|_{C(\bar{\Omega})}\},$$
$$\operatorname{supp}(\bar{u}^-) \subset \{x \in \bar{\Omega} : \bar{\varphi}(x, 0) = +\|\bar{\varphi}(0)\|_{C(\bar{\Omega})}\},$$

where $\bar{u} = \bar{u}^+ - \bar{u}^-$ is the Jordan decomposition of $\bar{u}$.

Conversely, if $\bar{u}$ is an element of U_α satisfying 1. or 2., then $\bar{u}$ is the solution to (P_α).

For completeness we will recapitulate the proof. In order to do so we give the following helpful results from [21, Lemma 2.6 and Lemma 2.7].

Lemma 4.5. *Given $u \in \mathcal{M}(\bar{\Omega})$, the solution $z_u \in L^r(0, T; W^{1,p}(\Omega))$ (for $p, r \in [1, 2)$ with $\frac{2}{r} + \frac{n}{p} > n + 1$) to*

$$\begin{cases} \partial_t z + Az &= 0, & \text{in } Q = \Omega \times (0, T), \\ z(x, 0) &= u, & \text{in } \bar{\Omega}, \\ \partial_n z(x, t) &= 0, & \text{on } \Sigma = \Gamma \times (0, T), \end{cases} \tag{4.5}$$

satisfies

$$\int_\Omega (\bar{y}(T) - y_\mathrm{d})\, z_u(T)\, dx = \int_{\bar{\Omega}} \bar{\varphi}(0)\, du.$$

Lemma 4.6. *For every $\epsilon > 0$ there exists a control $u \in L^2(\Omega)$, such that $\|y_u(T) - y_d\|_{L^2(\Omega)} < \epsilon$.*

We are now equipped to prove the optimality conditions.

Proof. (of Theorem 4.4)
Let $u \in U_\alpha$ arbitrary and denote by $z_{u-\bar{u}}$ the solution to (4.5), where u is replaced by $u - \bar{u}$. With Lemma 4.5 we get

$$\lim_{\rho \searrow 0} \frac{J(\bar{u} + \rho(u - \bar{u})) - J(\bar{u})}{\rho} = \int_\Omega (\bar{y}(T) - y_d)\, z_{u-\bar{u}}(T)\, dx = \int_{\bar{\Omega}} \bar{\varphi}(0)\, d(u - \bar{u}).$$

Now, (P_α) is a convex problem, therefore we have the following necessary and sufficient optimality condition for the solution $\bar{u} \in U_\alpha$:

$$\begin{aligned} \int_{\bar{\Omega}} \bar{\varphi}(0)\, d(u - \bar{u}) = J'(\bar{u})(u - \bar{u}) \ \geq 0 \qquad & \forall u \in U_\alpha, \text{ and thus} \\ -\int_{\bar{\Omega}} \bar{\varphi}(0)\, du \ \leq -\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} \qquad & \forall u \in U_\alpha. \end{aligned}$$

We can take the supremum over $u \in U_\alpha$ and get

$$\alpha \, \|\bar{\varphi}(0)\|_{C(\bar{\Omega})} = \sup_{u \in U_\alpha} -\int_{\bar{\Omega}} \bar{\varphi}(0)\, du = -\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u}.$$

If $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$, this identity is equivalent to

$$\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \|\bar{\varphi}(0)\|_{C(\bar{\Omega})} = -\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u}, \tag{4.6}$$

which in this case forms a necessary and sufficient optimality condition. Similarly to Lemma 3.8 we can conclude part 2 of the claim. Conversely, for $\bar{u} \in U_\alpha$ that satisfies part 2 of the claim, we can derive the equality (4.6), which shows that $\bar{u}$ solves (P_α) indeed.

We now consider the case $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$. For $\bar{y}(T) = y_d$ we have $J(\bar{u}) = 0 \leq J(u)$ for all $u \in U_\alpha$, since J is non-negative. Consequently $\bar{u}$ is the solution to (P_α).

Assume that $\bar{u}$ solves (P_α) and $\bar{y}(T) \neq y_d$ holds. From Lemma 4.6 we get the existence of an $u \in \mathcal{M}(\bar{\Omega})$ with $J(u) < J(\bar{u})$. Since $\bar{u}$ solves (P_α), we conclude that $u \notin U_\alpha$. Now take $\lambda \in \mathbb{R}$, such that

$$0 < \lambda < \min\left\{ \frac{\alpha - \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})}}{\|u - \bar{u}\|_{\mathcal{M}(\bar{\Omega})}}, 1 \right\}.$$

Define $v := \bar{u} + \lambda(u - \bar{u})$. It is easy to confirm that $v \in U_\alpha$. Furthermore, by convexity of J we have

$$J(v) = J(\lambda\, u + (1 - \lambda)\, \bar{u}) \leq \lambda \underbrace{J(u)}_{< J(\bar{u})} + (1 - \lambda)\, J(\bar{u}) < J(\bar{u}).$$

This is a contradiction to the optimality of $\bar{u}$, so if $\bar{u}$ is a solution, it must hold that $\bar{y}(T) = y_d$. Finally, from (4.4) we get $\bar{\varphi} = 0 \in Q$ and altogether part 1 of the claim. □

In some applications we may have a priori knowledge about the measure controls. This motivates the restriction of the admissible control set to positive controls $U_\alpha^+ := \{u \in \mathcal{M}^+(\bar{\Omega}) : \|u\|_{\mathcal{M}(\bar{\Omega})} \leq \alpha\}$, with $\|u\|_{\mathcal{M}(\bar{\Omega})} = u(\bar{\Omega})$. We then consider the problem

$$\min_{u \in U_\alpha^+} J(u) = \frac{1}{2} \|y_u(T) - y_d\|^2_{L^2(\Omega)}, \tag{P_α^+}$$

where y_u solves (4.1).

The properties of (P_α^+) have been derived in [21, Theorem 3.1]:

Theorem 4.7. *(P_α^+) has a unique solution. Let $\bar{u}$ be the unique solution of (P_α^+) with associated adjoint state $\bar{\varphi}$. Then, $\bar{u}$ is a solution of (P_α^+) if and only if*

$$\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} \leq \int_{\bar{\Omega}} \bar{\varphi}(0)\, du \qquad \forall\, u \in U_\alpha^+. \tag{4.7}$$

If $\bar{u}(\bar{\Omega}) = \alpha$ the following properties are fulfilled:

1. *Inequality* (4.7) *is equivalent to the identity*

$$\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} = \alpha\bar{\lambda} := \alpha \min_{x\in\bar{\Omega}} \bar{\varphi}(x,0), \tag{4.8}$$

   where $\bar{\lambda} \leq 0$.

2. *$\bar{u}$ is the solution of (P_α^+) if and only if*

$$\operatorname{supp}(\bar{u}) \subset \{x \in \bar{\Omega} : \bar{\varphi}(x,0) = \bar{\lambda}\}. \tag{4.9}$$

For completeness we give the proof here.

Proof. Existence and uniqueness of solutions is proven analogously to the proof of Theorem 4.2. Furthermore, as in the proof of Theorem 4.4 we get that $\bar{u} \in U_\alpha^+$ solves (P_α^+) if and only if

$$J'(\bar{u})(u-\bar{u}) = \int_{\bar{\Omega}} \bar{\varphi}(0)\, d(u-\bar{u}) \geq 0 \qquad \forall u \in U_\alpha^+,$$

which is equivalent to (4.7).
For the remainder of the proof we assume $\bar{u}(\bar{\Omega}) = \alpha$.
First, we prove part 1 of the claim. It holds

$$\alpha\bar{\lambda} = \int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} \leq \int_{\bar{\Omega}} \bar{\varphi}(0)\, du \qquad \forall u \in U_\alpha^+,$$

in particular for $u = 0$, so

$$\alpha\bar{\lambda} \leq 0 \quad \overset{\alpha>0}{\Longrightarrow} \quad \bar{\lambda} \leq 0.$$

Furthermore, (4.7) is equivalent to

$$\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} = \min_{u\in U_\alpha^+} \int_{\bar{\Omega}} \bar{\varphi}(0)\, du.$$

Now, take $x_0 \in \bar{\Omega}$, such that

$$\bar{\varphi}(x_0,0) = \bar{\lambda} := \min_{x\in\bar{\Omega}} \bar{\varphi}(x,0).$$

Then $\operatorname{argmin}_{u\in U_\alpha^+} \int_{\bar{\Omega}} \bar{\varphi}(0)\, du = \alpha\delta_{x_0}$, which in combination delivers

$$\begin{aligned}
\alpha\bar{\lambda} &= \int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} \\
&= \min_{u\in U_\alpha^+} \int_{\bar{\Omega}} \bar{\varphi}(0)\, du \\
&= \int_{\bar{\Omega}} \bar{\varphi}(0)\, d(\alpha\delta_{x_0}) \\
&= \alpha\bar{\varphi}(x_0,0) \\
&= \alpha \min_{x\in\bar{\Omega}} \bar{\varphi}(x,0).
\end{aligned}$$

Since the inverse implication is obvious, the proof of part 1 of the claim is herewith complete. We move on to proving part 2 of the claim and distinguish two cases to this end.

1^{st} *case:* $\bar{\lambda} = 0$
Then we have $0 = \min_{x\in\bar{\Omega}} \bar{\varphi}(x, 0)$, so $0 \leq \bar{\varphi}(x, 0)$ for all $x \in \bar{\Omega}$. From (4.8) we have

$$\int_{\bar{\Omega}} \bar{\varphi}(0)\, du = 0 \qquad \text{and thus} \qquad \operatorname{supp}(\bar{u}) \subset \left\{x \in \bar{\Omega} : \bar{\varphi}(x, 0) = 0\right\} = \left\{x \in \bar{\Omega} : \bar{\varphi}(x, 0) = \bar{\lambda}\right\}.$$

2^{nd} *case:* $\bar{\lambda} < 0$
We define $\psi(x) := -\min\{\bar{\varphi}(x, 0), 0\}$, which fulfills

$$0 \leq \psi(x) \leq -\bar{\lambda} \qquad \text{and} \qquad \|\psi\|_{C(\bar{\Omega})} = -\bar{\lambda}.$$

Now from (4.7) we know

$$-\int_{\bar{\Omega}} \bar{\varphi}(0)\, du \leq -\int_{\bar{\Omega}} \bar{\varphi}(0)\, d\bar{u} \leq \int_{\bar{\Omega}} \psi\, d\bar{u} \qquad \forall u \in U_\alpha^+.$$

In particular, if we insert $u = \alpha\delta_{x_0}$, where $\bar{\varphi}(x_0, 0) = \bar{\lambda}$, we get

$$-\int_{\bar{\Omega}} \bar{\varphi}(0)\, d(\alpha\delta_{x_0}) = -\alpha\bar{\varphi}(x_0, 0) = -\alpha\bar{\lambda} \leq \int_{\bar{\Omega}} \psi\, d\bar{u}.$$

Due to $\|\psi\|_{C(\bar{\Omega})} = -\bar{\lambda}$ and $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$ this gives

$$\|\psi\|_{C(\bar{\Omega})}\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \leq \int_{\bar{\Omega}} \psi\, d\bar{u}.$$

The converse inequality is well-known, so we have the equality

$$\|\psi\|_{C(\bar{\Omega})}\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \int_{\bar{\Omega}} \psi\, d\bar{u}.$$

Now, (4.9) can be proven analogous to Lemma 3.8.

It remains to see the converse implication. We have the non-negative measure $\bar{u}$ with $\bar{u}(\bar{\Omega}) = \alpha$, so (4.7) is a direct consequence of (4.9). □

We also give the following remark combined from [21, Remark 3.2 and Remark 3.3]:

Remark 4.8. *While in Theorem 4.4, we have $\bar{y}(T) = y_d$ and $\bar{\varphi} = 0 \in Q$ for an optimal control $\bar{u}$ with $\bar{u}(\bar{\Omega}) < \alpha$, this case is not a part of Theorem 4.7.*

We denote by y_0 the solution of (4.1) *corresponding to the control $u = 0$ and by z_u the solution of* (4.5)*. It holds $y_u(T) = z_u(T) + y_0(T)$ and by weak maximum principle $z_u \geq 0 \in Q$ for non-negative control u. Let $y_d \leq y_0(T)$, then we have for any $u \neq 0$*

$$J(0) = \frac{1}{2}\|y_0(T) - y_d\|^2_{L^2(\Omega)} < \frac{1}{2}\|z_u(T) - (y_d - y_0(T))\|^2_{L^2(\Omega)} = \frac{1}{2}\|y_u(T) - y_d\|^2_{L^2(\Omega)} = J(u).$$

Thus, the unique solution to (P_α^+) is given by $\bar{u} = 0$.

So even though $\bar{u}(\bar{\Omega}) = 0 < \alpha$, there exist cases with $\bar{y}(T) \neq y_d$ and consequently $\bar{\varphi} \neq 0 \in Q$.

4.3 Variational discretization

To discretize problems (P_α), (P_α^+) we define the space-time grid as follows: Define the partition $0 = t_0 < t_1 < \ldots < t_{N_\tau} = T$. For the temporal grid the interval I is split into subintervals $I_k = (t_{k-1}, t_k]$ for $k = 1, \ldots, N_\tau$. The temporal gridsize is denoted by $\tau = \max_{0\le k\le N_\tau} \tau_k$, where $\tau_k := t_k - t_{k-1}$. We assume that $\{I_k\}_k$ and $\{\mathcal{K}_h\}_h$ are quasi-uniform sequences of time grids and triangulations, respectively. For $K \in \mathcal{K}_h$ we denote by $\rho(K)$ the diameter of K, and $h := \max_{K\in\mathcal{K}_h} \rho(K)$. We set $\bar\Omega_h = \bigcup_{K\in\mathcal{K}_h} K$ and denote by Ω_h the interior and by Γ_h the boundary of $\bar\Omega_h$. We assume Ω to be polyhedral and that vertices on Γ_h are points on Γ. We then set up the space-time grid as $Q_h := \Omega_h \times (0, T)$.

We define the discrete spaces:

$$Y_h := \operatorname{span}\{e_{x_j} : 1 \le j \le N_h\}, \tag{4.10}$$

$$Y_\sigma := \operatorname{span}\{e_{x_j} \otimes \chi_k : 1 \le j \le N_h, 1 \le k \le N_\tau\}, \tag{4.11}$$

where χ_k is the indicator function of I_k and $\left(e_{x_j}\right)_{j=1}^{N_h}$ is the nodal basis formed by continuous piecewise linear functions satisfying $e_{x_j}(x_i) = \delta_{ij}$.

We choose the space Y_σ as our discrete state and test space in a dG(0) approximation of (4.1). The control space remains either U_α or U_α^+. This approximation scheme is equivalent to an implicit Euler time stepping scheme. To see this we recall that the elements $y_\sigma \in Y_\sigma$ can be represented as

$$y_\sigma = \textstyle\sum_{k=1}^{N_\tau} y_{k,h} \otimes \chi_k,$$

with $y_{k,h} := y_\sigma|_{I_k} \in Y_h$. Given a control $u \in U_\alpha$ for $k = 1, \ldots, N_\tau$ and all $z_h \in Y_h$ we thus end up with the variational discrete scheme

$$\begin{cases} (y_{k,h} - y_{k-1,h}, z_h)_{L^2} + a\,\tau_k \int_\Omega \nabla y_{k,h} \nabla z_h \, dx \\ \quad + \int_{I_k}\int_\Omega b(x,t)\nabla y_{k,h}\, z_h \; + c(x,t) y_{k,h}\, z_h \, dx\, dt = \int_{I_k}\int_\Omega f\, z_h \, dx\, dt, \\ y_{0,h} = y_{0h}, \end{cases} \tag{4.12}$$

where $y_{0h} \in Y_h$ is the unique element satisfying:

$$(y_{0h}, z_h)_{L^2} = \int_\Omega z_h \, du \qquad \forall\, z_h \in Y_h. \tag{4.13}$$

Here $(\cdot,\cdot)_{L^2}$ denotes the $L^2(\Omega)$ inner product. We assume that the discretization parameters h and τ are sufficiently small, such that there exists a unique solution to (4.12) for general functions b and c.

The variational discrete counterparts to (P_α) and (P_α^+) now read

$$\min_{u\in U_\alpha} J_\sigma(u) = \frac{1}{2}\|y_{u,\sigma}(T) - y_d\|^2_{L^2(\Omega_h)}, \tag{$P_{\alpha,\sigma}$}$$

and

$$\min_{u\in U_\alpha^+} J_\sigma(u) = \frac{1}{2}\|y_{u,\sigma}(T) - y_d\|^2_{L^2(\Omega_h)}, \tag{$P^+_{\alpha,\sigma}$}$$

respectively, where in both cases $y_{u,\sigma}$ for given u denotes the unique solution of (4.12). It is now straightforward to show that the optimality conditions for the problems $(P_{\alpha,\sigma})$ and $(P^+_{\alpha,\sigma})$ read like those for (P_α) and (P_α^+) with the adjoint φ replaced by $\varphi_{\bar u,\sigma} \in Y_h$ for given solution $\bar u$, the solution to the following system for $k = 1, \ldots, N_\tau$:

$$\begin{cases} -(\varphi_{k,h} - \varphi_{k-1,h}, z_h)_{L^2} + a\,\tau_k \int_\Omega \nabla \varphi_{k-1,h} \nabla z_h \, dx \\ \quad + \int_{I_k}\int_\Omega -\operatorname{div}(b(x,t)\varphi_{k-1,h})\, z_h \; + c(x,t)\varphi_{k-1,h}\, z_h \, dx\, dt = 0, \\ \varphi_{N_\tau,h} = \varphi_{N_\tau h}, \end{cases} \tag{4.14}$$

where $z_h \in Y_h$ and $\varphi_{N_\tau h} \in Y_h$ is the unique element satisfying:

$$(\varphi_{N_\tau h}, z_h)_{L^2} = \int_\Omega (y_{\bar{u},\sigma}(T) - y_d) z_h \, dx \qquad \forall\, z_h \in Y_h. \tag{4.15}$$

For details on the derivation of the optimality conditions we refer to Theorem 4.15 and Theorem 4.16, which will be proven after introducing a few helpful results.
This in particular implies that

$$\begin{aligned} \operatorname{supp}(\bar{u}^+) &\subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = -\|\varphi_{\bar{u},\sigma}(0)\|_\infty\}, \\ \operatorname{supp}(\bar{u}^-) &\subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = +\|\varphi_{\bar{u},\sigma}(0)\|_\infty\}. \end{aligned} \tag{4.16}$$

Analogously for $(P^+_{\alpha,\sigma})$, in the case $u(\bar{\Omega}) = \alpha$ we have the optimality condition

$$\operatorname{supp}(\bar{u}) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = \min_{x \in \bar{\Omega}} \varphi_{\bar{u},\sigma}(x,0)\}.$$

Since, in both cases, $\varphi_{\bar{u},\sigma}$ is a piecewise linear and continuous function, the extremal value in the generic case can only be attained at grid points, which leads to

$$\operatorname{supp}(\bar{u}) \subset \{x_j\}_{j=1}^{N_h}.$$

So, we derive the implicit discrete structure:

$$\bar{u} \in U_h \coloneqq \operatorname{span}\{\delta_{x_j} : 1 \le j \le N_h\},$$

where δ_{x_j} denotes a Dirac measure at gridpoint x_j. In the case of $(P^+_{\alpha,\sigma})$ we even know that all coefficients will be positive and hence we get

$$\bar{u} \in U_h^+ \coloneqq \left\{ \textstyle\sum_{j=1}^{N_h} u_j \delta_{x_j} : u_j \ge 0 \right\}.$$

Notice also that the natural pairing $\mathcal{M}(\bar{\Omega}) \times C(\bar{\Omega}) \to \mathbb{R}$ induces a pairing between Y_h and U_h in the discrete setting. Here we see the effect of the variational discretization concept: The choice for the discretization of the test space induces a natural discretization for the controls.

We note that the use of piecewise linear and continuous Ansatz- and test-functions in the variational discretization creates a setting, where the optimal control is supported on space grid points. However, it is possible to use piecewise quadratic and continuous Ansatz- and test-functions, so that the discrete adjoint variable can attain its extremal values not only on grid points, but anywhere. Calculating the location of these extremal values, then, would mean to determine the potential support of the optimal control - not limited to grid points anymore.

The following operator will be useful for the discussion of solutions to $(P_{\alpha,\sigma})$.

Lemma 4.9. *Let the linear operator Υ_h be defined as below:*

$$\Upsilon_h : \mathcal{M}(\bar{\Omega}) \to U_h \subset \mathcal{M}(\bar{\Omega}), \qquad \Upsilon_h u \coloneqq \sum_{j=1}^{N_h} \delta_{x_j} \int_\Omega e_{x_j} \, du.$$

Then for every $u \in \mathcal{M}(\bar{\Omega})$ and $\varphi_h \in Y_h$ the following properties hold.

$$\langle u, \varphi_h \rangle = \langle \Upsilon_h u, \varphi_h \rangle, \tag{4.17}$$

$$\|\Upsilon_h u\|_{\mathcal{M}(\bar{\Omega})} \le \|u\|_{\mathcal{M}(\bar{\Omega})}. \tag{4.18}$$

These results have been proven in [20, Proposition 4.1.]. Furthermore, it is obvious, for piecewise linear and continuous finite elements, that $\Upsilon_h(\mathcal{M}^+(\bar{\Omega})) \subset U_h^+$.

The mapping $u \mapsto y_{u,\sigma}(T)$ is in general not injective, hence the uniqueness of the solution cannot be concluded.

In the implicitly discrete setting however, we can prove uniqueness similarly as done in [14, Section 4.3.] and Theorem 3.12.

Theorem 4.10. *The problem $(P_{\alpha,\sigma})$ has at least one solution in $\mathcal{M}(\bar{\Omega})$ and there exists a unique solution $\bar{u} \in U_h$. Furthermore, for every solution $\hat{u} \in \mathcal{M}(\bar{\Omega})$ of $(P_{\alpha,\sigma})$ it holds $\Upsilon_h \hat{u} = \bar{u}$. Moreover, if $\bar{\varphi}_h(x_j) \neq \bar{\varphi}_h(x_k)$ for all neighboring finite element nodes $x_j \neq x_k$ of the finite element nodes $x_j (j = 1, \dots, N_h)$, problem $(P_{\alpha,\sigma})$ admits a unique solution, which is an element of U_h.*

Proof. The existence of solutions can be derived as for the continuous problem, see [21, Theorem 2.4.], since the control domain remains continuous. We include the details for the convenience of the reader.

The control domain U_α is bounded and weakly-* closed in $\mathcal{M}(\bar{\Omega})$. From Banach-Alaoglu-Bourbaki theorem we even know that it is weakly-* compact, see e.g. [10, Theorem 3.16.]. Hence, any minimizing sequence is bounded in $\mathcal{M}(\bar{\Omega})$ and any weak-* limit belongs to the control domain U_α. Using convergence properties from [21, Theorem 2.3.] we can conclude that any of these limits is a solution to $(P_{\alpha,\sigma})$.

Let $\hat{u} \in \mathcal{M}(\bar{\Omega})$ be a solution of $(P_{\alpha,\sigma})$ and $\bar{u} := \Upsilon_h \hat{u} \in U_h$. From (4.17) we have

$$y_{u,\sigma} = y_{\Upsilon_h u,\sigma} \quad \text{for all } u \in \mathcal{M}(\bar{\Omega}).$$

From this we deduce $J_\sigma(\bar{u}) = J_\sigma(\hat{u})$. Moreover (4.18) delivers

$$\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \leq \|\hat{u}\|_{\mathcal{M}(\bar{\Omega})},$$

so $\bar{u}$ is admissible, since $\hat{u} \in U_\alpha$. Altogether, this shows the existence of solutions in the discrete space U_h.

Since the mapping $u \mapsto y_{u,\sigma}(T)$ is injective for $u \in U_h$ - since we have $\dim(U_h) = \dim(Y_h)$, and $J_\sigma(u)$ is a quadratic function, we deduce strict convexity of $J_\sigma(u)$ on U_h. Furthermore $\left\{u \in U_h : \|u\|_{\mathcal{M}(\bar{\Omega})} = \sum_{j=1}^{N_h} |u_j| \leq \alpha\right\}$ is a closed and convex set, so we can conclude the uniqueness of the solution in the discrete space.

For every solution $\hat{u} \in \mathcal{M}(\bar{\Omega})$ of $(P_{\alpha,\sigma})$, the projection $\Upsilon_h \hat{u}$ is a discrete solution. Moreover, there exists only one discrete solution. So we deduce that all projections must coincide.

If now $\bar{\varphi}_h(x_j) \neq \bar{\varphi}_h(x_k)$ for all neighbors $k \neq j$, every solution u of $(P_{\alpha,\sigma})$ has its support in some of the finite element nodes of the triangulation, or vanish identically, and thus is an element of U_h. This shows the unique solvability of $(P_{\alpha,\sigma})$ in this case. □

Remark 4.11. *We note that the condition on the values of $\bar{\varphi}_h$ in the finite element nodes for guaranteeing uniqueness can be checked once the discrete adjoint solution is known. This condition is thus fully practical.*

For $(P^+_{\alpha,\sigma})$ we have a similar result like Theorem 4.10, which we state without proof, since it can be interpreted as a special case of Theorem 4.10 and can be proven analogously.

Theorem 4.12. *The problem $(P^+_{\alpha,\sigma})$ has at least one solution in $\mathcal{M}^+(\bar{\Omega})$ and there exists a unique solution $\bar{u} \in U^+_h$. Furthermore, for every solution $\hat{u} \in \mathcal{M}^+(\bar{\Omega})$ of $(P^+_{\alpha,\sigma})$ it holds $\Upsilon_h \hat{u} = \bar{u}$. Moreover, if $\bar{\varphi}_h(x_j) \neq \bar{\varphi}_h(x_k)$ for all neighboring finite element nodes $x_j \neq x_k$ of the finite element nodes $x_j (j = 1, \dots, N_h)$, problem $(P^+_{\alpha,\sigma})$ admits a unique solution, which is an element of U^+_h.*

Now, we introduce two useful lemmas.

Lemma 4.13. *Given $u \in \mathcal{M}(\bar{\Omega})$, the solution $z_{u,\sigma} \in Y_h$ to (4.12) with $f \equiv 0$ satisfies*

$$\int_\Omega (y_{u,\sigma}(T) - y_d) z_{u,\sigma}(T)\, dx = \int_\Omega \varphi_{u,\sigma}(0)\, du. \tag{4.19}$$

Proof. We take (4.12) with $f \equiv 0$ and test with $\varphi_{k,h}$, the components of $\varphi_{u,\sigma}$, for all $k = 1, \dots, N_\tau$. Similarly we take (4.14) and test this with $z_{k-1,h}$, the components of $z_{u,\sigma}$, for all $k = 1, \dots, N_\tau$. Now we can sum up the equations,

and since in both cases the right hand side is zero, we can equalize those sums. Furthermore, we can apply Gauß' theorem and drop all terms that appear on both sides. This leads to

$$\begin{aligned}
0 &= \sum_{k=1}^{N_\tau}(z_{k,h} - z_{k-1,h}, \varphi_{k,h}) - \sum_{k=1}^{N_\tau}(-\varphi_{k,h} + \varphi_{k-1,h}, z_{k-1,h}),\\
&= \sum_{k=1}^{N_\tau}(z_{k,h}, \varphi_{k,h}) - (z_{k-1,h}, \varphi_{k,h}) + \sum_{k=1}^{N_\tau}(\varphi_{k,h}, z_{k-1,h}) - (\varphi_{k-1,h}, z_{k-1,h}),\\
&= \sum_{k=1}^{N_\tau}(z_{k,h}, \varphi_{k,h}) - \sum_{k=0}^{N_\tau-1}(\varphi_{k,h}, z_{k,h}),\\
&= (z_{N_\tau,h}, \varphi_{N_\tau,h}) - (\varphi_{0,h}, z_{0,h}).
\end{aligned}$$

We have $z_{N_\tau,h} = z_{u,\sigma}(T) \in Y_h$ and $\varphi_{0,h} = \varphi_{u,\sigma}(0) \in Y_h$, so together with (4.13) and (4.15) we can deduce (4.19). □

Lemma 4.14. *For every $\epsilon > 0$ and h small enough, there exists a control $u \in L^2(\Omega)$, such that the solution $y_{u,\sigma}$ of* (4.12) *fulfills*

$$\|y_{u,\sigma}(T) - y_d\|_{L^2(\Omega_h)} < \epsilon. \tag{4.20}$$

Proof. Let $y_{d,\sigma}$ be the L^2-projection of y_d onto Y_h, and h small enough, such that

$$\|y_{u,\sigma}(T) - y_d\|_{L^2(\Omega_h)} \le \|y_{u,\sigma}(T) - y_{d,\sigma}\|_{L^2(\Omega_h)} + \underbrace{\|y_{d,\sigma} - y_d\|_{L^2(\Omega_h)}}_{<\epsilon}.$$

Let additionally τ be small enough, such that the scheme (4.12) has a unique solution. Then in every time-step we obtain a system of equations, where the matrix is an isomorphism on Y_h. Consequently the initial to final value map $y_{0h} \mapsto y_{u,\sigma}(T)$ is an isomorphism. Since $Y_h \subset L^2(\Omega_h)$ we can find $u \in L^2(\Omega)$, such that

$$\|y_{u,\sigma}(T) - y_{d,\sigma}\|_{L^2(\Omega_h)} = 0,$$

which completes the proof. □

Finally, we give the discrete version of Theorem 4.4 and Theorem 4.7. Both are proven very similarly to the continuous case.

Theorem 4.15. *Let $\bar{u}$ solve $(P_{\alpha,\sigma})$ with $y_{\bar{u},\sigma}$ and $\varphi_{\bar{u},\sigma}$ the associated discrete state and discrete adjoint state, respectively. Then for σ small enough,*

1. *if $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$, then $y_{\bar{u},\sigma}(T) = y_d$ and $\varphi_{\bar{u},\sigma} = 0 \in Q$.*
2. *if $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$, then*

$$\operatorname{supp}(\bar{u}^+) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = -\|\varphi_{\bar{u},\sigma}(0)\|_{C(\bar{\Omega})}\}, \tag{4.21}$$

$$\operatorname{supp}(\bar{u}^-) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = +\|\varphi_{\bar{u},\sigma}(0)\|_{C(\bar{\Omega})}\}, \tag{4.22}$$

where $\bar{u} = \bar{u}^+ - \bar{u}^-$ is the Jordan decomposition of $\bar{u}$.

Conversely, if $\bar{u}$ is an element of U_α satisfying 1. or 2., then $\bar{u}$ is the solution to $(P_{\alpha,\sigma})$.

Proof. Let $u \in U_\alpha$ arbitrary and denote by $z_{(u-\bar{u}),\sigma}$ the solution to (4.12) with $f \equiv 0$ and u replaced by $u - \bar{u}$. From Lemma 4.13 we get

$$\lim_{\rho \searrow 0} \frac{J_\sigma(\bar{u} + \rho(u - \bar{u})) - J_\sigma(\bar{u})}{\rho} = \int_\Omega (y_{\bar{u},\sigma}(T) - y_d) z_{(u-\bar{u}),\sigma}(T)\, dx = \int_\Omega \varphi_{\bar{u},\sigma}(0)\, d(u - \bar{u}).$$

Since $(P_{\alpha,\sigma})$ is a convex problem, the following variational inequality is a necessary and sufficient condition for optimality of a control $\bar{u} \in U_\alpha$:

$$\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d(u-\bar{u}) = J'_\sigma(\bar{u})(u-\bar{u}) \geq 0 \qquad \forall u \in U_\alpha.$$

By taking the supremum over $u \in U_\alpha$ in the above inequality, we deduce

$$\alpha \, \|\varphi_{\bar{u},\sigma}(0)\|_{C(\bar{\Omega})} = \sup_{u\in U_\alpha} \int_\Omega \varphi_{\bar{u},\sigma}(0)\, du = -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d\bar{u}.$$

Let $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} = \alpha$, then this is

$$\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} \|\varphi_{\bar{u},\sigma}(0)\|_{C(\bar{\Omega})} = -\int_\Omega \varphi_{\bar{u},\sigma}(0)\, d\bar{u}. \tag{4.23}$$

We now may conclude as in Lemma 3.8 to obtain (4.21) and (4.22). Also, if these conditions hold we get the equality (4.23), which is a necessary and sufficient condition for optimality of $\bar{u}$, so $\bar{u}$ solves $(P_{\alpha,\sigma})$.

Let us now study the case $\|\bar{u}\|_{\mathcal{M}(\bar{\Omega})} < \alpha$. If $y_{\bar{u},\sigma} = y_d$, then $J_\sigma(\bar{u}) = 0$ and since $J_\sigma(u) \geq 0$ for all $u \in U_\alpha$, we deduce that $\bar{u}$ solves $(P^+_{\alpha,\sigma})$. Now assume that $\bar{u}$ solves $(P^+_{\alpha,\sigma})$ and $y_{\bar{u},\sigma} \neq y_d$ holds. Then we have $J_\sigma(\bar{u}) > 0$. From Lemma 4.14 we know that for h small enough there exists an element $u \in \mathcal{M}(\bar{\Omega})$, such that $J_\sigma(u) < J_\sigma(\bar{u})$. Since $\bar{u}$ is a solution to $(P^+_{\alpha,\sigma})$, it must hold $u \notin U_\alpha$. Now take $\lambda \in \mathbb{R}$, such that

$$0 < \lambda < \min\left\{ \frac{\alpha - \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})}}{\|u-\bar{u}\|_{\mathcal{M}(\bar{\Omega})}}, 1 \right\}. \tag{4.24}$$

Then $v := \bar{u} + \lambda(u-\bar{u}) \in U_\alpha$ and by convexity of J_σ we get

$$J_\sigma(v) = J_\sigma(\lambda u + (1-\lambda)\bar{u}) \leq \lambda \underbrace{J_\sigma(u)}_{<J_\sigma(\bar{u})} + (1-\lambda) J_\sigma(\bar{u}) < J_\sigma(\bar{u}),$$

so that $\bar{u} \in U_\alpha$ can not be the solution of $(P^+_{\alpha,\sigma})$. Hence $y_{\bar{u},\sigma} = y_d$ must hold and from (4.15) we deduce $\varphi_{\bar{u},\sigma} = 0$. □

Theorem 4.16. *Let $\bar{u}$ solve $(P^+_{\alpha,\sigma})$ with associated discrete adjoint state $\varphi_{\bar{u},\sigma}$. Then, $\bar{u}$ is a solution of $(P^+_{\alpha,\sigma})$ if and only if*

$$\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} \leq \int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, du \qquad \forall\, u \in U^+_\alpha. \tag{4.25}$$

If $\bar{u}(\bar{\Omega}) = \alpha$ the following properties are fulfilled:

1. *Inequality* (4.25) *is equivalent to the identity*

$$\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} = \alpha\bar{\lambda} := \alpha \min_{x\in\bar{\Omega}} \varphi_{\bar{u},\sigma}(x,0), \tag{4.26}$$

   where $\bar{\lambda} \leq 0$.

2. *$\bar{u}$ is the solution of (P^+_α) if and only if*

$$\operatorname{supp}(\bar{u}) \subset \{x \in \bar{\Omega} : \varphi_{\bar{u},\sigma}(x,0) = \bar{\lambda}\}. \tag{4.27}$$

Proof. As in the proof of Theorem 4.15, we get that $\bar{u} \in U^+_\alpha$ solves $(P^+_{\alpha,\sigma})$, if and only if

$$J'_\sigma(\bar{u})(u-\bar{u}) = \int_\Omega \varphi_{\bar{u},\sigma}(0)\, d(u-\bar{u}) \geq 0 \qquad \forall u \in U^+_\alpha,$$

which is equivalent to the condition (4.25).

Now let $\bar{u}(\bar{\Omega}) = \alpha$. If $\bar{\lambda} = \min_{x\in\bar{\Omega}} \varphi_{\bar{u},\sigma}(x,0) > 0$, then take $u = 0 \in U_\alpha^+$ in (4.25) to see that in this case $\bar{u} = 0$ must hold. So we must have $\bar{\lambda} \leq 0$. Furthermore, we can equivalently write (4.25) as

$$\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} = \min_{u\in U_\alpha^+} \int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, du.$$

Take $x_0 \in \bar{\Omega}$, such that $\varphi_{\bar{u},\sigma}(x_0,0) = \bar{\lambda}$. Then $u = \alpha\delta_{x_0}$ achieves the minimum in the equation above and we get (4.26). The other direction of the equivalence is obvious and completes the proof of part 1.

In order to prove part 2, we look at two cases. First, let $\bar{\lambda} = 0$. By definition of $\bar{\lambda}$ this implies that $\varphi_{\bar{u},\sigma} \geq 0$ for all $x \in \bar{\Omega}$. So with (4.26) we get that $\bar{u}$ has support, where $\varphi_{\bar{u},\sigma}(x,0) = 0 = \bar{\lambda}$, in order for the integral to be zero.

The second case is $\bar{\lambda} < 0$. Define $\psi(x) := -\min\{\varphi_{\bar{u},\sigma}(x,0), 0\}$, then it holds $0 \leq \psi(x) \leq -\bar{\lambda}$ by definition of $\psi(x)$ and $\bar{\lambda}$. Furthermore $\|\psi\|_{C(\bar{\Omega})} = -\bar{\lambda}$. With (4.25) and $\psi(x) \geq -\varphi_{\bar{u},\sigma}(x,0)$, we find

$$\int_\Omega \psi(x)\, d\bar{u} \geq -\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, d\bar{u} \geq -\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, du \qquad \forall u \in U_\alpha^+.$$

Especially for $u = \alpha\delta_{x_0}$, we have

$$\int_\Omega \psi(x)\, d\bar{u} \geq -\int_\Omega \varphi_{\bar{u},\sigma}(x,0)\, d(\alpha\delta_{x_0}) = -\alpha\bar{\lambda} = \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})}\|\psi\|_{C(\bar{\Omega})}.$$

Furthermore, we obviously have

$$\int_\Omega \psi(x)\, d\bar{u} \leq \|\bar{u}\|_{\mathcal{M}(\bar{\Omega})}\|\psi\|_{C(\bar{\Omega})},$$

so we can deduce equality and as in Lemma 3.8 we then get (4.27).

The converse implication can be seen, since for a positive control $\bar{u} \in U_\alpha^+$ with $\bar{u}(\bar{\Omega}) = \alpha$, we can conclude (4.25) from the condition (4.27). □

4.4 Computational results

For the implementation we consider $b \equiv 0$ and $c \equiv 0$ in (4.2).

We will consider the case of positive sources first, since the implementation is straightforward, while the general case requires to handle absolute values in the constraints.

4.4.1 Positive sources (problem $(P^+_{\alpha,\sigma})$)

We recall the discrete state equation (4.12), which reduces to the following form, since $b \equiv 0$ and $c \equiv 0$, with $z_h \in Y_h$:

$$\begin{cases} (y_{k,h} - y_{k-1,h}, z_h)_{L^2} + a\,\tau_k \int_\Omega \nabla y_{k,h} \nabla z_h\, dx = \int_{I_k}\int_\Omega f\, z_h\, dx\, dt, \\ y_{0,h} = y_{0h}, \end{cases}$$

where $y_{0h} \in Y_h$, for given $u \in \mathcal{M}(\bar{\Omega})$, is the unique element satisfying:

$$(y_{0h}, z_h) = \int_\Omega z_h\, du \qquad \forall\, z_h \in Y_h.$$

Let the mass matrix $M_h = \left(\left(e_{x_j}, e_{x_k}\right)_{L^2}\right)_{j,k=1}^{N_h}$ and the stiffness matrix $A_h = \left(\int_{\bar{\Omega}} \nabla e_{x_j} \nabla e_{x_k}\right)_{j,k=1}^{N_h}$ corresponding to Y_h. We also notice that the matrix $\left(e_{x_j}, \delta_{x_k}\right)_{j,k=1}^{N_h}$ is the identity in $\mathbb{R}^{N_h\times N_h}$.

We represent the discrete state equation by the following operator $L : \mathbb{R}^{N_\sigma} \to \mathbb{R}^{N_\sigma}$:

$$\underbrace{\begin{pmatrix} M_h & & & 0 \\ -M_h & M_h + a\,\tau_1 A_h & & \\ & \ddots & \ddots & \\ 0 & & -M_h & M_h + a\,\tau_{N_\tau} A_h \end{pmatrix}}_{=:L} \underbrace{\begin{pmatrix} y_{0,h} \\ y_{1,h} \\ \vdots \\ y_{N_\tau,h} \end{pmatrix}}_{=:y_h} = \begin{pmatrix} u \\ \tau_1 M_h f_{1,h} \\ \vdots \\ \tau_{N_\tau} M_h f_{N_\tau,h} \end{pmatrix}. \tag{4.28}$$

In order to get $y_{N_\tau,h}$ from the vector y_h, a restriction matrix that we call R is applied. Let $f \equiv 0$ for simplicity, then we have

$$y_{N_\tau,h}(u) = R \cdot L^{-1} \begin{pmatrix} u \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

This illustrates the ill-posedness of the final time control problem. For long time horizon T, especially in combination with a big diffusion constant a, we observe computational challenges and a big condition number for the solver of the discrete state equation.

We can now formulate the following finite-dimensional formulation of the discrete problem $(P^+_{\alpha,\sigma})$:

$$\begin{aligned} \min_{u\in\mathbb{R}^{N_h}} \; & J(u) = \frac{1}{2}\left(y_{N_\tau,h}(u) - y_d\right)^\top M_h \left(y_{N_\tau,h}(u) - y_d\right), \\ \text{s.t.} \quad & \sum_{i=1}^{N_h} u_i - \alpha \le 0, \\ & -u_i \le 0, \qquad \forall\, i \in \{1,\dots,N_h\}. \end{aligned} \tag{P^+_h}$$

The corresponding Lagrangian function $\mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})$ with $\lambda^{(1)} \in \mathbb{R}$, $\lambda^{(2)} \in \mathbb{R}^{N_h}$ is defined by

$$\mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) := J(u) + \lambda^{(1)} \left(\sum_{i=1}^{N_h} u_i - \alpha \right) - \sum_{i=1}^{N_h} \lambda_i^{(2)} u_i.$$

All inequalities in (P^+_h) are strictly fulfilled for $u_i = \frac{\alpha}{N_h+1}$ for all $i \in \{1,\dots,N_h\}$, thus an interior point of the feasible set exists, and the Slater condition is satisfied (see Definition 2.29). Then the Karush-Kuhn-Tucker conditions (see Theorem 2.30) state that at the minimum u the following conditions hold:

1. $\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) = 0$,
2. $\lambda^{(1)} \left(\sum_{i=1}^{N_h} u_i - \alpha \right) = 0 \;\wedge\; \lambda^{(1)} \ge 0 \;\wedge\; \left(\sum_{i=1}^{N_h} u_i - \alpha \right) \le 0$,
3. $-\lambda_i^{(2)} u_i = 0 \;\wedge\; \lambda_i^{(2)} \ge 0 \;\wedge\; -u_i \le 0 \quad \forall\, i \in \{1,\dots,N_h\}$,

where 2. and 3. can be equivalently reformulated with an arbitrary $\kappa > 0$ by

$$\begin{aligned} N^{(1)}(u, \lambda^{(1)}) &:= \max\left\{0, \lambda^{(1)} + \kappa\left(\textstyle\sum_{i=1}^{N_h} u_i - \alpha\right)\right\} - \lambda^{(1)} = 0, \\ N^{(2)}(u, \lambda^{(2)}) &:= \max\{0, \lambda^{(2)} - \kappa u\} - \lambda^{(2)} = 0. \end{aligned}$$

We define

$$F(u, \lambda^{(1)}, \lambda^{(2)}) := \left(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) \quad N^{(1)}(u, \lambda^{(1)}) \quad N^{(2)}(u, \lambda^{(2)})\right)^\top,$$

and apply the semismooth Newton method to solve $F(u, \lambda^{(1)}, \lambda^{(2)}) = 0$.

We have

$$\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) = \partial_u J(u) + \lambda^{(1)} \mathbb{1}_{N_h} - \lambda^{(2)},$$

where $\mathbb{1}_{N_h} = (1, \dots, 1)^\top \in \mathbb{R}^{N_h}$, while the identity matrix of size $N_h \times N_h$ will be denoted by $\mathbb{I}_{N_h}$.

When setting up the matrix $DF = DF(u, \lambda^{(1)}, \lambda^{(2)})$, we always choose $\partial_x(\max\{0, g(x)\}) = \partial_x g(x)$ if $g(x) = 0$. This delivers

$$\begin{aligned} DF &:= \begin{pmatrix} \partial^2_{uu}\mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) & \partial_{\lambda^{(1)}}(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})) & \partial_{\lambda^{(2)}}(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})) \\ \partial_u N^{(1)}(u, \lambda^{(1)}) & \partial_{\lambda^{(1)}} N^{(1)}(u, \lambda^{(1)}) & \partial_{\lambda^{(2)}} N^{(1)}(u, \lambda^{(1)}) \\ \partial_u N^{(2)}(u, \lambda^{(2)}) & \partial_{\lambda^{(1)}} N^{(2)}(u, \lambda^{(2)}) & \partial_{\lambda^{(2)}} N^{(2)}(u, \lambda^{(2)}) \end{pmatrix} \\ &= \begin{pmatrix} \partial^2_{uu}\mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) & \partial_{\lambda^{(1)}}(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})) & \partial_{\lambda^{(2)}}(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})) \\ \partial_u N^{(1)}(u, \lambda^{(1)}) & \partial_{\lambda^{(1)}} N^{(1)}(u, \lambda^{(1)}) & 0 \\ \partial_u N^{(2)}(u, \lambda^{(2)}) & 0 & \partial_{\lambda^{(2)}} N^{(2)}(u, \lambda^{(2)}) \end{pmatrix}, \end{aligned}$$

with the entries:

$$\begin{aligned} \partial^2_{uu}\mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)}) &= \partial^2_{uu} J(u), \\ \partial_{\lambda^{(1)}}(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})) &= \mathbb{1}_{N_h}, \\ \partial_{\lambda^{(2)}}(\partial_u \mathscr{L}(u, \lambda^{(1)}, \lambda^{(2)})) &= -\mathbb{I}_{N_h}, \\ \partial_u N^{(1)}(u, \lambda^{(1)}) &= \begin{cases} \kappa\, \mathbb{1}^\top_{N_h}, & \lambda^{(1)} + \kappa\left(\sum_{i=1}^{N_h} u_i - \alpha\right) \geq 0, \\ 0, & \text{else}, \end{cases} \\ \partial_{\lambda^{(1)}} N^{(1)}(u, \lambda^{(1)}) &= \begin{cases} 0, & \lambda^{(1)} + \kappa\left(\sum_{i=1}^{N_h} u_i - \alpha\right) \geq 0, \\ -1, & \text{else}, \end{cases} \\ \partial_{u_j} N^{(2)}_i(u, \lambda^{(2)}) &= \begin{cases} -\kappa\, \delta_{ij}, & \lambda^{(2)}_i - \kappa u_i \geq 0, \\ 0, & \text{else}, \end{cases} \\ \partial_{\lambda^{(2)}_j} N^{(2)}_i(u, \lambda^{(2)}) &= \begin{cases} 0, & \lambda^{(2)}_i - \kappa u_i \geq 0, \\ -\delta_{ij}, & \text{else}, \end{cases} \end{aligned}$$

Numerical example

Let $\Omega = [0,1]$, $T = 1$ and $a = \frac{1}{100}$. This combination of T and a is chosen to avoid too big a condition number of the discrete state equation. We are working on an equidistant 20×20 space-time grid for this example.

To generate a desired state y_d, we choose $u_{\text{true}} = \delta_{0.5}$ and $f \equiv 0$, solve the state equation on a very fine space-time grid (1000×1000) and take the evaluation of the result in $t = T$ on the current grid Ω_h as desired state y_d (see Figure 4.1). Another option is to sample the associated state $y_{u_{\text{true}}}$ from the analytic solution with spacial Fourier modes (for more details see Appendix A.3) and then take $y_d = y_{u_{\text{true}}}(T)$.

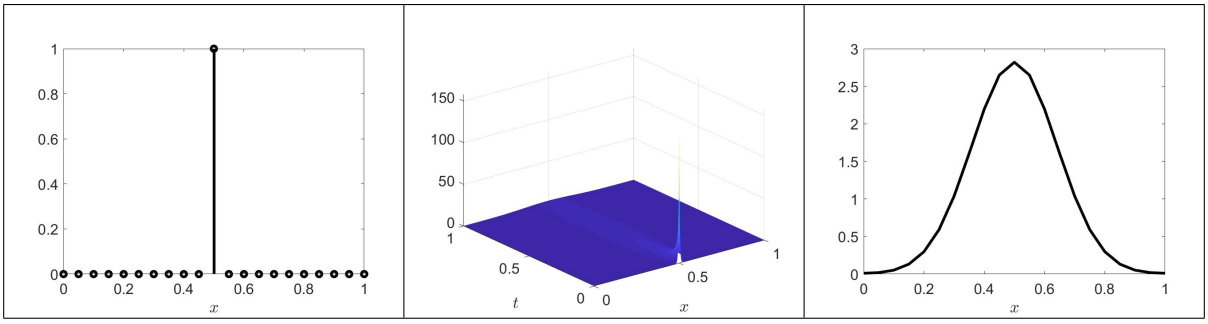

Figure 4.1: From left to right: true solution u_{true}, associated true state y_{true} in $Q = [0,1] \times [0,1]$ and desired state $y_d = y_{\text{true}}(T)$.

Either way, we can insert this y_d into our problem and solve for different values of α. Knowing the true solution u_{true}, we can compare our results to it. We also know $u_{\text{true}}(\Omega) = 1$ and $\text{supp}(u_{\text{true}}) = \{0.5\}$. We always start the algorithm with the control being identically zero and terminate when the residual is below 10^{-15}.

The first case we investigate is $\alpha = 0.1$ (see Figure 4.2). This α is smaller than the total variation of the true control and we observe $\bar{u}(\bar{\Omega}) = \alpha$. Furthermore $\bar{\lambda} = \min_{x \in \bar{\Omega}} \bar{\varphi}(0) \approx -35.859$ and we can verify the optimality conditions (4.26) and (4.27), since

$$\int_{\Omega} \bar{\varphi}_h(x,0)\, d\bar{u} \approx -3.5859 \approx \alpha\bar{\lambda},$$

and $\text{supp}(\bar{u}) = \{0.5\}$.

The second case we investigate is $\alpha = 1 = u_{\text{true}}(\bar{\Omega})$ (see Figure 4.3). The computed optimal control in this case has a total variation of $\bar{u}(\bar{\Omega}) = 1 = \alpha$ and we can again verify the sparsity $\text{supp}(\bar{u}) = \{0.5\}$. Furthermore $\bar{\lambda} = \min_{x \in \bar{\Omega}} \bar{\varphi}(0) \approx -0.0436$ and we can verify the optimality condition (4.26), since

$$\int_{\Omega} \bar{\varphi}_h(x,0)\, d\bar{u} \approx -0.0436 \approx \alpha\bar{\lambda}.$$

For cases with $\alpha > u_{\text{true}}(\bar{\Omega})$, we get similar results as in the case with $\alpha = 1$. In particular this means that we observe optimality conditions (4.26) and (4.27). Since we fixed $f \equiv 0$, we get $y_0(T) \equiv 0$ and therefore $y_d > y_0(T)$. Still, the properties that we found in the general case for $\bar{u}(\bar{\Omega}) < \alpha$: $\bar{y}(T) = y_d$ and $\varphi = 0 \in Q$ can not be observed (compare Figure 4.4 top). This is caused by the fact that the desired state y_d can not be reached on the coarse grid, so $\bar{y}(T) = y_d$ is not possible. Solving the problem with a desired state that has been projected onto the coarse grid, thus is reachable, delivers the expected properties $\bar{y}(T) = y_d$ and $\varphi = 0 \in Q$ (see Figure 4.4 bottom). For examples with $y_d \leq y_0(T)$ we can confirm Remark 4.8 and find the optimal solution $\bar{u} = 0$.

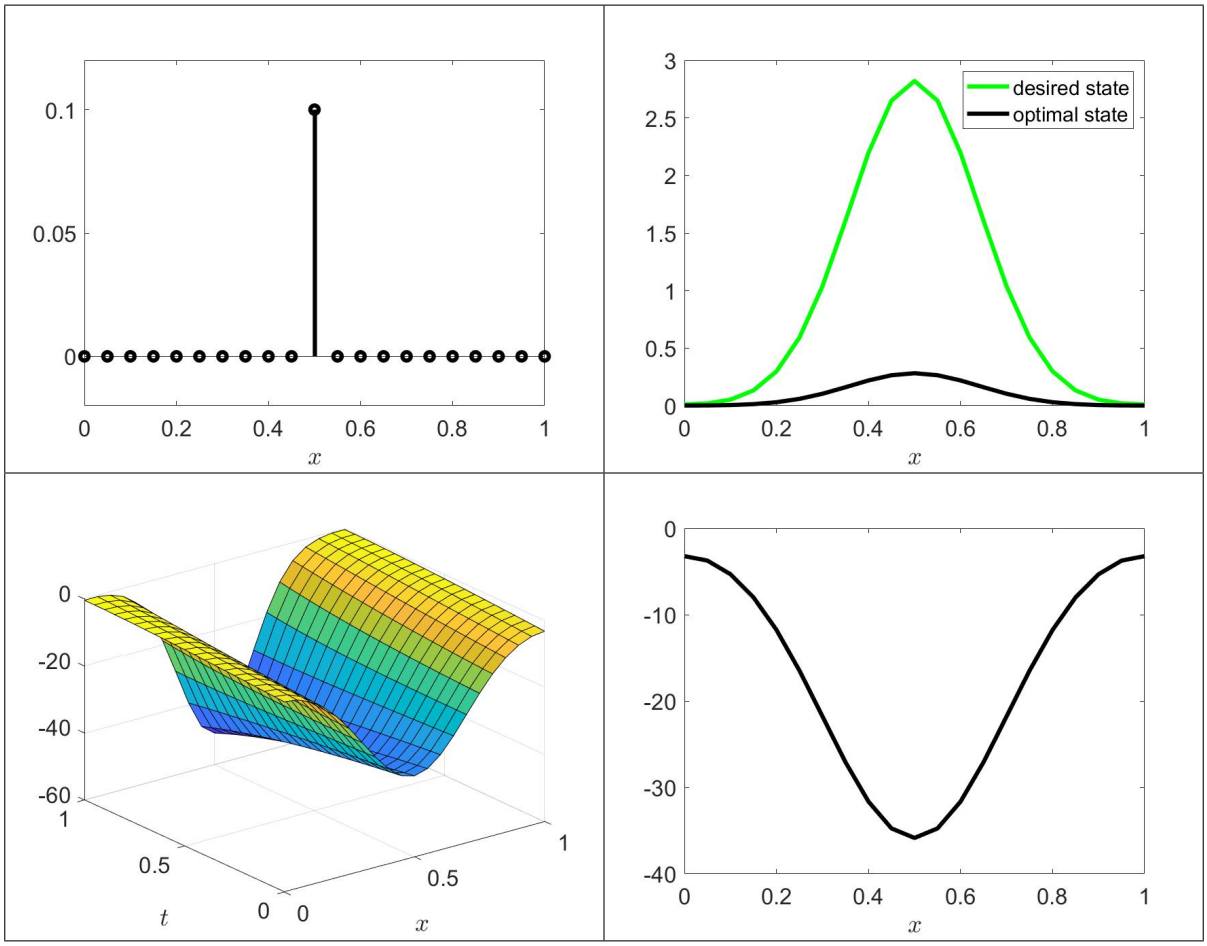

Figure 4.2: Solutions for $\alpha = 0.1$: from top left to bottom right: optimal control $\bar{u}$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 16 Newton steps.

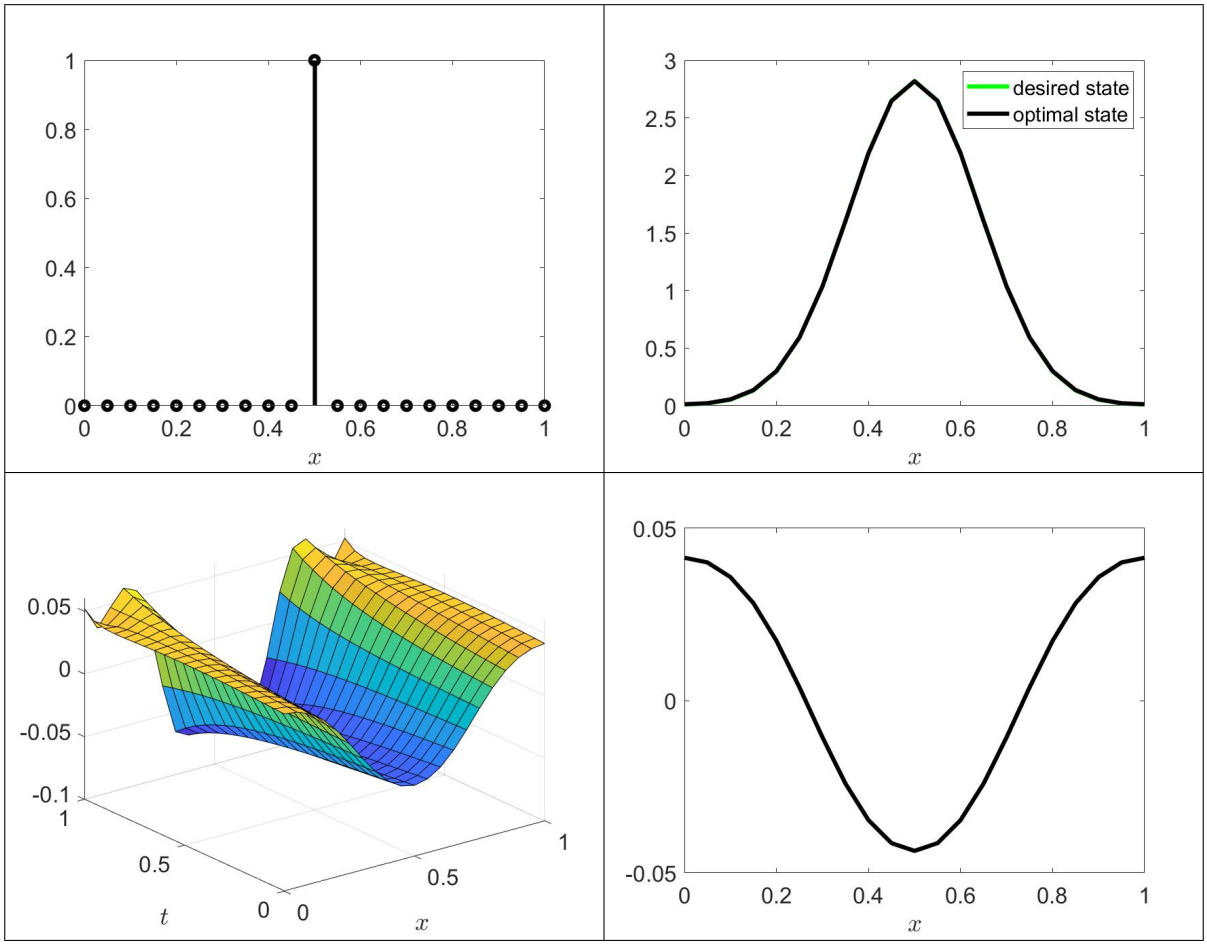

Figure 4.3: Solutions for $\alpha = 1$: from top left to bottom right: optimal control $\bar{u}$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, and associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 15 Newton steps.

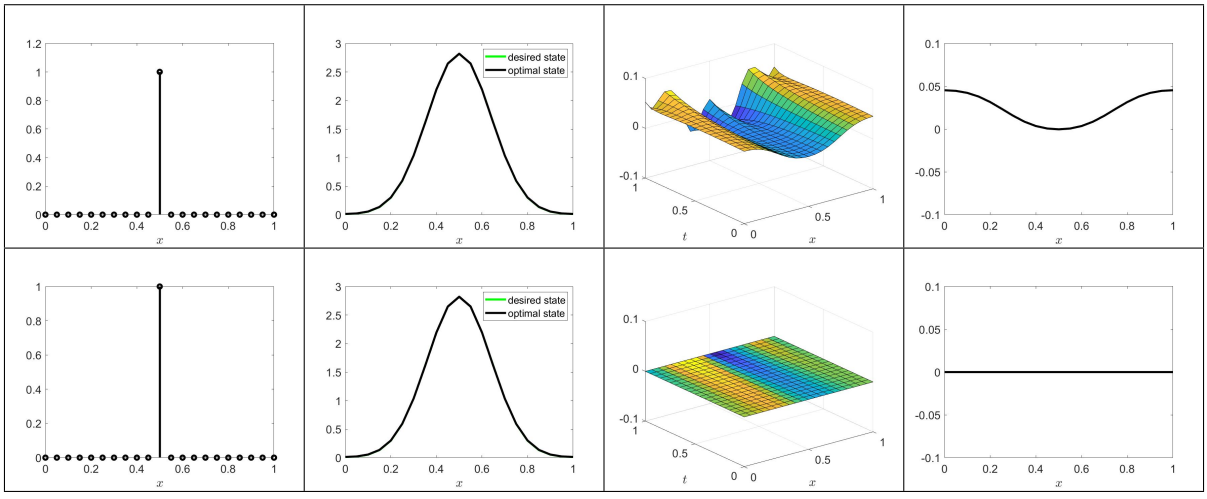

Figure 4.4: Solutions for $\alpha = 2$ with original desired state (top) and reachable desired state (bottom): from left to right: optimal control $\bar{u}$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 17 and 27 Newton steps, respectively.

4.4.2 The general case (problem $(P_{\alpha,\sigma})$)

Here, the source does not need to be positive. In the discrete problem we will decompose the control $u \in U_h$ into its positive and negative part, such that

$$u = u^+ - u^-, \qquad u^+ \geq 0, \quad u^- \geq 0.$$

We have the following finite-dimensional formulation of the discrete problem $(P_{\alpha,\sigma})$:

$$\begin{aligned} \min_{u^+,u^- \in \mathbb{R}^{N_h}} \; & J(u^+,u^-) = \frac{1}{2}\left(y_{N_\tau,h}(u^+,u^-) - y_d\right)^\top M_h \left(y_{N_\tau,h}(u^+,u^-) - y_d\right), \qquad (P_h) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{N_h} |u_i^+ - u_i^-| - \alpha \leq 0, \\ & -u_i^+ \leq 0 \quad \forall\, i, \\ & -u_i^- \leq 0 \quad \forall\, i, \end{aligned}$$

where $y_{N_\tau,h}(u^+,u^-)$ corresponds to solving (4.28) with $u = u^+ - u^-$ inserted into the right hand side of the equation. In order to allow taking second derivatives of the Lagrangian, we want to equivalently reformulate the absolute value in the first constraint. This can be done by adding the following constraint in our discrete problem:

$$u_i^+ u_i^- = 0 \quad \forall\, i. \tag{4.29}$$

and consequently the first constraint becomes

$$\left(\textstyle\sum_{i=1}^{N_h} u_i^+ + u_i^-\right) - \alpha \leq 0.$$

However, in the case $u_i^+ = u_i^- = 0$, the matrix in the Newton step will be singular. Since we want to handle sparse problems, this case will very likely occur, so we need to find a way to overcome this difficulty. Instead of adding an additional constraint, we could also add a penalty term that enforces $u_i^+ u_i^- = 0 \quad \forall\, i$ and consider the problem

$$\begin{aligned} \min_{u^+,u^- \in \mathbb{R}^{N_h}} \; & J(u^+,u^-) + \gamma (u^+)^\top u^-, \qquad (P_{h,\gamma}) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{N_h} u_i^+ + u_i^- - \alpha \leq 0, \\ & -u_i^+ \leq 0 \quad \forall\, i, \\ & -u_i^- \leq 0 \quad \forall\, i. \end{aligned}$$

For γ large enough the solutions of $(P_{h,\gamma})$ and (P_h) will coincide, i.e. the penalty function is exact. In [44, Theorem 4.6] and [73, Satz 18.5] it is specified that γ should be larger than the largest absolute value of the Karush-Kuhn-Tucker multipliers corresponding to the equality constraints (4.29), which are replaced. For clarification we formulate and prove this result in our problem setting:

Theorem 4.17. *Let $(\bar{u}^+, \bar{u}^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu})$ be a Karush-Kuhn-Tucker point of (P_h) with the additional constraint (4.29). With $\gamma \geq \max\{|\bar{\mu}_1|, \ldots, |\bar{\mu}_{N_h}|\}$ we get that $(\bar{u}^+, \bar{u}^-)$ is a global minimum of $(P_{h,\gamma})$.*

Proof. We define the Lagrangian $\mathscr{L}$ for problem (P_h) with constraint (4.29). Let $\lambda^{(1)} \in \mathbb{R}, \lambda^{(2)}, \lambda^{(3)} \in \mathbb{R}^{N_h}, \mu \in \mathbb{R}^{N_h}$, then:

$$\mathscr{L}(u^+,u^-,\lambda^{(1)},\lambda^{(2)},\lambda^{(3)},\mu) := J(u^+,u^-) + \underbrace{\lambda^{(1)}\left(\textstyle\sum_{i=1}^{N_h} u_i^+ - u_i^- - \alpha\right) - \sum_{i=1}^{N_h} \lambda_i^{(2)} u_i^+ - \sum_{i=1}^{N_h} \lambda_i^{(3)} u_i^-}_{=:\lambda^\top g} + \sum_{i=1}^{N_h} \mu_i u_i^+ u_i^-.$$

Now for all $(u^+, u^-) \in \mathbb{R}^{2N_h}$, since J is a convex C^1-function, we know

$$\begin{aligned}\mathscr{L}(u^+, u^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu}) \geq\ & \mathscr{L}(\bar{u}^+, \bar{u}^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu}) + \nabla_{u^+}\mathscr{L}(\bar{u}^+, \bar{u}^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu})^\top (u^+ - \bar{u}^+) \\ & + \nabla_{u^-}\mathscr{L}(\bar{u}^+, \bar{u}^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu})^\top (u^- - \bar{u}^-) \\ =\ & \mathscr{L}(\bar{u}^+, \bar{u}^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu}),\end{aligned}$$

by properties of Karush-Kuhn-Tucker points. So $(\bar{u}^+, \bar{u}^-)$ is a global minimizer of $\mathscr{L}(\cdot, \cdot, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu})$. Furthermore, we have $(\bar{u}^+)^\top \bar{u}^- = 0$, since $(\bar{u}^+, \bar{u}^-)$ fulfills constraint (4.29). This delivers

$$\begin{aligned}J(\bar{u}^+, \bar{u}^-) + \gamma(\bar{u}^+)^\top \bar{u}^- &= J(\bar{u}^+, \bar{u}^-) \\ &= J(\bar{u}^+, \bar{u}^-) + \sum_{i=1}^{N_h} \bar{\mu}_i \bar{u}_i^+ \bar{u}_i^- \\ &= \mathscr{L}(\bar{u}^+, \bar{u}^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu}) - (\bar{\lambda})^\top \bar{g} \\ &\leq \mathscr{L}(u^+, u^-, \bar{\lambda}^{(1)}, \bar{\lambda}^{(2)}, \bar{\lambda}^{(3)}, \bar{\mu}) - (\bar{\lambda})^\top \bar{g} \\ &= J(u^+, u^-) + \sum_{i=1}^{N_h} \bar{\mu}_i u_i^+ u_i^- \\ &\leq J(u^+, u^-) + \gamma(u^+)^\top u^-,\end{aligned}$$

where we used $\gamma \geq \max\{|\bar{\mu}_1|, \ldots, |\bar{\mu}_{N_h}|\}$. Thus $(\bar{u}^+, \bar{u}^-)$ is a global minimum of $(P_{h,\gamma})$. □

We will now work with $(P_{h,\gamma})$. We have the corresponding Lagrangian $\mathscr{L}_\gamma$ with $\lambda^{(1)} \in \mathbb{R}, \lambda^{(2)}, \lambda^{(3)} \in \mathbb{R}^{N_h}$:

$$\begin{aligned}\mathscr{L}_\gamma(u^+, u^-, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) :=& J(u^+, u^-) + \gamma(u^+)^\top u^- + \lambda^{(1)} \left(\textstyle\sum_{i=1}^{N_h} u_i^+ - u_i^- - \alpha\right) \\ & - \sum_{i=1}^{N_h} \lambda_i^{(2)} u_i^+ - \sum_{i=1}^{N_h} \lambda_i^{(3)} u_i^-.\end{aligned}$$

All inequalities in $(P_{h,\gamma})$ are strictly fulfilled for $u_i^+ = u_i^- = \frac{\alpha}{2(N_h+1)}$ for all $i \in \{1, \ldots, N_h\}$, so the Slater condition is satisfied (see Definition 2.29). By Karush-Kuhn-Tucker conditions (see Theorem 2.30) the following conditions in the minimum (u^+, u^-) must be fulfilled, where we directly reformulate the inequality conditions with an arbitrary $\kappa > 0$ as in the case with positive measures.

1. $\partial_{u^+}\mathscr{L}_\gamma(u^+, u^-, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) = 0$,
2. $\partial_{u^-}\mathscr{L}_\gamma(u^+, u^-, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) = 0$,
3. $N^{(1)}(u^+, u^-, \lambda^{(1)}) = \max\{0, \lambda^{(1)} + \kappa\left(\sum_{i=1}^{N_h} u_i^+ - u_i^- - \alpha\right)\} - \lambda^{(1)} = 0$,
4. $N^{(2)}(u^+, \lambda^{(2)}) = \max\{0, \lambda^{(2)} - \kappa u^+\} - \lambda^{(2)} = 0$,
5. $N^{(3)}(u^-, \lambda^{(3)}) = \max\{0, \lambda^{(3)} - \kappa u^-\} - \lambda^{(3)} = 0$.

We then apply the semismooth Newton method to solve

$$F(u^+, u^-, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) := \begin{pmatrix} \partial_{u^+}\mathscr{L}_\gamma(u^+, u^-, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) \\ \partial_{u^-}\mathscr{L}_\gamma(u^+, u^-, \lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}) \\ N^{(1)}(u^+, u^-, \lambda^{(1)}) \\ N^{(2)}(u^+, \lambda^{(2)}) \\ N^{(3)}(u^-, \lambda^{(3)}) \end{pmatrix} = 0.$$

We have

$$\partial_{u^+}\mathscr{L}_\gamma(u^+,u^-,\lambda^{(1)},\lambda^{(2)},\lambda^{(3)}) = \partial_{u^+}J(u^+,u^-) + \gamma u^- + \lambda^{(1)}\mathbb{1}_{N_h} - \lambda^{(2)},$$
$$\partial_{u^-}\mathscr{L}_\gamma(u^+,u^-,\lambda^{(1)},\lambda^{(2)},\lambda^{(3)}) = \partial_{u^-}J(u^+,u^-) + \gamma u^+ + \lambda^{(1)}\mathbb{1}_{N_h} - \lambda^{(3)}.$$

When setting up the matrix DF, we always make the choice $\partial_x(\max\{0,g(x)\}) = \partial_x g(x)$ if $g(x) = 0$. This delivers (in short notation):

$$DF := \begin{pmatrix} \partial^2_{u^+u^+}\mathscr{L}_\gamma & \partial_{u^-}\partial_{u^+}\mathscr{L}_\gamma & \partial_{\lambda^{(1)}}\partial_{u^+}\mathscr{L}_\gamma & \partial_{\lambda^{(2)}}\partial_{u^+}\mathscr{L}_\gamma & \partial_{\lambda^{(3)}}\partial_{u^+}\mathscr{L}_\gamma \\ \partial_{u^+}\partial_{u^-}\mathscr{L}_\gamma & \partial^2_{u^-}\mathscr{L}_\gamma & \partial_{\lambda^{(1)}}\partial_{u^-}\mathscr{L}_\gamma & \partial_{\lambda^{(2)}}\partial_{u^-}\mathscr{L}_\gamma & \partial_{\lambda^{(3)}}\partial_{u^-}\mathscr{L}_\gamma \\ \partial_{u^+}N^{(1)} & \partial_{u^-}N^{(1)} & \partial_{\lambda^{(1)}}N^{(1)} & \partial_{\lambda^{(2)}}N^{(1)} & \partial_{\lambda^{(3)}}N^{(1)} \\ \partial_{u^+}N^{(2)} & \partial_{u^-}N^{(2)} & \partial_{\lambda^{(1)}}N^{(2)} & \partial_{\lambda^{(2)}}N^{(2)} & \partial_{\lambda^{(3)}}N^{(2)} \\ \partial_{u^+}N^{(3)} & \partial_{u^-}N^{(3)} & \partial_{\lambda^{(1)}}N^{(3)} & \partial_{\lambda^{(2)}}N^{(3)} & \partial_{\lambda^{(3)}}N^{(3)} \end{pmatrix}$$
$$= \begin{pmatrix} \partial^2_{u^+u^+}\mathscr{L}_\gamma & \partial_{u^-}\partial_{u^+}\mathscr{L}_\gamma & \partial_{\lambda^{(1)}}\partial_{u^+}\mathscr{L}_\gamma & \partial_{\lambda^{(2)}}\partial_{u^+}\mathscr{L}_\gamma & 0 \\ \partial_{u^+}\partial_{u^-}\mathscr{L}_\gamma & \partial^2_{u^-}\mathscr{L}_\gamma & \partial_{\lambda^{(1)}}\partial_{u^-}\mathscr{L}_\gamma & 0 & \partial_{\lambda^{(3)}}\partial_{u^-}\mathscr{L}_\gamma \\ \partial_{u^+}N^{(1)} & \partial_{u^-}N^{(1)} & \partial_{\lambda^{(1)}}N^{(1)} & 0 & 0 \\ \partial_{u^+}N^{(2)} & 0 & 0 & \partial_{\lambda^{(2)}}N^{(2)} & 0 \\ 0 & \partial_{u^-}N^{(3)} & 0 & 0 & \partial_{\lambda^{(3)}}N^{(3)} \end{pmatrix},$$

with the entries

$$\partial^2_{u^+u^+}\mathscr{L}_\gamma = \partial^2_{u^+u^+}J,$$
$$\partial_{u^-}\partial_{u^+}\mathscr{L}_\gamma = \partial_{u^+}\partial_{u^-}\mathscr{L}_\gamma = \partial_{u^-}\partial_{u^+}J + \gamma\mathbb{I}_{N_h},$$
$$\partial_{\lambda^{(1)}}\partial_{u^+}\mathscr{L}_\gamma = \partial_{\lambda^{(1)}}\partial_{u^-}\mathscr{L}_\gamma = \mathbb{1}_{N_h},$$
$$\partial_{\lambda^{(2)}}\partial_{u^+}\mathscr{L}_\gamma = \partial_{\lambda^{(3)}}\partial_{u^-}\mathscr{L}_\gamma = -\mathbb{I}_{N_h},$$
$$\partial^2_{u^-}\mathcal{L}_\gamma = \partial^2_{u^-}J,$$
$$\partial_{u^+}N^{(1)} = \partial_{u^-}N^{(1)} = \begin{cases} \kappa\mathbb{1}^\top_{N_h}, & \lambda^{(1)} + \kappa\left(\sum_{i=1}^{N_h} u_i^+ + u_i^- - \alpha\right) \geq 0, \\ 0, & \text{else}, \end{cases}$$
$$\partial_{\lambda^{(1)}}N^{(1)} = \begin{cases} 0, & \lambda^{(1)} + \kappa\left(\sum_{i=1}^{N_h} u_i^+ + u_i^- - \alpha\right) \geq 0, \\ -1, & \text{else}, \end{cases}$$
$$\partial_{u_j^+}N_i^{(2)} = \begin{cases} -\kappa\delta_{ij}, & \lambda_i^{(2)} - \kappa u_i^+ \geq 0, \\ 0, & \text{else}, \end{cases}$$
$$\partial_{\lambda_j^{(2)}}N_i^{(2)} = \begin{cases} 0, & \lambda_i^{(2)} - \kappa u_i^+ \geq 0, \\ -\delta_{ij}, & \text{else}, \end{cases}$$
$$\partial_{u_j^-}N_i^{(3)} = \begin{cases} -\kappa\delta_{ij}, & \lambda_i^{(3)} - \kappa u_i^- \geq 0, \\ 0, & \text{else}, \end{cases}$$
$$\partial_{\lambda_j^{(3)}}N_i^{(3)} = \begin{cases} 0, & \lambda_i^{(3)} - \kappa u_i^- \geq 0, \\ -\delta_{ij}, & \text{else}. \end{cases}$$

Numerical example

Let $\Omega = [0,1]$, $T = 1$ and $a = \frac{1}{100}$. Again, this combination of T and a is chosen to avoid too big a condition number of the discrete state equation. We are working on a 20×20 space-time grid for this example. Positive parts of the measure are displayed by black circles and negative parts by red diamonds. We always start the algorithm with the control being identically zero and terminate when the residual is below 10^{-15}.

First example like described in Section 4.4.1, compare Figure 4.1. We found the following values to be suitable: The penalty parameter $\gamma = 70$ in $P_{h,\gamma}$ and the multiplier $\kappa = 2$ to reformulate the KKT-conditions.

The first case we investigate is $\alpha = 0.1$ (see Figure 4.5). This α is smaller than the total variation of the true control and we observe $\bar{u}^+(\bar{\Omega}) = \alpha, \bar{u}^-(\bar{\Omega}) = 0$.

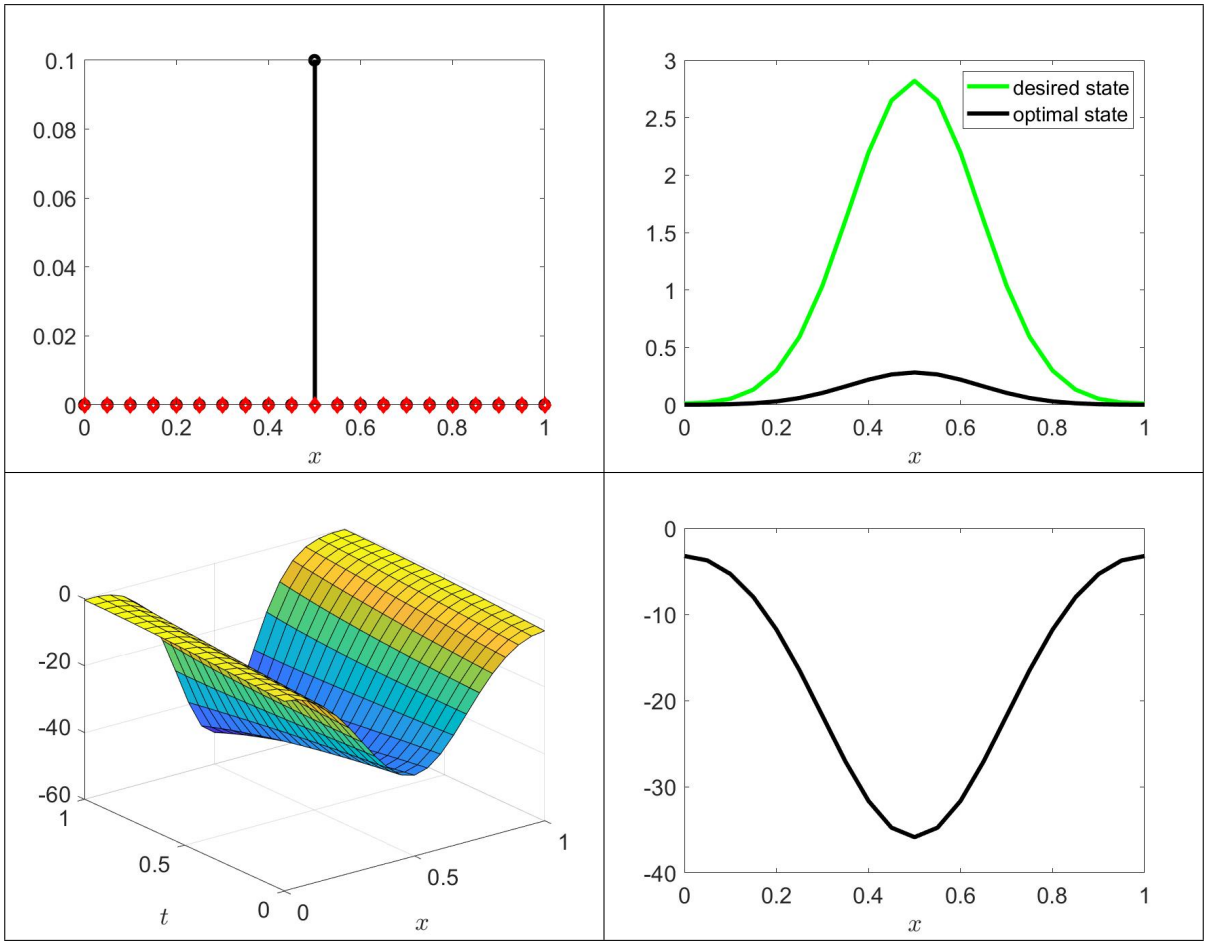

Figure 4.5: Solutions for $\alpha = 0.1$: from top left to bottom right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 11 Newton steps.

The second case we investigate is $\alpha = 1$ (see Figure 4.6). This α is equal to the total variation of the true control and we observe $\bar{u}^+(\bar{\Omega}) = \alpha, \bar{u}^-(\bar{\Omega}) = 1.8635 \cdot 10^{-20}$. These results are almost identical to the results in Section 4.4.1, where only positive measures were allowed (compare Figure 4.2 and 4.3).

The third case we investigate is $\alpha = 2$ (see Figure 4.7). This α is bigger than the total variation of the true control and we observe $\bar{u}^+(\bar{\Omega}) = 1.5, \bar{u}^-(\bar{\Omega}) = 0.5$. Furthermore $\bar{y}(T) \approx y_d$ (with an error of size 10^{-8}) and $\bar{\varphi} \approx 0 \in Q$. Since we allow positive and negative coefficients, the desired state can be reached on the coarse grid - different to the case of only positive sources, but as a payoff the sparsity of the optimal control is lost. As required, the complementarity condition has been fulfilled, i.e. $u_i^+ u_i^- = 0$ holds for all i. This however, comes at the cost of many iterations, since a big constant γ causes bad condition of our problem. As a remedy we implemented

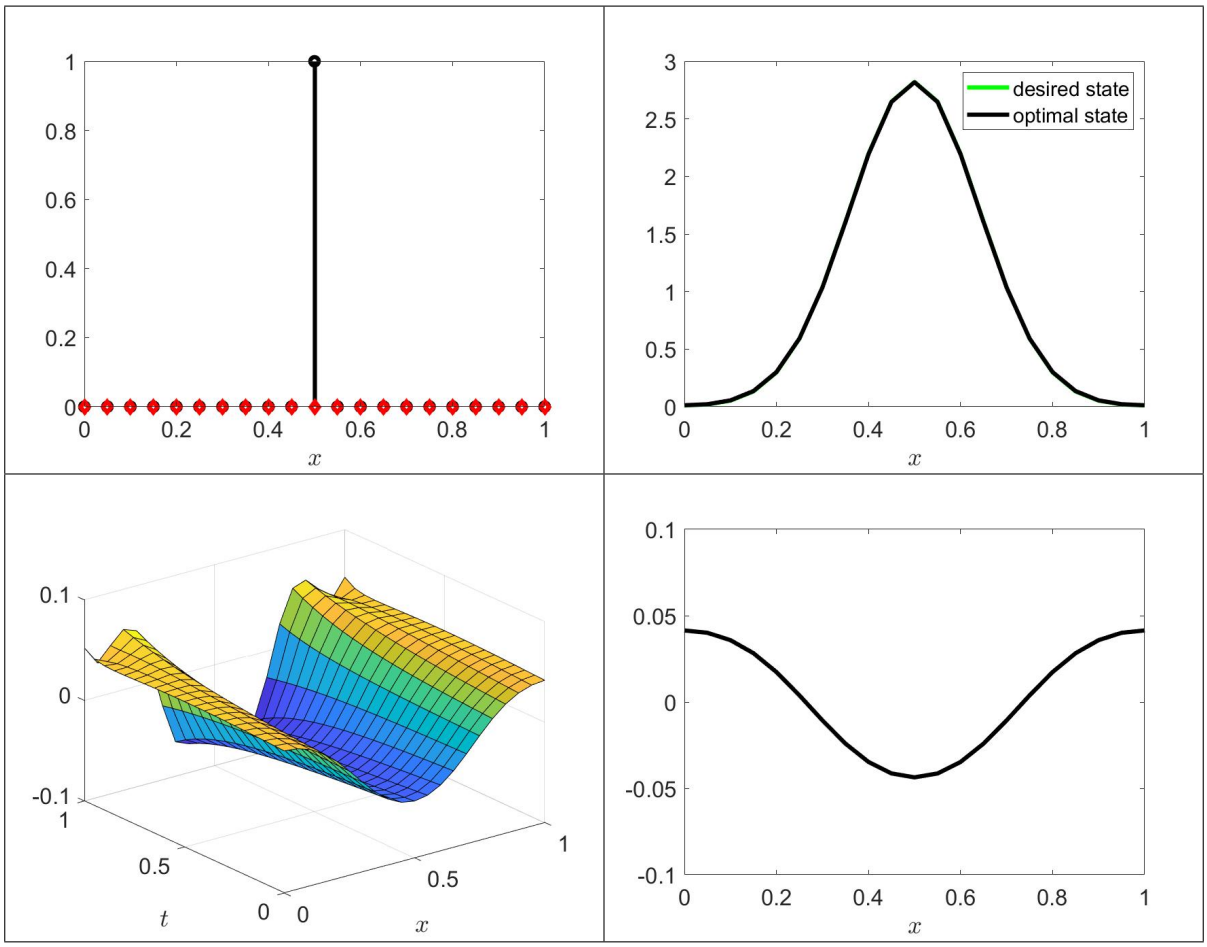

Figure 4.6: Solutions for $\alpha = 1$: from top left to bottom right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 64 Newton steps.

a γ-homotopy like e.g. in [14, Section 6], where we start with $\gamma = 1$, solve the problem using the semismooth Newton method and use this solution as a starting point for an increased γ until a solution satisfies the constraints. With a fixed $\gamma = 70$ we need almost 1000 Newton steps, with the the γ-homotopy, which terminates at $\gamma = 64$ in this setting, it takes 183 Newton steps.

As a comparison to the problem with only positive sources, we also solve the problem with the same reachable desired state as in Figure 4.4, i.e. the projection of the original desired state onto the coarse grid. Here, we also observe $\bar{y}(T) \approx y_d$ (with an error of size 10^{-12}) and $\bar{\varphi} \approx 0 \in Q$. Furthermore the optimal control is sparse with $\text{supp}(\bar{u}^+) = \{0.5\}$, only consists of a positive part and its total variation is $\bar{u}^+(\bar{\Omega}) = 1 < \alpha$. We fix $\gamma = 70$ and need 56 Newton steps in this case.

Furthermore, we solve this case on a finer space-time mesh (40×40) to compare the behavior of solutions (see Figure 4.8). We observe a higher iteration count: 255 Newton steps when employing a $\gamma-$homotopy, which terminates at $\gamma = 64$. In fact for any example, which we solved on two different meshes the solver needed more iterations on the finer grid. This is caused by the growing condition number of the partial differential equation solver, since it is a mapping from an initial measure control to the state at final time. We can also see a difference in the optimal controls in Figure 4.7 top and Figure 4.8, although comparable associated optimal state and adjoint are achieved.

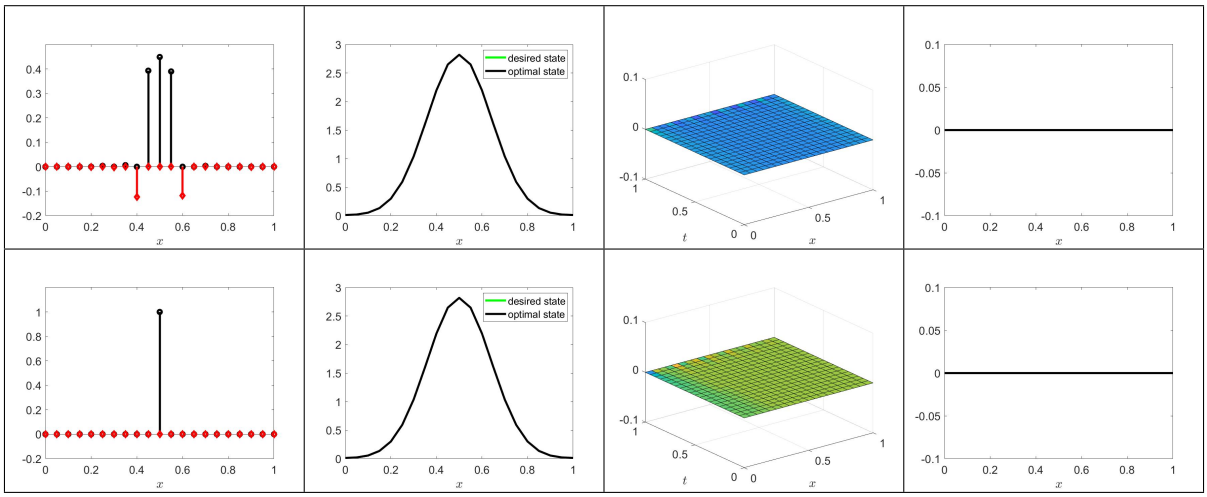

Figure 4.7: Solutions for $\alpha = 2$ with original desired state (top) and reachable desired state (bottom): from left to right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 183 and 56 Newton steps, respectively.

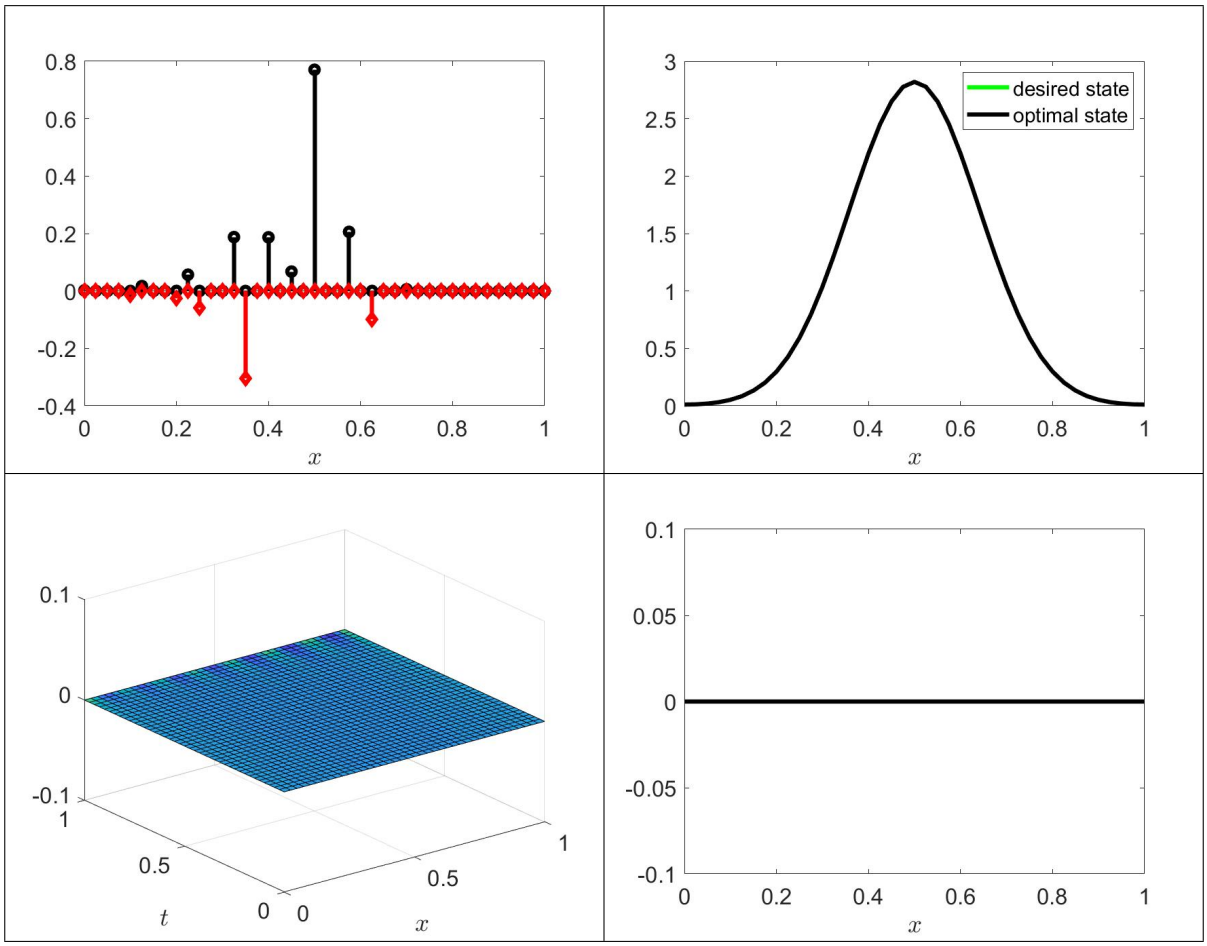

Figure 4.8: Solution for $\alpha = 2$ with original desired state on a 40×40 space-time grid: from top left to bottom right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 255 Newton steps.

The second example we want to look at is a measure consisting of a positive and a negative part. To generate a desired state y_d, we choose $u_{\text{true}} = \delta_{0.3} - 0.5 \cdot \delta_{0.8}$ and $f \equiv 0$, solve the state equation on a fine space-time grid (1000×1000) and take the evaluation of the result in $t = T$ on the current grid Ω_h as desired state y_d (see Figure 4.9).

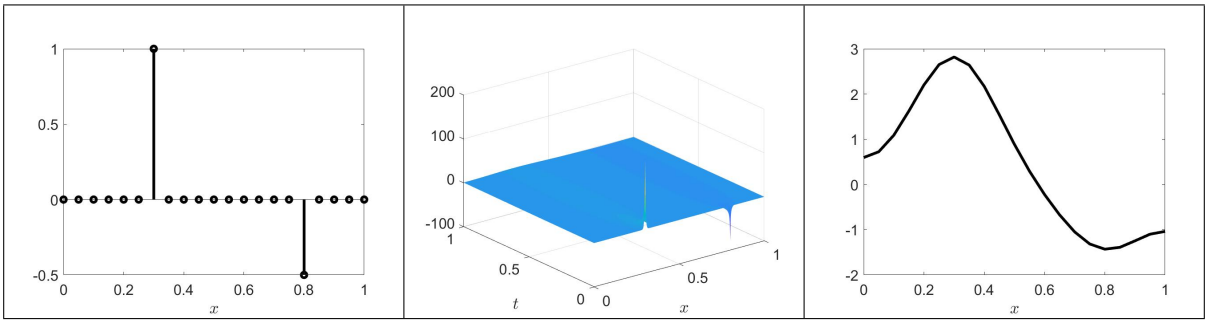

Figure 4.9: From left to right: true solution u_{true}, associated true state y_{true} in $Q = [0,1] \times [0,1]$ and desired state $y_d = y_{\text{true}}(T)$

The first case we investigate is $\alpha = 0.15$ (see Figure 4.10). This α is smaller than the total variation of the true control and we observe $\bar{u}^+(\bar{\Omega}) = 0.15, \bar{u}^-(\bar{\Omega}) = 1.2929 \cdot 10^{-16}$.

The second case we investigate is $\alpha = 1.5$ (see Figure 4.11). This α is equal to the total variation of the true control and we observe $\bar{u}^+(\bar{\Omega}) = 1.0001, \bar{u}^-(\bar{\Omega}) = 0.4999$. For both cases displayed in Figure 4.9 we fix $\gamma = 70$.

Again, we investigate as third case a setting, where $\alpha = 3 > 1.5 = \|u_{\text{true}}\|_{\mathcal{M}(\bar{\Omega})}$ (see Figure 4.12). We observe $\bar{u}^+(\bar{\Omega}) = 1.75, \bar{u}^-(\bar{\Omega}) = 1.25$. Here, $\bar{y}(T) \approx y_d$ (with an error of size 10^{-7}) and $\bar{\varphi} \approx 0 \in Q$ hold. The optimal control fulfills the complementarity condition, but we can not observe the same sparsity that was inherited by u_{true}. For this case we have to raise the fix γ to 100 and the computation took over 1700 Newton steps. Hence we employ a γ-homotopy again, which terminates at $\gamma = 64$ in this setting, and only need 137 Newton steps.

For comparison we project the desired state onto the coarse grid, such that it becomes reachable and then solve the problem again. Now we observe $\bar{u}^+(\bar{\Omega}) = 1, \bar{u}^-(\bar{\Omega}) = 0.5, \text{supp}(\bar{u}^+) = \{0.3\}, \text{supp}(\bar{u}^-) = \{0.8\}$, which are exactly the properties of u_{true}. Furthermore we see $\bar{y}(T) \approx y_d$ (with an error of size 10^{-14}) and $\bar{\varphi} \approx 0 \in Q$. We observe a reduction of Newton steps needed - the computation took 20 Newton steps with fixed $\gamma = 100$.

In summary, for all cases where $\alpha \leq \|u_{\text{true}}\|_{\mathcal{M}(\Omega)}$ holds, we observe the optimality conditions and sparsity structure, which we proved in Section 4.3. For the cases where $\alpha > \|u_{\text{true}}\|_{\mathcal{M}(\Omega)}$ holds, the bound on the total variation of the optimal control is higher than needed to achieve the desired state in the continuous setting. Here, we have constructed a desired state y_{d}, which is not supported in the grid points of the coarse grid we employ, therefore not being "reachable". So, the freedom of the optimal control to attain a higher total variation than the true control, leads to a better approximation of y_{d} in the general case with positive and negative parts of the measures, at the loss of sparsity. Projecting the given y_{d} onto the coarse grid, thus making the desired state "reachable", leads to a sparse optimal control.

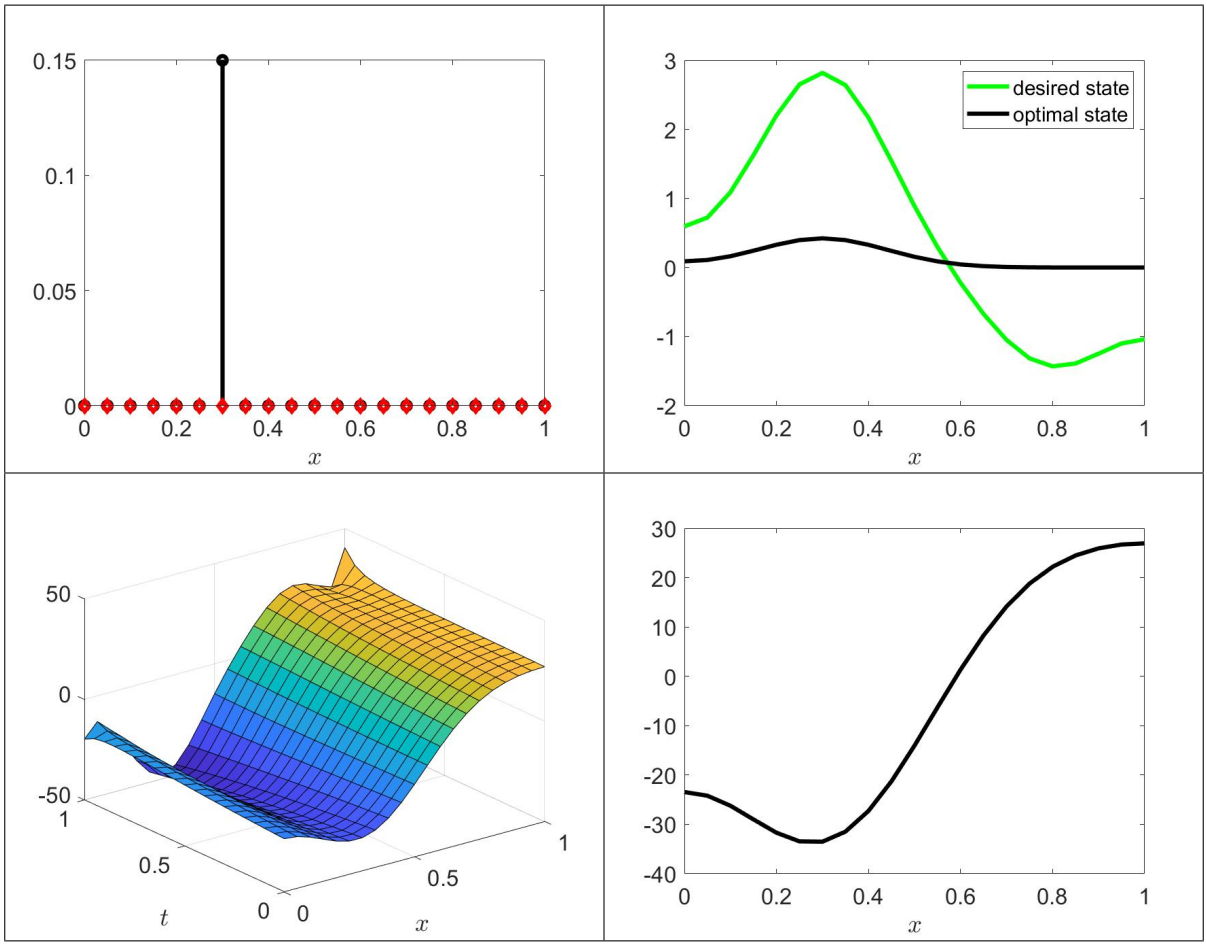

Figure 4.10: Solutions for $\alpha = 0.15$: from top left to bottom right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 29 Newton steps.

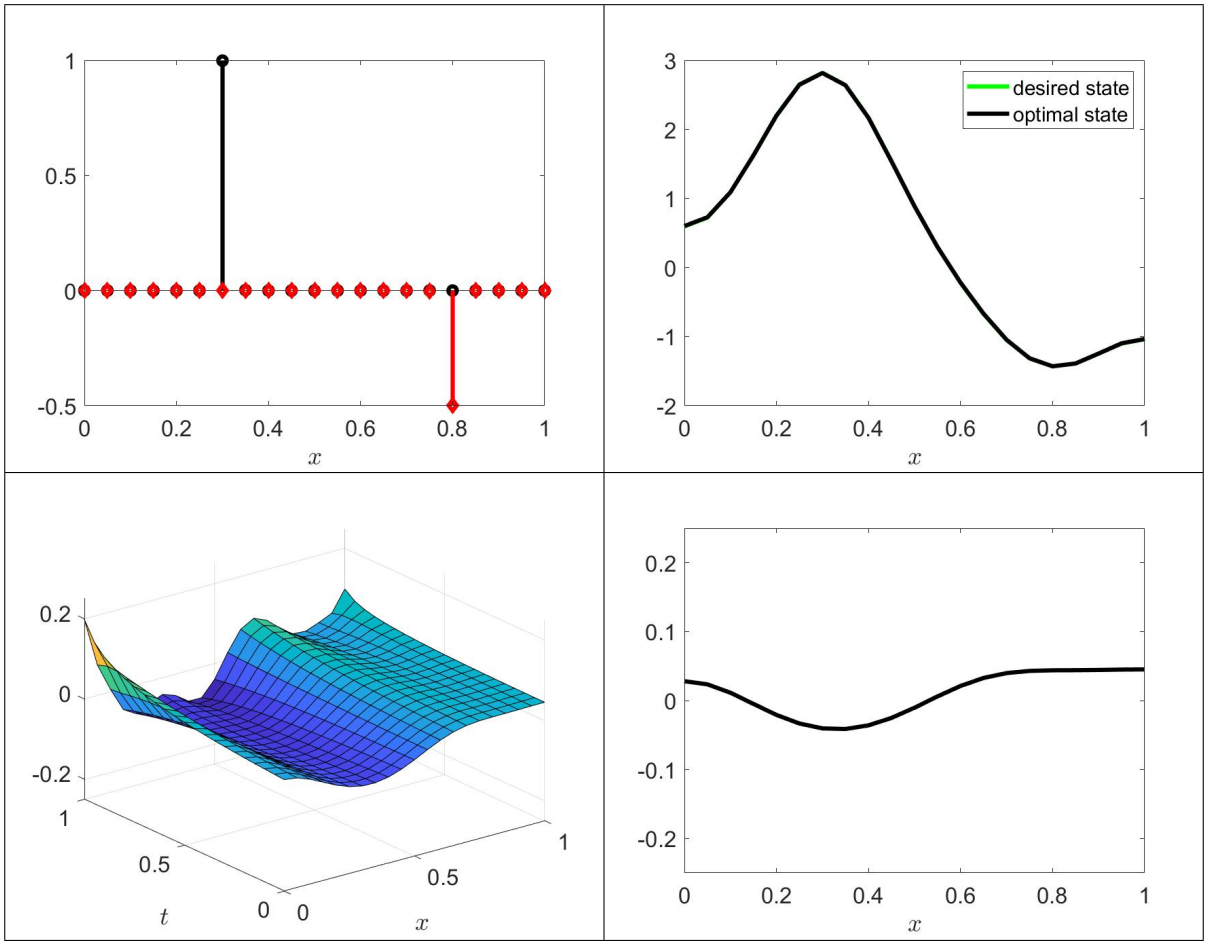

Figure 4.11: Solutions for $\alpha = 1.5$: from top left to bottom right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 44 Newton steps.

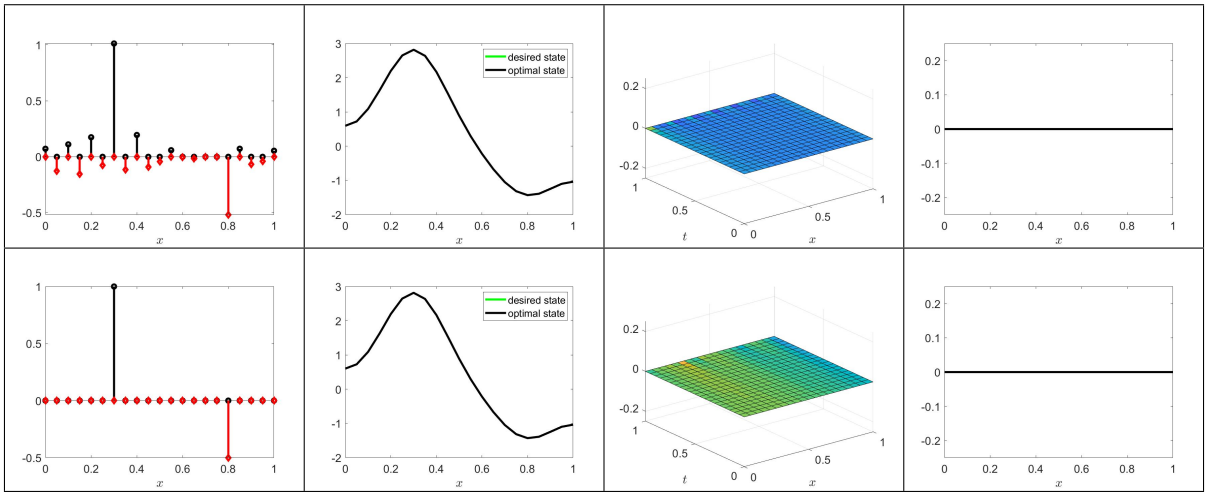

Figure 4.12: Solutions for $\alpha = 3$ with original desired state (top) and reachable desired state (bottom): from left to right: optimal control $\bar{u} = \bar{u}^+ - \bar{u}^-$ (solved with the semismooth Newton method), associated optimal state $\bar{y}$, associated adjoint $\bar{\varphi}$ on the whole space-time domain Q, associated adjoint $\bar{\varphi}$ at $t = 0$. Terminated after 137 and 20 Newton steps.

Chapter 5

Elliptic optimal control governed by functions of bounded variation

We structure this chapter as follows: We state the elliptic optimal control problem in Section 5.1. In Section 5.2 we introduce the mixed formulation of the state equation, prove existence of a unique solution to the elliptic optimal control problem and derive its optimality conditions and sparsity structure. We apply variational discretization to the problem in Section 5.3 and discuss the resulting structure of the non-discretized controls. Then, we proceed analogously to the analysis of the continuous problem by proving existence of a solution, deriving optimality conditions and sparsity structure. We also examine error estimates. Finally, in Section 5.4 we explain how to apply the semismooth Newton method to our problem, derive an optimization algorithm and then present computational results for two different examples. We compare our findings to the experiments from [43].

5.1 Problem formulation

We consider the optimal control problem:

$$\min_{u \in BV(\Omega)} J(u) \coloneqq \frac{1}{2}\|y - y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \alpha \|u'\|_{\mathcal{M}(\Omega)}, \tag{P}$$

where y satisfies the one-dimensional elliptic partial differential equation

$$\begin{cases} -y'' &= u, \qquad \text{in } \Omega, \\ y &= 0, \qquad \text{on } \Gamma. \end{cases} \tag{5.1}$$

Let $\Omega = (0,1)$ with boundary $\Gamma = \{0,1\}$, and the parameter $\alpha > 0$. We denote the control by $u \in BV(\Omega)$, the state by $y \in H^1_0(\Omega)$, and the desired state by $y_{\mathrm{d}} \in L^\infty(\Omega)$. We employ the BV-seminorm $\|u'\|_{\mathcal{M}(\Omega)}$ in the objective, since it favors piecewise constant controls that jump only a limited amount of times. This problem can be understood as a special case - i.e. with a special choice of elliptic partial differential equation - of the problem considered in [43].

5.2 Continuous optimality system

We begin by examining the state equation. The state y is supposed to solve (5.1) in the following weak sense:

Definition 5.1. *A function $y \in H^1_0(\Omega)$ is a solution to* (5.1)*, if it satisfies the identity*

$$\int_\Omega y'\, v'\, dx = \int_\Omega uv\, dx \qquad \forall v \in H^1_0(\Omega). \tag{5.2}$$

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

$$a(y, v) \coloneqq \int_\Omega y' \, v' \, dx,$$

and the linear functional $F \in (H_0^1(\Omega))^*$ as $F(v) \coloneqq (u, v)_{L^2(\Omega)}$. Taking this inner product is sensible, since we have $BV(\Omega) \hookrightarrow L^p(\Omega)$ continuously for $p \in [1, \infty]$ and compactly for $p \in [1, \infty)$ from Theorem 2.25 for $n = 1$. In particular $BV(\Omega) \hookrightarrow L^2(\Omega)$.

So, (5.2) can be written as

$$\text{Find } y \in H_0^1(\Omega) : \quad a(y, v) = F(v) \quad \forall v \in H_0^1(\Omega).$$

It is well known for this setting, that by the Lax-Milgram Theorem (see e.g. [52, Lemma 1.8]) we have a unique solution to (5.2).

However, in this work we want to take another approach: The state equation (5.1) can be written in mixed formulation. We remark that due to $n = 1$ in this setting the space (see e.g. [64])

$$H(\operatorname{div}; \Omega) \coloneqq \left\{ z \in L^2(\Omega) : \operatorname{div} z \in L^2(\Omega) \right\}$$

coincides with $H^1(\Omega)$. So, with $z \in H^1(\Omega)$ the mixed formulation reads:

$$\begin{cases} -z' &= u, & \text{in } \Omega, \\ z &= y', & \text{in } \Omega, \\ y &= 0, & \text{on } \Gamma. \end{cases} \tag{5.3}$$

This results in the weak formulation: Find $(y, z) \in H_0^1(\Omega) \times H^1(\Omega)$, such that

$$\int_\Omega zv + yv' \, dx = 0 \qquad \forall v \in H^1(\Omega), \tag{5.4a}$$

$$\int_\Omega wz' \, dx = -\int_\Omega wu \, dx \qquad \forall w \in L^2(\Omega). \tag{5.4b}$$

Here we write $(y, z) = \mathcal{S}(u)$ for the solution of (5.4). We know by [64, Theorem 1] that $\mathcal{S}(u)$ admits a unique solution $(y_u, z_u) \in H_0^1(\Omega) \times H^1(\Omega)$, where y_u solves (5.1) and $z = y'$.

Remark 5.2. *Introducing the mixed formulation gives rise to the opportunity of including $z = y'$ in the target functional.*

For simplicity in the following proofs we additionally introduce the control-to-state operator

$$S : BV(\Omega) \subset H^{-1}(\Omega) \coloneqq (H_0^1(\Omega))^* \to H_0^1(\Omega), \qquad u \mapsto y_u.$$

In this sense, it is meaningful to consider the reduced problem (P). We have the following regularity result (see [42, Lemma 2.2.])

Lemma 5.3. *Let $(y_u, z_u) = \mathcal{S}(u) \in H_0^1(\Omega) \times H^1(\Omega)$ with data $u \in L^2(\Omega)$. Then we have*

$$\|y_u\|_{L^2(\Omega)} + \|z_u\|_{H^1(\Omega)} \leq C \|u\|_{L^2(\Omega)}. \tag{5.5}$$

Furthermore, if Ω is convex, then $y_u \in H^2(\Omega)$ and

$$\|y_u\|_{H^2(\Omega)} + \|z_u\|_{H^1(\Omega)} \leq C \|u\|_{L^2(\Omega)}. \tag{5.6}$$

Especially, we get $\|Su\|_{H^2(\Omega)} \leq C \|u\|_{L^2(\Omega)}$.

The existence of a unique optimal control follows directly from [43, Theorem 2.2.], but we will give the result here for completeness.

Theorem 5.4. *Problem (P) admits a unique optimal control $\bar{u} \in BV(\Omega)$ with associated optimal state $\bar{y} \in H_0^1(\Omega)$.*

Proof. To prove the existence of a solution $\bar{u}$ we consider a minimizing sequence $(u_k)_k \in BV(\Omega)$, such that

$$\lim_{k\to\infty} J(u_k) = \inf_{u\in BV(\Omega)} J(u) =: \underline{\mathrm{J}},$$

and

$$J(u_k) \leq J(0) \qquad \forall k \in \mathbb{N}. \tag{5.7}$$

We will show boundedness of $(u_k)_k$ in the BV-norm $\|u_k\|_{BV(\Omega)} = \|u_k\|_{L^1(\Omega)} + \|u_k'\|_{\mathcal{M}(\Omega)}$. From (5.7) we have

$$\|u_k'\|_{\mathcal{M}(\Omega)} \leq \frac{J(0)}{\alpha}, \tag{5.8}$$

so it remains to show boundedness of $\left(\|u_k\|_{L^1(\Omega)}\right)_k$. By [3, Theorem 3.44] it holds for all $k \in \mathbb{N}$

$$\|u_k - \hat{u}_k\|_{L^1(\Omega)} \leq C_1 \|u_k'\|_{\mathcal{M}(\Omega)} \leq \frac{C_1 J(0)}{\alpha}, \tag{5.9}$$

with $\hat{u}_k := \frac{1}{|\Omega|}\int_\Omega u_k\,dx$ and C_1 depending on Ω only. Due to $\Omega = (0,1)$ we have

$$\|\hat{u}_k\|_{L^1(\Omega)} = |\hat{u}_k|.$$

Now, by reverse triangle inequality we see

$$\|u_k\|_{L^1(\Omega)} \leq \frac{C_1 J(0)}{\alpha} + |\hat{u}_k|.$$

This means we have to show boundedness of $(|\hat{u}_k|)_k$. Again by reverse triangle inequality, we have

$$\begin{aligned} |\hat{u}_k|\,\|S\,1\|_{L^2(\Omega)} &= \|S\,\hat{u}_k\|_{L^2(\Omega)} \\ &\leq \|S\,(\hat{u}_k - u_k)\|_{L^2(\Omega)} + \|S\,u_k\|_{L^2(\Omega)} \\ &\leq \|S\|_{\mathcal{L}(H^{-1}(\Omega),L^2(\Omega))}\|\hat{u}_k - u_k\|_{H^{-1}(\Omega)} + \|S\,u_k\|_{L^2(\Omega)}. \end{aligned} \tag{5.10}$$

From (5.7) we get

$$\begin{aligned} &\frac{1}{2}\|S\,u_k - y_\mathrm{d}\|^2_{L^2(\Omega)} + \alpha\|u_k'\|_{\mathcal{M}(\Omega)} \leq \frac{1}{2}\|y_\mathrm{d}\|^2_{L^2(\Omega)} \\ \Rightarrow\quad &\|S\,u_k - y_\mathrm{d}\|_{L^2(\Omega)} \leq \|y_\mathrm{d}\|_{L^2(\Omega)} \\ \Rightarrow\quad &\|S\,u_k\|_{L^2(\Omega)} \leq 2\,\|y_\mathrm{d}\|_{L^2(\Omega)} = 2\sqrt{2J(0)}. \end{aligned} \tag{5.11}$$

We have the embedding $L^1(\Omega) \hookrightarrow H^{-1}(\Omega)$, so that $\|u\|_{H^{-1}(\Omega)} \leq C_2\|u\|_{L^1(\Omega)}$, with constant $C_2 > 0$. Using this and combining (5.10) with (5.9) and (5.11), we derive

$$|\hat{u}_k| \leq \|S\,1\|^{-1}_{L^2(\Omega)}\left(\frac{C_1 C_2 J(0)}{\alpha}\|S\|_{\mathcal{L}(H^{-1}(\Omega),L^2(\Omega))} + 2\sqrt{2J(0)}\right),$$

where we use $S\,1 \neq 0$. Finally, we have for all $k \in \mathbb{N}$

$$\|u_k\|_{BV(\Omega)} \leq \frac{(C_1+1)J(0)}{\alpha} + \|S\,1\|^{-1}_{L^2(\Omega)}\left(\frac{C_1 C_2 J(0)}{\alpha}\|S\|_{\mathcal{L}(H^{-1}(\Omega),L^2(\Omega))} + 2\sqrt{2J(0)}\right). \tag{5.12}$$

Due to $BV(\Omega) \hookrightarrow L^1(\Omega)$ compactly, there exists a subsequence $(u_{k'})_{k'}$ of $(u_k)_k$ and a $\bar{u} \in L^1(\Omega)$, such that $u_{k'} \to \bar{u} \in L^1(\Omega)$ for $k' \to \infty$. For $u \in L^1(\Omega)$ the mapping $u \mapsto \frac{1}{2}\|Su - y_d\|^2_{L^2(\Omega)}$ is continuous and by [74, Theorem 5.2.1.] we have lower semicontinuity of $u \mapsto \|u'\|_{\mathcal{M}(\Omega)}$. Altogether

$$\underline{J} \leq J(\bar{u}) \leq \liminf_{k' \to \infty} J(u_{k'}) = \lim_{k \to \infty} J(u_k) = \underline{J},$$

which delivers existence of a solution.

From the injectivity of S we deduce strict convexity of $J(u)$, which delivers uniqueness of the solution. Assume there exist two solutions $u_1, u_2 \in BV(\Omega)$ of (P) with $u_1 \neq u_2$, then for $\lambda \in (0,1)$

$$J(\lambda u_1 + (1-\lambda)u_2) < \lambda J(u_1) + (1-\lambda)J(u_2) = J(u_1).$$

This contradicts u_1 being a solution, so $u_1 = u_2$ must hold. □

Similar to [43, Theorem 2.3.], but adapted to the mixed formulation of the state equation, we provide the following optimality conditions.

Theorem 5.5. *The control $\bar{u} \in BV(\Omega)$ with associated $(\bar{y}, \bar{z}) \in H^1_0(\Omega) \times H^1(\Omega)$ is optimal for the problem (P) if and only if there exists a unique tupel $(\bar{p}, \bar{q}) \in H^2(\Omega) \cap H^1_0(\Omega) \times H^1(\Omega)$, such that $(\bar{u}, \bar{y}, \bar{z}, \bar{p}, \bar{q})$ and the $H^3(\Omega)$ function $\bar{\Phi}(x) := \int_0^x \bar{p}(s)\,ds$ satisfy $\bar{\Phi}(1) = 0$ as well as*

$$\int_\Omega \bar{\Phi}\, d\bar{u}' = \alpha \|\bar{u}'\|_{\mathcal{M}(\Omega)}, \tag{5.13}$$

$$\|\bar{\Phi}\|_{C(\Omega)} \leq \alpha, \tag{5.14}$$

$$\int_\Omega \bar{z}v + \bar{y}v'\,dx = 0 \qquad \forall v \in H^1(\Omega), \tag{5.15}$$

$$\int_\Omega w\bar{z}'\,dx = -\int_\Omega w\bar{u}\,dx \qquad \forall w \in L^2(\Omega), \tag{5.16}$$

$$\int_\Omega \bar{q}v + \bar{p}v'\,dx = 0 \qquad \forall v \in H^1(\Omega), \tag{5.17}$$

$$\int_\Omega w\bar{q}'\,dx = -\int_\Omega w\,(\bar{y} - y_d)\,dx \qquad \forall w \in L^2(\Omega), \tag{5.18}$$

$$-(\bar{p}, u - \bar{u})_{L^2(\Omega)} \leq \alpha\left(\|u'\|_{\mathcal{M}(\Omega)} - \|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \qquad \forall u \in BV(\Omega). \tag{5.19}$$

In the proof we proceed analogously to the proof of [43, Theorem 2.3.].

Proof. By convex analysis (see e.g. [61]) the optimality of $\bar{u}$ is equivalent to

$$0 \in \partial J(\bar{u}),$$

where $\partial J(\bar{u})$ denotes the subdifferential of J at $\bar{u}$. By chain rule ([61, Proposition 3.28.]) and sum rule ([61, Theorem 3.30]), which we can apply since both summands of J are continuous on $BV(\Omega)$, we see

$$\begin{aligned} & 0 \in \partial\left(\frac{1}{2}\|S\bar{u} - y_d\|^2_{L^2(\Omega)} + \alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \\ \Rightarrow \quad & 0 \in S^*(S\bar{u} - y_d) + \partial\left(\alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \\ \Rightarrow \quad & -S^*(S\bar{u} - y_d) \in \partial\left(\alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)}\right). \end{aligned}$$

We recall $(\bar{y}, \bar{z}) = \mathcal{S}(\bar{u})$, which readily delivers (5.15) and (5.16). Now, define the adjoint state $\bar{p} := S^*(S\bar{u} - y_d)$, with $S\bar{u} = \bar{y}$ this gives

$$\begin{cases} -\bar{p}'' = \bar{y} - y_d & \text{in } \Omega, \\ \bar{p} = 0, & \text{on } \Gamma. \end{cases} \tag{5.20}$$

Formulating (5.20) in a mixed way with $\bar{q} \in H^1(\Omega)$, such that $\bar{q} = \bar{p}'$, similar to the procedure for the state equation, we see (5.17) and (5.18). The regularity of $\bar{p}$ follows from the convexity of Ω and Lemma 5.3, which then implies $\bar{\Phi} \in H^3(\Omega)$.

Furthermore, for the subdifferential we have the equivalence

$$-\bar{p} \in \partial\left(\alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \qquad \Leftrightarrow \qquad -(\bar{p}, u-\bar{u})_{L^2(\Omega)} \leq \alpha\left(\|u'\|_{\mathcal{M}(\Omega)} - \|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \quad \forall u \in BV(\Omega),$$

which gives (5.19). Inserting $u = 2\bar{u}$ and $u = 0$ into the above inequality delivers

$$-(\bar{p}, \bar{u})_{L^2(\Omega)} = \alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)}. \tag{5.21}$$

Also, for arbitrary $\tilde{u} \in BV(\Omega)$ we can insert $u = \tilde{u} + \bar{u}$ and $u = -\tilde{u} + \bar{u}$ into the same inequality to derive

$$\begin{aligned} -(\bar{p}, \tilde{u})_{L^2(\Omega)} &\leq \alpha\left(\|\tilde{u}' + \bar{u}'\|_{\mathcal{M}(\Omega)} - \|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \quad \leq \alpha\|\tilde{u}'\|_{\mathcal{M}(\Omega)}, \\ (\bar{p}, \tilde{u})_{L^2(\Omega)} &\leq \alpha\left(\|-\tilde{u}' + \bar{u}'\|_{\mathcal{M}(\Omega)} - \|\bar{u}'\|_{\mathcal{M}(\Omega)}\right) \leq \alpha\|\tilde{u}'\|_{\mathcal{M}(\Omega)}, \end{aligned}$$

which leads to

$$|(\bar{p}, u)_{L^2(\Omega)}| \leq \alpha\|u'\|_{\mathcal{M}(\Omega)} \qquad \forall u \in BV(\Omega). \tag{5.22}$$

From (5.22) with $u = 1$ we conclude $\bar{\Phi}(1) = \int_0^1 \bar{p}(s)\,ds = (\bar{p}, 1)_{L^2(\Omega)} = 0$.

By definition of $\bar{\Phi}$ and the generalized Green's formula for BV-functions ([4, Theorem 10.2.1.]) we have

$$-(\bar{p}, \bar{u})_{L^2(\Omega)} = -\int_\Omega \bar{\Phi}'\bar{u}\,dx = \int_\Omega \bar{\Phi}\,d\bar{u}'.$$

We equivalently reformulate (5.21) and (5.22):

$$\begin{aligned} &\int_\Omega \bar{\Phi}\,d\bar{u}' = \alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)}, \\ &\left|\int_\Omega \bar{\Phi}\,du'\right| \leq \alpha\|u'\|_{\mathcal{M}(\Omega)} \qquad \forall u \in BV(\Omega). \end{aligned}$$

The equality shows (5.13) and we insert $u = 1_{(x,1)} \in BV(\Omega)$, which denotes the characteristic function of the interval $(x, 1)$, with $u' = \delta_x$, into the inequality to see

$$\left|\bar{\Phi}(x)\right| = \left|\int_\Omega \bar{\Phi}\,d\delta_x\right| \leq \alpha\|\delta_x\|_{\mathcal{M}(\Omega)} = \alpha. \tag{5.23}$$

This shows (5.14) and completes the proof. □

Similar to the optimal control problems with measure control in Chapter 3 and Chapter 4 the problem inherits a sparsity structure. In this case the structure delivers information about the support of $\bar{u}'$ and not about the support of the optimal control itself. The support of $\bar{u}'$ indicates the location of the jumping points of the optimal control $\bar{u} \in BV(\Omega)$. We repeat the following result from [43, Corollary 1]:

Lemma 5.6. *If $\bar{u}$ is optimal for (P), then there hold*

$$\operatorname{supp}((\bar{u}')^+) \subset \left\{x \in \Omega : \bar{\Phi}(x) = \alpha\right\}, \tag{5.24}$$

$$\operatorname{supp}((\bar{u}')^-) \subset \left\{x \in \Omega : \bar{\Phi}(x) = -\alpha\right\}, \tag{5.25}$$

where $\bar{u}' = (\bar{u}')^+ - (\bar{u}')^-$ is the Jordan decomposition. Moreover, we have

$$\operatorname{supp}(\bar{u}') \subset \left\{x \in \Omega : |\bar{\Phi}(x)| = \alpha\right\} \subset \{x \in \Omega : \bar{p}(x) = 0\}. \tag{5.26}$$

Proof. Let $\hat{x} \in \Omega$, such that $\bar{\Phi}(\hat{x}) < \alpha$. By continuity of $\bar{\Phi}$ there exists an open neighborhood $U \subset \Omega$ of $\hat{x}$ and $\delta > 0$, such that $\bar{\Phi}(x) \leq \alpha - \delta$ for all x in U. Then, making use of Theorem 5.5, we see

$$\begin{aligned}
\alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)} &= \int_\Omega \bar{\Phi}\, d\bar{u}' \\
&= \int_\Omega \bar{\Phi}\, d(\bar{u}')^+ - \int_\Omega \bar{\Phi}\, d(\bar{u}')^- \\
&\leq \int_{\Omega\setminus U} \alpha\, d(\bar{u}')^+ + \int_U (\alpha - \delta)\, d(\bar{u}')^+ + \int_\Omega \alpha\, d(\bar{u}')^- \\
&= \int_\Omega \alpha\, d(\bar{u}')^+ + \int_\Omega \alpha\, d(\bar{u}')^- - \int_U \delta d(\bar{u}')^+ \\
&= \alpha\|\bar{u}'\|_{\mathcal{M}(\Omega)} - \delta(\bar{u}')^+(U).
\end{aligned}$$

Combined with non-negativity of $(\bar{u}')^+(U)$ it follows $(\bar{u}')^+(U) = 0$. So for any $\hat{x} \in \Omega$ with $\bar{\Phi}(\hat{x}) < \alpha$ we deduce $\hat{x} \notin \operatorname{supp}((\bar{u}')^+)$ and therewith (5.24) holds. To show (5.25) we proceed analogously for $\hat{x} \in \Omega$ with $\bar{\Phi}(\hat{x}) > -\alpha$.

Obviously, we have

$$\operatorname{supp}(\bar{u}') = \operatorname{supp}((\bar{u}')^+) \cup \operatorname{supp}((\bar{u}')^-) = \left\{x \in \Omega : |\bar{\Phi}(x)| = \alpha\right\}.$$

Since, $\|\bar{\Phi}\|_{C(\Omega)} \leq \alpha$ holds by Theorem 5.5, we conclude that $x \in \Omega$ with $|\bar{\Phi}(x)| = \alpha$ is a global minimum or maximum of the C^1-function $\bar{\Phi}$, so it satisfies $0 = \bar{\Phi}'(x) = \bar{p}(x)$ and (5.26) holds. □

5.3 Variational discretization

We employ variational discretization in order to achieve sparsity without discretizing the control u. Instead, via the piecewise constant discretization of the adjoint state p in combination with the optimality conditions for the variational discrete problem, the structure of the control u is induced. We will see that under a structural assumption u' is a sum of measures - without being discretized. This immediately delivers that the induced structure for the control u is to be piecewise constant.

Let $0 = x_0 < x_1 < \ldots < x_N = 1$ be a partition of $\bar{\Omega} = [0, 1]$. Then for $i = 1, \ldots, N$ we define the subintervals $I_i := (x_{i-1}, x_i)$ of size $h_i := x_i - x_{i-1}$ and define $h := \max_{1\leq i\leq N} h_i$ to be the mesh width. Let χ_i for $i = 1, \ldots, N$ be the indicator function of interval I_i, i.e.

$$\chi_i(x) = \begin{cases} 1, x \in I_i, \\ 0, \text{else.} \end{cases}$$

Let e_j for $j = 0, \ldots, N$ denote the hat functions, such that $e_j(x_i) = \delta_{ij}$ for $i, j = 0, \ldots, N$.

We introduce the discrete spaces

$$\begin{aligned}
P_0 &:= \operatorname{span}\{\chi_i : 1 \leq i \leq N\}, \\
P_1 &:= \operatorname{span}\left\{e_j : 0 \leq j \leq N\right\}.
\end{aligned}$$

Using these spaces we get the discrete formulation of (5.4): Find $y_h = \sum_{i=1}^N y_i\chi_i \in P_0$, and $z_h = \sum_{j=0}^N v_j e_j \in P_1$, such that

$$\int_\Omega z_h v_h + y_h v_h'\, dx = 0 \qquad \forall v_h \in P_1, \tag{5.27a}$$

$$\int_\Omega w_h z_h'\, dx = -\int_\Omega w_h u\, dx \qquad \forall w_h \in P_0. \tag{5.27b}$$

We write $(y_h, z_h) = \mathcal{S}_h(u)$ for the unique solution of the weak mixed formulation of the discrete state equation.

In the present case where $\Omega \subset \mathbb{R}$, the space of Raviart-Thomas elements of lowest order (see e.g. [5, 64]) coincides with the chosen space P_1. Furthermore, we stress that the control space remains $BV(\Omega)$, so the control u is not discretized. For simplicity we additionally introduce the discrete control-to-state operator

$$S_h : BV(\Omega) \to P_0, \qquad u \mapsto y_{u,h}.$$

The variational discrete counterpart of (P) then reads

$$\min_{u \in BV(\Omega)} J_h(u) \coloneqq \frac{1}{2}\|y_{u,h} - y_\mathrm{d}\|^2_{L^2(\Omega)} + \alpha\,\|u'\|_{\mathcal{M}(\Omega)}. \tag{P_vd}$$

We give the discrete counterpart of Theorem 5.4.

Theorem 5.7. *Problem (P_vd) admits an optimal control $\bar{u} \in BV(\Omega)$ with associated optimal state $\bar{y} \in P_0$. There exist $C, h_0 \in \mathbb{R}_{>0}$, such that for all $h \in (0, h_0]$ we have*

$$\|\bar{u}\|_{BV(\Omega)} \leq C \tag{5.28}$$

for any optimal control $\bar{u}$.

Since the control u remains continuous and it holds $S_h 1 \neq 0$, we can use the proof for existence of solutions from Theorem 5.4 verbatim. For the proof of the boundedness in the BV-norm we refer to [43, Theorem 3.5.]. We stress that uniqueness of the control is not given in this setting, since the control is not discretized, so the control-to-state operator is in general not injective.

Analogous to Theorem 5.5 from the continuous setting we derive the optimality conditions for (P_vd).

Theorem 5.8. *The control $\bar{u} \in BV(\Omega)$ with associated $\mathcal{S}_h(\bar{u}) = (\bar{y}_h, \bar{z}_h) \in P_0 \times P_1$ is optimal for the problem (P_vd) if and only if there exists a unique tupel $(\bar{p}_h, \bar{q}_h) \in P_0 \times P_1$, such that $(\bar{u}, \bar{y}_h, \bar{z}_h, \bar{p}_h, \bar{q}_h)$ and the P_1 function $\bar{\Phi}_h(x) \coloneqq \int_0^x \bar{p}_h(s)\,ds$ satisfy $\bar{\Phi}_h(1) = 0$ as well as*

$$\int_\Omega \bar{\Phi}_h \, d\bar{u}' = \alpha \|\bar{u}'\|_{\mathcal{M}(\Omega)}, \tag{5.29}$$

$$\|\bar{\Phi}_h\|_{C(\Omega)} \leq \alpha, \tag{5.30}$$

$$\int_\Omega \bar{z}_h v_h + \bar{y}_h v_h' \, dx = 0 \qquad \forall v_h \in P_1, \tag{5.31}$$

$$\int_\Omega w_h \bar{z}_h' \, dx = -\int_\Omega w_h \bar{u} \, dx \qquad \forall w_h \in P_0, \tag{5.32}$$

$$\int_\Omega \bar{q}_h v_h + \bar{p}_h v_h' \, dx = 0 \qquad \forall v_h \in P_1, \tag{5.33}$$

$$\int_\Omega w_h \bar{q}_h' \, dx = -\int_\Omega w_h \,(\bar{y}_h - y_\mathrm{d})\, dx \qquad \forall w_h \in P_0, \tag{5.34}$$

$$-(\bar{p}_h, u - \bar{u})_{L^2(\Omega)} \leq \alpha\,(\|u'\|_{\mathcal{M}(\Omega)} - \|\bar{u}'\|_{\mathcal{M}(\Omega)}) \qquad \forall u \in BV(\Omega). \tag{5.35}$$

Furthermore, we deduce a similar sparsity structure as in Lemma 5.6 for the continuous problem.

Lemma 5.9. *If $\bar{u}$ is optimal for (P_vd), then there hold*

$$\operatorname{supp}((\bar{u}')^+) \subset \left\{x \in \Omega : \bar{\Phi}_h(x) = \alpha\right\},$$
$$\operatorname{supp}((\bar{u}')^-) \subset \left\{x \in \Omega : \bar{\Phi}_h(x) = -\alpha\right\},$$

where $\bar{u}' = (\bar{u}')^+ - (\bar{u}')^-$ is the Jordan decomposition. Moreover, we have

$$\operatorname{supp}(\bar{u}') \subset \left\{x \in \Omega : |\bar{\Phi}_h(x)| = \alpha\right\} \subset \{x \in \Omega : \bar{p}_h(x) = 0\}. \tag{5.36}$$

These results can be proven as in the continuous case, so we refrain from giving the proofs here.
Even though the control is not discretized, we can deduct information about the structure of the control from the optimality conditions and the sparsity structure, especially properties (5.30) and (5.36). Let us make the following structural assumption:

Assumption 5.10. *Suppose that* $\left\{x \in \Omega : |\bar{\Phi}_h(x)| = \alpha\right\}$ *is finite.*

This assumption is fulfilled in the generic case that $\bar{\Phi}_h$ is not constant on any interval. In particular, under this assumption, the maximal absolute value $|\bar{\Phi}_h(x)| = \alpha$ of the P_1 function $\bar{\Phi}_h$ with bound $\|\bar{\Phi}_h\|_{C(\Omega)} \leq \alpha$ can only be attained at the grid points $\{x_i\}_{i=1}^N$, which gives

$$\operatorname{supp}(\bar{u}') \subset \{x_i\}_{i=1}^N . \tag{5.37}$$

Obviously not all grid points may be points, where $|\bar{\Phi}_h(x)|$ attains the value α, which we will later account for with the outer iteration of the algorithm. Due to this structure it is natural to express the optimal control $\bar{u}$ and its derivative as

$$\bar{u} = \bar{a}_h + \sum_{i=1}^N \bar{c}_h^i 1_{(x_i,1)}, \qquad \bar{u}' = \sum_{i=1}^N \bar{c}_h^i \delta_{x_i},$$

for suitable constants $\bar{a}_h \in \mathbb{R}, \bar{c}_h = (\bar{c}_h^1, \dots, \bar{c}_h^N)^\top \in \mathbb{R}^N$. So, it is obvious that $\bar{u} \in P_0$. We can determine the coefficients $\bar{a}_h$ and $\bar{c}_h$ by solving the finite-dimensional, convex optimization problem

$$\min_{a_h \in \mathbb{R}, c_h \in \mathbb{R}^N} \frac{1}{2}\|y_h - y_\mathrm{d}\|^2_{L^2(\Omega)} + \alpha \sum_{i=1}^N |c_h^i| \qquad \text{s.t.} \qquad (y_h, v_h) = \mathcal{S}_h(a_h + \sum_{i=1}^N c_h^i 1_{(x_i,1)}). \tag{P_h}$$

Analogous to [43, Definition 3.9. and Lemma 3.10.] we define a helpful operator and collect a few properties.

Lemma 5.11. *For $i = 1, \dots, N$ let the operator Υ_h be defined as below:*

$$\Upsilon_h : BV(\Omega) \to P_0, \qquad \Upsilon_h u_{|I_i} := \frac{1}{h_i} \int_{I_i} u(s)\, ds.$$

For any $u \in BV(\Omega)$ and $w_h \in P_0$ it holds

$$(u, w_h)_{L^2(\Omega)} = (\Upsilon_h u, w_h)_{L^2(\Omega)}, \tag{5.38}$$

$$\|u - \Upsilon_h u\|_{L^1(\Omega)} \leq h \|u'\|_{\mathcal{M}(\Omega)}, \tag{5.39}$$

$$\|(\Upsilon_h u)'\|_{\mathcal{M}(\Omega)} \leq \|u'\|_{\mathcal{M}(\Omega)}, \tag{5.40}$$

$$\|u - \Upsilon_h u\|_{L^\infty(\Omega)} \leq h \|u'\|_{L^\infty(\Omega)}, \qquad \text{provided that } u \in W^{1,\infty}(\Omega). \tag{5.41}$$

This proof has been collected from [17, Proposition 16] and [43, Lemma 3.10.].

Proof. The equality (5.38) is obvious, since

$$\int_{I_i} u\, dx = \int_{I_i} \Upsilon_h u\, dx$$

holds for all $i = 1, \dots, N$. Also, for $u \in C^1(\Omega)$ the inequality (5.39) is given. Now, for $u \in BV(\Omega)$ there exists a sequence $\{u_j\}_{j \in \mathbb{N}} \subset C^\infty(\Omega)$, such that

$$\|u - u_j\|_{L^1(\Omega)} + \left| \|u'\|_{\mathcal{M}(\Omega)} - \|u_j'\|_{\mathcal{M}(\Omega)} \right| \leq \frac{1}{j} \qquad \forall\, j \geq 1, \tag{5.42}$$

by [3, Remark 3.22.].

Then, we estimate

$$\begin{aligned}\|u - \Upsilon_h u\|_{L^1(\Omega)} &\leq \|u - u_j\|_{L^1(\Omega)} + \|u_j - \Upsilon_h u_j\|_{L^1(\Omega)} + \|\Upsilon_h u_j - \Upsilon_h u\|_{L^1(\Omega)} \\ &\leq \|u - u_j\|_{L^1(\Omega)} + h\|u_j'\|_{\mathcal{M}(\Omega)} + \|u_j - u\|_{L^1(\Omega)} \\ &\leq \frac{2}{j} + h\|u_j'\|_{\mathcal{M}(\Omega)}.\end{aligned}$$

For $j \to \infty$ we deduce (5.39). Next, we show (5.40) for $u \in C^\infty(\Omega)$. By continuity of u and the mean value theorem for integrals we know there exist points $\xi_i \in I_i, i = 1, \dots, N$, such that

$$\Upsilon_h u = \sum_{i=1}^N u(\xi_i)\chi_i.$$

For any element $v \in P_0$ with $v = \sum_{i=1}^N v_i\chi_i$ it holds

$$v' = \sum_{i=2}^N (v_i - v_{i-1})\delta_{x_{i-1}} \qquad \text{and} \qquad \|v'\|_{\mathcal{M}(\Omega)} = \sum_{i=2}^N |v_i - v_{i-1}|,$$

where δ_x denotes the Dirac measure concentrated in x. So, we have

$$\begin{aligned}\|(\Upsilon_h u)'\|_{\mathcal{M}(\Omega)} &= \sum_{i=2}^N |u(\xi_i) - u(\xi_{i-1})| \\ &\leq \sum_{i=2}^N \int_{\xi_{i-1}}^{\xi_i} |u'(x)|\,dx \\ &\leq \int_\Omega |u'(x)|\,dx \\ &= \|u'\|_{\mathcal{M}(\Omega)}.\end{aligned}$$

Now, let $u \in BV(\Omega)$ and again take a series $\{u_j\}_{j\in\mathbb{N}} \subset C^\infty(\Omega)$, which satisfies (5.42). From the convergence $u_j \to u \in L^1(\Omega)$ it obviously follows that $\Upsilon_h u_j \to \Upsilon_h u \in L^1(\Omega)$ by definition of Υ_h. We use [3, Proposition 3.6.], the fact that (5.40) holds for every u_j, and (5.42) to obtain

$$\|(\Upsilon_h u)'\|_{\mathcal{M}(\Omega)} \leq \liminf_{j\to\infty} \|(\Upsilon_h u_j)'\|_{\mathcal{M}(\Omega)} \leq \liminf_{j\to\infty} \|u_j'\|_{\mathcal{M}(\Omega)} = \|u'\|_{\mathcal{M}(\Omega)}.$$

This shows (5.40). Finally, to prove (5.41) we use that, given $u \in W^{1,\infty}(\Omega)$, by Rademacher's Theorem (see e.g. [3, Theorem 2.14.]) the control u is Lipschitz-continuous with Lipschitz-constant $\|u'\|_{L^\infty(\Omega)}$. So, for any $i = 1, \dots, N$ and arbitrary but fixed $x \in I_i$, we get

$$\begin{aligned}u(x) - \Upsilon_h u_{|I_i} &= u(x) - \frac{1}{h_i}\int_{I_i} u(s)\,ds \\ &\leq \frac{\|u'\|_{L^\infty(\Omega)}}{h_i}\int_{I_i} |x - s|\,ds \\ &\leq h_i\|u'\|_{L^\infty(\Omega)}.\end{aligned}$$

Employing $h = \max_{1\leq i\leq N} h_i$ we derive (5.41). □

In Theorem 5.7 we already saw that (P_{vd}) has at least one solution. We now examine uniqueness of the solution.

Theorem 5.12. *There exists a unique solution $\bar{u} \in P_0$ to problem (P_{vd}). Furthermore, for every solution $\hat{u} \in BV(\Omega)$ of (P_{vd}) it holds $\Upsilon_h \hat{u} = \bar{u}$.*

Proof. For $\bar{u} \in P_0$ the mapping $\bar{u} \mapsto (\bar{y}_h, \bar{z}_h)$ is injective, so that the quadratic term in J_h now delivers strict convexity of J_h on P_0. Therefore, uniqueness of solution in the discrete space is evident. Also, for every solution $\hat{u} \in BV(\Omega)$ to (P_{vd}) the projection $\Upsilon_h \hat{u}$ is a discrete solution and due to uniqueness of the discrete solution, all projections must coincide. □

This directly delivers the following result.

Lemma 5.13. *Under Assumption 5.10 problem* (P_{vd}) *admits a unique solution, which is an element of* P_0.

5.3.1 Error estimates

Here, we do not need a structural assumption, for now. Later on, when proving the convergence rate of the optimal control, we will discuss structural assumptions.

Error estimates for mixed finite elements applied to elliptic partial differential equations have been proven e.g. in [11, 36, 38, 42, 64], but we consider a partial differential equation with a function of bounded variation on the right hand side. Furthermore, there exist many other works on error estimates for the mixed formulation of elliptic problems in 2D and 3D, but we consider $\Omega = (0, 1)$.

Analogous to [26, Section 3] and [42, Section 3] we introduce interpolators for the mixed finite element method. As in [32] we define the standard $L^2(\Omega)$-orthogonal projection $P_h : L^2(\Omega) \to P_0$, which satisfies: for any $w \in L^2(\Omega)$

$$(w - P_h w, w_h) = 0 \qquad \forall w_h \in P_0.$$

Furthermore, we recall the Fortin projection (see [11, 32]), defined as $\Pi_h : H^1(\Omega) \to P_1$, which satisfies: for any $v \in H^1(\Omega)$)

$$(\mathrm{div}(v - \Pi_h v), w_h) = 0 \qquad \forall w_h \in P_0.$$

The following diagram then commutes:

$$\begin{array}{ccc} H^1(\Omega) & \xrightarrow{\mathrm{div}} & L^2(\Omega) \\ {\scriptstyle \Pi_h}\downarrow & & \downarrow{\scriptstyle P_h} \\ P_1 & \xrightarrow{\mathrm{div}} & P_0 \end{array}$$

i.e., $\mathrm{div}\,\Pi_h = P_h\,\mathrm{div} : H^1(\Omega) \to P_0$. We collect the following approximation properties, e.g. from [42, Section 3]:

$$\begin{aligned} \|w - P_h w\|_{L^p(\Omega)} &\le Ch\|w'\|_{L^p(\Omega)} && \text{for } w \in W^{1,p}(\Omega), \\ \|v - \Pi_h v\|_{L^p(\Omega)} &\le Ch\|v'\|_{L^p(\Omega)} && \text{for } v \in W^{1,p}(\Omega), \\ \|\,\mathrm{div}(v - \Pi_h v)\|_{L^2(\Omega)} &\le Ch\|(\mathrm{div}\, v)'\|_{L^2(\Omega)} && \text{for } \mathrm{div}\, v \in H^1(\Omega). \end{aligned}$$

With these interpolators we prove the following a priori error estimate for the state analogous to [42, Lemma 4.3.].

Theorem 5.14. *Let* $(\bar{y}, \bar{z}, \bar{p}, \bar{q})$ *be the solution of* (5.15)-(5.18) *and let* $(\bar{y}_h, \bar{z}_h)$ *be the solution of* (5.27)*. Furthermore, let* $\Omega \subset \mathbb{R}$ *be a bounded convex polygonal domain with Lipschitz boundary* Γ*. Problem* (P) *is solved by* $\bar{u} \in BV(\Omega)$*. Let* $\bar{y} \in H_0^1(\Omega)$*, and* $\bar{p} \in H^2(\Omega) \cap H_0^1(\Omega)$ *be the associated optimal state and adjoint state, respectively. Then, we have*

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \le Ch. \tag{5.43}$$

Proof. Consider the following elliptic equation:

$$\begin{cases} -\phi'' &= g, \qquad \text{in } \Omega, \\ \phi &= 0, \qquad \text{on } \Gamma, \end{cases} \tag{5.44}$$

where $g \in L^2(\Omega)$.

Consider the mixed weak formulation of the above problem, such that $(\phi, \psi) = \mathcal{S}(g)$, see (5.4). Since Ω is convex, we have as in (5.6) that $(\phi, \psi) \in H^2(\Omega) \times H^1(\Omega)$ and

$$\|\phi\|_{H^2(\Omega)} + \|\psi\|_{H^1(\Omega)} \leq C\|g\|_{L^2(\Omega)}. \tag{5.45}$$

Now, by employing the equalities from the mixed formulation, we see

$$\begin{aligned}(\bar{y} - \bar{y}_h, g) &= -(\bar{y} - \bar{y}_h, \psi') \\ &= -(\bar{y} - \bar{y}_h, \psi') - (\psi, \bar{z} - \bar{z}_h) - (\phi, (\bar{z} - \bar{z}_h)') \\ &= -(\psi - \Pi_h\psi, \bar{z} - \bar{z}_h) - (\phi - P_h\phi, (\bar{z} - \bar{z}_h)') - (\bar{y} - \bar{y}_h, (\psi - \Pi_h\psi)'),\end{aligned}$$

with P_h and Π_h the special interpolators introduced before. We also have

$$\|\bar{u}\|_{L^2(\Omega)} \leq C\|\bar{u}\|_{BV(\Omega)} \leq C.$$

This follows from the continuity of the embedding $BV(\Omega) \hookrightarrow L^2(\Omega)$ and the boundedness of the optimal control in the BV-norm, see the proof of Theorem 5.4. We use this to estimate

$$(\psi - \Pi_h\psi, \bar{z} - \bar{z}_h) \leq C\|\psi - \Pi_h\psi\|_{L^2(\Omega)}\|\bar{z} - \bar{z}_h\|_{H^1(\Omega)} \leq Ch\|\psi\|_{H^1(\Omega)}\|\bar{u}\|_{L^2(\Omega)} \leq Ch\|g\|_{L^2(\Omega)},$$

and

$$(\phi - P_h\phi, (\bar{z} - \bar{z}_h)') \leq C\|\phi - P_h\phi\|_{L^2(\Omega)}\|\bar{z} - \bar{z}_h\|_{H^1(\Omega)} \leq Ch\|\phi\|_{H^1(\Omega)}\|\bar{u}\|_{L^2(\Omega)} \leq Ch\|g\|_{L^2(\Omega)}.$$

Also, by the definition of Π_h and employing (5.6), we get

$$\begin{aligned}(\bar{y} - \bar{y}_h, (\psi - \Pi_h\psi)') &= (\bar{y} - P_h\bar{y}, (\psi - \Pi_h\psi)') + \underbrace{(P_h\bar{y} - \bar{y}_h, (\psi - \Pi_h\psi)')}_{=0} \\ &\leq C\|\bar{y} - P_h\bar{y}\|_{L^2(\Omega)}\|(\psi - \Pi_h\psi)'\|_{L^2(\Omega)} \\ &\leq Ch\|\bar{y}\|_{H^1(\Omega)}\|\psi\|_{H^1(\Omega)} \\ &\leq Ch\|\bar{u}\|_{L^2(\Omega)}\|g\|_{L^2(\Omega)} \\ &\leq Ch\|g\|_{L^2(\Omega)}.\end{aligned}$$

Altogether this delivers

$$(\bar{y} - \bar{y}_h, g) \leq Ch\|g\|_{L^2(\Omega)},$$

and therefore

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega)} = \sup_{g \in L^2(\Omega), g \neq 0} \frac{(\bar{y} - \bar{y}_h, g)}{\|g\|_{L^2(\Omega)}} \leq Ch.$$

□

We move on to establish an error estimate for the adjoint state. Let us remark that in the given problem setting, the solution operators for the mixed formulation of the state equation $\mathcal{S}$ and for the discrete state equation $\mathcal{S}_h$ coincide with the respective adjoint operators $\mathcal{S}^*$ and $\mathcal{S}_h^*$. Consequently, finite element error estimates that can be found in the literature for the mixed formulation of the state equation also apply to the mixed formulation of the adjoint state equation.

Theorem 5.15. *Let $\Omega \subset \mathbb{R}$ be a bounded convex polygonal domain with Lipschitz boundary Γ and problem (P) be solved by $\bar{u} \in BV(\Omega)$ with the associated optimal state $\bar{y} \in H_0^1(\Omega)$. Furthermore, let $(\bar{p}, \bar{q}) = \mathcal{S}^*(\bar{y} - y_\mathrm{d})$ and $(\bar{p}_h, \bar{q}_h) = \mathcal{S}_h^*(\bar{y}_h - y_\mathrm{d})$. Assume that $\bar{p} \in W^{1,\infty}(\Omega)$. Then, we have*

$$\|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} \leq Ch.$$

Proof. For notation purposes we write $\bar{p} = \mathcal{S}_1^*(\bar{y} - y_\mathrm{d})$ and $\bar{p}_h = \mathcal{S}_{h,1}^*(\bar{y}_h - y_\mathrm{d})$. With the properties of P_h, given desired state $y_\mathrm{d} \in L^\infty(\Omega)$, the continuous embedding $BV(\Omega) \hookrightarrow L^\infty(\Omega)$, and the bound $\|\bar{u}\|_{BV(\Omega)} \leq C$ from the proof of Theorem 5.4, we obtain

$$\begin{aligned}
\|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} &\leq \|\bar{p} - P_h\bar{p}\|_{L^\infty(\Omega)} + \|P_h\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} \\
&\leq Ch\|\bar{p}'\|_{L^\infty(\Omega)} + \|P_h\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} \\
&\leq Ch\|\bar{p}\|_{W^{1,\infty}(\Omega)} + \|P_h\mathcal{S}_1^*(\bar{y} - y_\mathrm{d}) - \mathcal{S}_{h,1}^*(\bar{y}_h - y_\mathrm{d})\|_{L^\infty(\Omega)} \\
&\leq Ch\|\bar{y} - y_\mathrm{d}\|_{L^\infty(\Omega)} + \|\mathcal{S}_{h,1}^*(\bar{y} - \bar{y}_h)\|_{L^\infty(\Omega)} \\
&\leq Ch(\|\bar{y}\|_{L^\infty(\Omega)} + \|y_\mathrm{d}\|_{L^\infty(\Omega)}) + \|\mathcal{S}_{h,1}^*(\bar{y} - \bar{y}_h) - \mathcal{S}_1^*(\bar{y} - \bar{y}_h)\|_{L^\infty(\Omega)} + \|\mathcal{S}_1^*(\bar{y} - \bar{y}_h)\|_{L^\infty(\Omega)} \\
&\leq Ch\|\bar{u}\|_{L^\infty(\Omega)} + Ch\|\bar{y} - \bar{y}_h\|_{L^\infty(\Omega)} + C\|\bar{y} - \bar{y}_h\|_{L^\infty(\Omega)} \\
&\leq Ch,
\end{aligned}$$

where we used [36, Corollary 6.1.] in the last two inequalities. □

Remark 5.16. *For higher dimensional mixed formulation approaches the error in the L^∞-norm is of order $O(h|\log h|)$. For further details we refer to [36].*

With the result for the adjoint state it is easy to see the following error estimate for the multiplier.

Lemma 5.17. *Let the conditions of Theorem 5.15 hold. Then, we have*

$$\|\bar{\Phi} - \bar{\Phi}_h\|_{L^\infty(\Omega)} \leq Ch.$$

Proof. By inserting the definitions $\bar{\Phi}(x) = \int_0^x \bar{p}(s)\,ds$ and $\bar{\Phi}_h = \int_0^x \bar{p}_h(s)\,ds$ it follows directly that

$$\|\bar{\Phi} - \bar{\Phi}_h\|_{L^\infty(\Omega)} \leq \|\bar{p} - \bar{p}_h\|_{L^1(\Omega)},$$

and due to $|\Omega| = 1$, we also have

$$\|\bar{p} - \bar{p}_h\|_{L^1(\Omega)} \leq \|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)}.$$

With $\|\bar{p} - \bar{p}_h\|_{L^\infty(\Omega)} \leq Ch$ the claim follows. □

Finally, under structural assumptions, we prove an error estimate for the control. Although the control is not discretized in the variational discretization approach, we will denote the solution to (P_vd) by $\bar{u}_h$ for clarity from here on. We remark that we can not expect a better convergence order than $\|\bar{u} - \bar{u}_h\|_{L^1(\Omega)} = O(h)$, because if we fix $\tilde{x} \in \Omega$ and consider $\bar{u} = 1_{(\tilde{x},1)}$, then it holds $\|\bar{u} - 1_{(x_i,1)}\|_{L^1(\Omega)} = |\tilde{x} - x_i| = O(h)$ for any node $x_i \in \Omega$. We will prove this order of convergence and see a numerical confirmation of our result in Section 5.4.

Let Assumption 5.10 hold and make the following additional assumption:

Assumption 5.18. *Suppose that $\left\{x \in \Omega : |\bar{\Phi}(x)| = \alpha\right\}$ is finite. Then there exists $m \in \mathbb{N}_0$, such that*

$$\left\{x \in \Omega : |\bar{\Phi}(x)| = \alpha\right\} = \{\hat{x}_1, \ldots, \hat{x}_m\},$$

with $m = 0$ indicating that these sets are empty.

From Lemma 5.6 we deduce that the support of $\bar{u}'$ is finite and we can express $\bar{u}$ as follows with $\bar{a} \in \mathbb{R}$ and $\bar{c} = (\bar{c}^1, \ldots, \bar{c}^m)^\top \in \mathbb{R}^m$:

$$\bar{u} = \bar{a} + \sum_{i=1}^{m} \bar{c}^i 1_{(\hat{x}_i,1)},$$

where some coefficients may be zero.

To obtain a convergence result, we need to estimate the difference in the jump points of the optimal control and the corresponding coefficients. We begin by analyzing the extremal points of the discrete multiplier $\bar{\Phi}_h$, which will deliver information about the support of the variational discrete optimal control $\bar{u}_h$.

For $i = 1, \dots, m$ we have $\hat{x}_i \in \Omega$, since $\bar{\Phi}(x) = 0$ for $x \in \Gamma$ and $|\bar{\Phi}(\hat{x}_i)| = \alpha > 0$. There exists $R > 0$, such that $B_R(\hat{x}_i) \subset \Omega$ and all $B_R(\hat{x}_i)$ are pairwise disjoint.

Let h_0, such that $h_0 < \frac{R}{2}$ and $m \leq N$ holds for all $h \in (0, h_0]$. Then for every $i = 1, \dots, m$ we find the closest grid points $x_{i,l}, x_{i,r}$, such that $x_{i,l} \leq \hat{x}_i \leq x_{i,r}$ and $x_{i,l}, x_{i,r} \in B_R(\hat{x}_i)$.

Now, with $|\bar{\Phi}_h(x)| \leq \alpha, \bar{\Phi}_h \in P_1$ and taking Assumption 5.10 into account, it holds either $|\bar{\Phi}_h(x_{i,l})| = \alpha$ or $|\bar{\Phi}_h(x_{i,r})| = \alpha$. So, we can find a unique grid point $x_{j(i)}$, with $j(i) \in \{1, \dots, N\}$ associated with $\hat{x}_i$ for every $i = 1, \dots, m$. And we have chosen h_0 small enough, such that every grid point is associated with at most one jump point $\hat{x}_i$. Since $x_{j(i)}$ is a neighboring node of $\hat{x}_i$, we always have that

$$|\hat{x}_i - x_{j(i)}| \leq h. \tag{5.46}$$

Furthermore, we show that $|\bar{\Phi}_h(x)| < \alpha$ for all $x \in \bar{\Omega} \setminus \cup_{i=1}^m B_R(\hat{x}_i)$, so that $\bar{u}_h$ can be represented as follows with $\bar{a}_h \in \mathbb{R}$ and $\bar{c}_h = (\bar{c}_h^{j(1)}, \dots, \bar{c}_h^{j(m)})^\top \in \mathbb{R}^m$:

$$\bar{u}_h = \bar{a}_h + \sum_{i=1}^m \bar{c}_h^{j(i)} 1_{(x_{j(i)}, 1)}.$$

It holds $|\bar{\Phi}_h(x)| \leq \alpha$ for all $x \in \bar{\Omega}$, so it is sufficient to show that $|\bar{\Phi}_h(x)| = \alpha$ can not be satisfied for $x \in \bar{\Omega} \setminus \cup_{i=1}^m B_R(\hat{x}_i)$. We know that $|\bar{\Phi}|$ is continuous on the compact set $\bar{\Omega} \setminus \cup_{i=1}^m B_R(\hat{x}_i)$, so it attains a maximum on this set. Since $|\bar{\Phi}(x)| \leq \alpha$ for all $x \in \bar{\Omega}$ holds and under Assumption 5.18 $|\bar{\Phi}(x)| = \alpha$ is not attained in $\bar{\Omega} \setminus \cup_{i=1}^m B_R(\hat{x}_i)$, we know that this maximum is smaller than α. Consequently, there exists $\epsilon > 0$, such that $|\bar{\Phi}(x)| \leq \alpha - \epsilon$ for all $x \in \bar{\Omega} \setminus \cup_{i=1}^m B_R(\hat{x}_i)$. With Lemma 5.17 we see $|\bar{\Phi}_h(x)| \leq \alpha - \frac{\epsilon}{2}$ for all $x \in \bar{\Omega} \setminus \cup_{i=1}^m B_R(\hat{x}_i)$, since $h < \frac{R}{2}$.

Next, we estimate the differences in the jump heights and the constant coefficient.

Lemma 5.19. *Let Assumption 5.10 and Assumption 5.18 hold. Then there exists $h_0 > 0$, such that for all $h \in (0, h_0]$ the coefficients of the optimal controls $\bar{u} = \bar{a} + \sum_{i=1}^m \bar{c}^i 1_{(\hat{x}_i, 1)}$ and $\bar{u}_h = \bar{a}_h + \sum_{i=1}^m \bar{c}_h^{j(i)} 1_{(x_{j(i)}, 1)}$ satisfy*

$$\sum_{i=1}^m |\bar{c}^i - \bar{c}_h^{j(i)}| \leq Ch, \tag{5.47}$$

$$|\bar{a} - \bar{a}_h| \leq Ch. \tag{5.48}$$

Proof. We know that there exists a $R > 0$, such that the balls $B_{\frac{3}{4}R}(\hat{x}_i)$ are contained in Ω and are pairwise disjoint for $i = 1, \dots, m$. For every $i = 1, \dots, m$ we proceed as follows: Consider a function $g \in C_c^\infty(\Omega)$, such that $g = 1$ on $B_{\frac{R}{2}}(\hat{x}_i)$ and $g = 0$ on $\bar{\Omega} \setminus \cup_{i=1}^m B_{\frac{3}{4}R}(\hat{x}_i)$. For h small enough we also have $x_{j(i)} \in B_{\frac{R}{2}}(\hat{x}_i)$ for every $i = 1, \dots, m$. We have

$$\bar{u}' = \sum_{i=1}^m \bar{c}^i \delta_{\hat{x}_i} \qquad \text{and} \qquad \bar{u}_h' = \sum_{i=1}^m \bar{c}_h^{j(i)} \delta_{x_{j(i)}},$$

so that by definition of g, definition of the distributional derivative, and the definition of the state equation we get for all $h \in (0, h_0]$

$$\begin{aligned} |\bar{c}^i - \bar{c}_h^{j(i)}| &= \left| \langle \bar{u}' - \bar{u}_h', g \rangle_{\mathcal{M}(\Omega), C(\Omega)} \right| \\ &= \left| -(\bar{u} - \bar{u}_h, g')_{L^2(\Omega)} \right| \\ &\leq \left| (\bar{u} - \bar{u}_h, P_h(g'))_{L^2(\Omega)} \right| + \left| (\bar{u} - \bar{u}_h, g' - P_h(g'))_{L^2(\Omega)} \right|. \end{aligned}$$

The second term can be estimated as follows:

$$\begin{aligned}\left|(\bar{u}-\bar{u}_h, g'-P_h(g'))_{L^2(\Omega)}\right| &\leq \|\bar{u}-\bar{u}_h\|_{L^2(\Omega)}\|g'-P_h(g')\|_{L^2(\Omega)}\\ &\leq \left(\|\bar{u}\|_{L^2(\Omega)}+\|\bar{u}_h\|_{L^2(\Omega)}\right)Ch\|g''\|_{L^2(\Omega)}\\ &\leq Ch,\end{aligned}$$

where we use the definition of P_h, that we have the compact embedding $BV(\Omega) \hookrightarrow L^2(\Omega)$, and the bounds $\|\bar{u}\|_{BV(\Omega)} \leq C$ and $\|\bar{u}_h\|_{BV(\Omega)} \leq C$. The latter bound has been proven in Theorem 5.7 for h small enough, so if necessary, we reduce h_0.

For the first term we use the definition of a, the definition of P_h and Theorem 5.14 to obtain

$$\begin{aligned}\left|(\bar{u}-\bar{u}_h, P_h(g'))_{L^2(\Omega)}\right| &= |a(\bar{y}-\bar{y}_h, P_h(g'))|\\ &\leq |a(\bar{y}, P_h(g')-g')| + |a(\bar{y}-\bar{y}_h, g')|\\ &= |(\bar{u}, P_h(g')-g')_{L^2(\Omega)}| + \left|\int_\Omega (\bar{y}-\bar{y}_h)g''\,dx\right|\\ &\leq \|\bar{u}\|_{L^2(\Omega)}\|P_h(g')-g'\|_{L^2(\Omega)} + \|\bar{y}-\bar{y}_h\|_{L^2(\Omega)}\|g''\|_{L^2(\Omega)}\\ &\leq Ch.\end{aligned}$$

Together we see $|\bar{c}^i - \bar{c}_h^{j(i)}| \leq Ch$ for every $i = 1, \dots, m$, which delivers (5.47).

To see (5.48), it suffices to adapt the proof of [43, Lemma 4.9.] to our setting and then insert the error estimate for the state

$$\|\bar{y}-\bar{y}_h\|_{L^2(\Omega)} \leq Ch$$

from Theorem 5.14. □

With the previous results we now have everything available to prove the convergence order $O(h)$ for the optimal control.

Theorem 5.20. *Let Assumption 5.10 and Assumption 5.18 hold. Then there exists $h_0 > 0$, such that for all $h \in (0, h_0]$ we have*

$$\|\bar{u}-\bar{u}_h\|_{L^1(\Omega)} \leq Ch.$$

Proof. We combine $|\Omega| = 1$, (5.46), (5.47), and (5.48) to get

$$\begin{aligned}\|\bar{u}-\bar{u}_h\|_{L^1(\Omega)} &= \int_\Omega \left|\bar{a}-\bar{a}_h + \sum_{i=1}^m \left(\bar{c}^i 1_{(\hat{x}_i,1)} - \bar{c}_h^{j(i)} 1_{(x_{j(i)},1)}\right)\right| dx\\ &\leq |\bar{a}-\bar{a}_h||\Omega| + \int_\Omega \left|\sum_{i=1}^m \bar{c}^i\left(1_{(\hat{x}_i,1)} - 1_{(x_{j(i)},1)}\right)\right| dx + \int_\Omega \left|\sum_{i=1}^m \left(\bar{c}^i - \bar{c}_h^{j(i)}\right) 1_{(x_{j(i)},1)}\right| dx\\ &\leq |\bar{a}-\bar{a}_h| + \sum_{i=1}^m |\bar{c}^i| \underbrace{\|1_{(\hat{x}_i,1)} - 1_{(x_{j(i)},1)}\|_{L^1(\Omega)}}_{=|\hat{x}_i - x_{j(i)}|} + \sum_{i=1}^m |\bar{c}^i - \bar{c}_h^{j(i)}|\,\|1_{(x_{j(i)},1)}\|_{L^1(\Omega)}\\ &\leq Ch.\end{aligned}$$

□

We remark that it may be possible to prove the convergence order for the optimal control in the L^1-norm only using Assumption 5.18. Then, Assumption 5.10 for the discrete multiplier $\bar{\Phi}_h$ would be obsolete.

5.4 Computational results

We can represent the mixed formulation of the discrete state equation (5.27) by the following matrix equation:

$$\begin{pmatrix} A & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \mathbf{z} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} 0 \\ -\mathbf{u} \end{pmatrix}, \tag{5.49}$$

where

$$A = \begin{pmatrix} \frac{1}{3}h_1 & \frac{1}{6}h_1 & 0 & \dots & 0 \\ \frac{1}{6}h_1 & \frac{1}{3}(h_1+h_2) & \frac{1}{6}h_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \frac{1}{6}h_{N-1} & \frac{1}{3}(h_{N-1}+h_N) & \frac{1}{6}h_N \\ 0 & \dots & 0 & \frac{1}{6}h_N & \frac{1}{3}h_N \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \qquad B = \begin{pmatrix} -1 & 0 & \dots & 0 \\ 1 & -1 & & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & & \ddots & -1 \\ 0 & \dots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(N+1)\times N},$$

with the vectors containing the coefficients : $\mathbf{z} = \left(z_0, \dots, z_N\right)^\top \in \mathbb{R}^{N+1}, \mathbf{y} = \left(y_1, \dots, y_N\right)^\top \in \mathbb{R}^N$, and the evaluation of the BV-function $\mathbf{u} = \left(u_1, \dots, u_N\right)^\top \in \mathbb{R}^N$. Here, $u_j := \int_{x_{j-1}}^{x_j} u$ for $j = 1, \dots, N$. With our knowledge about the structure of u we get $u_j = (a_h + \sum_{i=1}^{j-1} c_h^i)h_j$ for $j = 1, \dots, N$.

We use (5.49) to get

$$\begin{aligned} A\mathbf{z} + B\mathbf{y} = 0 &\Rightarrow \mathbf{z} = -A^{-1}B\mathbf{y}, \\ B^\top\mathbf{z} = -\mathbf{u} &\Rightarrow B^\top A^{-1}B\mathbf{y} = \mathbf{u} \\ &\Rightarrow \mathbf{y} = (B^\top A^{-1}B)^{-1}\mathbf{u}, \end{aligned}$$

and then insert this into (P_h) to obtain:

$$\min_{a_h\in\mathbb{R}, c_h\in\mathbb{R}^{N-1}} f(a_h, c_h) := \underbrace{\frac{1}{2}\|(B^\top A^{-1}B)^{-1}\mathbf{u} - y_\mathrm{d}\|^2_{L^2(\Omega)}}_{=:f_1(a_h,c_h)} + \alpha \sum_{i=1}^{N-1} |c_h^i|, \tag{$\hat{P}_h$}$$

where $\mathbf{u} = \mathbf{u}(a_h, c_h)$.

5.4.1 Semismooth Newton method

In the following we explain how to solve $(\hat{P}_h)$ by a semismooth Newton method: The representative vector of the adjoint $\mathbf{p}$ can be calculated using the following matrix equation:

$$\begin{pmatrix} A & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{y}_\mathrm{d} - \mathbf{y} \end{pmatrix}, \tag{5.50}$$

which gives

$$\mathbf{p} = (B^\top A^{-1}B)^{-1}(\mathbf{y} - \mathbf{y}_\mathrm{d}).$$

The optimality system for the unconstrained problem $(\hat{P}_h)$ then reads

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_h} f(a_h, c_h) = \frac{\partial}{\partial a_h} f_1(a_h, c_h), \\ 0 &= \frac{\partial}{\partial c_h^j} f_1(a_h, c_h) + \lambda_j && \forall j = 1, \dots, N-1, \\ 0 &= c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) && \forall j = 1, \dots, N-1, \end{aligned}$$

with an arbitrary $\gamma > 0$.

The third condition is a complementarity condition that originates from the generalized Jacobian

$$\alpha \frac{\partial}{\partial c_h^j} |c_h^j| \begin{cases} = \alpha, & c_h^j > 0, \\ = -\alpha, & c_h^j < 0, \\ \in [-\alpha, \alpha], & c_h^j = 0. \end{cases}$$

The following Lemma is a simplified version of [69, Lemma 2.2], which we give here for the sake of completeness.

Lemma 5.21. *The equation*

$$c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) = 0 \tag{5.51}$$

holds iff

$$(c_h^j, \lambda_j) \text{ satisfies } \begin{cases} \lambda_j = \alpha & \text{a.e. on } \{x \in \Omega : c_h^j > 0\}, \\ |\lambda_j| \leq \alpha & \text{a.e. on } \{x \in \Omega : c_h^j = 0\}, \\ \lambda_j = -\alpha & \text{a.e. on } \{x \in \Omega : c_h^j < 0\}. \end{cases} \tag{5.52}$$

Proof.
"(5.51) $\Rightarrow$ (5.52)"
Let $c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) = 0$, then we have the following cases:
1^{st} *case:* $c_h^j + \gamma(\lambda_j - \alpha) > 0$
This implies $c_h^j + \gamma(\lambda_j + \alpha) > 0$, so we have

$$\begin{aligned} & 0 = c_h^j - c_h^j - \gamma(\lambda_j - \alpha)) \\ \Rightarrow \qquad & \lambda_j = \alpha. \end{aligned}$$

So from $c_h^j + \gamma(\lambda_j - \alpha) > 0$ it is obvious that $c_h^j > 0$, which gives the first case of (5.52).
2^{nd} *case:* $c_h^j + \gamma(\lambda_j - \alpha) \leq 0$ *and* $c_h^j + \gamma(\lambda_j + \alpha) \geq 0$
From (5.51) we have directly $c_h^j = 0$ and therefore we get from the case assumption that

$$\lambda_j - \alpha \leq 0 \qquad \wedge \qquad \lambda_j + \alpha \geq 0 \qquad \Rightarrow \qquad |\lambda_j| \leq \alpha,$$

which gives the second case of (5.52).
3^{rd} *case:* $c_h^j + \gamma(\lambda_j + \alpha) < 0$
This implies $c_h^j + \gamma(\lambda_j - \alpha) < 0$, so we have

$$\begin{aligned} & 0 = c_h^j - c_h^j - \gamma(\lambda_j + \alpha) \\ \Rightarrow \qquad & \lambda_j = -\alpha. \end{aligned}$$

So from $c_h^j + \gamma(\lambda_j + \alpha) < 0$ it is obvious that $c_h^j < 0$, which gives the third case of (5.52).
"(5.51) $\Leftarrow$ (5.52)"
Here, we also look at three different cases:
1^{st} *case:* $\lambda_j = \alpha$ *and* $c_h^j > 0$

$$c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) = c_h^j - c_h^j - \gamma(\lambda_j - \alpha) = 0.$$

2^{nd} *case:* $|\lambda_j| \leq \alpha$ *and* $c_h^j = 0$

$$c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) = -\max(0, \gamma \underbrace{(\lambda_j - \alpha)}_{\leq 0}) - \min(0, \gamma \underbrace{(\lambda_j + \alpha)}_{\geq 0}) = 0.$$

3rd case: $\lambda_j = -\alpha$ *and* $c_h^j < 0$

$$c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) = c_h^j - c_h^j - \gamma(\lambda_j + \alpha) = 0.$$

So, we can verify (5.51) in all cases. □

We now get the following optimality system:

$$(B^\top A^{-1} B)\mathbf{y} - \mathbf{u} = 0, \tag{5.53}$$

$$(B^\top A^{-1} B)\mathbf{p} + \mathbf{y}_\mathrm{d} - \mathbf{y} = 0, \tag{5.54}$$

$$\left(\mathbf{y} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1} \frac{\partial}{\partial a_h}\mathbf{u}\right)_{L^2(\Omega)} = 0, \tag{5.55}$$

$$\left(\mathbf{y} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1} \frac{\partial}{\partial c_h^j}\mathbf{u}\right)_{L^2(\Omega)} + \lambda_j = 0 \qquad \forall\, j = 1, \dots, N-1, \tag{5.56}$$

$$c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) = 0 \qquad \forall\, j = 1, \dots, N-1. \tag{5.57}$$

Here, $\mathbf{u} = \mathbf{u}(a_h, c_h)$ with $\mathbf{u} = (u_1, \dots, u_N)^\top$ and $u_j = (a_h + \sum_{i=1}^{j-1} c_h^i)\, h_j$. Consequently, we get

$$\frac{\partial}{\partial a_h}\mathbf{u} = \begin{pmatrix} h_1 \\ \vdots \\ h_N \end{pmatrix} =: \mathbf{h} \qquad \text{and} \qquad \frac{\partial}{\partial c_h^j}\mathbf{u} = \begin{pmatrix} 0 & \dots & 0 & h_{j+1} & \dots & h_N \end{pmatrix}^\top =: \mathbf{h}_{j+1} \qquad \text{for } j = 1, \dots, N-1.$$

Plugging in the above derivatives and inserting (5.53) into (5.55) and (5.56), we get

$$F(a_h, c_h, \lambda) = \begin{pmatrix} \left((B^\top A^{-1} B)^{-1}\mathbf{u} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1}\mathbf{h}\right)_{L^2(\Omega)} \\ \left[\left((B^\top A^{-1} B)^{-1}\mathbf{u} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1}\mathbf{h}_{j+1}\right)_{L^2(\Omega)} + \lambda_j\right]_{j=1}^{N-1} \\ \left[c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha))\right]_{j=1}^{N-1} \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Then we choose the following matrix from the set of generalized Jacobian:

$$DF(a_h, c_h, \lambda) = \begin{pmatrix} DF_1 & DF_2 & 0 \\ DF_2^\top & DF_3 & DF_4 \\ 0 & DF_5 & DF_6 \end{pmatrix}.$$

The first derivatives are unique:

$$\begin{aligned}
DF_1 &= \frac{\partial}{\partial a_h}\left((B^\top A^{-1} B)^{-1}\mathbf{u} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1}\mathbf{h}\right)_{L^2(\Omega)} \\
&= \|(B^\top A^{-1} B)^{-1}\mathbf{h}\|^2_{L^2(\Omega)}, \\
DF_2 &= \left[\frac{\partial}{\partial c_h^j}\left((B^\top A^{-1} B)^{-1}\mathbf{u} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1}\mathbf{h}\right)_{L^2(\Omega)}\right]_{j=1}^{N-1} \\
&= \left[\left((B^\top A^{-1} B)^{-1}\mathbf{h}_{j+1}, (B^\top A^{-1} B)^{-1}\mathbf{h}\right)_{L^2(\Omega)}\right]_{j=1}^{N-1}, \\
DF_3 &= \left[\frac{\partial}{\partial c_h^k}\left(\left((B^\top A^{-1} B)^{-1}\mathbf{u} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1}\mathbf{h}_{j+1}\right)_{L^2(\Omega)} + \lambda_j\right)\right]_{j,k=1}^{N-1} \\
&= \left[\left((B^\top A^{-1} B)^{-1}\mathbf{h}_{k+1}, (B^\top A^{-1} B)^{-1}\mathbf{h}_{j+1}\right)_{L^2(\Omega)}\right]_{j,k=1}^{N-1},
\end{aligned}$$

$$DF_4 = \left[\frac{\partial}{\partial \lambda_k} \left(\left((B^\top A^{-1} B)^{-1} \mathbf{u} - \mathbf{y}_\mathrm{d}, (B^\top A^{-1} B)^{-1} \mathbf{h}_{j+1} \right)_{L^2(\Omega)} + \lambda_j \right) \right]_{j,k=1}^{N-1}$$
$$= \mathbf{I}_{N-1}.$$

In the following parts we make a choice:

$$DF_5 = \left[\frac{\partial}{\partial c_h^k} \left(c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) \right) \right]_{j,k=1}^{N-1}$$
$$= \begin{cases} \delta_{kj}, & \text{for } c_h^j + \gamma(\lambda_j - \alpha) \leq 0 \quad \text{and} \quad c_h^j + \gamma(\lambda_j + \alpha) \geq 0, \\ 0, & \text{else.} \end{cases}$$
$$DF_6 = \left[\frac{\partial}{\partial \lambda_k} \left(c_h^j - \max(0, c_h^j + \gamma(\lambda_j - \alpha)) - \min(0, c_h^j + \gamma(\lambda_j + \alpha)) \right) \right]_{j,k=1}^{N-1}$$
$$= \begin{cases} -\gamma\, \delta_{kj}, & \text{for } c_h^j + \gamma(\lambda_j - \alpha) > 0 \quad \text{or} \quad c_h^j + \gamma(\lambda_j + \alpha) < 0, \\ 0, & \text{else.} \end{cases}$$

5.4.2 Optimization algorithm

In contrast to [43] we know that the support of $\bar{u}_h'$ is a subset of the grid points $\{x_i\}_{i=1}^N$, so we don't need to approximate the support like it was done there. We start the algorithm with an empty support set and then update the set of support points in each outer iteration, where we will determine the grid points, in which the control is actually supported.

We define m_k as the cardinality of support points in iteration k and t_k the sorted vector of all support points in iteration k. The outer iteration should be terminated if the support points satisfy

$$m_k = m_{k-1} \qquad \text{and} \qquad \|t_k - t_{k-1}\|_2 \leq \epsilon. \tag{T_1}$$

Here, the second condition only needs to be checked if the first condition is fulfilled, to ensure that the support points are identical in both iterations. In [43] cycling of the outer iteration is reported. We also observe this and therefore insert a second set of termination conditions:

$$m_k = m_{k-1} = m_{k-2} \qquad \text{and} \qquad \|t_k - t_{k-2}\|_2 \leq \epsilon \qquad \text{and} \qquad f(u_h^k, y_h^k) < f(u_h^{k-1}, y_h^{k-1}). \tag{T_2}$$

Here, by $f(u_h^k, y_h^k)$ we denote the target function in $(\hat{P}_h)$. This leads to the following algorithm to solve $(\hat{P}_h)$:

Algorithm 5.22:

input : $m_0 \in \mathbb{R}, t_0 \in \mathbb{R}^{m_0}, \epsilon > 0$
for $k = 0, 1, \ldots$ **do**
  if (T_1) *or* (T_2) *holds* **then**
    $m := m_k$, $\bar{x}_h := t_k$,
    extract $(\bar{a}_h, \bar{c}_h)$ from u_h^k
    STOP
  Obtain (u_h^k, y_h^k, p_h^k) by solving $(\hat{P}_h)$.
  Compute $t_{k+1} \in \mathbb{R}^{m_{k+1}}$ from p_h^k.
output: $\bar{x}_h \in \mathbb{R}^m, (\bar{a}_h, \bar{c}_h) \in \mathbb{R}^{m+1}$

We initialize our algorithm with $\bar{a}_h = 0, \bar{c}_h = \{\}, \epsilon = 10^{-10}$ and solve $(\hat{P}_h)$ using the MATLAB routine 'fmincon' with the following choices: Algorithm: 'active-set'; MaxFunctionEvaluations: 10^5; MaxIterations: 10^4; FunctionTolerance: 10^{-12}, which will compute highly accurate solutions, since we want to display the order of convergence.

The MATLAB routine 'fmincon' with the above choices and the semismooth Newton method (see Subsection 5.4.1) executed up to tolerance 10^{-12} deliver similarly precise results to problem $(\hat{P}_h)$. We choose the stable and reliable routine 'fmincon' and accept the additional computation time.

5.4.3 Numerical Examples

As our first example, we consider [43, 5.3. Example 1] with known solution, which satisfies the optimality conditions as stated in Theorem 5.8, and has the following quantities:

- $c := 12 - 4\sqrt{8}$; $x_c := \frac{1}{2\pi}\arccos(\frac{c}{4})$;
- $\alpha := 10^{-5}$;
- $\bar{u} := 0.5 + 1_{(x_c,1)} - 2 \cdot 1_{(0.5,1)} + 1.5 \cdot 1_{(1-x_c,1)}$;
- $\bar{y} := \mathcal{S}(\bar{u}, 0)$;
- $\bar{\Phi}(x) := \frac{\alpha}{2c}\left[(1 - \cos(4\pi x)) - c(1 - \cos(2\pi x))\right]$;
- $\bar{p} := \bar{\Phi}'$;
- $y_\mathrm{d} := \bar{y} + \bar{p}''$.

In Figure 5.1 the approximated solutions on a grid with $h = \frac{1}{2048}$ are depicted.

In Figure 5.2 the errors between the known solutions and the solutions to the variationally discretized problem are displayed. We observe that the order of convergence is approximately h, except for $\|\bar{u} - \bar{u}_h\|_{L^2(Q)}$, which converges with a slower rate. These results align with our findings from Subsection 5.3.1.

In addition to plotting the errors, we also calculate the convergence order h^α for the refinement from some gridsize h_1 to some other gridsize h_2, see Table 5.1, by

$$\alpha = \frac{\log(\frac{e_{h_1}}{e_{h_2}})}{\log(\frac{h_1}{h_2})},$$

where e_{h_1} and e_{h_2} act as placeholders for the different errors we are examining, in particular: $\|\bar{u} - \bar{u}_h\|_{L^1(Q)}$, $\|\bar{u} - \bar{u}_h\|_{L^2(Q)}$, $\|\bar{y} - \bar{y}_h\|_{L^2(Q)}$, $\|\bar{p} - \bar{p}_h\|_{L^\infty(Q)}$, and $\|\bar{\Phi} - \bar{\Phi}_h\|_{L^\infty(Q)}$.

h_1	h_2	$\|\bar{u} - \bar{u}_h\|_{L^1(Q)}$	$\|\bar{u} - \bar{u}_h\|_{L^2(Q)}$	$\|\bar{y} - \bar{y}_h\|_{L^2(Q)}$	$\|\bar{p} - \bar{p}_h\|_{L^\infty(Q)}$	$\|\bar{\Phi} - \Phi_h\|_{L^\infty(Q)}$
0.25	0.125	0.1943	-0.0171	0.2403	-0.2224	-0.4324
0.125	0.0625	1.3436	0.7759	1.4444	1.1389	2.6278
0.0625	0.03125	1.1471	0.9368	1.7284	0.8966	1.5183
0.03125	0.015625	0.9982	0.4874	1.0286	1.0597	0.7761
0.015625	0.0078125	1.4732	2.7648	-0.1603	0.4420	-2.0774
0.0078125	0.00390625	0.0178	-2.3127	1.6393	1.4590	3.8838
0.00390625	0.001953125	0.9832	0.4948	0.9920	0.9936	1.3328
0.001953125	0.0009765625	0.9975	0.4975	0.9957	1.0184	0.3887
0.0009765625	0.00048828125	0.9353	0.4984	0.9025	0.9738	-0.6235
	mean	0.8989	0.4584	0.9790	0.8622	0.8216
	slope of best fit	0.9307	0.4854	1.0089	0.9241	0.9608

Table 5.1: Example 1: Convergence order (potency of gridsize h) of the respective errors when the grid is refined from gridsize h_1 to gridsize h_2.

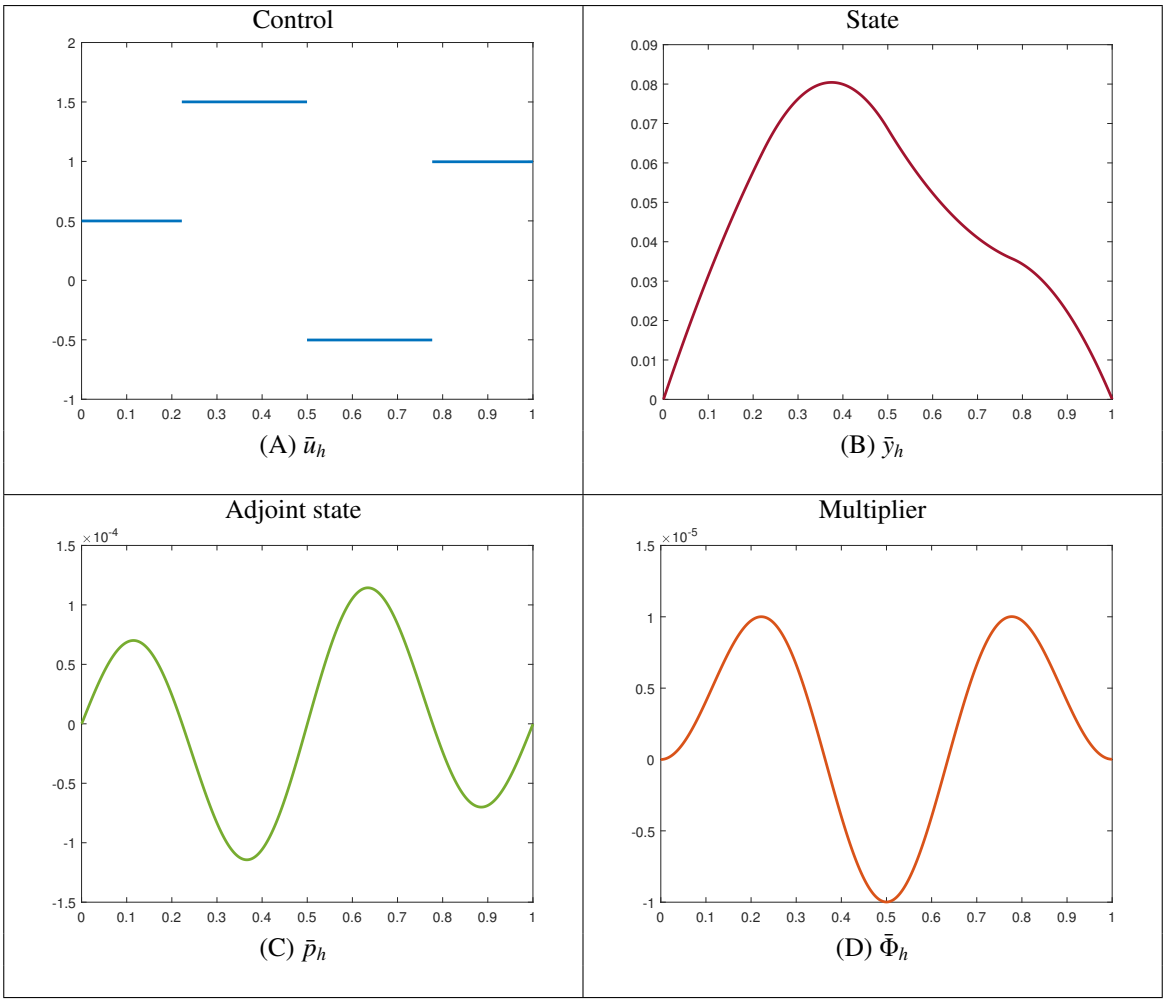

Figure 5.1: The variationally discrete solution to the data from Example 1 for $h = \frac{1}{2048}$. The inclusions in (5.36) are clearly visible.

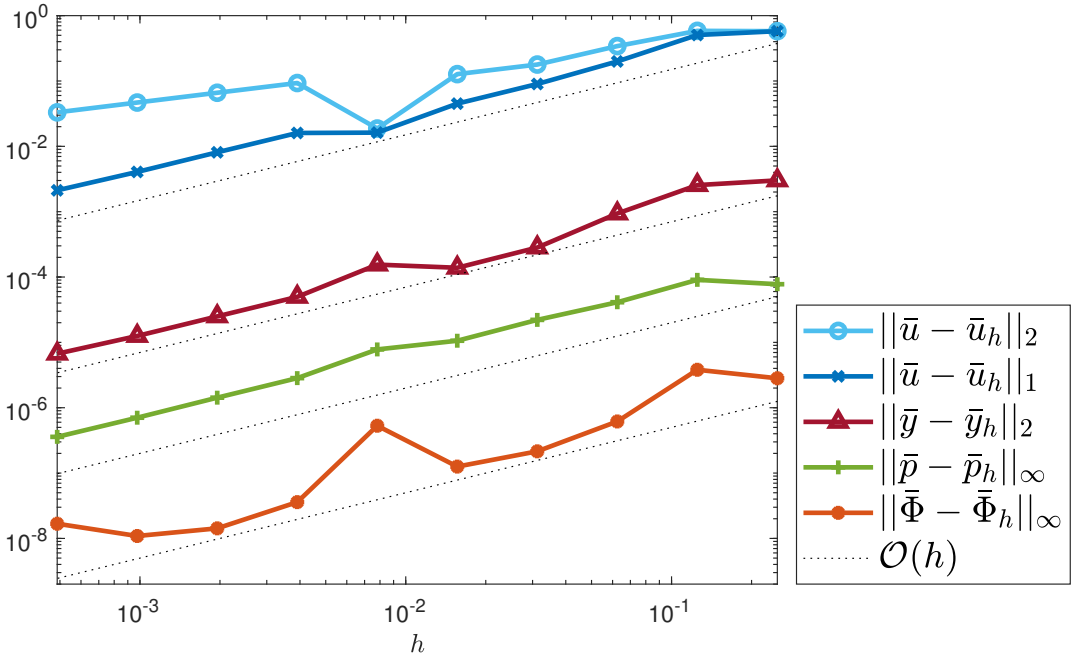

Figure 5.2: Example 1: Convergence plots of the errors of the solutions to the variationally discrete problem compared to the known exact solution.

As a second example we use [43, 5.4. Example 2] with unknown solution, $\alpha = 10^{-5}$ and $y_d(x) := 0.5\pi^{-2}(1 - \cos(2\pi x))$.

Since the solution is not known, we calculate a reference solution on the finest grid with a reasonable computing time, i.e. $h = \frac{1}{1024}$. The results displayed in Figure 5.3 are then used to approximate $\bar{u}, \bar{y}, \bar{p}, \bar{\Phi}$ for the calculation of the errors.

In Figure 5.2 the errors between the known solutions and the solutions to the variationally discretized problem are depicted. Again, we observe that the order of convergence is approximately h, except for $\|\bar{u} - \bar{u}_h\|_{L^2(Q)}$, which converges with a slower rate. This also aligns with the results from Subsection 5.3.1.

Furthermore, we calculate the convergence order h^α for the refinement from some gridsize h_1 to some other gridsize h_2 as explained before. The results are displayed in Table 5.2.

h_1	h_2	$\|\bar{u} - \bar{u}_h\|_{L^1(Q)}$	$\|\bar{u} - \bar{u}_h\|_{L^2(Q)}$	$\|\bar{y} - \bar{y}_h\|_{L^2(Q)}$	$\|\bar{p} - \bar{p}_h\|_{L^\infty(Q)}$	$\|\bar{\Phi} - \Phi_h\|_{L^\infty(Q)}$
0.25	0.125	0.3110	0.2454	1.0464	0.7448	1.5684
0.125	0.0625	0.9990	0.5319	1.0788	0.8119	1.6061
0.0625	0.03125	0.9763	0.4961	1.0147	1.0266	-0.2077
0.03125	0.015625	0.9348	0.4682	0.9376	0.9737	1.3400
0.015625	0.0078125	1.1204	0.5630	1.0757	1.1106	1.0238
0.0078125	0.00390625	0.7379	0.3679	0.6267	0.8702	0.3120
	mean	0.8466	0.4454	0.9633	0.9230	0.9404
	slope of best fit	0.9004	0.4679	0.9823	0.9450	0.9137

Table 5.2: Example 2: Convergence order (potency of gridsize h) of the respective errors when the grid is refined from gridsize h_1 to gridsize h_2.

Altogether, we are able to verify the results we show in Section 5.3, i.e. the inclusions from (5.36), the sparsity structure of the control, and the error estimates for control, state, adjoint state and multiplier.

In [43, Section 5] the same examples have been analyzed, but without employing a mixed formulation for the state equation. Under almost the same structural assumptions they get the following results: For a variational discretization approach with piecewise linear and continuous state and test functions they observe errors of the order $O(h^2)$. Additionally, for a full discretization with piecewise constant control and piecewise linear and continuous state and test functions they see errors of the order $O(h)$.

In comparison, we consider a variational discretization approach combined with a mixed formulation of the state equation discretized with lowest order Raviart Thomas elements, which corresponds to $(y_h, z_h) \in P_0 \times P_1$. We see that under the given structural assumption this leads to piecewise constant controls without discretizing the control, so we can not expect more than the order $O(h)$, which we have proven.

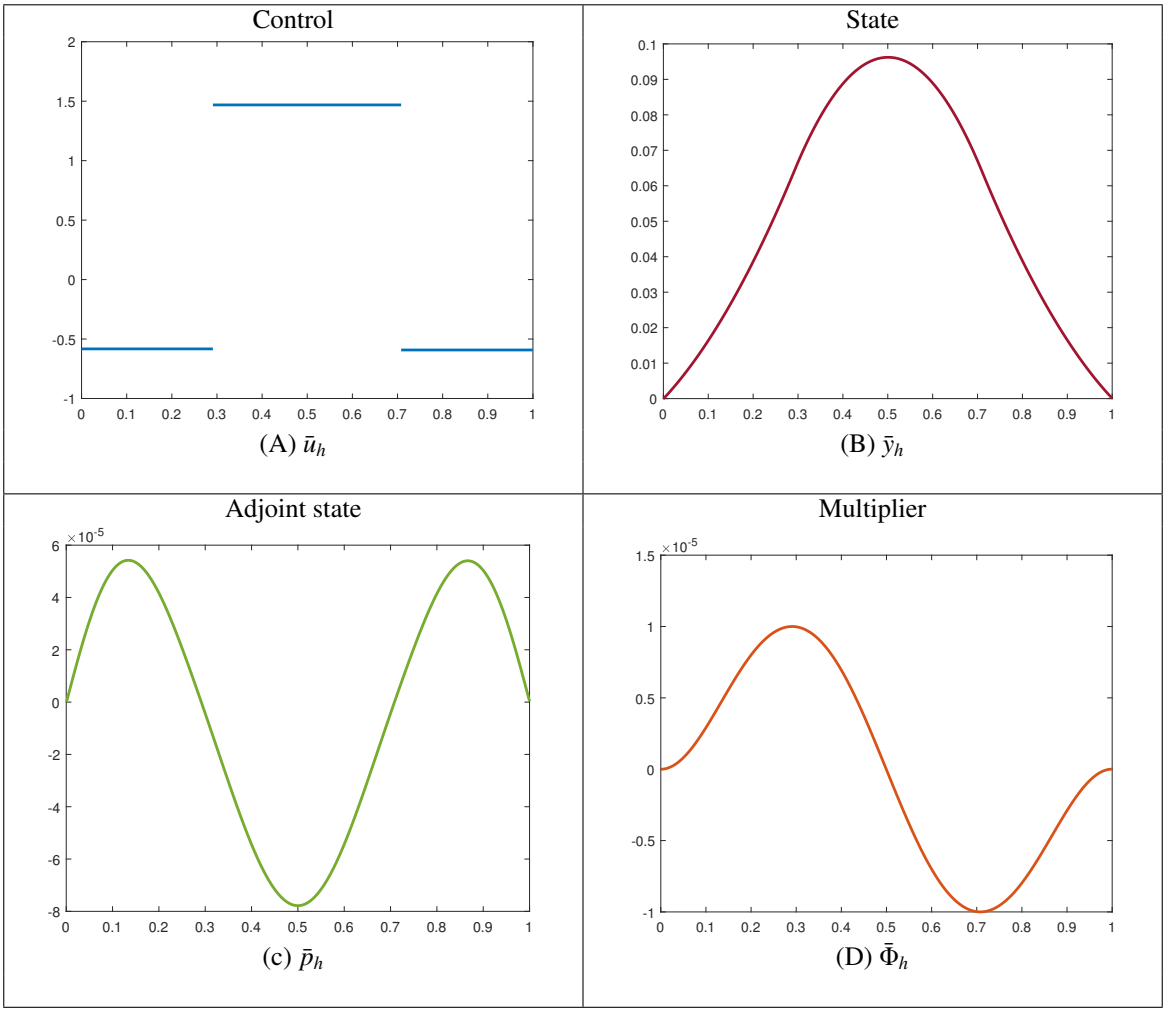

Figure 5.3: The variationally discrete solution to the data from Example 2 for $h = \frac{1}{1024}$. The inclusions in (5.36) are clearly visible.

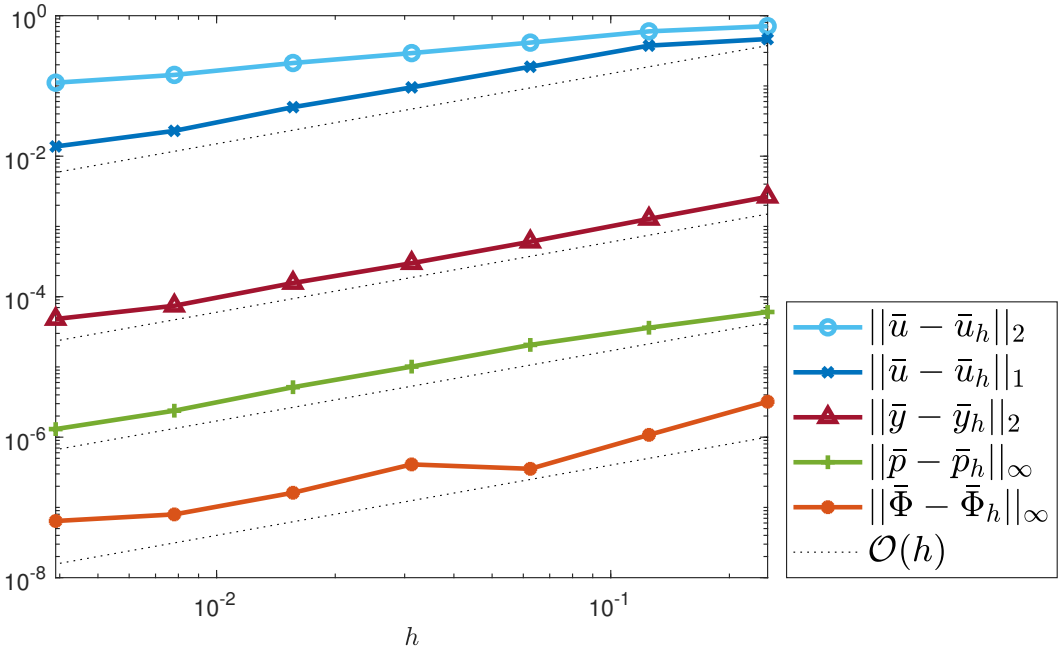

Figure 5.4: Example 2: Convergence plots of the errors of the solutions to the variationally discrete problem compared to the approximation of the exact solution. The reference solution is computed on a grid with $h = \frac{1}{1024}$.

Chapter 6

Conclusion

In this work we have applied the variational discretization approach to three different problem formulations.

In Chapter 3 we have analyzed a parabolic optimal control problem with space-time measure control and initial measure control. The variational discretization approach has been compared to a full discretization approach to illustrate the advantages of not discretizing the control. Here, we have seen the sparsity structure being retained better on the discrete level by the variational discretization than by the full discretization.

In the following Chapter 4 we have only considered variational discretization and no full discretization for comparison, because the parabolic optimal control problem with bounded initial measure control is closely related to the problem from Chapter 3, if we only consider an initial measure control u_0. The main difference is that we consider a regularization term in the cost functional for the initial control in Chapter 3, while we impose a bound on the measure norm in Chapter 4. Both problems admit a similar sparsity structure, so we can deduce from our results in Chapter 3 that also for the problem in Chapter 4 the variational discretization will perform better in retaining the sparsity structure than a full discretization. Additionally, since we consider an initial control, we only need to achieve sparsity in space at initial time, instead of sparsity in space-time.

For the elliptic optimal control problem governed by functions of bounded variation in Chapter 5 we have had results achieved by other discretization techniques available from [43] to compare our variational discretization approach to. In this case the sparsity structure is also retained in a full discretization approach, which is caused by the choice of spaces and the fact that we consider an elliptic control problem and not a parabolic one as in the earlier Chapters.

In fact, the main challenge in the parabolic setting in Chapter 3 with space-time control is to retain the sparsity in time. The approach we have taken for the parabolic optimal control problem delivers that the measure control is supported in grid points due to our choice of piecewise linear and continuous test functions. This limits the accuracy we can achieve on the discrete level. A remedy could be to consider piecewise quadratic test functions, such that the extremal points of the adjoint can also be attained in between grid points. From the sparsity structure we know that the controls support is a subset of those extremal points, so it would not be limited to grid points any more. This advantage comes at the cost of a more complicated discrete state equation, where we might need to add smoothing steps to avoid oscillations.

To justify our approach we have shown the convergence of the variationally discrete optimal control and state to their continuous counterparts in Chapter 3. This should be easy to adapt to the problem formulation in Chapter 4 although we have not presented this result. For the problem with BV-control in Chapter 5 however, we have proven convergence rates of order $O(h)$ for the optimal control in the L^1-norm, the optimal state in the L^2-norm and the optimal adjoint state and optimal multiplier in the L^∞-norm. It would be very interesting to also investigate the convergence rates for the problems in Chapter 3 and Chapter 4. Our computational results indicate that the convergence rates for the optimal control in the measure norm and the optimal state in the L^q-norm are of order $O(h)$, but we could also be seeing superconvergence effects due to our choice of example, where the true control was located on the grid. This remains to be studied.

Altogether, we have seen that the variational discretization approach can be tailored as needed for the problem and, depending on the choice of ansatz and test spaces, delivers an induced discrete structure of the not discretized optimal control. By observing the continuous sparsity structure and optimality system - especially the dependence of the optimal control on the optimal adjoint state - we get an indication on how to use variational discretization to achieve maximal sparsity of the control on the discrete level. For example, in Chapter 4 we have a parabolic optimal control problem, but the control only resides at initial time, so it suffices to achieve sparsity in space. We still need to make a choice how to discretize the test and ansatz spaces in time, but this will not affect the structure of the optimal control. Therefore, we can make the most simple choice: piecewise constant functions in time. In contrast, in Chapter 3 it was important for the retention of the sparsity structure of the space-time control to employ piecewise linear and continuous test functions in time.

Also, we remark that it is possible to apply the variational discretization approach to many other optimal control problems, where it is promising to be specifically beneficial for problems that admit a sparsity structure. In Section 1.3 we have presented a list of references that deal with such problems and for a lot of them it would be interesting to compare variational discretization to the respective full discretization technique considered in the reference. Also, there exist further problems not included in these references, which will be worth analyzing. For example, we are working on applying our approach to a parabolic optimal control problem with measure valued control in time, which can be viewed as a generalization of the impulse control for evolution equations.

We conclude this work by observing that we were indeed able to retain the respective sparsity structure of three different continuous problems on the discrete level by utilizing a variational discretization approach.

Appendix A

Appendix

A.1 Density argument 1

In order to see that

$$\{\psi \in C^2(\bar{I};\mathbb{R}) : \psi(T) = 0\} \otimes (C^2(\Omega) \cap W_0^{1,p}(\Omega)),$$

is dense in

$$W = \left\{ w \in W_2^{1,1}(Q) : w|_\Sigma = 0,\, w(T) = 0 \text{ and } -(\partial_t + \Delta)w \in L^p(Q) \right\},$$

we abbreviate $X \coloneqq W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$ and show the following two claims:

Claim I: $\{w \in C^k(\bar{I}; X) : w(T) = 0\}$ is dense in W.

This can be seen by constructing a suitable Friedrichs smoothing operator:
Choose a smooth function $\varphi\colon [0,\infty) \to \mathbb{R}$, with

- $0 \le \varphi(s) \le 1$ for all $s \in [0,\infty)$,
- $\operatorname{supp}(\varphi) \subset (0,2)$,
- $\int_0^\infty \varphi(s)\, ds = 1$.

For $\epsilon > 0$, put $\varphi_\epsilon(s) \coloneqq \frac{1}{\epsilon}\,\varphi(\frac{s}{\epsilon})$. Now for each function $v \in L^1_{\mathrm{loc}}([0,T]; Y)$ with values in any Banach space Y, define the smoothing $v_\epsilon \in L^1_{\mathrm{loc}}([0,T]; Y)$ by first extending v by 0 to a function $\tilde{v} \in L^1_{\mathrm{loc}}(\mathbb{R}; Y)$ and by putting

$$v_\epsilon(t) \coloneqq \int_{-\infty}^{\infty} \tilde{v}(t+s)\, \varphi_\epsilon(s)\, ds$$

For $w \in W \subset L^p(\bar{I}; X)$, the smoothing w_ϵ is a member of $\{\xi \in C^1(\bar{I}; X) : \xi(T) = 0\}$. Since $w_\epsilon(T) = 0$ holds by construction, we have only to show that $w_\epsilon \in C^1(\bar{I}; X)$. Indeed, for $\delta \in (-\epsilon, \epsilon)$, we have

$$\begin{aligned}
&\left\| w_\epsilon(t+\delta) - w_\epsilon(t) + \int_{-\infty}^{\infty} \tilde{w}(x, t+s)\, \varphi_\epsilon'(s)\, \delta\, ds \right\|_X \\
&= \left\| \int_{-\infty}^{\infty} \left((\tilde{w}(t+\delta+s) - \tilde{w}(t+s))\ \varphi_\epsilon(s) + \tilde{w}(t+s)\, \varphi_\epsilon'(s)\, \delta \right)\, ds \right\|_X \\
&= \left\| \int_{-\infty}^{\infty} \tilde{w}(t+s)\ \left(\varphi_\epsilon(s-\delta) - \varphi_\epsilon(s) + \varphi_\epsilon'(s)\, \delta \right)\, ds \right\|_X \\
&\le \int_{-\infty}^{\infty} \|\tilde{w}(t+s)\|_X\ |\varphi_\epsilon(s-\delta) - \varphi_\epsilon(s) + \varphi_\epsilon'(s)\, \delta|\, ds \\
&\le |T|^{1-1/p}\, \|w\|_{L^p(\bar{I};X)}\, \operatorname{Lip}(\varphi_\epsilon')\, \delta^2 \\
&\le |T|^{1-1/p}\, \|w\|_{W_p^{2,1}}\ \operatorname{Lip}(\varphi_\epsilon')\, \delta^2.
\end{aligned}$$

Thus, w_ϵ is differentiable with derivative

$$w_\epsilon'(t) = -\int_{-\infty}^{\infty} \tilde{w}(t+s)\,\varphi_\epsilon'(s)\,ds = \int_{-\infty}^{\infty} \tilde{w}'(t+s)\,\varphi_\epsilon(s)\,ds = (\partial_t w)_\epsilon.$$

Analogously, one sees that w_ϵ' is of class C^2 and thus continuous. Repeating this analysis, we find that $w_\epsilon \in C^k(\bar{I};X)$ for all $k \in \mathbb{N}$. Since $p < \infty$, standard results on convolution operators imply that w_ϵ converges to w in $L^p(\bar{I};X)$ for $\epsilon \searrow 0$. The same argument shows that $\partial_t w_\epsilon = (\partial_t w)_\epsilon$ converges to $\partial_t w$ in $L^p(\bar{I};L^p(\Omega))$ for $\epsilon \searrow 0$. Thus w_ϵ converges to w in W.

Claim II: $\{\xi \in C^2(\bar{I};\mathbb{R}) : \xi(T) = 0\} \otimes X$ is dense in $\{w \in C^2(\bar{I};X) : w(T) = 0\}$ and thus in W.

Let $w \in C^2(\bar{I};X)$ with $w(T) = 0$. By standard results in approximation theory (e.g., by cubic spline interpolation, see Theorem 1 in [2]), one may approximate w in $C^2(\bar{I};X)$ by functions of the form

$$w_N(t) \coloneqq \sum_{i=0}^{N} \varphi_{N,i}(t)\, w\left(i\,\frac{T}{N}\right) \quad \text{hence} \quad w_N \in C^2(\bar{I};\mathbb{R}) \otimes X$$

with suitable functions $\varphi_{N,i} \in C^2(\bar{I};\mathbb{R})$.

A.2 Density argument 2

Recall that we identified $\mathcal{M}(\bar{\Omega}_c)$ and $\mathcal{M}(\bar{Q}_c)$ with $\{u_0 \in \mathcal{M}(\Omega)\colon \operatorname{supp}(u_0) \subset \bar{\Omega}_c\}$ and $\{u \in \mathcal{M}(Q)\colon \operatorname{supp}(u) \subset \bar{Q}_c\}$, respectively. In this sense, the sets

$$\{f_0 \in C^\infty(\Omega)\colon \operatorname{supp}(f_0) \subset \bar{\Omega}_c\} \quad \text{and} \quad \{f \in C^\infty(Q)\colon \operatorname{supp}(f) \subset \bar{Q}_c\}$$

are dense in $\mathcal{M}(\bar{\Omega}_c)$ and $\mathcal{M}(\bar{Q}_c)$ with respect to the sequential weak* topology.

This can be seen by utilizing that $\bar{\Omega}_c$ and $\bar{Q}_c$ have to satisfy certain uniform cone conditions (because they are Lipschitz domains, see [1, Paragraph 4.8]) and by convolution against suitable Friedrichs mollifiers that are compactly supported in the interior of finite, convex cones.

In detail: Let $\mu \in \mathcal{M}(\bar{Q}_c)$. Let $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ be a smooth function with support in the unit ball satisfying $0 \le \varphi(x) \le 1$ for all $x \in \mathbb{R}^n$ and $\int_{\mathbb{R}^n} \varphi(x)\,dx = 1$. Then for $\epsilon > 0$, we put $\varphi_\epsilon(x) \coloneqq \frac{1}{\epsilon^n}\,\varphi(\frac{x}{\epsilon})$ and $f_\epsilon(x) \coloneqq \int_{\mathbb{R}^n} \varphi_\epsilon(y-x)\,d\mu(y)$. Some further =analysis shows that $f_\epsilon \in C^\infty(\mathbb{R}^n)$. Because $\bar{Q}_c \subset\subset Q$ is relatively compact, it has a positive distance $\epsilon_0 > 0$ to ∂Q and hence $\operatorname{supp}(f_\epsilon) \subset Q$ and thus $f_\epsilon \in C_0^\infty(Q)$ for all $0 < \epsilon < \epsilon_0$. Let $\psi \in C(\bar{Q}_c)$ and extend it continuously to $\mathbb{R}^n$ (this is possible because Q_c is assumed to be polyhedral or an extension domain). Then

$$\begin{aligned}\int_Q \psi(x)\,f_\epsilon(x)\,dx &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(x)\,\varphi_\epsilon(y-x)\,d\mu(y)\,dx \\ &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \psi(x)\,\varphi_\epsilon(y-x)\,dx\,d\mu(y) \\ &= \int_{\mathbb{R}^n} \psi_\epsilon(y)\,d\mu(y) = \int_Q \psi_\epsilon(y)\,d\mu(y)\end{aligned}$$

with $\psi_\epsilon(x) \coloneqq \int_{\mathbb{R}^n} \psi(x)\,\varphi_\epsilon(y-x)\,dx$. Now standard results on Friedrichs operators show $\psi_\epsilon \to \psi$ in $C(\bar{Q}_c)$. Hence we obtain

$$\textstyle\int_Q \psi(x)\,f_\epsilon(x)\,dx = \int_Q \psi_\epsilon(y)\,d\mu(y) \to \int_Q \psi(y)\,d\mu(y),$$

showing that $f_\epsilon(x)\,dx$ converges weak-$*$ to μ.

A.3 Fourier modes

In the sections on computational results, Section 3.5 and Section 4.4, we mention the option to sample the associated state from the analytic solution for the control u with spacial Fourier modes. A very nice representation of this method can be found in [12, Chapter 1.5]. We adapt it to our notation here:

We consider the diffusion equation

$$\partial_t y = a^2 \Delta y$$

for $0 < x < 1$ and $0 < t < \infty$ with general initial condition

$$y(x, 0) = u_0(x),$$

and either Dirichlet or Neumann boundary condition.

This equation can be seen as a special case of a Sturm-Liouville problem, which has the form

$$(py')' + (q + \lambda r)y = 0 \quad \text{for } 0 < x < 1$$

with linear homogeneous boundary conditions

$$\begin{aligned} c_1 y(a) + c_2 y'(a) &= 0, \\ c_3 y(b) + c_4 y'(b) &= 0. \end{aligned}$$

Here, p, q, r are functions of x and λ and c_1, c_2, c_3, c_4 are constants.

In the setting of Section 3.5 we have Dirichlet boundary conditions, i.e. $c_1 = c_3 = 1, c_2 = c_4 = 0$. By theorems about Sturm-Liouville problems it is known that in this case the functions

$$y_n(x) = \sin(n\pi x) \quad n = 1, 2, \dots$$

form a complete, orthogonal function system. Then the general initial condition can be represented with these functions:

$$u_0(x) = \sum_n A_n \sin(n\pi x).$$

This happens to be a Fourier-Sinus-transformation.

We calculate

$$\begin{aligned} \int_0^1 \sin(m\pi x) u(x)\, dx &= \sum_n A_n \int_0^1 \sin(m\pi x) \sin(n\pi x)\, dx \\ &= A_m \int_0^1 \sin^2(m\pi x)\, dx \\ &= \frac{A_m}{2} \end{aligned}$$

for all $m = 1, 2, \dots$ and derive

$$A_m = 2 \int_0^1 \sin(m\pi x) u(x)\, dx.$$

Altogether we get

$$y(x, t) = \sum_n A_n \exp(-n^2 \pi^2 a^2 t) \sin(n\pi x).$$

Now, the initial condition is given by a dirac measure: $u_0(x) = \delta_{x_0}$. Consequently, we can calculate the A_m as

$$A_m = 2 \int_0^1 \sin(m\pi x)\delta_{x_0}\, dx = 2\sin(m\pi x_0).$$

These can be inserted into $y(x,t)$ and then the state can be calculated. We want to remark that in Section 3.5 we consider δ_{x_0,t_0} with $t_0 > 0$. In this case we can set $\bar{y}(x,t) = 0$ for all $t < t_0$ and $\bar{y}(x,t) = y(x, t-t_0)$ for $t \geq t_0$.

In the setting of Section 4.4 we have Neumann boundary conditions, i.e. $c_1 = c_3 = 0, c_2 = c_4 = 1$ and we employ a system of cosine-functions to match these. The functions

$$y_1(x) = 1, \quad y_n(x) = \cos((n-1)\pi x), \quad n \geq 2$$

form a complete, orthogonal function system. Consequently we represent

$$u_0(x) = A_1 + \sum_{n\geq 2} A_n \cos((n-1)\pi x).$$

Then, we calculate

$$\int_0^1 1u(x)\, dx = A_1 \int_0^1 dx + \sum_{n\geq 2} A_n \int_0^1 \cos(n\pi x)\, dx = A_1$$

and for $m \geq 2$

$$\begin{aligned}\int_0^1 \cos((m-1)\pi x)u(x)\, dx &= A_1 \int_0^1 \cos((m-1)\pi x)\, dx + \sum_{n\geq 2} A_n \int_0^1 \cos((m-1)\pi x)\cos((n-1)\pi x)\, dx \\ &= A_m \int_0^1 \cos^2((m-1)\pi x)\, dx \\ &= \frac{A_m}{2}.\end{aligned}$$

This gives

$$\begin{aligned}A_1 &= \int_0^1 u(x)\, dx, \\ A_m &= 2\int_0^1 \cos((m-1)\pi x)u(x)\, dx \qquad \text{for } m \geq 2.\end{aligned}$$

Altogether in this case we have

$$y(x,t) = A_1 + \sum_{n\geq 2} A_n \exp(-n^2\pi^2 a^2 t)\cos((n-1)\pi x).$$

Again, we have $u_0(x) = \delta_{x_0}$ and

$$\begin{aligned}A_1 &= 1, \\ A_m &= 2\cos((m-1)\pi x_0) \qquad \text{for } m \geq 2.\end{aligned}$$

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Second. Vol. 140. Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, 2003, pp. xiv+305.

[2] J. H. Ahlberg and E. N. Nilson. "Convergence properties of the spline fit". In: *Journal of the Society for Industrial and Applied Mathematics* 11 (1963), pp. 95–104.

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. The Clarendon Press, Oxford University Press, 2000.

[4] H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. MPS-SIAM Series on Optimization, 2006.

[5] C. Bahriawati and C. Carstensen. "Three MATLAB implementations of the lowest-order Raviart-Thomas MFEM with a posteriori error control". In: *Computational methods in applied mathematics* 5.4 (2005), pp. 333–361.

[6] S. Bartels. "Total variation minimization with finite elements: convergence and iterative solution". In: *SIAM Journal on Numerical Analysis* 50.3 (2012), pp. 1162–1180.

[7] S. Bartels and M. Milicevic. "Iterative finite element solution of a constrained total variation regularized model problem". In: *Discrete & Continuous Dynamical Systems-S* 10.6 (2017), p. 1207.

[8] O. V. Besov, V. P. Il'in, and S. M. Nikol'skiĭ. *Integral representations of functions and imbedding theorems. Vol. I*. Translated from the Russian, Scripta Series in Mathematics, Edited by Mitchell H. Taibleson. V. H. Winston & Sons, Washington, D.C.; Halsted Press [John Wiley & Sons], New York-Toronto, Ont.-London, 1978, pp. viii+345.

[9] K. Bredies and D. Vicente. "A perfect reconstruction property for PDE-constrained total-variation minimization with application in Quantitative Susceptibility Mapping". In: *ESAIM: Control, Optimisation and Calculus of Variations* 25 (2019), p. 83.

[10] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011, pp. xiv+599.

[11] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer, 1991.

[12] G. Carrier and C. Pearson. *Partial Differential Equations: Theory and Technique*. Elsevier Science, 2014.

[13] E. Casas, C. Clason, and K. Kunisch. "Approximation of elliptic control problems in measure spaces with sparse solutions". In: *SIAM Journal on Control and Optimization* 50.4 (2012), pp. 1735–1752.

[14] E. Casas, C. Clason, and K. Kunisch. "Parabolic control problems in measure spaces with sparse solutions". In: *SIAM Journal on Control and Optimization* 51.1 (2013), pp. 28–63.

[15] E. Casas, R. Herzog, and G. Wachsmuth. "Analysis of spatio-temporally sparse optimal control problems of semilinear parabolic equations". In: *ESAIM: Control, Optimisation and Calculus of Variations* 23.1 (2017), pp. 263–295.

[16] E. Casas, P. I. Kogut, and G. Leugering. “Approximation of optimal control problems in the coefficient for the p-Laplace equation. I. Convergence result”. In: *SIAM Journal on Control and Optimization* 54.3 (2016), pp. 1406–1422.

[17] E. Casas, F. Kruse, and K. Kunisch. “Optimal control of semilinear parabolic equations by BV-functions”. In: *SIAM Journal on Control and Optimization* 55.3 (2017), pp. 1752–1788.

[18] E. Casas and K. Kunisch. “Analysis of optimal control problems of semilinear elliptic equations by bv-functions”. In: *Set-Valued and Variational Analysis* 27.2 (2019), pp. 355–379.

[19] E. Casas and K. Kunisch. “Optimal control of semilinear elliptic equations in measure spaces”. In: *SIAM Journal on Control and Optimization* 52.1 (2014), pp. 339–364.

[20] E. Casas and K. Kunisch. “Parabolic control problems in space-time measure spaces”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 22.2 (2016), pp. 355–370.

[21] E. Casas and K. Kunisch. “Using sparse control methods to identify sources in linear diffusion-convection equations”. In: *Inverse Problems* 35.11 (2019), p. 114002.

[22] E. Casas, K. Kunisch, and C. Pola. “Regularization by functions of bounded variation and applications to image enhancement”. In: *Applied Mathematics and Optimization* 40.2 (1999), pp. 229–257.

[23] E. Casas, M. Mateos, and A. Rösch. “Finite element approximation of sparse parabolic control problems”. In: *Mathematical Control and Related Fields* 7 (2017), pp. 393–417.

[24] E. Casas, M. Mateos, and A. Rösch. “Improved approximation rates for a parabolic control problem with an objective promoting directional sparsity”. In: *Computational Optimization and Applications* 70.1 (2018), pp. 239–266.

[25] E. Casas, B. Vexler, and E. Zuazua. “Sparse initial data identification for parabolic PDE and its finite element approximations”. In: *Mathematical Control and Related Fields* 5.3 (2015), pp. 377–399.

[26] Y. Chen and W. Liu. “Error estimates and superconvergence of mixed finite element for quadratic optimal control”. In: *International Journal of Numerical Analysis and Modeling* 3.3 (2006), pp. 311–321.

[27] C. Clason. *Nonsmooth Analysis and Optimization*. 2017. eprint: `arXiv:1708.04180`.

[28] C. Clason, F. Kruse, and K. Kunisch. “Total variation regularization of multi-material topology optimization”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 52.1 (2018), pp. 275–303.

[29] C. Clason and K. Kunisch. “A duality-based approach to elliptic control problems in non-reflexive Banach spaces”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 17.1 (2011), pp. 243–266.

[30] C. Clason and A. Schiela. “Optimal control of elliptic equations with positive measures”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 23.1 (2017), pp. 217–240.

[31] N. von Daniels, M. Hinze, and M. Vierling. “Crank-Nicolson time stepping and variational discretization of control-constrained parabolic optimal control problems”. In: *SIAM Journal on Control and Optimization* 53.3 (2015), pp. 1182–1198.

[32] J. Douglas and J. E. Roberts. “Global estimates for mixed methods for second order elliptic equations”. In: *Mathematics of computation* 44.169 (1985), pp. 39–52.

[33] I. Ekeland and R. Témam. *Convex analysis and variational problems*. English. Vol. 28. Classics in Applied Mathematics. Translated from the French. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999, pp. xiv+402.

[34] A. El Badia, T. Ha-Duong, and A. Hamdi. “Identification of a point source in a linear advection–dispersion–reaction equation: application to a pollution source problem”. In: *Inverse Problems* 21.3 (2005), p. 1121.

[35] L. C. Evans. *Partial differential equations*. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 1998, pp. xviii+662.

[36] L. Gastaldi and R. Nochetto. "Optimal L^∞-error estimates for nonconforming and mixed finite element methods of lowest order". In: *Numerische Mathematik* 50 (1987), pp. 587–611.

[37] J.-M. Ghidaglia. "Some backward uniqueness results". In: *Nonlinear Analysis: Theory, Methods & Applications* 10.8 (1986), pp. 777–790.

[38] V. Girault and P.-A. Raviart. "Finite element approximation of the Navier-Stokes equations". In: *Lecture Notes in Mathematics, Berlin Springer Verlag* (1979).

[39] C. Goll, R. Rannacher, and W. Wollner. "The damped Crank-Nicolson time-marching scheme for the adaptive solution of the Black-Scholes equation". In: *Journal of Computational Finance* 18 (June 2015), pp. 1–37.

[40] W. Gong. "Error estimates for finite element approximations of parabolic equations with measure data". In: *Mathematics of Computation* 82 (Jan. 2013), pp. 69–98.

[41] W. Gong, M. Hinze, and Z. Zhou. "A priori error analysis for finite element approximation of parabolic optimal control problems with pointwise control". In: *SIAM Journal on Control and Optimization* 52.1 (2014), pp. 97–119.

[42] W. Gong and N. Yan. "Mixed finite element method for Dirichlet boundary control problem governed by elliptic PDEs". In: *SIAM Journal on Control and Optimization* 49.3 (2011), pp. 984–1014.

[43] D. Hafemeyer, F. Mannel, I. Neitzel, and B. Vexler. "Finite element error estimates for one-dimensional elliptic optimal control by BV-functions". In: *Mathematical Control and Related Fields* (2019).

[44] S.-P. Han and O. L. Mangasarian. "Exact penalty functions in nonlinear programming". In: *Mathematical programming* 17.1 (1979), pp. 251–269.

[45] E. Herberg. "Variational discretization of parabolic control problems in space-time measure spaces". MA thesis. University of Hamburg, 2017.

[46] E. Herberg and M. Hinze. "Variational discretization approach applied to an optimal control problem with bounded measure controls". In: *arXiv preprint arXiv:2003.14380* (2020).

[47] E. Herberg, M. Hinze, and H. Schumacher. "Maximal discrete sparsity in parabolic optimal control with measures". In: *Mathematical Control and Related Fields* 10.4 (Dec. 2020), pp. 735–759.

[48] R. Herzog, G. Stadler, and G. Wachsmuth. "Directional sparsity in optimal control of partial differential equations". In: *SIAM Journal on Control and Optimization* 50.2 (2012), pp. 943–963.

[49] M. Hinze. "A variational discretization concept in control constrained optimization: the linear-quadratic case". In: *Computational Optimization and Applications* 30.1 (2005), pp. 45–61.

[50] M. Hinze, B. Kaltenbacher, and T. N. T. Quyen. "Identifying conductivity in electrical impedance tomography with total variation regularization". In: *Numerische Mathematik* 138.3 (2018), pp. 723–765.

[51] M. Hinze and T. N. T. Quyen. "Finite element approximation of source term identification with TV-regularization". In: *Inverse Problems* 35.12 (2019), p. 124004.

[52] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*. Vol. 23. Mathematical Modelling: Theory and Applications. Springer, New York, 2009, pp. xii+270.

[53] N. V. Krylov. *Lectures on elliptic and parabolic equations in Sobolev spaces*. Vol. 96. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2008, pp. xviii+357.

[54] K. Kunisch, K. Pieper, and B. Vexler. "Measure valued directional sparsity for parabolic optimal control problems". In: *SIAM Journal on Control and Optimization* 52.5 (2014), pp. 3078–3108.

[55] K. Kunisch, P. Trautmann, and B. Vexler. "Optimal control of the undamped linear wave equation with measure valued controls". In: *SIAM Journal on Control and Optimization* 54.3 (2016), pp. 1212–1244.

[56] Leykekhman, Dmitriy, Vexler, Boris, and Walter, Daniel. "Numerical analysis of sparse initial data identification for parabolic problems". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 54.4 (2020), pp. 1139–1180.

[57] Y. Li, S. Osher, and R. Tsai. "Heat source identification based on constrained minimization". In: *Inverse Problems and Imaging* 8.1 (2014), pp. 199–221.

[58] J. L. Lions. *Optimal control of systems governed by partial differential equations*. Springer, 1971.

[59] J. R. Munkres. *Topology*. Second edition of [MR0464128]. Prentice Hall Inc., Upper Saddle River, NJ, 2000, pp. xvi+537.

[60] J. A. Nitsche. "L_∞-convergence of finite element approximation". In: *Journées "Éléments Finis" (Rennes, 1975)*. Univ. Rennes, Rennes, 1975, p. 18.

[61] J. Peypouquet. *Convex optimization in normed spaces: theory, methods and examples*. Springer, 2015.

[62] K. Pieper and B. Vexler. "A priori error analysis for discretization of sparse elliptic optimal control problems in measure space". In: *SIAM Journal on Control and Optimization* 51.4 (2013), pp. 2788–2808.

[63] R. A. Polyak. "Complexity of the regularized Newton's method". In: *Pure and Applied Functional Analysis* 3.2 (2018), pp. 327–347.

[64] P.-A. Raviart and J.-M. Thomas. "A mixed finite element method for 2-nd order elliptic problems". In: *Mathematical aspects of finite element methods*. Springer, 1977, pp. 292–315.

[65] J. C. De los Reyes. *Numerical PDE-constrained optimization*. Springer, 2015.

[66] W. Rudin. *Real and complex analysis*. Third. McGraw-Hill Book Co., New York, 1987, pp. xiv+416.

[67] W. Schirotzek. *Nonsmooth analysis*. Universitext. Springer, Berlin, 2007, pp. xii+373.

[68] R. E. Showalter. *Monotone operators in Banach space and nonlinear partial differential equations*. Vol. 49. American Mathematical Soc., 2013.

[69] G. Stadler. "Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices". In: *Computational Optimization and Applications* 44.2 (2009), pp. 159–181.

[70] V. Thomée. *Galerkin finite element methods for parabolic problems*. Second. Vol. 25. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2006, pp. xii+370.

[71] P. Trautmann, B. Vexler, and A. Zlotnik. "Finite element error analysis for measure-valued optimal control problems governed by a 1D wave equation with variable coefficients". In: *Mathematical Control and Related Fields* 8.2 (2018), pp. 411–449.

[72] F. Tröltzsch. *Optimal control of partial differential equations: theory, methods, and applications*. Vol. 112. American Mathematical Society, 2010.

[73] M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung*. Springer-Verlag, 2012.

[74] W. P. Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*. Springer, New York, 1989.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt,

- dass ich die die eingereichte Dissertation selbstständig verfasst habe und alle von mir für die Arbeit benutzten Hilfsmittel und Quellen in der Arbeit angegeben sowie die Anteile etwaig beteiligter Mitarbeiterinnen oder Mitarbeiter sowie anderer Autorinnen oder Autoren klar gekennzeichnet habe;
- dass ich nicht die entgeltliche Hilfe von Vermittlungs- oder Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen habe;
- dass ich die Dissertation nicht in gleicher oder ähnlicher Form als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung im In- oder Ausland eingereicht habe;
- dass ich weder die gleiche noch eine andere Abhandlung in einem anderen Fachbereich oder einer anderen wissenschaftlichen Hochschule als Dissertation eingereicht habe;
- dass mir bewusst ist, dass ein Verstoß gegen einen der vorgenannten Punkte den Entzug des Doktortitels bedeuten und ggf. auch weitere rechtliche Konsequenzen haben kann.

Ort, Datum Evelyn Christin Herberg