A sequential quadratic Hamiltonian method for solving parabolic optimal control problems with discontinuous cost functionals

Tim Breitenbach* Alfio Borzì†

Abstract

A sequential quadratic Hamiltonian (SQH) method for solving control-constrained parabolic optimal control problems with continuous and discontinuous non-convex cost functionals is investigated. The solution to these problems is characterised by the Pontryagin's maximum principle, which is also the starting point for the development of a sequential quadratic Hamiltonian scheme. In a general setting that includes discontinuous and non-convex cost functionals, it is proved that the SQH method is well-defined; however, convergence to an optimal solution is proved only in the smooth case. Results of numerical experiments are presented that successfully validate the proposed optimisation framework and demonstrate its effectiveness and large applicability.

This is a preprint of the paper

Tim Breitenbach and Alfio Borzì ,
A Sequential Quadratic Hamiltonian Method for Solving Parabolic
Optimal Control Problems with Discontinuous Cost Functionals
appeared in Journal of Dynamical and Control Systems
https://doi.org/10.1007/s10883-018-9419-6

Keywords – Parabolic optimal control problems, discontinuous cost functionals, Pontryagin maximum principle, iterative scheme
MSC – 49K20, 49M05, 65K10

1 Introduction

Optimal control of parabolic models with cost functionals for which necessary first-order conditions can be reformulated as semi-smooth equations is a well developed modern research topic; see, e.g., [22, 45] and references therein. In this framework, optimal solutions are characterised by first-order

*Institut für Mathematik, Universität Würzburg, Emil-Fischer-Strasse 30, 97074 Würzburg, Germany; tim.breitenbach@mathematik.uni-wuerzburg.de
†Institut für Mathematik, Universität Würzburg, Emil-Fischer-Strasse 30, 97074 Würzburg, Germany; alfio.borzi@mathematik.uni-wuerzburg.de

optimality conditions that require semi-smoothness of the reduced cost functional allowing the development of different solution procedures like proximal methods [39] and semi-smooth Newton-methods [45]. However, in the case of cost functionals that are not Lipschitz continuous and in the much less investigated case of discontinuous cost functionals, the property of semi-smoothness is lost and the optimisation techniques mentioned above cannot be used, unless regularisation is considered at the cost of modifying the nature of the problem; see, e.g., [32].

We mainly focus on discontinuous and non-convex cost functionals as the most challenging benchmark that can be addressed by the method proposed in this work. In particular, we consider a cost functional given by the following: $\int_0^T \int_\Omega g\left(u\left(x,t\right)\right) dxdt$ where the control's cost is evaluated with the lower semi-continuous function

$$g\left(z\right) = \begin{cases} |z| & \text{if } |z| > s \\ 0 & \text{otherwise} \end{cases}$$

where $s > 0$.

Our numerical approach has obviously much larger applicability. Discontinuous cost functionals and, more generally, discontinuous variational problems appear already in the study of jet flows, cavity problems and in plasma physics [6, 29]. However their numerical realisation has been hindered by the lack of appropriate solvers. In this framework, the purpose of our work is to contribute to the field of non-smooth optimisation with partial differential equations (PDEs) by developing numerical tools that apply to non-regularised distributed parabolic optimal control problems with discontinuous and non-convex cost functionals. For this purpose, we deal with the optimal control theory based on the Pontryagin's maximum principle (PMP) [9, 17, 35] that was originally developed for control problems governed by ordinary differential equations (ODEs) and has been much less investigated in the context of time-dependent PDE models; see, e.g., [14, 30, 36, 42, 43]. In particular, we focus on the work [36] to characterise a solution to our parabolic control problems with a necessary optimality condition provided by the PMP. For these problems we briefly address the issue of existence of optimal controls and then turn our attention to their PMP characterisation.

Although the PMP principle represents a powerful theoretical tool, its use in the PDE context has been hindered by the lack of efficient numerical implementation. In fact, well-known direct and indirect methods used to solve ODE control problems are difficult to apply in the higher space-time dimensional setting of PDE problems and in the case of large-size ODE problems. It is the main purpose of the present work to address this issue by developing an efficient optimisation scheme that is consistent with the PMP framework. For this purpose, we notice that a natural approach to solve PDE models is by iterative methods that exploit the sparsity of discretized PDEs. Moreover, we point out that the PMP principle has a pointwise formulation and even its proof is by 'needle variations', and both are local in structure. Therefore it seems natural to consider iterative strategies that implement local updates of the control function pointwise in space and time.

For this reason, we focus on the iterative scheme first proposed in [38] and further discussed in [10, 40] to solve ODE control problems by an augmented Hamiltonian technique and discuss their extension to our PDE setting. Moreover, we also would like to mention the earlier works [26, 27] where different so-called successive iteration solvers based on the minimisation of the Pontryagin-Hamilton function are considered. Specifically, in [26] the Hamiltonian without augmentation is used to find an update for the control, but, as the authors mention, the issue of convergence

remains open for this approach. In any case, we have implemented the method in [26] and found that it has difficulty to cope with our cost functionals. In [27], modifications of the method in [26] are discussed that transform the state equation to obtain a weakly controlled problem or use a damping of the control update. The third proposed alternative is to restrict the change of the control to a short time window. In our opinion, the augmented Hamiltonian approach can be seen having the same purpose: keep the updates of the control conveniently small.

We remark that the iterative schemes in [26, 27] are designed so that the values of the state variable from the previous iteration are used while computing the update of the control in the new iteration sweep. This important feature is also characteristic of the method that we propose in this paper. Therefore we could say that our approach includes different aspects of the methods proposed in [38, 40] and in [26, 27].

In our approach, we pointwise minimise an augmented Hamiltonian function to find an update for the control that provides a better cost functional value. In doing this, we use the state function of the governing model from the previous iteration, thus avoiding to recalculate this function every time after a local control update as in [38, 40]. Our procedure results in a much smaller number of solving of the state equation, which is necessary since these calculations are very costly in the case of partial differential equations. In this way, we formulate a new efficient and robust iterative procedure that is able to solve discontinuous optimisation problems while not relying on regularisation techniques as in [21, 23, 24]. We would like to name our method the sequential quadratic Hamiltonian (SQH) scheme and show that it is able to solve discontinuous time-dependent parabolic optimal control problems with almost linear computational complexity. To the best of our knowledge, there is no other available methodology with similar capability. In this paper, we theoretically discuss the convergence properties of our iterative SQH method and demonstrate numerically its effectiveness.

In the next section, we formulate a class of parabolic optimal control problems and discuss the necessary functional estimates. In Section 3, we discuss the Pontryagin maximum principle for the chosen parabolic optimal control problems. We outline the proof of the PMP principle (Theorem 3.3), which is analogous to [36], providing a necessary optimality condition for our optimisation problems with a discontinuous cost of the control. In Section 4, we discuss our PMP-based SQH scheme. We discuss how our scheme provides updates to the control that correspond to monotonically decreasing cost functional values for our optimisation problem. Furthermore, in the case of a differentiable cost functional, we prove convergence of the SQH sequence to a local optimal solution in the sense that appropriate first-order optimality conditions are satisfied. The main results of Section 4 are Theorem 4.1 and Theorem 4.2. In Section 5, results of numerical experiments considering different cost functionals are presented that successfully demonstrate the almost optimal complexity of our SQH scheme and its robustness with respect to changes of the values of the optimisation parameters. In particular, we show that the solution obtained with the SQH scheme fulfils the PMP optimality condition. Furthermore, to allow a comparison with a well-known optimisation scheme as the non-linear conjugated gradient (NCG) method, we consider the case of a continuous cost functional and compare our iterative scheme with the NCG method. Notice that a direct comparison of the SQH scheme with the method in [38] results obviously in favour of our method since the latter requires a prohibitive large number of forward solves.

In order to provide all technical details supporting our work and to make this work self-contained as far as possible, we include an appendix. In the Appendix, we prove an L^∞-result for linear parabolic partial differential equations that is essential for the PMP characterization of

parabolic PDE control problems. We include this result since it is usually stated without proof by making references to [28] where the required result is embedded in a more general framework. Further, in the attempt to give a theoretical support to our framework with discontinuous cost functionals, we prove existence of a minimiser for this optimisation setting in the case of a compact admissible control set. A section of conclusion completes this work.

2 Formulation of the optimal control problem

In this section, we formulate our parabolic optimal control problems with discontinuous cost functionals and discuss existence of optimal controls. Our governing model is a heat equation with distributed control that is defined in the space-time cylinder $Q = \Omega \times (0,T)$, $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, where Ω is an open bounded domain with a smooth boundary. We choose homogeneous Dirichlet boundary conditions and an initial condition $y_0 \in L^\infty(\Omega)$.

For each $t \in (0,T)$, $T > 0$, the weak formulation of the resulting initial-boundary value problem is as follows: Find $y \in L^2(0,T;H_0^1(\Omega))$ and $y' \in L^2(0,T;H^{-1}(\Omega))$, that means, $y \in W(0,T) := \{y \in L^2(0,T;H_0^1(\Omega)) \mid y' \in L^2(0,T;H^{-1}(\Omega))\}$; see [44, Chapter 3], such that the following is satisfied

$$\begin{aligned} (y'(\cdot,t),v) + D(\nabla y(\cdot,t),\nabla v) &= (u(\cdot,t),v) \quad \text{in Q} \\ y(\cdot,0) &= y_0 \quad \text{on } \Omega \times \{t=0\} \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

for all $v \in H_0^1(\Omega)$. 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.

Requiring $u \in L^q(Q)$, $q > \frac{n}{2}+1$ if $n \geq 2$ and $q \geq 2$ if $n = 1$, we have that there exists an unique solution $y \in W(0,T)$ to (1), see [19, Chapter 7.1, Theorem 3] as $y_0, \in L^2(\Omega)$ and $u \in L^2(Q)$, see [1, Theorem 2.14]. However, for the aim of the Pontryagin maximum principle, this regularity result needs to be improved. For this purpose, we require $y_0 \in H_0^1(\Omega) \cap L^\infty(\Omega)$. Then we have $y \in L^2(0,T;H^2(\Omega)) \cap L^\infty(0,T;H_0^1(\Omega))$, see [19, Chapter 7 Theorem 5], such that we can apply Theorem 6.1 (Appendix) and have the following theorem.

Theorem 2.1. *Let y be the solution to (1). Then, y is essentially bounded by*

$$\|y\|_{L^\infty(Q)} \leq \|y_0\|_{L^\infty(\Omega)} + C\|u\|_{L^q(Q)},$$

where $C > 0$ is a constant.

Furthermore, we have that the control-to-state map, $S : L^q(Q) \to W(0,T)$, $u \mapsto y = S(u)$ is affine and continuous; see [44, (3.36)] and [1, Theorem 2.14] for the continuous embedding of $L^q(Q) \to L^2(Q)$. Moreover, the map S is continuous as a map $S : L^q(Q) \to L^2(Q)$, since $\|y\|^2_{L^2(Q)} = \int_0^T \|y\|^2_{L^2(\Omega)} dt \leq \int_0^T \|y\|^2_{H_0^1(\Omega)} dt = \|y\|_{L^2\left(0,T;H_0^1(\Omega)\right)}$.

Next, we discuss the following parabolic optimal control problem

$$\begin{aligned} &\min_{y,u} J(y,u) \\ \text{s.t. } &(y',v) + D(\nabla y, \nabla v) = (u,v) && \text{in Q} \\ &y(\cdot,0) = y_0 && \text{on } \Omega \times \{t=0\} \\ &y = 0 && \text{on } \partial\Omega \\ &u \in U_{ad}, \end{aligned} \tag{2}$$

where the cost functional J is given by

$$J(y,u) := J_c(y,u) + \gamma \int_Q g(u(x,t))\, dxdt. \tag{3}$$

In this functional, J_c represents a smooth functional objective as it appears in many control problems [11, 44]. We have

$$J_c(y,u) := \frac{1}{2}||y - y_d||^2_{L^2(Q)} + \frac{\alpha}{2}||u||^2_{L^2(Q)}, \qquad \alpha \geq 0. \tag{4}$$

In this case, the functional J_c models the task of driving the state y to track a desired state trajectory $y_d \in L^q(Q)$, 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

$$G(u) := \gamma \int_Q g(u(x,t))\, dxdt, \qquad \gamma \geq 0, \tag{5}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a non-negative and lower semi-continuous function.

In particular, we consider the case where

$$g(u) = \begin{cases} |u| & \text{if } |u| > s \\ 0 & \text{otherwise} \end{cases}$$

where $s > 0$. With this construction, we obtain a cost of the control that is zero if its value is below a given threshold and it measures a L^1 cost otherwise. Notice that with this choice the reduced cost functional $\hat{J}(u) := J(S(u), u)$ is discontinuous in $L^q(Q)$.

The admissible set of controls is defined as follows

$$U_{ad} := \{u \in L^q(Q) \mid u(x,t) \in K_U\}, \tag{6}$$

where K_U is a compact subset of $\mathbb{R}$.

In the case where G is a convex and continuous cost functional, existence of an optimal control is guaranteed [44]. However, in the case of discontinuous cost functionals the issue of existence of an optimal control is more delicate because the property of weakly lower semi-continuous is lost. However, existence of an optimal solution can be proved considering a set of admissible controls that is compact in $L^q(Q)$; see Theorem 6.2 in the Appendix. For our purpose, we assume that (2) admits a solution in U_{ad} given in (6).

3 The Pontryagin's maximum principle

In this section, we discuss the characterisation of optimal controls in U_{ad} in the framework of the Pontryagin's maximum principle in its easier variant with no final state constraints; see, e.g., [30, 36].

First, we illustrate the main theoretical steps in the derivation of the Pontryagin's maximum principle, also with the purpose to introduce essential concepts that are instrumental for the discussion on our SQH scheme. In the following, the notation $var1 \leftarrow var2$ means that the variable $var1$ is set equal to $var2$.

We formulate the following adjoint problem

$$\begin{aligned} \left(-p'\left(\cdot,t\right),v\right)+D\left(\nabla p\left(\cdot,t\right),\nabla v\right) &= \left(y\left(\cdot,t\right)-y_d\left(\cdot,t\right),v\right) && \text{in } Q \\ p\left(\cdot,T\right) &= 0 && \text{on } \Omega\times\{T=0\} \\ p &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{7}$$

This problem has the same structure as (1) after a transformation of the time variable $t := T-\tau$ and noticing that $y-y_d \in L^q\left(Q\right)$, see [1, Theorem 2.14]. Hence, there exists an unique $p \in L^2\left(0,T;H_0^1\left(\Omega\right)\right)$ and $p' \in L^2\left(0,T;H^{-1}\left(\Omega\right)\right)$ solving (7) for all $v \in H_0^1\left(\Omega\right)$.

Next, we define the Hamiltonian corresponding to (2) - (4) as follows

$$H\left(x,t,y,u,p\right)=\frac{1}{2}\left(y-y_d\right)^2+\frac{\alpha}{2}u^2+\gamma g\left(u\right)+p\,u, \tag{8}$$

where $H : \mathbb{R}^n \times \mathbb{R}_0^+ \times \mathbb{R} \times K_U \times \mathbb{R} \to \mathbb{R}$. If y, u and p are functions, then $H\left(x,t,y,u,p\right)$ stands short for $H\left(x,t,y(x,t),u(x,t),p(x,t)\right)$. We remark that throughout this work, we sometimes drop the arguments (x,t) of functions to save notational effort when the functional dependence is clear from the context.

The classical approach to prove the PMP principle is by the method of needle variation [17, 36]. For this purpose, let $S_k\left(x_0,t_0\right)$ be an open ball centered at $(x_0,t_0) \in Q$ with radius $s^k_{x_0,t_0}$ such that the Lebesgue measure of the ball tends to zero as $k \to \infty$, $\lim_{k\to\infty}\left|S_k\left(x_0,t_0\right)\right|=0$. Analogous to [36], we define the needle variation at (x_0,t_0) of an admissible control $u^* \in U_{ad}$ as follows

$$u_k\left(x,t\right) := \begin{cases} u^*\left(x,t\right) & \text{on } \; Q\backslash S_k\left(x_0,t_0\right) \\ u & \text{in } \; S_k\left(x_0,t_0\right)\cap Q \end{cases} \tag{9}$$

where $u \in K_U$. Notice that we consider a single needle variation as in [17, 30, 36].

We remark that the function $u_k \in U_{ad}$ for all $k \in \mathbb{N}$, for all $(x_0,t_0) \in Q$ and $u^* \in U_{ad}$. This can be seen as follows. The function $u_k = u^*\chi_{Z\backslash S_k(x_0,t_0)} + u\chi_{S_k(x_0,t_0)}$ is measurable for all $k \in \mathbb{N}$ and $(x_0,t_0) \in Q$ because the sum and the product of measurable functions is measurable, see [15, Proposition 2.1.7] and the characteristic function χ_A is measurable if and only if A is measurable, see [15, Example 2.1.2], the needle variation is measurable. As the image of the needle variation is in K_U almost everywhere and $\left(\int_Q \left|u_k\left(x,t\right)\right|^q dxdt\right)^{\frac{1}{q}} \leq \max\left(\left|u_a\right|,\left|u_b\right|\right)\left|Q\right|^{\frac{1}{q}}$ with $|Q|$ the Lebesgue measure of Q, the needle variation $u_k \in L^q\left(Q\right)$.

Next, we define the intermediate adjoint equation

$$\begin{aligned} \left(-\tilde{p}'\left(\cdot,t\right),v\right)+D\left(\nabla\tilde{p}\left(\cdot,t\right),\nabla v\right) &= \left(\frac{1}{2}\left(y_1\left(\cdot,t\right)+y_2\left(\cdot,t\right)\right)-y_d\left(\cdot,t\right),v\right) && \text{in } Q \\ \tilde{p}\left(\cdot,T\right) &= 0 && \text{on } \Omega\times\{T=0\}, \end{aligned} \tag{10}$$

with zero boundary conditions where y_1 is the solution to (1) for $u \leftarrow u_1$ and y_2 is the solution to (1) for $u \leftarrow u_2$. Analogously to (7), after setting $t := T - \tau$ and because $\frac{1}{2}(y_1 + y_2) - y_d \in L^q(Q)$, one can prove that the problem (10) has a unique solution $\tilde{p} \in L^2(0,T;H_0^1(\Omega))$ and $\tilde{p}' \in L^2(0,T;H^{-1}(\Omega))$. In addition, similarly to the forward equation (1), we also have $p, \tilde{p} \in L^2(0,T;H^2(\Omega)) \cap L^\infty(0,T;H_0^1(\Omega))$ as $p(\cdot,T) = \tilde{p}(\cdot,T) = 0$, and hence $p(\cdot,T), \tilde{p}(\cdot,T) \in L^\infty(\Omega) \cap H_0^1(\Omega)$. Thus, we can establish the following theorem.

Theorem 3.1. *The solution to (7) and the solution to (10) are essentially bounded.*

Proof. We consider the time transformation $t := T - \tau$ and set $p(t) \leftarrow p(T - \tau)$. Then, $\frac{\partial}{\partial t} p = -\frac{\partial}{\partial \tau} p$. As $y \in L^q(Q)$ according to Theorem 2.1 and $y_d \in L^q(Q)$, we can apply Theorem 2.1 again to the solutions of (7) and (10) as $p(\cdot,T) = 0$, and so $p(\cdot,T) \in L^\infty(\Omega) \cap H_0^1(\Omega)$. □

Now, we prove the following convergence result.

Theorem 3.2. *Let $u^* \in L^q(Q)$ and y^* be the solution to (1) for $u \leftarrow u^*$. Let p^* be the corresponding solution to (7) for $y \leftarrow y^*$. Let u_k be defined in (9), y_k be the solution to (1) for $u \leftarrow u_k$ as well as $\tilde{p}_k$ the corresponding solution to (10) for $y_1 \leftarrow y^*$, $y_2 \leftarrow y_k$. Then, y_k converges to y^* in $L^\infty(Q)$ and $\tilde{p}_k$ converges to p^* in $L^\infty(Q)$.*

Proof. We have $(x,t) \mapsto u^*(x,t) \in L^\infty(Q)$; therefore almost every point of Q is a Lebesgue point of $(x,t) \mapsto u^*(x,t)$, see [19, page 649, Theorem 6]. This means that

$$\left(\int_Q |u_k(x,t) - u^*(x,t)|^q dx dt \right)^{\frac{1}{q}} = \left(\int_{S_k(x_0,t_0)} |u - u^*(x,t)|^q dx dt \right)^{\frac{1}{q}} \underset{k \to \infty}{\longrightarrow} 0,$$

for almost every point (x_0,t_0) of Q as $(x,t) \mapsto u - u^*(x,t) \in L^q(Q)$. Therefore $||u_k - u||_{L^q(Q)} \to 0$ for $k \to \infty$. Taking the difference of the heat equation (1) with the two controls $u \leftarrow u^*$ and $u \leftarrow u_k$, we obtain

$$\begin{aligned} (z_k', v) + D(\nabla z_k, \nabla v) &= (u_k - u^*, v) && \text{in Q} \\ z_k(\cdot,0) &= 0 && \text{on } \Omega \times \{t = 0\}, \end{aligned}$$

where $z_k := y_k - y^*$. From Theorem 2.1, we have that $||z_k||_{L^\infty(Q)} \to 0$ for $k \to \infty$ because $||u_k - u||_{L^q(Q)} \to 0$ for $k \to \infty$, see [1, Theorem 2.14]. Similarly, consider $z_k = \tilde{p}_k - p^*$. Subtract the intermediate adjoint (10) for $y_1 \leftarrow y^*$, $y_2 \leftarrow y_k$ from the adjoint equation (7) with $y \leftarrow y^*$. Then, we obtain

$$\begin{aligned} (-z_k', v) + D(\nabla z_k, \nabla v) &= \left(\frac{1}{2}(y_k - y^*), v \right) && \text{in Q} \\ z_k(\cdot,T) &= 0 && \text{on } \Omega \times \{t = 0\}. \end{aligned}$$

Analogous to Theorem 2.1 and the proof of Theorem 3.1, we have that $||z_k||_{L^\infty(Q)} \to 0$ for $k \to \infty$ if $||y_k - y^*||_{L^{\frac{n}{2}+1}(Q)} \to 0$ for $k \to \infty$. Because of [1, Theorem 2.14], this is actually the case as already $||y_k - y^*||_{L^\infty(Q)} \to 0$ for $k \to \infty$ according to the above discussion. □

Next, we define the following function $F : \mathbb{R}^n \times \mathbb{R}_0^+ \times \mathbb{R} \times K_U \to \mathbb{R}$ as follows

$$F(x,t,y,u) := \frac{1}{2}(y - y_d)^2 + \frac{\alpha}{2}u^2 + \gamma g(u).$$

Notice that $F(x,t,y,u)$ stands short for $F(x,t,y(x,t),u(x,t))$ if y or u are functions and $J(y,u) = \int_Q F(x,t,y,u)\,dxdt$.

We remark that, in contrast to [36], we do not require that $F(x,t,y,\cdot)$ is continuous on $\mathbb{R}$. Next, we provide two basic results for proving the Pontryagin's maximum principle that we use to characterise a solution to (2). The proofs are analogous to the corresponding ones in [36, Section 4].

Lemma 3.1. *The following equation holds*

$$J(y_1,u_1) - J(y_2,u_2) = \int_Q \left(H(x,t,y_2,u_1,\tilde{p}) - H(x,t,y_2,u_2,\tilde{p})\right) dxdt,$$

where y_1 is the solution to (1) for $u \leftarrow u_1 \in L^q(Q)$, y_2 is the solution to (1) for $u \leftarrow u_2 \in L^q(Q)$ and $\tilde{p}$ is the solution to (10).

Lemma 3.2. *Let $u^* \in U_{ad}$ be an admissible control and $u \in K_U$. Furthermore, let u_k be defined as in (9), for all $k \in \mathbb{N}$, and y_k be the solution to (1) for $u \leftarrow u_k$. Then, the following holds*

$$\lim_{k\to\infty} \frac{1}{|S_k(x,t)|}\left(J(y_k,u_k) - J(y^*,u^*)\right) = H(x,t,y^*,u,p^*) - H(x,t,y^*,u^*,p^*), \tag{11}$$

for almost all $(x,t) \in Q$ where y^* is the solution to (1) for $u \leftarrow u^*$ and p^* is the corresponding solution to (7) for $y \leftarrow y^*$.

Theorem 3.3. *Let $(\bar{y},\bar{u},\bar{p})$ be an optimal solution to (2) where $\bar{y}$ is the solution to (1) for $u \leftarrow \bar{u}$ and $\bar{p}$ is the solution to (7) for $u \leftarrow \bar{u}$ and $y \leftarrow \bar{y}$. Then, the following holds*

$$H(x,t,\bar{y},\bar{u},\bar{p}) = \min_{u\in K_U} H(x,t,\bar{y},u,\bar{p}), \tag{12}$$

for almost every $(x,t) \in Q$.

4 A sequential quadratic Hamiltonian scheme

This section is devoted to the formulation and theoretical investigation of our sequential quadratic Hamiltonian (SQH) scheme for solving the parabolic optimal control problem (2) - (3). The starting point for the formulation of our SQH scheme is the idea of point-by-point (in a numerical grid defined later) implementation of (12), having in mind the results of Lemma 3.1 and Lemma 3.2. Notice that a similar idea has been successful in the Lagrange framework in the case of differentiable cost functionals, leading to the formulation of collective Gauss-Seidel schemes and efficient multigrid methods for optimality systems; see, e.g., [11]. As already discussed in the introduction, the SQH scheme represents a further development of the schemes proposed in [26, 27] and in [38, 40] in the context of ODE control problems. We remark that the SQH approach could formally be interpreted as a sequential quadratic programming method [13], which explains the

naming of our iterative procedure. This procedure is characterised by two important features: 1) a quadratic penalisation of the control's updates; 2) at the given iterate, the computation of the values of state variable after the control update at all points has been completed.

In the SQH method, the Hamiltonian (8) is augmented with the term $\epsilon\left(u\left(x,t\right)-v\left(x,t\right)\right)^2$. Thus, we define the following augmented Hamiltonian

$$K_\epsilon\left(x,t,y,u,v,p\right):=H\left(x,t,y,u,p\right)+\epsilon\left(u\left(x,t\right)-v\left(x,t\right)\right)^2, \tag{13}$$

where $K_\epsilon:\mathbb{R}^n\times\mathbb{R}_0^+\times\mathbb{R}\times K_U\times K_U\times\mathbb{R}\to\mathbb{R}$ and $\epsilon>0$. Notice that $K_\epsilon\left(x,t,y,u,v,p\right)$ stands short for $K_\epsilon\left(x,t,y\left(x,t\right),u\left(x,t\right),v\left(x,t\right),p\left(x,t\right)\right)$ if y, u, v or p are functions. Roughly speaking, the quadratic term $\epsilon\left(u\left(x,t\right)-v\left(x,t\right)\right)^2$ aims at penalising local control updates that differ too much from the current control value. This in turn prevents the corresponding state y to take values at (x,t) that differ too much from the current value, see Lemma 3.2. Therefore we can reasonably pursue to update the state variable after the control has been updated at all (x,t) points.

The basic idea in developing the SQH scheme is to minimise K_ϵ on K_U at each point (x,t) in some given order; e.g., lexicographically. For this purpose, there are several ways to calculate the element of K_U which minimizes K_ϵ at any given point of the space-time cylinder. First of all, one can discretize K_U and choose the corresponding minimising 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 the minimum of the augmented Hamiltonian up to a given tolerance. Third, one can use an analytical formula for the minima 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.

The main difference of our scheme with respect to the algorithm in [38, 40], and similar to [26], is that, in the minimisation process, we use $K_\epsilon\left(x,t,y^k,u,u^k,p^k\right)$ instead of $K_\epsilon\left(x,t,y^{k+1},u,u^k,p^k\right)$. In fact in [10, 38, 40] 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 minimising 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 minimisation 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 penalising 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 [38].

Algorithm 4.1 (SQH method)**.**

1. *Choose* $\epsilon>0$, $\kappa>0$, $\sigma>1$, $\zeta\in(0,1)$, $\eta\in(0,\infty)$, u^0, *compute* y^0 *and* p^0, *set* $k\leftarrow 0$
2. *Set*
$$\tilde{u}\left(x,t\right)=\underset{w\in K_U}{\operatorname{argmin}}\,K_\epsilon\left(x,t,y^k,w,u^k,p^k\right)$$
for all $(x,t)\in Q$.
3. *Calculate* $\tilde{y}$ *corresponding to* $\tilde{u}$ *and compute* $\tau:=\|\tilde{u}-u^k\|^2_{L^2(Q)}$.
4. *If*
$$J\left(\tilde{y},\tilde{u}\right)-J\left(y^k,u^k\right)>-\eta\,\tau\text{: Choose }\epsilon\leftarrow\sigma\epsilon$$
Else if

$J(\tilde{y},\tilde{u}) - J\left(y^k,u^k\right) \leq -\eta\,\tau$: *Choose* $\epsilon \leftarrow \zeta\epsilon$, $y^{k+1} \leftarrow \tilde{y}$, $u^{k+1} \leftarrow \tilde{u}$; *compute* p^{k+1} *corresponding to* y^{k+1} *and* u^{k+1} *and set* $k \leftarrow k+1$

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

Algorithm 4.1 works as follows. After choosing the problem's parameters and an initial guess for the control, we determine $\tilde{u}$ such that the augmented Hamiltonian is minimised for a given state, adjoint, current control and ϵ. If the resulting control $\tilde{u}$ and the corresponding $\tilde{y}$ do not minimise the cost functional more than $-\eta\tau$ with respect to the former values y^k and u^k, we increase ϵ and perform the minimisation of the resulting K_ϵ again. Else, we accept the new control function as well as the corresponding state, calculate the adjoint and diminish ϵ such that greater variations of the control value become more likely. If the convergence criterion $\tau < \kappa$ is not fulfilled, then in the SQH scheme the minimisation procedure is repeated. If the convergence criterion is fulfilled, then the algorithm stops and returns the last calculated control u^k.

Next, we prove that for given x,t,y,v,p and ϵ there exists a $u(x,t) \in K_U$ that minimises $K_\epsilon(x,t,y,u,v,p)$. Thus, Step 2 of Algorithm 4.1 is well posed. Later, we prove that there exists a ϵ sufficiently large such that the condition for sufficient decrease of the cost functional's value is satisfied, and $\|u^k - u^{k-1}\|^2$ decreases such that the convergence criterion is eventually satisfied. Hence, Step 4 in Algorithm 4.1 is well defined.

Concerning Step 2, we have the following.

Lemma 4.1. *The function* $K_\epsilon : \mathbb{R} \to \mathbb{R}$, $w \mapsto K_\epsilon(x,t,y,w,v,p)$ *attains a minimum for any* $(x,t,y,v,p) \in \mathbb{R}^n \times \mathbb{R}_0^+ \times \mathbb{R} \times K_U \times \mathbb{R}$ *and any* $\epsilon \in \mathbb{R}$.

Proof. As K_ϵ is bounded from below, there is a minimising sequence $(u_n)_{n\in\mathbb{N}} \subseteq K_U$ such that $\inf_{w\in K_U} K_\epsilon(x,t,y,w,v,p) = \lim_{n\to\infty} K_\epsilon(x,t,y,u_n,v,p)$ as $(K_\epsilon(x,t,y,u_n,v,p))_{n\in\mathbb{N}}$ is a monotonically decreasing sequence and thus converging [3, II Theorem 4.1]. As K_U is compact, there is a subsequence $K \subseteq \mathbb{N}$ such that $\lim_{k\to\infty} u_k = u$ with $u \in K_U$. Furthermore, we have with [3, II Theorem 5.7] and [18, Theorem 3.127]

$$\begin{aligned}
&\inf_{w\in K_U} K_\epsilon(x,t,y,w,v,p) = \lim_{k\to\infty} K_\epsilon(x,t,y,u_k,v,p) = \liminf_{k\to\infty} K_\epsilon(x,t,y,u_k,v,p) \\
&= \liminf_{k\to\infty} \left(\frac{1}{2}(y-y_d)^2 + \frac{\alpha}{2}u_k^2 + \gamma g(u_k) + pu_k + \epsilon(u_k - v)^2\right) \\
&\geq \frac{1}{2}(y-y_d)^2 + \frac{\alpha}{2}u^2 + \gamma g(u) + pu + \epsilon(u-v)^2 = K_\epsilon(x,t,y,u,v,p)
\end{aligned} \tag{14}$$

because of the lower semi-continuity of g. □

The question arises whether $\tilde{u}$, obtained in Step 2, is Lebesgue measurable. This is certainly the case if the function $(z,u) \mapsto K_\epsilon(z,y(z),u,v(z),p(z))$ is Lebesgue measurable in $z := (x,t)$ for each $u \in K_U$ and is continuous in u for each $z \in Q$. For this case, see [37, 14.29 Example, 14.37 Theorem].

If K_ϵ is only lower semi-continuous in u for each $z \in Q$, then, in general, we cannot guarantee that $\tilde{u}$ is Lebesgue measurable; see also the paragraph following [37, 14.28 Proposition]. However, in the case of $g(u) := \begin{cases} |u-d| & \text{for } |u-d| > s \\ 0 & \text{otherwise} \end{cases}$, $d \in \mathbb{R}$, $s > 0$, as considered in the section on

numerical experiments, we can prove that starting our SQH scheme with an initial guess u^0 that is Lebesgue measurable, we obtain iterates u^k that are Lebesgue measurable, see Section 5 for details. For the remaining part of this section, we assume that $\tilde{u}$, which is generated in Step 2 of Algorithm 4.1, is measurable.

In order to prove that, by increasing ϵ in Step 4 of Algorithm 4.1 ($\epsilon \leftarrow \sigma\epsilon$), a ϵ is obtained such that the condition for sufficient decrease is satisfied, we present the following lemma. A similar result is proved in [10].

Lemma 4.2. *Let $(\tilde{y}, \tilde{u})$ and $\left(y^k, u^k\right)$ be generated by Algorithm 4.1, $k \in \mathbb{N}_0$, and $\tilde{u}$, u^k be measurable; denote $\delta u := \tilde{u} - u^k$. Then, there is a $\theta > 0$ independent of ϵ such that for $\epsilon > 0$ currently chosen by Algorithm 4.1, the following holds*

$$J\left(\tilde{y}, \tilde{u}\right) - J\left(y^k, u^k\right) \leq -\left(\epsilon - \theta\right) \left\|\delta u\right\|_{L^2(Q)}^2. \tag{15}$$

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

Proof. We denote $\delta y := \tilde{y} - y^k$. In this proof, we have as in Algorithm 4.1 that $K_\epsilon\left(x, t, y^k, \tilde{u}, u^k, p^k\right) \leq K_\epsilon\left(x, t, y^k, w, u^k, p^k\right)$ for all $w \in K_U$, and thus

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

for all $(x, t) \in Q$. To obtain (15), we perform the following estimates where we use the Taylor expansion of the map $y \mapsto H\left(\cdot, \cdot, y, \cdot, \cdot\right)$, see [4, Chapter VII, Theorem 5.8 and Remark 5.9] and [4, Chapter VII, Theorem 5.2]

$$\begin{aligned}
&J\left(\tilde{y}, \tilde{u}\right) - J\left(y^k, u^k\right) \\
&= \int_Q F\left(x, t, \tilde{y}, \tilde{u}\right) - F\left(x, t, y^k, u^k\right) dxdt \\
&= \int_Q F\left(x, t, \tilde{y}, \tilde{u}\right) + p^k\tilde{u} - p^k\tilde{u} - F\left(x, t, y^k, u^k\right) - p^k u^k + p^k u^k dxdt \\
&= \int_Q H\left(x, t, \tilde{y}, \tilde{u}, p^k\right) - p^k\tilde{u} - H\left(x, t, y^k, u^k, p^k\right) + p^k u^k dxdt \\
&= \int_Q H\left(x, t, y^k + \delta y, \tilde{u}, p^k\right) - H\left(x, t, y^k, u^k, p^k\right) + p^k\left(u^k - \tilde{u}\right) dxdt \\
&= \int_Q H\left(x, t, y^k, \tilde{u}, p^k\right) + \left(y^k - y_d\right)\, \delta y + \frac{1}{2}\left(\delta y\right)^2 - H\left(x, t, y^k, u^k, p^k\right) dxdt \\
&+ \int_Q p^k\left(u^k - \tilde{u}\right) dxdt \\
&= \int_Q K_\epsilon\left(x, t, y^k, \tilde{u}, u^k, p^k\right) - \epsilon\left(\delta u\right)^2 - H\left(x, t, y^k, u^k, p^k\right) + \frac{1}{2}\left(\delta y\right)^2 dxdt \\
&+ \int_0^T -\left(\left(p^k\right)', \delta y\right) + D\left(\nabla p^k, \nabla\delta y\right) - \left(\left(\delta y\right)', p^k\right) - D\left(\nabla\delta y, \nabla p^k\right) dt \\
&\leq \int_Q -\epsilon\left(\delta u\right)^2 + \frac{1}{2}\left(\delta y\right)^2 dxdt.
\end{aligned}$$

Notice that in the first before last step, we use integration by parts [44, Theorem 3.11] using the fact that $\delta y(\cdot,0)=0$ and $p^k(\cdot,T)=0$, because of the initial condition for the state and the terminal condition for the adjoint, respectively. We have that $\|\delta y(\cdot,t)\|^2_{L^2(\Omega)} \leq c(D)\|\delta u(\cdot,t)\|^2_{L^2(\Omega)}$ for all $t\in[0,T]$, $c(D)>0$. This can be seen as follows; similar to [28, (6.3)]. Consider the difference between (1) with $u \leftarrow \tilde{u}$ and $y \leftarrow \tilde{y}$ and the same equation (1) but with $u \leftarrow u^k$ and $y \leftarrow y^k$. 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

$$\int_0^T \frac{1}{2}\frac{d}{dt}\|\delta y(\cdot,t)\|^2_{L^2(\Omega)} + D\|\nabla \delta y(\cdot,t)\|^2_{L^2(\Omega)}dt \leq \int_0^T \|\delta u(\cdot,t)\|^2_{L^2(\Omega)}\|\delta y(\cdot,t)\|^2_{L^2(\Omega)}dt,$$

according to [19, page 287, Theorem 3] and the Cauchy-Schwarz inequality, see [2, (2.2)]. Next, we have

$$\frac{1}{2}\left(\|\delta y(\cdot,T)\|^2_{L^2(\Omega)} - \|\delta y(\cdot,0)\|^2_{L^2(\Omega)}\right) + D\|\nabla \delta y\|^2_{L^2(Q)} \leq \hat{c}\int_0^T \|\delta u(\cdot,t)\|^2_{L^2(\Omega)}\|\nabla \delta y(\cdot,t)\|^2_{L^2(\Omega)}dt,$$

for some $\hat{c}>0$. Thus, as $\|\delta y(\cdot,0)\|^2_{L^2(\Omega)}=0$, we obtain

$$\|\nabla \delta y\|_{L^2(Q)} \leq \tilde{c}(D)\|\delta u\|_{L^2(Q)},$$

for some $\tilde{c}(D)\geq 0$. Furthermore, by the Poincaré inequality [2, (6.7)], we have for $\tilde{c}>0$

$$\|\delta y\|_{L^2(Q)} = \sqrt{\int_0^T \|\delta y\|^2_{L^2(\Omega)}dt} \leq \tilde{c}\sqrt{\int_0^T \|\nabla \delta y\|^2_{L^2(\Omega)}dt} = \tilde{c}\|\nabla \delta y\|_{L^2(Q)} \leq c(D)\|\delta u\|_{L^2(Q)}.$$

Thus, we have

$$\begin{aligned}\|\delta y\|^2_{L^2(Q)} &= \int_0^T \int_\Omega (\delta y)^2\, dxdt = \int_0^T \|\delta y(\cdot,t)\|^2_{L^2(\Omega)}dt \leq c(D)\int_0^T \|\delta u(\cdot,t)\|^2_{L^2(\Omega)}dt \\ &= c(D)\|\delta u\|^2_{L^2(Q)}.\end{aligned}$$

We conclude as follows

$$\int_Q -\epsilon(\delta u)^2 + \frac{1}{2}(\delta y)^2\, dxdt = -\epsilon\|\delta u\|^2_{L^2(Q)} + \frac{1}{2}\|\delta y\|^2_{L^2(Q)} \leq \left(-\epsilon + \frac{1}{2}c(D)\right)\|\delta u\|^2_{L^2(Q)},$$

which proves the claim with $\theta := \frac{c(D)}{2}$. □

Next, we prove a lemma stating that Algorithm 4.1 stops when u^k is a solution to (12).

Lemma 4.3. *Let y^k and u^k be generated by Algorithm 4.1, $k\in\mathbb{N}_0$, and u^k be measurable. If the iterate u^k is optimal, then Algorithm 4.1 stops, returning u^k.*

Proof. If u^k, $k \in \mathbb{N}_0$ is optimal, then we have that $H\left(x,t,y^k,u^k,p^k\right) = \min_{w\in K_U} H\left(x,t,y^k,w,p^k\right)$ for almost all $(x,t) \in Q$ and thus

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

for all $w \in K_U$ and for almost all $(x,t) \in Q$. 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 $\left(\tilde{x},\tilde{t}\right) \in Q$ where u^k is optimal and a $\tilde{u}$ with $\left(\tilde{u}\left(x,t\right)-u^k\left(x,t\right)\right)^2 > 0$ such that $K_\epsilon\left(\tilde{x},\tilde{t},y^k,\tilde{u},u^k,p^k\right) \leq K_\epsilon\left(\tilde{x},\tilde{t},y^k,u^k,u^k,p^k\right)$ in order to ensure that Algorithm 4.1 stays in its determined optimal solution u^k.

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

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

and the optimality of u^k that

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

and consequently $\epsilon\left(\tilde{u}\left(\tilde{x},\tilde{t}\right)-u^k\left(\tilde{x},\tilde{t}\right)\right)^2 \leq 0$. Algorithm 4.1 has updated the initial guess u^0 at most k times where ϵ is diminished by $\epsilon \leftarrow \zeta\epsilon$. Thus, we have that $\epsilon > 0$ and therefore $\left(\tilde{u}\left(\tilde{x},\tilde{t}\right)-u^k\left(\tilde{x},\tilde{t}\right)\right)^2 \leq 0$, which means that $\tilde{u} = u^k$ almost everywhere as the calculation holds for any $\left(\tilde{x},\tilde{t}\right) \in Q$ where u^k is optimal. Thus $\delta u = 0$ in the $L^2\left(Q\right)$ sense and Algorithm 4.1 stops and returns u^k. □

The following theorem states that the iteration over the Steps 2 to 4 in Algorithm 4.1 (no stopping criterion) generate 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 $\lim_{k\to\infty}\|u^k-u^{k-1}\|_{L^2(Q)} = 0$. A similar result is proved in [10]. In the view of Lemma 4.3, we assume for the rest of this section that no element of the sequence $\left(u^k\right)_{k\in\mathbb{N}}$ is optimal. Furthermore, we assume that each u^k, $k \in \mathbb{N}_0$, is measurable.

Theorem 4.1. *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 4.1 (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

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

Proof. Due to Lemma 4.2, we have that Algorithm 4.1 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 $-\left(\epsilon-\theta\right)\|u^{k+1}-u^k\|^2_{L^2(Q)}$.

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

$$J(\tilde{y},\tilde{u}) - J\left(y^k,u^k\right) \leq -(\epsilon-\theta)\|\tilde{u}-u^k\|^2_{L^2(Q)} \leq -\eta\|\tilde{u}-u^k\|^2_{L^2(Q)}. \tag{16}$$

Therefore there is an update after at least finitely many increases of ϵ in Step 4 of Algorithm 4.1 and we have that $u^{k+1} \leftarrow \tilde{u}$ with corresponding $\tilde{u}$. 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 k such that

$$-\rho\eta \leq J\left(y^{k+1},u^{k+1}\right) - J\left(y^k,u^k\right) \leq 0, \tag{17}$$

because any sequence bounded from below converges, see [3, Chapter II, Theorem 4.1, Theorem 6.1] for details.

Finally, as (16) also holds for u^{k+1} instead of $\tilde{u}$, we obtain from (16) and (17) the following

$$\rho\eta \geq -\left(J\left(y^{k+1},u^{k+1}\right) - J\left(y^k,u^k\right)\right) \geq \eta\|u^{k+1}-u^k\|^2_{L^2(Q)} \geq 0,$$

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

We remark that the result of Theorem 4.1 means that there exists a $\bar{k} \in \mathbb{N}$ such that the Algorithm 4.1 stops at the $\bar{k}$-th iteration where $\|u^{\bar{k}+1}-u^{\bar{k}}\|^2_{L^2(Q)} < \kappa$ as $\kappa > 0$.

Notice that, if $\tilde{u}$ determined in Algorithm 4.1 Step 2 is measurable, then due to the pointwise bounds, we have $\tilde{u} \in U_{ad}$ and thus especially $\left(u^k\right)_{k\in\mathbb{N}} \subseteq U_{ad}$.

In general, if the cost functional is discontinuous, we cannot prove that $\bar{u}$ returned by the SQH method represents the optimal control sought. On the other hand, assuming a continuously differentiable g, then we can prove that $\bar{u}$ satisfies the optimality condition $\int_Q \nabla J(\bar{u})(x,t)(w(x,t)-\bar{u}(x,t))\,dxdt \geq 0$ for all $w \in U_{ad}$; see [44, Lemma 2.21], where $J(\bar{u}) := J(y(\bar{u}),\bar{u})$.

In order to prove this fact, let us introduce the Euclidean projection $P_{K_U} : \mathbb{R} \to K_U$, see [8, Proposition 2.1.3 (Projection Theorem)], and the reduced gradient $\nabla J(u) := \alpha u + \gamma\frac{\partial}{\partial u}g(u) + p$, see [44].

Now, with an analogous calculation as in [10, Theorem 3.2], we can prove the following theorem.

Theorem 4.2. *Assume that $g : \mathbb{R} \to \mathbb{R}$ in (5) is continuously differentiable and there is a lower bound $\epsilon_0 > 0$ for ϵ. Then for each accumulation point $\bar{u}$ of the sequence $(u^n)_{n\in\mathbb{N}_0}$ generated as in Algorithm 4.1 (loop over Step 2 to Step 4) with $\lim_{\tilde{k}\to\infty}\|u^{\tilde{k}}-\bar{u}\|_{L^q(Q)} = 0$, $\tilde{k} \in \tilde{K} \subseteq \mathbb{N}$, there is a subsequence $\left(u^k\right)_{k\in K}$, $K \subseteq \tilde{K}$, such that*

$$\lim_{k\to\infty}\|u^k - P_{K_U}\left(u^k - \frac{1}{2\epsilon}\nabla J\left(u^k\right)\right)\|_{L^2(Q)} = 0,$$

where $\bar{u}$ fulfils the following optimality condition

$$\nabla J(\bar{u})(x,t)(w(x,t)-\bar{u}(x,t)) \geq 0,$$

for all $w \in U_{ad}$ and almost all $(x,t) \in Q$.

Proof. We remark that $\epsilon > 0$ for each iterate $u^{\tilde{k}}$, $\tilde{k} \in \tilde{K}$. As $u^{\tilde{k}+1}$ minimises $w \mapsto K_\epsilon\left(x,t,y^{\tilde{k}},w,u^{\tilde{k}},p^{\tilde{k}}\right)$ for all $(x,t) \in Q$ with $u^{\tilde{k}+1} \in K_U$, we have that

$$\begin{aligned}&\frac{\partial}{\partial u^{\tilde{k}+1}} K_\epsilon\left(x,t,y^{\tilde{k}},u^{\tilde{k}+1},u^{\tilde{k}},p^{\tilde{k}}\right)\left(w-u^{\tilde{k}+1}\right)\\&=\left(2\epsilon\left(u^{\tilde{k}+1}-u^{\tilde{k}}\right)+\frac{\partial}{\partial u^{\tilde{k}+1}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}+1},p^{k}\right)\right)\left(w-u^{\tilde{k}+1}\right)\geq 0,\end{aligned}$$

for all $w \in K_U$ and for all $(x,t) \in Q$, see [44, Lemma 2.21]. Equivalently, we can write

$$u^{\tilde{k}+1} = P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\frac{\partial}{\partial u^{\tilde{k}+1}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}+1},p^{\tilde{k}}\right)\right); \tag{18}$$

see [8, Proposition 2.1.3 (Projection Theorem)]. Additionally, we have

$$\nabla J\left(u^{\tilde{k}}\right) = \frac{\partial}{\partial u^{\tilde{k}}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}},p^{\tilde{k}}\right),$$

compare with (8). Starting from (18) and adding and subtracting equal terms, we have

$$\begin{aligned}u^{\tilde{k}} - P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\nabla J\left(u^{\tilde{k}}\right)\right) &= u^{\tilde{k}} - u^{\tilde{k}+1} + P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\frac{\partial}{\partial u^{\tilde{k}+1}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}+1},p^{\tilde{k}}\right)\right)\\&\quad - P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\frac{\partial}{\partial u^{\tilde{k}}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}},p^{\tilde{k}}\right)\right).\end{aligned}$$

Thus, using the triangle inequality and [8, Proposition 2.1.3 (Projection Theorem)] and $\epsilon > \epsilon_0$, we obtain

$$\begin{aligned}&\|u^{\tilde{k}} - P_{K_U}\left(u^{\tilde{k}} - \frac{1}{2\epsilon}\nabla J\left(u^{\tilde{k}}\right)\right)\|_{L^2(Q)}\\&\leq \|u^{\tilde{k}} - u^{\tilde{k}+1}\|_{L^2(Q)}\\&+ \frac{1}{2\epsilon_0}\|\frac{\partial}{\partial u^{\tilde{k}+1}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}+1},p^{\tilde{k}}\right) - \frac{\partial}{\partial u^{\tilde{k}}} H\left(x,t,y^{\tilde{k}},u^{\tilde{k}},p^{\tilde{k}}\right)\|_{L^2(Q)}\\&\leq \|u^{\tilde{k}} - u^{\tilde{k}+1}\|_{L^2(Q)} + \frac{1}{2\epsilon_0}\left(\alpha\|u^{\tilde{k}+1} - u^{\tilde{k}}\|_{L^2(Q)} + \gamma\|\frac{\partial}{\partial u^{\tilde{k}+1}} g\left(u^{\tilde{k}+1}\right) - \frac{\partial}{\partial u^{\tilde{k}}} g\left(u^{\tilde{k}}\right)\|_{L^2(Q)}\right).\end{aligned} \tag{19}$$

Now, we have the following estimates $\|y^{\tilde{k}} - \bar{y}\|_{L^2(Q)} \leq c\|u^{\tilde{k}} - \bar{u}\|_{L^2(Q)}$, and analogously $\|p^{\tilde{k}} - \bar{p}\|_{L^2(Q)} \leq c\|u^{\tilde{k}} - \bar{u}\|_{L^2(Q)}$, $c > 0$ where $\bar{y}$ is the solution to (1) for $u \leftarrow \bar{u}$ and $\bar{p}$ is the solution to (7) for $y \leftarrow \bar{y}$; see the proof of Lemma 4.2. For each accumulation point $\bar{u}$, there exists a subsequence within the sequence generated as in Algorithm 4.1 that strongly converges to $\bar{u}$ in $L^q(Q)$ according to our assumption. Using $\|\cdot\|_{L^2(Q)} \leq \|\cdot\|_{L^q(Q)}$; see [1, Theorem 2.14], and [7, Proposition 3.6, Remark 3.7], we obtain a subsequence, $\left(u^k\right)_{k\in K}$, $K \subseteq \mathbb{N}$, with the following pointwise convergence $\lim_{k\to\infty} u^k(x,t) = \bar{u}(x,t)$, $\lim_{k\to\infty} y^k(x,t) = \bar{y}(x,t)$ and $\lim_{k\to\infty} p^k(x,t) = \bar{p}(x,t)$ for almost all $(x,t) \in Q$. Consequently, we have

$$\lim_{k\to\infty} \nabla J\left(u^k\right) = \lim_{k\to\infty} \alpha u^k + \gamma\frac{\partial}{\partial u} g(u)\,|_{u=u^k} + p^k = \alpha\bar{u} + \gamma\frac{\partial}{\partial u} g(u)\,|_{u=\bar{u}} + \bar{p} = \nabla J(\bar{u}), \tag{20}$$

for almost every $(x,t) \in Q$. If we take the limit on both sides of (19), considering the pointwise converging subsequence, we obtain

$$\lim_{k\to\infty} \|u^k - P_{K_U}\left(u^k - \frac{1}{2\epsilon}\nabla J\left(u^k\right)\right)\|_{L^2(Q)} = 0, \tag{21}$$

where we use Theorem 4.1 for the first and second term and the dominated convergence theorem [7, Proposition 2.17], [7, 2.2 Measurable and Borel functions] and [3, III.3 Theorem 3.6] with the bounded image of u^k, $k \in K$ for the last term.

Next, we prove that $\nabla J\left(\bar{u}\right)(x,t)\left(w\left(x,t\right) - \bar{u}\left(x,t\right)\right) \geq 0$ for all $w \in U_{ad}$ for almost all $(x,t) \in Q$. 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 $(x,t) \in Q$. 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 $(x,t) \in Q$, see [8, 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

$$2\epsilon\left(v^k - u^k\right)\left(w - v^k\right) + \nabla J\left(u^k\right)\left(w - u^k\right) + \nabla J\left(u^k\right)\left(u^k - v^k\right) \geq 0. \tag{22}$$

Due to $|w - v^k| \leq 2\max\left(|u_a|, |u_b|\right)$ and the upper bound $\sigma\left(\eta + \theta\right)$ for ϵ because of (16) and Step 4 of Algorithm 4.1, and the fact that converging sequences are bounded [3, II, Theorem 1.10] combined with (21) and (20), we obtain the following by taking the limit in (22). We have

$$\nabla J\left(\bar{u}\right)\left(w - \bar{u}\right) \geq 0,$$

for all $w \in U_{ad}$ for almost all $(x,t) \in Q$; see [3, II Theorem 2.4] and [3, II Theorem 2.7]. □

This result also proves that $\int_Q \nabla J\left(\bar{u}\right)(x,t)\left(w\left(x,t\right) - \bar{u}\left(x,t\right)\right)dxdt \geq 0$ for all $w \in U_{ad}$, see [5, Chapter X Corollary 2.16]. Notice that, if g is strictly convex, then $\bar{u}$ is the optimal control sought.

We remark that the analysis above is performed at a functional level and independently of the discretisation used. However, for the numerical realisation of our optimisation scheme, we consider the following finite differences setting [25], where we assume that the control is approximated by a piecewise constant function.

We take a space-time cylinder $Q = \Omega \times (0,T)$ with $\Omega = (a,b)^n$, and define the following space-time grid

$$Q_{h,\triangle t} := \left\{\left(x_{i_1\ldots i_n}, t_m\right), \mid x_{i_1\ldots i_n} \in \Omega_h,\, t_m = m\,\triangle t,\, m \in \{1,\ldots,N_t\}\right\},$$

where

$$\Omega_h = \{(a + i_1 h, \ldots, a + i_n h) \in \mathbb{R}^n,\, i_j \in \{1, ..., N-1\},\, j \in 1, ..., n\}.$$

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 $(x_{i_1 \ldots i_n}, t_m)$ and $t_m = m \triangle t$ are ordered lexicographically.

In order to compute the state and adjoint variables, we approximate (1) and (7) using the implicit Euler scheme and finite differences. For the computation of the integrals appearing in J and for the integration of H (see below), we use the rectangle rule; see, e.g., [41].

5 Numerical experiments

In this section, we present results of numerical experiments to validate our optimal control formulation and the convergence performance of the SQH method.

For the lower semi-continuous function g in (3), we choose the following

$$g_{d,s}(u) := \begin{cases} |u - d| & \text{for } |u - d| > s \\ 0 & \text{otherwise}. \end{cases} \tag{23}$$

Notice that $G_{d,s}(u) := \int_Q g_{d,s}(u(x,t))\, dxdt$ measures zero costs as far as the control u is in the L^1 closed ball centered in $d \in \mathbb{R}$ with radius $s > 0$. If u is in the complement of this ball, then the cost given by $G_{d,s}$ is of L^1 type. It can be shown with an elementary calculation that $G_{d,s}$ is a discontinuous functional (e.g., consider the case of constant controls).

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, $\kappa = 10^{-6}$, $\zeta = \frac{3}{20}$, $\sigma = 50$ 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

$$y_d(x,t) = \begin{cases} 5 & \text{if } \bar{x}(t) - c \le x \le \bar{x}(t) + c \\ 0 & \text{else}, \end{cases} \tag{24}$$

where $\bar{x}(t) := x_0 + \frac{2}{5}(b-a)\sin\left(2\pi \frac{t}{T}\right)$, $x_0 = \frac{b+a}{2}$, and $c = \frac{7}{100}(b-a)$. We choose the cost functional J as in (3) - (4) with the desired trajectory (24) and set $d = 0$, $s = 1$.

The augmented Hamiltonian $K_\epsilon(x, t, y, u, v, p)$ is minimised as follows. As y and p are held fixed, we have

$$\tilde{K}_\epsilon(x, t, y, u, v, p) := \frac{\alpha}{2} u^2 + g_{0,1}(u) + pu + \epsilon(u - v)^2.$$

Its minimum can be exactly given by a case study as follows.

If $0 \le u \le s$, we have $\tilde{K}_\epsilon(x, t, y, u, v, p) = \frac{\alpha}{2} u^2 + pu + \epsilon(u - v)^2$ with its minimum at

$$u_1 := \min\left(\max\left(0, \frac{2\epsilon v - p}{2\epsilon + \alpha}\right), s\right).$$

If $s < u \le 10$, we have $\tilde{K}_\epsilon(x, t, y, u, v, p) := \frac{\alpha}{2} u^2 + \gamma u + pu + \epsilon(u - v)^2$ 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 $\tilde{K}_\epsilon$ over K_U is given by

$$u = \underset{w \in K_U}{\operatorname{argmin}} \tilde{K}_\epsilon(x,t,y,w,v,p) = \underset{w \in \{u_1,u_2\}}{\operatorname{argmin}} \tilde{K}_\epsilon(x,t,y,w,v,p).$$

Next, we remark that u, as a function, is Lebesgue measurable assuming that the last iterate v is also Lebesgue measurable. To illustrate this fact, denote with $z := (x,t)$, and notice that p is Lebesgue measurable since it is the solution to (10). Thus, we have that u_1 and u_2 are Lebesgue measurable functions; see [15, Proposition 2.1.4, Proposition 2.1.7]. Further, we have that $\tilde{K}_\epsilon(z,u_1(z)) := \tilde{K}_\epsilon(z,y(z),u_1(z),v(z),p(z))$ and
$\tilde{K}_\epsilon(z,u_1(z)) := \tilde{K}_\epsilon(z,y(z),u_1(z),v(z),p(z))$ are Lebesgue measurable according to Lemma 6.3 in the Appendix and because the sum and the product of Lebesgue measurable functions is Lebesgue measurable; see [15, Proposition 2.1.7].

Now, the function $u(z)$ is given by

$$u(z) := \begin{cases} u_1(z) & \text{if } \tilde{K}_\epsilon(z,u_1(z)) \leq \tilde{K}_\epsilon(z,u_2(z)) \\ u_2(z) & \text{if } \tilde{K}_\epsilon(z,u_1(z)) > \tilde{K}_\epsilon(z,u_2(z)) \end{cases}.$$

According to [15, Proposition 2.1.1] and the following paragraph, u is Lebesgue measurable if and only if the set $\{z \in Q |\ u(z) > c\}$ is Lebesgue measurable for any $c \in \mathbb{R}$. To show this fact, notice that the following holds

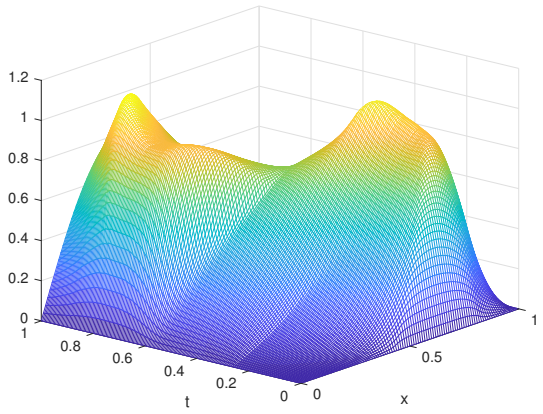
$$\begin{aligned} \{z \in Q |\ u(z) > c\} &= \left(\{z \in Q |\ u_1(z) > c\} \cap \left\{z \in Q |\ \tilde{K}_\epsilon(z,u_1(z)) \leq \tilde{K}_\epsilon(z,u_2(z))\right\}\right) \\ &\cup \left(\{z \in Q |\ u_2(z) > c\} \cap \left\{z \in Q |\ \tilde{K}_\epsilon(z,u_1(z)) > \tilde{K}_\epsilon(z,u_2(z))\right\}\right). \end{aligned} \tag{25}$$

Thus u is Lebesgue measurable, as the intersection and union of finite Lebesgue measurable sets is Lebesgue measurable, see [5, IX Theorem 5.1, Remark 1.1], if and only if the single sets are measurable. Now, we have that the sets $\{z \in Q |\ u_1(z) > c\}$ and $\{z \in Q |\ u_2(z) > c\}$ are Lebesgue measurable for any $c \in \mathbb{R}$ as u_1 and u_2 are Lebesgue measurable; further the set $\left\{z \in Q |\ \tilde{K}_\epsilon(z,u_1(z)) \leq \tilde{K}_\epsilon(z,u_2(z))\right\}$ is Lebesgue measurable, see [15, Proposition 2.1.3], and

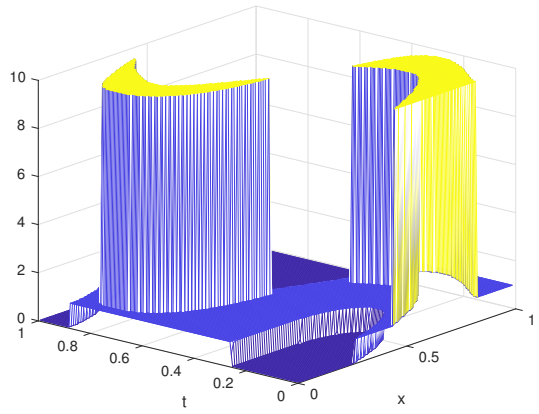
$$\left\{z \in Q |\ \tilde{K}_\epsilon(z,u_1(z)) > \tilde{K}_\epsilon(z,u_2(z))\right\} = Q \backslash \left\{z \in Q |\ \tilde{K}_\epsilon(z,u_1(z)) \leq \tilde{K}_\epsilon(z,u_2(z))\right\}$$

is Lebesgue measurable according to [5, IX Remark 1.1].

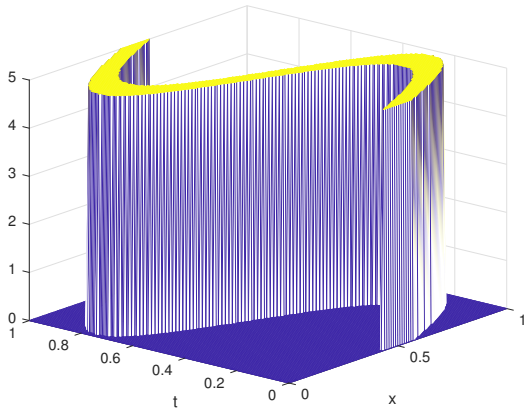
Next, having completed the theoretical discussion, we perform the first set of experiments using Algorithm 4.1 to solve our optimal control problem. The SQH algorithm converges in 29 iterations and we obtain the state and control functions depicted in Figure 1. The plot of the control function shows clearly the action of the discontinuous cost of the control given by $g_{0,1}$ 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 1: Optimal solution for the first experiment 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, in Table 1 we report the values of

$$\Delta H = \max_{(x,t)\in Q_{h,\triangle t}} \left(H\left(x,t,\bar{y},\bar{u},\bar{p}\right) - \min_{w\in K_U} H\left(x,t,\bar{y},w,\bar{p}\right) \right).$$

The value of ΔH gives a measure of optimality of the SQH solution $(\bar{y},\bar{u},\bar{p})$ and the results reported in Table 1 demonstrate how ΔH decreases as we refine the mesh size and the value of κ, thus demonstrating an improvement in accuracy of the PMP solution by refinement.

In Table 2, we report results that aim at showing the ratio of numbers of grid points $(x,t) \in Q_{h,\triangle t}$ where the optimality condition is satisfied to machine precision. For this purpose, in Table 2, we give the ratio of grid points where the following holds

$$H\left(x,t,\bar{y},\bar{u},\bar{p}\right) - \min_{w\in K_U} H\left(x,t,\bar{y},w,\bar{p}\right) \approx \text{eps},$$

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 $\kappa = 10^{-6}$.

$N_t \times N$ \ κ	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 1: Values of ΔH of the SQH solution with different choices of the value of κ.

$N_t \times N$ \ κ	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 2: Ratio of grid points at which the Pontryagin maximum principle is fulfilled to machine precision to the total number of grid points.

In the third experiment, we investigate the computational performance of Algorithm 4.1 with respect to different choices of the optimisation parameters. In Table 3, 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: Computational performance of Algorithm 4.1 with respect to different choices of values of the optimisation parameters.

The fourth numerical experiment deals with the complexity of Algorithm 4.1. Let $N_{gp} \in \mathbb{N}$ denote the total number of space-time grid points. We solve the same optimisation problem as in

Figure 1 using different meshes. The resulting CPU times are reported in Figure 2 and detailed in Table 4. In Figure 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.

$\frac{N}{100} \times \frac{N_t}{100}$	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 4: Data points for Figure 2.

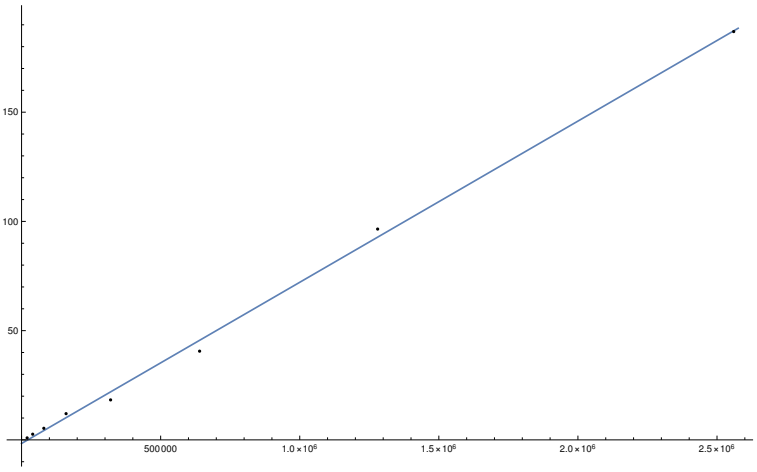


Figure 2: Computational complexity of Algorithm 4.1. The data points (dots) are fitted by a linear model.

Now in the fifth experiment, we use the same setting as 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 well-known projected Hager-Zhang-NCG (pNCG) method with Wolfe-Powell step-size strategy [11]. Additionally, we perform the comparison with a projected gradient method with Armijo step-size strategy (pGM). The minimum of the augmented Hamiltonian $K_\epsilon(x, t, y, u, v, p)$ 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 of the control function u between two 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 optimisation strategy. In Table 5, we see that the pNCG method in most cases outperforms our SQH method. On the other hand, one can see in Table 6 that the SQH method performs better than the pGM scheme.

For the case of $\alpha = 10^{-1}$, we take $\sigma = 2.1$ and $\zeta = 0.9$ in Algorithm 4.1 instead of $\sigma = 50$ and $\zeta = \frac{3}{20}$. We remark that the convergence performance of Algorithm 4.1 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 5: 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 6: Comparison of the SQH scheme with the pGM method.

For further illustration of our optimisation 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. In this experiment, the desired state is given by

$$y_d(x,t) = 5\sin\left(2\pi\frac{t}{T}\right);$$

see Figure 3.

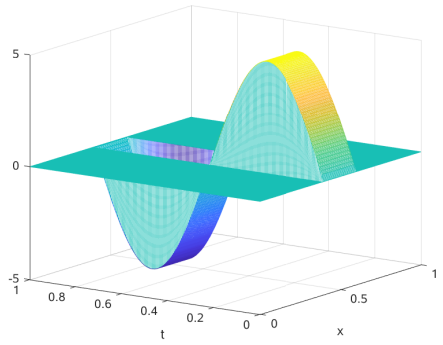
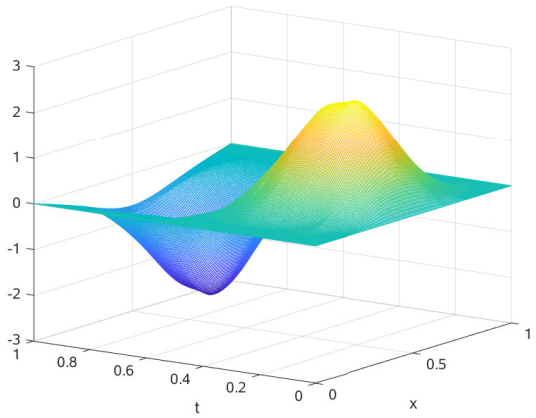


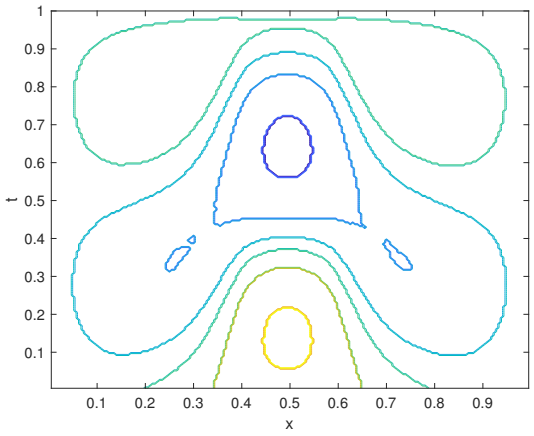
Figure 3: Desired function $y_d = 5\sin\left(2\pi\frac{t}{T}\right)$.

Further, we take $\alpha = 5 \cdot 10^{-3}$, $\gamma = 1 \cdot 10^{-3}$, $N = 200$ and $N_t = 200$. The parameters of Algorithm 4.1 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 4, where we clearly see how the admissible control values are taken by the control function.

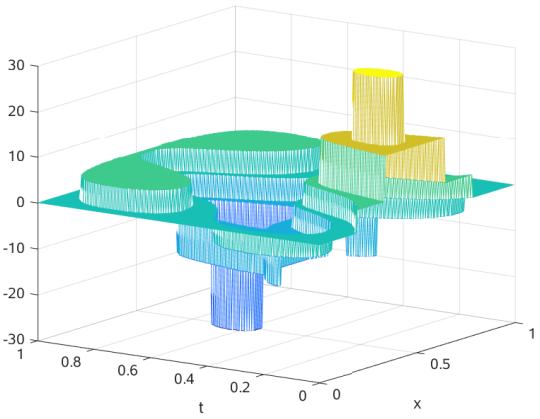
An analogous numerical test of optimality, as the one related to Table 2, provides the following result. We have that the inequality $0 \leq H\left(x,t,\bar{y},\bar{u},\bar{p}\right) - \min_{w\in K_U} H\left(x,t,\bar{y},w,\bar{p}\right) \leq 10^{-l}$ is fulfilled at 100% of the grid points for $l = 2$ and at 99.19% of the grid points for $l = 12$ with the returned values $(\bar{y},\bar{u},\bar{p})$ of the SQH method method, where the minimum of H over K_U is determined with a secant method. 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.



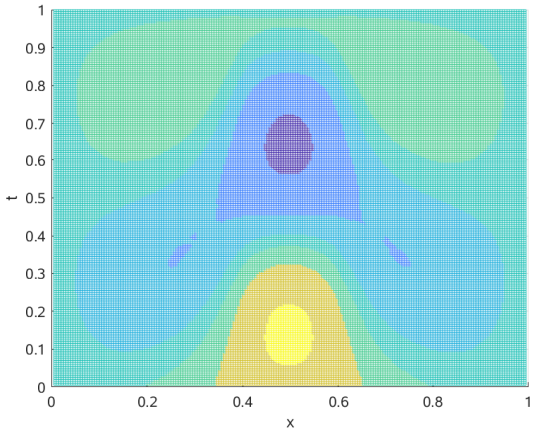
(a) The state y



(b) The control function u as contour plot



(c) The control function u



(d) The control function u viewed from above

Figure 4: Results with Algorithm 4.1 for the cost functional (3) with $g(\cdot) := |\cdot|^{\frac{1}{2}}$ and $K_U = \{-30, -15, -5, 0, 5, 15, 30\}$.

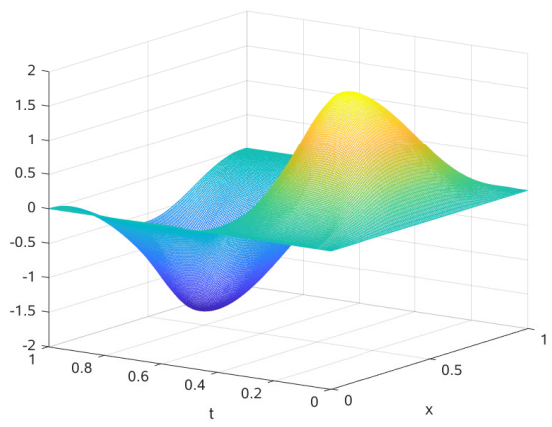
To conclude our series of experiments, we choose the following lower semi-continuous step function

$$g(z) := g_2(z) = \begin{cases} \frac{7}{2} & \text{for } |z| > 6 \\ 1 & \text{for } 3 < |z| \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

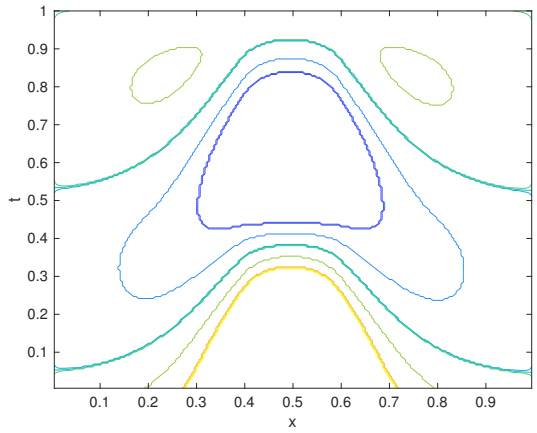
and $K_U = [-10, 10]$. In this case, while the control function may take a continuous set of values, the cost of the control is piecewise constant. The augmented Hamiltonian is minimised by a secant method as mentioned in Section 4. The problem's parameters are set $N = 200$ and $N_t = 200$, $\alpha = 0$, $\beta = 10^{-1}$, $\sigma = 50$, $\zeta = \frac{3}{20}$, $\eta = 10^{-9}$, $\kappa = 10^{-6}$, $u^0 = 0$ and the initial guess for $\epsilon = \frac{3}{5}$. The results for this case are depicted in Figure 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 2, provides the following result. We have that the

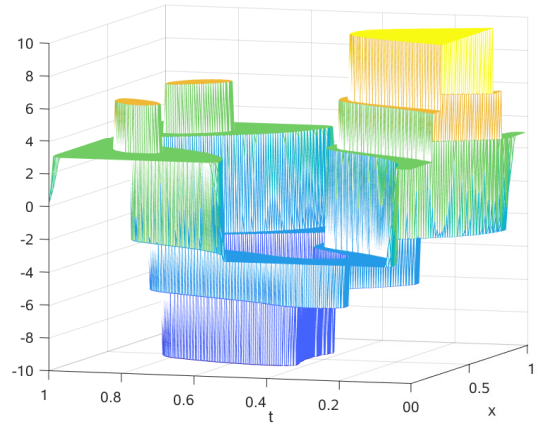
inequality $0 \leq H(x,t,\bar{y},\bar{u},\bar{p}) - \min_{w\in K_U} H(x,t,\bar{y},w,\bar{p}) \leq 10^{-l}$ is fulfilled at 100% of the grid points for $l = 2$ and at 99.53% of the grid points for $l = 12$ with the returned values $(\bar{y},\bar{u},\bar{p})$ of the SQH method, where the minimum of H over K_U is determined with a secant method.



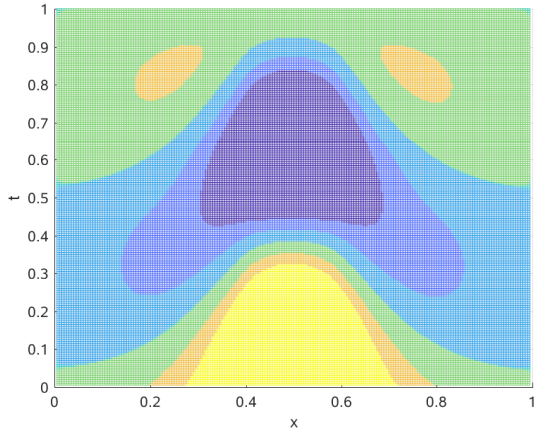
(a) The state y



(b) The control function u as a contour plot



(c) The control function u



(d) The control function u viewed from above

Figure 5: Results with Algorithm 4.1 for the cost functional (3) with $g = g_2$ and $K_U = [-10, 10]$.

6 Conclusion

This paper was devoted to the investigation of a sequential quadratic Hamiltonian (SQH) scheme for solving parabolic optimal control problems with discontinuous and non-convex cost functionals. The formulation of this scheme was inspired by the earlier works [26, 27] and [38, 40] that were proposed for solving smooth ODE control problems. However, while these methods cannot be applied in a PDE context because of a lack of robustness or prohibitive computational costs, it was shown that the SQH method is robust and has a computational performance that is typical of pointwise iterative schemes.

At the core of the SQH formulation was the characterisation of optimal controls by means

of the Pontryagin's maximum principle. Within this framework and in a general setting that included discontinuous and non-convex cost functionals, it was proved that the SQH method is well-defined. However, convergence to an optimal solution was proved only in the smooth case.

The efficiency and robustness of the proposed SQH scheme was successfully demonstrated by results of numerical experiments and the unmatched large applicability of the SQH method was illustrated considering different settings.

These encouraging results suggest further development and improvement of the SQH scheme. On the one hand, the investigation of this scheme to solve PDE control problems with state constraints and nonlinear control mechanisms. On the other hand, the acceleration of the SQH method by a multigrid strategy in order to obtain fast solvers for discontinuous and non-convex PDE control problems.

Acknowledgements

We are very grateful to many colleagues and the anonymous Referees who have supported our work through discussions, references, and all that. In particular, we thank F. Bonnans, E. Casas, C. Clason, A. Dmitruk, F. Petitta, M.I. Sumin, and F. Tröltzsch.

We especially thank Andrei Fursikov for continued support of our work.

Appendix

6.1 A L^∞ estimate

For our governing PDE model, we prove a L^∞ result that is essential in the Pontryagin maximum principle framework. However, we prove this estimate in a more general model setting as follows

$$\begin{aligned} \left(y'\left(\cdot,t\right),v\right)+B\left(y,v;t\right)&=\left(h\left(\cdot,t\right),v\right) \text{ in } \Omega\times\left(0,T\right)\\ y&=0 \text{ on } \partial\Omega\times\left[0,T\right]\\ y&=y_0 \text{ on } \Omega\times\left\{0\right\}, \end{aligned} \tag{26}$$

with bounded $\Omega\subseteq\mathbb{R}^n$, $T>0$ and $y'\left(\cdot,t\right):=\frac{\partial}{\partial t}y\left(\cdot,t\right)$ where $B\left(y,v;t\right):H_0^1\times H_0^1\times\mathbb{R}_0^+\to\mathbb{R}$ is a bilinear map with the coercivity condition $\beta\|y\left(\cdot,t\right)\|_{H_0^1(\Omega)}^2\leq B\left(y,y;t\right)$, $\beta>0$ and $B\left(-k,v;t\right)\leq 0$ for $k\geq 0$ if $v\geq 0$ for any $t\in\left[0,T\right]$. Furthermore, we require that $h\in L^q\left(Q\right)$, $q>\frac{n}{2}+1$, $y_0\in L^\infty\left(\Omega\right)$ and that (26) has a unique solution fulfilling $y\in L^2\left(0,T;H_0^1\left(\Omega\right)\right)\cap L^\infty\left(0,T;L^2\left(\Omega\right)\right)$ and $y'\in L^2\left(0,T;H^{-1}\left(\Omega\right)\right)$, such that (26) holds for almost all $t\in\left[0,T\right]$ and all $v\in H_0^1\left(\Omega\right)$, see [19, Chapter 7] for details. With the following lemma, we prepare for the proof of Theorem 6.1 below. This result and a similar proof can be found in [34] or [28, Chapter 7 Theorem 7.1, Corollary 7.1]. For the notation see [1].

Lemma 6.1. *Let $y\in L^q\left(0,T;W_0^{1,q}\left(\Omega\right)\right)\cap L^\infty\left(0,T;L^\rho\left(\Omega\right)\right)$, with $q\geq 1$, $\rho\geq 1$. Then $y\in L^\sigma\left(Q\right)$ with $\sigma=q\frac{n+\rho}{n}$ and there exists a constant $c>0$ with*

$$\int_Q|y\left(x,t\right)|^\sigma dxdt\leq c\|y\|_{L^\infty(0,T;L^\rho(\Omega))}^{\frac{\rho q}{n}}\int_Q|\nabla y\left(x,t\right)|^q dxdt.$$

Proof. By applying the Gagliardo-Nirenberg theorem for $\sigma := q\frac{\rho+n}{n} > 1$, see [33, Lecture II], we have

$$\left(\int_\Omega |y(x,t)|^\sigma dx\right)^{\frac{1}{\sigma}} \leq C\|\nabla y(\cdot,t)\|_{L^q(\Omega)}^{\frac{q}{\sigma}}\|y(\cdot,t)\|_{L^\rho(\Omega)}^{\left(1-\frac{q}{\sigma}\right)},$$

for all $t \in [0,T]$ and thus equivalently

$$\left(\int_\Omega |y(x,t)|^\sigma dx\right) \leq C^\sigma\|\nabla y(\cdot,t)\|_{L^q(\Omega)}^{q}\|y(\cdot,t)\|_{L^\rho(\Omega)}^{\left(1-\frac{q}{\sigma}\right)\sigma}.$$

By integrating over t, we obtain

$$\int_0^T\int_\Omega |y(x,t)|^\sigma dxdt \leq C^\sigma \int_0^T \|\nabla y(\cdot,t)\|_{L^q(\Omega)}^{q}\|y(\cdot,t)\|_{L^\rho(\Omega)}^{\left(1-\frac{q}{\sigma}\right)\sigma}dt.$$

Since $y \in L^\infty(0,T;L^\rho(\Omega))$, we have

$$\int_0^T\int_\Omega |y(x,t)|^\sigma dxdt \leq C^\sigma\|y\|_{L^\infty(0,T;L^\rho(\Omega))}^{\frac{\rho q}{n}} \int_0^T \|\nabla y(\cdot,t)\|_{L^q(\Omega)}^{q}dt.$$

Inserting the definition of σ on the right hand-side of this inequality, we obtain the statement of the lemma from the identity

$$\int_0^T \|\nabla y(\cdot,t)\|_{L^q(\Omega)}^{q}dt = \int_0^T\int_\Omega |\nabla y(x,t)|^q dxdt = \int_Q |\nabla y(x,t)|^q dxdt,$$

and $c := C^\sigma$. □

The next lemma is also used in the proof of Theorem 6.1. This lemma is proved in [46, Lemma 4.1.1].

Lemma 6.2. *Let $\varphi(t)$ be a nonnegative and nonincreasing 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 $d := M2^{\frac{\beta}{\beta-1}}(\varphi(k_0))^{\frac{\beta-1}{\alpha}}$.

Theorem 6.1. *The solution to (26) 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 $k > \|y_0\|_{L^\infty(\Omega)} \geq 0$. As $y(\cdot,t) - k \in H^1(\Omega)$ for any $t \in [0,T]$, it holds that $(y-k)_+(\cdot,t) := \max(y(\cdot,t)-k,0) \in H_0^1(\Omega)$ for any $t \in [0,T]$, see [16, Chapter 4, Proposition 6]. Then, we choose $v = (y-k)_+(\cdot,t)$ in (26) 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 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

$$\left((y-k)'_+,(y-k)_+\right) + \beta \| (y-k)_+(\cdot,t) \|^2_{H^1_0(\Omega)} \leq \left(h(\cdot,t),(y-k)_+(\cdot,t)\right), \tag{27}$$

for 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 $y'(\cdot,t) = (y(\cdot,t)-k)' = (y-k)'_+(\cdot,t)$ 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 [15, page 46] and

$$\int_0^T \| (y-k)_+(\cdot,t) \|^2_{H^1_0(\Omega)} dt \leq \int_0^T \| (y-k)(\cdot,t) \|^2_{H^1_0(\Omega)} dt = \int_0^T \|y\|^2_{H^1_0(\Omega)} dt < \infty,$$

and

$$\int_0^T \left((y-k)_+(\cdot,t),v\right)^2_{H^1_0(\Omega)} dt \leq \int_0^T \left((y-k)(\cdot,t),v\right)^2_{H^1_0(\Omega)} dt = \int_0^T \left(y(\cdot,t),v\right)^2_{H^1_0(\Omega)} < \infty,$$

for all $v \in H^1_0(\Omega)$, we obtain with [19, 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 (27) we get

$$\frac{1}{2}\frac{d}{dt}\| (y-k)_+(\cdot,t) \|^2_{L^2(\Omega)} + \beta \| (y-k)_+(\cdot,t) \|^2_{H^1_0(\Omega)} \leq \left(h(\cdot,t),(y-k)_+(\cdot,t)\right), \tag{28}$$

for any $t \in [0,T]$. By taking the absolute value of the right hand-side of (28), renaming the variable t into $\tilde{t}$ and integrating over it from 0 to t, we obtain

$$\begin{aligned}\frac{1}{2}\| (y-k)_+(\cdot,t) \|^2_{L^2(\Omega)} + \beta \int_0^t \| (y-k)_+(\cdot,\tilde{t}) \|^2_{H^1_0(\Omega)} d\tilde{t} &\leq \int_0^t \int_\Omega |h(x,\tilde{t})(y-k)_+(x,\tilde{t})| dx d\tilde{t} \\ &\leq \int_0^T \int_\Omega |h(x,\tilde{t})(y-k)_+(x,\tilde{t})| dx d\tilde{t},\end{aligned} \tag{29}$$

where, because of the definition of k, we have $\| (y-k)_+(\cdot,0) \|^2_{L^2(\Omega)} = 0$. From (29), it follows that

$$\frac{1}{2}\| (y-k)_+(\cdot,t) \|^2_{L^2(\Omega)} \leq \int_0^T \int_\Omega |h(x,\tilde{t})(y-k)_+(x,\tilde{t})| dx d\tilde{t}, \tag{30}$$

$$\beta \int_0^t \| (y-k)_+(\cdot,\tilde{t}) \|^2_{H^1_0(\Omega)} d\tilde{t} \leq \int_0^T \int_\Omega |h(x,\tilde{t})(y-k)_+(x,\tilde{t})| dx d\tilde{t}. \tag{31}$$

By the monotonicity of the square root and taking the supremum, we obtain from (30) that

$$\sqrt{\frac{1}{2}}\| (y-k)_+ \|_{L^\infty(0,T;L^2(\Omega))} \leq \sqrt{\int_0^T \int_\Omega |h\left(x,\tilde{t}\right)(y-k)_+\left(x,\tilde{t}\right)| dx d\tilde{t}}.$$

Further with this inequality and (31), we obtain the following

$$\tilde{C}\left(\| (y-k)_+ \|^2_{L^\infty(0,T;L^2(\Omega))} + \|\nabla (y-k)_+ \|^2_{L^2(Q)}\right) \leq \int_0^T \int_\Omega |h\left(x,t\right)(y-k)_+\left(x,t\right)| dx dt, \tag{32}$$

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 [7, (3.4)] and obtain

$$\begin{aligned}
&\| (y-k)_+ \|^{\frac{4}{n+2}}_{L^\infty(0,T;L^2(\Omega))} \|\nabla (y-k)_+ \|^{\frac{2n}{n+2}}_{L^2(Q)} \\
&\leq \frac{2n+4}{4}\left(\| (y-k)_+ \|^{\frac{4}{n+2}}_{L^\infty(0,T;L^2(\Omega))}\right)^{\frac{2n+4}{4}} + \frac{2n}{2n+4}\left(\|\nabla (y-k)_+ \|^{\frac{2n}{n+2}}_{L^2(Q)}\right)^{\frac{2n+4}{2n}} \\
&\leq \frac{2n+4}{4}\left(\| (y-k)_+ \|^2_{L^\infty(0,T;L^2(\Omega))} + \|\nabla (y-k)_+ \|^2_{L^2(Q)}\right).
\end{aligned}$$

This result and (32) imply the following

$$\left(\frac{4\tilde{C}}{2n+4}\right)^{\frac{n+2}{n}} \left(\| (y-k)_+ \|^{\frac{4}{n}}_{L^\infty(0,T;L^2(\Omega))} \|\nabla (y-k)_+ \|^2_{L^2(Q)}\right) \leq \left(\int_Q |h\left(x,t\right)(y-k)_+\left(x,t\right)| dx dt\right)^{\frac{n+2}{n}}. \tag{33}$$

Then by Lemma 6.1, we have that

$$\int_Q (y-k)_+^{2\frac{n+2}{n}} dx dt \leq c\| (y-k)_+ \|^{\frac{4}{n}}_{L^\infty(0,T;L^2(\Omega))} \|\nabla (y-k)_+ \|^2_{L^2(Q)},$$

with $c > 0$. This inequality and (33) imply the following

$$C\int_Q (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dx dt \leq \left(\int_Q |h\left(x,t\right)(y-k)_+\left(x,t\right)| dx dt\right)^{\frac{n+2}{n}} dx dt, \tag{34}$$

where $\bar{C} := \left(\frac{4\tilde{C}}{c(2n+4)}\right)^{\frac{n+2}{n}} > 0$. Consequently, we have

$$\bar{C}\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dx dt \leq \left(\int_{A_k} |h\left(x,t\right)(y-k)_+\left(x,t\right)| dx dt\right)^{\frac{n+2}{n}} dx dt, \tag{35}$$

where $A_k := \{(x,t) \in Q |\ y(x,t) > k\}$. The set A_k is measurable, see [15, Proposition 2.1.1 and page 42]. By estimating the right hand-side of (35) with Hölder's inequality, see [19, page 622], we obtain

$$\begin{aligned}
\bar{C}\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\, dx dt &\leq \left(\left(\int_{A_k} |h\left(x,t\right)|^{\frac{2n+4}{n+4}} dx dt\right)^{\frac{n+4}{2n+4}} \left(\int_{A_k} |\left(y-k\right)_+\left(x,t\right)|^{\frac{2n+4}{n}} dx dt\right)^{\frac{n}{2n+4}}\right)^{\frac{n+2}{n}} \\
&= \left(\int_{A_k} |h\left(x,t\right)|^{\frac{2n+4}{n+4}} dx dt\right)^{\frac{n+4}{2n}} \left(\int_{A_k} |\left(y-k\right)_+\left(x,t\right)|^{2\frac{n+2}{n}} dx dt\right)^{\frac{1}{2}}.
\end{aligned} \tag{36}$$

If $\int_{A_k} |\,(y-k)_+(x,t)\,|^{2\frac{n+2}{n}} dxdt > 0$, then (36) implies

$$\bar{C}\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt \leq \left(\int_{A_k} |h(x,t)|^{\frac{2n+4}{n+4}} dxdt\right)^{\frac{n+4}{n}}. \tag{37}$$

This is also true in the case of $\int_{A_k} |\,(y-k)_+(x,t)\,|^{2\frac{n+2}{n}} dxdt = 0$. We use Hölder's inequality again for the right hand-side of (37), see [19, page 622], and obtain the following

$$\begin{aligned}
&\bar{C}\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt \leq \left(\int_{A_k} |h(x,t)|^{\frac{2n+4}{n+4}} dxdt\right)^{\frac{n+4}{n}} \\
&\leq \left(\left(\int_{A_k} 1^{\frac{q(4+n)}{n(q-2)+4(q-1)}} dxdt\right)^{\frac{n(q-2)+4(q-1)}{q(4+n)}} \left(\int_{A_k} \left(|h(x,t)|^{\frac{2n+4}{n+4}}\right)^{q\frac{n+4}{2n+4}} dxdt\right)^{\frac{2n+4}{q(n+4)}}\right)^{\frac{n+4}{n}} \\
&= \left(\left(\int_{A_k} |h(x,t)|^q dxdt\right)^{\frac{1}{q}}\right)^{\frac{2n+4}{n}} |A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}} \|h\|^{\frac{2n+4}{n}}_{L^q(A_k)} \leq |A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}} \|h\|^{\frac{2n+4}{n}}_{L^q(Q)},
\end{aligned} \tag{38}$$

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

$$\int_{A_k} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt \geq \int_{A_m} (y-k)_+^{2\frac{n+2}{n}}(x,t)\,dxdt = \int_{A_m} (y-k)^{2\frac{n+2}{n}}(x,t)\,dxdt \geq (h-k)^{2\frac{n+2}{n}}|A_m|. \tag{39}$$

We combine (38) and (39) and obtain the following $(m-k)^{2\frac{n+2}{n}}|A_m| \leq \hat{C}\|h\|^{\frac{2n+4}{n}}_{L^q(Q)}|A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}}$, $\hat{C} := \frac{1}{\bar{C}}$. Therefore we have

$$|A_m| \leq \left(\frac{\hat{C}^{\frac{n}{2n+4}}\|h\|_{L^q(Q)}}{m-k}\right)^{\frac{2n+4}{n}} |A_k|^{\frac{n+4}{n}-\frac{2n+4}{qn}}. \tag{40}$$

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 6.2 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 (40) 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 $m \geq \|y_0\|_{L^\infty(\Omega)}$. Summarizing, 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 $(y+k)_- := \min(y+k, 0)$ 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 $\|y\|_{L^\infty(Q)} \leq C\|h\|_{L^q(Q)} + \|y_0\|_{L^\infty(\Omega)}$. □

6.2 Existence of a minimiser

Next, we give a proof of existence of a minimiser to the optimal control problem (2) on a compact set U of $L^q(Q)$. A compact set can, in general, be constructed with the Kolmogorov-M. Riesz-Fréchet theorem; see [12, Theorem 4.26] or [20]. Beyond the fact that a compact control set

may not be satisfactory in applications, we remark that the PMP characterisation of an optimal solution on such a set U is difficult. In fact, a needle variation can cause a smaller value of the cost functional than its optimal value on U. Thus the proof of Theorem 3.3 is not valid in this case.

Theorem 6.2. *Let U be a compact set. Then the optimal control problem (2) admits an optimal solution $\bar{u} \in U$.*

Proof. For the proof, we follow the reasoning in [31, 44]. First, the objective functional $J(y,u)$ is bounded from below. Therefore, there exists an infimum

$$\bar{J} := \inf_{u \in U} \hat{J}(u) := J(S(u), u),$$

and a minimizing sequence $(u_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty} \hat{J}(u_n) = \bar{J}$, see [3, II Theorem 4.1]. As U is compact and $(u_n)_{n\in\mathbb{N}} \subseteq U$, there exists a subsequence, still denoted by $(u_n)_{n\in\mathbb{N}}$ and $\bar{u} \in U$ with $||u_n - \bar{u}||_{L^q(\Omega)} \to 0$ for $n \to \infty$.

Now, we consider

$$\hat{J}(u) = \hat{J}_c(u) + G(u),$$

where $G(u)$ is given by (5) and $\hat{J}_c(u)$ is the continuous part of $\hat{J}(u)$. We have

$$\bar{J} = \liminf_{n\to\infty} \hat{J}(u_n) = \liminf_{n\to\infty} \left(\hat{J}_c(u_n) + G(u_n)\right) \geq \liminf_{n\to\infty} \hat{J}_c(u_n) + \liminf_{n\to\infty} G(u_n),$$

see for example [18, Theorem 3.127] for basic properties of the lim inf. As the control-to-state operator $S : L^q(\Omega) \to L^2(\Omega)$ is continuous, the functional $\hat{J}_c(u)$ is continuous from $L^q(\Omega)$ to $\mathbb{R}$, see [3, III Theorem 1.8]. Therefore we have

$$\liminf_{n\to\infty} \hat{J}_c(u_n) = \lim_{n\to\infty} \hat{J}_c(u_n) = \hat{J}_c(\bar{u}).$$

Next, we investigate the term $\liminf_{n\to\infty} G(u_n)$. From the strong $L^q(\Omega)$ convergence, we have that there is a subsequence of $(u_n)_{n\in\mathbb{N}}$, still denoted with $(u_n)_{n\in\mathbb{N}}$, which converges to $\bar{u}$ almost everywhere; i.e. there is a set $\mathring{\Omega}$ with $\Omega \backslash \mathring{\Omega}$ being a set of measure zero such that $u_n(x) \to \bar{u}(x)$ for all $x \in \mathring{\Omega}$ and $n \to \infty$, see [7, Proposition 3.6, Remark 3.7].

As lower semi-continuous functions are measurable and the composition of measurable functions is measurable [7, Proposition 2.2], we can consider the composition $f_n := g \circ u_n$ and using the Lemma of Fatou (see [7, Lemma 2.15]), we have

$$\liminf_{n\to\infty} \int_{\mathring{\Omega}} g(u_n(x))\,dx = \liminf_{n\to\infty} \int_{\mathring{\Omega}} f_n(x)\,dx \geq \int_{\mathring{\Omega}} \liminf_{n\to\infty} f_n(x)\,dx = \int_{\mathring{\Omega}} \liminf_{n\to\infty} g(u_n(x))\,dx. \quad (41)$$

If we define $a_n^x := u_n(x) \to \bar{u}(x) := \bar{a}^x$ for $n \to \infty$ for every $x \in \mathring{\Omega}$, then $(a_n^x)_{n\in\mathbb{N}}$ is a converging sequence in $\mathbb{R}$ for every $x \in \mathring{\Omega}$ converging to $\bar{a}^x$ for $n \to \infty$ and thus we have

$$\int_{\mathring{\Omega}} \liminf_{n\to\infty} g(u_n(x))\,dx = \int_{\mathring{\Omega}} \liminf_{n\to\infty} g(a_n^x)\,dx \geq \int_{\mathring{\Omega}} g(\bar{a}^x)\,dx = \int_{\mathring{\Omega}} g(\bar{u}(x))\,dx,$$

because of the lower semi-continuity of g. This gives

$$\liminf_{n\to\infty} \int_{\mathring{\Omega}} g(u_n(x))\,dx \geq \int_{\mathring{\Omega}} g(\bar{u}(x))\,dx,$$

and

$$\liminf_{n\to\infty} \int_\Omega g\left(u_n\left(x\right)\right) dx \geq \int_\Omega g\left(\bar{u}\left(x\right)\right) dx,$$

as both $\int_{\Omega\setminus\mathring{\Omega}} g\left(u_n\left(x\right)\right) dx = 0$ and $\int_{\Omega\setminus\mathring{\Omega}} g\left(\bar{u}\left(x\right)\right) dx = 0$; see [5, X Remark 4.4]. This proves the following result

$$\liminf_{n\to\infty} G\left(u_n\right) \geq G\left(\bar{u}\right).$$

Therefore we have $\bar{J} \geq \hat{J}_c\left(\bar{u}\right) + G\left(\bar{u}\right)$. Thus, the control $\bar{u} \in U$ is optimal. □

6.3 Measurability of composite functions

The next lemma states that the composition of a Lebesgue measurable function $u : Q \to \mathbb{R}$, $Q \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, with a lower semi-continuous function $g : \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable.

Lemma 6.3. *Let $u : Z \to \mathbb{R}$ be Lebesgue measurable and $g : \mathbb{R} \to \mathbb{R}$ lower semi-continuous. Then the composition $g \circ u : Z \to \mathbb{R}$ is Lebesgue measurable.*

Proof. By [15, Example 2.6.3], we have that u is Lebesgue measurable if and only if $u : (Z, \mathcal{M}) \to (\mathbb{R}, \mathcal{B})$ is measurable where $(Z, \mathcal{M})$ is a measurable space, $\mathcal{M}$ is the σ-algebra of the Lebesgue measurable subsets of $\mathbb{R}$ and $(\mathbb{R}, \mathcal{B})$ is a measurable space where $\mathcal{B}$ is the σ-algebra generated by the collection of open subsets of $\mathbb{R}$.

Next, we show that $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ is measurable, i.e. Borel measurable. We define for any constant $c \in \mathbb{R}$ the set $A := \{z \in \mathbb{R} |\ g\left(z\right) \leq c\}$. Let $\left(z_n\right)_{n\in\mathbb{N}} \subseteq A$ be a sequence with $\lim_{n\to\infty} z_n = \bar{z}$, then $c \geq \liminf_{n\to\infty} g\left(z_n\right) \geq g\left(\bar{z}\right)$, see [18, Theorem 3.127] for calculation rules of $\liminf$. This means that $\bar{z} \in A$ and thus A is closed. By [15, Proposition 1.1.4] we know that A belongs to $\mathcal{B}$ and thus by [15, Proposition 2.1.1 and page 42], we have that g is Borel measurable. Then with [15, Proposition 2.6.1], we have that $g \circ u : (Z, \mathcal{M}) \to (\mathbb{R}, \mathcal{B})$ is measurable, which means $g \circ u$ is Lebesgue measurable. □

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] H. W. Alt. *Linear Functional Analysis: An Application-Oriented Introduction*. Springer, 2016.

[3] H. Amann and J. Escher. *Analysis I*. Birkhäuser Basel, 2006.

[4] H. Amann and J. Escher. *Analysis II*. Birkhäuser, 2008.

[5] H. Amann and J. Escher. *Analysis III*. Birkhäuser Basel, 2009.

[6] A. Ambrosetti and R. E. Turner. Some discontinuous variational problems. *Diff. & Integral Equat*, 1:341–349, 1988.

[7] L. Ambrosio, G. Da Prato, and A. C. Mennucci. *Introduction to Measure Theory and Integration*. Edizioni della Normale, 2011.

[8] D. P. Bertsekas. *Nonlinear Programming*. Athena scientific Belmont, 1999.

[9] V. G. Boltyanskiĭ, R. V. Gamkrelidze, and L. S. Pontryagin. On the theory of optimal processes. *Dokl. Akad. Nauk SSSR (N.S.)*, 110:7–10, 1956.

[10] J. F. Bonnans. On an algorithm for optimal control using Pontryagin's maximum principle. *SIAM Journal on Control and Optimization*, 24(3):579–588, 1986.

[11] A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*, volume 8. SIAM, 2011.

[12] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer Science & Business Media, 2010.

[13] M. Burger and W. Mühlhuber. Iterative regularization of parameter identification problems by sequential quadratic programming methods. *Inverse Problems*, 18(4):943, 2002.

[14] E. 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.

[15] D. L. Cohn. *Measure Theory*. Springer, 2013.

[16] R. Dautray and J.-L. 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.

[17] A. V. Dmitruk and N. P. Osmolovskii. On the proof of Pontryagin's maximum principle by means of needle variations. *Journal of Mathematical Sciences*, 218(5):581–598, 2016.

[18] C. M. Dunn. *Introduction to Analysis*. CRC Press, 2017.

[19] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[20] H. Hanche-Olsen and H. Holden. The Kolmogorov–Riesz compactness theorem. *Expositiones Mathematicae*, 28(4):385–394, 2010.

[21] M. Hintermüller and T. 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.

[22] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer, Berlin, 2009.

[23] K. Ito and K. Kunisch. A variational approach to sparsity optimization based on Lagrange multiplier theory. *Inverse problems*, 30(1):015001, 2013.

[24] K. Ito and K. Kunisch. Optimal control with $L^p(\Omega)$, $p \in [0, 1)$, control cost. *SIAM Journal on Control and Optimization*, 52(2):1251–1275, 2014.

[25] B. S. Jovanović and E. Süli. *Analysis of Finite Difference Schemes*, volume 46. Springer London, London, 2014.

[26] I. A. Krylov and F. 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.

[27] I. A. Krylov and F. 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.

[28] O. A. Ladyzhenskaia, V. A. Solonnikov, and N. N. Ural'tseva. *Linear and Quasi-Linear Equations of Parabolic Type*, volume 23. American Mathematical Soc., 1988.

[29] R. Leitao and E. Teixeira. Regularity and geometric estimates for minima of discontinuous functionals. *Rev. Mat. Iberoam*, 31(1):69–108, 2015.

[30] X. Li and J. Yong. *Optimal Control Theory for Infinite Dimensional Systems*. Birkhäuser, 1995.

[31] J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1971.

[32] J. M. Martínez. Minimization of discontinuous cost functions by smoothing. *Acta Applicandae Mathematica*, 71(3):245–260, 2002.

[33] L. Nirenberg. On elliptic partial differential equations. In *Il principio di minimo e sue applicazioni alle equazioni funzionali*, pages 1–48. Springer, 2011.

[34] F. 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.

[35] L. S. Pontryagin, V. G. Boltyanskiĭ, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes*. John Wiley & Sons, New York-London, 1962.

[36] J.-P. Raymond and H. 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.

[37] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[38] Y. Sakawa and Y. Shindo. On global convergence of an algorithm for optimal control. *IEEE Transactions on Automatic Control*, 25(6):1149–1153, 1980.

[39] A. Schindele and A. Borzì. Proximal schemes for parabolic optimal control problems with sparsity promoting cost functionals. *International Journal of Control*, 90(11):2349–2367, 2017.

[40] Y. Shindo and Y. Sakawa. Local convergence of an algorithm for solving optimal control problems. *Journal of optimization theory and applications*, 46(3):265–293, 1985.

[41] J. Stoer, R. Bulirsch, R. H. Bartels, W. Gautschi, and C. Witzgall. *Introduction to Numerical Analysis*. Texts in applied mathematics. Springer, New York, 2002.

[42] M. 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.

[43] M. 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.

[44] F. Tröltzsch. *Optimal Control of Partial Differential Equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications.

[45] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2011.

[46] Z. Wu, J. Yin, and C. Wang. *Elliptic & Parabolic Equations*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2006.