Optimal control of semilinear elliptic PDEs with state constraints - numerical analysis and implementation

Dissertation with the aim of achieving a doctoral degree
at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Mathematics
of Universität Hamburg

submitted by
Ahmad Ahmad Ali

Hamburg, 2017

Referees: Prof. Dr. Michael Hinze
and Prof. Dr. Fredi Tröltzsch

Defense date: 13 June 2017

Eidesstattliche Versicherung

Declaration on oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.
I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Ort, Datum Place, date	Unterschrift Signature

Acknowledgements

Firstly, I would like to express my gratitude to my advisor Prof. Dr. Michael Hinze. Without his advice and continuous support, the thesis would not have been accomplished.
My sincere thanks also go to Prof. Dr. Klaus Deckelnick for his cooperation and advice in the first part of the thesis. The same thanks also go to Prof. Dr. Elisabeth Ullmann for her advice and cooperation in the second part of the thesis.
I would like also to thank my family: my parents, my brothers and my sisters for their support during my studies.

Contents

Chapter 1

Introduction

Many of the physical processes in the real world can be described mathematically by using partial differential equations (PDEs). For example heat flow, diffusion, fluid flow, elastic deformation and wave propagation and many other phenomena are modelled by PDEs. Modelling such processes in practical applications like in industry, medicine or engineering may necessitate knowing the optimal inputs in the considered models for some financial or safety reasons. This leads to establishing minimization problems with PDEs constraints. Solving such optimization problems prove challenging since the involved variables usually belong to infinite dimensional spaces and hence discretization concepts shall be developed. Therefore, it has been of interest to study optimal control problems of PDEs with probably additional constraints on the control/state variables, see [73, 45] for further details.

The thesis is divided into two parts. In the first part, we consider the optimal control problem

$$(\mathbb{P}) \quad \min_{u \in U_{ad}} J(u) := \frac{1}{2}\|y - y_0\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}$$

subject to the semilinear elliptic PDE

$$\begin{aligned} -\Delta y + \phi(y) &= u \quad \text{in } \Omega \subset \mathbb{R}^2, \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

and the pointwise state constraints

$$y_a(x) \leq y(x) \leq y_b(x), \quad x \in K \subset \Omega.$$

The precise assumptions on the data of the problem will be given in Section 2.1. Since the state equation is in general nonlinear, the control problem $(\mathbb{P})$ is nonconvex. Consequently, there may be several solutions of the necessary first order conditions. These can be examined further with the help of second order conditions. However, second order conditions can only decide if the given point is a local mimimum of $(\mathbb{P})$ and they in general don't provide any information about whether the given point is a global solution of the control problem.

Our aim in this part is to establish a sufficient condition that helps us to decide if a function $\bar{u}$ satisfying the necessary first order conditions is a global

solution of ($\mathbb{P}$). We establish such a condition in Theorem 2.3.5 under certain growth conditions of the nonlinearity ϕ. More precisely, this condition has the form $\|\bar{p}\|_{L^q(\Omega)} \leq \eta$ where q and η are constants depending only on the data of the problem and can be computed explicitly while $\bar{p}$ denotes the adjoint state associated with $\bar{u}$. It turns out that an analogous result can also be established for the variational discrete, (see [46]), counterpart ($\mathbb{P}_h$) of ($\mathbb{P}$), see Theorem 3.3.4. Furthermore, any sequence $(\bar{u})_{0<h\leq h_0}$ of global minima of ($\mathbb{P}_h$) that satisfies this condition uniformly converges to a global minimum of problem ($\mathbb{P}$) as the discretization parameter h tends to zero. Relying on these conditions, we can derive an error bound of order $O(h^{1-\varepsilon})$ for arbitrarily small $\varepsilon > 0$ for the convergence of the sequence $(\bar{u})_{0<h\leq h_0}$. Finally, we show that the condition $\|\bar{p}\|_{L^q(\Omega)} \leq \eta$ also implies the sufficient second order conditions derived in [20].

We give a brief overview of the literature on optimal control of semilinear PDEs with pointwise state constraints. For a broad overview, the interested reader is referred to the references of the respective citations. In [16] the analysis of semiliner elliptic control problems with pointwise state constraints and boundary controls is considered and the necessary first order conditions are established (compare [15] for the linear-quadratic case). See also [21] where the pointwise constraints are imposed on the gradient of the state and the controls are distributed in the spatial domain. Further analysis concerning the regularity of the optimal control and the associated multipliers as well as discussion of sufficient second order conditions can be found in [25] and in [24]. Second order conditions for finitely/infinitely many poitwise state constraints are established in [22, 19, 20], and compare [26] for the role of the second order conditions in PDE constrained control problems.

The finite element discretization of state constrained semilinear elliptic control problems with plain convergence, that is convergence without any rates of convergence, is considered in [23] and in [47] for a wider class of perturbations, including the finite element discretization, of the continuous control problems. On the other hand, established convergence rates can be found in [60] for finite dimensional controls and finitely many poitwise state constraints, and in [17] for control functions and finitely many pointwise state constraints. Just recently in [64] rates of convergence are derived when the controls are functions and the state constrains are imposed on infinitely many points in the domain. In [34] convergence rates are obtained for liner-quadratic control problems. In fact, we will use the discretization concepts considered there to discretize problem ($\mathbb{P}$). For the numerical analysis when the pointwise constraints are imposed only on the control variable we refer to [3, 18]. Finally, a detailed discussion of discretization concepts and error analysis in PDE-constrained control problems can be found in [48, 49] and[45, Chapter 3].

In the second part of the thesis we are interested in the control problem

$$(\mathrm{P}_\omega) \qquad \min_{(y,u)\in H_0^1(D)\times L^2(D)} J(y,u) = \frac{1}{2}\|y-z\|^2_{L^2(D)} + \frac{\alpha}{2}\|u\|^2_{L^2(D)}$$

subject to

$$\begin{aligned} -\nabla\cdot(a(\omega,x)\nabla y(\omega,x)) &= u(x) \quad \text{in } D\subset\mathbb{R}^d,\\ y(\omega,x) &= 0 \quad \text{on } \partial D, \end{aligned}$$

and

$$u_a \leq u(x) \leq u_b \qquad \text{for a.e. } x \in D,$$

where $a(\omega, x)$ is a random field defined on a given probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The precise assumptions on the data of the problem will be formulated in Section 4.1. Following the typical notation in the literature on PDEs with stochastic coefficients, the notation in this part will differ from and is independent of the one in the first part.

It is clear that for a fixed realization of the coefficient a, the control problem (P_ω) admits a unique solution. In fact, we will show in Theorem 4.3.2 that the mapping $u^* : \Omega \to L^2(D)$ where $u^*(\omega)$ is the solution of (P_ω) defines a $L^2(D)$-valued random variable.

Our aim is to compute the statistics of the mapping u^*, like the expected value $\mathbb{E}[u^*]$ or the variance $\mathrm{var}[u^*]$. This helps us to understand how random fluctuations in the state equation affect the optimal controls and provides practical information on the design of control devices subject to uncertain inputs. It is clear, however, that the quantity $\mathbb{E}[u^*]$ needs not be a solution of an optimal control problem in general and is not necessarily a robust control. More precisely, we will be interested in the numerical analysis of approximating $\mathbb{E}[u^*]$ by a multilevel Monte Carlo (MLMC) estimator, see [5, 32].

Various formulations for optimal control problems of PDEs with random coefficients have appeared in the literature to date. In what follows we give a quick overview and classify those formulations according to the type of the control (deterministic or stochastic) and the form of the cost functional to be minimized. In addition we comment on solvers for these problems.

Consider a cost functional $J = J(u, y(u), a)$ where u denotes the control, y denotes the state and a is some parameter associated with the PDE constraint. In our setting, a is a random function with realizations denoted by a_ω. We distinguish the following problem formulations:

(a) Mean-based control, see [9, 10, 11]: Replace a by its expected value $\mathbb{E}[a]$. Minimize $J(u, y(u), \mathbb{E}[a])$ by a deterministic optimal control.

(b) Individual or "pathwise" control, see [9, 11, 62, 63]: Fix a_ω, minimize $J(u, y(u), a_\omega)$ and obtain a realization $u^*(\omega)$ of a random field u^*. In a postprocessing step, compute the statistics of u^*, e.g. $\mathbb{E}[u^*]$.

(c) Averaged control, see [57, 79]: Control the averaged (expected) state by minimizing $J(u, \mathbb{E}[y(u)], a)$ using a deterministic optimal control.

(d) Robust deterministic control, see [10, 12, 37, 41, 42, 50, 53, 54, 55, 58, 67]: Minimize the expected cost $\mathbb{E}[J(u, y(u), a)]$ by a deterministic optimal control.

(e) Robust stochastic control, see [6, 7, 29, 30, 31, 56, 72]: Minimize the expected cost $\mathbb{E}[J(u, y(u), a)]$ by a stochastic optimal control.

The mean-based control problem (a) does not account for the uncertainties in the PDE and it is not clear if the resulting deterministic optimal control is robust with respect to the random fluctuations. The pathwise control problem (b) is highly modular and can be combined easily with sampling methods, e.g.

Monte Carlo or sparse grid quadrature. However, the expected value $\mathbb{E}[u^*]$ does not solve an optimal control problem and is in general not a robust control. The average control problem (c) introduced by Zuazua [79] seeks to minimize the distance of the expected state to a certain desired state. This is an interesting alternative to the robust control problem in (d) where the expected distance of the (random) state to a desired state is minimized. Since the cost functional in (c) uses a weaker error measure than the cost functional in (d) the average optimal control does not solve the robust control problem in general. *Stochastic* optimal controls in (e) are of limited practical use since controllers typically require a *deterministic* signal. This can, of course, be perturbed by a *known* mean-zero stochastic fluctuation which models the uncertainty in the controller response. For these reasons, deterministic, robust controls in (d) are perhaps most useful in practice and have attracted considerable attention compared to the other formulations. However, control problems in (d) or in (c) involve an infinite number of PDE constraints which are coupled by a single cost functional. The approximate solution of such problems is extremely challenging and requires much more computational resources than e.g. a deterministic control problem with a single deterministic PDE constraint. For this reason it is worthwhile to explore alternative problem formulations.

Note that the problem (P_ω) is of the form (b). Moreover, the expected value $\mathbb{E}[u^*]$ can be used as initial guess for the robust control problem in (d) if the variance $\mathrm{var}[u^*] = \mathbb{E}[u^* - \mathbb{E}[u^*]]^2$ is small. This is justified by the Taylor expansion

$$\mathbb{E}[\widehat{J}(u^*)] = \widehat{J}(\mathbb{E}[u^*]) + \frac{1}{2}\frac{d^2\widehat{J}}{du^2}(\mathbb{E}[u^*])\mathrm{var}[u^*] \ + \ \text{higher order moments},$$

where we have used the reduced cost functional $\widehat{J} = \widehat{J}(u)$ and the assumption that $\widehat{J}$ is smooth. However, we recall that $\mathbb{E}[u^*]$ is in general not the solution of an optimization problem.

The control problems (a)–(e) have been tackled by a variety of solver methodologies. We mention stochastic Galerkin approaches in [50, 56, 58, 67], stochastic collocation in [9, 11, 31, 53, 54, 55, 67, 72], low-rank, tensor-based methods in [6, 7, 37], and reduced basis/POD methods in [12, 29, 30, 41, 62, 63].

Part I

Chapter 2

Optimal Control of Semilinear Elliptic PDEs with State Constraints

In this chapter we consider an optimal control problem of a semilinear elliptic partial differential equation (PDE) where the nonlinearity is monotone and it satisfies certain growth conditions. The control is distributed in the domain Ω and the cost functional is of tracking type. Pointwise constraints on the control and the state are also considered.

The exposition in this chapter is as follows: in Section 2.1 we introduce the notation and set up the optimal control problem that we will consider and formulate the main assumptions on its data. In Section 2.2 we review some of the standard results about the state equation like the well-posdness of the PDE, the regularity of the solution and the differentiability of the control-to-state mapping. Section 2.3 is devoted to the study of the optimal control problem. We recall the associated first order necessary optimality conditions and we derive our main result which is a sufficient condition for global minima of the control problem. We also derive a sufficient condition that implies second order sufficient optimality conditions.

2.1 The Problem Setting

2.1.1 Notation

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. For $1 \leq p \leq \infty$ we denote by $L^p(\Omega)$ the usual Banach spaces of equivalence classes of Lebesgue measurable functions with norm $\|\cdot\|_{L^p(\Omega)}$. For $m \in \mathbb{N}_0$, we denote by $W^{m,p}(\Omega)$ the classical Sobolev spaces with norm $\|\cdot\|_{W^{m,p}(\Omega)}$. We denote by $W_0^{m,p}(\Omega)$ the closure of $C_0^\infty(\Omega)$ in $W^{m,p}(\Omega)$. In particular, the functions in $W_0^{1,p}(\Omega)$ vanish on $\partial\Omega$ in the sense of traces. In the case $p = 2$ we write $H^m(\Omega) := W^{m,2}(\Omega)$ and $H_0^m(\Omega) := W_0^{m,2}(\Omega)$.

We denote by $C^m(\bar{\Omega})$ the Banach spaces of functions whose derivatives up to order m are continuous in $\bar{\Omega}$ with norm $\|\cdot\|_{C^m(\bar{\Omega})}$. We write $C(\bar{\Omega}) := C^0(\bar{\Omega})$. For $0 < \beta \leq 1$, we denote by $C^{m,\beta}(\bar{\Omega})$ the classical spaces of β-Hölder continuous

functions with norm $\|\cdot\|_{C^{m,\beta}(\bar\Omega)}$. The closure of $C_0^\infty(\Omega)$ in $C(\bar\Omega)$ is denoted by $C_0(\Omega)$. In particular, $C_0(\Omega)$ is the space of all functions which are continuous in $\bar\Omega$ and vanish on $\partial\Omega$.

For a compact subset $K \subset \Omega$ or $K = \bar\Omega$ we denote by $\mathcal{M}(K)$ the space of all real regular Borel measures defined on K. We remark that $\mathcal{M}(K)$ can also be identified with the dual space of $C(K)$ and it is a Banach space for the norm

$$\|\mu\|_{\mathcal{M}(K)} = \sup_{f\in C(K),\|f\|_{C(K)}\le 1} \int_K f\, d\mu.$$

2.1.2 The Problem Setting

We consider the minimization problem

$$(\mathbb{P}) \qquad \min_{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)} J(y,u) = \frac{1}{2}\|y-y_0\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}$$

subject to

$$\begin{aligned} -\Delta y + \phi(y) &= u \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

and the *pointwise constraints*

$$\begin{aligned} u_a \le u(x) &\le u_b && \text{for a.e. } x\in\Omega, \\ y_a(x) \le y(x) &\le y_b(x) && \forall\, x\in K\subset\Omega, \end{aligned}$$

where we assume

- $\Omega \subset \mathbb{R}^2$ is a bounded, convex and polygonal domain.
- K is a (possibly empty) compact subset of Ω.
- $u_a \in \mathbb{R}\cup\{-\infty\}$ and $u_b \in \mathbb{R}\cup\{\infty\}$ with $u_a \le u_b$.
- $y_a, y_b \in C_0(\Omega)$ are given functions that satisfy $y_a(x) < y_b(x)$, $x\in K$.
- $y_0 \in L^2(\Omega)$ and $\alpha > 0$ are given.
- $\phi : \mathbb{R}\to\mathbb{R}$ is of class C^2 and monotonically increasing.
- There exist $r > 1$ and $M \ge 0$ such that

$$|\phi''(s)| \le M\phi'(s)^{\frac{1}{r}} \quad \text{for all } s\in\mathbb{R}, \tag{2.2}$$

  where ϕ' and ϕ'' denote the first and second derivative of ϕ, respectively.

We recall that the problem $(\mathbb{P})$ is called an *optimal control problem* where the function u is the *control*, y is the *state* and the semilinear elliptic PDE (2.1) is called the *state equation*.

A function ϕ satisfying the previous assumptions enjoys some properties that will be useful in our analysis. We summarize these properties in Appendix A.1. To have an example for such a function, consider $\phi(s) = |s|^{q-2}s$ for $q > 3$. Then we have

$$|\phi''(s)| = (q-1)(q-2)|s|^{q-3} = (q-2)(q-1)^{\frac{1}{q-2}}[\phi'(s)]^{\frac{q-3}{q-2}},$$

and thus, (2.2) is satisfied if we choose $r = \frac{q-2}{q-3}$ and $M = (q-2)(q-1)^{\frac{1}{q-2}}$.

2.2 The State Equation

In this section we will recall the classical results about the state equation (2.1) that will be relevant in our study of the optimal control problem ($\mathbb{P}$). In particular, we will investigate the well-posedness of (2.1), the regularity of the state variable and the differentiability of the control-to-state operator.

We begin by recalling the weak formulation of (2.1) which reads: for a given $u \in L^2(\Omega)$, find $y \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla v + \phi(y)v \, dx = \int_\Omega uv \, dx \quad \forall\, v \in H_0^1(\Omega). \tag{2.3}$$

If such a function $y \in H_0^1(\Omega)$ exists, it is called a weak solution to (2.1).

Theorem 2.2.1 *For every $u \in L^2(\Omega)$ the boundary value problem (2.1) admits a unique weak solution $y \in H_0^1(\Omega) \cap H^2(\Omega)$. Moreover, there exists $c > 0$ such that*

$$\|y\|_{H^2(\Omega)} \leq c\big(1 + \|u\|_{L^2(\Omega)}\big). \tag{2.4}$$

Proof: We divide the proof into two steps. In Step 1, we show that for every $u \in L^2(\Omega)$, there exists a unique solution $y \in H_0^1(\Omega)$ to (2.3). In Step 2, we show that the solution y belongs to $H_0^1(\Omega) \cap H^2(\Omega)$ and we verify the estimate (2.4).

Step 1:

We first observe that for a given $u \in L^2(\Omega)$, a function $y \in H_0^1(\Omega)$ is a solution to (2.3) if and only if it is a solution to following variational problem: find $y \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla y \cdot \nabla v + [\phi(y) - \phi(0)]v \, dx = \int_\Omega [u - \phi(0)]v \, dx \quad \forall\, v \in H_0^1(\Omega). \tag{2.5}$$

In the light of (A.1) and Lemma A.1.4, the superposition operator associated with ϕ maps $H_0^1(\Omega)$ into $L^t(\Omega)$ for $1 \leq t < \infty$ and it is also continuous. Hence, all the integrals in (2.5) are well defined.

Since ϕ is monotone, the existence and uniqueness of a solution $y \in H_0^1(\Omega)$ to (2.5) follows from applying *the main theorem on monotone operators* in a standard way, see for instance [77, Theorem 26.A] for this theorem. In fact, it is more convenient to apply this theorem to (2.5) instead of (2.3) which is why we introduced the former.

If we test (2.5) by the solution y, it follows from the Poincaré's inequality and the monotonicity of ϕ that

$$\begin{aligned}
\|y\|^2_{H^1(\Omega)} &\leq \int_\Omega |\nabla y|^2 + [\phi(y) - \phi(0)]y \, dx = \int_\Omega [u - \phi(0)]y \, dx \\
&\leq \|u - \phi(0)\|_{L^2(\Omega)} \|y\|_{L^2(\Omega)} \\
&\leq c\big(1 + \|u\|_{L^2(\Omega)}\big)\|y\|_{H^1(\Omega)},
\end{aligned}$$

which implies

$$\|y\|_{H^1(\Omega)} \leq c\big(1 + \|u\|_{L^2(\Omega)}\big). \tag{2.6}$$

Step 2:
Let $y \in H_0^1(\Omega)$ be the solution to (2.5) for a given $u \in L^2(\Omega)$. Let $f := u - \phi(y)$ and consider the boundary value problem

$$-\Delta w = f \quad \text{in } \Omega, \quad \text{and} \quad w = 0 \quad \text{on } \partial\Omega. \tag{2.7}$$

Since $f \in L^2(\Omega)$, it follows from [40, Theorem 4.4.3.7] that (2.7) admits a unique solution $w \in H_0^1(\Omega) \cap H^2(\Omega)$. Furthermore, according to [40, Theorem 4.3.1.4] there exists a constant $c > 0$ such that

$$\|w\|_{H^2(\Omega)} \leq c\big(\|f\|_{L^2(\Omega)} + \|w\|_{L^2(\Omega)}\big). \tag{2.8}$$

However, y is also the unique weak solution to (2.7) in $H_0^1(\Omega)$. Consequently, we conclude that $w = y$ and $y \in H_0^1(\Omega) \cap H^2(\Omega)$.
From the continuous embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ we have $y \in C(\bar{\Omega})$. Thus, from (2.8), Lemma A.1.3 and (2.6) we get

$$\begin{aligned}
\|y\|_{H^2(\Omega)} &\leq c\big(\|u - \phi(y)\|_{L^2(\Omega)} + \|y\|_{L^2(\Omega)}\big) \\
&\leq c\big(\|u\|_{L^2(\Omega)} + \|\phi(y) - \phi(0)\|_{L^2(\Omega)} + \|\phi(0)\|_{L^2(\Omega)} + \|y\|_{L^2(\Omega)}\big) \\
&\leq c\big(\|u\|_{L^2(\Omega)} + \|y\|_{L^2(\Omega)} + 1\big) \\
&\leq c\big(\|u\|_{L^2(\Omega)} + 1\big),
\end{aligned}$$

which is the estimate (2.4) and the proof is complete. ■

Remark 1 Since the domain Ω is assumed to be polygonal, and thus its boundary is non-smooth, it is essential to demand that Ω is convex to guarantee the H^2 regularity of the state variable y in Theorem 2.2.1. This high regularity of y is exploited frequently in our analysis and it simplifies it. Domains with sufficiently smooth boundaries guarantee this regularity for y even if they are nonconvex (see [40, Chapter 2]), however, such domains lead to some tedious analysis at the discrete level, which is why we avoid considering them in this work.

In the light of Theorem 2.2.1, we introduce the mapping

$$\mathcal{G} : L^2(\Omega) \to H_0^1(\Omega) \cap H^2(\Omega) \tag{2.9}$$

such that $y := \mathcal{G}(u)$ is the solution to (2.3) for a given $u \in L^2(\Omega)$. In the context of optimal control of PDEs, the mapping $\mathcal{G}$ is sometimes called *the control-to-state operator* since it assigns to each control u the corresponding state y. We state some of its properties in what follows for our further analysis.

Proposition 2.2.2 *The mapping $\mathcal{G}$ introduced in* (2.9) *is of class C^1 and for every $u, v \in L^2(\Omega)$ the first derivative $z := \mathcal{G}'(u)v$ is the solution of the linear elliptic boundary value problem*

$$\begin{aligned}
-\Delta z + \phi'(y)z &= v \quad \text{in } \Omega, \\
z &= 0 \quad \text{on } \partial\Omega,
\end{aligned} \tag{2.10}$$

where $y := \mathcal{G}(u)$.

Proof: To obtain the proof, we use the *Implicit Function Theorem*, see for instance [76, Theorem 4.B] and we argue like in the proof of [22, Theorem 2.5]. To begin, consider the mapping

$$F : H_0^1(\Omega) \cap H^2(\Omega) \times L^2(\Omega) \to L^2(\Omega), \qquad F(y,u) = -\Delta y + \phi(y) - u.$$

Notice that $F(y,u) = 0$ if and only if $y = y_u := \mathcal{G}(u)$ and that F is of class C^1 from the assumptions on ϕ. Moreover, it is easy to see that

$$F_y(y_u,u) : H_0^1(\Omega) \cap H^2(\Omega) \to L^2(\Omega), \qquad F_y(y_u,u)z = -\Delta z + \phi'(y_u)z$$

is bijective. Thus, we conclude from the implicit function theorem that the mapping $u \mapsto y_u := \mathcal{G}(u)$ is of class C^1. Differentiating with respect to u the following expression

$$F(\mathcal{G}(u),u) = 0$$

we deduce the PDE (2.10). This completes the proof. ■

Lemma 2.2.3 *Let $\mathcal{G}$ be the mapping introduced in* (2.9)*. Then there exists $c > 0$ depending only on Ω such that*

$$\|\mathcal{G}(u) - \mathcal{G}(v)\|_{L^2(\Omega)} \le c\|u - v\|_{L^2(\Omega)} \qquad \forall\, u, v \in L^2(\Omega).$$

Proof: The result is obtained by utilizing the Poincaré's inequality, the monotonicity of ϕ and the Cauchy-Schwarz inequality as follows. For a given $u, v \in L^2(\Omega)$, let $y_u := \mathcal{G}(u)$ and $y_v := \mathcal{G}(v)$. Then we have

$$\begin{aligned}
\frac{1}{c}\|y_u - y_v\|^2_{L^2(\Omega)} &\le \int_\Omega |\nabla(y_u - y_v)|^2\,dx \\
&\le \int_\Omega |\nabla(y_u - y_v)|^2 + [\phi(y_u) - \phi(y_v)](y_u - y_v)\,dx \\
&= \int_\Omega (u - v)(y_u - y_v)\,dx \\
&\le \|u - v\|_{L^2(\Omega)}\|y_u - y_v\|_{L^2(\Omega)}.
\end{aligned}$$

Here $c > 0$ is from the Poincaré's inequality and it depends only on Ω. Dividing both sides of the previous inequality by $\|y_u - y_v\|_{L^2(\Omega)}$ yields the desired result and the proof is complete. ■

Lemma 2.2.4 *Let $\mathcal{G}$ be the mapping introduced in* (2.9)*. Then for any $m > 0$ there exists $L(m) > 0$ such that*

$$\|\mathcal{G}(u_1) - \mathcal{G}(u_2)\|_{H^2(\Omega)} \le L(m)\|u_1 - u_2\|_{L^2(\Omega)}$$

for all u_1, $u_2 \in L^2(\Omega)$ with $\|u_1\|_{L^2(\Omega)}$, $\|u_2\|_{L^2(\Omega)} \le m$.

Proof: For a given number $m > 0$ choose u_1, $u_2 \in L^2(\Omega)$ with $\|u_1\|_{L^2(\Omega)}$, $\|u_2\|_{L^2(\Omega)} \le m$ and define $y_1 := \mathcal{G}(u_1)$ and $y_2 := \mathcal{G}(u_2)$. From Theorem 2.2.1 and the continuous embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ it follows that $\|y_1\|_{L^\infty(\Omega)}$, $\|y_2\|_{L^\infty(\Omega)} \le c_m$ for some $c_m > 0$ depending on m.

It is clear that the function $w := y_1 - y_2$ belongs to $H^2(\Omega)$ and it is a solution of

$$-\Delta w = f \quad \text{in } \Omega, \quad \text{and} \quad w = 0 \quad \text{on } \partial\Omega,$$

where $f := (u_1 - u_2) - [\phi(y_1) - \phi(y_2)]$. Notice that $f \in L^2(\Omega)$. Therefore, [40, Theorem 4.3.1.4] asserts the existence of a constant $c > 0$ such that

$$\|w\|_{H^2(\Omega)} \leq c\big(\|f\|_{L^2(\Omega)} + \|w\|_{L^2(\Omega)}\big).$$

We may now proceed while employing Lemma 2.2.3 and Lemma A.1.3 to obtain

$$\begin{aligned}\|y_1 - y_2\|_{H^2(\Omega)} &\leq c\big(\|u_1 - u_2\|_{L^2(\Omega)} + \|\phi(y_1) - \phi(y_2)\|_{L^2(\Omega)} + \|y_1 - y_2\|_{L^2(\Omega)}\big) \\ &\leq L(m)\big(\|u_1 - u_2\|_{L^2(\Omega)} + \|y_1 - y_2\|_{L^2(\Omega)}\big) \\ &\leq L(m)\|u_1 - u_2\|_{L^2(\Omega)},\end{aligned}$$

where $L(m) > 0$ is a constant depending on m. This completes the proof. ■

We would like to mention that the regularity of the state variable y plays an important role in the numerical analysis of the optimal control problem as we will see later. Therefore, it becomes an interesting task to investigate the regularity of the solution y to (2.3) if u is more regular than an $L^2(\Omega)$ function. For this reason, we derive the next result.

Theorem 2.2.5 *For a given $u \in W^{1,s}(\Omega)$ for some $1 < s < 2$, the solution y to (2.3) belongs to $W^{2,p}(\Omega) \cap H_0^1(\Omega)$ and there exists $c > 0$ such that*

$$\|y\|_{W^{2,p}(\Omega)} \leq c(\|u\|_{L^p(\Omega)} + 1), \tag{2.11}$$

where $p = \frac{2s}{2-s}$ for any $1 < s < s_\Omega := \min(2, \frac{2\theta_{\max}}{3\theta_{\max} - \pi})$ with $\theta_{\max} \in [\frac{\pi}{3}, \pi)$ being the maximum interior angle in Ω. For $\theta_{\max} = \frac{\pi}{3}$ we define $s_\Omega := 2$.

Proof: We first observe that for a given $u \in W^{1,s}(\Omega), 1 < s < 2$, the solution y to (2.3) belongs to $H^2(\Omega) \cap H_0^1(\Omega)$. This follows from Theorem 2.2.1 and the continuous embedding $W^{1,s}(\Omega) \hookrightarrow L^2(\Omega)$ for $1 < s < 2$.
Next, let $f := u - \phi(y)$ and consider the PDE

$$-\Delta w = f \text{ in } \Omega \quad \text{and} \quad w = 0 \text{ on } \partial\Omega. \tag{2.12}$$

Notice that $\phi(y) \in L^\infty(\Omega)$ from the continuous embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ and Lemma A.1.3. Moreover, $W^{1,s}(\Omega) \hookrightarrow L^p(\Omega)$ where $p = \frac{2s}{2-s}$ for $1 < s < 2$. This implies that $f \in L^p(\Omega)$ with $p = \frac{2s}{2-s}$ for $1 < s < 2$.
Consequently, we deduce from [40, Theorem 4.4.3.7] that there exists a unique $w \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ solution to (2.12) provided that

$$1 < s < s_\Omega := \min(2, \tfrac{2\theta_{\max}}{3\theta_{\max} - \pi})$$

where $\theta_{\max} \in [\frac{\pi}{3}, \pi)$ is the maximum interior angle in Ω. For $\theta_{\max} = \frac{\pi}{3}$ we define $s_\Omega := 2$. Furthermore, according to [40, Theorem 4.3.2.4], there exists $c > 0$ such that

$$\|w\|_{W^{2,p}(\Omega)} \leq c(\|\Delta w\|_{L^p(\Omega)} + \|w\|_{W^{1,p}(\Omega)}). \tag{2.13}$$

Since y is also the unique weak solution to (2.12) in $H_0^1(\Omega)$, we conclude that $w = y$ and thus $y \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$.
Finally, to obtain the estimate (2.11) we utilize (2.13), Lemma A.1.3, the continuous embedding $H^2(\Omega) \hookrightarrow W^{1,p}(\Omega)$ and (2.4) to get

$$\begin{aligned}\|y\|_{W^{2,p}(\Omega)} &\leq c\big(\|u\|_{L^p(\Omega)} + \|\phi(y) - \phi(0)\|_{L^p(\Omega)} + \|\phi(0)\|_{L^p(\Omega)} + \|y\|_{W^{1,p}(\Omega)}\big)\\ &\leq c\big(\|u\|_{L^p(\Omega)} + \|y\|_{L^p(\Omega)} + \|y\|_{H^2(\Omega)} + 1\big)\\ &\leq c\big(\|u\|_{L^p(\Omega)} + \|u\|_{L^2(\Omega)} + 1\big)\\ &\leq c\big(\|u\|_{L^p(\Omega)} + 1\big).\end{aligned}$$

Observe that $L^p(\Omega) \hookrightarrow L^2(\Omega)$ since $p = \frac{2s}{2-s} > 2$ for any $1 < s < 2$ and $\Omega \subset \mathbb{R}^2$ is bounded. This completes the proof. ■

2.3 The Optimal Control Problem (ℙ)

This section is devoted to the study of the optimal control problem (ℙ). In particular, we review the associated first order conditions and state our main result about global minima for the problem (ℙ). We also establish a condition that implies the second order sufficient conditions of problem (ℙ).

After we introduced the control-to-state operator $\mathcal{G}$ in (2.9), it will be convenient from now on to write $J(u)$ instead of $J(y,u)$. In this way, we can reformulate our optimal control problem as

$$(\mathbb{P}) \qquad \begin{array}{l}\min_{u\in U_{ad}} J(u) := \frac{1}{2}\|y - y_0\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)}\\ \text{subject to } y = \mathcal{G}(u) \text{ and } y|_K \in Y_{ad},\end{array}$$

where

$$\begin{aligned}U_{ad} &:= \{v \in L^2(\Omega) : u_a \leq v(x) \leq u_b \text{ a.e. in } \Omega\},\\ Y_{ad} &:= \{z \in C(K) : y_a(x) \leq z(x) \leq y_b(x) \text{ for all } x \in K\}.\end{aligned}$$

Notice that due to the embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ it becomes meaningful to impose pointwise constraints on the state variable y.

Definition 1 Let $F_{ad} := \{u \in U_{ad} : y|_K \in Y_{ad}, y := \mathcal{G}(u)\}$. Then a function $\bar{u} \in F_{ad}$ is called a *local minimum* (or a *local solution*) to the problem (ℙ) if there exists $\delta > 0$ such that

$$J(\bar{u}) \leq J(u) \quad \forall\, u \in F_{ad} \text{ and } \|\bar{u} - u\|_{L^2(\Omega)} \leq \delta.$$

If $\bar{u} \in F_{ad}$ satisfies

$$J(\bar{u}) \leq J(u) \quad \forall\, u \in F_{ad},$$

then $\bar{u}$ is called a *global minimum* (or a *global solution*) to the problem (ℙ).

Theorem 2.3.1 *Suppose that $F_{ad} := \{u \in U_{ad} : y|_K \in Y_{ad}, y := \mathcal{G}(u)\}$ is nonempty. Then the problem (ℙ) has at least one solution.*

Proof: The proof is classical, for instance, see the proof of [16, Theorem 5.1] for the details. ■

2.3.1 Necessary First Order Conditions for $(\mathbb{P})$

It is well known that to ensure the existence of *Lagrange multipliers* associated with a solution to a minimization problem in Banach spaces, one needs first to postulate some *constraint qualifications* at that solution, see for instance [78]. A typical constraint qualification for a local solution $\bar{u}$ of the problem $(\mathbb{P})$ is the *linearized Slater condition* which reads: there exist $u_0 \in U_{ad}$ and $\delta > 0$ such that

$$y_a(x) + \delta \leq \mathcal{G}(\bar{u})(x) + \mathcal{G}'(\bar{u})(u_0 - \bar{u})(x) \leq y_b(x) - \delta \qquad \forall\, x \in K. \tag{2.14}$$

The necessary first order optimality conditions for the problem $(\mathbb{P})$ are stated in the next result.

Theorem 2.3.2 *Let $\bar{u} \in U_{ad}$ be a local solution of the problem $(\mathbb{P})$ satisfying (2.14). Then there exist $\bar{p} \in W_0^{1,s}(\Omega)$ for $1 < s < 2$ and a regular Borel measure $\bar{\mu} \in \mathcal{M}(K)$ such that with $\bar{y} \in H_0^1(\Omega) \cap H^2(\Omega)$ there holds*

$$\int_\Omega \nabla \bar{y} \cdot \nabla v + \phi(\bar{y}) v \, dx = \int_\Omega \bar{u} v \, dx \quad \forall\, v \in H_0^1(\Omega), \qquad \bar{y}_{|K} \in Y_{ad}, \tag{2.15}$$

$$\begin{aligned} &\int_\Omega \bar{p}(-\Delta v) + \phi'(\bar{y})\bar{p} v \, dx \\ &\quad = \int_\Omega (\bar{y} - y_0) v \, dx + \int_K v \, d\bar{\mu} \quad \forall\, v \in H_0^1(\Omega) \cap H^2(\Omega), \end{aligned} \tag{2.16}$$

$$\int_\Omega (\bar{p} + \alpha \bar{u})(u - \bar{u}) \, dx \geq 0 \qquad \forall\, u \in U_{ad}, \tag{2.17}$$

$$\int_K (z - \bar{y}) \, d\bar{\mu} \leq 0 \qquad \forall\, z \in Y_{ad}. \tag{2.18}$$

Proof: The result follows from [16, Theorem 5.2] since the set Y_{ad} has a nonempty interior and the cost functional J is Gâteaux differentiable at $\bar{u}$. Here, the differentiability of J is a consequence of the chain rule and the differentiability of the control-to-state operator $\mathcal{G}$ as mentioned in Proposition 2.2.2. ■

It is worth pointing out that the variational inequality (2.17) implies a higher regularity for the optimal control $\bar{u}$ which in turn improves the regularity of the associated state $\bar{y}$. In fact, we have the following result.

Lemma 2.3.3 *Let $\bar{u}$ be a local solution of problem $(\mathbb{P})$ and let $\bar{y}$ be its associated state. Then $\bar{u} \in W^{1,s}(\Omega)$ for $1 < s < 2$ and $\bar{y} \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ such that*

$$\|\bar{y}\|_{W^{2,p}(\Omega)} \leq c(\|\bar{u}\|_{L^p(\Omega)} + 1), \tag{2.19}$$

for some $c > 0$, where $p = \frac{2s}{2-s}$ for $1 < s < s_\Omega := \min(2, \frac{2\theta_{\max}}{3\theta_{\max} - \pi})$ with $\theta_{\max} \in [\frac{\pi}{3}, \pi)$ being the maximum interior angle in Ω. For $\theta_{\max} = \frac{\pi}{3}$ we define $s_\Omega := 2$.

Proof: The result follows immediately from Theorem 2.2.5 if we show that the optimal control $\bar{u}$ admits the regularity $\bar{u} \in W^{1,s}(\Omega)$ for $1 < s < 2$. To achieve this, we recall that (2.17) is equivalent to

$$\bar{u}(x) = P_{U_{ad}}\big(-\tfrac{1}{\alpha}\bar{p}(x)\big) = \min\big(\max\big(u_a, -\tfrac{1}{\alpha}\bar{p}(x)\big), u_b\big) \quad \forall\, x \in \Omega,$$

where $P_{U_{ad}} : L^2(\Omega) \to U_{ad}$ is the L^2-projection into U_{ad}. Since $\bar{p} \in W_0^{1,s}(\Omega)$ for $1 < s < 2$ and $P_{U_{ad}}$ is a Lipschitz function, it follows from [51, Corollary A.6] that $\bar{u} \in W^{1,s}(\Omega)$ for $1 < s < 2$ as well. This completes the proof. ■

The next result states that the Lagrange multiplier $\bar{\mu}$ associated with the pointwise state constraints is concentrated at the set points in K where the state constraints are active. For a detailed proof of this result, see for instance [24].

Proposition 2.3.4 *Let $\bar{\mu} \in \mathcal{M}(K)$ and $\bar{y} \in C_0(\Omega)$ satisfy (2.18). Then there holds*

$$supp(\bar{\mu}_b) \subset \{x \in K : \bar{y}(x) = y_b(x)\},$$
$$supp(\bar{\mu}_a) \subset \{x \in K : \bar{y}(x) = y_a(x)\}.$$

where $\bar{\mu} = \bar{\mu}_b - \bar{\mu}_a$ with $\bar{\mu}_b, \bar{\mu}_a \geq 0$ is the Jordan decomposition of $\bar{\mu}$.

2.3.2 Global Minima for (ℙ)

Since the state equation is in general nonlinear, the optimal control problem (ℙ) is nonconvex and there may be several solutions of the necessary first order conditions (2.15)–(2.18). These can be examined further with the help of second order conditions but those will only give local information and usually do not allow a decision on whether the given point is a global minimum of (ℙ). Second order sufficient conditions for problem (ℙ) that are closest to the associated necessary ones can be found in [20]. In what follows, we provide a sufficient condition that help us to decide if a solution to (2.15)–(2.18) is a global minimum of (ℙ).

We begin by introducing the following constant:

$$\eta(\alpha, r) := \alpha^{\frac{\rho}{2}} C_q^{\frac{2-2r}{r}} M^{-1} \left(\frac{r-1}{2r-1} \right)^{\frac{1-r}{r}} q^{1/q} r^{1/r} \rho^{\rho/2} (2-\rho)^{\frac{\rho}{2}-1}. \tag{2.20}$$

Here, $q := \frac{3r-2}{r-1}$, $\rho := \frac{r+q}{rq}$, while M and r appear in (2.2). Furthermore, C_q is an upper bound on the optimal constant in the Gagliardo-Nirenberg inequality

$$\|f\|_{L^q} \leq C \|f\|_{L^2}^{\frac{2}{q}} \|\nabla f\|_{L^2}^{\frac{q-2}{q}} \quad (2 \leq q < \infty).$$

For our purposes it will be important to specify a constant C_q that is as sharp as possible. Theorem A.5.1 in Appendix A.5 will give three such bounds, two of which can be found in the literature, while the third is new to the best of our knowledge. Let us now formulate the main result of this section.

Theorem 2.3.5 *Suppose that $\bar{u} \in U_{ad}$, $\bar{y} \in H^2(\Omega) \cap H_0^1(\Omega)$, $\bar{p} \in W_0^{1,s}(\Omega)$ for $1 < s < 2$, $\bar{\mu} \in \mathcal{M}(K)$ is a solution of (2.15)–(2.18). If*

$$\|\bar{p}\|_{L^q(\Omega)} \leq \eta(\alpha, r), \tag{2.21}$$

then $\bar{u}$ is a global minimum for Problem (ℙ). If the inequality (2.21) is strict, then $\bar{u}$ is the unique global minimum. Here, $\eta(\alpha, r)$ and q are as defined in (2.20).

Proof: Let $u \in U_{ad}$ be a feasible control, $y = \mathcal{G}(u)$ the associated state with $y|_K \in Y_{ad}$. We have

$$J(u) - J(\bar{u}) = \frac{1}{2}\|y - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} + \alpha \int_\Omega \bar{u}(u - \bar{u})\,dx + \int_\Omega (\bar{y} - y_0)(y - \bar{y})\,dx =: (A)$$

Using $v := y - \bar{y}$ in (2.16) we get

$$\begin{aligned}(A) &= \frac{1}{2}\|y - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} + \alpha \int_\Omega \bar{u}(u - \bar{u})\,dx \\ &\quad + \int_\Omega \nabla \bar{p} \cdot \nabla(y - \bar{y}) + \phi'(\bar{y})\bar{p}(y - \bar{y})\,dx - \int_K (y - \bar{y})\,d\bar{\mu} \\ &\geq \frac{1}{2}\|y - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} + \alpha \int_\Omega \bar{u}(u - \bar{u})\,dx \\ &\quad + \int_\Omega \nabla \bar{p} \cdot \nabla(y - \bar{y}) + \phi'(\bar{y})\bar{p}(y - \bar{y})\,dx, \end{aligned} \tag{2.22}$$

by (2.18). Using (2.3) for y and $\bar{y}$ with test function $\bar{p}$ we get

$$\begin{aligned}\int_\Omega \nabla \bar{p} \cdot \nabla(y - \bar{y})\,dx &= \int_\Omega (u - \bar{u})\bar{p}\,dx - \int_\Omega (\phi(y) - \phi(\bar{y}))\bar{p}\,dx \\ &= \int_\Omega (u - \bar{u})\bar{p}\,dx \\ &\quad - \int_\Omega \bar{p}(y - \bar{y}) \int_0^1 \phi'(ty + (1 - t)\bar{y})\,dt\,dx.\end{aligned}$$

Using this in (2.22) and recalling (2.17) we arrive at

$$\begin{aligned}&J(u) - J(\bar{u}) \\ &\geq \frac{1}{2}\|y - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} + \int_\Omega (\alpha\bar{u} + \bar{p})(u - \bar{u})\,dx \\ &\quad - \int_\Omega \bar{p}(y - \bar{y}) \int_0^1 \phi'(ty + (1 - t)\bar{y}) - \phi'(\bar{y})\,dt\,dx \\ &\geq \frac{1}{2}\|y - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} - R(u), \end{aligned} \tag{2.23}$$

where

$$R(u) := \int_\Omega \bar{p}(y - \bar{y}) \int_0^1 \phi'(ty + (1 - t)\bar{y}) - \phi'(\bar{y})\,dt\,dx.$$

The aim is now to estimate $R(u)$. To begin, Lemma A.1.2 implies that

$$\begin{aligned}|R(u)| &\leq L_r \int_\Omega |\bar{p}||y - \bar{y}|^2 \bigg(\int_0^1 \phi'(ty + (1 - t)\bar{y})\,dt \bigg)^{\frac{1}{r}}\,dx \\ &= L_r \int_\Omega |\bar{p}||y - \bar{y}|^{\frac{2r-2}{r}} \bigg(\int_0^1 \phi'(ty + (1 - t)\bar{y})\,dt|y - \bar{y}|^2 \bigg)^{\frac{1}{r}}\,dx,\end{aligned}$$

where $L_r = M\Big(\frac{r-1}{2r-1}\Big)^{(r-1)/r}$. Next, Hölder's inequality with exponents

$$q = \frac{3r-2}{r-1}, \quad \frac{r(3r-2)}{2(r-1)^2} = \frac{qr}{2r-2} \text{ and } r$$

yields

$$\begin{aligned}|R(u)| &\le L_r \|\bar p\|_{L^q(\Omega)} \|y-\bar y\|_{L^q(\Omega)}^{\frac{2r-2}{r}} \\ &\quad \times \bigg(\int_\Omega \int_0^1 \phi'(ty+(1-t)\bar y)\,dt|y-\bar y|^2\,dx\bigg)^{\frac{1}{r}}.\end{aligned}$$

The Gagliardo–Nirenberg inequality $\|f\|_{L^q(\Omega)} \le C_q \|f\|_{L^2(\Omega)}^{\frac{2}{q}} \|\nabla f\|_{L^2(\Omega)}^{\frac{q-2}{q}}$, $f \in H_0^1(\Omega)$ (see Theorem A.5.1) together with the relation $\frac{2r-2}{r}(1-\frac{2}{q}) = \frac{2}{q}$ then implies

$$\begin{aligned}|R(u)| &\le L_r C_q^{\frac{2r-2}{r}} \|\bar p\|_{L^q(\Omega)} \|y-\bar y\|_{L^2(\Omega)}^{\frac{4r-4}{qr}} \|\nabla(y-\bar y)\|_{L^2(\Omega)}^{\frac{2}{q}} \\ &\quad \times \bigg(\int_\Omega \int_0^1 \phi'(ty+(1-t)\bar y)\,dt|y-\bar y|^2\,dx\bigg)^{\frac{1}{r}}.\end{aligned}$$

Applying Lemma A.4.2 with

$$\begin{aligned}&a := \int_\Omega |\nabla(y-\bar y)|^2\,dx, \quad b := \int_\Omega \int_0^1 \phi'(ty+(1-t)\bar y)\,dt|y-\bar y|^2\,dx, \\ &\lambda := \frac{1}{q}, \quad \mu := \frac{1}{r}\end{aligned}$$

we obtain

$$\begin{aligned}|R(u)| &\le L_r C_q^{\frac{2r-2}{r}} d_r \|\bar p\|_{L^q(\Omega)} \|y-\bar y\|_{L^2(\Omega)}^{\frac{4r-4}{qr}} \\ &\quad \times \bigg(\int_\Omega |\nabla(y-\bar y)|^2\,dx \\ &\quad + \int_\Omega \int_0^1 \phi'(ty+(1-t)\bar y)\,dt|y-\bar y|^2\,dx\bigg)^{\rho}, \end{aligned} \tag{2.24}$$

where

$$d_r = q^{-1/q} r^{-1/r} \rho^{-\rho}, \quad \rho = \frac{r+q}{rq}.$$

Using again (2.3) for y, $\bar y$, this time with test function $y-\bar y$ yields

$$\begin{aligned}&\int_\Omega |\nabla(y-\bar y)|^2\,dx + \int_\Omega \int_0^1 \phi'(ty+(1-t)\bar y)\,dt|y-\bar y|^2\,dx \\ &\quad \le \|u-\bar u\|_{L^2(\Omega)} \|y-\bar y\|_{L^2(\Omega)}.\end{aligned}$$

Inserting this estimate into (2.24) and observing that $\frac{4r-4}{qr} + \rho = 2-\rho$ we deduce

$$\begin{aligned}|R(u)| &\le L_r C_q^{\frac{2r-2}{r}} d_r \|\bar p\|_{L^q(\Omega)} \|y-\bar y\|_{L^2(\Omega)}^{2-\rho} \|u-\bar u\|_{L^2(\Omega)}^{\rho} \\ &= 2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r \|\bar p\|_{L^q(\Omega)} \bigg(\frac{1}{2}\|y-\bar y\|_{L^2(\Omega)}^2\bigg)^{1-\frac{\rho}{2}} \bigg(\frac{\alpha}{2}\|u-\bar u\|_{L^2(\Omega)}^2\bigg)^{\frac{\rho}{2}}.\end{aligned}$$

Applying again Lemma A.4.2, this time with the choices

$$a := \frac{1}{2}\|y-\bar{y}\|^2_{L^2(\Omega)}, \quad b := \frac{\alpha}{2}\|u-\bar{u}\|^2_{L^2(\Omega)}, \quad \lambda := 1-\frac{\rho}{2}, \quad \mu := \frac{\rho}{2},$$

we obtain

$$|R(u)| \leq 2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r e_r \|\bar{p}\|_{L^q(\Omega)} \left(\frac{1}{2}\|y-\bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u-\bar{u}\|^2_{L^2(\Omega)}\right), \quad (2.25)$$

where

$$e_r = \left(1-\frac{\rho}{2}\right)^{1-\frac{\rho}{2}} \left(\frac{\rho}{2}\right)^{\frac{\rho}{2}}.$$

Using (2.25) in (2.23) we get

$$\begin{aligned} &J(u) - J(\bar{u}) \\ &\geq \left(\frac{1}{2}\|y-\bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u-\bar{u}\|^2_{L^2(\Omega)}\right)\left(1 - 2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r e_r \|\bar{p}\|_{L^q(\Omega)}\right) \end{aligned}$$

so that $J(u) \geq J(\bar{u})$ provided that

$$\|\bar{p}\|_{L^q(\Omega)} \leq \left(2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r e_r\right)^{-1}. \quad (2.26)$$

By direct calculations, we have

$$2 d_r e_r = q^{-1/q} r^{-1/r} \rho^{-\rho/2} (2-\rho)^{1-\frac{\rho}{2}}.$$

Hence, using the above result and the value of L_r from Lemma A.1.2 we can rewrite (2.26) as

$$\|\bar{p}\|_{L^q(\Omega)} \leq \alpha^{\frac{\rho}{2}} C_q^{\frac{2-2r}{r}} M^{-1} \left(\frac{r-1}{2r-1}\right)^{\frac{1-r}{r}} q^{1/q} r^{1/r} \rho^{\rho/2} (2-\rho)^{\frac{\rho}{2}-1}$$

which is the desired result. ■

Remark 2 It is of interest to point out that Theorem 2.3.5 doesn't rely on the linearized Slater condtion (2.14). Basically, the theorem says that if there is a solution $(\bar{u}, \bar{y}, \bar{p}, \bar{\mu})$ to (2.15)–(2.18) such that (2.21) is satisfied, then the function $\bar{u}$ is a global minimum of the problem $(\mathbb{P})$.

2.3.3 Sufficient Second Order Conditions for $(\mathbb{P})$

In this section we show that the techniques used in the proof of Theorem 2.3.5 can also be used to establish a sufficient second order condition for a local minimum of problem $(\mathbb{P})$. To begin, we review briefly some material relevant to sufficient second order conditions.

We start by discussing the differentiability of the control-to-state operator. Recall that, according to Proposition 2.2.2, the mapping $\mathcal{G}$ introduced in (2.9) is of class C^1. However, establishing second order conditions requires the second derivative of $\mathcal{G}$. To guarantee that $\mathcal{G}$ is of class C^2 we require the following assumption on the nonlinearity ϕ:

- $\phi : \mathbb{R} \to \mathbb{R}$ is of class C^2 and for any $m > 0$ there exists $L(m) > 0$ such that

$$|\phi''(y_1) - \phi''(y_2)| \leq L(m)|y_1 - y_2| \quad \forall\, y_i \in \mathbb{R} \text{ with } |y_i| \leq m, \quad i = 1, 2.$$

Under the above assumption on ϕ we obtain the next result.

Proposition 2.3.6 *The mapping $\mathcal{G}$ introduced in (2.9) is of class C^2 and for every $u, v_1, v_2 \in L^2(\Omega)$ the second derivative $z := \mathcal{G}''(u)v_1v_2$ is the solution of*

$$\begin{aligned} -\Delta z + \phi'(y)z &= -\phi''(y)y_1y_2 \quad \text{in } \Omega, \\ z &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $y := \mathcal{G}(u)$ and $y_i := \mathcal{G}'(u)v_i$ for $i = 1, 2$.

Proof: The proof is analogous to that of Proposition 2.2.2. ■

To state the sufficient second order conditions for ($\mathbb{P}$), we first need to introduce the *Lagrange function* and the *cone of critical directions*.
The Lagrange function $\mathcal{L} : L^2(\Omega) \times (\mathcal{M}(K))^2 \to \mathbb{R}$ associated to the problem ($\mathbb{P}$) is defined by

$$\mathcal{L}(u, \mu_a, \mu_b) := J(u) + \int_K (y_a - \mathcal{G}(u))d\mu_a + \int_K (\mathcal{G}(u) - y_b)d\mu_b.$$

Notice that $\mathcal{L}$ is of class C^2 with respect to the first variable u. This is a consequence of the C^2 differentiability of $\mathcal{G}$ and the chain rule.
For any $\bar{u} \in U_{ad}$, $\bar{y} \in H^2(\Omega) \cap H_0^1(\Omega)$, $\bar{p} \in W_0^{1,s}(\Omega)$ for $1 < s < 2$, $\bar{\mu} \in \mathcal{M}(K)$ satisfying (2.15)–(2.18) the cone of critical directions is defined by

$$C_{\bar{u}} := \{h \in L^2(\Omega) : h \text{ satisfies (2.27)–(2.30)}\},$$

$$h(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = u_a, \\ \leq 0 & \text{if } \bar{u}(x) = u_b, \\ = 0 & \text{if } \bar{p} + \alpha\bar{u}(x) \neq 0, \end{cases} \tag{2.27}$$

$$z_h(x) \leq 0 \text{ if } \bar{y}(x) = y_b(x), \tag{2.28}$$

$$z_h(x) \geq 0 \text{ if } \bar{y}(x) = y_a(x), \tag{2.29}$$

$$\int_K |z_h(x)|\, d|\bar{\mu}|(x) = 0, \tag{2.30}$$

where $z_h := \mathcal{G}'(\bar{u})h$ and $|\bar{\mu}| = \bar{\mu}_b + \bar{\mu}_a$ such that $\bar{\mu} = \bar{\mu}_b - \bar{\mu}_a$.
The sufficient second order optimality conditions for problem ($\mathbb{P}$) are stated in the next result.

Theorem 2.3.7 *Suppose that $\bar{u} \in U_{ad}$, $\bar{y} \in H^2(\Omega) \cap H_0^1(\Omega)$, $\bar{p} \in W_0^{1,s}(\Omega)$ for $1 < s < 2$, $\bar{\mu} \in \mathcal{M}(K)$ satisfy (2.15)–(2.18). If*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2 > 0 \quad \forall\, h \in C_{\bar{u}} \setminus \{0\} \tag{2.31}$$

with $\bar{\mu} = \bar{\mu}_b - \bar{\mu}_a$, then there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$J(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} \leq J(u) \quad \forall\, u \in F_{ad} \text{ and } \|u - \bar{u}\|_{L^2(\Omega)} \leq \varepsilon,$$

where $F_{ad} := \{u \in U_{ad} : y|_K \in Y_{ad}, y := \mathcal{G}(u)\}$.

Proof: See [20, Theorem 4.3 & Section 5]. ■

In fact, it was mentioned in [20] that, under some regularity assumption, the inequality

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2 \geq 0 \quad \forall\, h \in C_{\bar{u}}$$

is expected to be a necessary condition for a local solution $\bar{u}$. This is at least the case when the state constraints are of integral type (see [22]) or when K is a finite set of points (see [19]).

We are now in a position to formulate a condition on $\bar{p}$ that serves as a sufficient second order condition for a local minimum of problem ($\mathbb{P}$).

Theorem 2.3.8 *Suppose that $\bar{u} \in U_{ad}$, $\bar{y} \in H^2(\Omega) \cap H^1_0(\Omega)$, $\bar{p} \in W^{1,s}_0(\Omega)$ for any $1 < s < 2$, $\bar{\mu} \in \mathcal{M}(K)$ is a solution of (2.15)–(2.18). If*

$$\|\bar{p}\|_{L^q(\Omega)} < 2\left(\frac{r-1}{2r-1}\right)^{\frac{r-1}{r}} \eta(\alpha, r), \tag{2.32}$$

then there exists $\delta > 0$ such that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2 \geq \delta\|v\|^2_{L^2(\Omega)} \qquad \text{for all } v \in L^2(\Omega),$$

where q and $\eta(\alpha, r)$ as defined in (2.20) and $\bar{\mu} = \bar{\mu}_b - \bar{\mu}_a$.

Proof: We divided the proof into two steps. In Step 1, we calculate the second derivative of the Lagrangian $\mathcal{L}$ with respect to u at $(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)$ in the arbitrarily chosen direction $v \in L^2(\Omega)$, where $\bar{\mu} = \bar{\mu}_b - \bar{\mu}_a$. In Step 2, we proceed by estimating the nonpositive terms in $\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2$.

Step 1:

By straightforward calculations, for any $v \in L^2(\Omega)$, we have

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2 = \int_\Omega y_v^2\, dx + \alpha \int_\Omega v^2\, dx + \int_\Omega (\bar{y} - y_0)y_{vv}\, dx + \int_K y_{vv}\, d(\bar{\mu}_b - \bar{\mu}_a),$$

where $\bar{y} = \mathcal{G}(\bar{u})$, $y_v := \mathcal{G}'(\bar{u})v$ and $y_{vv} := \mathcal{G}''(\bar{u})v^2$.

The task is now to rewrite the last two integrals in the previous identity. To this end, we test (2.16) by y_{vv} and use Proposition 2.3.6 to obtain

$$\begin{aligned}\int_\Omega (\bar{y} - y_0)y_{vv}\, dx + \int_K y_{vv}\, d(\bar{\mu}_b - \bar{\mu}_a) &= \int_\Omega \bar{p}(-\Delta y_{vv}) + \phi'(\bar{y})\bar{p}y_{vv}\, dx \\ &= -\int_\Omega \bar{p}y_v^2\phi''(\bar{y})\, dx.\end{aligned}$$

To summarize, for any $v \in L^2(\Omega)$, we have

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2 = \int_\Omega y_v^2 \, dx + \alpha \int_\Omega v^2 \, dx - \int_\Omega \bar{p} y_v^2 \phi''(\bar{y}) \, dx. \tag{2.33}$$

Step 2:
In view of (2.33) we only have to estimate the integral

$$R(v) := \int_\Omega \bar{p} y_v^2 \phi''(\bar{y}) dx.$$

Recall that, according to Proposition 2.2.2, $y_v \in H_0^1(\Omega)$ is the unique function satisfying

$$\int_\Omega \nabla y_v \cdot \nabla w + \phi'(\bar{y}) y_v w \, dx = \int_\Omega v w \, dx \quad \forall \, w \in H_0^1(\Omega). \tag{2.34}$$

We shall argue in a similar way to that used to estimate the term $R(u)$ in the proof of Theorem 2.3.5 and hence, we skip much of the details. In what follows, the constants ρ, d_r and e_r have the same value as in the proof of Theorem 2.3.5. To begin, (2.2), Hölder's inquality and the Gagliardo–Nirenberg inequality imply

$$\begin{aligned}
|R(v)| &\leq M \int_\Omega |\bar{p}| \, y_v^2 \phi'(\bar{y})^{\frac{1}{r}} dx \leq M \|\bar{p}\|_{L^q(\Omega)} \|y_v\|_{L^q(\Omega)}^{\frac{2r-2}{r}} \bigg(\int_\Omega \phi'(\bar{y}) y_v^2 dx \bigg)^{\frac{1}{r}} \\
&\leq M C_q^{\frac{2r-2}{r}} \|\bar{p}\|_{L^q(\Omega)} \|y_v\|_{L^2(\Omega)}^{\frac{4r-4}{qr}} \|\nabla y_v\|_{L^2(\Omega)}^{\frac{2}{q}} \bigg(\int_\Omega \phi'(\bar{y}) y_v^2 dx \bigg)^{\frac{1}{r}} \\
&\leq M C_q^{\frac{2r-2}{r}} d_r \|\bar{p}\|_{L^q(\Omega)} \|y_v\|_{L^2(\Omega)}^{\frac{4r-4}{qr}} \bigg(\int_\Omega |\nabla y_v|^2 dx + \int_\Omega \phi'(\bar{y}) y_v^2 dx \bigg)^{\rho}.
\end{aligned}$$

The last integral is estimated by testing (2.34) by y_v so that one obtains again analogous to the proof of Theorem 2.3.5

$$\begin{aligned}
|R(v)| &\leq M C_q^{\frac{2r-2}{r}} d_r \|\bar{p}\|_{L^q(\Omega)} \|y_v\|_{L^2(\Omega)}^{2-\rho} \|v\|_{L^2(\Omega)}^{\rho} \\
&= \alpha^{-\frac{\rho}{2}} M C_q^{\frac{2r-2}{r}} d_r \|\bar{p}\|_{L^q(\Omega)} \Big(\|y_v\|_{L^2(\Omega)}^2 \Big)^{1-\frac{\rho}{2}} \Big(\alpha \|v\|_{L^2(\Omega)}^2 \Big)^{\frac{\rho}{2}} \\
&\leq \alpha^{-\frac{\rho}{2}} M C_q^{\frac{2r-2}{r}} d_r e_r \|\bar{p}\|_{L^q(\Omega)} \Big(\|y_v\|_{L^2(\Omega)}^2 + \alpha \|v\|_{L^2(\Omega)}^2 \Big).
\end{aligned}$$

Combining this estimate with (2.33) we derive

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}_a, \bar{\mu}_b)v^2 \geq \Big(\|y_v\|_{L^2(\Omega)}^2 + \alpha \|v\|_{L^2(\Omega)}^2 \Big) \Big(1 - \alpha^{-\frac{\rho}{2}} M C_q^{\frac{2r-2}{r}} d_r e_r \|\bar{p}\|_{L^q(\Omega)} \Big)$$

and the result follows from (2.32) and the definition of $\eta(\alpha, r)$. ■

Remark 3 It is not difficult to see that $2\big(\frac{r-1}{2r-1}\big)^{\frac{r-1}{r}} > 1$ for $r > 1$ so that (2.32) is less restrictive than (2.21).

Chapter 3

Variational Discretization

This chapter is devoted to the discretization of the problem ($\mathbb{P}$) by the variational discretization concept introduced in [46]. We establish a sufficient condition for global minima of the resulting discrete problem ($\mathbb{P}_h$). We also carry out the convergence and error analysis of a sequence of discrete global minima of ($\mathbb{P}_h$) that satisfy our condition. We conclude by some numerical examples that verify our theoretical results.

The organization of this chapter is as follows: in Section 3.1, we introduce some finite element preliminaries that will be relevant in our study. In Section 3.2 we study the discretization of the state equation via continuous and piecewise linear finite elements. In Section 3.3 we consider the variational discretization of the control problem ($\mathbb{P}$) and establish a sufficient condition for global minima of the resulting discrete problem ($\mathbb{P}_h$). The convergence analysis of the discrete global minima of ($\mathbb{P}_h$) to those of ($\mathbb{P}$) is carried out in Section 3.4 while the associated error analysis is postponed to Section 3.6. In Section 3.5 we discuss some possible generalizations to the data of the control problem ($\mathbb{P}$). Finally, Section 3.7 is devoted for solving problem ($\mathbb{P}_h$) by the semismooth Newton's method while Section 3.8 contains the numerical verifications of our findings.

3.1 Finite Element Preliminaries

In this section we introduce some finite element preliminaries that will be relevant in our study of the discrete optimal control problem.

To begin, let $\{\mathcal{T}_h\}_{0<h\leq h_0}$ be a sequence of admissible triangulations of the polygonal domain $\Omega \subset \mathbb{R}^2$ such that for every $h > 0$ there holds

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

Here $h := \max_{T\in\mathcal{T}_h} \operatorname{diam}(T)$ is the maximum mesh size, where $\operatorname{diam}(T)$ stands for the diameter of the triangle T. We also assume that the sequence $\{\mathcal{T}_h\}_{0<h\leq h_0}$ is quasi-uniform in the sense that each $T \in \mathcal{T}_h$ is contained in a ball of radius $\gamma^{-1}h$ and contains a ball of radius γh for some $\gamma > 0$ independent of h. These assumptions on $\{\mathcal{T}_h\}_{0<h\leq h_0}$ should be valid throughout the whole chapter without further explicit mentioning unless otherwise stated.

On each triangulation $\mathcal{T}_h \in \{\mathcal{T}_h\}_{0<h\le h_0}$, we construct the spaces of linear finite elements:

$$X_h := \{v_h \in C(\bar{\Omega}) : v_h|_T \text{ is a linear polynomial on each } T \in \mathcal{T}_h\},$$
$$X_{h0} := \{v_h \in X_h : v_{h|\partial\Omega} = 0\}.$$

Functions from the space X_h satisfy the inverse estimate, see [13, Section 4.5],

$$\|v_h\|_{L^\infty(\Omega)} \le ch^{-1}\|v_h\|_{L^2(\Omega)} \quad \forall\, v_h \in X_h \tag{3.1}$$

and the discrete Sobolev inequality, see [13, Section 4.9],

$$\|v_h\|_{L^\infty(\Omega)} \le c(1+|\ln h|)^{\frac{1}{2}}\|v_h\|_{H^1(\Omega)} \quad \forall\, v_h \in X_h. \tag{3.2}$$

We define the Lagrange interpolation operator I_h by

$$I_h : C(\bar{\Omega}) \to X_h, \qquad I_h y := \sum_{i=1}^{n} y(x_i)\phi_i,$$

where $\{x_1, \ldots, x_n\}$ denote the nodes in the triangulation $\mathcal{T}_h$ and $\{\phi_1, \ldots, \phi_n\}$ the basis functions of the space X_h which satisfy $\phi_i(x_j) = \delta_{ij}$. Here δ_{ij} is the Kroneker delta function. The following estimates concerning the operator I_h can be found for instance in [13, Section 4.4].

$$\|y - I_h y\|_{L^\infty(\Omega)} \le ch^{2-\frac{2}{p}}\|y\|_{W^{2,p}(\Omega)} \quad \forall\, y \in W^{2,p}(\Omega),\ 1 < p \le \infty, \tag{3.3}$$
$$\|y - I_h y\|_{L^2(\Omega)} \le ch^2\|y\|_{H^2(\Omega)} \quad \forall\, y \in H^2(\Omega). \tag{3.4}$$

Let $R_h : H_0^1(\Omega) \to X_{h0}$ denote the Ritz projection defined by

$$\int_\Omega \nabla R_h y \cdot \nabla v_h dx = \int_\Omega \nabla y \cdot \nabla v_h dx \qquad \forall v_h \in X_{h0}. \tag{3.5}$$

Then, for any function $y \in W_0^{1,p}(\Omega) \cap W^{2,p}(\Omega)$ there holds (see [66])

$$\|y - R_h y\|_{W^{1,p}(\Omega)} \le ch\|y\|_{W^{2,p}(\Omega)}, \qquad 2 \le p \le \infty, \tag{3.6}$$
$$\|y - R_h y\|_{L^p(\Omega)} \le c_p h^2\|y\|_{W^{2,p}(\Omega)}, \quad 2 \le p < \infty. \tag{3.7}$$

Finally, we state the next lemma which requires the following assumption on Ω and its sequence of triangulations $\{\mathcal{T}_h\}_{0<h\le h_0}$.

Assumption 1 There is a convex polygonal domain $\tilde{\Omega}$ containing Ω, that is, $\Omega \subset \tilde{\Omega}$, such that for h_0 small enough each $\mathcal{T}_h \in \{\mathcal{T}_h\}_{0<h\le h_0}$ can be extended to a triangulation $\tilde{\mathcal{T}}_h$ of $\tilde{\Omega}$ such that the sequence $\{\tilde{\mathcal{T}}_h\}_{0<h\le h_0}$ is quasi-uniform with the same quasi-uniformity constant γ of $\{\mathcal{T}_h\}_{0<h\le h_0}$.

Lemma 3.1.1 *Suppose that Assumption 1 holds. Let $y \in H_0^1(\Omega) \cap C(\bar{\Omega})$ be given and let $y_h \in X_{h0}$ be the unique function satisfying*

$$\int_\Omega \nabla y_h \cdot \nabla v_h\, dx = \int_\Omega \nabla y \cdot \nabla v_h\, dx \quad \forall\, v_h \in X_{h0}.$$

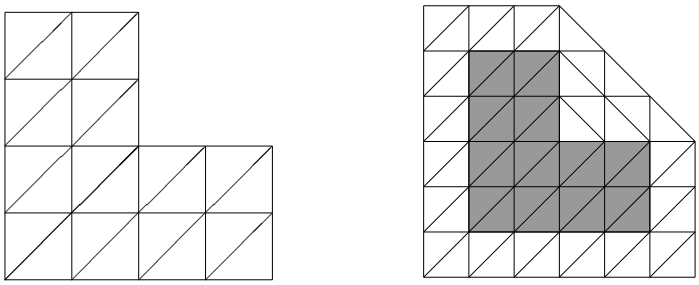
Then there holds

$$\|y - y_h\|_{L^\infty(\Omega)} \le c|\ln h| \inf_{\chi \in X_{h0}} \|y - \chi\|_{L^\infty(\Omega)}, \tag{3.8}$$

for some $c > 0$ independent of h.

Proof: See [69, Theorem 2]. ■

We point out that Lemma 3.1.1 is in general valid when Ω is a polygonal domain in $\mathbb{R}^2$ with maximal interior angle θ, $0 < \theta < 2\pi$. In Figure 3.1 we illustrate Assumption 1 when Ω is an L-shape domain.



(a) Ω and $\mathcal{T}_h$.

(b) $\tilde{\Omega}$ and $\tilde{\mathcal{T}}_h$. The inner L-shape highlighted region represents Ω.

Figure 3.1 Illustration of Assumption 1: for an L-shape domain Ω (left) together with its quasi-uniform triangulation $\mathcal{T}_h$ there exists a convex polygonal domain $\tilde{\Omega}$ (right) containing Ω and $\mathcal{T}_h$ can be extended to a triangulation $\tilde{\mathcal{T}}_h$ of $\tilde{\Omega}$ such that $\tilde{\mathcal{T}}_h$ is quasi-uniform with the same quasi-uniformity constant of $\mathcal{T}_h$.

3.2 The Discrete State Equation

In this section we discretize the state equation by means of continuous piecewise linear finite elements. We recall the relevant error estimates and improve the uniform convergence under certain conditions on the data. We also introduce the discrete control-to-state operator and discuss its differentiability.

We start by introducing the finite element discretization of (2.3) which reads: for a given $u \in L^2(\Omega)$, find $y_h \in X_{h0}$ such that

$$\int_\Omega \nabla y_h \cdot \nabla v_h + \phi(y_h) v_h \, dx = \int_\Omega u v_h \, dx \quad \forall \, v_h \in X_{h0}. \tag{3.9}$$

Theorem 3.2.1 *There exists a unique $y_h \in X_{h0}$ solution to* (3.9) *for a given $u \in L^2(\Omega)$.*

Proof: The result follows from the Brouwer fixed-point theorem and the monotonicity of ϕ. Compare the proof of [77, Theorem 26.A] for the details. ■

Analogously to (2.9), we introduce, for every $h > 0$, the mapping

$$\mathcal{G}_h : L^2(\Omega) \to X_{h0} \tag{3.10}$$

such that $y_h := \mathcal{G}_h(u)$ is the solution of (3.9) for a given $u \in L^2(\Omega)$. We sometimes call $\mathcal{G}_h$ the discrete control-to-state operator since it assigns to each control u a discrete state y_h.

Proposition 3.2.2 *The mapping $\mathcal{G}_h$ introduced in* (3.10) *is of class C^1 and for every $u, v \in L^2(\Omega)$ the first derivative $z_h := \mathcal{G}_h'(u)v \in X_{h0}$ is the unique function satisfying*

$$\int_\Omega \nabla z_h \cdot \nabla w_h + \phi'(y_h) z_h w_h \, dx = \int_\Omega v w_h \, dx \quad \forall\, w_h \in X_{h0},$$

where $y_h := \mathcal{G}_h(u)$.

Proof: The proof is along the lines of that of Proposition 2.2.2 if one considers the mapping $F : X_{h0} \times L^2(\Omega) \to X_{h0}^*$ defined by

$$F(y_h, u)\cdot = \int_\Omega \nabla y_h \cdot \nabla \cdot + \phi(y_h) \cdot \, dx - \int_\Omega u \cdot \, dx$$

where X_{h0}^* is the dual space of X_{h0}. ■

The next result shows the error in approximating the solution of (2.3) by the one of (3.9) in terms of the mesh size h.

Theorem 3.2.3 *For a given $u \in L^2(\Omega)$, let $y := \mathcal{G}(u)$ and let $y_h := \mathcal{G}_h(u)$, where $\mathcal{G}$ and $\mathcal{G}_h$ are as defined in* (2.9) *and* (3.10)*, respectively. Then, there exists $c > 0$ independent of h such that*

$$\|y - y_h\|_{L^2(\Omega)} \le c h^2 \big(\|u\|_{L^2(\Omega)} + 1\big), \tag{3.11}$$

$$\|y - y_h\|_{L^\infty(\Omega)} \le c h \big(\|u\|_{L^2(\Omega)} + 1\big). \tag{3.12}$$

Proof: The derivation of the estimate (3.11) can be found in [23, Theorem 2]. On the other hand, the estimate (3.12) can be deduced from (3.11) as follows:

$$\begin{aligned}
\|y - y_h\|_{L^\infty(\Omega)} &\le \|y - I_h y\|_{L^\infty(\Omega)} + \|I_h y - y_h\|_{L^\infty(\Omega)} \\
&\le c h \|y\|_{H^2(\Omega)} + c h^{-1} \|I_h y - y_h\|_{L^2(\Omega)} \\
&\le c h \|y\|_{H^2(\Omega)} + c h^{-1} \Big(\|I_h y - y\|_{L^2(\Omega)} + \|y - y_h\|_{L^2(\Omega)} \Big) \\
&\le c h \|y\|_{H^2(\Omega)} + c h^{-1} \Big(h^2 \|y\|_{H^2(\Omega)} + h^2 \big(\|u\|_{L^2(\Omega)} + 1\big) \Big) \\
&\le c h \big(\|u\|_{L^2(\Omega)} + 1\big),
\end{aligned}$$

where we used (3.1), (3.3), (3.4) and (2.4). ■

We remark that the finite element uniform convergence of the state equation plays a crucial rule in deriving error estimates for the numerical approximation of the problem ($\mathbb{P}$) if the pointwise state constraints are considered. For this reason, we establish the next theorem which asserts that for a better regularity of u and under certain conditions on Ω it is possible to improve the uniform estimate (3.12).

Theorem 3.2.4 *Suppose that Assumption 1 holds. For a given $u \in W^{1,s}(\Omega)$, for $1 < s < 2$, let $y := \mathcal{G}(u)$ and let $y_h := \mathcal{G}_h(u)$, where $\mathcal{G}$ and $\mathcal{G}_h$ are as defined in* (2.9) *and* (3.10)*, respectively. Then there exists $c > 0$ independent of h such that*

$$\|y - y_h\|_{L^\infty(\Omega)} \le c |\ln h| h^{2 - \frac{2}{p}} \big(\|u\|_{L^p(\Omega)} + 1\big), \tag{3.13}$$

where $p = \frac{2s}{2-s}$ *for* $1 < s < s_\Omega := \min(2, \frac{2\theta_{\max}}{3\theta_{\max}-\pi})$ *with* $\theta_{\max} \in [\frac{\pi}{3}, \pi)$ *being the maximum interior angle in* Ω. *For* $\theta_{\max} = \frac{\pi}{3}$ *we define* $s_\Omega := 2$.

Proof: We begin by considering

$$\|y - y_h\|_{L^\infty(\Omega)} \leq \underbrace{\|y - R_h y\|_{L^\infty(\Omega)}}_{\text{(I)}} + \underbrace{\|R_h y - y_h\|_{L^\infty(\Omega)}}_{\text{(II)}} \tag{3.14}$$

where $R_h y \in X_{h0}$ is the Ritz projection of y as introduced in (3.5).

To get an upper bound for the term (I), we apply Lemma 3.1.1, Theorem 2.2.5 and (3.3) to obtain

$$\begin{aligned}
\|y - R_h y\|_{L^\infty(\Omega)} &\leq c|\ln h| \inf_{\chi \in X_{h0}} \|y - \chi\|_{L^\infty(\Omega)} \\
&\leq c|\ln h| \|y - I_h y\|_{L^\infty(\Omega)} \\
&\leq c|\ln h| h^{2-\frac{2}{p}} \|y\|_{W^{2,p}(\Omega)} \\
&\leq c|\ln h| h^{2-\frac{2}{p}} (\|u\|_{L^p(\Omega)} + 1),
\end{aligned} \tag{3.15}$$

where p is as defined in Theorem 2.2.5.

To estimate (II), we first show that there exists $c > 0$ independent of h such that

$$\|R_h y - y_h\|_{H^1(\Omega)} \leq c\|y_h - y\|_{L^2(\Omega)}. \tag{3.16}$$

To achieve this, we observe that from the definitions of $R_h y$ and $y = \mathcal{G}(u)$ we have

$$\int_\Omega \nabla R_h y \cdot \nabla v_h \, dx = \int_\Omega u v_h \, dx - \int_\Omega \phi(y) v_h \, dx \quad \forall\, v_h \in X_{h0}.$$

Thus, subtracting $y_h = \mathcal{G}_h(u)$ from the previous equality and testing by $v_h := R_h y - y_h$ yields

$$\begin{aligned}
\int_\Omega |\nabla(R_h y - y_h)|^2 \, dx &= \int_\Omega [\phi(y_h) - \phi(y)](R_h y - y_h) \, dx \\
&\leq \|\phi(y_h) - \phi(y)\|_{L^2(\Omega)} \|R_h y - y_h\|_{L^2(\Omega)} \\
&\leq c\|y_h - y\|_{L^2(\Omega)} \|\nabla(R_h y - y_h)\|_{L^2(\Omega)},
\end{aligned} \tag{3.17}$$

where we used Lemma A.1.3 and the Poincaré's inequality. Notice that it follows from (3.12) that $\|y_h\|_{L^\infty(\Omega)}$ is uniformly bounded in h which implies that the constant c in (3.17) is independent of h. Dividing both sides of (3.17) by $\|\nabla(R_h y - y_h)\|_{L^2(\Omega)}$ and using again the Poincaré's inequality gives (3.16).
We are now in a position to estimate (II). For this purpose, we use (3.2), (3.16) and (3.11) to get

$$\begin{aligned}
\|R_h y - y_h\|_{L^\infty(\Omega)} &\leq c(1 + |\ln h|)^{\frac{1}{2}} \|R_h y - y_h\|_{H^1(\Omega)} \\
&\leq c(1 + |\ln h|)^{\frac{1}{2}} \|y_h - y\|_{L^2(\Omega)} \\
&\leq c(1 + |\ln h|)^{\frac{1}{2}} h^2 \big(\|u\|_{L^2(\Omega)} + 1\big).
\end{aligned} \tag{3.18}$$

Finally, combining (3.14), (3.15), (3.18) and observing that $L^p(\Omega) \hookrightarrow L^2(\Omega)$ and $(1 + |\ln h|)^{\frac{1}{2}} h^2 \leq |\ln h| h^{2-\frac{2}{p}}$ for h small enough gives the desired result and the proof is complete. ∎

Remark 4 The advantage of postulating Assumption 1 in Theorem 3.2.4 is to obtain the estimate (3.13) on the whole domain Ω. In other words, without this assumption we can only establish (3.13) on a subdomain of Ω. In the following steps, we summarise the main modifications that apply to the theorem and its proof if we drop this assumption.
Step 1. We consider the estimate from [70, Theorem 5.1], that is,

$$\|y - y_h\|_{L^\infty(\Omega_b)} \leq c\Big(|\ln h| \inf_{\chi \in X_{h0}} \|y - \chi\|_{L^\infty(\Omega_a)} + \|y - y_h\|_{L^2(\Omega_a)}\Big) \tag{3.19}$$

for some $\Omega_b \subset\subset \Omega_a \subset\subset \Omega$, where y and y_h are as defined in Lemma 3.1.1. We emphasise that (3.19) holds without requiring Assumption 1.
Step 2. We establish the inequality (3.14) on Ω_b instead of Ω. Then, we estimate the term (I) there using (3.19) instead of (3.8). The rest of the modifications in the proof are obvious.
Step 3. We may now drop Assumption 1 from the hypothesis of Theorem 3.2.4 and replace (3.13) by

$$\|y - y_h\|_{L^\infty(\Omega_b)} \leq c|\ln h| h^{2-\frac{2}{p}} \big(\|u\|_{L^p(\Omega)} + 1\big).$$

3.3 The Discrete Optimal Control Problem ($\mathbb{P}_h$)

In this section we consider the *variational discretization*, see [46], of Problem ($\mathbb{P}$). Then, we review the necessary first order conditions for the discrete control problem ($\mathbb{P}_h$). Finally, we derive an analogous result to Theorem 2.3.5 for the discrete problem ($\mathbb{P}_h$).

To begin, let us introduce, for $0 < h \leq h_0$, the following set of nodes:

$$\mathcal{N}_h := \{x_j \,|\, x_j \text{ is a vertex of } T \in \mathcal{T}_h, \text{ where } T \cap K \neq \emptyset\}.$$

We remark that $y_a(x_j) < y_b(x_j), x_j \in \mathcal{N}_h$ provided that h_0 is small enough. This follows from the fact that $\operatorname{dist}(x_j, K) \leq h, x_j \in \mathcal{N}_h$ and y_a, y_b are continuous functions with $y_a(x) < y_b(x), x \in K$.
The variational discretization of Problem ($\mathbb{P}$) now reads:

$$(\mathbb{P}_h) \qquad \begin{array}{c} \min_{u \in U_{ad}} J_h(u) := \frac{1}{2}\|y_h - y_0\|^2_{L^2(\Omega)} + \dfrac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \\ \text{subject to } y_h = \mathcal{G}_h(u),\ (y_h(x_j))_{x_j \in \mathcal{N}_h} \in Y^h_{ad}, \end{array}$$

where

$$Y^h_{ad} := \{(z_j)_{x_j \in \mathcal{N}_h} \,|\, y_a(x_j) \leq z_j \leq y_b(x_j), x_j \in \mathcal{N}_h\}.$$

We note that Problem ($\mathbb{P}_h$) is still an infinite dimensional optimization problem since the controls are sought in U_{ad}. Therefore, many of the techniques used in the analysis of ($\mathbb{P}$) can also be used for ($\mathbb{P}_h$). We have the next result whose proof is analogous to that of Theorem 2.3.1.

Theorem 3.3.1 *Suppose that there exists $u \in U_{ad}$ such that $(y_h(x_j))_{x_j \in \mathcal{N}_h} \in Y^h_{ad}$ where $y_h := \mathcal{G}_h(u)$. Then the problem ($\mathbb{P}_h$) has at least one solution.*

3.3.1 Necessary First Order Conditions for $(\mathbb{P}_h)$

To establish the necessary first order conditions of $(\mathbb{P}_h)$ at a local solution $\bar{u}_h$, we first assume the following linearized Slater condition: there exist $u_0 \in U_{ad}$ and $\delta > 0$ such that

$$y_a(x_j) + \delta \leq \mathcal{G}_h(\bar{u}_h)(x_j) + \mathcal{G}_h'(\bar{u}_h)(u_0 - \bar{u}_h)(x_j) \leq y_b(x_j) - \delta,\ x_j \in \mathcal{N}_h. \quad (3.20)$$

We state the optimality conditions of $(\mathbb{P}_h)$ in the next theorem.

Theorem 3.3.2 *Let $\bar{u}_h \in U_{ad}$ be a local solution of the problem $(\mathbb{P}_h)$ satisfying (3.20). Then there exist $\bar{p}_h \in X_{h0}$ and $\bar{\mu}_j \in \mathbb{R}, x_j \in \mathcal{N}_h$ such that with $\bar{y}_h \in X_{h0}$ there holds*

$$\int_\Omega \nabla \bar{y}_h \cdot \nabla v_h + \phi(\bar{y}_h) v_h \, dx = \int_\Omega \bar{u}_h v_h \, dx \quad \forall\, v_h \in X_{h0}, \qquad (\bar{y}_h(x_j))_{x_j \in \mathcal{N}_h} \in Y_{ad}^h \quad (3.21)$$

$$\int_\Omega \nabla \bar{p}_h \cdot \nabla v_h + \phi'(\bar{y}_h)\bar{p}_h v_h \, dx = \int_\Omega (\bar{y}_h - y_0) v_h \, dx + \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j v_h(x_j) \qquad \forall\, v_h \in X_{h0}, \quad (3.22)$$

$$\int_\Omega (\bar{p}_h + \alpha \bar{u}_h)(u - \bar{u}_h) \, dx \geq 0 \qquad \forall\, u \in U_{ad}, \quad (3.23)$$

$$\sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j (z_j - \bar{y}_h(x_j)) \leq 0 \qquad \forall\, (z_j)_{x_j \in \mathcal{N}_h} \in Y_{ad}^h. \quad (3.24)$$

Proof: The result follows from [16, Theorem 5.2] since the cost functional J_h is Gâteaux differentiable at $\bar{u}_h$ and Y_{ad}^h has a nonempty interior. The differentiability of J_h is deduced from that of the discrete control-to-state operator $\mathcal{G}_h$, according to Proposition 3.2.2, and the chain rule. ■

We note that the condition (3.23) is equivalent to the relation

$$\bar{u}_h(x) = P_{U_{ad}}\big(-\tfrac{1}{\alpha}\bar{p}_h(x)\big) = \min\big(\max\big(u_a, -\tfrac{1}{\alpha}\bar{p}_h(x)\big), u_b\big) \quad \forall\, x \in \Omega,$$

so that the control variable is implicitly discretized and (3.21)–(3.24) amounts to solving a nonlinear finite-dimensional system.

It will be convenient in the upcoming analysis to associate with the multipliers $(\bar{\mu}_j)_{x_j \in \mathcal{N}_h}$ from the optimality system (3.21)–(3.24) the measure $\bar{\mu}_h \in \mathcal{M}(\Omega)$ defined by

$$\bar{\mu}_h := \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j \delta_{x_j}, \quad (3.25)$$

where δ_{x_j} is the Dirac measure at x_j. We can easily deduce from (3.24) the following result about the support of the measure $\bar{\mu}_h$.

Proposition 3.3.3 *Let $\bar{\mu}_h \in \mathcal{M}(\Omega)$ be the measure introduced in (3.25) satisfying (3.24). Then there holds*

$$supp(\bar{\mu}_h^b) \subset \{x_j \in \mathcal{N}_h : \bar{y}_h(x_j) = y_b(x_j)\},$$
$$supp(\bar{\mu}_h^a) \subset \{x_j \in \mathcal{N}_h : \bar{y}_h(x_j) = y_a(x_j)\}.$$

where $\bar{\mu}_h = \bar{\mu}_h^b - \bar{\mu}_h^a$ *with* $\bar{\mu}_h^b, \bar{\mu}_h^a \geq 0$ *is the Jordan decomposition of* $\bar{\mu}_h$.

3.3.2 Global Minima for ($\mathbb{P}_h$)

We now derive an analogous result to Theorem 2.3.5 for the discrete control problem ($\mathbb{P}_h$).

Theorem 3.3.4 *Suppose that* $\bar{u}_h \in U_{ad}$, $\bar{y}_h \in X_{h0}$, $\bar{p}_h \in X_{h0}$, $(\bar{\mu}_j)_{x_j \in \mathcal{N}_h}$ *is a solution of* (3.21)–(3.24). *If*

$$\|\bar{p}_h\|_{L^q(\Omega)} \leq \eta(\alpha, r), \tag{3.26}$$

then $\bar{u}_h$ *is a global minimum for Problem* ($\mathbb{P}_h$). *If the inequality* (3.26) *is strict, then* $\bar{u}_h$ *is the unique global minimum. Here,* $\eta(\alpha, r)$ *and* q *are as defined in* (2.20).

Proof: The proof is obtained by arguing in almost the same way as in the proof of Theorem 2.3.5, that is, using the discrete counterpart of every continuous quantity there. To begin, let $u_h \in U_{ad}$ be a feasible control, $y_h = \mathcal{G}_h(u_h)$ the associated state with $(y_h(x_j))_{x_j \in \mathcal{N}_h} \in Y_{ad}^h$. We have

$$\begin{aligned} J_h(u_h) - J_h(\bar{u}_h) = \frac{1}{2}\|y_h - \bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h - \bar{u}_h\|^2_{L^2(\Omega)} + \alpha \int_\Omega \bar{u}_h(u_h - \bar{u}_h)\,dx \\ + \int_\Omega (\bar{y}_h - y_0)(y_h - \bar{y}_h)\,dx =: (A) \end{aligned}$$

Using $v_h := y_h - \bar{y}_h$ in (3.22) we get

$$\begin{aligned} (A) &= \frac{1}{2}\|y_h - \bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h - \bar{u}_h\|^2_{L^2(\Omega)} + \alpha \int_\Omega \bar{u}_h(u_h - \bar{u}_h)\,dx \\ &\quad + \int_\Omega \nabla \bar{p}_h \cdot \nabla(y_h - \bar{y}_h) + \phi'(\bar{y}_h)\bar{p}_h(y_h - \bar{y}_h)\,dx - \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j(y_h(x_j) - \bar{y}_h(x_j)) \\ &\geq \frac{1}{2}\|y_h - \bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h - \bar{u}_h\|^2_{L^2(\Omega)} + \alpha \int_\Omega \bar{u}_h(u_h - \bar{u}_h)\,dx \\ &\quad + \int_\Omega \nabla \bar{p}_h \cdot \nabla(y_h - \bar{y}_h) + \phi'(\bar{y}_h)\bar{p}_h(y_h - \bar{y}_h)\,dx, \end{aligned} \tag{3.27}$$

by (3.24). Using (3.9) for y_h and $\bar{y}_h$ with test function $\bar{p}_h$ we get

$$\begin{aligned} \int_\Omega \nabla \bar{p}_h \cdot \nabla(y_h - \bar{y}_h)\,dx &= \int_\Omega (u_h - \bar{u}_h)\bar{p}_h\,dx - \int_\Omega (\phi(y_h) - \phi(\bar{y}_h))\bar{p}_h\,dx \\ &= \int_\Omega (u_h - \bar{u}_h)\bar{p}_h\,dx \\ &\quad - \int_\Omega \bar{p}_h(y_h - \bar{y}_h) \int_0^1 \phi'(ty_h + (1-t)\bar{y}_h)\,dt\,dx. \end{aligned}$$

Using this in (3.27) and recalling (3.23) we arrive at

$$\begin{aligned}
&J_h(u_h) - J_h(\bar{u}_h) \\
&\quad \geq \frac{1}{2}\|y_h - \bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h - \bar{u}_h\|^2_{L^2(\Omega)} + \int_\Omega (\alpha\bar{u}_h + \bar{p}_h)(u_h - \bar{u}_h)\,dx \\
&\qquad - \int_\Omega \bar{p}_h(y_h - \bar{y}_h)\int_0^1 \phi'(ty_h + (1-t)\bar{y}_h) - \phi'(\bar{y}_h)\,dt\,dx \\
&\quad \geq \frac{1}{2}\|y_h - \bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u_h - \bar{u}_h\|^2_{L^2(\Omega)} - R_h(u_h), \qquad (3.28)
\end{aligned}$$

where

$$R_h(u_h) := \int_\Omega \bar{p}_h(y_h - \bar{y}_h)\int_0^1 \phi'(ty_h + (1-t)\bar{y}_h) - \phi'(\bar{y}_h)\,dt\,dx.$$

The aim is now to estimate $R_h(u_h)$. To begin, Lemma A.1.2 implies that

$$\begin{aligned}
|R_h(u_h)| &\leq L_r \int_\Omega |\bar{p}_h||y_h - \bar{y}_h|^2 \Big(\int_0^1 \phi'(ty_h + (1-t)\bar{y}_h)\,dt\Big)^{\frac{1}{r}}\,dx \\
&= L_r \int_\Omega |\bar{p}_h||y_h - \bar{y}_h|^{\frac{2r-2}{r}} \Big(\int_0^1 \phi'(ty_h + (1-t)\bar{y}_h)\,dt|y_h - \bar{y}_h|^2\Big)^{\frac{1}{r}}\,dx,
\end{aligned}$$

where $L_r = M\Big(\frac{r-1}{2r-1}\Big)^{(r-1)/r}$. Next, Hölder's inequality with exponents

$$q = \frac{3r-2}{r-1}, \quad \frac{r(3r-2)}{2(r-1)^2} = \frac{qr}{2r-2} \text{ and } r$$

yields

$$\begin{aligned}
|R_h(u_h)| &\leq L_r \|\bar{p}_h\|_{L^q(\Omega)} \|y_h - \bar{y}_h\|^{\frac{2r-2}{r}}_{L^q(\Omega)} \\
&\qquad \times \Big(\int_\Omega \int_0^1 \phi'(ty_h + (1-t)\bar{y}_h)\,dt|y_h - \bar{y}_h|^2\,dx\Big)^{\frac{1}{r}}.
\end{aligned}$$

The Gagliardo–Nirenberg inequality $\|f\|_{L^q(\Omega)} \leq C_q \|f\|^{\frac{2}{q}}_{L^2(\Omega)} \|\nabla f\|^{\frac{q-2}{q}}_{L^2(\Omega)}$, $f \in H^1_0(\Omega)$ (see Theorem A.5.1) together with the relation $\frac{2r-2}{r}(1 - \frac{2}{q}) = \frac{2}{q}$ then implies

$$\begin{aligned}
|R_h(u_h)| &\leq L_r C_q^{\frac{2r-2}{r}} \|\bar{p}_h\|_{L^q(\Omega)} \|y_h - \bar{y}_h\|^{\frac{4r-4}{qr}}_{L^2(\Omega)} \|\nabla(y_h - \bar{y}_h)\|^{\frac{2}{q}}_{L^2(\Omega)} \\
&\qquad \times \Big(\int_\Omega \int_0^1 \phi'(ty_h + (1-t)\bar{y}_h)\,dt|y_h - \bar{y}_h|^2\,dx\Big)^{\frac{1}{r}}.
\end{aligned}$$

Applying Lemma A.4.2 with

$$\begin{aligned}
&a := \int_\Omega |\nabla(y_h - \bar{y}_h)|^2\,dx, \quad b := \int_\Omega \int_0^1 \phi'(ty_h + (1-t)\bar{y}_h)\,dt|y_h - \bar{y}_h|^2\,dx, \\
&\lambda := \frac{1}{q}, \quad \mu := \frac{1}{r}
\end{aligned}$$

we obtain

$$\begin{aligned}|R_h(u_h)| &\le L_r C_q^{\frac{2r-2}{r}} d_r \|\bar p_h\|_{L^q(\Omega)} \|y_h - \bar y_h\|_{L^2(\Omega)}^{\frac{4r-4}{qr}} \\ &\quad \times \left(\int_\Omega |\nabla(y_h - \bar y_h)|^2 \, dx \right. \\ &\quad \left. + \int_\Omega \int_0^1 \phi'(t y_h + (1-t)\bar y_h) \, dt |y_h - \bar y_h|^2 \, dx \right)^\rho, \end{aligned} \tag{3.29}$$

where

$$d_r = q^{-1/q} r^{-1/r} \rho^{-\rho}, \quad \rho = \frac{r+q}{rq}.$$

Using again (3.9) for $y_h, \bar y_h$, this time with test function $y_h - \bar y_h$ yields

$$\begin{aligned}&\int_\Omega |\nabla(y_h - \bar y_h)|^2 \, dx + \int_\Omega \int_0^1 \phi'(t y_h + (1-t)\bar y_h) \, dt |y_h - \bar y_h|^2 \, dx \\ &\quad \le \|u_h - \bar u_h\|_{L^2(\Omega)} \|y_h - \bar y_h\|_{L^2(\Omega)}.\end{aligned}$$

Inserting this estimate into (3.29) and observing that $\frac{4r-4}{qr} + \rho = 2 - \rho$ we deduce

$$\begin{aligned}|R_h(u_h)| &\le L_r C_q^{\frac{2r-2}{r}} d_r \|\bar p_h\|_{L^q(\Omega)} \|y_h - \bar y_h\|_{L^2(\Omega)}^{2-\rho} \|u_h - \bar u_h\|_{L^2(\Omega)}^{\rho} \\ &= 2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r \|\bar p_h\|_{L^q(\Omega)} \left(\frac{1}{2} \|y_h - \bar y_h\|_{L^2(\Omega)}^2 \right)^{1-\frac{\rho}{2}} \left(\frac{\alpha}{2} \|u_h - \bar u_h\|_{L^2(\Omega)}^2 \right)^{\frac{\rho}{2}}.\end{aligned}$$

Applying again Lemma A.4.2, this time with the choices

$$a := \frac{1}{2} \|y_h - \bar y_h\|_{L^2(\Omega)}^2, \quad b := \frac{\alpha}{2} \|u_h - \bar u_h\|_{L^2(\Omega)}^2, \quad \lambda := 1 - \frac{\rho}{2}, \quad \mu := \frac{\rho}{2},$$

we obtain

$$|R_h(u_h)| \le 2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r e_r \|\bar p_h\|_{L^q(\Omega)} \left(\frac{1}{2} \|y_h - \bar y_h\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u_h - \bar u_h\|_{L^2(\Omega)}^2 \right), \tag{3.30}$$

where

$$e_r = \left(1 - \frac{\rho}{2}\right)^{1-\frac{\rho}{2}} \left(\frac{\rho}{2}\right)^{\frac{\rho}{2}}.$$

Using (3.30) in (3.28) we get

$$\begin{aligned}&J_h(u_h) - J_h(\bar u_h) \\ &\ge \left(\frac{1}{2} \|y_h - \bar y_h\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u_h - \bar u_h\|_{L^2(\Omega)}^2 \right) \left(1 - 2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r e_r \|\bar p_h\|_{L^q(\Omega)} \right)\end{aligned}$$

so that $J_h(u_h) \ge J_h(\bar u_h)$ provided that

$$\|\bar p_h\|_{L^q(\Omega)} \le \left(2\alpha^{-\frac{\rho}{2}} L_r C_q^{\frac{2r-2}{r}} d_r e_r \right)^{-1}. \tag{3.31}$$

By direct calculations, we have

$$2 d_r e_r = q^{-1/q} r^{-1/r} \rho^{-\rho/2} (2-\rho)^{1-\frac{\rho}{2}}.$$

Hence, using the above result and the value of L_r from Lemma A.1.2 we can rewrite (3.31) as

$$\|\bar{p}_h\|_{L^q(\Omega)} \leq \alpha^{\frac{\rho}{2}} C_q^{\frac{2-2r}{r}} M^{-1} \left(\frac{r-1}{2r-1}\right)^{\frac{1-r}{r}} q^{1/q} r^{1/r} \rho^{\rho/2} (2-\rho)^{\frac{\rho}{2}-1}$$

which is the desired result. ∎

Remark 5 We point out that we don't need the triangulation $\mathcal{T}_h$ of $\bar{\Omega}$ to be quasi-uniform in order to derive Theorem 3.3.4.

Remark 6 Notice that Theorem 3.3.4 doesn't require the linearized Slater condition (3.20). All what the theorem says is that if there exists a solution $(\bar{u}_h, \bar{y}_h, \bar{p}_h, \bar{\mu}_h)$ to (3.21)–(3.24) such that (3.26) is satisfied, then the function $\bar{u}_h$ is a global minimum of the problem $(\mathbb{P}_h)$. In practice this means that we first set up the system (3.21)–(3.24) and then try to solve it. If it has a solution such that (3.26) holds, then we have a global minimum of the problem $(\mathbb{P}_h)$ at hand.

3.4 Convergence Analysis

Since the quantities $\eta(\alpha, r)$ and $\|\bar{p}_h\|_{L^q(\Omega)}$ can be computed *explicitly*, Theorem 3.3.4 allows us to decide if a function $\bar{u}_h$ obtained from solving (3.21)–(3.24) is a global minimum of $(\mathbb{P}_h)$. A natural question which arises then is whether a sequence $(\bar{u}_h)_{0<h\leq h_0}$ of minima satisfying (3.26) uniformly in h converges to a global minimum of $(\mathbb{P})$ as the discretization parameter h tends to zero. We address this problem in this section.

To begin, let $\bar{u}_h \in U_{ad}$, $\bar{y}_h \in X_{h0}$, $\bar{p}_h \in X_{h0}$, $(\bar{\mu}_j)_{x_j \in \mathcal{N}_h}$ satisfy (3.21)–(3.24) as well as

$$\|\bar{p}_h\|_{L^q(\Omega)} \leq \eta(\alpha, r), \quad 0 < h \leq h_0. \tag{3.32}$$

As we mentioned earlier, it is convenient to introduce the measure $\bar{\mu}_h \in \mathcal{M}(\Omega)$ by

$$\bar{\mu}_h := \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j \delta_{x_j}.$$

Since $K \subset \Omega$, $\operatorname{dist}(x_j, K) \leq h$, $x_j \in \mathcal{N}_h$ and $y_a(x) < y_b(x)$, $x \in K$ there exists a compact set $\tilde{K} \subset \Omega$, $\delta > 0$ and $0 < h_1 \leq h_0$ such that $K \subset \tilde{K}$ and

$$\begin{aligned} y_a(x) &< y_b(x), \qquad x \in \tilde{K}, \\ \mathcal{N}_h &\subset \tilde{K}, \qquad 0 < h \leq h_1, \\ y_a(x) + \delta \leq \frac{1}{2}(y_a(x) + y_b(x)) &\leq y_b(x) - \delta, \qquad x \in \tilde{K}. \end{aligned}$$

For the existence of such a compact set $\tilde{K}$, see for instance [68, Theorem 2.7]. Since $C_0(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in $C(\bar{\Omega})$, there exists a function $w \in C_0^\infty(\Omega)$ approximating $\frac{1}{2}(y_a + y_b) \in C_0(\Omega)$ uniformly such that

$$y_a(x) + \frac{\delta}{2} \leq w(x) \leq y_b(x) - \frac{\delta}{2}, \quad x \in \tilde{K}. \tag{3.33}$$

Let $R_h w$ denote the Ritz projection of w as introduced in (3.5). Then it follows from the continuous embedding $W^{1,p}(\Omega) \hookrightarrow C(\bar{\Omega})$, $2 < p < \infty$, and the estimate (3.6) that

$$\|w - R_h w\|_{C(\bar{\Omega})} \leq c\|w - R_h w\|_{W^{1,p}(\Omega)} \leq ch\|w\|_{W^{2,p}(\Omega)}.$$

From the previous uniform convergence we may assume after choosing h_1 smaller if necessary that

$$y_a(x) + \frac{\delta}{4} \leq R_h w(x) \leq y_b(x) - \frac{\delta}{4}, \qquad x \in \tilde{K}. \tag{3.34}$$

Our first step in the convergence analysis are uniform bounds on the optimal control $\bar{u}_h$ as well as on its state $\bar{y}_h$ and $\bar{\mu}_h$.

Lemma 3.4.1 *Let $\bar{u}_h \in U_{ad}$, $\bar{y}_h$, $\bar{p}_h \in X_{h0}$ and $(\bar{\mu}_j)_{x_j \in \mathcal{N}_h}$ be a solution of* (3.21)–(3.24) *satisfying* (3.32). *Then there exists a constant $C > 0$, which is independent of h, such that*

$$\|\bar{u}_h\|_{L^2(\Omega)},\ \|\bar{y}_h\|_{H^1(\Omega)},\ \|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})} \leq C.$$

Proof: To begin, fix a function $u_0 \in U_{ad}$. Inserting u_0 into (3.23) we infer

$$\begin{aligned}
\alpha\|\bar{u}_h\|^2_{L^2(\Omega)} &\leq \int_\Omega u_0(\alpha\bar{u}_h + \bar{p}_h)dx - \int_\Omega \bar{u}_h\bar{p}_h dx \\
&\leq \|u_0\|_{L^2(\Omega)}\big(\alpha\|\bar{u}_h\|_{L^2(\Omega)} + \|\bar{p}_h\|_{L^2(\Omega)}\big) + \|\bar{u}_h\|_{L^2(\Omega)}\|\bar{p}_h\|_{L^2(\Omega)} \\
&\leq \frac{\alpha}{2}\Big(\|u_0\|_{L^2(\Omega)} + \frac{1}{\alpha}\|\bar{p}_h\|_{L^2(\Omega)}\Big)^2 + \frac{\alpha}{2}\|\bar{u}_h\|^2_{L^2(\Omega)} + \|u_0\|_{L^2(\Omega)}\|\bar{p}_h\|_{L^2(\Omega)}.
\end{aligned}$$

Since $q = \frac{3r-2}{r-1} \geq 3$ we deduce with the help of (3.32)

$$\|\bar{u}_h\|_{L^2(\Omega)} \leq C\big(\|u_0\|_{L^2(\Omega)} + \|\bar{p}_h\|_{L^2(\Omega)}\big) \leq C\big(\|u_0\|_{L^2(\Omega)} + \|\bar{p}_h\|_{L^q(\Omega)}\big) \leq C.$$

Testing (3.21) with $\bar{y}_h$, using the monotonicity of ϕ and Poincaré's inequality in a similar way to that of deriving (2.6) we infer

$$\|\bar{y}_h\|_{H^1(\Omega)} \leq C\big(1 + \|\bar{u}_h\|_{L^2(\Omega)}\big) \leq C. \tag{3.35}$$

Furthermore, (A.1), (A.2) along with the continuous embedding $H^1(\Omega) \hookrightarrow L^t(\Omega)$ for all $1 \leq t < \infty$ yield

$$\|\phi(\bar{y}_h)\|_{L^2(\Omega)},\ \|\phi'(\bar{y}_h)\|_{L^2(\Omega)} \leq C. \tag{3.36}$$

In order to verify the uniform boundedness of $\|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})}$ we first observe that (3.24) implies

$$\bar{y}_h(x_j) = \begin{cases} y_b(x_j), & \text{if } \bar{\mu}_j > 0, \\ y_a(x_j), & \text{if } \bar{\mu}_j < 0. \end{cases}$$

As a result we deduce with the help of (3.34)

$$\begin{aligned}\frac{\delta}{4}\|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})} &= \frac{\delta}{4}\sum_{x_j\in\mathcal{N}_h}|\bar{\mu}_j| \\ &= \frac{\delta}{4}\sum_{x_j\in\mathcal{N}_h:\bar{\mu}_j>0}\bar{\mu}_j + \frac{\delta}{4}\sum_{x_j\in\mathcal{N}_h:\bar{\mu}_j<0}-\bar{\mu}_j \\ &\le \frac{\delta}{4}\sum_{x_j\in\mathcal{N}_h:\bar{\mu}_j>0}\bar{\mu}_j\frac{4}{\delta}\big(y_b(x_j)-R_hw(x_j)\big) \\ &\quad + \frac{\delta}{4}\sum_{x_j\in\mathcal{N}_h:\bar{\mu}_j<0}\bar{\mu}_j\frac{4}{\delta}\big(y_a(x_j)-R_hw(x_j)\big) \\ &= \sum_{x_j\in\mathcal{N}_h}\bar{\mu}_j\big(\bar{y}_h(x_j)-R_hw(x_j)\big).\end{aligned}$$

Using $v_h = \bar{y}_h - R_hw$ in (3.22) we may continue

$$\begin{aligned}\frac{\delta}{4}\|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})} &\le \int_\Omega \nabla\bar{p}_h\cdot\nabla\bar{y}_h\,dx - \int_\Omega\nabla\bar{p}_h\cdot\nabla R_hw\,dx \\ &\quad + \int_\Omega\phi'(\bar{y}_h)\bar{p}_h(\bar{y}_h-R_hw)\,dx - \int_\Omega(\bar{y}_h-y_0)(\bar{y}_h-R_hw)\,dx \\ &\equiv \sum_{i=1}^4 S_i. \end{aligned}\tag{3.37}$$

If we let $v_h = \bar{p}_h$ in (3.21) we obtain with the help of (3.36) and (3.32)

$$|S_1| = \left|\int_\Omega(\bar{u}_h-\phi(\bar{y}_h))\bar{p}_h dx\right| \le \big(\|\bar{u}_h\|_{L^2(\Omega)}+\|\phi(\bar{y}_h)\|_{L^2(\Omega)}\big)\|\bar{p}_h\|_{L^2(\Omega)} \le C.$$

Next, the definition of the Ritz projection and integration by parts yields

$$S_2 = -\int_\Omega\nabla\bar{p}_h\cdot\nabla w\,dx = \int_\Omega\bar{p}_h\Delta w\,dx$$

so that

$$|S_2| \le \|\bar{p}_h\|_{L^2(\Omega)}\|\Delta w\|_{L^2(\Omega)} \le C.$$

Hölder's inequality along with (3.36), (3.32) and (3.35) implies that

$$|S_3| \le \|\phi'(\bar{y}_h)\|_{L^2(\Omega)}\|\bar{p}_h\|_{L^q(\Omega)}\|\bar{y}_h-R_hw\|_{L^{\frac{2q}{q-2}}(\Omega)} \le C\|\bar{y}_h-R_hw\|_{H^1(\Omega)} \le C.$$

Notice that the uniform boundedness of $\|R_hw\|_{H^1(\Omega)}$ can be seen from (3.6). Finally,

$$|S_4| \le \big(\|\bar{y}_h\|_{L^2(\Omega)}+\|y_0\|_{L^2(\Omega)}\big)\big(\|\bar{y}_h\|_{L^2(\Omega)}+\|R_hw\|_{L^2(\Omega)}\big) \le C.$$

Inserting the above estimates into (3.37) yields the bound on $\|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})}$. ■

We are now in a position to formulate the main theorem in this section:

Theorem 3.4.2 *Suppose that* $(\bar{u}_h, \bar{y}_h, \bar{p}_h, \bar{\mu}_h)_{0<h\le h_1}$ *is a sequence satisfying* (3.21)–(3.24) *as well as* (3.32). *Then*

$$\bar{u}_h \to \bar{u} \text{ in } L^2(\Omega) \text{ for a subsequence } h \to 0,$$

where $\bar{u}$ *is a global minimum for Problem* $(\mathbb{P})$. *If*

$$\|\bar{p}_h\|_{L^q(\Omega)} \le \kappa\eta(\alpha, r), \quad 0 < h \le h_1, \tag{3.38}$$

for some $0 < \kappa < 1$, *then* $\bar{u}$ *is the unique global solution of* $(\mathbb{P})$ *and the whole sequence* $(\bar{u}_h)_{0<h\le h_1}$ *converges to* $\bar{u}$.

Proof: From Lemma 3.4.1, we deduce the existence of a subsequence $h \to 0$ and $\bar{u} \in L^2(\Omega)$, $\bar{y} \in H_0^1(\Omega)$, $\bar{p} \in L^q(\Omega)$, $\bar{\mu} \in \mathcal{M}(\tilde{K})$ such that

$$\bar{u}_h \rightharpoonup \bar{u} \quad \text{in } L^2(\Omega), \tag{3.39}$$

$$\bar{y}_h \rightharpoonup \bar{y} \quad \text{in } H_0^1(\Omega), \tag{3.40}$$

$$\bar{p}_h \rightharpoonup \bar{p} \quad \text{in } L^q(\Omega), \tag{3.41}$$

$$\bar{\mu}_h \rightharpoonup \bar{\mu} \quad \text{in } \mathcal{M}(\tilde{K}). \tag{3.42}$$

Our aim is to show that $(\bar{u}, \bar{y}, \bar{p}, \bar{\mu})$ is a solution of (2.15)–(2.18). Firstly, since U_{ad} is closed and convex, it is weakly sequentially closed and thus $\bar{u} \in U_{ad}$. To show that $\bar{y} = \mathcal{G}(\bar{u})$, consider the Ritz projection $R_h v \in X_{h0}$ for any $v \in H_0^1(\Omega)$. Then, testing (3.21) by $R_h v$ gives

$$\int_\Omega \nabla \bar{y}_h \cdot \nabla R_h v + \phi(\bar{y}_h) R_h v \, dx = \int_\Omega \bar{u}_h R_h v \, dx. \tag{3.43}$$

Using (3.39), (3.40), the convergence $R_h v \to v$ in $H_0^1(\Omega)$ from (3.6), (A.1) together with Lemma A.1.4 we can pass to the limit $h \to 0$ in (3.43) to obtain $\bar{y} = \mathcal{G}(\bar{u})$.
Next, the fact that $\operatorname{dist}(x_j, K) \le h$, $x_j \in \mathcal{N}_h$ implies that $\operatorname{supp}(\bar{\mu}) \subset K$. Combining this with the bound $\|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})} \le C$ we infer that

$$\int_{\tilde{K}} z^h \, d\bar{\mu}_h \to \int_K z \, d\bar{\mu} \quad \text{as } h \to 0 \tag{3.44}$$

for every sequence $(z^h)_{0<h\le h_1} \subset C(\tilde{K})$ converging uniformly to z on $\tilde{K}$.
Our next claim is that

$$\bar{y}_h \to \bar{y} \qquad \text{uniformly in } \bar{\Omega}. \tag{3.45}$$

To see this, denote by $y^h \in H^2(\Omega) \cap H_0^1(\Omega)$ the solution of

$$-\Delta y^h = \bar{u}_h - \phi(\bar{y}_h) \text{ in } \Omega, \quad y^h = 0 \text{ on } \partial\Omega.$$

We deduce from Lemma 3.4.1 and (3.36) that $(y^h)_{0<h\le h_1}$ is bounded in $H^2(\Omega)$, so that there exists a further subsequence and a function $\hat{y} \in H^2(\Omega) \cap H_0^1(\Omega)$ with

$$y^h \rightharpoonup \hat{y} \text{ in } H^2(\Omega), \quad y^h \to \hat{y} \text{ in } C(\bar{\Omega}).$$

In the light of (A.1) and Lemma A.1.4 we have $\bar{u}_h - \phi(\bar{y}_h) \rightharpoonup \bar{u} - \phi(\bar{y})$ in $L^2(\Omega)$ from which we find that $-\Delta\hat{y} = -\Delta\bar{y}$ a.e. in Ω. Hence $\hat{y} = \bar{y}$ and $y^h \to \bar{y}$ in $C(\bar{\Omega})$. On the other hand, the definition of y^h implies that $\bar{y}_h = R_h y^h$, so that the standard estimates (3.3),(3.4),(3.1),(3.7) imply

$$\begin{aligned}
\|\bar{y}_h - \bar{y}\|_{L^\infty(\Omega)} &\le \|R_h y^h - y^h\|_{L^\infty(\Omega)} + \|y^h - \bar{y}\|_{L^\infty(\Omega)} \\
&\le \|R_h y^h - I_h y^h\|_{L^\infty(\Omega)} + \|I_h y^h - y^h\|_{L^\infty(\Omega)} + \|y^h - \bar{y}\|_{L^\infty(\Omega)} \\
&\le Ch^{-1}\|R_h y^h - I_h y^h\|_{L^2(\Omega)} + Ch\|y^h\|_{H^2(\Omega)} + \|y^h - \bar{y}\|_{L^\infty(\Omega)} \\
&\le Ch^{-1}\Big(\|R_h y^h - y^h\|_{L^2(\Omega)} + \|y^h - I_h y^h\|_{L^2(\Omega)}\Big) \\
&\quad + Ch\|y^h\|_{H^2(\Omega)} + \|y^h - \bar{y}\|_{L^\infty(\Omega)} \\
&\le Ch\|y^h\|_{H^2(\Omega)} + \|y^h - \bar{y}\|_{L^\infty(\Omega)} \to 0 \quad \text{as } h \to 0,
\end{aligned}$$

since $\|y^h\|_{H^2(\Omega)} \le C$. This proves (3.45).
Let us check that $\bar{y}_{|K} \in Y_{ad}$. For a fixed point $x \in K$ we can choose a sequence $(x_{j_h})_{0<h\le h_1}$ such that $x_{j_h} \in \mathcal{N}_h$ and $|x_{j_h} - x| \le h$. Since $y_a(x_{j_h}) \le \bar{y}_h(x_{j_h}) \le y_b(x_{j_h})$ and

$$\begin{aligned}
|\bar{y}_h(x_{j_h}) - \bar{y}(x)| &\le |\bar{y}_h(x_{j_h}) - \bar{y}(x_{j_h})| + |\bar{y}(x_{j_h}) - \bar{y}(x)| \\
&\le \|\bar{y}_h - \bar{y}\|_{L^\infty(\Omega)} + |\bar{y}(x_{j_h}) - \bar{y}(x)|,
\end{aligned}$$

we obtain $y_a(x) \le \bar{y}(x) \le y_b(x)$ by passing to the limit $h \to 0$ and using (3.45).

Next, let us fix $z \in Y_{ad}$. By Lemma A.3.1 we extend z to a function $\tilde{z} \in C(\tilde{K})$ satisfying $y_a(x) \le \tilde{z}(x) \le y_b(x)$, $x \in \tilde{K}$. We obtain from (3.24), (3.44) and (3.45)

$$0 \ge \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j(\tilde{z}(x_j) - \bar{y}_h(x_j)) = \int_{\tilde{K}} (\tilde{z} - \bar{y}_h) d\bar{\mu}_h \to \int_K (z - \bar{y}) d\bar{\mu},$$

which yields (2.18).

In order to derive (2.16) we fix $v \in H^2(\Omega) \cap H_0^1(\Omega)$ and insert $v_h = R_h v$ into (3.22), i.e.

$$\int_\Omega \nabla\bar{p}_h \cdot \nabla R_h v + \phi'(\bar{y}_h)\bar{p}_h R_h v \, dx = \int_\Omega (\bar{y}_h - y_0) R_h v \, dx + \int_{\tilde{K}} R_h v \, d\bar{\mu}_h.$$

Using the definition of R_h and integration by parts we may write

$$\int_\Omega \nabla\bar{p}_h \cdot \nabla R_h v \, dx = \int_\Omega \nabla\bar{p}_h \cdot \nabla v \, dx = \int_\Omega \bar{p}_h(-\Delta v) \, dx$$

so that (2.16) follows from passing to the limit $h \to 0$ taking into account (3.41), (3.45), Lemma A.1.3 and (3.44).

Our next goal is to show that $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$. Inserting $\bar{u}$ into (3.23) and rearranging we infer

$$\alpha\|\bar{u}_h\|^2_{L^2(\Omega)} \le \int_\Omega \bar{u}(\alpha\bar{u}_h + \bar{p}_h) dx - \int_\Omega \bar{u}_h\bar{p}_h dx. \tag{3.46}$$

The second integral can be rewritten with the help of (3.21) and (3.22), namely

$$\begin{aligned}\int_\Omega \bar{u}_h\bar{p}_h dx &= \int_\Omega \nabla\bar{y}_h\cdot\nabla\bar{p}_h dx + \int_\Omega \phi(\bar{y}_h)\bar{p}_h dx\\ &= -\int_\Omega \phi'(\bar{y}_h)\bar{p}_h\bar{y}_h dx + \int_\Omega (\bar{y}_h - y_0)\bar{y}_h dx + \int_{\tilde{K}} \bar{y}_h d\bar{\mu}_h\\ &\quad + \int_\Omega \phi(\bar{y}_h)\bar{p}_h dx.\end{aligned}$$

This relation allows us to pass to the limit in a similar way as above to give

$$\begin{aligned}\int_\Omega \bar{u}_h\bar{p}_h dx &\to -\int_\Omega \phi'(\bar{y})\bar{p}\bar{y}dx + \int_\Omega(\bar{y}-y_0)\bar{y}dx + \int_K \bar{y}d\bar{\mu} + \int_\Omega\phi(\bar{y})\bar{p}dx\\ &= \int_\Omega(-\Delta\bar{y})\bar{p}dx + \int_\Omega\phi(\bar{y})\bar{p}dx = \int_\Omega \bar{u}\bar{p}dx,\end{aligned}$$

where we used (2.16) and the fact that $\bar{y} = \mathcal{G}(\bar{u})$. We can now pass to the limit in (3.46) and deduce that

$$\limsup_{h\to 0}\|\bar{u}_h\|^2_{L^2(\Omega)} \le \|\bar{u}\|^2_{L^2(\Omega)}.$$

Since $\|\bar{u}\|^2_{L^2(\Omega)} \le \liminf_{h\to 0}\|\bar{u}_h\|^2_{L^2(\Omega)}$ we infer that $\|\bar{u}_h\|_{L^2(\Omega)} \to \|\bar{u}\|_{L^2(\Omega)}$, which together with the fact that $\bar{u}_h \rightharpoonup \bar{u}$ in $L^2(\Omega)$ implies that $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$.

Combining this with the weak convergence $\bar{p}_h \rightharpoonup \bar{p}$ in $L^2(\Omega)$, one can pass to the limit in (3.23) to obtain

$$\int_\Omega (\bar{p} + \alpha\bar{u})(u-\bar{u})\,dx \ge 0 \quad \forall u \in U_{ad},$$

which is (2.17). In conclusion we see that $(\bar{u},\bar{y},\bar{p},\bar{\mu})$ is a solution of (2.15)–(2.18). Furthermore, the lower semicontinuity of the L^q-norm implies that

$$\|\bar{p}\|_{L^q(\Omega)} \le \liminf_{h\to 0}\|\bar{p}_h\|_{L^q(\Omega)} \le \eta(\alpha,r)$$

and we infer from Theorem 2.3.5, that $\bar{u}$ is a global minimum of Problem $(\mathbb{P})$. If (3.38) holds, then $\bar{p}$ satisfies $\|\bar{p}\|_{L^q(\Omega)} \le \kappa\eta(\alpha,r) < \eta(\alpha,r)$ and $\bar{u}$ is the unique global minimum of $(\mathbb{P})$. A standard argument then shows that the whole sequence $(\bar{u}_h)_{0<h\le h_1}$ converges to $\bar{u}$. ■

We conclude this section by the following general remarks.

Remark 7 The proof of Theorem 3.4.2 shows that if $(\bar{u}_h,\bar{y}_h,\bar{p}_h,\bar{\mu}_h)_{0<h\le h_1}$ satisfies (3.21)–(3.24) with

$$\|\bar{p}_h\|_{L^q(\Omega)} \le \kappa\eta(\alpha,r)$$

for some $0<\kappa<1$, then there exists $(\bar{u},\bar{y},\bar{p},\bar{\mu})$ satisfying (2.15)–(2.18) with

$$\|\bar{p}\|_{L^q(\Omega)} \le \kappa\eta(\alpha,r)$$

and $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$ as $h\to 0$, where $\bar{u}_h$ and $\bar{u}$ are the unique global minima for Problem $(\mathbb{P}_h)$ and Problem $(\mathbb{P})$, respectively. Importantly, neither (3.20) nor (2.14) are required.

Remark 8 1. It is well known that (3.21)–(3.24) can be rewritten equivalently as a system of semi-smooth equations and thus can be solved by a semi-smooth Newton method, see for instance [35, 44, 74]. In particular, we can avoid the use of relaxation methods such as Moreau-Yosida relaxation, interior point methods, or Lavrentiev-type regularization. We will come back to this point again in Section 3.7 where we will develop an algorithm to solve (3.21)–(3.24) by the semismooth Newton's method.

2. Since we solve (3.21)–(3.24) in practice on the computer, we consider $\bar{u}_h$ a global minimum if the inequality (3.26) is satisfied up to machine precision. Here, the integral $\|\bar{p}_h\|_{L^q}$ on the left hand side of this inequality is assumed to be calculated exactly. However, this assumption can be achieved easily whenever q is an integer because in this case the function $|\bar{p}_h|^q$ restricted to every triangle in the mesh is a (possibly piecewise) polynomial of degree q. Hence, one can use an appropriate quadrature rule to evaluate such an integral exactly.

3.5 Generalizations

It is possible to obtain the previous results, like Theorem 3.3.4 and Theorem 3.4.2, if the choices $K = \bar{\Omega}$, nonlinearity of the type $\phi(x, y)$ and a domain $\Omega \subset \mathbb{R}^3$ are considered for the Problem $(\mathbb{P})$. In this section we discuss the main changes that apply to the analysis of the previous results if these choices are considered.

3.5.1 The Case $K = \bar{\Omega}$

Here we briefly outline how the convergence analysis can be carried out for the case $K = \bar{\Omega}$, provided that the bounds $y_a, y_b \in C(\bar{\Omega})$ satisfy in addition to $y_a < y_b$ in $\bar{\Omega}$ the compatibility condition $y_a < 0 < y_b$ on $\partial\Omega$.

The main change concerns the construction of the function $w \in C_0^\infty(\Omega)$ at the beginning of Section 3.4. From the above assumptions on y_a, y_b we infer the existence of $\delta > 0$ and $\varepsilon > 0$ such that

$$y_a(x) + \delta \le \frac{1}{2}(y_a(x) + y_b(x)) \le y_b(x) - \delta, \quad x \in \bar{\Omega},$$
$$y_a(x) + \delta \le 0 \le y_b(x) - \delta, \quad x \in \bar{\Omega},\ \mathrm{dist}(x, \partial\Omega) \le \varepsilon.$$

According to Urysohn's lemma, there exists $\chi \in C_0(\Omega), 0 \le \chi \le 1$ such that

$$\chi(x) = \begin{cases} 0 & \text{if } \mathrm{dist}(x, \partial\Omega) \le \frac{\varepsilon}{2}, \\ 1 & \text{if } \mathrm{dist}(x, \partial\Omega) \ge \varepsilon. \end{cases}$$

Then one can easily see that

$$y_a(x) + \delta \le \frac{1}{2}(y_a(x) + y_b(x))\chi(x) \le y_b(x) - \delta, \quad x \in \bar{\Omega}$$

and applying a smoothing procedure to $x \mapsto \frac{1}{2}(y_a + y_b)\chi \in C_0(\Omega)$ gives the desired function $w \in C_0^\infty(\Omega)$ satisfying (3.33) with $\tilde{K} = \bar{\Omega}$. The remainder of the analysis in Section 3.4 can be carried out as before if one chooses $\mathcal{N}_h := \{x_j | x_j \text{ is a vertex of } T \in \mathcal{T}_h, x_j \notin \partial\Omega\}$.

3.5.2 The Nonlinearity $\phi(x, y)$

All the results derived so far remain valid for a nonlinearity of the type $\phi : \Omega \times \mathbb{R} \to \mathbb{R}$, $\phi = \phi(x, y)$ provided that it satisfies the following assumptions:

- For any fixed $y \in \mathbb{R}$, $\phi(\cdot, y)$ is measurable with respect to $x \in \Omega$ and for almost all $x \in \Omega$, $\phi(x, \cdot)$ is of class C^2 with respect to $y \in \mathbb{R}$.
- For almost all $x \in \Omega$ and all $y \in \mathbb{R}$ it holds that $\phi_y(x, y) \geq 0$, where ϕ_y denotes the partial derivative of ϕ with respect to y.
- There exists $c > 0$ such that

$$|\phi(x, 0)| + |\phi_y(x, 0)| \leq c \quad \text{for a.e. } x \in \Omega.$$

- There exist $r > 1$ and $M \geq 0$ such that

$$|\phi_{yy}(x, y)| \leq M \phi_y(x, y)^{\frac{1}{r}} \quad \text{for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}. \tag{3.47}$$

Analogously to the proof of Lemma A.1.1 one can use the bound on $\phi(x, 0)$ and $\phi_y(x, 0)$ in order to show that

$$\begin{aligned} &\phi_y(x, y) \leq c_1(1 + |y|^{r_1}) \quad \text{for a.e. } x \in \Omega,\ y \in \mathbb{R}, \quad r_1 = \frac{r}{r-1}; \\ &|\phi(x, y)| \leq c_0(1 + |y|^{r_0}) \quad \text{for a.e. } x \in \Omega,\ y \in \mathbb{R}, \quad r_0 = \frac{2r-1}{r-1}. \end{aligned}$$

In particular, ϕ is locally Lipschitz with respect to y uniformly in $x \in \Omega$ so that the corresponding semilinear PDE is well-posed.

For instance, if $\phi(x, y) = a(x)|y|^{q-2}y$, where $q > 3$ and $a \in L^\infty(\Omega)$ with $a(x) \geq 0$ a.e. in Ω, then

$$|\phi_{yy}(x, y)| = (q-2)[(q-1)a(x)]^{\frac{1}{q-2}}[(q-1)a(x)|y|^{q-2}]^{\frac{q-3}{q-2}}.$$

Hence, (3.47) holds with $r = \frac{q-2}{q-3}$ and $M = (q-2)(q-1)^{\frac{1}{q-2}} \|a\|_{L^\infty(\Omega)}^{\frac{1}{q-2}}$.

3.5.3 The 3D case

It is possible to obtain a result similar to Theorem 3.3.4 for the case without state constraints if the domain $\Omega \subset \mathbb{R}^3$ is bounded, convex and polyhedral. The proof of this result will be the same as that of Theorem 3.3.4 except that instead of Theorem A.5.1, one needs to use the Gagliardo–Nirenberg interpolation inequality in $\mathbb{R}^3$ which reads

$$\|f\|_{L^q(\mathbb{R}^3)} \leq G N_q \|f\|_{L^2(\mathbb{R}^3)}^{1-\theta} \|\nabla f\|_{L^2(\mathbb{R}^3)}^{\theta} \quad \forall\, f \in H^1(\mathbb{R}^3), \quad 2 \leq q \leq 6, \quad \theta := \tfrac{3}{2} - \tfrac{3}{q}.$$

An upper bound for GN_q can be found, for instance, in [75]. However, since our analysis relies on the choice $q := \frac{3r-2}{r-1}$, the condition $2 \leq q \leq 6$ leads to a restriction on r, namely, $r \geq \frac{4}{3}$. The convergence analysis can then be carried out in almost the same way as in the 2D case. For the uniform convergence of the Ritz projection when $\Omega \subset \mathbb{R}^3$ is a convex polyhedral domain we refer to the work in [59].

In the case of state constraints the adjoint variable p in general only belongs to $W^{1,s}(\Omega)$ with $s \in [1, \frac{3}{2})$, and hence $p \in L^q(\Omega)$ for $1 \leq q < 3$. In particular, the choice $q := \frac{3r-2}{r-1}$ is no longer possible and an extension of our approach to the 3D case would require a better regularity result for p. A corresponding result which gives $p \in H^1(\Omega) \cap L^\infty(\Omega)$ has recently been obtained in [24] under mild restrictions on the bounds y_a, y_b. Namely, for a linear elliptic state equation the state bounds should satisfy

$$\begin{aligned} &y_a, y_b \in C(\bar{\Omega}), \\ &y_a(x) < y_b(x) \quad \forall\, x \in \bar{\Omega}, \\ &y_a(x) < 0 < y_b(x) \quad \forall\, x \in \partial\Omega, \\ &\mathcal{A}y_a, \mathcal{A}y_b \in L^\infty(\Omega), \end{aligned}$$

where $\mathcal{A}$ is the differential operator associated with the state equation which is assumed to be uniformly elliptic with coefficient funtions bounded in the domain $\Omega \subset \mathbb{R}^n$ for $n = 2, 3$ with Lipschitz boundary $\partial\Omega$. We point out that the state equation satisfies a homogeneous Dirichlet boundary condition and the control is distributed in Ω. If, in addition, control constraints of box type with bounds $u_a, u_b \in \mathbb{R}$ and $u_a < u_b$ are considered, then one should require

$$u_a < \mathcal{A}y_b \text{ and } u_b > \mathcal{A}y_a \text{ in } \Omega.$$

For a semilinear elliptic state equation the state bounds should, in addition to the previous conditions, satisfy

$$\phi(\cdot, y_a(\cdot)),\ \phi(\cdot, y_b(\cdot)) \in L^\infty(\Omega),$$

where $\phi : \Omega \times \mathbb{R} \to \mathbb{R}$ is the nonlinearity of the PDE which has classical assumptions, for example those in Section 3.5.2, that guarantee the well-posedness of the state equation and the continuity of the state variable in $\bar{\Omega}$.

3.6 Error Analysis

Recall that, according to Remark 7, if $(\bar{u}_h, \bar{y}_h, \bar{p}_h, \bar{\mu}_h)_{0<h\leq h_1}$ satisfies (3.21)–(3.24) with

$$\|\bar{p}_h\|_{L^q(\Omega)} \leq \kappa\eta(\alpha, r) \qquad \text{for some } 0 < \kappa < 1, \tag{3.48}$$

then there exists $(\bar{u}, \bar{y}, \bar{p}, \bar{\mu})$ satisfying (2.15)–(2.18) with

$$\|\bar{p}\|_{L^q(\Omega)} \leq \kappa\eta(\alpha, r) \tag{3.49}$$

and $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$ as $h \to 0$, where $\bar{u}_h$ and $\bar{u}$ are the unique global minima for Problem $(\mathbb{P}_h)$ and Problem $(\mathbb{P})$, respectively. Our goal now is to investigate the error in approximating $\bar{u}$ by $\bar{u}_h$.

To start, let us introduce the auxiliary functions $\tilde{y} \in H^2(\Omega) \cap H_0^1(\Omega)$, $\tilde{y}_h \in$

X_{h0}, $\tilde{p}_h \in X_{h0}$ such that

$$\int_\Omega \nabla\tilde{y}\cdot\nabla v + \phi(\tilde{y})v\,dx = \int_\Omega \bar{u}_h v\,dx \qquad \forall\, v \in H_0^1(\Omega), \tag{3.50}$$

$$\int_\Omega \nabla\tilde{y}_h\cdot\nabla v_h + \phi(\tilde{y}_h)v_h\,dx = \int_\Omega \bar{u}v_h\,dx \qquad \forall\, v_h \in X_{h0}, \tag{3.51}$$

$$\begin{aligned}&\int_\Omega \nabla\tilde{p}_h\cdot\nabla v_h + \phi'(\bar{y})\tilde{p}_h v_h\,dx\\ &= \int_\Omega (\bar{y}-y_0)v_h\,dx + \int_K v_h\,d\bar{\mu} \qquad \forall\, v_h \in X_{h0}.\end{aligned} \tag{3.52}$$

Then, the following error estimates hold

$$\|\bar{y}_h - \tilde{y}\|_{L^2(\Omega)} \le ch^2\big(\|\bar{u}_h\|_{L^2(\Omega)}+1\big), \tag{3.53}$$

$$\|\tilde{y}_h - \bar{y}\|_{L^2(\Omega)} \le ch^2\big(\|\bar{u}\|_{L^2(\Omega)}+1\big), \tag{3.54}$$

$$\|\tilde{y}_h - \bar{y}\|_{L^\infty(\Omega)} \le c|\ln h|h^{2-\frac{2}{p}}\big(\|\bar{u}\|_{L^p(\Omega)}+1\big), \tag{3.55}$$

$$\|\tilde{p}_h - \bar{p}\|_{L^2(\Omega)} \le ch\big(\|\bar{y}-y_0\|_{L^2(\Omega)} + \|\bar{\mu}\|_{\mathcal{M}(K)}\big). \tag{3.56}$$

The estimates (3.53) and (3.54) follow from Theorem 3.2.3. Notice that $\|\bar{u}_h\|_{L^2(\Omega)}$ is uniformly bounded in h according to Lemma 3.4.1. On the other hand, (3.55) follows from Theorem 3.2.4 since $\bar{u} \in W^{1,s}(\Omega)$, $1 < s < 2$, by Lemma 2.3.3. Finally, the estimate (3.56) is a consequence of [14, Theorem 3].

To guarantee a high order of convergence for the sequence $(\bar{u}_h)_{0<h\le h_1}$ when the pointwise state constraints are considered, we make the next assumptions. Firstly, we require the bounds y_a, y_b to be regular enough, namely,

- $y_a, y_b \in C_0(\Omega)\cap W^{2,\infty}(\Omega)$ such that $y_a(x) < y_b(x)$, $x \in K$.

Next, we make an assumption on $\nabla\bar{y}$ at the set of points in K where the state constraints are active. This assumption shall be mentioned explicitly wherever it is needed.

Assumption 2 For the optimal state $\bar{y}$ there holds

$$\begin{aligned}\nabla\bar{y}(x) &= \nabla y_b(x) \qquad \forall\, x \in K : \bar{y}(x) = y_b(x),\\ \nabla\bar{y}(x) &= \nabla y_a(x) \qquad \forall\, x \in K : \bar{y}(x) = y_a(x).\end{aligned}$$

Notice that Assumption 2 is satisfied automatically at the state constraints active points that belong to the interior of K, denoted by $\operatorname{int} K$. To see this, consider the function $f_b : K \to \mathbb{R}$ defined by $f_b := \bar{y} - y_b$. If $f_b(x^*) = 0$ for some $x^* \in K$, then x^* is a global maximum for the function f_b since $f_b \le 0$ in K. Consequently, the first order necessary optimality condition at x^* reads $\nabla f_b(x^*) = 0$ provided that $x^* \in \operatorname{int} K$. On the other hand, if $x^* \in \partial K$, then $\nabla f_b(x^*) \neq 0$ in general.

Since the optimal state $\bar{y}$ is usually not known a priori, the previous assumption might be practically restrictive. For this reason, we make the next assumption which guarantees the same convergence order for $(\bar{u}_h)_{0<h\le h_1}$ as the one that Assumption 2 guarantees but it is relatively easier to be fulfilled.

Assumption 3 For every $h > 0$ there exists a set of triangles $\mathbb{T} \subset \mathcal{T}_h$ such that

$$K = \bigcup_{T\in\mathbb{T}} \bar{T}.$$

Observe that if the set K is polygonal, the previous assumption can be easily satisfied.

We can now formulate our main theorem of this section which establishes the error estimate of approximating the unique global minimum $\bar{u}$ of problem $(\mathbb{P})$ by the sequence $(\bar{u}_h)_{0<h\leq h_1}$ of the unique global minima of the corresponding problems $(\mathbb{P}_h)$.

Theorem 3.6.1 *Suppose that Assumption 1 holds. Let $(\bar{u}_h)_{0<h\leq h_1}$ be obtained from solving (3.21)–(3.24) such that (3.48) is satisfied. Then there exists $c>0$ independent of h such that*

$$\|\bar{u}_h-\bar{u}\|_{L^2(\Omega)}+\|\bar{y}_h-\bar{y}\|_{H^1(\Omega)}\leq ch^{\frac{1}{2}},$$

where $\bar{u}$ is the unique global minimum of Problem $(\mathbb{P})$. Here, $\bar{y}$ and $\bar{y}_h$ are the optimal states corresponding to $\bar{u}$ and $\bar{u}_h$, respectively. If in addition Assumption 2 or Assumption 3 holds, then

$$\|\bar{u}_h-\bar{u}\|_{L^2(\Omega)}+\|\bar{y}_h-\bar{y}\|_{H^1(\Omega)}\leq c\sqrt{|\ln h|}h^{\frac{3}{2}-\frac{1}{s}} \tag{3.57}$$

for any $1<s<s_\Omega:=\min(2,\frac{2\theta_{\max}}{3\theta_{\max}-\pi})$ with $\theta_{\max}\in[\frac{\pi}{3},\pi)$ being the maximum interior angle in Ω. For $\theta_{\max}=\frac{\pi}{3}$ we define $s_\Omega:=2$.

Proof: Throughout the proof, the sequence $(\bar{u}_h,\bar{y}_h,\bar{p}_h,\bar{\mu}_h)_{0<h\leq h_1}$ and $(\bar{u},\bar{y},\bar{p},\bar{\mu})$ are exactly as described at the beginning of this section.

Testing (2.17) with $\bar{u}_h$ and (3.23) with $\bar{u}$ and adding the resulting inequalities gives

$$\int_\Omega(\bar{p}_h-\bar{p})(\bar{u}-\bar{u}_h)-\alpha(\bar{u}_h-\bar{u})^2\,dx\geq 0$$

from which we obtain

$$\begin{aligned}\alpha\|\bar{u}-\bar{u}_h\|^2_{L^2(\Omega)}&\leq\int_\Omega(\bar{u}-\bar{u}_h)(\bar{p}_h-\bar{p})\,dx\\&=\underbrace{\int_\Omega(\tilde{p}_h-\bar{p})(\bar{u}-\bar{u}_h)\,dx}_{S_1}+\int_\Omega(\bar{p}_h-\tilde{p}_h)(\bar{u}-\bar{u}_h)\,dx.\end{aligned}\tag{3.58}$$

We see that from (3.21) and (3.51) with the choice $v_h=\bar{p}_h-\tilde{p}_h$ we have

$$\begin{aligned}\int_\Omega(\bar{p}_h-\tilde{p}_h)(\bar{u}-\bar{u}_h)\,dx&=\int_\Omega[\phi(\tilde{y}_h)-\phi(\bar{y}_h)](\bar{p}_h-\tilde{p}_h)\,dx\\&+\int_\Omega\nabla(\tilde{y}_h-\bar{y}_h)\cdot\nabla(\bar{p}_h-\tilde{p}_h)\,dx\\&=\underbrace{\int_\Omega(\bar{y}_h-\bar{y})(\tilde{y}_h-\bar{y}_h)\,dx}_{S_2}+\underbrace{\int_{\tilde{K}}(\tilde{y}_h-\bar{y}_h)\,d\bar{\mu}_h-\int_K(\tilde{y}_h-\bar{y}_h)\,d\bar{\mu}}_{S_3}\\&+\underbrace{\int_\Omega[\phi(\tilde{y}_h)-\phi(\bar{y}_h)](\bar{p}_h-\tilde{p}_h)\,dx-\int_\Omega[\phi'(\bar{y}_h)\bar{p}_h-\phi'(\bar{y})\tilde{p}_h](\tilde{y}_h-\bar{y}_h)\,dx}_{S_4},\end{aligned}$$

where we utilized (3.22) and (3.52) with the test function $v_h = \tilde{y}_h - \bar{y}_h$ to rewrite the term containing the gradients in the first equality. Consequently, adding the terms S_2, S_3, S_4 to S_1 in (3.58) gives

$$\alpha\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \leq \sum_{i=1}^{4} S_i. \tag{3.59}$$

Estimating S_1: To obtain an upper bound for S_1 we use the Cauchy-Schwarz inequality, Lemma A.4.3 for some $\epsilon > 0$ and (3.56) to get

$$\begin{aligned} S_1 &\leq \|\tilde{p}_h - \bar{p}\|_{L^2(\Omega)}\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \\ &\leq \frac{1}{2\alpha\epsilon}\|\tilde{p}_h - \bar{p}\|^2_{L^2(\Omega)} + \frac{\alpha\epsilon}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \\ &\leq \frac{c}{\alpha\epsilon}h^2\big(\|\bar{y} - y_0\|^2_{L^2(\Omega)} + \|\bar{\mu}\|^2_{\mathcal{M}(K)}\big) + \frac{\alpha\epsilon}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \\ &= \frac{\alpha\epsilon}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} + O(h^2). \end{aligned}$$

Estimating S_2: we argue in a similar way to that of estimating S_1 and we use (3.54) to obtain

$$\begin{aligned} S_2 &= -\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \int_\Omega (\bar{y}_h - \bar{y})(\tilde{y}_h - \bar{y})\,dx \\ &\leq -\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \|\bar{y}_h - \bar{y}\|_{L^2(\Omega)}\|\tilde{y}_h - \bar{y}\|_{L^2(\Omega)} \\ &\leq -\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{\epsilon}{2}\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{1}{2\epsilon}\|\tilde{y}_h - \bar{y}\|^2_{L^2(\Omega)} \\ &\leq (\frac{\epsilon}{2} - 1)\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{c}{\epsilon}h^4\big(\|\bar{u}\|^2_{L^2(\Omega)} + 1\big) \\ &= (\frac{\epsilon}{2} - 1)\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + O(h^4). \end{aligned}$$

Estimating S_3: Before we start estimating S_3, we first introduce some notation. Let us define $f_b := \bar{y} - y_b$ and $f_a := y_a - \bar{y}$. It is clear that $f_b, f_a \leq 0$ in K and that $f_b, f_a \in W^{2,p}(\Omega) \hookrightarrow C^{1,1-\frac{2}{p}}(\bar{\Omega})$ since $\bar{y} \in W^{2,p}(\Omega)$ according to Lemma 2.3.3. Next, we recall the decomposition $\bar{\mu}_h = \bar{\mu}^b_h - \bar{\mu}^a_h$ from Proposition 3.3.3 and we introduce, for $0 < h \leq h_1$, the sets $\mathcal{A}_h$, $\mathcal{B}_h \subset \mathcal{N}_h$ by

$$\begin{aligned} \mathcal{A}_h &:= \{x_j \in \operatorname{supp}(\bar{\mu}^a_h) \setminus K : f_a(x_j) \geq 0\}, \\ \mathcal{B}_h &:= \{x_j \in \operatorname{supp}(\bar{\mu}^b_h) \setminus K : f_b(x_j) \geq 0\}. \end{aligned}$$

We have the following claims:

$$\mathcal{A}_h \neq \emptyset \quad \forall\, h > 0 \Rightarrow \exists\, x \in \partial K : f_a(x) = 0, \tag{3.60}$$

$$\mathcal{B}_h \neq \emptyset \quad \forall\, h > 0 \Rightarrow \exists\, x \in \partial K : f_b(x) = 0. \tag{3.61}$$

To show (3.60), suppose that $f_a(x) < 0$ for all $x \in \partial K$. Then, from the continuity of f_a and the fact that $\operatorname{dist}(x_j, \partial K) \leq h$ for any $x_j \in \mathcal{A}_h$ we deduce that $\mathcal{A}_h = \emptyset$ for sufficiently small h, which is a contradiction. The claim (3.61) can be shown analogously.

Thanks to (3.60), we can assign to each $x_j \in \mathcal{A}_h$ a point $x^*_j \in \partial K$ such that $|x_j - x^*_j| \leq h$ and $f_a(x^*_j) = 0$. Notice that if Assumption 2 holds, then

$\nabla f_a(x_j^*) = 0$ as well. Since $f_a \in C^{1,1-\frac{2}{p}}(\bar{\Omega})$ as explained at the beginning, we conclude from Lemma A.2.1 that there exists a constant $c > 0$ independent of h such that

$$f_a(x_j) = f_a(x_j) - f_a(x_j^*) \leq c|x_j - x_j^*|^\gamma \leq ch^\gamma \quad \forall\, x_j \in \mathcal{A}_h, \tag{3.62}$$

where we define

$$\gamma := \begin{cases} 2 - \frac{2}{p}, & \text{if Assumption 2 holds,} \\ 1, & \text{otherwise.} \end{cases}$$

Using a similar argumentation, we have

$$f_b(x_j) = f_b(x_j) - f_b(x_j^*) \leq c|x_j - x_j^*|^\gamma \leq ch^\gamma \quad \forall\, x_j \in \mathcal{B}_h. \tag{3.63}$$

We are now ready to estimate S_3 and we start by establishing an upper bound for the first integral in S_3.
Using $\bar{\mu}_h = \bar{\mu}_h^b - \bar{\mu}_h^a$, Lemma 3.4.1, the estimate (3.55) as well as (3.62) and (3.63), we have

$$\begin{aligned}
\int_{\tilde{K}} (\tilde{y}_h - \bar{y}_h)\, d\bar{\mu}_h &= \int_{\tilde{K}} (\tilde{y}_h - y_b)\, d\bar{\mu}_h^b - \int_{\tilde{K}} (\tilde{y}_h - y_a)\, d\bar{\mu}_h^a \\
&= \int_{\tilde{K}} (\tilde{y}_h - \bar{y})\, d\bar{\mu}_h^b + \int_{\tilde{K}} f_b\, d\bar{\mu}_h^b + \int_{\tilde{K}} (\bar{y} - \tilde{y}_h)\, d\bar{\mu}_h^a + \int_{\tilde{K}} f_a\, d\bar{\mu}_h^a \\
&\leq \|\tilde{y}_h - \bar{y}\|_{L^\infty(\tilde{K})} \|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})} + \sum_{x_j \in \mathcal{B}_h} f_b(x_j)\bar{\mu}_j + \sum_{x_j \in \mathcal{A}_h} f_a(x_j)\bar{\mu}_j \\
&\leq c|\ln h| h^{2-\frac{2}{p}} \big(\|\bar{u}\|_{L^p(\Omega)} + 1\big) + ch^\gamma \sum_{x_j \in \mathcal{B}_h} \bar{\mu}_j + ch^\gamma \sum_{x_j \in \mathcal{A}_h} \bar{\mu}_j \\
&\leq O(|\ln h| h^{2-\frac{2}{p}}) + ch^\gamma \|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})}.
\end{aligned}$$

Observe that if Assumption 3 holds, then $\mathcal{N}_h \subset K$ and

$$f_a(x_j) \leq 0 \quad \forall\, x_j \in \operatorname{supp}(\bar{\mu}_h^a) \quad \text{and} \quad f_b(x_j) \leq 0 \quad \forall\, x_j \in \operatorname{supp}(\bar{\mu}_h^b).$$

This implies that the summation terms containing f_a and f_b in the previous estimate can be dropped and then one ends up with the order $O(|\ln h| h^{2-\frac{2}{p}})$. Consequently, we conclude that

$$\int_{\tilde{K}} (\tilde{y}_h - \bar{y}_h)\, d\bar{\mu}_h = \begin{cases} O(|\ln h| h^{2-\frac{2}{p}}), & \text{if Assumption 2} \\ & \text{or Assumption 3 holds,} \\ O(h), & \text{otherwise.} \end{cases} \tag{3.64}$$

To estimate the second integral in S_3 we use the decomposition $\bar{\mu} = \bar{\mu}_b - \bar{\mu}_a$ from Proposition 2.3.4, the fact that $I_h y_a \leq \bar{y}_h \leq I_h y_b$ in K, the interpolation

error (3.3) and the uniform estimate (3.55) to obtain

$$\begin{aligned}
\int_K (\bar{y}_h - \tilde{y}_h)\, d\bar{\mu} &= \int_K (\bar{y}_h - \tilde{y}_h)\, d\bar{\mu}_b - \int_K (\bar{y}_h - \tilde{y}_h)\, d\bar{\mu}_a \\
&= \int_K (\bar{y}_h - I_h y_b + I_h y_b - y_b + y_b - \bar{y} + \bar{y} - \tilde{y}_h)\, d\bar{\mu}_b \\
&\quad + \int_K (\tilde{y}_h - \bar{y} + \bar{y} - y_a + y_a - I_h y_a + I_h y_a - \bar{y}_h)\, d\bar{\mu}_a \\
&\leq \int_K (I_h y_b - y_b)\, d\bar{\mu}_b + \int_K (\bar{y} - \tilde{y}_h)\, d\bar{\mu}_b + \int_K (\tilde{y}_h - \bar{y})\, d\bar{\mu}_a + \int_K (y_a - I_h y_a)\, d\bar{\mu}_a \\
&\leq \|\bar{\mu}\|_{\mathcal{M}(K)} \Big(\|\tilde{y}_h - \bar{y}\|_{L^\infty(K)} + \|y_a - I_h y_a\|_{L^\infty(K)} + \|y_b - I_h y_b\|_{L^\infty(K)} \Big) \\
&\leq c\|\bar{\mu}\|_{\mathcal{M}(K)} \Big(|\ln h| h^{2-\frac{2}{p}} (\|\bar{u}\|_{L^p(\Omega)} + 1) + h^2 \|y_a\|_{W^{2,\infty}(\Omega)} + h^2 \|y_b\|_{W^{2,\infty}(\Omega)} \Big) \\
&\leq c|\ln h| h^{2-\frac{2}{p}} \|\bar{\mu}\|_{\mathcal{M}(K)} \Big(\|\bar{u}\|_{L^p(\Omega)} + \|y_a\|_{W^{2,\infty}(\Omega)} + \|y_b\|_{W^{2,\infty}(\Omega)} + 1 \Big) \\
&= O(|\ln h| h^{2-\frac{2}{p}}),
\end{aligned} \tag{3.65}$$

where we used that $h^2 \leq |\ln h| h^{2-\frac{2}{p}}$ for sufficiently small h.
Combining (3.64) and (3.65) yields

$$S_3 = \begin{cases} O(|\ln h| h^{2-\frac{2}{p}}), & \text{if Assumption 2 or Assumption 3 holds,} \\ O(h), & \text{otherwise.} \end{cases}$$

Estimating S_4: To have an upper bound for S_4, we first rewrite S_4 as

$$\begin{aligned}
S_4 = &\underbrace{\int_\Omega [\phi(\tilde{y}_h) - \phi(\bar{y}_h) + \phi'(\bar{y}_h)(\bar{y}_h - \tilde{y}_h)]\bar{p}_h\, dx}_{S_{4.1}} + \underbrace{\int_\Omega [\phi(\tilde{y}) - \phi(\bar{y}) + \phi'(\bar{y})(\bar{y} - \tilde{y})]\bar{p}\, dx}_{S_{4.2}} \\
&+ \underbrace{\int_\Omega [\phi(\tilde{y}) - \phi(\bar{y}) + \phi'(\bar{y})(\bar{y} - \tilde{y})](\tilde{p}_h - \bar{p})\, dx}_{S_{4.3}} + \underbrace{\int_\Omega [\phi(\bar{y}) - \phi(\tilde{y}_h) + \phi'(\bar{y})(\tilde{y}_h - \bar{y})]\tilde{p}_h\, dx}_{S_{4.4}} \\
&+ \underbrace{\int_\Omega [\phi(\bar{y}_h) - \phi(\tilde{y}) + \phi'(\bar{y})(\tilde{y} - \bar{y}_h)]\tilde{p}_h\, dx}_{S_{4.5}}.
\end{aligned}$$

To estimate $S_{4.1}$, we first use the same argumentation used for estimating $R_h(u_h)$ in the proof of Theorem 3.3.4. Then, we apply (3.48), the triangle inequality, Lemma A.4.3 for some $\epsilon > 0$ and (3.54) to obtain

$$\begin{aligned}
S_{4.1} &= \int_\Omega \bar{p}_h(\tilde{y}_h - \bar{y}_h) \int_0^1 \phi'\big(\tilde{y}_h t + (1-t)\bar{y}_h\big) - \phi'(\bar{y}_h)\, dt\, dx \\
&\leq \eta(\alpha, r)^{-1} \|\bar{p}_h\|_{L^q(\Omega)} \Big(\frac{1}{2}\|\tilde{y}_h - \bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \Big) \\
&\leq \kappa \Big(\frac{1+\epsilon}{2}\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{1+\epsilon^{-1}}{2}\|\tilde{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \Big) \\
&\leq \kappa \Big(\frac{1+\epsilon}{2}\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \Big) + \kappa c(1+\epsilon^{-1}) h^4 \big(\|\bar{u}\|^2_{L^2(\Omega)} + 1 \big) \\
&= \kappa \Big(\frac{1+\epsilon}{2}\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega)} \Big) + O(h^4).
\end{aligned}$$

In a similar way, but this time while using (3.49) and (3.53), we have for $S_{4.2}$

$$\begin{aligned}
S_{4.2} &= \int_\Omega \bar{p}(\tilde{y}-\bar{y}) \int_0^1 \phi'\big(\tilde{y}t+(1-t)\bar{y}\big) - \phi'(\bar{y})\,dt\,dx \\
&\leq \eta(\alpha,r)^{-1}\|\bar{p}\|_{L^q(\Omega)}\Big(\frac{1}{2}\|\tilde{y}-\bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)}\Big) \\
&\leq \kappa\Big(\frac{1+\epsilon}{2}\|\bar{y}_h-\bar{y}\|^2_{L^2(\Omega)} + \frac{1+\epsilon^{-1}}{2}\|\tilde{y}-\bar{y}_h\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)}\Big) \\
&\leq \kappa\Big(\frac{1+\epsilon}{2}\|\bar{y}_h-\bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)}\Big) + \kappa c(1+\epsilon^{-1})h^4\big(\|\bar{u}_h\|^2_{L^2(\Omega)}+1\big) \\
&= \kappa\Big(\frac{1+\epsilon}{2}\|\bar{y}_h-\bar{y}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)}\Big) + O(h^4).
\end{aligned}$$

To have an upper bound for $S_{4.3}$ we apply the Cauchy-Schwarz inequality, Lemma A.1.3, Lemma 2.2.3, Lemma A.4.3 for some $\epsilon > 0$ and (3.56) to get

$$\begin{aligned}
S_{4.3} &\leq \|\phi(\tilde{y})-\phi(\bar{y})+\phi'(\bar{y})(\bar{y}-\tilde{y})\|_{L^2(\Omega)}\|\tilde{p}_h-\bar{p}\|_{L^2(\Omega)} \\
&\leq \big(L(m)+\|\phi'(\bar{y})\|_{L^\infty(\Omega)}\big)\|\tilde{y}-\bar{y}\|_{L^2(\Omega)}\|\tilde{p}_h-\bar{p}\|_{L^2(\Omega)} \\
&\leq c\|\bar{u}_h-\bar{u}\|_{L^2(\Omega)}\|\tilde{p}_h-\bar{p}\|_{L^2(\Omega)} \\
&\leq \frac{\alpha\epsilon}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)} + \frac{c^2}{2\alpha\epsilon}\|\tilde{p}_h-\bar{p}\|^2_{L^2(\Omega)} \\
&\leq \frac{\alpha\epsilon}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)} + \frac{c}{\alpha\epsilon}h^2\big(\|\bar{y}-y_0\|^2_{L^2(\Omega)} + \|\bar{\mu}\|^2_{\mathcal{M}(K)}\big) \\
&= \frac{\alpha\epsilon}{2}\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)} + O(h^2),
\end{aligned}$$

where $L(m) > 0$ is a constant from Lemma A.1.3.
Arguing like above and using (3.54), we obtain for $S_{4.4}$

$$\begin{aligned}
S_{4.4} &\leq \big(L(m)+\|\phi'(\bar{y})\|_{L^\infty(\Omega)}\big)\|\tilde{y}_h-\bar{y}\|_{L^2(\Omega)}\|\tilde{p}_h\|_{L^2(\Omega)} \\
&\leq ch^2\big(\|\bar{u}\|_{L^2(\Omega)}+1\big) \\
&= O(h^2).
\end{aligned}$$

Notice that $\|\tilde{p}_h\|_{L^2(\Omega)}$ is uniformly bounded for sufficiently small h as it can be seen from (3.56).
In a similar way and while using (3.53) this time, we have for $S_{4.5}$

$$\begin{aligned}
S_{4.5} &\leq \big(L(m)+\|\phi'(\bar{y})\|_{L^\infty(\Omega)}\big)\|\tilde{y}-\bar{y}_h\|_{L^2(\Omega)}\|\tilde{p}_h\|_{L^2(\Omega)} \\
&\leq ch^2\big(\|\bar{u}_h\|_{L^2(\Omega)}+1\big) \\
&= O(h^2).
\end{aligned}$$

Gathering the estimates for $S_{4.1},\dots,S_{4.5}$, we conclude that S_4 can be bounded by

$$S_4 \leq \kappa(1+\epsilon)\|\bar{y}_h-\bar{y}\|^2_{L^2(\Omega)} + (\kappa\alpha+\frac{\alpha\epsilon}{2})\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)} + O(h^2).$$

The Final Step: Inserting the estimates of the terms $S_1,\dots,S_4$ into (3.59) yields

$$\begin{aligned}
\alpha\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)} \leq \big(\kappa(1+\epsilon)+(\frac{\epsilon}{2}-1)\big)\|\bar{y}_h-\bar{y}\|^2_{L^2(\Omega)} + \alpha(\kappa+\epsilon)\|\bar{u}_h-\bar{u}\|^2_{L^2(\Omega)} \\
+ \begin{cases} O(|\ln h|h^{2-\frac{2}{p}}), & \text{if Assumption 2 or Assumption 3 holds,} \\ O(h), & \text{otherwise.} \end{cases}
\end{aligned}$$

Since $\kappa < 1$, choosing $\epsilon > 0$ to be small enough in the above expression yields the existence of $c > 0$ independent of h such that

$$c\|\bar{u}_h - \bar{u}\|^2_{L^2(\Omega)} + c\|\bar{y}_h - \bar{y}\|^2_{L^2(\Omega)} = \begin{cases} O(|\ln h| h^{2-\frac{2}{p}}), & \text{if Assumption 2} \\ & \text{or Assumption 3 holds,} \\ O(h), & \text{otherwise,} \end{cases}$$

from which we obtain the desired estimate for the control. It remains to establish an upper bound for $\|\bar{y}_h - \bar{y}\|_{H^1(\Omega)}$. To this end, let $R_h\bar{y}$ denote the Ritz projection of $\bar{y}$. Then, from the triangle inequality and (3.6) we obtain

$$\begin{aligned} \|\bar{y}_h - \bar{y}\|_{H^1(\Omega)} &\leq \|\bar{y}_h - R_h\bar{y}\|_{H^1(\Omega)} + \|R_h\bar{y} - \bar{y}\|_{H^1(\Omega)} \\ &\leq \|\bar{y}_h - R_h\bar{y}\|_{H^1(\Omega)} + ch\|\bar{y}\|_{H^2(\Omega)}. \end{aligned} \tag{3.66}$$

We now derive an upper bound for $\|\bar{y}_h - R_h\bar{y}\|_{H^1(\Omega)}$. Firstly, from the definition of $R_h\bar{y}$ and the weak formulation of $\bar{y}$ we have

$$\int_\Omega \nabla R_h\bar{y} \cdot \nabla w_h \, dx = \int_\Omega \nabla\bar{y} \cdot \nabla w_h \, dx = \int_\Omega \bar{u}w_h \, dx - \int_\Omega \phi(\bar{y})w_h \, dx \quad \forall\, w_h \in X_{h0}.$$

From this and (3.21) we get

$$\int_\Omega \nabla(R_h\bar{y} - \bar{y}_h) \cdot \nabla w_h \, dx = \int_\Omega (\bar{u} - \bar{u}_h)w_h \, dx + \int_\Omega [\phi(\bar{y}_h) - \phi(\bar{y})]w_h \, dx \quad \forall\, w_h \in X_{h0}.$$

Using $w_h = R_h\bar{y} - \bar{y}_h$ in the previous variational equation and then applying the Cauchy-Schwarz inequality, Lemma A.1.3 and the Poincaré's inequality we obtain

$$\begin{aligned} \int_\Omega |\nabla(R_h\bar{y} - \bar{y}_h)|^2 \, dx &= \int_\Omega (\bar{u} - \bar{u}_h)(R_h\bar{y} - \bar{y}_h) \, dx \\ &\quad + \int_\Omega [\phi(\bar{y}_h) - \phi(\bar{y})](R_h\bar{y} - \bar{y}_h) \, dx \\ &\leq \big(\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} + \|\phi(\bar{y}_h) - \phi(\bar{y})\|_{L^2(\Omega)}\big)\|R_h\bar{y} - \bar{y}_h\|_{L^2(\Omega)} \\ &\leq c_2\big(\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} + c_1\|\bar{y}_h - \bar{y}\|_{L^2(\Omega)}\big)\|\nabla(R_h\bar{y} - \bar{y}_h)\|_{L^2(\Omega)}. \end{aligned}$$

Dividing both sides of the previous inequality by $\|\nabla(R_h\bar{y} - \bar{y}_h)\|_{L^2(\Omega)}$ and using again the Poincaré's inequality yields

$$\|\bar{y}_h - R_h\bar{y}\|_{H^1(\Omega)} \leq c_2\big(\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} + c_1\|\bar{y}_h - \bar{y}\|_{L^2(\Omega)}\big), \tag{3.67}$$

where c_1, $c_2 > 0$ are constants coming from Lemma A.1.3 and the Poincaré's inequality, respectively, and they are independent of h. Here, c_1 is independent of h because $\|\bar{y}_h\|_{L^\infty(\Omega)}$ is uniformly bounded as it can be seen from the uniform convergence (3.45).
Finally, from (3.66) and (3.67) we have

$$\|\bar{y}_h - \bar{y}\|_{H^1(\Omega)} \leq c\big(\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} + \|\bar{y}_h - \bar{y}\|_{L^2(\Omega)} + h\|\bar{y}\|_{H^2(\Omega)}\big)$$

from which we obtain the desired estimate after recalling the bounds that we established for $\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)}$ and $\|\bar{y}_h - \bar{y}\|_{L^2(\Omega)}$. ■

The reason for postulating Assumption 1 in the hypothesis of Theorem 3.6.1 is that we use the estimate (3.55), which follows from Theorem 3.2.4 which requires this assumption, to bound the integrals

$$\int_{\tilde{K}} (\tilde{y}_h - \bar{y})\, d\bar{\mu}_h \quad \text{and} \quad \int_K (\tilde{y}_h - \bar{y})\, d\bar{\mu} \tag{3.68}$$

that appear in the term S_3 in the proof of the theorem. However, since both of the previous integrals are on subdomains of Ω, Remark 4 suggests the next one.

Remark 9 Assumption 1 in the hypothesis of Theorem 3.6.1 is superfluous as it can be seen from the following steps.
Step 1. We replace (3.55) by

$$\|\tilde{y}_h - \bar{y}\|_{L^\infty(\Omega_b)} \leq c|\ln h| h^{2-\frac{2}{p}} \big(\|\bar{u}\|_{L^p(\Omega)} + 1\big), \tag{3.69}$$

where $\Omega_b \subset\subset \Omega$ is as described in Remark 4.
Step 2. If we choose Ω_b in a way such that $\tilde{K} \subset\subset \Omega_b$, then the first integral in (3.68) can be bounded as follows

$$\int_{\tilde{K}} (\tilde{y}_h - \bar{y})\, d\bar{\mu}_h \leq \|\tilde{y}_h - \bar{y}\|_{L^\infty(\tilde{K})} \|\bar{\mu}_h\|_{\mathcal{M}(\tilde{K})} \leq c\|\tilde{y}_h - \bar{y}\|_{L^\infty(\Omega_b)}.$$

The second integral in (3.68) can be estimated in a similar way since $K \subset\subset \tilde{K}$.

We would like to comment on the work in [64] where the error estimate of the finite element discretization of the problem $(\mathbb{P})$ is also considered there but it was established using a different approach. In particular, it is considered there the same class of problems as $(\mathbb{P})$ with the exceptions that the nonlinearity ϕ need not satisfy (2.2) and that the control bounds are dropped. In fact, with the absence of the control constraints it becomes easy to verify that there exists a Slater point in any given $L^2(\Omega)$-neighbourhood of a local solution $\bar{u}$ of $(\mathbb{P})$. This Slater point plays an essential rule in the overall analysis there. The discrete counter part $(\mathbb{P}_h)$ of $(\mathbb{P})$ is obtained via discretizing the state and the control variables by means of piecewise linear and continuous finite elements. Under the assumption that a local solution $\bar{u}$ of $(\mathbb{P})$ satisfies a quadratic growth condition, that is, there exist constants $\delta > 0$ and $\beta > 0$ such that

$$\beta\|u - \bar{u}\|^2_{L^2(\Omega)} \leq J(u) - J(\bar{u})$$

for all feasible controls u of $(\mathbb{P})$ with $\|u - \bar{u}\|_{L^2(\Omega)} \leq \delta$ where J is the cost functional, it was shown that there exists a sequence $(\bar{u}_h)_{h>0}$ of local solutions of $(\mathbb{P}_h)$ such that $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$ as $h \to 0$ with order $O(h^{2-d/2-\varepsilon})$, for arbitrarily small $\varepsilon > 0$ in space dimensions $d = 2, 3$ for a convex polygonal or polyhedron domain Ω. This was achieved by first considering an auxiliary discrete problem $(\mathbb{P}_h^r)$ which is basically the same as $(\mathbb{P}_h)$ but the controls u_h of $(\mathbb{P}_h^r)$ are sought in a $L^2(\Omega)$-neighbourhood of the local solution $\bar{u}$ of $(\mathbb{P})$, that is, $\|u_h - \bar{u}\|_{L^2(\Omega)} \leq r$ for a fixed $r > 0$ but sufficiently small such that u_h is a feasible test function in the previous quadratic growth condition. Next, it was shown that $(\mathbb{P}_h^r)$ admits at least one global solution $\bar{u}_h^r$ and that any sequence of global solutions $(\bar{u}_h^r)_{h>0}$ converges to $\bar{u}$ in $L^2(\Omega)$ with order $O(h^{2-d/4})$. Since a global solution of $(\mathbb{P}_h^r)$ is a local solution of $(\mathbb{P}_h)$, the existence of the sequence

$(\bar{u}_h)_{h>0}$ of the local solutions of $(\mathbb{P}_h)$ is now verified and then the corresponding first necessary optimality conditions can be formulated which in turn are used to improve the convergence rate to $O(h^{2-d/2-\varepsilon})$. We remark that this convergence rate, namely $O(h^{2-d/2-\varepsilon})$, for the controls was obtained only when the state constraints are prescribed in a compact subdomain $K \subset \Omega$ and it was mentioned there that the case $K = \bar{\Omega}$ will lead to a lower order of convergence.

3.6.1 The Case $K = \bar{\Omega}$

Recall that in Section 3.5.1 we mentioned that the choice $K = \bar{\Omega}$ is possible for Problem $(\mathbb{P})$ provided that the bounds y_a, $y_b \in C(\bar{\Omega})$ satisfy in addition to $y_a < y_b$ in $\bar{\Omega}$ the compatibility condition $y_a < 0 < y_b$ on $\partial\Omega$. In the following, we discuss how this choice affects Theorem 3.6.1 and its proof.

First of all, the set $\mathcal{N}_h$ should be in this case defined as

$$\mathcal{N}_h := \{x_j | x_j \text{ is a vertex of } T \in \mathcal{T}_h, x_j \notin \partial\Omega\}.$$

Next, to insure a high order of convergence the bounds y_a, y_b should admit a higher regularity, namely,

- $y_a, y_b \in W^{2,\infty}(\Omega)$ such that $y_a < y_b$ in $\bar{\Omega}$ and $y_a < 0 < y_b$ on $\partial\Omega$.

Theorem 3.6.2 *If $K = \bar{\Omega}$ such that the previous settings hold, then the estimate* (3.57) *in Theorem 3.6.1 holds without requiring Assumption 2 nor Assumption 3.*

Proof: The proof is exactly the same as the one of Theorem 3.6.1 except for small modifications in the analysis of the term S_3. More precisely, one should use $\bar{\Omega}$ in place of $\tilde{K}$ and K there. Then, the analysis is the same as when Assumption 3 is satisfied. ■

We now address the question that whether Assumption 1 is actually needed to obtain the convergence order $O(\sqrt{|\ln h|}h^{\frac{3}{2}-\frac{1}{s}})$ for the controls if the case $K = \bar{\Omega}$ is considered.
Firstly, the integrals in (3.68) become

$$\int_{\bar{\Omega}} (\tilde{y}_h - \bar{y})\, d\bar{\mu}_h \quad \text{and} \quad \int_{\bar{\Omega}} (\tilde{y}_h - \bar{y})\, d\bar{\mu}.$$

At first glance, one should use the estimate (3.55) to bound the previous integrals since they are over the whole domain Ω and thus Assumption 1 is needed. However, using the fact that y_a, $y_b \in C(\bar{\Omega})$, $y_a < y_b$ in $\bar{\Omega}$ and $y_a < 0 < y_b$ on $\partial\Omega$, it is possible to show that there exists $\Omega_c \subset\subset \Omega$ such that $\text{supp}(\bar{\mu}) \subset \Omega_c$ and $\text{supp}(\bar{\mu}_h) \subset \Omega_c$ for h small enough, see [24, Corollary 5.4] for the proof. Consequently, it is enough to consider the estimate (3.69) to bound the previous integrals provided that Ω_b is chosen in a way such that $\Omega_c \subset\subset \Omega_b$. For instance, for the first integral we have

$$\begin{aligned}\int_{\bar{\Omega}} (\tilde{y}_h - \bar{y})\, d\bar{\mu}_h = \int_{\bar{\Omega}_c} (\tilde{y}_h - \bar{y})\, d\bar{\mu}_h &\leq \|\tilde{y}_h - \bar{y}\|_{L^\infty(\Omega_c)} \|\bar{\mu}_h\|_{\mathcal{M}(\bar{\Omega}_c)} \\ &\leq c\|\tilde{y}_h - \bar{y}\|_{L^\infty(\Omega_b)}.\end{aligned}$$

In summary, we have the next remark.

Remark 10 Theorem 3.6.2 is valid without the need for postulating Assumption 1.

We conclude this section by deriving an error bound for the uniform convergence of the optimal state $\bar{y}_h$ associated with the unique global minimum $\bar{u}_h$ of problem $(\mathbb{P}_h)$ to the optimal state $\bar{y}$ associated with the unique global minimum $\bar{u}$ of problem $(\mathbb{P})$.

Theorem 3.6.3 *Under the hypothesis of Theorem 3.6.1, apart from Assumption 1, let $\bar{y}_h$ be the state associated with the unique global minimum $\bar{u}_h$ of $(\mathbb{P}_h)$. Then there exists a constant $c > 0$ independent of h such that*

$$\|\bar{y}_h - \bar{y}\|_{L^\infty(\Omega)} \leq ch^{\frac{1}{2}},$$

where $\bar{y}$ is the state associated with the unique global minimum $\bar{u}$ of $(\mathbb{P})$. Moreover, if in addition either Assumption 2 or Assumption 3 is satisfied, or if $K = \bar{\Omega}$ under the hypothesis of Theorem 3.6.2, then

$$\|\bar{y}_h - \bar{y}\|_{L^\infty(\Omega)} \leq c\sqrt{|\ln h|}h^{\frac{3}{2}-\frac{1}{s}},$$

for any $1 < s < s_\Omega$ where s_Ω is as defined in Theorem 3.6.1.

Proof: Since $\bar{y} = \mathcal{G}(\bar{u})$ and $\bar{y}_h = \mathcal{G}_h(\bar{u}_h)$, we get from the continuous embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$, Lemma 2.2.4 and (3.12) that

$$\begin{aligned}\|\mathcal{G}_h(\bar{u}_h) - \mathcal{G}(\bar{u})\|_{L^\infty(\Omega)} &\leq \|\mathcal{G}_h(\bar{u}_h) - \mathcal{G}(\bar{u}_h)\|_{L^\infty(\Omega)} + \|\mathcal{G}(\bar{u}_h) - \mathcal{G}(\bar{u})\|_{L^\infty(\Omega)} \\ &\leq ch\big(\|\bar{u}_h\|_{L^2(\Omega)} + 1\big) + c\|\bar{u}_h - \bar{u}\|_{L^2(\Omega)},\end{aligned}$$

from which the result follows after recalling Theorem 3.6.1 and Theorem 3.6.2. The dispensing with Assumption 1 is justified in Remark 9 and in Remark 10. ■

3.7 Implementation Issues

Our aim in this section is to develop an algorithm to solve the nonlinear system (3.21)–(3.24) numerically by the semismooth Newton's method.

To start, let $\bar{\mu}_h$ be the measure defined in (3.25) and let $\bar{\mu}_h = \bar{\mu}_h^b - \bar{\mu}_h^a$ be its Jordan decomposition where we define

$$\bar{\mu}_h^b := \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j^b \delta_{x_j} \quad \text{and} \quad \bar{\mu}_h^a := \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j^a \delta_{x_j},$$

where $\bar{\mu}_j^b$, $\bar{\mu}_j^a \geq 0$ are real numbers for any corresponding $x_j \in \mathcal{N}_h$ and δ_{x_j} is the Dirac measure at x_j. Then, it is well known that the inequality (3.24) is equivalent to the system

$$\bar{\mu}_j^b \geq 0, \quad \bar{y}_h(x_j) \leq y_b(x_j),\ x_j \in \mathcal{N}_h \text{ and } \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j^b(\bar{y}_h(x_j) - y_b(x_j)) = 0,$$

$$\bar{\mu}_j^a \geq 0, \quad \bar{y}_h(x_j) \geq y_a(x_j),\ x_j \in \mathcal{N}_h \text{ and } \sum_{x_j \in \mathcal{N}_h} \bar{\mu}_j^a(y_a(x_j) - \bar{y}_h(x_j)) = 0,$$

or, equivalently,

$$\bar{\mu}_j^b = \max\Big(0, \bar{\mu}_j^b + c_j\big(\bar{y}_h(x_j) - y_b(x_j)\big)\Big), \quad x_j \in \mathcal{N}_h, \text{ for any } c_j > 0,$$
$$\bar{\mu}_j^a = \max\Big(0, \bar{\mu}_j^a + c_j\big(y_a(x_j) - \bar{y}_h(x_j)\big)\Big), \quad x_j \in \mathcal{N}_h, \text{ for any } c_j > 0,$$

where the function $\max(0,\cdot)$ is defined by

$$\max(0,x) = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Also, we recall that the inequality (3.23) is equivalent to

$$\bar{u}_h(x) = P_{[u_a,u_b]}\big(-\tfrac{1}{\alpha}\bar{p}_h(x)\big) = \min\big(\max\big(u_a, -\tfrac{1}{\alpha}\bar{p}_h(x)\big), u_b\big) \quad \forall\, x \in \Omega. \tag{3.70}$$

Consequently, we see that the system (3.21)–(3.24) can be rewritten as a set of equations for five unknowns, namely, u_h, y_h, p_h, μ_h^b and μ_h^a where $\mu_h := \mu_h^b - \mu_h^a$. In fact, if we use the equation of u_h in the one of y_h, we are left with only four unknowns, namely, y_h, p_h, μ_h^b and μ_h^a. The control u_h can be recovered easily via (3.70) once we have computed p_h. Our task now is to write the equations for the unknowns y_h, p_h, μ_h^b and μ_h^a in matrix-vector form and then establish an algorithm to solve them by the semismooth Newtons method. To achieve this, we first introduce some notation.

Let $\{x_1, \ldots, x_n\}$ be the set of the inner nodes of the triangulation $\mathcal{T}_h$ so that $n \in \mathbb{N}$ is the dimension of the space X_{h0} and let $\{\varphi_1, \ldots, \varphi_n\}$ be the corresponding basis functions. Moreover, let $m \in \mathbb{N}$ be the number of the nodes in the set $\mathcal{N}_h$. For simplicity and without loss of generality, we assume that the nodes in $\mathcal{N}_h$ are the first m nodes in the set $\{x_1, \ldots, x_n\}$. In other words, the constraints on the discrete state variable y_h are imposed only in the nodes x_j, for $j = 1, \ldots, m$. Notice that for the case $K \subset \Omega$ we have $m \leq n$ while if $K = \bar{\Omega}$, then $m = n$. Next, we define the following one-to-one correspondence between the discrete objects and their nodal representations

$$\mathbf{y} = [y_j]_{j=1}^n \in \mathbb{R}^n \Leftrightarrow y_h = \sum_{j=1}^n y_j \varphi_j,$$
$$\mathbf{p} = [p_j]_{j=1}^n \in \mathbb{R}^n \Leftrightarrow p_h = \sum_{j=1}^n p_j \varphi_j,$$
$$\boldsymbol{\mu}_b = [\mu_j^b]_{j=1}^m \in \mathbb{R}^m \Leftrightarrow \mu_h^b = \sum_{j=1}^m \mu_j^b \delta_{x_j},$$
$$\boldsymbol{\mu}_a = [\mu_j^a]_{j=1}^m \in \mathbb{R}^m \Leftrightarrow \mu_h^a = \sum_{j=1}^m \mu_j^a \delta_{x_j}.$$

We also introduce the mappings

$$\mathbb{R}^n \ni \mathbf{y} \mapsto \boldsymbol{\Phi}(\mathbf{y}) := \left[\int_\Omega \phi(y_h)\varphi_j \, dx\right]_{j=1}^n \in \mathbb{R}^n, \tag{3.71}$$

$$\mathbb{R}^n \ni \mathbf{y} \mapsto \boldsymbol{\Phi}'(\mathbf{y}) := \left[\int_\Omega \phi'(y_h)\varphi_i\varphi_j \, dx\right]_{i,j=1}^n \in \mathbb{R}^{n\times n}, \tag{3.72}$$

$$\mathbb{R}^n \ni \mathbf{p} \mapsto \mathbf{f}(\mathbf{p}) := \left[\int_\Omega P_{[u_a,u_b]}\big(-\tfrac{1}{\alpha}p_h\big)\varphi_j \, dx\right]_{j=1}^n \in \mathbb{R}^n. \tag{3.73}$$

Notice that $\mathbf{\Phi}'$ is the derivative of the map $\mathbf{\Phi}$. Next, we define

$$\mathbf{y}_a := [y_a(x_j)]_{j=1}^m, \qquad \mathbf{y}_b := [y_b(x_j)]_{j=1}^m, \qquad \mathbf{y}_0 := \left[\int_\Omega y_0\varphi_j\,dx\right]_{j=1}^n,$$

$$\mathbf{M} := \left[\int_\Omega \varphi_i\varphi_j\,dx\right]_{i,j=1}^n, \quad \mathbf{A} := \left[\int_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,dx\right]_{i,j=1}^n.$$

Using the previous notation, solving (3.21)–(3.24) is equivalent to solving for $(\mathbf{y},\mathbf{p},\boldsymbol{\mu}_b,\boldsymbol{\mu}_a)\in\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^m$ the following system

$$\mathbf{A}\mathbf{y}+\mathbf{\Phi}(\mathbf{y})=\mathbf{f}(\mathbf{p}), \tag{3.74}$$

$$\mathbf{A}\mathbf{p}+\mathbf{\Phi}'(\mathbf{y})\mathbf{p}=\mathbf{M}\mathbf{y}-\mathbf{y}_0+\mathbf{E}\boldsymbol{\mu}_b-\mathbf{E}\boldsymbol{\mu}_a, \tag{3.75}$$

$$\boldsymbol{\mu}_b=\max\big(0,\boldsymbol{\mu}_b+c(\mathbf{E}^\top\mathbf{y}-\mathbf{y}_b)\big), \tag{3.76}$$

$$\boldsymbol{\mu}_a=\max\big(0,\boldsymbol{\mu}_a+c(\mathbf{y}_a-\mathbf{E}^\top\mathbf{y})\big). \tag{3.77}$$

Here, $\mathbf{E}\in\mathbb{R}^{n\times m}$ is a matrix defined by

$$\mathbf{E}:=\begin{bmatrix}\mathbf{I}_m\\ \mathbf{0}_{r\times m}\end{bmatrix}$$

where $\mathbf{I}_m\in\mathbb{R}^{m\times m}$ is the identity matrix while $\mathbf{0}_{r\times m}\in\mathbb{R}^{r\times m}$ is the zero matrix with $r:=n-m$. Notice that $\mathbf{E}$ is the embedding of $\mathbb{R}^m$ into $\mathbb{R}^n$ by zero extension, that is,

$$\mathbb{R}^m\ni\mathbf{x}\mapsto\mathbf{E}\mathbf{x}=\begin{bmatrix}\mathbf{x}\\ 0\end{bmatrix}\in\mathbb{R}^n.$$

The presence of $\mathbf{E}$ makes the addition of the terms on the right hand side of (3.75) feasible. It is clear that in the case $m=n$, $\mathbf{E}$ becomes the identity matrix.

On the other hand, $\mathbf{E}^\top\in\mathbb{R}^{m\times n}$ is the transpose of $\mathbf{E}$ and it functions as the projection of $\mathbb{R}^n$ into $\mathbb{R}^m$ by considering the first m components, that is,

$$\mathbf{E}^\top=\begin{bmatrix}\mathbf{I}_m & \mathbf{0}_{m\times r}\end{bmatrix}$$

and

$$\mathbb{R}^n\ni\mathbf{z}\mapsto\mathbf{E}^\top\mathbf{z}=\begin{bmatrix}z_1\\ \vdots\\ z_m\end{bmatrix}\in\mathbb{R}^m.$$

Recall that, by convention, the constraints on the variable $\mathbf{y}$ are imposed only on its first m components.

The quantity c in (3.76) and in (3.77) can be chosen to be simply an arbitrary positive constant $c>0$, or a diagonal matrix $c=\operatorname{diag}(c_1,\dots,c_m)$ with arbitrary real numbers $c_j>0$ for $j=1,\dots m$. We emphasize that the choice for c in (3.76) need not be the same as the one in (3.77). Finally, for a given $\mathbf{x}\in\mathbb{R}^m$, $\max(0,\mathbf{x})$ is understood in a component wise sense, that is,

$$\max(0,\mathbf{x}):=\big[\max(0,x_j)\big]_{j=1}^m.$$

Remark 11 One could avoid using the matrix $\mathbf{E}$ and its transpose $\mathbf{E}^\top$ in (3.75)–(3.77) by simply considering $\boldsymbol{\mu}_b$, $\boldsymbol{\mu}_a$ to be elements from $\mathbb{R}^n$ instead of $\mathbb{R}^m$ and by replacing (3.76) and (3.77) by

$$\boldsymbol{\mu}_b=\max\big(0,\boldsymbol{\mu}_b+c(\mathbf{y}-\tilde{\mathbf{y}}_b)\big),\ \text{ and }\ \boldsymbol{\mu}_a=\max\big(0,\boldsymbol{\mu}_a+c(\tilde{\mathbf{y}}_a-\mathbf{y})\big), \tag{3.78}$$

respectively, where $\tilde{\mathbf{y}}_b$, $\tilde{\mathbf{y}}_a \in \mathbb{R}^n$ are the extensions of $\mathbf{y}_b$, $\mathbf{y}_a \in \mathbb{R}^m$ defined as

$$\tilde{\mathbf{y}}_b := \begin{bmatrix} \mathbf{y}_b \\ \infty \end{bmatrix}, \qquad \tilde{\mathbf{y}}_a := \begin{bmatrix} \mathbf{y}_a \\ -\infty \end{bmatrix}.$$

It is clear from (3.78) that $\mu_j^b = \mu_j^a = 0$, for $j = m+1, \dots, n$, which is exactly the same effect of $\mathbf{E}$ on the elements of $\mathbb{R}^m$. However, from practical point of view, this strategy might lead to an unnecessary increase in the dimension of the system (3.74)–(3.77) especially if $m \ll n$ which is the case when the state constraints are imposed on a relatively very small subset K of the domain Ω.

It is convenient in what follows to introduce the function

$$G : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$$

by

$$G(\mathbf{y}, \mathbf{p}, \boldsymbol{\mu}_b, \boldsymbol{\mu}_a) := \begin{bmatrix} \mathbf{A}\mathbf{y} + \boldsymbol{\Phi}(\mathbf{y}) - \mathbf{f}(\mathbf{p}) \\ \mathbf{A}\mathbf{p} + \boldsymbol{\Phi}'(\mathbf{y})\mathbf{p} - \mathbf{M}\mathbf{y} + \mathbf{y}_0 - \mathbf{E}\boldsymbol{\mu}_b + \mathbf{E}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b - \mathbf{r}_b(\mathbf{y}, \boldsymbol{\mu}_b) \\ \boldsymbol{\mu}_a - \mathbf{r}_a(\mathbf{y}, \boldsymbol{\mu}_a) \end{bmatrix},$$

where

$$\begin{aligned} \mathbf{r}_b(\mathbf{y}, \boldsymbol{\mu}_b) &:= \max\big(0, \boldsymbol{\mu}_b + c(\mathbf{E}^\top \mathbf{y} - \mathbf{y}_b)\big), \\ \mathbf{r}_a(\mathbf{y}, \boldsymbol{\mu}_a) &:= \max\big(0, \boldsymbol{\mu}_a + c(\mathbf{y}_a - \mathbf{E}^\top \mathbf{y})\big). \end{aligned}$$

Then, solving (3.74)–(3.77) is equivalent to solving the equation

$$G(\mathbf{y}, \mathbf{p}, \boldsymbol{\mu}_b, \boldsymbol{\mu}_a) = 0. \tag{3.79}$$

Since the functions $\max(0, \cdot)$ and $P_{[u_a, u_b]}(\cdot)$ are not differentiable in a classical sense, the function G is nonsmooth and hence the Newton's method can't be applied directly to solve (3.79). Instead, we utilize the semismooth Newton's method, which uses the generalized Jacobian of G, to solve (3.79), see for instance [45, Chapter 2] for more details about the semimooth Newton's method.

The generalized Jacobian of G at a given point $(\mathbf{y}, \mathbf{p}, \boldsymbol{\mu}_b, \boldsymbol{\mu}_a) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ reads

$$\begin{aligned} &DG(\mathbf{y}, \mathbf{p}, \boldsymbol{\mu}_b, \boldsymbol{\mu}_a) = \\ &= \begin{bmatrix} \mathbf{A} + \boldsymbol{\Phi}'(\mathbf{y}) & -D\mathbf{f}(\mathbf{p}) & 0 & 0 \\ \boldsymbol{\Phi}''(\mathbf{y})\mathbf{p} - \mathbf{M} & \mathbf{A} + \boldsymbol{\Phi}'(\mathbf{y}) & -\mathbf{E} & \mathbf{E} \\ -D_{\mathbf{y}}\mathbf{r}_b(\mathbf{y}, \boldsymbol{\mu}_b) & 0 & \mathbf{I}_m - D_{\boldsymbol{\mu}_b}\mathbf{r}_b(\mathbf{y}, \boldsymbol{\mu}_b) & 0 \\ -D_{\mathbf{y}}\mathbf{r}_a(\mathbf{y}, \boldsymbol{\mu}_a) & 0 & 0 & \mathbf{I}_m - D_{\boldsymbol{\mu}_a}\mathbf{r}_a(\mathbf{y}, \boldsymbol{\mu}_a) \end{bmatrix}. \end{aligned}$$

Here, $D\mathbf{f}$ denotes the generalized Jacobian of $\mathbf{f}$ and it is given by

$$D\mathbf{f}(\mathbf{p}) = \Big[\int_\Omega -\tfrac{1}{\alpha} \partial P_{[u_a, u_b]}\big(-\tfrac{1}{\alpha} p_h\big) \varphi_i \varphi_j \, dx \Big]_{i,j=1}^n \tag{3.80}$$

where $\partial P_{[u_a,u_b]}(\cdot)$ denotes the generalized derivative of $P_{[u_a,u_b]}(\cdot)$, that is,

$$\partial P_{[u_a,u_b]}(x) := \begin{cases} \{0\} & x < u_a \text{ or } x > u_b, \\ \{1\} & u_a < x < u_b, \\ [0,1] & x = u_a \text{ or } x = u_b. \end{cases} \tag{3.81}$$

Furthermore, $\mathbf{\Phi}''$ is the second derivative of the mapping $\mathbf{\Phi}$. More precisely, we have

$$\mathbf{\Phi}''(\mathbf{y})\mathbf{p} = \left[\int_\Omega \phi''(y_h) p_h \varphi_i \varphi_j \, dx \right]_{i,j=1}^n. \tag{3.82}$$

Finally, we have

$$\begin{aligned} D_{\mathbf{y}}\mathbf{r}_b(\mathbf{y},\boldsymbol{\mu}_b) &= cD\max\big(0, \boldsymbol{\mu}_b + c(\mathbf{E}^\top\mathbf{y} - \mathbf{y}_b)\big)\mathbf{E}^\top, \\ D_{\boldsymbol{\mu}_b}\mathbf{r}_b(\mathbf{y},\boldsymbol{\mu}_b) &= D\max\big(0, \boldsymbol{\mu}_b + c(\mathbf{E}^\top\mathbf{y} - \mathbf{y}_b)\big), \\ D_{\mathbf{y}}\mathbf{r}_a(\mathbf{y},\boldsymbol{\mu}_a) &= -cD\max\big(0, \boldsymbol{\mu}_a + c(\mathbf{y}_a - \mathbf{E}^\top\mathbf{y})\big)\mathbf{E}^\top, \\ D_{\boldsymbol{\mu}_a}\mathbf{r}_a(\mathbf{y},\boldsymbol{\mu}_a) &= D\max\big(0, \boldsymbol{\mu}_a + c(\mathbf{y}_a - \mathbf{E}^\top\mathbf{y})\big), \end{aligned}$$

where, for a given $\mathbf{x} \in \mathbb{R}^m$, $D\max(0,\mathbf{x})$ is a diagonal matrix given by

$$D\max(0,\mathbf{x}) = \operatorname{diag}\Big(\partial\max(0,x_1), \ldots, \partial\max(0,x_m)\Big)$$

with $\partial\max(0,\cdot)$ being the generalized derivative of $\max(0,\cdot)$, that is,

$$\partial\max(0,x) := \begin{cases} \{0\} & x < 0, \\ \{1\} & x > 0, \\ [0,1] & x = 0. \end{cases} \tag{3.83}$$

Notice that (3.81) and (3.83) are set-valued mappings which implies that there might be several choices for $DG(\mathbf{y},\mathbf{p},\boldsymbol{\mu}_b,\boldsymbol{\mu}_a)$ at a given point $(\mathbf{y},\mathbf{p},\boldsymbol{\mu}_b,\boldsymbol{\mu}_a) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. In general, to guarantee the convergence of the semismooth Newton's method to a solution $(\bar{\mathbf{y}},\bar{\mathbf{p}},\bar{\boldsymbol{\mu}}_b,\bar{\boldsymbol{\mu}}_a)$ of the system (3.79), i.e., $G(\bar{\mathbf{y}},\bar{\mathbf{p}},\bar{\boldsymbol{\mu}}_b,\bar{\boldsymbol{\mu}}_a) = 0$, we assume that for all points $(\mathbf{y},\mathbf{p},\boldsymbol{\mu}_b,\boldsymbol{\mu}_a)$ sufficiently close to the solution $(\bar{\mathbf{y}},\bar{\mathbf{p}},\bar{\boldsymbol{\mu}}_b,\bar{\boldsymbol{\mu}}_a)$ any choice for $DG(\mathbf{y},\mathbf{p},\boldsymbol{\mu}_b,\boldsymbol{\mu}_a)$ is invertible, see [45, Chapter 2] for further details.

We are now ready to state the algorithm for solving the system (3.79) by the semismooth Newton's method.

Algorithm 1 (Semismooth Newton's method)

0. Choose $(\mathbf{y}^0,\mathbf{p}^0,\boldsymbol{\mu}_b^0,\boldsymbol{\mu}_a^0) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ (sufficiently close to the solution $(\bar{\mathbf{y}},\bar{\mathbf{p}},\bar{\boldsymbol{\mu}}_b,\bar{\boldsymbol{\mu}}_a)$).

For $k = 0, 1, 2, \ldots$:

1. Obtain $(\delta\mathbf{y}^k,\delta\mathbf{p}^k,\delta\boldsymbol{\mu}_b^k,\delta\boldsymbol{\mu}_a^k) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ by solving

$$DG(\mathbf{y}^k,\mathbf{p}^k,\boldsymbol{\mu}_b^k,\boldsymbol{\mu}_a^k) \begin{bmatrix} \delta\mathbf{y}^k \\ \delta\mathbf{p}^k \\ \delta\boldsymbol{\mu}_b^k \\ \delta\boldsymbol{\mu}_a^k \end{bmatrix} = -G(\mathbf{y}^k,\mathbf{p}^k,\boldsymbol{\mu}_b^k,\boldsymbol{\mu}_a^k).$$

2. Set $\mathbf{y}^{k+1} = \delta\mathbf{y}^k + \mathbf{y}^k$, $\mathbf{p}^{k+1} = \delta\mathbf{p}^k + \mathbf{p}^k$, $\boldsymbol{\mu}_b^{k+1} = \delta\boldsymbol{\mu}_b^k + \boldsymbol{\mu}_b^k$, $\boldsymbol{\mu}_a^{k+1} = \delta\boldsymbol{\mu}_a^k + \boldsymbol{\mu}_a^k$.

We conclude this section by a general remark.

Remark 12 1. For a given $\mathbf{p} \in \mathbb{R}^n$, the vector $\mathbf{f}(\mathbf{p})$ introduced in (3.73) can be written as

$$\begin{aligned}\mathbf{f}(\mathbf{p}) &= \left[\int_{\mathcal{A}(\mathbf{p})} u_a \varphi_j \, dx\right]_{j=1}^n - \frac{1}{\alpha}\left[\int_{\mathcal{I}(\mathbf{p})} p_h \varphi_j \, dx\right]_{j=1}^n + \left[\int_{\mathcal{B}(\mathbf{p})} u_b \varphi_j \, dx\right]_{j=1}^n \\ &= \mathbf{M}_{\mathcal{A}(\mathbf{p})}\mathbf{u}_a - \frac{1}{\alpha}\mathbf{M}_{\mathcal{I}(\mathbf{p})}\mathbf{p} + \mathbf{M}_{\mathcal{B}(\mathbf{p})}\mathbf{u}_b,\end{aligned}$$

where

$$\begin{aligned}&\mathcal{A}(\mathbf{p}) := \{x \in \Omega : -\tfrac{1}{\alpha}p_h(x) \le u_a\}, \quad \mathcal{B}(\mathbf{p}) := \{x \in \Omega : -\tfrac{1}{\alpha}p_h(x) \ge u_b\}, \\ &\mathcal{I}(\mathbf{p}) := \{x \in \Omega : u_a < -\tfrac{1}{\alpha}p_h(x) < u_a\},\end{aligned}$$

and

$$\begin{aligned}&\mathbf{u}_a := [u_a]_{j=1}^n, \quad \mathbf{u}_b := [u_b]_{j=1}^n, \quad \mathbf{M}_{\mathcal{I}(\mathbf{p})} := \left[\int_{\mathcal{I}(\mathbf{p})} \varphi_i \varphi_j \, dx\right]_{i,j=1}^n, \\ &\mathbf{M}_{\mathcal{A}(\mathbf{p})} := \left[\int_{\mathcal{A}(\mathbf{p})} \varphi_i \varphi_j \, dx\right]_{i,j=1}^n, \quad \mathbf{M}_{\mathcal{B}(\mathbf{p})} := \left[\int_{\mathcal{B}(\mathbf{p})} \varphi_i \varphi_j \, dx\right]_{i,j=1}^n.\end{aligned}$$

In a similar way, for the matrix $D\mathbf{f}(\mathbf{p})$ defined in (3.80) we have

$$D\mathbf{f}(\mathbf{p}) = -\frac{1}{\alpha}\left[\int_{\mathcal{I}(\mathbf{p})} \varphi_i \varphi_j \, dx\right]_{i,j=1}^n = -\frac{1}{\alpha}\mathbf{M}_{\mathcal{I}(\mathbf{p})}.$$

Since p_h is a continuous and piecewise linear function on $\bar{\Omega}$, the sets $\mathcal{A}(\mathbf{p})$, $\mathcal{B}(\mathbf{p})$ and $\mathcal{I}(\mathbf{p})$ are polygonal and thus the entries of the matrices $\mathbf{M}_{\mathcal{A}(\mathbf{p})}$, $\mathbf{M}_{\mathcal{B}(\mathbf{p})}$ and $\mathbf{M}_{\mathcal{I}(\mathbf{p})}$ can be computed exactly. The determination of the sets $\mathcal{A}(\mathbf{p})$, $\mathcal{B}(\mathbf{p})$ and $\mathcal{I}(\mathbf{p})$ is a very classical task in variational discretization, see [46], and it can be achieved simply by looping over the triangles in the mesh and comparing $-\frac{1}{\alpha}p_h$ with the bounds u_a and u_b.

2. It is also possible to compute the entries of $\mathbf{\Phi}(\mathbf{y})$, $\mathbf{\Phi}'(\mathbf{y})$ and $\mathbf{\Phi}''(\mathbf{y})\mathbf{p}$ defined by (3.71), (3.72) and (3.82), respectively, exactly for certain types of the nonlinearity ϕ. For example, for the choices $\phi(s) := s^3$ and $\phi(s) := s^5$, which we will consider for our numerical examples in Section 3.8, the functions $\phi(y_h)$, $\phi'(y_h)$ and $\phi''(y_h)$ restricted to each triangle in the mesh are polynomials.

3.8 Numerical Examples

In this section we consider variational discretization of the problem ($\mathbb{P}$) for different choices of the nonlinearity ϕ and the data y_0, u_a, u_b, y_a, y_b, α, while $\Omega := (0,1) \times (0,1)$ is kept fixed in all considered examples. The numerical solution of the corresponding systems (3.21)–(3.24) is performed with the semismooth Newton method proposed in Section 3.7.

3.8.1 Examples with Unique Global Minima

In this part all the computations are performed on a uniform triangulation of $\bar{\Omega}$ with mesh size $h = 2^{-5}\sqrt{2}$.

Example 1 We start with the example presented in [64, Section 7], where the nonlinearity is given by $\phi(s) = s^3$, and unilateral state constraints of the form $y \geq y_a$ are considered, with

$$y_a(x) := -\frac{1}{2} + \frac{1}{2}\min(x_1 + x_2, 1 + x_1 - x_2, 1 - x_1 + x_2, 2 - x_1 - x_2).$$

The desired state is given by $y_0 := -1$. It is easy to see that the nonlinearity satisfies (2.2) with $r = 2$ and $M = 2\sqrt{3}$. Hence, in view of Theorem 3.3.4 we have $q = 4$ and a control $\bar{u}_h$ obtained from solving (3.21)–(3.24) is a global minimum of $(\mathbb{P}_h)$ if the associated adjoint state $\bar{p}_h$ satisfies

$$\|\bar{p}_h\|_{L^4(\Omega)} \leq 5^{-\frac{5}{8}} 3^{\frac{3}{8}} \sqrt{2} C_4^{-1} \alpha^{\frac{3}{8}} =: \eta(\alpha),$$

where $C_4 = C_4^{(3)} \approx 0.648027075$ is the constant from Theorem A.5.1.

The numerical findings for this example with varying α are presented in Table 3.1 and they are illustrated graphically for the case $\alpha = 10^{-2}$ in Figure 3.2. It is clear that $\bar{u}_h$ is a global minimum for the given values of α, and that it is most likely that the related continuous problem admits a unique global solution for all values of $\alpha > 0$. We will consider this example again in the next subsection to investigate the convergence rates.

Table 3.1 Example 1 from [64, Section 7]: The values of $\|\bar{p}_h\|_{L^4}$, $\eta(\alpha)$ and $J(\bar{u}_h)$ for different values of α.

α	$\|\bar{p}_h\|_{L^4}$	$\eta(\alpha)$	$J(\bar{u}_h)$
1.0e-06	1.782757474150e-04	6.776197632762e-03	3.151530945981e-01
1.0e-05	6.588913644647e-04	1.606889689070e-02	3.337051126085e-01
1.0e-04	2.476579707110e-03	3.810535956559e-02	3.680280814272e-01
1.0e-03	9.826749689797e-03	9.036204771862e-02	4.264978699222e-01
1.0e-02	2.919304462314e-02	2.142821839497e-01	4.866764267990e-01
1.0e-01	3.012357097331e-02	5.081431366100e-01	4.986253239639e-01
1.0e+00	3.021974315533e-02	1.204997272869e+00	4.998620947749e-01
1.0e+01	3.022939247330e-02	2.857498848277e+00	4.999862050864e-01
1.0e+02	3.023035772702e-02	6.776197632762e+00	4.999986204647e-01
1.0e+03	3.023045425561e-02	1.606889689070e+01	4.999998620460e-01

We now examine a series of numerical examples with different constraints and desired states, where we consider the nonlinearities $\phi(s) := s^3$ and $\phi(s) := s^5$. For the desired state y_0 we make the following two choices

$$\textbf{A1}: \quad y_0(x) := 2\sin(2\pi x_1)\sin(2\pi x_2),$$
$$\textbf{A2}: \quad y_0(x) := 60 + 160(x_1(x_1 - 1) + x_2(x_2 - 1)).$$

We note that in choice **A1** the desired state y_0 vanishes on the boundary $\partial\Omega$ of the domain, i.e. it satisfies the boundary condition of the state equation and

thus is reachable, while in choice **A2** it is not reachable. As constraints we consider the cases
Case 1 (unconstrained problem)

$$u_b = -u_a = \infty,\ y_b = -y_a = \infty.$$

Case 2 (constrained control)

$$u_a = -5, u_b = 5,\ y_b = -y_a = \infty.$$

Case 3 (constrained state)

$$u_b = -u_a = \infty,\ y_a = -1, y_b = 1.$$

Example 2 Let us first consider $\phi(s) := s^3$ with η specified as above. The numerical findings are summarized in Figure 3.3, where we compare $\|\bar{p}_h\|_{L^4}$ with $\eta(\alpha)$. We conclude that unique global solutions exist more likely in the case where the desired state is reachable.

Example 3 In this example we consider $\phi(s) := s^5$. We see that (2.2) is satisfied with

$$r = \frac{4}{3} \quad \text{and} \quad M = \frac{20}{5^{\frac{3}{4}}}.$$

Hence, in view of Theorem 3.3.4 we have $q = 6$ and a control $\bar{u}_h$ obtained from solving (3.21)–(3.24) is a global minimum if the associated adjoint state $\bar{p}_h$ satisfies

$$\|\bar{p}_h\|_{L^6(\Omega)} \leq \frac{11^{\frac{11}{24}}}{13^{\frac{13}{24}} 2^{\frac{1}{6}} \sqrt{3}} C_6^{-\frac{1}{2}} \alpha^{\frac{11}{24}} =: \eta(\alpha),$$

where $C_6 = C_6^{(1)} \approx 0.610888$ is the constant from Theorem A.5.1.

The numerical findings are summarized in Figure 3.4. Again we can conclude that problems with reachable desired states more likely admit unique global solutions.

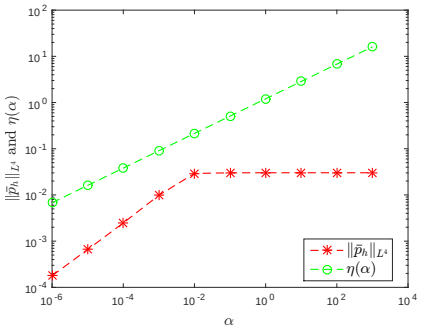
3.8.2 Convergence Rates

We now examine numerically the error bounds established in Theorem 3.6.1 and in Theorem 3.6.3. For this purpose, we consider again Example 1 with $\alpha = 10^{-2}$.

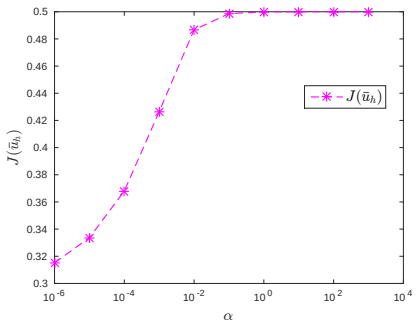
The numerical computations, which are illustrated in Figure 3.2, show that the state constraints are active at one point, namely $\tilde{x} := (\frac{1}{2}, \frac{1}{2})$, and the corresponding multiplier is approximately given by

$$\bar{\mu}_h^a = 0.3386\, \delta_{\tilde{x}},$$

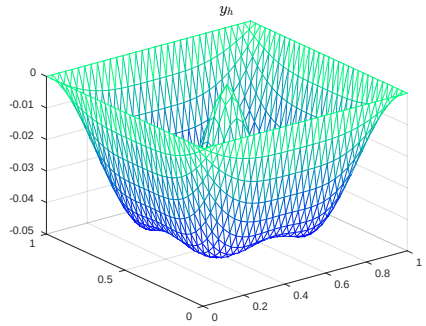
where $\delta_{\tilde{x}}$ is a Dirac measure at $\tilde{x}$. We can easily find a polygonal subdomain $K \subset\subset \Omega$ that contains the active point $\tilde{x}$ so that Assumption 3 holds. Recall that Assumption 1 can be dropped in the light of Remark 9. Consequently, we are expecting the bound $\sqrt{|\ln h|} h^{\frac{3}{2} - \frac{1}{s}}$, or equivalently $h^{1-\varepsilon}$ for arbitrarily small $\varepsilon > 0$, for the computed errors according to Theorem 3.6.1 and Theorem 3.6.3.



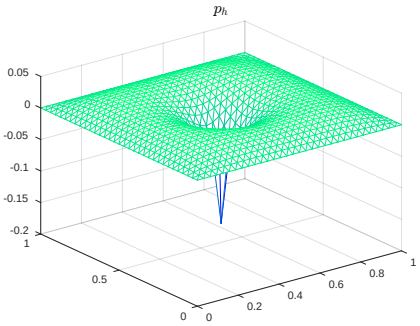
(a) $\|\bar{p}_h\|_{L^4}$ and $\eta(\alpha)$ vs. α.



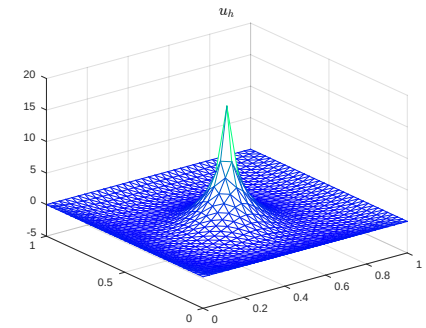
(b) $J(\bar{u}_h)$ vs. α.



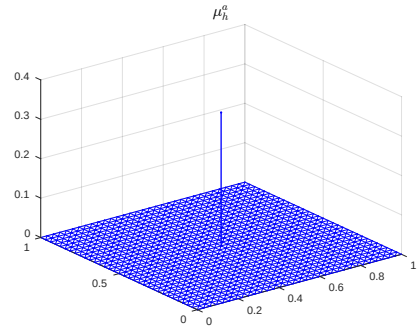
(c) The optimal state $\bar{y}_h$.



(d) The adjoint state $\bar{p}_h$.

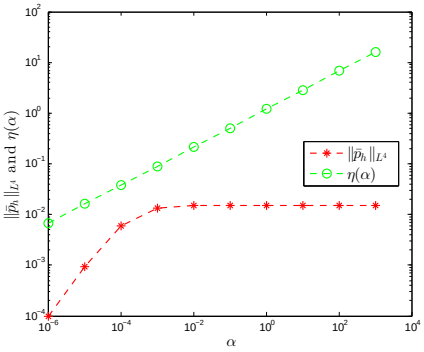


(e) The optimal control $\bar{u}_h$.

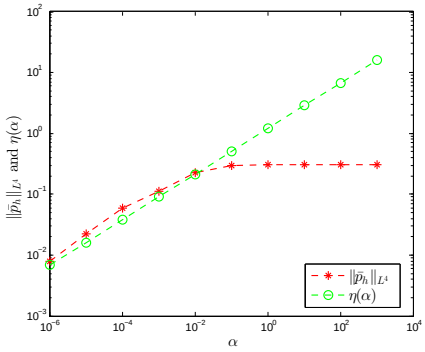


(f) The multiplier $\bar{\mu}_h^a = 0.3386\,\delta_{\tilde{x}}$, $\tilde{x} := (\frac{1}{2}, \frac{1}{2})$.

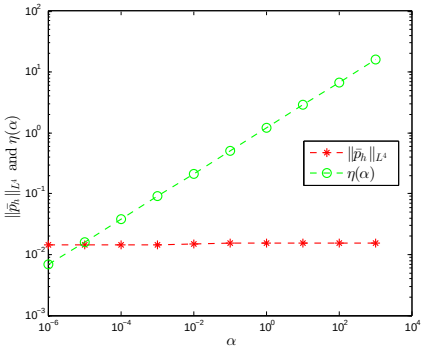
Figure 3.2 Example 1 from [64, Section 7]: The values of $\|\bar{p}_h\|_{L^4}$, $\eta(\alpha)$ and $J(\bar{u}_h)$ vs. α. The optimal state $\bar{y}_h$, the optimal control $\bar{u}_h$, the adjoint state $\bar{p}_h$ and the multiplier $\bar{\mu}_h^a$ for $\alpha = 10^{-2}$.



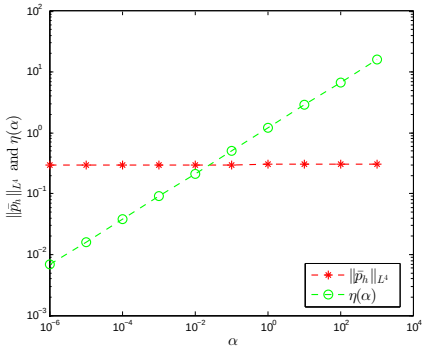
(a) Case 1 with A1



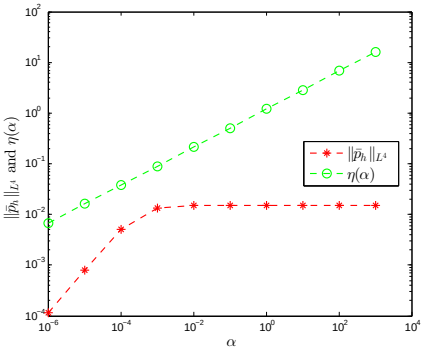
(b) Case 1 with A2



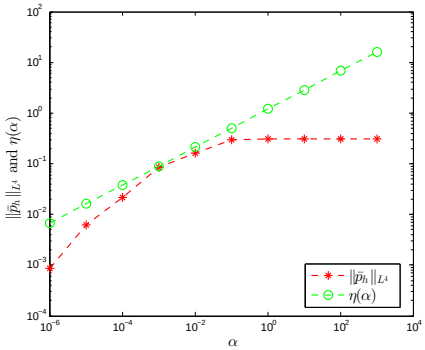
(c) Case 2 with A1



(d) Case 2 with A2

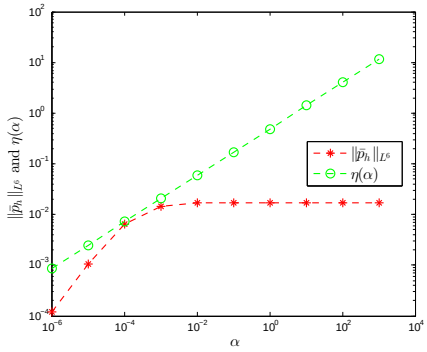


(e) Case 3 with A1

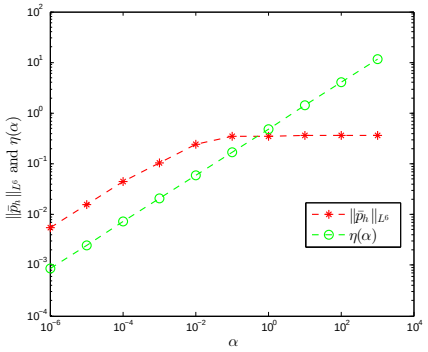


(f) Case 3 with A2

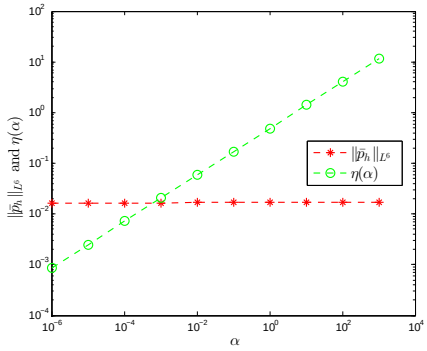
Figure 3.3 Results for $\phi(s) = s^3$



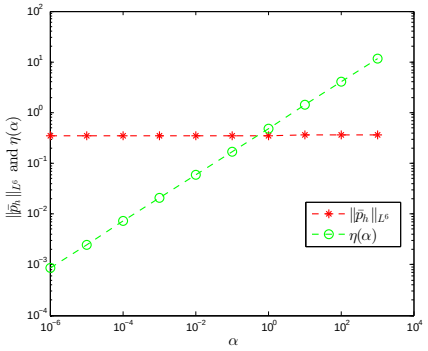
(a) Case 1 with A1



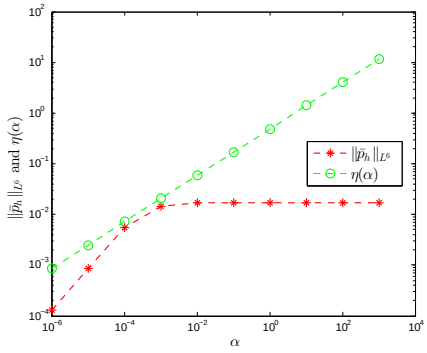
(b) Case 1 with A2



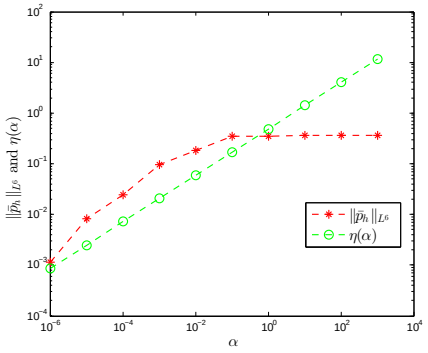
(c) Case 2 with A1



(d) Case 2 with A2



(e) Case 3 with A1



(f) Case 3 with A2

Figure 3.4 Results for $\phi(s) = s^5$

To deduce the convergence rates numerically, we compute the experimental order of convergence (EOC) which is defined as

$$\text{EOC} := \frac{\log E(h_i) - \log E(h_{i-1})}{\log h_i - \log h_{i-1}},$$

where E is a given positive error functional and h_{i-1}, h_i are two consecutive mesh sizes. For our experiment, we consider the error functionals

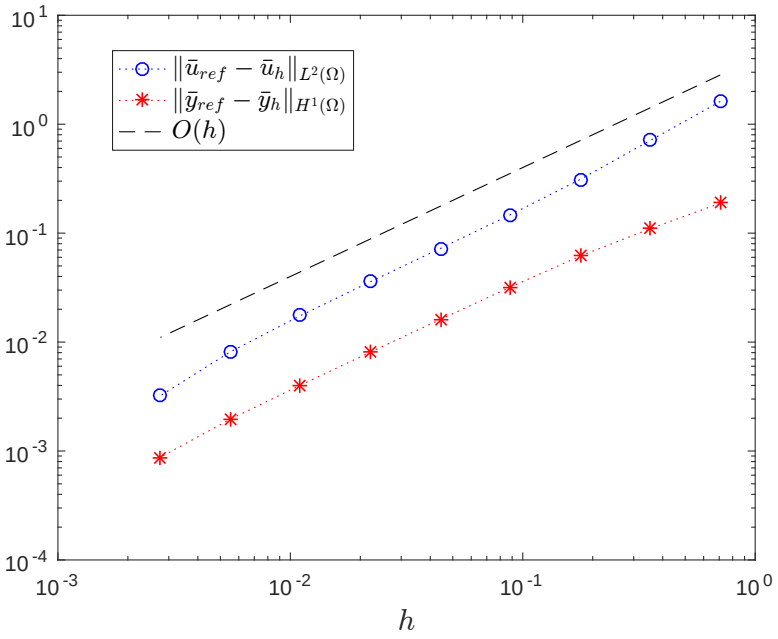
$$\begin{aligned} E_{u_{L2}}(h_i) &:= \|\bar{u}_{ref} - \bar{u}_{h_i}\|_{L^2(\Omega)}, \\ E_{y_{H1}}(h_i) &:= \|\bar{y}_{ref} - \bar{y}_{h_i}\|_{H^1(\Omega)}, \\ E_{y_{L2}}(h_i) &:= \|\bar{y}_{ref} - \bar{y}_{h_i}\|_{L^2(\Omega)}, \\ E_{y_{L\infty}}(h_i) &:= \|\bar{y}_{ref} - \bar{y}_{h_i}\|_{L^\infty(\Omega)}, \end{aligned}$$

and denote the corresponding experimental orders of convergence by $\text{EOC}_{u_{L2}}$, $\text{EOC}_{y_{H1}}$, $\text{EOC}_{y_{L2}}$ and $\text{EOC}_{y_{L\infty}}$, respectively. Furthermore, we consider the sequence of mesh sizes $h_i = 2^{-i}\sqrt{2}$, for $i = 1, \ldots, 9$. Since we don't have the exact solution of Example 1 at hand, we consider the numerical solution computed at mesh size $h_{10} = 2^{-10}\sqrt{2}$ to be the reference solution, that is, we define $\bar{u}_{ref} := \bar{u}_{h_{10}}$ and $\bar{y}_{ref} := \bar{y}_{h_{10}}$.

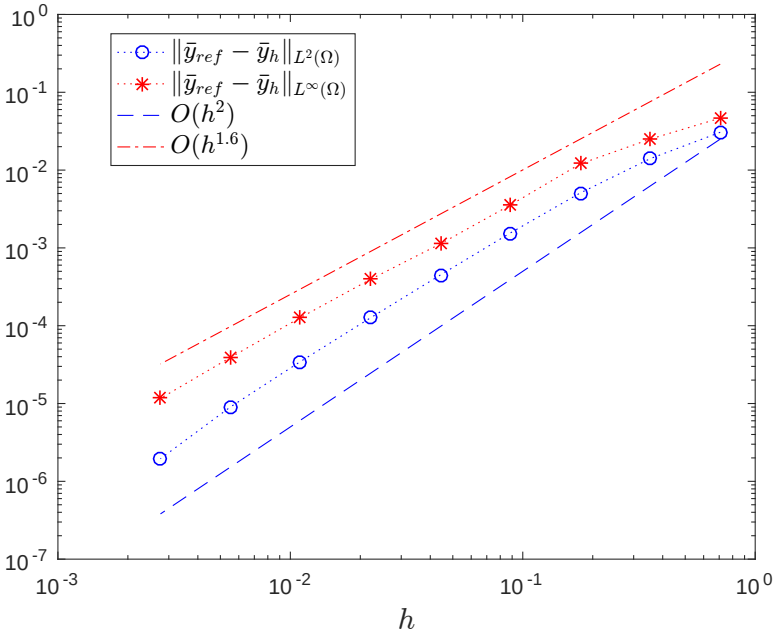
In Table 3.2 we report the values of the previous error functionals at different mesh sizes. The plots of these values versus the corresponding mesh sizes are illustrated in Figure 3.5. The computed values of the associated EOC are presented in Table 3.3.

From the numerical findings we see that as the mesh size h decreases the errors $E_{u_{L2}}(h)$ and $E_{y_{H1}}(h)$ behave like $O(h)$ which indicates that the convergence rate, namely $O(h^{1-\varepsilon})$ for arbitrarily small $\varepsilon > 0$, predicted in Theorem 3.6.1 is optimal. On the other hand, for $E_{y_{L2}}(h)$ and $E_{y_{L\infty}}(h)$ we see the behaviour $O(h^2)$ and $O(h^{1.6})$, respectively, from which we conclude that the error bounds for the discrete optimal state in the spaces $L^2(\Omega)$ and $L^\infty(\Omega)$ which are deduced from the error bound of the discrete optimal control via the Lipschitz continuity of the control-to-state map are not sharp.

In fact, the $O(h^2)$ behaviour of $E_{y_{L2}}(h)$ could be explained in the light of the work [65] where it was shown there that for an elliptic control problem with finitely many pointwise inequality state constraints the error of the discrete optimal state in $L^2(\Omega)$ is of order h^{4-d} up to logarithmic factor in $d = 2$ or $d = 3$ space dimensions when the control problem is discretized by piecewise linear and continuous finite elements.



(a) $E_{u_{L2}}(h)$ and $E_{y_{H1}}(h)$ v.s. h.



(b) $E_{y_{L2}}(h)$ and $E_{y_{L\infty}}(h)$ v.s. h.

Figure 3.5 Errors for the optimal control and its state of Example 1 with $\alpha = 10^{-2}$ versus the mesh size.

Table 3.2 Errors for the optimal control and its state of Example 1 with $\alpha = 10^{-2}$.

$h/\sqrt{2}$	$E_{u_{L2}}(h)$	$E_{y_{H1}}(h)$	$E_{y_{L2}}(h)$	$E_{y_{L\infty}}(h)$
2^{-1}	1.6151338799381	0.1881784634445	0.0305960611799	0.0465574255352
2^{-2}	0.7094890598326	0.1098931145143	0.0140288901847	0.0250260522632
2^{-3}	0.3114790933874	0.0616536701645	0.0050825086327	0.0120171679655
2^{-4}	0.1475243025114	0.0319521773619	0.0015547890304	0.0035520156534
2^{-5}	0.0723799608480	0.0161391228724	0.0004482149627	0.0011256351724
2^{-6}	0.0357734199802	0.0080807758796	0.0001259623551	0.0003940261492
2^{-7}	0.0174753747282	0.0040194139919	0.0000345959726	0.0001271705046
2^{-8}	0.0081450867872	0.0019615865181	0.0000090359657	0.0000390725086
2^{-9}	0.0032211298694	0.0008772740328	0.0000019527627	0.0000116353242

Table 3.3 EOC for the optimal control and its state of Example 1 with $\alpha = 10^{-2}$.

Levels	$\text{EOC}_{u_{L2}}$	$\text{EOC}_{y_{H1}}$	$\text{EOC}_{y_{L2}}$	$\text{EOC}_{y_{L\infty}}$
1-2	1.186801	0.776001	1.124945	0.895581
2-3	1.187645	0.833842	1.464788	1.058334
3-4	1.078183	0.948273	1.708822	1.758387
4-5	1.027290	0.985352	1.794456	1.657899
5-6	1.016702	0.997996	1.831198	1.514376
6-7	1.033565	1.007509	1.864317	1.631527
7-8	1.101321	1.034964	1.936853	1.702538
8-9	1.338363	1.160921	2.210162	1.747642

Part II

Chapter 4

Optimal Control of Elliptic PDEs with Random Coefficients

In this chapter we consider a parameter-dependent optimal control problem. More precisely, we minimize a quadratic cost functional subject to a linear elliptic PDE with homogeneous Dirichlet boundary condition where the diffusion coefficient is a random field defined on a given probability space. Our aim is to carry out the error analysis and the computational cost associated to the computation of the expectation of the solutions of this problem.

The exposition in this chapter is organized as follows: in Section 4.1, we establish the notation and the problem setting. In Section 4.2, we study the state equation. In Section 4.3, we study the optimal control problem and construct the map that assigns to each realization of the diffusion coefficient the corresponding solution of the control problem.

4.1 The Problem Setting

4.1.1 Notation

In the sequel, $(\Omega, \mathcal{A}, \mathbb{P})$ denotes a probability space, where Ω is a sample space, $\mathcal{A} \subset 2^{\Omega}$ is a σ-algebra and $\mathbb{P} : \mathcal{A} \to [0,1]$ is a probability measure. Given a Banach space $(X, \|\cdot\|_X)$, the space $L^p(\Omega, X)$ is the set of strongly measurable functions $v : \Omega \to X$ such that $\|v\|_{L^p(\Omega,X)} < \infty$, where

$$\|v\|_{L^p(\Omega,X)} := \begin{cases} \left(\int_\Omega \|v(\omega)\|_X^p d\mathbb{P}(\omega) \right)^{1/p} & \text{for } 1 \leq p < \infty, \\ \operatorname{esssup}_{\omega\in\Omega} \|v(\omega)\|_X & \text{for } p = \infty. \end{cases}$$

For $L^p(\Omega, \mathbb{R})$ we write $L^p(\Omega)$. For a bounded Lipschitz domain $D \subset \mathbb{R}^d$, the spaces $C^0(\bar{D})$ and $C^1(\bar{D})$ are the usual spaces of uniformly continuous functions and continuously differentiable functions, respectively, with their standard norms. The space $C^t(\bar{D})$ with $0 < t < 1$ denotes the space of Hölder continuous

functions with the norm

$$\|v\|_{C^t(\bar{D})} := \sup_{x\in\bar{D}} |v(x)| + \sup_{x,y\in\bar{D},x\neq y} \frac{|v(x)-v(y)|}{|x-y|^t}.$$

For $k \in \mathbb{N}$, the space $H^k(D)$ is the classical Sobolev space, on which we define the norm and semi-norm, respectively, as

$$\|v\|_{H^k(D)} := \left(\sum_{|\alpha|\leq k} \int_D |D^\alpha v|^2 \, dx\right)^{1/2} \quad \text{and} \quad |v|_{H^k(D)} := \left(\sum_{|\alpha|=k} \int_D |D^\alpha v|^2 \, dx\right)^{1/2}.$$

Recall that, for bounded domains $D \subset \mathbb{R}^d$, the norm $\|\cdot\|_{H^k(D)}$ and semi-norm $|\cdot|_{H^k(D)}$ are equivalent on the subspace $H_0^k(D)$ of $H^k(D)$. For $r := k + s$ with $0 < s < 1$ and $k \in \mathbb{N}$, we denote by $H^r(D)$ the space of all functions $v \in H^k(D)$ such that $\|v\|_{H^r(D)} < \infty$, where the norm $\|\cdot\|_{H^r(D)}$ is defined by

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

Finally, for any two positive quantities a and b, we write

$$a \lesssim b$$

to indicate that $\frac{a}{b}$ is uniformly bounded by a positive constant independent of the realization $\omega \in \Omega$ or the discretization parameter h.

4.1.2 The Problem Setting

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. For $\mathbb{P}$-a.s. $\omega \in \Omega$, we consider the following optimal control problem

$$\min_{(y,u)\in H_0^1(D)\times L^2(D)} J(y,u) = \frac{1}{2}\|y - z\|^2_{L^2(D)} + \frac{\alpha}{2}\|u\|^2_{L^2(D)} \tag{4.1}$$

subject to

$$\begin{aligned} -\nabla \cdot (a(\omega, x)\nabla y(\omega, x)) &= u(x) \quad \text{in } D, \\ y(\omega, x) &= 0 \quad \text{on } \partial D, \end{aligned} \tag{4.2}$$

and

$$u_a \leq u(x) \leq u_b \qquad \text{for a.e. } x \in D,$$

where we assume

- $D \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a bounded, convex and polyhedral domain.
- $u_a \in \mathbb{R} \cup \{-\infty\}$ and $u_b \in \mathbb{R} \cup \{\infty\}$ with $u_a \leq u_b$.
- $z \in L^2(D)$ is a given deterministic function and $\alpha > 0$ is a given real number.

The differential operators $\nabla\cdot$ and ∇ are with respect to $x \in D$. The σ-algebra $\mathcal{A}$ associated with Ω is generated by the random variables $\{a(\cdot, x)\colon x \in D\}$. Let us formally define for all $\omega \in \Omega$,

$$a_{\min}(\omega) := \min_{x\in\bar{D}} a(\omega, x) \qquad \text{and} \qquad a_{\max}(\omega) := \max_{x\in\bar{D}} a(\omega, x).$$

We make the following assumptions on the coefficient a.

A1. $a \in L^p(\Omega, C^t(\overline{D}))$ for some $0 < t \leq 1$ and for all $p \in [1, \infty)$,

A2. $a_{\min} \geq 0$ almost surely and $a_{\min}^{-1} \in L^p(\Omega)$, for all $p \in [1, \infty)$.

From Assumption A1 we see that the quantities $a_{\min}$ and $a_{\max}$ are well defined and that $a_{\max} \in L^p(\Omega)$ for all $p \in [1, \infty)$ since $a_{\max}(\omega) = \|a(\omega, \cdot)\|_{C^0(\bar{D})}$. Moreover, together with Assumption A2, we have $a_{\min}(\omega) > 0$ and $a_{\max}(\omega) < \infty$ for almost all $\omega \in \Omega$.

An example for a coefficient a that satisfies Assumptions A1–A2 is a log-normal random field $a(\omega, x) = e^{g(\omega,x)}$ where $g : \Omega \times \bar{D} \to \mathbb{R}$ is a Gaussian field with a Hölder-continuous mean function $\bar{g}(x) := \mathbb{E}[g(\omega, x)]$ and a Lipschitz continuous covariance function

$$C(x, y) := \mathbb{E}\Big[\big(g(\omega, x) - \bar{g}(x)\big)\big(g(\omega, y) - \bar{g}(y)\big)\Big], \qquad x, y \in \bar{D}.$$

For a detailed exposition of log-normal random fields that satisfy Assumptions A1–A2, we refer the reader to [27, Section 2.3].

4.2 The State Equation

We start by recalling the weak formulation of (4.2), parametrized by $\omega \in \Omega$, which reads: for a given $u \in L^2(D)$, find $y_\omega \in H_0^1(D)$ such that

$$\int_D a_\omega \nabla y_\omega \cdot \nabla v \, dx = \int_D uv \, dx \quad \forall\, v \in H_0^1(D). \tag{4.3}$$

If such a function $y_\omega \in H_0^1(D)$ exists, it is called a weak solution to (4.2). Here and in what follows, we use the subscript ω to indicate the dependence on the random parameter ω, for instance, $y_\omega := y(\omega, \cdot)$ and $a_\omega := a(\omega, \cdot)$.

The next result, which is a special case of [71, Theorem 2.1], is about the existence, uniqueness and the regularity of the solution to (4.3).

Theorem 4.2.1 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, the boundary value problem (4.2) admits a unique weak solution $y_\omega \in H_0^1(D) \cap H^{1+s}(D)$ for every $u \in L^2(D)$. Moreover, it holds*

$$|y_\omega|_{H^1(D)} \lesssim \frac{\|u\|_{L^2(D)}}{a_{\min}(\omega)}, \tag{4.4}$$

and

$$\|y_\omega\|_{H^{1+s}(D)} \lesssim C_{4.2.1}(\omega)\|u\|_{L^2(D)}, \tag{4.5}$$

for all $0 < s < t$ except $s = \frac{1}{2}$, where

$$C_{4.2.1}(\omega) := \frac{a_{\max}(\omega)\|a_\omega\|^2_{C^t(\bar{D})}}{a_{\min}(\omega)^4}.$$

Furthermore, $y \in L^p(\Omega, H^{1+s}(D))$, for all $p \in [1,\infty)$. If $t = 1$, then $y \in L^p(\Omega, H^2(D))$ and the bound (4.5) holds with $s = 1$.

Proof: Let us define, for $\mathbb{P}$-a.s. $\omega \in \Omega$, the bilinear form $b_\omega : H_0^1(D) \times H_0^1(D) \to \mathbb{R}$ by

$$b_\omega(y, v) := \int_D a_\omega \nabla y \cdot \nabla v \, dx. \tag{4.6}$$

Then, from Assumptions A1–A2, we have

$$\begin{aligned} |b_\omega(y,v)| &\le a_{\max}(\omega)|y|_{H^1(D)}|v|_{H^1(D)} \quad \forall\, y, v \in H_0^1(D), \\ b_\omega(y,y) &\ge a_{\min}(\omega)|y|^2_{H^1(D)} \quad \forall\, y \in H_0^1(D). \end{aligned} \tag{4.7}$$

Hence, according to the Lax-Milgram lemma, for a given $u \in L^2(D)$ there exists a unique $y_\omega \in H_0^1(D)$ such that

$$b_\omega(y_\omega, v) = \int_D uv \, dx \quad \forall\, v \in H_0^1(D), \qquad \text{and} \qquad |y_\omega|_{H^1(D)} \lesssim \frac{\|u\|_{L^2(D)}}{a_{\min}(\omega)}.$$

The $H^{1+s}(D)$ regularity of the solution y_ω and the estimate (4.5) were shown in [71, Theorem 2.1]. From (4.5), Assumptions A1–A2 and the Hölder's inequality, it follows that $y \in L^p(\Omega, H^{1+s}(D))$, for all $p \in [1,\infty)$. ■

Thanks to Theorem 4.2.1, we may now introduce for $\mathbb{P}$-a.s. $\omega \in \Omega$ the mapping

$$S_\omega : L^2(D) \to H_0^1(D) \tag{4.8}$$

such that $y_\omega := S_\omega u$ is the weak solution of (4.2) for a given right hand side $u \in L^2(D)$ and a realization a_ω. The mapping S_ω is sometimes referred to as the control-to-state operator since it assigns to a given control u its associated state y_ω. Obviously, S_ω is bounded and linear.

4.3 The Optimal Control Problem

In this section we rewrite the problem (4.1) into its reduced form (P_ω) and discuss the existence of a unique global solution to (P_ω) for a given $\omega \in \Omega$. Then we construct a map that assigns to a given $\omega \in \Omega$ the solution to (P_ω). We show that this map is a $L^2(D)$-valued random variable and establish some properties of it.

By means of the control-to-state operator S_ω introduced in (4.8), the problem (4.1) is equivalent to, for $\mathbb{P}$-a.s. $\omega \in \Omega$,

$$(\mathrm{P}_\omega) \qquad \min_{u \in U_{ad}} J_\omega(u) := \frac{1}{2}\|S_\omega u - z\|^2_{L^2(D)} + \frac{\alpha}{2}\|u\|^2_{L^2(D)}$$

where

$$U_{ad} := \{u \in L^2(D) : u_a \le u(x) \le u_b \text{ for a.e. } x \in D\} \tag{4.9}$$

is the set of admissible controls, which is closed and convex.

Theorem 4.3.1 *Suppose that U_{ad} is non-empty. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there exists a unique global solution for the Problem (P_ω).*

Proof: For a fixed $\omega \in \Omega$, the Problem (P_ω) is a deterministic minimization problem. The existence and uniqueness of a global solution to (P_ω) follows by a classical argument, see for instance [45, Section 1.5] for a detailed proof. ■

Owing to Theorem 4.3.1, we can now introduce the map

$$u^* : \Omega \to L^2(D) \text{ such that } u^*(\omega) := \underset{u \in U_{ad}}{\arg\min}\, J_\omega(u) \quad \text{for } \mathbb{P}\text{-a.s. } \omega \in \Omega. \tag{4.10}$$

In other word, for a given $\omega \in \Omega$, $u^*(\omega)$ is the solution of (P_ω). In what follows, we show that u^* is indeed a $L^2(D)$-valued random variable, i.e., it is a measurable map between Ω and $L^2(D)$, and then we establish some properties of it.

Theorem 4.3.2 *The map $u^* : \Omega \to L^2(D)$ introduced in* (4.10) *is measurable.*

Proof: Throughout the proof, we write $J(\omega, u)$ instead of $J_\omega(u)$ for any $(\omega, u) \in \Omega \times L^2(D)$ for convenience.

Recall that for a fixed control $u \in L^2(D)$ the map $\Omega \ni \omega \mapsto S_\omega u \in L^2(D)$ is measurable. This implies, by [33, Proposition 1.2], that $\Omega \ni \omega \mapsto \|S_\omega u - z\|_{L^2(D)} \in \mathbb{R}$ is measurable as well. From this, we can easily see that the map $J : \Omega \times L^2(D) \to \mathbb{R}$ defined in Problem (P_ω) is Carathéodory, i.e., for every $u \in L^2(D)$, $J(\cdot, u)$ is measurable and for every $\omega \in \Omega$, $J(\omega, \cdot)$ is continuous. Since J is Carathéodory, we deduce from [4, Theorem 8.2.11] that the set valued map R defined by

$$\Omega \ni \omega \mapsto R(\omega) := \{u \in U_{ad} : J(\omega, u) = \min_{v \in U_{ad}} J(\omega, v)\},$$

is measurable. Consequently, according to [4, Theorem 8.1.3], there exists a measurable selection of R, that is to say, there exists a measurable map $f : \Omega \to L^2(D)$ satisfying

$$\forall\, \omega \in \Omega, \quad f(\omega) \in R(\omega).$$

However, since Theorem 4.3.1 guarantees that for every $\omega \in \Omega$, the set $R(\omega)$ has a unique element, which we denote by $u^*(\omega)$, we conclude that the map $\Omega \ni \omega \mapsto u^*(\omega) \in L^2(D)$ is the measurable selection of R. This is the desired conclusion. ■

Theorem 4.3.3 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$ and let u^* be the map defined in* (4.10)*. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there holds*

$$\|u^*(\omega)\|_{L^2(D)} \lesssim C_{4.3.3}(\omega) := \sqrt{\left(\frac{2}{\alpha a_{\min}(\omega)^2} + 1\right)\|u\|^2_{L^2(D)} + \frac{2}{\alpha}\|z\|^2_{L^2(D)}} \tag{4.11}$$

for any $u \in U_{ad}$. Moreover, $u^* \in L^p(\Omega, L^2(D))$ for all $p \in [1, \infty)$.

Proof: We begin by establishing the bound in (4.11). From the optimality of $u^*(\omega)$ together with the estimate (4.4) it follows that for any $u \in U_{ad}$ there

holds

$$\frac{\alpha}{2}\|u^*(\omega)\|^2_{L^2(D)} \leq J_\omega(u^*(\omega)) \leq J_\omega(u) = \frac{1}{2}\|S_\omega u - z\|^2_{L^2(D)} + \frac{\alpha}{2}\|u\|^2_{L^2(D)}$$
$$\lesssim \left(\frac{1}{a_{\min}(\omega)^2} + \frac{\alpha}{2}\right)\|u\|^2_{L^2(D)} + \|z\|^2_{L^2(D)},$$

from which we obtain the desired result.
From (4.11) and Assumption A2 it follows that $u^* \in L^p(\Omega, L^2(D))$ for all $p \in [1, \infty)$. This completes the proof. ■

Remark 13 Under the hypothesis of Theorem 4.3.3, if moreover we assume that U_{ad} is bounded or $0 \in U_{ad}$, then we have $u^* \in L^\infty(\Omega, L^2(D))$. We can see this from the following.
If U_{ad} is bounded, then u^* is uniformly bounded in ω since $u_a \leq u^*(\omega) \leq u_b$ and the bounds u_a, u_b are finite and independent of ω.
On the other hand, if $0 \in U_{ad}$, then choosing $u = 0$ in (4.11) gives

$$\|u^*(\omega)\|_{L^2(D)} \lesssim \sqrt{\frac{2}{\alpha}}\|z\|_{L^2(D)}$$

which is also a uniform bound with respect to ω.

Remark 14 The estimate in (4.11) holds for any $u \in U_{ad}$. In the proof of Theorem 5.3.4 we will see that it is desirable to choose $u \in U_{ad}$ such that the constant $C_{4.3.3}(\omega)$ in (4.11) is small. This can be achieved by using the projection of 0 onto U_{ad}, that is $u := \min\{\max\{0, u_a\}, u_b\}$ in (4.11).

Chapter 5

Variational Discretization

In this chapter we are interested in approximating the expectation of the solutions of the control problem (P_ω), $\omega \in \Omega$, using the Monte Carlo finite element method. In particular, we discretize the control problem using the variational discretization concept developed in [46] to obtain discrete optimal controls. Then, we take the sample average for these discrete solutions.

This chapter is organized as follows: Section 5.1 contains the finite element setting. In Section 5.2 we discretize the state equation using piecewise linear and continuous finite elements. In Section 5.3, we apply the variational discretization to the control problem (P_ω), $\omega \in \Omega$, and construct the map that assigns to each $\omega \in \Omega$ the solution of the corresponding discrete control problem. In Section 5.4, we study the Monte Carlo finite element method applied to the problem (P_ω), $\omega \in \Omega$. Particularly, we will investigate the error and the computational cost associated to the single-level and multilevel Monte Carlo finite element method. Finally, we verify our theoretical findings numerically in Section 5.5.

5.1 Finite Element Preliminaries

Let $\{\mathcal{T}_{h_l}\}_{l\geq 0}$ be a sequence of triangulations of the domain D such that $\mathcal{T}_{h_l}$ is obtained from an initial coarse triangulation $\mathcal{T}_{h_0}$ via l successive uniform refinements, that is to say, $h_l = \frac{1}{2}h_{l-1} = 2^{-l}h_0$ for $l = 1, 2, \ldots$, where $h_l := \max_{T\in\mathcal{T}_{h_l}} \operatorname{diam}(T)$ denotes the maximum mesh size of $\mathcal{T}_{h_l}$. Here, $\operatorname{diam}(T)$ stands for the diameter of the triangle T. In addition, for any triangulation $\mathcal{T}_{h_l}$ we assume that

$$\bar{D} = \bigcup_{T\in\mathcal{T}_{h_l}} \bar{T}.$$

We point out that a sequence of triangulations generated as above is qausi-uniform, see [13, Remark 4.4.17].

On each $\mathcal{T}_{h_l}$, for $l = 0, 1, \ldots$, we construct the space of linear finite elements X_{h_l} defined by

$$X_{h_l} := \{v \in C^0(\bar{D}) : v \text{ is a linear polynomial on each } T \in \mathcal{T}_{h_l} \text{ and } v_{|\partial D} = 0\}.$$

It is clear that for these spaces there holds

$$X_{h_0} \subset X_{h_1} \subset \cdots \subset X_{h_l} \subset \cdots .$$

Moreover, we have $\dim(X_{h_l}) = N_l$ where N_l is the number of the inner nodes in the triangulation $\mathcal{T}_{h_l}$ and $\dim(X_{h_l})$ is the dimension of the space X_{h_l}.

The above notation will be used in Section 5.4. On the other hand, the results that we will establish in Section 5.2–5.3 are still valid if the sequence of triangulations of D is assumed to be only regular, in other words, the sequence needs not to be generated via uniform refinements of $\mathcal{T}_{h_0}$. Therefore, in these two sections we drop the subscript $l = 1, 2, \ldots$, and we use the notation $\{\mathcal{T}_h\}_{0<h\leq h_0}$, h, X_h instead of $\{\mathcal{T}_{h_l}\}_{l\geq 0}$, h_l, X_{h_l}, respectively. Here, $\{\mathcal{T}_h\}_{0<h\leq h_0}$ is regular in the sense that there exists $\rho > 0$ such that for all $T \in \mathcal{T}_h$ and for all $h \in (0, h_0]$,

$$\operatorname{diam}(B_T) \geq \rho \operatorname{diam}(T),$$

where $\operatorname{diam}(B_T)$ is the diameter of the largest ball contained in T.

Finally, we state the next lemma which will be useful in our analysis.

Lemma 5.1.1 *Let $v \in H_0^1(D) \cap H^{1+s}(D)$ for some $0 < s \leq 1$. Then*

$$\inf_{v_h \in X_h} |v - v_h|_{H^1(D)} \lesssim \|v\|_{H^{1+s}(D)} h^s,$$

where the hidden constant is independent of v and h.

Proof: See [43, Lemma 8.4.9]. ■

5.2 The Discrete State Equation

In this section we approximate the solution of (4.3) by continuous piecewise linear finite elements and then estimate the resulting error.

For a given $\omega \in \Omega$, the finite element discretization of (4.3) reads: find $y_{\omega,h} \in X_h$ such that

$$\int_D a_\omega \nabla y_{\omega,h} \cdot \nabla v_h \, dx = \int_D u v_h \, dx \quad \forall\, v_h \in X_h. \tag{5.1}$$

Theorem 5.2.1 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there exists a unique solution $y_{\omega,h} \in X_h$ to (5.1). Moreover,*

$$|y_{\omega,h}|_{H^1(D)} \lesssim \frac{\|u\|_{L^2(D)}}{a_{\min}(\omega)}. \tag{5.2}$$

Proof: The result follows from applying the Lax-Milgram lemma as in the proof of Theorem 4.2.1. ■

In the light of Theorem 5.2.1, we introduce for $\mathbb{P}$-a.s. $\omega \in \Omega$ the mapping

$$S_{\omega,h} : L^2(D) \to X_h \tag{5.3}$$

such that $y_{\omega,h} := S_{\omega,h} u$ is the solution of (5.1) for a given $u \in L^2(D)$ and a given realization a_ω. Sometimes, $S_{\omega,h}$ is referred to as the discrete control-to-state operator. Notice that the operator $S_{\omega,h}$ is bounded and linear.

In the next two theorems, we estimate the error in approximating the solution of (4.3) by the one of (5.1).

Theorem 5.2.2 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$. For a given $u \in L^2(D)$, let $y_\omega := S_\omega u$ and let $y_{\omega,h} := S_{\omega,h} u$ where S_ω and $S_{\omega,h}$ are as defined in* (4.8) *and* (5.3)*, respectively. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there holds*

$$|y_\omega - y_{\omega,h}|_{H^1(D)} \lesssim C_{5.2.2}(\omega) \|u\|_{L^2(D)} h^s, \tag{5.4}$$

for all $0 < s < t$ except $s = \frac{1}{2}$, where

$$C_{5.2.2}(\omega) := \Big(\frac{a_{\max}^3(\omega)}{a_{\min}^9(\omega)} \Big)^{\frac{1}{2}} \|a_\omega\|^2_{C^t(\bar{D})}.$$

If $t = 1$, the above estimate holds with $s = 1$.

Proof: For a given $\omega \in \Omega$, let $b_\omega : H_0^1(D) \times H_0^1(D) \to \mathbb{R}$ be the bilinear form introduced in (4.6). From Assumptions A1–A2, we conclude that $b_\omega(\cdot, \cdot)$ defines an inner product over $H_0^1(D)$ and its induced norm $|v|_{b_\omega} := (b_\omega(v, v))^{\frac{1}{2}}$, for any $v \in H_0^1(D)$, satisfies

$$\sqrt{a_{\min}(\omega)} |v|_{H^1(D)} \lesssim |v|_{b_\omega} \lesssim \sqrt{a_{\max}(\omega)} |v|_{H^1(D)}. \tag{5.5}$$

Using the Galerkin orthogonality

$$b_\omega(y_\omega - y_{\omega,h}, v_h) = 0 \qquad \forall\, v_h \in X_h,$$

it is a classical task to show that

$$|y_\omega - y_{\omega,h}|_{b_\omega} = \min_{v_h \in X_h} |y_\omega - v_h|_{b_\omega}.$$

Thus, from the previous relation and (5.5) we get

$$|y_\omega - y_{\omega,h}|_{H^1(D)} \lesssim \sqrt{\frac{a_{\max}(\omega)}{a_{\min}(\omega)}} \inf_{v_h \in X_h} |y_\omega - v_h|_{H^1(D)}.$$

Recalling Theorem 4.2.1 and Lemma 5.1.1, we may continue

$$\begin{aligned} |y_\omega - y_{\omega,h}|_{H^1(D)} &\lesssim \sqrt{\frac{a_{\max}(\omega)}{a_{\min}(\omega)}} \|y_\omega\|_{H^{1+s}(D)} h^s \\ &\lesssim \sqrt{\frac{a_{\max}(\omega)}{a_{\min}(\omega)}} C_{4.2.1}(\omega) \|u\|_{L^2(D)} h^s, \end{aligned}$$

which is the desired result. ∎

Theorem 5.2.3 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$. For a given $u \in L^2(D)$, let $y_\omega := S_\omega u$ and let $y_{\omega,h} := S_{\omega,h} u$ where S_ω and $S_{\omega,h}$ are as defined in* (4.8) *and* (5.3)*, respectively. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there holds*

$$\|y_{\omega,h} - y_\omega\|_{L^2(D)} \lesssim C_{5.2.3}(\omega) \|u\|_{L^2(D)} h^{2s}, \tag{5.6}$$

for all $0 < s < t$ except $s = \frac{1}{2}$, where

$$C_{5.2.3}(\omega) := \Big(\frac{a_{\max}^7(\omega)}{a_{\min}^{17}(\omega)} \Big)^{\frac{1}{2}} \|a_\omega\|^4_{C^t(\bar{D})}.$$

Moreover,

$$\|y_h - y\|_{L^p(\Omega,L^2(D))} \leq ch^{2s}, \quad \text{for all } p \in [1,\infty), \tag{5.7}$$

with $c > 0$ independent of ω and h. If $t = 1$, the above two estimates hold with $s = 1$.

Proof: The key idea is a duality argument. For a given $\omega \in \Omega$, let $e_\omega := y_{\omega,h} - y_\omega$ and let $\tilde{y}_\omega \in H_0^1(D)$ be the solution of the problem

$$b_\omega(\tilde{y}_\omega, v) = (e_\omega, v)_{L^2(D)} := \int_D e_\omega v \, dx \quad \forall\, v \in H_0^1(D),$$

where b_ω is the bilinear form defined in (4.6). Recall that $\tilde{y}_\omega \in H_0^1(D) \cap H^{1+s}(D)$ according to Theorem 4.2.1 and that we have the Galerkin orthogonality

$$b_\omega(e_\omega, v_h) = 0 \quad \forall\, v_h \in X_h.$$

Consequently, we obtain

$$\begin{aligned}
\|e_\omega\|^2_{L^2(D)} &= (e_\omega, e_\omega)_{L^2(D)} = b_\omega(\tilde{y}_\omega, e_\omega) = b_\omega(\tilde{y}_\omega - v_h, e_\omega) \\
&\leq a_{\max}(\omega)|\tilde{y}_\omega - v_h|_{H^1(D)}|e_\omega|_{H^1(D)} && \text{(using (4.7))} \\
&\leq a_{\max}(\omega)\|\tilde{y}_\omega\|_{H^{1+s}(D)}|e_\omega|_{H^1(D)}h^s && \text{(using Lemma 5.1.1)} \\
&\lesssim a_{\max}(\omega)C_{4.2.1}(\omega)\|e_\omega\|_{L^2(D)}|e_\omega|_{H^1(D)}h^s && \text{(using (4.5))} \\
&\lesssim a_{\max}(\omega)C_{4.2.1}(\omega)\|e_\omega\|_{L^2(D)}C_{5.2.2}(\omega)\|u\|_{L^2(D)}h^{2s}. && \text{(using (5.4))}
\end{aligned}$$

Dividing both sides of the previous inequality by $\|e_\omega\|_{L^2(D)}$ gives the estimate (5.6) from which we get (5.7) after applying the Hölder's inequality together with the Assumptions A1–A2. This completes the proof. ■

We conclude this section by the next remark.

Remark 15 The order of convergence $O(h^{2s})$ in the estimate (5.6) is obtained while assuming that the integrals in (5.1) are computed exactly. In general, those integrals can't be computed exactly, instead, they are approximated by quadrature which introduces another sort of error that one should consider. However, it is still possible to achieve the order $O(h^{2s})$ in (5.6) even with quadrature provided that the function $a(\omega,\cdot)$ belongs to at least $C^{2s}(\bar{D})$ as it was explained in [27, Section 3.3]. It is important to mentioned this at this stage because all the upcoming error estimates related to the optimal control problem are heavily depending on (5.6).

5.3 The Discrete Optimal Control Problem

In this section we discretize the problem (P_ω) using the variational discretization approach developed in [46]. Then, we construct the discrete counter part of the map u^* introduced in (4.10) and carry out the associated error analysis.

Using the discrete control-to-state operator $S_{\omega,h}$ introduced in (5.3), the variational discretization of (P_ω), for $\mathbb{P}$-a.s. $\omega \in \Omega$, reads

$$(\mathrm{P}_{\omega,h}) \qquad \min_{u\in U_{ad}} J_{\omega,h}(u) := \frac{1}{2}\|S_{\omega,h}u - z\|^2_{L^2(D)} + \frac{\alpha}{2}\|u\|^2_{L^2(D)}.$$

Notice that problem $(\mathrm{P}_{\omega,h})$ is again an optimization problem in infinite dimensions as the controls are still sought in U_{ad}. Thus all techniques we used previously to study problem (P_ω) can also be used for $(\mathrm{P}_{\omega,h})$.

Theorem 5.3.1 *Suppose that U_{ad} is non-empty. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there exists a unique global solution for the Problem $(\mathrm{P}_{\omega,h})$.*

Proof: The result can be established analogously to Theorem 4.3.1. ■

We may introduce in the light of the previous theorem, for a given mesh size h, the map

$$u_h^* : \Omega \to L^2(D) \text{ such that } u_h^*(\omega) := \underset{u \in U_{ad}}{\arg\min}\, J_{\omega,h}(u) \quad \text{for } \mathbb{P}\text{-a.s. } \omega \in \Omega. \tag{5.8}$$

That is to say, for a given $\omega \in \Omega$, $u_h^*(\omega)$ is the solution of $(\mathrm{P}_{\omega,h})$. The next result tells us that u_h^* is a $L^2(D)$-valued random variable.

Theorem 5.3.2 *The map $u_h^* : \Omega \to L^2(D)$ introduced in (5.8) is measurable.*

Proof: The proof is analogous to that of Theorem 4.3.2. ■

Theorem 5.3.3 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$ and let u_h^* be the map defined in (5.8). Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there holds*

$$\|u_h^*(\omega)\|_{L^2(D)} \lesssim C_{5.3.3}(\omega) := \sqrt{\left(\frac{2}{\alpha a_{\min}(\omega)^2} + 1\right)\|u\|_{L^2(D)}^2 + \frac{2}{\alpha}\|z\|_{L^2(D)}^2} \tag{5.9}$$

for any $u \in U_{ad}$. Moreover, $u_h^* \in L^p(\Omega, L^2(D))$ for all $p \in [1, \infty)$.

Proof: Let us firstly verify the bound in (5.9). From the optimality of $u_h^*(\omega)$ and the estimate (5.2) we obtain for any $u \in U_{ad}$

$$\begin{aligned}\frac{\alpha}{2}\|u_h^*(\omega)\|_{L^2(D)}^2 \leq J_{\omega,h}(u_h^*(\omega)) \leq J_{\omega,h}(u) &= \frac{1}{2}\|S_{\omega,h}u - z\|_{L^2(D)}^2 + \frac{\alpha}{2}\|u\|_{L^2(D)}^2 \\ &\lesssim \left(\frac{1}{a_{\min}(\omega)^2} + \frac{\alpha}{2}\right)\|u\|_{L^2(D)}^2 + \|z\|_{L^2(D)}^2,\end{aligned}$$

from which we get the desired bound.
From (5.9) and Assumption A2 it follows that $u_h^* \in L^p(\Omega, L^2(D))$ for all $p \in [1, \infty)$. This completes the proof. ■

Remark 16 If U_{ad} is bounded or if $0 \in U_{ad}$, then $u_h^* \in L^\infty(\Omega, L^2(D))$ can be shown analogously to Remark 13.

We now show that the map u_h^* converges to u^* in $L^p(\Omega, L^2(D))$ as the discretization parameter h tends to zero and we derive the corresponding error estimate.

Theorem 5.3.4 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$ and let u^* and u_h^* be the maps defined in (4.10) and (5.8), respectively. Then, for $\mathbb{P}$-a.s. $\omega \in \Omega$, there holds*

$$\|u^*(\omega) - u_h^*(\omega)\|_{L^2(D)} \lesssim C_{5.3.4}(\omega) h^{2s}, \tag{5.10}$$

for all $0 < s < t$ *except* $s = \frac{1}{2}$*, where*

$$C_{5.3.4}(\omega) := \frac{1}{\alpha}\sqrt{\alpha\|u^*(\omega)\|^2_{L^2(D)} + \max(1, a_{\min}^{-1}(\omega))^2\big(\|u^*(\omega)\|_{L^2(D)} + \|z\|_{L^2(D)}\big)^2}\times C_{5.2.3}(\omega).$$

Moreover,

$$\|u^* - u_h^*\|_{L^p(\Omega,L^2(D))} \leq C(\alpha, z, U_{ad})h^{2s}, \quad \text{for all } p \in [1,\infty), \tag{5.11}$$

with $C(\alpha, z, U_{ad}) > 0$ *independent of* ω *and* h *and depending only on the deterministic data* α, z *and* U_{ad}*. If* $t = 1$*, the above two estimates hold with* $s = 1$*.*

Proof: For a given $\omega \in \Omega$, the realizations $u^*(\omega)$ and $u_h^*(\omega)$ are the solutions of (P_ω) and $(\mathrm{P}_{\omega,h})$, respectively. Hence, according to [45, Theorem 3.4] we have

$$\begin{aligned}\alpha\|u^*(\omega) - u_h^*(\omega)\|^2_{L^2(D)} &\leq \|(S_\omega - S_{\omega,h})u_\omega\|^2_{L^2(D)} \\ &\quad + \frac{1}{\alpha}\|(S_\omega - S_{\omega,h})(y_\omega - z)\|^2_{L^2(D)},\end{aligned} \tag{5.12}$$

where we define here and subsequently $u_\omega := u^*(\omega)$ and $y_\omega := S_\omega u_\omega$.
The first term in (5.12) can be estimated with the help of Theorem 5.2.3, to obtain

$$\|(S_\omega - S_{\omega,h})u_\omega\|_{L^2(D)} \lesssim C_{5.2.3}(\omega)\|u_\omega\|_{L^2(D)}h^{2s}.$$

The second term can be bounded using again Theorem 5.2.3 and (4.4) as follows:

$$\begin{aligned}\|(S_\omega - S_{\omega,h})(y_\omega - z)\|_{L^2(D)} &\lesssim C_{5.2.3}(\omega)\|y_\omega - z\|_{L^2(D)}h^{2s} \\ &\lesssim C_{5.2.3}(\omega)\big(\|y_\omega\|_{L^2(D)} + \|z\|_{L^2(D)}\big)h^{2s} \\ &\lesssim C_{5.2.3}(\omega)\left(\frac{\|u_\omega\|_{L^2(D)}}{a_{\min}(\omega)} + \|z\|_{L^2(D)}\right)h^{2s} \\ &\lesssim C_{5.2.3}(\omega)\max(1, a_{\min}^{-1}(\omega))\big(\|u_\omega\|_{L^2(D)} + \|z\|_{L^2(D)}\big)h^{2s}.\end{aligned}$$

Inserting the above estimates into (5.12) gives the estimate (5.10) from which one obtains (5.11) after using the Hölder's inequality together with Assumptions A1–A2 as well as recalling (4.11). This completes the proof. ■

5.4 Monte Carlo FE Methods

In this section we study the approximation of the expectation of the map u^* defined in (4.10) by Monte Carlo finite element methods. In particular, we will carry out the error and the cost analysis for the single-level MC FE and the multilevel MC FE methods. To begin, we first make a quick review for Monte Carlo methods.

For a given $u \in L^2(\Omega, L^2(D))$, the expectation or the expected value of u is

$$\mathbb{E}[u] := \int_\Omega u(\omega)\, d\mathbb{P}(\omega).$$

The Monte Carlo (MC) estimator for $\mathbb{E}[u]$ is the sample average

$$E_M[u] := \frac{1}{M}\sum_{i=1}^{M} u(\omega_i),$$

where $u(\omega_i)$, $i = 1, \dots, M$, are M independent identically distributed samples of u. Notice that, for a fixed M, the estimator $E_M[u]$ can be interpreted as a $L^2(D)$-valued random variable. The next result gives the *statistical error* associated with this estimator.

Theorem 5.4.1 *Let $u \in L^2(\Omega, L^2(D))$. Then, for any $M \in \mathbb{N}$, we have*

$$\|\mathbb{E}[u] - E_M[u]\|_{L^2(\Omega, L^2(D))} \leq M^{-\frac{1}{2}} \|u\|_{L^2(\Omega, L^2(D))}.$$

Proof: Using the fact that $u(\omega_i)$, $i = 1, \dots, M$, are independent, identically distributed random samples, we obtain

$$\begin{aligned}
\mathbb{E}\big[\|\mathbb{E}[u] - E_M[u]\|^2_{L^2(D)}\big] &= \mathbb{E}\Big[\Big\|\mathbb{E}[u] - \frac{1}{M}\sum_{i=1}^{M} u(\omega_i)\Big\|^2_{L^2(D)}\Big] \\
&= \frac{1}{M^2}\mathbb{E}\Big[\Big\|\sum_{i=1}^{M}\big(\mathbb{E}[u] - u(\omega_i)\big)\Big\|^2_{L^2(D)}\Big] \\
&= \frac{1}{M^2}\sum_{i=1}^{M}\mathbb{E}\Big[\Big\|\mathbb{E}[u] - u(\omega_i)\Big\|^2_{L^2(D)}\Big] \\
&= \frac{1}{M}\mathbb{E}\big[\|\mathbb{E}[u] - u\|^2_{L^2(D)}\big] \\
&= \frac{1}{M}\Big(\mathbb{E}\big[\|u\|^2_{L^2(D)}\big] - \|\mathbb{E}[u]\|^2_{L^2(D)}\Big) \\
&\leq \frac{1}{M}\mathbb{E}\big[\|u\|^2_{L^2(D)}\big].
\end{aligned}$$

Taking the square root of both sides of the previous inequality gives the desired result. ∎

Lemma 5.4.2 *Let u^* be the map introduced in* (4.10)*. Then, for any $M \in \mathbb{N}$, there holds*

$$\|\mathbb{E}[u^*] - E_M[u^*]\|_{L^2(\Omega, L^2(D))} \leq M^{-\frac{1}{2}} \|u^*\|_{L^2(\Omega, L^2(D))}.$$

Proof: The result is a direct consequence of Theorem 5.4.1 as $u^* \in L^2(\Omega, L^2(D))$ according to Theorem 4.3.3. ∎

The previous lemma shows that $\mathbb{E}[u^*]$ can by approximated by $E_M[u^*]$ which is the sample average of M independent realizations $u^*(\omega_i)$, $i = 1, \dots, M$. However, evaluating u^* at a given $\omega \in \Omega$ can be a difficult task in practice since u^* is usually not known explicitly. To overcome this difficulty, we consider sampling from $u^*_{h_l}$ at a given refinement level $l \in \mathbb{N}$ instead of sampling from u^* directly. Recall that $u^*_{h_l}$ is the finite element approximation of u^* introduced in (5.8).

In the rest of this section we study two approaches to approximate $\mathbb{E}[u^*]$ while using $u^*_{h_l}$, namely, we will study the single-level and the multilevel Monte Carlo finite element methods. Before we start, we make the following assumption on the computational cost of computing $u^*_{h_l}$.

A3. For a given $\omega \in \Omega$, the computational cost $\mathcal{C}_l$ of computing $u^*_{h_l}(\omega)$ such that (5.10) holds is asymptotically, as $l \to \infty$, bounded by

$$\mathcal{C}_l \lesssim h_l^{-\gamma} \approx N_l^{\frac{\gamma}{d}}$$

for some real number $\gamma > 0$, where $N_l = \dim(X_{h_l})$ and d is the dimension of the computational domain D.

In the previous assumption we adopt the convention that $h_l \approx N_l^{-\frac{1}{d}}$. It is worth to mention that the ideal value of γ in the above assumption would be $\gamma = d$, in this case for instance, doubling the number of unknowns N_l should result in doubling the computational cost $\mathcal{C}_l$.

5.4.1 Single-Level Monte Carlo FE Method

For a given refinement level $l \in \mathbb{N}$, the single-level MC FE estimator for $\mathbb{E}[u^*]$ is

$$E_M[u^*_{h_l}] := \frac{1}{M}\sum_{i=1}^{M} u^*_{h_l}(\omega_i), \tag{5.13}$$

where $u^*_{h_l}(\omega_i)$, $i = 1, \ldots, M$ are M independent samples of $u^*_{h_l}$. The error bound associated with this estimator is stated in the next theorem.

Theorem 5.4.3 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$. Then*

$$\|\mathbb{E}[u^*] - E_M[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))} \leq C(\alpha, z, U_{ad})(M^{-\frac{1}{2}} + h_l^{2s})$$

for all $0 < s < t$ except $s = \frac{1}{2}$ where $C(\alpha, z, U_{ad}) > 0$ is a constant independent of h_l and depending only on the data α, z and on U_{ad}. If $t = 1$, the above estimate holds with $s = 1$.

Proof: We start the proof by using the triangle inequality to obtain

$$\|\mathbb{E}[u^*]-E_M[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))} \leq \|\mathbb{E}[u^*]-\mathbb{E}[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))}+\|\mathbb{E}[u^*_{h_l}]-E_M[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))}.$$

The task is now to estimate the two terms on the right hand side of the previous inequality. The estimate for the first term follows from Theorem 5.3.4 after utilizing the Cauchy-Schwarz inequality. In fact, we have

$$\begin{aligned}\|\mathbb{E}[u^*] - \mathbb{E}[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))} &= \|\mathbb{E}[u^* - u^*_{h_l}]\|_{L^2(\Omega,L^2(D))} = \|\mathbb{E}[u^* - u^*_{h_l}]\|_{L^2(D)} \\ &\leq \mathbb{E}[\|u^* - u^*_{h_l}\|_{L^2(D)}] \leq \|u^* - u^*_{h_l}\|_{L^2(\Omega,L^2(D))} \\ &\leq C(\alpha, z, U_{ad})h_l^{2s}.\end{aligned}$$

For the second term, we use Lemma 5.4.2 to obtain

$$\begin{aligned}\|\mathbb{E}[u^*_{h_l}] - E_M[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))} &\leq M^{-\frac{1}{2}}\|u^*_{h_l}\|_{L^2(\Omega,L^2(D))} \\ &\leq C(\alpha, z, U_{ad})M^{-\frac{1}{2}},\end{aligned}$$

where we used the fact that $\|u^*_{h_l}\|_{L^2(\Omega,L^2(D))}$ is uniformly bounded with respect to h_l according to (5.9). Combining the estimates of the two terms gives the desired result. ■

We can see from Theorem 5.4.3 that the error that comes from using (5.13) as an approximation to $\mathbb{E}[u^*]$ can be split into two parts; a statistical part which is of order $M^{-1/2}$ and a discretization part of order h_l^{2s}. This suggests that there should be some coupling between the number of samples M and the mesh size h_l in order to achieve a certain overall error. We provide such coupling in the next theorem and establish the corresponding error and computational cost.

Theorem 5.4.4 *Let Assumptions A1–A3 hold for some $0 < t \leq 1$. Then, the MC estimator* (5.13) *with the following choice of number of samples*

$$M = O(h_l^{-4s}),$$

yields the error bound

$$\|\mathbb{E}[u^*] - E_M[u^*_{h_l}]\|_{L^2(\Omega,L^2(D))} \leq C(\alpha, z, U_{ad})h_l^{2s} \tag{5.14}$$

for all $0 < s < t$ except $s = \frac{1}{2}$, with a total computational cost $\mathcal{C}_l$ which is asymptotically, as $l \to \infty$, bounded by

$$\mathcal{C}_l \lesssim h_l^{-\gamma-4s}, \tag{5.15}$$

for some $C(\alpha, z, U_{ad}) > 0$ depending on the data α, z and on U_{ad}. If $t = 1$, the above estimates (5.14) *and* (5.15) *hold with $s = 1$.*

Proof: The estimate (5.14) follows from Theorem 5.4.3 after choosing $M \approx h_l^{-4s}$. To obtain the bound (5.15), it is enough to multiply the computational cost of one sample, that is $h_l^{-\gamma}$, from Assumption A3 by the total number of samples $M \approx h_l^{-4s}$. ■

Remark 17 It is important to point out that the previous theorem is valid provided that $h_l \leq \tilde{h}$, where $\tilde{h}$ is a mesh size such that (5.10) is satisfied for all $h \leq \tilde{h}$. Since, by assumption, we have $h_l = 2^{-l}h_0$, $l = 0, 1, \ldots$, we can either choose $h_0 = \tilde{h}$ or l large enough to satisfy $h_l \leq \tilde{h}$.

5.4.2 Multilevel Monte Carlo FE Method

We start by observing that at a given refinement level $L \in \mathbb{N}$ the random variable $u^*_{h_L}$ can be written as

$$u^*_{h_L} = \sum_{l=0}^{L}(u^*_{h_l} - u^*_{h_{l-1}}),$$

where $u^*_{h_{-1}} := 0$. The linearity of the expectation operator $\mathbb{E}[\cdot]$ implies

$$\mathbb{E}[u^*_{h_L}] = \sum_{l=0}^{L}\mathbb{E}[u^*_{h_l} - u^*_{h_{l-1}}]. \tag{5.16}$$

If we approximate $\mathbb{E}[u^*_{h_l} - u^*_{h_{l-1}}]$ in (5.16) by the single-level Monte Carlo estimator (5.13) with a number of samples M_l that depends on the refinement level l, we obtain the so called multilevel MC FE estimator for $\mathbb{E}[u^*]$, that is,

$$E^L[u^*_{h_L}] := \sum_{l=0}^{L} E_{M_l}[u^*_{h_l} - u^*_{h_{l-1}}], \tag{5.17}$$

where the samples over all levels, $l = 0, \ldots, L$, are independent of each other. The next theorem gives the error bound associated with the estimator (5.17).

Theorem 5.4.5 *Let Assumptions A1–A2 hold for some $0 < t \leq 1$. Then*

$$\|\mathbb{E}[u^*] - E^L[u^*_{h_L}]\|_{L^2(\Omega,L^2(D))} \leq C(\alpha, z, U_{ad})\Big(h_L^{2s} + \sum_{l=0}^{L} M_l^{-\frac{1}{2}} h_l^{2s}\Big), \qquad (5.18)$$

for all $0 < s < t$ except $s = \frac{1}{2}$, where $C(\alpha, z, U_{ad}) > 0$ depends on the data α, z and on U_{ad}. If $t = 1$, the above estimate holds with $s = 1$.

Proof: Throughout the proof, we use the notation $\|\cdot\|_V := \|\cdot\|_{L^2(\Omega,L^2(D))}$. We begin by using the triangle inequality and recalling (5.16), (5.17) to obtain

$$\begin{aligned}\|\mathbb{E}[u^*] - E^L[u^*_{h_L}]\|_V &\leq \|\mathbb{E}[u^*] - \mathbb{E}[u^*_{h_L}]\|_V + \|\mathbb{E}[u^*_{h_L}] - E^L[u^*_{h_L}]\|_V \\ &\leq I + II, \end{aligned} \qquad (5.19)$$

where we define

$$I := \|\mathbb{E}[u^*] - \mathbb{E}[u^*_{h_L}]\|_V \quad \text{and} \quad II := \sum_{l=0}^{L} \|\mathbb{E}[u^*_{h_l} - u^*_{h_{l-1}}] - E_{M_l}[u^*_{h_l} - u^*_{h_{l-1}}]\|_V.$$

To estimate the term I, it is enough to argue like in the proof of Theorem 5.4.3 to obtain

$$\|\mathbb{E}[u^*] - \mathbb{E}[u^*_{h_L}]\|_V \leq C(\alpha, z, U_{ad}) h_L^{2s}.$$

On the other hand, the term II can be bounded by utilizing Theorem 5.4.1, the triangle inequality, Theorem 5.3.4 and the fact that $h_{l-1} = 2h_l$ to get

$$\begin{aligned}\sum_{l=0}^{L} \|\mathbb{E}[u^*_{h_l} - u^*_{h_{l-1}}] - E_{M_l}[u^*_{h_l} - u^*_{h_{l-1}}]\|_V &\leq \sum_{l=0}^{L} M_l^{-\frac{1}{2}} \|u^*_{h_l} - u^*_{h_{l-1}}\|_V \\ &\leq \sum_{l=0}^{L} M_l^{-\frac{1}{2}} \Big(\|u^*_{h_l} - u^*\|_V + \|u^* - u^*_{h_{l-1}}\|_V\Big) \\ &\leq C(\alpha, z, U_{ad}) \sum_{l=0}^{L} M_l^{-\frac{1}{2}} (h_l^{2s} + h_{l-1}^{2s}) \\ &= C(\alpha, z, U_{ad}) \sum_{l=0}^{L} M_l^{-\frac{1}{2}} (1 + 2^{2s}) h_l^{2s}. \end{aligned}$$

Inserting the bounds of the terms I, II in (5.19) gives

$$\|\mathbb{E}[u^*] - E^L[u^*_{h_L}]\|_V \leq C(\alpha, z, U_{ad})\Big(h_L^{2s} + \sum_{l=0}^{L} M_l^{-\frac{1}{2}} h_l^{2s}\Big),$$

which is the desired result and the proof is complete. ■

The previous theorem holds for any choice of $\{M_l\}_{l=0}^L$ in (5.17), where M_l is the number of samples over the refinement level l. However, it is desirable that $\{M_l\}_{l=0}^L$ is chosen in such a way that the statistical error and the discretization error in (5.18) are balanced. The next theorem suggests a choice for $\{M_l\}_{l=0}^L$ such that the overall error in (5.18) is of order h_L^{2s} and it gives the associated computational cost.

Theorem 5.4.6 *Let Assumptions A1–A3 hold for some $0 < t \leq 1$. Then, the MLMC estimator (5.17) with the following choice of $\{M_l\}_{l=0}^L$ where*

$$M_l = \begin{cases} O(h_L^{-4s} h_l^{\frac{\gamma+4s}{2}}), & 4s > \gamma \\ O\big((L+1)^2 h_L^{-4s} h_l^{\frac{\gamma+4s}{2}}\big), & 4s = \gamma \\ O(h_L^{-\frac{\gamma+4s}{2}} h_l^{\frac{\gamma+4s}{2}}), & 4s < \gamma \end{cases}$$

yields the error bound

$$\|\mathbb{E}[u^*] - E^L[u^*_{h_L}]\|_{L^2(\Omega,L^2(D))} \leq C(\alpha, z, U_{ad}) h_L^{2s} \tag{5.20}$$

for all $0 < s < t$ except $s = \frac{1}{2}$, with a total computational cost $\mathcal{C}_L$ which is asymptotically, as $L \to \infty$, bounded by

$$\mathcal{C}_L \lesssim \begin{cases} h_L^{-4s}, & 4s > \gamma \\ (L+1)^3 h_L^{-4s}, & 4s = \gamma \\ h_L^{-\gamma}, & 4s < \gamma. \end{cases} \tag{5.21}$$

Here, $C(\alpha, z, U_{ad}) > 0$ depends on the data α, z and on U_{ad}. If $t = 1$, the above estimates (5.20) and (5.21) hold with $s = 1$.

Proof: We give the proof only for the case $4s > \gamma$; the other two cases $4s = \gamma$ and $4s < \gamma$ can be treated analogously. To verify the estimate (5.20) it is enough to utilize Theorem 5.4.5 together with the choice

$$M_l \approx h_L^{-4s} h_l^{\frac{\gamma+4s}{2}}, \quad l = 0, \ldots, L, \tag{5.22}$$

as well as the approximation $h_l \approx 2^{-l}$ to obtain

$$\begin{aligned} \|\mathbb{E}[u^*] - E^L[u^*_{h_L}]\|_{L^2(\Omega,L^2(D))} &\leq C(\alpha, z, U_{ad}) \Big(h_L^{2s} + \sum_{l=0}^{L} M_l^{-\frac{1}{2}} h_l^{2s} \Big) \\ &= C(\alpha, z, U_{ad}) \Big(h_L^{2s} + h_L^{2s} \sum_{l=0}^{L} h_l^{\frac{4s-\gamma}{4}} \Big) \\ &\approx C(\alpha, z, U_{ad}) \Big(h_L^{2s} + h_L^{2s} \sum_{l=0}^{L} 2^{-(\frac{4s-\gamma}{4})l} \Big) \\ &= C(\alpha, z, U_{ad}) h_L^{2s} \left(1 + \frac{2^{-(\frac{4s-\gamma}{4})(L+1)} - 1}{2^{-\frac{4s-\gamma}{4}} - 1} \right) \\ &\lesssim C(\alpha, z, U_{ad}) h_L^{2s}, \qquad \text{as } L \to \infty. \end{aligned}$$

It remains to verify the asymptotic upper bound for the total computational cost (5.21). To achieve this, we see that from Assumption A3 together with the choice (5.22) and $h_l \approx 2^{-l}$, we have

$$\begin{aligned} \mathcal{C}_L = \sum_{l=0}^{L} M_l \mathcal{C}_l \lesssim \sum_{l=0}^{L} h_L^{-4s} h_l^{\frac{\gamma+4s}{2}} h_l^{-\gamma} = h_L^{-4s} \sum_{l=0}^{L} h_l^{\frac{4s-\gamma}{2}} &\approx h_L^{-4s} \sum_{l=0}^{L} 2^{-(\frac{4s-\gamma}{2})l} \\ \approx h_L^{-4s} \frac{2^{-(\frac{4s-\gamma}{2})(L+1)} - 1}{2^{-(\frac{4s-\gamma}{2})} - 1} &\lesssim h_L^{-4s}, \qquad \text{as } L \to \infty, \end{aligned}$$

which is the desired result. ∎

Remark 18 It should be emphasized that in Theorem 5.4.5 and in Theorem 5.4.6, the mesh size h_0 of the initial triangulation $\mathcal{T}_{h_0}$ should be chosen in such a way that (5.10) is satisfied for all $h \leq h_0$.

By comparing the total cost of MC in (5.15) and MLMC in (5.21) we see that the multilevel estimator achieves the same accuracy as classical Monte Carlo at a fraction of the cost. For example, to achieve the error bound h_L^{2s} using the MC estimator we need computations with the cost bound $h_L^{-\gamma-4s}$. On the other hand, we can get the same error bound with computations whose cost is bounded by $h_L^{-\gamma}$ if we use the MLMC estimator provided that $4s < \gamma$. Note that the bound $h_L^{-\gamma}$ is the largest possible cost bound for the MLMC estimator. Nevertheless, it is still smaller than the cost bound of the MC estimator by a factor of h_L^{-4s}.

We remark that the hidden constant in $O(\cdot)$ in the sequence $\{M_l\}_{l=0}^L$ from Theorem 5.4.6 plays a crucial rule in determining the size of the statistical error. This can be seen in (5.18) where it is clear that the larger the value of the constant in $O(\cdot)$, the smaller the statistical error. In order to obtain a minimal choice of $\{M_l\}_{l=0}^L$ we adapt the strategy presented in [52, Remark 4.11], that is, $\{M_l\}_{l=0}^L$ is chosen to be the solution of the following minimization problem

$$\text{(PN)} \qquad \min_{x \in \mathcal{M}_{ad}} \mathcal{J}(x) := \sum_{l=0}^L x_l \mathcal{C}_l \tag{5.23}$$

where

$$\mathcal{M}_{ad} := \Big\{(x_0, \dots, x_L) \in \mathbb{N}^{L+1} : x_l \geq 1 \text{ for } l = 0, \dots, L \text{ and } \sum_{l=0}^L x_l^{-\frac{1}{2}} h_l^{2s} \leq c_0 h_L^{2s}, \text{ for a fixed } c_0 > 0\Big\}.$$

Here, $\mathcal{C}_l$ is the computational cost of one sample at level l. The problem (PN) is a convex minimization problem. Moreover, for a fixed $c_0 > 0$, the set $\mathcal{M}_{ad}$ is non-empty because it contains $\{M_l\}_{l=0}^L$ from Theorem 5.4.6 provided that the hidden constant in $O(\cdot)$ is large enough. Consequently, if x^* is the solution of (PN), then we have

$$\mathcal{J}(x^*) \leq \mathcal{J}(\{M_l\}_{l=0}^L) = \sum_{l=0}^L M_l \mathcal{C}_l =: \mathcal{C}_L.$$

In other words, the solution of (PN) satisfies (5.21).

Remark 19 Observe that the admissible set of controls U_{ad} is a *convex* set. However, it is clear that the MLMC estimate for $\mathbb{E}[u^*]$ in (5.17) is in general not admissible since the corrections in (5.17) are computed using different realizations of the random coefficient. In contrast, the classical MC estimate in (5.13) is always admissible since it is a convex combination of admissible controls. This has already been observed in the context of random obstacle problems [8, 52]. Nevertheless, one may obtain an admissible approximation to $\mathbb{E}[u^*]$ using the MLMC estimator (5.17) by considering $P_{U_{ad}}(E^L[u^*_{h_L}])$ where $P_{U_{ad}} : L^2(D) \to U_{ad}$ is the projection into U_{ad}. Recall that $P_{U_{ad}}$ is nonexpansive and since $\mathbb{E}[u^*] \in U_{ad}$ it follows that

$$\|\mathbb{E}[u^*] - P_{U_{ad}}(E^L[u^*_{h_L}])\|_{L^2(D)} \leq \|\mathbb{E}[u^*] - E^L[u^*_{h_L}]\|_{L^2(D)}.$$

Hence, $P_{U_{ad}}(E^L[u^*_{h_L}])$ has at least the same rate of convergence as $E^L[u^*_{h_L}]$.

5.4.3 Variational Discrete Controls

We discuss now some general properties of variational discrete controls that will be useful to know if we aim to compute their expected value.

To begin, we consider again problem ($\mathrm{P}_{\omega,h}$) from Section 5.3 and let us denote by $\mathcal{N}_h$ the set of the nodes in the considered triangulation $\mathcal{T}_h$ of the domain D, that is,

$$\mathcal{N}_h := \{x_1, \ldots, x_n\},$$

where $x_i \in D$ for $i = 1, \ldots, n$ are the vertices of the triangles in $\mathcal{T}_h$.

It is well known that, for a given $\omega \in \Omega$, the solution of ($\mathrm{P}_{\omega,h}$), which we shall denote by $u_h(\omega)$, satisfies the following optimality condition

$$u_h(\omega) = P_{U_{ad}}\big(-\frac{1}{\alpha}p_{\omega,h}\big) \tag{5.24}$$

where $P_{U_{ad}} : L^2(D) \to U_{ad}$ is the projection into U_{ad} and $p_{\omega,h} := S_{\omega,h}(y_{\omega,h} - z)$ is the adjoint state while $y_{\omega,h} := S_{\omega,h}u_h(\omega)$ denotes the state associated with the solution $u_h(\omega)$. Note that in our case the operator $S_{\omega,h}$ is self-adjoint.

It is clear from (5.24) that $u_h(\omega) \notin X_h$ in general and hence the function $u_h(\omega) : D \to \mathbb{R}$ can't be simply characterized by its values at the nodes in $\mathcal{N}_h$. For this reason, we should also determine the active set of $u_h(\omega)$, or equivalently, the set of points

$$\mathcal{N}_{vd}(\omega) := \{\tilde{x}_1, \ldots, \tilde{x}_m\}$$

where $\tilde{x}_i \in D$ for $i = 1, \ldots, m$ are the intersection points of the active set boundary with the edges of the triangles in $\mathcal{T}_h$. Note that the boundary of the active set is polygon since $p_{\omega,h}$ is continuous and piecewise linear over each triangle in $\mathcal{T}_h$, see Figure 5.1 for an illustration of such an active set. Moreover, the points in $\mathcal{N}_{vd}(\omega)$ need not belong to $\mathcal{N}_h$ in general and that they depend on the realization $\omega \in \Omega$.

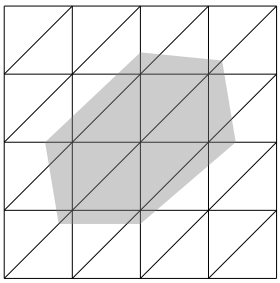


Figure 5.1 The active set of a variational discrete control (highlighted in grey). The boundary of the active set is polygon whose line segments need not coincide with the edges of the triangles in the mesh.

In conclusion, the function $u_h(\omega) : D \to \mathbb{R}$ is fully determined if we know its values on the set of points

$$\mathcal{N}_h \cup \mathcal{N}_{vd}(\omega).$$

In particular, if we would like to perform the sum

$$\frac{1}{M}\sum_{i=1}^{M} u_h(\omega_i) \tag{5.25}$$

for a given number of samples M, then we need to know the set of points

$$\mathcal{N}_h \cup \bigcup_{i\in\{1,\dots,M\}} \mathcal{N}_{vd}(\omega_i). \tag{5.26}$$

We observe that, for a fixed mesh size h, the number of points in the set (5.26) doesn't have a uniform bound with respect to M. In other words, as $M \to \infty$ the number of points in (5.26) might also tend to infinity even if the mesh size h is kept fixed. This reflects the fact that the variational discrete controls are indeed infinite dimensional functions even if they can be computed on computers.

Performing the sum (5.25) might be a tedious task in practice, especially for large values of M, and it might cause storage problems in computers because of (5.26). However, a remedy for this is to consider an approximation of (5.25) in X_h, that is to say,

$$\frac{1}{M}\sum_{i=1}^{M} \Pi_h\big(u_h(\omega_i)\big)$$

where $\Pi_h : L^2(D) \to X_h$ is a projection or an interpolation operator.

5.5 Numerical Example

In this section we verify numerically the assertion of Theorem 5.4.6, namely, the order of convergence (5.20) and the upper bound (5.21) for the computational cost. For this purpose, we consider the optimal control problem

$$\min_{u\in U_{ad}} J_\omega(u) = \frac{1}{2}\|y_\omega - z\|^2_{L^2(D)} + \frac{\alpha}{2}\|u\|^2_{L^2(D)} \tag{5.27}$$

subject to

$$\begin{aligned} -\nabla\cdot(a(\omega,x)\nabla y(\omega,x)) &= u(x) \quad \text{in } D,\\ y(\omega,x) &= 0 \quad \text{on } \partial D, \end{aligned}$$

where we define $D := (-0.5, 0.5) \times (-0.5, 0.5) \subset \mathbb{R}^2$ and $U_{ad} := L^2(D)$. The data is chosen as follows:

$$\begin{aligned} \alpha &= 10^{-2},\\ z(x) &= \sin(2\pi x_1)\cos(\pi x_2),\\ a(\omega,x) &= e^{\kappa(\omega,x)}, \end{aligned} \tag{5.28}$$

with the random field κ defined by

$$\begin{aligned} \kappa(x,\omega) :=& 0.84\cos(0.42\pi x_1)\cos(0.42\pi x_2)Y_1(\omega) + 0.45\cos(0.42\pi x_1)\sin(1.17\pi x_2)Y_2(\omega)\\ &+0.45\sin(1.17\pi x_1)\cos(0.42\pi x_2)Y_3(\omega) + 0.25\sin(1.17\pi x_1)\sin(1.17\pi x_2)Y_4(\omega), \end{aligned}$$

where $Y_i \sim N(0,1)$, $i = 1,\dots,4$, are independent normally distributed random variables. In fact, the random field κ approximates a *Gaussian* random field with zero mean and covariance function $C(x,\widetilde{x}) = e^{-\|x-\widetilde{x}\|_1}, x, \widetilde{x} \in D$, where $\|\cdot\|_1$ denotes the l_1-norm in $\mathbb{R}^2$. The terms in κ are the four leading terms in the associated Karhunen-Loève expansion, see [38] for more details. As a consequence, the random field a in (5.28) is a (truncated) lognormal field.

Assumptions A1–A2 are satisfied for all $t < 1/2$ for any lognormal random field a where $\log(a)$ has a Lipschitz continuous, isotropic covariance function and a mean function in $C^t(\overline{D})$, see [27, Proposition 2.4]. The property $1/a_{\min} \in L^p(\Omega)$ for all $p \in [1, \infty)$ is proved in [28, Proposition 2.3]. In our example the covariance function of κ is in fact analytic in $\overline{D} \times \overline{D}$. This gives realizations of $\kappa = \log(a)$ (and thus a) which belong to $C^1(\overline{D})$ almost surely. Hence Assumption A1 is satisfied for $t = 1$.

For a given realization of the coefficient $a(\omega, x)$, the problem in (5.27) is discretized by means of the variational discretization as described in Section 5.3. To compute the solution, which we denote by $u_{\omega,h}$, of the discrete control problem at a given mesh size h and realization ω, we solve the corresponding first order necessary conditions, which reads: there exist a state $y_{\omega,h} \in X_h$ and an adjoint state $p_{\omega,h} \in X_h$ such that

$$\begin{aligned}
\int_D a_\omega \nabla y_{\omega,h} \cdot \nabla v_h \, dx &= \int_D u_{\omega,h} v_h \, dx \quad \forall\, v_h \in X_h, \\
\int_D a_\omega \nabla p_{\omega,h} \cdot \nabla v_h \, dx &= \int_D (y_{\omega,h} - z) v_h \, dx \quad \forall\, v_h \in X_h, \\
p_{\omega,h} + \alpha u_{\omega,h} &= 0.
\end{aligned}$$

All the computations are done using a Matlab implementation running on 3.40 GHz 4×Intel Core i5-3570 processor with 7.8 GByte of RAM. For solving the linear system corresponding to the previous optimality conditions we use the Matlab backslash operation. Note that the linear system is a 3×3 block system of order $O(N_l)$ with sparse blocks. Hence the backslash costs about $O(N_l^{1.5}) = O(h_l^{-1.5d})$ operations in d-dimensional space. In summary, the cost to obtain one sample of the optimal control is $O(h_l^{-1.5d})$ and hence Assumption A3 is satisfied with $\gamma = 1.5 \times d$. We mention that it is possible to achieve the ideal value $\gamma = d$ by using a multigrid based method (see e.g. [39]).

5.5.1 FE Convergence Rate and Computational Cost

Observe that Theorem 5.4.6 requires the values of γ and s a priori. These can be estimated easily via numerical computations as illustrated in Figure 5.2. The value of γ for our solver can be deduced from Figure 5.2a where we plot the average cost (CPU-time in seconds) of computing u_{h_l}, the approximate solution of (5.27) for a given realization $a(\omega, x)$, versus the number of degrees of freedom N_l in the mesh when $h_l = 2^{-l}$ for $l = 0, \ldots, 8$. We see in the figure that the asymptotic behavior of the average cost is $O(N_l^{1.2})$ and thus $\gamma \approx 2.4$. This is slightly better than $\gamma = 1.5 \times d = 3$ which we expect in 2D space. Here, the average cost at a given N_l is considered to be the average of the total CPU-time in seconds required to solve (5.27) for 500 independent realizations of the coefficient $a(\omega, x)$ at the given mesh size h_l. To confirm that the cost per sample does not vary significantly across the realizations a_ω we plot the CPU-time in seconds with respect to N_l for individual realizations of a_ω in Figure 5.3.

The value of s can be obtained from Figure 5.2b where we plot $E_{500}[\|u_{\omega,h^*} - u_{\omega,h_l}\|_{L^2(D)}]$ versus $h_l^{-1} = 2^l$ for $l = 0, \ldots, 7$. Here, $E_{500}[\cdot]$ denotes the sample average of 500 independent samples. Furthermore, u_{ω,h_l} is the approximate solution of (5.27) at a given mesh size h_l and realization of $a(\omega, x)$. The control u_{ω,h^*} with $h^* := 2^{-8}$ is considered to be the reference solution since the exact

solution of (5.27) is not available at hand. We see clearly from the plot that the asymptotic behavior of the error is $O(h_l^2)$ as $h_l \to 0$, and thus $s = 1$. In fact, this quadratic order of convergence should be expected since the realizations of (5.28) belong to $C^t(\overline{D})$ with $t = 1$ and according to Theorem 5.3.4 we have $s = 1$ if $t = 1$. Furthermore, we observe that the error enters the asymptotic regime when the mesh size is $h_2 = 2^{-2}$ or smaller. This suggests that in the MLMC estimator (5.16) one should choose the mesh size h_0 of the coarsest level to be $h_0 = 2^{-2}$. Finally, for all the experiments used in Figure 5.2, the triangulation of the domain D when $l = 0$ consists of four triangles with only one degree of freedom located at the origin.

5.5.2 Multilevel Monte Carlo Simulation

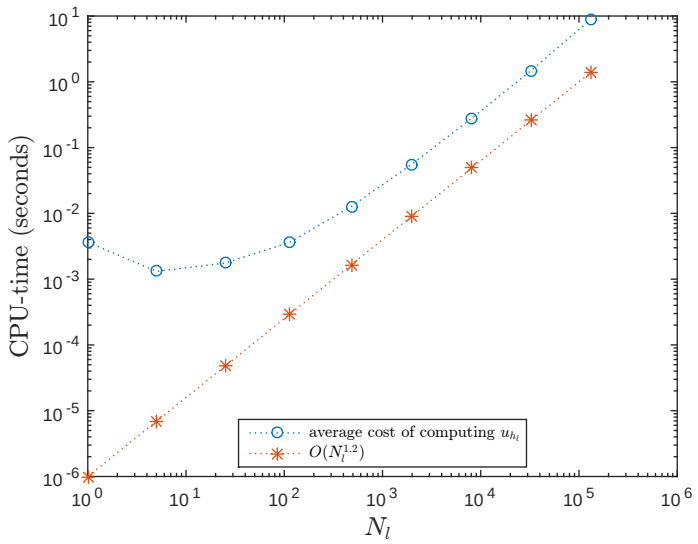
Having estimated the values of γ and s, we are in a position to verify the error estimate (5.20) and the upper bound for the computational cost in (5.21) for the MLMC estimator $E^L[u^*_{h_L}]$, where $u^*_{h_L}$ is the random variable associated to (5.27) as defined in (5.8). To this end, let $\{\mathcal{T}_{h_l}\}_{l=0}^L$, for $L = 1, \ldots, 5$, be sequences of triangulations of the domain D as described in Section 5.1. Here, we choose the mesh size h_0 of the initial coarse triangulation $\mathcal{T}_{h_0}$ to be $h_0 = 2^{-2}$ (the reason for this choice of h_0 is explained in the previous subsection). Since the expected value $\mathbb{E}[u^*]$ is not known explicitly, we consider the MLMC estimator $E^{L^*}[u^*_{h_{L^*}}]$ to be the reference expected value with $L^* = 6$ and $h_{L^*} = 2^{-8}$.

It is clear that the asymptotic behavior of the error $\mathbb{E}[u^*] - E^L[u^*_{h_L}]$ in the $L^2(\Omega, L^2(D))$-norm and in the $L^2(D)$-norm is the same. To simplify the computations we thus calculate the error in the $L^2(D)$-norm. Finally, for $L = 1, \ldots, 5$, we obtain the sequence $\{M_l\}_{l=0}^L$ of number of samples per refinement level l through solving (5.23) with the choice $c_0 = \frac{1}{2}$ by the `fmincon` function from the Matlab Optimization Toolbox. We round non-integer values in the sequence using the ceiling function. In Table 5.1, we report the sequences $\{M_l\}_{l=0}^L$, for $L = 1, \ldots, 5$, used in computing $E^L[u^*_{h_L}]$.

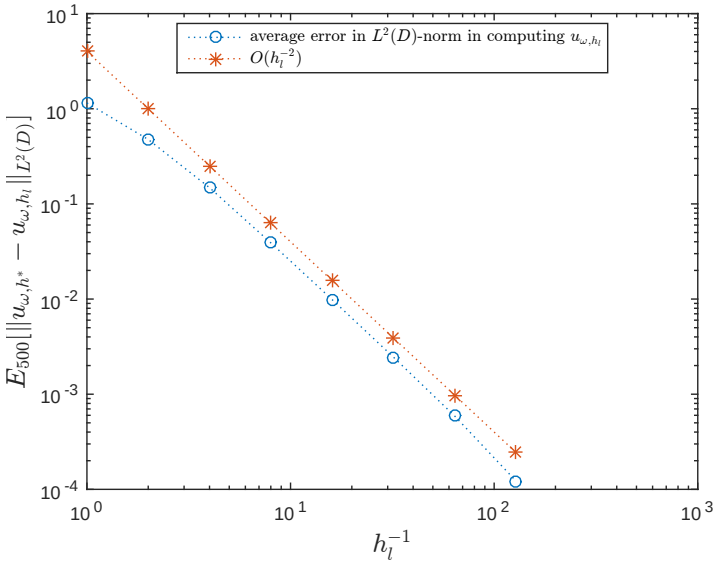
Figure 5.4a represents the plot of the CPU-time (in seconds) of computing $E^L[u^*_{h_L}]$ vs. the number of degrees of freedom N_L in a triangulation with mesh size $h_L = 2^{-(2+L)}$, for $L = 1, \ldots, 5$. It is clear from the figure that the computational cost is asymptotically bounded by $O(N_L^2)$ as $L \to \infty$. Since $N_L = O(h_L^{-2})$, this confirms the theoretical cost bound in (5.21) in the case $4s > \gamma$ (recall that $s = 1$ and $\gamma \approx 2.4$). In fact, the theoretical cost bound is sharp in this case.

Note that we did not verify the cost bound for the MC estimator in (5.15) due to limited computational time. In our example the MC estimator requires $O(h_L^{-2.4})$ more operations on level L than the MLMC estimator to achieve the same accuracy.

In addition we report the error associated with $E^L[u^*_{h_L}]$ in Figure 5.4b. We can see clearly that the best fitting curve for the error behaves like $O(h_L^2)$ as $L \to \infty$. This is predicted by (5.20).



(a) The average cost (CPU-time in seconds) of computing u_{h_l} the approximate solution of (5.27) for a given realization of $a(\omega, x)$ vs. the number of degrees of freedom N_l when $h_l = 2^{-l}$ for $l = 0, \ldots, 8$.



(b) The average error $E_{500}[\|u_{\omega,h^*} - u_{\omega,h_l}\|_{L^2(D)}]$ vs. $h_l^{-1} = 2^l$ for $l = 0, \ldots, 7$, where u_{ω,h_l} is the approximate solution of (5.27) for a given realization $a(\omega, x)$ and u_{ω,h^*} the reference solution with $h^* = 2^{-8}$.

Figure 5.2 The computations of s and γ for the estimate (5.10) and Assumption A3, respectively.

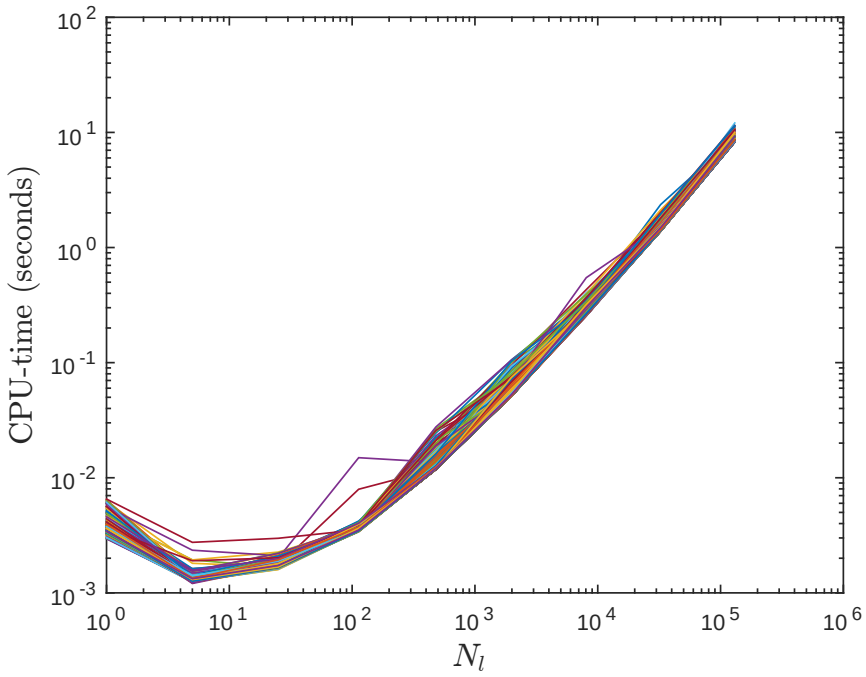
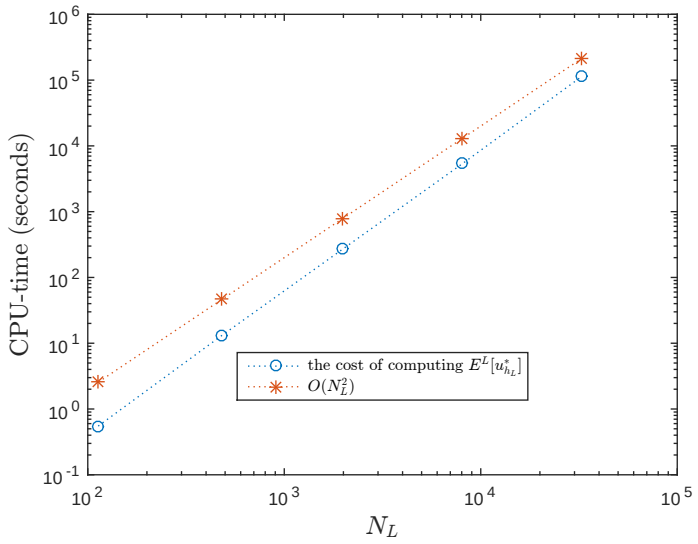
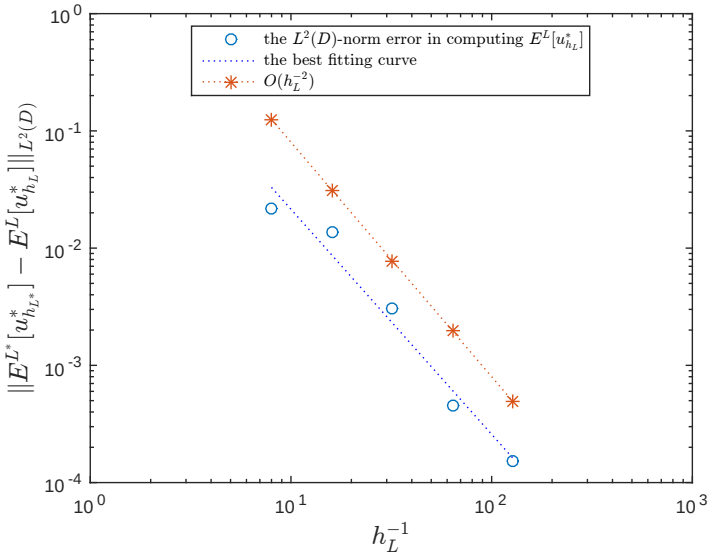


Figure 5.3 The cost (CPU-time in seconds) of computing u_{h_l} the approximate solution of (5.27) for a given realization of $a(\omega, x)$ vs. the number of degrees of freedom N_l when $h_l = 2^{-l}$ with $l = 0, \ldots, 8$ for 500 independent realizations.



(a) The cost (CPU-time in seconds) of computing $E^L[u^*_{h_L}]$ vs. the number of degrees of freedom N_L when $h_L = 2^{-(L+2)}$ for $L = 1, \ldots, 5$.



(b) The error $\|E^{L^*}[u^*_{h_{L^*}}] - E^L[u^*_{h_L}]\|_{L^2(D)}$ vs. $h_L^{-1} = 2^{(L+2)}$ for $L = 1, \ldots, 5$, where $E^{L^*}[u^*_{h_{L^*}}]$ with $L^* = 6$ is the reference expected value.

Figure 5.4 The error order of convergence and the computational cost upper bound for the MLMC estimator $E^L[u^*_{h_L}]$, where $u^*_{h_L}$ is the discrete random variable associated to (5.27) as defined in (5.8).

Table 5.1 The sequences $\{M_l\}_{l=0}^{L}$, for $L = 1, \ldots, 5$, used in computing $E^L[u^*_{h_L}]$, where M_l is the number of samples over a refinement level l which has a mesh size $h_l = 2^{-(2+l)}$.

L	M_0	M_1	M_2	M_3	M_4	M_5
1	183	24				
2	4815	631	83			
3	102258	13387	1753	230		
4	1948277	255053	33390	4372	573	
5	34878076	4565950	597737	78251	10244	1342

Conclusion

We have seen in the first part that it is possible to establish a sufficient condition for global minima of a certain class of optimal control problems of semilinear elliptic PDEs with pointwise constraints on the state and/or the control variables provided that the nonlinearity in the PDE satisfies certain growth conditions. This sufficient condition can also give information about the uniqueness of the global solutions. Moreover, one can establish in an analogous way to the continuous setting a similar condition for the variational discrete control problem. It turns out that a sequence of discrete unique global minima satisfying this condition uniformly converges strongly to the unique global minimum of the corresponding continuous control problem as the discretization parameter tends to zero. A rate of convergence for the sequence of the discrete unique global minima can be established using this sufficient condition as well. The numerical experiments show that this convergence rate is optimal. In addition, we managed to compute the unique global minima for several examples. The results of this part are partially published in [1].

In the second part we considered optimal control problems of elliptic PDEs with stochastic coefficients. The task was to compute the expected value of the optimal controls corresponding to the different realizations of the random coefficient of the state equation utilizing the finite element Monte Carlo and multilevel Monte Carlo methods and to carry out the associated error analysis. However, the computed expected value needs not to be an optimal control in general. The results of this part are published in [2].

Zusammenfassung

Im ersten Teil betrachten wir ein Optimalsteuerungsproblem mit semilinearer, elliptischer, partieller Differentialgleichung sowie punktweisen Restriktionen an Zustand und/oder Steuerung. Eine hinreichende Bedingung für globale Minima der Optimierungsaufgabe wird bewiesen, wenn die Nichtlinearität der semilinearen PDE bestimmte Wachstumsbedingungen erfüllt. Die gleiche Bedingung gilt auch für das diskrete Gegenstück zur Optimierungsaufgabe. Wir haben gezeigt, dass eine Folge der globalen Minima der diskreten Optimierungsprobleme gegen ein globales Minimum der stetigen Optimierungsaufgabe konvergiert, wenn die Folge diese Bedingung erfüllt. Zusätzlich haben wir mit der Hilfe dieser Bedingung eine Fehlerabschätzung und eine hinreichende Bedingung zweiter Ordnung für lokale Minima bewiesen. Die Ergebnisse dieses Teils sind teilweise in [1] veröffentlicht.

Im zweiten Teil untersuchen wir ein elliptisches Steuerungsproblem, wobei die PDE eine Zufallsvariable als Koeffizient besitzt. Unser Ziel ist den Erwartungswert der optimalen Steuerungen mit der Monte Carlo und Multilevel Monte Carlo Finite Elemente Methode zu berechnen und die verbundene Analyse durchzuführen. Dieser Erwartungswert ist im Allgemeinen keine globale Steuerung. Die Ergebnisse dieses Teils sind in [2] veröffentlicht.

Appendix A

A.1 Properties of ϕ

Lemma A.1.1 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be of class C^2 and monotonically increasing such that (2.2) is satisfied. Then*

$$\phi'(s) \le c_1\big(1 + |s|^{r_1}\big),\ s \in \mathbb{R}, \quad r_1 = \frac{r}{r-1}, \tag{A.1}$$

$$|\phi(s)| \le c_0\big(1 + |s|^{r_0}\big),\ s \in \mathbb{R}, \quad r_0 = \frac{2r-1}{r-1}, \tag{A.2}$$

where $c_1, c_0 > 0$ depending on r, M and $\phi'(0)$.

Proof: We will show (A.1) and (A.2) for $s \ge 0$. The case $s \le 0$ can be treated analogously. To this end we see that (2.2) implies

$$\frac{r}{r-1}\Big|\frac{d}{dt}\big(\phi'(t) + \epsilon\big)^{\frac{r-1}{r}}\Big| \le M$$

for any $\epsilon > 0$, from which after integrating on $[0, s]$ for some $s \ge 0$, we deduce that

$$\int_0^s \frac{d}{dt}\big(\phi'(t) + \epsilon\big)^{\frac{r-1}{r}}\, dt \le \frac{r-1}{r} M \int_0^s 1\, dt.$$

Evaluating the integrals in the previous inequality and taking the limit $\epsilon \to 0^+$ yields

$$\phi'(s) \le c_1(1 + s^{\frac{r}{r-1}}) \quad \forall\, s \ge 0,$$

where $c_1 > 0$ depending on r, M and $\phi'(0)$. This gives (A.1). Integrating the previous inequality on $[0, s]$ once more gives

$$\phi(s) \le c(1 + s + s^{\frac{2r-1}{r-1}}) \quad \forall\, s \ge 0,$$

where $c > 0$. Considering the cases $s \le 1$ and $s \ge 1$ we see that the previous estimate implies

$$\phi(s) \le c_0(1 + s^{\frac{2r-1}{r-1}}) \quad \forall\, s \ge 0,$$

for $c_0 > 0$ chosen appropriately. This gives (A.2) and the proof is complete. ■

Lemma A.1.2 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be of class C^2 and monotonically increasing such that (2.2) is satisfied. Then we have for $a, b \in \mathbb{R}$*

$$\left|\int_0^1 \phi'(ta + (1-t)b) - \phi'(b)\, dt\right| \le |a-b| L_r \left(\int_0^1 \phi'(ta + (1-t)b)\, dt\right)^{\frac{1}{r}},$$

where

$$L_r := M\left(\frac{r-1}{2r-1}\right)^{\frac{r-1}{r}}.$$

Proof: We start by noticing that

$$\begin{aligned}\int_0^1 \phi'(ta+(1-t)b)-\phi'(b)\,dt &= \int_0^1\int_0^t \phi''(\tau a+(1-\tau)b)(a-b)\,d\tau\,dt\\ &= (a-b)\int_0^1 (1-t)\phi''(ta+(1-t)b)\,dt.\end{aligned}$$

Therefore, taking the absolute value and using (2.2) we get

$$\begin{aligned}\left|\int_0^1 \phi'\big(ta+(1-t)b\big)-\phi'(b)\,dt\right| &\le |a-b|M\int_0^1 (1-t)\phi'(ta+(1-t)b)^{\frac{1}{r}}\,dt\\ &\le |a-b|M\|1-t\|_{L^{r'}(0,1)}\\ &\quad\times\left(\int_0^1 \phi'(ta+(1-t)b)\,dt\right)^{\frac{1}{r}},\end{aligned}$$

where $\frac{1}{r}+\frac{1}{r'}=1$. It is easy to see that

$$\|1-t\|_{L^{r'}(0,1)} = \left(\frac{1}{r'+1}\right)^{\frac{1}{r'}} = \left(\frac{r-1}{2r-1}\right)^{\frac{r-1}{r}}.$$

Denoting $M\|1-t\|_{L^{r'}(0,1)}$ by L_r completes the proof. ■

Lemma A.1.3 *Let $\phi:\mathbb{R}\to\mathbb{R}$ be of class C^1. Then for any $1\le r\le\infty$, there holds*

$$\|\phi(v)-\phi(w)\|_{L^r(\Omega)} \le L(m)\|v-w\|_{L^r(\Omega)}$$

for all $v,w\in L^\infty(\Omega)$ such that $\|v\|_{L^\infty(\Omega)},\|w\|_{L^\infty(\Omega)}\le m$. Here $L(m)>0$ is a constant depending on $m>0$.

Proof: The result is a direct consequence of [73, Lemma 4.11]. ■

Lemma A.1.4 *Let $\Omega\subset\mathbb{R}^2$ be open and bounded. Let $\phi:\mathbb{R}\to\mathbb{R}$ be of class C^1 such that its first derivative ϕ' satisfies*

$$|\phi'(s)|\le c\big(1+|s|^r\big),\ s\in\mathbb{R},\quad \text{for some } c>0 \text{ and } r>1.$$

Then for any $1\le t<\infty$ there holds

$$\|\phi(v)-\phi(w)\|_{L^t(\Omega)} \le L(m)\|v-w\|_{H^1(\Omega)}$$

for all v, $w\in H_0^1(\Omega)$ with $\|v\|_{H^1(\Omega)}$, $\|w\|_{H^1(\Omega)}\le m$. Here, $L(m)>0$ is a constant depending on $m>0$. Moreover, if $(y_k)\subset H_0^1(\Omega)$ such that $y_k\rightharpoonup y$, then $\phi(y_k)\to\phi(y)$ in $L^t(\Omega)$ for $1\le t<\infty$.

Proof: Let $v, w \in H_0^1(\Omega)$ be given such that $\|v\|_{H^1(\Omega)}, \|w\|_{H^1(\Omega)} \leq m$ for some constant $m > 0$. Then, for a.e. $x \in \Omega$, it follows from the mean value theorem and the assumption on $|\phi'|$ that

$$\begin{aligned}
|\phi(v(x)) - \phi(w(x))| &= \left| \int_0^1 \phi'(tv(x) + (1-t)w(x))(v(x) - w(x))\, dt \right| \\
&\leq |v(x) - w(x)| \int_0^1 |\phi'(tv(x) + (1-t)w(x))|\, dt \\
&\leq c|v(x) - w(x)| \int_0^1 \big(1 + |tv(x) + (1-t)w(x)|^r\big)\, dt \\
&\leq c|v(x) - w(x)| \int_0^1 \big(1 + t|v(x)|^r + (1-t)|w(x)|^r\big)\, dt \\
&\leq c|v(x) - w(x)|\big(1 + |v(x)|^r + |w(x)|^r\big),
\end{aligned}$$

Note that $r > 1$ and the function $|\cdot|^r$ is thus convex. Moreover, $\phi(v(\cdot))$ and $\phi(w(\cdot))$ are measurable by the continuity of ϕ. Taking the norm $\|\cdot\|_{L^t(\Omega)}$ for some $1 \leq t < \infty$ of both sides of the previous inequality gives

$$\|\phi(v) - \phi(w)\|_{L^t(\Omega)} \leq c\||v - w|(1 + |v|^r + |w|^r)\|_{L^t(\Omega)}.$$

For the given $r > 1$ and a given $1 \leq t < \infty$, we choose a real number p such that $p > rt$ and we define $q := \frac{tp}{p - rt}$. Applying the generalization of Hölder's inequality with the exponents

$$\frac{1}{q} + \frac{1}{(\frac{p}{r})} = \frac{1}{t}$$

to the right hand side of the previous inequality results in

$$\begin{aligned}
\|\phi(v) - \phi(w)\|_{L^t(\Omega)} &\leq c\||v - w|(1 + |v|^r + |w|^r)\|_{L^t(\Omega)} \\
&\leq c\|v - w\|_{L^q(\Omega)}\|1 + |v|^r + |w|^r\|_{L^{\frac{p}{r}}(\Omega)} \\
&\leq c\|v - w\|_{L^q(\Omega)}\big(1 + \|v\|_{L^p(\Omega)}^r + \|w\|_{L^p(\Omega)}^r\big) \\
&\leq c\|v - w\|_{H^1(\Omega)}\big(1 + \|v\|_{H^1(\Omega)}^r + \|w\|_{H^1(\Omega)}^r\big) \\
&\leq L(m)\|v - w\|_{H^1(\Omega)},
\end{aligned}$$

where we utilized the continuous (which is also compact) embedding $H_0^1(\Omega) \hookrightarrow L^s(\Omega)$, $1 \leq s < \infty$. Here, $L(m)$ is a positive constant depending on m.

The intermediate steps in the above inequality suggest that if $y_k \rightharpoonup y$ in $H_0^1(\Omega)$, then $\phi(y_k) \to \phi(y)$ in $L^t(\Omega)$ for $1 \leq t < \infty$ because then $y_k \to y$ in $L^q(\Omega)$ from the compact embedding $H_0^1(\Omega) \hookrightarrow L^q(\Omega)$ with q being as above. ■

A.2 Hölder Continuous Functions

Lemma A.2.1 *Let $\Omega \subset \mathbb{R}^n$ be open. For $f \in C^{1,\beta}(\bar{\Omega})$ for some $0 < \beta \leq 1$ there holds*

$$|f(x) - f(y)| \leq c|x - y|^\gamma \tag{A.3}$$

for some $c > 0$ independent of $|x - y|$ where

$$\gamma = \begin{cases} 1 & \forall\, x, y \in \bar{\Omega}, \\ 1 + \beta & \forall\, x, y \in \bar{\Omega} \text{ with } \nabla f(x) = 0. \end{cases}$$

Proof: The estimate (A.3) holds with $\gamma = 1$ for any $x, y \in \bar{\Omega}$ since the function f is Lipschitz continuous. On the other hand, if $\nabla f(x) = 0$ we see from the Mean Value Theorem that

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \nabla f(x + t(y - x)) \cdot (y - x)\, dt \\ &= \int_0^1 [\nabla f(x + t(y - x)) - \nabla f(x)] \cdot (y - x)\, dt \\ &\leq c|x - y|^{1+\beta} \end{aligned}$$

where we used the fact that $\nabla f \in [C^{0,\beta}(\bar{\Omega})]^n$. ■

A.3 Tietze's Extension Theorem

Lemma A.3.1 *Let $K, \tilde{K} \subset \mathbb{R}^n$ be compact sets such that $K \subset \tilde{K}$ and let $y_a, y_b \in C(\tilde{K})$ satisfy $y_a(x) \leq y_b(x)$, $x \in \tilde{K}$. Then for a given $z \in C(K)$ with $y_a(x) \leq z(x) \leq y_b(x)$, $x \in K$ there exists $\tilde{z} \in C(\tilde{K})$ such that $\tilde{z}|_K = z$ and $y_a(x) \leq \tilde{z}(x) \leq y_b(x)$, $x \in \tilde{K}$.*

Proof: Let $z \in C(K)$ be given. Then the Tietze's extension theorem, see for example [68, Theorem 20.4], asserts that there exists a compactly supported continuous function $\hat{z} \in C_c(\mathbb{R}^n)$ such that $\hat{z}|_K = z$. Next, define $\tilde{z} : \tilde{K} \to \mathbb{R}$ by

$$\tilde{z}(x) := \max\Big(y_a(x), \min\Big(\hat{z}(x), y_b(x)\Big)\Big).$$

It is clear that $\tilde{z} \in C(\tilde{K})$, $\tilde{z}(x) = z(x)$, $x \in K$, and $y_a(x) \leq \tilde{z}(x) \leq y_b(x)$, $x \in \tilde{K}$. Hence, $\tilde{z}$ is the desired function and the proof is complete. ■

A.4 Young's Inequality

Lemma A.4.1 (Young's inequality) *Let $x, y \geq 0$, $p, q > 1$, $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

Lemma A.4.2 *We have for $a, b \geq 0, \lambda, \mu > 0$ that*

$$a^\lambda b^\mu \leq \frac{\lambda^\lambda \mu^\mu}{(\lambda + \mu)^{\lambda+\mu}} (a + b)^{\lambda+\mu}.$$

Proof: Apply Young's inequality to $p = \frac{\lambda+\mu}{\lambda}$, $q = \frac{\lambda+\mu}{\mu}$ and $x = (pa)^{\frac{1}{p}}$, $y = (qb)^{\frac{1}{q}}$. ■

Lemma A.4.3 *Let $a, b \geq 0$, $\epsilon > 0$, $p, q > 1$, $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$ab \leq \frac{\epsilon a^p}{p} + \frac{\epsilon^{1-q} b^q}{q}.$$

Proof: Apply Young's inequality to $x = a$, $y = \dfrac{b}{\epsilon}$. ■

A.5 Gagliardo–Nirenberg Inequality

Theorem A.5.1 (Gagliardo–Nirenberg interpolation inequality)
For $2 \leq q < \infty$ we define $\mu = 1 - \frac{2}{q}$ as well as

$$GN_q := \sup_{f \in H^1(\mathbb{R}^2), f \neq 0} \frac{\|f\|_{L^q(\mathbb{R}^2)}}{\|f\|^{1-\mu}_{L^2(\mathbb{R}^2)} \|\nabla f\|^{\mu}_{L^2(\mathbb{R}^2)}}.$$

Then $GN_q \leq C_q := \min(C_q^{(1)}, C_q^{(2)}, C_q^{(3)})$, where

$$C_q^{(1)} = \big(\mu C_{2,2\mu}\big)^{-\mu}, \quad \text{if } q \geq 4; \tag{A.4}$$

$$C_q^{(2)} = \frac{1}{\sqrt{\mu^\mu (1-\mu)^{1-\mu}}} \left(2\pi B\left(1, \frac{2(1-\mu)}{2\mu}\right)\right)^{\mu/2} k_B\left(\frac{4}{2+2\mu}\right); \tag{A.5}$$

$$C_q^{(3)} = \left(\frac{1}{\pi}\right)^{\frac{q-2}{2q}} \prod_{j=2}^{\infty} \left(\frac{2^j}{2^j + q - 2}\right)^{\frac{2^j+2-q}{2^j q}}. \tag{A.6}$$

Here,

$$C_{2,s} = 2^{1/s} \left(\frac{2-s}{s-1}\right)^{(s-1)/s} \left(2\pi B\left(\frac{2}{s}, 3 - \frac{2}{s}\right)\right)^{1/2},\ 1 < s < 2;\ C_{2,1} = 2\sqrt{\pi};$$

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad a, b > 0$$

$$k_B(p) = \left(\frac{p}{2\pi}\right)^{1/p} \left(\frac{p'}{2\pi}\right)^{-1/p'}, \quad \frac{1}{p} + \frac{1}{p'} = 1.$$

Proof: The bounds (A.4) and (A.5) can be found in the paper [75] by Veling. We remark that $GN_q = \lambda_{2,\mu}^{-1}$, where $\lambda_{2,\mu}$ is defined in [75, (1.7)]. The estimate (A.4) is [75, (1.31)] (note that $\mu \geq \frac{1}{2} \Leftrightarrow q \geq 4$), while (A.5) is [75, (1.42),(1.43)], where the latter bound has been proved by Nasibov in [61].

Let us now turn to the proof of (A.6). To begin, we claim that for all $k \in \mathbb{N}_0$

$$\|f\|_{L^q} \leq \left(\frac{1}{\pi}\right)^{\frac{1}{2}(1-\frac{q_k}{q})} \prod_{j=2}^{k+1} \left(\frac{2^j}{2^j + q - 2}\right)^{\frac{2^j+2-q}{2^j q}} \|f\|^{\frac{q_k}{q}}_{L^{q_k}} \|\nabla f\|^{1-\frac{q_k}{q}}_{L^2}, \tag{A.7}$$

where

$$q_k = 2^{-k}\Big(q + 2(2^k - 1)\Big).$$

The inequality clearly holds for $k = 0$. Suppose that (A.7) is true for some $k \in \mathbb{N}_0$. We infer from Theorem 1 in [36] for the case $d = 2$ that

$$\|f\|_{L^{2p}} \leq A \|f\|^{1-\theta}_{L^{p+1}} \|\nabla f\|^{\theta}_{L^2}, \qquad 1 < p < \infty. \tag{A.8}$$

Here,

$$A = \left(\frac{y(p-1)^2}{4\pi}\right)^{\frac{\theta}{2}} \left(\frac{2y-2}{2y}\right)^{\frac{1}{2p}} \left(\frac{\Gamma(y)}{\Gamma(y-1)}\right)^{\frac{\theta}{2}} \quad \text{with} \quad \theta = \frac{2(p-1)}{4p},$$
$$y = \frac{p+1}{p-1}.$$

Using the formula for y and observing that $\Gamma(y) = (y-1)\Gamma(y-1)$, the expression for A can be simplified to

$$A = \left(\frac{1}{\pi}\right)^{\frac{\theta}{2}} \left(\frac{p+1}{2}\right)^{\frac{\theta}{2}-\frac{1}{2p}}.$$

We apply (A.8) for $p = \frac{1}{2}q_k$ and obtain

$$\|f\|_{L^{q_k}} \le A \|f\|^{1-\theta}_{L^{\frac{1}{2}q_{k+1}}} \|\nabla f\|^{\theta}_{L^2}, \tag{A.9}$$

where

$$A = \left(\frac{1}{\pi}\right)^{\frac{\theta}{2}} \left(\frac{\frac{1}{2}q_k+1}{2}\right)^{\frac{\theta}{2}-\frac{1}{q_k}} \quad \text{and} \quad \theta = \frac{q_k-2}{2q_k}.$$

Since $\frac{1}{2}q_k + 1 = q_{k+1}$ we find that

$$A = \left(\frac{1}{\pi}\right)^{\frac{\theta}{2}} \left(\frac{q_{k+1}}{2}\right)^{\frac{\theta}{2}-\frac{1}{q_k}} \quad \text{and} \quad \theta = 1 - \frac{q_{k+1}}{q_k},$$

which, inserted into (A.9) yields

$$\|f\|_{L^{q_k}} \le \left(\frac{1}{\pi}\right)^{\frac{\theta}{2}} \left(\frac{q_{k+1}}{2}\right)^{\frac{\theta}{2}-\frac{1}{q_k}} \|f\|^{1-\theta}_{L^{q_{k+1}}} \|\nabla f\|^{\theta}_{L^2}. \tag{A.10}$$

Using the induction hypothesis we infer

$$\|f\|_{L^q} \le \left(\frac{1}{\pi}\right)^{\frac{1}{2}(1-\frac{q_k}{q})+\frac{\theta}{2}\frac{q_k}{q}} \left(\frac{q_{k+1}}{2}\right)^{\left(\frac{\theta}{2}-\frac{1}{q_k}\right)\frac{q_k}{q}} \times \prod_{j=2}^{k+1} \left(\frac{2^j}{2^j+q-2}\right)^{\frac{2^j+2-q}{2^j q}} \|f\|^{(1-\theta)\frac{q_k}{q}}_{L^{q_{k+1}}} \|\nabla f\|^{1-\frac{q_k}{q}+\theta\frac{q_k}{q}}_{L^2}.$$

Elementary calculations show that

$$\frac{1}{2}\left(1-\frac{q_k}{q}\right) + \frac{\theta}{2}\frac{q_k}{q} = \frac{1}{2}\left(1-\frac{q_{k+1}}{q}\right),$$
$$\left(\frac{q_{k+1}}{2}\right)^{\left(\frac{\theta}{2}-\frac{1}{q_k}\right)\frac{q_k}{q}} = \left(\frac{2^{k+2}}{2^{k+2}+q-2}\right)^{\frac{2^{k+2}+2-q}{2^{k+2}q}},$$
$$(1-\theta)\frac{q_k}{q} = \frac{q_{k+1}}{q},$$
$$1 - \frac{q_k}{q} + \theta\frac{q_k}{q} = 1 - \frac{q_{k+1}}{q},$$

which implies (A.7) for $k+1$. The result now follows by sending $k \to \infty$ in (A.7) and by observing that $\lim_{k\to\infty} q_k = 2$. ∎

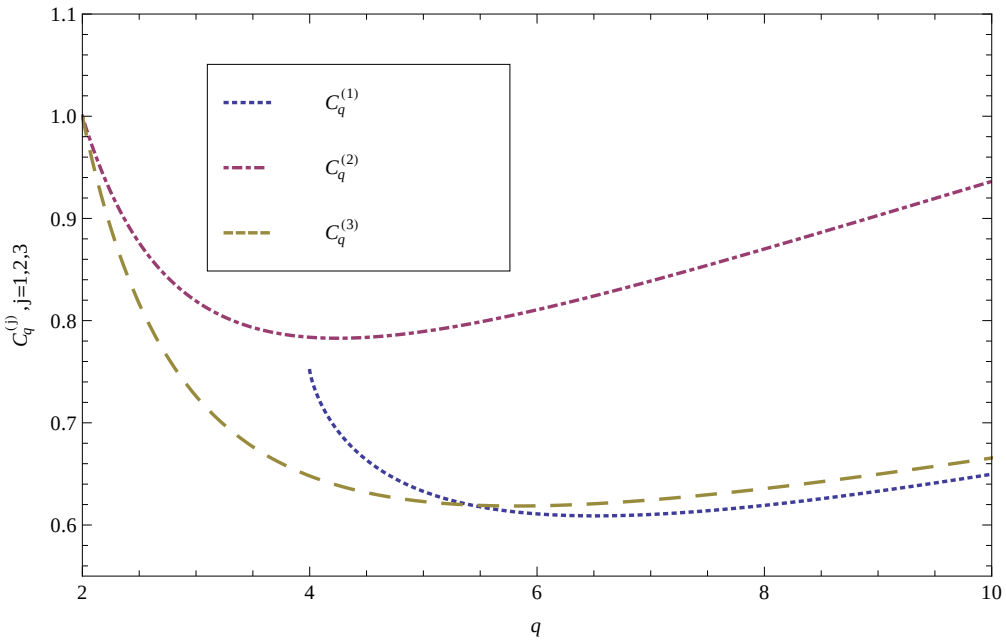


Figure A.1 The values of the constants $C_q^{(2)}, C_q^{(3)}$ over the range $2 \leq q \leq 10$ and $C_q^{(1)}$ over $4 \leq q \leq 10$.

Figure A.1 illustrates the values of the constants (A.4)–(A.6) for a certain range of q, namely, $2 \leq q \leq 10$ for $C_q^{(2)}, C_q^{(3)}$ and $4 \leq q \leq 10$ for $C_q^{(1)}$. We can see clearly that the values of $C_q^{(1)}$ are smaller than those of $C_q^{(2)}, C_q^{(3)}$ for approximately $q \geq 6$. In order to derive a computable upper bound on $C_q^{(3)}$ we note that $\frac{2^j}{2^j+q-2} \leq 1$ for $j \in \mathbb{N}$, and therefore

$$C_q^{(3)} \leq \left(\frac{1}{\pi}\right)^{\frac{q-2}{2q}} \prod_{j=2}^{k-1} \left(\frac{2^j}{2^j+q-2}\right)^{\frac{2^j+2-q}{2^j q}}, \ k \geq k_0,$$

where $k_0 \geq 2$ is chosen so large that $2^{k_0} + 2 - q \geq 0$. In our calculations we used $k - 1 = 200$. All the computations of these constants are done using *Mathematica* 8.

Bibliography

[1] Ahmad Ahmad Ali, Klaus Deckelnick, and Michael Hinze. Global minima for semilinear optimal control problems. *Computational Optimization and Applications*, 65(1):261–288, 2016.

[2] Ahmad Ahmad Ali, Elisabeth Ullmann, and Michael Hinze. Multilevel monte carlo analysis for optimal control of elliptic pdes with random coefficients. *SIAM/ASA J. Uncertainty Quantification, accepted for publication.*

[3] Nadir Arada, Eduardo Casas, and Fredi Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem. *Computational Optimization and Applications*, 23(2):201–229, 2002.

[4] Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis*. Springer Science+Business Media, 2009.

[5] Andrea Barth, Christoph Schwab, and Nathaniel Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011.

[6] Peter Benner, Akwum Onwuta, and Martin Stoll. Block-diagonal preconditioning for optimal control problems constrained by PDEs with uncertain inputs. Technical Report MPIMD/15-05, Max Planck Institute Magdeburg, April 2015. Accepted for SIAM J. Matrix Anal. Appl.

[7] Peter Benner, Akwum Onwuta, and Martin Stoll. Low-rank solvers for unsteady Stokes-Brinkman optimal control problem with random data. Technical Report MPIMD/15-10, Max Planck Institute Magdeburg, Juli 2015. Accepted for Comput. Methods Appl. Mech. Engrg.

[8] Claudio Bierig and Alexey Chernov. Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems. *Numer. Math.*, 130(4):579–613, 2015.

[9] A. Borzì. Multigrid and sparse-grid schemes for elliptic control problems with random coefficients. *Comput. Vis. Sci.*, 13(4):153–160, 2010.

[10] A. Borzì, V. Schulz, C. Schillings, and G. von Winckel. On the treatment of distributed uncertainties in PDE-constrained optimization. *GAMM-Mitt.*, 33(2):230–246, 2010.

[11] A. Borzì and G. von Winckel. Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients. *SIAM J. Sci. Comput.*, 31(3):2172–2192, 2009.

[12] A. Borzì and G. von Winckel. A POD framework to determine robust controls in PDE optimization. *Comput. Vis. Sci.*, 14(3):91–103, 2011.

[13] Susanne Brenner and Ridgway Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2007.

[14] Eduardo Casas. L^2 Estimates for the Finite Element Method for the Dirichlet Problem with Singular Data. *Numerische Mathematik*, 47(4):627–632, 1985.

[15] Eduardo Casas. Control of an elliptic problem with pointwise state constraints. *SIAM Journal on Control and Optimization*, 24(6):1309–1318, 1986.

[16] Eduardo Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM Journal on Control and Optimization*, 31(4):993–1006, 1993.

[17] Eduardo Casas. Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints. *ESAIM: Control, Optimisation and Calculus of Variations*, 8:345–374, 2002.

[18] Eduardo Casas. Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems. *Advances in Computational Mathematics*, 26(1-3):137–153, 2007.

[19] Eduardo Casas. Necessary and sufficient optimality conditions for elliptic control problems with finitely many pointwise state constraints. *ESAIM: Control, Optimisation and Calculus of Variations*, 14(3):575–589, 2008.

[20] Eduardo Casas, Juan Carlos De Los Reyes, and Fredi Tröltzsch. Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints. *SIAM Journal on Optimization*, 19(2):616–643, 2008.

[21] Eduardo Casas and Luis Alberto Fernández. Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state. *Applied Mathematics and Optimization*, 27(1):35–56, 1993.

[22] Eduardo Casas and Mariano Mateos. Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints. *SIAM journal on control and optimization*, 40(5):1431–1454, 2002.

[23] Eduardo Casas and Mariano Mateos. Uniform convergence of the FEM. Applications to state constrained control problems. *Comput. Appl. Math.*, 21(1):67–100, 2002. Special issue in memory of Jacques-Louis Lions.

[24] Eduardo Casas, Mariano Mateos, and Boris Vexler. New regularity results and improved error estimates for optimal control problems with state constraints. *ESAIM. Control, Optimisation and Calculus of Variations*, 20(3):803, 2014.

[25] Eduardo Casas and Fredi Tröltzsch. Recent advances in the analysis of pointwise state-constrained elliptic optimal control problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 16(3):581–600, 2010.

[26] Eduardo Casas and Fredi Tröltzsch. Second order optimality conditions and their role in pde control. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 117(1):3–44, 2015.

[27] J. Charrier, R. Scheichl, and A. L. Teckentrup. Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. *SIAM J. Numer. Anal.*, 51(1):322–352, 2013.

[28] Julia Charrier. Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(1):216–246, 2012.

[29] P. Chen, A. Quarteroni, and G. Rozza. Multilevel and weighted reduced basis methods for stochastic optimal control problems constrained by Stokes equation. *Numer. Math.* Published online 05 July 2015.

[30] Peng Chen and Alfio Quarteroni. Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):364–396, 2014.

[31] Peng Chen, Alfio Quarteroni, and Gianluigi Rozza. Stochastic optimal Robin boundary control problems of advection-dominated elliptic equations. *SIAM J. Numer. Anal.*, 51(5):2700–2722, 2013.

[32] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, 2011.

[33] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*, volume 152 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, second edition, 2014.

[34] Klaus Deckelnick and Michael Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM Journal on Numerical Analysis*, 45(5):1937–1953, 2007.

[35] Klaus Deckelnick and Michael Hinze. A finite element approximation to elliptic control problems in the presence of control and state constraints. *Hamburger Beiträge zur Angewandten Mathematik*, 2007-01, 2007.

[36] Manuel Del Pino and Jean Dolbeault. Best constants for Gagliardo-Nirenberg inequalities and applications to nonlinear diffusions. *J. Math. Pures Appl. (9)*, 81(9):847–875, 2002.

[37] Sebastian Garreis and Michael Ulbrich. Constrained Optimization with Low-Rank Tensors and Applications to Parametric Problems with PDEs. Technical Report 5301, Optimization Online, January 2016.

[38] Roger G. Ghanem and Pol D. Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[39] Wei Gong, Hehu Xie, and Ningning Yan. A multilevel correction method for optimal controls of elliptic equations. *SIAM J. Sci. Comput.*, 37(5):A2198–A2221, 2015.

[40] Pierre Grisvard. *Elliptic problems in nonsmooth domains*, volume 69. SIAM, 2011.

[41] Max Gunzburger and Ju Ming. Optimal control of stochastic flow over a backward-facing step using reduced-order modeling. *SIAM J. Sci. Comput.*, 33(5):2641–2663, 2011.

[42] Max D. Gunzburger, Hyung-Chun Lee, and Jangwoon Lee. Error estimates of stochastic optimal Neumann boundary control problems. *SIAM J. Numer. Anal.*, 49(4):1532–1552, 2011.

[43] W. Hackbusch. *Elliptic differential equations: theory and numerical treatment*. Number 18. Springer Science+Business Media, 1992.

[44] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2002.

[45] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.

[46] Michael Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Computational Optimization and Applications*, 30(1):45–61, 2005.

[47] Michael Hinze and Christian Meyer. Stability of semilinear elliptic optimal control problems with pointwise state constraints. *Computational Optimization and Applications*, 52(1):87–114, 2012.

[48] Michael Hinze and Arnd Rösch. Discretization of optimal control problems. In *Constrained Optimization and Optimal Control for Partial Differential Equations*, pages 391–430. Springer, 2012.

[49] Michael Hinze and Fredi Tröltzsch. Discrete concepts versus error analysis in pde-constrained optimization. *GAMM-Mitteilungen*, 33(2):148–162, 2010.

[50] L. S. Hou, J. Lee, and H. Manouzi. Finite element approximations of stochastic optimal control problems constrained by stochastic elliptic PDEs. *J. Math. Anal. Appl.*, 384(1):87–103, 2011.

[51] David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and Their Applications*, volume 31. Siam, 1980.

[52] Ralf Kornhuber, Christoph Schwab, and Maren-Wanda Wolf. Multilevel Monte Carlo finite element methods for stochastic elliptic variational inequalities. *SIAM J. Numer. Anal.*, 52(3):1243–1268, 2014.

[53] D. P. Kouri. A multilevel stochastic collocation algorithm for optimization of PDEs with uncertain coefficients. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):55–81, 2014.

[54] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty. *SIAM J. Sci. Comput.*, 35(4):A1847–A1879, 2013.

[55] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders. Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty. *SIAM J. Sci. Comput.*, 36(6):A3011–A3029, 2014.

[56] Angela Kunoth and Christoph Schwab. Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs. *SIAM J. Control Optim.*, 51(3):2442–2471, 2013.

[57] Martin Lazar and Enrique Zuazua. Averaged control and observation of parameter-depending wave equations. *C. R. Math. Acad. Sci. Paris*, 352(6):497–502, 2014.

[58] Hyung-Chun Lee and Jangwoon Lee. A stochastic Galerkin method for stochastic control problems. *Commun. Comput. Phys.*, 14(1):77–106, 2013.

[59] Dmitriy Leykekhman and Boris Vexler. Finite Element Pointwise Results on Convex Polyhedral Domains. *SIAM J. Numer. Anal.*, 54(2):561–587, 2016.

[60] Pedro Merino, Fredi Tröltzsch, and Boris Vexler. Error estimates for the finite element approximation of a semilinear elliptic control problem with state constraints and finite dimensional control space. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(1):167–188, 2010.

[61] S. M. Nasibov. On optimal constants in some Sobolev inequalities and their applications to the nonlinear Schrödinger equation. *Soviet Math. Dokl.*, 40:110–115, (1990). translation of Dokl. Akad. Nauk SSSR 307:538-542 (1989).

[62] Federico Negri, Andrea Manzoni, and Gianluigi Rozza. Reduced basis approximation of parametrized optimal flow control problems for the Stokes equations. *Comput. Math. Appl.*, 69(4):319–336, 2015.

[63] Federico Negri, Gianluigi Rozza, Andrea Manzoni, and Alfio Quarteroni. Reduced basis method for parametrized elliptic optimal control problems. *SIAM J. Sci. Comput.*, 35(5):A2316–A2340, 2013.

[64] Ira Neitzel, Johannes Pfefferer, and Arnd Rösch. Finite element discretization of state-constrained elliptic optimal control problems with semilinear state equation. *SIAM Journal on Control and Optimization*, 53(2):874–904, 2015.

[65] Ira Neitzel and Winnifried Wollner. A Priori L^2-Discretization Error Estimates for the State in Elliptic Optimization Problems with Pointwise Inequality State Constraints. 2016. INS Preprint No. 1606.

[66] Rolf Rannacher and Ridgway Scott. Some optimal error estimates for piecewise linear finite element approximations. *mathematics of computation*, 38(158):437–445, 1982.

[67] Eveline Rosseel and Garth N. Wells. Optimal control with stochastic PDE constraints and uncertain controls. *Comput. Methods Appl. Mech. Engrg.*, 213/216:152–167, 2012.

[68] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[69] Alfred H Schatz. A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains. I. *Mathematics of Computation*, 34(149):77–91, 1980.

[70] Alfred H Schatz and Lars B Wahlbin. Interior maximum norm estimates for finite element methods. *Mathematics of Computation*, 31(138):414–442, 1977.

[71] A. L. Teckentrup, R. Scheichl, M. B. Giles, and E. Ullmann. Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. *Numer. Math.*, 125(3):569–600, 2013.

[72] Hanne Tiesler, Robert M. Kirby, Dongbin Xiu, and Tobias Preusser. Stochastic collocation for optimal control problems with stochastic PDE constraints. *SIAM J. Control Optim.*, 50(5):2659–2682, 2012.

[73] Fredi Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010.

[74] Michael Ulbrich. Semismooth newton methods for operator equations in function spaces. *SIAM Journal on Optimization*, 13(3):805–841, 2002.

[75] E. J. M. Veling. Lower bounds for the infimum of the spectrum of the Schrödinger operator in $\mathbb{R}^N$ and the Sobolev inequalities. *JIPAM. J. Inequal. Pure Appl. Math.*, 3(4):Article 63, 22 pp. (electronic), 2002.

[76] Eberhard Zeidler. *Nonlinear Functional Analysis and its Applications I*. Springer-Verlag, New York, 1986. Fixed-Point Theorems, Translated from the German by Peter R. Wadsack.

[77] Eberhard Zeidler. *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*. Springer-Verlag, New York, 1990.

[78] Jochem Zowe and Stanisław Kurcyusz. Regularity and stability for the mathematical programming problem in banach spaces. *Applied mathematics and Optimization*, 5(1):49–62, 1979.

[79] Enrique Zuazua. Averaged control. *Automatica J. IFAC*, 50(12):3077–3087, 2014.