A sequential quadratic Hamiltonian scheme for solving optimal control problems with non-smooth cost functionals

Tim Breitenbach

Dissertation

Würzburg, 2019

cc BY NC SA

Contents

Summary **11**

1 Introduction **13**

2 An SQH framework for ODE optimal control problems **19**

2.1 The formulation of ODE optimal control problems . . . 19

2.2 The characterization by the Pontryagin maximum principle . . . 21

2.3 Convergence analysis of the SQH scheme . . . 28

2.4 Optimal quantum control . . . 47

2.5 Optimal tumor therapy . . . 57

3 An SQH framework for PDE optimal control problems **69**

3.1 Parabolic and elliptic optimal control problems . . . 70

3.2 The characterization by the Pontryagin maximum principle . . . 72

3.3 Convergence analysis of the SQH scheme . . . 78

3.4 Numerical experiments . . . 88

3.4.1 Application to a linear parabolic case . . . 90

3.4.2 Application to a linear elliptic case . . . 99

3.4.3 Application to a bilinear elliptic case . . . 101

3.4.4 Application to a bilinear parabolic case . . . 104

3.4.5 Application to a non-linear elliptic case . . . 106

3.4.6 Application to a state-constrained optimal control problem . . . 108

3.4.7 Application to an elliptic optimal control problem with L^1-tracking term . . . 112

4 An SQH framework for Fokker-Planck control problems **117**

4.1 From Random walk to Fokker-Planck . . . 117

4.1.1 A random walk with reflecting and absorbing barriers . . . 117

4.1.2 Optimal control of a random walk with jumps . . . 122

4.1.3 Derivation of the Fokker-Planck equation . . . 128

4.1.4 Numerical experiments . . . 135

4.2 The formulation of Fokker-Planck optimal control problems . . . 139

4.3 The characterization by the Pontryagin maximum principle . . . 145

4.4 Numerical schemes . . . 150

4.5 Numerical experiments . . . 161

5 Appendix **165**

5.1 General auxiliary results . . . 165

5.2 Ordinary differential equations . . . 168

5.3 Partial differential equations . . . 171

5.3.1 Elliptic partial differential equations . . . 171

5.3.2 Parabolic partial differential equations 174
5.3.3 Fokker-Planck equation 179
5.4 PMP sufficient conditions for an optimal solution 182
5.5 Discussion of the Assumptions A.1) to A.6) from Chapter 2 184
5.6 Description of the provided MATLAB files 184

Danksagung

Zunächst möchte ich gerne erwähnen, was mir beim Erarbeiten der präsentierten Resultate sehr hilfreich war und so zum Gelingen meiner Arbeit beigetragen hat. Auf diesem Wege möchte ich meinen Dank für die Unterstützung zum Ausdruck bringen, die ich erfahren durfte.

Als erstes möchte ich mich bei Alfio bedanken, der mir die Möglichkeit zur Promotion bei ihm eröffnet hat und sich in der ganzen Zeit sehr um meine Ausbildung gekümmert hat. Hier ist vor allem sein unermüdlicher Einsatz zu erwähnen, mich beim nachvollziehbaren schriftlichen Darstellen meiner Ergebnisse voranzubringen. Weiterhin waren seine kritischen Kommentare über meine Arbeit ein Gewinn für mich, denn diese haben mich oft zu ihrer Verbesserung angestoßen.

Als nächstes möchte ich mich für das große Verständnis meiner Freunde und Familie bedanken, wenn ich Zeit für mich gebraucht habe oder den Fokus für eine gewisse Zeit auf die Arbeit gelegt habe, wenn sie mich sehr beschäftigt hat.

Besonders sei das Engagement meiner Eltern erwähnt, die mich so oft sie können unterstützen und es mir damit ermöglichen, den Großteil meiner Zeit für die Entwicklung meiner Fähigkeiten einsetzen zu können.

Dann möchte ich mich für das sehr kollegiale Arbeitsklima am Institut bedanken, sodass man auf kurzem Wege fachlichen Austausch bekommt. Besonders möchte ich mich bei meinen Mitdoktorandinnen und Doktoranden bedanken, die sich die Zeit genommen haben, mit mir fachliche und andere Diskussionen zu führen, die für mich oft sehr fruchtbar waren.

Schließlich möchte ich noch erwähnen, dass ich mir bewusst bin, dass zum Gelingen einer solchen Arbeit nicht ausschließlich eigenes Vermögen und Leistungsbereitschaft wichtig sind, sondern neben Gesundheit auch glückliche Zufälle einen Beitrag leisten, die einem erst Gelegenheit geben, zu zeigen was man kann. Für diese Dissertation gab es den folgenden Zufall, als Alfio von einer Konferenz ein Paper mitgebracht hat, in dem eine Grundversion des Schemas veröffentlicht war, dessen Weiterentwicklung Thema dieser Arbeit ist. Dies war die Initialzündung für die Ausarbeitung und Entwicklung des präsentierten Schemas, die zu ergiebigen Resultaten geführt hat. Illustrativ lässt sich das glückliche Ereignis so beschreiben, dass Alfio einen Fisch an Land gezogen hat, den ich dann zu einem Gericht zubereiten konnte.

Zusammenfassend bin ich sehr dankbar in ein Umfeld gekommen zu sein, in dem es mir möglich war, meine Talente und Fähigkeiten weiter entwickeln zu können und mein Wirken zu Ergebnissen geführt hat, aus denen ich die vorliegende Dissertation erstellen konnte.

Sich zu mühen und mit dem Widerstande zu kämpfen, ist dem Menschen Bedürfnis, wie dem Maulwurf das Graben.

— Arthur Schopenhauer —

Zusammenfassung

Diese Dissertation handelt von einem neuen so genannten sequentiellen quadratischen Hamilton (SQH) iterativen Schema um Optimalsteuerungsprobleme mit Differentialmodellen und Kostenfunktionalen, die von glatt bis zu unstetig und nicht-konvex reichen, zu lösen. Dieses Schema basiert auf dem Pontryagin Maximumprinzip (PMP), welches notwendige Optimalitätsbedingungen für eine optimale Lösung zur Verfügung stellt.

In diesem Rahmen wird eine Hamiltonfunktion definiert, die ihr Minimum punktweise an der optimalen Lösung des entsprechenden Optimalsteuerungsproblems annimmt. In diesem SQH Schema wird diese Hamiltonfunktion durch einen quadratischen Strafterm erweitert, der aus der aktuellen Steuerungsfunktion und der Steuerungsfunktion aus der vorherigen Iteration besteht. Das Herzstück des SQH Schemas ist die punktweise Minimierung dieser erweiterten Hamiltonfunktion um eine Aktualisierung der Steuerungsfunktion zu bestimmen. Da das PMP keine Differenzierbarkeit in Bezug auf das Steuerungsfunktionsargument verlangt, kann das SQH Schema dazu benutzt werden, Optimalsteuerungsprobleme mit sowohl glatten als auch nicht-konvexen oder sogar unstetigen Kostenfunktionalen zu lösen.

Das Hauptergebnis dieser Dissertation ist die Formulierung eines robusten und effizienten SQH Schemas und eines Rahmens, in dem die Konvergenzanalyse des SQH Schemas ausgeführt werden kann. In diesem Rahmen bedeutet Konvergenz des Schemas, dass die berechnete Lösung die PMP Bedingung erfüllt.

Die steuernden Differentialmodelle der betrachteten Optimalsteuerungsprobleme sind gewöhnliche Differentialgleichungen (ODEs) und partielle Differentialgleichungen (PDEs). Im PDE Fall werden elliptische und parabolische Gleichungen, sowie die Fokker-Planck (FP) Gleichung betrachtet. Für sowohl den ODE als auch den PDE Fall werden Annahmen formuliert, für die bewiesen werden kann, dass eine Lösung eines Optimalsteuerungsproblems das PMP erfüllen muss. Die erhaltenen Resultate sind für die Diskussion der Konvergenzanalyse des SQH Schemas essentiell. Diese Analyse hat zwei Teile. Der erste ist die Wohlgestelltheit des Schemas, was bedeutet, dass alle Schritte des Schemas ausgeführt werden können und ein Ergebnis in endlicher Zeit liefern. Der zweite Teil ist die PMP Konsistenz der Lösung. Das bedeutet, dass die Lösung des SQH Schemas die PMP Bedingungen erfüllt.

Im ODE Fall werden die folgenden Resultate erhalten, die die Wohlgestelltheit des Schemas und die PMP Konsistenz der entsprechenden Lösung darlegen. Lemma 7 legt die Existenz eines punktweisen Minimums der erweiterten Hamiltonfunktion dar. Lemma 11 beweist die Existenz eines Gewichtes des quadratischen Strafterms, sodass die Minimierung der entsprechenden erweiterten Hamiltonfunktion zu einer Kontrollaktualisierung führt, die den Wert des Kostenfunktionals verringert. Lemma 12 legt dar, dass das SQH Schema stehen bleibt falls eine Iterierte PMP optimal ist. Satz 13 beweist die Kostenfunktional verringernden Eigenschaften der SQH Steuerungsfunktionsaktualisierung. Das Hauptresultat ist in Satz 14 gegeben, welches die punktweise Konvergenz des SQH Schemas gegen eine PMP konsistente Lösung darlegt. Das SQH-Verfahren wird in diesem ODE Rahmen auf zwei Optimalsteuerungsprobleme angewendet. Das erste ist ein optimales Quantensteuerungsproblem, bei dem gezeigt wird, dass das SQH-Verfahren viel schneller zu einer optimalen Lösung konvergiert als ein globalisiertes Newton-Verfahren. Das zweite Optimalsteuerungsproblem ist ein optimales Tumorbehandlungsproblem mit einem System gekoppelter hochgradig nicht-linearer Zustandsgleichungen, die das Tumorwachstum beschreiben. Es wird gezeigt, dass der Rahmen, in dem die Konvergenz des SQH Schemas bewiesen wird, auf diesen hochgradig nicht-linearen Fall anwendbar ist.

Als nächstes wird der Fall von PDE Optimalsteuerungsprobleme betrachtet. Zunächst wird ein allgemeiner Rahmen diskutiert, in dem eine Lösung des entsprechenden Optimalsteuerungsproblem die PMP Bedingungen erfüllt. In diesem Fall werden viele theoretische Abschätzungen in Satz 59 und Satz 64 bewiesen, die insbesondere die essentielle Beschränktheit von Zustands- und Adjungiertenvariablen beweisen. Die Schritte für die Konvergenzanalyse des SQH Schemas sind analog zu denen des ODE Falls und führen zu Satz 27, der die PMP Konsistenz der Lösung, erhalten durch das SQH Schemas, darlegt. Dieser Rahmen wird auf verschiedene elliptische und parabolische Optimalsteuerungsprobleme angewendet, die lineare und bilineare Steuerungsmechanismen beinhalten, genauso wie nicht-lineare Zustandsgleichungen. Darüber hinaus wird das SQH-Verfahren zum Lösen eines zustandsbeschränkten Optimalsteuerungsproblems in einer erweiterten Formulieren diskutiert. Es wird in Satz 30 gezeigt, dass wenn man das Gewicht des Erweiterungsterms, der die Verletzung der Zustandsbeschränkung bestraft, erhöht, das Maß dieser Zustandsbeschränkungsverletzung durch die entsprechende Lösung gegen null konvergiert. Weiterhin wird ein Optimalsteuerungsproblem mit einem nicht-glatten L^1-Zielverfolgungsterm und einer nicht-glatten Zustandsgleichung untersucht. Für diesen Zweck wird eine adjungierte Gleichung definiert und das SQH-Verfahren wird benutzt um das entsprechende Optimalsteuerungsproblem zu lösen.

Der letzte Teil dieser Dissertation ist einer Klasse von FP Modellen gewidmet, die auf bestimmte stochastische Prozesse bezogen sind. Die Diskussion beginnt mit dem Fokus auf Random Walks bei dem auch Sprünge mit enthalten sind. Dieser Rahmen erlaubt die Herleitung eines diskreten FP Modells, das einem kontinuierlichen FP Modell mit Sprüngen und Randbedingungen entspricht, die sich zwischen absorbierend bis komplett reflektierend bewegen. Diese Diskussion erlaubt die Betrachtung der Driftsteuerung, die aus einer anisotropen Wahrscheinlichkeit für die Schritte des Random Walks resultiert. Danach werden zwei Drift-Diffusionsprozesse und die entsprechenden FP Modelle mit zwei verschiedenen Steuerungsstrategien für ein Optimalsteuerungsproblem mit Erwartungswertfunktional betrachtet. In der ersten Strategie hängen die Steuerungsfunktionen von der Zeit ab und in der zweiten hängen die Steuerungsfunktionen von Ort und Zeit ab. In beiden Fällen wird eine Lösung zum entsprechendem Optimalsteuerungsproblem mit den PMP Bedingungen charakterisiert, dargestellt in Satz 48 und Satz 49. Die Wohlgestelltheit des SQH Schemas ist in beiden Fällen gezeigt und weitere Bedingungen, die die Konvergenz des SQH Schemas zu einer PMP konsistenten Lösung sicherstellen, werden diskutiert. Der Fall einer Ort und Zeit abhängigen Steuerungsstrategie führt auf eine spezielle Struktur der entsprechenden PMP Bedingungen, die in einem weiteren Lösungsverfahren ausgenutzt werden, dem sogenannten direkten Hamiltonfunktionsverfahren (DH).

Summary

This thesis deals with a new so-called sequential quadratic Hamiltonian (SQH) iterative scheme to solve optimal control problems with differential models and cost functionals ranging from smooth to discontinuous and non-convex. This scheme is based on the Pontryagin maximum principle (PMP) that provides necessary optimality conditions for an optimal solution.

In this framework, a Hamiltonian function is defined that attains its minimum pointwise at the optimal solution of the corresponding optimal control problem. In the SQH scheme, this Hamiltonian function is augmented by a quadratic penalty term consisting of the current control function and the control function from the previous iteration. The heart of the SQH scheme is to minimize this augmented Hamiltonian function pointwise in order to determine a control update. Since the PMP does not require any differentiability with respect to the control argument, the SQH scheme can be used to solve optimal control problems with both smooth and non-convex or even discontinuous cost functionals.

The main achievement of the thesis is the formulation of a robust and efficient SQH scheme and a framework in which the convergence analysis of the SQH scheme can be carried out. In this framework, convergence of the scheme means that the calculated solution fulfills the PMP condition.

The governing differential models of the considered optimal control problems are ordinary differential equations (ODEs) and partial differential equations (PDEs). In the PDE case, elliptic and parabolic equations as well as the Fokker-Planck (FP) equation are considered. For both the ODE and the PDE cases, assumptions are formulated for which it can be proved that a solution to an optimal control problem has to fulfill the PMP. The obtained results are essential for the discussion of the convergence analysis of the SQH scheme. This analysis has two parts. The first one is the well-posedness of the scheme which means that all steps of the scheme can be carried out and provide a result in finite time. The second part part is the PMP consistency of the solution. This means that the solution of the SQH scheme fulfills the PMP conditions.

In the ODE case, the following results are obtained that state well-posedness of the SQH scheme and the PMP consistency of the corresponding solution. Lemma 7 states the existence of a pointwise minimum of the augmented Hamiltonian. Lemma 11 proves the existence of a weight of the quadratic penalty term such that the minimization of the corresponding augmented Hamiltonian results in a control updated that reduces the value of the cost functional. Lemma 12 states that the SQH scheme stops if an iterate is PMP optimal. Theorem 13 proves the cost functional reducing properties of the SQH control updates. The main result is given in Theorem 14, which states the pointwise convergence of the SQH scheme towards a PMP consistent solution. In this ODE framework, the SQH method is applied to two optimal control problems. The first one is an optimal quantum control problem where it is shown that the SQH method converges much faster to an optimal solution than a globalized Newton method. The second optimal control problem is an optimal tumor treatment problem with a system of coupled highly non-linear state equations that describe the tumor growth. It is shown that the framework in which the convergence of the SQH scheme is proved is applicable for this highly non-linear case.

Next, the case of PDE control problems is considered. First a general framework is discussed in which a solution to the corresponding optimal control problem fulfills the PMP conditions. In this case, many theoretical estimates are presented in Theorem 59 and Theorem 64 to prove in particular the essential boundedness of the state and adjoint variables. The steps for the convergence analysis of the SQH scheme

are analogous to that of the ODE case and result in Theorem 27 that states the PMP consistency of the solution obtained with the SQH scheme. This framework is applied to different elliptic and parabolic optimal control problems, including linear and bilinear control mechanisms, as well as non-linear state equations. Moreover, the SQH method is discussed for solving a state-constrained optimal control problem in an augmented formulation. In this case, it is shown in Theorem 30 that for increasing the weight of the augmentation term, which penalizes the violation of the state constraint, the measure of this state constraint violation by the corresponding solution converges to zero. Furthermore, an optimal control problem with a non-smooth L^1-tracking term and a non-smooth state equation is investigated. For this purpose, an adjoint equation is defined and the SQH method is used to solve the corresponding optimal control problem.

The final part of this thesis is devoted to a class of FP models related to specific stochastic processes. The discussion starts with a focus on random walks where also jumps are included. This framework allows a derivation of a discrete FP model corresponding to a continuous FP model with jumps and boundary conditions ranging from absorbing to totally reflecting. This discussion allows the consideration of the drift-control resulting from an anisotropic probability of the steps of the random walk. Thereafter, in the PMP framework, two drift-diffusion processes and the corresponding FP models with two different control strategies for an optimal control problem with an expectation functional are considered. In the first strategy, the controls depend on time and in the second one, the controls depend on space and time. In both cases a solution to the corresponding optimal control problem is characterized with the PMP conditions, stated in Theorem 48 and Theorem 49. The well-posedness of the SQH scheme is shown in both cases and further conditions are discussed that ensure the convergence of the SQH scheme to a PMP consistent solution. The case of a space and time dependent control strategy results in a special structure of the corresponding PMP conditions that is exploited in another solution method, the so-called direct Hamiltonian (DH) method.

Chapter 1

Introduction

One central topic of optimal control theory is the investigation of necessary optimality conditions that a solution to an optimal control problem hast to fulfill. In this thesis, we consider (among others) optimal control problems having the following structure

$$\begin{aligned} &\min_{y,u} J\left(y,u\right) := \int_Z \left(h\left(y\right) + g\left(u\right)\right) dx \\ &\text{such that } c\left(y,u\right) = 0 \text{ and } u \in U_{ad} \end{aligned} \tag{1.1}$$

where the state y and the control u can depend on time or space variables or both depending on the differential model $c\left(y,u\right) = 0$. The function h determines the objective of the state y at each point of the domain Z and g determines the cost of the control u at each point of Z. The task is to find a control in the admissible set U_{ad} such that the cost functional J is minimized subject to the governing differential model $c\left(y,u\right) = 0$. Such an optimal control is called a solution to (1.1). For this purpose, necessary conditions that such a control has to fulfill are exploited for designing efficient numerical solution algorithms. For example, a common approach for the first-order characterization of a solution to optimal control problems is the Lagrange approach [48, 55, 95, 54] where first and second order methods are used for the numerical calculation of a solution [96, 19]. However, in this framework the calculation of a subdifferential is necessary, and the existence of a subdifferential requires some smoothness of the problem like directionally differentiability or convexity of the reduced cost functional [96, 82, 58, 10].

In this thesis, we focus on the Pontryagin maximum principle (PMP) [67, 81, 16, 77, 42, 28, 92, 93, 41, 35, 80, 95] that provides an alternative characterization of a solution to an optimal control problem. In this framework, a Hamiltonian function is defined that attains its minimum pointwise at the optimum with respect to all possible values of the control. Thus a derivative of the cost functional with respect to the control argument is in general not necessary in order to characterize a solution. This is the starting point for our considerations in this thesis where we use the minimum of a Hamiltonian function instead of a gradient or elements of a subdifferential in order to obtain a PMP consistent scheme. This allows us to consider not only optimal control problems with smooth but also with non-convex and even discontinuous cost functionals without the need of regularization techniques as in [70, 53, 56, 57].

In this thesis, we consider different cost functionals where the cost of the controls can be continuous and convex or just lower semi-continuous [3]. If the corresponding cost functional is weakly lower semi-continuous, then the existence of an optimal solution can be obtained with variational techniques, see [95] for instance. However, difficulties arise in the case of cost functions that are only lower semi-continuous as the following discussion illustrates.

Consider the lower semi-continuous function

$$g : \mathbb{R} \to \mathbb{R}, \quad z \mapsto g\left(z\right) := \begin{cases} 1 & \text{if } z \neq 0 \\ 0 & \text{if } z = 0 \end{cases},$$

which is associated with the so-called L^0-norm $\int_\Omega g\left(u\left(x\right)\right)dx$ of a function $u:\Omega\to\mathbb{R}$ on the open set $\Omega\subseteq\mathbb{R}^n$, $n\in\mathbb{N}$. We choose $\Omega=(0,1)$ and

$$u\left(x\right):=\begin{cases}1 & \text{if } x\in\left[\frac{1}{2}+k,1+k\right)\\ 0 & \text{if } x\in\left(0+k,\frac{1}{2}+k\right)\end{cases},\ k\in\mathbb{N}_0$$

for $x\in\mathbb{R}$. Then we choose a sequence $u_m\left(x\right):=u\left(mx\right)$, $m\in\mathbb{N}$ for $x\in\Omega$. According to [7, Proposition 1] or [59, A property of the mean value], we have that the sequence $\left(u_m\right)_{m\in\mathbb{N}}$ weakly converges to the mean value of u on Ω, which is $\frac{1}{2}$ representing the weak limit denoted with $\bar{u}$. Then we have, since $\bar{u}$ is a constant function on Ω, that $\int_0^1 g\left(\bar{u}\left(x\right)\right)dx=1$. Furthermore, we have by a direct calculation that $\int_0^1 g\left(u_m\left(x\right)\right)dx=\frac{1}{2}$. This means that we have a sequence $\left(u_m\right)_{m\in\mathbb{N}}$ weakly converging to $\bar{u}$ which contradicts the condition of weakly lower semi-continuity as follows

$$\frac{1}{2}=\liminf_{m\to\infty}\int_0^1 g\left(u_m\left(x\right)\right)dx<\int_0^1 g\left(\bar{u}\left(x\right)\right)dx=1.$$

Consequently in the case of a so-called L^0-cost functional, the proof of existence of an optimal solution is not possible with a direct variational technique.

We remark that for g being lower semi-continuous, the existence of a minimizer can be proven on a compact admissible set, see [21, Theorem A.2]. However, in order to characterize this solution with the PMP condition, the technique of needle variation is required in this thesis. The values of a function and of its needle variation differ at most on a ball centered at an arbitrary point of the domain where the values of the needle variation are set to a constant value from an admissible set of values on this ball. In order to apply the technique of needle variation, it is necessary that all needle variations of all admissible controls, that means the needle variation of any admissible control at any point of the domain and any radius of the ball, are included in the admissible set. Thus the compactness of the admissible set contradicts the accommodation of all needle variations of all admissible controls.

Since the focus of our work is the PMP characterization of optimal controls and their computation by our PMP-based optimization solver, we consider bounded convex and closed admissible control sets in Lebesgue spaces (which are not compact) and assume existence of optimal controls in these sets.

However, notice that our approach covers all cases where continuous and convex cost functionals appear, and in these cases, we prove existence of optimal solutions in the sets mentioned above. On the other hand, our work provides a framework to address control problems that are beyond the continuous and convex cases, for which a PMP characterization is possible and the related optimization procedure constructs a minimizing sequence that converges to a point satisfying the PMP condition.

Next, we give an overview of existing work about numerical schemes based on the PMP and show how our presented scheme is related to them. In particular, we refer to the work [17, 40, 85, 88, 97] and further [27, 62, 63, 69, 73, 87] where in both cases a Hamiltonian function is pointwise minimized in order to calculate the next iterate of the control function. We consider two major variants in the class of schemes using the PMP. The first one, the so-called successive iteration scheme, originates in [62, 63] where the Hamiltonian function from the PMP is pointwise minimized. For any pointwise update of the control, the state variable from the previous iteration is used. This strategy results in an efficient calculation method because the number of solving the state equation is kept small. However, this method is not robust with respect to its convergence behavior since the minimization of the Hamiltonian is independent of the control-to-state map and thus does not consider the variation of the state with respect to the control. Furthermore, there is a lack of convergence theory for this class of PMP based schemes. The second major variant of PMP based methods originates in [85, 88] where the Hamiltonian function is augmented with a quadratic penalization term consisting of the difference between the current control value and the one from the previous iteration. Furthermore, the minimization of the augmented Hamiltonian depends on the control-to-state map since with any pointwise update of the control the state variable is updated. For this class of schemes, there is convergence theory available, see [17]. The theory says that for smooth cost functionals the corresponding scheme converges to an optimal solution that fulfills the variational inequality corresponding to the optimality condition of the Lagrange approach. However, in this latter framework it is required to update the state after each pointwise update of the control. Consequently, this class suffers from a large computational effort for minimizing the augmented Hamiltonian. In the present thesis, we develop a new scheme that combines the advantages of both methods mentioned above.

In our method, that we call sequential quadratic Hamiltonian (SQH) method, an augmented Hamiltonian is considered and is pointwise minimized where the state variable from the last iteration is used and a quadratic penalization term consisting of the difference between the current control value and the one from the previous iteration is included. This ensures that the update of the control is sufficiently small such that the state variable of the previous iteration is still a good approximation for the currently updated control while by fixing the state variable the minimization process avoids to be too computationally expensive. In this setting, we provide a framework that allows to prove convergence of this scheme to a PMP consistent solution, which means that the obtained solution fulfills the PMP conditions for optimality. Besides the theoretical investigation of the convergence of the SQH scheme, we also show that the results from the corresponding algorithm can be numerically checked for optimality with the PMP and thus validate the proposed framework.

In Chapter 2, we consider optimal control problems governed by ordinary differential equations (ODEs). We set up a general framework in which we prove the PMP characterization of an optimal solution and the convergence of our SQH scheme. The characterization with the PMP follows essentially the reasoning in [81]. The corresponding result is Theorem 5. Next, we prove the well-posedness of our scheme and its convergence to a PMP consistent solution. The main result of this chapter is Theorem 14, which states the convergence of the SQH method to a PMP consistent control. For this purpose, we formulate an assumption in (2.32) that ensures a sufficient descent of the augmented Hamiltonian in each sweep of the SQH method such that the iterates converge pointwise to a solution that is PMP optimal. This condition replaces the requirement of differentiability of the augmented Hamiltonian with respect to the control argument. It is shown that all requirements of this theorem are fulfilled for several cost functionals including smooth L^2- and non-smooth L^1-functionals for which we prove existence of optimal controls. Next, we demonstrate the applicability of our framework, defined in Section 2.1, with two cases and show that all our requirements are fulfilled to obtain convergence of our SQH method. The first case is an optimal quantum control problem with a bilinear control mechanism. The second case is an optimal tumor control problem where the tumor growth is modeled with highly non-linear coupled state equations. Both optimal control problems have non-smooth cost functionals. We show existence of an optimal solution and prove that both cases are included in our theoretical framework such that we have the PMP characterization of an optimal solution and the convergence of the SQH method. For the quantum optimal control problem, we choose L^2-cost and L^1-cost terms where we compare the numerical performance of the SQH method with the performance of a globalized Newton method. In this experiment, the SQH scheme converges much faster than the globalized Newton method where the output of both methods are PMP optimal. Then we replace the L^1-cost term by an L^0-cost term and show that the L^0-cost term appears advantageous with respect to the total convergence time of the SQH method. In the second application, we consider a model for optimal tumor treatment consisting of anti-angiogenesis and irradiation. In this section, we apply the SQH method to an optimal control problem with a system of coupled highly non-linear state equations that model the dynamics of tumor growth. We verify that this optimal control problem is covered by our theoretical framework such that also in this case convergence of our SQH method is proved and verified by our numerical PMP test.

In Chapter 3, we extend the framework of Chapter 2 to optimal control problems governed by elliptic and parabolic partial differential equations (PDEs). With the same procedure as in the previous chapter we show how to characterize a solution to an optimal control problem with the PMP and perform the convergence analysis of the SQH method in the PDE case. For this purpose, we define a general framework that holds for both the elliptic and the parabolic cases and prove convergence of our SQH method to a PMP consistent solution. For this analysis, we prove L^∞-estimates for the solution of the governing state model and for the corresponding adjoint equation. The discussion about the PMP characterization of solutions to our PDE control problems results in Theorem 25. The convergence analysis is analogous to the ODE case, although more involved, and the convergence to a PMP consistent solution is proved in Theorem 27. Notice that, for the case of a differentiable control cost, we also prove in Theorem 29 that the SQH iterates converge to the solution of the variational inequality characterizing an optimal control in the Lagrange framework.

Next, we demonstrate the applicability of this framework to seven different PDE optimal control problems. In particular in the parabolic case, we consider a linear and a bilinear control mechanism. In the elliptic case, we have a linear and a bilinear control mechanism and in addition we consider a non-linear state equation with distributed control. We show that in these cases the requirements of our framework are fulfilled. Furthermore, we consider a state-constrained optimal control problem and an optimal control problem with an L^1-tracking term with a non-smooth state equation and show how to apply the SQH method in these cases. In these problems, we have non-convex and discontinuous cost functionals. We discuss the PMP optimality of solutions to the considered optimal control problems (assuming they exist) and demonstrate how to obtain these solutions by the SQH method. We observe the convergence of the SQH method to PMP consistent solutions. Furthermore, we investigate the performance of the SQH method compared with a projected gradient method (pGM) and a projected non-linear conjugated gradient (pNCG) method in the case of smooth cost functionals, showing that the SQH method is much faster than the pGM and almost comparable with the pNCG method. In addition, we investigate the numerical complexity of the SQH method that is linear in these cases.

We remark that our PMP framework is also appropriate to solve mixed-integer PDE control problems and we demonstrate the applicability of the SQH method to optimal control problems where the values of the control are in a discrete set. This demonstrates that the numerical treatment of mixed-integer problems, like in [52], is also in the scope of our framework, since the minimum of the augmented Hamiltonian in this case is given by an array search.

In the case of state-constrained optimal control problems, we formulate an alternative optimal control problem where the state constraint is replaced by an augmented objective. This means that we add a penalization term to the objective that restricts the violation of the bounds of the constraint. The solution to the augmented optimal control problem is characterized by the PMP and it is shown that the SQH method converges to a PMP consistent solution. In this case, in Theorem 30, we prove that increasing the penalization parameter decreases the discrepancy between the state bounds and the state. In the case of an L^1-tracking term, a solution to the corresponding optimal control problem cannot be characterized by the PMP within our framework. However, we demonstrate how to define an adjoint equation in this case and we observe that the solution of the SQH method results in a reduction of the cost functional and fulfills the PMP test.

In Chapter 4, we focus on a drift-controlled Fokker-Planck (FP) equation. The FP equation models the evolution of the probability density function of stochastic processes. We start our investigation by considering a random walk with jumps and different boundary conditions ranging from absorbing to totally reflecting. This results in different discrete FP models from which we derive continuous FP equations. This discussion illustrates the macroscopic model parameters of the Fokker-Planck equation by the model parameters of a microscopic random walk. In particular, this investigation allows a discussion of the drift-control mechanism that results from an anisotropic probability for the steps of the random walk. Next, we formulate continuous FP optimal control problems with two different control mechanisms. In the first one, the controls are time dependent and the space dependency is explicitly given by a bilinear control structure. In the second one, the drift is a space and time depended optimal control. We characterize in both cases a solution to the optimal control problems with the PMP, resulting in Theorem 48 and Theorem 49. Furthermore, we show that the SQH method is well-defined and discuss the convergence of the scheme. We validate our results with Monte-Carlo simulations.

In the Appendix, we provide technical results that we use for our discussion and proofs in this thesis. Specifically, we discuss the measurability of functions that are pointwise determined as the result of an arg min-function, which is the basic procedure in the SQH method. Furthermore, we prove L^∞-results for elliptic and parabolic PDEs. In addition, we discuss that the adjoint variable corresponding to our FP optimal control problem is bounded in the L^2-norm by the L^2-norm of the corresponding controls. Then, we discuss a sufficient PMP based condition for an optimal solution. This condition is given in (5.34) and is a growth condition with respect to the control argument of the Hamiltonian. In Corollary 69, we focus on the special case where the optimal control problem consists of an L^2-tracking term and a distributed control mechanism. We show that any pair of a state and a control variable that fulfills the corresponding necessary PMP optimality conditions is a solution to this optimal control problem. Then we discuss the general assumptions made in Section 2.1 with respect to replacing single assumptions by alternatives, in particular the weakening of these assumptions, in the case that we only aim at characterizing a solution with the PMP. In Section 5.6, we describe the numerical codes that implement the SQH scheme in the numerical experiments.

The results presented in this thesis are partly based on the following publications:

- [21]: Tim Breitenbach and Alfio Borzì, "A sequential quadratic Hamiltonian method for solving parabolic optimal control problems with discontinuous cost functionals", Journal of Dynamical and Control Systems (2018), pp. 1–33.

- [20]: Tim Breitenbach, Mario Annunziato and Alfio Borzì, "On the Optimal Control of a Random Walk with Jumps and Barriers", Methodology and Computing in Applied Probability 20, 1 (2018), pp. 435–462.

- [47]: Melina-Lorén Kienle Garrido, Tim Breitenbach, Kurt Chudej and Alfio Borzì, "Modeling and Numerical Solution of a Cancer Therapy Optimal Control Problem", Applied Mathematics 9 (2018), pp. 985–1004.

- [25]: Tim Breitenbach, Mario Annunziato and Alfio Borzì, "On the optimal control of random walks", IFAC-PapersOnLine 49, 8 (2016), pp. 248–253.

- [22]: Tim Breitenbach and Alfio Borzì, "On the SQH scheme to solve non-smooth PDE optimal control problems”, Journal of Numerical Functional Analysis and Optimization (2019), pp. 1–43, 2019.

- [24]: Tim Breitenbach and Alfio Borzì, "A sequential quadratic Hamiltonian scheme for solving non-smooth quantum control problems with sparsity”, submitted to the Journal of Computational and Applied Mathematics, 2019.

- [23]: Tim Breitenbach and Alfio Borzì, "The Pontryagin maximum principle for solving Fokker-Planck optimal control problems”, submitted to the Journal of Computational Optimization and Applications, 2019.

Chapter 2

An SQH framework for ODE optimal control problems

In this chapter, we discuss the characterization of solutions to optimal control problems governed by ordinary differential equations (ODEs) in the framework of the Pontryagin maximum principle (PMP) and their computation by a sequential quadratic Hamiltonian (SQH) method. The applicability of the SQH method for solving a quantum optimal control problem with a bilinear control mechanism and for solving an optimal tumor treatment problem with a highly non-linear ODE model both with non-smooth L^1-cost functionals is demonstrated. The existence of optimal solutions, their characterization with the PMP conditions for optimality and the convergence of the SQH method to a PMP consistent solution are proved.

2.1 The formulation of ODE optimal control problems

Consider the following initial value problem

$$\begin{aligned} y'(t) &= f\left(t, y\left(t\right), u\left(t\right)\right) \text{ for } t \in (0,T) \\ y\left(0\right) &= y_0 \end{aligned} \tag{2.1}$$

on the interval $[0,T]$, $T > 0$ with $y : \mathbb{R} \to \mathbb{R}^n$, $t \mapsto y\left(t\right)$, the time derivative $y' := \frac{d}{dt}y$ and $y_0 \in \mathbb{R}^n$, $n \in \mathbb{N}$, the initial value of the state. The control function $u : \mathbb{R} \to \mathbb{R}^m$, $t \mapsto u\left(t\right)$, $m \in \mathbb{N}$ can be chosen from the following admissible set of controls

$$U_{ad} := U_{ad}^1 \times ... \times U_{ad}^m$$

with $U_{ad}^j := \left\{u \in L^2\left(0,T\right) \mid u\left(t\right) \in K_U^j \text{ a.e.}\right\}$, K_U^j a compact set in $\mathbb{R}$, $j \in \{1,...,m\}$ and $K_U := K_U^1 \times ... \times K_U^m$. We assume that (2.1) is uniquely solvable in the sense of [90, Definition C.2.1] on the interval $[0,T]$ for any $u \in U_{ad}$. Therefore there exists a function $y : [0,T] \to \mathbb{R}^n$, $t \mapsto y\left(t\right)$ that is absolutely continuous on $[0,T]$ and fulfills the following integral equation

$$y\left(t\right) = y_0 + \int_0^t f\left(\tilde{t}, y\left(\tilde{t}\right), u\left(\tilde{t}\right)\right) d\tilde{t} \tag{2.2}$$

for any $t \in [0,T]$ and any chosen $u \in U_{ad}$, see [90, Definition C.2.1]. We call such a function a global solution to (2.1). The definition of absolute continuity [90, page 471] also implies that any component function $y_i : [0,T] \to \mathbb{R}$, $i \in \{1,...,n\}$ is absolutely continuous. Furthermore, we define the set $I \subseteq \mathbb{R}^n$ as the convex hull [10, Section 3.1] of the union of all images from each solution y to (2.1) for any $u \in U_{ad}$, given by

$$I := \text{conv}\left\{y\left([0,T]\right) \subseteq \mathbb{R}^n \middle| \, y \text{ solves (2.1) for } u \in U_{ad}\right\}.$$

Our purpose is to investigate the following optimal control problem

$$\begin{aligned} \min_{y,u} J\left(y,u\right) := \int_0^T \left(h\left(y\left(t\right)\right) + g\left(u\left(t\right)\right)\right) dt + F\left(y\left(T\right)\right) \\ y' = f\left(t, y\left(t\right), u\left(t\right)\right) \quad \text{for } t \in \left(0,T\right) \\ y\left(0\right) = y_0 \\ u \in U_{ad} \end{aligned} \tag{2.3}$$

where we have $g : K_U \to \mathbb{R}$, $u \mapsto g\left(u\right)$ a lower semi-continuous function. This means that for any sequence $\left(z_k\right)_{k\in\mathbb{N}} \subseteq K_U$ converging to $z \in K_U$ it holds that $\liminf_{k\to\infty} g\left(z_k\right) \geq g\left(z\right)$. Furthermore, we require that g is bounded from below. Additionally, we assume that the functions $h : I \to \mathbb{R}$, $y \mapsto h\left(y\right)$ and $F : I \to \mathbb{R}$, $y \mapsto F\left(y\right)$ are bounded from below.

Next, we discuss the existence of solutions to (2.3). If the cost functional of the optimal control problem is weakly lower semi-continuous, then the existence of a minimizer can be proven with the variational technique, see [95]. An example is a continuous and convex cost functional where the right hand-side of the ODE implements a linear or bilinear control mechanism, as in Section 2.5. In this section, the variational technique is applied to the case of optimal control problems governed by a highly non-linear system of ODEs with a bilinear control mechanism and a convex and continuous cost functional with L^2-L^1-costs of the control. However, our requirement for g being lower semi-continuous may result in cost functionals that are not weakly lower semi-continuous, as discussed in the Introduction. In this case, we assume that (2.3) is well defined and admits a solution.

In the following, we formulate further requirements that we underlie for our analysis. For this purpose, we need first and second derivatives of h, F, f with respect to y in the finite dimension framework as in [4, VII, VII.4] where we denote with D_y the first and with D_{yy} the second derivative with respect to the variable y. This notation also holds throughout the chapter. The L^∞-norm for any vector valued function $\tilde{\zeta} : \mathbb{R} \to \mathbb{R}^{\tilde{n}}$, $\tilde{n} \in \mathbb{N}$ is defined by $\|\tilde{\zeta}\|_{L^\infty} := \max_{i=1,\dots,\tilde{n}} \|\tilde{\zeta}_i\|_{L^\infty(0,T)}$ where $\tilde{\zeta}_i : \mathbb{R} \to \mathbb{R}$ is the i-th component of $\tilde{\zeta}$ and $\|\cdot\|_{L^\infty(0,T)}$ is the L^∞-norm for a real valued function, see [5, X.4] for a definition. Next we remark that integration over a vector valued function is componentwise defined and we have that $\|\int_0^t \tilde{\zeta}\left(\tilde{t}\right) d\tilde{t}\|_{L^\infty} \leq \int_0^t \|\tilde{\zeta}\left(\tilde{t}\right)\|_{L^\infty} d\tilde{t}$ for any $t \in [0,T]$, see [5, X Theorem 2.11]. The L^p-norm is defined as follows $\|\tilde{\zeta}\|_{L^p} := \left(\sum_{i=1}^n \|\tilde{\zeta}_i\|_{L^p(0,T)}^p\right)^{\frac{1}{p}}$ where $\|\tilde{\zeta}_i\|_{L^p(0,T)} := \left(\int_0^T |\tilde{\zeta}_i\left(t\right)|^p dt\right)^{\frac{1}{p}}$ is the L^p-norm for a real valued function, see [5, X.4] for a definition with $p \in (0,\infty)$. The assumptions are given as follows.

A.1) The functions $h : I \to \mathbb{R}$, $y \mapsto h\left(y\right)$, $F : I \to \mathbb{R}$, $y \mapsto F\left(y\right)$ and $f : I \to \mathbb{R}^n$, $y \mapsto f\left(t,y,u\right)$ are twice continuously differentiable for every $u \in K_U$ and for any $t \in [0,T]$.

A.2) The functions $f : [0,T] \times I \times K_U \to \mathbb{R}^n$, $\left(t,y,u\right) \mapsto f\left(t,y,u\right)$, $D_y f : [0,T] \times I \times K_U \to \mathbb{R}^{n\times n}$, $\left(t,y,u\right) \mapsto D_y f\left(t,y,u\right)$ and $D_{yy} f : [0,T] \times I \times K_U \to \mathbb{R}^{n\times n\times n}$, $\left(t,y,u\right) \mapsto D_{yy} f\left(t,y,u\right)$ are Borel measurable on $[0,T] \times I \times K_U$.

A.3) For almost all $t_0 \in (0,T)$ and for any y solving (2.1) where $u \in U_{ad}$ and any $\tilde{u} \in U_{ad}$ there exists an open set $E\left(t_0\right) \subseteq (0,T)$ containing t_0 such that $\int_{E(t_0)} f_i\left(t, y\left(t\right), \tilde{u}\left(t\right)\right) dt < \infty$ for all $i \in \{1,...,n\}$.

A.4) The function $f : I \times K_U \to \mathbb{R}^n$, $\left(y,u\right) \mapsto f\left(t,y,u\right)$ is continuous for all $t \in [0,T]$.

A.5) There exists a constant $L > 0$ such that the functions $f_i : K_U \to \mathbb{R}$, $u \mapsto f_i\left(t,y,u\right)$, $\frac{\partial}{\partial y_l} f_i : K_U \to \mathbb{R}$, $u \mapsto \frac{\partial}{\partial y_l} f_i\left(t,y,u\right)$ for $l,i \in \{1,...,n\}$ are Lipschitz continuous with $|f_i\left(t,y,u_1\right) - f_i\left(t,y,u_2\right)| \leq L\sum_{j=1}^m |\left(u_1\right)_j - \left(u_2\right)_j|$ and $|\frac{\partial}{\partial y_l} f_i\left(t,y,u_1\right) - \frac{\partial}{\partial y_l} f_i\left(t,y,u_2\right)| \leq L\sum_{j=1}^m |\left(u_1\right)_j - \left(u_2\right)_j|$ for any fixed $y \in I$ and $t \in [0,T]$. That means the Lipschitz constant L is independent of all $t \in [0,T]$, all $u \in K_U$ and all $y \in I$.

A.6) There exists a constant $c > 0$ such that $\| \frac{\partial}{\partial y_l} f_i(\cdot, y, u) \|_{L^\infty} \leq c$, $\| \frac{\partial}{\partial y_l} h(y) \|_{L^\infty} \leq c$, $\| \frac{\partial}{\partial y_l} F(y) \|_{L^\infty} \leq c$, $\| \frac{\partial^2}{\partial y_l \partial y_\ell} f_i(\cdot, y, u) \|_{L^\infty} \leq c$, $\| \frac{\partial^2}{\partial y_l \partial y_\ell} h(y) \|_{L^\infty} \leq c$ and $\| \frac{\partial^2}{\partial y_l \partial y_\ell} F(y) \|_{L^\infty} \leq c$ for all $i, l, \ell \in \{1, ..., n\}$ and all $y \in I$ and all $u \in K_U$.

Since h is twice continuously differentiable the functions $\frac{\partial}{\partial y_l} h \circ y : [0, T] \to \mathbb{R}$, $t \mapsto \frac{\partial}{\partial y_l} h(y(t))$, $l \in \{1, ..., n\}$ and $\frac{\partial^2}{\partial y_l \partial y_\ell} h \circ y : [0, T] \to \mathbb{R}$, $t \mapsto \frac{\partial^2}{\partial y_l \partial y_\ell} h(y(t))$, $l, i \in \{1, ..., n\}$, are Lebesgue measurable for any Lebesgue measurable function y. This can be seen with a similar proof to Lemma 51 as continuous function are Borel measurable [36, Examples 2.1.2] and the function $y : [0, T] \to \mathbb{R}^n$ is Lebesgue measurable since any component function y_i is Lebesgue measurable [36, Example 2.6.5] due to its continuity [36, page 42]. Then the composite functions are also measurable [36, Proposition 2.6.1]. Analogously the vector-valued function $(t, y, u) : [0, T] \to \mathbb{R}^{1 \times n \times m}$, $t \mapsto (t, y(t), u(t))$ is Lebesgue measurable as any component function is Lebesgue measurable [36, Example 2.6.5].

Remark 1. As the Assumption A.2) might be considered to be quite technical we remark that if the functions $(t, y, u) \mapsto f(t, y, u)$, $(t, y, u) \mapsto D_y f(t, y, u)$ and $(t, y, u) \mapsto D_{yy} f(t, y, u)$ are continuous on $[0, T] \times I \times K_U$, then Assumption A.2) is fulfilled [36, Examples 2.1.2]. By the continuity of $(t, y, u) \mapsto f(t, y, u)$ also Assumption A.4) is then fulfilled.

In order to check Assumption A.5) and Assumption A.6), we usually face the problem that the estimations depend on the state variable y. Therefore the usual way of checking Assumption A.5) and Assumption A.6) is to use the argument that continuous functions take their minimum and maximum on a compact set [3, III Corollary 3.8]. For this purpose, the boundedness and closedness of K_U can be obtained by construction of the admissible set. The boundedness of I can be proved as follows. The existence of a constant $K > 0$ such that $\|y\|_{L^\infty} < K$ for all (y, u) fulfilling (2.1) with $u \in U_{ad}$ is sufficient. For example, for $f(u) = u$, $f(y, u) = y + u$ or $f(y, u) = uy$, Gronwall's Lemma, see Lemma 57 in the appendix, guarantees the required boundedness for $u(t) \in K_U$ for almost all $t \in (0, T)$ with K_U bounded analogous to (2.55). However, any argument that ensures a solution that is bounded by a fixed constant for all $u \in U_{ad}$ is useful for this purpose. We show an example for this in Section 2.5 where the boundedness of the state is shown with an argument different from Gronwall's Lemma.

We remark that the triple (t, y, u) in the assumptions above reduces to the tuple (y, u) if the right hand-side does not explicitly depend on t.

In the rest of this chapter, we write $h_y(y) := (D_y h(y))^T$ for the transposed of the first derivative of h with respect to y, $f_y(t, y, u) := D_y f(t, y, u)$ the first partial derivative of f with respect to y for any $t \in [0, T]$. We remark that the possibly explicit time dependency of f is often neglected for the rest of this chapter, especially in proofs, in order to save notational effort.

2.2 The characterization by the Pontryagin maximum principle

In the next step, we characterize a solution to (2.3) by the PMP that provides necessary characteristics that a solution to (2.3) has to fulfill. For this purpose, we define the Hamiltonian function $H : \mathbb{R} \times \mathbb{R}^n \times K_U \times \mathbb{R}^n \to \mathbb{R}$ as follows

$$H(t, y, u, p) := h(y) + g(u) + p^T f(t, y, u) \tag{2.4}$$

where $(\cdot)^T$ is the transposed of a vector in $\mathbb{R}^n$. The adjoint equation for (2.3) is given by

$$-p'(t) = h_y(y(t)) + f_y(t, y(t), u(t))^T p(t) \tag{2.5}$$

with the terminal condition $p(T) = (D_y F(y(T)))^T$, $p : \mathbb{R} \to \mathbb{R}^n$, $t \mapsto p(t)$ and $p' := \frac{d}{dt} p$ the derivative with respect to t where (y, u) solves the initial value problem (2.1) with $u \in U_{ad}$. There exists a unique solution to the linear inhomogeneous differential equation (2.5) for the interval $[0, T]$ according to Theorem 55 in the Appendix. In addition from Theorem 55 we have that any solution p to (2.5) is absolutely continuous and thus the L^∞-norm $\|p\|_{L^\infty}$ is well defined as continuous functions take their maximum value [3, III Corollary 3.8] on a compact set.

The arguments that are used for the characterization of a solution to (2.3) are analogous to [81, Section 4] or [21, Section 3]. Crucial for our proof that a solution to (2.3) fulfills the PMP is the needle variation of a function $u^* \in U_{ad}$ that is given by

$$u_k(t) := \begin{cases} u & t \in S_k(t_0) \cap [0,T] \\ u^*(t) & t \in [0,T] \backslash S_k(t_0) \end{cases} \tag{2.6}$$

where $u \in K_U$, $S_k(t_0)$ an interval centered at $t_0 \in [0,T]$ whose measure, denoted by $|S_k(t_0)|$, goes to zero for k to infinity. Furthermore, the intermediate adjoint equation is given by

$$-\tilde{p}' = \tilde{h}(y_1, y_2) + \tilde{f}(y_1, y_2, u_1)^T \tilde{p} \tag{2.7}$$

with

$$\tilde{p}(T) = \tilde{F}(y_1, y_2) := \left(\int_0^1 D_y F(y_2(T) + \theta(y_1(T) - y_2(T)))\, d\theta \right)^T,$$

$$\tilde{f}(y_1, y_2, u_1) := \int_0^1 f_y(y_2 + \theta(y_1 - y_2), u_1)\, d\theta$$

where the integration is also componentwise and

$$\tilde{h}(y_1, y_2) := \int_0^1 h_y(y_2 + \theta(y_1 - y_2))\, d\theta.$$

According to Lemma 56 the functions $\tilde{F}$, $\tilde{f}$ and $\tilde{h}$ are well defined and there exists a unique absolutely continuous solution to (2.7) on the interval $[0,T]$.

Next, we prove a convergence property of the intermediate adjoint equation. In the following, the notation $var1 \leftarrow var2$ means that the variable $var1$ is replaced by $var2$ in the corresponding equation.

Lemma 2. *Let $u^* \in L^2(0,T)$ and y^* be the solution to the initial value problem (2.1) for $u \leftarrow u^*$ and p^* be the solution to (2.5) for $y \leftarrow y^*$ and $u \leftarrow u^*$. Let u_k be defined in (2.6), y_k be the solution to (2.1) for $u \leftarrow u_k$ and p_k the solution to (2.7) for $y_1 \leftarrow y_k$, $y_2 \leftarrow y^*$ and $u_1 \leftarrow u_k$. Then*

$$\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty} = 0$$

and

$$\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty} = 0$$

for almost all $t_0 \in (0,T)$.

Proof. As y^* and y_k solve (2.1), a subtraction provides

$$\begin{aligned} y_k - y^* &= \int_0^t f(y_k, u_k) - f(y^*, u^*)\, d\tilde{t} = \int_0^t f(y_k, u_k) - f(y^*, u_k) + f(y^*, u_k) - f(y^*, u^*)\, d\tilde{t} \\ &= \int_0^t \int_0^1 f_y(y^* + \theta(y_k - y^*), u_k)\, d\theta\, (y_k - y^*) + f(y^*, u_k) - f(y^*, u^*)\, d\tilde{t} \end{aligned} \tag{2.8}$$

where we use the fundamental theorem of calculus [4, VI 4.13] for $\theta \mapsto f(y^* + \theta(y_k - y^*), u_k)$ since $y \mapsto f_y(y,u)$ is continuous for every $u \in K_U$ due to Assumption A.1), see [78, Chapter 5 Theorem 6] and that the composite function of continuous functions is continuous [3, III Theorem 1.8]. From (2.8), we obtain that

$$(y_k)_i - (y^*)_i = \int_0^t \sum_{l=1}^n \int_0^1 \frac{\partial}{\partial y_l} f_i(y, u_k)\,|_{y=y^*+\theta(y_k-y^*)} d\theta\, ((y_k)_l - (y^*)_l) + f_i(y^*, u_k) - f_i(y^*, u^*)\, dt$$

and thus

$$|\left(y_k\right)_i-\left(y^*\right)_i| \leq \int_0^t \sum_{l=1}^n \int_0^1 |\frac{\partial}{\partial y_l} f_i\left(y, u_k\right)|_{y=y^*+\theta\left(y_k-y^*\right)}| d\theta\left(|\left(y_k\right)_l-\left(y^*\right)_l|\right)+|f_i\left(y^*, u_k\right)-f_i\left(y^*, u^*\right)| dt.$$

Consequently from Assumption A.5) and Assumption A.6) we obtain

$$\sum_{i=1}^n|\left(y_k\right)_i-\left(y^*\right)_i| \leq \int_0^t \sum_{i=1}^n\left(c \sum_{l=1}^n|\left(y_k\right)_l-\left(y^*\right)_l|+L \sum_{l=1}^n|\left(u_k\right)_l-\left(u^*\right)_l|\right) dt$$

which gives the following

$$\sum_{i=1}^n|\left(y_k\right)_i-\left(y^*\right)_i| \leq \int_0^t n c \sum_{i=1}^n|\left(y_k\right)_i-\left(y^*\right)_i| dt+n L\left\|u_k-u^*\right\|_{L^1} .$$

By Lemma 57, we have that

$$\sum_{i=1}^n|\left(y_k\right)_i-\left(y^*\right)_i| \leq\left(n L+n^2 L c T \mathrm{e}^{n c T}\right)\left\|u_k-u^*\right\|_{L^1} \tag{2.9}$$

which holds for all $t \in[0, T]$. Next, by the definition of the needle variation (2.6), we have that

$$\left\|u_k-u^*\right\|_{L^1}=\sum_{i=1}^n \int_0^T|\left(u_k\right)_i-\left(u^*\right)_i| dt=\sum_{i=1}^n \int_{S_k\left(t_0\right) \cap[0, T]}|u_i-\left(u^*\right)_i| dt. \tag{2.10}$$

From (2.10) and [15, Theorem 5.6.2] we obtain

$$\lim_{k \rightarrow \infty}\left\|u_k-u^*\right\|_{L^1}=0 \tag{2.11}$$

for almost all $t_0 \in(0, T)$. From (2.9) and (2.11), we have for all $i \in\{1, \ldots, n\}$ that

$$\lim_{k \rightarrow \infty}|\left(y_k\right)_i(t)-\left(y^*\right)_i(t)|=0 \tag{2.12}$$

for each $t \in[0, T]$ for almost all $t_0 \in(0, T)$. This implies that $\lim_{k \rightarrow \infty}\left\|y_k-y^*\right\|_{L^{\infty}}=0$ for almost all $t_0 \in(0, T)$.

In the next step, we consider the difference of the solution p_k to (2.7) and the solution p^* to (2.5), which is transformed into an initial value problem by $\tau:=T-t$, as follows

$$\begin{aligned}
& p_k-p^* \\
& =\int_0^t \tilde{h}\left(y_k, y^*\right)+\tilde{f}\left(y_k, y^*, u_k\right)^T p_k-h_y\left(y^*\right)-f_y\left(y^*, u^*\right)^T p^* d \tilde{t} \\
& =\int_0^t \int_0^1 h_y\left(y^*+\theta\left(y_k-y^*\right)\right)-h_y\left(y^*\right) d \theta+\int_0^1 f_y\left(y^*+\theta\left(y_k-y^*\right), u_k\right)^T p_k-f_y\left(y^*, u^*\right)^T p^* d \theta d \tilde{t} \\
& =\int_0^t \int_0^1 h_y\left(y^*+\theta\left(y_k-y^*\right)\right)-h_y\left(y^*\right) d \theta d \tilde{t}+\int_0^t \int_0^1 f_y\left(y^*+\theta\left(y_k-y^*\right), u_k\right)^T\left(p_k-p^*\right) d \theta d \tilde{t} \\
& +\int_0^t \int_0^1\left(f_y\left(y^*+\theta\left(y_k-y^*\right), u_k\right)^T-f_y\left(y^*+\theta\left(y_k-y^*\right), u^*\right)^T\right) p^* d \theta d \tilde{t} \\
& +\int_0^t \int_0^1\left(f_y\left(y^*+\theta\left(y_k-y^*\right), u^*\right)^T-f_y\left(y^*, u^*\right)^T\right) p^* d \theta d \tilde{t} .
\end{aligned} \tag{2.13}$$

For each component $i \in \{1,...,n\}$ of (2.13), we obtain

$$\begin{aligned}
&(p_k)_i - (p^*)_i \\
&= \int_0^t \int_0^1 \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*} d\theta d\tilde{t} \\
&+ \int_0^t \int_0^1 \sum_{l=1}^{n} \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*+\theta(y_k-y^*)} \left(\left(p_k\right)_l - \left(p^*\right)_l\right) d\theta d\tilde{t} \\
&+ \int_0^t \int_0^1 \sum_{l=1}^{n} \left(\frac{\partial}{\partial y_i} f_l\left(y, u_k\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y^*+\theta(y_k-y^*)}\right) \left(p^*\right)_l d\theta d\tilde{t} \\
&+ \int_0^t \int_0^1 \sum_{l=1}^{n} \left(\frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*}\right) \left(p^*\right)_l d\theta d\tilde{t}
\end{aligned}$$

and consequently

$$\begin{aligned}
&|\left(p_k\right)_i - \left(p^*\right)_i| \\
&\leq \int_0^t \int_0^1 |\frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*}| d\theta d\tilde{t} \\
&+ \int_0^t \int_0^1 \sum_{l=1}^{n} |\frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*+\theta(y_k-y^*)}| \left(|\left(p_k\right)_l - \left(p^*\right)_l|\right) d\theta d\tilde{t} \\
&+ \int_0^t \int_0^1 \sum_{l=1}^{n} |\left(\frac{\partial}{\partial y_i} f_l\left(y, u_k\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y^*+\theta(y_k-y^*)}\right)| \left(|\left(p^*\right)_l|\right) d\theta d\tilde{t} \\
&+ \int_0^t \int_0^1 \sum_{l=1}^{n} \left(|\frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*}|\right) \left(|\left(p^*\right)_l|\right) d\theta d\tilde{t}.
\end{aligned} \tag{2.14}$$

Now we prepare the application of the dominated convergence theorem [36, Theorem 2.4.5]. The functions $(\theta, t) \mapsto \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*(t)+\theta(y_k(t)-y^*(t))}$, $(\theta, t) \mapsto \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*(t)}$, $(\theta, t) \mapsto \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*(t)+\theta(y_k(t)-y^*(t))}$ and $(\theta, t) \mapsto \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*(t)}$ are measurable as discussed in the proofs of Theorem 55, Lemma 56 and Lemma 52. According to Assumption A.1) the functions $y \mapsto \frac{\partial}{\partial y_i} h\left(y\right)$, $i = 1, ..., n$, are continuous and thus bounded on a compact set [3, III Corollary 3.8]. Also as y^* is continuous there exists a compact set $B^* \subseteq \mathbb{R}^n$ such that the image $y^*\left([0,T]\right) \subseteq B^*$, see [3, III Theorem 3.6] and [3, Theorem 3.2]. Because of the uniform pointwise convergence of y_k for almost all $t_0 \in (0,T)$, see (2.9) and (2.12), there exists a compact ball $B \subseteq \mathbb{R}^n$ with $B^* \subseteq B$ such that the image $\left(y^* + \theta\left(y_k - y^*\right)\right)\left([0,T]\right) \subseteq B$. Therefore the functions

$$(\theta, t) \mapsto |\frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*(t)+\theta(y_k(t)-y^*(t))} - \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*(t)}|$$

for all $i \in \{1, ..., n\}$ are bounded. The functions

$$(\theta, t) \mapsto |\frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*(t)+\theta(y_k(t)-y^*(t))} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*(t)}|,$$

$i, l \in \{1, ..., n\}$, are bounded by Assumption A.6). For any fixed $(\theta, t) \in [0,1] \times [0,T]$, we have the pointwise limit

$$\lim_{k\to\infty} \left(|\frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*(t)+\theta(y_k(t)-y^*(t))} - \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*(t)}|\right) = 0$$

and

$$\lim_{k\to\infty} \left(|\frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*(t)+\theta(y_k(t)-y^*(t))} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*(t)}|\right) = 0$$

due to the continuity of $y \mapsto \frac{\partial}{\partial y_i} h\left(y\right)$ and $y \mapsto \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)$ for all $i, l \in \{1, ..., n\}$ and almost all $t_0 \in (0, T)$, see [3, III Theorem 1.4]. Furthermore, we have that

$$|\frac{\partial}{\partial y_i} f_l\left(y, u_k\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*+\theta(y_k-y^*)}| \leq L \sum_{l=1}^{n} |\left(u_k\right)_l - \left(u^*\right)_l|$$

by Assumption A.5). We define the functions

$$\xi_i^k := \int_0^T \int_0^1 |\frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} h\left(y\right)|_{y=y^*}| d\theta dt$$

and

$$\psi_i^k := \int_0^T \int_0^1 \sum_{l=1}^{n} \left(|\frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*+\theta(y_k-y^*)} - \frac{\partial}{\partial y_i} f_l\left(y, u^*\right)|_{y=y^*}| \right) \left(|\left(p^*\right)_l| \right) d\theta dt.$$

Then we have by summing (2.14) over i that

$$\begin{aligned} &\sum_{i=1}^{n} |\left(p_k\right)_i - \left(p^*\right)_i| \\ &\leq \sum_{i=1}^{n} \left(\xi_i^k + \psi_i^k\right) + nL \left(\max_{t \in [0,T]} \sum_{l=1}^{n} |\left(p^*\right)_l\left(t\right)| \right) \|u_k - u^*\|_{L^1} + cn \int_0^t \sum_{i=1}^{n} \left(|\left(p_k\right)_l - \left(p^*\right)_l| \right) d\tilde{t} \end{aligned}$$

and consequently by Lemma 57, we have that

$$\sum_{i=1}^{n} |\left(p_k\right)_i - \left(p^*\right)_i| \leq \left(1 + cnT\mathrm{e}^{cnT}\right) \Phi^k$$

for almost all $t_0 \in (0, T)$ with the definition

$$\Phi^k := \sum_{i=1}^{n} \left(\xi_i^k + \psi_i^k\right) + nL \left(\max_{t \in [0,T]} \sum_{l=1}^{n} |\left(p^*\right)_l\left(t\right)| \right) \|u_k - u^*\|_{L^1}$$

where we remark that the fixed function p^* is a continuous function and thus bounded on $[0, T]$, see [3, III Corollary 3.8]. By the dominated convergence theorem [36, Theorem 2.4.5], the calculation rules for the limit [3, II Theorem 2.2] and (2.11), we have that

$$\lim_{k \to \infty} |\left(p_k\right)_i\left(t\right) - \left(p^*\right)_i\left(t\right)| = 0$$

for all $i \in \{1, ..., n\}$ and all $t \in [0, T]$ and almost all $t_0 \in (0, T)$. This implies that $\lim_{k \to \infty} \|p_k - p^*\|_{L^\infty} = 0$ for almost all $t_0 \in (0, T)$. □

Now, we can go on with the proof that a solution to (2.3) fulfills the PMP. For this purpose, we need the following lemma.

Lemma 3. *Let (y_1, u_1) and (y_2, u_2) solve the initial value problem (2.1). Then, it holds that*

$$J\left(y_1, u_1\right) - J\left(y_2, u_2\right) = \int_0^T H\left(t, y_2, u_1, \tilde{p}\right) - H\left(t, y_2, u_2, \tilde{p}\right) dt$$

where $\tilde{p}$ solves (2.7).

Proof. Because of the continuity of f_y and h_y in the state argument, we apply the fundamental theorem of calculus [4, VI 4.13] and thus we obtain pointwise

$$\begin{aligned} f\left(y_1,u_1\right)-f\left(y_2,u_1\right)&=f\left(y_2+\theta\left(y_1-y_2\right),u_1\right)|_{\theta=1}-f\left(y_2+\theta\left(y_1-y_2\right),u_1\right)|_{\theta=0}\\ &=\int_0^1\frac{d}{d\theta}f\left(y_2+\theta\left(y_1-y_2\right),u_1\right)d\theta=\int_0^1 f_y\left(y_2+\theta\left(y_1-y_2\right),u_1\right)\left(y_1-y_2\right)d\theta\\ &=\tilde{f}\left(y_1,y_2,u_1\right)\left(y_1-y_2\right)\end{aligned}$$

with the chain rule [4, VII Theorem 3.3]. Analogously, we have

$$h\left(y_1\right)-h\left(y_2\right)=\tilde{h}\left(y_1,y_2\right)^T\left(y_1-y_2\right)$$

and

$$F\left(y_1\left(T\right)\right)-F\left(y_2\left(T\right)\right)=\tilde{F}\left(y_1,y_2\right)^T\left(y_1\left(T\right)-y_2\left(T\right)\right).$$

Next, we obtain

$$\begin{aligned} &J\left(y_1,u_1\right)-J\left(y_2,u_2\right)=\int_0^T h\left(y_1\right)+g\left(u_1\right)-h\left(y_2\right)-g\left(u_2\right)dt+F\left(y_1\left(T\right)\right)-F\left(y_2\left(T\right)\right)\\ &=\int_0^T h\left(y_2\right)+g\left(u_1\right)-h\left(y_2\right)+h\left(y_1\right)-h\left(y_2\right)-g\left(u_2\right)+\tilde{p}^Tf\left(y_2,u_1\right)-\tilde{p}^Tf\left(y_2,u_1\right)dt\\ &+\int_0^T\tilde{p}^Tf\left(y_2,u_2\right)-\tilde{p}^Tf\left(y_2,u_2\right)dt+F\left(y_1\left(T\right)\right)-F\left(y_2\left(T\right)\right)\\ &=\int_0^T H\left(y_2,u_1,\tilde{p}\right)-H\left(y_2,u_2,\tilde{p}\right)+h\left(y_1\right)-h\left(y_2\right)dt\\ &+\int_0^T\tilde{p}^T\left(f\left(y_2,u_2\right)-f\left(y_1,u_1\right)+f\left(y_1,u_1\right)-f\left(y_2,u_1\right)\right)dt+F\left(y_1\left(T\right)\right)-F\left(y_2\left(T\right)\right)\\ &=\int_0^T H\left(y_2,u_1,\tilde{p}\right)-H\left(y_2,u_2,\tilde{p}\right)+\left(y_1-y_2\right)^T\left(\tilde{h}\left(y_1,y_2\right)+\tilde{f}\left(y_1,y_2,u_1\right)^T\tilde{p}\right)dt\\ &+\int_0^T\tilde{p}^T\left(f\left(y_2,u_2\right)-f\left(y_1,u_1\right)\right)dt+F\left(y_1\left(T\right)\right)-F\left(y_2\left(T\right)\right)\\ &=\int_0^T H\left(y_2,u_1,\tilde{p}\right)-H\left(y_2,u_2,\tilde{p}\right)-\left(y_1-y_2\right)^T\tilde{p}'-\tilde{p}^T\left(y_1'-y_2'\right)dt+\tilde{F}\left(y_1,y_2\right)\left(y_1\left(T\right)-y_2\left(T\right)\right)\\ &=\int_0^T H\left(y_2,u_1,\tilde{p}\right)-H\left(y_2,u_1,\tilde{p}\right)dt\\ &+y_2\left(T\right)^T\tilde{F}\left(y_1,y_2\right)-y_1\left(T\right)^T\tilde{F}\left(y_1,y_2\right)+\tilde{F}\left(y_1,y_2\right)^T\left(y_1\left(T\right)-y_2\left(T\right)\right)\end{aligned}$$

where we use the partial integration [36, Corollary 6.3.9] in the third to last line. □

The next lemma is given as follows and relates the differences of the cost functionals to the difference of Hamiltonian functions.

Lemma 4. *Let $u^*\in U_{ad}$ and $u\in K_U$. Furthermore let u_k be defined as in (2.6) for all $k\in\mathbb{N}$ and y_k be the solution to (2.1) for $u\leftarrow u_k$. Then, the following holds*

$$\lim_{k\to\infty}\frac{1}{|S_k\left(t_0\right)|}\left(J\left(y_k,u_k\right)-J\left(y^*,u^*\right)\right)=H\left(t_0,y^*,u,p^*\right)-H\left(t_0,y^*,u^*,p^*\right)$$

for almost all $t_0\in(0,T)$ where y^* is the solution to (2.1) for $u\leftarrow u^*$ and p^* is the corresponding solution to (2.5) for $y\leftarrow y^*$ and $u\leftarrow u^*$.

Proof. With Lemma 3, we have

$$\begin{aligned} & J\left(y_k, u_k\right) - J\left(y^*, u^*\right) \\ & = \int_0^T H\left(y^*, u_k, p_k\right) - H\left(y^*, u^*, p_k\right) dt = \int_{S_k(t_0)\cap[0,T]} H\left(y^*, u, p_k\right) - H\left(y^*, u^*, p_k\right) dt \\ & = \int_{S_k(t_0)\cap[0,T]} H\left(y^*, u, p^*\right) - H\left(y^*, u^*, p^*\right) + \left(p_k - p^*\right)^T f\left(y^*, u\right) + \left(p^* - p_k\right)^T \left(f\left(y^*, u^*\right)\right) dt \end{aligned} \tag{2.15}$$

where p_k is the solution to (2.7) with $u_1 \leftarrow u_k$, $y_1 \leftarrow y_k$ and $y_2 \leftarrow y^*$. We multiply both sides of (2.15) with $\frac{1}{|S_k(t_0)|}$ and apply the limit for k on both sides. Then we obtain

$$\lim_{k\to\infty} \frac{1}{|S_k\left(t_0\right)|} \left(J\left(y_k, u_k\right) - J\left(y^*, u^*\right)\right) = H\left(y^*, u, p^*\right) - H\left(y^*, u^*, p^*\right)$$

because with Lemma 2 and Assumption A.3), for k sufficiently large and thus $S_k\left(t_0\right)$ is sufficiently small, we have that

$$\begin{aligned} & \lim_{k\to\infty} \frac{1}{|S_k\left(t_0\right)|} \left| \int_{S_k(t_0)\cap[0,T]} \left(p_k - p^*\right)^T f\left(y^*, u\right) dt \right| \\ & \leq \lim_{k\to\infty} \left(\|p - p^*\|_{L^\infty} \frac{1}{|S_k\left(t_0\right)|} \int_{S_k(t_0)\cap[0,T]} \sum_{i=1}^n |f_i\left(y^*, u\right)| dt \right) = 0 \end{aligned}$$

and analogously

$$\lim_{k\to\infty} \frac{1}{|S_k\left(t_0\right)|} \left| \int_{S_k(t_0)\cap[0,T]} \left(p^* - p_k\right)^T f\left(y^*, u^*\right) dt \right| = 0$$

for almost all $t_0 \in (0,T)$ considering the limit rules [3, II Remark 2.1 (a)], [3, II Theorem 2.4], [3, Theorem 1.10] and the mean value theorem [15, Theorem 5.6.2]. We remark that the union of countably many null sets is a null set, see [5, IX Remark 2.5 (b)]. □

Now, we have the following theorem that characterizes a solution to (2.3).

Theorem 5. *Let $(\bar{y}, \bar{u})$ be a solution to (2.3). Then it holds that*

$$H\left(t, \bar{y}, \bar{u}, \bar{p}\right) = \min_{w\in K_U} H\left(t, \bar{y}, w, \bar{p}\right) \tag{2.16}$$

for almost all $t \in (0,T)$ where $\bar{p}$ is a solution to (2.5) with $y \leftarrow \bar{y}$ and $u \leftarrow \bar{u}$.

Proof. As we have that $J\left(\tilde{y}, \tilde{u}\right) \geq J\left(\bar{y}, \bar{u}\right)$ for all $(\tilde{y}, \tilde{u})$ solving (2.1) with $\tilde{u} \in U_{ad}$, we especially have that $J\left(y_k, u_k\right) \geq J\left(\bar{y}, \bar{u}\right)$ for any solution (y_k, u_k) to (2.1) as $u_k \in U_{ad}$. This can be seen as follows. The sum and the product of measurable functions is measurable, see [36, Proposition 2.1.7]. The needle variation (2.6) can be written as $u_k = u^* \chi_{[0,T]\setminus S_k(t_0)} + u \chi_{S_k(t_0)\cap[0,T]}$. Since the characteristic function χ_A is measurable if and only if A is measurable, see [36, Example 2.1.2] the needle variation is Lebesgue measurable, as $[0,T] \setminus S_k\left(t_0\right)$ and $S_k\left(t_0\right) \cap [0,T]$ are Lebesgue measurable, see [36, Theorem 1.3.6]. Furthermore it is pointwise $u_k \in K_U$ and thus we have by

$$\begin{aligned} \sum_{j=1}^m \int_0^T \left(\left(u_k\right)_j(t)\right)^2 dt &= \sum_{j=1}^m \int_{[0,T]\setminus S_k(t_0)} \left(\left(u^*\right)_j(t)\right)^2 dt + \int_{S_k(t_0)\cap[0,T]} u_j^2 dt \\ &\leq \sum_{j=1}^m \left(\int_0^T \left(\left(u^*\right)_j(t)\right)^2 dt + u_j^2 |S_k\left(t_0\right)| \right) \end{aligned}$$

the L^2-integrability since $u^* \in U_{ad}$ and u_j are real numbers for all $j \in \{1, ..., m\}$. Then we have from $J(y_k, u_k) - J(\bar{y}, \bar{u}) \geq 0$ that $\frac{1}{|S_k(t_0)|}(J(y_k, u_k) - J(\bar{y}, \bar{u})) \geq 0$ and consequently

$$0 \leq \lim_{k\to\infty} \frac{1}{|S_k(t_0)|}(J(y_k, u_k) - J(\bar{y}, \bar{u})) = H(t_0, \bar{y}, u, \bar{p}) - H(t_0, \bar{y}, \bar{u}, \bar{p})$$

see [3, II Theorem 2.7] and Lemma 4 for almost all $t_0 \in (0, T)$. From this we conclude, by renaming t_0 into t, that $H(t, \bar{y}, \bar{u}, \bar{p}) \leq H(t, \bar{y}, u, \bar{p})$ for almost all $t \in (0, T)$ and all $u \in K_U$ which is equivalent to

$$H(t, \bar{y}, \bar{u}, \bar{p}) = \min_{w\in K_U} H(t, \bar{y}, w, \bar{p}).$$

□

2.3 Convergence analysis of the SQH scheme

In this section, we discuss the sequential quadratic Hamiltonian (SQH) scheme for optimal control problems governed by ODEs in the framework of Section 2.1. This section is based on [22, Section 4] and [22, Section 3]. As already discussed in the introduction, the SQH scheme represents an advancement of the schemes proposed in [62, 63] and [85, 88] in the context of ODE control problems. This procedure is characterized by two important features. First, a quadratic pointwise penalization of the control's updates. Second, the computation of the state variable after the control's update at all points has been completed.

In the SQH method, the Hamiltonian (2.4) is augmented with the term $\epsilon(u(t) - v(t))^2$ where

$$(u(t) - v(t))^2 := \sum_{j=1}^{m} (u_j(t) - v_j(t))^2.$$

Thus we define the following augmented Hamiltonian

$$K_\epsilon(t, y, u, v, p) := H(t, y, u, p) + \epsilon(u(t) - v(t))^2 \tag{2.17}$$

where $K_\epsilon : \mathbb{R} \times \mathbb{R}^n \times K_U \times K_U \times \mathbb{R}^n \to \mathbb{R}$, $\epsilon > 0$. We use the notation

$$K_\epsilon(t, y, u, v, p) := K_\epsilon(t, y(t), u(t), v(t), p(t))$$

whenever an argument of K_ϵ is a function instead of a number.

Specifically, the quadratic term $\epsilon(u(t) - v(t))^2$ aims at penalizing local control updates that differ too much from the current control value. This in turn prevents the corresponding state y to take values at t that differ too much from the current value, see Lemma 4. Therefore we can reasonably pursue to update the state variable after the control has been updated at all grid points.

The basic idea in developing the SQH scheme is to minimize K_ϵ over K_U at each point $t \in [0, T]$ in some given order, for instance lexicographically. For this purpose, there are several ways to calculate the elements of K_U which minimize K_ϵ at the corresponding grid points. First of all, one can discretize K_U and choose the corresponding minimizing value of K_ϵ by array search in the resulting discretized set and assign this value to the control. Second, one can apply a secant method in the set K_U to find a minimum of the augmented Hamiltonian up to a given tolerance. Third, one can use an analytical formula for a minimum in K_U, if available. From these comments, we notice that the first approach can also be used if the set K_U is a discrete set, and in this case an application can be mixed-integer optimal control problems without the need for relaxation as in [52] for instance.

The main difference of our scheme with respect to the algorithm in [85, 88] and similar to [62], is that, in the minimization process, we use $K_\epsilon(t, y^k, u, u^k, p^k)$ instead of $K_\epsilon(t, y^{k+1}, u, u^k, p^k)$. In fact in [17, 85, 88] an update of the state y is computed after each local pointwise update of the control, whereas in the SQH scheme the state y^k of the previous iteration is used while minimizing K_ϵ. This approach provides a great computational advantage since the update of the state variable is a very costly procedure in large-size problems. Furthermore, the implementation of the minimization of K_ϵ becomes much easier since it involves only the control function.

Notice that the weight ϵ plays an essential role to attain convergence of the proposed scheme while penalizing large control updates. Our SQH scheme is given in detail in the following algorithm. The strategy for the adaptive changing of ϵ is based on that given in [85]. The scheme is implemented by the following algorithm.

Algorithm 2.1 (SQH method)

1. Choose $\epsilon > 0$, $\kappa > 0$, $\sigma > 1$, $\zeta \in (0,1)$, $\eta \in (0,\infty)$, $u^0 \in U_{ad}$, compute y^0 by (2.1) for $u \leftarrow u^0$ and p^0 by (2.5) for $y \leftarrow y^0$ and $u \leftarrow u^0$, set $k \leftarrow 0$

2. Set
$$u\left(t\right) = \underset{w \in K_U}{\arg\min}\, K_\epsilon\left(t, y^k, w, u^k, p^k\right)$$
for all $t \in [0,T]$

3. Calculate y by (2.1) for u and $\tau := \sum_{j=1}^m \|u_j - \left(u^k\right)_j\|_{L^2(0,T)}^2$

4. If $J\left(y,u\right) - J\left(y^k,u^k\right) > -\eta\tau$: Choose $\epsilon \leftarrow \sigma\epsilon$
Else:
Choose $\epsilon \leftarrow \zeta\epsilon$, $y^{k+1} \leftarrow y$, $u^{k+1} \leftarrow u$, calculate p^{k+1} by (2.5) for $y \leftarrow y^{k+1}$ and $u \leftarrow u^{k+1}$, set $k \leftarrow k+1$

5. If $\tau < \kappa$: STOP and return u^k
Else go to 2.

We remark that Step 2 in Algorithm 2.1 can also be formulated as: Choose $u \in K_U$ such that

$$K_\epsilon\left(t, y^k, u, u^k, p^k\right) \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$$

for all $w \in K_U$ and all $t \in [0,T]$.

In the following we explain the different steps of Algorithm 2.1. After choosing the problem's parameters and an initial guess for the control, we determine u such that the augmented Hamiltonian is minimized for a given state, adjoint, current control and ϵ. If the resulting control u and the corresponding y do not minimize the cost functional more than $-\eta\tau$ with respect to the former values y^k and u^k, we increase ϵ and perform the minimization of the resulting K_ϵ again. Else, we accept the new control function as well as the corresponding state, calculate the adjoint and decrease ϵ such that greater variations of the control value become more likely and thus accelerate the determination of an optimal control. If the convergence criterion $\tau < \kappa$ is not fulfilled, then in the SQH scheme the minimization procedure is repeated. If the convergence criterion is fulfilled, then the algorithm stops and returns the last calculated control u^k.

Remark 6. The concept of augmenting the Hamiltonian, where the state variable from the previous iteration is used for the calculation of an update for the control function, is valuable to solve time discrete optimal control problems that are used for the training of neuronal networks as in [66, Section 3 to 4 and C] for instance. In this reference, they report the need of an update resulting from minimizing the Hamiltonian that differs not too drastically from the previous control variable.

Next, we prove that for given t, y, v, p and ϵ there exists a $u \in K_U$ that minimizes $K_\epsilon\left(t, y, u, v, p\right)$. Thus, Step 2 of Algorithm 2.1 is well posed. Later, we prove that there exists an ϵ sufficiently large such that the condition for sufficient decrease of the cost functional's value is satisfied and $\|u^k - u^{k-1}\|_{L^2}^2$

decreases such that the convergence criterion is eventually satisfied. Hence, Step 4 in Algorithm 2.1 is well defined.

Concerning Step 2, we have the following.

Lemma 7. *The function $K_\epsilon : \mathbb{R}^m \to \mathbb{R}$, $w \mapsto K_\epsilon(t,y,w,v,p)$ attains a minimum for any $(t,y,v,p) \in [0,T] \times I \times K_U \times \mathbb{R}^n$ and any $\epsilon \in \mathbb{R}$.*

Proof. We have that

$$w \mapsto K_\epsilon(t,y,w,v,p) = h(y) + g(w) + p^T f(t,y,w) + \epsilon (w-v)^2$$

is bounded from below. This can be seen as follows. We just consider the terms depending on w since the others are constants with respect to the minimization in w. Since $w \mapsto f_i(t,y,w)$, $i \in \{1,...,n\}$, is continuous for any fixed t and y due to Assumption A.4) and K_U is compact, we have that the set $\{f_i(t,y,w) \in \mathbb{R} |\, w \in K_U\}$, $i \in \{1,...,n\}$, is compact, see [3, III Theorem 3.6], and thus bounded, see [3, III Theorem 3.2]. Consequently for any fixed $p \in \mathbb{R}^n$, the function

$$w \mapsto p^T f(t,y,w) = \sum_{i=1}^{n} p_i f_i(t,y,w)$$

is bounded from below by a constant. The function $w \mapsto \epsilon(w-v)^2 = \epsilon \sum_{j=1}^m (w_j - v_j)^2$ is continuous and thus it is bounded with an analogous reasoning as above for any $\epsilon \in \mathbb{R}$. We remark that for the special case that we just consider $\epsilon \geq 0$ the term $\epsilon(w-v)^2$ is bounded from below by zero. The function $w \mapsto g(w)$ is bounded from below by our requirement.

Thus there is a lower bound for $w \mapsto K_\epsilon(t,y,w,v,p)$ and a biggest lower bound

$$d := \inf_{w \in K_U} K_\epsilon(t,y,w,v,p) := \inf \{K_\epsilon(t,y,w,v,p) \in \mathbb{R} |\, w \in K_U\}$$

exists since any subset of $\mathbb{R}$ bounded from below has an infimum, see [3, I Theorem 10.4, Theorem 10.1]. Consequently for any given number $\tilde{\epsilon}_l > 0$, monotonically decreasing for increasing $l \in \mathbb{N}$, there is a u_l with

$$d \leq K_\epsilon(t,y,u_l,v,p) \leq d + \tilde{\epsilon}_l. \tag{2.18}$$

If this was not the case, that means if there was an $\tilde{l}$ such that $d + \tilde{\epsilon}_{\tilde{l}} < K_\epsilon(t,y,w,v,p)$ for all $w \in K_U$, then it would contradict d being the biggest lower bound which would be at least $d + \tilde{\epsilon}_{\tilde{l}}$ in this case.

By applying the limit on both sides of (2.18), we have for the minimizing sequence $(u_l)_{l \in \mathbb{N}} \subseteq K_U$ that

$$\inf_{w \in K_U} K_\epsilon(t,y,w,v,p) = \lim_{l \to \infty} K_\epsilon(t,y,u_l,v,p),$$

see [3, Theorem 2.9]. As K_U is compact, there is an index set $K \subseteq \mathbb{N}$ such that for the corresponding subsequence $(u_k)_{k \in K}$ it holds $\lim_{k \to \infty} u_k = u$ with $u \in K_U$. Furthermore, we have with [3, II Theorem 5.7] and [43, Theorem 3.127] the following

$$\begin{aligned} &\inf_{w \in K_U} K_\epsilon(t,y,w,v,p) = \lim_{k \to \infty} K_\epsilon(t,y,u_k,v,p) = \liminf_{k \to \infty} K_\epsilon(t,y,u_k,v,p) \\ &= \liminf_{k \to \infty} \left(h(y) + g(u_k) + p^T f(t,y,u_k) + \epsilon(u_k - v)^2 \right) \\ &\geq h(y) + g(u) + p^T f(t,y,u) + \epsilon(u-v)^2 = K_\epsilon(t,y,u,v,p) \end{aligned} \tag{2.19}$$

because of the lower semi-continuity of g and the continuity of f, see Assumption A.4) and [3, III Theorem 1.4]. □

The issue if u, obtained in Step 2 of Algorithm 2.1, is Lebesgue measurable is discussed in Section 5.1 where we see that for all cases considered in this thesis this is certainly the case.

For the following analysis, we need some auxiliary results which are given and proved in the following. In the next lemma, we show that the adjoint variable is bounded by a constant for all (y, u) solving (2.1) with $u \in U_{ad}$. Furthermore it is made an extensive use of Gronwall's inequality that is denoted in Lemma 57 in the Appendix.

Lemma 8. *For the solution p to (2.5) there exists a constant $C_1 > 0$ such that $\|p\|_{L^\infty} < C_1$ for all (y, u) solving (2.1) with $u \in U_{ad}$.*

Proof. By the transformation $\tau := T - t$, $\hat{p}(\tau) := p(T-\tau)$, $\hat{y}(\tau) := y(T-\tau)$ and

$$\hat{f}_y(\hat{y}, \hat{u}) := f_y\left(y(T-\tau), u(T-\tau)\right)$$

we obtain an initial value problem

$$\hat{p}' = h_y(\hat{y}) + \hat{f}_y(\hat{y}, \hat{u})\, \hat{p},$$

$\hat{p}(0) = \left(D_y F(\hat{y}(0))\right)^T$ with $\hat{p}'(\tau) := \frac{\partial}{\partial \tau}\hat{p}(\tau) = \frac{\partial}{\partial \tau} p(T-\tau) = -p'$ from (2.5). Its solution fulfills

$$\hat{p}(\tau) = \left(D_y F(\hat{y}(0))\right)^T + \int_0^\tau h_y(\hat{y}(\tilde{\tau})) + \hat{f}_y(\hat{y}(\tilde{\tau}), \hat{u}(\tilde{\tau}))^T \hat{p}(\tilde{\tau})\, d\tilde{\tau},$$

see the proof of Theorem 55. Consequently for each component $i \in \{1, ..., n\}$, we obtain

$$\hat{p}_i(\tau) = \frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}(0)} + \int_0^\tau \frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}(\tilde{\tau})} + \sum_{l=1}^n \frac{\partial}{\partial y_i} \hat{f}_l(y, \hat{u}(\tilde{\tau}))\,|_{y=\hat{y}(\tau)} \hat{p}_l(\tilde{\tau})\, d\tilde{\tau}$$

and thus by taking the absolute value we have the following

$$|\hat{p}_i(\tau)| \leq |\frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}(0)}| + \int_0^\tau |\frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}(\tilde{\tau})}| + \sum_{l=1}^n |\frac{\partial}{\partial y_i} \hat{f}_l(y, \hat{u}(\tilde{\tau}))\,|_{y=\hat{y}(\tau)}||\hat{p}_l(\tilde{\tau})|\, d\tilde{\tau}. \tag{2.20}$$

Adding up both sides of (2.20) provides

$$\sum_{i=1}^n |\hat{p}_i(\tau)| \leq \sum_{i=1}^n \left(|\frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}(0)}| + \int_0^\tau |\frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}(\tilde{\tau})}| + \sum_{l=1}^n |\frac{\partial}{\partial y_i} \hat{f}_l(y, \hat{u}(\tilde{\tau}))\,|_{y=\hat{y}(\tau)}||\hat{p}_l(\tilde{\tau})|\, d\tilde{\tau} \right).$$

Due to Assumption A.6), we have that

$$\sum_{i=1}^n |\hat{p}_i(\tau)| \leq \tilde{C} + nc \int_0^\tau |\hat{p}_l(\tilde{\tau})|\, d\tilde{\tau}$$

where $\tilde{C} := nc + cnT$. With Gronwall's inequality, see Lemma 57 in the Appendix, we obtain that

$$\sum_{i=1}^n |\hat{p}_i(\tau)| \leq \tilde{C} + \tilde{C} nc \int_0^T \exp(ncT)\, d\tilde{\tau} = \tilde{C}\left(1 + ncT \exp(ncT)\right)$$

where the right hand-side is independent of t and thus $\hat{p}$ is bounded since each component is bounded by $C_1 := \tilde{C}\left(1 + ncT \exp(ncT)\right)$. By backsubstitution we have that also p is bounded. □

In the next lemma, we have a boundedness result for two different solutions to the initial value problem (2.1).

Lemma 9. *There exists a constant $C_2 > 0$ such that for any two solutions (y_1, u_1) and (y_2, u_2) to the initial value problem (2.1) with $y_1(0) = y_2(0) = y_0$ with $u_1, u_2 \in U_{ad}$, it holds*

$$|\left(y_1\right)_i(t) - \left(y_2\right)_i(t)| \leq C_2 \|u_1 - u_2\|_{L^1}$$

for all $t \in [0, T]$ and all $i \in \{1, ..., n\}$.

Proof. We have that

$$y_1(t) - y_2(t) = \int_0^t f\left(y_1\left(\tilde{t}\right), u_1\left(\tilde{t}\right)\right) - f\left(y_2\left(\tilde{t}\right), u_2\left(\tilde{t}\right)\right) d\tilde{t}$$

and thus for all $i \in \{1, ..., n\}$ it holds by taking the absolute value that

$$\begin{aligned}
&|\left(y_1\right)_i(t) - \left(y_2\right)_i(t)| \\
&\leq \int_0^t |f_i\left(y_1(t), u_1(t)\right) - f_i\left(y_1(t), u_2(t)\right)| + |f_i\left(y_1(t), u_2(t)\right) - f_i\left(y_2(t), u_2(t)\right)| d\tilde{t} \\
&\leq \int_0^T L \sum_{j=1}^m |\left(u_1\right)_j(t) - \left(u_2\right)_j(t)| dt \\
&+ \int_0^t \sum_{l=1}^n \int_0^1 |\frac{\partial}{\partial y_l} f_i\left(y, u_2\left(\tilde{t}\right)\right)|_{y=y_2(\tilde{t})+\theta\left(y_1(\tilde{t})-y_2(\tilde{t})\right)}| d\theta \left(|\left(y_1\right)_l\left(\tilde{t}\right) - \left(y_2\right)_l\left(\tilde{t}\right)|\right) d\tilde{t} \\
&\leq L\|u_1 - u_2\|_{L^1} + c\int_0^t \sum_{l=1}^n \left(|\left(y_1\right)_l\left(\tilde{t}\right) - \left(y_2\right)_l\left(\tilde{t}\right)|\right) d\tilde{t}
\end{aligned} \tag{2.21}$$

due to Assumption A.5), Assumption A.6) and with the application of the fundamental theorem of calculus [4, VI 4.13]. By summing both sides of (2.21) over i, we obtain

$$\sum_{i=1}^n |\left(y_1\right)_i(t) - \left(y_2\right)_i(t)| \leq nL\|u_1 - u_2\|_{L^1} + cn\int_0^t \sum_{l=1}^n \left(|\left(y_1\right)_l\left(\tilde{t}\right) - \left(y_2\right)_l\left(\tilde{t}\right)|\right) d\tilde{t}$$

With the Gronwall's inequality, see Lemma 57 in the Appendix, we obtain that

$$\sum_{i=1}^n |\left(y_1\right)_i(t) - \left(y_2\right)_i(t)| \leq \left(nL + n^2 LcT \exp(cnT)\right) \|u_1 - u_2\|_{L^1}$$

where $C_2 \coloneqq nL + n^2 LcT \exp(cnT)$. □

In the next lemma, we have a boundedness result for two different solutions to the adjoint equation (2.5).

Lemma 10. *There exists a constant $C_3 > 0$ such that for any two solutions (p_1, y_1, u_1) and (p_2, y_2, u_2) to (2.5) with $u_1, u_2 \in U_{ad}$ the following holds*

$$|\left(p_1\right)_i(t) - \left(p_2\right)_i(t)| \leq C_3 \|u_1 - u_2\|_{L^1}$$

for all $t \in [0, T]$ and all $i \in \{1, ..., n\}$.

Proof. Analogously to the proof of Lemma 8, we have the initial value problem

$$\begin{aligned}
&\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau) \\
&= \frac{\partial}{\partial y_i} F(y)|_{y=\hat{y}_1(0)} - \frac{\partial}{\partial y_i} F(y)|_{y=\hat{y}_2(0)} + \frac{\partial}{\partial y_i} h(y)|_{y=\hat{y}_1(\tilde{\tau})} - \frac{\partial}{\partial y_i} h(y)|_{y=\hat{y}_2(\tilde{\tau})} d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)} \left(\hat{p}_1\right)_l(\tilde{\tau}) - \sum_{l=1}^n \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_2(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)} \left(\hat{p}_2\right)_l(\tilde{\tau}) d\tilde{\tau}
\end{aligned}$$

for the difference of each component $i \in \{1, ..., n\}$ of $\hat{p}_1$ and $\hat{p}_2$ which gives by adding and subtracting corresponding terms the following

$$\begin{aligned}
&\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau) \\
&= \frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}_1(0)} - \frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}_2(0)} + \int_0^\tau \frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}_1(\tilde{\tau})} - \frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}_2(\tilde{\tau})} d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)} \left(\left(\hat{p}_1\right)_l(\tilde{\tau}) - \left(\hat{p}_2\right)_l(\tilde{\tau})\right) d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \left(\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)} - \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)} \right) \left(\hat{p}_2\right)_l(\tilde{\tau})\, d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \left(\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)} - \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_2(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)} \right) \left(\hat{p}_2\right)_l(\tilde{\tau})\, d\tilde{\tau}.
\end{aligned}$$

Consequently, by taking the absolute value, we have that

$$\begin{aligned}
&|\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau)| \\
&\leq |\frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}_1(0)} - \frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}_2(0)}| + \int_0^\tau |\frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}_1(\tilde{\tau})} - \frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}_2(\tilde{\tau})}| d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n |\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)}| \left(|\left(\hat{p}_1\right)_l(\tilde{\tau}) - \left(\hat{p}_2\right)_l(\tilde{\tau})|\right) d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \left(|\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)} - \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)}| \right) |\left(\hat{p}_2\right)_l(\tilde{\tau})|\, d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \left(|\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)} - \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_2(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)}| \right) |\left(\hat{p}_2\right)_l(\tilde{\tau})|\, d\tilde{\tau}.
\end{aligned}$$

and thus

$$\begin{aligned}
&|\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau)| \\
&\leq \int_0^1 \sum_{\ell=1}^n |\frac{\partial}{\partial y_\ell}\frac{\partial}{\partial y_i} F(y)\,|_{y=\hat{y}_2(0)+\theta(\hat{y}_1(0)-\hat{y}_2(0))}| d\theta \left(|\left(\hat{y}_1\right)_\ell(0) - \left(\hat{y}_2\right)_\ell(0)|\right) \\
&+ \int_0^\tau \int_0^1 \sum_{\ell=1}^n |\frac{\partial}{\partial y_\ell}\frac{\partial}{\partial y_i} h(y)\,|_{y=\hat{y}_2(\tilde{\tau})+\theta(\hat{y}_1(\tilde{\tau})-\hat{y}_2(\tilde{\tau}))}| d\theta \left(|\left(\hat{y}_1\right)_\ell(\tilde{\tau}) - \left(\hat{y}_2\right)_\ell(\tilde{\tau})|\right) d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n |\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)}| \left(|\left(\hat{p}_1\right)_l(\tilde{\tau}) - \left(\hat{p}_2\right)_l(\tilde{\tau})|\right) d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \left(\int_0^1 \sum_{\ell=1}^n |\frac{\partial}{\partial y_\ell}\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_1(\tau)+\theta(\hat{y}_1(\tau)-\hat{y}_2(\tau))}| d\theta \left(|\left(\hat{y}_1\right)_\ell(\tilde{\tau}) - \left(\hat{y}_2\right)_\ell(\tilde{\tau})|\right) \right) |\left(\hat{p}_2\right)_l(\tilde{\tau})|\, d\tilde{\tau} \\
&+ \int_0^\tau \sum_{l=1}^n \left(|\frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_1(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)} - \frac{\partial}{\partial y_i} \hat{f}_l\left(y, \hat{u}_2(\tilde{\tau})\right)|_{y=\hat{y}_2(\tau)}| \right) |\left(\hat{p}_2\right)_l(\tilde{\tau})|\, d\tilde{\tau}.
\end{aligned}$$

with the application of the fundamental theorem of calculus [4, VI 4.13], the triangle inequality for the Riemann integral [4, VI Theorem 4.3] and the Lebesgue integral [5, X Remark 2.1 e)]. With Lemma 8 to estimate $|\left(\hat{p}_2\right)_l(\tilde{\tau})|$ for all $l \in \{1,...,n\}$, Assumption A.6) and Assumption A.5), we obtain

$$\begin{aligned}
&|\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau)| \\
&\leq c\sum_{\ell=1}^{n} |\left(\hat{y}_1\right)_\ell(0) - \left(\hat{y}_2\right)_\ell(0)| + c\int_0^\tau \sum_{\ell=1}^{n} |\left(\hat{y}_1\right)_\ell(\tilde{\tau}) - \left(\hat{y}_2\right)_\ell(\tilde{\tau})| d\tilde{\tau} + c\int_0^\tau \sum_{l=1}^{n} \left(|\left(\hat{p}_1\right)_l(\tilde{\tau}) - \left(\hat{p}_2\right)_l(\tilde{\tau})|\right) d\tilde{\tau} \\
&+ C_1 cn \int_0^\tau \sum_{\ell=1}^{n} |\left(\hat{y}_1\right)_\ell(\tilde{\tau}) - \left(\hat{y}_2\right)_\ell(\tilde{\tau})| d\tilde{\tau} + C_1 L \int_0^\tau \sum_{j=1}^{m} |\left(\hat{u}_1\right)_j(\tilde{\tau}) - \left(\hat{u}_2\right)_j(\tilde{\tau})| d\tilde{\tau}
\end{aligned}$$

and thus by Lemma 9 to estimate $|\left(y_1\right)_\ell(\tilde{\tau}) - \left(y_2\right)_\ell(\tilde{\tau})|$ for all $\ell \in \{1,...,n\}$ and the transformation formula [5, X Examples 6.6 b)], we have the following

$$\begin{aligned}
&|\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau)| \\
&\leq cnC_2\|u_1 - u_2\|_{L^1} + cnTC_2\|u_1 - u_2\|_{L^1} + c\int_0^\tau \sum_{l=1}^{n} \left(|\left(\hat{p}_1\right)_l(\tilde{\tau}) - \left(\hat{p}_2\right)_l(\tilde{\tau})|\right) d\tilde{\tau} \qquad (2.22) \\
&+ C_1 cn^2 TC_2\|u_1 - u_2\|_{L^1} + C_1 L\|u_1 - u_2\|_{L^1}.
\end{aligned}$$

By summing up both sides of (2.22) over all $i \in \{1,...,n\}$ we obtain the following

$$\sum_{i=1}^{n} |\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau)| \leq \tilde{C}\|u_1 - u_2\|_{L^1} + cn\int_0^\tau \sum_{i=1}^{n} \left(|\left(\hat{p}_1\right)_i(\tilde{\tau}) - \left(\hat{p}_2\right)_i(\tilde{\tau})|\right) d\tilde{\tau}$$

where $\tilde{C} := cn^2C_2 + cn^2TC_2 + C_1cn^3TC_2 + C_1nL$. By Gronwall's inequality, see Lemma 57 in the Appendix, we have the following

$$\sum_{i=1}^{n} |\left(\hat{p}_1\right)_i(\tau) - \left(\hat{p}_2\right)_i(\tau)| \leq \tilde{C}\left(1 + cnT\mathrm{e}^{cnT}\right)\|u_1 - u_2\|_{L^1}$$

where $C_3 := \tilde{C}\left(1 + cnT\mathrm{e}^{cnT}\right)$. Thus we have the statement of the lemma after a backsubstitution to p_1 and p_2 for each component. □

Next, we state a lemma concerning the minimizing property of the SQH iterates. Specifically, if in one iterate no sufficient decrease of J is achieved, then it is possible to improve the descent by choosing a larger ϵ in K_ϵ. A similar result can be found in [85, 17]. This lemma proves that, by increasing ϵ in Step 4 of Algorithm 2.1 ($\epsilon \leftarrow \sigma\epsilon$), an ϵ is obtained such that the condition for sufficient decrease is satisfied. A similar result is proved in [17].

Lemma 11. *Let (y,u) and $\left(y^k, u^k\right)$ be generated by the SQH method, $k \in \mathbb{N}_0$. Then, there is a $\theta \geq 0$ independent of ϵ such that for the $\epsilon > 0$ currently chosen by the SQH method and with corresponding $\delta u := u - u^k$* the following holds

$$J(y,u) - J\left(y^k, u^k\right) \leq -(\epsilon - \theta)\|\delta u\|_{L^2}^2.$$

In particular, $J(y,u) - J\left(y^k, u^k\right) \leq 0$ for $\epsilon \geq \theta$.

Proof. We define $\delta y := y - y^k$, $\delta p := p - p^k$ and $\delta u^2 := \sum_{j=1}^{m}\left(\left(\delta u\right)_j\right)^2$ where $\left(\delta u\right)_j$ is the j-th component of the vector function δu. We also remark that we do not note the functional dependency on t for notational reasons. We use from Algorithm 2.1 that u is determined such that $K_\epsilon\left(t, y^k, u, u^k, p^k\right) \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$ for all $w \in K_U$ and thus especially

$$K_\epsilon\left(t, y^k, u, u^k, p^k\right) \leq K_\epsilon\left(t, y^k, u^k, u^k, p^k\right) = H\left(t, y^k, u^k, p^k\right)$$

for all $t \in [0,T]$. Then we have with (2.4) that

$$\begin{aligned} & J(y,u) - J\left(y^k,u^k\right) = \int_0^T h(y) + g(u) - h\left(y^k\right) - g\left(u^k\right) dt + F(y(T)) - F\left(y^k(T)\right) \\ & = \int_0^T H(y,u,p) - p^T f(y,u) - H\left(y^k,u^k,p^k\right) + \left(p^k\right)^T f\left(y^k,u^k\right) dt + F(y(T)) - F\left(y^k(T)\right) \\ & = \int_0^T H(y,u,p) - H\left(y^k,u,p^k\right) + \epsilon\delta u^2 + H\left(y^k,u,p^k\right) - H\left(y^k,u^k,p^k\right) - \epsilon\delta u^2 dt \\ & + \int_0^T -p^T f(y,u) + \left(p^k\right)^T f\left(y^k,u^k\right) dt + F(y(T)) - F\left(y^k(T)\right) \\ & \leq \int_0^T H(y,u,p) - H\left(y^k,u,p^k\right) - \epsilon\delta u^2 - p^T f(y,u) + \left(p^k\right)^T f\left(y^k,u^k\right) dt + \tilde{F}\left(y,y^k\right)^T \delta y(T) \end{aligned} \tag{2.23}$$

where we apply the fundamental theorem of calculus [4, VI 4.13] in the last line with

$$\tilde{F}\left(y,y^k\right) := \left(\int_0^1 D_y F\left(y^k(T) + \theta\left(y(T) - y^k(T)\right)\right) d\theta\right)^T.$$

Furthermore, we have with the Taylor series, see [4, VII Corollary 5.5, Theorem 5.8], with the symmetry of the second derivative [4, VII Theorem 5.2] that

$$\begin{aligned} & H(y,u,p) - H\left(y^k,u,p^k\right) = H(y,u,p) - H(y-\delta y,u,p-\delta p) \\ & = D_y H(y,u,p)\,\delta y + D_p H(y,u,p)\,\delta p - \frac{1}{2}\delta y^T D_{yy} H(y,u,p)\,\delta y - \delta p^T D_{yp} H(y,u,p)\,\delta y \\ & + R_2(H,y,p;\delta y,\delta p) \\ & = D_y h(y)\,\delta y + p^T D_y f(y,u)\,\delta y + f(y,u)^T \delta p - \frac{1}{2}\delta y^T D_{yy} h(y)\,\delta y - \frac{1}{2} p^T \delta y^T D_{yy} f(y,u)\,\delta y \\ & - \delta p^T D_y f(y,u)\,\delta y + R_2(H,y,p;\delta y,\delta p) \end{aligned} \tag{2.24}$$

where $D_y\cdot$, $D_p\cdot$ is the first derivative and $D_{yy}\cdot$, $D_{yp}\cdot$ is the second derivative with respect to the corresponding variable, see [4, VII, VII.4] for details. In addition, we have with (2.5) and [36, Corollary 6.3.9] that

$$\begin{aligned} \int_0^T D_y h(y)\,\delta y + p^T D_y f(y,u)\,\delta y dt &= \int_0^T -\left(p'\right)^T \delta y dt \\ &= \int_0^T p^T f(y,u) - p^T f\left(y^k,u^k\right) dt - D_y F(y(T))\,\delta y(T) \end{aligned} \tag{2.25}$$

since $\delta y(0) = 0$ and thus starting from (2.23) using (2.24) with (2.25) we obtain

$$\begin{aligned} & J(y,u) - J\left(y^k,u^k\right) \\ & \leq \int_0^T -\epsilon\delta u^2 + \left(p^k\right)^T f\left(y^k,u^k\right) + R_2(H,y,p;\delta y,\delta p)\,dt + \left(\tilde{F}\left(y,y^k\right) - D_y F(y(T))\right)\delta y(T) \\ & + \int_0^T -p^T f\left(y^k,u^k\right) + f(y,u)^T \delta p - \frac{1}{2}\delta y^T D_{yy} h(y)\,\delta y - \frac{1}{2} p^T \delta y^T D_{yy} f(y,u)\,\delta y - \delta p^T D_y f(y,u)\,\delta y dt \\ & = \int_0^T -\epsilon\delta u^2 + \delta p^T \left(f(y,u) - f\left(y^k,u^k\right)\right) - \frac{1}{2}\delta y^T D_{yy} h(y)\,\delta y - \frac{1}{2} p^T \delta y^T D_{yy} f(y,u)\,\delta y dt \\ & + \int_0^T -\delta p^T D_y f(y,u)\,\delta y + R_2(H,y,p;\delta y,\delta p)\,dt + \left(\tilde{F}\left(y,y^k\right)^T - D_y F(y(T))\right)\delta y(T). \end{aligned} \tag{2.26}$$

Further, we have with

$$f(y,u) - f\left(y^k,u^k\right) = f(y,u) - f\left(y,u^k\right) + f\left(y,u^k\right) - f\left(y^k,u^k\right)$$

and from (2.26) the following

$$\begin{aligned} & J(y,u) - J\left(y^k,u^k\right) \\ & \le \int_0^T -\epsilon\delta u^2 + \delta p^T\left(f(y,u) - f\left(y,u^k\right) + f\left(y,u^k\right) - f\left(y^k,u^k\right)\right) - \frac{1}{2}\delta y^T D_{yy}h(y)\,\delta y dt \\ & + \int_0^T -\frac{1}{2}p^T\delta y^T D_{yy}f(y,u)\,\delta y dt - \delta p^T D_y f(y,u)\,\delta y + R_2(H,y,p;\delta y,\delta p)\,dt \\ & + \left(\int_0^1 D_yF\left(y^k(T) + \theta\delta y(T)\right) - D_yF(y(T))\,d\theta\right)\delta y(T). \end{aligned} \tag{2.27}$$

Consequently by the fundamental theorem of calculus [4, VI 4.13] we obtain from (2.27) the following

$$\begin{aligned} & J(y,u) - J\left(y^k,u^k\right) \\ & \le -\epsilon\|\delta u\|_{L^2}^2 + \int_0^T \sum_{i=1}^n \delta p_i\left(f_i(y,u) - f_i\left(y,u^k\right)\right) + \sum_{i=1}^n\sum_{l=1}^n \delta p_i \int_0^1 \frac{\partial}{\partial y_l} f_i\left(y,u^k\right)|_{y=y^k+\theta\delta y} d\theta\delta y_l dt \\ & + \int_0^T -\frac{1}{2}\sum_{i=1}^n\sum_{l=1}^n \delta y_l \frac{\partial^2}{\partial y_l\partial y_i} h(y)\,\delta y_i - \frac{1}{2}\sum_{i=1}^n\sum_{l=1}^n\sum_{\ell=1}^n p_i\delta y_l \frac{\partial^2}{\partial y_l\partial y_\ell} f_i\delta y_\ell - \sum_{i=1}^n\sum_{l=1}^n \delta p_i \frac{\partial}{\partial y_l} f_i(y,u)\,\delta y_l dt \\ & + \int_0^T R_2(H,y,p;\delta y,\delta p)\,dt \\ & + \sum_{i=1}^n\sum_{l=1}^n \delta y_l(T)\int_0^1\int_0^1 \frac{\partial}{\partial y_l\partial y_i} F\left(y(T) + \hat{\theta}((\theta-1)\,\delta y(T))\right) d\hat{\theta}\,(1-\theta)\,d\theta\delta y_i(T) \end{aligned} \tag{2.28}$$

where we use that it holds

$$\begin{aligned} & \int_0^1 D_yF\left(y^k(T) + \theta\delta y(T)\right) - D_yF(y(T))\,d\theta \\ & = \int_0^1\int_0^1 D_{yy}F\left(y(T) + \hat{\theta}\left(y^k(T) + \theta\delta y(T) - y(T)\right)\right) d\hat{\theta}\,(\theta-1)\,\delta y(T)\,d\theta \\ & = \int_0^1\int_0^1 D_{yy}F\left(y(T) + \hat{\theta}((\theta-1)\,\delta y(T))\right) d\hat{\theta}\,(\theta-1)\,\delta y(T)\,d\theta \end{aligned}$$

also due to the fundamental theorem of calculus [4, VI 4.13].

Now by Assumptions A.5) and A.6), Lemma 8, Lemma 9 and Lemma 10, we obtain from (2.28) the following

$$\begin{aligned} & J(y,u) - J\left(y^k,u^k\right) \\ & \le -\epsilon\|\delta u\|_{L^2}^2 + nLC_3\|\delta u\|_{L^1}^2 + n^2TcC_3C_2\|\delta u\|_{L^1}^2 + \frac{1}{2}n^2TcC_2^2\|\delta u\|_{L^1}^2 + \frac{1}{2}n^3TcC_1C_2^2\|\delta u\|_{L^1}^2 \\ & + n^2TcC_3C_2\|\delta u\|_{L^1}^2 + n^2cC_2^2\|\delta u\|_{L^1}^2 + \int_0^T R_2(H,y,p;\delta y,\delta p)\,dt \end{aligned} \tag{2.29}$$

where we use the triangle inequality for the Riemann integral [4, VI Theorem 4.3] and the Lebesgue integral [5, X Remark 2.1 e)]. Next, by [1, Theorem 2.14], we have that there is a constant $\tilde{c} > 0$ such that

$\|\delta u\|_{L^1} \leq \tilde{c}\|\delta u\|_{L^2}$. Furthermore the Taylor remainder $R_2\left(H, y, p; \delta y, \delta p\right)$ is estimated by the remainder formula [4, VII Theorem 5.8] and the boundedness of the second derivatives analogously to the calculation which are done for the second derivatives for (2.29). Consequently from (2.29), we obtain the existence of a constant $\theta > 0$ such that

$$J\left(y, u\right) - J\left(y^k, u^k\right) \leq -\epsilon\|\delta u\|_{L^2}^2 + \theta\|\delta u\|_{L^2}^2$$

which proofs the claim. □

Next, we prove a lemma stating that Algorithm 2.1 stops when u^k, $k \in \mathbb{N}_0$ is a solution to (2.16).

Lemma 12. *Let y^k and u^k be generated by Algorithm 2.1, $k \in \mathbb{N}_0$. If the iterate u^k is optimal, then Algorithm 2.1 stops, returning u^k.*

Proof. If u^k, $k \in \mathbb{N}_0$, is optimal, then we have, according to Theorem 5, that

$$H\left(t, y^k, u^k, p^k\right) = \min_{w \in K_U} H\left(t, y^k, w, p^k\right)$$

for almost all $t \in (0, T)$ and thus

$$\begin{aligned} K_\epsilon\left(t, y^k, u^k, u^k, p^k\right) &= H\left(t, y^k, u^k, p^k\right) \leq H\left(t, y^k, w, p^k\right) \\ &\leq H\left(t, y^k, w, p^k\right) + \epsilon\left(w - u^k\left(t\right)\right)^2 = K_\epsilon\left(t, y^k, w, u^k, p^k\right) \end{aligned}$$

for all $w \in K_U$ and for almost all $t \in (0, T)$. That means that an optimal solution is always among those candidates being selected by our algorithm. On the other hand, once having an optimal solution u^k, we have to exclude that there is a $\tilde{t} \in [0, T]$ where u^k is optimal and a $\tilde{u}$ with $\left(\tilde{u}\left(\tilde{t}\right) - u^k\left(\tilde{t}\right)\right)^2 > 0$ such that $K_\epsilon\left(\tilde{t}, y^k, \tilde{u}, u^k, p^k\right) \leq K_\epsilon\left(\tilde{t}, y^k, u^k, u^k, p^k\right)$ in order to ensure that Algorithm 2.1 stays in its determined optimal solution u^k.

Suppose $K_\epsilon\left(\tilde{t}, y^k, \tilde{u}, u^k, p^k\right) \leq K_\epsilon\left(\tilde{t}, y^k, u^k, u^k, p^k\right)$. First, we have, because of the optimality of u^k, that $H\left(\tilde{t}, y^k, u^k, p^k\right) \leq H\left(\tilde{t}, y^k, w, p^k\right)$ for all $w \in K_U$, especially for $w = \tilde{u}\left(\tilde{t}\right)$. Then, we conclude from

$$K_\epsilon\left(\tilde{t}, y^k, \tilde{u}, u^k, p^k\right) \leq K_\epsilon\left(\tilde{t}, y^k, u^k, u^k, p^k\right)$$

and the optimality of u^k that

$$\begin{aligned} &H\left(\tilde{t}, y^k, u^k, p^k\right) + \epsilon\left(\tilde{u}\left(\tilde{t}\right) - u^k\left(\tilde{t}\right)\right)^2 \leq H\left(\tilde{t}, y^k, \tilde{u}, p^k\right) + \epsilon\left(\tilde{u}\left(\tilde{t}\right) - u^k\left(\tilde{t}\right)\right)^2 \\ &= K_\epsilon\left(\tilde{t}, y^k, \tilde{u}, u^k, p^k\right) \leq K_\epsilon\left(\tilde{t}, y^k, u^k, u^k, p^k\right) = H\left(\tilde{t}, y^k, u^k, p^k\right) \end{aligned}$$

and consequently $\epsilon\left(\tilde{u}\left(\tilde{t}\right) - u^k\left(\tilde{t}\right)\right)^2 \leq 0$. Algorithm 2.1 has updated the initial guess u^0 at most k times where ϵ is decreased by $\epsilon \leftarrow \zeta\epsilon$. Thus, we have that $\epsilon > 0$ and therefore $\left(\tilde{u}\left(\tilde{t}\right) - u^k\left(\tilde{t}\right)\right)^2 \leq 0$, which means that $\tilde{u} = u^k$ almost everywhere since the calculation holds for any $\tilde{t} \in [0, T]$ where u^k is optimal. Thus $\delta u = 0$ in the $L^2\left(0, T\right)$ sense and Algorithm 2.1 stops and returns u^k. □

The following theorem states that the iteration over the Steps 2 to 4 in Algorithm 2.1 (no stopping criterion) generates sequences $\left(u^k\right)_{k \in \mathbb{N}_0}$ and $\left(y^k\right)_{k \in \mathbb{N}_0}$ such that the cost functional $J\left(y^k, u^k\right)$ monotonically decreases with $\lim_{k \to \infty} \|u^{k+1} - u^k\|_{L^2} = 0$. A similar result is proved in [17]. In the next two theorems we analyze the convergence properties of these sequences $\left(u^k\right)_{k \in \mathbb{N}_0}$ and $\left(y^k\right)_{k \in \mathbb{N}_0}$. Considering Lemma 12, we assume for the rest of this section that no element of the sequence $\left(u^k\right)_{k \in \mathbb{N}_0}$ is optimal.

Theorem 13. *Let the sequence $\left(y^k\right)_{k\in\mathbb{N}_0}$ and $\left(u^k\right)_{k\in\mathbb{N}_0}$ be generated as in Algorithm 2.1 (loop over Step 2 to Step 4). Then, the sequence of cost functional values $J\left(y^k,u^k\right)$ monotonically decreases with*

$$\lim_{k\to\infty}\left(J\left(y^{k+1},u^{k+1}\right)-J\left(y^k,u^k\right)\right)=0$$

and

$$\lim_{k\to\infty}\|u^{k+1}-u^k\|_{L^2}=0.$$

Proof. Due to Lemma 11, we have that Algorithm 2.1 determines $\epsilon>\theta$ in finitely many steps and we obtain an update of the control that reduces the value of the cost functional by at least $-\left(\epsilon-\theta\right)\|u^{k+1}-u^k\|_{L^2}^2$. This is seen as follows.

If the update is rejected because $J\left(y,u\right)-J\left(y^k,u^k\right)>-\eta\|u-u^k\|_{L^2}^2$, then ϵ is further increased until $\epsilon-\theta\geq\eta$ and thus

$$J\left(y,u\right)-J\left(y^k,u^k\right)\leq-\left(\epsilon-\theta\right)\|u-u^k\|_{L^2}^2\leq-\eta\|u-u^k\|_{L^2}^2. \tag{2.30}$$

Therefore there is an update after at least finitely many increases of ϵ in Step 4 of Algorithm 2.1 and we have that $u^{k+1}\leftarrow u$ with the corresponding u. Then we always have $J\left(y^{k+1},u^{k+1}\right)\leq J\left(y^k,u^k\right)$ and thus the sequence of iterates $J\left(y^k,u^k\right)$ monotonically decreases.

As the cost functional is bounded from below, we have for any $\rho>0$ the existence of a $k\in\mathbb{N}_0$ such that

$$-\rho\eta\leq J\left(y^{k+1},u^{k+1}\right)-J\left(y^k,u^k\right)\leq0 \tag{2.31}$$

because any sequence bounded from below converges and any converging sequence is a Cauchy sequence, see [3, II Theorem 4.1, Theorem 6.1] for details, whose definition is used in (2.31).

Finally, as (2.30) also holds for u^{k+1} instead of u, we obtain from (2.30) and (2.31) the following

$$\rho\eta\geq-\left(J\left(y^{k+1},u^{k+1}\right)-J\left(y^k,u^k\right)\right)\geq\eta\|u^{k+1}-u^k\|_{L^2}^2\geq0$$

for k sufficiently large and thus $0\leq\|u^{k+1}-u^k\|_{L^2}^2\leq\rho$ for k sufficiently large. As $\rho>0$ can be chosen arbitrarily small, we have $\lim_{k\to\infty}\|u^{k+1}-u^k\|_{L^2}=0$. □

From Theorem 13, we obtain that Algorithm 2.1 is well defined for $\kappa>0$. This means that there is an iteration number $\bar{k}\in\mathbb{N}_0$ such that $\|u^{\bar{k}+1}-u^{\bar{k}}\|_{L^2}<\kappa$ and consequently Algorithm 2.1 stops in finitely many steps; see Step 4 in Algorithm 2.1 and Lemma 11.

Notice that u determined in Algorithm 2.1 Step 2 is measurable and, due to the pointwise bounds, we have $u\in U_{ad}$. Thus especially $\left(u^k\right)_{k\in\mathbb{N}_0}\subseteq U_{ad}$.

So far we have proven that Algorithm 2.1 is well defined, that means stops in finite time and generates iterates u^k, $k\in\mathbb{N}_0$, that minimize the cost functional as long as no iterate is optimal in the sense of Theorem 5. The content of the next theorem is to prove the convergence of the iterates u^k to a limit that fulfills (2.16) at least up to a tolerance that can be chosen arbitrarily small. This limit is called a PMP consistent solution of the SQH method. For this purpose, we state a further assumption to guarantee convergence of the SQH method to a PMP consistent solution. Instead of a differentiability assumption of the Hamiltonian with respect to the control argument, we have the following inequality to prove a convergence theorem for the SQH method. We discuss our theoretical result represented by the following theorem that states the convergence of the SQH method to the PMP solution, characterized by (2.16), without any differentiability assumptions on the Hamiltonian function with respect to the control argument u. Therefore this result can be applied to optimal control problems with discontinuous and non-convex cost functionals. In our discussion, the assumption of differentiability of H with respect to the control is replaced by the requirement that for any iterate u^k, $k \in \mathbb{N}_0$ and for any ϵ chosen by Algorithm 2.1 there exists an $r \geq \epsilon$ such that

$$K_\epsilon\left(t, y^k, u^{k+1}, u^k, p^k\right) + r\left(w - u^{k+1}(t)\right)^2 \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right) \tag{2.32}$$

is fulfilled for all $w \in K_U$ and for all $t \in [0,T]$. This condition ensures sufficient descent of the cost functional. In addition (2.32) implicates that $u^{k+1}(t)$ is the global minimum of the function $w \mapsto K_\epsilon\left(t, y^k, w, u^k, p^k\right)$. If this was not the case and if there was another control function $\tilde{u}$ such that for one $\tilde{t} \in [0,T]$ it held $u^{k+1}\left(\tilde{t}\right) \neq \tilde{u}\left(\tilde{t}\right)$ and $K_\epsilon\left(\tilde{t}, y^k, u^{k+1}, u^k, p^k\right) = K_\epsilon\left(\tilde{t}, y^k, \tilde{u}, u^k, p^k\right)$, then from (2.32) for $w = \tilde{u}\left(\tilde{t}\right)$, we would have that $r\left(\tilde{u}\left(\tilde{t}\right) - u^{k+1}\left(\tilde{t}\right)\right)^2 \leq 0$ and thus $\tilde{u}\left(\tilde{t}\right) = u^{k+1}\left(\tilde{t}\right)$ in contradiction to $u^{k+1}\left(\tilde{t}\right) \neq \tilde{u}\left(\tilde{t}\right)$ since $r > 0$ because $\epsilon > 0$.

A further implication of (2.32) is that if u in Step 2 of Algorithm 2.1 is determined such that it equals u^k and thus $u^{k+1} = u^k$, then we have that $\left(y^k, u^k, p^k\right)$ is already optimal in the sense of (2.16), which means that

$$H\left(t, y^k, u^k, p^k\right) = \min_{w \in K_U} H\left(t, y^k, w, p^k\right)$$

for almost every $t \in (0,T)$ since we have from (2.32) the following

$$H\left(t, y^k, u^k, p^k\right) + \epsilon\left(u^k(t) - u^k(t)\right)^2 + r\left(w - u^k(t)\right)^2 \leq H\left(t, y^k, u^k, w, p^k\right) + \epsilon\left(w - u^k(t)\right)^2$$

for all $w \in K_U$ and for all $t \in [0,T]$ from which the PMP optimality condition (2.16) follows because $\epsilon \geq r$ and $u^{k+1} = u^k$. Regarding Theorem 13, this means that if (2.32) holds, an update of an iterate u^k strictly decreases the cost functional value with respect to this iterate or u^k is already PMP optimal if the update provides the same cost functional value as u^k.

In Example 17, we show that (2.32) is fulfilled for an L^2-cost functional and in Example 18, we verify (2.32) for an L^1-cost functional. Certain discontinuous functionals are discussed later in the thesis as an L^0-cost functional or an L^1- like cost functional. First we prove the following theorem that states the PMP consistency of the result of our SQH method. For this purpose we extract subsequences from the sequence of iterates. In order to denote the elements of this subsequence we use an infinite index set, denoted for instance with K, that is a subset of $\mathbb{N}_0$, which is denoted with $K \subseteq \mathbb{N}_0$.

Theorem 14. *Let the sequence $\left(u^n\right)_{n \in \mathbb{N}_0}$ be generated as in Algorithm 2.1 (loop over Step 2 to Step 4) and let (2.32) hold. Then for any subsequence $\left(u^k\right)_{k \in K}$, $k \in K \subseteq \mathbb{N}_0$, with the property*

$$\lim_{K \ni k \to \infty} u^k(t) = \bar{u}(t)$$

for almost all $t \in (0,T)$, it holds that $\bar{u} \in U_{ad}$ and

$$H\left(t, \bar{y}, \bar{u}, \bar{p}\right) = \min_{w \in K_U} H\left(t, \bar{y}, w, \bar{p}\right)$$

for almost all $t \in (0,T)$ where $\bar{y}$ solves (2.1) with $u \leftarrow \bar{u}$ and $\bar{p}$ is the corresponding adjoint variable solving (2.5) for $y \leftarrow \bar{y}$ and $u \leftarrow \bar{u}$.

Furthermore, for almost each $t \in (0,T)$ and any $\mu > 0$, there exists a $\bar{k} \in \mathbb{N}$ and an index set $K_1 \subseteq \mathbb{N}_0$ and a $\bar{k} \in K_1$ such that

$$H\left(t, y^{k+1}, u^{k+1}, p^{k+1}\right) \leq H\left(t, y^{k+1}, w, p^{k+1}\right) + \mu \tag{2.33}$$

for all $w \in K_U$ and for all $k \geq \bar{k}$ with $k \in K_1$.

Proof. The iterates u^n, $n \in \mathbb{N}_0$ are measurable, see Section 5.1. Thus $\bar{u}$ is measurable, see [5, X Theorem 1.14]. Since $u^k(t) \in K_U$ for almost all $t \in (0,T)$ we have that $\bar{u}(t) \in K_U$ for almost all $t \in (0,T)$, see [3, II Theorem 2.7]. Since K_U is bounded we have that $\bar{u} \in U_{ad}$ because $\bar{u}$ is integrable.

Next we construct a subsequence having all the properties that we need for the proof. Because of the boundedness of

$$|u_j^k(t) - \bar{u}_j(t)| \leq 2 \max_{j=1,\dots,m} \max_{u_j \in K_U^j} |u_j|$$

for almost all $t \in (0,T)$, $k \in K$ and all $j \in \{1,...,m\}$, we have by the dominated convergence theorem [36, Theorem 2.4.5] that

$$\lim_{k\to\infty} \|u^k - \bar{u}\|_{L^1} = 0.$$

Due to Lemma 9 and Lemma 10, we have $\lim_{k\to\infty} y^k(t) = \bar{y}(t)$ and $\lim_{k\to\infty} p^k(t) = \bar{p}(t)$ for almost all $t \in (0,T)$. According to Theorem 13, we have for the sequence $(u^n)_{n\in\mathbb{N}_0}$ that it holds $\lim_{n\to\infty} \|u^{n+1} - u^n\|_{L^2} = 0$. Consequently we have for any subsequence and thus for the sequence $\left(u^k\right)_{k\in K}$ that

$$\lim_{k\to\infty} \|u^{k+1} - u^k\|_{L^2} = 0$$

since u^{k+1} is the following element of u^k in the sequence $(u^n)_{n\in\mathbb{N}_0}$. Then by [6, Proposition 3.6, Remark 3.7], there exists a subsequence with the index set $K_1 \subseteq K$ such that

$$\lim_{K_1 \ni k\to\infty} \left(u^{k+1}(t) - u^k(t)\right) = 0$$

for almost all $t \in (0,T)$ where all the other convergence properties above remain since any subsequence of a converging sequence also converges, see [3, II Theorem 1.15]. From this we can also conclude that

$$\lim_{K_1 \ni k\to\infty} u^{k+1}(t) = \lim_{K_1 \ni k\to\infty} \left(u^{k+1}(t) - u^k(t)\right) + \lim_{K_1 \ni k\to\infty} u^k(t) = \bar{u}(t) \tag{2.34}$$

for almost all $t \in (0,T)$ where we use the calculation rules for the limit [3, II Theorem 2.2]. Analogously we have with Lemma 9, Lemma 10 and [1, Theorem 2.14], that $\lim_{K_1 \ni k\to\infty} \left(y^{k+1}(t) - y^k(t)\right) = 0$, $\lim_{K_1 \ni k\to\infty} \left(p^{k+1}(t) - p^k(t)\right) = 0$ and thus

$$\lim_{K_1 \ni k\to\infty} y^{k+1}(t) = \bar{y}(t)$$

and

$$\lim_{K_1 \ni k\to\infty} p^{k+1}(t) = \bar{p}(t)$$

for almost all $t \in (0,T)$.

As the control u^k, $k \in K_1$, is an element of $(u^n)_{n\in\mathbb{N}_0}$, it is determined by Algorithm 2.1 such that due to (2.32) the following holds

$$K_\epsilon\left(t, y^k, u^{k+1}, u^k, p^k\right) + r\left(w - u^{k+1}(t)\right)^2 \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$$

with $r \geq \epsilon$ for all $w \in K_U$, for all $k \in K_1$ and all $t \in [0,T]$ which gives by inserting the definition of K_ϵ the following

$$\begin{aligned} &H\left(t, y^k, u^{k+1}, p^k\right) + \epsilon\left(u^{k+1}(t) - u^k(t)\right)^2 + r\left(w - u^{k+1}(t)\right)^2 \\ &\leq H\left(t, y^k, w, p^k\right) + \epsilon\left(w - u^k(t)\right)^2 \end{aligned} \tag{2.35}$$

for all $w \in K_U$, for all $k \in K_1$ and all $t \in [0,T]$.

Now, we consider (2.35) where it also holds due to our assumption $r \geq \epsilon$ that

$$H\left(t, y^k, u^{k+1}, p^k\right) + \epsilon\left(u^{k+1}(t) - u^k(t)\right)^2 + \epsilon\left(w - u^{k+1}(t)\right)^2 \leq H\left(t, y^k, w, p^k\right) + \epsilon\left(w - u^k(t)\right)^2$$

and thus by inserting

$$\left(w - u^{k+1}(t)\right)^2 = \left(w - u^k(t)\right)^2 + \left(u^k(t) - u^{k+1}(t)\right)^2 + 2\left(w - u^k(t)\right)^T\left(u^k(t) - u^{k+1}(t)\right)$$

we obtain

$$\begin{aligned} &H\left(t, y^k, u^{k+1}, p^k\right) + 2\epsilon\left(u^{k+1}(t) - u^k(t)\right)^2 + 2\epsilon\left(w - u^k(t)\right)^T\left(u^k(t) - u^{k+1}(t)\right) \\ &\leq H\left(t, y^k, w, p^k\right) \end{aligned} \tag{2.36}$$

for all $w \in K_U$, for all $k \in K_1$ and all $t \in (0, T)$. Then, by using the definition of the Hamiltonian, (2.36) becomes the following

$$\begin{aligned} &h\left(y^k(t)\right) + g\left(u^{k+1}(t)\right) + \left(p^k\right)^T(t) f\left(t, y^k, u^{k+1}\right) + 2\epsilon\left(u^{k+1}(t) - u^k(t)\right)^2 \\ &+ 2\epsilon\left(w - u^k(t)\right)^T\left(u^k(t) - u^{k+1}(t)\right) \leq h\left(y^k(t)\right) + g(w) + \left(p^k\right)^T(t) f\left(t, y^k, w\right) \end{aligned} \tag{2.37}$$

for all $w \in K_U$, for all $k \in K_1$ and all $t \in (0, T)$. Next, we have that ϵ is bounded from below by 0 and from above by $\sigma(\eta + \theta)$ due to (2.30) and the definition of Step 4 in Algorithm 2.1. The boundedness of ϵ guarantees that the corresponding terms in (2.37) go to zero for k to infinity, see [3, Theorem 2.4, Theorem 6.1] since $u^k(t) - u^{k+1}(t)$ converges pointwise to zero for $k \in K_1$ and $\left(w - u^k(t)\right)$ is also bounded as $w, u^k \in K_U$ for all $k \in K_1$. This connection is exploited in the next step. Now, as g is lower semi-continuous, we apply the lim inf on both sides of (2.37) where $k \in K_1$ and recall that whenever a lim exists the corresponding lim inf equals the lim, see [3, Theorem 5.7] and recall the calculation rules for a sum of lim inf [43, Theorem 3.127]. Further, we set $u^{k+1}(t) =: a^{k+1} \to \bar{a} := \bar{u}(t)$ for $k \to \infty$ and for fixed t. Then we have

$$\liminf_{K_1 \ni k \to \infty} g\left(u^{k+1}(t)\right) = \liminf_{K_1 \ni k \to \infty} g\left(a^{k+1}\right) \geq g(\bar{a}) = g(\bar{u}(t))$$

for almost all $t \in (0, T)$ according to (2.34).We obtain for the left-hand side of (2.37) the following

$$\begin{aligned} &\liminf_{K_1 \ni k \to \infty}\Bigg(h\left(y^k(t)\right) + g\left(u^{k+1}(t)\right) + \left(p^k\right)^T(t) f\left(t, y^k, u^{k+1}\right) + 2\epsilon\left(u^{k+1}(t) - u^k(t)\right)^2 \\ &+2\epsilon\left(w - u^k(t)\right)^T\left(u^k(t) - u^{k+1}(t)\right)\Bigg) \geq h(\bar{y}(t)) + g(\bar{u}(t)) + \bar{p}^T(t) f(t, \bar{y}, \bar{u}) = H(t, \bar{y}, \bar{u}, \bar{p}) \end{aligned}$$

where we use the continuity of f according to Assumption A.4). For the right-hand side of (2.37), we have

$$\begin{aligned} &\liminf_{K_1 \ni k \to \infty}\left(h\left(y^k(t)\right) + g(w) + p^k(t) f\left(t, y^k, w\right)\right) \\ &= \lim_{K_1 \ni k \to \infty}\left(h\left(y^k(t)\right) + g(w) + \left(p^k\right)^T(t) f\left(t, y^k, w\right)\right) \\ &= h(\bar{y}(t)) + g(w) + \bar{p}^T(t) f(t, \bar{y}, w) = H(t, \bar{y}, w, \bar{p}) \end{aligned}$$

where we also use the continuity for f, see Assumption A.4) and recall that differentiable functions are continuous, see [78, 1 The Rules of Differentiation]. Consequently, we obtain the optimality condition

$$H(t, \bar{y}, \bar{u}, \bar{p}) \leq H(t, \bar{y}, w, \bar{p})$$

for all $w \in K_U$ and almost all $t \in (0,T)$.

In order to prove (2.33), we consider (2.35) inserting the assumption $r \geq \epsilon$ and obtain

$$\begin{aligned}
H\left(t, y^{k+1}, u^{k+1}, p^{k+1}\right) &\leq H\left(t, y^{k+1}, w, p^{k+1}\right) \\
&+ \left|\left(p^k\right)^T (t)\, f\left(t, y^k, u^{k+1}\right) - \left(p^{k+1}\right)^T (t)\, f\left(t, y^{k+1}, u^{k+1}\right)\right| \\
&+ \left|\left(p^k\right)^T (t)\, f\left(t, y^k, w\right) - \left(p^{k+1}\right)^T (t)\, f\left(t, y^{k+1}, w\right)\right| \\
&+ \epsilon \left|\left(\left(w - u^k(t)\right)^2 - \left(u^{k+1}(t) - u^k(t)\right)^2 - \left(w - u^{k+1}(t)\right)^2\right)\right|
\end{aligned} \tag{2.38}$$

by adding and subtracting corresponding terms. Now, for almost all $t \in (0,T)$, by continuity, especially Assumption A.4), and $k \in K_1$, it follows the result (2.33) if k is sufficiently large such that the last three terms in (2.38) are smaller than the given μ using the boundedness of ϵ and [3, II Theorem 2.4]. □

Remark. If Algorithm 2.1 determines a u in Step 2 with $u = u^n$ and consequently $u^{n+1} = u$, then by (2.32) and the discussion following it, we have that u^n already fulfills the PMP optimality condition (2.16). On the other hand, considering just a loop over Step 2 to Step 4, we have that, by Lemma 12, the iterates following u^n generated by Algorithm 2.1 equal u^n up to a set of measure zero since the union of countably many null sets is a null set, see [5, IX Remark 2.5 (b)]. This means that the iterates of the SQH method generate a constant sequence in the $L^2(0,T)$-sense with the limit u^n. Since this also means pointwise convergence of the iterates almost everywhere, the requirements of Theorem 14 are fulfilled and the PMP optimality of u^n is consistent with the statement of Theorem 14.

Remark 15. We have that Theorem 14 holds in a similar formulation for any subsequence $\left(u^k\right)_{k\in K}$, $K \subseteq \mathbb{N}$, of $\left(u^n\right)_{n\in\mathbb{N}_0}$ with the property that $\lim_{K\ni k\to\infty} \|u^k - \bar{u}\|_{L^2} = 0$ with $\bar{u} \in U_{ad}$ where we drop the assumption $\lim_{K\ni k\to\infty} u^k(t) = \bar{u}(t)$ for almost all $t \in (0,T)$, see Theorem 27.

Remark 16. If we consider (2.33) on a set of a finite number of elements $F_N \subsetneq [0,T]$ of the interval $[0,T]$ and a fixed $\mu > 0$, then there is a $\bar{k} \in \mathbb{N}$ such that (2.33) holds for all $t \in F_N$. This can be seen by applying the calculation following (2.38) for each $t \in F_N$ and then choosing the largest $\bar{k}$.

From the proof of Theorem 14 and Remark 16, we obtain the following corollary. As shown in the proof of Theorem 14, there exists a subsequence within the sequence of iterates $\left(u^n\right)_{n\in\mathbb{N}_0}$ such that the pointwise convergence

$$\lim_{k\to\infty} \left(u^{k+1}(t) - u^k(t)\right) = 0, \ \lim_{k\to\infty} \left(y^{k+1}(t) - y^k(t)\right) = 0 \text{ and } \lim_{k\to\infty} \left(p^{k+1}(t) - p^k(t)\right) = 0,$$

$k \in K \subseteq \mathbb{N}_0$ holds for almost all $t \in (0,T)$. We denote with $\tilde{\Omega} \subseteq [0,T]$ the subset of $[0,T]$ on which the pointwise convergence above holds everywhere. That means the set $\tilde{\Omega}$ equals $[0,T]$ up to a set of measure zero. Then the corollary is given as follows.

Corollary. *Let the sequence $\left(u^n\right)_{n\in\mathbb{N}_0}$ be generated as in Algorithm 2.1 (loop over Step 2 to Step 4), let (2.32) hold and let f be bounded as follows $\|f(\cdot, y, u)\| \leq \tilde{c}$, $\tilde{c} > 0$, for all $y \in I$ and $u \in K_U$. For any set of a finite number of elements $F_N \subsetneq [0,T]$ and any $\tilde{\mu} > 0$, there exists an iterate $u^{\tilde{n}}$, $\tilde{n} \in \mathbb{N}_0$, fulfilling*

$$H\left(t, y^{\tilde{n}}, u^{\tilde{n}}, p^{\tilde{n}}\right) \leq H\left(t, y^{\tilde{n}}, w, p^{\tilde{n}}\right) + \tilde{\mu}$$

for all $w \in K_U$ and for any $t \in F_N$.

Proof. From (2.38) by adding and subtracting corresponding terms and not explicitly noting the time dependence to save notational effort, we obtain the following

$$\begin{aligned}
&H\left(y^{k+1},u^{k+1},p^{k+1}\right) \leq H\left(y^{k+1},w,p^{k+1}\right)\\
&+|\left(p^k-p^{k+1}\right)^T f\left(y^k,u^{k+1}\right)+\left(p^{k+1}\right)^T\left(f\left(y^k,u^{k+1}\right)-f\left(y^{k+1},u^{k+1}\right)\right)|\\
&+|\left(p^k-p^{k+1}\right)^T f\left(y^k,w\right)+\left(p^{k+1}\right)^T\left(f\left(y^k,w\right)-f\left(y^{k+1},w\right)\right)|\\
&+\epsilon|\left(-2wu^k+\left(u^k\right)^2+2wu^{k+1}-\left(u^{k+1}\right)^2-\left(u^{k+1}-u^k\right)^2\right)|\\
&\leq 2\tilde{c}\sum_{i=1}^n|p_i^k-p_i^{k+1}|+2c^2\sum_{i=1}^n|y_i^k-y_i^{k+1}|\\
&+\epsilon\left(2w|u^{k+1}-u^k|+|u^k+u^{k+1}||u^k-u^{k+1}|+|u^{k+1}-u^k|^2\right)
\end{aligned}$$

for any $t\in\tilde{\Omega}$ where we use the fundamental theorem of calculus [4, VI 4.13] and the chain rule [4, VII Theorem 3.3]. Consequently from the inequality above, due to the pointwise convergence mentioned just before this corollary and the boundedness of K_U, for any $t\in F_N$, there is a $\hat{k}_t\in K$ such that

$$H\left(y^{\hat{k}_t+1},u^{\hat{k}_t+1},p^{\hat{k}_t+1}\right)\leq H\left(y^{\hat{k}_t+1},w,p^{\hat{k}_t+1}\right)+\tilde{\mu}$$

is fulfilled for all $w\in K_U$. Since there are only finitely many elements in F_N, we can take choose $\tilde{n}:=\max\left\{\hat{k}_t|\ t\in F_N\right\}+1$ to obtain the desired statement. □

With the following two examples, we show how to verify (2.32), which is the central assumption for the proof of the convergence result Theorem 14.

Example 17. Consider a one dimensional linear optimal control problem on $[0,T]$ for $u_a\leq u\leq u_b$, $u_a<u_b$, with the cost functional

$$J(y,u):=\frac{1}{2}\|y-y_d\|_{L^2}^2+\frac{\alpha}{2}\|u\|_{L^2}^2,$$

$\alpha>0$, and the constraint $y'=u$ with an initial value. The corresponding Hamiltonian is given by $H(t,y,u,p):=\frac{1}{2}(y-y_d)^2+\frac{\alpha}{2}u^2+pu$. Then the augmented Hamiltonian is given by

$$K_\epsilon(t,y,u,v,p):=\frac{1}{2}(y-y_d)^2+\frac{\alpha}{2}u^2+pu+\epsilon(u-v)^2.$$

In Algorithm 2.1, we have $K_\epsilon\left(t,y^k,u^{k+1},u^k,p^k\right)$ with

$$K_\epsilon\left(t,y^k,u^{k+1},u^k,p^k\right)\leq K_\epsilon\left(t,y^k,w,u^k,p^k\right) \tag{2.39}$$

for all $w\in K_U$. Next, we show that (2.32) holds. For the example, we assume $w\neq u^{k+1}$ because in the case $w=u^{k+1}$, we have that (2.32) is fulfilled with equality. Next, we show that (2.32) also holds for $w\neq u^{k+1}$.

If for the minimum u^{k+1} of the function $w\mapsto K_\epsilon\left(t,y^k,w,u^k,p^k\right)$ it holds that $u_a<u^{k+1}<u_b$, then we have by $\frac{\partial}{\partial u}K_\epsilon\left(t,y^k,u,u^k,p^k\right)|_{u=u^{k+1}}=0$, see [4, VII Theorem 3.13], that

$$p^k=-\alpha u^{k+1}-2\epsilon\left(u^{k+1}-u^k\right). \tag{2.40}$$

From (2.39) we obtain the following

$$\frac{\alpha}{2}\left(u^{k+1}\right)^2 + p^k u^{k+1} + \epsilon\left(u^{k+1}-u^k\right)^2 + r\left(u^{k+1}-w\right)^2 \leq \frac{\alpha}{2}w^2 + p^k w + \epsilon\left(w-u^k\right)^2 \tag{2.41}$$

and with (2.40) it consequently holds that

$$\begin{aligned}&-\frac{\alpha}{2}\left(u^{k+1}\right)^2 - \epsilon\left(u^{k+1}\right)^2 + \epsilon\left(u^k\right)^2 + r\left(u^{k+1}-w\right)^2\\ &\leq \frac{\alpha}{2}w^2 - \frac{\alpha}{2}wu^{k+1} - 2\epsilon wu^{k+1} + 2\epsilon u^k w + \epsilon w^2 - 2\epsilon wu^k + \epsilon\left(u^k\right)^2.\end{aligned}$$

This is equivalent to

$$r\left(u^{k+1}-w\right)^2 \leq \left(\frac{\alpha}{2}+\epsilon\right)\left(u^{k+1}-w\right)^2$$

which is fulfilled for $r \leq \frac{\alpha}{2}+\epsilon$. Finally, if we set $r = \frac{\alpha}{2}+\epsilon$, then $r \geq \epsilon$ in the case of $u_a < u^{k+1} < u_b$.

From (2.40), we have that the minimum of the function $w \mapsto K_\epsilon\left(t, y^k, w, u^k, p^k\right)$ is given by $\frac{2\epsilon u^k - p^k}{\alpha+2\epsilon}$. If $u^{k+1} = u_b$, then we have $u^{k+1} \leq \frac{2\epsilon u^k - p^k}{\alpha+2\epsilon}$ and thus

$$p^k \leq 2\epsilon u^k - (\alpha+2\epsilon)\, u^{k+1} \tag{2.42}$$

and if $u^{k+1} = u_a$, then we have $u^{k+1} \geq \frac{2\epsilon u^k - p^k}{\alpha+2\epsilon}$ and thus

$$p^k \geq 2\epsilon u^k - (\alpha+2\epsilon)\, u^{k+1}. \tag{2.43}$$

From (2.41), we obtain

$$\frac{\alpha}{2}\left(u^{k+1}\right)^2 + p^k\left(u^{k+1}-w\right) + \epsilon\left(u^{k+1}-u^k\right)^2 + r\left(u^{k+1}-w\right)^2 \leq \frac{\alpha}{2}w^2 + \epsilon\left(w-u^k\right)^2. \tag{2.44}$$

Noticing the sign of $\left(u^{k+1}-w\right)$, we increase the left hand-side of (2.44) by inserting the estimation for p^k, (2.42) and (2.43), in both cases. If (2.44) is still fulfilled for $r \geq \epsilon$, then (2.41) is fulfilled in particular. Inequality (2.44) is fulfilled for $r \leq \frac{\alpha}{2}+\epsilon$ and thus setting $r := \frac{\alpha}{2}+\epsilon$, (2.32) is satisfied with $r \geq \epsilon$.

Notice that Example 17 also holds for a bilinear case where the term pu in the augmented Hamiltonian is replaced by pyu and thus p^k is replaced by $p^k y^k$ in the calculations. See also the following example that is performed with a bilinear control problem.

Example 18. In this example, we consider a one dimensional bilinear optimal control problem on $[0,T]$ for $u_a \leq u \leq u_b$, $u_a < 0 < u_b$, with the cost functional

$$J\left(y,u\right) := \frac{1}{2}\|y-y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 + \beta\|u\|_{L^1},$$

$\alpha, \beta \geq 0$, and the constraint $y' = uy$ with an initial value. The corresponding Hamiltonian is given by $H\left(t,y,u,p\right) := \frac{1}{2}\left(y-y_d\right)^2 + \frac{\alpha}{2}u^2 + \beta|u| + puy$. Then the augmented Hamiltonian is given by

$$K_\epsilon\left(t,y,u,v,p\right) := \frac{1}{2}\left(y-y_d\right)^2 + \frac{\alpha}{2}u^2 + \beta|u| + puy + \epsilon\left(u-v\right)^2.$$

In Algorithm 2.1, we have $K_\epsilon\left(t, y^k, u^{k+1}, u^k, p^k\right)$ with $K_\epsilon\left(t, y^k, u^{k+1}, u^k, p^k\right) \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$ for all $w \in K_U$. Now, we show that (2.32) holds. In the following we consider each case $0 < u^{k+1} \leq u_b$, $u_a \leq u^{k+1} < 0$ and $u^{k+1} = 0$ separately.

We start with the case that $0 < u^{k+1} < u_b$. In this case, analogous like in Example 17, we have

$$\alpha u^{k+1} + \beta + p^k y^k + 2\epsilon\left(u^{k+1}-u^k\right) = 0,$$

thus

$$p^k y^k = -\alpha u^{k+1} - \beta - 2\epsilon \left(u^{k+1} - u^k\right) \tag{2.45}$$

and

$$u^{k+1} = \frac{2\epsilon u^k - \beta - p^k y^k}{\alpha + 2\epsilon}. \tag{2.46}$$

We insert (2.45) into (2.32) and obtain that

$$\begin{aligned} &\frac{\alpha}{2}\left(u^{k+1}\right)^2 + \beta u^{k+1} + p^k y^k u^{k+1} + \epsilon \left(u^{k+1} - u^k\right)^2 + r\left(w - u^{k+1}\right)^2 \\ &\leq \frac{\alpha}{2} w^2 + \beta |w| + p^k y^k w + \epsilon \left(w - u^k\right)^2 \end{aligned} \tag{2.47}$$

which equivalently gives by inserting (2.45) the following

$$r\left(w - u^{k+1}\right) \leq \left(\epsilon + \frac{\alpha}{2}\right)\left(w - u^{k+1}\right)^2 + \beta\left(|w| - w\right) \tag{2.48}$$

similar as in Example 17. For $r = \epsilon + \frac{\alpha}{2} \geq \epsilon$, we have that (2.48) is fulfilled since if $w > 0$, then it is $\beta\left(w - w\right) = 0$ and if $w < 0$, then we have that $\beta\left(-w - w\right) \geq 0$.

In the case that $u^{k+1} = u_b$, we have from (2.46) that $u^{k+1} \leq \frac{2\epsilon u^k - \beta - p^k y^k}{\alpha + 2\epsilon}$. Consequently we have that

$$p^k y^k \leq 2\epsilon u^k - \beta - \left(\alpha + 2\epsilon\right) u^{k+1}.$$

Since $w - u^{k+1} = w - u_b \leq 0$ and $p^k y^k$ is replaced by an expression that is greater, the expression $\left(w - u^{k+1}\right) p^k y^k$ is replaced by a term that is smaller. However, it is the same as discussed for (2.47) where $p^k y^k$ is replaced by (2.45). Therefore we obtain again (2.48), which holds for $r = \epsilon + \frac{\alpha}{2}$, and thus (2.47) holds in particular.

The case $u_a < u^{k+1} < 0$ is analogous to the case $0 < u^{k+1} < u_b$ where we have $\beta\left(|w| + w\right) \geq 0$ instead of $\beta\left(|w| - w\right) \geq 0$. Furthermore the same reasoning as in the case $u^{k+1} = u_b$ holds for the case that $u^{k+1} = u_a$ where $u^{k+1} \geq \frac{2\epsilon u^k + \beta - p^k y^k}{\alpha + 2\epsilon}$ and thus

$$p^k y^k \geq 2\epsilon u^k + \beta - \left(\alpha + 2\epsilon\right) u^{k+1}.$$

In contrast to the case above where $u^{k+1} = u_b$, we have that $w - u^{k+1} = w - u_a \geq 0$ and consequently the expression $\left(w - u^{k+1}\right) p^k y^k$ is again replaced by a term that is smaller and the argumentation is analogous.

Next, we have the case where $u^{k+1} = 0$. The fact that $u^{k+1} = 0$ means that

$$K_\epsilon\left(t, y^k, 0, u^{k,} p^k\right) \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right) \tag{2.49}$$

for all $w \in K_U$. Now, we perform a preliminary discussion that shows that the minimum

$$w_1 = \frac{2\epsilon u^k - \beta - p^k y^k}{\alpha + 2\epsilon}$$

of the parabola

$$w \mapsto K_\epsilon\left(t, y^k, w, u^k, p^k\right) = \frac{\alpha}{2} w^2 + \beta w + p^k y^k w + \epsilon\left(w - u^k\right)^2 \tag{2.50}$$

with the property (2.49) is not positive, that means $w_1 \leq 0$ and that the minimum

$$w_2 = \frac{2\epsilon u^k + \beta - p^k y^k}{\alpha + 2\epsilon}$$

of the parabola

$$w \mapsto K_\epsilon\left(t, y^k, w, u^k, p^k\right) = \frac{\alpha}{2} w^2 - \beta w + p^k y^k w + \epsilon\left(w - u^k\right)^2 \tag{2.51}$$

with the property (2.49) is not negative, that means $w_2 \geq 0$.

According to (2.49), the situation corresponding to (2.50) can be translated into $ax^2 + bx + c \geq c$ for all $x > 0$ with $a > 0$ and the situation corresponding to (2.51) can be translated into $ax^2 + bx + c \geq c$ for all $x < 0$ with $a > 0$. We start with the case associated with (2.51). Then we have for the minimum $\tilde{x}$ that it holds $\tilde{x} \leq 0$. This can be seen es follows. The inequality $ax^2 + bx + c \geq c$ for all $x > 0$ is equivalent to $ax + b \geq 0$ for all $x > 0$. The minimum of the function $x \mapsto ax^2 + bx + c$ is characterized by the root $\tilde{x}$ of the first derivative such that it holds $2a\tilde{x} + b = 0$. If we now assumed that $\tilde{x} > 0$, then it would follow that $b < 0$ in order to fulfill this equality. In addition we have that $ax + b \geq 0$ holds for all $x > 0$, that means in particular it has to hold $a\tilde{x} + b \geq 0$, and $2a\tilde{x} + b = 0$ at the same time. Inserting the equation into the inequality provides the contradiction $\frac{b}{2} \geq 0$ to $b < 0$ as discussed before. Analogous for the case (2.51) where it holds that $ax^2 + bx + c \geq c$ for all $x < 0$ with $a > 0$. Then we have that the minimum is not negative.

Inserting the definition of K_ϵ, Inequality (2.49) is given by

$$\epsilon \left(u^k\right)^2 \leq \left(\frac{\alpha}{2} + \epsilon\right) w^2 + \beta|w| + p^k y^k w - 2\epsilon u^k w + \epsilon \left(u^k\right)^2 .$$

Now we conclude from the preliminary discussion in the situation (2.50) with $w > 0$ for the corresponding function, which is given by

$$w \mapsto \left(\frac{\alpha}{2} + \epsilon\right) w^2 + \left(\beta + p^k y^k - 2\epsilon u^k\right) w + \epsilon \left(u^k\right)^2$$

with its minimum $w_1 = \frac{2\epsilon u^k - \beta - p^k y^k}{\alpha + 2\epsilon}$, that it holds

$$\frac{2\epsilon u^k - \beta - p^k y^k}{\alpha + 2\epsilon} \leq 0.$$

Consequently, we have that

$$p^k y^k \geq 2\epsilon u^k - \beta. \tag{2.52}$$

Analogous in the situation (2.51) with $w < 0$ for the function

$$w \mapsto \left(\frac{\alpha}{2} + \epsilon\right) w^2 + \left(-\beta + p^k y^k - 2\epsilon u^k\right) w + \epsilon \left(u^k\right)^2$$

we have that $p^k y^k \leq 2\epsilon u^k + \beta$.

Specifically, we have to show that $K_\epsilon\left(t, y^k, 0, u^{k,} p^k\right) + rw^2 \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$, or equivalently that

$$\epsilon \left(u^k\right)^2 + rw^2 \leq \frac{\alpha}{2} w^2 + \beta|w| + p^k y^k w + \epsilon \left(w - u^k\right)^2 \tag{2.53}$$

is fulfilled for an $r \geq \epsilon$ in order to show (2.32). For the case that $w > 0$, we have with an analogous consideration as above in the present example using (2.52) the following

$$rw^2 \leq \left(\frac{\alpha}{2} + \epsilon\right) w^2 + \beta w + 2\epsilon u^k w - \beta w - 2\epsilon u^k w$$

which is true for $r = \epsilon + \frac{\alpha}{2}$ and thus (2.53) is true in particular. Analogous the case where $w < 0$.

Finally, we conclude that in all cases we can choose $r = \epsilon + \frac{\alpha}{2}$ and thus (2.32) is fulfilled for the considered L^1-cost functional.

Notice that Example 18 also holds in the case of a linear control framework where $y' = u$. For this purpose, the term $p^k y^k$ is replaced by p^k. Further, we remark that the calculation in Example 18 also holds in the case of $\alpha = 0$, that means a pure L^1-functional.

2.4 Optimal quantum control

In this section, which is based on [24], we apply the SQH method to an optimal quantum control problem both with an L^1-cost term and with an L^0-cost term of the control. We start with the L^1-cost term and similar to [32], we consider the problem given by

$$\begin{aligned}\min_{y,u} J(y,u) &:= \frac{1}{2}\sum_{i=1}^{n}\left(y_i(T)-(y_d)_i\right)^2+\frac{\alpha}{2}\|u\|_{L^2(0,T)}^2+\beta\|u\|_{L^1(0,T)}\\ \text{s.t. } y' &= (A+uB)\,y,\ t\in(0,T)\\ y(0) &= y_0\\ u &\in U_{ad}\end{aligned}\tag{2.54}$$

where $A,B\in\mathbb{R}^{n\times n}$ are skew-symmetric matrices, $K_U=[\underline{u},\overline{u}]$ with $\underline{u}=-60$, $\overline{u}=60$ and $U_{ad}\subseteq L^2(0,T)$. According to our framework from Section 2.1 we have that $h(y)=0$, $g(u):=\frac{\alpha}{2}u^2+\beta|u|$, which is lower semi-continuous, $F(y(T)):=\frac{1}{2}\sum_{i=1}^{n}\left(y_i(T)-(y_d)_i\right)^2$, $f(y,u):=(A+uB)\,y$ and $u:[0,T]\to K_U$. A proof of existence of a solution to (2.54) can be found in [32] where the reasoning is similar to the proof of Lemma 21.

Next, we check the Assumptions A.1) to A.6). The functions $y\mapsto 0$, $y(T)\mapsto\frac{1}{2}\sum_{i=1}^{n}\left(y_i(T)-(y_d)_i\right)^2$ and $y\mapsto(A+uB)\,y$ are quadratic or linear, respectively and thus Assumption A.1) is fulfilled. By Remark 1, Assumption A.2) and Assumption A.4) are fulfilled. Assumption A.3) is fulfilled since $u\in L^2(0,T)$ and the unique solution y of the state equation is absolutely continuous, see [90, C.4]. For the remaining two assumptions, we use the boundedness of the state to show them. By Gronwall's inequality, see Lemma 57, we have that each component of y is bounded as follows. We have by the definition of a solution (2.2) that

$$y_i(t)=y_i(0)+\int_0^t\sum_{l=1}^{n}\left(A_{il}+u\left(\tilde{t}\right)B_{il}\right)y_l\left(\tilde{t}\right)d\tilde{t}$$

and by taking the absolute value and summing up over all $i\in\{1,...,n\}$ we obtain

$$\sum_{i=1}^{n}|y_i(t)|\le\sum_{i=1}^{n}|y_i(0)|+\sum_{i=1}^{n}\int_0^t\sum_{l=1}^{n}|\left(A_{il}+u\left(\tilde{t}\right)B_{il}\right)||y_l\left(\tilde{t}\right)|d\tilde{t}.\tag{2.55}$$

Consequently by the boundedness of u we have from (2.55) the following

$$\sum_{i=1}^{n}|y_i(t)|\le c_1+c_2\int_0^T\sum_{i=1}^{n}|y_l\left(\tilde{t}\right)|d\tilde{t}$$

for two constants $c_1,c_2>0$ which provides the desired boundedness result for each component of y, see Lemma 57. Then Assumption A.5) and Assumption A.6) immediately hold. Thus, in the view of Example 18, the theoretical results from Section 2.2 and Section 2.3 are applicable.

In order to validate a numerical solution (y,u,p) from the SQH method for optimality, we define the number

$$\triangle H(t):=\left(H(t,y,u,p)-\min_{w\in K_U}H(t,y,w,p)\right).$$

The number $N_{\%}^{\iota}$ is the part of the grid points in % at which the inequality $0\le\triangle H\le 10^{-\iota}$, $\iota\in\mathbb{N}$, is fulfilled. The parameter for Algorithm 2.1 are chosen as follows. We have $\kappa=10^{-15}$, $\zeta=0.8$, $\sigma=2$, $\eta=10^{-9}$, the initial guess $\epsilon=0.005$ and the control $u^0=0$.

In our experiment, we consider $T = 0.008$, $\alpha = 2^{-9}$ and

$$A = 2\pi \cdot 483 \cdot \begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B = 2\pi \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$y_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad y_d = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

We discretize the interval $[0, T]$ equidistantly with a step size $\triangle t = 10^{-5}$. The state equation and the adjoint equation are solved with a modified Crank-Nicholson scheme [34, Supplementary material, (129)]. The augmented Hamiltonian is given by

$$K_\epsilon\left(t, y, u, v, p\right) := \frac{\alpha}{2} u^2 + \beta |u| + p^T \left(\left(A + uB\right) y\right) + \epsilon \left(u - v\right)^2. \tag{2.56}$$

By a case study, we have that its pointwise minimum is given by

$$u = \underset{\tilde{u} \in \{u_1, u_2\}}{\arg\min}\; K_\epsilon\left(t, y, \tilde{u}, v, p\right)$$

where

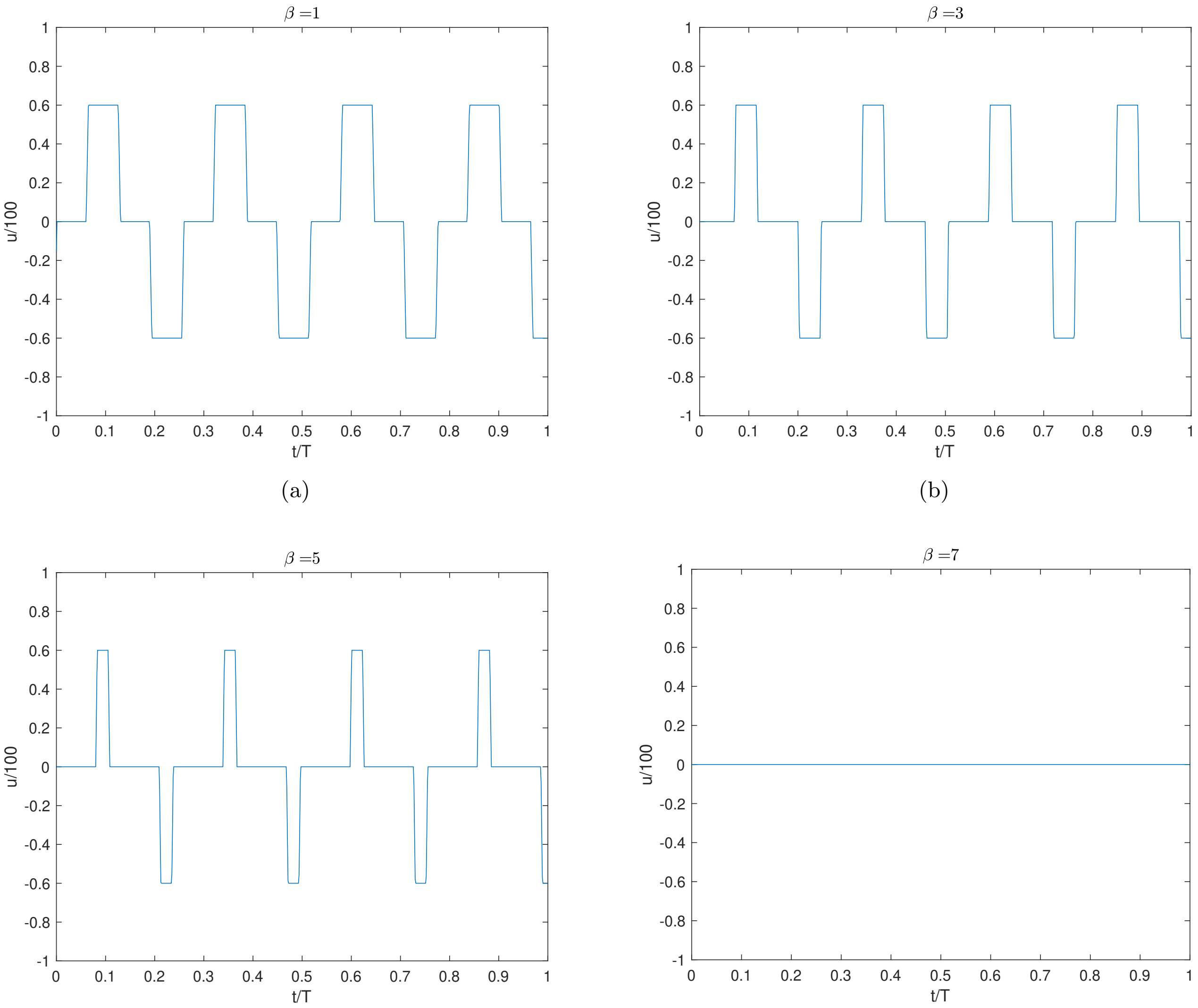
$$u_1 = \min\left(\max\left(0, \frac{2\epsilon v - p^T B y - \beta}{2\epsilon + \alpha}\right), \overline{u}\right)$$

and

$$u_2 = \min\left(\max\left(\underline{u}, \frac{2\epsilon v - p^T B y - \beta}{2\epsilon + \alpha}\right), 0\right)$$

with a similar calculation as that one starting on page 166 in the Appendix. In Figure 2.1, we show the control for different β where we see that the control becomes sparser the higher β is. In Table 2.1 we give the corresponding numerical optimality check, which validates Theorem 14 in the view of Example 18 and give the CPU time for the calculation.

We remark that this numerical experiment with identical $\triangle t$ as above can be performed with the LONE code [33, 32] that is based on a globalized semi-smooth Newton method. The figures corresponding to the experiment depicted in Figure 2.1 look identical, however the calculation time for $\beta = 1$, $\beta = 3$ and $\beta = 5$ ranges between 13 and 46 seconds. The reason for this is that the gradient method that globalizes the semi-smooth Newton method takes a long time until it is sufficiently close to the solution where the semi-smooth Newton method converges. Analogous, as for the solution from the SQH method, we perform the same numerical test with the solution from the LONE code to check for optimality. We have that for $\beta = 1$, $N_{\%}^{2} = 99.75\%$, $N_{\%}^{15} = 86.64\%$, for $\beta = 3$, $N_{\%}^{2} = 99.88\%$, $N_{\%}^{15} = 78.53\%$, for $\beta = 5$, $N_{\%}^{2} = 99.88\%$, $N_{\%}^{15} = 82.90\%$ and for $\beta = 7$, $N_{\%}^{2} = 100\%$, $N_{\%}^{15} = 100\%$ which supports the statement that the solutions from both methods are identical, compare with Table 2.1.



(a)

(b)

(c)

(d)

Figure 2.1: Optimal controls for different β.

β	$\frac{N^{2}_{\%}}{\%}$	$\frac{N^{3}_{\%}}{\%}$	$\frac{N^{4}_{\%}}{\%}$	$\frac{N^{5}_{\%}}{\%}$	$\frac{N^{8}_{\%}}{\%}$	$\frac{N^{15}_{\%}}{\%}$	CPU time/s	# iteration	# update
1	100	99.50	93.63	93.01	93.01	85.89	0.76	55	28
3	100	100	100	99.88	96.13	79.40	0.66	47	25
5	100	100	98.25	96.50	95.51	83.02	0.56	40	18
7	100	100	100	100	100	100	0.02	1	0

Table 2.1: Numerical optimality and CPU time in seconds required for the calculation for different β. The number of total iterations of Algorithm 2.1 is denoted by # iteration and the number of updates on the initial guess is denoted by # update.

Next, we consider an optimal control problem that is similar to (2.54) where the L^1-norm is replaced by the so called L^0-norm. We have

$$\begin{aligned} \min_{y,u} J(y,u) &:= \frac{1}{2}\sum_{i=1}^{n}\left(y_i(T)-(y_d)_i\right)^2+\left(\frac{\alpha}{2}\|u\|^2_{L^2(0,T)}+\beta\|u\|_{L^0(0,T)}\right)\\ \text{s.t. } y' &= (A+uB)\,y,\ t\in(0,T)\\ y(0) &= y_0\\ u &\in U_{ad}\end{aligned} \tag{2.57}$$

where

$$\|u\|_{L^0(0,T)} := \int_0^T |u(t)|_0 dt,\quad |u(t)|_0 := \begin{cases}0 & \text{if } u(t)=0\\ 1 & \text{else}\end{cases}$$

and $K_U := [\underline{u},\overline{u}]$ with $\underline{u} < 0 < \overline{u}$. We assume that there exists a solution to (2.57). The corresponding augmented Hamiltonian is given by

$$K_\epsilon(t,y,u,v,p) := \frac{\alpha}{2}u^2+\beta|u|_0+p^T\left((A+uB)\,y\right)+\epsilon\,(u-v)^2\,.$$

As $|\cdot|_0$ is lower semi-continuous, we can apply the theoretical results from Section 2.2 and Section 2.3 with an analogous consideration as above if we can show that for any iterate u^k, $k\in\mathbb{N}_0$ and for any ϵ chosen by Algorithm 2.1 there exists an $r\geq\epsilon$ such that

$$K_\epsilon\left(t,y^k,u^{k+1},u^k,p^k\right)+r\left(w-u^{k+1}(t)\right)^2\leq K_\epsilon\left(t,y^k,w,u^k,p^k\right) \tag{2.58}$$

is fulfilled for all $w\in K_U$ and for all $t\in[0,T]$. In fact, we can show this if we make a further assumption denoted in the following lemma.

Lemma 19. *We consider the optimal control problem given by (2.57). Let $\alpha>0$ and $\beta\geq 0$ be given. If for all iterations u^k, $k\in\mathbb{N}_0$, generated by Algorithm 2.1 it holds pointwise that either $u^k=0$ or there exists a constant $\tilde{\theta}>0$ with $|u^k|>\tilde{\theta}>0$,*

$$-\frac{\alpha}{2}\tilde{\theta}^2+\beta\leq 0 \tag{2.59}$$

and $\tilde{\theta}\leq\min\left(|\underline{u}|,|\overline{u}|\right)$, then there exists a constant $d>0$ such that (2.58) is fulfilled if $|\underline{u}|,|\overline{u}|\leq d$.

Proof. First we investigate the connection between p^k, u^k and u^{k+1} pointwise. An analogous consideration as in Example 17 or Example 18 where we use that u^{k+1} minimizes $w\mapsto K_\epsilon\left(t,y^k,w,v,p^k\right)$ and thus $0=\frac{\partial}{\partial w}K_\epsilon\left(t,y^k,w,v,p^k\right)$ if $0<u^{k+1}<\overline{u}$ or $\underline{u}<u^{k+1}<0$, provides the following

$$\left(p^k\right)^T By^k=-\alpha u^{k+1}-2\epsilon\left(u^{k+1}-u^k\right). \tag{2.60}$$

Furthermore if $u^{k+1}=\overline{u}$, then

$$\left(p^k\right)^T By^k\leq 2\epsilon u^k-(\alpha+2\epsilon)\,\overline{u} \tag{2.61}$$

and if $u^{k+1}=\underline{u}$, then $\left(p^k\right)^T By^k\geq 2\epsilon u^k-(\alpha+2\epsilon)\,\underline{u}$.

In the case that $u^{k+1}=0$, we have for the value $\hat{u}=\frac{2\epsilon u^k-p^k}{\alpha+2\epsilon}$ where $\tilde{u}\mapsto K\left(t,y^k,\tilde{u},u^k,p^k\right)$ would take its minimum if $\hat{u}\neq 0$, that $K_\epsilon\left(t,y^k,0,u^k,p^k\right)\leq K\left(t,y^k,\frac{2\epsilon u^k-p^k}{\alpha+2\epsilon},u^k,p^k\right)$ which is with $b:=\alpha+2\epsilon$

equivalent to

$$\begin{aligned}
\epsilon\left(u^k\right)^2 &\le \frac{\alpha}{2}\left(\frac{2\epsilon u^k-\left(p^k\right)^T By^k}{\alpha+2\epsilon}\right)^2+\beta+p^k\frac{2\epsilon u^k-\left(p^k\right)^T By^k}{\alpha+2\epsilon}+\epsilon\left(\frac{2\epsilon u^k-\left(p^k\right)^T By^k}{\alpha+2\epsilon}-u^k\right)^2\\
&=\frac{\alpha}{2b^2}\left(4\epsilon^2\left(u^k\right)^2-4\epsilon u^k\left(p^k\right)^T By^k+\left(\left(p^k\right)^T By^k\right)^2\right)+\beta+\frac{2}{b}\epsilon u^k\left(p^k\right)^T By^k\\
&\quad-\frac{1}{b}\left(\left(p^k\right)^T By^k\right)^2+\frac{\epsilon}{b^2}\left(\left(\left(p^k\right)^T By^k\right)^2+2\alpha u^k\left(p^k\right)^T By^k+\alpha^2\left(u^k\right)^2\right)\\
&=\epsilon\left(u^k\right)^2-\frac{1}{2b}\left(\left(p^k\right)^T By^k\right)^2+\frac{2\epsilon u^k}{b}\left(p^k\right)^T By^k-\frac{2\epsilon^2}{b}\left(u^k\right)^2+\beta
\end{aligned}$$

and this in turn is equivalent to

$$0\le-\frac{1}{2}\left(\left(p^k\right)^T By^k\right)^2+2\epsilon u^k\left(p^k\right)^T By^k-2\epsilon^2\left(u^k\right)^2+b\beta. \tag{2.62}$$

From (2.62) we obtain that it holds

$$2\epsilon u^k-\sqrt{2\left(\alpha+2\epsilon\right)\beta}\le\left(p^k\right)^T By^k\le 2\epsilon u^k+\sqrt{2\left(\alpha+2\epsilon\right)\beta}. \tag{2.63}$$

If $\frac{2\epsilon u^k-p^k}{\alpha+2\epsilon}\ge\overline{u}$, then we have $K_\epsilon\left(t,y^k,0,u^k,p^k\right)\le K\left(t,y^k,\overline{u},u^k,p^k\right)$ which is equivalent to

$$0\le-\left(\frac{\alpha}{2}+\epsilon\right)\overline{u}+\frac{\alpha}{2}\overline{u}^2+\beta+\left(p^k\right)^T By^k\overline{u}+\epsilon\overline{u}^2-2\epsilon\overline{u}u^k$$

and in turn

$$\left(p^k\right)^T By^k\ge 2\epsilon u^k-\left(\frac{\alpha}{2}+\epsilon\right)\overline{u}-\frac{\beta}{\overline{u}}. \tag{2.64}$$

Analogous we have for $\frac{2\epsilon u^k-p^k}{\alpha+2\epsilon}\le\underline{u}$ that $K_\epsilon\left(t,y^k,0,u^k,p^k\right)\le K\left(t,y^k,\underline{u},u^k,p^k\right)$ and thus

$$\left(p^k\right)^T By^k\le 2\epsilon u^k-\left(\frac{\alpha}{2}+\epsilon\right)\underline{u}-\frac{\beta}{\underline{u}}.$$

Next we show that for $r=\epsilon$ the condition (2.58) holds for all $w\in K_U$ and $t\in[0,T]$. We start with the case that $u^{k+1}=\overline{u}$. Then we have from

$$K_\epsilon\left(t,y^k,\overline{u},u^k,p^k\right)+r\left(w-\overline{u}\left(t\right)\right)^2\le K_\epsilon\left(t,y^k,w,u^k,p^k\right)$$

that

$$\frac{\alpha}{2}\overline{u}^2+\left(p^k\right)^T By^k\left(\overline{u}-w\right)+\epsilon\left(\overline{u}-u^k\right)^2+\epsilon\left(w-\overline{u}\right)^2\le\frac{\alpha}{2}w^2+|w|_0+\epsilon\left(w-u^k\right)^2. \tag{2.65}$$

As $\overline{u}-w\ge0$, the left hand-side of (2.65) increases if we estimate $\left(p^k\right)^T By^k$ with (2.61). If

$$\frac{\alpha}{2}\overline{u}^2+\left(2\epsilon u^k-\left(\alpha+2\epsilon\right)\overline{u}\right)\left(\overline{u}-w\right)+\epsilon\left(\overline{u}-u^k\right)^2+\epsilon\left(w-\overline{u}\right)^2\le\frac{\alpha}{2}w^2+|w|_0+\epsilon\left(w-u^k\right)^2 \tag{2.66}$$

holds, then (2.58) holds in particular. A direct calculation from (2.66) shows that $0\le\frac{\alpha}{2}\left(w-\overline{u}\right)^2$ if $w\neq0$ which is true and that $0\le\frac{\alpha}{2}\overline{u}^2-\beta$ if $w=0$ which is also true as $|\overline{u}|\ge\tilde{\theta}$. An analogous calculation holds for the case that $u^{k+1}=\underline{u}$.

Also analogous for the case that $0 < u^{k+1} < \overline{u}$ or $\underline{u} < u^{k+1} < 0$ we have the following. By (2.60), we obtain from

$$K_\epsilon\left(t, y^k, u^{k+1}, u^k, p^k\right) + r\left(w - u^{k+1}(t)\right)^2 \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$$

with $r = \epsilon$ that

$$-\frac{\alpha}{2}\left(u^{k+1}\right)^2 + \beta \leq \frac{\alpha}{2}w^2 + \beta|w|_0 - \alpha u^{k+1} w$$

which is true in the case $w \neq 0$ and if $w = 0$ then it is true by our assumption as $|u^{k+1}| > \tilde{\theta}$ with $-\frac{\alpha}{2}\tilde{\theta}^2 + \beta \leq 0$.

Finally, we consider the case that $u^{k+1} = 0$. Then from

$$K_\epsilon\left(t, y^k, 0, u^k, p^k\right) + rw^2 \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$$

for $r = \epsilon$ we obtain the following

$$0 \leq \frac{\alpha}{2}w^2 + \beta|w|_0 + \left(p^k\right)^T B y^k w - 2\epsilon u^k w \tag{2.67}$$

which is true for $w = 0$. If $w > 0$ there are two sub cases. Either

$$2\epsilon u^k - \sqrt{2(\alpha + 2\epsilon)\beta} \leq \left(p^k\right)^T B y^k,$$

see (2.63) or

$$2\epsilon u^k - \left(\frac{\alpha}{2} + \epsilon\right)\overline{u} - \frac{\beta}{\overline{u}} \leq \left(p^k\right)^T B y^k,$$

see (2.64). In the first case we have that $0 \leq \frac{\alpha}{2}w^2 + \beta - \sqrt{2(\alpha + 2\epsilon)\beta}w$ which is true if

$$w \leq \frac{\sqrt{2(\alpha + 2\epsilon)\beta} - \sqrt{2(\alpha + 2\epsilon)\beta - 2\alpha\beta}}{\alpha} = \frac{\sqrt{2(\alpha + 2\epsilon)\beta} - \sqrt{4\epsilon\beta}}{\alpha}.$$

The function

$$\epsilon \mapsto \frac{\sqrt{2(\alpha + 2\epsilon)\beta} - \sqrt{4\epsilon\beta}}{\alpha} = \frac{2\beta}{\sqrt{2(\alpha + 2\epsilon)\beta} + \sqrt{4\epsilon\beta}}$$

is monotonically decreasing. We have that ϵ is bounded from above by $\sigma(\eta + \theta)$ due to (2.30) and the definition of Step 4 in Algorithm 2.1 and thus this function takes a minimum with the value $\frac{2\beta}{\sqrt{2(\alpha+2\overline{\epsilon})\beta}+\sqrt{4\overline{\epsilon}\beta}}$ where we denote this upper bound of ϵ by $\overline{\epsilon}$. In the second case we have from (2.67) that

$$K_\epsilon\left(t, y^k, 0, u^k, p^k\right) + rw^2 \leq K_\epsilon\left(t, y^k, w, u^k, p^k\right)$$

with $r = \epsilon$ which is true if

$$w \leq \frac{\frac{\alpha}{2}\overline{u} + \epsilon\overline{u} + \frac{\beta}{\overline{u}} - \sqrt{\left(\frac{\alpha}{2}\overline{u} + \epsilon\overline{u} + \frac{\beta}{\overline{u}}\right)^2 - 2\alpha\beta}}{\alpha} = \frac{2\beta}{\frac{\alpha}{2}\overline{u} + \epsilon\overline{u} + \frac{\beta}{\overline{u}} + \sqrt{\left(\frac{\alpha}{2}\overline{u} + \epsilon\overline{u} + \frac{\beta}{\overline{u}}\right)^2 - 2\alpha\beta}}.$$

We remark that if $\left(\frac{\alpha}{2}\overline{u} + \epsilon\overline{u} + \frac{\beta}{\overline{u}}\right)^2 - 2\alpha\beta < 0$, then this case has no restriction to w. Analogously to the case above, we have that the smallest value is taken at $\overline{\epsilon}$ with the value

$$\frac{2\beta}{\frac{\alpha}{2}\overline{u} + \overline{\epsilon}\overline{u} + \frac{\beta}{\overline{u}} + \sqrt{\left(\frac{\alpha}{2}\overline{u} + \overline{\epsilon}\overline{u} + \frac{\beta}{\overline{u}}\right)^2 - 2\alpha\beta}}.$$

Both cases are fulfilled if we choose

$$w \leq d := \min\left(\frac{2\beta}{\sqrt{2\left(\alpha+2\overline{\epsilon}\right)\beta}+\sqrt{4\overline{\epsilon}\beta}}, \frac{2\beta}{\frac{\alpha}{2}\overline{u}+\overline{\epsilon u}+\frac{\beta}{\overline{u}}+\sqrt{\left(\frac{\alpha}{2}\overline{u}+\overline{\epsilon u}+\frac{\beta}{\overline{u}}\right)^2-2\alpha\beta}}\right).$$

Analogously for $w < 0$ where we have that $w \geq -d$. Consequently the statement of the lemma is proved if $K_U \subseteq [-d, d]$. □

For the numerical discussion we take the same example as above for (2.54), that means we solve (2.57) with the same T, $\overline{u}$, $\underline{u}$, A, B, y_0 and y_d and the same parameters for the SQH method $\kappa = 10^{-15}$, $\zeta = 0.8$, $\sigma = 2$, $\eta = 10^{-9}$, the initial guess $\epsilon = 0.005$ and the control $u^0 = 0$. The pointwise minimum of the augmented Hamilton is given by

$$u = \underset{\tilde{u}\in\{0,u_1,u_2\}}{\arg\min}\ K_\epsilon\left(t, y, \tilde{u}, v, p\right)$$

where

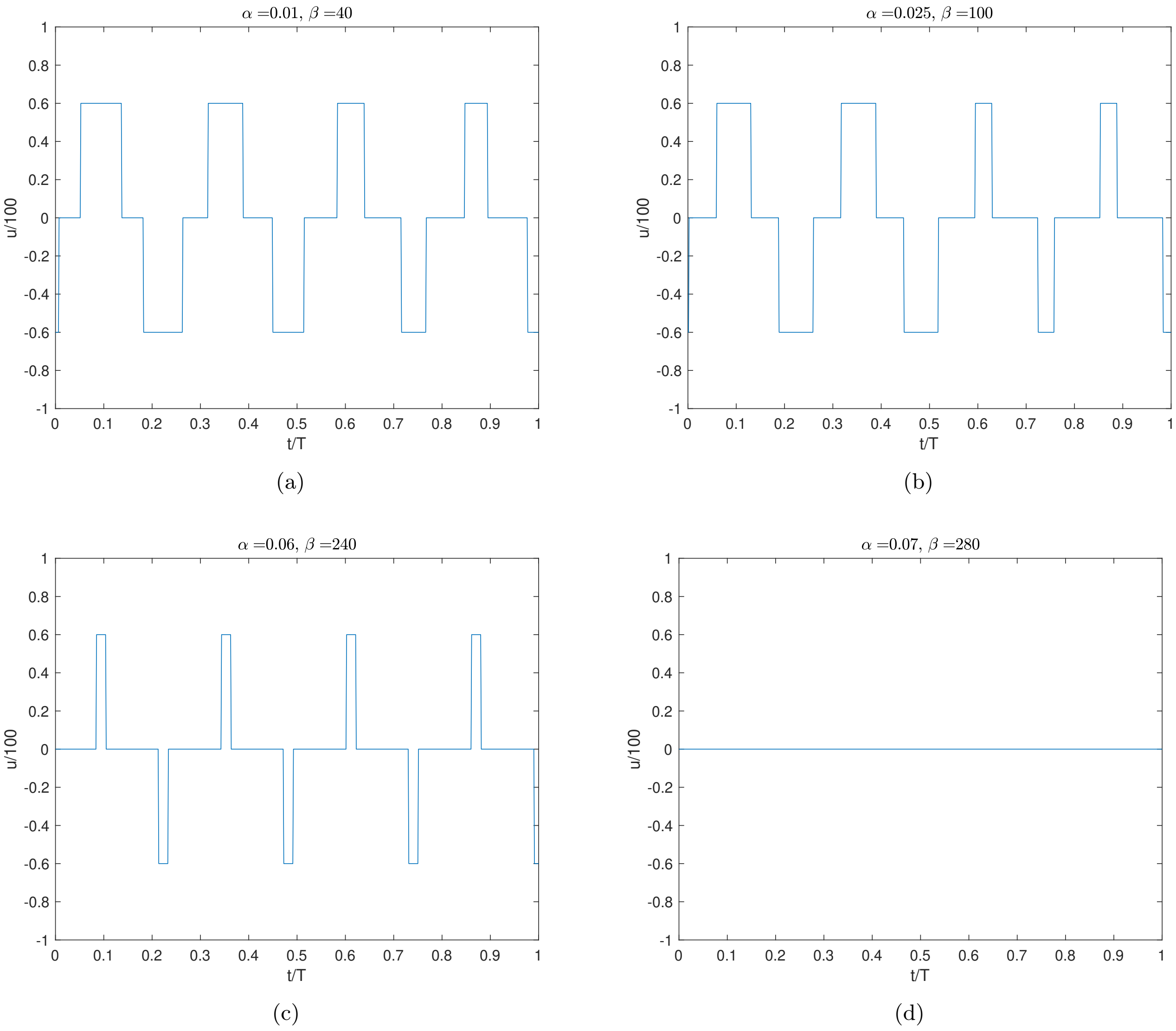
$$u_1 = \min\left(\max\left(0, \frac{2\epsilon v - p^T B y}{\alpha+2\epsilon}\right), \overline{u}\right)$$

and

$$u_2 = \min\left(\max\left(\underline{u}, \frac{2\epsilon v - p^T B y}{\alpha+2\epsilon}\right), 0\right).$$

We have that u determined in Step 2 of Algorithm 2.1 is Lebesgue measurable, see the discussion in the Appendix starting on page 166.

We discretize $[0, T]$ with subintervals of size $\triangle t = 10^{-5}$ for our experiment where we solve (2.57) for fixed $\tilde{\alpha}$ and $\tilde{\beta}$ that are multiplied by a constant $\varkappa > 0$ such that $\alpha = \varkappa\tilde{\alpha}$ and $\beta = \varkappa\tilde{\beta}$. This simultaneous alteration of α and β is motivated by (2.59) where α is supposed to become not too small compared with β. We choose $\tilde{\alpha} = 5 \cdot 10^{-3}$ and $\tilde{\beta} = 20$. In Figure 2.2 we show the results of the SQH method for different α and β and in Table 2.2 we give the corresponding numerical optimality and the calculation time. This experiment demonstrates that for appropriately chosen α and β a fast convergence is achieved while still sparse solutions are provided. Furthermore, we see that the assumption from Lemma 19 that each iterate of Algorithm 2.1 is pointwise bounded away from zero by a constant, if not zero, seems to be justified and reasonable.



(a)

(b)

(c)

(d)

Figure 2.2: Optimal controls for different α and β.

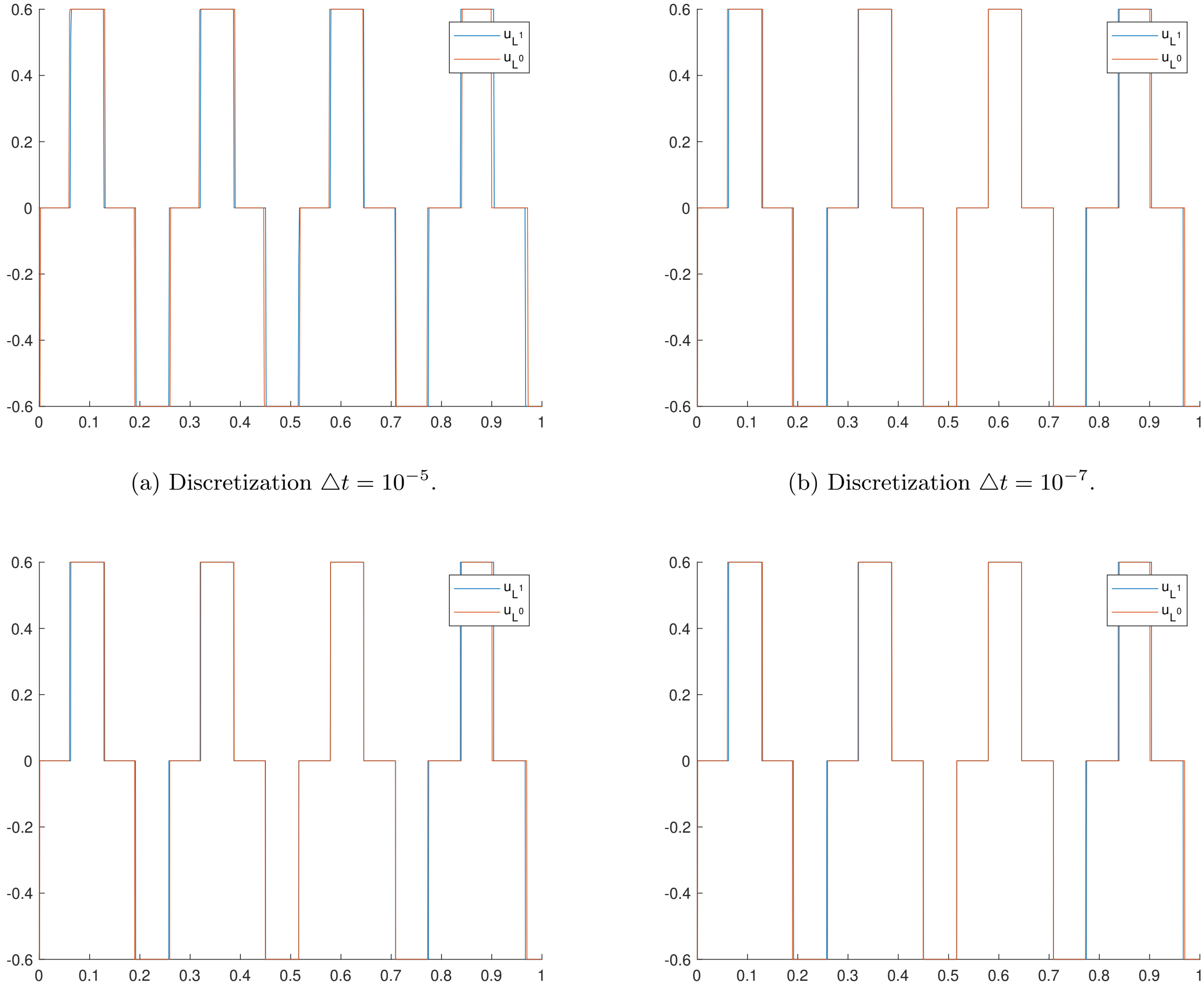
$\varkappa$	$\frac{N^2_{\%}}{\%}$	$\frac{N^3_{\%}}{\%}$	$\frac{N^4_{\%}}{\%}$	$\frac{N^5_{\%}}{\%}$	$\frac{N^8_{\%}}{\%}$	$\frac{N^{15}_{\%}}{\%}$	CPU time/s	# iteration	# update
2	87.39	87.39	87.39	87.39	87.39	87.39	0.40	26	19
5	85.14	85.14	85.14	85.14	85.14	85.14	0.34	22	16
12	85.02	85.02	85.02	85.02	85.02	85.02	0.05	3	2
14	100	100	100	100	100	100	0.02	1	0

Table 2.2: Numerical optimality and CPU time in seconds required for the calculation of a solution to (2.57) for different $\varkappa$ where $\alpha = \varkappa \cdot 5 \cdot 10^{-3}$ and $\beta = \varkappa \cdot 20$. The number of total iterations of Algorithm 2.1 is denoted by # iteration and the number of updates on the initial guess is denoted by # update.

In the next experiment, we demonstrate that both optimal control problems (2.54) and (2.57) provide the same controls that steer the quantum system to a state with an identical distance to the desired state. For this purpose, we compare the solution to (2.54) with the one to (2.57) where the weights of the controls' cost terms are chosen such that each state has the same distance to the desired final state according to

$\frac{1}{2}\sum_{i=1}^{n}\left(y_i\left(T\right)-\left(y_d\right)_i\right)^2=0.0267$ for a discretization $\triangle t=10^{-7}$. This provides solutions for both optimal control problems (2.54) and (2.57) that steer the quantum control system equally close to the desired state what we use as our reference in this experiment. We choose $\alpha=0$ and $\beta=1$ in (2.54) (referred to as L^1-problem with L^1-solution) and $\alpha=1.47\cdot10^{-2}$ and $\beta=42$ in (2.57) (referred to as L^0-problem with L^0-solution). The rest of the parameters is set as discussed before. In this experiment, we compare the L^1-solution with the L^0-solution. For this purpose, we plot the corresponding solutions in one figure and consider their numerical optimality and CPU time that is needed for the calculation. In Figure 2.3 and Table 2.3 we can see the results for different discretizations. We have that the solutions equal each other. However, the calculation time for the solution to (2.57) is only almost about a third of the time that is needed for the calculation of the solution to (2.54) in the cases with a very fine discretization (from $\triangle t=10^{-7}$). Consequently from a numerical point of view it can be advantageous to use an L^0-problem for the calculation of sparse solutions since the SQH method converges faster while providing identical results compared to the ones from the corresponding L^1-problem which includes to obtain each a state whose distance to the desired state is the same.

We remark that instead of solving a pure L^1-problem we can increase the weight of the L^2-cost term α and decrease the weight of the L^1-cost term such that $\frac{1}{2}\sum_{i=1}^{n}\left(y_i\left(T\right)-\left(y_d\right)_i\right)^2=0.0267$ is still fulfilled and the solutions have a bang-bang structure to obtain a faster convergence for the L^1-problem. Although this provides an improvement for the L^1-problem, the L^0-problem can be solved much faster.



(a) Discretization $\triangle t = 10^{-5}$.

(b) Discretization $\triangle t = 10^{-7}$.

(c) Discretization $\triangle t = 10^{-8}$.

(d) Discretization $\triangle t = 10^{-9}$.

Figure 2.3: Optimal controls for different discretizations where the time is on the abscissa and the value of the control is on the ordinate.

$\triangle t$		CPU-time/s	$\frac{N^{2}_{\%}}{\%}$	$\frac{N^{15}_{\%}}{\%}$	# iteration	# update
10^{-5}	L^1	0.87	100	93.13	59	36
	L^0	0.70	84.89	84.89	49	27
10^{-6}	L^1	17.1	100	93.24	116	84
	L^0	7.4	82.83	82.81	51	35
10^{-7}	L^1	116.7	100	93.11	78	62
	L^0	42.8	90.78	90.76	29	22
10^{-8}	L^1	1274.8	100	93.07	85	68
	L^0	427.2	91.07	91.06	29	22
10^{-9}	L^1	13841.8	100	93.07	93	75
	L^0	4261.5	91.07	91.06	29	22

Table 2.3: Comparison of the computation time and the numerical optimality of the solution to the L^1-problem with the solution to the L^0-problem. The number of total iterations of Algorithm 2.1 is denoted by # iteration and the number of updates on the initial guess is denoted by # update.

2.5 Optimal tumor therapy

In this section, we investigate a non-linear model of ODEs modeling tumor growth and check that the corresponding optimal control problem fulfills the Assumptions A.1) to A.6). This section is based on [47]. We start our discussion by illustrating some modeling issues.

We investigate a mathematical model for cancer development and treatment resulting from a combination of two complementary mathematical models. Both models consider the dynamics between the tumor volume p and the carrying capacity q. One of the most commonly used models for tumor growth is based on the Gompertzian growth law as follows

$$\dot{p} = p\left(a - \xi \log\left(p\right)\right), \qquad\qquad a > \xi > 0.$$

While the proliferation rate a of the cells is constant, the death rate $\xi \log(p)$ increases with a growing tumor volume p. The value q is given by

$$q = \exp\left(\frac{a}{\xi}\right).$$

The carrying capacity is a measure of how much tissue of the tumor is sufficiently vascularized such that its cells are well supplied with nutrients and oxygen. Using this normalized carrying capacity, we obtain the following relation for the tumor growth and the carrying capacity

$$\dot{p} = \xi p\left(\tfrac{a}{\xi} - \ln\left(p\right)\right) = \xi p\left[\ln\left(\exp\left(\tfrac{a}{\xi}\right)\right) - \ln\left(p\right)\right] = -\xi p \ln\left(\tfrac{p}{q}\right). \tag{2.68}$$

For $p < q$ the tumor grows ($\dot{p} > 0$) until $p = q$. For $p > q$, the tumor shrinks ($\dot{p} < 0$) again until $p = q$ is reached. See [3, IV Theorem 2.5] for the connection of the derivative of a function and its growth behavior.

Next, we switch from a constant carrying capacity to a time-varying carrying capacity q. While the equation for the tumor growth (2.68) stays the same, we have to develop an expression for the dynamics of the carrying capacity. The basic idea is a combination of stimulatory (T_S) and inhibitory (T_I) effects as follows

$$\dot{q} = T_S\left(p, q\right) - T_I\left(p, q\right).$$

A modeling issue is the choice of T_S and T_I. For this reason, we consider the model proposed in [51] as follows

$$\dot{q} = bp - dp^{\frac{2}{3}}q \tag{2.69}$$

with the birth rate $b > 0$ and the death rate $d > 0$. This is a well-recognized mathematical model for time-varying carrying capacity. However, it couples the tumor volume variable to the carrying capacity.

On the other hand, a model of time-varying carrying capacity that does not involve the tumor volume explicitly is described in [44]. This model is computationally convenient since p does not appear in the equation. We have

$$\dot{q} = bq^{\frac{2}{3}} - dq^{\frac{4}{3}}. \tag{2.70}$$

Based on validation with real data [44, 51], both models appear promising in the quest of an accurate description of tumor growth. For this reason, we consider a combination of the two models (2.69) and (2.70) as follows

$$\dot{q} = \varkappa \left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + (1-\varkappa)\left(bp - dp^{\frac{2}{3}}q\right)$$

where $\varkappa \in (0,1)$. Together with the equation for the tumor growth (2.68), we obtain the following differential system that models the evolution of the tumor volume and of the carrying capacity of the vasculature. We have

$$\begin{aligned} \dot{p} &= -\xi p \ln\left(\frac{p}{q}\right) \\ \dot{q} &= \varkappa \left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + (1-\varkappa)\left(bp - dp^{\frac{2}{3}}q\right). \end{aligned} \tag{2.71}$$

Next, we introduce two control mechanisms in (2.71) that represent the treatment of cancer by anti-angiogenesis and radiotherapy, respectively [86].

In our case angiogenesis is a process where a growing tumor develops its own blood vessels, which provide the tumor with oxygen and nutrients. The anti-angiogenesis therapy is an indirect treatment since it does not fight the tumor cells directly but influences the tumor's micro-environment, in particular the vasculature. The lack of oxygen and nutrients will force the tumor to shrink.

To model this treatment, we introduce a control u that takes its values in

$$K_U := [0, \overline{u}],$$

$\overline{u} > 0$, and represents the dose of the anti-angiogenic medicine. With the anti-angiogenic elimination parameter $\gamma > 0$, we can augment the equation for q in (2.71) as follows

$$\dot{q} = \varkappa \left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + (1-\varkappa)\left(bp - dp^{\frac{2}{3}}q\right) - \gamma q u.$$

Hence, our model for an anti-angiogenetic mono-therapy is given by

$$\dot{p} = -\xi p \ln\left(\frac{p}{q}\right), \quad \dot{q} = \varkappa \left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + (1-\varkappa)\left(bp - dp^{\frac{2}{3}}q\right) - \gamma q u. \tag{2.72}$$

The anti-angiogenic treatment influences the carrying capacity of the vasculature q, but as q appears in the equation for p, it also influences the tumor volume p.

Radiotherapy is a treatment that uses ionizing radiation to kill cancer cells. To model this treatment, we introduce the control w, which represents the dose of radiation and takes its values in

$$K_W := [0, \overline{w}],$$

$\overline{w} > 0$. Following a model described in [98], the damage that is done to the tumor by radiation is modeled as follows

$$-p(t)\left(\alpha + \beta \int_0^t w(s)\, e^{-\rho(t-s)} ds\right) w(t) \tag{2.73}$$

with the radiosensitive parameters $\alpha, \beta > 0$ depending on the treated tissue and the tissue repair rate $\rho > 0$. To simplify the expression above, we introduce the function

$$r(t) := \int_0^t w(s)\, e^{-\rho(t-s)} ds.$$

This is the solution to a linear ODE given by

$$\dot{r} = -\rho r + w, \quad r(0) = 0$$

which is obtained by the Leibniz rule for parameter integrals [4, VII Corollary 6.8]. Hence, from (2.73) the term that quantifies the damage done to the tumor can be written as follows

$$-(\alpha + \beta r)\, pw.$$

Now, we have to take into account that the radiation has also a damaging effect on the healthy tissues. Specifically, the damage on the carrying capacity of the vasculature q is given by

$$-(\varsigma + \delta r)\, qw.$$

Notice that the radiosensitive parameters $\varsigma, \delta > 0$ have different values, because malignant and healthy tissues have different characteristics.

Summarizing, our controlled model of cancer's development and treatment is given by

$$\begin{aligned} \dot{p} &= -\xi p \ln\left(\tfrac{p}{q}\right) - (\alpha + \beta r)\, pw \\ \dot{q} &= \varkappa \left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + (1 - \varkappa)\left(bp - dp^{\frac{2}{3}} q\right) - \gamma q u - (\varsigma + \delta r)\, qw \cdot \\ \dot{r} &= -\rho r + w \end{aligned} \tag{2.74}$$

This model is completely specified by giving the values of the parameters appearing in it. These values are specified in Table 2.4, see [11] and [44].

	Description	**Value**	**Unit**
ξ	Parameter for tumor growth	0.084	$[\mathrm{day}^{-1}]$
b	Tumor-induced stimulation parameter	5.85	$[\mathrm{day}^{-1}]$
d	Tumor-induced inhibition parameter	0.00873	$[\mathrm{mm}^{-2}\mathrm{day}^{-1}]$
γ	Antiangiogenic elimination parameter	0.15	$[\frac{\mathrm{kg}}{\mathrm{mg(dose)}}]\mathrm{day}^{-1}$
α	Radiosensitive parameter for tumor	0.7	$[\mathrm{Gy}^{-1}]$
β	Radiosensitive parameter for tumor	0.14	$[\mathrm{Gy}^{-2}]$
ς	Radiosensitive parameter for healthy tissue	0.136	$[\mathrm{Gy}^{-1}]$
δ	Radiosensitive parameter for healthy tissue	0.086	$[\mathrm{Gy}^{-2}]$
θ	Healthy tissue parameter	0.5	$[\mathrm{day}^{-1}]$
ρ	Tumor repair rate	$\frac{\ln(2)}{0.02}$	$[\mathrm{day}^{-1}]$

Table 2.4: Model parameters.

Usually, in the context of optimal control of cancer development models, the objective of the control is to minimize the volume of the tumor at final time, that means $p(T)$. See [86] for a detailed discussion of this setting.

Now, while we keep this objective, we introduce additional terms that include a reduction of the tumor volume p along the time evolution and L^2- and L^1-norms of the controls u and w. With respect to the side effects of anti-angiogenetic medicine and radiotherapy, it is reasonable to have penalty terms for the corresponding controls.

We define our optimal control problem with anti-angiogenesis and radiotherapy as follows

$$\begin{aligned} &\min_{p,q,r,u,w} J\left(\left(p,q,r\right),\left(u,w\right)\right) \\ &:= \frac{\tilde{\sigma}}{2}\int_0^T p^2\left(t\right)dt + \frac{\vartheta}{2}p^2\left(T\right) + \frac{\nu_u}{2}\|u\|^2_{L^2(0,T)} + \mu_u\|u\|_{L^1(0,T)} + \frac{\nu_w}{2}\|w\|^2_{L^2(0,T)} + \mu_w\|w\|_{L^1(0,T)} \end{aligned} \tag{2.75}$$

subject to

$$\begin{aligned} \dot{p}\left(t\right) &= -\xi p\left(t\right)\ln\left(\tfrac{p(t)}{q(t)}\right) - \left(\alpha+\beta r\left(t\right)\right)pw\left(t\right) \\ \dot{q}\left(t\right) &= \varkappa\left(bq^{\frac{2}{3}}\left(t\right) - dq^{\frac{4}{3}}\left(t\right)\right) + \left(1-\varkappa\right)\left(bp\left(t\right) - dp^{\frac{2}{3}}\left(t\right)q\left(t\right)\right) - \gamma q\left(t\right)u\left(t\right) - \left(\varsigma+\delta r\left(t\right)\right)q\left(t\right)w\left(t\right) \\ \dot{r}\left(t\right) &= -\rho r\left(t\right) + w\left(t\right) \end{aligned} \tag{2.76}$$

with the initial conditions

$$p(0) = p_0 > 0, \quad q(0) = q_0 > 0, \quad r(0) = 0$$

and the functions $u, w \in L^2\left(0,T\right)$ take their values in K_U and K_W, respectively. The admissible sets of controls are defined by

$$U_{ad} := \left\{u \in L^2\left(0,T\right) \mid u\left(t\right) \in K_U \text{ a.e.}\right\}, \quad W_{ad} := \left\{w \in L^2\left(0,T\right) \mid w\left(t\right) \in K_W \text{ a.e.}\right\}. \tag{2.77}$$

The parameters $\tilde{\sigma}, \vartheta, \nu_u, \mu_u, \nu_w, \mu_w \geq 0$ can be chosen differently to obtain different settings. In the following we drop the arguments of the function in order to simplify notation.

In the next step, we analyze the system of ODEs (2.76) with respect to the existence of a unique global solution and its boundedness. Subsequently we discuss the existence of a solution to (2.75) subject to (2.76).

For the investigation of (2.76) with respect to a unique global solution we define the set

$$I := \left\{(p,q,r) \in \mathbb{R}^3 \middle| \ \underline{p} \leq p \leq \overline{p}, \underline{q} \leq q \leq \overline{q}, 0 \leq r \leq \overline{r}\right\}$$

where

$$\underline{p} := \min\left(p_0, \underline{q}\exp\left(-\frac{\left(\alpha+\beta\overline{r}\right)\overline{w}}{\xi}\right)\right) > 0,$$

$$\overline{p} := \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}, q_0, p_0\right),$$

$$\underline{q} := \min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}}, q_0\right)\exp\left(\left(\gamma\overline{u} + \left(\varsigma+\delta\overline{r}\right)\overline{w} + \left(1-\varkappa\right)d\overline{p}^{\frac{2}{3}}\right)T\right) > 0,$$

$$\overline{q} := \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}, q_0\right)$$

and $\overline{r} := \frac{\overline{w}}{\rho}$.

In the following lemma we prove existence of a unique global solution to (2.76).

Lemma 20. *There exists a unique global solution $(p,q,q):[0,T]\to\mathbb{R}$, $t\mapsto(p(t),q(t),r(t))$ to (2.76) in the sense of (2.2) for any $u\in U_{ad}$ and $w\in W_{ad}$. Furthermore, the solution stays in I and thus any component of the solution (p,q,r) is bounded by a constant that holds for all $u\in U_{ad}$ and $w\in W_{ad}$, defined in (2.77).*

Proof. According to [90, Proposition C.3.6] there exists a unique global solution $(p,q,r):[0,T]\to\mathbb{R}^3$, $t\mapsto(p(t),q(t),r(t))$ to (2.76) if we prove that for the corresponding fixed initial values $p_0,q_0>0$, $r(0)=0$, the local solution stays in I on its maximum existence interval. For the proof, we define the right hand-side of (2.76) by the function $f:[0,T]\times(0,\infty)^2\times\mathbb{R}\to\mathbb{R}^3$, $(t,p,q,r)\mapsto f(t,p,q,r)$ as follows

$$f(t,p,q,r):=\begin{pmatrix}-\xi p\ln\left(\frac{p}{q}\right)-(\alpha+\beta r)\,pw\\ \varkappa\left(bq^{\frac{2}{3}}-dq^{\frac{4}{3}}\right)+(1-\varkappa)\left(bp-dp^{\frac{2}{3}}q\right)-\gamma qu-(\varsigma+\delta r)\,qw\\ -\rho r+w\end{pmatrix}\tag{2.78}$$

for any fixed controls $u\in U_{ad}$ and $w\in W_{ad}$. In the following we prove that the global solution $(p,q,r):[0,T]\to\mathbb{R}^3$, $t\mapsto(p(t),q(t),r(t))$ to (2.76) exists and stays in I for any chosen control $u\in U_{ad}$ and $w\in W_{ad}$ in the right hand-side f.

We have that the function $f(\cdot,p,q,r):[0,T]\to\mathbb{R}^3$ is measurable for each fixed $(p,q,r)\in(0,\infty)^2\times\mathbb{R}$, the function $f(t,\cdot):(0,\infty)^2\times\mathbb{R}\to\mathbb{R}^3$ is continuous for each fixed $t\in[0,T]$. The function $f(t,\cdot):(0,\infty)^2\times\mathbb{R}\to\mathbb{R}^3$ is locally Lipschitz continuous for any $(p,q,r)\in(0,\infty)^2\times\mathbb{R}$, which means that there is a $\varrho>0$ with a ball $B_\varrho(p,q,r)$ centered at (p,q,r) such that the function $f(t,\cdot):B_\varrho(p,q,r)\to\mathbb{R}^3$ is Lipschitz continuous with a Lipschitz constant $\tilde{L}>0$ possibly depending on (p,q,r) for each $t\in[0,T]$, due to the continuous differentiability of $f(t,\cdot)$ in an environment of (p,q,r) and the pointwise boundedness of the controls u,w. Since each component $t\mapsto f_i(t,p_0,q_0,0)$, $i\in\{1,2,3\}$ is integrable on $[0,T]$, since the controls are integrable, we have that there exists a unique local solution for any initial value $(p_0,q_0,0)$ with a maximum existence interval $\left[0,\hat{t}\right)$, $\hat{t}>0$, see [90, Theorem 54].

Now we show that the local solution stays in I on its maximum existence interval $\left[0,\hat{t}\right)$, $\hat{t}>0$. We start to investigate f_3. Since $-\rho r+w$ is a linear inhomogeneous equation, we have, according to [90, page 487], that there is a global and unique solution on $[0,T]$. According to [3, IV Theorem 2.5 (i)], the function r grows if and only if $-\rho r+w\geq 0$ which is equivalent to $r\leq\frac{w}{\rho}$. That means the biggest value for r is $\frac{\overline{w}}{\rho}$ and since $r(0)=0$ and $w\geq 0$, we have that $0\leq r(t)\leq\frac{\overline{w}}{\rho}$ for all $[0,T]$.

The same reasoning holds for the first equation f_1. However, before we discuss this, we remark that there is an interval $\left[0,\tilde{t}\right)\subseteq\left[0,\hat{t}\right)$, $\tilde{t}>0$ where the functions $p(t),q(t)>0$, $t\in\left[0,\tilde{t}\right)$, for the following reason. Since for the initial values it holds that $p_0,q_0>0$, we have by the continuity of p and q that for example for $\frac{\min(p_0,q_0)}{2}$, there is a distance $\tilde{t}>0$ from zero such that it holds $p(t),q(t)>0$ for all $t\in\left[0,\tilde{t}\right)$. The following discussion holds for the interval $\left[0,\tilde{t}\right)$ where we have that $p,q>0$.

For $p_0>0$, the function p decrease if and only if $-\xi p\ln\left(\frac{p}{q}\right)-(\alpha+\beta r)\,pw\leq 0$. Since $p,q>0$, we have that

$$\ln\left(\frac{p}{q}\right)\geq-\frac{(\alpha+\beta r)\,w}{\xi}$$

from which it follows that

$$p\geq q\exp\left(-\frac{(\alpha+\beta r)\,w}{\xi}\right)\geq q\exp\left(-\frac{(\alpha+\beta\overline{r})\,\overline{w}}{\xi}\right).$$

That means if p is smaller than $q\exp\left(-\frac{(\alpha+\beta\overline{r})\overline{w}}{\xi}\right)>0$, the function p always increases and thus

$$p>\min\left(p_0,q\exp\left(-\frac{(\alpha+\beta\overline{r})\,\overline{w}}{\xi}\right)\right)>0.\tag{2.79}$$

In the other direction, the function p can at most increase if $p \leq q$. If it holds that $p > q$, then p can only decrease and thus $p \leq p_0$.

Now we are able to estimate p depending on q and consequently we investigate f_2 next. We have that $bq^{\frac{2}{3}} - dq^{\frac{4}{3}} = q^{\frac{2}{3}}\left(b - dq^{\frac{2}{3}}\right) \geq 0$, equivalently $b - dq^{\frac{2}{3}} \geq 0$ since $q^{\frac{2}{3}} \geq 0$ for all $q \in \mathbb{R}$, for

$$q \leq \left(\frac{b}{d}\right)^{\frac{3}{2}}.$$

For the second term we have that $bp - dp^{\frac{2}{3}}q \geq 0$ for $q \leq \frac{b}{d}p^{\frac{1}{3}}$ since $p > 0$. From the investigation of f_1, we have that $p \leq q$ or $p \leq p_0$ and thus we have that the biggest value that can be chosen for p is $\max\left(p_0, q\right)$. Thus we have $q \leq \frac{b}{d}p^{\frac{1}{3}} = \frac{b}{d}q^{\frac{1}{3}}$ or $q \leq \frac{b}{d}p^{\frac{1}{3}} = \frac{b}{d}p_0^{\frac{1}{3}}$. For the growth of q, we must at least have that one of the two terms, $bq^{\frac{2}{3}} - dq^{\frac{4}{3}}$ or $bp - dp^{\frac{2}{3}}q$, is greater than zero. According to our calculation, we have that for $q > \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}\right)$ both terms are negative and cause a decreasing of q. Thus $\max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}\right)$ is the biggest value for q that we can observe if $q_0 \leq \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}\right)$ and if $q_0 > \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}\right)$, then q_0 is the biggest value that we can observe for q. Therefore we have that

$$\overline{q} = \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}, q_0\right).$$

Since if $p \leq q$, then p grows until $p = q$ or if $p > q$, then p only decreases, we obtain

$$\overline{p} = \max\left(\frac{b}{d}p_0^{\frac{1}{3}}, \left(\frac{b}{d}\right)^{\frac{3}{2}}, q_0, p_0\right).$$

Now, we estimate the lower bound for q. For this purpose, we show that the term $bq^{\frac{2}{3}} - dq^{\frac{4}{3}}$ is non-negative if q is below a certain threshold such that we can focus just on the terms where q is linearly included. We have that the term $bq^{\frac{2}{3}} - dq^{\frac{4}{3}}$ is non-negative if $q \leq \left(\frac{b}{d}\right)^{\frac{3}{2}}$. For the second term, we have that $bp - dp^{\frac{2}{3}}q \geq -d\overline{p}^{\frac{2}{3}}q$. That means if $q \leq \left(\frac{b}{d}\right)^{\frac{3}{2}}$, then we have the following estimation. It holds

$$\begin{aligned}
q\left(\bar{t}\right) &= q_0 + \int_0^{\bar{t}} \varkappa\left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + (1-\varkappa)\left(bp - dp^{\frac{2}{3}}q\right) - \gamma qu - (\varsigma + \delta r)\, qw dt \\
&\geq q_0 + \int_0^{\bar{t}} -\gamma qu - (\varsigma + \delta r)\, qw - (1-\varkappa)\, dp^{\frac{2}{3}} q dt \\
&\geq q_0 + \int_0^{\bar{t}} \left(-\gamma\overline{u} - (\varsigma + \delta\overline{r})\, \overline{w} - (1-\varkappa)\, d\overline{p}^{\frac{2}{3}}\right) q dt \\
&\geq q_0 - \int_0^{\bar{t}} \bar{K} q dt
\end{aligned} \tag{2.80}$$

where $\bar{t} \in \left[0, \tilde{t}\right)$ and $\bar{K} := \left(\gamma\overline{u} + (\varsigma + \delta\overline{r})\, \overline{w} + (1-\varkappa)\, d\overline{p}^{\frac{2}{3}}\right)$. If we choose $\epsilon > 0$ such that $q_0 - \epsilon > 0$, then the function $\tilde{q}\left(t\right) = \min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}}, (q_0 - \epsilon)\right) \exp\left(-\bar{K}t\right)$ fulfills the following integral equation

$$\tilde{q}\left(\bar{t}\right) = \min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}}, (q_0 - \epsilon)\right) - \int_0^{\bar{t}} \bar{K}\tilde{q}\left(t\right) dt \tag{2.81}$$

where it holds that

$$\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},(q_0-\epsilon)\right)-\int_0^{\bar{t}}\bar{K}\tilde{q}\left(t\right)dt<q_0-\int_0^{\bar{t}}\bar{K}\tilde{q}\left(t\right)dt \tag{2.82}$$

since if $\left(\frac{b}{d}\right)^{\frac{3}{2}}\geq(q_0-\epsilon)$, then

$$\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},(q_0-\epsilon)\right)=q_0-\epsilon<q_0$$

and if $\left(\frac{b}{d}\right)^{\frac{3}{2}}<(q_0-\epsilon)$, then

$$\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},(q_0-\epsilon)\right)=\left(\frac{b}{d}\right)^{\frac{3}{2}}<q_0-\epsilon<q_0.$$

From (2.81) and (2.82), we obtain the following

$$\tilde{q}\left(\bar{t}\right)<q_0-\int_0^{\bar{t}}\bar{K}\tilde{q}\left(t\right)dt \tag{2.83}$$

Furthermore we have

$$\tilde{q}\left(0\right)=\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},(q_0-\epsilon)\right)\leq q_0-\epsilon<q_0=q\left(0\right)$$

and consequently we can apply [65, Theorem 5.1.1] to (2.80) and (2.83) and obtain

$$q\left(\bar{t}\right)>\tilde{q}\left(\bar{t}\right)=\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},(q_0-\epsilon)\right)\exp\left(-K\bar{t}\right)$$

for all $\bar{t}\in\left[0,\tilde{t}\right)$. Because the argument holds for all $\epsilon>0$ sufficiently small, we have that $q\left(\bar{t}\right)\geq\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},q_0\right)\exp\left(-K\bar{t}\right)$ for all $\bar{t}\in\left[0,\tilde{t}\right)$. Since

$$\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},q_0\right)\exp\left(-\bar{K}t\right)\geq\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},q_0\right)\exp\left(-\bar{K}T\right)>0$$

for all $t\in[0,T]$ we have that $q\left(t\right)>0$ not only for all $t\in\left[0,\tilde{t}\right)$ but also for all $t\in\left[0,\hat{t}\right)$. This means that $q\left(t\right)$ is bounded from below by

$$\underline{q}:=\min\left(\left(\frac{b}{d}\right)^{\frac{3}{2}},q_0\right)\exp\left(\left(\gamma\overline{u}+\left(\varsigma+\delta\overline{r}\right)\overline{w}+\left(1-\varkappa\right)d\overline{p}^{\frac{2}{3}}\right)T\right)$$

for all $t\in\left[0,\hat{t}\right)$. From $\underline{q}$, the lower bound of q and (2.79), we have the lower bound of p which is given by

$$\underline{p}:=\min\left(p_0,\underline{q}\exp\left(-\frac{\left(\alpha+\beta\overline{r}\right)\overline{w}}{\xi}\right)\right).$$

Consequently the calculation above holds also on the maximum existence interval $\left[0,\hat{t}\right]$ since $\underline{q}$ holds for all $t\in[0,T]$. This means that our local solution stays in I. Concluding, by [90, Proposition C.3.6], we have that there exists a global solution $(p,q,r):[0,T]\to\mathbb{R}^3$, $t\mapsto\left(p\left(t\right),q\left(t\right),r\left(t\right)\right)$ to (2.76). Furthermore, this solution stays in I and is essentially bounded by a constant that holds for all $u\in U_{ad}$ and all $w\in W_{ad}$, given by $\max\left(\overline{p},\overline{q},\overline{r}\right)$. □

Next, the existence of an optimal solution to (2.75) subject to (2.76) is proved in the following lemma.

Lemma 21. *There exists a solution $u \in U_{ad}$ and $w \in W_{ad}$ to (2.75) subject to (2.76), where U_{ad} and W_{ad} are defined in (2.77).*

Proof. The existence of an optimal solution to (2.75) subject to (2.76) can be shown as follows. We know that any solution $x = (p, q, r)$ to (2.76) for

$$u \in U_{ad} = \left\{u \in L^2\left(0, T\right) | 0 \leq u\left(t\right) \leq \overline{u} \text{ a.e.}\right\}, \quad w \in W_{ad} = \left\{w \in L^2\left(0, T\right) | 0 \leq u\left(t\right) \leq \overline{w} \text{ a.e.}\right\}$$

is an element of

$$X = \{x \in \left(L^2(0, T)\right)^3 | x = (p, q, r)\,,\ 0 \leq p \leq \bar{p},\ 0 \leq q \leq \bar{q},\ 0 \leq r \leq \frac{\overline{w}}{\rho}\}$$

where U_{ad}, W_{ad}, X are convex, bounded and closed.

Then we choose a sequence $(u_n, w_n)_{n\in\mathbb{N}} \subseteq U_{ad} \times W_{ad}$ with corresponding solution $(x_n)_{n\in\mathbb{N}} \subseteq X$ to (2.76) for (u_k, w_k) instead of (u, w), which minimizes the cost functional J. Such a minimizing sequence always exists as follows. The cost functional J is bounded from below and thus the infimum

$$d := \inf_{(u,w)\in U_{ad}\times W_{ad}} J\left(x\left(u, w\right), u, w\right) := \inf\left\{(u, w) \in U_{ad} \times W_{ad} |\ J\left(x\left(u, w\right), u, w\right)\right\}$$

exists, see [3, I Theorem 10.4, Theorem 10.1], where $x\left(u, w\right)$ is the solution to (2.76) corresponding to $(u, w) \in U_{ad} \times W_{ad}$. Consequently for any given number $\epsilon_n > 0$, monotonically decreasing for increasing $n \in \mathbb{N}$, there is a u_n with

$$d \leq J\left(x\left(u_n, w_n\right), u_n, w_n\right) \leq d + \epsilon_n.$$

If this was not the case, that means if there was an $\tilde{n}$ such that $d + \epsilon_{\tilde{n}} < J\left(x\left(u, w\right), u, w\right)$ for all $(u, w) \in U_{ad} \times W_{ad}$, then it would contradict d being the biggest lower bound which would be at least $d + \epsilon_{\tilde{n}}$ in this case.

By applying the limit on both sides of the last inequlity, we have for the minimizing sequence $(u_n, w_n)_{n\in\mathbb{N}} \subseteq U_{ad} \times W_{ad}$ that

$$\inf_{(u,w)\in U_{ad}\times W_{ad}} J\left(x\left(u, w\right), u, w\right) = \lim_{n\to\infty} J\left(x\left(u_n, w_n\right), u_n, w_n\right),$$

see [3, Theorem 2.9]. Since the set $U_{ad} \times W_{ad}$ is weakly sequentially compact, see [95, Theorem 2.11] and due to the reflexivity of $L^2\left(0, T\right)$ spaces [95, Section 2.4], there exists an index set $K_1 \subseteq \mathbb{N}$ such that for the corresponding subsequence it holds that $(u_k, w_k) \rightharpoonup (u^*, w^*)$ for $K_1 \ni k \to \infty$ and the weak limit $(u^*, w^*) \in U_{ad} \times W_{ad}$ exists.

In the next step, we show that there is an index set $K_2 \subseteq K_1$ such that the corresponding subsequence of solutions $(x_k)_{k\in K_2}$ with $x_k = (p_k, q_k, r_k) \subseteq X$ to (2.76) for (u_k, w_k) instead of (u, w), $k \in K_2$ converges uniformly pointwise to x^* which is the corresponding solution to (2.76) for (u^*, w^*) instead of (u, w). We have that $(x_k)_{k\in K_1} \subseteq X$ and thus $(x_k)_{k\in K_1}$ is a bounded sequence in X. Since in addition each component f_i, $i \in \{1, 2, 3\}$, of the right hand-side f defined in (2.78) is continuous as a function $f_i\left(t, \cdot\right) : I \to \mathbb{R}$, $(p, q, r) \mapsto f\left(t, p, q, r\right)$ for any fixed $t \in [0, T]$ and thus is bounded on I, see [3, III Corollary 3.8], we have that the function $f_i\left(t, p\left(t\right), q\left(t\right), r\left(t\right)\right) : [0, T] \to \mathbb{R}$ is bounded because of the pointwise bounded controls and that $\left(p\left(t\right), q\left(t\right), r\left(t\right)\right) \in I$ for any $t \in [0, T]$, see the discussion above. This means that the derivative of p, q, r with respect to t is essentially pointwise bounded. Consequently the sequence $(x_k)_{k\in K_1}$ is a bounded sequence in $\left(H^1\left(0, T\right)\right)^3$. From this it follows, that there exists an index set $K_2 \subseteq K_1$ such that $x_k \rightharpoonup x^*$ and $\dot{x}_k \rightharpoonup \frac{d}{dt}x^*$ in $\left(H^1\left(0, T\right)\right)^3$ for $K_2 \ni k \to \infty$, see [95, Theorem 2.11]. This can be seen by taking the following continuous linear functionals on $\left(H^1\left(0, T\right)\right)^3$. For $x_k \rightharpoonup x^*$ take the L^2-scalar product for instance and for $\dot{x}_k \rightharpoonup \frac{d}{dt}x^*$ take the L^2-scalar product for $\dot{x}_k$ instead of x_k. Moreover, due to the compact embedding of $\left(H^1(0,T)\right)^3$ into the space of the continuous functions $(C(0,T))^3$ with the maximum norm, denoted with $\mathcal{C}$, see [26, Theorem 8.8], we have the existence of an index set $K_3 \subseteq K_2$ such that we have the strong convergence for the corresponding subsequence $x_k \to x^*$ for $K_3 \ni k \to \infty$ with $x^* \in \mathcal{C}$, since $\mathcal{C}$ is a Banach space equipped with the maximum norm, see [75, page 65].

Next we prove that for the limit u^* and w^* with corresponding x^* the constraint (2.76) is fulfilled. We have that

$$0 = \int_0^T \left(\dot{p}_k + \xi p_k \ln\left(\tfrac{p_k}{q_k}\right) + (\alpha + \beta r_k)\, p_k w_k\right) \varphi dt$$

for all $\varphi \in C_c^\infty(0,T)$, the space of arbitrarily often differentiable functions with compact support, since u_k, w_k with corresponding x_k fulfill (2.76). According to the definition of weak convergence, we have that for the following continuous linear functional on $H^1(0,T)$ it holds

$$\lim_{K_3 \ni k \to \infty} \int_0^T \dot{p}_k \varphi dt = \int_0^T \frac{d}{dt} p^* \varphi dt.$$

Since the continuous part of the equation is bounded, we have that

$$\lim_{K_3 \ni k \to \infty} \int_0^T \xi p_k \ln\left(\tfrac{p_k}{q_k}\right) \varphi dt = \int_0^T \lim_{K_3 \ni k \to \infty} \xi p_k \ln\left(\tfrac{p_k}{q_k}\right) \varphi dt = \int_0^T \xi p^* \ln\left(\tfrac{p^*}{q^*}\right) \varphi dt$$

by the dominated convergence theorem [36, Theorem 2.4.5]. Next, we have that

$$\begin{aligned} &\lim_{K_3 \ni k \to \infty} \int_0^T (\alpha + \beta r_k)\, p_k w_k \varphi dt \\ &= \lim_{K_3 \ni k \to \infty} \left(\int_0^T ((\alpha + \beta r_k)\, p_k - (\alpha + \beta r^*)\, p^*)\, w_k \varphi dt + \int_0^T (\alpha + \beta r^*)\, p^* w_k \varphi dt \right). \end{aligned}$$

As $(\alpha + \beta r_k)\, p_k - (\alpha + \beta r^*)\, p^*$ is bounded we have by the dominated convergence theorem [36, Theorem 2.4.5] that

$$\begin{aligned} &\lim_{K_3 \ni k \to \infty} \int_0^T ((\alpha + \beta r_k)\, p_k - (\alpha + \beta r^*)\, p^*)\, w_k \varphi dt \\ &= \int_0^T \varphi \lim_{K_3 \ni k \to \infty} (((\alpha + \beta r_k)\, p_k - (\alpha + \beta r^*)\, p^*)\, w_k) dt = 0 \end{aligned}$$

since w_k is bounded for any $k \in \mathbb{N}$, see [3, II Theorem 2.4 (i)]. Because w_k is weakly converging in $L^2(0,T)$, we have that the continuous linear functional $\int_0^T (\alpha + \beta r^*)\, p^* \cdot \varphi dt$ on $L^2(0,T)$ converges as follows

$$\lim_{K_3 \ni k \to \infty} \int_0^T (\alpha + \beta r^*)\, p^* w_k \varphi dt = \int_0^T (\alpha + \beta r^*)\, p^* w^* \varphi dt.$$

Summarizing, we obtain that

$$0 = \int_0^T \left(\frac{d}{dt} p^* + \xi p^* \ln\left(\tfrac{p^*}{q^*}\right) + (\alpha + \beta r^*)\, p^* w^* \right) \varphi dt$$

for all $\varphi \in C_c^\infty(0,T)$. Then it holds that

$$\frac{d}{dt} p^* = -\xi p^* \ln\left(\tfrac{p^*}{q^*}\right) - (\alpha + \beta r^*)\, p^* w^*$$

due to the fundamental lemma of calculus of variation, see [49, Chapter 1 Lemma 3], which means that the first equation of (2.76) is fulfilled for (x^*, u^*, w^*). Analogously we argue for the equations for q and r, that means for the second and the third equation of the right hand-side of (2.76). Consequently, we have shown that our weak limits (x^*, u^*, w^*) fulfill the constraint (2.76).

In the final step we have to see that this weak limit is a minimizer sought. Our objective $J : \mathcal{C} \times U_{ad} \times W_{ad} \to \mathbb{R}$, $(x, u, w) \mapsto J(x, u, w)$ is convex and continuous as a map from $\mathcal{C} \times U_{ad} \times W_{ad}$ to $\mathbb{R}$ for the following reason. The convexity follows from the non-negativity of p, the convexity of norms and the Jensen inequality, see [72, Proposition 824]. The continuity follows from the composition of continuous functions [3, III Theorem 1.8] as follows. First we the composition of the norm and the square of a function which are both continuous maps. Second we have that

$$\begin{aligned}
&\|p_1 - p_2\|_{L^2(0,T)} \\
&= \sqrt{\int_0^T (p_1(t) - p_2(t))^2 \, dt} \leq \sqrt{T} |\max_{t\in[0,T]} (p_1(t) - p_2(t))| \\
&\leq \sqrt{T} \max_{t\in[0,T]} |p_1(t) - p_2(t)| = \sqrt{T} \|p_1 - p_2\|_{L^\infty(0,T)}
\end{aligned}$$

and

$$\begin{aligned}
&|p_1^2(T) - p_2^2(T)| \\
&= |(p_1(T) + p_2(T))(p_1(T) - p_2(T))| \leq 2\overline{p}(p_1(T) - p_2(T)) \\
&\leq 2\overline{p} \max_{t\in[0,T]} |p_1(t) - p_2(t)| = 2\overline{p} \|p_1 - p_2\|_{L^\infty(0,T)}
\end{aligned}$$

for any two solutions p_1, p_2 of (2.76) which proves the continuity as a map from $\mathcal{C}$ to $\mathbb{R}$ of the terms of the functional J which include p. Consequently the functional $J : \mathcal{C} \times U_{ad} \times W_{ad} \to \mathbb{R}$, $(x, u, w) \mapsto J(x, u, w)$ is weakly lower semi-continuous, see [95, Theorem 2.12]. Since the sequence $(x_k)_{k\in K_3}$ converges strongly in $\mathcal{C}$, it follows that the sequence $(x_k)_{k\in K_3}$ also converges weakly in $\mathcal{C}$ with $x_k \rightharpoonup x^*$ for $K_3 \ni k \to \infty$, see [95, Subsection 3.4.3]. This means we write $(x_k, u_k, w_k) \rightharpoonup (x^*, u^*, w^*)$. Thus, since the subsequence is still a minimizing sequence, we have that

$$\inf_{(u,w)\in U_{ad}\times W_{ad}} J(x(u,w), u, w) = \lim_{K_3\ni k\to\infty} J(x_k, u_k, w_k) = \liminf_{K_3\ni k\to\infty} J(x_k, u_k, w_k) \geq J(x^*, u^*, w^*)$$

where the first equality sign is true due to the discussion above and for the second equality sign we recall that whenever a lim exists the corresponding lim inf equals lim, see [3, Theorem 5.7]. This proves that (x^*, u^*, w^*) is a minimizer sought. □

Next, we show that our Assumptions A.1) to A.6) are fulfilled. For this purpose, we define the right hand-side of (2.76) by

$$f(y, u, w) := \begin{pmatrix} -\xi p \ln\left(\frac{p}{q}\right) - (\alpha + \beta r) p w \\ \varkappa\left(b q^{\frac{2}{3}} - d q^{\frac{4}{3}}\right) + (1 - \varkappa)\left(b p - d p^{\frac{2}{3}} q\right) - \gamma q u - (\varsigma + \delta r) q w \\ -\rho r + w \end{pmatrix}$$

with $y := (p, q, r)$. We have that A.1) is fulfilled for $f(y, u, w)$, $h(p) := \frac{\tilde{\sigma}}{2} p^2$, $F(p) := \frac{\vartheta}{2} p^2(T)$. Since the function $(y, u, w) \mapsto f(y, u, w)$, $(y, u, w) \mapsto D_y f(y, u, w)$, $(y, u, w) \mapsto D_y f(y, u, w)$ are continuous on $I \times K_U \times K_W$, we have that A.2) and A.4) are fulfilled, see Remark 1. Since I is bounded, as shown in the proof of Lemma 20, we have that A.3) and A.5) hold. Due to the twice continuous differentiability of h and F, we have that also A.6) holds. Because the control mechanism in the present model is bilinear, we can adopt the calculation from Example 18 to show that (2.32) holds. Therefore the theoretical framework of Section 2.2 and Section 2.3 holds for the presented optimal control problem (2.75) subject to (2.76).

According to Section 2.2 the Hamiltonian is given by

$$\begin{aligned}
&H\left(t,p,q,r,\lambda_1,\lambda_2,\lambda_3,u,w\right)\\
&= \tfrac{\tilde{\sigma}}{2}p^2 + \tfrac{\nu_u}{2}u^2 + \tfrac{\nu_w}{2}w^2 + \mu_u|u| + \mu_w|w| - \lambda_1\left(\xi p\ln\left(\tfrac{p}{q}\right) + \left(\alpha+\beta r\right)pw\right)\\
&+ \lambda_2\left(\varkappa\left(bq^{\frac{2}{3}} - dq^{\frac{4}{3}}\right) + \left(1-\varkappa\right)\left(bp - dp^{\frac{2}{3}}q\right) - \gamma qu - \left(\varsigma+\delta r\right)qw\right) + \lambda_3\left(-\rho r + w\right)
\end{aligned}$$

and the corresponding adjoint equations are given by

$$\begin{aligned}
\dot{\lambda_1} &= -\tilde{\sigma}p + \lambda_1\left(\xi\ln\left(\tfrac{p}{q}\right) + \xi + \left(\alpha+\beta r\right)w\right) - \lambda_2\left(1-\varkappa\right)\left(b - \tfrac{2}{3}dp^{-\frac{1}{3}}q\right)\\
\dot{\lambda_2} &= -\lambda_1\xi\frac{p}{q} - \lambda_2\left(\varkappa\left(\tfrac{2}{3}bq^{-\frac{1}{3}} - \tfrac{4}{3}dq^{\frac{1}{3}}\right) + \left(1-\varkappa\right)\left(-dp^{\frac{2}{3}}\right) - \gamma u - \left(\varsigma+\delta r\right)w\right)\\
\dot{\lambda_3} &= \lambda_1\beta pw + \lambda_2\delta qw + \lambda_3\rho
\end{aligned} \tag{2.84}$$

with the terminal conditions

$$\begin{aligned}
\lambda_1\left(T\right) &= \vartheta p\left(T\right)\\
\lambda_2\left(T\right) &= 0\\
\lambda_3\left(T\right) &= 0
\end{aligned}$$

where p,q,r,u and w fulfill (2.76). We remark, since our optimal control problem fulfills the Assumptions A.1) to A.6), the adjoint equation (2.84) has a unique global solution on $[0,T]$, see Theorem 55.

For our SQH method, the augmented Hamiltonian is defined as follows

$$K_\epsilon\left(t,p,q,r,\lambda,u,w,\hat{u},\hat{w}\right) := H\left(t,p,q,r,\lambda_1,\lambda_2,\lambda_3,u,w\right) + \epsilon\left(\left(u\left(t\right)-\hat{u}\left(t\right)\right)^2 + \left(w\left(t\right)-\hat{w}\left(t\right)\right)^2\right)$$

where $\epsilon > 0$ and $K_\epsilon : \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \times K_U \times K_W \times K_U \times K_W \to \mathbb{R}$. The pointwise minimum is given by

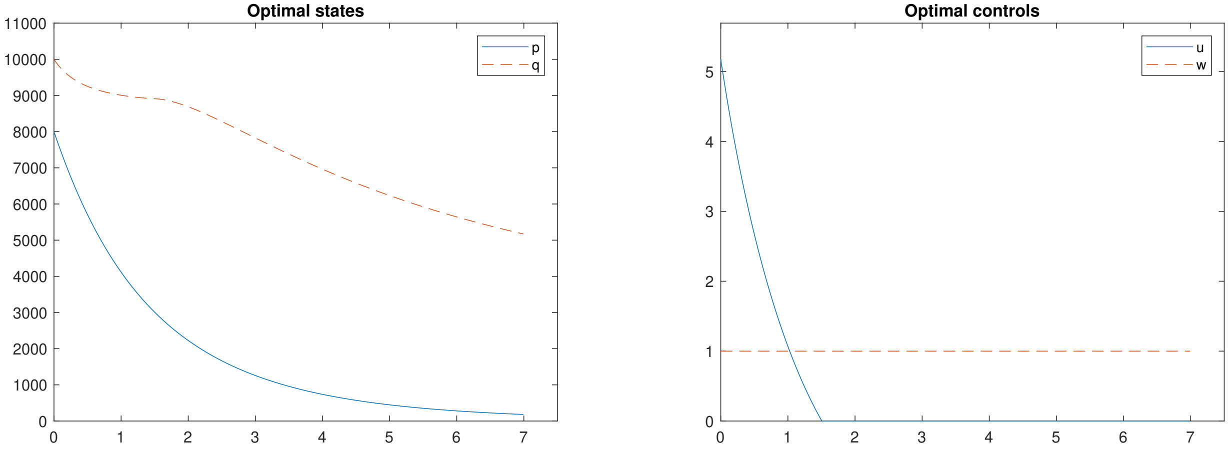
$$\begin{aligned}
u\left(t\right) &= \min\left(\overline{u}, \max\left(0, \frac{-\mu_u + \lambda_2\left(t\right)\gamma q\left(t\right) + 2\epsilon\hat{u}\left(t\right)}{\nu_u + 2\epsilon}\right)\right)\\
w\left(t\right) &= \min\left(\overline{w}, \max\left(0, \frac{-\mu_w + \lambda_1\left(t\right)\left(\alpha+\beta r\left(t\right)\right)p\left(t\right) + \lambda_2\left(t\right)\left(\varsigma+\delta r\left(t\right)\right)q\left(t\right) - \lambda_3\left(t\right) + 2\epsilon\hat{w}\left(t\right)}{\nu_w + 2\epsilon}\right)\right)
\end{aligned}$$

where the calculation is analogous as the one starting on page 166 in the Appendix.

For our numerical illustration, we choose the model constants from Table 2.4 and the following constants. We have $\varkappa = 0.5$, $\mu_u = \mu_w = 10^4$ and $\nu_u = \nu_w = 5\cdot 10^3$, $\tilde{\sigma} = \frac{1}{5}$, $\vartheta = 1$, $T = 7$, $p_0 = 8000$, $q_0 = 10000$ and $r_0 = 0$ for the optimal control problem. For the SQH method, we choose $\kappa = 10^{-7}$, $\sigma = 50$, $\zeta = 0.15$, $\eta = 10^{-7}$ and the initial guess $\epsilon = \frac{1}{10}$ and the initial guess for the controls $u^0 = w^0 = 0$. Our step size $dt = \frac{7}{1000}$ and we use an explicit Euler scheme to solve the state and adjoint equation. Algorithm 2.1 converges after 15 sweeps of Step 2 to Step 4 in 0.06 seconds. The results are depicted in Figure 2.4 where about $p\left(T\right) = 180$, $q\left(T\right) = 5200$ and the inequality

$$0 \leq \left(H\left(t,p,q,r,\lambda_1,\lambda_2,\lambda_3,u,w\right) - \min_{w\in K_U} H\left(t,p,q,r,\lambda_1,\lambda_2,\lambda_3,u,w\right)\right) \leq 10^{-15}$$

is fulfilled for 78% of the grid points where $p,q,r,\lambda_1,\lambda_2,\lambda_3,u$ and w are the return values of the SQH method. This result validates the convergence of Algorithm 2.1 to a solution to (2.75) that is optimal in the PMP sense (2.16).



(a) On the ordinate we have the values for p and q and on the abscissa, we have the time in days.

(b) On the ordinate we have the values for u and w and on the abscissa, we have the time in days.

Figure 2.4: Results from the optimal control problem (2.75) calculated by the SQH method.

Chapter 3

An SQH framework for PDE optimal control problems

In this chapter, similar to [21, 22], we discuss optimal control problems that are governed by partial differential equations (PDEs), specifically elliptic and parabolic PDEs. We characterize an optimal solution with the Pontryagin maximum principle (PMP) and perform the convergence analysis of the sequential quadratic Hamiltonian (SQH) method with which we solve the PDE control problems in this chapter. In order to give an overview over the considered problems we shortly list them below.

P.1) $\min_{y,u} J\left(y,u\right) := \int_Q \frac{1}{2}\left(y\left(x,t\right) - y_d\left(x,t\right)\right)^2 + g_1\left(u\left(x,t\right)\right) dxdt$
subject to $\left(y'\left(\cdot,t\right),v\right) + D\left(\nabla y\left(\cdot,t\right),\nabla v\right) = \left(u\left(\cdot,t\right),v\right)$, $y\left(\cdot,0\right) = y_0$, $u \in U_{ad}$;

P.2) $\min_{y,u} J\left(y,u\right) := \int_\Omega \frac{1}{2}\left(y\left(x\right) - y_d\left(x\right)\right)^2 + g_1\left(u\left(x\right)\right) dx$ subject to $\left(\nabla y,\nabla v\right) = \left(u,v\right)$, $u \in U_{ad}$;

P.3) $\min_{y,u} J\left(y,u\right) := \int_\Omega \frac{1}{2}\left(y\left(x\right) - y_d\left(x\right)\right)^2 + g_1\left(u\left(x\right)\right) dx$ subject to $\left(\nabla y,\nabla v\right) + \left(uy,v\right) = \left(\tilde{f},v\right)$, $u \in U_{ad}$;

P.4) $\min_{y,u} J\left(y,u\right) := \int_Q \frac{1}{2}\left(y\left(x,t\right) - y_d\left(x,t\right)\right)^2 + g_1\left(u\left(x,t\right)\right) dxdt$
subject to $\left(y'\left(\cdot,t\right),v\right) + D\left(\nabla y\left(\cdot,t\right),\nabla v\right) + \left(u\left(\cdot,t\right)y\left(\cdot,t\right),v\right) = \left(f\left(\cdot,t\right),v\right)$, $y\left(\cdot,0\right) = y_0$, $u \in U_{ad}$;

P.5) $\min_{y,u} J\left(y,u\right) := \int_\Omega \frac{1}{2}\left(y\left(x\right) - y_d\left(x\right)\right)^2 + g_1\left(u\left(x\right)\right) dx$ subject to $\left(\nabla y,\nabla v\right) + \left(y^3,v\right) = \left(u,v\right)$, $u \in U_{ad}$;

P.6) $\min_{y,u} J\left(y,u\right) := \int_\Omega \frac{1}{2}\left(y\left(x\right) - y_d\left(x\right)\right) + g_1\left(u\left(x\right)\right) dx$ subject to $\left(\nabla y,\nabla v\right) = \left(u,v\right)$, $y \leq \xi$, $u \in U_{ad}$;

P.7) $\min_{y,u} J\left(y,u\right) := \int_\Omega \left|y\left(x\right) - y_d\left(x\right)\right| + g_2\left(u\left(x\right)\right) dx$
subject to $\left(\nabla y,\nabla v\right) + \left(\max\left(y,0\right),v\right) = \left(u,v\right)$, $u \in U_{ad}$;

where, in all cases, we choose homogeneous Dirichlet boundary conditions and we take

$$g_1\left(z\right) := \begin{cases} |z| & \text{if } |z| > s \\ 0 & \text{else} \end{cases}, s > 0 \qquad \text{and} \qquad g_2\left(z\right) := \ln\left(1 + |z|\right) \tag{3.1}$$

with $z \in \mathbb{R}$. In the following, we use the notation $var1 \leftarrow var2$ which means that the variable $var1$ is replaced by $var2$ in the corresponding equation. Furthermore the L^p-norm for a vector valued function $\tilde{\zeta}$ with n components is defined by $\|\tilde{\zeta}\|^p_{L^p(Z_i)} := \sum_{l=1}^n \|\tilde{\zeta}_l\|^p_{L^p(Z_i)}$, $\|\tilde{\zeta}_l\|_{L^p(Z_i)} := \left(\int_{Z_i} |\tilde{\zeta}_l\left(z\right)|^p dz\right)^{\frac{1}{p}}$, see [5, X.4] for a definition with $p \in (0,\infty)$.

3.1 Parabolic and elliptic optimal control problems

In this section, we formulate classes of elliptic and parabolic optimal control problems. In order to distinguish between the elliptic and the parabolic case, we introduce the index $i \in \{\mathrm{e}, \mathrm{p}\}$, where e refers to the elliptic case and p to the parabolic case. We denote $Z_{\mathrm{e}} \coloneqq \Omega \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$ and $Z_{\mathrm{p}} \coloneqq \Omega \times (0, T)$, $T > 0$ where Ω is an open and bounded domain. Further, we define a vector of controls $u \coloneqq (u_1, ..., u_m)$ where each u_j, $j \in \{1, ..., m\}$, $m \in \mathbb{N}$, is an element of the following admissible set of controls

$$U_{ad}^j = \left\{u_j \in L^q\left(Z_i\right) \mid u_j\left(z\right) \in K_U^j \text{ a.e. in } Z_i\right\}$$

with $K_U^j \subseteq \mathbb{R}$ compact, $q \geq 2$ and $U_{ad} \coloneqq U_{ad}^1 \times ... \times U_{ad}^m$, $K_U \coloneqq K_U^1 \times ... \times K_U^m$.

For our cases, we define

$$U_{ad} \coloneqq \left\{L^q\left(\Omega\right) \mid u\left(x\right) \in K_U \text{ a.e. in } \Omega\right\}$$

in the elliptic case where $q \geq \max\left(2, \frac{n}{2} + 1\right)$ and $K_U \subsetneq \mathbb{R}$ is a compact set, specified later. Analogous we define

$$U_{ad} \coloneqq \left\{L^q\left(Q\right) \mid u\left(x, t\right) \in K_U \text{ a.e. in } Q\right\}$$

in the parabolic case where $Q \coloneqq \Omega \times (0, T)$ and we have that $q > \frac{n}{2} + 1$ for $n \geq 2$ and $q \geq 2$ for $n = 2$. Also in this case $K_U \subsetneq \mathbb{R}$ is a compact set, specified later.

Next, we formulate our PDE constraint (governing model) in weak form as in [45, page 296] for the elliptic case and as in [45, page 351/352] for the parabolic case.

Consider the bilinear form for the parabolic case as follows: $B : \mathcal{H}\times\mathcal{H}\times[0, T] \to \mathbb{R}$, $(y, v; t) \mapsto B\left(y, v; t\right)$ where the function space $\mathcal{H}$, a set of functions mapping Z_i to $\mathbb{R}$, has to be chosen accordingly. Then, the weak formulation of a parabolic equation is given by

$$\left(y'\left(\cdot, t\right), v\right) + B\left(y, v; t\right) = \int_{\Omega} f\left(x, t, y, u\right) v\left(x\right) dx \tag{3.2}$$

for almost every $t \in [0, T]$ and all $v \in \mathcal{H}$ where $(\cdot, \cdot)$ is the $L^2\left(\Omega\right)$ scalar product, $y' \coloneqq \frac{\partial}{\partial t} y$ and $f : \mathbb{R}^n \times \mathbb{R}_0^+ \times \mathbb{R} \times K_U \to \mathbb{R}$. Notice that $f\left(z, y, u\right) \coloneqq f\left(z, y\left(z\right), u\left(z\right)\right)$, $z \in Z_i$, whenever an argument of f is a function instead of a number. We implicitly assume for the rest of this chapter that if functions do not depend on time, we refer to the elliptic case and if we consider an elliptic equation, then functions do not depend on time t. This implies in the elliptic case that $\left(y'\left(\cdot, t\right), v\right) = 0$ for all $v \in \mathcal{H}$. We require that (3.2) is well defined and that there is a unique solution $y : Z_i \to \mathbb{R}$, $z \mapsto y\left(z\right)$ to (3.2), that means that y fulfills (3.2) for almost all $t \in [0, T]$ and all $v \in \mathcal{H}$.

Now, in view of this requirement, we consider P.1) to P.7) and define $\mathcal{H} \coloneqq H_0^1\left(\Omega\right)$ and $y_0 \in H_0^1\left(\Omega\right) \cap L^\infty\left(\Omega\right)$ in the parabolic case where we also assume a smooth boundary for Ω. For P.1), we have the bilinear form $B_1\left(y, v; t\right) \coloneqq D\left(\nabla y\left(\cdot, t\right), v\right)$ and the right-hand side $f_1 = u$ where the model of P.1) has a unique solution $y \in L^2\left(0, T, H^2\left(\Omega\right)\right) \cap L^\infty\left(0, T; H_0^1\left(\Omega\right)\right)$, see [45, 7.1 Theorem 5]. For the governing model of P.2), we have the bilinear form $B_2\left(y, v\right) \coloneqq \left(\nabla y, \nabla v\right)$ and the right-hand side $f_2 = u$ where the constraint of P.2) has a unique solution $y \in H_0^1\left(\Omega\right)$, see [45, 6.2 Theorem 1] and [1, Theorem 2.14]. Analogously for P.3) with $B_3 \coloneqq B_1$ and $f_3 \coloneqq \tilde{f} - uy$ and for P.4) with $B_4 \coloneqq B_2$ and $f_4 \coloneqq f_3$, we have a unique solution for the corresponding constraint where we require that $K_U \subseteq \mathbb{R}_0^+$ and $\tilde{f} \in L^q\left(Z_i\right)$. In P.5), the constraint for $B_5 \coloneqq B_1$ and $f_5 \coloneqq u - y^3$ has a unique solution $y \in H_0^1\left(\Omega\right)$, see [29]. The case P.6) is analogous to P.2) where we additionally require that $y \leq \xi$, $\xi \in \mathbb{R}$. The constraint of P.7) for $B_6 \coloneqq B_1$ and $f_6 \coloneqq u - \max\left(y, 0\right)$ has a unique solution $y \in H_0^1\left(\Omega\right)$, see [31].

In the next step, we formulate our general optimal control problem as follows

$$\begin{aligned} &\min_{y,u} J\left(y, u\right) \coloneqq \int_{Z_i} h\left(y\left(z\right)\right) + g\left(u\left(z\right)\right) dz \\ &\text{s.t. } \int_{\Omega} y'\left(x, t\right) v\left(x\right) dx + B\left(y, v; t\right) = \int_{\Omega} f\left(x, t, y\left(x, t\right), u\left(x, t\right)\right) v\left(x\right) dx \\ &u \in U_{ad} \end{aligned} \tag{3.3}$$

where $z := (x,t)$ for the parabolic case and $z := x$ for the elliptic case. We assume J to be bounded from below and require that (3.3) is well posed. In particular, we assume $g : K_U \to \mathbb{R}$, $z \mapsto g(z)$ to be bounded from below and lower semi-continuous analogous to Section 2.1.

We remark that, analogous to Chapter 2, in the case of an L^2-L^1-functional, existence of an optimal solution to (3.3), denoted with $(\bar{y}, \bar{u})$, can be proved. This case is included in our framework presented in this section. Notice that the scope of our SQH scheme and the corresponding framework is beyond this case. For the purpose of this thesis, we assume the existence of an optimal solution to (3.3) in order to focus on its characterization in the PMP framework and the convergence analysis of the SQH scheme under the assumptions listed below.

Before we define the corresponding PMP necessary optimality conditions that a solution to (3.3) must fulfill, we introduce the adjoint bilinear form $B^* : \mathcal{H} \times \mathcal{H} \times [0,T] \to \mathbb{R}$, $(p,v,t) \mapsto B^*(p,v;t)$, where we require that

$$B^*(p,v;t) = B(v,p;t) \tag{3.4}$$

holds for almost every $t \in [0,T]$ and all $v \in \mathcal{H}$. The adjoint equation for (3.3) is, analogous to [81, Theorem 2.1], defined as follows

$$-\left(p'(\cdot,t), v\right) + B^*(p,v;t) = \int_\Omega \left(\frac{\partial}{\partial y} h(y)\,|_{y=y(x,t)} + \frac{\partial}{\partial y} f(x,t,y,u)\,|_{y=y(x,t)} p(x,t)\right) v(x)\,dx \tag{3.5}$$

with $p(\cdot,T) = 0$ where y is the solution to (3.2) and $p' := \frac{\partial}{\partial t}p$. We require that there exists a unique solution $p : Z_i \to \mathbb{R}$, $z \mapsto p(z)$ such that (3.5) holds for almost all $t \in [0,T]$ and all $v \in \mathcal{H}$. We remark that for P.1) to P.5) it is shown in Section 3.2 that the corresponding adjoint equations are uniquely solvable.

Crucial for the PMP and later for the SQH method is the Hamiltonian $H : Z_i \times \mathbb{R} \times K_U \times \mathbb{R} \to \mathbb{R}$, $(z,y,u,p) \mapsto H(z,y,u,p)$ that is given by

$$H(z,y,u,p) := h(y) + g(u) + pf(z,y,u)\,. \tag{3.6}$$

Now, we can formulate the necessary optimality conditions given by the PMP. If $\bar{p}$ is the solution to (3.5) where $\bar{y}$ is inserted for y, we write $y \leftarrow \bar{y}$, and $\bar{u}$ is inserted for u, we write $u \leftarrow \bar{u}$, then we have that

$$H(z,\bar{y},\bar{u},\bar{p}) = \min_{w \in K_U} H(z,\bar{y},w,\bar{p}) \tag{3.7}$$

for almost all $z \in Z_i$. Notice that we use the notation $H(z,y,u,p) := H(z,y(z),u(z),p(z))$ whenever an argument of H is a function instead of a number.

We additionally assume that a solution to (3.3) fulfills the PMP (3.7). For the analysis of the SQH method, we make the following assumptions, that are used in the corresponding proofs. For this purpose, we define the set $I \subseteq \mathbb{R}$ as the convex hull [10, Section 3.1] of the union of all images from each solution y to (3.2) for any $u \in U_{ad}$, given by

$$I := \operatorname{conv}\{y(Z_i) \subseteq \mathbb{R} |\ y \text{ solves (3.2) for } u \in U_{ad}\}\,.$$

The assumptions are given as follows.

A.1) The functions $h : I \to \mathbb{R}$, $v \mapsto h(v)$ and $f : I \to \mathbb{R}$, $v \mapsto f(z,v,u)$ are supposed to be twice continuously differentiable for all $u \in K_U$ and for almost all $z \in K_i$.

Furthermore, we require the existence of a constant $c > 0$ such that the following holds

A.2) $\|\delta y\|_{L^2(Z_i)} \le c\|\delta u\|_{L^2(Z_i)}$, $\|\delta p\|_{L^2(Z_i)} \le c\|\delta u\|_{L^2(Z_i)}$;

A.3) The function $f : K_U \to \mathbb{R}$, $u \mapsto f(z,y,u)$ is Lipschitz continuous with $|f(z,y,u_1) - f(z,y,u_2)| \le c\sum_{j=1}^m |(u_1)_j - (u_2)_j|$ for any fixed $y \in I$ and $z \in Z_i$. That means the Lipschitz constant c is independent of all $z \in Z_i$, all $u \in K_U$ and all $y \in I$;

A.4) $\|p\|_{L^\infty(Z_i)} \leq c$ for any solution (y,u) to (3.2) for $u \in U_{ad}$;

A.5) $\|\frac{\partial}{\partial y} f\left(\cdot, y, u\right)\|_{L^\infty(Z_i)} \leq c$, $\|\frac{\partial^2}{\partial y^2} f\left(\cdot, y, u\right)\|_{L^\infty(Z_i)} \leq c$, $\|\frac{\partial^2}{\partial y^2} h\left(y\right)\|_{L^\infty(Z_i)} \leq c$ for all $y \in I$ and $u \in K_U$;

where $\delta y := y_1 - y_2$, $\delta u := u_1 - u_2$, $\delta p := p_1 - p_2$ and $\|\delta u\|^2_{L^2(Z_i)} := \sum_{j=1}^m \|\delta u_j\|^2_{L^2(Z_i)}$ where (y_ℓ, u_ℓ), $\ell = 1, 2$, are solutions to (3.2) and (y_ℓ, u_ℓ, p_ℓ) are solutions to (3.5) . Additionally we require the following.

A.6) The function $f : I \times K_U \to \mathbb{R}$, $(y,u) \mapsto f\left(z, y, u\right)$ is continuous for all $z \in Z_i$.

We remark that the Assumptions A.1) to A.6) are fulfilled by P.1) to P.5), see Section 3.4. The cases P.6) and P.7) are not covered by our PMP characterization that is similar to [81]. Nevertheless, we show how to apply the SQH method to these problems.

We can show that optimal solutions to our problems P.1) to P.5) can be characterized by the PMP with an analogous calculation as in [81]. The next section is devoted to this purpose.

3.2 The characterization by the Pontryagin maximum principle

In this section, we formulate a class of optimal control problems, including P.1) to P.5), where a solution to the corresponding optimal control problem (3.3) fulfills the necessary equation (3.7). For this purpose, we choose $\mathcal{H} = H_0^1\left(\Omega\right)$. The argumentation in the present section is similar to the one in Section 2.2. Therefore, we recall the definition of the needle variation of a function $u^* \in U_{ad}$ that is given by

$$u_k\left(z\right) := \begin{cases} u & z \in S_k\left(z_0\right) \cap Z_i \\ u^*\left(z\right) & z \in Z_i \backslash S_k\left(z_0\right) \end{cases} \tag{3.8}$$

where $u \in K_U$ and $S_k\left(z_0\right)$ is a ball centered at $z_0 \in Z_i$ whose measure, denoted by $|S_k\left(z_0\right)|$, goes to zero for k to infinity.

We remark that the function $u_k \in U_{ad}$ for all $k \in \mathbb{N}$, for all $z \in Z_i$ and $u^* \in U_{ad}$. This can be seen as follows. The function $u_k = u^* \chi_{Z_i \backslash S_k(z_0)} + u \chi_{S_k(z_0)}$ is measurable for all $k \in \mathbb{N}$ and $z_0 \in Z_i$ because the sum and the product of measurable functions is measurable, see [36, Proposition 2.1.7] and the characteristic function χ_A is measurable if and only if A is measurable, see [36, Example 2.1.2]. Consequently the needle variation is measurable. The integrability is seen by

$$\begin{aligned} \sum_{j=1}^m \int_{Z_i} \left(\left(u_k\right)_j\left(z\right)\right)^q dz &= \sum_{j=1}^m \int_{Z_i \backslash S_k(z_0)} \left(\left(u^*\right)_j\left(z\right)\right)^q dz + \int_{S_k(z_0) \cap Z_i} u_j^q dz \\ &\leq \sum_{j=1}^m \left(\int_{Z_i} \left(\left(u^*\right)_j\left(z\right)\right)^q dz + u_j^q |S_k\left(z_0\right)|\right) \end{aligned} \tag{3.9}$$

where we have the L^q-integrability since $u^* \in U_{ad}$ and u_j are real numbers for all $j \in \{1, ..., m\}$. As the image of the needle variation is in K_U almost everywhere the needle variation it holds $u_k \in L^q\left(Z_i\right)$.

Also in the PDE case we define the intermediate adjoint equation, analogous to [81, (22)] which is needed for the PMP characterization, by

$$-\left(\tilde{p}'\left(\cdot, t\right), v\right) + B^*\left(\tilde{p}, v; t\right) = \int_\Omega \left(\tilde{h}\left(y_1, y_2\right) + \tilde{f}\left(x, t, y_1, y_2, u_1\right) p\left(x, t\right)\right) v\left(x\right) dx \tag{3.10}$$

with $\tilde{p}\left(\cdot, T\right) = 0$ where y_1 is the solution to (3.2) for $u \leftarrow u_1$, y_2 is the solution to (3.2) for $u \leftarrow u_2$ and $\tilde{p}' := \frac{\partial}{\partial t}\tilde{p}$ with

$$\tilde{f}\left(x, t, y_1, y_2, u_1\right) := \int_0^1 \frac{\partial}{\partial y} f\left(x, t, y, u_1\right)|_{y=y_2+\theta(y_1-y_2)} d\theta$$

and

$$\tilde{h}(y_1, y_2) := \int_0^1 \frac{\partial}{\partial y} h(y)\,|_{y=y_2+\theta(y_1-y_2)} d\theta.$$

Also, we require that there exists a unique solution $\tilde{p} : Z_i \to \mathbb{R}$, $z \mapsto p(z)$ such that (3.10) holds for almost all $t \in [0, T]$ and all $v \in \mathcal{H}$. This implicitly includes that $\tilde{f}$ and $\tilde{h}$ are well defined for any solution y_1, y_2 to (3.2) for $u_1, u_2 \in U_{ad}$.

In the next step, we have to the following lemma, analogous to Lemma 3, which is used for the PMP characterization of a solution to (3.3).

Lemma 22. *Let (y_1, u_1) and (y_2, u_2) solve (3.2). Then, it holds that*

$$J(y_1, u_1) - J(y_2, u_2) = \int_{Z_i} H(z, y_2, u_1, \tilde{p}) - H(z, y_2, u_2, \tilde{p})\, dz$$

where $\tilde{p}$ solves (3.10).

Proof. Because of the continuity of $\frac{\partial}{\partial y} f$ and $\frac{\partial}{\partial y} h$ in the state argument, we apply the fundamental theorem of calculus [4, VI 4.13] and thus we obtain pointwise

$$\begin{aligned} f(z, y_1, u_1) - f(z, y_2, u_1) &= f(z, y_2 + \theta(y_1 - y_2), u_1)\,|_{\theta=1} - f(z, y_2 + \theta(y_1 - y_2), u_1)\,|_{\theta=0} \\ &= \int_0^1 \frac{\partial}{\partial y} f(z, y, u_1)\,|_{y=y_2+\theta(y_1-y_2)} (y_1 - y_2)\, d\theta = \tilde{f}(z, y_1, y_2, u_1)(y_1 - y_2) \end{aligned}$$

with the chain rule [4, VII Theorem 3.3]. Analogously, we have

$$h(y_1) - h(y_2) = \tilde{h}(y_1, y_2)(y_1 - y_2).$$

Next, we obtain

$$\begin{aligned} &J(y_1, u_1) - J(y_2, u_2) = \int_{Z_i} h(y_1) + g(u_1) - h(y_2) - g(u_2)\, dz \\ &= \int_{Z_i} h(y_2) + g(u_1) - h(y_2) + h(y_1) - h(y_2) - g(u_2) + \tilde{p}f(z, y_2, u_1) - \tilde{p}f(z, y_2, u_1)\, dz \\ &+ \int_{Z_i} \tilde{p}f(z, y_2, u_2) - \tilde{p}f(z, y_2, u_2)\, dz \\ &= \int_{Z_i} H(z, y_2, u_1, \tilde{p}) - H(z, y_2, u_2, \tilde{p}) + h(y_1) - h(y_2)\, dz \\ &+ \int_{Z_i} \tilde{p}\left(f(z, y_2, u_2) - f(z, y_1, u_1) + f(z, y_1, u_1) - f(z, y_2, u_1)\right) dz \\ &= \int_{Z_i} H(z, y_2, u_1, \tilde{p}) - H(z, y_2, u_2, \tilde{p}) + \left(\tilde{h}(y_1, y_2) + \tilde{p}\tilde{f}(z, y_1, y_2, u_1)\right)(y_1 - y_2)\, dz \\ &+ \int_{Z_i} \tilde{p}\left(f(z, y_2, u_2) - f(z, y_1, u_1)\right) dz \\ &= \int_{Z_i} H(z, y_2, u_1, \tilde{p}) - H(z, y_2, u_2, \tilde{p})\, dz + \int_0^T -\tilde{p}'(y_1 - y_2) + B^*[\tilde{p}, y_1 - y_2; t]\, dt \\ &- \int_0^T \tilde{p}y_1' + B[y_1, \tilde{p}; t] - \tilde{p}y_2' - B[y_2, \tilde{p}; t]\, dt \\ &= \int_{Z_i} H(z, y_2, u_1, \tilde{p}) - H(z, y_2, u_1, \tilde{p})\, dz \end{aligned}$$

where we use the partial integration [95, Theorem 3.11] in the third to last line and (3.4). □

Lemma 23. *Let $u^* \in U_{ad}$ and $u \in K_U$. Furthermore let u_k be defined as in (3.8) for all $k \in \mathbb{N}$ and y_k be the solution to (3.2) for $u \leftarrow u_k$. Let y^* be the solution to (3.2) for $u \leftarrow u^*$, p^* be the corresponding solution to (3.5) for $y \leftarrow y^*$ and $u \leftarrow u^*$ and p_k be the solution to (3.10) with $u_1 \leftarrow u_k$, $y_1 \leftarrow y_k$ and $y_2 \leftarrow y^*$. If*

$$\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)}$$

and f is bounded on $Z_i \times I \times K_U$, then the following holds

$$\lim_{k\to\infty} \frac{1}{|S_k(z_0)|} \left(J(y_k, u_k) - J(y^*, u^*)\right) = H(z, y^*, u, p^*) - H(z, y^*, u^*, p^*)$$

for almost all $z_0 \in Z_i$.

Proof. With Lemma 22, we have

$$\begin{aligned} J(y_k, u_k) - J(y^*, u^*) &= \int_{Z_i} H(z, y^*, u_k, p_k) - H(z, y^*, u^*, p_k)\, dz \\ &= \int_{S_k(z_0)\cap Z_i} H(z, y^*, u, p_k) - H(z, y^*, u^*, p_k)\, dz \qquad (3.11) \\ &= \int_{S_k(z_0)\cap Z_i} H(z, y^*, u, p^*) - H(z, y^*, u^*, p^*) + (p_k - p^*) f(z, y^*, u) + (p^* - p_k)(f(z, y^*, u^*))\, dz. \end{aligned}$$

We multiply both sides of (3.11) with $\frac{1}{|S_k(z_0)|}$ and apply the limit for k to both sides. Then we obtain

$$\lim_{k\to\infty} \frac{1}{|S_k(z_0)|} \left(J(y_k, u_k) - J(y^*, u^*)\right) = H(z, y^*, u, p^*) - H(z, y^*, u^*, p^*)$$

because with our two requirements that it holds $\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)}$ and that f is bounded on $Z_i \times I \times K_U$, we have that

$$\begin{aligned} &\lim_{k\to\infty} \frac{1}{|S_k(t_0)|} \left| \int_{S_k(t_0)\cap Z_i} (p_k - p^*) f(y^*, u)\, dz \right| \\ &\leq \lim_{k\to\infty} \left(\|p_k - p^*\|_{L^\infty} \frac{1}{|S_k(z_0)|} \int_{S_k(z_0)\cap Z_i} |f_i(z, y^*, u)|\, dz \right) = 0 \end{aligned}$$

and analogously

$$\lim_{k\to\infty} \frac{1}{|S_k(z_0)|} \left| \int_{S_k(z_0)\cap Z_i} (p^* - p_k) f(z, y^*, u^*)\, dz \right| = 0$$

for almost all $z_0 \in Z_i$ considering the limit rules [3, II Remark 2.1 (a)], [3, II Theorem 2.4], [3, Theorem 1.10] and the mean value theorem [15, Theorem 5.6.2]. We remark that the union of countably many null sets is a null set, see [5, IX Remark 2.5 (b)]. □

Remark 24. For the proof of Lemma 23 it is sufficient that $z \mapsto f(z, y(z), u(z))$ is locally integrable for all y solving (3.2) and any $u \in U_{ad}$. However, using the boundedness of f on $Z_i \times I \times K_U$ is reasonable for our purpose since it is fulfilled for P.1) to P.5) as Z_i, I and K_U are bounded. Especially the boundedness of I is shown below.

Now, we have the following theorem that characterizes a solution to (3.3).

Theorem 25. *Let $(\bar{y}, \bar{u})$ be a solution to (3.3). Then under the assumptions of Lemma 23 it holds that*

$$H(z, \bar{y}, \bar{u}, \bar{p}) = \min_{w\in K_U} H(z, \bar{y}, w, \bar{p}) \qquad (3.12)$$

for almost all $z \in Z_i$ where $\bar{p}$ is a solution to (3.5) with $y \leftarrow \bar{y}$ and $u \leftarrow \bar{u}$.

Proof. As we have that $J(\tilde{y},\tilde{u}) \geq J(\bar{y},\bar{u})$ for all $(\tilde{y},\tilde{u})$ solving (3.3) with $\tilde{u} \in U_{ad}$, we especially have that

$$J(y_k, u_k) \geq J(\bar{y},\bar{u}) \tag{3.13}$$

for any solution (y_k, u_k) to (3.2) as $u_k \in U_{ad}$. This can be seen as follows. The sum and the product of measurable functions is measurable, see [36, Proposition 2.1.7]. The needle variation (3.8) can be written as $u_k = u^* \chi_{Z_i \backslash S_k(z_0)} + u\chi_{S_k(z_0)\cap Z_i}$. Since the characteristic function χ_A is measurable if and only if A is measurable, see [36, Example 2.1.2] the needle variation is Lebesgue measurable, since $Z_i \backslash S_k(z_0)$ and $S_k(z_0) \cap Z_i$ are Lebesgue measurable, see [36, Theorem 1.3.6]. Furthermore it holds pointwise $u_k \in K_U$ almost everywhere and thus we have by

$$\begin{aligned}\sum_{j=1}^{m} \int_{Z_i} \left((u_k)_j(z)\right)^q dz &= \sum_{j=1}^{m} \int_{Z_i \backslash S_k(z_0)} \left((u^*)_j(z)\right)^q dz + \int_{S_k(z_0)\cap Z_i} u_j^q dz \\ &\leq \sum_{j=1}^{m} \left(\int_{Z_i} \left((u^*)_j(z)\right)^q dz + u_j^q |S_k(z_0)|\right)\end{aligned}$$

the L^q-integrability since $u^* \in U_{ad}$ and u_j are real numbers for all $j \in \{1, ..., m\}$. Then we have from (3.13) that it holds $J(y_k,u_k) - J(\bar{y},\bar{u}) \geq 0$ and consequently $\frac{1}{|S_k(z_0)|}(J(y_k,u_k) - J(\bar{y},\bar{u})) \geq 0$. Thus we obtain

$$0 \leq \lim_{k\to\infty} \frac{1}{|S_k(z_0)|}(J(y_k,u_k) - J(\bar{y},\bar{u})) = H(z_0,\bar{y},u,\bar{p}) - H(z_0,\bar{y},\bar{u},\bar{p}), \tag{3.14}$$

see Lemma 23 and [3, II Theorem 2.7] for almost all $z_0 \in Z_i$. From (3.14) and by renaming z_0 into z, we conclude that $H(z,\bar{y},\bar{u},\bar{p}) \leq H(z,\bar{y},u,\bar{p})$ for almost all $z \in Z_i$ and all $u \in K_U$ which is equivalent to

$$H(z,\bar{y},\bar{u},\bar{p}) = \min_{w\in K_U} H(z,\bar{y},w,\bar{p}).$$

□

For the last part of this section we denote with u_k the needle variation defined in (3.8) and y_k the solution to (3.2) for $u \leftarrow u_k$ as well as p^* the solution to (3.5) with $y \leftarrow y^*$ and $u \leftarrow u^*$ where y^* is the solution to (3.2) for $u \leftarrow u^* \in U_{ad}$. Furthermore we denote with p_k the solution to (3.10) with $u_1 \leftarrow u_k$, $y_1 \leftarrow y_k$ and $y_2 \leftarrow y^*$. Summarizing if we can show that (3.5) and (3.10) are uniquely solvable and $\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$, then we have that Theorem 25 holds for a solution to an optimal control problem of the class (3.3).

We remark that a general proof of the PMP characterization of solutions to an optimal control problem constraint by semi-linear parabolic PDEs can be found in [81].

Next, we specify our cases. We use that $K_U \subseteq \mathbb{R}$ is a compact set and for the bilinear cases, that means P.3) and P.4) in addition $K_U \subseteq \mathbb{R}_0^+$. Furthermore, we have that

$$U_{ad} = \{u \in L^q(Z_i) \,|\, u(z) \in K_U\}$$

and $q = 2$ for $n = 1$ and for $n \geq 2$ we have $q > \frac{n}{2} + 1$ in the parabolic case. In the elliptic case we have $q = 2$ for $n = 1$ and $q \geq \frac{n}{2} + 1$ for $n \geq 2$. Then we require that $y_d \in L^q(Z_i)$. For P.1) to P.5), we start to argue that I is bounded. This is in fact the case considering Theorem 64, Theorem 65, Remark 61, Theorem 59 and Theorem 60.

Next, we define the following terms. If we write the "linear case", we refer to P.1) and P.2). If we write the "bilinear case", we mean P.3) and P.4) and if we write the "non-linear case", we refer to P.5). Now we formulate the corresponding adjoint equations according to (3.5) where we always have that $p(\cdot,T) = 0$. In the following of the section, we have that $D > 0$ for the parabolic case and $D = 1$ in the elliptic case. For the linear case, we have

$$-(p'(\cdot,t), v) + D(\nabla p(\cdot,t), \nabla v) = \int_\Omega (y(x,t) - y_d(x,t))\, v(x)\, dx. \tag{3.15}$$

For the bilinear case, we have

$$-\left(p'\left(\cdot,t\right),v\right)+D\left(\nabla p\left(\cdot,t\right),\nabla v\right)=\int_{\Omega}\left(\left(y\left(x,t\right)-y_d\left(x,t\right)\right)-u\left(x,t\right)p\left(x,t\right)\right)v\left(x\right)dx \tag{3.16}$$

and for the non-linear case, we have

$$\left(\nabla p,\nabla v\right)=\int_{\Omega}\left(\left(y\left(x\right)-y_d\left(x\right)\right)-3y^2\left(x\right)p\left(x\right)\right)v\left(x\right)dx. \tag{3.17}$$

Next, we formulate the intermediate adjoint equations according to (3.10) where $\tilde{p}\left(\cdot,T\right)=0$. For the linear case, we have

$$-\left(\tilde{p}'\left(\cdot,t\right),v\right)+D\left(\nabla\tilde{p}\left(\cdot,t\right),\nabla v\right)=\int_{\Omega}\left(\frac{1}{2}\left(y_1\left(x,t\right)+y_2\left(x,t\right)\right)-y_d\left(x,t\right)\right)v\left(x\right)dx. \tag{3.18}$$

For the bilinear case, we have that

$$-\left(\tilde{p}'\left(\cdot,t\right),v\right)+D\left(\nabla\tilde{p}\left(\cdot,t\right),\nabla v\right)=\int_{\Omega}\left(\frac{1}{2}\left(y_1\left(x,t\right)+y_2\left(x,t\right)\right)-y_d\left(x,t\right)-u_1\left(x,t\right)p\left(x,t\right)\right)v\left(x\right)dx \tag{3.19}$$

and for the non-linear case we have

$$\left(\nabla\tilde{p},\nabla v\right)=\int_{\Omega}\left(\frac{1}{2}\left(y_1\left(x\right)+y_2\left(x\right)\right)-y_d\left(x\right)-\left(y_1^2\left(x\right)+y_1\left(x\right)y_2\left(x\right)+y_2^2\left(x\right)\right)\tilde{p}\left(x\right)\right)v\left(x\right)dx \tag{3.20}$$

where it holds that

$$0\leq\int_0^1 3\left(y_2\left(x\right)+\theta\left(y_1\left(x\right)-y_2\left(x\right)\right)\right)^2d\theta=y_1^2\left(x\right)+y_1\left(x\right)y_2\left(x\right)+y_2^2\left(x\right).$$

Since $y,y_1,y_2\in L^\infty\left(Z_i\right)$ for P.1) to P.5), see discussion above starting on page 75, and with [1, Theorem 2.14], we have that (3.15) and (3.18) admit a unique solution in $H_0^1\left(\Omega\right)$ for the elliptic case or in $L^2\left(0,T,H^2\left(\Omega\right)\right)\cap L^\infty\left(0,T;H_0^1\left(\Omega\right)\right)$ in the parabolic case, respectively, with analogous arguments as in the discussion about the linear case for the state equation (3.2), see Section 3.1. The other cases, that means (3.16), (3.17), (3.19) and (3.20), can be discussed as the bilinear case for the state equation (3.2) since $u,u_1,y^2\geq 0$ and $y_1^2\left(x\right)+y_1\left(x\right)y_2\left(x\right)+y_2^2\left(x\right)\geq 0$ where we also have a unique solution in $H_0^1\left(\Omega\right)$ for the elliptic case or in $L^2\left(0,T,H^2\left(\Omega\right)\right)\cap L^\infty\left(0,T;H_0^1\left(\Omega\right)\right)$ in the parabolic case, respectively.

For the same reason, the corresponding results Theorem 64, Theorem 65, Remark 61, Theorem 59 and Theorem 60 also hold for (3.15) to (3.20) and thus we have the corresponding boundedness results for the adjoint and intermediate adjoint variable.

Since $|u-u^*|^q$ is integrable on every ball that is contained in Z_i, we have that

$$\begin{aligned}\lim_{k\to\infty}\|u_k-u^*\|_{L^q(Z_i)}^q&=\lim_{k\to\infty}\int_{Z_i}|u_k\left(z\right)-u^*\left(z\right)|^q dz=\lim_{k\to\infty}\int_{S_k(z_0)\cap Z_i}|u-u^*\left(z\right)|^q dz\\&=\lim_{k\to\infty}S_k\left(z_0\right)\frac{1}{S_k\left(z_0\right)}\int_{S_k(z_0)\cap Z_i}|u-u^*\left(z\right)|^q dz\\&=\lim_{k\to\infty}S_k\left(z_0\right)\lim_{k\to\infty}\frac{1}{S_k\left(z_0\right)}\int_{S_k(z_0)\cap Z_i}|u-u^*\left(z\right)|^q dz\\&=0\cdot|u-u^*\left(z_0\right)|^q=0\end{aligned}$$

for almost all $z_0\in Z_i$ according to the mean value theorem [15, Theorem 5.6.2] and the limit rules [3, II Theorem 2.4].

Next we prove the condition

$$\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)} = 0$$

for almost all $z_0 \in Z_i$ for P.1) to P.5). We start proving that in the linear, in the bilinear and in the non-linear case we have that

$$\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(Z_i)} = 0$$

for almost all $z_0 \in Z_i$. This is proved by subtracting the corresponding state equation (3.2) for $u \leftarrow u^*$ from (3.2) for $u \leftarrow u_k$. In the linear case we have

$$\left((y_k - y^*)', v\right) + D\left(\nabla\left(y_k - y^*\right), \nabla v\right) = \left(u_k - u^*, v\right)$$

where we obtain with Theorem 59 in the elliptic case or Theorem 64 in the parabolic case that $\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$. For the bilinear case, we that

$$\left((y_k - y^*)', v\right) + D\left(\nabla\left(y_k - y^*\right), \nabla v\right) + \left(u_k y_k - u^* y^*, v\right) = 0$$

which is equivalently given by

$$\left((y_k - y^*)', v\right) + D\left(\nabla\left(y_k - y^*\right), \nabla v\right) + \left(u_k\left(y_k - y^*\right), v\right) = \left(y^*\left(u^* - u_k\right), v\right)$$

where we have according to Theorem 60 in the elliptic or Theorem 65 in the parabolic case with $\tilde{f} \leftarrow u^* - u_k$ that $\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$. In the non-linear case we have that

$$\left(\nabla\left(y_k - y^*\right), \nabla v\right) + \left(y_k^3 - (y^*)^3, v\right) = \left(u_k - u^*, v\right)$$

which is equivalently given by

$$\left(\nabla\left(y_k - y^*\right), \nabla v\right) + \left(\int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta \left(y_k - y^*\right), v\right) = \left(u_k - u^*, v\right)$$

using the fundamental theorem of calculus [4, VI 4.13] for $\theta \mapsto \left(y^* + \theta\left(y_k - y^*\right)\right)^3$. Since

$$\int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta \geq 0,$$

we have by Remark 61 for $y \leftarrow y_k - y^*$ that $\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(\Omega)} = 0$ for almost all $z_0 \in Z_i$ in the non-linear case.

In the next step, we prove for the linear, the bilinear and the non-linear case that $\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$ where we have that $\tilde{p} \leftarrow p_k$. In the linear case, we have that

$$-\left((p_k - p^*)', v\right) + D\left(\nabla\left(p_k - p^*\right), \nabla v\right) = \left(\frac{1}{2}\left(y_k - y^*\right), v\right)$$

where we have with the same argumentation as above that $\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$ where we use that $\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$ and [1, Theorem 2.14]. In the bilinear case, we have that

$$-\left((p_k - p^*)', v\right) + D\left(\nabla\left(p_k - p^*\right), \nabla v\right) = \left(\frac{1}{2}\left(y_k - y^*\right) - u_k p_k + u^* p^*\right)$$

which is equivalently given by

$$-\left((p_k - p^*)', v\right) + D\left(\nabla\left(p_k - p^*\right), \nabla v\right) + \left(u_k\left(p_k - p^*\right), v\right) = \left(\frac{1}{2}\left(y_k - y^*\right) + p^*\left(u^* - u_k\right), v\right).$$

By using the triangle inequality [36, Proposition 3.3.3], we have with an analogous discussion as for the difference of the state equations that $\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$ in the bilinear case. In the non-linear case, we have

$$\begin{aligned}
&\left(\nabla\left(p_k - p^*\right), \nabla v\right)\\
&= \left(\frac{1}{2}\left(y_k - y^*\right) - \int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta p_k + 3\left(y^*\right)^2 p^*, v\right)\\
&= \left(\frac{1}{2}\left(y_k - y^*\right) - \int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta\left(p_k - p^*\right) - \int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta p^* + 3\left(y^*\right)^2 p^*, v\right)\\
&= \left(\frac{1}{2}\left(y_k - y^*\right) - \int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta\left(p_k - p^*\right) - \int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 - 3\left(y^*\right)^2 d\theta p^*, v\right)\\
&= \left(\frac{1}{2}\left(y_k - y^*\right) - \int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta\left(p_k - p^*\right) - \int_0^1 6y^*\theta\left(y_k - y^*\right) + 3\theta^2\left(y_k - y^*\right)^2 d\theta p^*, v\right)
\end{aligned}$$

which equivalently gives

$$\begin{aligned}
&\left(\nabla\left(p_k - p^*\right), \nabla v\right) + \left(\int_0^1 3\left(y^* + \theta\left(y_k - y^*\right)\right)^2 d\theta\left(p_k - p^*\right), v\right)\\
&= \left(\left(\frac{1}{2} + 3y^*\right)\left(y_k - y^*\right) + \left(y_k - y^*\right)^2, v\right).
\end{aligned} \tag{3.21}$$

With the triangle inequality [36, Proposition 3.3.3] and that since $y_k - y^* \in L^\infty\left(Z_i\right)$ the function $y_k - y^* \in L^{2q}\left(Z_i\right)$, we obtain from (3.21) that $\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(Z_i)} = 0$ for almost all $z_0 \in Z_i$ in the non-linear case with Remark 61 and [1, Theorem 2.14].

3.3 Convergence analysis of the SQH scheme

In this section, we discuss convergence of the SQH scheme in the PDE case to a PMP solution. For this purpose, we define the following augmented Hamiltonian

$$K_\epsilon\left(z, y, u, v, p\right) := H\left(z, y, u, p\right) + \epsilon\left(u\left(z\right) - v\left(z\right)\right)^2 \tag{3.22}$$

where $K_\epsilon : Z_i \times \mathbb{R} \times K_U \times K_U \times \mathbb{R} \to \mathbb{R}$ with $\epsilon > 0$ and

$$\left(u\left(z\right) - v\left(z\right)\right)^2 := \sum_{j=1}^{m}\left(u_j\left(z\right) - v_j\left(z\right)\right)^2.$$

We use the notation $K_\epsilon\left(z, y, u, v, p\right) := K_\epsilon\left(z, y\left(z\right), u\left(z\right), v\left(z\right), p\left(z\right)\right)$ whenever an argument of K_ϵ is a function instead of a number.

The SQH scheme is implemented as follows.

Algorithm 3.1 (SQH method)

```
1. Choose $\epsilon > 0$, $\kappa > 0$, $\sigma > 1$, $\zeta \in (0,1)$, $\eta \in (0,\infty)$, $u^0 \in U_{ad}$, compute $y^0$ by (3.2) for $u \leftarrow u^0$ and $p^0$ by (3.5) for $y \leftarrow y^0$ and $u \leftarrow u^0$, set $k \leftarrow 0$
2. Choose $u \in K_U$ such that
   $K_\epsilon\left(z, y^k, u, u^k, p^k\right) \leq K_\epsilon\left(z, y^k, w, u^k, p^k\right)$
   for all $w \in K_U$ and all $z \in Z_i$
3. Calculate $y$ by (3.2) for $u$ and $\tau := \|u - u^k\|^2_{L^2(Z_i)}$
4. If $J(y,u) - J\left(y^k, u^k\right) > -\eta\tau$: Choose $\epsilon \leftarrow \sigma\epsilon$
   Else:
   Choose $\epsilon \leftarrow \zeta\epsilon$, $y^{k+1} \leftarrow y$, $u^{k+1} \leftarrow u$, calculate $p^{k+1}$ by (3.5) for $y \leftarrow y^{k+1}$ and $u \leftarrow u^{k+1}$, set $k \leftarrow k+1$
5. If $\tau < \kappa$: STOP and return $u^k$
   Else go to 2.
```

The controls u obtained in Step 2 of Algorithm 3.1 are measurable for all the problems considered in this thesis. For a detailed discussion about the measurability of u, obtained in Step 2 of Algorithm 3.1, see the Appendix page 166 and the following pages.

The description of the single steps of Algorithm 3.1 is as in Section 2.3 for the ODE case. In particular, Lemma 7, which says that K_ϵ attains a minimum, holds analogously in this case. Also the next lemma has an equivalent in the ODE case, which is Lemma 11. However, since the assumptions in the present chapter differ from Chapter 2 the proof is different and that is why we present it here.

Lemma 26. *Let (y,u) and $\left(y^k,u^k\right)$ be generated by Algorithm 3.1, $k \in \mathbb{N}_0$, denote $\delta u := u - u^k$. Then there is a $\theta > 0$ independent of ϵ such that for the $\epsilon > 0$ currently chosen by Algorithm 3.1, the following holds*

$$J(y,u) - J\left(y^k, u^k\right) \leq -(\epsilon - \theta)\,\|\delta u\|^2_{L^2(Z_i)}.$$

In particular, $J(y,u) - J\left(y^k,u^k\right) \leq 0$ for $\epsilon \geq \theta$.

Proof. We define $(\delta u)^2 := \sum_{j=1}^m (\delta u_j(z))^2$, $\delta y := y(z) - y^k(z)$ and $\delta p := p(z) - p^k(z)$ where p is calculated by (3.5) for y and u. Furthermore, to save notational effort, we note $H := H(z,y,u,p)$ or $H^k := H\left(z,y^k,u^k,p^k\right)$ and drop the functional dependency of the functions y, y^k, u, u^k, p and p^k as well as we write $f := f(z,y,u)$, $f^k := f\left(z,y^k,u^k\right)$, $h^k := h\left(y^k\right)$ and $h := h(y)$ for all $k \in \mathbb{N}_0$. We use from Algorithm 3.1 that u is determined such that

$$K_\epsilon\left(z, y^k, u, u^k, p^k\right) \leq K_\epsilon\left(z, y^k, w, u^k, p^k\right)$$

for all $w \in K_U$ and thus it holds in particular that

$$K_\epsilon\left(z, y^k, u, u^k, p^k\right) \leq K_\epsilon\left(z, y^k, u^k, u^k, p^k\right) = H\left(z, y^k, u^k, p^k\right)$$

for all $z \in Z_i$. We start the proof as follows

$$\begin{aligned}
&J(y,u) - J\left(y^k,u^k\right) = \int_{Z_i} h(y) + g(u) - h\left(y^k\right) - g\left(u^k\right) dz = \int_{Z_i} H - pf - H^k + p^k f^k dz \\
&= \int_{Z_i} H - H\left(z,y^k,u,p^k\right) + H\left(z,y^k,u,p^k\right) + \epsilon\left(\delta u\right)^2 - H^k - \epsilon\left(\delta u\right)^2 dz \\
&- \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt \\
&\leq \int_{Z_i} H - H\left(z,y^k,u,p^k\right) - \epsilon\left(\delta u\right)^2 dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt.
\end{aligned} \tag{3.23}$$

Next, we estimate the term

$$|\int_{Z_i} H - H\left(z,y^k,u,p^k\right) dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt|.$$

For this purpose, we first consider

$$\int_0^T \left(\delta y', \delta p\right) + B\left(\delta y, \delta p; t\right) dt = \int_{Z_i} \left(f\left(z,y,u\right) - f\left(z,y^k,u^k\right)\right) \delta p dz$$

which can be estimated as follows

$$\begin{aligned}
&|\int_0^T \left(\delta y', \delta p\right) + B\left(\delta y, \delta p; t\right) dt| \leq \int_{Z_i} |f\left(z,y,u\right) - f\left(z,y^k,u^k\right)||\delta p| dz \\
&= \int_{Z_i} |f\left(z,y,u\right) - f\left(z,y^k,u\right) + f\left(z,y^k,u\right) - f\left(z,y^k,u^k\right)||\delta p| dz \\
&\leq \int_{Z_i} \int_0^1 |\frac{\partial}{\partial y} f\left(z,y,u\right)|_{y=y^k+\theta\left(y-y^k\right)}| d\theta |\delta y||\delta p| + c\sum_{j=1}^m |\delta u_j||\delta p| dz \\
&\leq c\|\delta y\|_{L^2(Z_i)}\|\delta p\|_{L^2} + c\|\delta u\|_{L^2(Z_i)}\|\delta p\|_{L^2(Z_i)} \\
&\leq \left(c^3 + c^2\right)\|\delta u\|^2_{L^2(Z_i)}
\end{aligned} \tag{3.24}$$

using the fundamental theorem of calculus [4, VI 4.13] for $\theta \mapsto f\left(z, y^k + \theta\left(y - y^k\right), u^k\right)$ and Assumption A.2), Assumption A.3), Assumption A.5) and the Cauchy-Schwarz inequality [2, Lemma 2.2]. Using the Taylor formula [4, VII Theorem 5.8] and with the symmetry of the second derivative [4, VII Theorem 5.2], we obtain by (3.24) the following

$$\begin{aligned}
&|\int_{Z_i} H - H\left(z, y^k, u, p^k\right) dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt| \\
&= |\int_{Z_i} H - H\left(z, y - \delta y, u, p - \delta p\right) dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt| \\
&= |\int_{Z_i} \frac{\partial}{\partial y} H \delta y + \frac{\partial}{\partial p} H \partial p dz - \frac{1}{2}\int_{Z_i} \frac{\partial^2}{\partial y^2} H \left(\delta y\right)^2 + 2\frac{\partial^2}{\partial y \partial p} H \delta y \delta p dz + \int_{Z_i} R_2\left(H, y, p; \delta y, \delta p\right) dz \\
&- \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt| \\
&= |\int_{Z_i} \frac{\partial}{\partial y} h \delta y + p\frac{\partial}{\partial y} f \delta y + f \delta p dz - \frac{1}{2}\int_{Z_i} \left(\frac{\partial^2}{\partial y^2} h + p\frac{\partial^2}{\partial y^2} f\right) \left(\delta y\right)^2 + 2\frac{\partial}{\partial y} f \delta y \delta p dz \qquad (3.25) \\
&+ \int_{Z_i} R_2\left(H, y, p; \delta y, \delta p\right) dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt| \\
&= |\int_0^T - \left(\left(p\right)', \delta y\right) + B^*\left(p, \delta y; t\right) dt - \int_0^T \left(\left(\delta y\right)', p^k\right) + B\left(\delta y, p^k; t\right) dt \\
&- \frac{1}{2}\int_{Z_i} \left(\frac{\partial^2}{\partial y^2} h + p\frac{\partial^2}{\partial y^2} f\right) \left(\delta y\right)^2 + 2\frac{\partial}{\partial y} f \delta y \delta p dz + \int_{Z_i} R_2\left(H, y, p; \delta y, \delta p\right) dz| \\
&\leq \left(c^2 + \frac{9}{2}c^3 + \frac{3}{2}c^4\right) \|\delta u\|^2_{L^2(Z_i)}
\end{aligned}$$

where we use the partial integration rule [95, Theorem 3.11], the Cauchy-Schwarz inequality [2, Lemma 2.2] in the last inequality for the term $\int_{Z_i} 2\frac{\partial}{\partial y} f \delta y \delta p dz$ and that the Taylor remainder $R_2\left(H, y, p; \delta y, \delta p\right)$ is estimated by the remainder formula [4, VII Theorem 5.8] and the boundedness of the second derivatives analogously to the calculation which are done for the second derivatives in (3.25). Combining (3.23) and (3.25), we obtain

$$\begin{aligned}
&J\left(y, u\right) - J\left(y^k, u^k\right) \\
&\leq \int_{Z_i} H - H\left(z, y^k, u, p^k\right) - \epsilon \left(\delta u\right)^2 dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\delta y', p^k\right) + B\left(\delta y, p^k\right) dt \\
&\leq |\int_{Z_i} H - H\left(z, y^k, u, p^k\right) dz - \int_{Z_i} \delta p f dz - \int_0^T \left(\delta y', p^k\right) + B\left(\delta y, p^k\right) dt| - \int_{Z_i} \epsilon \left(\delta u\right)^2 dz \\
&\leq \left(c^2 + \frac{9}{2}c^3 + \frac{3}{2}c^4\right) \|\delta u\|^2_{L^2(Z_i)} - \int_{Z_i} \epsilon \left(\delta u\right)^2 dz = \left(\theta - \epsilon\right) \|\delta u\|^2_{L^2(Z_i)}
\end{aligned}$$

where $\theta := c^2 + \frac{9}{2}c^3 + \frac{3}{2}c^4$. □

For the investigation of the sequence $\left(y^k\right)_{k\in\mathbb{N}_0}$ and $\left(u^k\right)_{k\in\mathbb{N}_0}$ generated by the iterated Steps 2 to 4 of Algorithm 3.1 (no stopping criterion), Lemma 12, which says that Algorithm 3.1 stops if an iterate u^k is optimal in the sense of (3.7), and Theorem 13, stating a minimizing property of the sequence $\left(u^k\right)_{k\in\mathbb{N}_0}$, hold in the corresponding form as well in the PDE case.

Next, also in the present case, we have the requirement that for any iterate u^k, $k \in \mathbb{N}_0$ and for any ϵ chosen by Algorithm 3.1 there exists an $r \geq \epsilon$ such that

$$K_\epsilon\left(z, y^k, u^{k+1}, u^k, p^k\right) + r\left(w - u^{k+1}\left(z\right)\right)^2 \leq K_\epsilon\left(z, y^k, w, u^k, p^k\right) \tag{3.26}$$

is fulfilled for all $w \in K_U$ and for all $z \in Z_i$. We show in Lemma 28 that this condition is always satisfied for P.1) to P.5) with some further assumptions.

The next theorem has also an equivalent in Section 2.3, which is Theorem 14. Due to some differences in the requirements in the PDE case compared to the ODE case we give the proof of the following theorem.

Theorem 27. *Let the sequence $(u^n)_{n\in\mathbb{N}_0}$ be generated as in Algorithm 3.1 (loop over Step 2 to Step 4) and let (3.26) hold. Then for any subsequence $\left(u^k\right)_{k\in K}$, $K\subseteq\mathbb{N}_0$ with the property*

$$\lim_{k\to\infty}\|u^k-\bar{u}\|_{L^2(Z_i)}=0$$

it holds that $\bar{u}\in U_{ad}$ and

$$H\left(z,\bar{y},\bar{u},\bar{p}\right)=\min_{w\in K_U}H\left(z,\bar{y},w,\bar{p}\right)$$

for almost all $z\in Z_i$ where $\bar{y}$ solves (3.2) with $u\leftarrow\bar{u}$ and $\bar{p}$ is the corresponding adjoint variable solving (3.5) for $y\leftarrow\bar{y}$ and $u\leftarrow\bar{u}$.

Furthermore, for almost each $z\in Z_i$ and any $\mu>0$, there exists an index set $\tilde{K}\subseteq K$ and a $\bar{k}\in\tilde{K}$ such that

$$H\left(z,y^{k+1},u^{k+1},p^{k+1}\right)\leq H\left(z,y^{k+1},w,p^{k+1}\right)+\mu \tag{3.27}$$

for all $w\in K_U$ and for all $k\geq\bar{k}$ with $k\in\tilde{K}$.

Proof. We construct a subsequence having all the properties that we need for the proof. By [6, Proposition 3.6, Remark 3.7], we have that there exists an index set $K_1\subseteq K$ such that

$$\lim_{K_1\ni k\to\infty}u^k\left(z\right)=\bar{u}\left(z\right),\quad\lim_{K_1\ni k\to\infty}y^k\left(z\right)=\bar{y}\left(z\right)\text{ and }\lim_{K_1\ni k\to\infty}p^k\left(z\right)=\bar{p}\left(z\right)$$

for almost all $z\in Z_i$ due to $\lim_{k\to\infty}\|u^k-\bar{u}\|_{L^2(Z_i)}=0$, Assumption A.2) and since any subsequence of a converging sequence also converges, see [3, II Theorem 1.15].

The iterates u^k, $k\in K_1$ are measurable, see the discussion below Algorithm 3.1. Thus $\bar{u}$ is measurable, see [5, X Theorem 1.14]. Since $u^k\left(z\right)\in K_U$ for almost all $z\in Z_i$ we have that $\bar{u}\left(z\right)\in K_U$ for almost all $z\in Z_i$, see [3, II Theorem 2.7]. Since K_U is bounded we have that $\bar{u}\in U_{ad}$ because $\bar{u}$ is integrable.

Because Theorem 13 holds analogously for the PDE case from which we have

$$\lim_{n\to\infty}\|u^{n+1}-u^n\|_{L^2(Z_i)}=0,$$

we have

$$\lim_{K_1\ni k\to\infty}\|u^{k+1}-u^k\|_{L^2(Z_i)}=0$$

since u^{k+1} is the following element of u^k in the sequence $(u^n)_{n\in\mathbb{N}_0}$. Consequently by [6, Proposition 3.6, Remark 3.7], we have a subsequence $K_2\subseteq K_1$ such that

$$\lim_{K_2\ni k\to\infty}u^{k+1}\left(z\right)-u^k\left(z\right)=0$$

for almost all $z\in Z_i$ where all the other properties above remain since any subsequence of a converging sequence also converges, see [3, II Theorem 1.15]. From this we can also conclude that

$$\lim_{K_2\ni k\to\infty}u^{k+1}\left(z\right)=\lim_{K_2\ni k\to\infty}\left(u^{k+1}\left(z\right)-u^k\left(z\right)\right)+\lim_{K_2\ni k\to\infty}u^k\left(z\right)=\bar{u}\left(z\right) \tag{3.28}$$

for almost all $z\in Z_i$ where we use the calculation rules for the limit [3, II Theorem 2.2]. Analogous we have with Assumption A.2) another index set $K_3\subseteq K_2$ such that $\lim_{K_3\ni k\to\infty}y^{k+1}\left(z\right)-y^k\left(z\right)=0$, $\lim_{K_3\ni k\to\infty}p^{k+1}\left(z\right)-p^k\left(z\right)=0$ and thus

$$\lim_{K_3\ni k\to\infty}y^{k+1}\left(z\right)=\bar{y}\left(z\right)\text{ and }\lim_{K_3\ni k\to\infty}p^{k+1}\left(z\right)=\bar{p}\left(z\right)$$

for almost all $z \in Z_i$.

As the control u^k, $k \in K_3$ is an element of $\left(u^n\right)_{n\in\mathbb{N}_0}$, the control u^k is determined by Algorithm 3.1 such that due to (3.26) the following holds

$$K_\epsilon\left(z, y^k, u^{k+1}, u^k, p^k\right) + r\left(w - u^{k+1}(z)\right)^2 \leq K_\epsilon\left(z, y^k, w, u^k, p^k\right)$$

for all $w \in K_U$, for all $k \in \mathbb{N}_0$ and all $z \in Z_i$ which is equivalent to

$$H\left(z, y^k, u^{k+1}, p^k\right) + \epsilon\left(u^{k+1}(z) - u^k(z)\right)^2 + r\left(w - u^{k+1}(z)\right)^2 \leq H\left(z, y^k, w, p^k\right) + \epsilon\left(w - u^k(z)\right)^2. \tag{3.29}$$

Now, we consider (3.29) where it also holds due to our assumption $r \geq \epsilon$ that

$$H\left(z, y^k, u^{k+1}, p^k\right) + \epsilon\left(u^{k+1}(z) - u^k(z)\right)^2 + \epsilon\left(w - u^{k+1}(z)\right)^2 \leq H\left(z, y^k, w, p^k\right) + \epsilon\left(w - u^k(z)\right)^2$$

and thus by inserting

$$\left(w - u^{k+1}(z)\right)^2 = \left(w - u^k(z)\right)^2 + \left(u^k(t) - u^{k+1}(z)\right)^2 + 2\left(w - u^k(z)\right)^T\left(u^k(z) - u^{k+1}(z)\right)$$

we obtain

$$H\left(z, y^k, u^{k+1}, p^k\right) + 2\epsilon\left(u^{k+1}(z) - u^k(z)\right)^2 + 2\epsilon\left(w - u^k(z)\right)^T\left(u^k(z) - u^{k+1}(z)\right) \leq H\left(z, y^k, w, p^k\right) \tag{3.30}$$

for all $w \in K_U$, for all $k \in \mathbb{N}_0$ and all $z \in Z_i$. Then (3.30) is equivalent to

$$\begin{aligned} &h\left(y^k(z)\right) + g\left(u^{k+1}(z)\right) + p^k(z) f\left(z, y^k, u^{k+1}\right) + 2\epsilon\left(u^{k+1}(z) - u^k(z)\right)^2 \\ &+ 2\epsilon\left(w - u^k(z)\right)^T\left(u^k(z) - u^{k+1}(z)\right) \leq h\left(y^k(z)\right) + g(w) + p^k(z) f\left(z, y^k, w\right) \end{aligned} \tag{3.31}$$

for all $w \in K_U$, for all $k \in \mathbb{N}_0$ and all $z \in Z_i$. Next, we have that ϵ is bounded from below by 0 and from above by $\sigma(\eta + \theta)$ analogous to the proof of Theorem 14. The boundedness of ϵ guarantees that the corresponding terms go to zero for k to infinity, see [3, Theorem 2.4, Theorem 6.1] since $u^k(z) - u^{k+1}(z)$ converges pointwise for $k \in K_3$ and $\left(w - u^k(z)\right)$ is also bounded as $w, u^k \in K_U$ for all $k \in \mathbb{N}_0$. This connection is exploited in the next step. Since g is lower semi-continuous, we apply the lim inf on both sides of the last inequality (3.31) with $k \in K_3$ and recall that whenever a lim exists the corresponding lim inf equals lim, see [3, Theorem 5.7] and the calculation rules for a sum of lim inf [43, Theorem 3.127]. Further, we set $u^{k+1}(z) =: a^{k+1} \to \bar{a} := \bar{u}(z)$ for $K_3 \ni k \to \infty$ and we have

$$\liminf_{K_3 \ni k \to \infty} g\left(u^{k+1}(z)\right) = \liminf_{K_3 \ni k \to \infty} g\left(a^{k+1}\right) \geq g(\bar{a}) = g(\bar{u}(z))$$

for almost all $z \in Z_i$. We obtain for the left-hand side of (3.31) the following

$$\begin{aligned} &\liminf_{K_3 \ni k \to \infty}\Bigg(h\left(y^k(z)\right) + g\left(u^{k+1}(z)\right) + p^k(z) f\left(z, y^k, u^{k+1}\right) + 2\epsilon\left(u^{k+1}(z) - u^k(z)\right)^2 \\ &+2\epsilon\left(w - u^k(z)\right)^T\left(u^k(z) - u^{k+1}(z)\right)\Bigg) \geq h(\bar{y}(z)) + g(\bar{u}(z)) + \bar{p}(z) f(z, \bar{y}, \bar{u}) = H(z, \bar{y}, \bar{u}, \bar{p}) \end{aligned}$$

where we use the continuity of f according to Assumption A.6). For the right-hand side of (3.31), we have

$$\begin{aligned} &\liminf_{K_3 \ni k \to \infty}\left(h\left(y^k(z)\right) + g(w) + p^k(z) f\left(z, y^k, w\right)\right) = \lim_{K_3 \ni k \to \infty}\left(h\left(y^k(z)\right) + g(w) + p^k(z) f\left(z, y^k, w\right)\right) \\ &= h(\bar{y}(z)) + g(w) + \bar{p}(z) f(z, \bar{y}, w) = H(z, \bar{y}, w, \bar{p}) \end{aligned}$$

where we also use the continuity for f, see Assumption A.6) and recall that differentiable functions are continuous, see [78, 1 The Rules of Differentiation]. Consequently, we obtain the optimality condition

$$H\left(z,\bar{y},\bar{u},\bar{p}\right) \leq H\left(z,\bar{y},w,\bar{p}\right)$$

for all $w \in K_U$ and almost all $z \in Z_i$.

In order to prove (3.27), we consider (3.29) inserting the assumption $r \geq \epsilon$ and obtain

$$\begin{aligned} H\left(z,y^{k+1},u^{k+1},p^{k+1}\right) &\leq H\left(z,y^{k+1},w,p^{k+1}\right) \\ &+ \left|p^k\left(z\right)f\left(z,y^k,u^{k+1}\right) - p^{k+1}\left(z\right)f\left(z,y^{k+1},u^{k+1}\right)\right| \\ &+ \left|p^k\left(z\right)f\left(z,y^k,w\right) - p^{k+1}\left(z\right)f\left(z,y^{k+1},w\right)\right| \\ &+ \epsilon\left|\left(\left(w-u^k\left(z\right)\right)^2 - \left(u^{k+1}\left(z\right)-u^k\left(z\right)\right)^2 - \left(w-u^{k+1}\left(z\right)\right)^2\right)\right| \end{aligned} \tag{3.32}$$

by adding and subtracting corresponding terms. Now, by continuity, especially Assumption A.6) and $k \in \tilde{K} := K_3$ it follows the result (3.27) where the last three terms in (3.32) are smaller than any given $\mu > 0$ if k is sufficiently large using the boundedness of ϵ and [3, Theorem 2.4, Theorem 6.1]. □

We remark that an analogous result corresponding to the corollary on page 42 also holds in this case and states the existence of an iterate within the sequence of iterates of Algorithm 3.1 which fulfills the PMP optimality condition for any given tolerance.

In the next lemma, we see that for a further assumption (3.26) is fulfilled. We remark that Example 17 and Example 18 also hold in the present PDE case with an analogous calculation which shows that (3.26) is fulfilled for L^2- and L^1-cost functionals.

Lemma 28. *We consider an augmented Hamiltonian given by*

$$K_\epsilon\left(z,y,u,v,p\right) := \frac{\alpha}{2}u^2 + g\left(u\right) + pf\left(z,y\right) + \epsilon\left(u-v\right)^2$$

where we only include the terms depending on u and $u_a \leq u \leq u_b$, $u_a < 0 < u_b$, with $\alpha > 0$,

$$g\left(u\right) := \beta \begin{cases} |u| & |u| > s \\ 0 & |u| \leq s \end{cases},$$

$s, \beta > 0$. If for all iterations u^k, $k \in \mathbb{N}_0$, generated by Algorithm 3.1, it either holds that $|u^k| \leq s$ or $|u^k| > \theta > s$ with $\frac{\alpha}{2}\left(s-\theta\right)^2 - \beta s \geq 0$, then (3.26) is fulfilled.

Proof. In Algorithm 3.1, we have $K_\epsilon\left(z,y^k,u^{k+1},u^k,p^k\right)$ with $K_\epsilon\left(z,y^k,u^{k+1},u^k,p^k\right) \leq K_\epsilon\left(z,y^k,w,u^k,p^k\right)$ for all $w \in K_U$. We show that (3.26) is fulfilled for all $w \in K_U$. We assume $w \neq u^{k+1}$ because in the case $w = u^{k+1}$, we have that

$$K_\epsilon\left(z,y^k,u^{k+1},u^k,p^k\right) \leq K_\epsilon\left(z,y^k,w,u^k,p^k\right) \tag{3.33}$$

is fulfilled with equality. For u^{k+1}, we have three cases, $s < u^{k+1} \leq u_b$, $|u^{k+1}| \leq s$, $u_b < u^{k+1} < s$ where each is discussed in the following.

If $s < u^{k+1} < u_b$, then we have, analogous to Example 17 or Example 18, that

$$p^k f\left(y^k\right) = 2\epsilon u^k - 2\epsilon u^{k+1} - \alpha u^{k+1} - \beta \tag{3.34}$$

and

$$u^{k+1} = \frac{2\epsilon u^k - p^k f\left(y^k\right) - \beta}{\alpha + 2\epsilon}. \tag{3.35}$$

Then we have from (3.33) that

$$\begin{aligned}&\frac{\alpha}{2}\left(u^{k+1}\right)^2+\beta u^{k+1}+p^k f\left(y^k\right)u^{k+1}+\epsilon\left(u^{k+1}-u^k\right)^2+r\left(w-u^{k+1}\right)^2\\&\leq\frac{\alpha}{2}w^2+g\left(w\right)+p^k f\left(y^k\right)w+\epsilon\left(w-u^k\right)^2\end{aligned}$$

which is equivalent to

$$r\left(w-u^{k+1}\right)^2\leq\left(\frac{\alpha}{2}+\epsilon\right)\left(w-u^{k+1}\right)^2+g\left(w\right)-\beta w$$

inserting (3.34). Now if we choose $r=\epsilon$, we have

$$0\leq\frac{\alpha}{2}\left(w-u^{k+1}\right)^2+g\left(w\right)-\beta w. \tag{3.36}$$

If $w>s$, we have that $g\left(w\right)-\beta w=\beta w-\beta w=0$. If $w\leq 0$, we have that $-\beta w\geq 0$ if additionally $w\geq -s$ or we have that $-\beta w-\beta w\geq 0$ if additionally $w<-s$. If $0<w\leq s$, we have that $\left(w-u^{k+1}\right)^2\geq\left(s-\theta\right)^2$. Due to our requirement that $\frac{\alpha}{2}\left(s-\theta\right)-\beta s\geq 0$, we have that (3.36) is fulfilled. Similar to Example 17 or Example 18 the same arguments hold for the case that $u^{k+1}=u_b$.

If $u_a<u^{k+1}<-s$, the discussion results also in (3.36) except that we have $g\left(w\right)+\beta w$ instead of $g\left(w\right)-\beta w$. If $w\geq 0$, we have that $\beta w+\beta w\geq 0$ if $w>s$ or we have that $\beta w\geq 0$ if in addition $w\leq s$. If $w<-s$, we have that $g\left(w\right)+\beta w=-\beta w+\beta w=0$. If $-s\leq w<0$, we have that $g\left(w\right)+\beta w=\beta w\geq-\beta s$ and $\left(w-u^{k+1}\right)^2\geq\left(\theta-s\right)^2$. Due to our requirement that $\frac{\alpha}{2}\left(s-\theta\right)^2-\beta s\geq 0$, we have that (3.36) is fulfilled. Analogous to the case where $u^{k+1}=u_b$ in Example 17 or Example 18, the discussion holds for the case $u^{k+1}=u_a$.

If $-s<u^{k+1}<s$, then we obtain with the same calculations that

$$0\leq\frac{\alpha}{2}\left(w-u^{k+1}\right)^2+g\left(w\right) \tag{3.37}$$

which is always true as $g\left(w\right)\geq 0$ for all $w\in K_U$. The cases $u^{k+1}=s$ or $u^{k+1}=-s$ result also in (3.37) with analogous arguments as above or in Example 17 or Example 18. Concluding, for all values of $u^{k+1}\in K_U$, we have shown that (3.26) is fulfilled. □

We conclude this section with a convergence result for the case that g is continuously differentiable. For this purpose, we restrict ourselves to the case of $K_U\subseteq[u_a,u_b]\subseteq\mathbb{R}$, $u_a,u_b\in\mathbb{R}$. However, we remark that with similar arguments the same holds for any bounded set $K_U\subseteq\mathbb{R}^m$, $m\in\mathbb{N}$. Furthermore the discussion also holds in the framework of ODEs described in Chapter 2. We consider an optimal control problem given by (3.3) with the Hamiltonian function given by (3.6) and the augmented Hamiltonian function given by (3.22) where g is continuously differentiable with respect to the control argument u. We set $f\left(\cdot,y,u\right)=u$ in order to focus on the main arguments. We remark that the following proof can be done analogously if $\left(y,u\right)\mapsto\frac{\partial}{\partial u}f\left(z,y,u\right)$ is continuous for any fixed $z\in Z$.
With our simplifications the reduced gradient of the corresponding optimal control problem is given by

$$\nabla J\left(u\right)=\frac{\partial}{\partial u}g\left(u\right)+p \tag{3.38}$$

with $J\left(\bar{u}\right):=J\left(y\left(\bar{u}\right),\bar{u}\right)$ where y is the solution to (3.2) and p is the solution to (3.5). We investigate the sequence $\left(u^n\right)_{n\in\mathbb{N}_0}$ that is generated in a loop over Step 2 to Step 4 in Algorithm 3.1, referred to as the SQH method. We show that the variational inequality

$$\nabla J\left(\bar{u}\right)\left(z\right)\left(w\left(z\right)-\bar{u}\left(z\right)\right)\geq 0 \tag{3.39}$$

is fulfilled for all $w \in U_{ad}$ and almost all $z \in Z_i$ where $\bar{u}$ is the limit of a subsequence $\left(u^k\right)_{k\in K}$, $K \subseteq \mathbb{N}_0$ with the property

$$\lim_{k\to\infty} \|u^k - \bar{u}\|_{L^2(Z_i)} = 0.$$

From (3.39) by integration, see [5, X Corollary 2.16], we have that $\bar{u}$ satisfies the optimality condition

$$\int_{Z_i} \nabla J\left(\bar{u}\right)(z)\left(w\left(z\right) - \bar{u}\left(z\right)\right) dz \geq 0$$

for all $w \in U_{ad}$, see [95, Lemma 2.21].

In order to prove this fact, we use the Euclidean projection $P_{K_U} : \mathbb{R} \to K_U$, see [12, Proposition 2.1.3 (Projection Theorem)]. Now, with an analogous calculation as in [17, Theorem 3.2], we prove the following theorem where some technical parts are similar to the proof of Theorem 27.

Theorem 29. *Assume that g is continuously differentiable with respect to u and there is a lower bound $\epsilon_0 > 0$ for ϵ. Then for each accumulation point $\bar{u}$ of the sequence $\left(u^n\right)_{n\in\mathbb{N}_0}$ generated in the SQH method (loop over Step 2 to Step 4) with the property*

$$\lim_{\tilde{k}\to\infty} \|u^{\tilde{k}} - \bar{u}\|_{L^2(Z_i)} = 0,$$

$\tilde{k} \in \tilde{K} \subseteq \mathbb{N}$, there is a subsequence $\left(u^k\right)_{k\in K}$, $K \subseteq \tilde{K}$, such that

$$\lim_{k\to\infty} \|u^k - P_{K_U}\left(u^k - \frac{1}{2\epsilon}\nabla J\left(u^k\right)\right)\|_{L^2(Z_i)} = 0$$

where $\bar{u}$ fulfills the following optimality condition

$$\nabla J\left(\bar{u}\right)(z)\left(w\left(z\right) - \bar{u}\left(z\right)\right) \geq 0$$

for all $w \in U_{ad}$ and almost all $z \in Z_i$.

Proof. We remark that $\epsilon > 0$ for each iterate $u^{\tilde{k}}$, $\tilde{k} \in \tilde{K}$. As $u^{\tilde{k}+1}$ minimizes $w \mapsto K_\epsilon\left(z, y^{\tilde{k}}, w, u^{\tilde{k}}, p^{\tilde{k}}\right)$ for all $z \in Z$ with $u^{\tilde{k}+1} \in K_U$, we have that

$$\begin{aligned}
&\frac{\partial}{\partial u^{\tilde{k}+1}} K_\epsilon\left(z, y^{\tilde{k}}, u^{\tilde{k}+1}, u^{\tilde{k}}, p^{\tilde{k}}\right)\left(w - u^{\tilde{k}+1}\right)\\
&= \left(2\epsilon\left(u^{\tilde{k}+1} - u^{\tilde{k}}\right) + \frac{\partial}{\partial u^{\tilde{k}+1}} H\left(z, y^{\tilde{k}}, u^{\tilde{k}+1}, p^{\tilde{k}}\right)\right)\left(w - u^{\tilde{k}+1}\right) \geq 0
\end{aligned}$$

for all $w \in K_U$ and for all $z \in Z$, see [95, Lemma 2.21]. Equivalently, we can write

$$u^{\tilde{k}+1} = P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\frac{\partial}{\partial u^{\tilde{k}+1}} H\left(z, y^{\tilde{k}}, u^{\tilde{k}+1}, p^{\tilde{k}}\right)\right) \tag{3.40}$$

see [12, Proposition 2.1.3 (Projection Theorem)]. Additionally, we have

$$\nabla J\left(u^{\tilde{k}}\right) = \frac{\partial}{\partial u^{\tilde{k}}} H\left(\cdot, y^{\tilde{k}}, u^{\tilde{k}}, p^{\tilde{k}}\right)$$

compare (3.6) with (3.38). Starting from (3.40) and adding and subtracting equal terms, we have

$$\begin{aligned}
u^{\tilde{k}} - P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\nabla J\left(u^{\tilde{k}}\right)\right) &= u^{\tilde{k}} - u^{\tilde{k}+1} + P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\frac{\partial}{\partial u^{\tilde{k}+1}} H\left(\cdot, y^{\tilde{k}}, u^{\tilde{k}+1}, p^{\tilde{k}}\right)\right)\\
&\quad - P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\frac{\partial}{\partial u^{\tilde{k}}} H\left(\cdot, y^{\tilde{k}}, u^{\tilde{k}}, p^{\tilde{k}}\right)\right).
\end{aligned}$$

Thus, using the triangle inequality, the projection theorem [12, Proposition 2.1.3 (Projection Theorem)] and $\epsilon > \epsilon_0$, we obtain

$$\begin{aligned}
&\|u^{\tilde{k}} - P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\nabla J\left(u^{\tilde{k}}\right)\right)\|_{L^2(Z_i)} \\
&\leq \|u^{\tilde{k}} - u^{\tilde{k}+1}\|_{L^2(Z_i)} \\
&+ \frac{1}{2\epsilon_0}\|\frac{\partial}{\partial u^{\tilde{k}+1}}H\left(\cdot, y^{\tilde{k}}, u^{\tilde{k}+1}, p^{\tilde{k}}\right) - \frac{\partial}{\partial u^{\tilde{k}}}H\left(\cdot, y^{\tilde{k}}, u^{\tilde{k}}, p^{\tilde{k}}\right)\|_{L^2(Z_i)} \\
&\leq \|u^{\tilde{k}} - u^{\tilde{k}+1}\|_{L^2(Z_i)} + \frac{1}{2\epsilon_0}\left(\alpha\|u^{\tilde{k}+1} - u^{\tilde{k}}\|_{L^2(Z_i)} + \gamma\|\frac{\partial}{\partial u^{\tilde{k}+1}}g\left(u^{\tilde{k}+1}\right) - \frac{\partial}{\partial u^{\tilde{k}}}g\left(u^{\tilde{k}}\right)\|_{L^2(Z_i)}\right).
\end{aligned} \tag{3.41}$$

Now, we have the following estimates

$$\|y^{\tilde{k}} - \bar{y}\|_{L^2(Z_i)} \leq c\|u^{\tilde{k}} - \bar{u}\|_{L^2(Z_i)} \text{ and } \|p^{\tilde{k}} - \bar{p}\|_{L^2(Z_i)} \leq c\|u^{\tilde{k}} - \bar{u}\|_{L^2(Z_i)},\ c > 0, \tag{3.42}$$

see Assumption A.2), where $\bar{y}$ is the solution to (3.2) for $\bar{u}$ instead of u and $\bar{p}$ is the solution to (3.5) for $\bar{y}$ instead of y and for $\bar{u}$ instead of u. By our assumption, there exists a subsequence within the sequence $(u^n)_{n\in\mathbb{N}_0}$ that strongly converges to $\bar{u}$ in $L^2(Z_i)$. Using [6, Proposition 3.6, Remark 3.7], (3.42) and since any subsequence of a converging sequence also converges, see [3, II Theorem 1.15], we obtain a subsequence, $\left(u^k\right)_{k\in K}$, $K_1 \subseteq \tilde{K}$, with the following pointwise convergence

$$\lim_{k\to\infty} u^k(z) = \bar{u}(z),\ \lim_{k\to\infty} y^k(z) = \bar{y}(z) \text{ and } \lim_{k\to\infty} p^k(z) = \bar{p}(z)$$

for almost all $z \in Z$ and $k \in K_1$. Consequently, we have

$$\lim_{K_1\ni k\to\infty} \nabla J\left(u^k\right) = \lim_{K_1\ni k\to\infty}\left(\alpha u^k + \gamma\frac{\partial}{\partial u}g(u)|_{u=u^k} + p^k\right) = \alpha\bar{u} + \gamma\frac{\partial}{\partial u}g(u)|_{u=\bar{u}} + \bar{p} = \nabla J(\bar{u}) \tag{3.43}$$

for almost every $z \in Z$.

For the next step, we need some preparations. Because Theorem 13 holds analogously for the PDE case from which we have $\lim_{n\to\infty}\|u^{n+1} - u^n\|_{L^2(Z_i)} = 0$, we have

$$\lim_{K_1\ni k\to\infty}\|u^{k+1} - u^k\|_{L^2(Z_i)} = 0$$

since u^{k+1} is the following element of u^k in the sequence $(u^n)_{n\in\mathbb{N}_0}$. Consequently by [6, Proposition 3.6, Remark 3.7], we have an index set $K_2 \subseteq K_1$ such that $\lim_{K_2\ni k\to\infty} u^{k+1}(z) - u^k(z) = 0$ for almost all $z \in Z_i$ where all the other properties above remain since any subsequence of a converging sequence also converges, see [3, II Theorem 1.15]. From this we can also conclude that

$$\lim_{K_2\ni k\to\infty} u^{k+1}(z) = \lim_{K_2\ni k\to\infty}\left(u^{k+1}(z) - u^k(z)\right) + \lim_{K_2\ni k\to\infty} u^k(z) = \bar{u}(z) \tag{3.44}$$

for almost all $z \in Z_i$ where we use the calculation rules for the limit [3, II Theorem 2.2].

If we take the limit on both sides of (3.41), considering the pointwise converging subsequence $\left(u^k\right)_{k\in K_2}$, we obtain

$$\lim_{K_2\ni k\to\infty}\|u^k - P_{K_U}\left(u^k - \frac{1}{2\epsilon}\nabla J\left(u^k\right)\right)\|_{L^2(Z_i)} = 0 \tag{3.45}$$

where we use that $\lim_{K_2\ni k\to\infty}\|u^{k+1} - u^k\|_{L^2(Z_i)} = 0$ with the same reasoning as in the beginning of the paragraph above for the first and second term and for the last term the dominated convergence theorem

[6, Proposition 2.17]. The dominated convergence theorem can be applied to the measurable functions $g \circ u^n$, $n \in \mathbb{N}_0$, see [6, 2.2 Measurable and Borel functions] because the pointwise limit

$$\lim_{K_2 \ni k \to \infty} \frac{\partial}{\partial u} g(u) |_{u=u^k(z)} = \lim_{K_2 \ni k \to \infty} \frac{\partial}{\partial u} g(u) |_{u=u^{k+1}(z)} = \frac{\partial}{\partial u} g(u) |_{u=\bar{u}(z)}$$

holds due to the continuity of $\frac{\partial}{\partial u} g(u)$ and the pointwise convergence of u^k, $k \in K_2$ and because the integrability of $g \circ u^k$, $k \in K_2$ holds since [3, III.3 Theorem 3.6] with the bounded image of u^k ensures an upper bound for $g \circ u^k$ that holds for all $k \in K_2$.

Next, we prove that $\nabla J(\bar{u})(z)(w(z) - \bar{u}(z)) \geq 0$ for all $w \in U_{ad}$ for almost all $z \in Z$. For this purpose, we start with

$$v^k := P_{K_U}\left(u^k - \frac{1}{2\epsilon} \nabla J\left(u^k\right)\right)$$

for almost every $z \in Z$. This is equivalent to

$$\left(v^k - u^k + \frac{1}{2\epsilon} \nabla J\left(u^k\right)\right)\left(w - v^k\right) \geq 0$$

for all $w \in U_{ad}$ for almost all $z \in Z$, see [12, Proposition 2.1.3 (Projection Theorem)]. Then we have

$$\left(v^k - u^k\right)\left(w - v^k\right) + \frac{1}{2\epsilon} \nabla J\left(u^k\right)\left(w - v^k\right) \geq 0.$$

Adding and subtracting u^k, we obtain

$$2\epsilon\left(v^k - u^k\right)\left(w - v^k\right) + \nabla J\left(u^k\right)\left(w - u^k\right) + \nabla J\left(u^k\right)\left(u^k - v^k\right) \geq 0. \tag{3.46}$$

From (3.45) there exists a subsequence $K \subseteq K_2$ with

$$\lim_{K \ni k \to \infty} u^k - v^k = 0,$$

see [6, Proposition 3.6, Remark 3.7]. Due to $|w - v^k| \leq 2 \max(|u_a|, |u_b|)$ and the upper bound $\sigma(\eta + \theta)$ for ϵ because of (2.30) and Step 4 of the SQH method and the fact that converging sequences are bounded [3, II Theorem 1.10] combined with (3.45) and (3.43), we obtain by taking the limit in (3.46) for $k \in K$ the following

$$\nabla J(\bar{u})(w - \bar{u}) \geq 0$$

for all $w \in U_{ad}$ for almost all $z \in Z$, see [3, II Theorem 2.4] and [3, II Theorem 2.7]. □

3.4 Numerical experiments

In this section, we present results of numerical experiments with the SQH method applied to the different control problems P.1) to P.7). The purpose of these experiments is to validate the theoretical results and the computational performance of the SQH scheme for PDE constraint optimal control problems. In particular, we demonstrate that by decreasing the tolerance in the SQH stopping criterion, the fulfillment of the PMP optimality condition by the returned solution improves as expected. Furthermore, we plot the optimal solutions to the given problems and show the convergence history of the SQH scheme in terms of reduction of the value of the cost functional.

We remark that the analysis of the SQH method in Section 3.3 is performed at a functional level and independently of the discretization used. However, for the numerical realization of our optimization scheme, we consider the following finite differences setting [60]. We take a space-time cylinder $Q = \Omega \times (0, T)$ in the parabolic case with $\Omega = (a, b)^n$ and define the following space-time grid

$$Q_{h,\triangle t} := \{(x_{i_1 \dots i_n}, t_m), \mid x_{i_1 \dots i_n} \in \Omega_h, \, t_m = m \triangle t, \, m \in \{1, ..., N_t\}\}$$

where

$$\Omega_h = \left\{(a + i_1 h, \dots, a + i_n h) \in \mathbb{R}^n,\ i_j \in \{1, ..., N-1\},\ j \in \{1, ..., n\}\right\}.$$

In the elliptic case we use the domain Ω_h. The space and time mesh-sizes are given by $h := \frac{b-a}{N}$, $\triangle t := \frac{T}{N_t}$. We assume that the grid points $(x_{i_1 \dots i_n}, t_m)$ and $t_m = m \triangle t$ are ordered lexicographically.

In order to compute the state and adjoint variables, we approximate (3.2) and (3.5) using the implicit Euler scheme and finite differences in the parabolic case. For the elliptic case, we use a 5-point finite-difference discretization of the Laplacian. For the computation of the integrals appearing in J, we use the rectangle rule, see for example [91].

In the case of elliptic problems, we choose the domain $\Omega = (0,1) \times (0,1)$ and use a mesh size $\triangle x = \frac{1}{50}$ if not otherwise stated. In the parabolic cases, we have $Q = (0,1) \times (0,1)$ with a step size $\triangle x = \frac{1}{50}$ and $\triangle t = \frac{1}{100}$. Alterations from this are noted at the corresponding site.

To solve linear problems (or subproblems), we use the MATLAB backslash operator. Non-linear problems as $P.5)$ and $P.7)$ are solved by a Gauss-Seidel-Picard iteration [18] with a tolerance on the discrete L^2-norm of the residuum of 10^{-8}. Specifically, for $P.5)$ the state variable is updated pointwise within a loop over the interior points of the domain with

$$y(i,j) = \frac{1}{4}\left(y(i+1,j) + y(i-1,j) + y(i,j+1) + y(i,j-1) + h^2\left(u(i-1,j-1) - y^3(i,j)\right)\right).$$

For $P.7)$ we use

$$y(i,j) = \frac{y(i+1,j) + y(i-1,j) + y(i,j+1) + y(i,j-1)}{4} + u(i-1,j-1) \cdot \begin{cases} \frac{\triangle x^2}{4+\triangle x^2} & \text{if } y(i,j) \geq 0 \\ \frac{\triangle x^2}{4} & \text{else} \end{cases}.$$

The minimization of the augmented Hamiltonian in Step 2 of Algorithm 3.1 can be performed by a secant method or by an analytical formula that solves the one-dimensional (since u is scalar valued in our cases) minimization problem.

In the attempt to provide an overall view and comparison of the SQH performance with all test cases, we present results concerning PMP optimality of the SQH solution in Table 3.1. In this table, N_{up} denotes the total number of updates that are made by the SQH method to the initial guess of the control on the given grid starting with the same initial guess of the control and "iter" is the number of total sweeps, which means Step 2 to Step 5 in Algorithm 3.1.

To measure PMP optimality, we define the function

$$\triangle H(z) := H(z, y, u, p) - \min_{w \in K_U} H(z, y, w, p)$$

where y, u and p are the return values from the SQH method upon convergence. As a measure of optimality, we report the number $N_{\%}^{l}$ that is the percentage of the grid points at which the inequality

$$0 \leq \triangle H \leq 10^{-l},$$

$l \in \mathbb{N}$, is fulfilled. This is to verify the PMP optimality (3.7) up to a tolerance of the solution returned by the SQH method, at least on a subset of grid points. Corresponding to this solution, in Table 3.1, we also give the value $\max_{z \in Z_i} \triangle H(z)$ for illustration.

Table 3.1 provides an overview for the optimality results of the SQH method for the optimal control problems P.1) to P.7). In most cases, PMP optimality of over 90% is achieved with very stringent tolerance ($l = 8, 12$). We remark that in the case $P.6)$, we solve an augmented problem, see Subsection 3.4.6 for details.

The results of Table 3.1 correspond to the following choice of parameters: $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$, $\kappa = 10^{-8}$, $u^0 = 0$ and the initial guess $\epsilon = \frac{1}{150}$.

	N_{up}	iter	$\max_{z\in Z_i} \triangle H(z)$	$\frac{N_{\%}^{2}}{\%}$	$\frac{N_{\%}^{4}}{\%}$	$\frac{N_{\%}^{6}}{\%}$	$\frac{N_{\%}^{8}}{\%}$	$\frac{N_{\%}^{12}}{\%}$
P.1)	40	58	$2.74 \cdot 10^{-3}$	100	99.8776	99.8776	99.8776	99.8776
P.2)	26	40	$3.42 \cdot 10^{-3}$	100	99.3336	99.2503	99.1670	99.1670
P.3)	436	645	$2.88 \cdot 10^{-3}$	100	90.9204	90.8372	90.8372	90.8372
P.4)	173	255	$3.11 \cdot 10^{-5}$	100	100	97.2245	96.8776	96.8776
P.5)	235	348	$7.75 \cdot 10^{-3}$	100	97.0012	91.5452	90.9621	90.9621
P.6)	864	1281	$2.05 \cdot 10^{-2}$	93.7526	87.4636	83.7151	83.3819	83.2153
P.7)	283	419	$3.00 \cdot 10^{-1}$	76.2371	75.3393	75.3393	75.3393	75.3393

Table 3.1: Numerical investigation of optimality of the SQH solution to the problems P.1) to P.7) with $\kappa = 10^{-8}$.

3.4.1 Application to a linear parabolic case

In this subsection, we discuss a parabolic optimal control problem that is governed by the heat equation. The weak formulation of the heat equation is given as follows. For each $t \in (0,T)$, $T > 0$, the resulting initial-boundary value problem is given by: Find $y \in L^2\left(0,T;H_0^1(\Omega)\right)$ and $y' \in L^2\left(0,T;H^{-1}(\Omega)\right)$, that means, $y \in W(0,T) := \left\{y \in L^2\left(0,T;H_0^1(\Omega)\right) \mid y' \in L^2\left(0,T;H^{-1}(\Omega)\right)\right\}$, see [95, Chapter 3], such that the following is satisfied

$$\begin{aligned} \left(y'(\cdot,t),v\right) + D\left(\nabla y(\cdot,t),\nabla v\right) = &\quad (u(\cdot,t),v) && \text{in Q} \\ y(\cdot,0) = &\quad y_0 && \text{on } \Omega \times \{t=0\} \\ y = &\quad 0 && \text{on } \partial\Omega \end{aligned} \tag{3.47}$$

for all $v \in H_0^1(\Omega)$ where Ω has a smooth boundary. In this setting, $y : Q \to \mathbb{R}$ denotes the state variable and $u : Q \to \mathbb{R}$ denotes the control. We denote with $(\cdot,\cdot)$ the scalar product in $L^2(\Omega)$, $D > 0$ is the diffusion coefficient, $y' := \frac{\partial}{\partial t} y(x,t)$ and ∇ denotes the $L^2(\Omega)$ gradient. Next, we discuss the following parabolic optimal control problem

$$\begin{aligned} &\min_{y,u} J(y,u) \\ \text{s.t.}\quad \left(y',v\right) + D(\nabla y,\nabla v) &= (u,v) && \text{in Q} \\ y(\cdot,0) &= y_0 && \text{on } \Omega \times \{t=0\} \\ y &= 0 && \text{on } \partial\Omega \\ u &\in U_{ad} \end{aligned} \tag{3.48}$$

where the cost functional J is given by

$$J(y,u) := J_c(y,u) + \gamma \int_Q g(u(x,t))\,dxdt. \tag{3.49}$$

In this functional, J_c represents a smooth functional objective as it appears in many control problems [19, 95]. We have

$$J_c(y,u) := \frac{1}{2}||y - y_d||_{L^2(Q)}^2 + \frac{\alpha}{2}||u||_{L^2(Q)}^2, \qquad \alpha \geq 0. \tag{3.50}$$

In this case, the functional J_c models the task of driving the state y to track a desired state trajectory $y_d \in L^q(Q)$, $q > \frac{n}{2} + 1$ if $n \geq 2$ and $q \geq 2$ if $n = 1$, while keeping small the $L^2(Q)$-cost of the control.

In addition to J_c, we have a possibly discontinuous cost functional given by

$$G(u) := \gamma \int_Q g(u(x,t))\,dxdt, \qquad \gamma \geq 0 \tag{3.51}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a non-negative and lower semi-continuous function.

In particular, we consider the case where

$$g(u) = \begin{cases} |u| & \text{if } |u| > s \\ 0 & \text{otherwise} \end{cases}, \quad s > 0. \tag{3.52}$$

With this construction, we obtain a cost of the control that is zero if its value is below a given threshold and it measures an L^1 cost otherwise.

The admissible set of controls is defined as follows

$$U_{ad} := \{u \in L^q(Q) \mid u(x,t) \in K_U\} \tag{3.53}$$

where K_U is a compact subset of $\mathbb{R}$.

To show that the control cost G is discontinuous as a map from U_{ad} to $\mathbb{R}$, consider constant controls and choose $\bar{u} \equiv s$ and $u_\epsilon = \bar{u} + \epsilon$ with $\epsilon > 0$. We have that $\|u_\epsilon - \bar{u}\|_{L^p(Q)} \to 0$ for $\epsilon \to 0$ for every $p \geq 1$. On the other hand, we have a discontinuity as the following demonstrates

$$|\int_Q (g(u_\epsilon) - g(\bar{u}))dxdt| = |\int_Q (s+\epsilon)dxdt| = \int_Q (s+\epsilon)dxdt > sQ > 0$$

for any fixed $s > 0$.

In the case where G is a convex and continuous cost functional, existence of an optimal control is guaranteed [95]. However, in the case of discontinuous cost functionals the issue of existence of an optimal control is more delicate. For this reason, as mentioned in Section 3.1, we assume the existence of a solution to (3.48) in U_{ad} and focus on the numerical treatment of the problem. Notice that any solution to (3.48) can be characterized with the PMP as discussed in Section 3.2.

We recall that the corresponding adjoint problem according to (3.5) is given by

$$\begin{aligned} \left(-p'(\cdot,t), v\right) + D(\nabla p(\cdot,t), \nabla v) &= (y(\cdot,t) - y_d(\cdot,t), v) && \text{in } Q \\ p(\cdot,T) &= 0 && \text{on } \Omega \times \{T = 0\} \\ p &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{3.54}$$

This problem has the same structure as (3.47) after a transformation of the time variable $\tau := T - t$ and noticing that $y - y_d \in L^q(Q)$, see [1, Theorem 2.14]. Hence, there exists a unique $p \in L^2\left(0,T; H_0^1(\Omega)\right)$ and $p' \in L^2\left(0,T; H^{-1}(\Omega)\right)$ solving (3.54) for all $v \in H_0^1(\Omega)$.

Next, we define the Hamiltonian corresponding to (3.48) - (3.50) according to (3.6) as follows

$$H(x,t,y,u,p) = \frac{1}{2}(y - y_d)^2 + \frac{\alpha}{2}u^2 + \gamma g(u) + pu \tag{3.55}$$

where $H : \mathbb{R}^n \times \mathbb{R}_0^+ \times \mathbb{R} \times K_U \times \mathbb{R} \to \mathbb{R}$.

Next, we show that our requirements A.1) to A.6) are fulfilled. We have that $h(y) = \frac{1}{2}(y - y_d)^2$ and $f(y,u) = u$ fulfill A.1) and A.6). For A.2), we consider the difference between (3.47) with $u \leftarrow u_1$ and $y \leftarrow y_1$ and the same equation (3.47) but with $u \leftarrow u_2$ and $y \leftarrow y_2$ with $\delta y = y_1 - y_2$ and $\delta u = u_1 - u_2$. We assume that $u_1 \neq u_2$ and thus $y_1 \neq y_2$ due to the unique solvability of (3.47). We obtain

$$\int_0^T \int_\Omega \delta y'(x,t) v(x) + D \nabla \delta y(x,t) \nabla v(x)\, dxdt = \int_0^T \int_\Omega \delta u(x,t) v(x)\, dxdt$$

from which we have

$$\int_0^T \frac{1}{2}\frac{d}{dt}\|\delta y(\cdot,t)\|_{L^2(\Omega)}^2 + D\|\nabla \delta y(\cdot,t)\|_{L^2(\Omega)}^2 dt \leq \int_0^T ||\delta u(\cdot,t)\|_{L^2(\Omega)}^2 \|\delta y(\cdot,t)\|_{L^2(\Omega)}^2 dt$$

according to [45, page 287, Theorem 3] and the Cauchy-Schwarz inequality, see [2, Lemma 2.2]. Next, we have

$$\frac{1}{2}\left(\|\delta y\left(\cdot,T\right)\|^2_{L^2(\Omega)}-\|\delta y\left(\cdot,0\right)\|^2_{L^2(\Omega)}\right)+D\|\nabla\delta y\|^2_{L^2(Q)}\leq\hat{c}\int_0^T||\delta u\left(\cdot,t\right)\|^2_{L^2(\Omega)}\|\nabla\delta y\left(\cdot,t\right)\|^2_{L^2(\Omega)}dt \quad (3.56)$$

for some $\hat{c}>0$ with the Cauchy-Schwarz inequality, see [2, Lemma 2.2] for the right hand-side

$$\int_0^T||\delta u\left(\cdot,t\right)\|^2_{L^2(\Omega)}\|\nabla\delta y\left(\cdot,t\right)\|^2_{L^2(\Omega)}dt.$$

Thus, as $\|\delta y\left(\cdot,0\right)\|^2_{L^2(\Omega)}=0$, we obtain from (3.56) the following

$$\|\nabla\delta y\|_{L^2(Q)}\leq\tilde{c}\left(D\right)\|\delta u\|_{L^2(Q)}$$

for some $\tilde{c}\left(D\right)>0$ since $\|\nabla\delta y\|_{L^2(Q)}\neq 0$. Furthermore, by the Poincaré inequality [2, 6.7], for $\tilde{c}>0$ it holds that

$$\|\delta y\|_{L^2(Q)}=\sqrt{\int_0^T\|\delta y\left(\cdot,t\right)\|^2_{L^2(\Omega)}dt}\leq\tilde{c}\sqrt{\int_0^T\|\nabla\delta y\left(\cdot,t\right)\|^2_{L^2(\Omega)}dt}=\tilde{c}\|\nabla\delta y\|_{L^2(Q)}\leq c\left(D\right)\|\delta u\|_{L^2(Q)}.$$

The same calculation holds for the adjoint variable. For this purpose, we replace δy by $\delta p=p_1-p_2$ where p_1 solves (3.54) for $y\leftarrow y_1$ and correspondingly p_2. Furthermore we replace δu by δy and use the result $\|\delta y\|_{L^2(Q)}\leq c\left(D\right)\|\delta u\|_{L^2(Q)}$ in order to obtain

$$\|\delta p\|_{L^2(Q)}\leq c\left(D\right)^2\|\delta u\|_{L^2(Q)}.$$

Assumption A.3) is fulfilled immediately and Assumption A.4) is fulfilled due to the boundedness discussion of the adjoint variable in Section 3.2. Since $\frac{\partial^2}{\partial y^2}h\left(y\right)=1$, Assumption A.5) is also fulfilled. Now, we have checked that our considered case fits to our theoretical framework of Section 3.1. Next we come to the numerical investigation.

In our numerical experiments, we consider $\Omega=(a,b)$ with $a=0$, $b=1$ and $T=1$. The initial guess for the control and the initial value y_0 for the state is the zero function. Furthermore, the parameter in Algorithm 3.1 are chosen as follows $\kappa=10^{-6}$, $\zeta=\frac{3}{20}$, $\sigma=50$ and $\eta=10^{-7}$. The initial value of ϵ equals $\frac{3}{5}$. The numerical parameters are set as follows, $N=100$, $N_t=200$, $D=\frac{1}{5}$ and if not otherwise stated $\alpha=10^{-5}$, $\gamma=10^{-1}$. Furthermore, we have, $K_U=[0,10]$ and

$$y_d\left(x,t\right)=\begin{cases}5 & \text{if } \bar{x}\left(t\right)-c\leq x\leq\bar{x}\left(t\right)+c\\ 0 & \text{else,}\end{cases} \quad (3.57)$$

where $\bar{x}\left(t\right):=x_0+\frac{2}{5}\left(b-a\right)\sin\left(2\pi\frac{t}{T}\right)$, $x_0=\frac{b+a}{2}$ and $c=\frac{7}{100}\left(b-a\right)$. We choose $s=1$ in (3.52).

The augmented Hamiltonian

$$K_\epsilon\left(x,t,y,u,v,p\right):=\frac{1}{2}\left(y-y_d\right)^2+\frac{\alpha}{2}u^2+\gamma g\left(u\right)+pu+\epsilon\left(u-v\right)^2 \quad (3.58)$$

is minimized as follows. Its minimum can be exactly given by a case study.

If $0\leq u\leq s$, we have

$$K_\epsilon\left(x,t,y,u,v,p\right)=\frac{1}{2}\left(y-y_d\right)^2+\frac{\alpha}{2}u^2+pu+\epsilon\left(u-v\right)^2. \quad (3.59)$$

If the minimum u of (3.59) is in $0<u<s$, we have that $0=\frac{\partial}{\partial u}K_\epsilon\left(x,t,y,u,v,p\right)$. If the minimum is outside of $0<u<s$, then the minimum is at 0 or s. Consequently in the case $0\leq u\leq s$ the minimum is analytically given by

$$u_1:=\min\left(\max\left(0,\frac{2\epsilon v-p}{2\epsilon+\alpha}\right),s\right).$$

Analogous in the case if $s < u \leq 10$. We have

$$K_\epsilon\left(x,t,y,u,v,p\right) := \frac{1}{2}\left(y-y_d\right)^2 + \frac{\alpha}{2}u^2 + \gamma u + pu + \epsilon\left(u-v\right)^2$$

with its minimum at

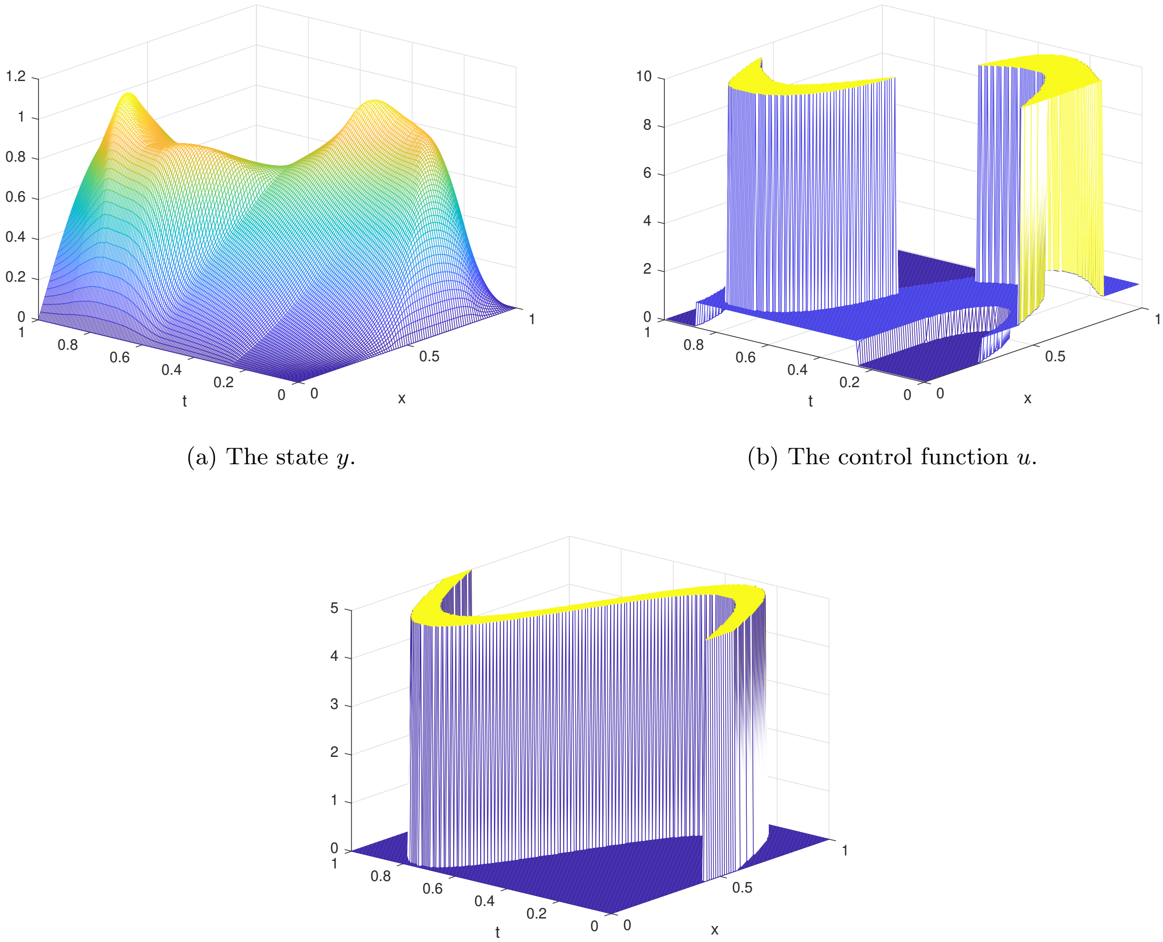
$$u_2 := \min\left(\max\left(s, \frac{2\epsilon v - (p+\gamma)}{2\epsilon+\alpha}\right), 10\right).$$

Then the minimum of K_ϵ defined in (3.58) over K_U is given by

$$u = \operatorname*{arg\,min}_{w\in K_U} K_\epsilon\left(x,t,y,w,v,p\right) = \operatorname*{arg\,min}_{w\in\{u_1,u_2\}} K_\epsilon\left(x,t,y,w,v,p\right)$$

since a minimum over K_U is in $0 \leq u \leq s$ or $s < u \leq 10$.

We perform the first set of experiments using Algorithm 3.1 to solve our optimal control problem. The SQH algorithm converges in 29 iterations and we obtain the state and control functions depicted in Figure 3.1. The plot of the control function shows clearly the action of the discontinuous cost of the control given by g in (3.52) and the presence of the control's upper bound at 10.



(a) The state y.

(b) The control function u.

(c) The desired function y_d.

Figure 3.1: Optimal solution for the first experimental setting.

With the second experiment, we present results to investigate how well the solution of the SQH method satisfies the optimality condition given by the PMP. For this purpose, we have

$$\Delta H = H\left(x,t,\bar{y},\bar{u},\bar{p}\right) - \min_{w\in K_U} H\left(x,t,\bar{y},w,\bar{p}\right)$$

and give in Table 3.2 the ratio of numbers of grid points $(x,t) \in Q_{h,\triangle t}$ where the optimality condition (3.7) is satisfied to machine precision. These entries give a measure of optimality of the SQH solution $(\bar{y},\bar{u},\bar{p})$ and demonstrate an improvement in accuracy of the PMP solution by refinement of κ.

For this purpose, in Table 3.2, we give the ratio of grid points where the following holds

$$0 \leq H\left(x,t,\bar{y},\bar{u},\bar{p}\right) - \min_{w\in K_U} H\left(x,t,\bar{y},w,\bar{p}\right) \leq \text{eps}$$

with eps the machine precision given by $2.2\cdot 10^{-16}$ in our case. We see that, independently of the mesh size, at almost all grid points the PMP condition is fulfilled to machine precision, already for $\kappa = 10^{-6}$. In Table 3.3 we report the values of $\max_{(x,t)\in Q_{h,\triangle t}} \Delta H$. The results reported in Table 3.3 demonstrate how $\max_{(x,t)\in Q_{h,\triangle t}} \Delta H$ decreases as we refine the mesh size and the value of κ.

κ $N_t \times N$	10^{-1}	10^{-3}	10^{-6}	10^{-11}	10^{-16}
100×200	0	0.9973	0.9988	0.9995	0.9998
200×400	$6.28\cdot 10^{-5}$	0.9966	0.9998	0.9998	0.9998
400×800	$6.70\cdot 10^{-4}$	0.9934	0.9981	0.9998	0.9998
800×1600	$1.59\cdot 10^{-3}$	0.9868	0.9998	0.9998	0.9998

Table 3.2: Ratio of grid points at which the Pontryagin maximum principle is fulfilled to machine precision to the total number of grid points.

κ $N_t \times N$	10^{-1}	10^{-3}	10^{-6}	10^{-11}	10^{-16}
100×200	3.43	$9.00\cdot 10^{-3}$	$5.68\cdot 10^{-3}$	$1.27\cdot 10^{-3}$	$7.29\cdot 10^{-4}$
200×400	3.42	$5.34\cdot 10^{-3}$	$5.17\cdot 10^{-4}$	$5.17\cdot 10^{-4}$	$5.17\cdot 10^{-4}$
400×800	3.41	$1.06\cdot 10^{-2}$	$6.89\cdot 10^{-3}$	$6.70\cdot 10^{-4}$	$6.70\cdot 10^{-4}$
800×1600	3.41	$1.13\cdot 10^{-2}$	$3.93\cdot 10^{-7}$	$1.82\cdot 10^{-10}$	$7.08\cdot 10^{-11}$

Table 3.3: Values of $\max_{(x,t)\in Q_{h,\triangle t}} \Delta H$ of the SQH solution with different choices of the value of κ.

In the third experiment, we investigate the computational performance of Algorithm 3.1 with respect to different choices of the optimization parameters. In Table 3.4, we report the total number of iterations and corresponding CPU times for convergence with different values of α and γ. Notice that a similar computational effort is required in all cases. Further, we see that the value of the cost functional decreases if α and γ decrease and this is also true for $\|y - y_d\|_{L^2(Q)}$.

α	γ	k	CPU time/s	J	$\|\|y - y_d\|\|_{L^2(Q)}$
10^{-1}	10^{-5}	14	0.5	1.64	1.766037
10^{-3}	10^{-5}	43	1.5	1.33	1.621753
10^{-5}	10^{-5}	57	2.0	1.31	1.621513
0	10^{-5}	63	2.2	1.31	1.621513
0	0	62	2.1	1.31	1.621513
10^{-5}	0	57	1.9	1.31	1.621513
10^{-5}	10^{-3}	51	2.0	1.32	1.621521
10^{-5}	10^{-2}	39	1.3	1.34	1.622160
10^{-5}	10^{-1}	29	1.0	1.52	1.661420

Table 3.4: Computational performance of Algorithm 3.1 with respect to different choices of values of the optimization parameters.

The fourth numerical experiment deals with the complexity of Algorithm 3.1. Let $N_{gp} \in \mathbb{N}$ denote the total number of space-time grid points. We solve the same optimization problem as in Figure 3.1 using different meshes. The resulting CPU times are reported in Figure 3.2 and detailed in Table 3.5. In Figure 3.2, on the abscissa, we have the number of total grid points N_{gp} and on the ordinate the CPU time (sec) required for convergence. Notice that the data points are fitted by a linear model. We remark that this is reasonable since the complexity of Step 2 of Algorithm 3.1 scales linearly and the state and adjoint problems can also be solved with linear complexity, see [19].

$\frac{N}{100} \times \frac{N_t}{100}$	1×2	2×2	2×4	4×4	4×8	8×8	8×16	16×16
CPU time/s	0.9	2.6	5.3	12.0	18.3	40.6	96.5	186.9

Table 3.5: Data points for Figure 3.2.

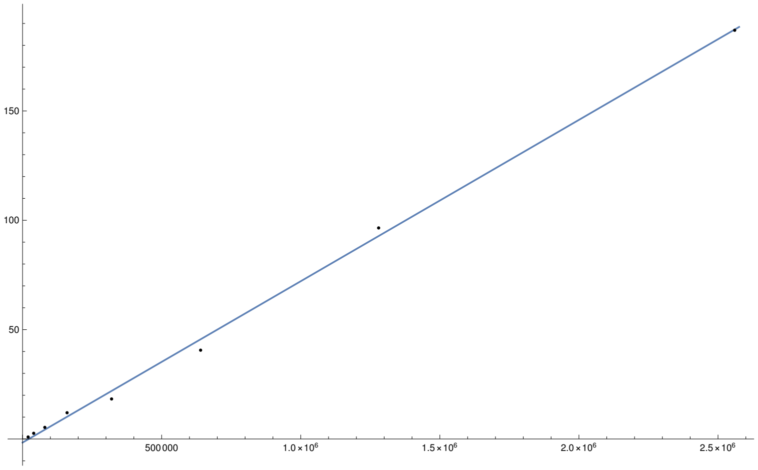


Figure 3.2: Computational complexity of Algorithm 3.1. The data points (dots) from Table 3.5 are fitted by a linear model. On the abscissa we have the number of gird points and on the ordinate the corresponding CPU time is plotted in seconds.

Now in the fifth experiment, we use the same setting like for the investigation of the computational complexity of our algorithm, but choosing $\gamma = 0$. With this choice the discontinuity in the cost of the control is removed and we can compare our SQH scheme with the projected Hager-Zhang-NCG (pNCG) method with Wolfe-Powell step-size strategy [19]. Additionally, we perform the comparison with a projected gradient method with Armijo step-size strategy (pGM). The minimum of the augmented Hamiltonian $K_\epsilon(x,t,y,u,v,p)$ defined in (3.58) with $\gamma = 0$ is given by $u = \frac{2\epsilon v - p}{\alpha + 2\epsilon}$. Furthermore, in the attempt to have the same convergence criterion for all methods, we stop the different iterative procedures if the square of the discrete L^2-norm of the difference between two control functions u of two successive iterations is less than 10^{-6}.

The purpose of this comparison is to address the question of how the SQH scheme performs in the case of continuous cost functionals with respect to a standard optimization strategy. In Table 3.6, we see that the pNCG method in most cases outperforms our SQH method. On the other hand, one can see in Table 3.7 that the SQH method performs better than the pGM scheme. We remark that for the SQH method no step size strategy, like Armijo or Wolfe-Powell, is necessary which makes its implementation easier.

For the case of $\alpha = 10^{-1}$, we take $\sigma = 2.1$ and $\zeta = 0.9$ in Algorithm 3.1 instead of $\sigma = 50$ and $\zeta = \frac{3}{20}$. We remark that the convergence performance of Algorithm 3.1 depends on the choice of σ and ζ whose convenient choice of values may result from numerical experience, as in the setting of different linesearch methods.

α	$N_{gp} = N \times N_t$	SQH		pNCG	
		CPU time/s	number iteration	CPU time/s	number iteration
10^{-1}	200×400	0.7	23	1.6	15
10^{-1}	400×800	2.8	23	3.6	15
10^{-1}	800×1600	11.6	23	12.2	15
10^{-3}	200×400	1.0	33	1.1	8
10^{-3}	400×800	3.9	33	2.6	8
10^{-3}	800×1600	18.6	40	8.6	8
10^{-5}	200×400	1.4	44	1.1	7
10^{-5}	400×800	6.8	58	2.5	7
10^{-5}	800×1600	24.5	54	30.1	49
10^{-7}	200×400	1.7	61	1.0	7
10^{-7}	400×800	7.2	60	2.4	7
10^{-7}	800×1600	19.2	42	7.9	7

Table 3.6: Comparison of the SQH scheme with the pNCG method.

α	$N_{gp} = N \times N_t$	SQH		pGM	
		CPU time/s	number iteration	CPU time/s	number iteration
10^{-1}	200×400	0.7	23	1.8	40
10^{-1}	400×800	2.8	23	3.6	40
10^{-1}	800×1600	11.6	23	12.7	40
10^{-2}	200×400	0.8	23	8.5	272
10^{-2}	400×800	2.9	24	23.9	272
10^{-2}	800×1600	11.9	24	86.6	272
10^{-3}	200×400	1.0	33	20.3	679
10^{-3}	400×800	3.9	33	58.6	675
10^{-3}	800×1600	18.6	40	214.6	675

Table 3.7: Comparison of the SQH scheme with the pGM method.

For further illustration of our optimization framework, we perform the sixth experiment with

$$g(z) := g_1(z) = |z|^{\frac{1}{2}},$$

which is a lower semi-continuous non-convex function. Moreover, we choose a discrete

$$K_U = \{-30, -15, -5, 0, 5, 15, 30\}$$

that models the fact that the control function u may take only a finite set of values. This is intended to demonstrate the easy applicability of the SQH scheme to this kind of optimal control problems, known as mixed-integer problems [52]. In this experiment, the desired state is given by

$$y_d\left(x,t\right) = 5\sin\left(2\pi\frac{t}{T}\right),$$

see Figure 3.3.

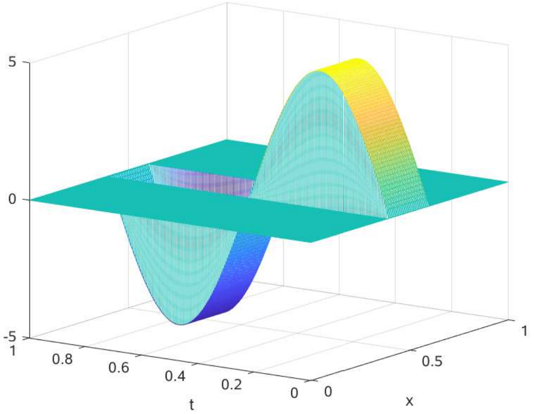


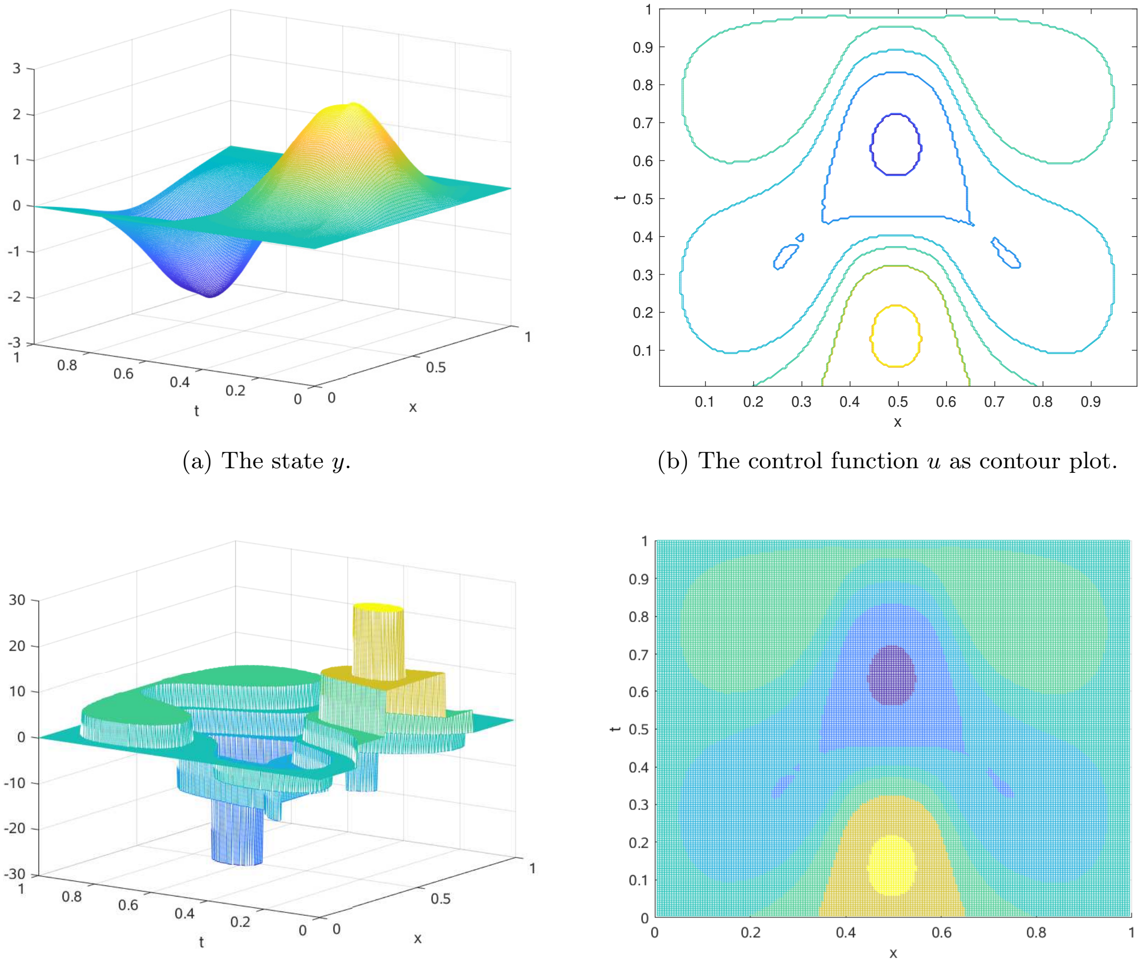
Figure 3.3: Desired function $y_d = 5\sin\left(2\pi\frac{t}{T}\right)$.

Further, we take $\alpha = 5\cdot 10^{-3}$, $\gamma = 1\cdot 10^{-3}$, $N = 200$ and $N_t = 200$. The parameters of Algorithm 3.1 are set as follows. We have $\sigma = 1.1$, $\zeta = 0.5$, $\eta = 10^{-9}$, $\kappa = 10^{-6}$, $u^0 = 0$ and the initial guess for ϵ is given by $\frac{3}{5}\cdot 10^{-7}$. The results are depicted in Figure 3.4 where we clearly see how the admissible control values are taken by the control function.

An analogous numerical test of optimality, as the one related to Table 3.2, provides the following result. We have that the inequality

$$0 \leq H\left(x,t,\bar{y},\bar{u},\bar{p}\right) - \min_{w\in K_U} H\left(x,t,\bar{y},w,\bar{p}\right) \leq 10^{-l}$$

is fulfilled at 100% of the grid points for $l = 2$ and at 99.29% of the grid points for $l = 12$ with the returned values $(\bar{y},\bar{u},\bar{p})$ of the SQH method where the minimum of H over K_U is determined with a direct search. We remark that, for $\alpha = 0$, the cost functional consists only of the control cost $|\cdot|^{\frac{1}{2}}$, which promotes sparse bang-bang solutions. For this reason, the $L^2\left(Q\right)$-cost is included to ensure that the control also takes intermediate values in K_U.



(a) The state y.

(b) The control function u as contour plot.

(c) The control function u.

(d) The control function u viewed from above.

Figure 3.4: Results with Algorithm 3.1 for the cost functional (3.49) with $g(\cdot) := |\cdot|^{\frac{1}{2}}$ and $K_U = \{-30, -15, -5, 0, 5, 15, 30\}$.

To conclude our series of experiments, we choose the following lower semi-continuous step function

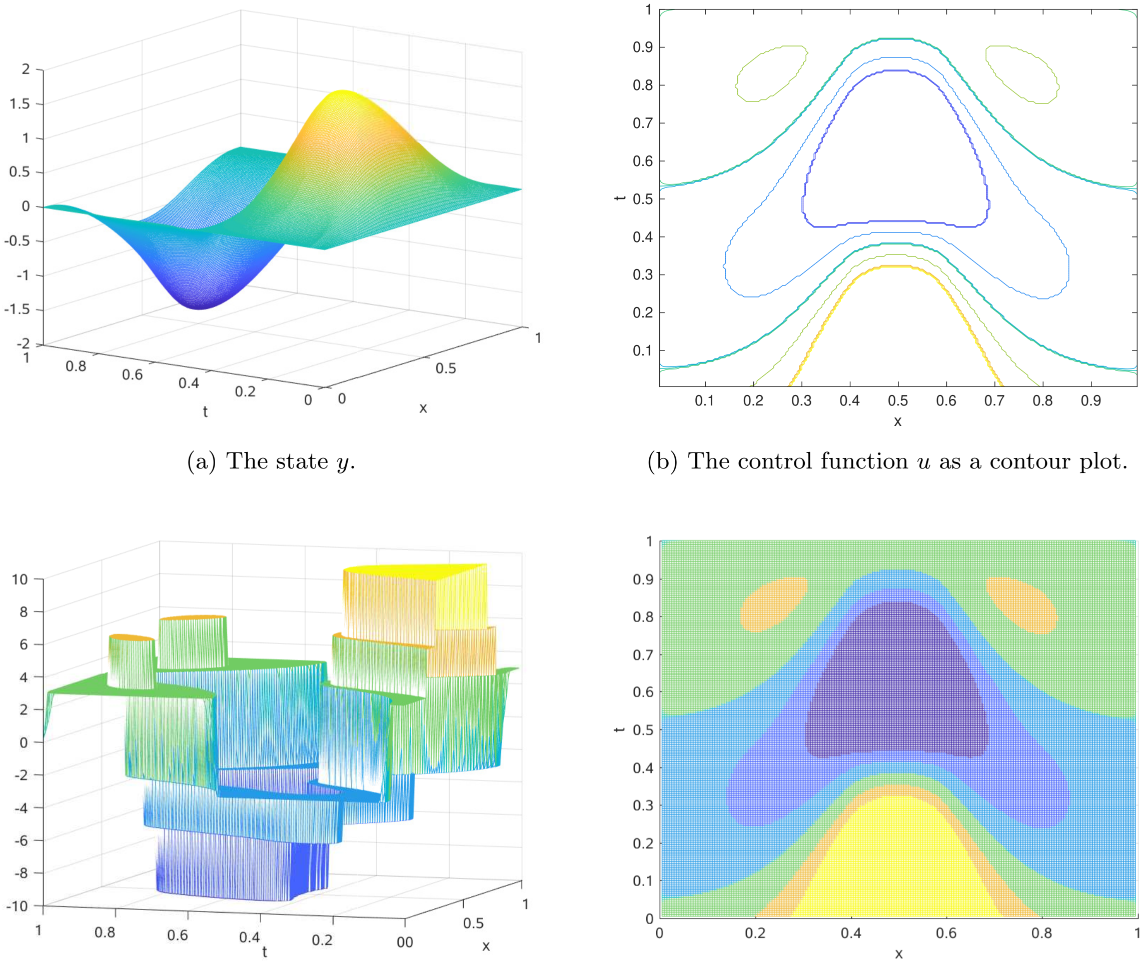
$$g(z) := g_2(z) = \begin{cases} \frac{7}{2} & \text{for } |z| > 6 \\ 1 & \text{for } 3 < |z| \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

and $K_U = [-10, 10]$. In this case, while the control function may take a continuous set of values, the cost of the control is piecewise constant. The augmented Hamiltonian is minimized by a secant method. The problem's parameters are set $N = 200$ and $N_t = 200$, $\alpha = 0$, $\beta = 10^{-1}$, $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$, $\kappa = 10^{-6}$, $u^0 = 0$ and the initial guess $\epsilon = \frac{3}{5}$. The results for this case are depicted in Figure 3.5 where one can see the stepwise structure of the control.

Besides the reduction of the functional to an observed minimum value, an analogous numerical test of optimality, as the one related to Table 3.2, provides the following result. We have that the inequality

$$0 \leq H(x, t, \bar{y}, \bar{u}, \bar{p}) - \min_{w \in K_U} H(x, t, \bar{y}, w, \bar{p}) \leq 10^{-l}$$

is fulfilled at 100% of the grid points for $l = 2$ and at 99.53% of the grid points for $l = 12$ with the returned values $(\bar{y}, \bar{u}, \bar{p})$ of the SQH method where the minimum of H over K_U is determined with a secant method.



(a) The state y.

(b) The control function u as a contour plot.

(c) The control function u.

(d) The control function u viewed from above.

Figure 3.5: Results with Algorithm 3.1 for the cost functional (3.49) with $g = g_2$ and $K_U = [-10, 10]$.

3.4.2 Application to a linear elliptic case

In this subsection, we consider a linear elliptic optimal control problem, given by $P.2$), with distributed control and a discontinuous cost functional.

We choose $Z_e = \Omega := (0,1) \times (0,1)$ and consider the following optimal control problem:

Find $y \in H_0^1(\Omega)$ and $u \in U_{ad}$ with $K_U = [0, 100]$ such that

$$\begin{aligned} \min_{y,u} J(y,u) := \int_\Omega \frac{1}{2} \left(y(x) - y_d(x)\right)^2 + g(u(x))\, dx \\ (\nabla y, \nabla v) = (u, v) \\ u \in U_{ad} \end{aligned} \tag{3.60}$$

for all $v \in H_0^1(\Omega)$ where

$$g(z) := \begin{cases} \beta|z| & \text{if } |z| > 20 \\ 0 & \text{else} \end{cases}, \quad \beta = 10^{-3}$$

and $y_d(x) := \sin(2\pi x_1)\cos(2\pi x_2) + 1$.

We have $h\left(y\right) \coloneqq \frac{1}{2}\left(y - y_d\right)^2$ and $f\left(x, y, u\right) \coloneqq u$ in terms of the general framework of Section 3.1 and the Hamiltonian is given by

$$H\left(x, y, u, p\right) \coloneqq \frac{1}{2}\left(y - y_d\right)^2 + g\left(u\right) + pu$$

according to (3.6). Corresponding to (3.5), we have the following adjoint problem

$$\left(\nabla p, \nabla v\right) = \left(y - y_d, v\right)$$

for all $v \in H_0^1\left(\Omega\right)$ where $p \in H_0^1\left(\Omega\right)$. We remark that $\|\delta y\|_{L^2(\Omega)} \leq \tilde{c}\|\nabla \delta y\|$, $\|\delta p\|_{L^2(\Omega)} \leq \tilde{c}\|\nabla \delta p\|$, $\tilde{c} \geq 0$ because of the Poincaré inequality [2, 6.7] and thus $\|\delta y\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$ and $\|\delta p\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$ because of the Cauchy-Schwarz inequality, see [2, Lemma 2.2] for a constant $c > 0$.

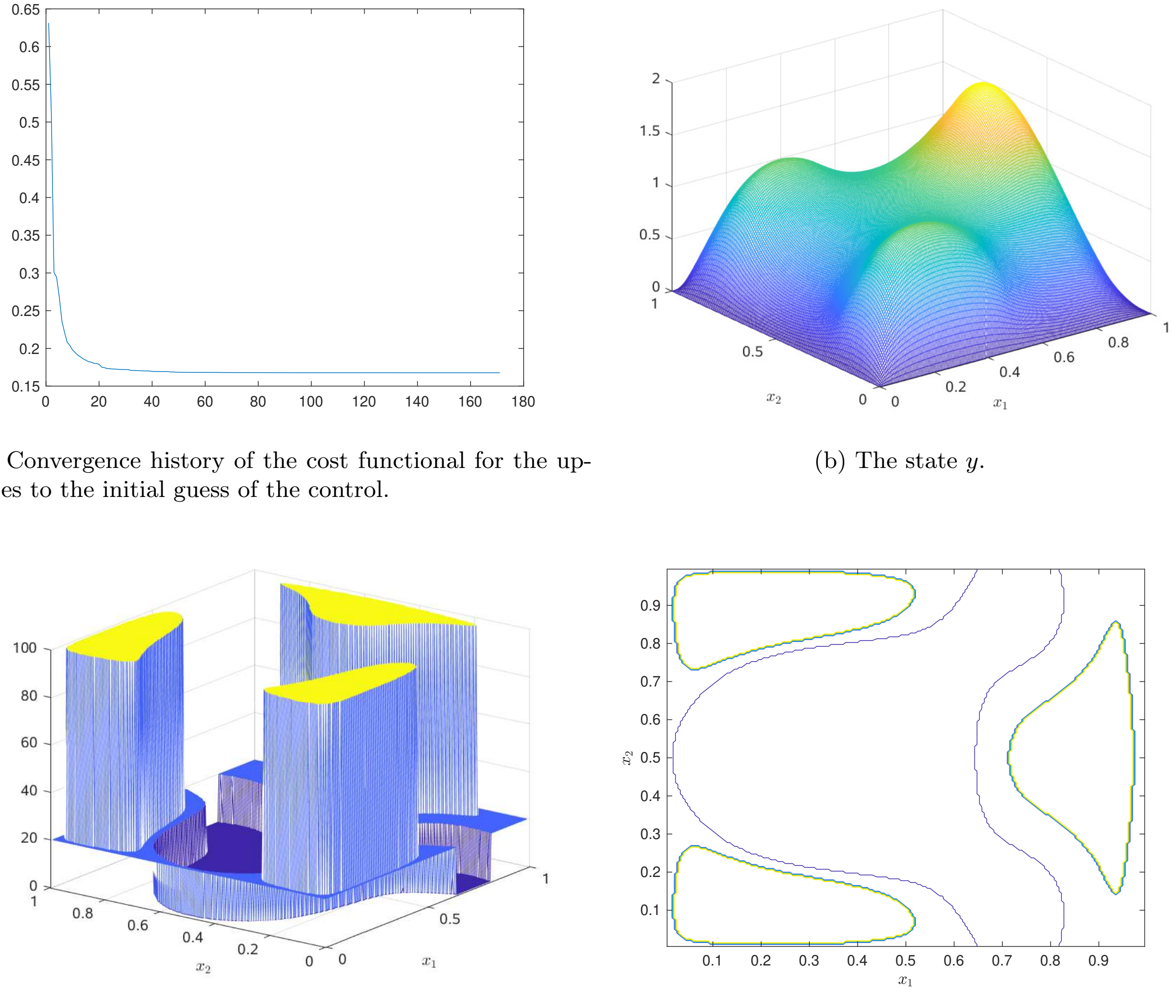
We have $\|p\|_{L^\infty(\Omega)} \leq c$ for any solution (y, u) to the state equation with $u \in U_{ad}$, see Section 3.2. Furthermore, we have that A.3) and A.6) are fulfilled and since the derivative $\frac{\partial^2}{\partial y^2}h = 1$, $\frac{\partial}{\partial y}f = \frac{\partial^2}{\partial y^2}f = 0$, we see that P.2) fits to our theoretical framework of Section 3.1 for which the analysis of the SQH method is performed in Section 3.3.

In Figure 3.6, we depict the optimal solution obtained with Algorithm 3.1 for the elliptic optimal control problem (3.60). In this case, the parameters are as follows. The initial guess for ϵ equals $\frac{1}{150}$ and we have $u^0 = 0$, $\kappa = 10^{-6}$, $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$. The domain Ω is discretized with an equidistant mesh with size $\triangle x = \frac{1}{200}$.

Although we have not checked that all the requirements of Lemma 28 hold, we observe convergence of the SQH method to a PMP consistent solution according to Theorem 27. We denote with $(\bar{y}, \bar{u}, \bar{p})$ the solution to which Algorithm 3.1 converges. The inequality

$$H\left(x, \bar{y}, \bar{u}, \bar{p}\right) - \min_{w \in K_U} H\left(x, \bar{y}, w, \bar{p}\right) \leq \text{eps},$$

eps $= 2.2 \cdot 10^{-16}$, is fulfilled at 37.39% of the grid points for $\kappa = 10^{-1}$, at 86.23% of the grid points for $\kappa = 10^{-3}$, at 99.15% of the grid points for $\kappa = 10^{-4}$, at 99.93% of the grid points for $\kappa = 10^{-6}$ and at 99.95% of the grid points for $\kappa = 10^{-8}$. Further results are given in Table 3.1.



(a) Convergence history of the cost functional for the updates to the initial guess of the control.

(b) The state y.

(c) The optimal control u.

(d) The optimal control u as a contour plot.

Figure 3.6: Results for the elliptic optimal control problem $P.2)$.

3.4.3 Application to a bilinear elliptic case

In this subsection, we consider the bilinear elliptic control problem $P.3)$ with $K_U \subseteq \mathbb{R}_0^+$ that is given by

$$\begin{aligned} \min_{y,u} J\left(y,u\right) := \int_\Omega \frac{1}{2}\left(y\left(x\right)-y_d\left(x\right)\right)^2 + g\left(u\left(x\right)\right)dx \\ \left(\nabla y, \nabla v\right) + \left(uy, v\right) = \left(\tilde{f}, v\right) \\ u \in U_{ad} \end{aligned} \tag{3.61}$$

for all $v \in H_0^1\left(\Omega\right)$, $y_d \in L^q\left(\Omega\right)$ and g is specified below. From (3.6), the corresponding Hamiltonian for (3.61) is given by

$$H\left(x,y,u,p\right) = \frac{1}{2}\left(y-y_d\right)^2 + g\left(u\right) + p\tilde{f} - uyp. \tag{3.62}$$

According to (3.5), the adjoint problem is as follows: Find $p \in H_0^1\left(\Omega\right)$ such that

$$\left(\nabla p, \nabla v\right) + \left(up, v\right) = \left(y - y_d, v\right) \tag{3.63}$$

holds for all $v \in H_0^1(\Omega)$. Now, we check that Assumptions A.1) to A.6) are fulfilled for our optimal control problem (3.61). Notice that all solutions to the state equation and (3.63) are essentially bounded by a constant for all $u \in U_{ad}$, see the discussion starting on page 75.

We have that $\|\nabla \delta y\|_{L^2(\Omega)} \leq \tilde{c}\|\delta u\|_{L^2(\Omega)}$, $\tilde{c} \geq 0$ and $\|\delta y\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$ as follows. If we define $\delta u := u_1 - u_2$ and $\delta y := y_1 - y_2$, then we obtain from taking the difference of the state equation of (3.61) for two different pairs (y_ℓ, u_ℓ), $\ell \in \{1,2\}$ the following

$$(\nabla (y_1 - y_2), \nabla v) + (u_1 y_1, v) - (u_2 y_2, v) = 0,$$

equivalently

$$(\nabla (y_1 - y_2), \nabla v) + (u_2 (y_1 - y_2), v) = (-(u_1 - u_2) y_1, v). \tag{3.64}$$

By choosing $v = y_1 - y_2$, we have from (3.64) the following

$$\|\nabla (y_1 - y_2)\|_{L^2(\Omega)}^2 \leq \|y_1\|_{L^\infty(\Omega)} \|u_1 - u_2\|_{L^2(\Omega)} \|y_1 - y_2\|_{L^2(\Omega)} \tag{3.65}$$

with the Cauchy-Schwarz inequality, see [2, Lemma 2.2] and that $u_2 \geq 0$ almost everywhere. If we use the Poincaré inequality [2, 6.7], we obtain $\|\delta y\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$.

To discuss the boundedness of the adjoint, we first subtract (3.63) for $(y,u,p) \leftarrow (y_2,u_2,p_2)$ from (3.63) for $(y,u,p) \leftarrow (y_1,u_1,p_1)$ and obtain

$$(\nabla \delta p, \nabla v) + (u_2 \delta p, v) = (\delta y, v) - (p_1 (u_1 - u_2), v)$$

where $\delta p := p_1 - p_2$.

Because $u_2 \geq 0$ almost everywhere and $p_1 \in L^\infty(\Omega)$, we have that $\|\nabla \delta p\|_{L^2(\Omega)} \leq \tilde{c}\|\delta u\|_{L^2(\Omega)}$, $\tilde{c} > 0$ if we choose $v = \delta p \in H_0^1(\Omega)$ and use the Cauchy-Schwarz inequality, see [2, Lemma 2.2] with $\|\delta y\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$. By the Poincaré inequality [2, 6.7] we obtain

$$\|\delta p\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$$

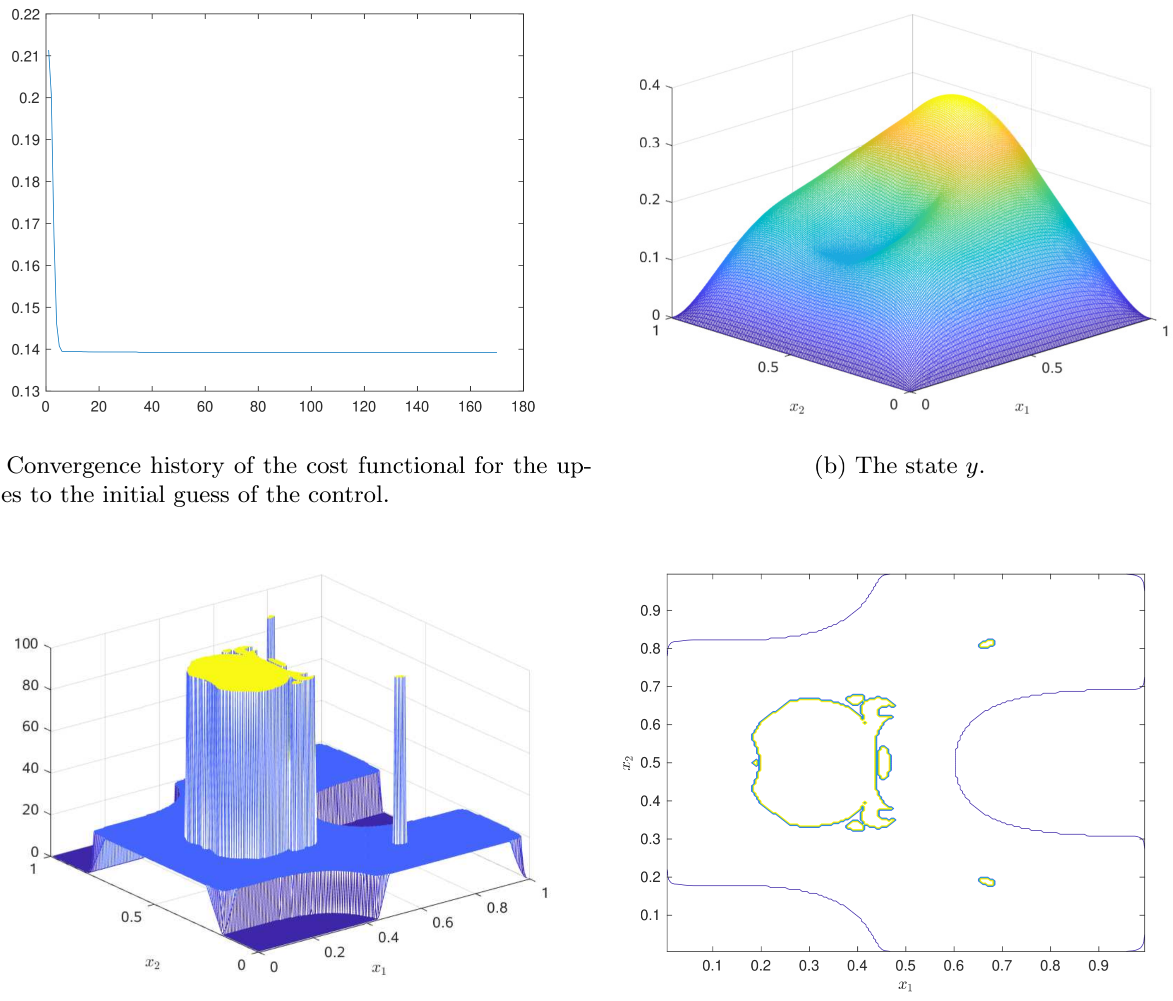
for a $c > 0$. Consequently we have checked A.2).

As $f = \tilde{f} - uy$, we have that A.1), A.5) and A.6) are fulfilled. By the essential boundedness of any solution to the state equation by a constant for all $u \in U_{ad}$, we have that A.3) is fulfilled. Now we have checked that our theoretical framework of Section 3.3 fits to our case (3.61).

Next, we choose $\Omega = (0,1) \times (0,1)$ and

$$g(z) := \begin{cases} \beta|z| & \text{if } |z| > 20 \\ 0 & \text{else} \end{cases}, \quad \beta = 10^{-3}$$

and $y_d(x) := \sin(2\pi x_1)\cos(2\pi x_2)$. In Figure 3.7, we depict the solution obtained with Algorithm 3.1 solving (3.61) with $\tilde{f} = 10$. The parameters are as follows. The initial guess for ϵ equals $\frac{1}{150}$ and we have $u^0 = 0$, $\kappa = 10^{-6}$, $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$, $K_U = [0,100]$. The domain Ω is discretized with an equidistant mesh with size $\triangle x = \frac{1}{200}$.



(a) Convergence history of the cost functional for the updates to the initial guess of the control.

(b) The state y.

(c) The optimal control u.

(d) The optimal control u as a contour plot.

Figure 3.7: Solution to the elliptic bilinear optimal control problem $P.3$).

Notice that the two peaks of the control appear also at finer discretization, for example $\triangle x = \frac{1}{500}$, and thus they are not numerical artefacts.

Although we have not checked that all the requirements of Lemma 28 hold, we observe convergence of the SQH method to a PMP consistent solution according to Theorem 27. We denote with $(\bar{y}, \bar{u}, \bar{p})$ the solution obtained with Algorithm 3.1. The inequality

$$H\left(x, \bar{y}, \bar{u}, \bar{p}\right) - \min_{w \in K_U} H\left(x, \bar{y}, w, \bar{p}\right) \leq \text{eps},$$

eps $= 2.2 \cdot 10^{-16}$, is fulfilled at 73.31% of the grid points for $\kappa = 10^{-1}$, at 84.28% of the grid points for $\kappa = 10^{-3}$, at 90.80% of the grid points for $\kappa = 10^{-6}$, at 94.21% of the grid points for $\kappa = 10^{-10}$. See also Table 3.1 for additional results.

3.4.4 Application to a bilinear parabolic case

In this subsection, we consider the bilinear parabolic control problem P.4) with $K_U \subseteq \mathbb{R}_0^+$ and we present numerical results for the bilinear parabolic control problem P.4). We have

$$\begin{aligned} &\min_{y,u} \int_0^T \int_\Omega \frac{1}{2} \left(y\left(x,t\right) - y_d\left(x,t\right)\right)^2 + g\left(u\left(x,t\right)\right) dxdt \\ \text{s.t. } &\left(y'\left(\cdot,t\right),v\right) + D\left(\nabla y\left(\cdot,t\right),\nabla v\right) + \left(u\left(\cdot,t\right)y\left(\cdot,t\right),v\right) = \left(\tilde{f}\left(\cdot,t\right),v\right) \text{ in } Q \text{ for all } v \in H_0^1\left(\Omega\right) \\ &u \in U_{ad} \end{aligned}$$

for almost all $t \in (0,T)$ with $y\left(x,t\right) = 0$ for $x \in \partial\Omega$ and $y\left(0,x\right) = 0$ for $x \in \Omega$. Further, we have $D = \frac{1}{5}$, $T = 1$, $\Omega = (0,1)$, $Q = (0,1) \times (0,1)$, $K_U = [0,15]$

$$y_d\left(x,t\right) = \begin{cases} \frac{1}{2} & \text{if } \bar{x}\left(t\right) - \frac{7}{100} \leq x \leq \bar{x}\left(t\right) + \frac{7}{100} \\ 0 & \text{else} \end{cases}$$

where $\bar{x}\left(t\right) := \frac{1}{2} + \frac{2}{5}\sin\left(2\pi t\right)$,

$$g\left(z\right) := \begin{cases} \beta|z| & \text{if } |z| > 10 \\ 0 & \text{else} \end{cases}, \quad \beta = 10^{-4}$$

and we have that $\tilde{f}$ is a constant function with value 1.

In this case, the Hamiltonian is given by

$$H\left(x,t,y,u,p\right) := \frac{1}{2}\left(y - y_d\right)^2 + g\left(u\right) + p\tilde{f} - uyp$$

and the adjoint problem is given by

$$-\left(p'\left(\cdot,t\right),v\right) + D\left(\nabla p\left(\cdot,t\right),\nabla v\right) = \int_\Omega \left(\left(y\left(x,t\right) - y_d\left(x,t\right)\right) - u\left(x,t\right)p\left(x,t\right)\right)v\left(x\right)dx$$

for all $v \in H_0^1\left(\Omega\right)$ with $p\left(\cdot,T\right) = 0$ which has a unique and by a constant essentially bounded solution $p \in L^2\left(0,T,H^2\left(\Omega\right)\right) \cap L^\infty\left(0,T;H_0^1\left(\Omega\right)\right)$ for all $u \in U_{ad}$ according to Section 3.2. Therefore A.1) to A.6) are proved as in the elliptic bilinear case in Subsection 3.4.3 where we only consider A.2) closer. With the same arguments as for the elliptic case in Subsection 3.4.3 we obtain for any $t \in (0,T)$ that

$$\|\delta y\left(\cdot,t\right)\|_{L^2(\Omega)}^2 \leq c^2 \|\delta u\left(\cdot,t\right)\|_{L^2(\Omega)}^2$$

for a constant $c > 0$ which gives by integrating over t the following

$$\int_0^T \int_\Omega \delta y^2\left(x,t\right)dxdt \leq c^2 \int_0^T \int_\Omega \delta u^2\left(x,t\right)dxdt.$$

By extracting a root on both sides we have that

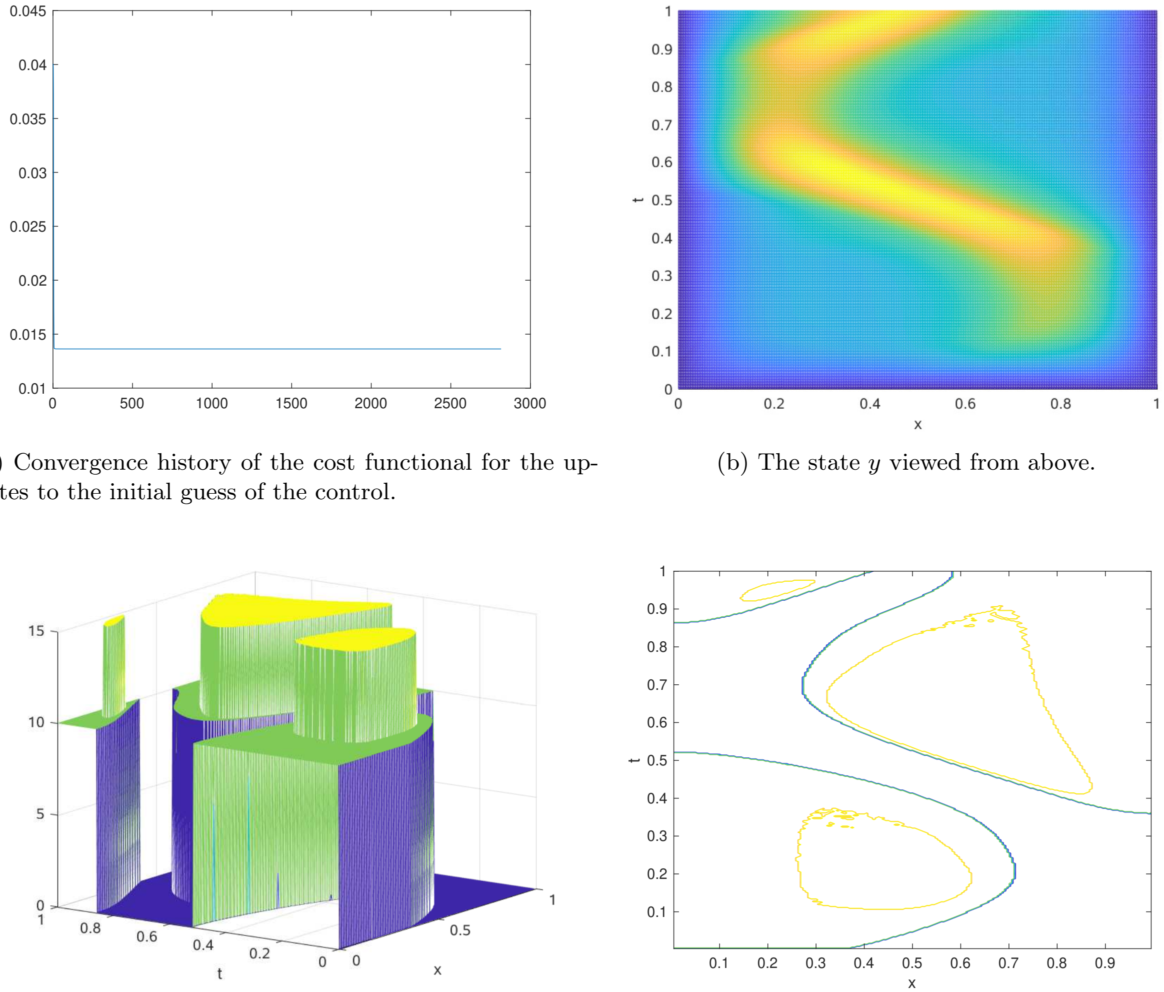
$$\|\delta y\|_{L^2(Q)} \leq c\|\delta u\|_{L^2(Q)}.$$

Analogously we obtain

$$\|\delta p\|_{L^2(Q)} \leq \tilde{c}\|\delta u\|_{L^2(Q)}$$

for a constant $\tilde{c} > 0$ which means that A.2) is fulfilled and thus our theoretical framework of Section 3.1 fits to P.4).

The parameters for the numerical experiment are as follows. The initial guess for $\epsilon = \frac{3}{5}$ and for u is the zero function. The parameters are set as follows $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-12}$, $\kappa = 10^{-12}$. The discretization is equidistant in time and space with $\triangle t = \frac{1}{400}$ and $\triangle x = \frac{1}{200}$. The results are presented in Figure 3.8.



(a) Convergence history of the cost functional for the updates to the initial guess of the control.

(b) The state y viewed from above.

(c) The control u.

(d) A contour plot of the control u.

Figure 3.8: Numerical results for the parabolic bilinear optimal control problem $P.4$).

Also in the present parabolic case, we have that although the requirements of Lemma 28 are not checked if they hold, we observe convergence of the SQH method to a PMP consistent solution according to Theorem 27. We denote with $(\bar{y}, \bar{u}, \bar{p})$ the solution obtained by Algorithm 3.1. The inequality

$$H\left(x, t, \bar{y}, \bar{u}, \bar{p}\right) - \min_{w \in K_U} H\left(x, t, \bar{y}, w, \bar{p}\right) \leq \text{eps},$$

$\text{eps} = 2.2 \cdot 10^{-16}$, is fulfilled at 89.27% of the grid points for $\kappa = 10^{-4}$, at 93.64% of the grid points for $\kappa = 10^{-6}$, at 97.06% of the grid points for $\kappa = 10^{-8}$, at 97.59% of the grid points for $\kappa = 10^{-10}$ and at 98.04% of the grid points for $\kappa = 10^{-12}$.

3.4.5 Application to a non-linear elliptic case

In this subsection, we discuss $P.5)$ that is given by

$$\begin{aligned} \min_{y,u} J(y,u) := \int_\Omega \frac{1}{2}\left(y(x) - y_d(x)\right)^2 + g(u(x))\,dx \\ (\nabla y, \nabla v) + \left(y^3, v\right) = (u,v) \\ u \in U_{ad} \end{aligned} \tag{3.66}$$

where we choose $\Omega := (0,1) \times (0,1)$,

$$g(z) := \begin{cases} \beta|z| & \text{if } |z| > 20 \\ 0 & \text{else} \end{cases}, \quad \beta = 10^{-3},$$

$K_U := [-100, 100]$ and $y_d(x) := \sin(2\pi x_1)\cos(2\pi x_2)$.

We have $h(y) := \frac{1}{2}(y - y_d)^2$ and $f(x,y,u) := u - y^3$ in the general setting of Section 3.1 and we define the following Hamiltonian

$$H(x,y,u,p) := \frac{1}{2}(y - y_d)^2 + g(u) + p\left(u - y^3\right)$$

according to (3.6). Corresponding to (3.5), we have the following adjoint problem for $p \in H_0^1(\Omega)$ given by

$$(\nabla p, \nabla v) + \left(3y^2 p, v\right) = (y - y_d, v)$$

for all $v \in H_0^1(\Omega)$. Since $y^4 \geq 0$ and $3y^2p^2 \geq 0$ we have that

$$\|\delta y\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$$

and

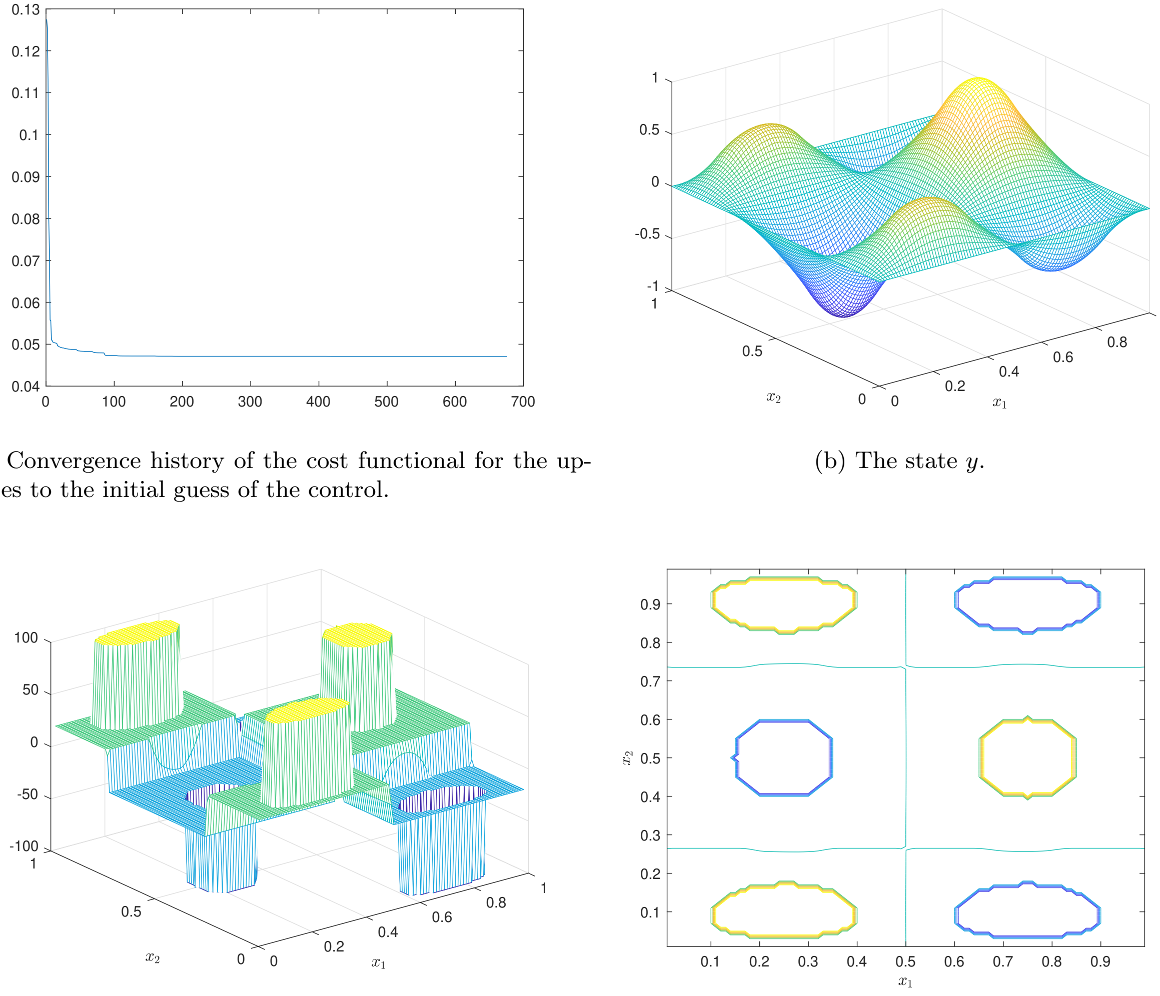
$$\|\delta p\|_{L^2(\Omega)} \leq c\|\delta u\|_{L^2(\Omega)}$$

for a constant $c > 0$ analogous to Subsection 3.4.2 where the linear elliptic case is discussed. We have that A.1), A.3) and A.6) are fulfilled. Furthermore, we have

$$\|p\|_{L^\infty(\Omega)} \leq c$$

for any solution (y,u) to the state equation with $u \in U_{ad}$, see Section 3.2. Since the derivatives $\frac{\partial^2}{\partial y^2}h = 1$, $\frac{\partial}{\partial y}f = 3y^2$, $\frac{\partial}{\partial y}f = 6y$ are bounded, because y is essentially bounded by a constant for all $u \in U_{ad}$, see page 75, we have that $P.5)$ fits to our theoretical framework of Section 3.1 for which the analysis of the SQH method is performed in Section 3.3.

The parameters in Algorithm 3.1 are chosen as follows. The initial guess for ϵ equals $\frac{1}{150}$ and we have $u^0 = 0$, $\kappa = 10^{-12}$, $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$. The domain Ω is discretized with a mesh of size $\triangle x = \frac{1}{100}$. The non-linear PDE is solved until the L^2-norm of its residuum is less than 10^{-6}. The results are shown in Figure 3.9 where one can see the action of the bounds on the control and of the discontinuous control costs.



(a) Convergence history of the cost functional for the updates to the initial guess of the control.

(b) The state y.

(c) The optimal control u.

(d) The optimal control u as a contour plot.

Figure 3.9: Solution to the non-linear elliptic optimal control problem $P.5$).

Next, although we have not checked that all the requirements of Lemma 28 hold, we observe convergence of the SQH method to a PMP consistent solution according to Theorem 27. We denote with $(\bar{y}, \bar{u}, \bar{p})$ the solution obtained with Algorithm 3.1. The inequality

$$H\left(x, t, \bar{y}, \bar{u}, \bar{p}\right) - \min_{w \in K_U} H\left(x, t, \bar{y}, w, \bar{p}\right) \leq 10^{-7}$$

is fulfilled at 37.68% of the grid points for $\kappa = 10^{-2}$, is fulfilled at 79.88% of the grid points for $\kappa = 10^{-4}$, at 94.34% of the grid points for $\kappa = 10^{-6}$, at 94.83% of the grid points for $\kappa = 10^{-8}$ and at 95.25% of the grid points for $\kappa = 10^{-12}$. Notice that we use the smaller tolerance 10^{-7} instead of the machine precision $2.2 \cdot 10^{-16}$ since the state equation is only solved to a tolerance of 10^{-6} and not exactly as in the linear case.

3.4.6 Application to a state-constrained optimal control problem

In this subsection, we discuss $P.6)$ that is given by

$$\begin{aligned} &\min_{y,u} J(y,u) := \int_\Omega h(y(x)) + g(u(x))\, dx \\ &\text{s.t. } (\nabla y, \nabla v) = (u,v) \\ &y \leq \xi \\ &u \in U_{ad} \end{aligned} \tag{3.67}$$

for all $v \in H_0^1(\Omega)$ where $h(y) := \frac{1}{2}(y - y_d)^2$,

$$g(u) := \begin{cases} \beta|u| & \text{if } |u| > 20 \\ 0 & \text{else} \end{cases}, \quad \beta = 10^{-3},$$

$K_U := [-100, 100]$, $\xi \in \mathbb{R}$ and we assume that (3.67) admits a solution denoted with $(\bar{y}, \bar{u})$.

In this case, PMP optimality involves multipliers that are only implicitly characterized by inequalities, see for example [28]. For this reason, the computation of solutions to (3.67) is a delicate issue. We follow the idea of augmented Lagrangian [61, Section 3] and transform the optimal control problem (3.67) into the following

$$\begin{aligned} &\min_{y,u} J(y,u;\xi,\gamma) := \int_\Omega h_\xi(y(x);\gamma) + g(u(x))\, dx \\ &\text{s.t. } (\nabla y, \nabla v) = (u,v) \\ &u \in U_{ad} \end{aligned} \tag{3.68}$$

where $h_\xi(y;\gamma) := h(y) + \gamma(\max(0, y-\xi))^3$, $\gamma \geq 0$.

Since we require $h_\xi(y;\gamma)$ to be twice continuously differentiable in order to fulfill Assumption A.1), we choose $(\max(0, y-\xi))^3$. The differentiability of $(\max(0, y-\xi))^3$ can be shown as follows. First, we have that

$$\begin{aligned} &\frac{\partial}{\partial y}(\max(0, y-\xi))^2 \\ &= \begin{cases} \frac{\partial}{\partial y}(y-\xi)^2 & \text{if } y-\xi > 0 \\ \begin{cases} 0 & ,h<0 \\ \lim_{h\to 0}\frac{h^2}{h} & ,h>0 \end{cases} & \text{if } y-\xi = 0 \\ \frac{\partial}{\partial y} 0 & \text{if } y-\xi < 0 \end{cases} = \begin{cases} 2(y-\xi) & \text{if } y-\xi > 0 \\ 0 & \text{if } y-\xi = 0 \\ 0 & \text{if } y-\xi < 0 \end{cases} = 2\max(0, y-\xi) \end{aligned} \tag{3.69}$$

is differentiable and thus

$$(\max(0, y-\xi))^3 = \left((\max(0, y-\xi))^2\right)^{\frac{3}{2}}$$

is differentiable due to the chain rule, see [4, VII Theorem 3.3], where the first derivative is given by

$$\frac{\partial}{\partial y}(\max(0, y-\xi))^3 = \begin{cases} 3(y-\xi)^2 & \text{if } y-\xi > 0 \\ 0 & \text{if } y-\xi = 0 \\ 0 & \text{if } y-\xi < 0 \end{cases} = 3(\max(0, y-\xi))^2$$

and according to (3.69) the second derivative is given as follows

$$\frac{\partial^2}{\partial y^2}(\max(0, y-\xi))^3 = 3\frac{\partial}{\partial y}(\max(0, y-\xi))^2 = 6\max(0, y-\xi).$$

Furthermore the second derivative of $(\max(0, y - \xi))^3$ with respect to y is continuous since

$$\lim_{y \to \xi} 6\,(y - \xi) = 0.$$

We assume that (3.68) admits a solution for any $\gamma \geq 0$. Analogously to (3.60), we have that a solution to (3.68) is characterized by the PMP as follows. The adjoint is given by

$$(\nabla p, \nabla v) = \left(y - y_d + 3\gamma\,(\max(0, y - \xi))^2, v \right) \tag{3.70}$$

for all $v \in H_0^1(\Omega)$. Because $y \in H_0^1(\Omega)$ and thus measurable, see [1, page 60], we have that $\max(0, y - \xi)$ is measurable, see [36, Proposition 2.1.4] and with Lemma 51, we have that $(\max(0, y - \xi))^2$ is measurable. We have that $y \in L^\infty(\Omega)$, see the discussion on page 75, from which it follows that A.5) is fulfilled and since we require that $y_d \in L^q(\Omega)$, we have that $y - y_d + 3\,(\max(0, y - \xi))^2 \in L^q(\Omega)$. From this it follows with an analogous discussion as for (3.15) that (3.70) is uniquely solvable in $H_0^1(\Omega)$ and A.4) is fulfilled. For the characterization with the PMP, we need the intermediate adjoint equation according to (3.10) that is given by

$$(\nabla \tilde{p}, \nabla v) = \left(\frac{1}{2}(y_1 + y_2) - y_d + 3\gamma \int_0^1 (\max(0, y_2 + \theta\,(y_1 - y_2) - \xi))^2\, d\theta, v \right)$$

for all $v \in H_0^1(\Omega)$ where y_1 solves the state equation for $u \leftarrow u_1$ and y_2 solves the state equation for $u \leftarrow u_2$. Next, we show that the intermediate adjoint is well defined. For this purpose, we have to see that the function

$$x \mapsto \int_0^1 (\max(0, y_2(x) + \theta\,(y_1(x) - y_2(x)) - \xi))^2\, d\theta$$

is measurable. We have that according to Lemma 52 the functions $(\theta, x) \mapsto \theta$, $(\theta, x) \mapsto y_1(x)$, $(\theta, x) \mapsto y_2(x)$ are Lebesgue measurable on $[0,1] \times \Omega$. Then we have that $(\theta, x) \mapsto y_2(x) + \theta\,(y_1(x) - y_2(x)) - \xi$ is measurable on $[0,1] \times \Omega$, see [36, Proposition 2.1.7] and consequently

$$(\theta, x) \mapsto (\max(0, y_2(x) + \theta\,(y_1(x) - y_2(x)) - \xi))^2,$$

see [36, Proposition 2.1.4] and Lemma 51. Then by Tonelli's Theorem [5, X Theorem 6.7 ii)], we obtain that

$$x \mapsto \int_0^1 (\max(0, y_2(x) + \theta\,(y_1(x) - y_2(x)) - \xi))^2\, d\theta$$

is measurable and due to the discussion in Section 3.2 we have $y_1, y_2 \in L^\infty(\Omega)$ from which it follows that

$$x \mapsto \int_0^1 (\max(0, y_2(x) + \theta\,(y_1(x) - y_2(x)) - \xi))^2\, d\theta \in L^\infty(\Omega).$$

Analogous to the adjoint equation, the intermediate equation is uniquely solvable in $H_0^1(\Omega)$. To finalize the PMP characterization we consider the difference of the adjoint equation for $y \leftarrow y_2$ and the intermediate adjoint equation that is given by

$$(\nabla(\tilde{p} - p), \nabla v) = \left(\frac{1}{2}(y_1 - y_2) + 3\gamma \int_0^1 (\max(0, y_2 + \theta\,(y_1 - y_2) - \xi))^2 - (\max(0, y_2 - \xi))^2\, d\theta, v \right)$$

where

$$\begin{aligned} &(\max(0, y_2 + \theta\,(y_1 - y_2) - \xi))^2 - (\max(0, y_2 - \xi))^2 \\ &= 2 \int_0^1 \max(0, y_2 + \eta\,(\theta\,(y_1 - y_2)) - \xi)\, d\eta\,(\theta\,(y_1 - y_2)) \end{aligned}$$

with the fundamental theorem of calculus [4, VI 4.13]. With this equations we can conclude that

$$\|\tilde{p} - p\|_{L^\infty(\Omega)} \leq C\|y_1 - y_2\|_{L^q(\Omega)}$$

for a constant $C > 0$ since y_1, y_2 are bounded by a constant for all $u \in U_{ad}$, see Section 3.2. From this it follows, analogous to Section 3.2, the PMP characterization of a solution to (3.68). Analogously we consider the difference of (3.70) for different y_1, y_2 which are solutions to the state equation for the corresponding $u_1, u_2 \in U_{ad}$. We obtain

$$\begin{aligned}(\nabla (p_1 - p_2), \nabla v) &= \left(y_1 - y_2 + 3\gamma \left((\max (0, y_1 - \xi))^2 - (\max (0, y_2 - \xi))^2\right), v\right)\\ &= \left(y_1 - y_2 + 6\gamma \int_0^1 \max (0, y_2 - \theta (y_1 - y_2))\, d\theta\, (y_1 - y_2), v\right)\end{aligned}$$

which provides A.2) since y_1, y_2 are essentially bounded by a constant for all $u_1, u_2 \in U_{ad}$, see discussion on page 75 and $\|y_1 - y_2\|_{L^2(\Omega)} \leq c\|u_1 - u_2\|_{L^2(\Omega)}$, $c > 0$, see Subsection 3.4.2. Consequently, the theoretical framework of Section 3.3 holds for (3.68).

Henceforth, we use our SQH scheme to solve (3.68) for increasing γ denoted by $\gamma = \gamma_k$ for increasing k. Let (y_k, u_k) be a corresponding solution to (3.68) and $(\bar{y}, \bar{u})$ be a solution to (3.67). We show in the next theorem that increasing γ_k improves the solution to (3.68) with respect to the original task of solving the state-constrained optimal control problem (3.67). Specifically, we have that the measure of the state violation by the corresponding state y_k goes to zero for increasing k. Summarizing, we show that solving (3.68) with the SQH method provides a solution that fulfills the state constraint up to a tolerance depending on γ_k and results in a value $J(y_k, u_k)$ that is smaller than $J(\bar{y}, \bar{u})$. In addition if g is a square function, it can be proven that for increasing γ_k the sequence (y_k, u_k) converges to $(\bar{y}, \bar{u})$, see for example [61, Lemma 3.6] for details.

Theorem 30. *Let* $\lim_{k\to\infty} \gamma_k = \infty$*, let* (y_k, u_k) *be a corresponding solution to* (3.68) *with*

$$M_k := \{x \in \Omega |\ y_k(x) > \xi\}$$

and let $(\bar{y}, \bar{u})$ *be a solution to* (3.67)*. Then, we have that*

$$\lim_{k\to\infty} \int_{M_k} (y_k(x) - \xi)^3\, dx = 0$$

and

$$J(y_k, u_k) = \int_\Omega h(y_k(x)) + g(u_k(x))\, dx \leq J(\bar{y}, \bar{u})$$

for all $k \in \mathbb{N}$.

Proof. First, the set M_k is measurable as y_k is measurable, see [5, X Theorem 1.9] for details. Thus integration over M_k is well defined. We have that

$$J(\bar{y}, \bar{u}) = \int_\Omega h(\bar{y}(x)) + g(\bar{u}(x)) + \gamma (\max (0, \bar{y}(x) - \xi))^3\, dx = J(\bar{y}, \bar{u}; \xi, \gamma)$$

as $\bar{y} \leq \xi$ and thus for an optimal solution (y_k, u_k) to (3.68) it holds that

$$J(y_k, u_k; \xi, \gamma) \leq J(\bar{y}, \bar{u}). \tag{3.71}$$

By inserting the definition of $J(y_k, u_k; \xi, \gamma)$, see (3.68), we have from (3.71) the following

$$\int_\Omega h(y_k(x)) + g(u_k(x)) + \gamma_k (\max (0, y_k(x) - \xi))^3\, dx \leq J(\bar{y}, \bar{u}). \tag{3.72}$$

Now if we assume that there is an $\epsilon > 0$ such that

$$\int_\Omega \left(\max\left(0, y_k\left(x\right) - \xi\right)\right)^3 dx = \int_{M_k} \left(y_k\left(x\right) - \xi\right)^3 dx > \epsilon$$

for all $k \in \mathbb{N}$, then we have a contradiction to (3.72) due to the lower boundedness of h and g. Also from (3.72) we have that

$$J\left(\bar{y}, \bar{u}\right) \geq \int_\Omega h\left(y_k\left(x\right)\right) + g\left(u_k\left(x\right)\right) + \gamma_k\left(\max\left(0, y_k\left(x\right) - \xi\right)\right)^3 dx \geq \int_\Omega h\left(y_k\left(x\right)\right) + g\left(u_k\left(x\right)\right) dx.$$

□

We remark that the arguments of the proof of Theorem 30 are not restricted to the elliptic optimal control problem (3.67) but also hold in the general framework of Section 3.1. That means that they also hold for a state-constrained optimal control problem corresponding to (3.3) for h and g bounded from below.

For our numerical experiment, we choose the initial guess $\epsilon = \frac{1}{150}$ and we have $u^0 = 0$, $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$, $\xi = \frac{3}{5}$, $\triangle x = \frac{1}{100}$, $\kappa = 10^{-10}$, $K_U = [-100, 100]$ and $y_d\left(x\right) := \sin\left(2\pi x_1\right)\cos\left(2\pi x_2\right)$. The minimum of the augmented Hamiltonian

$$K_\epsilon\left(x, y, u, v, p\right) := \frac{1}{2}\left(y - y_d\right)^2 + \gamma\left(\max\left(0, y - \xi\right)\right)^3 + g\left(u\right) + pu + \epsilon\left(u - v\right)^2$$

in Step 2 of Algorithm 3.1 is determined pointwise with an exact formula as follows. Analogous to the discussion in Subsection 3.4.1, the candidates at which a minimum of the augmented Hamiltonian is located are given by

$$u_1 = \min\left(\max\left(20, \frac{2\epsilon u^k\left(x\right) - p^k\left(x\right) - \beta}{\alpha + 2\epsilon}\right), 100\right),$$

$$u_2 = \min\left(\max\left(-100, \frac{2\epsilon u^k\left(x\right) - p^k\left(x\right) + \beta}{\alpha + 2\epsilon}\right), -20\right)$$

or

$$u_3 = \min\left(\max\left(-20, \frac{2\epsilon u^k\left(x\right) - p^k\left(x\right)}{\alpha + 2\epsilon}\right), 20\right).$$

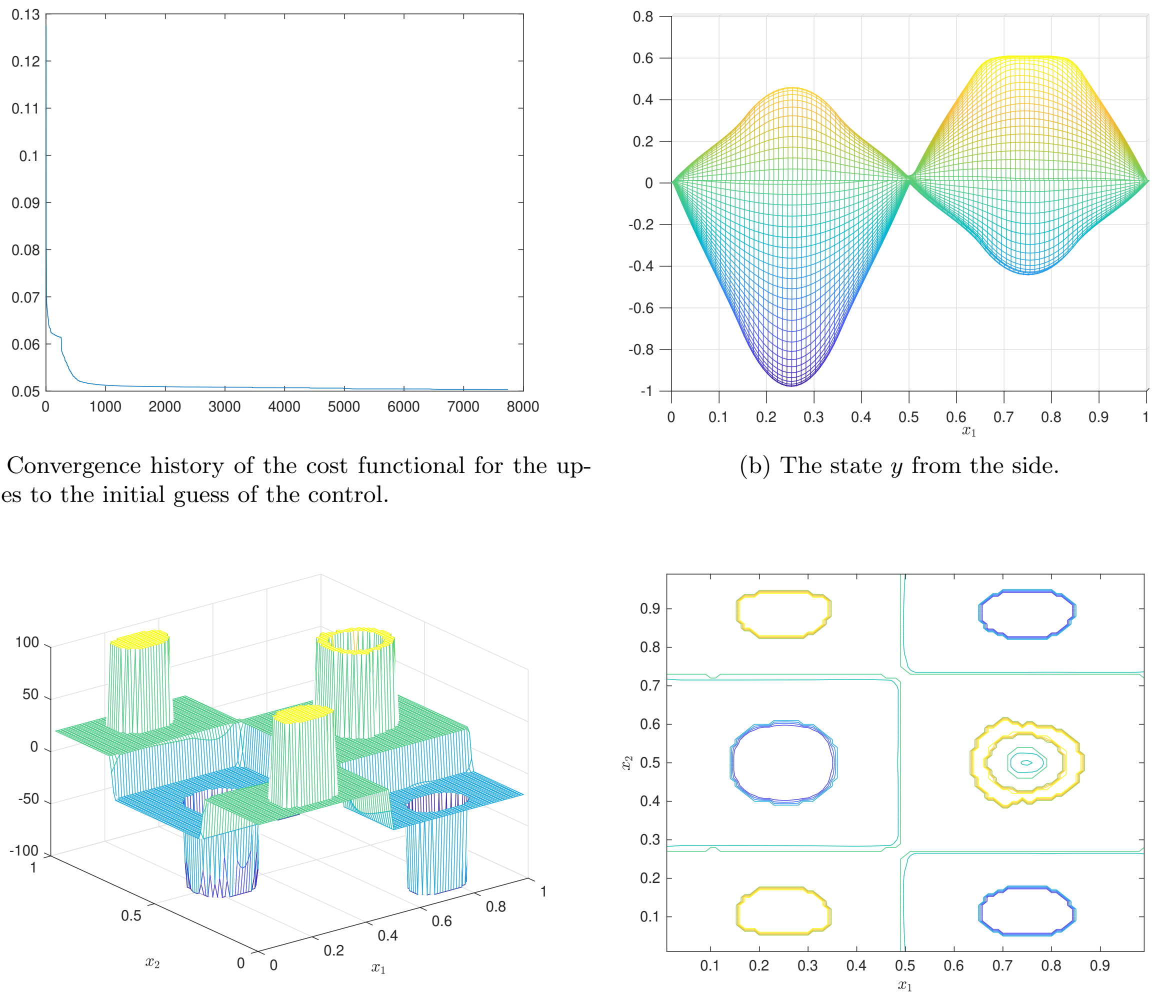
Consequently, the update for the control is pointwise given by

$$u\left(x\right) = \underset{w \in \{u_1, u_2, u_3\}}{\arg\min}\ K_\epsilon\left(x, y^k, w, u^k, p^k\right).$$

In Table 3.8, we show results that validate Theorem 30. We can see that for increasing γ the maximum of the state variable y converges to the upper bound of the state. Additionally, the measure of the set M_k where the state variable violates the upper bound becomes smaller when γ increases. According to Theorem 30 the quantity $\int_{M_k}\left(y_k\left(x\right) - \xi\right)^3 dx$ converges to zero for increasing γ. In Figure 3.10, we depict the state and the control for $\gamma = 100000$.

γ	$\max_{x\in\Omega} y\left(x\right)$	$\int_{M_k}\left(y_k\left(x\right) - \frac{3}{5}\right)^3 dx$	$\lvert M_k\rvert$	$J\left(y, u; \frac{3}{5}\right)$
1	0.8218	$1.7843 \cdot 10^{-4}$	0.0461	0.0474
10	0.7125	$2.1182 \cdot 10^{-5}$	0.0383	0.0480
100	0.6543	$1.3580 \cdot 10^{-6}$	0.0289	0.0484
1000	0.6237	$7.9157 \cdot 10^{-8}$	0.0185	0.0487
10000	0.6081	$2.9827 \cdot 10^{-9}$	0.0137	0.0489
100000	0.6032	$1.1971 \cdot 10^{-10}$	0.0104	0.0503

Table 3.8: Results that numerically validate Theorem 30 where (y, u) is obtained with the SQH method.



(a) Convergence history of the cost functional for the updates to the initial guess of the control.

(b) The state y from the side.

(c) The control u.

(d) The control u as a contour plot.

Figure 3.10: Solution to the optimal control problem (3.68) corresponding to $P.6)$ for $\gamma = 100000$.

Similar as for the previous cases, we have the following for the solution to (3.68) with $\gamma = 100000$. The inequality

$$H\left(x, t, \bar{y}, \bar{u}, \bar{p}\right) - \min_{w \in K_U} H\left(x, t, \bar{y}, w, \bar{p}\right) \leq 2.2 \cdot 10^{-16}$$

is fulfilled at 6.45% of the grid points for $\kappa = 10^{-4}$, at 76.53% of the grid points for $\kappa = 10^{-8}$ and at 81.16% of the grid points for $\kappa = 10^{-12}$. Consequently we obtain a PMP consistent solution from the SQH method according to Theorem 27 although we have not checked that all the requirements of Lemma 28 hold.

3.4.7 Application to an elliptic optimal control problem with L^1-tracking term

In this subsection, we consider the non-smooth optimal control problem $P.7)$ that is given by

$$\begin{aligned} \min_{y,u} J\left(y, u\right) &:= \int_\Omega \left|y\left(x\right) - y_d\left(x\right)\right| + g\left(u\left(x\right)\right) dx \\ \left(\nabla y, \nabla v\right) + \left(\max\left(0, y\right), v\right) &= \left(u, v\right) \\ u &\in U_{ad} \end{aligned} \tag{3.73}$$

for all $v \in H_0^1(\Omega)$ where $\Omega \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, is open and bounded, $y_d \in L^1(\Omega)$, $g : \mathbb{R} \to \mathbb{R}$, $z \mapsto g(z)$ lower semi-continuous and non-negative with $\int_\Omega g(u(x))\,dx < \infty$ for all $u \in U_{ad}$. In the experiment, we choose

$$g(z) := \beta \ln(1 + |z|) \quad \beta > 0.$$

The characterization of a solution to (3.73) with the PMP is in general not possible with the technique in [81] for the following reason. We define $h(y) := |y - y_d|$. Then, because of the Lipschitz continuity of h, we have the existence of a function h' such that

$$\begin{aligned} h(y_1(x)) - h(y_2(x)) &= h(y_2(x) + \theta(y_1(x) - y_2(x)))\,\big|_{\theta=0}^{1} \\ &= \int_0^1 h'(y)\,\big|_{y=y_2(x)+\theta(y_1(x)-y_2(x))} d\theta\,(y_1(x) - y_2(x)) \end{aligned}$$

almost everywhere on Ω, see [6, Theorem 7.3], and we define

$$\tilde{h}(y_1, y_2) := \int_0^1 h'(y)\,\big|_{y=y_2(x)+\theta(y_1(x)-y_2(x))} d\theta.$$

In order to apply the technique used in [81, Proposition 4.4], we need the existence of a function $p^* : \Omega \to \mathbb{R}$, $z \mapsto p^*(z)$ such that

$$\lim_{k\to\infty} \|p_k - p^*\|_{L^\infty(\Omega)} = 0$$

where p_k is the solution to the intermediate adjoint equation (3.10). This is usually proved by subtracting the adjoint equation (3.5) from the intermediate adjoint equation (3.10) where it is necessary to define a pointwise limit

$$\lim_{k\to\infty} \tilde{h}(y_k, y^*)(x) = h'(y^*)(x)$$

almost everywhere on Ω with

$$\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(\Omega)} = 0$$

pointwise almost everywhere on Ω where y_k is the solution to the state equation for $u \leftarrow u_k$ defined in (3.8) and y^* is the solution to the state equation for $u \leftarrow u^*$. Since

$$\lim_{k\to\infty} \|y_k - y^*\|_{L^\infty(\Omega)} = 0,$$

we can start our considerations with the $\bar{k} \in \mathbb{N}$ such that if $y^*(x) > y_d(x)$, then $y_k(x) > y_d(x)$ and if $y^*(x) < y_d(x)$, then $y_k(x) < y_d(x)$ for almost all $x \in \Omega$ and all $k \geq \bar{k}$. In our case where $h(y) = |y - y_d|$, we choose pointwise

$$h'(y) := \begin{cases} 1 & \text{if } y \geq y_d \\ -1 & \text{else} \end{cases}$$

which gives

$$h'(y)\,\big|_{y=y^*+\theta(y_k-y^*)} = \begin{cases} 1 & \text{if } y_k > y_d \text{ and } y^* > y_d \\ 1 & \text{if } y_k = y_d \text{ and } y^* > y_d \\ 1 & \text{if } y_k > y_d \text{ and } y^* = y_d \\ 1 & \text{if } y_k = y_d \text{ and } y^* = y_d \\ -1 & \text{if } y_k < y_d \text{ and } y^* = y_d \\ -1 & \text{if } y_k = y_d \text{ and } y^* < y_d \\ -1 & \text{if } y_k < y_d \text{ and } y^* < y_d \end{cases}$$

using the estimation that

$$y^* + \theta (y_k - y^*) = (1-\theta) y^* + \theta y_k \begin{cases} > (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k > y_d \text{ and } y^* > y_d \\ > (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k = y_d \text{ and } y^* > y_d \\ > (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k > y_d \text{ and } y^* = y_d \\ = (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k = y_d \text{ and } y^* = y_d \\ < (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k < y_d \text{ and } y^* = y_d \\ < (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k = y_d \text{ and } y^* < y_d \\ < (1-\theta) y_d + \theta y_d = y_d & \text{if } y_k < y_d \text{ and } y^* < y_d \end{cases}.$$

Then we pointwise have that

$$\tilde{h} (y_k, y^*) = \begin{cases} 1 & \text{if } y_k > y_d \text{ and } y^* > y_d \\ 1 & \text{if } y_k = y_d \text{ and } y^* > y_d \\ 1 & \text{if } y_k > y_d \text{ and } y^* = y_d \\ 1 & \text{if } y_k = y_d \text{ and } y^* = y_d \\ -1 & \text{if } y_k < y_d \text{ and } y^* = y_d \\ -1 & \text{if } y_k = y_d \text{ and } y^* < y_d \\ -1 & \text{if } y_k < y_d \text{ and } y^* < y_d \end{cases}$$

such that it pointwise holds

$$|y_k - y_d| - |y^* - y_d| = \tilde{h} (y_k, y^*) (y_k - y^*)$$

since we pointwise have

$$|y_k - y_d| - |y^* - y_d| = \begin{cases} y_k - y^* & \text{if } y_k > y_d \text{ and } y^* > y_d \\ 0 - y^* + y_d = y_k - y^* & \text{if } y_k = y_d \text{ and } y^* > y_d \\ y_k - y_d - 0 = y_k - y^* & \text{if } y_k > y_d \text{ and } y^* = y_d \\ 0 - 0 = y_k - y_d & \text{if } y_k = y_d \text{ and } y^* = y_d \\ -(y_k - y_d) - 0 = -(y_k - y^*) & \text{if } y_k < y_d \text{ and } y^* = y_d \\ 0 - (-(y^* - y_d)) = -(y_k - y^*) & \text{if } y_k = y_d \text{ and } y^* < y_d \\ -(y_k - y_d) + (y^* - y_d) = -(y_k - y^*) & \text{if } y_k < y_d \text{ and } y^* < y_d \end{cases}.$$

In the case of $y(x) = y_d(x)$ on a set M of measure non-zero and $\lim_{k\to\infty} y_k(x) = y(x)$ for $x \in M$, then we do not know if the sign of $y_k(x) - y(x)$, $x \in M$, changes along the sequence $(y_k)_{k\in\mathbb{N}}$ and we cannot extract a subsequence with constant sign of $y_k(x) - y(x)$ as there are uncountable many elements in M. Consequently, the required limit $\lim_{k\to\infty} \tilde{h}(y_k, y)(x)$ does not exist in general and the proof used so far does not work.

Nevertheless, we apply our SQH method implemented in Algorithm 3.1 to $P.7$). For this purpose, we consider the following Hamiltonian

$$H(x, y, u, p) = |y - y_d| + \beta \ln(1 + |u|) + pu - p \max(0, y)$$

and the following adjoint equation

$$\int_\Omega \nabla p(x) \nabla v(x) + h_2(y(x)) p(x) v(x) \, dx = \int_\Omega h_1(y(x)) v(x) \, dx \tag{3.74}$$

where

$$h_1\left(y\left(x\right)\right) := \begin{cases} 1 & \text{if } y\left(x\right) \geq y_d\left(x\right) \\ -1 & \text{else} \end{cases}$$

is from the discussion above and

$$h_2\left(y\left(x\right)\right) := \begin{cases} 1 & \text{if } y\left(x\right) \geq 0 \\ 0 & \text{else} \end{cases}$$

can be obtained with a similar investigation such that

$$\max\left(0, y_k\right) - \max\left(0, y^*\right)$$

can be written in terms of

$$\max\left(0, y_k\right) - \max\left(0, y^*\right) = \tilde{h}_2\left(y_k, y^*\right)\left(y_k - y^*\right)$$

for a function $\tilde{h}_2$. Summarizing, we have that the formal definition of (3.74) is motivated by the case where our technique of the proof for characterizing a solution to an optimal control problem with the PMP, as discussed above in this subsection, can be applied due to the available differentiability in this case. We insert the functions h_1 and h_2 into the adjoint equation at these sites where in the smooth case the corresponding available derivatives would be.

The functions h_1 and h_2 are bounded and measurable, see [36, 2.1 Measurable Functions] and therefore elements of $L^\infty\left(\Omega\right)$. Thus (3.74) is uniquely solvable, see [45, Theorem 3 on page 301] with

$$\|p\|_{L^\infty(\Omega)} \leq c,$$

$c > 0$ as $h_2 \geq 0$, see the discussion starting on page 75.

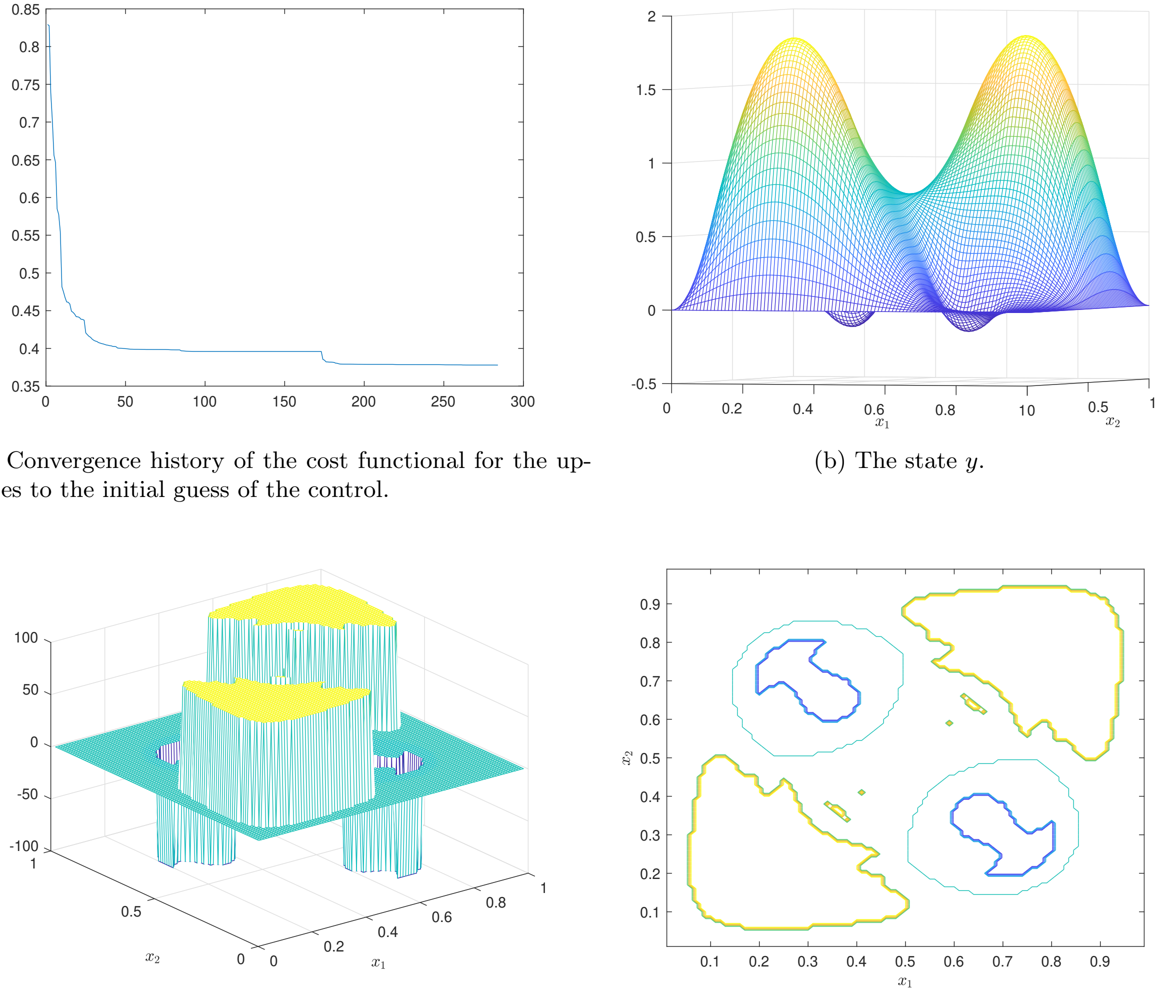
We consider $P.7$) with $\Omega := (0,1) \times (0,1)$, $y_d\left(x\right) := \sin\left(2\pi x_1\right)\sin\left(2\pi x_2\right) + \frac{8}{10}$, $K_U = [-100, 100]$ and $\beta = 9 \cdot 10^{-2}$. In our finite differences framework, the non-linear equation

$$-\Delta y + \max\left(0, y\right) = u$$

is solved by a Picard iteration until the L^2-norm of its residuum is less than 10^{-6}. The initial guess for the control equals zero and ϵ equals $\frac{1}{150}$. The parameters are given by $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$ and $\kappa = 10^{-8}$ where Ω is equidistantly discretized with $\triangle x = \frac{1}{100}$.

Results of this experiment are shown in Figure 3.11. Notice the fast reduction of the value of the cost functional in the first few iterations. This shows that the SQH method also works well in the case of a problem with an L^1 tracking term and non-smooth PDE constraints with respect to its capability of improving the initial guess of the control such that the cost functional takes smaller values.

Notice that although it is not proved that (3.7) is necessary for a solution to (3.73) and thus especially the theoretical framework of Section 3.1 does not have to hold in this case, the numerical optimality of the solution returned by the SQH is fulfilled even for small tolerances in more than 75% of the grid points, see Table 3.1.



(a) Convergence history of the cost functional for the updates to the initial guess of the control.

(b) The state y.

(c) The control u.

(d) The control u as a contour plot.

Figure 3.11: Results of the non-smooth optimal control problem $P.7$) for $\kappa = 10^{-8}$.

Chapter 4

An SQH framework for Fokker-Planck control problems

This chapter presents results related to [25, 20] and [23]. We discuss optimal control problems that are governed by the Fokker-Planck (FP) equation. In particular, we formulate an optimal control problem with a mean value objective for the state and a mean value cost term for the controls. The functions that determine the costs are assumed to be bounded from below and lower semi-continuous. Furthermore we characterize a solution to our optimal control problem with the Pontryagin maximum principle (PMP) and discuss the convergence analysis of the sequential quadratic Hamiltonian (SQH) method. We start our discussion by a preliminary investigation of the Fokker-Planck equation from a random walk (RW).

4.1 From Random walk to Fokker-Planck

The FP equation can be used to calculate the distribution of a stochastic process like a RW with infinitesimal small time steps. For this purpose, we start with introducing a RW in the following subsection and restrict ourselves to the one dimensional case to focus on the basic ideas. Moreover we consider a RW with jumps and different boundary conditions and investigate how these frameworks result in different FP models.

4.1.1 A random walk with reflecting and absorbing barriers

In this subsection, we introduce RW models in a bounded discrete space with different kind of barriers.

A RW consists of evolution paths given by a sequence of random steps at subsequent time instants on a grid [37]. These paths can be conveniently described as the evolution of the conditional probability distribution of the state X_{n+m} (position at the time-step $n+m$, $n, m \in \mathbb{N}$) of the RW and it is governed by the Chapman-Kolmogorov (CK) equation, see [37],

$$P\left(X_{m+n} = y | X_0 = x\right) = \sum_{z \in \mathbb{X}} P\left(X_{m+n} = y | X_m = z\right) P\left(X_m = z | X_0 = x\right) \tag{4.1}$$

where $\mathbb{X}$ is the state space and $P(\cdot|\cdot)$ denotes the conditional probability between two state configurations.

Specifically, consider a family of random variables $Z_i : \Omega \to \{j \triangle x |\ j \in \mathbb{Z}\}$, $i \in \mathbb{N}$ where Ω is the abstract space of elementary events and $\triangle x \in \mathbb{R}^+$ is the Euclidean distance between two nearest neighbour states. We define the random walk as follows

$$X_n \coloneqq x_0 + \sum_{i=1}^{n} Z_i \in \mathbb{R}$$

for all $n \in \mathbb{N}$ where $x_0 \in \mathbb{R}$ is the initial value of the random walk. Our state space is given by

$$\mathbb{X}_{x_0}^{\triangle x} := \{x_j \in \mathbb{R} |\ x_j = x_0 + j\triangle x, j \in \mathbb{Z}\} \cap [a,b]$$

where $a = x_0 + j_a \triangle x$ and $b = x_0 + j_b \triangle x$ with $j_a, j_b \in \mathbb{Z}$ and $j_a < 0 < j_b$ such that the initial point $x_0 \in \mathbb{X}_{x_0}^{\triangle x}$. We denote the total number of states with $N := \frac{b-a}{\triangle x} + 1$. The time space is given by

$$\mathbb{T}_{t_0}^{\triangle t} := \left\{t_n \in \mathbb{R}_0^+ |\ t_n = t_0 + n\triangle t, n \in \mathbb{N}_0\right\} \cap [0,T]$$

where the time horizon $T = t_0 + n_T \triangle t$, $n_T \in \mathbb{N}$ with the total number of time steps $N_t = \frac{T-t_0}{\triangle t}$. We denote with $p\left(x,t_n\right)$, resp. $q\left(x,t_n\right)$, the discrete transition probabilities to move from x to $x+\triangle x$, resp. x to $x - \triangle x$. The random variables Z_i include the RW of $\pm\triangle x$ steps and the jumps. The random variables Z_i are pair wise mutually independent, but spatially and time dependent. The process is called compound Poisson process as the random variables Z_i are composed of a RW characterized by p, q and a RW with jumps. We have a jump rate $\lambda \geq 0$ such that $\lambda\triangle t$ represents the probability that a jump occurs in the time interval $\triangle t$. To weight the jump length, we utilize a function $g : \mathbb{R} \to \mathbb{R}$ with compact support, that means that there is $x_c \in \mathbb{R}_0^+$ such that $g\left(x\right) = 0$ for all $|x| > x_c$. Additionally, we require that $x_c \leq b - a$ in order to avoid multi-reflections within one jump. Finally, the conditional probability related to the Poisson jump process is given by

$$\begin{aligned}
P\left(Z_n = j\triangle x | X_{n-1} = x\right) &= \lambda\triangle t\triangle x g\left(j\triangle x\right) \text{ for all } j \geq 2 && (4.2)\\
P\left(Z_n = \triangle x | X_{n-1} = x\right) &= p\left(x,t_{n-1}\right) + \lambda\triangle t\triangle x g\left(\triangle x\right) && (4.3)\\
P\left(Z_n = 0 | X_{n-1} = x\right) &= 1 - p\left(x,t_{n-1}\right) - q\left(x,t_{n-1}\right) - \lambda\triangle t\triangle x \sum_{\substack{j=-\infty\\ j\neq 0}}^{\infty} g\left(j\triangle x\right) && (4.4)\\
P\left(Z_i = -\triangle x | X_{n-1} = x\right) &= q\left(x,t_{n-1}\right) + \lambda\triangle t\triangle x g\left(-\triangle x\right) && (4.5)\\
P\left(Z_i = -j\triangle x | X_{n-1} = x\right) &= \lambda\triangle t\triangle x g\left(-j\triangle x\right) \text{ for all j} \leq -2 && (4.6)
\end{aligned}$$

where $t_n := n\triangle t + t_0$ and $\triangle t, t_0 \in \mathbb{R}_0^+$, $n \in \mathbb{N}_0$ with the period $\triangle t$ between two steps.

In order to have a well defined transition probability, the following constraints have to be satisfied

$$0 \leq p\left(x,t_n\right) \leq 1,\ 0 \leq q\left(x,t_n\right) \leq 1,\ 0 \leq \lambda\triangle t\triangle x \sum_{j=-\infty}^{\infty} g\left(j\triangle x\right) \leq 1$$

and

$$0 \leq p\left(x,t_n\right) + q\left(x,t_n\right) + \lambda\triangle t\triangle x \sum_{j=-\infty}^{\infty} g\left(j\triangle x\right) \leq 1.$$

From now on whenever clear from the context, we write t instead of t_n to simplify the notation.

Next, we turn our attention to the modeling of different barriers at the boundaries a and b.

Our random walk with absorbing barriers is defined as follows

$$X_n := \begin{cases} X_{n-1} + Z_n & \text{if } a < X_{n-1} < b \\ X_{n-1} & \text{if } X_{n-1} \leq a \\ X_{n-1} & \text{if } X_{n-1} \geq b \end{cases} \tag{4.7}$$

for all $n \in \mathbb{N}$. Roughly speaking, the meaning of this absorbing barrier is that if the process steps out of the domain, then it can never return into the domain.

Our random walk with reflecting barriers is defined as follows

$$X_n := \begin{cases} a - Z_n - \left(X_{n-1} - a\right) & \text{if } \left(X_{n-1} + Z_n < a\right) \\ b - Z_n - \left(X_{n-1} - b\right) & \text{if } \left(X_{n-1} + Z_n > b\right) \\ X_{n-1} + Z_n & \text{else} \end{cases} \tag{4.8}$$

for all $n \in \mathbb{N}$. This model corresponds to a wall where the process is being reflected such that the total length of the jump is preserved.

Now, we derive the CK equation for our random walk with absorbing/reflecting barriers. We denote with $f : \mathbb{X}_{x_0}^{\triangle x} \times \mathbb{T}_{t_0}^{\triangle t} \to \mathbb{R}_0^+$ a discrete function representing the discrete occupation probability 'density' of the random walk process at the location x at time t. That is, $\triangle x f(x,t)$ is the probability that the process is located in x at the time t. For convenience, we extend f to zero for all x beyond an absorbing barrier.

Now, we discuss the evolution model for f. For this purpose, consider all contribution to $f(x, t + \triangle t)$ from the states at time t.

We start considering a RW model with absorbing barriers. In this case, the evolution of f, similar to (4.1), is as follows

$$\begin{aligned} f(x, t+\triangle t) &= p(x-\triangle x, t) f(x-\triangle x, t) + q(x+\triangle x, t) f(x+\triangle x, t) \\ &\quad + \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j \neq 0}}^{\infty} \left(g(j\triangle x) f(x - j\triangle x, t)\right) \\ &\quad + \left(1 - p(x,t) - q(x,t) - \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g(j\triangle x)\right) f(x,t) \end{aligned} \tag{4.9}$$

for $a < x < b$.

For simplicity, we also refer to (4.9) as the CK equation.

In the following, we discuss the CK equation for f in the case of a combination of absorbing and reflecting barriers. For this purpose, we introduce two continuous weight functions $\kappa_a, \kappa_b : \mathbb{R}_0^+ \to [0,1]$ that are able to modulate the type of barrier from absorbing $\kappa_a \equiv \kappa_b \equiv 0$ to $\kappa_a \equiv \kappa_b \equiv 1$. In all other cases, these functions model a combination of absorbing and reflecting barriers.

With this setting, additional terms appear in the CK equation (4.9) that depend on the position x where the process takes place. To obtain the corresponding Chapman-Kolmogorov equation for f, we have to add terms to the right hand-side of (4.9) which start at $\tilde{x}$, are reflected at a reflecting barrier and thus contribute to x at $t + \triangle t$. The traveled distance of a process reflected at a is given by $j\triangle x = -(x-a) - (\tilde{x} - a)$, equivalently, $\tilde{x} = 2a - x - j\triangle x$ where $j < 0$. If a process is reflected at b, we have $j\triangle x = b - x + b - \tilde{x}$, equivalently, $\tilde{x} = 2b - x - j\triangle x$ where $j > 0$. After this preparation, we illustrate the evolution equation for f depending on x.

We distinguish five cases: $x = a$, $x = a + \triangle x$, $a + \triangle x < x < b - \triangle x$, $x = b - \triangle x$, $x = b$.

Consider the case $x = a$. We have

$$\begin{aligned} f(a, t+\triangle t) &= \kappa_a(t) \left(q(a+\triangle x, t) f(a+\triangle x, t) + \lambda \triangle t \triangle x \sum_{j=-\infty}^{-1} \left(g(j\triangle x) f(a - j\triangle x, t)\right) \right) \\ &\quad + \left(1 - \kappa_a(t) \left(p(a,t) + q(a,t) + \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g(j\triangle x) \right)\right) f(a,t) \\ &\quad + \kappa_a(t) \kappa_b(t) \lambda \triangle t \triangle x \sum_{j=1}^{\infty} \left(g(j\triangle x) f(2b - a - j\triangle x, t)\right). \end{aligned} \tag{4.10}$$

Next, in the case $x = a + \triangle x$, we have

$$\begin{aligned} f\left(a+\triangle x, t+\triangle t\right) &= \left(\kappa_a\left(t\right) q\left(a,t\right) + p\left(a,t\right)\right) f\left(a,t\right) + q\left(a+2\triangle x,t\right) f\left(a+2\triangle x,t\right) \\ &\quad + \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} \left(g\left(j\triangle x\right) f\left(\left(a+\triangle x\right) - j\triangle x, t\right)\right) \\ &\quad + \kappa_a\left(t\right) \lambda \triangle t \triangle x \sum_{j=-\infty}^{-1} \left(g\left(j\triangle x\right) f\left(a - \triangle x - j\triangle x, t\right)\right) \\ &\quad + \kappa_b\left(t\right) \lambda \triangle t \triangle x \sum_{j=1}^{\infty} \left(g\left(j\triangle x\right) f\left(2b - a - \triangle x - j\triangle x, t\right)\right) \\ &\quad + \left(1 - p\left(a+\triangle x,t\right) - q\left(a+\triangle x,t\right) - \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)\right) f\left(a+\triangle x,t\right). \end{aligned} \tag{4.11}$$

In the case $a + \triangle x < x < b - \triangle x$, we have

$$\begin{aligned} f\left(x, t+\triangle t\right) &= p\left(x-\triangle x,t\right) f\left(x-\triangle x,t\right) + q\left(x+\triangle x,t\right) f\left(x+\triangle x,t\right) \\ &\quad + \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} \left(g\left(j\triangle x\right) f\left(x - j\triangle x, t\right)\right) \\ &\quad + \kappa_a\left(t\right) \lambda \triangle t \triangle x \sum_{j=-\infty}^{-1} \left(g\left(j\triangle x\right) f\left(2a - x - j\triangle x, t\right)\right) \\ &\quad + \kappa_b\left(t\right) \lambda \triangle t \triangle x \sum_{j=1}^{\infty} \left(g\left(j\triangle x\right) f\left(2b - x - j\triangle x, t\right)\right) \\ &\quad + \left(1 - p\left(x,t\right) - q\left(x,t\right) - \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)\right) f\left(x,t\right). \end{aligned} \tag{4.12}$$

In the case $x = b - \triangle x$, we have

$$\begin{aligned} f\left(b-\triangle x, t+\triangle t\right) &= p\left(b-2\triangle x,t\right) f\left(b-2\triangle x,t\right) + \left(\kappa_b\left(t\right) p\left(b,t\right) + q\left(b,t\right)\right) f\left(b,t\right) \\ &\quad + \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} \left(g\left(j\triangle x\right) f\left(x - j\triangle x, t\right)\right) \\ &\quad + \kappa_a\left(t\right) \lambda \triangle t \triangle x \sum_{j=-\infty}^{-1} \left(g\left(j\triangle x\right) f\left(2a - b + \triangle x - j\triangle x, t\right)\right) \\ &\quad + \kappa_b\left(t\right) \lambda \triangle t \triangle x \sum_{j=1}^{\infty} \left(g\left(j\triangle x\right) f\left(b + \triangle x - j\triangle x, t\right)\right) \\ &\quad + \left(1 - p\left(b-\triangle x,t\right) - q\left(b-\triangle x,t\right) - \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)\right) f\left(b-\triangle x,t\right). \end{aligned} \tag{4.13}$$

In the last case $x = b$, we obtain

$$\begin{aligned} f\left(b, t+\triangle t\right) &= \kappa_b\left(t\right)\left(p\left(b-\triangle x, t\right) f\left(b-\triangle x, t\right) + \lambda \triangle t \triangle x \sum_{j=1}^{\infty}\left(g\left(j\triangle x\right) f\left(b-j\triangle x, t\right)\right)\right) \\ &\quad + \left(1-\kappa_b\left(t\right)\left(p\left(b,t\right)+q\left(b,t\right)+\lambda\triangle t\triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)\right)\right) f\left(b,t\right) \qquad (4.14) \\ &\quad + \kappa_b\left(t\right)\kappa_a\left(t\right)\lambda\triangle t\triangle x \sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right) f\left(2a-b-j\triangle x, t\right)\right). \end{aligned}$$

Remark 31. If both barriers are reflecting and $\triangle x \sum_{j=0}^{N} f\left(a+j\triangle x, 0\right) = 1$, then $\triangle x \sum_{j=0}^{N} f\left(a+j\triangle x, t\right) = 1$ for all $t \in \mathbb{T}_{t_0}^{\triangle t}$. That is the CK evolution is conservative. In the presence of an absorbing barrier, we have $\triangle x \sum_{j=0}^{N} f\left(a+j\triangle x, t\right) \leq 1$, since there is the possibility that the random walk process leaves the domain.

Our next step is to formulate (4.10) to (4.14) using a matrix representation in the sense that

$$f\left(\cdot, t+\triangle t\right) = \mathcal{P} f\left(\cdot, t\right)$$

for all $x \in \mathbb{X}_{x_0}^{\triangle x}$ and $\mathcal{P} \in \mathbb{R}^{N\times N}$.

We define $r\left(x,t\right) \coloneqq 1 - p\left(x,t\right) - q\left(x,t\right) - \lambda\triangle t\triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)$ and, for ease of notation, we omit the time dependence of the transition matrix $\mathcal{P} = \mathcal{P}\left(t\right)$. The first row of $\mathcal{P}$ is given by

$$\begin{pmatrix} 1-\kappa_a\left(t\right)\left(p\left(a,t\right)+q\left(a,t\right)+\lambda\triangle t\triangle x\sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)\right) \\ \kappa_a\left(t\right)\left(q\left(a+\triangle x,t\right)+\lambda\triangle t\triangle x\left(g\left(-\triangle x\right)\right)\right) \\ \vdots \\ \lambda\triangle t\triangle x\kappa_a\left(t\right)\left(\kappa_b\left(t\right)g\left(b-a\right)+g\left(-\left(b-a\right)\right)\right) \end{pmatrix}^{\mathrm{T}}$$

where $\left(\cdot\right)^{T}$ means transpose. The second row is given by

$$\begin{pmatrix} \kappa_a\left(t\right)q\left(a,t\right)+p\left(a,t\right)+\lambda\triangle t\triangle x\left(g\left(\triangle x\right)+\kappa_a\left(t\right)g\left(-\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-\triangle x\right)\right) \\ r\left(a+\triangle x,t\right)+\lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-2\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-2\triangle x\right)\right) \\ q\left(a+2\triangle x,t\right)+\lambda\triangle t\triangle x\left(g\left(-\triangle x\right)+\kappa_a\left(t\right)g\left(-3\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-3\triangle x\right)\right) \\ \vdots \\ \lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-\left(b-a\right)-\triangle x\right)+\kappa_b\left(t\right)g\left(\left(b-a\right)-\triangle x\right)+g\left(-\left(b-a\right)+\triangle x\right)\right) \end{pmatrix}^{\mathrm{T}}.$$

For the j-th row, $j \in \{3, ..., N-2\}$, and $x = a + j\triangle x$, we have the following

$$\begin{pmatrix} \lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-\left(j-1\right)\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-\left(j-1\right)\triangle x\right)+g\left(j\triangle x\right)\right) \\ \lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-\left(j-1\right)\triangle x-\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-\left(j-1\right)\triangle x-\triangle x\right)+g\left(j\triangle x-\triangle x\right)\right) \\ \lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-\left(j-1\right)\triangle x-2\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-\left(j-1\right)\triangle x-2\triangle x\right)+g\left(j\triangle x-2\triangle x\right)\right) \\ \vdots \\ p\left(x-\triangle x,t\right)+\lambda\triangle t\triangle x\left(g\left(\triangle x\right)+\kappa_a\left(t\right)g\left(-2\left(j-1\right)\triangle x+\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-2\left(j-1\right)\triangle x+\triangle x\right)\right) \\ r\left(x,t\right)+\lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-2\left(j-1\right)\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-2\left(j-1\right)\triangle x\right)\right) \\ q\left(x+\triangle x,t\right)+\lambda\triangle t\triangle x\left(g\left(-\triangle x\right)+\kappa_a\left(t\right)g\left(-2\left(j-1\right)\triangle x-\triangle x\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-2\left(j-1\right)\triangle x-\triangle x\right)\right) \\ \vdots \\ \lambda\triangle t\triangle x\left(\kappa_a\left(t\right)g\left(-\left(j-1\right)\triangle x-\left(b-a\right)\right)+\kappa_b\left(t\right)g\left(2\left(b-a\right)-\left(j-1\right)\triangle x-\left(b-a\right)\right)+g\left(\left(b-a\right)-j\triangle x\right)\right) \end{pmatrix}^{\mathrm{T}}.$$

The second to last row is given by

$$\begin{pmatrix} \lambda \triangle t \triangle x \left(\kappa_a\left(t\right) g\left(-\left(b-a\right)+\triangle x\right)+\kappa_b\left(t\right) g\left(\left(b-a\right)+\triangle x\right)+g\left(\left(b-a\right)-\triangle x\right)\right) \\ \vdots \\ p\left(b-2\triangle x, t\right)+\lambda \triangle t \triangle x \left(g\left(\triangle x\right)+\kappa_a\left(t\right) g\left(-2\left(b-a\right)+3\triangle x\right)+\kappa_b\left(t\right) g\left(3\triangle x\right)\right) \\ r\left(b-\triangle x, t\right)+\lambda \triangle t \triangle x \left(\kappa_a\left(t\right) g\left(-2\left(b-a\right)+2\triangle x\right)+\kappa_b\left(t\right) g\left(2\triangle x\right)\right) \\ q\left(b,t\right)+\kappa_b\left(t\right) p\left(b,t\right)+\lambda \triangle t \triangle x \left(g\left(-\triangle x\right)+\kappa_a\left(t\right) g\left(-2\left(b-a\right)+\triangle x\right)+\kappa_b\left(t\right) g\left(\triangle x\right)\right) \end{pmatrix}^{\mathrm{T}}$$

and the last row reads as follows

$$\begin{pmatrix} \lambda \triangle t \triangle x \kappa_b\left(t\right)\left(\kappa_a\left(t\right) g\left(-\left(b-a\right)\right)+g\left(b-a\right)\right) \\ \vdots \\ \kappa_b\left(t\right)\left(p\left(b-\triangle x, t\right)+\lambda \triangle t \triangle x g\left(\triangle x\right)\right) \\ 1-\kappa_b\left(t\right)\left(p\left(b,t\right)+q\left(b,t\right)+\lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)\right) \end{pmatrix}^{\mathrm{T}}$$

where $b-a=\triangle x\left(N-1\right)$.

4.1.2 Optimal control of a random walk with jumps

In this subsection, we include controls into the transition matrix $\mathcal{P}=\mathcal{P}\left(u\right)$ in order to first extend and second validate our modeling framework with numerical experiments later. In the following, we define a RW optimal control problem. We prove existence of an optimal control and derive the optimality conditions that characterize it.

For ease of notation, we refer to the case with absorbing boundaries, that is, $\kappa_a=0$ and $\kappa_b=0$. However, notice that our discussion refers to a general transition matrix.

In the case of absorbing barriers, we have $f\left(a,t\right)=f\left(b,t\right)=0$. Therefore the number of states is given by $N=\frac{b-a}{\triangle x}-1$, and $t\geq t_0$.

We insert the control mechanism in $p\left(x,t\right)$ and $q\left(x,t\right)$, thus allowing a control mechanism in the drift and in the diffusion. This fact will become evident when considering of the FP equation in Subsection 4.1.3.

Our control function u enters in the transition probabilities as follows

$$p\left(x,t\right):=\frac{1}{2}\left(s\left(x,t\right)+\frac{\triangle x}{D}u\left(x,t\right)\right) \tag{4.15}$$

and

$$q\left(x,t\right):=\frac{1}{2}\left(s\left(x,t\right)-\frac{\triangle x}{D}u\left(x,t\right)\right) \tag{4.16}$$

with a constant

$$D:=\frac{\left(\triangle x\right)^2}{\triangle t}\in\mathbb{R}^+ \tag{4.17}$$

with given $s:\mathbb{X}_{x_0}^{\triangle x}\times\mathbb{T}_{t_0}^{\triangle t}\to\mathbb{R}^+$ and $u:\mathbb{X}_{x_0}^{\triangle x}\times\mathbb{T}_{t_0}^{\triangle t}\to\mathbb{R}$.

Since we wish to provide a detailed discussion of all quantities that enter our computational framework and, in particular, present these quantities in vector notation, we focus on the case where the control function depends only on time, i.e. $u=u\left(t\right)$ and s is constant. The same discussion can be repeated in the general case $u=u\left(x,t\right)$ and $s=s\left(x,t\right)$ with additional notational effort.

Now, we insert (4.15) and (4.16) into (4.9) and obtain the following

$$f(x,t+\triangle t) = (s+\triangle x u(t)) f(x-\triangle x,t) + (s-\triangle x u(t)) f(x+\triangle x,t) \\ + \lambda\triangle t\triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} \left(g(j\triangle x) f(x-j\triangle x,t)\right) + \left(1-s-\lambda\triangle t\triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g(j\triangle x)\right) f(x,t).$$

From $0 \le p(x,t) + q(x,t) + \lambda\triangle t\triangle x \sum_{j=-\infty}^{\infty} g(j\triangle x) \le 1$, we have $0 \le s \le 1 - \lambda\triangle t\triangle x \sum_{j=-\infty}^{\infty} g(j\triangle x)$. From $0 \le p(x,t) \le 1$, we have $-\frac{D}{\triangle x}s \le u(t) \le \frac{D}{\triangle x}(2-s)$ and from $0 \le q(x,t) \le 1$, we have $-\frac{D}{\triangle x}(2-s) \le u(t) \le \frac{D}{\triangle x}s$. We conclude that the admissible set of controls is given by

$$U_{ad}(s) := \{u(t) \in \mathbb{R} |\ \underline{u}_s \le u(t) \le \overline{u}_s\}$$

where

$$\underline{u}_s := -\frac{D}{\triangle x}s \text{ and } \overline{u}_s := \frac{D}{\triangle x}s. \tag{4.18}$$

Next, we define the two objective functionals $J_{\mathrm{t}}(f,u)$ and $J_{\mathrm{c}}(f,u)$ that model the purpose of the control and its cost. Let $\alpha, \beta, \gamma \in \mathbb{R}_0^+$, we have

$$J_{\mathrm{t}}(f,u) := \frac{\alpha}{2}\triangle x\triangle t \left(\sum_{n=1}^{N_t-1} \sum_{x=a+\triangle x}^{b-\triangle x} \left(f(x,t_n) - f_d(x,t_n)\right)^2\right) + \frac{\beta}{2}\triangle x \left(\sum_{x=a+\triangle x}^{b-\triangle x} \left(f(x,T) - f_d(x,T)\right)^2\right) \\ + \frac{\gamma}{2}\triangle x\triangle t \left(\sum_{n=0}^{N_t-1} \sum_{x=a+\triangle x}^{b-\triangle x} f(x,t_n) u^2(t_n)\right) \tag{4.19}$$

where $f_d : \mathbb{X}_{x_0}^{\triangle x} \times \mathbb{T}_{t_0}^{\triangle t} \to \mathbb{R}_0^+$ with $\triangle x \sum_{x=a+\triangle x}^{b-\triangle x} f_d(x,t) \le 1$ represents a desired trajectory and a target at final time. The purpose of the first term in the functional models the requirement that the evolving discrete occupation probability density f follows as close as possible the desired trajectory given by f_d. The second term requires that f reaches a target configuration as close as possible to $f_d(\cdot,T)$. The last term models the expectation costs of the control.

Further, we consider the following functional J_{c}, which has the structure of a statistical expectation functional. We have

$$J_{\mathrm{c}}(f,u) := \alpha\triangle x\triangle t \left(\sum_{n=1}^{N_t-1} \sum_{x=a+\triangle x}^{b-\triangle x} f(x,t_n) c(x,t_n)\right) + \beta\triangle x \left(\sum_{x=a+\triangle x}^{b-\triangle x} f(x,T)\Psi(x)\right) \\ + \frac{\gamma}{2}\triangle x\triangle t \left(\sum_{n=0}^{N_t-1} \sum_{x=a+\triangle x}^{b-\triangle x} f(x,t_n) u^2(t_n)\right) \tag{4.20}$$

where $c : \mathbb{X}_{x_0}^{\triangle x} \times \mathbb{T}_{t_0}^{\triangle t} \to \mathbb{R}$ and $\Psi : \mathbb{X}_{x_0}^{\triangle x} \to \mathbb{R}$ are given discrete functions. The first two terms model the expectation of a running gain function c and of a pay-off Ψ, while the last term represents the expectation of the cost of the control.

Our optimal control problems are defined as follows

$$\min_{f,u} J_{\Phi}(f,u) \text{ s.t. } f,u \text{ fulfill (4.9) and } u \in U_{ad}(s) \tag{4.21}$$

for $\Phi \in \{\mathrm{t},\mathrm{c}\}$.

In order to formulate these problems with the matrix-vector setting above, we explicitly refer to our discrete functions as vectors as follows $f := \begin{pmatrix} f^1 \\ \vdots \\ f^{N_t} \end{pmatrix} \in \mathbb{R}^{NN_t}$, where $f^n := \begin{pmatrix} f\left(a+\triangle x, t_n\right) \\ \vdots \\ f\left(b-\triangle x, t_n\right) \end{pmatrix} \in \mathbb{R}^N$ for all $n \in \{1, ..., N_t\}$ and $u := \begin{pmatrix} u^0 \\ \vdots \\ u^{N_t-1} \end{pmatrix} \in \mathbb{R}^{N_t}$, where $u^n = u\left(t_n\right)$ for all $n \in \{0, ..., N_t-1\}$. According to (4.9) and for a given $f^0 \geq 0$, where $\geq$ works componentwise, we have

$$\begin{pmatrix} f^1 \\ \vdots \\ f^{N_t} \end{pmatrix} = \begin{pmatrix} \mathcal{P}\left(u^0\right) & & \\ & \ddots & \\ & & \mathcal{P}\left(u^{N_t-1}\right) \end{pmatrix} \begin{pmatrix} f^0 \\ \vdots \\ f^{N_t-1} \end{pmatrix}$$

where the transition matrix $\mathcal{P}\left(u^n\right) \in \mathbb{R}^{N\times N}$, for (4.10) to (4.14) including (4.15), (4.16) and $\kappa_a = \kappa_b = 0$, is given as follows

$$\begin{aligned} &\mathcal{P}\left(u^n\right) \\ &= \begin{pmatrix} r & \frac{1}{2}s - \frac{\triangle x}{2D}u\left(t_n\right) + \tilde{g}\left(-\triangle x\right) & & \cdots \\ \frac{1}{2}s + \frac{\triangle x}{2D}u\left(t_n\right) + \tilde{g}\left(\triangle x\right) & \ddots & \ddots & \ddots \\ & \ddots & \ddots & s + \tilde{g}\left(-\triangle x\right) \\ \vdots & \ddots & \frac{1}{2}s + \frac{\triangle x}{2D}u\left(t_n\right) + \tilde{g}\left(\triangle x\right) & r \end{pmatrix} \end{aligned}$$

with $r = 1 - s - \lambda \triangle t \triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g\left(j\triangle x\right)$ and $\tilde{g}\left(j\triangle x\right) = \lambda \triangle t \triangle x g\left(j\triangle x\right)$ for all $j \in \mathbb{Z}$.

Furthermore, we define

$$F\left(f,u\right) := \begin{pmatrix} f^1 \\ \vdots \\ f^{N_t} \end{pmatrix} - \begin{pmatrix} \mathcal{P}\left(u^0\right) & & \\ & \ddots & \\ & & \mathcal{P}\left(u^{N_t-1}\right) \end{pmatrix} \begin{pmatrix} f^0 \\ \vdots \\ f^{N_t-1} \end{pmatrix} \in \mathbb{R}^{N_t N}. \tag{4.22}$$

With (4.22), the optimal control problem (4.21) is formulated as follows

$$\min_{f,u} J_\Phi\left(f,u\right) \text{ s.t. } F\left(f,u\right) = 0 \text{ and } u \in U_{ad}\left(s\right). \tag{4.23}$$

Although for a given u and f^0 the problem $F\left(f,u\right) = 0$ is readily solvable by matrix multiplication of $f^{n+1} = \mathcal{P}\left(u^n\right) f^n$ for all $n \in \{0, ..., N_t-1\}$, see (4.22), we prefer to discuss it in the framework of the implicit function theorem [4], because this approach allows to prove smoothness of the control-to-state map and thus existence of an optimal control. Furthermore in this way, we obtain the gradient of the reduced functional.

Next, we prove existence of a solution $f = f\left(u\right)$ to $F\left(f,u\right) = 0$ for a given u and the Fréchet differentiability of $f\left(u\right)$ with respect to u, see [4, 78] for details.

The Fréchet derivative of $F\left(f,u\right)$ with respect to f is given by

$$D_f F\left(f,u\right) = \begin{pmatrix} \mathbb{1} & & & \\ -\mathcal{P}\left(u^1\right) & \mathbb{1} & & \\ & -\mathcal{P}\left(u^2\right) & \ddots & \\ & & -\mathcal{P}\left(u^{N_t-1}\right) & \mathbb{1} \end{pmatrix} \in \mathbb{R}^{N_t N \times N_t N}, \tag{4.24}$$

which is a lower triangular matrix where $\mathbb{1} \in \mathbb{R}^{N\times N}$ is the identity. The determinant of $D_f F\left(f,u\right)$ is one and thus invertible. Applying the implicit function theorem to $F\left(f,u\right) = 0$, we obtain that the map

$u \mapsto f(u)$ is infinitely differentiable in a certain neighborhood of each fixed u and the Fréchet derivative $D_u f(u)$ of $f(u)$ with respect to u is given by

$$D_u f(u) = -\left(D_f F(f,u)\right)^{-1} D_u F(f,u) \in \mathbb{R}^{N_t N \times N_t} \tag{4.25}$$

where $D_u F(f,u) \in \mathbb{R}^{N_t N \times N_t}$ denotes the Fréchet derivative of $F(f,u)$ with respect to u, which is given by

$$D_u F(f,u) = -\begin{pmatrix} D_{u^0} & & & \\ & D_{u^1} & & \\ & & \ddots & \\ & & & D_{u^{N_t-1}} \end{pmatrix} \in \mathbb{R}^{N_t N \times N_t} \tag{4.26}$$

where

$$D_{u^n} := \frac{\triangle x}{2D} \begin{pmatrix} -f(a+2\triangle x, t_n) \\ f(a+\triangle x, t_n) - f(a+3\triangle x, t_n) \\ \vdots \\ f(b-3\triangle x, t_n) - f(b-\triangle x, t_n) \\ f(b-2\triangle x, t_n) \end{pmatrix} \in \mathbb{R}^{N\times 1}$$

for all $n \geq 0$.

Using the mapping $u \mapsto f(u)$, we define the reduced objective functional

$$\hat{J}(u) := J(f(u), u) \tag{4.27}$$

for any differentiable functional $J(f,u)$. The vector for the adjoint equation (4.31) is denoted with $\zeta := \begin{pmatrix} \zeta^1 \\ \vdots \\ \zeta^{N_t} \end{pmatrix} \in \mathbb{R}^{N_t N}$, $\zeta^n \in \mathbb{R}^N$ for all $n \in \{1, ..., N_t\}$. We prove the following theorem.

Theorem 32. *Let $J(f,u)$ be differentiable with respect to $f \in \mathbb{R}^{NN_t}$ and $u \in \mathbb{R}^{N_t}$. Then, the optimal control problem*

$$\begin{cases} \min_u \hat{J}(u) \\ u \in U_{ad}(s) \end{cases} \tag{4.28}$$

with $\hat{J}(u)$ defined in (4.27) has a solution where

$$U_{ad}(s) := \left\{ u \in \mathbb{R}^{N_t} \middle| \, \underline{u}_s \leq u^n \leq \overline{u}_s, \; n \in \{0, ...N_t - 1\} \right\} \tag{4.29}$$

with $\underline{u}_s$ and $\overline{u}_s$ defined in (4.18) and $\leq, \geq$ are meant componentwise. The reduced gradient is given by

$$\nabla_u \hat{J}(u) = \left(D_u F(f,u)\right)^{\mathrm{T}} \zeta + \nabla_u J(f,u) \in \mathbb{R}^{N_t} \tag{4.30}$$

where ζ solves

$$\left(D_f F(f,u)\right)^{\mathrm{T}} \zeta = -\nabla_f J(f,u) \in \mathbb{R}^{N_t N}. \tag{4.31}$$

An optimal solution $u^{\mathrm{opt}} \in \mathbb{R}^{N_t}$ of (4.28) is characterized by

$$\left(\nabla_u \hat{J}\left(u^{\mathrm{opt}}\right)\right)^{\mathrm{T}} \left(u - u^{\mathrm{opt}}\right) \geq 0 \tag{4.32}$$

for all $u \in U_{ad}(s)$ where $U_{ad}(s)$ is defined in (4.29).

Proof. The set $U_{ad}(s)$ is compact because it is closed and bounded [78] and the function is differentiable because its a composition of the differentiable function $f(u)$, see the discussion following (4.23) and the differentiable target functional, see (4.27). Therefore $\hat{J}$ is continuous and thus function $\hat{J}(u)$ takes its minimum on $U_{ad}(s)$, see e.g. [78].

The Fréchet derivative $D_u\hat{J}(u)$ is given by

$$\begin{aligned} D_u\hat{J}(u) &= D_fJ(f,u)\,D_uf + D_uJ(f,u) \\ &= D_fJ(f,u)\left(-\left(D_fF(f,u)\right)^{-1}D_uF(f,u)\right) + D_uJ(f,u) \end{aligned} \tag{4.33}$$

where we used (4.25). We define $\zeta^{\mathrm{T}} := -D_fJ(f,u)\left(D_fF(f,u)\right)^{-1}$, equivalently, we have $\zeta^{\mathrm{T}}D_fF(f,u) = -D_fJ(f,u)$. We obtain the reduced gradient $\nabla_u\hat{J}(u)$ from (4.33). We obtain

$$\nabla_u\hat{J}(u) = \left(D_uF(f,u)\right)^{\mathrm{T}}\zeta + \nabla_uJ(f,u)$$

where ζ solves

$$\left(D_fF(f,u)\right)^{\mathrm{T}}\zeta = -\nabla_fJ(f,u)\,.$$

The set $U_{ad}(s)$ is convex. By [68, page 178] and the differentiability of $\hat{J}(u)$, we obtain the necessary optimality condition (4.32). □

Remark 33. Equation (4.31) is equivalent to

$$\begin{pmatrix} \zeta^1 - \left(\mathcal{P}\left(u^1\right)\right)^{\mathrm{T}}\zeta^2 \\ \vdots \\ \zeta^{N_t-1} - \left(\mathcal{P}\left(u^{T-1}\right)\right)^{\mathrm{T}}\zeta^{N_t} \\ \zeta^{N_t} \end{pmatrix} = -\nabla_fJ(f,u)$$

which is a backward equation for ζ with the terminal value $\zeta^{N_t} = -\nabla_{f^{N_t}}J(f,u) \in \mathbb{R}^N$. The solution to the forward CK equation $F(f,u) = 0$ for all time steps is given by

$$\begin{pmatrix} f^1 \\ \vdots \\ f^{N_t} \end{pmatrix} = \begin{pmatrix} \mathcal{P}\left(u^0\right)f^0 \\ \vdots \\ \mathcal{P}\left(u^{N_t-1}\right)f^{N_t-1} \end{pmatrix}$$

for any given $f^0 \geq 0$. Analogously, we have

$$\nabla_u\hat{J}(u) = \nabla_uJ(f,u) - \begin{pmatrix} \left(D_{u^0}\right)^{\mathrm{T}}\zeta^1 \\ \vdots \\ \left(D_{u^{N_t-1}}\right)^{\mathrm{T}}\zeta^{N_t} \end{pmatrix}$$

according to (4.30).

Corollary 34. *The optimal control problem* (4.28) *with* $J = J_{\mathrm{t}}(f,u)$ *or* $J = J_{\mathrm{c}}(f,u)$ *has a solution.*

Proof. The functionals $J_{\mathrm{t}}(f,u)$ and $J_{\mathrm{c}}(f,u)$ are infinitely differentiable with respect to f and u. Thus, we apply Theorem 32 and prove the claim. □

Further calculation proves the following corollaries. To use a compact notation, we denote with $e_N \in \mathbb{R}^N$ a unit vector where each entry is one.

Corollary 35. *The gradient of* $J_{\mathrm{t}}(f,u)$ *given in* (4.19) *with respect to* f *and* u *is given by*

$$\begin{aligned}&\nabla_f J_{\mathrm{t}}(f,u)\\&=\alpha\triangle x\triangle t\begin{pmatrix} f^1-f_d^1\\ \vdots\\ f^{N_t-1}-f_d^{N_t-1}\\ 0\\ \vdots\\ 0\end{pmatrix}+\beta\triangle x\begin{pmatrix}0\\ \vdots\\ 0\\ f^{N_t}-f_d^{N_t}\end{pmatrix}+\frac{\gamma}{2}\triangle x\triangle t\begin{pmatrix}u^2(t_1)e_N\\ \vdots\\ u^2(t_{N_t-1})e_N\\ 0\\ \vdots\\ 0\end{pmatrix}\in\mathbb{R}^{N_tN}\end{aligned}$$

and

$$\nabla_u J_{\mathrm{t}}(f,u)=\gamma\triangle x\triangle t\begin{pmatrix}u^0\sum_{x=a+\triangle x}^{b-\triangle x}f(x,t_0)\\ \vdots\\ u^{N_t-1}\sum_{x=a+\triangle x}^{b-\triangle x}f(x,t_{N_t-1})\end{pmatrix}\in\mathbb{R}^{N_t}.$$

Corollary 36. *The gradient of* $J_{\mathrm{c}}(f,u)$ *given in* (4.20) *with respect to* f *and* u *is given by*

$$\nabla_f J_{\mathrm{c}}(f,u)=\alpha\triangle x\triangle t\begin{pmatrix}c^1\\ \vdots\\ c^{N_t-1}\\ 0\\ \vdots\\ 0\end{pmatrix}+\beta\triangle x\begin{pmatrix}0\\ \vdots\\ 0\\ \Psi\end{pmatrix}+\frac{\gamma}{2}\triangle x\triangle t\begin{pmatrix}u^2(t_1)e_N\\ \vdots\\ u^2(t_{N_t-1})e_N\\ 0\\ \vdots\\ 0\end{pmatrix}\in\mathbb{R}^{N_tN}$$

and

$$\nabla_u J_{\mathrm{c}}(f,u)=\gamma\triangle x\triangle t\begin{pmatrix}u^0\sum_{x=a+\triangle x}^{b-\triangle x}f(x,t_0)\\ \vdots\\ u^{N_t-1}\sum_{x=a+\triangle x}^{b-\triangle x}f(x,t_{N_t-1})\end{pmatrix}\in\mathbb{R}^{N_t}$$

where $c^n:=\begin{pmatrix}c(a+\triangle x,t_n)\\ \vdots\\ c(b-\triangle x,t_n)\end{pmatrix}\in\mathbb{R}^N$ *for all* $n\geq 1$ *and* $\Psi:=\begin{pmatrix}\Psi(a+\triangle x)\\ \vdots\\ \Psi(b-\triangle x)\end{pmatrix}\in\mathbb{R}^N$.

For the discussion that follows, we remark that, in the case $J=J_{\mathrm{c}}(f,u)$ and the control constraints are not active, the optimality system represents a sufficient condition for an optimal solution to (4.28). This result follows from the fact that ζ in (4.31) is independent of f and thus the reduced gradient defined in (4.30) depends linearly on f. Let u be optimal and such that the control constraints are not active and f be such that $F(f,u)=0$. Then the following holds

$$\nabla_u\hat{J}_{\mathrm{c}}(u)=0,\qquad\text{that is}\qquad\left(-\frac{\triangle x}{2D}M_\zeta+\gamma\triangle x\triangle tM_u\right)\begin{pmatrix}f^0\\ \vdots\\ f^{N_t-1}\end{pmatrix}=0$$

where

$$M_\zeta=\begin{pmatrix}m_\zeta^1&&&\\&m_\zeta^2&&\\&&\ddots&\\&&&m_\zeta^{N_t}\end{pmatrix}\in\mathbb{R}^{N_t\times NN_t},\ m_\zeta^n=\begin{pmatrix}\zeta_2^n&\zeta_3^n-\zeta_1^n&\cdots&\zeta_N^n-\zeta_{N-2}^n&-\zeta_{N-1}^n\end{pmatrix}\in\mathbb{R}^{1\times N},$$

$\zeta_j^n := \zeta\left(a + j\triangle x, t_n\right)$ with $j \in \{1, ..., N\}$, $N = \frac{b-a}{\triangle x} - 1$, for $n \in \{1, ...N_t\}$ and

$$M_u = \begin{pmatrix} m_u^0 & & \\ & \ddots & \\ & & m_u^{N_t-1} \end{pmatrix} \in \mathbb{R}^{N_t \times N_t N},\ m_u^n = u\left(t_n\right) e_N^{\mathrm{T}} \in \mathbb{R}^{2N \times N}$$

for all $n \in \{0, ..., N_t - 1\}$ and thus as sufficient condition for optimality $-\frac{\triangle x}{2D} M_\zeta + \gamma \triangle x \triangle t M_u = 0$. From that, we express each entry of u by the relation $M_u = \frac{1}{2\gamma D \triangle t} M_\zeta$ and substitute u in the adjoint equation (4.31) by the corresponding expressions with ζ. The solution of the nonlinear system arising from the adjoint equation (4.31) is inserted in M_ζ again to obtain a solution for u form $M_u = \frac{1}{2\gamma D \triangle t} M_\zeta$.

4.1.3 Derivation of the Fokker-Planck equation

In this subsection, we present two ways of deriving the Fokker-Planck equation from the RW to illustrate the connection between the controlled CK equation and the controlled FP equation. We have $x \in \mathbb{X}_{x_0}^{\triangle x}$ and $t \in \mathbb{T}_{t_0}^{\triangle t}$.

First, as in [37], we interpret (4.2) to (4.6) as the conditional probability of discrete jumps of the process between the lattice points within the interval $\triangle t$ and investigate the limit for $\triangle x, \triangle t \to 0$ of our CK equation (4.9) to obtain the FP equation. We define the diffusion $\sigma^2\left(x, t\right)$ and the drift $\mu\left(x, t\right)$ of the FP equation as follows

$$\sigma^2\left(x, t\right) := \frac{\left(\triangle x\right)^2}{\triangle t}\left(p\left(x, t\right) + q\left(x, t\right)\right), \tag{4.34}$$

$$\mu\left(x, t\right) := \frac{\triangle x}{\triangle t}\left(p\left(x, t\right) - q\left(x, t\right)\right) \tag{4.35}$$

and we have the constant D defined in (4.17). Inserting (4.15) and (4.16) into (4.34) and (4.35), we have $\sigma^2\left(x, t\right) = Ds\left(x, t\right)$ and $\mu\left(x, t\right) = u\left(x, t\right)$. Notice that we identify the diffusion with $s\left(x, t\right)$, up to a constant D, and the drift with the control function $u\left(x, t\right)$.

Next, we require that f, σ, μ and g are defined in the continuous case and that f, σ^2 are twice differentiable with respect to x, f is once differentiable with respect to t, μ is once differentiable with respect to x, and the limit $\lim_{\triangle x, \triangle t \to 0} \lambda \triangle x \sum_{j=-\infty}^{\infty} g\left(x - y_j\right)\left(f\left(y_j, t\right) - f\left(x, t\right)\right)$ exists for all $x, y \in \mathbb{R}$ and is equal to $\lambda \int_{-\infty}^{\infty} g\left(x - y\right)\left(f\left(y, t\right) - f\left(x, t\right)\right) dy$ where $y_j := x - j\triangle x$ and $y_{j+1} - y_j = -\triangle x$, see [78] for details about the Riemann sum. From (4.34) and (4.35), we obtain

$$p\left(x, t\right) = \frac{\triangle t}{2\left(\triangle x\right)^2}\sigma^2\left(x, t\right) + \frac{\triangle t}{2\triangle x}\mu\left(x, t\right) \text{ and } q\left(x, t\right) = \frac{\triangle t}{2\left(\triangle x\right)^2}\sigma^2\left(x, t\right) - \frac{\triangle t}{2\triangle x}\mu\left(x, t\right) \tag{4.36}$$

and insert (4.36) into (4.9) and perform the limit $\lim_{\triangle x, \triangle t \to 0}$ with (4.17) as follows

$$\begin{aligned}
&\frac{\partial f}{\partial t}\left(x, t\right) \\
&= \lim_{\triangle x, \triangle t \to 0} \frac{f\left(x, t + \triangle t\right) - f\left(x, t\right)}{\triangle t} \\
&= \lim_{\triangle x, \triangle t \to 0}\left(\frac{\sigma^2\left(x + \triangle x, t\right) f\left(x + \triangle x, t\right) - 2\sigma^2\left(x, t\right) f\left(x, t\right) + \sigma^2\left(x - \triangle x, t\right) f\left(x - \triangle x, t\right)}{2\left(\triangle x\right)^2}\right) \\
&- \lim_{\triangle x, \triangle t \to 0}\left(\frac{\mu\left(x + \triangle x, t\right) f\left(x + \triangle x, t\right) - \mu\left(x - \triangle x, t\right) f\left(x - \triangle x, t\right)}{2\triangle x}\right) \\
&+ \lim_{\triangle x, \triangle t \to 0} \lambda \triangle x \sum_{j=-\infty}^{\infty} g\left(x - y_j\right)\left(f\left(y_j, t\right) - f\left(x, t\right)\right) \\
&= \frac{\partial^2}{\partial x^2}\left(\frac{\sigma^2\left(x, t\right)}{2} f\left(x, t\right)\right) - \frac{\partial}{\partial x}\left(u\left(x, t\right) f\left(x, t\right)\right) - \lambda \int_{\infty}^{-\infty} g\left(x - y\right)\left(f\left(y, t\right) - f\left(x, t\right)\right) dy.
\end{aligned}$$

Furthermore, we have $0 \leq p(x,t) + q(x,t) \leq 1$, thus $0 \leq s(x,t) \leq 1$ independent of the value of $\triangle x$. To fulfill $\sigma^2(x,t) = Ds(x,t)$, we set

$$D = \sup_{x,t\in\mathbb{R}} \sigma^2(x,t)\,.$$

Therefore, for a given $\triangle x$ and diffusion σ^2, the time step $\triangle t$ is given by (4.17). In correspondence to $s(x,t)$, we obtain a drift $u(x,t)$ in the limit $\triangle x \to 0$ according to (4.15) and (4.16) such that the condition $0 \leq p(x,t) \leq 1$ and $0 \leq q(x,t) \leq 1$ are fulfilled. Using the compact support of g, that means that $g(x-y) = 0$ if $|x-y| > x_c$, we finally obtain the following FP equation

$$\frac{\partial f}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(u(x,t)\,f(x,t)\right) - \frac{\partial^2}{\partial x^2}\left(\frac{\sigma^2(x,t)}{2}f(x,t)\right) - \lambda\int_{x-x_c}^{x+x_c} g(x-y)\left(f(y,t)-f(x,t)\right)dy = 0 \tag{4.37}$$

with the boundary conditions $f(x,t) = 0$ for $x \leq a$ and $b \leq x$.

We complete this section deriving the FP equation for a combination of absorbing and reflecting boundary conditions. This derivation is more involved and requires the introduction of the concept flux of the probability density.

Another approach from the discrete Chapman-Kolmogorov equation (4.9) to the Fokker-Planck equation is to consider a flux of probability density between the states which are connected pair wise by tubes in which the probability density flows according to (4.2) to (4.6). We interpret $p(x,t)\,f(x,t)$, $q(x,t)\,f(x,t)$ and $\lambda\triangle t\triangle x g(j\triangle x)\,f(x,t)$, $j \in \mathbb{Z}$ as microscopic particle flux densities where the ” + ” sign in front of them means that the flux points in increasing x direction and the "-" sign means that the flux points in decreasing x direction. We define the particle flux density between $x + \triangle x$ and x at t by

$$\begin{aligned} j^+(x,t) &:= \frac{\triangle x}{\triangle t}\left(p(x,t)\,f(x,t) - q(x+\triangle x,t)\,f(x+\triangle x,t)\right) \\ &- \lambda(\triangle x)^2 \sum_{j=-\infty}^{-1} g(j\triangle x)\sum_{\tilde{j}=0}^{j+1} f\left(x+\triangle x-\tilde{j}\triangle x,t\right) + \lambda(\triangle x)^2\sum_{j=1}^{\infty} g(j\triangle x)\sum_{\tilde{j}=-1}^{j} f\left(x-\triangle x-\tilde{j}\triangle x,t\right). \end{aligned} \tag{4.38}$$

The quantity $j^+(x,t)$ considers all incoming or outgoing fluxes as far as they go through x from or to $x+\triangle x$. As mentioned before, we have $f(b+j\triangle x,t) = 0$ for all $j \in \mathbb{N}_0$. The particle flux density between x and $x-\triangle x$ at t is defined by

$$\begin{aligned} j^-(x,t) &:= \frac{\triangle x}{\triangle t}\left(p(x-\triangle x,t)\,f(x-\triangle x,t) - q(x,t)\,f(x,t)\right) \\ &- \lambda(\triangle x)^2 \sum_{j=-\infty}^{-1} g(j\triangle x)\sum_{\tilde{j}=-1}^{j} f\left(x-\triangle x-\tilde{j}\triangle x,t\right) + \lambda(\triangle x)^2\sum_{j=1}^{\infty} g(j\triangle x)\sum_{\tilde{j}=0}^{j+1} f\left(x+\triangle x-\tilde{j}\triangle x,t\right). \end{aligned} \tag{4.39}$$

The quantity $j^-(x,t)$ considers all incoming or outgoing fluxes as far as they go through x from or to $x-\triangle x$. As mentioned before $f(a-j\triangle x,t) = 0$ for all $j \in \mathbb{N}_0$. The total particle flux density at (x,t) is defined by

$$\jmath(x,t) := \frac{j^+(x,t) + j^-(x,t)}{2}, \tag{4.40}$$

which is an arithmetic mean of $j^+(x,t)$ and $j^-(x,t)$. The next definitions are made to reproduce the discrete Fokker-Planck equation described in Subsection 4.1.1.

The discrete divergence of the flux is defined as follows

$$\nabla_x^{\mathcal{D}} \jmath(x,t) := \frac{j^+(x,t) - j^-(x,t)}{\triangle x}. \tag{4.41}$$

Next, we set the discrete flux equation at (x,t)

$$\frac{f(x,t+\triangle t) - f(x,t)}{\triangle t} = -\nabla_x^{\mathcal{D}} \jmath(x,t). \tag{4.42}$$

If the discrete occupation probability density at (x,t) gets greater when time increases, then $\nabla_x^{\mathcal{D}} \jmath(x,t) < 0$ or if the discrete occupation probability density gets smaller when time increases, then $\nabla_x^{\mathcal{D}} \jmath(x,t) > 0$.

Remark 37. Inserting (4.38) and (4.39) into (4.42), we have that (4.42) is equivalent to (4.9) using that

$$\begin{aligned}
&\sum_{j=-\infty}^{-1} \triangle x g(j\triangle x) \left(\sum_{\tilde{j}=0}^{j+1} f\left(x + \triangle x - \tilde{j}\triangle x, t\right) - \sum_{\tilde{j}=-1}^{j} f\left(x - \triangle x - \tilde{j}\triangle x, t\right) \right) \\
&\quad = \sum_{j=-\infty}^{-1} \triangle x g(j\triangle x) \left(\sum_{\tilde{j}=1}^{j} f\left(x - \tilde{j}\triangle x, t\right) - \sum_{\tilde{j}=0}^{j-1} f\left(x - \tilde{j}\triangle x, t\right) \right) \\
&\qquad = f(x - j\triangle x, t) - f(x,t),
\end{aligned}$$

$$\begin{aligned}
&\sum_{j=1}^{\infty} \triangle x g(j\triangle x) \left(\sum_{\tilde{j}=-1}^{j} f\left(x - \triangle x - \tilde{j}\triangle x, t\right) - \sum_{\tilde{j}=0}^{j+1} f\left(x + \triangle x - \tilde{j}\triangle x, t\right) \right) \\
&\quad \sum_{j=1}^{\infty} \triangle x g(j\triangle x) \left(\sum_{\tilde{j}=0}^{j-1} f\left(x - \tilde{j}\triangle x, t\right) - \sum_{\tilde{j}=1}^{j} f\left(x - \tilde{j}\triangle x, t\right) \right) \\
&\qquad f(x,t) - f(x - j\triangle x, t)
\end{aligned}$$

and $f(x - j\triangle x, t) - f(x,t) = 0$ for $j = 0$.

In the next lemma, we prove that in connection with (4.42) the function $f(x,t)$ is non negative for all $x \in \mathbb{X}_{x_0}^{\triangle x}$ and $t \in \mathbb{T}_{t_0}^{\triangle t}$ once starting with $f(x,t_0) \geq 0$ for all $x \in \mathbb{X}_{x_0}^{\triangle x}$ where $\geq$ works componentwise.

Lemma 38. *If $f(x,t_0) \geq 0$ for all $x \in \mathbb{X}_{x_0}^{\triangle x}$, then $f(x,t) \geq 0$ for all $t \in \mathbb{T}_{t_0}^{\triangle t}$ and $x \in \mathbb{X}_{x_0}^{\triangle x}$ following* (4.42).

Proof. The proof is done by complete induction. We have

$$\frac{f(x,t_0+\triangle t) - f(x,t_0)}{\triangle t} = -\nabla_x^{\mathcal{D}} \jmath(x,t_0),$$

or equivalently,

$$\begin{aligned}
f(x,t_0+\triangle t) &= p(x-\triangle x,t_0) f(x-\triangle x,t_0) + q(x+\triangle x,t_0) f(x+\triangle x,t_0) \\
&\quad + \lambda\triangle t\triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} \left(g(j\triangle x) f(x - j\triangle x, t_0) \right) \\
&\quad + \underbrace{\left(1 - p(x,t_0) - q(x,t_0) - \lambda\triangle t\triangle x \sum_{\substack{j=-\infty \\ j\neq 0}}^{\infty} g(j\triangle x) \right)}_{\geq 0} f(x,t_0) \geq 0.
\end{aligned}$$

Assuming that $f\left(x,t\right)\geq 0$ for one $t\in\mathbb{T}_{t_0}^{\triangle t}$, we have

$$\begin{aligned} f\left(x,t+\triangle t\right) &= p\left(x-\triangle x,t\right)f\left(x-\triangle x,t\right)+q\left(x+\triangle x,t\right)f\left(x+\triangle x,t\right)\\ &\quad +\lambda\triangle t\triangle x\sum_{\substack{j=-\infty\\ j\neq 0}}^{\infty}\left(g\left(j\triangle x\right)f\left(x-j\triangle x,t\right)\right)\\ &\quad +\underbrace{\left(1-p\left(x,t\right)-q\left(x,t\right)-\lambda\triangle t\triangle x\sum_{\substack{j=-\infty\\ j\neq 0}}^{\infty}g\left(j\triangle x\right)\right)}_{\geq 0}f\left(x,t\right)\geq 0. \end{aligned}$$

□

Remark 39. Inserting (4.34) and (4.35) into (4.40), we have

$$\jmath\left(x,t\right)=\frac{\mu\left(x,t\right)}{2}f\left(x,t\right)-\frac{1}{2}\left(\frac{\sigma^2\left(x+\triangle x,t\right)f\left(x+\triangle x,t\right)-\sigma^2\left(x-\triangle x,t\right)f\left(x-\triangle x,t\right)}{2\triangle x}\right)\tag{4.43}$$

$$+\frac{\mu\left(x+\triangle x,t\right)}{4}f\left(x+\triangle x,t\right)+\frac{\mu\left(x-\triangle x,t\right)}{4}f\left(x-\triangle x,t\right)\tag{4.44}$$

$$-\frac{1}{2}\lambda\triangle x\left(\sum_{j=-\infty}^{-1}\triangle x g\left(j\triangle x\right)\sum_{\tilde{\jmath}=0}^{j+1}f\left(x+\triangle x-\tilde{\jmath}\triangle x,t\right)\right)\tag{4.45}$$

$$-\frac{1}{2}\lambda\triangle x\left(\sum_{j=-\infty}^{-1}\triangle x g\left(j\triangle x\right)\sum_{\tilde{\jmath}=-1}^{j}f\left(x-\triangle x-\tilde{\jmath}\triangle x,t\right)\right)\tag{4.46}$$

$$+\frac{1}{2}\lambda\triangle x\left(\sum_{j=1}^{\infty}\triangle x g\left(j\triangle x\right)\sum_{\tilde{\jmath}=0}^{j+1}f\left(x+\triangle x-\tilde{\jmath}\triangle x,t\right)\right)\tag{4.47}$$

$$+\frac{1}{2}\lambda\triangle x\left(\sum_{j=1}^{\infty}\triangle x g\left(j\triangle x\right)\sum_{\tilde{\jmath}=-1}^{j}f\left(x-\triangle x-\tilde{\jmath}\triangle x,t\right)\right)\tag{4.48}$$

for all $x\in\mathbb{X}_{x_0}^{\triangle x}$ and $t\in\mathbb{T}_{t_0}^{\triangle t}$.

If f and σ^2 are differentiable with respect to x, μ is continuous and (4.45) to (4.48) are Riemann sums, see [78], where $y_j:=x-j\triangle x$ and $\tilde{y}_{\tilde{\jmath}}:=x-\tilde{\jmath}\triangle x$, then, similarly to [9], we have that

$$\lim_{\triangle x\to 0}\jmath\left(x,t\right)=\mu\left(x,t\right)f\left(x,t\right)-\frac{\partial}{\partial x}\left(\frac{\sigma^2\left(x,t\right)}{2}f\left(x,t\right)\right)-\lambda\int_{-\infty}^{\infty}g\left(x-y\right)\int_x^y f\left(\tilde{y},t\right)d\tilde{y}dy.$$

We have $y_{j+1}-y_j=-\triangle x$ and $\tilde{y}_{\tilde{\jmath}-1}-\tilde{y}_{\tilde{\jmath}}=\triangle x$. According to the definition of the Riemann sum, see [78], the limit for $\triangle x\to 0$ of (4.45) is given by $-\frac{1}{2}\lambda\int_{\infty}^{0}-g\left(x-y\right)\int_x^y f\left(\tilde{y},t\right)d\tilde{y}dy$ and of (4.46) is given by $-\frac{1}{2}\lambda\int_{\infty}^{0}-g\left(x-y\right)\int_x^y f\left(\tilde{y},t\right)d\tilde{y}dy$. We have $\tilde{y}_{\tilde{\jmath}+1}-\tilde{y}_{\tilde{\jmath}}=-\triangle x$ and thus the limit of (4.47) for $\triangle x\to 0$ is given by $\frac{1}{2}\lambda\int_{0}^{-\infty}-g\left(x-y\right)\int_x^y -f\left(\tilde{y},t\right)d\tilde{y}dy$ and of (4.48) is given by $\frac{1}{2}\lambda\int_{0}^{-\infty}-g\left(x-y\right)\int_x^y -f\left(\tilde{y},t\right)d\tilde{y}dy$.

Now, we give $j_{\text{ref}}^{+}\left(x,t\right)$ and $j_{\text{ref}}^{-}\left(x,t\right)$ such that we obtain (4.10) to (4.14) inserting $j_{\text{ref}}^{+}\left(x,t\right)$ and $j_{\text{ref}}^{-}\left(x,t\right)$ into (4.42) for all $x\in\mathbb{X}_{x_0}^{\triangle x}$ and $t\in\mathbb{T}_{t_0}^{\triangle t}$. For this purpose, we add the contributions from the reflecting barrier to $j^{+}\left(x,t\right)$ and $j^{-}\left(x,t\right)$. For $x=a$, we have

$$j_{\text{ref}}^{+}\left(a,t\right) :=\kappa_a\left(t\right)\left(j^{+}\left(a,t\right)-\lambda\kappa_b\left(t\right)\left(\triangle x\right)^2\sum_{j=1}^{\infty}g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(2b-a-\tilde{j}\triangle x,t\right)\right)$$

and

$$j_{\text{ref}}^{-}\left(a,t\right) :=\kappa_a\left(t\right)\left(j^{-}\left(a,t\right)-\lambda\kappa_b\left(t\right)\left(\triangle x\right)^2\sum_{j=1}^{\infty}g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(2b-a-\tilde{j}\triangle x,t\right)\right).$$

For $x=a+\triangle x$, we obtain

$$\begin{aligned}j_{\text{ref}}^{+}\left(a+\triangle x,t\right) :=&j^{+}\left(a+\triangle x,t\right)-\kappa_b\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=1}^{\infty}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(2b-a-\triangle x-\tilde{j}\triangle x,t\right)\right)\\&+\kappa_a\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(a-\triangle x-\tilde{j}\triangle x,t\right)\right)\end{aligned}$$

and

$$\begin{aligned}j_{\text{ref}}^{-}\left(a+\triangle x,t\right) :=&j^{-}\left(a+\triangle x,t\right)+\kappa_a\left(t\right)\frac{\triangle x}{\triangle t}q\left(a,t\right)f\left(a,t\right)\\&-\kappa_b\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=1}^{\infty}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(2b-a-\triangle x-\tilde{j}\triangle x,t\right)\right)\\&+\kappa_a\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(a-\triangle x-\tilde{j}\triangle x,t\right)\right).\end{aligned}$$

If $a+\triangle x<x<b-\triangle x$, then

$$\begin{aligned}j_{\text{ref}}^{+}\left(x,t\right) :=& j^{+}\left(x,t\right)-\kappa_b\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=1}^{\infty}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(2b-x-\tilde{j}\triangle x,t\right)\right)\\&+\kappa_a\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(2a-x-\tilde{j}\triangle x,t\right)\right)\end{aligned}\tag{4.49}$$

and

$$\begin{aligned}j_{\text{ref}}^{-}\left(x,t\right) :=& j^{-}\left(x,t\right)-\kappa_b\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=1}^{\infty}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(2b-x-\tilde{j}\triangle x,t\right)\right)\\&+\kappa_a\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(2a-x-\tilde{j}\triangle x,t\right)\right).\end{aligned}\tag{4.50}$$

For $x = b - \triangle x$, we have

$$\begin{aligned}\jmath^{+}_{\text{ref}}\left(b-\triangle x,t\right) \coloneqq \jmath^{+}\left(b-\triangle x,t\right) &- \kappa_b\left(t\right)\frac{\triangle x}{\triangle t}p\left(b,t\right)f\left(b,t\right)\\ &- \kappa_b\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=1}^{\infty}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(b+\triangle x-\tilde{j}\triangle x,t\right)\right)\\ &+ \kappa_a\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(2a-b+\triangle x-\tilde{j}\triangle x,t\right)\right)\end{aligned}$$

and

$$\begin{aligned}\jmath^{-}_{\text{ref}}\left(b-\triangle x,t\right) \coloneqq& \jmath^{-}\left(b-\triangle x,t\right) - \kappa_b\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=1}^{\infty}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(b+\triangle x-\tilde{j}\triangle x,t\right)\right)\\ &+ \kappa_a\left(t\right)\lambda\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(2a-b+\triangle x-\tilde{j}\triangle x,t\right)\right).\end{aligned}$$

For $x = b$, we obtain

$$\jmath^{+}_{\text{ref}}\left(b,t\right) \coloneqq \kappa_b\left(t\right)\left(\jmath^{+}\left(b,t\right)+\lambda\kappa_a\left(t\right)\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j-1}f\left(2a-b-\tilde{j}\triangle x,t\right)\right)\right)$$

and

$$\jmath^{-}_{\text{ref}}\left(b,t\right) \coloneqq \kappa_b\left(t\right)\left(\jmath^{-}\left(b,t\right)+\lambda\kappa_a\left(t\right)\left(\triangle x\right)^2\sum_{j=-\infty}^{-1}\left(g\left(j\triangle x\right)\sum_{\tilde{j}=0}^{j}f\left(2a-b-\tilde{j}\triangle x,t\right)\right)\right).$$

In the limit for $\triangle x \to 0$, we obtain from (4.40) with (4.49) and (4.50) that

$$\begin{aligned}&\jmath^{c}\left(x,t\right) \coloneqq \lim_{\triangle x\to 0}\jmath\left(x,t\right)\\ &= \mu\left(x,t\right)f\left(x,t\right)-\frac{\partial}{\partial x}\left(\frac{\sigma^2\left(x,t\right)}{2}f\left(x,t\right)\right)-\lambda\int_{-\infty}^{\infty}g\left(x-y\right)\int_{x}^{y}f\left(\tilde{y},t\right)d\tilde{y}dy \qquad (4.51)\\ &+\kappa_a\left(t\right)\lambda\int_{x}^{\infty}g\left(x-y\right)\int_{x}^{y}f\left(2\left(a-x\right)+\tilde{y}\right)d\tilde{y}dy+\kappa_b\left(t\right)\lambda\int_{-\infty}^{x}g\left(x-y\right)\int_{x}^{y}f\left(2\left(b-x\right)+\tilde{y}\right)d\tilde{y}dy\end{aligned}$$

analogously to Remark 39. Inserting (4.49) and (4.50) into (4.42), we have

$$\begin{aligned}\frac{\partial f}{\partial t}\left(x,t\right) = -\frac{\partial}{\partial x}\left(\mu\left(x,t\right)f\left(x,t\right)\right)+\frac{\partial^2}{\partial x^2}\left(\frac{\sigma^2\left(x,t\right)}{2}f\left(x,t\right)\right)+\lambda\int_{-\infty}^{\infty}g\left(x-y\right)\left(f\left(y,t\right)-f\left(x,t\right)\right)dy\\ +\kappa_a\left(t\right)\lambda\int_{x}^{\infty}g\left(x-y\right)f\left(2\left(a-x\right)+y\right)dy+\kappa_b\left(t\right)\lambda\int_{-\infty}^{x}g\left(x-y\right)f\left(2\left(b-x\right)+y\right)dy\end{aligned} \qquad (4.52)$$

analogously to (4.37).

In the next lemma we have that the loss of probability can be expressed by the flux at the boundary.

Lemma 40. *Considering* (4.10) *to* (4.14) *for* $g = 0$*, it holds*

$$\begin{aligned}
&\frac{\triangle x}{\triangle t}\left(\sum_{x=a}^{b} f\left(x, t+\triangle t\right) - \sum_{x=a}^{b} f\left(x,t\right)\right)\\
&= \kappa_a\left(t\right)\left(\frac{\sigma^2\left(a+\triangle x, t\right) f\left(a+\triangle x, t\right) - \sigma^2\left(a,t\right) f\left(a,t\right)}{2\triangle x} - \frac{1}{2}\left(\mu\left(a+\triangle x,t\right) f\left(a+\triangle x, t\right) + \mu\left(a,t\right) f\left(a,t\right)\right)\right)\\
&- \left(\frac{\sigma^2\left(a+\triangle x, t\right) f\left(a+\triangle x, t\right) - \sigma^2\left(a,t\right) f\left(a,t\right)}{2\triangle x} - \frac{1}{2}\left(\mu\left(a+\triangle x,t\right) f\left(a+\triangle x, t\right) + \mu\left(a,t\right) f\left(a,t\right)\right)\right)\\
&+ \frac{\sigma^2\left(b,t\right) f\left(b,t\right) - \sigma^2\left(b-\triangle x, t\right) f\left(b-\triangle x,t\right)}{2\triangle x} - \frac{1}{2}\left(\mu\left(b,t\right) f\left(b,t\right) + \mu\left(b-\triangle x,t\right) f\left(b-\triangle x, t\right)\right)\\
&- \kappa_b\left(t\right)\left(\frac{\sigma^2\left(b,t\right) f\left(b,t\right) - \sigma^2\left(b-\triangle x, t\right) f\left(b-\triangle x,t\right)}{2\triangle x} - \frac{1}{2}\left(\mu\left(b,t\right) f\left(b,t\right) + \mu\left(b-\triangle x,t\right) f\left(b-\triangle x, t\right)\right)\right)
\end{aligned}$$

for $x \in \mathbb{X}_{x_0}^{\triangle x}$.

Proof. A straightforward calculation gives

$$\begin{aligned}
&\frac{\triangle x}{\triangle t}\left(\sum_{x=a}^{b} f\left(x, t+\triangle t\right) - \sum_{x=a}^{b} f\left(x,t\right)\right)\\
&= \frac{\triangle x}{\triangle t}\left(\kappa_a\left(t\right) q\left(a+\triangle x,t\right) f\left(a+\triangle x,t\right) - \kappa_a\left(t\right) p\left(a,t\right) f\left(a,t\right) + p\left(a,t\right) f\left(a,t\right) - q\left(a+\triangle x,t\right) f\left(a+\triangle x, t\right)\right)\\
&+ \frac{\triangle x}{\triangle t}\left(q\left(b,t\right) f\left(b,t\right) - p\left(b-\triangle x,t\right) f\left(b-\triangle x,t\right) + \kappa_b\left(t\right) p\left(b-\triangle x,t\right) f\left(b-\triangle x,t\right) - \kappa_b\left(t\right) q\left(b,t\right) f\left(b,t\right)\right).
\end{aligned}$$

By inserting (4.36), we obtain the claim. □

Remark 41. If $\sigma^2 f$ is differentiable with respect to x, f is differentiable with respect to t and μf is continuous, then we have

$$\begin{aligned}
&\int_a^b \frac{\partial f\left(x,t\right)}{\partial t} dx\\
&= \left(1-\kappa_a\left(t\right)\right)\left(\mu\left(a\right) f\left(a\right) - \frac{\partial}{\partial x}\left(\frac{\sigma^2\left(a,t\right)}{2} f\left(a,t\right)\right)\right)\\
&+ \left(\kappa_b\left(t\right)-1\right)\left(\mu\left(b\right) f\left(b\right) - \frac{\partial}{\partial x}\left(\frac{\sigma^2\left(b,t\right)}{2} f\left(b,t\right)\right)\right)
\end{aligned}$$

in the limit $\triangle x \to 0$ according to Lemma 40.

Remark 42. For $\kappa_a \equiv \kappa_b \equiv 1$, we have $\jmath^c\left(a,t\right) = \jmath^c\left(b,t\right) = 0$. We perform the calculation for $x = a$. It is analogous for $x = b$. The main part is to see that both $\int_{-\infty}^{a} g\left(a-y\right) \int_a^y f\left(\tilde{y},t\right) d\tilde{y}dy = 0$ and $\int_{-\infty}^{a} g\left(a-y\right) \int_a^y f\left(2\left(b-a\right)+\tilde{y}\right) d\tilde{y}dy = 0$. With a transformation of variables, see [78] and the compact support of g, we obtain $\int_0^{x_c} g\left(\tilde{z}\right) \int_a^{a-\tilde{z}} f\left(\tilde{y},t\right) d\tilde{y}d\tilde{z} = 0$ and $\int_0^{x_c} g\left(\tilde{z}\right) \int_{2(b-a)+a}^{2(b-a)+a-\tilde{z}} f\left(z\right) dz d\tilde{z} = 0$ because $x_c \leq b-a$ and $f\left(x,t\right) = 0$ outside the domain by definition.

The discussion above motivates the following definition of the boundary conditions for (4.52) originating from our microscopic model

$$\jmath^c\left(a,t\right) := \left(1-\kappa_a\left(t\right)\right)\left(\mu\left(a\right) f\left(a\right) - \frac{\partial}{\partial x}\left(\frac{\sigma^2\left(a,t\right)}{2} f\left(a,t\right)\right) - \lambda \int_a^{\infty} g\left(a-y\right) \int_a^y f\left(\tilde{y},t\right) d\tilde{y}dy\right)$$

and

$$\jmath^c\left(b,t\right) := \left(\kappa_b\left(t\right)-1\right)\left(\mu\left(a\right) f\left(a\right) - \frac{\partial}{\partial x}\left(\frac{\sigma^2\left(a,t\right)}{2} f\left(a,t\right)\right) + \int_{-\infty}^{b} g\left(b-y\right) \int_b^y f\left(\tilde{y},t\right) d\tilde{y}dy\right)$$

where we have zero flux conditions if $\kappa_a \equiv \kappa_b \equiv 1$. If $\kappa_a \equiv 0$ or $\kappa_b \equiv 0$, then the corresponding flux boundary condition is substituted by the zero boundary condition, see (4.37). We introduce boundary flux operators

$$B_a(f,t) \coloneqq j^c(a,t) \tag{4.53}$$

and

$$B_b(f,t) \coloneqq j^c(b,t)\,. \tag{4.54}$$

According to (4.52), (4.53) and (4.54), we summarize the discussion above with the following model originating from our microscopic model using the compact support of g, that means that $g(x-y)=0$ if $|x-y|>x_c$

$$\begin{aligned}\frac{\partial f}{\partial t}(x,t) = &-\frac{\partial}{\partial x}\left(\mu(x,t)\,f(x,t)\right) + \frac{\partial^2}{\partial x^2}\left(\frac{\sigma^2(x,t)}{2}f(x,t)\right) + \lambda\int_{x-x_c}^{x+x_c} g(x-y)\left(f(y,t)-f(x,t)\right)dy\\ &+\kappa_a(t)\,\lambda\int_{x}^{x+x_c} g(x-y)\,f(2(a-x)+y)\,dy + \kappa_b(t)\,\lambda\int_{x-x_c}^{x} g(x-y)\,f(2(b-x)+y)\,dy\end{aligned}$$

for $x \in (a,b)$,

$$B_a(f,t) = \gamma_a(t)$$

for $x = a$ and

$$B_b(f,t) = \gamma_b(t)$$

for $x = b$ where $\gamma_a, \gamma_b : \mathbb{R}_0^+ \to \mathbb{R}$ are given functions that means given boundary values of B_a and B_b and $f(x,t)=0$ outside the domain for all $t \geq 0$.

4.1.4 Numerical experiments

In this subsection, we present results of numerical experiments to validate our control framework. On the one hand, we show that our control framework is robust with respect to the choice of the parameters characterizing the RW model and the optimal control settings. Specifically, we consider control problems corresponding to the two cost functionals J_{t} with two absorbing barriers and J_c with two reflecting barriers. On the other hand, we aim at demonstrating that the RW control solution converges to the solution of a FP control problem.

First, we solve the optimal control problem (4.28) for $J = J_t$ defined in (4.19) with the following RW setting: $\triangle t_i = T\left(\frac{1}{4}\right)^{4+i}$ and $\triangle x_i = \left(\frac{1}{2}\right)^{4+i}\sqrt{DT}$, $i \in \{1,2,3\}$. We choose $T = D = 1$ and $\alpha = \beta = 1$ and $\gamma = \frac{1}{10}$ and $s = \frac{1}{10}$ with two absorbing barriers at $a = 0$ and $b = 1$. For all $t \in (0,T]$, we choose the trajectory-target function as follows

$$f_d(x,t) = \begin{cases} 5 & \text{if } \bar{x}(t) - \frac{1}{10} \leq x \leq \bar{x}(t) + \frac{1}{10}\\ 0 & \text{else}\end{cases}$$

where $\bar{x}(t) \coloneqq x_0 + \left(\frac{b-a}{2} - \frac{3}{20}\right)\sin\left(2\pi\frac{t}{T}\right)$ and $x_0 = \frac{b+a}{2}$. The initial state f^0 is given by

$$f^0(x) = \begin{cases} \frac{5}{2} & \text{if } \frac{b+a}{2} - \frac{1}{5} \leq x \leq \frac{b+a}{2} + \frac{1}{5}\\ 0 & \text{else}\end{cases}\,. \tag{4.55}$$

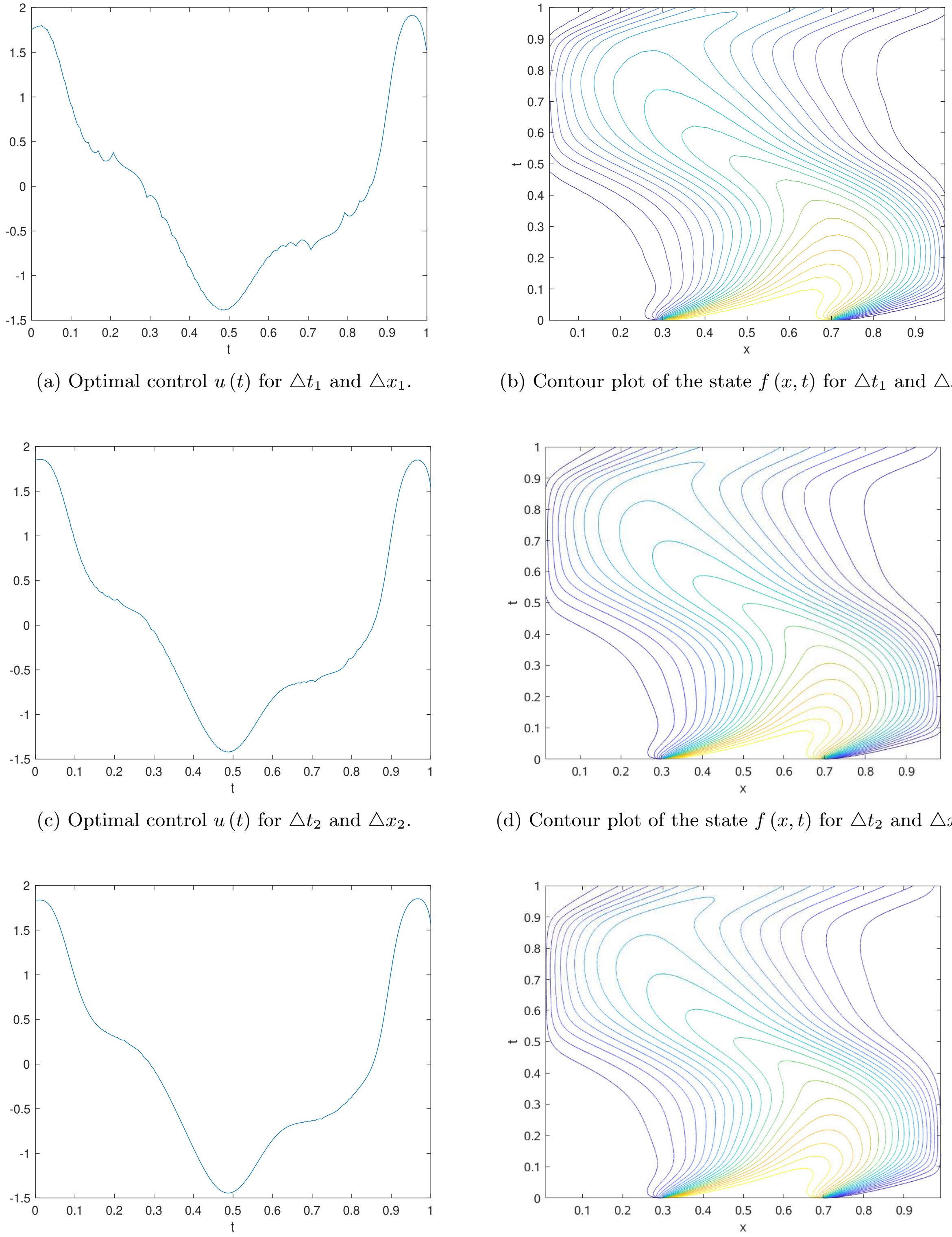
We choose a jump rate $\lambda = 1$ and the length of the jump distribution is modeled with the following function

$$g(x) = \begin{cases} 10 & \text{if } -\frac{1}{10} \leq x \leq \frac{1}{10} \\ 0 & \text{else} \end{cases}. \tag{4.56}$$

To solve our optimization problems, we use a nonlinear conjugate gradient (NCG) scheme in the variant proposed by Hager and Zhang [50]. To test the accuracy of the computed reduced gradient, we perform several gradient tests comparing with a centered difference quotient of the reduced cost functional and obtain convergence of quadratic order. Specifically, we compare our gradient $\nabla \hat{J}$ applied to a variation δu with the value of the finite-difference formula $\frac{\hat{J}(u+\theta\delta u)-\hat{J}(u-\theta\delta u)}{2\theta}$ for different $\theta > 0$ as $\theta \to 0$. In all experiments, the iterative NCG scheme starts with an initial guess for the control given by $u^0 = 0 \in \mathbb{R}^{N_t}$ and the stopping criterion is given by $\sqrt{dt \nabla_u \hat{J}(u)^{\mathrm{T}} \nabla_u \hat{J}(u)} < 10^{-8}$. This stopping criterion can be used as far as the control is within its bounds, see (4.29), which is always our case.

In the Figure 4.1, we present results of numerical experiments with three different space- and time-step sizes for the RW model.

Notice that f is set equal to zero at the absorbing barriers for all times and it follows the desired trajectory given by f_d. Furthermore, the control (drift) u behaves as it could be expected. As we refine the space- and time-step sizes, we see a pointwise convergence of the optimal control solutions.



(a) Optimal control $u(t)$ for $\triangle t_1$ and $\triangle x_1$.

(b) Contour plot of the state $f(x,t)$ for $\triangle t_1$ and $\triangle x_1$.

(c) Optimal control $u(t)$ for $\triangle t_2$ and $\triangle x_2$.

(d) Contour plot of the state $f(x,t)$ for $\triangle t_2$ and $\triangle x_2$.

(e) Optimal control $u(t)$ for $\triangle t_3$ and $\triangle x_3$.

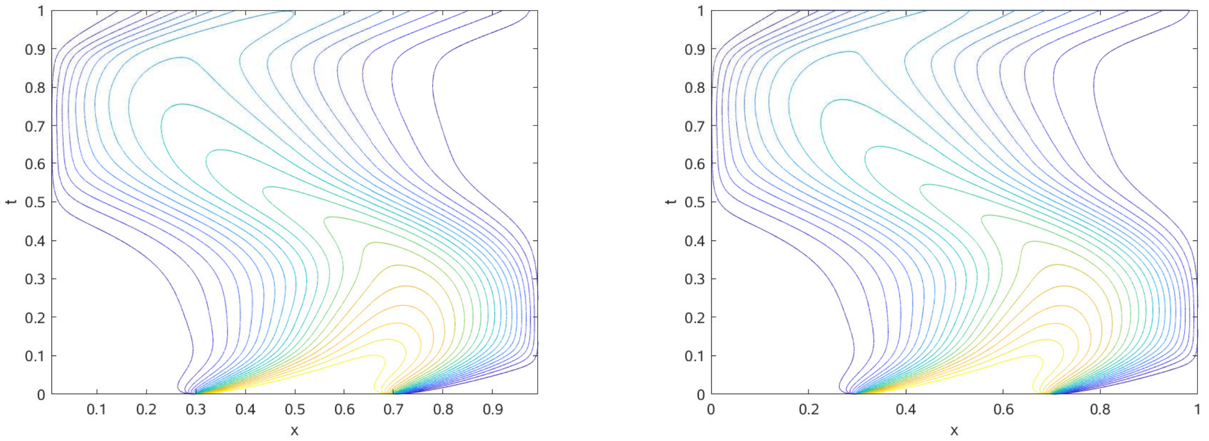
(f) Contour plot of the state $f(x,t)$ for $\triangle t_3$ and $\triangle x_3$.

Figure 4.1: Optimal control $u(t)$ and corresponding state $f(x,t)$ with absorbing barriers refining space- and time-step size.

Another experiment in the same numerical optimization setting as above for $\triangle t_3$ and $\triangle x_3$ is to solve the Fokker-Planck equation

$$\frac{\partial f\left(x,t\right)}{\partial t}=-u\left(t\right)\frac{\partial f\left(x,t\right)}{\partial x}+\frac{s}{2}\frac{\partial^{2} f\left(x,t\right)}{\partial x^{2}} \tag{4.57}$$

arising from the Chapman-Kolmogorov framework shown in Subsection 4.1.3 for $g=0$ with a Chang-Cooper method, see [30] with the optimal control u calculated from the discrete RW model by the Hager-Zhang-NCG method. In Figure 4.2, we see a contour plot of f. In 4.2a, the state f corresponds to the optimal control u, both calculated with the Hager-Zhang-NCG method. In 4.2b, the state f is calculated with the Chang-Cooper method from the continuous equation (4.57). Both figures of the state look the same since our discrete random walk is a discretization of the Fokker-Planck equation (4.57) according to Subsection 4.1.3.



(a) Contour plot of the state $f\left(x,t\right)$ calculated by the Hager-Zhang-NCG method from the discrete RW model.

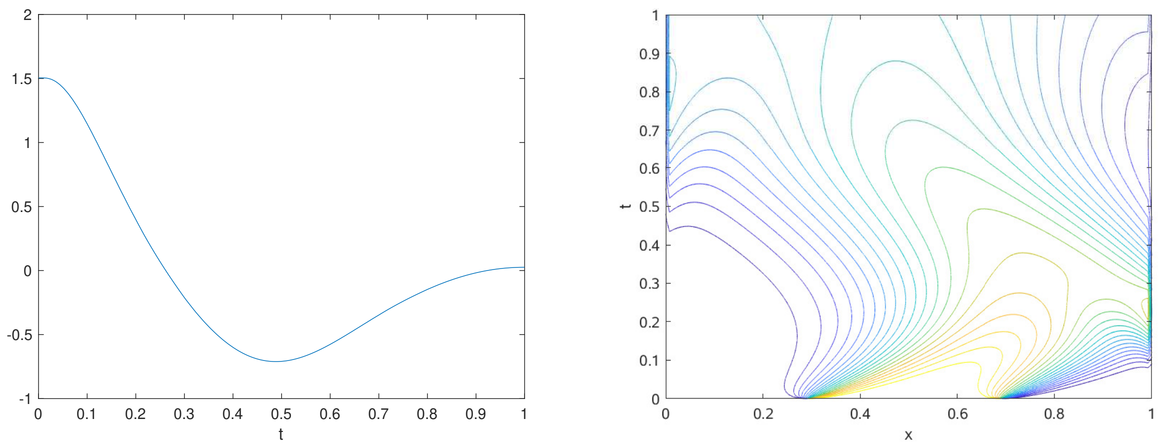
(b) Contour plot of the state $f\left(x,t\right)$ calculated by the Chang-Cooper method from the Fokker-Planck equation (4.57).

Figure 4.2

Our last numerical experiment considers the target functional J_c defined in (4.20) with reflecting barriers at $a=0$ and $b=1$ where $D=T=1$ and $N=\frac{b-a}{\triangle x}+1$. The discretization is $\triangle x_3$ and $\triangle t_3$. The diffusion $s=\frac{1}{10}$ and the function g is given as in (4.56). Furthermore, $\alpha=\beta=1$, $\gamma=\frac{1}{10}$ and $\lambda=1$. The initial value $u^0=0\in\mathbb{R}_t^N$ and f^0 is given as in (4.55). For $t\in(0,T]$, the function c is defined as follows

$$c\left(x,t\right)=\begin{cases}0 & \text{if } \bar{x}\left(t\right)-\frac{1}{10}\leq x\leq\bar{x}\left(t\right)+\frac{1}{10}\\ 100 & \text{else}\end{cases}$$

where $\bar{x}\left(t\right):=x_0+\left(\frac{b-a}{2}-\frac{7}{20}\right)\sin\left(2\pi\frac{t}{T}\right)$ and $x_0=\frac{b+a}{2}$. The results of the experiment are presented in Figure 4.3 where we use a nonlinear conjugate gradient (NCG) scheme in the variant proposed by Hager and Zhang [50] to calculate our optimization problem.



(a) Optimal control $u(t)$.

(b) Contour plot of the state $f(x,t)$.

Figure 4.3: Numerical experiment with reflecting barriers at $a = 0$ and $b = 1$.

4.2 The formulation of Fokker-Planck optimal control problems

In this section, we formulate optimal control problems governed by the FP equation. Our FP equation results from the consideration of stochastic Itô models that are continuous-time stochastic processes described by the following stochastic differential equation (SDE)

$$\begin{aligned} dX(t) &= b(X(t),t)\,dt + \sigma(X(t),t)\,dW(t) \\ X(0) &= X_0 \end{aligned} \tag{4.58}$$

where the state variable $X(t) \in \Omega \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$ is subject to deterministic infinitesimal increments driven by the vector valued drift function b and to random increments proportional to a multi-dimensional Wiener process $dW(t) \in \mathbb{R}^m$, $m \in \mathbb{N}$, with stochastically independent components. We denote with $\Omega \subseteq \mathbb{R}^n$ an open bounded domain with smooth boundary and measure $|\Omega|$. We assume that there are absorbing barriers on $\partial\Omega$.

The FP equation associated to (4.58) is given by

$$\partial_t f(x,t) - \sum_{i,j=1}^{n} \partial^2_{x_i x_j}\left(a_{ij}(x,t)\, f(x,t)\right) + \sum_{i=1}^{n} \partial_{x_i}\left(b_i(x,t)\, f(x,t)\right) = 0 \tag{4.59}$$

$$f(x,0) = f_0(x) \tag{4.60}$$

where f denotes the PDF of the process, f_0 represents the initial PDF of the initial state of the process X_0 and hence $f_0(x) \geq 0$ with $\int_\Omega f_0(x)dx = 1$. In general $\sigma \in \mathbb{R}^{n\times m}$ is a matrix which results in the diffusion coefficients a_{ij}. However, in our discussion, we assume that the diffusion coefficient a_{ij} are constants all with the value $\frac{\sigma^2}{2} > 0$ such that we have $a_{ij} = \frac{\sigma^2}{2}$. Both the stochastic process (4.58) and the FP equation (4.59) are considered in the time interval $[0,T]$ and Ω represents also the domain where f is defined. We denote with $Q := \Omega \times (0,T)$. Corresponding to absorbing barriers, we have homogeneous Dirichlet boundary conditions for f on $\partial\Omega$, $t \in [0,T]$.

For the intention of this thesis, our starting point is given by the following two stochastic processes. The first one is given by

$$dX(t) = \left(v(t) + w(t) \circ X(t)\right) dt + \sigma dW(t) \tag{4.61}$$

where $\circ$ denotes the Hadamard product $x \circ w := \begin{pmatrix} x_1 w^1 \\ \vdots \\ x_n w^n \end{pmatrix}$ with $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \Omega$ and $w = \begin{pmatrix} w^1 \\ \vdots \\ w^n \end{pmatrix} \in \mathbb{R}^n$. Our second SDE model is as follows

$$dX\left(t\right) = u\left(X(t), t\right) dt + \sigma dW\left(t\right). \tag{4.62}$$

In both models, the drift represents the control function. However, in (4.61), the dependence of the control on the state X is explicitly given and the controls sought are $v, w : [0, T] \to \mathbb{R}^n$. We refer to this case as the open-loop control mechanism since there is no feedback of the actual position of the stochastic process by space dependency of control functions. Notice that, from the SDE point of view, this mechanism includes the linear and bilinear control cases that appear in many applications. On the other hand, in (4.62), the control function $u : Q \to \mathbb{R}^n$ is intended to define a closed-loop control mechanism for the stochastic process. In contrast to the open-loop process above, the position of the stochastic process, in addition to time, is also used to design a control strategy.

For the open-loop problem, the FP equation in its weak formulation is given by

$$\begin{aligned} &\int_\Omega \left(f'\left(x,t\right)\varphi\left(x\right) + \frac{\sigma^2}{2}\nabla f\left(x,t\right)\cdot\nabla\varphi\left(x\right) + \nabla\left(\left(v\left(t\right) + x\circ w\left(t\right)\right) f\left(x,t\right)\right)\varphi\left(x\right)\right) dx = 0 \\ &f\left(\cdot, 0\right) = f_0 \end{aligned} \tag{4.63}$$

for all $\varphi \in H_0^1\left(\Omega\right)$ and for almost all $t \in (0, T)$ where the dot $\cdot$ denotes the Euclidean scalar product, ∇ denotes the gradient with respect to the Euclidean scalar product in $\mathbb{R}^n$, the divergence of a vector-valued function $y = \begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix}$ is denoted with $\nabla y := \sum_{i=1}^n \frac{\partial}{\partial x_i} y^i$ and the partial derivative with respect to t is denoted with $f' := \frac{\partial}{\partial t} f$.

The admissible control sets are given as follows

$$V_{ad}^i := \left\{ v \in L^q\left(0,T\right) \mid v\left(t\right) \in K_V^i \text{ a.e. in } \left(0,T\right)\right\}$$

and

$$W_{ad}^i := \left\{ w \in L^q\left(0,T\right) \mid w\left(t\right) \in K_W^i \text{ a.e. in } \left(0,T\right)\right\},$$

$i \in \{1, ..., n\}$, $q \geq 2$ where K_V^i, K_W^i are compact subsets of $\mathbb{R}$. Hence we have that

$$K_V := K_V^1 \times ... \times K_V^n \text{ and } K_W := K_W^1 \times ... \times K_W^n$$

and

$$V_{ad} := V_{ad}^1 \times ... \times V_{ad}^n \text{ and } w \in W_{ad} := W_{ad}^1 \times ... \times W_{ad}^n.$$

We remark that for any function $y \in \left(L^q\left(0,T\right)\right)^n$, we have that $\|y\|_{L^q(0,T)}^q := \sum_{i=1}^n \|y^i\|_{L^q(0,T)}^q$ and analogously for any function $y \in \left(L^\infty\left(0,T\right)\right)^n$, we have that $\|y\|_{L^\infty(0,T)} := \max_{i=1,\dots,n} \|y^i\|_{L^\infty(0,T)}$.

Remark. We remark that (4.63) describes a decoupled stochastic process. This means that this stochastic process decomposes into n one dimensional stochastic processes that do not depend on each other. While the discussion in the following is for a general case, this decoupling can be used to accelerate numerical calculations. Actually, we have to solve n one dimensional equations of the form

$$f_i'\left(t,x_i\right) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2} f_i\left(t,x_i\right) + \frac{\partial}{\partial x_i}\left(\left(v_i\left(t\right) + x_i w_i\left(t\right)\right) f_i\left(t,x_i\right)\right) = 0$$

for $i \in \{1, ..., n\}$. This is computationally advantageous compared with solving the total FP equation

$$f'\left(t,x\right) - \frac{\sigma^2}{2}\Delta f\left(t,x\right) + \nabla\left(\left(v\left(t\right) + x\circ w\left(t\right)\right) f\left(t,x\right)\right) = 0$$

in the n dimensional space Ω with respect to the scaling of the number of mesh grid points. The solution of the total FP equation is then assembled by $f(t,x) = \prod_{i=1}^{n} f_i(t,x_1)$ since by putting this approach into the total FP equation, we obtain

$$\begin{aligned}
&f'(t,x) - \frac{\sigma^2}{2}\Delta f(t,x) + \nabla\left(\left(v(t) + x \circ w(t)\right) f(t,x)\right) \\
&= \sum_{i=1}^{n}\left(\left(f_i'(t,x_i) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial x_i^2} f_i(t,x_i)\right)\prod_{j=1,j\neq i}^{n} f_j(t,x_j) + \frac{\partial}{\partial x_i}\left(\left(v_i(t) + x_i w_i(t)\right)\prod_{j=1}^{n} f_j(t,x_j)\right)\right) \\
&= \sum_{i=1}^{n}\left(\frac{\partial}{\partial x_i}\left(\left(v_i(t) + x_i w_i(t)\right)\prod_{j=1}^{n} f_j(t,x_j)\right) - \left(\frac{\partial}{\partial x_i}\left(\left(v_i(t) + x_i w_i(t)\right) f_i(t,x_i)\right)\right)\prod_{j=1,j\neq i}^{n} f_j(t,x_j)\right) \\
&= \sum_{i=1}^{n}\left(w_i(t)\prod_{j=1}^{n} f_j(t,x_j) + \left(v_i(t) + x_i w_i(t)\right)\frac{\partial}{\partial x_i} f_i(t,x_i)\prod_{j=1,j\neq i}^{n} f_j(t,x_j)\right) \\
&- \sum_{i=1}^{n}\left(w_i(t)\prod_{j=1}^{n} f_j(t,x_j) + \left(v_i(t) + x_i w_i(t)\right)\frac{\partial}{\partial x_i} f_i(t,x_i)\prod_{j=1,j\neq i}^{n} f_j(t,x_j)\right) \\
&= 0
\end{aligned}$$

with the product rule [3, IV Theorem 1.6] and using that each f_i, $i \in \{1,...,n\}$, solves the corresponding one dimensional equation.

An essential result for our analysis is the following theorem that states a specific boundedness result. For our purpose, we remark that, since the drift is differentiable with respect to x, we have

$$\begin{aligned}
&\nabla\left(\left(v(t) + x \circ w(t)\right) f(x,t)\right) = \sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(\left(v^i(t) + x_i w^i(t)\right) f(x,t)\right) \\
&= \sum_{i=1}^{n}\left(v^i(t) + x_i w^i(t)\right)\frac{\partial}{\partial x_i} f(x,t) + w^i(t) f(x,t) = \left(v(t) + x \circ w(t)\right)\cdot\nabla f(x,t) + \sum_{i=1}^{n} w^i(t) f(x,t)
\end{aligned} \tag{4.64}$$

with the product rule [2, 4.25]. This fact motivates the proof of the following theorem.

Theorem 43. *Consider the following parabolic problem*

$$\begin{aligned}
\left(y',\varphi\right) + a\left(\nabla y, \nabla\varphi\right) + \left(b\cdot\nabla y, \varphi\right) + \left(cy,\varphi\right) &= \left(h,\varphi\right) \text{ in } \Omega\times(0,T) \\
y &= 0 \text{ on } \partial\Omega\times[0,T] \\
y &= y_0 \text{ on } \Omega\times\{0\}
\end{aligned} \tag{4.65}$$

for all $\varphi \in H_0^1(\Omega)$ and almost all $t \in (0,T)$ where $(\cdot,\cdot)$ is the $L^2(\Omega)$-scalar product, $a, T > 0$, $b \in (L^\infty(Q))^n$, $c \in L^\infty(Q)$, $y_0 \in L^\infty(\Omega)$ and $h \in L^q(Q)$ where $q > \frac{n}{2} + 1$ for $n \geq 2$ and $q \geq 2$ for $n = 1$ with n the dimension of the bounded domain Ω such that $y \in L^2\left(0,T; H_0^1(\Omega)\right) \cap L^\infty\left(0,T; L^2(\Omega)\right)$ solves (4.65). Then, we have that

$$\|y\|_{L^\infty(Q)} \leq C\left(\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(\Omega)}\right)$$

where $C := C\left(\Omega, a, T, \|b\|_{L^\infty(Q)}, \|c\|_{L^\infty(Q)}\right) > 0$.

Proof. The proof uses Theorem 64. First, we define the bilinear map

$$B(y,\varphi;t) := a\int_\Omega \nabla y(x,t)\cdot\nabla\varphi(x) + b(x,t)\cdot\nabla y(x,t)\,\varphi(x) + c(x,t)\,y(x,t)\,\varphi(x)\,dx.$$

In order to apply Theorem 64, we define an auxiliary problem where the corresponding bilinear map fulfills the coercivity condition.

For this purpose, we set $\hat{y} := \mathrm{e}^{-\eta t} y$ for any $\eta \geq 0$ where y solves (4.65). Then, we multiply both sides of (4.65) with $\mathrm{e}^{-\eta t}$ and obtain

$$\int_\Omega \mathrm{e}^{-\eta t} y'(x,t)\,\varphi(x)\,dx + \mathrm{e}^{-\eta t} B\left(y(\cdot,t),\varphi\right) = \int_\Omega \mathrm{e}^{-\eta t} h(x,t)\,\varphi(x)\,dx$$

from which we obtain, by inserting the definition of $\hat{y}$, the following

$$\int_\Omega \hat{y}'(x,t)\,\varphi(x)\,dx + B\left(\hat{y}(\cdot,t),\varphi;t\right) + \int_\Omega \eta\hat{y}(x,t)\,\varphi(x)\,dx = \int_\Omega \hat{h}(x,t)\,\varphi(x)\,dx \tag{4.66}$$

where $\hat{h} := h\mathrm{e}^{-\eta\cdot} \in L^q(Q)$ because of the boundedness of $t \mapsto \mathrm{e}^{-\eta t}$ over $[0,T]$. Now, from (4.66), we have that

$$\int_\Omega -\hat{y}'(x,t)\,\varphi(x)\,dx + \hat{B}\left(\hat{y}(\cdot,t),\varphi\right) = \int_\Omega \hat{h}(x,t)\,\varphi(x)\,dx, \tag{4.67}$$

which is uniquely solvable with $\hat{y} = 0$ on $\partial\Omega \times [0,T]$ and $\hat{y} = \mathrm{e}^{-\eta 0} y_0 = y_0$ on $\Omega \times \{0\}$ where

$$\hat{B}\left(\hat{y},\varphi;t\right) := B\left(\hat{y}(\cdot,t),\varphi;t\right) + \int_\Omega \eta\hat{y}(x,t)\,\varphi(x)\,dx,$$

see [45, Section 7.1 Thoerem 3] with $\hat{y} \in L^2\left(0,T;H_0^1(\Omega)\right) \cap L^\infty\left(0,T;L^2(\Omega)\right)$ since $t \mapsto \mathrm{e}^{-\eta t}$ is bounded over $[0,T]$. Then we have the following result

$$\begin{aligned} &a\|\hat{y}(\cdot,t)\|^2_{H_0^1(\Omega)} \\ &= a\int_\Omega \nabla\hat{y}(x,t)\cdot\nabla\hat{y}(x,t)\,dx = \hat{B}\left(\hat{y},\hat{y};t\right) - \int_\Omega b(x,t)\cdot\nabla\hat{y}(x,t)\,\hat{y}(x,t) + \left(c(x,t)+\eta\right)\hat{y}^2(x,t)\,dx. \end{aligned} \tag{4.68}$$

From (4.68), we obtain

$$\begin{aligned} &a\|\hat{y}(\cdot,t)\|^2_{H_0^1(\Omega)} + \int_\Omega \eta\hat{y}^2(x,t)\,dx \\ &= \hat{B}\left(\hat{y},\hat{y};t\right) - \int_\Omega b(x,t)\cdot\nabla\hat{y}(x,t)\,\hat{y}(x,t) + c(x,t)\,\hat{y}^2(x,t)\,dx \\ &\leq \hat{B}\left(\hat{y},\hat{y};t\right) + \|b\|_{L^\infty(Q)}\left(\epsilon\int_\Omega \nabla\hat{y}(x,t)\cdot\nabla\hat{y}(x,t)\,dx + \frac{n}{4\epsilon}\int_\Omega \hat{y}^2(x,t)\,dx\right) + \|c\|_{L^\infty(Q)}\int_\Omega \hat{y}^2(x,t)\,dx \end{aligned} \tag{4.69}$$

with

$$\begin{aligned} &|\int_\Omega b(x,t)\cdot\nabla\hat{y}(x,t)\,\hat{y}(x,t)\,dx| \leq \|b\|_{L^\infty(Q)}\sum_{i=1}^n\int_\Omega |\frac{\partial}{\partial x_i}\hat{y}_i(x,t)\,||\hat{y}(x,t)\,|dx \\ &\leq \|b\|_{L^\infty(Q)}\sum_{i=1}^n\left(\int_\Omega \epsilon|\frac{\partial}{\partial x_i}\hat{y}_i(x,t)\,|^2 + \frac{1}{4\epsilon}|\hat{y}(x,t)\,|^2\right)dx \\ &= \|b\|_{L^\infty(Q)}\int_\Omega \epsilon\nabla\hat{y}(x,t)\cdot\nabla\hat{y}(x,t) + \frac{n}{4\epsilon}|\hat{y}(x,t)\,|^2dx \end{aligned}$$

where we use the Cauchy inequality, see [45, page 622], for $\epsilon > 0$.

We assume that $\|b\|_{L^\infty(Q)} \neq 0$ and choose $\epsilon := \frac{a}{2\|b\|_{L^\infty(Q)}}$. From (4.69), we have that

$$\frac{a}{2}\|\hat{y}(\cdot,t)\|^2_{H_0^1(\Omega)} + \int_\Omega \eta\hat{y}^2(x,t)\,dx \leq \hat{B}\left(\hat{y},\hat{y};t\right) + \frac{\|b\|^2_{L^\infty(Q)} + 2a\|c\|_{L^\infty(Q)}}{2a}\int_\Omega \hat{y}^2(x,t)\,dx$$

which gives

$$a\|\hat{y}\left(\cdot,t\right)\|^2_{H^1_0(\Omega)} \leq \hat{B}\left(\hat{y},\hat{y};t\right) \tag{4.70}$$

for $\eta \geq \frac{\|b\|^2_{L^\infty(Q)}+2a\|c\|_{L^\infty(Q)}}{2a}$.

If $\|b\|_{L^\infty(Q)} = 0$, then from (4.69) we obtain that (4.70) holds for $\eta \geq \|c\|_{L^\infty(Q)}$.

Consequently, we choose

$$\eta \geq \frac{\|b\|^2_{L^\infty(Q)} + 2a\|c\|_{L^\infty(Q)}}{2a}.$$

Since it holds that $\hat{B}\left(-k,\varphi;t\right) \leq 0$ for $k \geq 0$ if $\varphi \geq 0$ for any $t \in [0,T]$, we can apply Theorem 64 to the following parabolic problem

$$\begin{aligned} \left(\hat{y}',\varphi\right) + \hat{B}\left(\hat{y},\varphi;t\right) = \left(\hat{h},\varphi\right) &\text{ in } \Omega\times(0,T)\\ \hat{y} = 0 &\text{ on } \partial\Omega\times[0,T]\\ \hat{y} = y_0 &\text{ on } \Omega\times\{0\} \end{aligned}$$

and obtain

$$\|\hat{y}\|_{L^\infty(Q)} \leq \hat{C}\|\hat{h}\|_{L^\infty(Q)} + \|y_0\|_{L^\infty(\Omega)} \tag{4.71}$$

for a constant $\hat{C} > 0$. Thus, from (4.71) we have

$$\begin{aligned} \|y\|_{L^\infty(Q)} = \|\mathrm{e}^{\eta\cdot}\hat{y}\|_{L^\infty(Q)} \leq \mathrm{e}^{\eta T}\|\hat{y}\|_{L^\infty} &\leq \mathrm{e}^{\eta T}\hat{C}\|\hat{h}\|_{L^2(Q)} + \mathrm{e}^{\eta T}\|y_0\|_{L^\infty(\Omega)}\\ &\leq \hat{C}\mathrm{e}^{\eta T}\|h\|_{L^2(Q)} + \mathrm{e}^{\eta T}\|y_0\|_{L^\infty(\Omega)} \end{aligned}$$

where $C := \max\left(\hat{C}\mathrm{e}^{\eta T},\mathrm{e}^{\eta T}\right)$. □

The FP equation (4.63) with $v \in V_{ad}$ and $w \in W_{ad}$, considered as a parabolic equation in the framework of Theorem 43, is uniquely solvable with $f \in L^2\left(0,T;H^1_0\left(\Omega\right)\right)$ and $f' \in L^2\left(0,T;H^{-1}\left(\Omega\right)\right)$, see [1, Theorem 2.14], [45, Section 7.1, Theorem 3 and Theorem 4] for $f_0 \in L^2\left(\Omega\right)$.

However, in order to obtain the desired regularity, we require $f_0 \in L^\infty\left(\Omega\right) \cap H^1_0\left(\Omega\right)$ such that we have $f \in L^2\left(0,T;H^2\left(\Omega\right)\right) \cap L^\infty\left(0,T;H^1_0\left(\Omega\right)\right)$, see [45, Section 7.1 Theorem 5] where the proof for the relevant part of [45, Section 7.1 Theorem 5] is also applicable with our assumptions. From these results and Theorem 43, we conclude the following. A similar result can be found in [13].

Theorem 44. *For the solution to (4.63), it holds*

$$\|f\|_{L^\infty(Q)} \leq C\|f_0\|_{L^\infty(\Omega)} \tag{4.72}$$

where $C := C\left(\Omega,\sigma,T,\|v + x\circ w\|_{L^\infty(Q)},\|\sum_{i=1}^n w^i\|_{L^\infty(Q)}\right) > 0$.

Proof. In view of (4.64), we choose $b\left(x,t\right) = v\left(t\right) + x\circ w\left(t\right)$, $c = \sum_{i=1}^n w^i\left(t\right)$ and apply Theorem 43 since in our framework it is assumed that $f_0 \in L^\infty\left(\Omega\right) \cap H^1_0\left(\Omega\right)$. Thus it holds $f \in L^2\left(0,T;H^2\left(\Omega\right)\right) \cap L^\infty\left(0,T;H^1_0\left(\Omega\right)\right)$. □

Having completed our discussion on the FP model, we formulate our optimal control problem corresponding to (4.61) as follows

$$\begin{aligned} &\min_{f,v,w} J\left(f,v,w\right) := \int_0^T\int_\Omega G\left(v,w\right)\left(x,t\right)f\left(x,t\right)dxdt + \int_\Omega F\left(x\right)f\left(x,T\right)dx\\ &\text{s.t. } \int_\Omega\left(f'\left(x,t\right)\varphi\left(x\right) + \frac{\sigma^2}{2}\nabla f\left(x,t\right)\cdot\nabla\varphi\left(x\right) + \nabla\left(\left(v\left(t\right) + x\circ w\left(t\right)\right)f\left(x,t\right)\right)\varphi\left(x\right)\right)dx = 0\\ &\qquad \text{a.e. in } (0,T) \text{ for all } \varphi \in H^1_0\left(\Omega\right)\\ &f\left(\cdot,0\right) = f_0\\ &v \in V_{ad},\ w \in W_{ad} \end{aligned} \tag{4.73}$$

with

$$G\left(v,w\right)\left(x,t\right) := -A\left(x,t\right) + \alpha g_{s_1}\left(v\left(t\right)\right) + \beta g_{s_2}\left(w\left(t\right)\right)$$

where $s_1, s_2 \geq 0$, $A \in L^q\left(0,T;W^{1,q}\left(\Omega\right)\right) \cap L^\infty\left(Q\right)$ a non-negative function, $q \geq 2$ for $n=1$ and $q > \frac{n}{2}+1$ for $n \geq 2$, $\alpha, \beta \geq 0$. The lower boundedness of the cost functional is ensured by the boundedness of A and f, see (4.72). The functions that determine the costs of the controls are given by

$$g_s : \mathbb{R}^n \to \mathbb{R},\ z \mapsto g_s\left(z\right) := \sum_{i=1}^{n} g_s^i\left(z^i\right).$$

The functions are assumed to be lower semi-continuous for all $z \in \mathbb{R}^n$ with z^i is the i-th component of z, $g_s^i : \mathbb{R} \to \mathbb{R}$, $z^i \mapsto g_s^i\left(z^i\right)$, such that $G\left(v,w\right) \in L^\infty\left(Q\right)$ and $F \in L^\infty\left(\Omega\right) \cap H_0^1\left(\Omega\right)$.

Remark 45. The existence of a solution to (4.73) can be shown as in [46, 5 Existence of optimal controls]. However, we need to modify the arguments since we multiply the state with functions of the controls. According to [46, Remark 5.1], the control-to-state map of our FP equation is sequentially continuous as a map from $V_{ad} \times W_{ad}$ to $L^2\left(Q\right)$. This means that any weakly converging sequence of the controls results in a strongly converging subsequence of the state. Assume that $\left(v_k, w_k\right)_{k\in K}$, $\left(f_k\right)_{k\in K}$, $K \subseteq \mathbb{N}$ is a minimizing sequence for the functional J where $\left(\bar{v}, \bar{w}\right)$ is the weak limit of $\left(v_k, w_k\right)_{k\in K}$ and $\bar{f}$ is the limit of the sequence $\left(f_k\right)_{k\in K}$. If we further assume that the functions g_s^i are convex and Lipschitz continuous, then the functional

$$\int_Q \bar{f} G\left(v,w\right)\left(x,t\right) dxdt$$

is convex since $\bar{f} \geq 0$ and continuous as a map from $V_{ad} \times W_{ad}$ to $\mathbb{R}$ since it holds

$$\begin{aligned}&\int_Q \bar{f}|g_s^i\left(\zeta^1\left(t\right)\right) - g_s^i\left(\zeta^2\left(t\right)\right)|dxdt \leq |\Omega|LC\|f_0\|_{L^\infty(\Omega)} \int_0^T |\zeta^1\left(t\right) - \zeta^2\left(t\right)|dt \\ &= L|\Omega|C\|f_0\|_{L^\infty(\Omega)}\|\zeta^1 - \zeta^2\|_{L^1(0,T)} \leq c\|\zeta^1 - \zeta^2\|_{L^q(0,T)}\end{aligned}$$

for any function $\zeta^1, \zeta^2 \in L^q\left(0,T\right)$, see Theorem 44 and the embedding [1, Theorem 2.14]. Consequently the functional $\int_Q \bar{f} G\left(v,w\right)\left(x,t\right) dxdt$ is weakly lower semi-continuous, see [95, Theorem 2.12]. Then we have that with the calculation rules for a sum of lim inf [43, Theorem 3.127] that the following holds

$$\begin{aligned}&\liminf_{k\to\infty} \int_Q G\left(v_k,w_k\right)\left(x,t\right) f_k\left(x,t\right) dxdt \\ &= \liminf_{k\to\infty} \left(\int_Q G\left(v_k,w_k\right)\left(x,t\right) \bar{f}\left(x,t\right) dxdt + \int_Q G\left(v_k,w_k\right)\left(x,t\right)\left(f_k - \bar{f}\right)\left(x,t\right) dxdt\right) \\ &\geq \liminf_{k\to\infty} \left(\int_Q G\left(v_k,w_k\right)\left(x,t\right) \bar{f}\left(x,t\right) dxdt\right) + \liminf_{k\to\infty} \left(\int_Q G\left(v_k,w_k\right)\left(x,t\right)\left(f_k - \bar{f}\right)\left(x,t\right) dxdt\right) \\ &= \liminf_{k\to\infty} \left(\int_Q G\left(v_k,w_k\right)\left(x,t\right) \bar{f}\left(x,t\right) dxdt\right) + \lim_{k\to\infty} \left(\int_Q G\left(v_k,w_k\right)\left(x,t\right)\left(f_k - \bar{f}\right)\left(x,t\right) dxdt\right) \\ &\geq \int_Q G\left(\bar{v},\bar{w}\right)\left(x,t\right) \bar{f}\left(x,t\right) dxdt\end{aligned}$$

since the lim exists being equal to zero, see

$$\lim_{k\to\infty} |\int_Q G\left(v_k,w_k\right)\left(x,t\right)\left(f_k - \bar{f}\right)\left(x,t\right) dxdt| \leq d \lim_{k\to\infty} \|f_k - \bar{f}\|_{L^2(Q)} = 0,$$

$d > 0$ due to the boundedness of G and the lim inf equals the lim, see [3, Theorem 5.7].

Our second FP optimal control problem corresponds to the stochastic process (4.62) and is given by

$$\begin{aligned}
&\min_{f,u} J(f,u) := \int_0^T \int_\Omega G(u)(x,t)\, f(x,t)\, dxdt + \int_\Omega F(x)\, f(x,T)\, dx \\
&\text{s.t. } \int_\Omega \left(f'(x,t)\,\varphi(x) + \frac{\sigma^2}{2}\nabla f(x,t)\cdot\nabla\varphi(x) + \nabla\cdot\left(u(x,t)\, f(x,t)\right)\varphi(x)\right) dx = 0 \\
&\qquad \text{a.e. in } (0,T) \text{ for all } \varphi \in H_0^1(\Omega) \\
&f(\cdot,0) = f_0 \\
&u \in U_{ad}
\end{aligned} \tag{4.74}$$

where we choose the following admissible set of controls

$$U_{ad} := \{u \in (L^q(Q))^n \mid u(x,t) \in K_U \text{ a.e. on } Q\}$$

with $u_{\min}, u_{\max} \in \mathbb{R}$, $u_{\min} < u_{\max}$, $q \geq 2$ and $K_U := [u_{\min}, u_{\max}]^n$. For G, it holds the same as in the case of (4.73) and is specified later.

Notice that, with $u \in (L^q(Q))^n$, $2 < q \leq \infty$, the well-posedness of the forward FP problem can be shown, see [46], and also in this case, an L^∞ bound for the PDF, analogous to Theorem 44, can be shown based on [13, Theorem 3.1]. The discussion of existence of an optimal solution is the same as in the case (4.73) above.

In the following section, we discuss the PMP characterization of a solution to (4.73) and (4.74).

4.3 The characterization by the Pontryagin maximum principle

Next, we discuss the characterization of a solution to (4.73) in the PMP framework. We define the Hamiltonian function $H : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \times K_V \times K_W \times \mathbb{R}^n \to \mathbb{R}$ as follows

$$H(x,t,f,v,w,\zeta) := G(v,w)\, f + \zeta \cdot (v + x \circ w)\, f.$$

We remark that if f, v, w, ζ are functions, we write short

$$H(x,t,f,v,w,\zeta) := H(x,t,f(x,t),v(x,t),w(x,t),\zeta(x,t)).$$

Later the place holder $\zeta \in \mathbb{R}^n$ will be filled with the space derivative of the solution to the adjoint equation that is given by

$$\begin{aligned}
&\int_\Omega \left(-p'(x,t)\,\varphi(x) + \frac{\sigma^2}{2}\nabla p(x,t)\cdot\nabla\varphi(x) - (v(t) + x\circ w(t))\cdot\nabla p(x,t)\,\varphi(x)\right) dx \\
&= \int_\Omega \left(G(v,w)(x,t)\,\varphi(x)\right) dx
\end{aligned} \tag{4.75}$$

for all $\varphi \in H_0^1(\Omega)$ with $p(\cdot,T) = F(\cdot)$. The adjoint equation (4.75) is uniquely solvable. This is shown as follows. After a time transformation $\tilde{t} := T - t$ and since $G(v,w) \in L^q(Q)$, see [1, Theorem 2.14], for which we consider the boundedness of v, w and the measurability of G, see Lemma 51, for any $v \in V_{ad}$ and $w \in W_{ad}$, we have the existence of a unique solution to (4.75) analogously to (4.63). Furthermore, by the proof of [45, Section 7.1 Theorem 5], we have that $p \in L^2\left(0,T;H^2(\Omega)\right) \cap L^\infty\left(0,T;H_0^1(\Omega)\right)$ and $p' \in L^2\left(0,T;L^2(\Omega)\right)$. Therefore, we have the following theorem.

Theorem 46. *For the solution to (4.75), it holds*

$$\|p\|_{L^\infty(Q)} \leq C\left(\|G(v,w)\|_{L^q(Q)} + \|F\|_{L^\infty(\Omega)}\right)$$

for $C := C\left(\Omega,\sigma,T,\max_{i=1,\ldots,n}\|v^i + x_i w^i\|_{L^\infty(Q)}\right) > 0.$

Proof. As $F \in L^q(Q)$, due to the pointwise boundedness of v, w and $\left(v^i(t) + x_i w^i(t)\right) \in L^\infty(Q)$ for all $i \in \{1, ..., n\}$, we can apply Theorem 43 to obtain the desired result after a transformation of time according to $\tilde{t} := T - t$. □

Additionally, we have that $p \in L^q\left(0, T; W_0^{1,q}\right)$, see [14].

Notice that in the adjoint FP problem (4.75) the solution of the forward FP problem does not appear. This is due to the linearity of our cost functional with respect to the state f.

The next step in order to characterize a solution to (4.73) in the PMP framework is the following lemma that provides a direct relation between the values of cost functional at different triples (f, v, w) and the values of the corresponding Hamiltonian. We have that (f_1, v_1, w_1) solves (4.63) for f_1 instead of f, we write $f \leftarrow f_1$, with v_1 instead of v, we write $v \leftarrow v_1$ and with w_1 instead of w, we write $w \leftarrow w_1$. We remark that the PMP characterization simplifies due to the fact that the cost functional only depends linearly on the state variable f. A consequence is that we can perform the corresponding proofs without the definition of an intermediate adjoint as it is done in Chapter 2, Chapter 3 or [81, 21] for instance.

Lemma 47. *Let (f_1, v_1, w_1) solve (4.63) for $(f, v, w) \leftarrow (f_1, v_1, w_1)$ and let (f_2, v_2, w_2) solve (4.63) for $(f, v, w) \leftarrow (f_2, v_2, w_2)$. Then, we have that*

$$J(f_1, v_1, w_1) - J(f_2, v_2, w_2) = \int_0^T \int_\Omega H(x, t, f_2, v_1, w_1, \nabla p_1) - H(x, t, f_2, v_2, w_2, \nabla p_1)\, dxdt$$

where p_1 is given by (4.75) for $v \leftarrow v_1$ and $w \leftarrow w_1$ and J is defined in (4.73).

Proof. In order to save notational effort, we drop the functions' dependency with respect to x, t. We have

$$\begin{aligned} &J(f_1, v_1, w_1) - J(f_2, v_2, w_2) \\ &= \int_0^T \int_\Omega G(v_1, w_1) f_1 dxdt + \int_\Omega F f_1(\cdot, T)\, dx - \int_0^T \int_\Omega G(v_2, w_2) f_2 dxdt - \int_\Omega F f_2(\cdot, T)\, dx \\ &= \int_0^T \int_\Omega G(v_1, w_1) f_2 + G(v_1, w_1)(f_1 - f_2) - G(v_2, w_2) f_2 dxdt + \int_\Omega F(f_1 - f_2)(\cdot, T)\, dx \end{aligned} \tag{4.76}$$

and

$$\begin{aligned} &\int_0^T \int_\Omega G(v_1, w_1)(f_1 - f_2) \\ &= \int_0^T \int_\Omega -p_1'(f_1 - f_2) + \frac{\sigma^2}{2} \nabla p_1 \cdot \nabla(f_1 - f_2) - (v_1 + x \circ w_1) \cdot \nabla p_1 (f_1 - f_2)\, dx \\ &= \int_0^T \int_\Omega \left(f_1' - f_2'\right) p_1 + \frac{\sigma^2}{2}(\nabla f_1 - \nabla f_2) \cdot \nabla p_1 + (\nabla((v_1 + x \circ w_1) f_1) - \nabla((v_1 + x \circ w_1) f_2)) p_1 dxdt \\ &- \int_\Omega F(f_1 - f_2)(\cdot, T)\, dx \\ &= \int_0^T \int_\Omega \nabla((v_2 + x \circ w_2) f_2) p_1 - \nabla((v_1 + x \circ w_1) f_2) p_1 dxdt - \int_\Omega F(f_1 - f_2)(\cdot, T)\, dx \end{aligned} \tag{4.77}$$

by partial integration with respect to t [95, Satz 3.11], partial integration with respect to x , see Lemma

53 (third line) and with (4.63) (fifth line). Combining (4.76) and (4.77), we obtain

$$\begin{aligned}
&J\left(f_1,v_1,w_1\right)-J\left(f_2,v_2,w_2\right)\\
&=\int_0^T\int_\Omega G\left(v_1,w_1\right)f_2-\nabla\left(\left(v_1+x\circ w_1\right)f_2\right)p_1-G\left(v_2,w_2\right)f_2+\nabla\left(\left(v_2+x\circ w_2\right)f_2\right)p_1dxdt\\
&=\int_0^T\int_\Omega G\left(v_1,w_1\right)f_2+\left(\left(v_1+x\circ w_1\right)f_2\right)\cdot\nabla p_1-G\left(v_2,w_2\right)f_2-\left(\left(v_2+x\circ w_2\right)f_2\right)\cdot\nabla p_1dxdt\\
&=\int_0^T\int_\Omega H\left(x,t,f_2,v_1,w_1,\nabla p_1\right)-H\left(x,t,f_2,v_2,w_2,\nabla p_1\right)dxdt.
\end{aligned}$$

□

The next step for the PMP characterization of a solution to (4.73) is to introduce the concept of needle variation. For this purpose, we define the needle variation for any $\tilde{v}\in V_{ad}$ and $\tilde{w}\in W_{ad}$ with $t_0\in(0,T)$ and with $S_k\left(t_0\right)$, a ball centered in t_0 and for its measure $|S_k\left(t_0\right)|$ it holds that $\lim_{k\to\infty}|S_k\left(t_0\right)|=0$, as follows

$$v_k\left(t\right):=\begin{cases}\tilde{v}\left(t\right) & \text{if } t\in(0,T)\setminus S_k\left(t_0\right)\\ v & \text{if } t\in S_k\left(t_0\right)\cap(0,T)\end{cases},\quad w_k\left(t\right):=\begin{cases}\tilde{w}\left(t\right) & \text{if } t\in(0,T)\setminus S_k\left(t_0\right)\\ w & \text{if } t\in S_k\left(t_0\right)\cap(0,T)\end{cases}$$

where $v\in K_V$ and $w\in K_W$.

Now, we can state the PMP characterization of an optimal control to (4.73).

Theorem 48. *Let $\left(\bar{f},\bar{v},\bar{w}\right)$ be a solution to (4.73). Then it holds that*

$$\int_\Omega H\left(x,t,\bar{f},\bar{v},\bar{w},\nabla\bar{p}\right)dx=\min_{v\in K_V,w\in K_W}\int_\Omega H\left(x,t,\bar{f},v,w,\nabla\bar{p}\right)dx \tag{4.78}$$

for almost all $t\in(0,T)$ where $\bar{p}$ is the solution to (4.75) for $v\leftarrow\bar{v}$ and $w\leftarrow\bar{w}$.

Proof. Since $v_k\in V_{ad}$ and $w_k\in W_{ad}$ for all $k\in\mathbb{N}$, we have with Lemma 47 that for any $k\in\mathbb{N}$ the following holds

$$\begin{aligned}
0&\leq\frac{1}{|S_k\left(t_0\right)|}\left(J\left(f_k,v_k,w_k\right)-J\left(\bar{f},\bar{v},\bar{w}\right)\right)\\
&=\frac{1}{|S_k\left(t_0\right)|}\left(\int_0^T\int_\Omega H\left(x,t,\bar{f},v_k,w_k,\nabla p_k\right)-H\left(x,t,\bar{f},\bar{v},\bar{w},\nabla p_k\right)dxdt\right)\\
&=\frac{1}{|S_k\left(t_0\right)|}\left(\int_{S_k(t_0)}\int_\Omega H\left(x,t,\bar{f},v,w,\nabla\bar{p}\right)-H\left(x,t,\bar{f},\bar{v},\bar{w},\nabla\bar{p}\right)dxdt\right)\\
&+\frac{1}{|S_k\left(t_0\right)|}\left(\int_{S_k(t_0)}\int_\Omega\left(\nabla p_k-\nabla\bar{p}\right)\cdot\left(v+x\circ w\right)\bar{f}+\left(\nabla\bar{p}-\nabla p_k\right)\cdot\left(\bar{v}+x\circ\bar{w}\right)\bar{f}dxdt\right)\\
&=\frac{1}{|S_k\left(t_0\right)|}\left(\int_{S_k(t_0)}\int_\Omega H\left(x,t,\bar{f},v,w,\nabla\bar{p}\right)-H\left(x,t,\bar{f},\bar{v},\bar{w},\nabla\bar{p}\right)dxdt\right)\\
&-\frac{1}{|S_k\left(t_0\right)|}\left(\int_{S_k(t_0)}\int_\Omega\left(p_k-\bar{p}\right)\nabla\left(\left(v+x\circ w\right)\bar{f}\right)+\left(\bar{p}-p_k\right)\nabla\left(\left(\bar{v}+x\circ\bar{w}\right)\bar{f}\right)dxdt\right)
\end{aligned} \tag{4.79}$$

for all $v\in K_V$ and $w\in K_W$ where $\left(f_k,v_k,w_k\right)$ solves (4.63) for $(f,v,w)\leftarrow\left(f_k,v_k,w_k\right)$.

Next, we prove that

$$\lim_{k\to\infty}\|p_k-\bar{p}\|_{L^\infty(Q)}=0.$$

We subtract (4.75) for $v \leftarrow v_k$ and $w \leftarrow w_k$ from (4.75) for $v \leftarrow \bar{v}$ and $w \leftarrow \bar{w}$ and obtain

$$\begin{aligned}&\int_\Omega -\delta p'\left(x,t\right)\varphi\left(x\right)+\frac{\sigma^2}{2}\nabla\delta p\left(x\right)\cdot\nabla\varphi\left(x\right)-\left(\bar{v}\left(t\right)+x\circ\bar{w}\left(t\right)\right)\cdot\nabla\bar{p}\left(x,t\right)\varphi\left(x\right)dx\\&+\int_\Omega\left(v_k\left(t\right)+x\circ w_k\left(t\right)\right)\cdot\nabla p^k\left(x,t\right)\varphi\left(x\right)dx=\int_\Omega\left(G\left(\bar{v},\bar{w}\right)\left(x,t\right)-G\left(v_k,w_k\right)\left(x,t\right)\right)\varphi\left(x\right)dx\end{aligned}$$

where $\delta p \coloneqq \bar{p} - p^k$ and thus

$$\begin{aligned}&\int_\Omega -\delta p'\left(x,t\right)\varphi\left(x\right)+\frac{\sigma^2}{2}\nabla\delta p\left(x\right)\cdot\nabla\varphi\left(x\right)-\left(v_k\left(t\right)+x\circ w_k\left(t\right)\right)\cdot\nabla\delta p\left(x,t\right)\varphi\left(x\right)dx\\&=\int_\Omega\left(v_k\left(t\right)+x\circ w_k\left(t\right)-\left(\bar{v}\left(t\right)+x\circ\bar{w}\left(t\right)\right)\right)\cdot\nabla\bar{p}\left(x,t\right)\varphi\left(x\right)dx\\&+\int_\Omega\left(G\left(\bar{v},\bar{w}\right)\left(x,t\right)-G\left(v_k,w_k\right)\left(x,t\right)\right)\varphi\left(x\right)dx.\end{aligned}\tag{4.80}$$

From (4.80) and Theorem 46, we have that

$$\lim_{k\to\infty}\|p_k-\bar{p}\|_{L^\infty(Q)}=0$$

if

$$\lim_{k\to\infty}\|\left(v_k+\left(\cdot\right)\circ w_k-\left(\bar{v}+\left(\cdot\right)\circ\bar{w}\right)\right)\cdot\nabla\bar{p}\|_{L^q(Q)}=0$$

and

$$\lim_{k\to\infty}\|G\left(\bar{v},\bar{w}\right)-G\left(v_k,w_k\right)\|_{L^q(Q)}=0.$$

For the first term, we have the following

$$\begin{aligned}&\int_Q|\left(\left(v_k\left(t\right)+x\circ w_k\left(t\right)\right)-\left(\bar{v}\left(t\right)+x\circ\bar{w}\left(t\right)\right)\right)\cdot\nabla\bar{p}\left(x,t\right)|^q dxdt\\&\leq c\int_0^T\int_\Omega\sum_{i=1}^n|\frac{\partial}{\partial x_i}\bar{p}\left(x,t\right)|^q dxdt\leq c\int_0^T\|\nabla\bar{p}\left(\cdot,t\right)\|_{L^2(\Omega)}^q dt\leq c\|\bar{p}\|_{L^q(0,T;W^{1,q}(\Omega))}^q\end{aligned}$$

for a constant $c>0$ due to $p\in L^q\left(0,T;W^{1,q}\left(\Omega\right)\right)$, $q>\frac{n}{2}+1$, see [14]. This means that the function

$$t\mapsto\int_\Omega|\left(v_k\left(t\right)+x\circ w_k\left(t\right)-\left(\bar{v}\left(t\right)+x\circ\bar{w}\left(t\right)\right)\right)\cdot\nabla\bar{p}\left(x,t\right)|^q dx$$

is measurable, see [5, X Theorem 6.7], Lemma 52 and as the product and sum of Lebesgue measurable functions is Lebesgue measurable [36, Proposition 2.1.7] and integrable, see [5, Theorem 6.11, Theorem 6.9]. Consequently we can apply the Average Value Theorem [78, Theorem 51] in order to obtain

$$\begin{aligned}&\lim_{k\to\infty}\|\left(v_k+\left(\cdot\right)\circ w_k-\left(\bar{v}+\left(\cdot\right)\circ\bar{w}\right)\right)\cdot\nabla\bar{p}\|_{L^q(Q)}\\&=\lim_{k\to\infty}\int_{S_k(t_0)}\int_\Omega|\left(v\left(t\right)+x\circ w\left(t\right)-\left(\bar{v}\left(t\right)+x\circ\bar{w}\left(t\right)\right)\right)\cdot\nabla\bar{p}\left(x,t\right)|^q dxdt=0\end{aligned}$$

for almost all $t_0\in(0,T)$. Further, since

$$\|G\left(\bar{v},\bar{w}\right)-G\left(v_k,w_k\right)\|_{L^\infty(Q)}<c$$

for all $\bar{v}\in V_{ad}$ and $\bar{w}\in W_{ad}$, we analogously have that

$$\lim_{k\to\infty}\|G\left(\bar{v},\bar{w}\right)-G\left(v_k,w_k\right)\|_{L^2(Q)}=\lim_{k\to\infty}\int_{S_k(t_0)}\int_\Omega|G\left(\bar{v},\bar{w}\right)-G\left(v,w\right)|^2 dxdt=0$$

for almost all $t_0 \in (0, T)$.

Now, we have that the last line in (4.79) goes to zero for $k \to \infty$ due to

$$\lim_{k\to\infty} \|p_k - \bar{p}\|_{L^\infty(Q)} = 0$$

and due to the Average Value Theorem [78, Theorem 51]. For the application of the Average Value Theorem, we need that the function

$$t \mapsto \int_\Omega \nabla \left(\left(v\left(t\right) + x \circ w\left(t\right) \right) \bar{f}\left(x,t\right) \right) dx$$

is locally integrable. This is the case because we have the boundedness

$$\begin{aligned}&\int_Q |\nabla \left(\left(v\left(t\right) + x \circ w\left(t\right) \right) \bar{f}\left(x,t\right) \right)| dx \\ &= \int_Q | \left(\sum_{i=1}^n w^i\left(t\right) \bar{f}\left(x,t\right) \right) + \left(v\left(t\right) + x \circ w\left(t\right) \right) \cdot \nabla \bar{f}\left(x,t\right) | dxdt \le c\|\bar{f}\|_{L^2\left(0,T;H_0^1(\Omega)\right)}\end{aligned}$$

using the Poincaré inequality [2, 6.7] and the measurability of the function

$$t \mapsto \int_\Omega \nabla \left(\left(v\left(t\right) + x \circ w\left(t\right) \right) \bar{f}\left(x,t\right) \right) dx$$

since we have (4.64), [5, X Theorem 6.7], Lemma 52 and as the product and sum of Lebesgue measurable functions is Lebesgue measurable [36, Proposition 2.1.7]. Because

$$(x,t) \mapsto H\left(x,t,\bar{f},v,w,\nabla\bar{p}\right) - H\left(x,t,f_2,v_2,w_2,\nabla\bar{p}\right)$$

is measurable on Q, see [36, Proposition 2.1.7] and an element of $L^1\left(Q\right)$, we apply Fubini's Theorem [5, Theorem 6.11, Theorem 6.9] and obtain

$$t \mapsto \int_\Omega H\left(x,t,\bar{f},v,w,\nabla\bar{p}\right) - H\left(x,t,\bar{f},\nabla\bar{f},\bar{v},\bar{w},\nabla\bar{p}\right) dx \in L^1\left(0,T\right).$$

Thus we conclude with the Average Value Theorem [78, Theorem 51] the following

$$0 \le \int_\Omega \left(H\left(x,t,\bar{f},v,w,\nabla\bar{p}\right) - H\left(x,t,\bar{f},\bar{v},\bar{w},\nabla\bar{p}\right) \right) dx$$

by taking the limit over k on both sides of the inequality (4.79) for all $v \in K_V$ and $w \in K_W$ and for almost all $t \in (0,T)$, renaming t_0 into t. □

We conclude this section by characterize the solution to (4.74) where we proceed similar to the case discussed above.

The adjoint FP problem for (4.74) is given by

$$\begin{aligned}&\int_\Omega \left(-p'\left(x,t\right)\varphi\left(x\right) + \frac{\sigma^2}{2}\nabla p\left(x,t\right)\cdot\nabla\varphi\left(x\right) - u\left(x,t\right)\cdot\nabla p\left(x,t\right)\varphi\left(x\right) \right) dx \\ &= \int_\Omega \left(G\left(u\right)\left(x,t\right)\varphi\left(x\right) \right) dx, \\ &p\left(\cdot,T\right) = F\left(\cdot\right)\end{aligned} \tag{4.81}$$

for all $\varphi \in H_0^1\left(\Omega\right)$. The proof of a unique solution to (4.81) can be done analogous to (4.75). Further, since $u \in L^\infty\left(Q\right)$, we can apply Theorem 43 to obtain an L^∞ bound for the solution of the adjoint problem that is analogous to that of Theorem 46. Notice that, also in this case, in the adjoint FP problem the state f does not appear.

The PMP Hamiltonian corresponding to (4.74) is given by

$$H\left(x,t,f,u,\zeta\right) \coloneqq \left(G\left(u\right)+\zeta\cdot u\right)f. \tag{4.82}$$

Analogous to the discussion for (4.73) we have that a solution to (4.74) is characterized by the PMP as follows.

Theorem 49. *Let $\left(\bar{f},\bar{u}\right)$ be a solution to* (4.74)*. Then it holds that*

$$H\left(x,t,\bar{f},\bar{u},\nabla\bar{p}\right) = \min_{u\in K_U} H\left(x,t,\bar{f},u,\nabla\bar{p}\right) \tag{4.83}$$

for almost all $(x,t)\in Q$ where $\bar{p}$ is the solution to (4.81) *for $u\leftarrow\bar{u}$.*

Proof. We have that $p\in L^q\left(0,T;W_0^{1,q}\right)$ due to the regularity of the right hand-side of (4.81), see [14] and [83, Proposition 8.35]. By [46, Theorem 3.1], we have that $f\in L^2\left(0,T;H_0^1\left(\Omega\right)\right)$. Then the proofs of Lemma 47 and Theorem 48 can be done analogously, where the corresponding control terms are replaced by the control of (4.74). Going step by step through the mentioned proofs, we can apply the same arguments to the control u. □

Now, notice that our FP equation is uniformly parabolic and in this case the PDF is almost everywhere non-negative. Therefore if it holds

$$\left(G\left(\bar{u}\right)+\nabla\bar{p}\cdot\bar{u}\right) = \min_{u\in K_U}\left(G\left(u\right)+\nabla\bar{p}\cdot u\right) \tag{4.84}$$

for almost all $(x,t)\in Q$ and $\bar{p}$ is the solution to (4.81) for $u\leftarrow\bar{u}$, then (4.83) is fulfilled.

4.4 Numerical schemes

In this section, starting from the PMP characterization of solutions to our FP control problems (4.73) and (4.74), we discuss two numerical solution procedures. In the first case, we implement the iterative sequential quadratic Hamiltonian (SQH) method. For its convergence analysis, we need the following further assumption on the function G. We have to require that the function $G : K_V\times K_W\to\mathbb{R}$, $(v,w)\mapsto G\left(\cdot,\cdot,v,w\right)$ is Lipschitz continuous, that means

$$|G\left(\cdot,\cdot,v_1,w_1\right)-G\left(\cdot,\cdot,v_2,w_2\right)| \leq L\sum_{i=1}^{n}\left(|\left(v_1\right)^i-\left(v_2\right)^i|+|\left(w_1\right)^i-\left(w_2\right)^i|\right)$$

for a Lipschitz constant $L>0$ independent of $(x,t)\in Q$. The need for this requirement comes from the fact that we consider a product of the state and functions of the control in our cost functional, in contrast to the cases in the chapters before. The consequence is that G is on the right hand-side of our adjoint equation.

In the second case, which is (4.74), we exploit the special structure of the resulting optimality system consisting of (4.81) and (4.84) to formulate a non-iterative solution procedure for determining an optimal control. In fact, this is already known in literature to solve the Hamilton-Jacobi-Bellman equation [99], however, for clarity of presentation we call this procedure the direct Hamiltonian (DH) method.

Before discussing the implementation of both methods, we illustrate the numerical approximation of the FP and adjoint FP problems. For this purpose, let us consider the two-dimensional case, $n=2$. We define a sequence of uniform grids $\{\Omega_h\}_{h>0}$ given by

$$\Omega_h = \left\{(x,y)\in\mathbb{R}^2 : (x_i,y_j)=(ih,jh)\,,\ i,j\in\{0,...,N_x\}\right\}\cap\Omega$$

where N_x represents the number of grid points in each direction and h is the spatial mesh size. We assume that Ω is a square and h is chosen such that the boundaries of Ω coincide with the grid points. Let $\delta t = \frac{T}{N_t}$ be the time step and N_t denotes the number of time steps. Define

$$Q_{h,\delta t} = \{(x_i, y_j, t_m) : (x_i, y_j) \in \Omega_h,\ t_m = m\delta t,\ 0 \leq m \leq N_t\}.$$

On this grid, $\phi^m_{i,j}$ represents the value of a grid function in Ω_h at (x_i, y_j) and time t_m.

For the space discretization of the FP equation, we consider a second-order accurate scheme which guarantees positivity of the PDF and, in the case of reflecting barriers, it should provide conservation of the total probability. These are the essential features of the Chang-Cooper (CC) scheme [8, 30, 71].

The first step in the formulation of the CC scheme is to consider the flux form of the FP equation (4.59) by defining the flux

$$F_i\left(x,t\right) = \sum_{j=1}^{2} \partial_{x_j}\left(a_{ij}\left(x,t\right) f\left(x,t\right)\right) - b_i\left(x,t\right) f\left(x,t\right), \qquad i = 1, 2. \tag{4.85}$$

Thus, the FP equation becomes $\partial_t f = \nabla F$. The CC method is a finite-volume scheme where the term ∇F at time t_m is approximated as follows

$$\nabla F = \frac{1}{h}\left\{\left(F^m_{i+\frac{1}{2},j} - F^m_{i-\frac{1}{2},j}\right) + \left(F^m_{i,j+\frac{1}{2}} - F^m_{i,j-\frac{1}{2}}\right)\right\}$$

with $F^m_{i+\frac{1}{2},j}$ and $F^m_{i,j+\frac{1}{2}}$ the representation of the flux in the i-th and j-th direction, respectively. To compute these flux terms, Chang and Cooper proposed to use a linear convex combination of values of f at the cells sharing the same edge. For example, considering the edge between the grid points i, j and $i+1, j$, we have

$$f^m_{i+1/2,j} = \left(1 - \delta^j_i\right) f^m_{i+1,j} + \delta^j_i f^m_{i,j}$$

where the value of δ^j_i is specified below. This approach was motivated in [30] by the need to guarantee positive solutions that preserve equilibrium configuration.

Now, focusing on our cases with diagonal diffusion, we have

$$F^m_{i+\frac{1}{2},j} = \left[\left(1 - \delta^j_i\right) B^m_{i+\frac{1}{2},j} + \frac{\sigma^2}{2h}\right] f^m_{i+1,j} - \left[\frac{\sigma^2}{2h} - \delta^j_i B^m_{i+\frac{1}{2},j}\right] f^m_{i,j} \tag{4.86}$$

and

$$F^m_{i,j+\frac{1}{2}} = \left[\left(1 - \delta^i_j\right) B^m_{i,j+\frac{1}{2}} + \frac{\sigma^2}{2h}\right] f^m_{i,j+1} - \left[\frac{\sigma^2}{2h} - \delta^i_j B^m_{i,j+\frac{1}{2}}\right] f^m_{i,j} \tag{4.87}$$

where $B^m_{i+\frac{1}{2},j} = -b_1(x_{i+\frac{1}{2}}, y_j, t_m)$ and $B^m_{i,j+\frac{1}{2}} = -b_2(x_i, y_{j+\frac{1}{2}}, t_m)$, see (4.59). Further, the linear-combination parameters are given by

$$\begin{aligned} \delta^j_i &= \frac{1}{w^j_i} - \frac{1}{\exp\left(w^j_i\right) - 1}, \qquad & w^j_i &= \frac{2hB^m_{i+\frac{1}{2},j}}{\sigma^2}, \\ \delta^i_j &= \frac{1}{w^i_j} - \frac{1}{\exp\left(w^i_j\right) - 1}, \qquad & w^i_j &= \frac{2hB^m_{i,j+\frac{1}{2}}}{\sigma^2}. \end{aligned} \tag{4.88}$$

This specification completes our illustration of the CC scheme. For the time discretization, we employ a backward Euler scheme as follows

$$\frac{f^m_{i,j} - f^{m-1}_{i,j}}{\delta t} = \frac{1}{h}\left(F^m_{i+\frac{1}{2},j} - F^m_{i-\frac{1}{2},j}\right) + \frac{1}{h}\left(F^m_{i,j+\frac{1}{2}} - F^m_{i,j-\frac{1}{2}}\right), \qquad m = 1, \ldots, N_t. \tag{4.89}$$

The numerical analysis of this scheme is presented in [71] where it is proved that the resulting numerical solution is $O\left(\delta t + h^2\right)$ accurate.

Now, concerning the adjoint FP equation, it has been proved in [8, 84] that the transpose of (4.89) provides an $O\left(\delta t + h^2\right)$ accurate approximation of the adjoint FP equation. This approximation is as follows

$$\begin{aligned} p_{i,j}^{m-1} &= S\left(p^m, u^m\right) \\ &:= p_{i,j}^m + \frac{\delta t}{h}\left(K_{i-\frac{1}{2},j}^m p_{i-1,j}^m - R_{i+\frac{1}{2}.j}^m p_{i,j}^m - K_{i-\frac{1}{2},j}^m p_{i,j}^m + R_{i+\frac{1}{2},j}^m p_{i+1,j}^m\right) \\ &+ \frac{\delta t}{h}\left(K_{i,j-\frac{1}{2}}^m p_{i,j-1}^m - R_{i.j+\frac{1}{2}}^m p_{i,j}^m - K_{i,j-\frac{1}{2}}^m p_{i,j}^m + R_{i,j+\frac{1}{2}}^m p_{i,j+1}^m\right) + \delta t\, G\left(b^m\right) \end{aligned} \tag{4.90}$$

where

$$K_{i+\frac{1}{2},j}^m = \left(1 - \delta_i^j\right) B_{i+\frac{1}{2},j}^m + \frac{\sigma^2}{h}, \quad K_{i-\frac{1}{2},j}^m = \left(1 - \delta_{i-1}^j\right) B_{i-\frac{1}{2},j}^m + \frac{\sigma^2}{h},$$

$$K_{i,j+\frac{1}{2}}^m = \left(1 - \delta_j^i\right) B_{i,j+\frac{1}{2}}^m + \frac{\sigma^2}{h}, \quad K_{i,j-\frac{1}{2}}^m = \left(1 - \delta_{j-1}^i\right) B_{i,j-\frac{1}{2}}^m + \frac{\sigma^2}{h},$$

$$R_{i+\frac{1}{2},j}^m = -\delta_i^j B_{i+\frac{1}{2},j}^m + \frac{\sigma^2}{h}, \quad R_{i-\frac{1}{2},j}^m = -\delta_{i-1}^j B_{i-\frac{1}{2},j}^m + \frac{\sigma^2}{h},$$

$$R_{i,j+\frac{1}{2}}^m = -\delta_j^i B_{i,j+\frac{1}{2}}^m + \frac{\sigma^2}{h}, \quad R_{i,j-\frac{1}{2}}^m = -\delta_{j-1}^i B_{i,j-\frac{1}{2}}^m + \frac{\sigma^2}{h}.$$

Now, we discuss our numerical SQH optimization scheme. This scheme results from the combination of two PMP-based strategies for solving optimal control problems governed by dynamical systems. On the one hand, we refer to the iterative scheme proposed in [85] and on the other hand, we refer to the method presented in [62].

The SQH method to solve (4.73) is given by the following scheme where the augmented Hamiltonian $K_\epsilon : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \times K_V \times K_V \times K_W \times K_W \times \mathbb{R}^n \to \mathbb{R}$ is defined as follows

$$K_\epsilon\left(x, t, f, v, \tilde{v}, w, \tilde{w}, \zeta\right) := H\left(x, t, f, v, w, \zeta\right) + \epsilon\left(\left(v\left(t\right) - \tilde{v}\left(t\right)\right)^2 + \left(w\left(t\right) - \tilde{w}\left(t\right)\right)^2\right)$$

where $v^2 := \sum_{i=1}^n \left(v^i\right)^2$ for any vector $v \in \mathbb{R}^n$. The quadratic term, which augments the Hamiltonian H, aims at penalizing the update of the control. This is supposed to keep the update sufficiently small that the current state f is still an approximation in correspondence of the new control. The value of the weight ϵ of the quadratic term is adapted according to the capability of the control to minimizes the cost functional such that the control updates are chosen as large as possible depending on the strategy for choosing ϵ.

Algorithm 4.1 (SQH method)

```
1. Choose $\epsilon > 0$, $\kappa > 0$, $\hat{\sigma} > 1$, $\zeta \in (0,1)$, $\eta \in (0,\infty)$, $v^0, w^0$, compute $f^0$ by (4.63) and $p^0$ by (4.75) for $v \leftarrow v^0$, $w \leftarrow w^0$, set $k \leftarrow 0$
2. Choose $v \in K_V$ and $w \in K_W$ such that
   $\int_\Omega K_\epsilon\left(x,t,f^k,v,v^k,w,w^k,\nabla p^k\right)dx \leq \int_\Omega K_\epsilon\left(x,t,f^k,\hat{v},v^k,\hat{w},w^k,\nabla p^k\right)dx$
   for all $\hat{v} \in K_V$ and $\hat{w} \in K_W$ and for all $t \in [0,T]$
3. Calculate $f$ by (4.63) for $v, w$ and $\tau_1 := \|v - v^k\|^2_{L^2(0,T)}$, $\tau_2 := \|w - w^k\|^2_{L^2(0,T)}$
4. If $J(f,v,w) - J\left(f^k,v^k,w^k\right) > -\eta\left(\tau_1+\tau_2\right)$: Choose $\epsilon \leftarrow \hat{\sigma}\epsilon$
   Else: Choose $\epsilon \leftarrow \zeta\epsilon$, $f^{k+1} \leftarrow f$, $v^{k+1} \leftarrow v$, $w^{k+1} \leftarrow w$, calculate $p^{k+1}$ by (4.75) for $v \leftarrow v^{k+1}$ and $w \leftarrow w^{k+1}$, set $k \leftarrow k+1$
5. If $\tau_1 + \tau_2 < \kappa$: STOP and return $v^k$ and $w^k$
   Else go to 2.
```

The SQH method determines $v(t)$ and $w(t)$ such that the function

$$(\hat{v},\hat{w}) \mapsto \int_\Omega K_\epsilon\left(x,t,f^k,\hat{v},v^k,\hat{w},w^k,\nabla p^k\right)$$

is minimized for all $\hat{v} \in K_V$, $\hat{w} \in K_W$ for any $t \in [0,T]$. We assume that the functions v, w, which are determined in Step 2 of Algorithm 4.1, are measurable, see the discussion starting on page 166 for details. Next, we calculate f with v and w and check if the triple (f,v,w) reduces the value of the cost functional by at least $\eta(\tau_1+\tau_2)$. If not, ϵ is increased in order to obtain a smaller deviation of the control. If the triple (f,v,w) reduces the value of the cost functional by at least $\eta(\tau_1+\tau_2)$, we accept the values (f,v,w) as our next iterate, decrease ϵ to possibly obtain a larger update for the control in the following sweep and calculate the adjoint variable. Specifically, we have that f^k is a solution to (4.63) for $v \leftarrow v^k$, $w \leftarrow w^k$ and p^k is a solution to (4.75) for $v \leftarrow v^k$, $w \leftarrow w^k$. We repeat this steps until the convergence criterion is fulfilled where the algorithm stops and the last accepted control v^k, w^k is returned.

The SQH method to solve (4.73) is well defined since analogous results as Lemma 7 and Lemma 12 hold with the same arguments. These results guarantee the existence of a minimum of $(\hat{v},\hat{w}) \mapsto \int_\Omega K_\epsilon\left(x,t,f^k,\hat{v},v^k,\hat{w},w^k,\nabla p^k\right)$ even for lower semi-continuous functions g_s^i and that Algorithm 4.1 stops if v^k and w^k are optimal in the PMP sense (4.78). We remark that the minimization in Step 2 of Algorithm 4.1 is only in t instead of (x,t). However, the scaling of the calculation effort is the same since we have to integrate K_ϵ over the space instead of performing the minimization of K_ϵ in (x,t).

The Step 4 of Algorithm 4.1 is well posed. This means that increasing ϵ, at most finitely times, will end up in an update for the control that minimizes the cost functional value by at least $\eta(\tau_1+\tau_2)$. The argument is provided by the following lemma.

Lemma 50. *Let (f,v,w) and $\left(f^k,v^k,w^k\right)$ be generated by Algorithm 4.1, $k \in \mathbb{N}_0$. Then, there is a $\theta_k > 0$ independent of ϵ in each iteration such that for any ϵ currently chosen by Algorithm 4.1, the following holds*

$$J(f,v,w) - J\left(f^k,v^k,w^k\right) \leq -\left(\epsilon|\Omega| - \theta_k\right)\left(\|\delta v\|^2_{L^2(0,T)} + \|\delta w\|^2_{L^2(0,T)}\right)$$

with $\delta v := v - v^k$ and $\delta w = w - w^k$. In particular $J(f,v,w) - J\left(f^k,v^k,w^k\right) \leq 0$ for $\epsilon \geq \theta_k$.

Proof. We drop the arguments of the functions for notational effort. We define $\delta f := f - f^k$, $\delta p := p - p^k$, $\delta\nabla f := \nabla f - \nabla f^k$ with $\nabla\delta f = \nabla\left(f - f^k\right) = \delta\nabla f$ and $\delta\nabla p := \nabla p - \nabla p^k$ with $\nabla\delta p = \nabla\left(p - p^k\right) = \delta\nabla p$.

Furthermore, we write $H^k \coloneqq H\left(x,t,f^k,v^k,w^k,\nabla p^k\right)$, $H \coloneqq H\left(x,t,f,v,w,\nabla p\right)$, $G^k \coloneqq G\left(x,t,v^k,w^k\right)$ and $G \coloneqq G\left(x,t,v,w\right)$. With Lemma 53, we have

$$\begin{aligned}
&J\left(f,v,w\right) - J\left(f^k,v^k,w^k\right) = \int_Q Gf - G^k f^k dxdt + \int_\Omega F\delta f dx \\
&= \int_Q H - \nabla p \cdot \left(\left(v + x\circ w\right) f\right) - H^k + \nabla p^k \cdot \left(\left(v^k + x \circ w^k\right) f^k\right) dxdt + \int_\Omega F\delta f dx \\
&= \int_Q H - H\left(x,t,f^k,v,w,\nabla p^k\right) + H\left(x,t,f^k,v,w,\nabla p^k\right) - H^k dxdt \\
&+ \int_Q -\nabla p \cdot \left(\left(v + x\circ w\right) f\right) + \nabla p^k \cdot \left(\left(v^k + x \circ w^k\right) f^k\right) dxdt \\
&+ \int_Q \epsilon\left(\left(\delta v\right)^2 + \left(\delta w\right)^2\right) - \epsilon\left(\left(\delta v\right)^2 + \left(\delta w\right)^2\right) dxdt + \int_\Omega F\delta f dx \\
&\le \int_Q -\epsilon\left(\left(\delta v\right)^2 + \left(\delta w\right)^2\right) + H - H\left(x,t,f^k,v,w,\nabla p^k\right) - \delta\nabla p \cdot \left(\left(v + x\circ w\right) f\right) dxdt \qquad (4.91)\\
&+ \int_Q -\nabla p^k \cdot \left(\left(v + x\circ w\right) f\right) + \nabla p^k \cdot \left(\left(v^k + x \circ w^k\right) f^k\right) dxdt + \int_\Omega F\delta f dx \\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2 + \left(\delta w\right)^2\right) + H - H\left(x,t,f^k,v,w,\nabla p^k\right) - \delta\nabla p \cdot \left(\left(v + x\circ w\right) f\right) dxdt \\
&+ \int_Q -f'p^k - \frac{\sigma^2}{2}\nabla f \cdot \nabla p^k + p^k\left(f^k\right)' + \frac{\sigma^2}{2}\nabla f^k \cdot \nabla p^k dxdt + \int_\Omega F\delta f dx \\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2 + \left(\delta w\right)^2\right) + H - H\left(x,t,f^k,v,w,\nabla p^k\right) - \delta\nabla p \cdot \left(\left(v + x\circ w\right) f\right) dxdt \\
&- \int_Q p^k\left(\delta f\right)' + \frac{\sigma^2}{2}\nabla\delta f \cdot \nabla p^k dxdt + \int_\Omega F\delta f dx,
\end{aligned}$$

since we have by Algorithm 4.1 that

$$\int_\Omega K_\epsilon\left(x,t,f^k,v,v^k,w,w^k,\nabla p^k\right) dx \le \int_\Omega K_\epsilon\left(x,t,f^k,\hat{v},v^k,\hat{w},w^k,\nabla p^k\right) dx$$

for all $\hat{v} \in K_V$ and $\hat{w} \in K_W$ and for all $t \in [0,T]$ from which it follows that

$$\int_0^T \int_\Omega K_\epsilon\left(x,t,f^k,v,v^k,w,w^k,\nabla p^k\right) dx \le \int_0^T \int_\Omega K_\epsilon\left(x,t,f^k,\hat{v},v^k,\hat{w},w^k,\nabla p^k\right) dx$$

for all $\hat{v} \in K_V$ and $\hat{w} \in K_W$, see [5, X Corollary 2.16 ii)] and thus

$$\begin{aligned}
0 &\ge \int_0^T \int_\Omega K_\epsilon\left(x,t,f^k,v,v^k,w,w^k,\nabla p^k\right) - K_\epsilon\left(x,t,f^k,v^k,v^k,w^k,w^k,\nabla p^k\right) dx \\
&= \int_Q H\left(x,t,f^k,v,w,\nabla p^k\right) + \epsilon\left(\left(\delta v\right)^2 + \left(\delta w\right)^2\right) - H^k dxdt.
\end{aligned}$$

Now, we estimate the term $\int_Q H - H\left(x,t,f^k,v,w,\nabla p^k\right) dxdt$. We have by the Taylor formula [4, Chapter VII, Theorem 5.8] and with the symmetry of the second derivative [4, Chapter VII, Theorem 5.2] the following

$$\begin{aligned}
&\int_Q H - H\left(x,t,f^k,v,w,\nabla p^k\right) dxdt = \int_Q H - H\left(x,t,f-\delta f,v,w,\nabla p-\delta\nabla p\right) dxdt\\
&= \int_Q \frac{\partial}{\partial f} H\delta f + \partial_{\nabla p} H\cdot\delta\nabla p - \frac{1}{2}\left(2\left(v+x\circ w\right)\delta f\delta\nabla p\right) dxdt\\
&= \int_Q \left(G+\nabla p\cdot\left(v+x\circ w\right)\right)\delta f + \left(v+x\circ w\right)\cdot f\delta\nabla p - \left(v+x\circ w\right)\cdot\delta f\delta\nabla p dxdt\\
&= \int_Q -p'\delta f + \frac{\sigma^2}{2}\nabla p\cdot\nabla\delta f - \left(v+x\circ w\right)\cdot\nabla p\delta f + \nabla p\cdot\left(v+x\circ w\right)\delta f + \left(v+x\circ w\right)\cdot f\delta\nabla p dxdt\\
&- \int_Q \left(v+x\circ w\right)\cdot\delta f\delta\nabla p dxdt
\end{aligned}\tag{4.92}$$

where

$$\partial_{\nabla p} H := \begin{pmatrix} \frac{\partial}{\partial r_1} H\left(x,t,f,v,w,r\right)|_{r=\nabla p}\\ \vdots\\ \frac{\partial}{\partial r_n} H\left(x,t,f,v,w,r\right)|_{r=\nabla p}\end{pmatrix},\ r = \begin{pmatrix} r_1\\ \vdots\\ r_n\end{pmatrix}\in\mathbb{R}^n.$$

We combine (4.91) and (4.92) and obtain by partial integration with respect to t, see [95, Satz 3.11], in (4.92) the following

$$\begin{aligned}
&J\left(f,v,w\right) - J\left(f^k,v^k,w^k\right)\\
&\leq \int_Q -\epsilon\left(\left(\delta v\right)^2+\left(\delta w\right)^2\right) + \delta f' p dxdt - \int_\Omega F\delta f dx + \int_Q \frac{\sigma^2}{2}\nabla p\cdot\nabla\delta f - \left(v+x\circ w\right)\cdot\nabla p\delta f dxdt\\
&+ \int_Q \left(v+x\circ w\right)\cdot\nabla p\delta f + \delta\nabla p\cdot\left(v+x\circ w\right)f - \left(v+x\circ w\right)\cdot\delta f\delta\nabla p - \delta\nabla p\cdot\left(v+xw\right) f dxdt\\
&+ \int_Q -p^k\delta f' - \frac{\sigma^2}{2}\nabla p^k\cdot\nabla\delta f dxdt + \int_\Omega F\delta f dx\\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2+\left(\delta w\right)^2\right) + \delta f'\delta p + \frac{\sigma^2}{2}\nabla\delta p\cdot\nabla\delta f - \left(v+x\circ w\right)\cdot\delta f\delta\nabla p dxdt\\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2+\left(\delta w\right)^2\right) + f'\delta p + \frac{\sigma^2}{2}\nabla f\cdot\nabla\delta p + \nabla\left(\left(v+x\circ w\right)f\right)\delta p - \nabla\left(\left(v+x\circ w\right)f\right)\delta p dxdt\\
&+ \int_Q -\left(f^k\right)'\delta p - \frac{\sigma^2}{2}\nabla f^k\cdot\nabla\delta p - \nabla\left(\left(v^k+x\circ w^k\right)f^k\right)\delta p + \nabla\left(\left(v^k+x\circ w^k\right)f^k\right)\delta p dxdt\\
&- \int_Q \left(v+x\circ w\right)\cdot\delta f\delta\nabla p dxdt\\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2+\left(\delta w\right)^2\right) - \nabla\left(\left(v+x\circ w\right)f\right)\delta p + \nabla\left(\left(v^k+x\circ w^k\right)f^k\right)\delta p dxdt\\
&- \int_Q \left(v+x\circ w\right)\cdot\delta f\delta\nabla p dxdt\\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2+\left(\delta w\right)^2\right) + \nabla\left(\left(v^k+x\circ w^k\right)f^k\right)\delta p - \nabla\left(\left(v+x\circ w\right)f^k\right)\delta p dxdt\\
&= \int_Q -\epsilon\left(\left(\delta v\right)^2+\left(\delta w\right)^2\right) + \left(v^k - v + x\circ\left(w^k-w\right)\right)\cdot\nabla f^k\delta p + \sum_{i=1}^n\left(\left(w^k\right)^i - \left(w\right)^i\right) f^k\delta p dxdt.
\end{aligned}\tag{4.93}$$

Next, estimate the terms in the last line of (4.93) by δv and δw. We have by the Cauchy-Schwarz inequality, see [2, Lemma 2.2] and the Jensen inequality, see [72, Proposition 824] the following

$$\begin{aligned}
&\sum_{i=1}^{n}\int_Q |\left(\left(w^k\right)^i-(w)^i\right) f^k\delta p|dxdt \le \|f^k\|_{L^\infty(Q)}\sum_{i=1}^{n}\| \left(\delta w\right)^i \|_{L^2(Q)}\|\delta p\|_{L^2(Q)}\\
&= \|f^k\|_{L^\infty(Q)}\sqrt{\left(\sum_{i=1}^{n}\| \left(\delta w\right)^i \|_{L^2(Q)}\right)^2}\|\delta p\|_{L^2(Q)}\\
&\le \sqrt{n}\|f^k\|_{L^\infty(Q)}\sqrt{\left(\sum_{i=1}^{n}\| \left(\delta w\right)^i \|^2_{L^2(Q)}\right)}\|\delta p\|_{L^2(Q)}\\
&\le \sqrt{n}\|f^k\|_{L^\infty(Q)}|\Omega|\|\delta w\|_{L^2(0,T)}\hat{\theta}^k\left(\|\delta v\|_{L^2(0,T)}+\|\delta w\|_{L^2(0,T)}\right)\\
&\le \sqrt{n}\|f^k\|_{L^\infty(Q)}\hat{\theta}^k|\Omega|\left(\frac{1}{2}\|\delta v\|^2_{L^2(0,T)}+\frac{3}{2}\|\delta w\|^2_{L^2(0,T)}\right)\\
&\le \frac{3}{2}\sqrt{n}\|f^k\|_{L^\infty(Q)}\hat{\theta}^k|\Omega|\left(\|\delta v\|^2_{L^2(0,T)}+\|\delta w\|^2_{L^2(0,T)}\right)
\end{aligned}\tag{4.94}$$

where $|\Omega|$ is the measure of Ω, $\|f^k\|_{L^\infty(Q)}$ exists according to Theorem 44 and we use Cauchy's inequality, see [45, page 622] and where $\|\delta p\|_{L^2(Q)}$ is estimated with Lemma 66. With the Jensen inequality, see [72, Proposition 824], and Lemma 66 we estimate the following

$$\begin{aligned}
&\int_Q |\left(v^k-v+x\circ\left(w^k-w\right)\right)\cdot\nabla f^k\delta p|dxdt\\
&\le \int_0^T\sum_{i=1}^{n}\left(\left(|\left(v^k\right)^i-v^i|+c|\left(\left(w^k\right)^i-w^i\right)|\right)\right)\int_\Omega|\frac{\partial}{\partial x_i}f^k\delta p|dxdt\\
&\le \int_0^T\left(\sum_{i=1}^{n}\left(|\left(\left(v^k\right)^i-v^i|+c|\left(\left(w^k\right)^i-w^i\right)\right)|\right)\right)\|\frac{\partial}{\partial x_i}f^k\left(\cdot,t\right)\|_{L^2(\Omega)}\|\delta p\left(\cdot,t\right)\|_{L^2(\Omega)}dt\\
&\le \sum_{i=1}^{n}\int_0^T\left(\left(|\left(\left(v^k\right)^i-v^i|+c|\left(\left(w^k\right)^i-w^i\right)\right)|\right)\right)\|f^k\left(\cdot,t\right)\|_{H_0^1(\Omega)}\|\delta p\left(\cdot,t\right)\|_{L^2(\Omega)}dt\\
&\le \|f^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}\sum_{i=1}^{n}\left(\|\left(\delta v\right)^i\|_{L^2(0,T)}+c\|\left(\delta w\right)^i\|_{L^2(0,T)}\right)\|\|\delta p\left(\cdot,t\right)\|_{L^2(\Omega)}\|_{L^2(0,T)}\\
&\le \bar{\theta}^k\left(\sqrt{\left(\sum_{i=1}^{n}\|\left(\delta v\right)^i\|_{L^2(0,T)}\right)^2}+\sqrt{\left(\sum_{i=1}^{n}\|\left(\delta w\right)^i\|_{L^2(0,T)}\right)^2}\right)\left(\|\delta v\|_{L^2(0,T)}+\|\delta w\|_{L^2(0,T)}\right)\\
&\le \bar{\theta}^k\sqrt{n}\left(\|\delta v\|_{L^2(0,T)}+\|\delta w\|_{L^2(0,T)}\right)\left(\|\delta v\|_{L^2(0,T)}+\|\delta w\|_{L^2(0,T)}\right)\\
&\le 2\sqrt{n}\bar{\theta}^k\left(\|\delta v\|^2_{L^2(0,T)}\|+\|\delta w\|^2_{L^2(0,T)}\right)
\end{aligned}\tag{4.95}$$

where $c:=\max_{i=1,\dots,n}\max_{x\in\Omega}|x_i|$, $|\Omega|$ is the measure of Ω, $\bar{\theta}^k:=\|f^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}\hat{\theta}^k\max\left(1,c\right)$ and

$$\|\|\delta p\left(\cdot,t\right)\|_{L^2(\Omega)}\|_{L^2(0,T)}=\sqrt{\int_0^T\int_\Omega\delta p^2\left(x,t\right)dxdt}=\|\delta p\|_{L^2(Q)}.$$

We obtain

$$J\left(f,v,w\right)-J\left(f^k,v^k,w^k\right)\le-\left(\epsilon|\Omega|-\theta_k\right)\left(\|\delta v\|^2_{L^2((0,T))}+\|\delta w\|^2_{L^2((0,T))}\right)$$

with $\theta_k:=\max\left(\frac{3}{2}\sqrt{n}\|f^k\|_{L^\infty(Q)}\hat{\theta}^k|\Omega|,2\sqrt{n}\bar{\theta}^k\right)$ where there is an upper bound for $\|f^k\|_{L^\infty(Q)}$ for all $k\in\mathbb{N}_0$ due to Theorem 44. □

Notice that in contrast to the analogous results from the previous chapters, Lemma 11 and Lemma 26, the existing threshold θ_k in the present case, for which we have a minimization of the cost functional if exceeded by ϵ, depends on k where in Lemma 11 and Lemma 26 the corresponding constant θ holds for all k. The reason for this is that the norms $\|f^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ and $\|p^k\|^2_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ exist, see the discussion of existence of a solution to the state and adjoint equation, but it is not proved if $\|f^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ and $\|p^k\|^2_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ are bounded by a constant that holds for all $v \in V_{ad}$ and all $w \in W_{ad}$ as it is the case for $\|f^k\|_{L^\infty(Q)}$ or $\|p^k\|_{L^\infty(Q)}$ for instance, see Theorem 44 or Theorem 46. This causes a dependency of θ_k on $\|f^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ and $\|p^k\|^2_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$, see (4.95) and (5.32) for details and the lack of an upper bound for θ_k in the present formulation.

A consequence of the lack of an upper bound for θ_k for all $k \in \mathbb{N}_0$ is that the investigation of the properties of a sequence generated by a loop over Step 2 to Step 4 of Algorithm 4.1 is more delicate now. While the cost functional minimizing properties stated in Theorem 13 still hold with the same proof, the proof of Theorem 14 or analogous Theorem 27, where the pointwise convergence to a PMP consistent solution is proved, cannot be applied, since they require an upper bound for ϵ that holds for all iterations of the SQH method. This upper bound for ϵ is given if θ_k has an upper bound that holds for all iterations of the SQH method, see below (2.37) for instance. However, if there is an upper bound for θ_k for all $k \in \mathbb{N}_0$ and the estimations $\|f - f^k\|_{L^2(Q)} \leq \tilde{C}\|u - u^k\|_{L^2(0,T)}$, $\|p - p^k\|_{L^2(Q)} \leq \tilde{C}\|u - u^k\|_{L^2(0,T)}$, $\tilde{C} > 0$ hold, then the proof of Theorem 14 or analogous Theorem 27 also hold in the FP case. These estimation can also be proved with a similar calculation as the one starting on page 91 if we have that $\|f^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ and $\|p^k\|^2_{L^\infty\left(0,T;H_0^1(\Omega)\right)}$ are bounded by a constant for all $k \in \mathbb{N}_0$, for all $v \in V_{ad}$ and for all $w \in W_{ad}$.

Summarizing the discussion, we have that each sweep of Algorithm 4.1 is well defined and performs improvements to an initial guess of the control, if it is not already optimal, such that the cost functional value decreases.

Next, we discuss the numerical solution of (4.74). The SQH scheme corresponding to (4.74) is given below where a similar discussion holds as in the case above. The augmented Hamiltonian is defined by

$$K_\epsilon\left(x,t,f,u,\tilde{u},\nabla p\right) := H\left(x,t,f,u,\nabla p\right) + \epsilon\left(u\left(x,t\right) - \tilde{u}\left(x,t\right)\right)^2 \tag{4.96}$$

where H is given by (4.82).

Algorithm 4.2 (SQH method)

```
1. Choose $\epsilon > 0$, $\kappa > 0$, $\hat{\sigma} > 1$, $\zeta \in (0,1)$, $\eta \in (0,\infty)$, $v^0, w^0$, compute $f^0$ by the state equation of
   (4.74) and $p^0$ by (4.81) for $v \leftarrow v^0$, $w \leftarrow w^0$, set $k \leftarrow 0$

2. Choose $u \in K_U$ such that

   $K_\epsilon\left(x,t,f^k,u,u^k,\nabla p^k\right) \leq K_\epsilon\left(x,t,f^k,\hat{u},u^k,\nabla p^k\right)$

   for all $\hat{u} \in K_U$ and for all $t \in [0,T]$

3. Calculate $f$ by the state equation of (4.74) for $u$ and $\tau := \|u - u^k\|^2_{L^2(Q)}$

4. If $J\left(f,u\right) - J\left(f^k,u^k\right) > -\eta\tau$: Choose $\epsilon \leftarrow \hat{\sigma}\epsilon$
   Else: Choose $\epsilon \leftarrow \zeta\epsilon$, $f^{k+1} \leftarrow f$, $u^{k+1} \leftarrow u$ calculate $p^{k+1}$ by (4.81) for $u \leftarrow u^{k+1}$, set $k \leftarrow k+1$

5. If $\tau < \kappa$: STOP and return $u^k$
   Else go to 2.
```

A closer look on the necessary optimality conditions corresponding to (4.74) reveals that we do not need to compute the solution of the forward FP problem, but it is sufficient to calculate the adjoint FP equation together with the PMP optimality condition, see (4.84). Therefore a natural approach to solve this problem is to consider an explicit (backward) time approximation of (4.81) starting from the terminal condition, and implementing the minimization of $v \mapsto G(v) + \nabla \bar{p} \cdot v$, before proceeding to the next time step. It is known [99] that such an explicit scheme may suffer from instabilities. However, in our setting and with a sufficiently small time step, the following solution procedure appears stable.

Algorithm 4.3 Direct Hamiltonian method

```
1. Set $p^{N_t} = F$
2. For $m = N_t, ..., 0$ do:
    (a) Set $u^m := \arg\min_{v \in K_U} G(v) + \nabla_h p^m \cdot v$
    (b) Set $p^{m-1} := S(p^m, u^m)$
```

Notice that in this algorithm all equalities are meant for all space grid points i, j.

We remark that in the finite-volume CC scheme, and its adjoint, the drift (control) is placed on the cell edges. In our second setting of (4.74), the first component of the control field u_1 is placed on the vertical edges and normal to it. On the other hand, the second component u_2 is placed on the horizontal edges and normal to it. Thus in this setting the product $\nabla_h p^m \cdot u$ is approximated as follows

$$\nabla_h p^m \cdot u|_{ij} = u_1\left(x_{i+1/2}, y_j, t_m\right) \frac{p^m_{i+1,j} - p^m_{i,j}}{h} + u_2\left(x_i, y_{j+1/2}, t_m\right) \frac{p^m_{i,j+1} - p^m_{i,j}}{h}$$

Notice that this is a second-order approximation of the continuous product $\nabla_h p^m \cdot u$.

We remark that in both the SQH and the DH schemes, the pointwise values of the optimal controls are obtained by solving finite-dimensional optimization problems in the respective compact sets K_V, K_W and K_U, respectively. These problems can be solved by many optimization schemes, including direct search. However, in many cases it is possible to determine the solution by a case study and, if this is the case, the optimization procedure becomes very fast. To illustrate this advantage of our PMP-based procedure, we discuss two specific non-smooth cost functionals that are also considered in our numerical experiments below.

Let us consider the optimal control problem (4.73) with $\frac{\sigma^2}{2} = \frac{1}{32}$. We choose $\Omega = (-5,5) \times (-5,5)$ with a uniform space discretization $h = \frac{1}{10}$ and on the interval $[0,T]$ with $T = 1$, we set the time steps $\delta t = \frac{1}{100}$.

In our cost functional, we choose

$$G(x,t,v,w) = -A(x,t) + \alpha g_{s_1}(v) + \beta g_{s_2}(w), \qquad F = 0$$

where $\alpha = \beta = \frac{1}{2} \cdot 10^{-2}$ and

$$A(x,t) := \begin{cases} \mathrm{e}^{\frac{\rho^2}{|x - x_d(t)|^2 - \rho^2}} & \text{if } |x - x_d(t)| < \rho \\ 0 & \text{else} \end{cases} \tag{4.97}$$

with $|\cdot|$ the Euclidean norm according to $|y| := \sqrt{(y^1)^2 + ... + (y^n)^2}$. We have $\rho = 1$ and the desired trajectory $x_d : \mathbb{R} \to \mathbb{R}^2$ is given by the spiral curve

$$x_d(t) = \begin{pmatrix} \frac{2t}{T} \cos\left(2\pi \frac{t}{T}\right) \\ \frac{2t}{T} \sin\left(2\pi \frac{t}{T}\right) \end{pmatrix}. \tag{4.98}$$

The control costs are determined by the function

$$g_s(z) := \max\left(0, |z^1| - s\right) + \max\left(0, |z^2| - s\right)$$

where $s_1 = \frac{3}{5}$, $s_2 = \frac{3}{10}$ in G. The admissible values of the controls are given by the intervals $K_V^1 = K_V^2 = [v_{\min}, v_{\max}]$ and $K_W^1 = K_W^2 = [-w_{\min}, w_{\max}]$ where $v_{\min} = -2$, $v_{\max} = 2$, $w_{\min} = -1$ and $w_{\max} = 1$.

The initial condition for the forward FP problem is given by the following Boltzmann-like distribution

$$f_0(x) = d|x - x_0|^2 \mathrm{e}^{-4|x-x_0|^2} \tag{4.99}$$

with $x_0 = (-2.1, -3.1)$. The constant $d > 0$ is set such that $\int_\Omega f_0(x)\, dx = 1$.

A discussion of existence of a solution to the corresponding optimal control problem can be done analogous to Remark 45 due to Lemma 67 that states the Lipschitz continuity and the convexity of the function $z^i \mapsto \max\left(0, |z^i| - s\right)$, $i = 1, 2$.

The parameters for the SQH Algorithm 4.1 are set as follows. The initial guess $\epsilon = 10^{-2}$ and the initial values of v and w are zero. Furthermore, we have $\eta = 10^{-7}$, $\hat{\sigma} = 50$, $\zeta = \frac{3}{20}$ and $\kappa = 10^{-10}$.

Next, we can discuss how to find the point-wise minimum in Step 2 of the SQH Algorithm 4.1. We have

$$\begin{aligned}
&\int_\Omega K_\epsilon\left(x, t, f, v, \tilde{v}, w, \tilde{w}, \nabla p\right) dx \\
&= \int_\Omega G(v, w) f(x,t) + \nabla p(x,t) \cdot (v + x \circ w) f(x,t) + \epsilon \left((v - \tilde{v}(t))^2 + (w - \tilde{w}(t))^2 \right) dx \\
&= G(v,w)\, a(t) + \sum_{i=1}^{2} v^i b_i(t) + \sum_{i=1}^{2} w^i c_i(t) + \epsilon |\Omega| \left(\sum_{i=1}^{n} \left(v^i - \tilde{v}^i(t)\right)^2 + \sum_{i=1}^{n} \left(w^i - \tilde{w}^i(t)\right)^2 \right)
\end{aligned}$$

where $|\Omega|$ is the measure of Ω, $a(t) := \int_\Omega f(x,t)\, dx$ and

$$b_i(t) := \int_\Omega \frac{\partial}{\partial x_i} p(x,t) f(x,t)\, dx, \quad c_i(t) := \int_\Omega x_i \frac{\partial}{\partial x_i} p(x,t) f(x,t)\, dx, \quad i = \{1, 2\}.$$

We remark that due to the zero boundary conditions, which means the homogeneous Dirichlet boundary conditions, it holds that $0 \leq a \leq 1$. Since

$$\max(0, |z| - s) = \begin{cases} z - s & \text{if } z > s \\ -z - s & \text{if } z < -s\ , \\ 0 & \text{if } |z| \leq s \end{cases}$$

the pointwise minimum of $(v, w) \mapsto \int_\Omega K_\epsilon\left(x, t, f, v, \tilde{v}, w, \tilde{w}, \nabla p\right)$ is given by the following case study where we use the differentiability of $(v, w) \mapsto \int_\Omega K_\epsilon\left(x, t, f, v, \tilde{v}, w, \tilde{w}, \nabla p\right)$ in the intervals $(v_{\min}, -s)$, $(-s, s)$, $(s, v_{\max})$ for v and analogously for w.

Similar to the discussion starting on page 166, we see that the minimization problem in the given intervals reduces to the evaluation of the integral of the augmented Hamiltonian on a discrete set of points as follows

$$\begin{aligned}
(v(t), w(t)) &= \underset{\hat{v} \in K_V, \hat{w} \in K_W}{\arg\min} \int_\Omega K_\epsilon\left(x, t, f, \hat{v}, \tilde{v}, \hat{w}, \tilde{w}, \nabla p\right) dx \\
&= \underset{\hat{v} \in \tilde{K}_V(t), \hat{w} \in \tilde{K}_W(t)}{\arg\min} \int_\Omega K_\epsilon\left(x, t, f, \hat{v}, \tilde{v}, \hat{w}, \tilde{w}, \nabla p\right) dx
\end{aligned}$$

where

$$\tilde{K}_V(t) := \tilde{K}_V^1(t) \times \tilde{K}_V^2(t), \quad \tilde{K}_W := \tilde{K}_W^1(t) \times K_W^2(t),$$

$$\tilde{K}_V^i(t) := \left\{v_1^i(t), v_2^i(t), v_3^i(t)\right\}, \quad \tilde{K}_W^i := \left\{w_1^i(t), w_2^i(t), w_3^i(t)\right\}, \quad i = \{1,2\},$$

with

$$v_1^i(t) = \min\left(\max\left(s_1, \frac{2\epsilon|\Omega|\tilde{v}^i(t) - \alpha a(t) - b_i(t)}{2|\Omega|\epsilon}\right), v_{\max}\right),$$

$$v_2^i(t) = \min\left(\max\left(v_{\min}, \frac{2\epsilon|\Omega|\tilde{v}^i(t) - \alpha a(t) - b_i(t)}{2|\Omega|\epsilon}\right), -s_1\right),$$

$$v_3^i(t) = \min\left(\max\left(-s_1, \frac{2\epsilon|\Omega|\tilde{v}^i(t) - b_i(t)}{2|\Omega|\epsilon}\right), s_1\right),$$

$$w_1^i(t) = \min\left(\max\left(s_2, \frac{2\epsilon|\Omega|\tilde{w}^i(t) - \beta a(t) - c_i(t)}{2|\Omega|\epsilon}\right), w_{\max}\right),$$

$$w_2^i(t) = \min\left(\max\left(w_{\min}, \frac{2\epsilon|\Omega|\tilde{w}^i(t) - \beta a(t) - c_i(t)}{2|\Omega|\epsilon}\right), -s_2\right)$$

and

$$w_3^i(t) = \min\left(\max\left(-s_2, \frac{2\epsilon|\Omega|\tilde{w}^i(t) - c_i(t)}{2|\Omega|\epsilon}\right), s_2\right)$$

for any $t \in [0,T]$ and $i = \{1,2\}$ since the minimum is either in the inner of the corresponding intervals where the derivative with respect to v or w equals zero or on the boundary of the intervals, see [3, IV Remark 2.2 (b)].

Next, we consider our second optimal control problem (4.74). We choose the same domain Ω and $T = 1$, with the same discretization in space, and for the time discretization, we take $\delta t = \frac{1}{500}$. Now, our cost functional is given by

$$G(u(x,t)) = -A(x,t) + \frac{\alpha}{2}\left(u_1^2(x,t) + u_2^2(x,t)\right) + \beta\left(|u_1(x,t)| + |u_2(x,t)|\right)$$

where A is as above and $\alpha = 10^{-5}$, $\beta = 10^{-3}$. The admissible set of values of the control is given by the interval $K_U = [u_{\min}, u_{\max}]^2$ with $u_{\min} = -10$, $u_{\max} = 10$. We have $\sigma = 1$.

Since K_U is compact, we have that the square function $z \mapsto z^2 : [u_{\min}, u_{\max}] \to \mathbb{R}$ is Lipschitz-continuous as follows

$$|z_1^2 - z_2^2| = |z_1 + z_2||z_1 - z_2| \leq 2\max\left(|u_{\min}|, |u_{\max}|\right)|z_1 - z_2|.$$

The Lipschitz continuity for the absolute value $z \mapsto |z| : [u_{\min}, u_{\max}] \to \mathbb{R}$ follows from the reversed triangle inequality, see [3, Corollary 8.11]. The convexity of the square function follows from Jensen's inequality, see [72, Proposition 824] and the convexity of the absolute value from the triangle inequality [3, Theorem 8.10]. Consequently a discussion of existence of a solution to the corresponding optimal control problem can be done analogous to Remark 45.

The parameters for the SQH method, Algorithm 4.2, are given in this case as follows. We have that the initial guess for $\epsilon = 0$ and the initial guess of the control $u = 0$. We have $\hat{\sigma} = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$, $\kappa = 10^{-10}$.

We can also calculate the points where the augmented Hamiltonian (4.96) used in the SQH scheme, see Algorithm 4.2, can attain (pointwise) its minimum value. The formulas are given by

$$\begin{aligned} u &= \underset{v \in K_U}{\arg\min}\left(G(v) + \nabla_h p \cdot v\right) f + \epsilon\left(v - \tilde{u}\right)^2 \\ &= \underset{v_1 \in \left\{v_1^1, v_1^2\right\},\, v_2 \in \left\{v_2^1, v_2^2\right\}}{\arg\min}\left(\frac{\alpha}{2}\left(v_1^2 + v_2^2\right) + \beta\left(|v_1| + |v_2|\right) + v_1 \nabla_h^1 p + v_2 \nabla_h^2 p\right) f + \epsilon\left(v - \tilde{u}\right)^2 \end{aligned}$$

where

$$v_i^1 = \min\left(\max\left(0, \frac{2\epsilon\tilde{u}_i(x,t) - \nabla_h^i p(x,t) f(x,t) - \beta f(x,t)}{2\epsilon + \alpha f(x,t)}\right), u_{\max}\right)$$

and

$$v_i^2 = \min\left(\max\left(u_{\min}, \frac{2\epsilon\tilde{u}_i(x,t) - \nabla_h^i p(x,t) f(x,t) + \beta f(x,t)}{2\epsilon + \alpha f(x,t)}\right), 0\right).$$

Also for the DH method, Algorithm 4.3, we can determine a priori the set of points where the involved Hamiltonian can take a minimum. Specifically, we have

$$\begin{aligned} u &= \underset{v\in K_U}{\arg\min}\left(G(v) + \nabla_h p \cdot v\right) \\ &= \underset{v_1\in\left\{v_1^1,v_1^2\right\},\, v_2\in\left\{v_2^1,v_2^2\right\}}{\arg\min} \frac{\alpha}{2}\left(v_1^2+v_2^2\right) + \beta\left(|v_1|+|v_2|\right) + v_1\nabla_h^1 p + v_2\nabla_h^2 p \end{aligned}$$

where

$$v_i^1 = \min\left(\max\left(0, \frac{-\nabla_h^i p(x,t) - \beta}{\alpha}\right), u_{\max}\right)$$

and

$$v_i^2 = \min\left(\max\left(u_{\min}, \frac{-\nabla_h^i p(x,t) + \beta}{\alpha}\right), 0\right)$$

for $i = \{1,2\}$. We show that (4.74) can be solved with both methods in order to control the corresponding stochastic process.

4.5 Numerical experiments

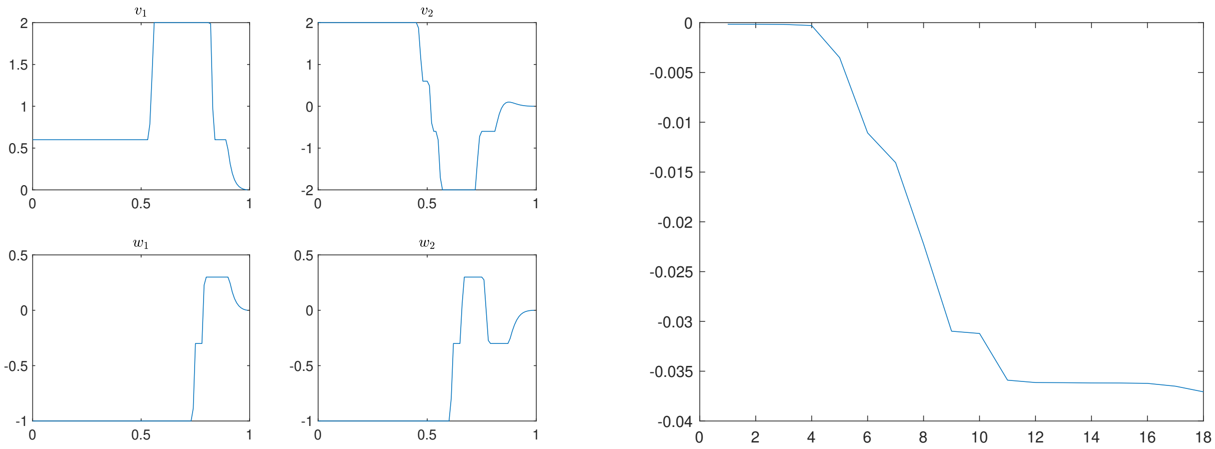
In this section, we report results of numerical experiments that validate the Fokker-Planck optimization framework and the ability of the resulting controls to drive the related stochastic processes. Our numerical experiments are performed with the setting specified in the previous section and in the same order.

Concerning the first goal, we would like to demonstrate that our optimization procedure is able to provide a solution that satisfies the PMP optimality conditions discussed in the previous sections. For this purpose, we define a measure of PMP optimality of the numerical solution as follows

$$\triangle H(t) := \int_\Omega H(x,t,f,v,w,\nabla p)\,dx - \min_{\tilde{v}\in K_V,\tilde{w}\in K_W}\int_\Omega H(x,t,f,\tilde{v},\tilde{w},\nabla p)\,dx$$

where f,v,w,p represent the output of Algorithm 4.1. The number $N_\%^l$, $l\in\mathbb{N}$, gives the percentage of grid points where $0 \le \triangle H(t) \le 10^{-l}$ is fulfilled.

For the first FP control problem (4.73), we obtain the optimal controls shown in Figure 4.4a. We can see the effect of the thresholds s_1 for v and s_2 for w. In Figure 4.4b, we plot the convergence history of the cost functional. The numerical PMP test for the initial guess gives $N_\%^1 = 100\%$, $N_\%^2 = 100$, $N_\%^3 = 32\%$, $N_\%^4 = 24\%$, $N_\%^5 = 20\%$, $N_\%^6 = 17\%$, $N_\%^{12} = 5\%$ and for the result of Algorithm 4.1, the numerical PMP test gives $N_\%^1 = 100\%$, $N_\%^2 = 100\%$, $N_\%^3 = 97\%$, $N_\%^4 = 81\%$, $N_\%^5 = 72\%$, $N_\%^6 = 67\%$ and $N_\%^{12} = 67\%$. These results indicate that the solution obtained with the SQH method is PMP optimal in the sense of (4.78).



(a) The time curves of the components of the controls.

(b) Reduction of the value of the cost functional during the SQH iterative process.

Figure 4.4: Results of the first numerical experiment.

In correspondence to the controls v, w depicted in Figure 4.4a, we perform a Monte-Carlo simulation with the stochastic process (4.61) driven by these controls, starting from $(x_1, x_2) = (-2.1, -3.1)$. The resulting paths are plotted in Figure 4.5 where we see that the mean value of the state f (trajectory with circles) is steered towards the desired trajectory x_d (dashed line) and starts following it when coming close. Similarly, we see that the stochastic trajectories (solid lines) are close to the mean value of f.

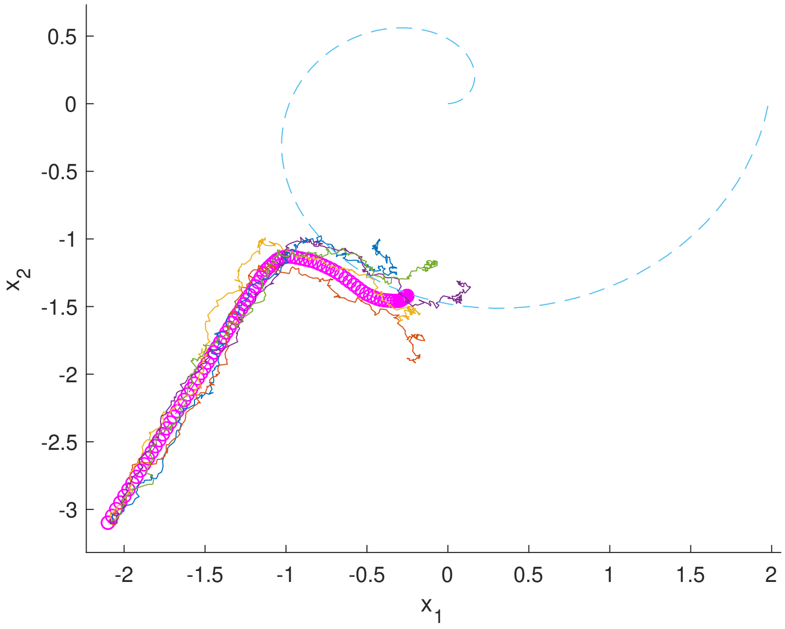
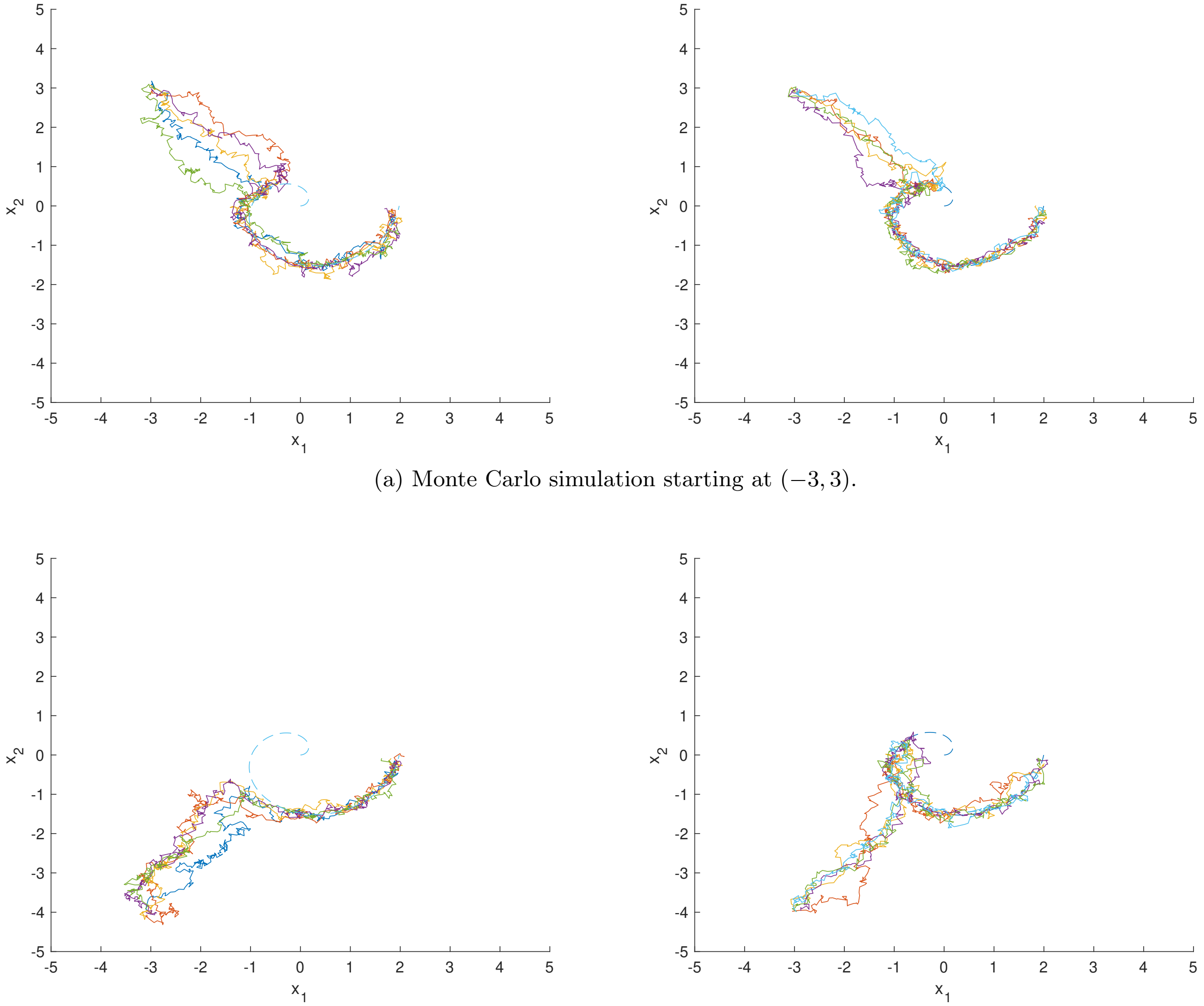


Figure 4.5: Monte-Carlo simulation (solid lines) with the controls of Figure 4.4a. The circles correspond to the mean value of f and the dashed line is the desired trajectory x_d.

In the next experiment, we consider our second control problem (4.74) to compute optimal control fields that are used in a Monte-Carlo simulation with (4.62). For comparison, this problem is solved using both the SQH and the DH scheme. The resulting controls are implemented in the corresponding stochastic process (4.62) where we consider different initial conditions in Ω. In Figure 4.6, we plot some stochastic trajectories obtained in this framework and using the controls resulting from the SQH scheme and DH scheme. In any case, we can see that the controlled stochastic trajectories are steered towards our desired trajectory x_d.



(a) Monte Carlo simulation starting at $(-3, 3)$.

(b) Monte Carlo simulation starting at $(-3, -4)$.

Figure 4.6: Monte-Carlo simulation with the controls obtained with the DH method (left) and with the SQH method (right).

Chapter 5

Appendix

In the Appendix we give supplementary results that are used in the thesis or have the purpose to support the discussion in the thesis.

5.1 General auxiliary results

Let $\mathcal{Z} \subseteq \mathbb{R}^N$ be a set with $N \in \mathbb{N}$. The following lemma states that the composition of a Lebesgue measurable function $u : \mathcal{Z} \to \mathbb{R}$ with a lower semi-continuous function $g : \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable. The result can also be found in [21].

Lemma 51. *Let $u : \mathcal{Z} \to \mathbb{R}^n$, $n \in \mathbb{N}$ be Lebesgue measurable and $g : \mathbb{R}^n \to \mathbb{R}$ be lower semi-continuous. Then the composition $g \circ u : \mathcal{Z} \to \mathbb{R}$ is Lebesgue measurable.*

Proof. By [36, Example 2.6.3] and [36, Example 2.6.5], we have that u is Lebesgue measurable if and only if each component function $u_i : (\mathcal{Z}, \mathcal{M}) \to (\mathbb{R}^n, \mathcal{B})$, $i \in \{1, ..., n\}$ is measurable where $(\mathcal{Z}, \mathcal{M})$ is a measurable space [36, page 8], $\mathcal{M}$ is the σ-algebra of the Lebesgue measurable subsets of $\mathcal{Z}$ and $(\mathbb{R}^n, \mathcal{B})$ is a measurable space where $\mathcal{B}$ is the σ-algebra generated by the collection of open subsets of $\mathbb{R}^n$.

Next, we show that $g : (\mathbb{R}^n, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ is measurable, that means Borel measurable. We define for any constant $c \in \mathbb{R}$ the set

$$A := \{z \in \mathbb{R}^n |\ g(z) \leq c\}.$$

Let $(z_m)_{m \in \mathbb{N}} \subseteq A$ be a sequence with $\lim_{m \to \infty} z_m = \bar{z}$, then

$$c \geq \liminf_{m \to \infty} g(z_m) \geq g(\bar{z}),$$

see [43, Theorem 3.127] for the calculation rules of lim inf. This means that $\bar{z} \in A$ and thus A is closed. By [36, Proposition 1.1.4] we know that A belongs to $\mathcal{B}$ and thus by [36, Proposition 2.1.1 and page 42], we have that g is Borel measurable. Then with [36, Proposition 2.6.1], we have that $g \circ u : (\mathcal{Z}, \mathcal{M}) \to (\mathbb{R}, \mathcal{B})$ is measurable, which means $g \circ u$ is Lebesgue measurable. □

Lemma 52. *Let $f : \mathbb{R}^n \to \mathbb{R}^l$, $x \mapsto f(x)$ be a measurable function with $n, l \in \mathbb{N}$. Then the function $\tilde{f} : \mathbb{R}^{n+m} \to \mathbb{R}^l$, $(x, y) \mapsto \tilde{f}(x, y) := f(x)$ is measurable.*

Proof. The function $\tilde{f}$ is measurable if and only if each component function $\tilde{f}_i$, $i \in \{1, ..., l\}$ is measurable, see [5, X Theorem 1.7 ii)]. The set

$$\left\{(x, y) \in \mathbb{R}^{n+m} |\ \tilde{f}_i(x, y) < t\right\} = \left\{(x, y) \in \mathbb{R}^{n+m} |\ f_i(x) < t\right\} = \{x \in \mathbb{R}^n |\ f_i(x) < t\} \times \mathbb{R}^m$$

is given by $A_i \times \mathbb{R}^m$ for any $t \in \mathbb{R}$ where

$$\{x \in \mathbb{R}^n |\ f_i(x) < t\} =: A_i \subseteq \mathbb{R}^n$$

is a measurable set for any $i \in \{1, ..., l\}$ according to [5, X Theorem 1.9]. According to [5, IX Corollary 1.18] we have that $A \times \mathbb{R}^m$ is a measurable set of $\mathbb{R}^{n+m}$. Thus the function $\tilde{f}$ is measurable, see [5, X Theorem 1.7 ii)], since any component function is measurable, see [5, X Theorem 1.9]. □

Next we discuss if the control function in Step 2 of Algorithm 2.1 is Lebesgue measurable. The following discussion also holds for all the other SQH formulations in this thesis. The Lebesgue measurability of the controls certainly holds in the case if the function $(z, u) \mapsto K_\epsilon (z, y (z), u, v (z), p (z))$ is Lebesgue measurable in $z \in \mathcal{Z}$ for each $u \in K_U$ and continuous in u for each $z \in \mathcal{Z}$, see [82, 14.29 Example, 14.37 Theorem]. An example for this case is

$$K_\epsilon^1 (z, y (z), u, v (z), p (z)) := (y (z) - y_d (z))^2 + \frac{\alpha}{2} u^2 + \beta |u| + p (z) u + \epsilon (u - v (z))^2$$

which is the corresponding augmented Hamiltonian for an optimal control problem with distributed control, tracking term and an L^2- and L^1-cost term for the control where $y_d \in L^2 (\mathcal{Z})$ and $\alpha, \beta > 0$.

If K_ϵ is only lower semi-continuous in $u \in K_U$ for each $z \in \mathcal{Z}$ where K_U is a compact interval containing $\overline{u}$ as the upper and $\underline{u}$ as the lower bound, then, in general, we cannot guarantee that u is Lebesgue measurable, see the paragraph following [82, 14.28 Proposition]. However, in the case of

$$K_\epsilon^2 (z, y (z), u, v (z), p (z)) := h (y (z)) + \frac{\alpha}{2} u^2 + g (u) + p (z) f (z, y (z), u) + \epsilon (u - v (z))^2$$

where $\alpha \geq 0$, $z \mapsto h (y (z))$ and $z \mapsto f (z, y (z), u)$ is Lebesgue measurable, the partial derivative $f_u (z, y) := \frac{\partial}{\partial u} f (z, y, u)$ of f with respect to u is independent of u and

$$g (u) := \begin{cases} \gamma |u| & \text{for } |u| > s \\ 0 & \text{otherwise} \end{cases}, \quad s, \gamma > 0,$$

we prove that starting our SQH scheme with an initial guess u^0 that is Lebesgue measurable, we obtain that any u is Lebesgue measurable as follows.

The augmented Hamiltonian $K_\epsilon^2 (z, y, u, v, p)$ is minimized as follows. Its minimum, denoted by u, can exactly be given by a case study according to [3, IV Remark 2.2 (b)] as follows.

If $-s \leq u \leq s$, we have the minimum at

$$u_1 (z) := \min \left(\max \left(-s, \frac{2\epsilon v (z) - p (z) f_u (z, y)}{2\epsilon + \alpha} \right), s \right).$$

If $s < u \leq \overline{u}$, we have the minimum at

$$u_2 (z) := \min \left(\max \left(s, \frac{2\epsilon v (z) - (p (z) f_u (z, y) + \gamma)}{2\epsilon + \alpha} \right), \overline{u} \right).$$

If $\underline{u} \leq u < -s$, we have the minimum at

$$u_3 (z) := \min \left(\max \left(\underline{u}, \frac{2\epsilon v (z) - (p (z) f_u (z, y) - \gamma)}{2\epsilon + \alpha} \right), -s \right).$$

Then the minimum of K_ϵ^2 over K_U is given by

$$u (z) = \operatorname*{arg\,min}_{w \in K_U} K_\epsilon^2 (z, y (z), w, v (z), p (z)) = \operatorname*{arg\,min}_{w \in \{u_1, u_2, u_3\}} K_\epsilon^2 (z, y (z), w, v (z), p (z)).$$

Next, we prove that u, as a function, is Lebesgue measurable assuming that the last iterate represented by v is also Lebesgue measurable. Notice that the adjoint variable p, as a solution of a well-posed equation, is always Lebesgue measurable in this thesis. Thus, we have that $z \mapsto u_1(z)$, $z \mapsto u_2(z)$ and $z \mapsto u_3(z)$ are Lebesgue measurable functions, see [36, Proposition 2.1.4, Proposition 2.1.7]. Further, we have that

$$K_\epsilon^1(z) := K_\epsilon^2(z, y(z), u_1(z), v(z), p(z)),\ K_\epsilon^2(z) := K_\epsilon^2(z, y(z), u_2(z), v(z), p(z))$$

and

$$K_\epsilon^3(z) := K_\epsilon^2(z, y(z), u_3(z), v(z), p(z))$$

are Lebesgue measurable according to Lemma 51 and because the sum and the product of Lebesgue measurable functions are Lebesgue measurable, see [36, Proposition 2.1.7].

Now, the function $z \mapsto u(z)$ is given by

$$u(z) := \begin{cases} u_1(z) & \text{if } K_\epsilon^1(z) \leq K_\epsilon^2(z) \text{ and } K_\epsilon^1(z) \leq K_\epsilon^3(z) \\ u_2(z) & \text{if } K_\epsilon^2(z) \leq K_\epsilon^3(z) \text{ and } K_\epsilon^2(z) < K_\epsilon^1(z) \\ u_3(z) & \text{if } K_\epsilon^3(z) < K_\epsilon^2(z) \text{ and } K_\epsilon^3(z) < K_\epsilon^1(z) \end{cases}$$

for the following reason. There are three cases. First, the value of K_ϵ for the corresponding control is strictly the minimum. In this case, the corresponding branch is taken to set the value of $u(z)$. Second, it is $K_\epsilon^1(z) = K_\epsilon^2(z) = K_\epsilon^3(z)$. In this case, we have $u(z) = u_1(z)$. Third, we have that two values of K_ϵ^i, $i \in \{1, 2, 3\}$ are equal and are strictly smaller than the third one. Then we have three sub cases. First, $K_\epsilon^1(z) = K_\epsilon^2(z)$, that means $K_\epsilon^1(z) \leq K_\epsilon^2(z)$ and $K_\epsilon^1(z) < K_\epsilon^3(z)$ which is covered by the first branch $u(z) = u_1(z)$. Second, $K_\epsilon^2(z) = K_\epsilon^3(z)$, that means that $K_\epsilon^2(z) \leq K_\epsilon^3(z)$ and $K_\epsilon^2(z) < K_\epsilon^1(z)$ which is covered by the second branch $u(z) = u_2(z)$. Third, $K_\epsilon^1(z) = K_\epsilon^3(z)$, that means $K_\epsilon^1(z) \leq K_\epsilon^3(z)$ and $K_\epsilon^1(z) < K_\epsilon^2(z)$ which is covered by the first branch $u(z) = u_1(z)$. According to [36, Proposition 2.1.1] and the following paragraph, u is Lebesgue measurable if and only if the set $\{z \in \mathcal{Z} |\ u(z) > c\}$ is Lebesgue measurable for any $c \in \mathbb{R}$. To show this fact, notice that the following holds

$$\begin{aligned} &\{z \in \mathcal{Z} |\ u(z) > c\} \\ &= \left(\{z \in \mathcal{Z} |\ u_1(z) > c\} \cap \left\{z \in \mathcal{Z} |\ K_\epsilon^1(z) \leq K_\epsilon^2(z)\right\} \cap \left\{z \in \mathcal{Z} |\ K_\epsilon^1(z) \leq K_\epsilon^3(z)\right\}\right) \\ &\cup \left(\{z \in \mathcal{Z} |\ u_2(z) > c\} \cap \left\{z \in \mathcal{Z} |\ K_\epsilon^2(z) \leq K_\epsilon^3(z)\right\} \cap \left\{z \in \mathcal{Z} |\ K_\epsilon^2(z) < K_\epsilon^1(z)\right\}\right) \\ &\cup \left(\{z \in \mathcal{Z} |\ u_3(z) > c\} \cap \left\{z \in \mathcal{Z} |\ K_\epsilon^3(z) < K_\epsilon^2(z)\right\} \cap \left\{z \in \mathcal{Z} |\ K_\epsilon^3(z) < K_\epsilon^1(z)\right\}\right) \end{aligned} \tag{5.1}$$

Thus u is Lebesgue measurable, as the intersection and the union of finitely many Lebesgue measurable sets are Lebesgue measurable, see [5, IX Theorem 5.1, Remark 1.1], if and only if the single sets are measurable which in fact they are as follows. We have that the sets $\{z \in \mathcal{Z} |\ u_i(z) > c\}$, $i \in \{1, 2, 3\}$ are Lebesgue measurable for any $c \in \mathbb{R}$ as u_i is Lebesgue measurable for any $i \in \{1, 2, 3\}$. Further the sets $\left\{z \in \mathcal{Z} |\ K_\epsilon^i(z) \leq K_\epsilon^{\tilde{i}}(z)\right\}$ and $\left\{z \in Q |\ K_\epsilon^i(z) < K_\epsilon^{\tilde{i}}(z)\right\}$, $i, \tilde{i} \in \{1, 2, 3\}$ are Lebesgue measurable, see [36, Proposition 2.1.3].

In the case of a so-called L^0-norm where g is given by

$$g(u) := \begin{cases} \gamma & \text{if } |u| \neq 0 \\ 0 & \text{otherwise} \end{cases}, \quad \gamma > 0,$$

the calculation is analogous. The only difference is in giving the possible minima u_1, u_2 and u_3 what we do in the following.

If $0 < u \leq \overline{u}$ or $\underline{u} \leq u < 0$, we have the minimum at

$$u_1(z) := \min\left(\max\left(0, \frac{2\epsilon v(z) - p(z) f_u(z, y)}{2\epsilon + \alpha}\right), \overline{u}\right),$$

$$u_2(z) := \min\left(\max\left(\underline{u}, \frac{2\epsilon v(z) - p(z) f_u(z, y)}{2\epsilon + \alpha}\right), 0\right)$$

or if $u = 0$, we have the minimum at

$$u_3(z) := 0.$$

The next lemma is a version of partial integration that we need for the thesis. In this lemma $\cdot$ denotes the Euclidean scalar product and ∇ the divergence or the gradient with respect to the Euclidean scalar product depending on if its argument is a vector-valued or real-valued function.

Lemma 53. *Let $\phi_1, ..., \phi_n, f \in H_0^1(\Omega, \mathbb{R})$. Then we have that*

$$\int_\Omega \nabla(\phi(x)) f(x)\, dx = -\int_\Omega \phi(x) \cdot \nabla f(x)\, dx$$

where $\phi := \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix}$.

Proof. We take a sequence $\|f_n - f\|_{H_0^1(\Omega)} \to 0$ for $n \to \infty$ with $f_n \in C_c^\infty(\Omega)$ for all $n \in \mathbb{N}$ where $C_c^\infty(\Omega)$ is the space of all arbitrarily differentiable functions on Ω with a compact support. This is always possible, as $H_0^1(\Omega)$ is the closure of $C_c^\infty(\Omega)$, see [26, 9.4]. By the definition of the weak derivative, see [26, 9.1], we have

$$\int_\Omega \nabla(\phi(x)) f_n(x)\, dx = \int_\Omega \sum_{i=1}^n (\partial_{x_i}\phi_i(x)) f_n(x)\, dx = -\int_\Omega \phi_i(x)\, \partial_{x_i} f_n(x)\, dx = -\int_\Omega \phi(x) \nabla f_n(x)\, dx.$$

Now by the Cauchy-Schwarz inequality [2, Lemma 2.2] and the Poincaré inequality [2, 6.7] we have that

$$\lim_{n\to\infty} \int_\Omega \nabla\phi(x)(f_n(x) - f(x))\, dx = 0$$

and

$$\lim_{n\to\infty} \int_\Omega \phi(x) \nabla(f_n(x) - f(x))\, dx = 0.$$

□

5.2 Ordinary differential equations

We give a result that ensures a unique global solution to an initial value problem of the form (2.1). According to [90, Theorem 54] and [90, Proposition C.3.8] we have the following theorem where the symbols are defined as in Chapter 2.

Theorem 54. *For any given $u \in U_{ad}$, let $f : [0,T] \times \mathbb{R}^n \times K_U \to \mathbb{R}^n$, $t \mapsto \tilde{f}(t,y) := f(t,y,u(t))$ be measurable for any fixed $y \in \mathbb{R}^n$. Furthermore, let a locally integrable function ϕ_1 exist such that*

$$\|\tilde{f}(t,y_1) - \tilde{f}(t,y_2)\| \le \phi_1(t)\|y_1 - y_2\| \tag{5.2}$$

holds for a norm defined on $\mathbb{R}^n$, for each $t \in [0,T]$ and any $y_1, y_2 \in \mathbb{R}^n$. If $\tilde{f}$ is locally integrable in t, that means that for each fixed y_0 there exists a locally integrable function $\phi_2 : [0,T] \to \mathbb{R}^+$ with

$$\|\tilde{f}(t,y_0)\| \le \phi_2(t) \tag{5.3}$$

for almost all t, then there is a unique absolutely continuous function $y : [0,T] \to \mathbb{R}^n$ fulfilling the integral equation

$$\begin{aligned} y(t) &= y_0 + \int_0^t \tilde{f}\left(\tilde{t}, y\left(\tilde{t}\right)\right) d\tilde{t}. \\ &= y_0 + \int_0^t f\left(\tilde{t}, y\left(\tilde{t}\right), u\left(\tilde{t}\right)\right) d\tilde{t}. \end{aligned}$$

By the virtue of Theorem 54 we now prove that (2.5) is uniquely solvable by the Assumption A.6) of Chapter 2.

Theorem 55. *The adjoint equation (2.5) with* $p(T) = (D_y F(y(T)))^T$ *has a unique absolutely continuous solution* $t \mapsto p(t)$ *on the interval* $[0,T]$ *where* y *is the solution to the initial value problem (2.1) for the corresponding* $u \in U_{ad}$.

Proof. By the transformation $\tau := T - t$, $\hat{p}(\tau) := p(T-\tau)$, $\hat{y}(\tau) := y(T-\tau)$, $\hat{u}(\tau) := u(T-\tau)$ and $\hat{f}_y(\tau, \hat{y}(\tau), \hat{u}(\tau)) := f_y(T-\tau, y(T-\tau), u(T-\tau))$ we obtain an initial value problem

$$\hat{p}'(\tau) = h_y(\hat{y}(\tau)) + \hat{f}_y(\tau, \hat{y}(\tau), \hat{u}(\tau))^T \hat{p}(\tau),$$

$\hat{p}(0) = (D_y F(\hat{y}(0)))^T$ with $\hat{p}'(\tau) := \frac{\partial}{\partial \tau}\hat{p}(\tau) = \frac{\partial}{\partial \tau} p(T-\tau) = -\frac{\partial}{\partial t} p(t) = -p'(t)$. Now we have that $\hat{y} : [0,T] \to \mathbb{R}^n$, $\tau \mapsto \hat{y}(\tau)$, $\hat{p} : [0,T] \to \mathbb{R}^n$, $\tau \mapsto \hat{p}(\tau)$, $\hat{u}[0,T] \to \mathbb{R}^m$, $\tau \mapsto \hat{u}(\tau)$, $\hat{f}_y : [0,T] \to \mathbb{R}^{n\times n}$, $\tau \mapsto \hat{f}_y(\tau, \hat{y}(\tau), \hat{u}(\tau))$ and $\hat{h}_y := h_y \circ \hat{y} : [0,T] \to \mathbb{R}^{n\times 1}$, $\tau \mapsto \hat{h}_y(\hat{y}(\tau))$ are still Lebesgue measurable due to the translation invariance of the Lebesgue measure [5, IX Corollary 5.23] and that continuous functions are Lebesgue measurable [36, page 42]. Consequently the right hand-side of the adjoint equation

$$\tau \mapsto \Xi(\tau, \hat{p}) := h_y(\hat{y}(\tau)) + \hat{f}_y(\tau, \hat{y}(\tau), \hat{u}(\tau))^T \hat{p}$$

is Lebesgue measurable for each fixed $\hat{p} \in \mathbb{R}^n$, as the product and sum of Lebesgue measurable functions is Lebesgue measurable [36, Proposition 2.1.7]. We show that the rest of the requirements of Theorem 54 are fulfilled. The Lipschitz continuity (5.2) results from Assumption A.6) as follows. We have

$$\|\Xi(\tau, \hat{p}_1) - \Xi(\tau, \hat{p}_2)\| = \|\hat{f}_y(\tau, \hat{y}, \hat{u})(\hat{p}_1 - \hat{p}_2)\| \leq c_1 \left(\max_{i=1,\dots,n} \sum_{l=1}^{n} |\frac{\partial}{\partial y_l} \hat{f}_i(\tau, \hat{y}(\tau), \hat{u}(\tau))| \right) \|\hat{p}_1 - \hat{p}_2\|$$

with $c_1 > 0$ by the equivalence of norms of a finite vector space [75, Theorem 4.9] and the formula for the L^1 matrix norm, see [39, page 22 and 23]. As the absolute value of a Lebesgue measurable function is Lebesgue measurable [36, page 46] and the $\max(\cdot,\cdot)$ of two Lebesgue measurable functions is Lebesgue measurable [36, Proposition 2.1.4], we have by the Lebesgue measurability of $\tau \mapsto \frac{\partial}{\partial y_l}\hat{f}_i(\tau, \hat{y}(\tau), \hat{u}(\tau))$ for all $i, l \in \{1, ..., n\}$ and by the Lebesgue measurability of sums [36, Proposition 2.1.7] that (5.2) holds for

$$\phi_1(\tau) := c_1 \left(\max_{i=1,\dots,n} \sum_{l=1}^{n} \frac{\partial}{\partial y_l} \hat{f}_i(\tau, \hat{y}(\tau), \hat{u}(\tau)) \right).$$

The integrability of ϕ_1 holds due to the boundedness assumption A.6) of $\frac{\partial}{\partial y_l}\hat{f}_i(\tau, \hat{y}, \hat{u})$ for all $i, l \in \{1, ..., n\}$. The calculation for (5.3) is similar due to the linearity of Ξ in $\hat{p}$. For fixed $\hat{p}$, we have with [39, Definition 1.7] that

$$\begin{aligned}\|h_y(\hat{y}(\tau)) + \hat{f}_y(\tau, \hat{y}(\tau), \hat{u}(\tau))^T \hat{p}\| &\leq \|h_y(\hat{y}(\tau))\| + \|\hat{f}_y(\tau, \hat{y}(\tau), \hat{u}(\tau))^T \hat{p}\| \\ &\leq c_2 \left(\sum_{i=1}^{n} |\frac{\partial}{\partial y_i} h(\hat{y}(\tau))| + \max_{i=1,\dots,n} \sum_{l=1}^{n} |\frac{\partial}{\partial y_l} \hat{f}_i(\tau, \hat{y}(\tau), \hat{u}(\tau))| \right)\end{aligned}$$

with $c_2 > 0$ where

$$\phi_2(\tau) := c_2 \left(\sum_{i=1}^{n} |\frac{\partial}{\partial y_i} h(\hat{y}(\tau))| + \max_{i=1,\dots,n} \sum_{l=1}^{n} |\frac{\partial}{\partial y_l} \hat{f}_i(\tau, \hat{y}(\tau), \hat{u}(\tau))| \right)$$

is measurable and integrable with analogous arguments as for (5.2). However, we remark that for the integrability of $\tau \mapsto |\frac{\partial}{\partial y_i} h(\hat{y}(\tau))|$ Assumption A.6) is not necessary. The measurability and the integrability follow from [36, Theorem 2.5.4] due to the continuity of $\frac{\partial}{\partial y_i} h$ for all $i \in \{1, ..., n\}$, because of the continuity of the solution to (2.1) and that the composite function of continuous functions is continuous [3, III Theorem 1.8]. Thus the composition is measurable since continuous functions are Lebesgue measurable [36, page 42] and is bounded on the compact interval $[0, T]$, see [3, III Corollary 3.8]. By Theorem 54 and the backsubstitution $p(t) = p(T - \tau) = \hat{p}(\tau)$ we obtain the unique absolutely continuous solution to (2.5). □

Lemma 56. *The intermediate adjoint equation (2.7) with $\tilde{p}(T) = \tilde{F}(y_1, y_2)$ has a unique absolutely continuous solution $t \mapsto \tilde{p}(t)$ on the interval $[0, T]$ where y_1, y_2 are the solutions to the initial value problem (2.1) for the corresponding $u_1, u_2 \in U_{ad}$.*

Proof. The proof is basically the same as for Theorem 55. We first have to check that

$$\tilde{F}_i(y_1, y_2) := \int_0^1 \frac{\partial}{\partial y_l} F(y)\,|_{y = y_2(T) + \theta(y_1(T) - y_2(T))} d\theta,$$

$$\tilde{f}_{il}(t, y_1, y_2, u_1) := \int_0^1 \frac{\partial}{\partial y_l} f_i(t, y, u_1(t))\,|_{y = y_2(t) + \theta(y_1(t) - y_2(t))} d\theta$$

and

$$\tilde{h}_i(y_1, y_2) := \int_0^1 \frac{\partial}{\partial y_l} h(y)\,|_{y = y_2(t) + \theta(y_1(t) - y_2(t))} d\theta,$$

$i, l \in \{1, ..., n\}$ are well defined for any $t \in [0, T]$. This is the case since the functions

$$\theta \mapsto \frac{\partial}{\partial y_l} F(y)\,|_{y = y_2(T) + \theta(y_1(T) - y_2(T))},$$

$$\theta \mapsto \frac{\partial}{\partial y_l} f_i(t, y, u_1(t))\,|_{y = y_2(t) + \theta(y_1(t) - y_2(t))}$$

and

$$\theta \mapsto \frac{\partial}{\partial y_l} h(y)\,|_{y = y_2(t) + \theta(y_1(t) - y_2(t))}$$

are continuous and composite functions of continuous functions are continuous [3, III Theorem 1.8], thus bounded [3, III Corollary 3.8] and integrable [36, Theorem 2.5.4] where we use [10, Proposition 3.4] which gives that $y_2(t) + \theta(y_1(t) - y_2(t)) \in I$ for any $\theta \in [0, 1]$.

Next we have to see that $t \mapsto \tilde{h}(y_1(t), y_2(t))$ and $t \mapsto \tilde{f}(t, y_1(t), y_2(t))$ are measurable. According to Lemma 52, we have that the function $(\theta, t) \mapsto \theta$, $(\theta, t) \mapsto t$, $(\theta, t) \mapsto y_1(t)$, $(\theta, t) \mapsto y_2(t)$ and $(\theta, t) \mapsto u_1(t)$ is Lebesgue measurable on $[0, 1] \times [0, T]$. As the sum and the product of measurable functions is measurable [36, Proposition 2.1.7], applied to each component [36, Example 2.6.5], we have that $(\theta, t) \mapsto \frac{\partial}{\partial y_l} f_i(t, y, u_1(t))\,|_{y = y_2(t) + \theta(y_1(t) - y_2(t))}$ and $(\theta, t) \mapsto \frac{\partial}{\partial y_l} h(y)\,|_{y = y_2(t) + \theta(y_1(t) - y_2(t))}$ are measurable on $[0, 1] \times [0, T]$.

Next, by [36, Proposition 2.1.4, page 46], Tonelli's Theorem [5, X Theorem 6.7 ii)] and that the sum of measurable functions is measurable [36, Proposition 2.1.7], we have that $t \mapsto \tilde{h}_i(y_1(t), y_2(t))$ and $t \mapsto \tilde{f}_{il}(t, y_1(t), y_2(t))$ is Lebesgue measurable. We remark that $t \mapsto \tilde{h}(y_1(t), y_2(t))$ is integrable for any continuous y_1 and y_2 since $(\theta, t) \mapsto h_y(y_2(t) + \theta(y_1(t) - y_2(t)))$ is continuous [3, III Theorem 1.5, Theorem 1.8], thus bounded [3, III Corollary 3.8] on the compact set $[0, 1] \times [0, T]$ and Lebesgue integrable. By Fubini's Theorem [5, X Theorem 6.9], we have that $t \mapsto \tilde{h}_i(y_1(t), y_2(t))$. The integrability of $t \mapsto \tilde{f}_{il}(t, y_1(t), y_2(t))$ follows from the boundedness in Assumption A.6). □

For the characterization of a solution to the optimal control problem (2.3) Gronwall's lemma is extensively used. For this purpose, we write it down from [94, Lemma 2.7].

Lemma 57. *Suppose that $\psi : \mathbb{R} \to \mathbb{R}$ fulfills*

$$\psi(t) \leq \mathit{\Lambda}(t) + \int_0^t \mathit{\Gamma}(\tilde{t})\, \psi(\tilde{t})\, d\tilde{t},\ t \in [0,T]$$

with $\mathit{\Lambda}(t) \in \mathbb{R}$ and $\mathit{\Gamma}(t) \geq 0$. Then it holds

$$\psi(t) \leq \mathit{\Lambda}(t) + \int_0^t \mathit{\Lambda}(\tilde{t})\, \mathit{\Gamma}(\tilde{t}) \exp\left(\int_s^t \mathit{\Gamma}(\hat{t})\, d\hat{t}\right) d\tilde{t}$$

with $t \in [0,T]$.

A further useful result for technical calculations for the used solution concept of ODEs where a solution is only almost everywhere differentiable is the following formulation of the fundamental theorem of calculus, see [79, Theorem 6.3.6].

5.3 Partial differential equations

In this section, we provide basic results that are used for the analysis of the PMP and the SQH method in a partial differential equations (PDEs) framework. We start with a general result and give specific results in the corresponding subsection about elliptic PDEs or parabolic PDEs, respectively.

The following lemma is used for both elliptic and parabolic PDE analysis. This lemma is proved in [100, Lemma 4.1.1].

Lemma 58. *Let $\varphi(t)$ be a non-negative and non-increasing function on $[k_0,\infty)$ satisfying*

$$\varphi(m) \leq \left(\frac{M}{m-k}\right)^{\alpha} (\varphi(k))^{\beta},\ \forall m > k \geq k_0$$

for some constants $M > 0$, $\alpha > 0$ and $\beta > 1$. Then there exists a $d > 0$ such that $\varphi(m) = 0$ for all $m \geq k_0 + d$. It is sufficient for this statement to choose $d := M 2^{\frac{\beta}{\beta-1}} (\varphi(k_0))^{\frac{\beta-1}{\alpha}}$.

5.3.1 Elliptic partial differential equations

In this subsection, we prove an L^∞-result for elliptic partial differential equations that is essential for this thesis. Specifically in the PMP framework including the numerical treatment with the SQH method, the L^∞ boundedness of the solution to the corresponding PDE is crucial.

Let $\Omega \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$ be an open set. We provide a results for the elliptic PDEs that is useful for this purpose. We have the following

$$\begin{aligned} B(y,v) &= (h,v) \text{ in } \Omega \\ y &= 0 \text{ on } \partial\Omega \end{aligned} \tag{5.4}$$

for all $v \in H_0^1(\Omega)$ where $B(y,v) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is a bilinear map with the coercivity condition

$$\beta \|y\|^2_{H_0^1(\Omega)} \leq B(y,y),\ \beta > 0$$

and

$$B(-k,v) \leq 0 \text{ for } k \geq 0$$

if $v \geq 0$ and $h \in L^q(\Omega)$, $q \geq \frac{n}{2} + 1$. We assume that (5.4) has a unique solution $y \in H_0^1(\Omega)$. Then the following theorem holds.

Theorem 59. *If there exists a unique solution to (5.4), then the initial value problem (5.4) has an essentially bounded solution for which it holds*

$$\|y\|_{L^\infty(\Omega)} \leq C\|h\|_{L^q(\Omega)}$$

where $C > 0$.

Proof. The proof is based on [100, Theorem 4.2.1] or [22]. We assume that h is not the zero function. In the case of $h = 0$, the solution $y = 0$ solves (5.4) and thus the statement of the theorem is true. We choose the constant $k \geq 0$. We have that $y - k \in H^1(\Omega)$ due to the linearity of the weak differentiation operation [45, 5.2 Theorem 1] and that the derivative of the constant function k fulfills the definition of weak derivative [45, page 243]. Furthermore there is no other weak derivative because of the uniqueness of the weak derivative [45, page 243] and therefore we have that

$$(y-k)_+ := \max(y-k, 0) \in H_0^1(\Omega),$$

see [38, Chapter 4, Proposition 6]. Then, we choose $v = (y-k)_+$ in (5.4) and obtain the following

$$B\left(y-k, (y-k)_+\right) \leq \left(h, (y-k)_+\right)$$

where we use that

$$B\left(y, (y-k)_+\right) \geq B\left(y, (y-k)_+\right) + B\left(-k, (y-k)_+\right) = B\left(y-k, (y-k)_+\right)$$

and thus

$$\beta \| (y-k)_+ \|^2_{H_0^1(\Omega)} \leq \left(h, (y-k)_+\right) \tag{5.5}$$

as $(y-k)_+ = 0$ if $y - k \leq 0$ and $B\left(y-k, (y-k)_+\right) = B\left((y-k)_+, (y-k)_+\right)$ if $y - k > 0$ and

$$\beta \| (y-k)_+ \|^2_{H_0^1(\Omega)} \leq B\left((y-k)_+, (y-k)_+\right)$$

due to the coercivity assumption for B. We remark that the function $(y-k)_+ \in H_0^1(\Omega)$ is also an element of $L^p(\Omega)$ with $\| (y-k)_+ \|_{L^p(\Omega)} \leq M \| (y-k)_+ \|_{H_0^1(\Omega)}$, $M > 0$ where

$$2 \leq p \begin{cases} \leq \infty & \text{for } n = 1 \\ < \infty & \text{for } n = 2 \\ \leq \frac{2n}{n-2} & \text{for } n \geq 3 \end{cases}, \tag{5.6}$$

see the Sobolev embedding theorem [1, Theorem 4.12], especially [1, Theorem 4.12 Part III]. This implies

$$\| (y-k)_+ \|^2_{L^p(\Omega)} \leq \tilde{\beta} \int_\Omega h(x)(y-k)_+(x)\, dx \tag{5.7}$$

with $\tilde{\beta} > 0$. Next, we define

$$A_k := \{x \in \Omega |\ y(x) > k\}$$

which is measurable, see [5, X Theorem 1.9] and $|A_k(t)|$ is the measure of $A_k(t)$. Due to $(y(x) - k)_+ = 0$ for $x \in \Omega \backslash A_k$, we consequently have from (5.7) the following

$$\| (y-k)_+ \|^2_{L^p(A_k)} \leq \tilde{\beta} \int_{A_k} h(x)(y-k)_+(x)\, dx. \tag{5.8}$$

In the next step, we have the estimate by Hölder's inequality, see [5, X Theorem 4.2]

$$\| (y-k)_+ \|^2_{L^p(A_k)} \leq \tilde{\beta} \left(\int_{A_k} |h(x)|^{\frac{n}{2}+1} dx \right)^{\frac{1}{\frac{n}{2}+1}} \left(\int_{A_k} (y-k)_+^{\frac{2+n}{n}}(x)\, dx \right)^{\frac{n}{2+n}}$$

which can be applied since $(y-k)_+ \in L^{\frac{n+2}{n}}(\Omega)$. This is true because in the case $n=1$ and $n=2$, we have $(y-k)_+ \in L^p$, $2 \le p < \infty$ and in the case $n \ge 3$, we have $\frac{2n}{n-2} \ge \frac{2+n}{n}$, which is true for all $n \ge 3$ since equivalently $n^2 \ge -4$, and with the L^p-embedding [1, Theorem 2.14] consequently

$$\| (y-k)_+ \|^2_{L^p(A_k)} \le \tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} \left(\int_{A_k} (y-k)_+^{\frac{2+n}{n}}(x)\, dx \right)^{\frac{n}{2+n}}. \tag{5.9}$$

We apply Hölder's inequality again with $\frac{1}{\tilde{p}} + \frac{1}{\tilde{q}} = 1$, thus for a given $\tilde{p}$ we have $\tilde{q} = \frac{\tilde{p}}{\tilde{p}-1}$, and we obtain the following

$$\| (y-k)_+ \|^2_{L^p(A_k)} \le \tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} \left(\int_{A_k} 1dx \right)^{\frac{\tilde{p}-1}{\tilde{p}} \frac{n}{2+n}} \left(\int_{A_k} (y-k)_+^{\frac{2+n}{n}\tilde{p}}(x)\, dx \right)^{\frac{n}{\tilde{p}(2+n)}} \tag{5.10}$$

We choose $p = \frac{2+n}{n}\tilde{p}$ and conclude from (5.10) for $\| (y-k)_+ \|_{L^{\frac{2+n}{n}\tilde{p}}(A_k)} > 0$ the following

$$\left(\int_{A_k} | (y-k)_+(x) |^{\frac{2+n}{n}\tilde{p}} dx \right)^{\frac{n}{\tilde{p}(2+n)}} = \| (y-k)_+ \|_{L^{\frac{2+n}{n}\tilde{p}}(A_k)} \le \tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} \left(\int_{A_k} 1dx \right)^{\frac{\tilde{p}-1}{\tilde{p}} \frac{n}{2+n}}, \tag{5.11}$$

which is also true in the case $\| (y-k)_+ \|_{L^{\frac{2n}{n-2}}(A_k)} = 0$.

Furthermore, for $m > k$, we have that $A_m \subseteq A_k$. Additionally it is $y > m$ on A_m and thus $y \ge y-k > m-k$ on A_m due to $k \ge 0$. Since $y-k = (y-k)_+$ on A_m, we obtain

$$\int_{A_k} | (y-k)_+(x) |^{\frac{2+n}{n}\tilde{p}} dx \ge \int_{A_m} (y-k)^{\frac{2+n}{n}\tilde{p}}(x)\, dx \ge (m-k)^{\frac{2+n}{n}\tilde{p}} \int_{A_m} 1dx. \tag{5.12}$$

We combine (5.12) with (5.11) and obtain $(m-k)\, |A_m|^{\frac{n}{\tilde{p}(2+n)}} \le \tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} |A_k|^{\frac{\tilde{p}-1}{\tilde{p}} \frac{n}{2+n}}$ and equivalently

$$|A_m| \le \left(\frac{\tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)}}{m-k} \right)^{\frac{2+n}{n}\tilde{p}} |A_k|^{\tilde{p}-1}. \tag{5.13}$$

In order to apply Lemma 58, we need that $\tilde{p}-1 > 1$ and that p fulfills (5.6). For the case $n=1$ and $n=2$ we can choose any $\tilde{p} > 2$, for example $\tilde{p} = 3$. For the case that $n \ge 3$, we have to ensure that $\frac{2+n}{n}\tilde{p} \le \frac{2n}{n-2}$, which is $\tilde{p} \le \frac{2n^2}{n^2-4}$. Since the expression $\frac{2n^2}{n^2-4} > 2$ for $n \ge 3$ is equivalent to $0 > -8$ and thus always true, we can choose $\tilde{p} = \frac{2n^2}{n^2-4}$ in the case of $n \ge 3$. Then we also have that $\frac{2+n}{n}\tilde{p} > 0$. By applying Lemma 58, we obtain that $|A_m| = 0$ for $m \ge \tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} 2^{\frac{\tilde{p}-1}{\tilde{p}-2}} |\Omega|^{\frac{n(\tilde{p}-2)}{\tilde{p}(2+n)}}$ where $|\Omega|$ is the measure of Ω. This means that the set where

$$y > \tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} 2^{\frac{\tilde{p}-1}{\tilde{p}-2}} |\Omega|^{\frac{n(\tilde{p}-2)}{\tilde{p}(2+n)}}$$

is of measure zero. With the same arguments, we have for $(y+k)_- := \min(y+k, 0)$ and $A_k := \{x \in \Omega |\ y < -k\}$ that the set where $y < -\tilde{\beta} \|h\|_{L^{\frac{n}{2}+1}(\Omega)} 2^{\frac{\tilde{p}-1}{\tilde{p}-2}} |\Omega|^{\frac{n(\tilde{p}-2)}{\tilde{p}(2+n)}}$ is of measure zero. Consequently, we obtain that $\|y\|_{L^\infty(\Omega)} \le C \|h\|_{L^{\frac{n}{2}+1}(\Omega)}$ with $C := \tilde{\beta} 2^{\frac{\tilde{p}-1}{\tilde{p}-2}} |\Omega|^{\frac{n(\tilde{p}-2)}{\tilde{p}(2+n)}}$. □

For $P.3$) this result above holds also immediately if we assume $K_U \subseteq \mathbb{R}_0^+$ because then we have that $-ukv \le 0$ for $v \ge 0$ and we can continue the proof of Theorem 59 from (5.5) to obtain the following theorem.

Theorem 60. *For the solution y to a bilinear elliptic boundary value problem as in P.3) with $K_U \subseteq \mathbb{R}_0^+$, we have*

$$\|y\|_{L^\infty(\Omega)} \leq d\|\tilde{f}\|_{L^q(\Omega)}$$

with $d > 0$ for any right-hand side $\tilde{f} \in L^q(\Omega)$.

Remark 61. For P.5), we have the following consideration such that the proof of Theorem 59 can be followed from (5.5) in order to obtain a corresponding boundedness result as in Theorem 59 for P.5). We have that

$$\left(\nabla y, \nabla (y-k)_+\right) \leq \left(\nabla y, \nabla (y-k)_+\right) + \left(y^3, (y-k)_+\right)$$

as $y^3 (y-k)_+ \geq 0$ due to $(y-k)_+ = 0$ if $y \leq k \geq 0$. With similar considerations the $L^\infty(\Omega)$-result is proved for (P.7)) as $\max(0, y) \geq 0$.

5.3.2 Parabolic partial differential equations

In this subsection, which is based on the appendix of [21] we prove an L^∞-result that is essential in the Pontryagin maximum principle framework of this thesis. We take the following framework

$$\begin{aligned} \left(y'(\cdot,t), v\right) + B(y, v; t) = (h(\cdot,t), v) \text{ in } \Omega \times (0,T) \\ y = 0 \text{ on } \partial\Omega \times [0,T] \\ y = y_0 \text{ on } \Omega \times \{0\} \end{aligned} \tag{5.14}$$

for all $v \in H_0^1(\Omega)$with bounded $\Omega \subseteq \mathbb{R}^n$, $T > 0$ and $y'(\cdot,t) := \frac{\partial}{\partial t} y(\cdot,t)$ where $B(y,v;t) : H_0^1(\Omega) \times H_0^1(\Omega) \times \mathbb{R}_0^+ \to \mathbb{R}$ is a bilinear map with the coercivity condition

$$\beta \|y(\cdot,t)\|^2_{H_0^1(\Omega)} \leq B(y,y;t), \ \beta > 0$$

and

$$B(-k, v; t) \leq 0 \text{ for } k \geq 0$$

if $v \geq 0$ for any $t \in [0,T]$. Furthermore, we require that $h \in L^q(Q)$, $q > \frac{n}{2} + 1$ for $n \geq 2$ and $q \geq 2$ for $n = 1$, $y_0 \in L^\infty(\Omega)$ and that (5.14) has a unique solution fulfilling

$$y \in L^2\left(0,T; H_0^1(\Omega)\right) \cap L^\infty\left(0,T; L^2(\Omega)\right) \text{ and } y' \in L^2\left(0,T; H^{-1}(\Omega)\right)$$

such that (5.14) holds for almost all $t \in (0,T)$ and all $v \in H_0^1(\Omega)$, see [45, Chapter 7] for details. With the following lemmas, we prepare for the proof of Theorem 64 below. This result and a similar proof can be found in [76] or [64, Chapter 7 Theorem 7.1, Corollary 7.1]. For the notation, see [1]. We start with the Gagliardo-Nirenberg theorem which can be found in a more general formulation in [74, Lecture II, (2.2)].

Lemma 62. *Let $y \in H^1(\Omega)$. Then following inequality holds*

$$\|y\|_{L^{2\frac{n}{n+\rho}}(\Omega)} \leq \|\nabla y\|^{\frac{n}{n+\rho}}_{L^2(\Omega)} \|y\|^{1-\frac{n}{n+\rho}}_{L^2(\Omega)}$$

with $\rho \geq 1$.

Lemma 63. *Let $y \in L^2\left(0,T; W_0^{1,2}(\Omega)\right) \cap L^\infty(0,T; L^\rho(\Omega))$ with $\rho \geq 1$. Then $y \in L^\sigma(Q)$ with $\sigma = 2\frac{n+\rho}{n}$ and there exists a constant $c > 0$ with*

$$\int_Q |y(x,t)|^\sigma dxdt \leq c\|y\|^{\frac{2\rho}{n}}_{L^\infty(0,T;L^\rho(\Omega))} \|\nabla y\|^2_{L^2(Q)}.$$

Proof. By applying Lemma 62 (Gagligardo-Nierenberg) for $\sigma := 2\frac{\rho+n}{n} > 1$, we have

$$\left(\int_\Omega |y(x,t)|^\sigma dx\right)^{\frac{1}{\sigma}} \leq C\|\nabla y(\cdot,t)\|_{L^2(\Omega)}^{\frac{2}{\sigma}} \|y(\cdot,t)\|_{L^\rho(\Omega)}^{\left(1-\frac{2}{\sigma}\right)}$$

for all $t \in [0,T]$ and thus equivalently

$$\left(\int_\Omega |y(x,t)|^\sigma dx\right) \leq C^\sigma \|\nabla y(\cdot,t)\|_{L^2(\Omega)}^{2} \|y(\cdot,t)\|_{L^\rho(\Omega)}^{\left(1-\frac{2}{\sigma}\right)\sigma}.$$

By integrating over t, we obtain

$$\int_0^T \int_\Omega |y(x,t)|^\sigma dxdt \leq C^\sigma \int_0^T \|\nabla y(\cdot,t)\|_{L^2(\Omega)}^{2} \|y(\cdot,t)\|_{L^\rho(\Omega)}^{\left(1-\frac{2}{\sigma}\right)\sigma} dt.$$

Since $y \in L^\infty(0,T;L^\rho(\Omega))$, we have

$$\int_0^T \int_\Omega |y(x,t)|^\sigma dxdt \leq C^\sigma \|y\|_{L^\infty(0,T;L^\rho(\Omega))}^{\frac{2\rho}{n}} \int_0^T \|\nabla y(\cdot,t)\|_{L^2(\Omega)}^{2} dt.$$

Inserting the definition of σ on the right hand-side of this inequality, we obtain the statement of the lemma from the identity

$$\int_0^T \|\nabla y(\cdot,t)\|_{L^2(\Omega)}^2 dt = \int_0^T \int_\Omega |\nabla y(x,t)|^2 dxdt = \int_Q |\nabla y(x,t)|^2 dxdt = \|\nabla y\|_{L^2(Q)}^2$$

and $c := C^\sigma$. □

Theorem 64. *Assuming there exists a unique solution to (5.14) fulfilling*

$$y \in L^2\left(0,T;H_0^1(\Omega)\right) \cap L^\infty\left(0,T;L^2(\Omega)\right) \text{ and } y' \in L^2\left(0,T;H^{-1}(\Omega)\right),$$

then the solution is essentially bounded with

$$\|y\|_{L^\infty(Q)} \leq C\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(\Omega)}$$

where $C > 0$.

Proof. We choose $k > \|y_0\|_{L^\infty(\Omega)} \geq 0$. We have that $y(\cdot,t) - k \in H^1(\Omega)$ for any $t \in [0,T]$ due to the linearity of the weak differentiation operation [45, 5.2 Theorem 1] and that the derivative of the constant function k fulfills the definition of weak derivative [45, page 243]. Furthermore there is no other weak derivative because of the uniqueness of the weak derivative [45, page 243] and therefore it holds that

$$(y-k)_+(\cdot,t) := \max(y(\cdot,t) - k, 0) \in H_0^1(\Omega)$$

for almost any $t \in (0,T)$, see [38, Chapter 4, Proposition 6]. Then, we choose $v = (y-k)_+(\cdot,t)$ in (5.14) and obtain

$$\left(y'(\cdot,t), (y-k)_+(\cdot,t)\right) + B\left(y-k, (y-k)_+; t\right) \leq \left(h(\cdot,t), (y-k)_+(\cdot,t)\right)$$

for almost any $t \in (0,T)$ where we use

$$B\left(y, (y-k)_+; t\right) \geq B\left(y, (y-k)_+; t\right) + B\left(-k, (y-k)_+; t\right) = B\left(y-k, (y-k)_+; t\right)$$

for any $t \in [0,T]$ and thus with the coercivity condition

$$\left((y-k)_+', (y-k)_+\right) + \beta\|(y-k)_+(\cdot,t)\|_{H_0^1(\Omega)}^2 \leq \left(h(\cdot,t), (y-k)_+(\cdot,t)\right) \tag{5.15}$$

for almost any $t \in (0,T)$. Notice that $(y-k)_+(\cdot,t) = 0$ if $y-k \leq 0$ and therefore

$$B\left(y-k,(y-k)_+;t\right) = B\left((y-k)_+,(y-k)_+;t\right)$$

and

$$\left(y'(\cdot,t),(y-k)_+(\cdot,t)\right) = \left((y(\cdot,t)-k)',(y-k)_+(\cdot,t)\right) = \left((y-k)'_+(\cdot,t),(y-k)_+(\cdot,t)\right)$$

due to the bilinearity and also in the case $y-k>0$ as $(y-k)_+(\cdot,t) = (y(\cdot,t)-k)$. Next, as $(y-k)_+$ is measurable, see [36, page 46] and

$$\int_0^T \|(y-k)_+(\cdot,t)\|^2_{H_0^1(\Omega)}dt \leq \int_0^T \|(y-k)(\cdot,t)\|^2_{H_0^1(\Omega)}dt = \int_0^T \|y\|^2_{H_0^1(\Omega)}dt < \infty$$

and

$$\int_0^T \left((y-k)_+(\cdot,t),v\right)^2_{H_0^1(\Omega)}dt \leq \int_0^T \left((y-k)(\cdot,t),v\right)^2_{H_0^1(\Omega)}dt = \int_0^T \left(y(\cdot,t),v\right)^2_{H_0^1(\Omega)} < \infty$$

for all $v \in H_0^1(\Omega)$, we obtain with [45, 5.9 Theorem 3] the following

$$\left((y-k)'_+,(y-k)_+\right) = \frac{1}{2}\frac{d}{dt}\|(y-k)_+(\cdot,t)\|^2_{L^2(\Omega)}.$$

Thus with (5.15) we get

$$\frac{1}{2}\frac{d}{dt}\|(y-k)_+(\cdot,t)\|^2_{L^2(\Omega)} + \beta\|(y-k)_+(\cdot,t)\|^2_{H_0^1(\Omega)} \leq \left(h(\cdot,t),(y-k)_+(\cdot,t)\right) \tag{5.16}$$

for almost any $t \in (0,T)$. By taking the absolute value of the right hand-side of (5.16), renaming the variable t into $\tilde{t}$ and integrating over it from 0 to t, we obtain

$$\begin{aligned}\frac{1}{2}\|(y-k)_+(\cdot,t)\|^2_{L^2(\Omega)} + \beta\int_0^t \|(y-k)_+\left(\cdot,\tilde{t}\right)\|^2_{H_0^1(\Omega)}d\tilde{t} &\leq \int_0^t\int_\Omega |h\left(x,\tilde{t}\right)(y-k)_+\left(x,\tilde{t}\right)|dxd\tilde{t} \\ &\leq \int_0^T\int_\Omega |h\left(x,\tilde{t}\right)(y-k)_+\left(x,\tilde{t}\right)|dxd\tilde{t}\end{aligned} \tag{5.17}$$

where, because of the definition of k, we have $\|(y-k)_+(\cdot,0)\|^2_{L^2(\Omega)} = 0$. From (5.17), it follows that

$$\frac{1}{2}\|(y-k)_+(\cdot,t)\|^2_{L^2(\Omega)} \leq \int_0^T\int_\Omega |h\left(x,\tilde{t}\right)(y-k)_+\left(x,\tilde{t}\right)|dxd\tilde{t}, \tag{5.18}$$

$$\beta\int_0^t \|(y-k)_+\left(\cdot,\tilde{t}\right)\|^2_{H_0^1(\Omega)}d\tilde{t} \leq \int_0^T\int_\Omega |h\left(x,\tilde{t}\right)(y-k)_+\left(x,\tilde{t}\right)|dxd\tilde{t}. \tag{5.19}$$

By the monotonicity of the square root and taking the supremum over t, we obtain from (5.18) that

$$\sqrt{\frac{1}{2}}\|(y-k)_+\|_{L^\infty(0,T;L^2(\Omega))} \leq \sqrt{\int_0^T\int_\Omega |h\left(x,\tilde{t}\right)(y-k)_+\left(x,\tilde{t}\right)|dxd\tilde{t}}.$$

Further with this inequality and (5.19), we obtain the following

$$\tilde{C}\left(\|(y-k)_+\|^2_{L^\infty(0,T;L^2(\Omega))} + \|\nabla(y-k)_+\|^2_{L^2(Q)}\right) \leq \int_0^T\int_\Omega |h(x,t)(y-k)_+(x,t)|dxdt \tag{5.20}$$

for $\tilde{C} := \min\left\{\frac{1}{4}, \frac{\beta}{2}\right\} > 0$ and renaming $\tilde{t}$ into t. Then we can apply Young's inequality, see [6, (3.4)] and obtain

$$\begin{aligned}
&\| (y-k)_+ \|_{L^\infty(0,T;L^2(\Omega))}^{\frac{4}{n+2}} \|\nabla (y-k)_+ \|_{L^2(Q)}^{\frac{2n}{n+2}} \\
&\leq \frac{2n+4}{4} \left(\| (y-k)_+ \|_{L^\infty(0,T;L^2(\Omega))}^{\frac{4}{n+2}} \right)^{\frac{2n+4}{4}} + \frac{2n}{2n+4} \left(\|\nabla (y-k)_+ \|_{L^2(Q)}^{\frac{2n}{n+2}} \right)^{\frac{2n+4}{2n}} \\
&\leq \frac{2n+4}{4} \left(\| (y-k)_+ \|_{L^\infty(0,T;L^2(\Omega))}^{2} + \|\nabla (y-k)_+ \|_{L^2(Q)}^{2} \right).
\end{aligned}$$

This result and (5.20) imply the following

$$\left(\frac{4\tilde{C}}{2n+4}\right)^{\frac{n+2}{n}} \left(\| (y-k)_+ \|_{L^\infty(0,T;L^2(\Omega))}^{\frac{4}{n}} \|\nabla (y-k)_+ \|_{L^2(Q)}^{2} \right) \leq \left(\int_Q |h(x,t)(y-k)_+(x,t)| dxdt \right)^{\frac{n+2}{n}}. \tag{5.21}$$

Then, due to $y \in L^2\left(0,T; H_0^1(\Omega)\right) \cap L^\infty\left(0,T; L^2(\Omega)\right)$ and $(y-k)_+(\cdot,t) \in H_0^1(\Omega)$ for almost any $t \in (0,T)$, see [38, Chapter 4, Proposition 6], we have that

$$\int_Q (y-k)_+^{2\frac{n+2}{n}} dxdt \leq c \| (y-k)_+ \|_{L^\infty(0,T;L^2(\Omega))}^{\frac{4}{n}} \|\nabla (y-k)_+ \|_{L^2(Q)}^{2} \tag{5.22}$$

with $c > 0$ by Lemma 63. Inequality (5.22) and (5.21) imply the following

$$\bar{C} \int_Q (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dxdt \leq \left(\int_Q |h(x,t)(y-k)_+(x,t)| dxdt \right)^{\frac{n+2}{n}} dxdt \tag{5.23}$$

where $\bar{C} := \frac{1}{c}\left(\frac{4\tilde{C}}{(2n+4)}\right)^{\frac{n+2}{n}} > 0$. Consequently, we have

$$\bar{C} \int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dxdt \leq \left(\int_{A_k} |h(x,t)(y-k)_+(x,t)| dxdt \right)^{\frac{n+2}{n}} dxdt \tag{5.24}$$

where

$$A_k := \{(x,t) \in Q |\ y(x,t) > k\}.$$

The set A_k is measurable, see [36, Proposition 2.1.1 and page 42]. By estimating the right hand-side of (5.24) with Hölder's inequality, see [5, X Theorem 4.2], we obtain

$$\begin{aligned}
&\bar{C} \int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dxdt \\
&\leq \left(\left(\int_{A_k} |h(x,t)|^{\frac{2n+4}{n+4}} dxdt \right)^{\frac{n+4}{2n+4}} \left(\int_{A_k} |(y-k)_+(x,t)|^{\frac{2n+4}{n}} dxdt \right)^{\frac{n}{2n+4}} \right)^{\frac{n+2}{n}} \\
&= \left(\int_{A_k} |h(x,t)|^{\frac{2n+4}{n+4}} dxdt \right)^{\frac{n+4}{2n}} \left(\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dxdt \right)^{\frac{1}{2}}.
\end{aligned} \tag{5.25}$$

If $\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dxdt > 0$, then (5.25) implies

$$\bar{C} \int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dxdt \leq \left(\int_{A_k} |h(x,t)|^{\frac{2n+4}{n+4}} dxdt \right)^{\frac{n+4}{n}}. \tag{5.26}$$

This is also true in the case of $\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt = 0$. We use Hölder's inequality again for the right hand-side of (5.26), see [5, X Theorem 4.2], and obtain the following

$$\begin{aligned}
&\bar{C}\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt \leq \left(\int_{A_k} |h(x,t)|^{\frac{2n+4}{n+4}}dxdt\right)^{\frac{n+4}{n}} \\
&\leq \left(\left(\int_{A_k} 1^{\frac{q(4+n)}{n(q-2)+4(q-1)}}dxdt\right)^{\frac{n(q-2)+4(q-1)}{q(4+n)}} \left(\int_{A_k} \left(|h(x,t)|^{\frac{2n+4}{n+4}}\right)^{q\frac{n+4}{2n+4}} dxdt\right)^{\frac{2n+4}{q(n+4)}}\right)^{\frac{n+4}{n}} \qquad (5.27)\\
&= \left(\left(\int_{A_k} |h(x,t)|^q dxdt\right)^{\frac{1}{q}}\right)^{\frac{2n+4}{n}} |A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}} \|h\|_{L^q(A_k)}^{\frac{2n+4}{n}} \leq |A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}} \|h\|_{L^q(Q)}^{\frac{2n+4}{n}}
\end{aligned}$$

where $|A_k|$ is the measure of A_k. Now, if we take $m > k$, then we have $A_m \subseteq A_k$. Additionally, we have that $y > m$ on A_m and thus $y \geq y-k > m-k$ on A_m since $k > \|y_0\|_{L^\infty(\Omega)} \geq 0$. Due to $y-k = (y-k)_+$ on A_m, we obtain

$$\begin{aligned}
&\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt \\
&\geq \int_{A_m} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt = \int_{A_m} (y-k)^{2\frac{n+2}{n}}(x,t)\,dxdt \geq (h-k)^{2\frac{n+2}{n}}|A_m|.
\end{aligned} \qquad (5.28)$$

We combine (5.27) and (5.28) and obtain the following

$$(m-k)^{2\frac{n+2}{n}}|A_m| \leq \hat{C}\|h\|_{L^q(Q)}^{\frac{2n+4}{n}}|A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}}$$

with $\hat{C} := \frac{1}{\bar{C}}$. Therefore we have

$$|A_m| \leq \left(\frac{\hat{C}^{\frac{n}{2n+4}}\|h\|_{L^q(Q)}}{m-k}\right)^{\frac{2n+4}{n}} |A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}}. \qquad (5.29)$$

Now, we consider the case that $\|h\|_{L^q(Q)} > 0$. We have that $\frac{2n+4}{n} > 0$ for $n \geq 1$ and $\frac{n+4}{n} - \frac{2n+4}{qn} > 1$ since $q > \frac{n}{2}+1$. Therefore, we apply Lemma 58 and obtain that $|A_m| = 0$ for all $m \geq C\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(\Omega)}$, $C := \hat{C}^{\frac{n}{2n+4}} 2^{\frac{4+2n-4q-nq}{4+2n-4q}} |Q|^{\frac{2q-n-2}{2q+nq}}$ where $|Q|$ is the measure of Q. If $\|h\|_{L^q(Q)} = 0$, then we have from (5.29) that $A_m = 0$ for any $m > k$ and any $k > \|y_0\|_{L^\infty(Q)}$. Therefore in the limit for $m \to k$ and $k \to \|y_0\|_{L^\infty(\Omega)}$, we have that $|A_m| = 0$ for $m \geq \|y_0\|_{L^\infty(\Omega)}$. If there was a number $\epsilon > 0$ such that the statement did not hold, then we would choose $\epsilon > k > 0$ in contradiction to the already proved statement. Concluding, this means that the set A_m where the function y is such that

$$y > C\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(Q)}$$

has measure zero.

In the same way, if we follow the reasoning above for

$$(y+k)_- := \min(y+k, 0) \text{ and } A_k := \{(x,t) \in Q | \; y < -k\},$$

we obtain that the set $A_m = \{(x,t) \in Q | \; y < -m\}$ where the function y is such that

$$y < -\left(C\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(Q)}\right)$$

has measure zero. Therefore, we obtain $\|y\|_{L^\infty(Q)} \leq C\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(\Omega)}$. □

For $P.4$), we have the following theorem similar to Theorem 65 that holds immediately, assuming that $K_U \subseteq \mathbb{R}_0^+$ considering step (5.15) since $uy\,(y-k)_+ \geq 0$ because $(y-k)_+ = 0$ if $y \leq k \geq 0$.

Theorem 65. *For the solution y to a bilinear parabolic boundary value problem as in P.4) with $K_U \subseteq \mathbb{R}_0^+$, we have*

$$\|y\|_{L^\infty(Q)} \leq d\|\tilde{f}\|_{L^q(Q)} + \|y_0\|_{L^\infty(\Omega)}$$

with $d > 0$ for any right-hand side $\tilde{f} \in L^q\,(Q)$.

5.3.3 Fokker-Planck equation

The following Lemma states that the L^2-norm of the adjoint equation (4.75) is bounded by the L^2-norm of the controls. The Lemma is used in the proof of Lemma 50.

Lemma 66. *The solution to (4.80), where it holds that $\delta p := p^k - p$ and p is a solution to (4.75) for v, w and p^k is a solution to (4.75) for $v \leftarrow v^k$, $w \leftarrow w^k$, is bounded by*

$$\|\delta p\|_{L^2(Q)} \leq \hat{\theta}^k \left(\|\delta v\|_{L^2(0,T)} + \|\delta w\|_{L^2(0,T)}\right)$$

where $\hat{\theta}^k := \mathrm{e}^{(\eta+1)T}\left(\tilde{\theta}^k + 2nL|\Omega|\right)T$ and $\delta v = v^k - v$, $\delta w = w^k - w$.

Proof. We start from (4.80) for $\delta p := p^k - p$ and perform a transformation of time $\tau := T - t$ where we still denote $\tau \mapsto \delta p^k\,(\cdot, T-\tau)$ by $t \mapsto \delta p^k\,(\cdot, t)$. Then, we obtain

$$\begin{aligned}
&\int_\Omega \delta p'\,(x,t)\,\varphi\,(x) + \frac{\sigma^2}{2}\,(\nabla \delta p\,(x)) \cdot \nabla\varphi\,(x) - \left(v^k\,(t) + x \circ w^k\,(t)\right) \cdot \nabla\delta p\,(x,t)\,\varphi\,(x)\,dx \\
&= \int_\Omega \left(v\,(t) + x \circ w\,(t) - \left(v^k\,(t) + x \circ w^k\,(t)\right)\right) \cdot \nabla p^k\,(x,t)\,\varphi\,(x)\,dx \\
&+ \int_\Omega \left(G\left(v^k, w^k\right)(x,t) - G\,(v,w)\,(x,t)\right)\varphi\,(x)\,dx.
\end{aligned} \tag{5.30}$$

Next we choose $\hat{p}^k\,(\cdot,t) := \mathrm{e}^{-\eta t}p^k\,(\cdot,t)$, $\hat{p}^{k+1} := \mathrm{e}^{-\eta t}p^{k+1}\,(\cdot,t)$ and $\delta\hat{p}\,(\cdot,t) := \mathrm{e}^{-\eta t}\delta p\,(\cdot,t)$, $\eta \geq 0$ and insert $\delta\hat{p}$ for φ into (5.30). Then we obtain, with the same reasoning as in the proof of Theorem 43, for η sufficiently large that

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\delta\hat{p}\,(\cdot,t)\,\|^2_{L^2(\Omega)} \\
&\leq \|\left((v\,(t) + x \circ w\,(t)) - \left(v^k\,(t) + x \circ w^k\,(t)\right)\right) \cdot \nabla\hat{p}^k\,(\cdot,t)\,\|_{L^2(\Omega)}\|\delta\hat{p}\,(\cdot,t)\,\|_{L^2(\Omega)} \\
&+ \|G\left(v^k, w^k\right)(\cdot,t) - G\,(v,w)\,(\cdot,t)\,\|_{L^2(\Omega)}\|\delta\hat{p}\,(\cdot,t)\,\|_{L^2(\Omega)}
\end{aligned}$$

with the Cauchy-Schwarz inequality, see [2, Lemma 2.2] and with [45, Section 5.9 Theorem 3] for

$$\frac{1}{2}\frac{d}{dt}\|\delta\hat{p}\,(\cdot,t)\,\|^2_{L^2(\Omega)} = \int_\Omega \delta\hat{p}'\,(x,t)\,\delta\hat{p}\,(x,t)\,dx.$$

With Cauchy's inequality, see [45, page 622], we have the following

$$\begin{aligned}
&\frac{d}{dt}\|\delta\hat{p}\,(\cdot,t)\,\|^2_{L^2(\Omega)} \\
&\leq \|\left((v\,(t) + x \circ w\,(t)) - \left(v^k\,(t) + x \circ w^k\,(t)\right)\right) \cdot \nabla\hat{p}^k\,(\cdot,t)\,\|^2_{L^2(\Omega)} + \|\delta\hat{p}\,(\cdot,t)\,\|^2_{L^2(\Omega)} \\
&+ \|G\left(v^k, w^k\right)(\cdot,t) - G\,(v,w)\,(\cdot,t)\,\|^2_{L^2(\Omega)} + \|\delta\hat{p}\,(\cdot,t)\,\|^2_{L^2(\Omega)}.
\end{aligned} \tag{5.31}$$

In order to apply Gronwall's inequality, see [45, page 624] or Lemma 57 for instance, we perform the following estimations. For this purpose, we estimate the term

$$\| \left(\left(v\left(t\right) + x \circ w\left(t\right) \right) - \left(v^k\left(t\right) + x \circ w^k\left(t\right) \right) \right) \cdot \nabla \hat{p}^k\left(\cdot, t\right) \|_{L^2(\Omega)}^2.$$

We have the following

$$\begin{aligned}
&\| \left(\left(v\left(t\right) + x \circ w\left(t\right) \right) - \left(v^k\left(t\right) + x \circ w^k\left(t\right) \right) \right) \cdot \nabla \hat{p}^k\left(\cdot, t\right) \|_{L^2(\Omega)}^2 \\
&= \int_\Omega \left(\sum_{i=1}^n \left(\left(v^i\left(t\right) + x_i w^i\left(t\right) \right) - \left(\left(v^k\right)^i\left(t\right) + x_i \left(w^k\right)^i\left(t\right) \right) \right) \frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) \right)^2 dx \\
&= \int_\Omega \left(\sum_{i=1}^n \left[\left(v^i\left(t\right) - \left(v^k\right)^i\left(t\right) \right) \frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) + x_i \left(w^i\left(t\right) - \left(w^k\right)^i\left(t\right) \right) \frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) \right] \right)^2 dx \\
&\leq 2 \int_\Omega \left(\sum_{i=1}^n \left(v^i\left(t\right) - \left(v^k\right)^i\left(t\right) \right) \frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) \right)^2 dx \\
&+ 2 \int_\Omega \left(\sum_{i=1}^n x_i \left(w^i\left(t\right) - \left(w^k\right)^i\left(t\right) \right) \frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) \right)^2 dx \\
&\leq 2n \sum_{i=1}^n \int_\Omega \left(v^i\left(t\right) - \left(v^k\right)^i\left(t\right) \right)^2 \left(\frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) \right)^2 dx \\
&+ 2n \sum_{i=1}^n \int_\Omega x_i^2 \left(w^i\left(t\right) - \left(w^k\right)^i\left(t\right) \right)^2 \left(\frac{\partial}{\partial x_i} \hat{p}^k\left(x,t\right) \right)^2 dx \\
&\leq 2n \|\hat{p}^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}^2 \sum_{i=1}^n \left(v^i\left(t\right) - \left(v^k\right)^i\left(t\right) \right)^2 \\
&+ 2n \|\hat{p}^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}^2 \left(\max_{i=1,\dots,n} \max_{x\in\Omega} |x_i|^2 \right) \sum_{i=1}^n \left(w^i\left(t\right) - \left(w^k\right)^i\left(t\right) \right)^2 \\
&\leq \tilde{\theta}^k \left(\sum_{i=1}^n \left(\left(v^i\left(t\right) - \left(v^k\right)^i\left(t\right) \right)^2 + \left(w^i\left(t\right) - \left(w^k\right)^i\left(t\right) \right)^2 \right) \right)
\end{aligned} \tag{5.32}$$

with the Jensen inequality, see [72, Proposition 824] and

$$\tilde{\theta}^k := 2n \|p^k\|_{L^\infty\left(0,T;H_0^1(\Omega)\right)}^2 \max\left(1, \max_{i=1,\dots,n} \max_{x\in\Omega} |x_i|^2 \right).$$

Furthermore, we have with Jensen's inequality and our Lipschitz assumption for G that

$$\begin{aligned}
&\|G\left(v^k, w^k\right)\left(\cdot,t\right) - G\left(v,w\right)\left(\cdot,t\right)\|_{L^2(\Omega)}^2 = \int_\Omega \left(G\left(v^k, w^k\right)\left(\cdot,t\right) - G\left(v,w\right)\left(x,t\right) \right)^2 dx \\
&\leq \int_\Omega 2nL^2 \left(\sum_{i=1}^n \left(\left(\left(v^k\right)^i\left(t\right) - v^i\left(t\right) \right)^2 + \left(\left(w^k\right)^i\left(t\right) - w^i\left(t\right) \right)^2 \right) \right) \\
&= 2nL|\Omega| \left(\sum_{i=1}^n \left(\left(\left(v^k\right)^i\left(t\right) - v^i\left(t\right) \right)^2 + \left(\left(w^k\right)^i\left(t\right) - w^i\left(t\right) \right)^2 \right) \right)
\end{aligned} \tag{5.33}$$

where $|\Omega|$ is the measure of Ω. Putting (5.32) and (5.33) into (5.31), we obtain with Gronwall's inequality, see [45, page 624] or Lemma 57 for example, the following

$$\begin{aligned}
&\|\delta p\|^2_{L^2(Q)} = \|\mathrm{e}^{\eta \cdot}\delta \hat{p}\|^2_{L^2(Q)} \leq \mathrm{e}^{\eta T} \int_0^T \int_\Omega \delta \hat{p}\left(x,t\right)^2 dxdt = \mathrm{e}^{\eta T} \int_0^T \|\delta \hat{p}\left(\cdot,t\right)\|^2_{L^2(\Omega)} dt \\
&\leq \mathrm{e}^{\eta T} \int_0^T \mathrm{e}^{2t} \int_0^t \left(\tilde{\theta}^k + 2nL|\Omega|\right) \sum_{i=1}^n \left(\left(\left(v^k\right)^i \left(\tilde{t}\right) - v^i\left(\tilde{t}\right)\right)^2 \right) d\tilde{t}dt \\
&+ \mathrm{e}^{\eta T} \int_0^T \mathrm{e}^{2t} \int_0^t \left(\tilde{\theta}^k + 2nL|\Omega|\right) \left(\sum_{i=1}^n \left(\left(w^k\right)^i \left(\tilde{t}\right) - w^i\left(\tilde{t}\right)\right)^2 \right) d\tilde{t}dt \\
&\leq \left(\hat{\theta}^k\right)^2 \left(\|v^k - v\|^2_{L^2(0,T)} + \|w^k - w\|^2_{L^2(0,T)}\right) \\
&\leq \left(\hat{\theta}^k\right)^2 \left(\|v^k - v\|^2_{L^2(0,T)} + 2\|v^k - v\|_{L^2(0,T)}\|w^k - w\|_{L^2(0,T)} + \|w^k - w\|^2_{L^2(0,T)}\right) \\
&= \left(\hat{\theta}^k\right)^2 \left(\|v^k - v\|_{L^2(0,T)} + \|w^k - w\|_{L^2(0,T)}\right)^2
\end{aligned}$$

since $\delta \hat{p}^k\left(0\right) = F\left(T\right) - F\left(T\right) = 0$ where $\left(\hat{\theta}^k\right)^2 := \mathrm{e}^{(\eta+2)T}\left(\tilde{\theta}^k + 2nL|\Omega|\right)T$. □

The following lemma states that the function $z \mapsto \max\left(0, |z| - s\right)$, $s > 0$ is Lipschitz continuous and convex.

Lemma 67. *The function* $z \mapsto \max\left(0, |z| - s\right) : \mathbb{R} \to \mathbb{R}$, $s > 0$, *is Lipschitz continuous with Lipschitz constant equal one and is convex.*

Proof. We start proving the Lipschitz continuity where the Lipschitz constant equals one by a case study. For this purpose, we need the reversed triangle inequality, see [3, Corollary 8.11]. We have to see that

$$|\max\left(0, |z_1| - s\right) - \max\left(0, |z_2| - s\right)| \leq |z_1 - z_2|.$$

If $|z_1| \geq s$ and $|z_2| \geq s$, then we have

$$||z_1| - s - |z_2| + s| = ||z_1| - |z_2|| \leq |z_1 - z_2|.$$

If $|z_1| < s$, equivalently $-|z_1| > -s$ and $|z_2| \geq s$, then we have that

$$|0 - |z_2| + s| = ||z_2| - s| \leq ||z_2| - |z_1|| = ||z_1| - |z_2|| \leq |z_1 - z_2|.$$

If $|z_1| \geq s$ and $|z_2| < s$, equivalently $-|z_2| > -s$ then we have

$$||z_1| - s| \leq ||z_1| - |z_2|| \leq |z_1 - z_2|.$$

If $|z_1| < s$ and $|z_2| < s$, then we have that

$$|0 - 0| = 0 \leq |z_1 - z_2|.$$

The convexity can be seen as follows. We have to prove that

$$\max\left(0, |\left(1-\lambda\right) z_1 + \lambda z_2| - s\right) \leq \left(1 - \lambda\right) \max\left(0, |z_1| - s\right) + \lambda \max\left(0, |z_2| - s\right)$$

for all $\lambda \in [0,1]$. We have that

$$\begin{aligned}
&\max\left(0, |\left(1-\lambda\right) z_1 + \lambda z_2| - s\right) \leq \max\left(0, \left(1-\lambda\right)|z_1| + \lambda |z_2| - s\right) \\
&= \max\left(0, \left(1-\lambda\right)\left(|z_1| - s\right) + \lambda\left(|z_2| - s\right)\right)
\end{aligned}$$

for all $\lambda \in [0,1]$ since replacing a number in one of the two arguments of max by a bigger one, the result of max also is greater or equal. This argument is also used for the following where we prove the convexity by a case study.
If we have that $|z_1| \geq s$ and $|z_2| \geq s$, then we have that

$$\begin{aligned} \max\left(0, (1-\lambda)\left(|z_1| - s\right) + \lambda\left(|z_2| - s\right)\right) &= (1-\lambda)\left(|z_1| - s\right) + \lambda\left(|z_2| - s\right) \\ &= (1-\lambda)\max\left(0, |z_1| - s\right) + \lambda \max\left(0, |z_2| - s\right). \end{aligned}$$

If we have that $|z_1| < s$ and $|z_2| \geq s$, then we have that

$$\begin{aligned} \max\left(0, (1-\lambda)\left(|z_1| - s\right) + \lambda\left(|z_2| - s\right)\right) &\leq \max\left(0, \lambda\left(|z_2| - s\right)\right) = \lambda\left(|z_2| - s\right) \\ &= (1-\lambda)\max\left(0, |z_1| - s\right) + \lambda \max\left(0, |z_2| - s\right). \end{aligned}$$

If $|z_1| \geq s$ and $|z_2| < s$, then we have that

$$\begin{aligned} \max\left(0, (1-\lambda)\left(|z_1| - s\right) + \lambda\left(|z_2| - s\right)\right) &\leq \max\left(0, (1-\lambda)\left(|z_1| - s\right)\right) = (1-\lambda)\left(|z_1| - s\right) \\ &= (1-\lambda)\max\left(0, |z_1| - s\right) + \lambda \max\left(0, |z_2| - s\right). \end{aligned}$$

If $|z_1| < s$ and $|z_2| < s$, then we have that

$$\max\left(0, (1-\lambda)\left(|z_1| - s\right) + \lambda\left(|z_2| - s\right)\right) = 0 = (1-\lambda)\max\left(0, |z_1| - s\right) + \lambda \max\left(0, |z_2| - s\right).$$

□

5.4 PMP sufficient conditions for an optimal solution

In this section, we refer to Chapter 3. The the results also hold for Chapter 2 or Chapter 4 with analogous arguments.
We show that the condition

$$H\left(z, \bar{y}, \bar{u}, \bar{p}\right) + r\left(w - \bar{u}\right)^2 \leq H\left(z, \bar{y}, w, \bar{p}\right) \tag{5.34}$$

for a triple $(\bar{y}, \bar{u}, \bar{p})$ and the constant $r \geq 0$ sufficiently large serves as a sufficient condition for a solution to (3.3). The idea for the present formulation (5.34) can be found in [89].

Theorem 68. *Let Assumptions A.1) to A.6) from Chapter 3 be fulfilled. Let $(\bar{y}, \bar{u})$ solve the state equation (3.2) and $\bar{p}$ solve the corresponding adjoint equation (3.5) for $\bar{y}$ instead of y and $\bar{u}$ instead of u. Assume that*

$$|\frac{\partial}{\partial \bar{y}} f\left(z, \bar{y}, u\right) - \frac{\partial}{\partial \bar{y}} f\left(z, \bar{y}, \bar{u}\right)| \leq \tilde{c} \sum_{j=1}^{m} |u_j - \bar{u}_j|$$

holds for almost all $z \in Z_i$ where $\tilde{c} > 0$ is a constant. Let $(\bar{y}, \bar{u}, \bar{p})$ fulfill

$$H\left(z, \bar{y}, \bar{u}, \bar{p}\right) + r\left(w - \bar{u}\right)^2 \leq H\left(z, \bar{y}, w, \bar{p}\right) \tag{5.35}$$

for all $w \in K_U$ and for almost all $z \in Z_i$ with $r \geq \tilde{c}c\sqrt{m} + \frac{5}{2}c^3 + \frac{1}{2}c^4$ where c, m are given in Section 3.1. Then $(\bar{y}, \bar{u})$ is a solution to (3.2), that is, $J(y,u) \geq J(\bar{y}, \bar{u})$ for all (y,u) solving (3.2) with $u \in U_{ad}$.

Proof. Notice that the notation is analogous to the one of the proof of Lemma 26 with $\delta y := y - \bar{y}$ and $\delta u := u - \bar{u}$ where we do not show the functions dependency on z. We have

$$
\begin{aligned}
&J\left(y,u\right)-J\left(\bar{y},\bar{u}\right)=\int_{Z_i} h\left(y\right)+g\left(u\right)-h\left(\bar{y}\right)-g\left(\bar{u}\right)dz\\
&=\int_{Z_i} H\left(z,y,u,\bar{p}\right)-\bar{p}f\left(z,y,u\right)-H\left(z,\bar{y},\bar{u},\bar{p}\right)+\bar{p}f\left(z,\bar{y},\bar{u}\right)dz\\
&=\int_{Z_i} H\left(z,\bar{y},u,\bar{p}\right)+\frac{\partial}{\partial y}H\left(z,\bar{y},u,\bar{p}\right)\delta y+\frac{1}{2}\left(\frac{\partial^2}{\partial\bar{y}^2}H\left(z,\bar{y},u,\bar{p}\right)\left(\delta y\right)^2\right)dz+\int_{Z_i} R_2\left(H,\bar{y};\delta y\right)dz\\
&-\int_{Z_i} H\left(z,\bar{y},\bar{u},\bar{p}\right)dz-\int_0^T\left(\delta y'\left(\cdot,t\right),\bar{p}\left(\cdot,t\right)\right)+B\left(\delta y,\bar{p};t\right)dt\\
&\geq r\int_{Z_i}\delta u^2dz+\frac{\partial}{\partial\bar{y}}h\left(\bar{y}\right)\delta y+\bar{p}\frac{\partial}{\partial\bar{y}}f\left(\bar{y},u\right)\delta y+\frac{1}{2}\left(\frac{\partial^2}{\partial\bar{y}^2}h\left(\bar{y}\right)\left(\delta y\right)^2+\bar{p}\frac{\partial^2}{\partial\bar{y}^2}f\left(\bar{y},u\right)\left(\delta y\right)^2\right)\\
&+\int_{Z_i} R_2\left(H,\bar{y};\delta y\right)dz-\int_0^T\left(\delta y'\left(\cdot,t\right),\bar{p}\left(\cdot,t\right)\right)+B\left(\delta y,\bar{p};t\right)dt\\
&=r\int_{Z_i}\delta u^2dz+\bar{p}\frac{\partial}{\partial\bar{y}}f\left(\bar{y},u\right)\delta y-\bar{p}\frac{\partial}{\partial\bar{y}}f\left(\bar{y},\bar{u}\right)\delta y+\frac{1}{2}\left(\frac{\partial^2}{\partial\bar{y}^2}h\left(\bar{y}\right)\left(\delta y\right)^2+\bar{p}\frac{\partial^2}{\partial\bar{y}^2}f\left(\bar{y},u\right)\left(\delta y\right)^2\right)\\
&+\int_{Z_i} R_2\left(H,\bar{y};\delta y\right)dz-\int_0^T\left(\delta y'\left(\cdot,t\right),\bar{p}\left(\cdot,t\right)\right)+B\left(\delta y,\bar{p};t\right)dt\\
&+\int_0^T-\left(\bar{p}'\left(\cdot,t\right),\delta y\left(\cdot,t\right),\right)+B^*\left(\bar{p},\delta y;t\right)dt\\
&\geq r\|\delta u\|^2_{L^2(Z_i)}-\tilde{c}c\sqrt{m}\|\delta u\|^2_{L^2(Z_i)}-\frac{1}{2}\left(c+c^2\right)c^2\|\delta u\|^2_{L^2(Z_i)}-\left(c+c^2\right)c^2\|\delta u\|_{L^2(Z_i)}
\end{aligned}
\tag{5.36}
$$

for all $u\in U_{ad}$ where we use the partial integration rule [95, Theorem 3.11], the Cauchy-Schwarz inequality [2, Lemma 2.2] in the last inequality and the estimation for $R_2\left(H,\bar{y};\delta y\right)$ as in the proof of Lemma 26, the equality

$$
\int_0^T B^*\left(\bar{p},\delta y;t\right)-B\left(\delta y,\bar{p};t\right)dt=\int_0^T B\left(\delta y,\bar{p};t\right)-B\left(\delta y,\bar{p};t\right)dt=0
$$

and the following estimation

$$
\begin{aligned}
&\int_{Z_i}\frac{\partial}{\partial\bar{y}}f\left(\bar{y},u\right)\delta y-\frac{\partial}{\partial\bar{y}}f\left(\bar{y},\bar{u}\right)\delta y dx\geq-\int_{Z_i}\tilde{c}\sum_{j=1}^m|u_j-\bar{u}_j||\delta y|dx\\
&\geq-\tilde{c}\sum_{j=1}^m\|\delta u_j\|_{L^2(Z_i)}\|\delta y\|_{L^2(Z_i)}\geq-\tilde{c}c\|\delta u\|_{L^2(Z_i)}\sqrt{\left(\sum_{j=1}^m\|\delta u_j\|_{L^2(Z_i)}\right)^2}\\
&\geq-\tilde{c}c\sqrt{m}\|\delta u\|_{L^2(Z_i)}\sqrt{\sum_{j=1}^m\|\delta u_j\|^2_{L^2(Z_i)}}=-\tilde{c}c\sqrt{m}\|\delta u\|^2_{L^2(Z_i)}
\end{aligned}
$$

with the Cauchy-Schwarz inequality [2, Lemma 2.2] and the Jensen inequality, see [72, Proposition 824]. □

If we consider an optimal control problem corresponding to Example 17 or Example 18, which holds in the ODE as well as in the PDE case, and do the same calculation for $\epsilon=0$, then we obtain that (5.34) holds for any $r\in\left[0,\frac{\alpha}{2}\right]$. Consequently, we know that if α is sufficiently large such that r can be chosen larger than $\tilde{c}c\sqrt{m}+\frac{5}{2}c^3+\frac{1}{2}c^4$, then any triple $\left(\bar{y},\bar{u},\bar{p}\right)$ which is PMP optimal, that means fulfills (3.7) ($r=0$ in (5.34)) is a solution to the considered optimal control problem according to Theorem 68 supposing that the other assumptions of Theorem 68 are fulfilled as well. Further we have the following corollary for a special case which includes optimal control problems with a distributed control, that means a linear control-to-state map, and a quadratic function h, which means that $\frac{\partial^2}{\partial y^2}h\left(\bar{y}\right) \geq 0$, and $R_2\left(H, \bar{y}; \delta y\right) = 0$ in (5.36). The proof of the following corollary is analogous to the one of Theorem 68 inserting the corresponding further assumptions.

Corollary 69. *Let Assumptions* A.1) *to* A.6) *from Chapter 3 be fulfilled. Let* $(\bar{y}, \bar{u})$ *solve the state equation (3.2) and* $\bar{p}$ *solve the corresponding adjoint equation (3.5) for* $\bar{y}$ *instead of* y *and* $\bar{u}$ *instead of* u*. If the function* f *does not depend on* y *and* h *is a quadratic function with* $\frac{\partial^2}{\partial y^2}h \geq 0$*, then we can choose* $r = 0$ *in* (5.35) *and thus the necessary condition*

$$H\left(z, \bar{y}, \bar{u}, \bar{p}\right) = \min_{w \in K_U} H\left(z, \bar{y}, w, \bar{p}\right) \tag{5.37}$$

is sufficient for $(\bar{y}, \bar{u})$ *to be a solution to* (3.3).

5.5 Discussion of the Assumptions A.1) to A.6) from Chapter 2

The Assumptions A.1) to A.6) guarantee that a solution to (2.3) can be characterized with the PMP, see Theorem 5 and the convergence analysis of the SQH can be performed, which means that the iterates of the SQH scheme converge to a PMP consistent solution, see Theorem 14.

However, for just the characterization with the PMP less assumptions are required to perform the corresponding proofs. The requirements formulated in Assumption A.1) can be weakened. It is sufficient if the functions $h : I \to \mathbb{R}$, $y \mapsto h\left(y\right)$, $F : I \to \mathbb{R}$, $y \mapsto F\left(y\right)$ and $f : I \to \mathbb{R}^n$, $y \mapsto f\left(t, y, u\right)$ are once continuously differentiable for every $u \in K_U$ and for any $t \in [0, T]$. Furthermore, it is made use of the condition that $\|\frac{\partial}{\partial y_l}f_i\left(\cdot, y, u\right)\|_{L^\infty} \leq c$ for all $l, i \in \{1, ..., n\}$, see the proof of Lemma 2 where also Assumption A.5) is needed. The local integrability of f, see Assumption A.3), is needed in the proof of Lemma 4. The measurability of the corresponding functions, see Assumption A.2), is obligatory to have a well-defined integrals.

If we have that $\|\frac{\partial}{\partial y_l}f_i\left(\cdot, y, u\right)\|_{L^\infty} = 0$ for all $l, i \in \{1, ..., n\}$, all $y \in I$ and all $u \in K_U$, then we can do without the assumption that $\|\frac{\partial}{\partial y_l}h\left(y\right)\|_{L^\infty} \leq c$ and $\|\frac{\partial}{\partial y_l}F\left(y\right)\|_{L^\infty} \leq c$ for all $l \in \{1, ..., n\}$ and all $y \in I$ which is needed for the boundedness result of the adjoint variable in Lemma 8. This result in turn is only needed if $\|\frac{\partial}{\partial y_l}f_i\left(\cdot, y, u\right)\|_{L^\infty} > 0$ for one $l, i \in \{1, ..., n\}$, one $y \in I$ or one $u \in K_U$ as we can see in the proof of Lemma 10 where the boundedness of the difference of two adjoint variables is shown. In any case, we need $\|\frac{\partial^2}{\partial y_l \partial y_\ell}h\left(y\right)\|_{L^\infty} \leq c$ and $\|\frac{\partial^2}{\partial y_l \partial y_\ell}F\left(y\right)\|_{L^\infty} \leq c$ for all $l, \ell \in \{1, ..., n\}$ and all $y \in I$. For Lemma 11, we only need that the adjoint variable is bounded if $\|\frac{\partial^2}{\partial y_l \partial y_\ell}f_i\left(\cdot, y, u\right)\|_{L^\infty} > 0$ for one $i, l, \ell \in \{1, ..., n\}$, one $y \in I$ or one $u \in K_U$.

The compactness of K_U is needed twice in this thesis. First it is needed to ensure that the subproblem where we minimize K_ϵ in Step 2 in Algorithm 2.1 has a solution, see the proof of Lemma 7. That means that if the function $K_\epsilon : \mathbb{R}^m \to \mathbb{R}$, $w \mapsto K_\epsilon\left(t, y, w, v, p\right)$ has a global minimum, for example because it is quadratic, then we can do without the requirement that K_U has to be compact for Lemma 7. For Theorem 14, only the boundedness of K_U is needed to ensure pointwise convergence of the augmented Hamiltonian. However, the boundedness or compactness of K_U is not needed for the characterization of an optimal control with the PMP.

5.6 Description of the provided MATLAB files

In this section, we describe the provided MATLAB files that are used for the implementations of the SQH method for the corresponding numerical experiments. All the files necessary for the calculations of an experiment are zipped together in a zip-file. To start the calculations execute the corresponding main file with MATLAB.

In the file SQH_QC_L1.zip we have the codes for the L^1-experiment of Figure 2.1. The main file is the PMP_QC.m. In the LONE.zip is the LONE Code for the comparison between the implemented globalized Newton method with the SQH method in Section 2.4. The file L1SKRYN.m has to be executed and in the file Test2GL.m we can set the problems parameters. In the file SQH_OC_L0.zip we have the codes for the L^0-experiment of Figure 2.2. The main file is the PMP_OC_L0.m.

In the file SQH_therapy.zip we have the codes for the experiment where the results are depicted in Figure 2.4. The main file is SQH_therapy.m

The codes of the file SQH_P1.zip are set for the experiment shown in Figure 3.1. The main file is SQH_P1.m. In the folder gradient methods, we give the implemented projected gradient and projected nonlinear conjugated gradient method to obtain the results in Table 3.6 and Table 3.7.

Further with the file SQH_integer.zip the Figure 3.4 is created. The main file is SQH_integer.m. In order to obtain the results shown in Figure 3.5, we use the codes of the file SQH_stepCost.zip. The main file is SQH_stepCost.m.

In the file SQH_P2.zip we have the code that is set such that we get the results depicted in Figure 3.6. The main file is SQH_P2.m.

The code of SQH_P3.zip calculates the results of Figure 3.7. The main file is SQH_P3.m.

In the file SQH_P4.zip we have the code for the results depicted in Figure 3.8. The main file is SQH_P4.m.

The results of Figure 3.9 are obtained with the code of the file SQH_P5.zip. The main file is SQH_P5.m.

The code of SQH_P6.zip is set such that it calculates the results shown in Figure 3.10. The main file is SQH_P6.m.

In the file SQH_P7.zip we have the codes to obtain the results depicted in Figure 3.11. The main file is SQH_P7.m.

The file optRWkonNCGe.zip contains one file that is used for the experiment depicted in Figure 4.1.

In the file optRWkonNCGcc.zip we have the files that are used for the experiment shown in Figure 4.2. The main file is optRWkonNCGcc.m.

The file optRWJc.zip contains one file that is used to perform the calculations whose results are depicted in Figure 4.3.

In the file SQH_FP.zip we have the codes to obtain the results depicted in Figure 4.4. The main file is SQH_FP.m. With the file TEST_CONTROLLED_MC_2D.m we perform the corresponding Monte-Carlo simulation and plot the result into the figure with the mean value obtained by the execution of SQH_PF.m, see Figure 4.5. The diffusion of the random walk is set in the file model.m.

In the file SQH_FP_u.zip, we have the codes that implement the SQH method for the results depicted in Figure 4.6. The output of the main file SQH_FP_u.m is a .mat-file containing the optimal control vector field. The file TEST_CONTROLLED_MC_2D.m performs a random walk with this control vector field, resulting in Figure 4.6. The starting point for the random walk is set in the file TEST_CONTROLLED_MC_2D.m and the diffusion is set in the file model.m.

The description for the DH method is analogous to SQH_PF_u.zip where the codes can be found in DH.zip and the main file is DH.m.

Bibliography

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and applied mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] Hans W. Alt. *Linear Functional Analysis: An Application-Oriented Introduction.* Springer, 2016.

[3] Herbert Amann and Joachim Escher. *Analysis I.* Birkhäuser Basel, 2006.

[4] Herbert Amann and Joachim Escher. *Analysis II.* Birkhäuser, 2008.

[5] Herbert Amann and Joachim Escher. *Analysis III.* Birkhäuser Basel, 2009.

[6] Luigi Ambrosio, Giuseppe Da Prato, and Andrea C.G. Mennucci. *Introduction to Measure Theory and Integration.* Edizioni della Normale, 2011.

[7] Andrey A. Amosov. Weak convergence for a class of rapidly oscillating functions. *Mathematical Notes*, 62(1):122–126, 1997.

[8] Mario Annunziato and Alfio Borzì. A Fokker-Planck control framework for multidimensional stochastic processes. *J. Comput. Appl. Math.*, 237(1):487–507, 2013.

[9] Mario Annunziato and Hanno Gottschalk. Calibration of Lévy processes using optimal control of Kolmogorov equations with periodic boundary conditions. *Mathematical Modeling and Analysis*, 23(3):390–390, 2018.

[10] Heinz H. Bauschke and Patrick L. Combettes. Convex analysis and monotone operator theory in Hilbert spaces. 2017.

[11] Nicola Bellomo and Luigi Preziosi. Modelling and mathematical problems related to tumor evolution and its interaction with the immune system. *Mathematical and Computer Modelling*, 32(3-4):413–452, 2000.

[12] Dimitri P. Bertsekas. *Nonlinear programming.* Athena scientific Belmont, 1999.

[13] Stefano Bianchini, Maria Colombo, Gianluca Crippa, and Laura V. Spinolo. Optimality of integrability estimates for advection–diffusion equations. *Nonlinear Differential Equations and Applications NoDEA*, 24(4):33, 2017.

[14] Lucio Boccardo, Andrea Dall'Aglio, Thierry Gallouët, and Luigi Orsina. Existence and regularity results for some nonlinear parabolic equations. *Adv. Math. Sci. Appl*, 9(2):1017–1031, 1999.

[15] Vladimir I. Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

[16] Vladimir G. Boltyanskiĭ, Revaz V. Gamkrelidze, and Lev S. Pontryagin. On the theory of optimal processes. *Dokl. Akad. Nauk SSSR (N.S.)*, 110:7–10, 1956.

[17] Joseph F. Bonnans. On an algorithm for optimal control using Pontryagin's maximum principle. *SIAM Journal on Control and Optimization*, 24(3):579–588, 1986.

[18] Alfio Borzì and Karl Kunisch. The numerical solution of the steady state solid fuel ignition model and its optimal control. *SIAM Journal on Scientific Computing*, 22(1):263–284, 2000.

[19] Alfio Borzì and Volker Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*, volume 8. SIAM, 2011.

[20] Tim Breitenbach, Mario Annunziato, and Alfio Borzì. On the optimal control of a random walk with jumps and barriers. *Methodology and Computing in Applied Probability*, 20(1):435–462, 2018.

[21] Tim Breitenbach and Alfio Borzì. A sequential quadratic Hamiltonian method for solving parabolic optimal control problems with discontinuous cost functionals. *Journal of Dynamical and Control Systems*, pages 1–33, 2018.

[22] Tim Breitenbach and Alfio Borzì. On the SQH scheme to solve nonsmooth PDE optimal control problems. *Numerical Functional Analysis and Optimization*, pages 1–43, 2019.

[23] Tim Breitenbach and Alfio Borzì. The Pontryagin maximum principle for solving Fokker-Planck optimal control problems, submitted for publication. 2019.

[24] Tim Breitenbach and Alfio Borzì. A sequential quadratic Hamiltonian scheme for solving non-smooth quantum control problems with sparsity, submitted for publication. 2019.

[25] Tim H. Breitenbach, Mario Annunziato, and Alfio Borzì. On the optimal control of random walks. *IFAC-PapersOnLine*, 49(8):248–253, 2016.

[26] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[27] Roland Bulirsch, Josef Stoer, and Klaus Well. *Optimal Control: Calculus of Variations, Optimal Control Theory, and Numerical Methods*. International series of numerical mathematics. Birkhäuser Basel, 1993.

[28] Eduardo Casas. Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations. *SIAM Journal on Control and Optimization*, 35(4):1297–1327, 1997.

[29] Eduardo Casas. Second order analysis for bang-bang control problems of PDEs. *SIAM Journal on Control and Optimization*, 50(4):2355–2372, 2012.

[30] J.S. Chang and G. Cooper. A practical difference scheme for Fokker-Planck equations. *Journal of Computational Physics*, 6(1):1 – 16, 1970.

[31] Constantin Christof, Christian Meyer, Stephan Walther, and Christian Clason. Optimal control of a non-smooth semilinear elliptic equation. *Mathematical Control & Related Fields*, 8(1):247–276, 2018.

[32] Gabriele Ciaramella and Alfio Alfio Borzì. Quantum optimal control problems with a sparsity cost functional. *Numerical Functional Analysis and Optimization*, 37(8):938–965, 2016.

[33] Gabriele Ciaramella and Alfio Borzì. A LONE code for the sparse control of quantum systems. *Computer Physics Communications*, 200:312–323, 2016.

[34] Gabriele Ciaramella, Alfio Borzì, Gunther Dirr, and Daniel Wachsmuth. Newton methods for the optimal control of closed quantum spin systems. *SIAM Journal on Scientific Computing*, 37(1):A319–A346, 2015.

[35] Francis Clarke. *Functional analysis, calculus of variations and optimal control*, volume 264. Springer Science & Business Media, 2013.

[36] Donald L. Cohn. *Measure theory.* Springer, 2013.

[37] David R. Cox and Hilton D. Miller. *The theory of stochastic processes.* John Wiley & Sons, Inc., New York, 1965.

[38] Robert Dautray and Jacques-Louis Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 2.* Springer-Verlag, Berlin, 1988. Functional and variational methods, With the collaboration of Michel Artola, Marc Authier, Philippe Bénilan, Michel Cessenat, Jean Michel Combes, Hélène Lanchon, Bertrand Mercier, Claude Wild and Claude Zuily, Translated from the French by Ian N. Sneddon.

[39] James W. Demmel. *Applied numerical linear algebra*, volume 56. Siam, 1997.

[40] David Devadze and Vakhtang Beridze. Methods of numerical solution of optimal control problems based on the Pontryagin maximum principle. *Journal of Mathematical Sciences*, 206(4):348–356, 2015.

[41] Andrei V. Dmitruk. On the development of Pontryagin's maximum principle in the works of A. Ya. Dubovitskii and A.A. Milyutin. *Control and Cybernetics*, 38(4A):923–957, 2009.

[42] Andrei V. Dmitruk and Nikolai P. Osmolovskii. On the proof of Pontryagin's maximum principle by means of needle variations. *Journal of Mathematical Sciences*, 218(5):581–598, 2016.

[43] Corey M. Dunn. *Introduction to Analysis.* CRC Press, 2017.

[44] Ayla Ergun, Kevin Camphausen, and Lawrence M Wein. Optimal scheduling of radiotherapy and angiogenic inhibitors. *Bulletin of mathematical biology*, 65(3):407–424, 2003.

[45] Lawrence C. Evans. *Partial Differential Equations.* Graduate studies in mathematics. American Mathematical Society, 1998.

[46] Arthur Fleig and Roberto Guglielmi. Optimal control of the Fokker–Planck equation with space-dependent controls. *Journal of Optimization Theory and Applications*, 174(2):408–427, 2017.

[47] Melina-Lorén Kienle Garrido, Tim Breitenbach, Kurt Chudej, and Alfio Borzì. Modeling and numerical solution of a cancer therapy optimal control problem. *Applied Mathematics*, 9:985–1004, 2018.

[48] Matthias Gerdts. *Optimal control of ODEs and DAEs.* Walter de Gruyter, 2011.

[49] Mariano Giaquinta and Stefan Hildebrandt. *Calculus of Variations I.* Springer-Verlag Berlin Heidelberg, 2004.

[50] William W. Hager and Hongchao Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on Optimization*, 16(1):170–192, 2005.

[51] Philip Hahnfeldt, Dipak Panigrahy, Judah Folkman, and Lynn Hlatky. Tumor development under angiogenic signaling. *Cancer research*, 59(19):4770–4775, 1999.

[52] Falk M. Hante. On the relaxation gap for PDE mixed-integer optimal control problems. *PAMM*, 16(1):783–784, 2016.

[53] Michael Hintermüller and Tao Wu. Nonconvex TV^q-models in image restoration: analysis and a trust-region regularization–based superlinearly convergent solver. *SIAM Journal on Imaging Sciences*, 6(3):1385–1415, 2013.

[54] Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*, volume 23. Springer Science & Business Media, 2008.

[55] Kazufumi Ito and Karl Kunisch. *Lagrange multiplier approach to variational problems and applications*, volume 15. Siam, 2008.

[56] Kazufumi Ito and Karl Kunisch. A variational approach to sparsity optimization based on lagrange multiplier theory. *Inverse problems*, 30(1):015001, 2013.

[57] Kazufumi Ito and Karl Kunisch. Optimal control with $L^p(\Omega)$, $p \in [0, 1)$, control cost. *SIAM Journal on Control and Optimization*, 52(2):1251–1275, 2014.

[58] Johannes Jahn. *Introduction to the theory of nonlinear optimization.* Springer Science & Business Media, 2007.

[59] Vasilii V. Jikov, Sergei M. Kozlov, and Olga A. Oleinik. *Homogenization of differential operators and integral functionals.* Springer Science & Business Media, 2012.

[60] Boško S. Jovanović and Endre Süli. *Analysis of Finite Difference Schemes*, volume 46. Springer London, London, 2014.

[61] Veronika Karl and Daniel Wachsmuth. An augmented Lagrange method for elliptic state constrained optimal control problems. *Computational Optimization and Applications*, 69(3):857–880, 2018.

[62] Ivan A. Krylov and Feliks L. Chernous'ko. On a method of successive approximations for the solution of problems of optimal control. *USSR Computational Mathematics and Mathematical Physics*, 2(6):1371–1382, 1963.

[63] Ivan A. Krylov and Feliks L. Chernous'ko. An algorithm for the method of successive approximations in optimal control problems. *USSR Computational Mathematics and Mathematical Physics*, 12(1):15–38, 1972.

[64] Olga A. Ladyzhenskaia, Vsevolod A. Solonnikov, and Nina N. Ural'tseva. *Linear and Quasi-Linear Equations of Parabolic Type*, volume 23. American Mathematical Soc., 1988.

[65] Vangipuram Lakshmikantham and Srinivasa Leela. *Differential and Integral Inequalities: Theory and Applications: Volume I: Ordinary Differential Equations.* Academic press, 1969.

[66] Qianxiao Li and Shuji Hao. An optimal control approach to deep learning and applications to discrete-weight neural networks. *arXiv preprint arXiv:1803.01299*, 2018.

[67] Xunjing Li and Jiongmin Yong. *Optimal Control Theory for Infinite Dimensional Systems.* Optimal Control Theory for Infinite Dimensional Systems. Birkhäuser, 1995.

[68] David G. Luenberger. *Optimization by vector space methods.* John Wiley & Sons, 1969.

[69] Alexey A. Lyubushin. Modifications of the method of successive approximations for solving optimal control problems. *USSR Computational Mathematics and Mathematical Physics*, 22(1):29–34, 1982.

[70] José M. Martínez. Minimization of discontinuous cost functions by smoothing. *Acta Applicandae Mathematica*, 71(3):245–260, 2002.

[71] Masoumeh Mohammadi and Alfio Borzì. Analysis of the Chang-Cooper discretization scheme for a class of Fokker-Planck equations. *J. Numer. Math.*, 23(3):271–288, 2015.

[72] Vicente Montesinos, Peter Zizler, and Václav Zizler. *An introduction to modern analysis.* Springer, 2015.

[73] Mikhail S. Nikol'ski. On the convergence problem of a certain variant of the successive-approximation method for solving optimal control problems. *Journal of Mathematical Sciences*, 139(5):6902–6908, 2006.

[74] Louis Nirenberg. On elliptic partial differential equations. In *Il principio di minimo e sue applicazioni alle equazioni funzionali*, pages 1–48. Springer, 2011.

[75] Sergei Ovchinnikov. *Functional Analysis: An Introductory Course.* Universitext. Springer International Publishing, 2018.

[76] Francesco Petitta. *A Not So Long Introduction to the Weak Theory of Parabolic Problems with Singular Data.* 2007. Lecture nodes of a course held in Granada October-December. University of Roma 1, http://www1.mat.uniroma1.it/people/orsina/EDP/EDP02.pdf.

[77] Lev S. Pontryagin, Vladimir G. Boltyanskiĭ, Revaz V. Gamkrelidze, and Evgenii F. Mishchenko. *The Mathematical Theory of Optimal Processes.* John Wiley & Sons, New York-London, 1962.

[78] Charles C. Pugh. *Real Mathematical Analysis.* Undergraduate Texts in Mathematics. Springer, Cham, second edition, 2015.

[79] Inder K. Rana. *An introduction to measure and integration*, volume 45. American Mathematical Soc., 2002.

[80] Jean-Pierre Raymond and Hasnaa Zidani. Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls. *SIAM Journal on Control and Optimization*, 36(6):1853–1879, 1998.

[81] Jean-Pierre Raymond and Hasnaa Zidani. Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations. *Applied Mathematics and Optimization. An International Journal with Applications to Stochastics*, 39(2):143–177, 1999.

[82] Ralph T. Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[83] Tomáš Roubíček. *Nonlinear partial differential equations with applications*, volume 153. Springer Science & Business Media, 2013.

[84] Souvik Roy, Mario Annunziato, Alfio Borzì, and Christian Klingenberg. A Fokker–Planck approach to control collective motion. *Computational Optimization and Applications*, 69(2):423–459, 2018.

[85] Yoshiyuki Sakawa and Yuji Shindo. On global convergence of an algorithm for optimal control. *IEEE Transactions on Automatic Control*, 25(6):1149–1153, 1980.

[86] Heinz Schättler and Urszula Ledzewicz. *Optimal Control for Mathematical Models of Cancer Therapies.* Springer, 2015.

[87] Ernest Scheiber. Chernousko-Lyubushin version of the succesive approximation method for optimal control problem revisited. *Bulletin of the Transilvania University of Brasov Vol*, 6(55), 2013.

[88] Yuji Shindo and Yoshiyuk Sakawa. Local convergence of an algorithm for solving optimal control problems. *Journal of optimization theory and applications*, 46(3):265–293, 1985.

[89] Ilya Shvartsman and Zuhra Mingaleeva. Second-order optimality conditions for singular Pontryagin local minimizers. *Numerical Functional Analysis and Optimization*, 35(7-9):1245–1257, 2014.

[90] Eduardo D. Sontag. *Mathematical control theory, volume 6 of Texts in Applied Mathematics.* Springer-Verlag, New York, 1998.

[91] Josef Stoer, Roland Bulirsch, Richard H. Bartels, Walter Gautschi, and Christoph Witzgall. *Introduction to Numerical Analysis.* Texts in applied mathematics. Springer, New York, 2002.

[92] Mikhail I. Sumin. Suboptimal control of systems with distributed parameters: normality properties and a dual subgradient method. *Computational Mathematics and Mathematical Physics*, 37(2):158–174, 1997.

[93] Mikhail I. Sumin. The first variation and Pontryagin's maximum principle in optimal control for partial differential equations. *Computational Mathematics and Mathematical Physics*, 49(6):958–978, 2009.

[94] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Society Providence, 2012.

[95] Fredi Tröltzsch. *Optimal Control of Partial Differential Equations*, volume 112 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2010. Theory, methods and applications.

[96] Michael Ulbrich. *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, volume 11. SIAM, 2011.

[97] Oleg V. Vasiliev and Aleksandr I. Tyatyushkin. A method for solving optimal control problems that is based on the maximum principle. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 21(6):1376–1384, 1981.

[98] Lawrence M. Wein, Jonathan E. Cohen, and Joseph T. Wu. Dynamic optimization of a linear–quadratic model with incomplete repair and volume-dependent sensitivity and repopulation. *International Journal of Radiation Oncology* Biology* Physics*, 47(4):1073–1083, 2000.

[99] Jan H. Witte and Christoph Reisinger. A penalty method for the numerical solution of Hamilton-Jacobi-Bellman (HJB) equations in finance. *SIAM Journal on Numerical Analysis*, 49(1):213–231, 2011.

[100] Zhuoqun Wu, Jingxue Yin, and Chunpeng Wang. *Elliptic & Parabolic Equations.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2006.