A PRIORI FINITE ELEMENT ERROR ANALYSIS FOR OPTIMAL CONTROL OF THE OBSTACLE PROBLEM

CHRISTIAN MEYER AND OLIVER THOMA

Abstract. An optimal control problem governed by an unilateral obstacle problem is considered. The problem is discretzed by using linear finite elements for the state and the obstacle and a variational discrete approach for the control. Based on strong stationarity and a quadratic growth condition we establish a priori error estimates which turn out to be quasi-optimal under additional assumptions on the data. The theoretical findings are illustrated by two numerical tests.

1 Introduction

In the present paper, we consider the following optimal control problem governed by the obstacle problem:

$$\left.\begin{aligned} \min \quad & J(y,u) := \frac{1}{2}\,\|y-y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\,\|u\|^2_{L^2(\Omega)}, \\ \text{s.t.} \quad & y \in K, \quad \int_\Omega \nabla y \cdot \nabla v\,dx \geq \int_\Omega u(v-y)\,dx \quad \forall\, v \ \in K, \end{aligned}\right\} \tag{P}$$

with

$$K := \{y \in H^1_0(\Omega) |\, y(x) \geq \psi(x) \text{ a.e. in } \Omega\}. \tag{1.1}$$

The precise assumptions on the data Ω, y_d, ν and ψ will be made below.

Our aim is to derive a priori error estimates for the finite element (FE) discretization of (P). Up to the authors' best knowledge, error estimates for the discretization of optimal control problems governed by variational inequalities (VIs) have not been addressed so far in literature. In contrast to this, there is a multitude of contributions in the field of a priori error analysis for optimal control of PDEs. We only refer to Falk [1973], Geveci [1979], Arada et al. [2002], Rösch [2004], Hinze [2005], Deckelnick and Hinze [2007], Meyer [2008] and the references therein. Compared to the PDE-constrained case, optimal control problems subject to VIs provide a lack of regularity, since the associated solution operator is in general not Gâteaux-differentiable. This complicates the derivation of necessary and sufficient optimality conditions. A remedy is to use modified stationarity concepts such as the so called strong stationarity which is the most rigorous concept and has been derived by Mignot and Puel [1984] for the case of (P). Furthermore, in Kunisch and Wachsmuth [2009] the strong stationarity conditions are used to establish second-order sufficient conditions. Other stationarity concepts are Clarke-, Bouligand- and Mordukhovich-stationarity which can be derived e.g. by regularization techniques, see for instance Ito and Kunisch [2000], Hintermüller [2001], Hintermüller and Kopacka [2009], Jarušek et al. [2010].

Our error analysis is based on the strong stationarity conditions and yields a convergence rate of order $1-\varepsilon$ for the L^2-norm of the control and the H^1-norm of

This work was supported by a DFG grant within the Priority Program SPP 1253 (*Optimization with Differential Equations*), which is gratefully acknowledged.

the state in case of a two-dimensional domain under additional assumptions on the data. Since (P) is a nonlinear and thus non-convex optimal control problem, one can only expect this convergence rate to hold locally. In view of the limited regularity of the optimal solution, caused by a lack of regularity of the Lagrange multipliers, this approximation rate can be seen to be optimal.

The paper is organized as follows: After stating the precise assumptions and some well known results in Section 2, we introduce the discretization of (P) in Section 3. Afterwards Section 4 is devoted to the derivation of strong stationarity for the discrete version of (P), which turns out to be the discrete counterpart to the infinite dimensional strong stationarity conditions derived in Mignot and Puel [1984]. These stationarity conditions allow to prove some uniform boundedness result for the discrete solutions, which is essential for the convergence analysis presented in Section 5. Another important ingredient for the error analysis is a quadratic growth condition which holds under the second-order sufficient condition established in Kunisch and Wachsmuth [2009]. The general convergence result of Section 5 requires an a priori estimate for the L^2-error of the FE-discretization of the obstacle problem. In contrast to the Poisson problem, the Nitsche trick for the obstacle problem is unkown in general, and Section 6 shows how to employ L^∞-error estimates to circumvent this difficulty in the two dimensional case. The associated L^∞-error estimates for the obstacle problem are derived in Appendix A based on a technique introduced in Nitsche [1977]. Finally in Section 7 two numerical examples are presented that illustrate the theoretical findings.

2 Preliminaries

2.1. Notation and assumptions. As usual the dual of $H_0^1(\Omega)$ w.r.t. the L^2-inner product is denoted by $H^{-1}(\Omega)$. The dual pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ is denoted by the symbol $\langle ., . \rangle$. Moreover, we introduce the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ by

$$a(y, v) := \int_\Omega \nabla y \cdot \nabla v \, dx \qquad \forall y, v \in H_0^1(\Omega). \tag{2.1}$$

The coercivity constant of a will be denoted by α, i.e.

$$a(v, v) \geq \alpha \|v\|^2_{H^1(\Omega)} \quad \forall \, v \in H_0^1(\Omega).$$

Given two bounded sets $A, B \subset \mathbb{R}^d$, $d = 2, 3$, we use the notation $A \subset\subset B$ if A is strictly contained in B, i.e. $\operatorname{dist}(\bar{A}, \partial B) > 0$. Finally, throughout the paper, c and C denote positive generic constants.

The standing assumptions which are assumed throughout the paper are as follows:

Assumption 2.1. *We impose the following hypotheses on the data in* (P)*:*

- *(i) The domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is assumed to be convex and its boundary Γ is supposed to be a polygon (polyhedron).*
- *(ii) The desired state satisfies $y_d \in L^2(\Omega)$, and $\nu > 0$ is a fixed real number.*
- *(iii) The lower bound ψ fulfills $\psi \in W^{2,\infty}(\Omega)$ and $\psi \leq 0$ a.e. on Γ.*

For simplicity we do not consider a discretization of the desired state y_d. If however y_d provides sufficient regularity, e.g. $y_d \in H^2(\Omega)$, then a discretization of y_d can easily be incorporated into the analysis.

2.2. Known results. In the following we collect some known results concerning the obstacle problem as well as for the optimal control problem (P). The first lemma covers existence and uniqueness for the obstacle problem. The proof is standard and can for instance be found in Kinderlehrer and Stampacchia [1980].

Lemma 2.2. *For every $u \in H^{-1}(\Omega)$ there exists a unique solution $y \in K$ of*

$$a(y, v - y) \geq \langle u,\, v - y \rangle \quad \forall\, v \ \in K. \tag{2.2}$$

Moreover, the associated solution operator $S : H^{-1}(\Omega) \to K \subset H_0^1(\Omega)$, mapping u to y, is globally Lipschitz continuous with Lipschitz constant $L = 1/\alpha$, where α is the coercivity constant of a.

Remark 2.3. *We point out that S is in general* not *Gâteaux-differentiable (unless the biactive set has measure zero), cf. Mignot [1976]. Therefore, the optimal control problem* (P) *is non-smooth and represents an mathematical program with equilibrium constraints (MPEC) in function space.*

In the subsequent, S will be considered with different domains and ranges, for simplicity denoted by the same symbol. Based on Lemma 2.2, we introduce the reduced functional $f : L^2(\Omega) \to \mathbb{R}$ by $f(u) := J(S(u), u)$ and the associated reduced problem, which is equivalent to (P) by construction:

$$\text{(P)} \quad \Leftrightarrow \quad \min_{u \in L^2(\Omega)} f(u).$$

Proposition 2.4. *Problem* (P) *admits a solution $\bar{u} \in L^2(\Omega)$.*

Proof. The proof follows standard arguments: since $J \geq 0$, there exists an infimal sequence of controls which is uniformly bounded in $L^2(\Omega)$ due to $\nu > 0$. Thus there is a subsequence converging strongly in $H^{-1}(\Omega)$, and since $S : H^{-1}(\Omega) \to H_0^1(\Omega)$ is continuous, we have strong convergence of the states. The weak lower semicontinuity of J then allows passage to the limit. □

Since S is nonlinear, one cannot expect the solution of (P) to be unique.

3 Discretization

Now, we turn to the discretization of (P). First, let us introduce a family of meshes $\{\mathcal{T}_h\}$ with mesh size $h > 0$. The mesh $\mathcal{T}_h$ consists of open cells T (triangles, tetrahedra) such that

$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h} \overline{T}, \tag{3.1}$$

and is assumed to be regular, see e.g. Brenner and Scott [1994]. The *mesh size* is defined by

$$h := \max_{T \in \mathcal{T}_h} h_T \quad \text{with} \quad h_T := \operatorname{diam}(T). \tag{3.2}$$

For each $T \in \mathcal{T}_h$, we associate the diameter of the largest ball contained in T, denoted by R_T. We impose the following regularity assumptions on $\{\mathcal{T}_h\}$:

Assumption 3.1. *There exist two positive constants ρ and R such that*

$$\frac{h_T}{R_T} \leq R\,, \quad \frac{h}{h_T} \leq \rho \tag{3.3}$$

hold for all cells $T \in \bigcup_{h>0} \mathcal{T}_h$.

For the discretization of (P) we employ the following standard piecewise linear FE space:

$$V_h = \{v \in C(\overline{\Omega}) : v|_T \in \mathcal{P}_1 \ \forall T \in \mathcal{T}_h\}, \quad V_h^0 = V_h \cap H_0^1(\Omega). \tag{3.4}$$

By $\{x_1, ..., x_n\}$ we denote the interior nodes of the domain, while $\{x_{n+1}, ..., x_{n+m}\}$ are the nodes on $\Gamma = \partial\Omega$. Moreover, by $\{\varphi_1, ..., \varphi_{n+m}\}$ we denote the nodal basis of V_h, i.e. $\varphi_i \in V_h$, $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, ..., n+m$. Then $V_h^0 = \operatorname{span}(\varphi_1, ..., \varphi_n)$ and $\dim(V_h^0) = n$. Given a function $v_h \in V_h^0$, we denote by $\boldsymbol{v} = (v_i)_{i=1}^n \in \mathbb{R}^n$ its vector of coefficients w.r.t. the basis $\{\varphi_1, ..., \varphi_n\}$, i.e. $v_h = \sum_{i=1}^n v_i \varphi_i$ (and analogously for a function in V_h).

This leads to the following discrete version of (2.2):

$$y_h \in K_h, \quad a(y_h, v_h - y_h) \geq \langle u,\, v_h - y_h\rangle \quad \forall\, v_h \in K_h, \tag{3.5}$$

with

$$K_h := \{v_h \in V_h^0 : v_h(x) \geq (I_h\psi)(x)\ \forall\, x \in \Omega\}.$$

Here, $I_h : C(\bar{\Omega}) \to V_h$ denotes the standard Lagrange interpolation operator. Note that, thanks to $W^{2,\infty}(\Omega) \hookrightarrow C(\bar{\Omega})$ for $d = 2, 3$, we have that $I_h\psi$ is well defined. If we set $\boldsymbol{\psi} := \big(\psi(x_i)\big)_{i=1}^n$, then it follows

$$K_h = \{v_h \in V_h^0 : \boldsymbol{v} \geq \boldsymbol{\psi}\}.$$

Here and in the following, inequalities of the form $\boldsymbol{a} \geq \boldsymbol{b}$ with vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ are to be understood in a componentwise sense, i.e., $a_i \geq b_i$ for all $i = 1, ..., n$. Note in this context, that $\psi|_\Gamma \leq 0$ so that $(I_h\psi)|_\Gamma \leq 0$ and $v_h \geq (I_h\psi)$ is automatically fulfilled on Γ if $\boldsymbol{v} \geq \boldsymbol{\psi}$ (which is a condition for the nodes in the interior only). In this way, (3.5) can be rewritten as

$$\begin{aligned} &\sum_{i,j=1}^n y_i a(\varphi_i, \varphi_j)(v_j - y_i) \geq \sum_{j=1}^n \langle u,\, \varphi_j\rangle (v_j - y_i) \quad \forall\, \boldsymbol{v} \in \mathbb{R}^n \text{ with } \boldsymbol{v} \geq \boldsymbol{\psi} \\ &\quad \Leftrightarrow \quad \boldsymbol{y}^T \mathcal{A}(\boldsymbol{v} - \boldsymbol{y}) \geq \boldsymbol{b}_u^T(\boldsymbol{v} - \boldsymbol{y}) \quad \forall\, \boldsymbol{v} \in \mathbb{R}^n \text{ with } \boldsymbol{v} \geq \boldsymbol{\psi}. \end{aligned} \tag{3.6}$$

Herein $\boldsymbol{b}_u := (\langle u,\, \varphi_i\rangle)_{i=1}^n \in \mathbb{R}^n$ and

$$\mathcal{A} \in \mathbb{R}^{n\times n}, \quad \mathcal{A}_{ij} := a(\varphi_i, \varphi_j) \tag{3.7}$$

denotes the stiffness matrix.

Lemma 3.2. *For all $u \in H^{-1}(\Omega)$, (3.5) and (3.6) admit a unique solution $y_h \in K_h$ $\boldsymbol{y} \in \mathbb{R}^n$, respectively.*

Proof. The proof follows standard arguments. Due to convexity, (3.5) is the necessary and sufficient optimality condition of

$$\left.\begin{array}{ll} \min & \dfrac{1}{2}\, a(v, v) - \langle u,\, v\rangle \\ \text{s.t.} & v \in K_h. \end{array}\right\} \tag{L_h}$$

To prove that (L_h) possesses a unique solution, observe that its objective is bounded from below by $-1/2\|u\|^2_{H^{-1}(\Omega)}$ giving in turn the existence of a minimizing sequence $\{y_n\} \subset K_h$. Because of the radial unboundedness of the objective, there is subsequence, converging weakly to some y_h in $H_0^1(\Omega)$. Since V_h is a closed subspace and K_h is therefore convex and closed, thus weakly closed, we have $y_h \in K_h$. The weak lower semicontinuity of the objective then implies optimality of y_h, and uniqueness follows from strict convexity. □

The above result allows to define the solution operator of (3.5) as follows:

$$S_h : H^{-1}(\Omega) \to V_h \subset H_0^1(\Omega), \qquad u \mapsto y_h. \tag{3.8}$$

As in case of S, we will consider S_h with different domains and ranges again denoted by the same symbol. Using the solution operator S_h, we define the following *variational discretization* of (P):

$$\min_{u \in L^2(\Omega)} f_h(u) := \frac{1}{2}\|S_h(u) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)}. \tag{P_h}$$

Observe that the control u is not discretized. However, as we will see in Corollary 4.2 each local optimum of (P_h) is an element of V_h such that is suffices to restrict the controls to the set V_h. In this way, a fully discrete optimization problem is obtained, cf. Remark 4.3.

The operator S_h is globally Lipschitz continuous as the following result shows:

Lemma 3.3. *Let $y_h^{(1)} \in K_h$ and $y_h^{(2)} \in K_h$ solve problem (3.5) for given $u_1 \in H^{-1}(\Omega)$ and $u_2 \in H^{-1}(\Omega)$, respectively. Then, there holds*

$$\|S_h(u_1) - S_h(u_2)\|_{H^1(\Omega)} \le L\,\|u_1 - u_2\|_{H^{-1}(\Omega)}, \tag{3.9}$$

where L is the Lipschitz constant of S and thus independent of h.

Proof. The proof is straight forward. We set $y_h^{(1)} = S_h(u_1)$ and $y_h^{(2)} = S_h(u_2)$ and insert $v_h = y_h^{(1)} \in K_h$ and $v_h = y_h^{(2)} \in K_h$ in (3.5) with $u = u_2$ and $u = u_1$, respectively. Adding the arising inequalities and using coercivity of a then yields

$$\alpha\|y_h^{(1)} - y_h^{(2)}\|^2_{H^1(\Omega)} \le a(y_h^{(1)} - y_h^{(2)}, y_h^{(1)} - y_h^{(2)}) \le \|u_1 - u_2\|_{H^{-1}(\Omega)}\,\|y_h^{(1)} - y_h^{(2)}\|_{H^1(\Omega)},$$

giving the assertion with $L = 1/\alpha$, which is exactly the Lipschitz constant of S. □

Using the above result, an argument completely analogous to the proof of Proposition 2.4 yields the following

Proposition 3.4. *Problem (P_h) has a solution.*

Again, due to the nonlinearity of S_h, one cannot expect the solution to be unique.

4 Strong stationarity for the discrete problem

In the sequel we will establish first-order necessary optimality condtions for (P_h). These conditions imply that local solutions to (P_h) are discrete functions so that the variational discrete problem (P_h) is equivalent to a fully discrete optimization problem, see Corollary 4.2 and Remark 4.3 below. To this end, we reformulate (P_h) in terms of a mathematical problem with complementarity constraints (MPCC). We first introduce the discrete slack variable by

$$\boldsymbol{\xi} := \mathcal{A}\boldsymbol{y} - \boldsymbol{b}_u \in \mathbb{R}^n.$$

Then, by standard arguments, the discrete variational inequality (3.6) can be rewritten in form of a complementarity system

$$\begin{aligned} \mathcal{A}\boldsymbol{y} &= \boldsymbol{b}_u + \boldsymbol{\xi} \\ \boldsymbol{y} \ge \boldsymbol{\psi}, \quad \boldsymbol{\xi}^T(\boldsymbol{y} - \boldsymbol{\psi}) &= 0, \quad \boldsymbol{\xi} \ge 0. \end{aligned} \tag{4.1}$$

Hence, the variational discrete optimal control problem (P_h) is equivalent to the following MPCC:

$$\left.\begin{array}{ll} \min & \frac{1}{2}\,\boldsymbol{y}^T\mathcal{M}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{b}_d + \frac{\nu}{2}\,\|u\|^2_{L^2(\Omega)} \\ \text{s.t.} & \mathcal{A}\boldsymbol{y} = (\langle u,\,\varphi_i\rangle)_{i=1}^n + \boldsymbol{\xi} \\ & \min(\boldsymbol{y}-\boldsymbol{\psi},\boldsymbol{\xi}) = \boldsymbol{0} \end{array}\right\} \qquad (\mathbf{P}_h)$$

with $\boldsymbol{b}_d := \left(\int_\Omega y_d\varphi_j dx\right)_{j=1}^n$. Moreover, $\mathcal{M}\in\mathbb{R}^{n\times n}$, $\mathcal{M}_{ij} = \int_\Omega \varphi_i\varphi_j\,dx$, is the mass matrix and the function $\min : \mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}^n$ is defined by $\min(\boldsymbol{a},\boldsymbol{b})_i = \min\{a_i, b_i\}$, $i = 1, ..., n$. Employing the equivalence of (P_h) to $(\mathbf{P}_h)$ we now establish first-order optimality conditions for (P_h). From finite dimensional MPECs it is well known that standard Karush-Kuhn-Tucker (KKT) conditions cannot be expected and the most rigorous first-order conditions are given by the so-called strong stationarity conditions, cf. e.g. Scheel and Scholtes [2000]. The next theorem states these conditions for the case of (P_h).

Theorem 4.1. *Let $\bar u_h \in L^2(\Omega)$ be a local optimal solution of (P_h) with associated state $\bar y_h = \sum_{i=1}^n \bar y_i\varphi_i \in V_h$ and slack variable $\bar{\boldsymbol{\xi}}\in\mathbb{R}^n$. Then there exist an adjoint state $p_h = \sum_{i=1}^n p_i\varphi_i \in V_h$ and a multiplier $\boldsymbol{\mu}\in\mathbb{R}^n$ such that the following strong stationarity system is fulfilled:*

$$\mathcal{A}\bar{\boldsymbol{y}} = \left(\int_\Omega \bar u_h\varphi_j dx\right)_{j=1}^n + \bar{\boldsymbol{\xi}} \tag{4.2a}$$

$$\bar{\boldsymbol{\xi}} \ge 0, \quad \bar{\boldsymbol{\xi}}^T(\bar{\boldsymbol{y}} - \boldsymbol{\psi}) = 0, \quad \bar{\boldsymbol{y}} \ge \boldsymbol{\psi} \tag{4.2b}$$

$$\mathcal{A}^T\boldsymbol{p} = \mathcal{M}\bar{\boldsymbol{y}} - \left(\int_\Omega y_d\varphi_j dx\right)_{j=1}^n + \boldsymbol{\mu} \tag{4.2c}$$

$$(\bar{\boldsymbol{y}} - \boldsymbol{\psi}) * \boldsymbol{\mu} = \boldsymbol{0}, \quad \bar{\boldsymbol{\xi}} * \boldsymbol{p} = \boldsymbol{0} \tag{4.2d}$$

$$\mu_k \le 0,\ p_k \ge 0 \quad \forall\, k\in\{1,...,n\} \text{ with } \bar y_k - \psi(x_k) = \bar\xi_k = 0 \tag{4.2e}$$

$$\nu\bar u_h(x) + p_h(x) = 0 \quad \text{a.e. in } \Omega, \tag{4.2f}$$

where $*$ denotes the Hadamard product and $\bar{\boldsymbol{y}} = (\bar y_i)_{i=1}^n$ and $\boldsymbol{p} = (p_i)_{i=1}^n$.

Corollary 4.2. *By (4.2f) we find that, if $\bar u_h \in L^2(\Omega)$ is a local optimal solution of (P_h), then $\bar u_h$ is a discrete function, i.e. $\bar u_h \in V_h^0$.*

Remark 4.3. *In view of Corollary 4.2, it suffices to search the optimum of (P_h) in V_h^0. Consequently, after a suitable discretization of the data, (P_h) is a fully discretized optimal control problem which can be solved by algorithms for finite dimensional MPECs.*

Proof of Theorem 4.1. Since the complementarity constraint in $(\mathbf{P}_h)$ only involves $\boldsymbol{y}$ and $\boldsymbol{\xi}$, thus finite dimensional variables, the proof is completely along the lines of the theory for finite dimensional MPECs, cf. for instance Scheel and Scholtes [2000]. Nevertheless, for convenience of the reader, we recall the arguments applied to $(\mathbf{P}_h)$. To this end, let $(\bar{\boldsymbol{y}},\bar{\boldsymbol{\xi}},\bar u_h)\in\mathbb{R}^n\times\mathbb{R}^n\times L^2(\Omega)$ be an arbitary but fixed local minimum of $(\mathbf{P}_h)$. The proof will be performed in three steps. First we establish Fritz-John type conditions based on Clarke's subdifferential calculus. Afterwards these conditions will be tightened to C-stationarity and in a third step to strong stationarity conditions.

Step (1): Fritz-John conditions
The objective in $(\mathbf{P}_h)$ and the function $h : (\boldsymbol{y},\boldsymbol{\xi},u) \mapsto \mathcal{A}\boldsymbol{y} - (\langle u,\,\varphi_i\rangle)_{i=1}^n - \boldsymbol{\xi}$ are clearly continuously Fréchet-differentiable from $\mathbb{R}^n\times L^2(\Omega)$ to $\mathbb{R}$ and $\mathbb{R}^n\times\mathbb{R}^n\times L^2(\Omega)$ to $\mathbb{R}^n$, respectively. Moreover, the function $\min : \mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}^n$ is globally

Lipschitz continuous. Therefore, [Clarke, 1976, Theorem 1 and Corollary 1] implies the existence of nonvanishing multipliers $(\alpha, \boldsymbol{p}, \boldsymbol{\sigma}) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$ so that $\alpha \geq 0$ and

$$\alpha\,\nu \int_\Omega \bar{u}_h\, v\, dx + \boldsymbol{p}^T \left(\int_\Omega v\, \varphi_i\, dx \right)_{i=1}^n = 0 \quad \forall\, v \in L^2(\Omega) \tag{4.3a}$$

$$\begin{pmatrix} \alpha\, \mathcal{M}\bar{\boldsymbol{y}} - \alpha\, \boldsymbol{b}_d - \mathcal{A}^T \boldsymbol{p} \\ \boldsymbol{p} \end{pmatrix} + \sum_{i=1}^n \sigma_i \boldsymbol{g}_i = 0 \tag{4.3b}$$

with

$$\boldsymbol{g}_i \in \partial_{(\boldsymbol{y},\boldsymbol{\xi})} \min\left\{ \bar{y}_i - \psi(x_i), \bar{\xi}_i \right\} = \begin{cases} (\boldsymbol{e}_i, \boldsymbol{0}), & \text{if } \bar{y}_i = \psi(x_i),\ \bar{\xi}_i > 0 \\ \operatorname{conv}\{(\boldsymbol{e}_i, \boldsymbol{0}), (\boldsymbol{0}, \boldsymbol{e}_i)\}, & \text{if } \bar{y}_i - \psi(x_i) = \bar{\xi}_i = 0 \\ (\boldsymbol{0}, \boldsymbol{e}_i), & \text{if } \bar{y}_i > \psi(x_i),\ \bar{\xi}_i = 0. \end{cases}$$

Herein, $\boldsymbol{e}_i \in \mathbb{R}^n$, $i = 1, ..., n$ denotes the i-th unit vector, $\partial_{(\boldsymbol{y},\boldsymbol{\xi})} \min$ is the Clarke subdifferential, and conv denotes the convex hull. Hence, for all $i = 1, ..., n$, there exists $\lambda_i \in [0, 1]$ so that

$$\boldsymbol{g}_i = \lambda_i \begin{pmatrix} \boldsymbol{e}_i \\ \boldsymbol{0} \end{pmatrix} + (1 - \lambda_i) \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{e}_i \end{pmatrix} \quad \text{and} \quad \lambda_i\big((\bar{y}_i - \psi(x_i)\big) = (1 - \lambda_i)\bar{\xi}_i = 0. \tag{4.4}$$

Now define

$$\mu_i := \lambda_i\, \sigma_i \quad \text{and} \quad \omega_i := (1 - \lambda_i)\sigma_i. \tag{4.5}$$

Then (4.3b) implies

$$\alpha \mathcal{M}\bar{\boldsymbol{y}} - \alpha \boldsymbol{b}_d - \mathcal{A}^T \boldsymbol{p} + \boldsymbol{\mu} = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{p} + \boldsymbol{\omega} = \boldsymbol{0}.$$

Moreover, in view of (4.4), we have $(\bar{y}_i - \psi(x_i))\mu_i = \sigma_i \lambda_i (\bar{y}_i - \psi(x_i)) = 0$ and $\bar{\xi}_i p_i = -\sigma_i (1 - \lambda_i)\bar{\xi}_i = 0$ which already gives (4.2d). Furthermore, by construction of $\boldsymbol{\mu}$ and $\boldsymbol{p}$ there holds

$$\mu_i\, p_i = -\lambda_i (1 - \lambda_i)\sigma_i^2 \leq 0 \quad \forall\, i = 1, ..., n.$$

In summary, we obtain the following optimality conditions:

$$\alpha \mathcal{M}\bar{\boldsymbol{y}} - \alpha \boldsymbol{b}_d - \mathcal{A}^T \boldsymbol{p} + \boldsymbol{\mu} = 0 \tag{4.6a}$$

$$\alpha \nu \bar{u}_h(x) + p_h(x) = 0 \quad \text{a.e. in } \Omega \tag{4.6b}$$

$$\alpha \geq 0 \tag{4.6c}$$

$$\boldsymbol{\mu} * (\bar{\boldsymbol{y}} - \boldsymbol{\psi}) = \boldsymbol{0}, \quad \boldsymbol{p} * \bar{\boldsymbol{\xi}} = \boldsymbol{0} \tag{4.6d}$$

$$\boldsymbol{\mu} * \boldsymbol{p} \leq \boldsymbol{0}, \tag{4.6e}$$

where (4.6b) follows from (4.3a) with $p_h := \sum_{i=1}^n p_i \varphi_i$.

Step (2): C-stationarity
Next we show that $\alpha > 0$ to obtain a qualified optimality system. We argue by contradiction and suppose $\alpha = 0$. Then (4.6b) yields

$$\sum_{i=1}^n p_i \varphi_i(x) = 0 \quad \text{a.e. in } \Omega \quad \Rightarrow \quad \boldsymbol{p} = 0.$$

Consequently, $\boldsymbol{\mu} = 0$ obtains by (4.6a), and (4.5) implies $\sigma_i = \omega_i + \lambda_i \sigma_i = -p_i + \mu_i = 0$ for all $i = 1, ..., n$. Thus we have $(\alpha, \boldsymbol{p}, \boldsymbol{\sigma}) = (0, \boldsymbol{0}, \boldsymbol{0})$, but according to [Clarke, 1976, Theorem 1], $(\alpha, \boldsymbol{p}, \boldsymbol{\sigma})$ must not vanish, which yields the desired contradiction.

Thus we have $\alpha > 0$ and appropriate scaling yields $\alpha = 1$ so that the following C-stationarity conditions are obtained:

$$\mathcal{A}^T \boldsymbol{p} = \mathcal{M}\bar{\boldsymbol{y}} - \boldsymbol{b}_d + \boldsymbol{\mu} \tag{4.7a}$$

$$\nu \bar{u}_h(x) + p_h(x) = 0 \quad \text{a.e. in } \Omega \tag{4.7b}$$

$$\boldsymbol{\mu} * (\bar{\boldsymbol{y}} - \boldsymbol{\psi}) = \boldsymbol{0}, \quad \boldsymbol{p} * \bar{\boldsymbol{\xi}} = \boldsymbol{0} \tag{4.7c}$$

$$\boldsymbol{\mu} * \boldsymbol{p} \leq \boldsymbol{0}. \tag{4.7d}$$

Thus, we have already verified (4.2c), (4.2d), and (4.2f). Due to (4.7b), we have $\bar{u}_h \in V_h^0$, (see also Corollary 4.2). Therefore, $\bar{\boldsymbol{u}} \in \mathbb{R}^n$ defined through $\bar{u}_h = \sum_{i=1}^n \bar{u}_i \varphi_i$ is a local minimum of the following finite dimensional MPCC

$$\left.\begin{aligned} \min \quad & \frac{1}{2}\, \boldsymbol{y}^T \mathcal{M} \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{b}_d + \frac{\nu}{2}\, \boldsymbol{u}^T \mathcal{M} \boldsymbol{u} \\ \text{s.t.} \quad & \mathcal{A}\boldsymbol{y} = \mathcal{M}\boldsymbol{u} + \boldsymbol{\xi} \\ & \min(\boldsymbol{y} - \boldsymbol{\psi}, \boldsymbol{\xi}) = \boldsymbol{0}. \end{aligned}\right\} \tag{MPCC}$$

Step (3): Strong stationarity
It remains to prove the sign condition (4.2e) on the multipliers $\boldsymbol{\mu}$ and $\boldsymbol{p}$. To this end, assume the contrary, i.e. there is an index $\ell \in \{1, ..., n\}$ with $\bar{y}_\ell - \psi(x_\ell) = \bar{\xi}_\ell = 0$ and w.l.o.g. $\mu_\ell > 0$ (the case $p_\ell < 0$ can be discussed analogously). Let us consider the following standard quadratic program:

$$\left.\begin{aligned} \min \quad & \frac{1}{2}\, \boldsymbol{y}^T \mathcal{M} \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{b}_d + \frac{\nu}{2}\, \boldsymbol{u}^T \mathcal{M} \boldsymbol{u} \\ \text{s.t.} \quad & \mathcal{A}\boldsymbol{y} = \mathcal{M}\boldsymbol{u} + \boldsymbol{\xi} \\ & y_i = \psi(x_i) \quad \forall\, i \in \{1, ..., n\} \text{ with } \bar{y}_i = \psi(x_i) \text{ and } i \neq \ell \\ & \xi_i = 0 \quad \forall\, i \in \{1, ..., n\} \text{ with } \bar{\xi}_i = 0 \\ & y_\ell \geq \psi(x_\ell) \\ & y_i \geq \psi(x_i) \quad \forall\, i \in \{1, ..., n\} \text{ with } \bar{y}_i > \psi(x_i) \\ & \xi_i \geq 0 \quad \forall\, i \in \{1, ..., n\} \text{ with } \bar{\xi}_i > 0. \end{aligned}\right\} \tag{QP$_\ell$}$$

Because of $\bar{y}_\ell - \psi(x_\ell) = \bar{\xi}_\ell = 0$, the feasible set of (QP$_\ell$) is contained in the one of (MPCC). Since moreover $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{\xi}})$ is feasible for (QP$_\ell$), $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{\xi}})$ is a local minimum of (QP$_\ell$). Due to the linearity of the constraints in (QP$_\ell$), there thus exist Lagrange multipliers $\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\mu}} \in \mathbb{R}^n$ such that the following KKT-system holds true:

$$\mathcal{A}^T \tilde{\boldsymbol{p}} = \mathcal{M}\bar{\boldsymbol{y}} - \boldsymbol{b}_d + \tilde{\boldsymbol{\mu}} \tag{4.8a}$$

$$\nu \bar{\boldsymbol{u}} + \tilde{\boldsymbol{p}} = 0 \tag{4.8b}$$

$$\tilde{\mu}_\ell \leq 0, \quad \tilde{\mu}_\ell \left(\bar{y}_\ell - \psi(x_\ell)\right) = 0 \tag{4.8c}$$

$$\tilde{\mu}_i \leq 0, \quad \tilde{\mu}_i \left(\bar{y}_i - \psi(x_i)\right) = 0 \quad \forall\, i \in \{1, ..., n\} \text{ with } \bar{y}_i > \psi(x_i) \tag{4.8d}$$

$$\tilde{p}_i \geq 0, \quad \tilde{p}_i\, \bar{\xi}_i = 0 \quad \forall\, i \in \{1, ..., n\} \text{ with } \bar{\xi}_i > 0. \tag{4.8e}$$

In view of (4.7b) and (4.8b) we find $\tilde{\boldsymbol{p}} = \boldsymbol{p}$ giving in turn $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$ by (4.7a) and (4.8a). Hence (4.8c) yields a contradiction to the assumption $\mu_\ell > 0$. □

Based on the results of Mignot [1976], Mignot and Puel [1984] established strong stationarity conditions for the case $\psi \equiv 0$. The associated analysis readily transfers to (P) giving the following result:

Theorem 4.4. *Let $\bar u \in L^2(\Omega)$ be a local optimum of* (P) *with associated state $\bar y \in H_0^1(\Omega)$. Then there exist $p \in H_0^1(\Omega)$ and $\mu \in H^{-1}(\Omega)$ so that*

$$\bar y \in K, \quad a(\bar y, v - \bar y) \geq \int_\Omega \bar u\,(v - \bar y)\,dx \quad \forall\, v \in K \tag{4.9a}$$

$$p \in \mathcal{S}_{\bar y}, \quad a(v, p) \leq \int_\Omega (\bar y - y_d) v\,dx \quad \forall\, v \in \mathcal{S}_{\bar y} \tag{4.9b}$$

$$p(x) + \nu\,\bar u(x) = 0 \quad \textit{a.e. in } \Omega, \tag{4.9c}$$

where the set $\mathcal{S}_{\bar y}$ is defined (up to sets of measure zero) by

$$\mathcal{S}_{\bar y} := \{v \in H_0^1(\Omega) :\ v(x) \geq 0 \ \textit{a.e. where } \bar y(x) = \psi(x),\ a(\bar y, v) = \langle \bar u,\, v\rangle\}.$$

Now assume that $\psi \in H_0^1(\Omega)$ such that $\{v - \psi : v \in K\} = \{v \in H_0^1(\Omega) : v \geq 0 \text{ a.e. in } \Omega\}$ becomes a convex cone. Then a slack variable $\xi \in H^{-1}(\Omega)$ can be introduced so that (2.2) is equivalent to the following complementarity system:

$$\begin{aligned} a(\bar y, v) &= \int_\Omega \bar u\, v\,dx + \langle \bar\xi,\, v\rangle \quad \forall\, v \in H_0^1(\Omega) \\ \bar\xi \geq 0, \quad \langle \bar\xi,\, \bar y - \psi\rangle &= 0, \quad \bar y(x) \geq \psi(x) \quad \text{a.e. in } \Omega. \end{aligned} \tag{4.10}$$

The same holds true for arbitrary ψ (satisfying the conditions in Assumption 2.1), provided that $\bar y \in H^2(\Omega)$. In this case, we define $-\Delta \bar y - \bar u =: \bar\xi \in L^2(\Omega)$. The density of $H^1(\Omega)$ in $L^2(\Omega)$ then gives that (2.2) is equivalent to $\langle \bar\xi,\, v - (\bar y - \psi)\rangle \geq 0$ for all $v \in \tilde K$ with $\tilde K = \{v \in L^2(\Omega) : v \geq 0 \text{ a.e. in } \Omega\}$, which is again a convex cone. Thus, (2.2) is again equivalent to (4.10). Hence, in these cases, (4.9) implies the following

Corollary 4.5. *Under the additional assumption that $\psi|_\Gamma = 0$ or $\bar y \in H^2(\Omega)$, every local optimum $\bar u$ of the infinite dimensional problem* (P) *satisfies the following* strong stationarity conditions*:*

There exist $\xi \in H^{-1}(\Omega)$, $p \in H_0^1(\Omega)$, and $\mu \in H^{-1}(\Omega)$ so that

$$a(\bar y, v) = \int_\Omega \bar u\, v\,dx + \langle \bar\xi,\, v\rangle \quad \forall\, v \in H_0^1(\Omega) \tag{4.11a}$$

$$\bar\xi \geq 0, \quad \langle \bar\xi,\, \bar y - \psi\rangle = 0, \quad \bar y(x) \geq \psi(x) \quad \textit{a.e. in } \Omega \tag{4.11b}$$

$$a(v, p) = \int_\Omega (\bar y - y_d) v\,dx + \langle \mu,\, v\rangle \quad \forall\, v \in H_0^1(\Omega) \tag{4.11c}$$

$$p \in \mathcal{S}_{\bar y}, \quad \langle \mu,\, v\rangle \leq 0 \quad \forall\, v \in \mathcal{S}_{\bar y} \tag{4.11d}$$

$$p(x) + \nu\,\bar u(x) = 0 \quad \textit{a.e. in } \Omega, \tag{4.11e}$$

where $\mathcal{S}_{\bar y}$ is defined (up to sets of measure zero) by

$$\mathcal{S}_{\bar y} := \{v \in H_0^1(\Omega) :\ v(x) \geq 0 \ \textit{a.e. where } \bar y(x) = \psi(x),\ \langle \xi,\, v\rangle = 0\}.$$

It is easily seen that (4.2) *just represents the finite dimensional counterpart to* (4.11).

Remark 4.6. *It is well known that $S : L^2(\Omega) \to H^2(\Omega)$ in case of convex and polygonally (polyhedrally) bounded domains, see e.g. Kinderlehrer and Stampacchia [1980] (cf. also Theorem 6.4 below). Thus $\bar y \in H^2(\Omega)$ is always fulfilled if Assumption 2.1(i) holds true. Hence, under Assumption 2.1, the strong stationarity conditions in form of* (4.11) *are valid.*

5 Convergence analysis

For the derivation of a priori error estimates, we require the following assumption on the finite element approximation of the obstacle problem (2.2):

Assumption 5.1 (FE error for the obstacle problem)**.** *There are constants $C > 0$ and $\alpha \geq 1$, independent of h, such that the following estimates hold*

$$\|S_h(u) - S(u)\|_{H^1(\Omega)} \leq C\, h \left(\|u\|_{L^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\right) \tag{5.1a}$$

$$\|S_h(u) - S(u)\|_{L^2(\Omega)} \leq C\, h^\alpha \left(\|u\|_{H^1(\Omega)} + \|\psi\|_{W^{2,\infty}(\Omega)}\right), \tag{5.1b}$$

with S and S_h are the solution operators of (2.2) *and* (3.5)*, respectively.*

Remark 5.2. *It is well known that H^2-regularity of the obstacle problem gives an FE a priori estimate of the form* (5.1a)*, cf. e.g. Falk [1974]. This regularity is for instance guaranteed in case of smooth boundaries or convex and polygonally (polyhedrally) bounded domains, since the obstacle problem is as smooth as the associated Poisson problem, see Kinderlehrer and Stampacchia [1980] and Section 6. In contrast to this, the verification of* (5.1b) *for $\alpha > 1$ is much more delicate (of course $\alpha = 1$ is evident). An adaptation of the well known Nitsche trick from elliptic PDEs to the obstacle problem is only known for very particular cases, cf. Natterer [1976]. A remedy to overcome this difficulty is to employ L^∞-error estimates as done in Section 6 below.*

In the following, let $\bar{u} \in L^2(\Omega)$ be a fixed local optimum of (P).

Assumption 5.3 (Quadratic growth condition)**.** *There are $\varepsilon, \delta > 0$ such that*

$$f(\bar{u}) \leq f(u) - \delta\, \|u - \bar{u}\|^2_{L^2(\Omega)} \qquad \forall\, u \in B_\varepsilon(\bar{u}), \tag{5.2}$$

where $B_\varepsilon(\bar{u})$ denotes the ball of radius ε around $\bar{u}$ in the topology of $L^2(\Omega)$.

Remark 5.4. *In Kunisch and Wachsmuth [2009] it is shown that Assumption 5.3 is satisfied if $\bar{u}$ satisfies the following* second-order sufficient optimality condition*:*

The couple $(\bar{u}, \bar{y}) \in L^2(\Omega), H^1_0(\Omega)$ is strongly stationary*, i.e., there exists $\xi \in H^{-1}(\Omega)$, $p \in H^1_0(\Omega)$, and $\mu \in H^{-1}(\Omega)$ so that* (4.11) *holds true. Moreover, there is a constant $\tau > 0$ such that*

$$p(x) \geq 0 \quad \text{a.e. in } \{x \in \Omega : 0 < \bar{y}(x) < \tau\} \tag{5.3}$$

and, in addition to (4.11d)*, μ satisfies*

$$\mu \leq 0 \quad \Leftrightarrow \quad \langle \mu,\, v\rangle \leq 0 \ \forall\, v \in H^1_0(\Omega) \text{ with } v(x) \geq 0 \text{ a.e. in } \Omega, \tag{5.4}$$

i.e., a sign condition of the multiplier μ does not only hold on the biactive set, but in the whole domain.

Lemma 5.5. *Let* (5.1a) *hold true and suppose that $\bar{u}$ satisfy Assumption 5.3. Then there is a sequence of locally optimal solutions to* (P_h)*, denoted by $\{\bar{u}_h\}_{h>0}$, with $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$ as $h \to 0$.*

Proof. The proof follows standard arguments that can be found e.g. in CasasTroe Casas and Tröltzsch [2002]. Nevertheless, we shortly recall the proof for convenience of the reader. By Assumption 5.3, $\bar{u}$ is an isolated local optimum of (P) and the area of local optimality is the closed ball $B_\varepsilon(\bar{u})$. We introduce the following auxiliary problem:

$$\left.\begin{aligned} \min \quad & f_h(u) = \frac{1}{2}\|S_h(u) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)}, \\ \text{s.t.} \quad & u \in B_\varepsilon(\bar{u}). \end{aligned}\right\} \tag{$\mathrm{P}_h^{(\varepsilon)}$}$$

Since $B_\varepsilon(\bar u)$ is convex and closed, the existence of globally optimal solutions to $(\mathrm{P}_h^{(\varepsilon)})$ follows completely analogously to the proof of Proposition 3.4. For each $h>0$, we consider a fixed global optimum of $(\mathrm{P}_h^{(\varepsilon)})$, denoted by $\bar u_h$. Due to optimality of $\bar u_h$, the following estimate holds for all $u \in B_\varepsilon(\bar u)$:

$$\frac{1}{2}\|S_h(\bar u_h) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|\bar u_h\|^2_{L^2(\Omega)} \le \frac{1}{2}\|S_h(u) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)}. \quad (5.5)$$

Lemma 3.3 yields $\|S_h(u)\|_{H^1_0(\Omega)} \le L\,\|u\|_{H^{-1}(\Omega)}$ with L independent of h. Therefore, by inserting $\bar u$ in (5.5), the sequence $\{\bar u_h\}_{h>0}$ is uniformly bounded in $L^2(\Omega)$ and thus, there exists a weakly convergent subsequence, for simplicity denoted by the same symbol, i.e., $\bar u_h \rightharpoonup \tilde u$ in $L^2(\Omega)$. As we will see in the following the weak limit is unique so that a known argument gives the convergence of the whole sequence, which justifies this notation. By compact embedding, we have $\bar u_h \to \tilde u$ in $H^{-1}(\Omega)$. In view of Lemma 3.3 and Assumption 5.1, we therefore obtain that

$$\|S_h(\bar u_h) - S(\tilde u)\|_{H^1(\Omega)} \le \|S_h(\bar u_h) - S_h(\tilde u)\|_{H^1(\Omega)} + \|S_h(\tilde u) - S(\tilde u)\|_{H^1(\Omega)} \to 0 \quad (5.6)$$

as $h \to 0$. Moreover, thanks to $S_h(\bar u) \to S(\bar u)$ and $f_h(\bar u_h) \le f_h(\bar u)$ by optimality, we obtain

$$f(\bar u) \ge \limsup_{h\to 0} f_h(\bar u_h) \ge \liminf_{h\to 0} f_h(\bar u_h) \ge f(\tilde u) \ge f(\bar u). \quad (5.7)$$

Here, we used (5.6), the weak lower semicontinuity of $\|\,.\,\|^2_{L^2(\Omega)}$, and the local optimality of $\bar u$ in $B_\varepsilon(\bar u)$ for the last two inequalities. As $\bar u$ is an isolated local optimum, there holds $f(\bar u) < f(u)$ for all $u \in B_\varepsilon(\bar u)$ with $u \ne \bar u$. Since $B_\varepsilon(\bar u)$ is convex and closed, hence weakly closed, we have $\tilde u \in B_\varepsilon(\bar u)$ and consequently, $\tilde u = \bar u$. Finally, (5.6) and (5.7), i.e. $f_h(\bar u_h) \to f(\bar u)$, yield $\|\bar u_h\|_{L^2(\Omega)} \to \|\bar u\|_{L^2(\Omega)}$ and together with weak convergence, this gives strong convergence. Since $\bar u$ is an isolated local optimum, the limit is unique so that the whole sequence converges.

It remains to verify that $\bar u_h$ is a local solution of (P_h), if $h>0$ is chosen sufficiently small. For this reason, we choose $u \in L^2(\Omega)$ arbitrarily with $\|u - \bar u_h\|_{L^2(\Omega)} < \frac{\varepsilon}{2}$. The convergence of $\{\bar u_h\}$ to $\bar u$ now yields

$$\|u - \bar u\|_{L^2(\Omega)} \le \|u - \bar u_h\|_{L^2(\Omega)} + \|\bar u_h - \bar u\|_{L^2(\Omega)} < \varepsilon,$$

provided that h is sufficiently small. Therefore, $u \in B_\varepsilon(\bar u)$ is feasible for $(\mathrm{P}_h^{(\varepsilon)})$. Since u was chosen arbitrarily and since $\bar u_h$ is a global solution of $(\mathrm{P}_h^{(\varepsilon)})$, one finally obtains

$$\frac{1}{2}\|S_h(u) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)} \ge \frac{1}{2}\|S_h(\bar u_h) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|\bar u_h\|^2_{L^2(\Omega)},$$

for all u with $\|u - \bar u_h\|_{L^2(\Omega)} < \frac{\varepsilon}{2}$, and thus local optimality of $\bar u_h$ for (P_h). □

Before we can prove a convergence rate for $\|\bar u - \bar u_h\|_{L^2(\Omega)}$, we need the following

Lemma 5.6. *Let $\{\bar u_h\}$ be the sequence of Lemma 5.5, i.e. $\bar u_h \to \bar u$ in $L^2(\Omega)$. Then $\{\bar u_h\}$ is uniformly bounded in $H^1(\Omega)$, i.e. there exists a constant $c>0$, independent of h, such that $\|\bar u_h\|_{H^1(\Omega)} \le c$ for all $h>0$.*

Proof. Since $\bar u_h$ is a local optimum of (P_h) Theorem 4.1 yields the existence of multipliers $\boldsymbol{p}, \boldsymbol{\mu} \in \mathbb{R}^n$ so that the strong stationarity system (4.2) is fulfilled. Testing

(4.2c) with $\boldsymbol{p} \in \mathbb{R}^n$ yields

$$\begin{aligned}\alpha \|p_h\|^2_{H^1(\Omega)} &\le a(p_h, p_h)\\ &= \boldsymbol{p}^T \mathcal{A}\boldsymbol{p}\\ &= \boldsymbol{p}^T \Big(\mathcal{M}\bar{\boldsymbol{y}} - \Big(\int_\Omega y_d \varphi_j dx \Big)_{j=1}^n + \boldsymbol{\mu} \Big)\\ &= \int_\Omega (\bar{y}_h - y_d) p_h \, dx + \boldsymbol{p}^T \boldsymbol{\mu} \le \big(\|\bar{y}_h\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)} \big) \|p_h\|_{H^1(\Omega)},\end{aligned}$$

where we used that $\boldsymbol{p}^T \boldsymbol{\mu} \le 0$ by (4.2d) and (4.2e), cf. also (4.7d). Thus, in view of (4.2f) we arrive at

$$\|\bar{u}_h\|_{H^1(\Omega)} = \frac{1}{\nu} \|p_h\|_{H^1(\Omega)} \le \frac{1}{\nu\alpha} \big(\|\bar{y}_h\|_{L^2(\Omega)} + \|y_d\|_{L^2(\Omega)} \big). \tag{5.8}$$

Analogously to (5.6), Assumption 5.1, the Lipschitz continuity of S_h by Lemma 3.3, and the convergence of $\{\bar{u}_h\}$ give the convergence of the discrete states in $L^2(\Omega)$. Hence $\{\bar{y}_h\}$ is uniformly bounded in $L^2(\Omega)$ giving in turn the claimed uniform boundedness of $\{\bar{u}_h\}$ by (5.8). □

Now we are in the position to prove our main result which reads as follows:

Theorem 5.7. *Let Assumption 5.1 hold and suppose that $\bar{u}$ satisfies Assumption 5.3. Then, there is a sequence $\{\bar{u}_h\}$ of local solutions of (P_h) tending strongly to $\bar{u}$ in $L^2(\Omega)$ as $h \to 0$, and there is a constant $C > 0$, independent of h, so that*

$$\|\bar{u}_h - \bar{u}\|_{L^2(\Omega)} \le C\, h^\beta \quad \text{with} \quad \beta = \min\{1, \alpha/2\}, \tag{5.9}$$

provided that h is sufficiently small.

Proof. By Lemma 5.5, there is a sequence of local solutions to (P_h) converging strongly to $\bar{u}$ giving the first assertion of the theorem. Moreover, as the proof of Lemma 5.5 shows, these local solutions represent global solutions of $(\mathrm{P}_h^{(\varepsilon)})$ for $h > 0$ sufficiently small, and consequently

$$f_h(\bar{u}_h) \le f_h(\bar{u}). \tag{5.10}$$

Furthermore, again for $h > 0$ sufficiently small, we have $\bar{u}_h \in B_\varepsilon(\bar{u})$ such that the quadratic growth condition in Assumption 5.3 implies

$$\begin{aligned}\delta \|\bar{u}_h - \bar{u}\|^2_{L^2(\Omega)} &\le f(\bar{u}_h) - f_h(\bar{u}_h) + f_h(\bar{u}) - f(\bar{u}) + f_h(\bar{u}_h) - f_h(\bar{u})\\ &\le |f(\bar{u}_h) - f_h(\bar{u}_h)| + |f(\bar{u}) - f_h(\bar{u})|,\end{aligned} \tag{5.11}$$

where we used (5.10) for the last estimate. For $|f(\bar{u}_h) - f_h(\bar{u}_h)|$, Assumption 5.1 yields

$$\begin{aligned}&|f(\bar{u}_h) - f_h(\bar{u}_h)|\\ &\quad= \frac{1}{2} \Big| \|S(\bar{u}_h) - y_d\|^2_{L^2(\Omega)} - \|S_h(\bar{u}_h) - S(\bar{u}_h) + S(\bar{u}_h) - y_d\|^2_{L^2(\Omega)} \Big|\\ &\quad= \frac{1}{2} \Big| \|S_h(\bar{u}_h) - S(\bar{u}_h)\|^2_{L^2(\Omega)} + 2(S_h(\bar{u}_h) - S(\bar{u}_h), S(\bar{u}_h) - y_d)_{L^2(\Omega)} \Big|\\ &\quad\le \frac{1}{2} \|S_h(\bar{u}_h) - S(\bar{u}_h)\|^2_{L^2(\Omega)} + \|S_h(\bar{u}_h) - S(\bar{u}_h)\|_{L^2(\Omega)} \|S(\bar{u}_h) - y_d\|_{L^2(\Omega)}\\ &\quad\le C \Big(h^2 \big(\|\bar{u}_h\|_{L^2(\Omega)} + \|\psi\|_{H^2(\Omega)} \big)^2\\ &\qquad\qquad + h^\alpha \big(\|\bar{u}_h\|_{H^1(\Omega)} + \|\psi\|_{W^{2,\infty}(\Omega)} \big) \|S(\bar{u}_h) - y_d\|_{L^2(\Omega)} \Big).\end{aligned}$$

Since $\{\bar{u}_h\}$ is uniformly bounded in $H^1(\Omega)$ by Lemma 5.6, the Lipschitz continuity of $S : L^2(\Omega) \to H^1(\Omega)$ implies the uniform boundedness of $\|S(\bar{u}_h) - y_d\|_{L^2(\Omega)}$ and we end up with

$$|f(\bar{u}_h) - f_h(\bar{u}_h)| \leq C\, h^\alpha.$$

Applying the same argument for $|f(\bar{u}) - f_h(\bar{u})|$ completes the proof. □

For the states, the estimate

$$\begin{aligned}\|S_h(\bar{u}_h) - S(\bar{u})\|_{H^1(\Omega)} &\leq \|S_h(\bar{u}_h) - S_h(\bar{u})\|_{H^1(\Omega)} + \|S_h(\bar{u}) - S(\bar{u})\|_{H^1(\Omega)} \\ &\leq L\, \|\bar{u}_h - \bar{u}\|_{H^{-1}(\Omega)} + C\, h\, \|\bar{u}\|_{L^2(\Omega)}\end{aligned}$$

obtains by (5.1a) and the Lipschitz continuity of S_h. Therefore, Theorem 5.7 implies the following

Corollary 5.8. *Let Assumptions 5.1 and 5.3 be satisfied. Then there is a sequence of optimal states, denoted by $\{\bar{y}_h\}$, that converges strongly in $H^1(\Omega)$ to $\bar{y} = S(\bar{u})$ and satisfies*

$$\|\bar{y}_h - \bar{y}\|_{H^1(\Omega)} \leq C\, h^\beta \quad \textit{with} \quad \beta = \min\{1, \alpha/2\},$$

provided that $h > 0$ is sufficiently small.

6 A specific setting

In the upcoming section, we quantify the value for α in (5.1b) under fairly restrictive assumptions. These assumptions are caused by the fact that we employ L^∞-error estimates for the obstacle problem (2.2) to overcome the difficulties described in Remark 5.2. The most restrictive assumption is given in form of a variant of the discrete maximum principle which we introduce in the sequel. Given an index set $\mathbb{J} \subset \{1, ..., n+m\}$, we define

$$\begin{aligned}\overline{\mathbb{J}} &:= \{i \in \{1, ..., n+m\} :\ \exists\, j \in \mathbb{J} \text{ with } x_i \in \omega_j\} \\ \underline{\mathbb{J}} &:= \{j \in \mathbb{J} : 1 \leq j \leq n\}\,,\end{aligned} \tag{6.1}$$

with $\omega_j = \operatorname{supp} \varphi_j$. Recall in this context that n is the number of nodes of the triangulation in the interior of Ω and $n+m$ is the overall number of nodes. Thus $\underline{\mathbb{J}}$ contains the indices of $\mathbb{J}$ whose the nodes lie in the interior of Ω. Note in addition that $\mathbb{J} \subset \overline{\mathbb{J}}$ by construction. Now we are in the position to formulate our particular form of the discrete maximum principle:

Assumption 6.1 (Discrete maximum principle)**.** *Let $\mathbb{J} \subset \{1, ..., n+m\}$ and $v_h \in V_h^0$ be given. If*

$$a(v_h, \varphi_i) \leq 0 \quad \forall\, i \in \underline{\mathbb{J}}, \tag{6.2}$$

then

$$v_j \leq \max\{0, m\} \quad \textit{for all } j \in \overline{\mathbb{J}},\ \textit{where } m = \max_{i \in \overline{\mathbb{J}} \setminus \underline{\mathbb{J}}} v_i. \tag{6.3}$$

Remark 6.2. *The above assumption is for instance fulfilled if the stiffness matrix defined in* (3.7) *is a weakly diagonally dominant M-matrix, see Appendix B. This is of course a fairly restrictive hypothesis on the discretization.*

Based on the discrete maximum principle two error estimates for the FE discretization of the obstacle problem are proven in Appendix A, based on a technique introduced by Nitsche [1977]. These results require $W^{2,p}$-regularity of the obstacle problem which is addressed in the next theorems for two particular cases:

Theorem 6.3. *Let $\Omega \subset \mathbb{R}^2$ be a convex and polygonally bounded domain, where the largest angle less or equal $\pi/2$. Moreover, let $u \in L^p(\Omega)$ and $\psi \in W^{2,p}(\Omega)$ with $2 \le p < \infty$ be given. Then the solution $y \in H_0^1(\Omega)$ of (2.2) satisfies $y \in W^{2,p}(\Omega)$ and there is a constant C, depending only on Ω and p, such that*

$$\|y\|_{W^{2,p}(\Omega)} \le C \left(\|u\|_{L^p(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)} \right).$$

Proof. Using a regularization approach it is shown in [Kinderlehrer and Stampacchia, 1980, Section IV.2] that the solution operator S of (2.2) is as smooth as the solution operator of the associated Poisson problem, i.e.

$$\begin{aligned} -\Delta \tilde{y} &= u \quad \text{in } \Omega \\ \tilde{y} &= 0 \quad \text{on } \Gamma \end{aligned} \tag{6.4}$$

(provided that the obstacle is sufficiently smooth). To be more precise, in [Kinderlehrer and Stampacchia, 1980, Lemma 2.2 and Thm. 2.3] the authors showed that the solution of (2.2) fulfills

$$\begin{aligned} \|y\|_{W^{2,p}(\Omega)} &\le C_p \left(\|u\|_{L^p(\Omega)} + \| \max(-\Delta\psi - u, 0)\|_{L^p(\Omega)} \right) \\ &\le 2\, C_p \left(\|u\|_{L^p(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)} \right), \end{aligned}$$

provided that the solution of (6.4) satisfies

$$\|\tilde{y}\|_{W^{2,p}(\Omega)} \le C_p \, \|u\|_{L^p(\Omega)}. \tag{6.5}$$

However, under the assumptions on Ω, (6.5) indeed holds for every $p < \infty$ with a constant $C_p > 0$ depending only on p and Ω, as shown by Grisvard [1992]. □

The next theorem addresses improved interior regularity and in this way weakens the restrictive conditions on Ω in the Theorem 6.3.

Theorem 6.4. *Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a convex domain with polygonal (polyhedral) boundary. Moreover, assume $u \in L^p(\Omega)$ and $\psi \in W^{2,p}(\Omega)$ with some $2 \le p < \infty$. Then for every subdomain $\Omega' \subset\subset \Omega$, strictly contained in Ω, we have*

$$\|y\|_{H^2(\Omega)} + \|y\|_{W^{2,p}(\Omega')} \le C \left(\|u\|_{L^p(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)} \right)$$

with some constant $C > 0$ depending only on Ω, Ω', and p.

Proof. The derivation of the H^2-regularity result exactly follows the lines of the proof of Theorem 6.3 having in mind that the required regularity of Γ implies H^2-regularity for the Poisson problem, see e.g. Grisvard [1992]. Concerning the higher interior regularity, we first prove the result for the Poisson problem. Once the interior regularity for the Poisson problem is established, regularization again yields the assertion for the obstacle problem. Since $\operatorname{dist}(\Omega', \partial\Omega) > 0$, there is a subset Ω'' with smooth boundary such that $\Omega' \subset\subset \Omega'' \subset\subset \Omega$. Now, let $\zeta \in C_0^\infty(\Omega)$ be given with $\zeta \in [0,1]$ in Ω, $\zeta \equiv 1$ in Ω', and $\zeta \equiv 0$ in $\Omega \setminus \Omega''$. For the existence of such a function we refer to [Gajewski et al., 1974, Remark 1.3]. Moreover, denote as above by $\tilde{y} \in H_0^1(\Omega)$ the solution of the Poisson problem (6.4). Then integration by parts on Ω'' yields for every $v \in C_0^\infty(\Omega'')$

$$\begin{aligned} \int_{\Omega''} \nabla(\zeta\, \tilde{y}) \cdot \nabla v \, dx &= \int_{\Omega''} \zeta\, \nabla\tilde{y} \cdot \nabla v \, dx + \int_{\Omega''} \tilde{y}\, \nabla\zeta \cdot \nabla v \, dx \\ &= \int_{\Omega''} \zeta\, u\, v \, dx - \int_{\Omega''} v\, \nabla\zeta \cdot \nabla\tilde{y} + \int_{\Omega''} \tilde{y}\, \nabla\zeta \cdot \nabla v \, dx \\ &= \int_{\Omega''} \zeta\, u\, v \, dx + \int_{\Omega''} \tilde{y}\, v\, \triangle\zeta \, dx, \end{aligned}$$

so that $\eta := \zeta\tilde{y} \in H_0^1(\Omega'')$ solves the following Poisson problem on Ω''

$$\begin{aligned} -\triangle\eta &= \zeta\, u + \tilde{y}\,\triangle\zeta \quad \text{in } \Omega'' \\ \eta &= 0 \qquad\qquad\quad \text{on } \partial\Omega''. \end{aligned} \tag{6.6}$$

Hence the classical Calderon-Zygmund theory for smooth domains yields

$$\begin{aligned} \|\eta\|_{W^{2,p}(\Omega'')} &\le c\,\|\zeta\, u + \tilde{y}\,\triangle\zeta\|_{L^p(\Omega'')} \\ &\le c\,\big(\|\zeta\|_{L^\infty(\Omega'')}\,\|u\|_{L^p(\Omega)} + \|\tilde{y}\|_{L^p(\Omega'')}\,\|\zeta\|_{C^2(\Omega'')}\big) \\ &\le c\,\big(\|u\|_{L^p(\Omega)} + \|\tilde{y}\|_{H^2(\Omega)}\big), \end{aligned}$$

where we used the continuous embedding $H^2(\Omega) \hookrightarrow L^p(\Omega)$ for $d = 2, 3$. Again in view of Grisvard [1992], one has $\|\tilde{y}\|_{H^2(\Omega)} \le c\,\|u\|_{L^2(\Omega)}$, which together with $\zeta \equiv 1$ in Ω' implies

$$\|\tilde{y}\|_{W^{2,p}(\Omega')} \le c\,\|u\|_{L^p(\Omega)}.$$

Now, one can again apply the regularization argument of Kinderlehrer and Stampacchia [1980] to obtain the result for the obstacle problem. □

Together with the continuous embeddings

$$H^1(\Omega) \hookrightarrow \begin{cases} L^p(\Omega) \text{ for all } p < \infty, & \text{if } d = 2 \\ L^6(\Omega), & \text{if } d = 3, \end{cases}$$

the following two propositions are an immediate consequence of Theorems A.6 and 6.3 and Theorems A.8 and 6.4, respectively.

Proposition 6.5. *Suppose that $\Omega \subset \mathbb{R}^2$ is polygonally bounded with a largest angle less or equal $\pi/2$. Suppose moreover that all meshes from $\{\mathcal{T}_h\}_{h>0}$ are such that the discrete maximum principle is satisfied for every $h > 0$. Then, for all $p < \infty$, the estimate*

$$\|S_h(u) - S(u)\|_{L^2(\Omega)} \le C\,h^{2-2/p}\,|\log h|\,\big(\|u\|_{H^1(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)}\big) \tag{6.7}$$

holds with a constant C, depending on p and Ω but not on h.

Proposition 6.6. *Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ be a convex domain with polygonal (polyhedral) boundary. Assume further that the bound ψ satisfies $\psi \in W^{2,p}(\Omega)$ with $\psi(x) < 0$ for all $x \in \Gamma$. If all meshes from $\{\mathcal{T}_h\}_{h>0}$ are such that the discrete maximum principle is satisfied for every $h > 0$, then there is a constant C, depending on p and Ω but not on h, such that*

$$\|S_h(u) - S(u)\|_{L^2(\Omega)} \le C\,h^{2-d/p}\,|\log h|\,\big(\|u\|_{H^1(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)}\big) \tag{6.8}$$

with $p < \infty$ if $d = 2$ and $p = 6$ if $d = 3$.

Note that the restrictive conditions on the largest angle of Γ in Proposition 6.5 are avoided in Proposition 6.6 by using $\psi|_\Gamma < 0$ and interior regularity, cf. also the proof of Theorem A.8. If we choose $\varepsilon < d/(2p)$, then (6.7) and (6.8) imply

$$\|S_h(u) - S(u)\|_{L^2(\Omega)} \le C\,h^{2-2\varepsilon}\,\big(\|u\|_{H^1(\Omega)} + \|\psi\|_{W^{2,\infty}(\Omega)}\big), \tag{6.9}$$

with $C > 0$, depending on ε but not on h. Thus (5.1b) in Assumption 5.1 is satisfied with $\alpha = 2 - 2\varepsilon$. With this results at hand, Theorem 5.7 and Corollary 5.8 finally yield the following

Corollary 6.7. *Let $\Omega \subset \mathbb{R}^2$ be a convex domain with polygonal boundary. Furthermore, let $\psi \in W^{2,\infty}(\Omega)$ be given and assume that at least one of the following conditions is fulfilled*

(i) $\psi(x) < 0$ for all $x \in \Gamma$
(ii) the largest angle in Γ is less or equal $\pi/2$.

Assume in addition that all meshes are such that the associated stiffness matrices satisfy the discrete maximum principle in Assumption 6.1. Then, for every local solution $\bar{u}$ of (P) *satisfying the quadratic growth condition in Assumption 5.3, there exists a sequence of local solutions $\{\bar{u}_h\}$ to* (P_h) *that converges to $\bar{u}$ and satisfies*

$$\|\bar{u}_h - \bar{u}\|_{L^2(\Omega)} + \|\bar{y}_h - \bar{y}\|_{H^1(\Omega)} \leq C\, h^{1-\varepsilon} \quad \text{for all } \varepsilon > 0$$

provided that h is sufficiently small with a constant C depending on ε but not on h. Herein, $\bar{y} = S(\bar{u})$ and $\bar{y}_h = S_h(\bar{u}_h)$ denote the associated states.

Remark 6.8. *Due to the generic regularity of the optimal control which is in $H^1(\Omega)$ by* (4.11e), (4.11c), *and the generic regularity of $\mu \in H^{-1}(\Omega)$, the above result can be seen to be nearly optimal as the estimate is up to an ε as good as the interpolation error.*

Remark 6.9. *Note that the assertion of Corollary 6.7 is limited to the two dimensional case. In view of Proposition 6.6, the above theory would only yield a convergence rate of* $3/4$ *in the three-dimensional case. Of course the situation would change if L^2-error estimates known for elliptic equations would also hold for the obstacle problem* (2.2). *However, due to a lack of regularity of the dual problem, the Nitsche trick does not seem to be applicable here.*

In Section 7 two examples are presented that satisfy the hypotheses of Corollary 6.7.

7 Numerical examples

In the following, we test the presented error analysis with two different examples. Both examples are constructed such that

- $\Omega = (0,1)^2$
- $\{\mathcal{T}_h\}_{h>0}$ only consists of Friedrich-Keller triangulations.

Thus the boundary Γ satisfies condition *(ii)* in Corollary 6.7. Moreover, it is well known that the stiffness matrix for the Poisson problem in case of a Friedrich-Keller triangulation of the unit square is irreducibly diagonally dominant and hence a weakly diagonally dominant M-matrix. Therefore, according to Lemma B.1 and Remark B.2, the discrete maximum principle in Assumption 6.1 is satisfied. Thus the conditions of Corollary 6.7 are fulfilled and we expect a theoretical convergence rate of $1-\varepsilon$ for all $\varepsilon > 0$, provided that the respective exact solutions satisfy the quadratic growth condition from Assumption 5.3.

To enable the construction of exact solutions, we consider a slightly modified optimal control problem given by

$$\left.\begin{aligned} \min \quad & \frac{1}{2}\,\|y - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\,\|u - u_d\|^2_{L^2(\Omega)} \\ \text{s.t.} \quad & a(y, v-y) \geq \langle u, v-y\rangle, \quad \forall\, v \in K \text{ and } y \in K, \end{aligned}\right\} \tag{P*}$$

where we add a data $u_d \in L^2(\Omega)$. It is straight forward to see that the above error analysis is not affected by this modification provided that u_d is sufficiently smooth. Note that the strong stationarity conditions for (P*) are equivalent to the ones of (P), i.e. the system in (4.11), except (4.11e), which has to be replaced by

$$p(x) + \nu\big(\bar{u}(x) - u_d(x)\big) = 0 \quad \text{f.a.a. } x \in \Omega. \tag{7.1}$$

The discretized optimal control problems are solved numerically by adding the biquadratic penalization term $\frac{1}{4\gamma}\int_\Omega \max\{-\gamma y, 0\}^4\,dx$ to the objective of the lower level problem (L_h). The arising penalized problems are solved with Newton's

method, which is applicable due to the smoothness of $\max^4$. The penalty parameter γ is increased until the iterates do not change any more. In both examples the Tikhonov parameter ν is set to $\nu = 1$, and for the bound we choose $\psi \equiv 0$.

7.1. Example 1. To construct a locally optimal solution, we first define a solution to the strong stationarity system (4.11a)–(4.11d) and (7.1) and afterwards verify the second-order sufficient conditions (5.3) and (5.4). We start by defining the state and the adjoint state by

$$\bar{y}(x_1, x_2) = \begin{cases} y_1(x_1) \cdot y_2(x_2), & \text{in } (0, 0.5) \times (0, 0.8) \\ 0, & \text{otherwise} \end{cases}$$

with

$$y_1(x_1) = -4096\, x_1^6 + 6144\, x_1^5 - 3072\, x_1^4 + 512\, x_1^3$$
$$y_2(x_2) = -244.140625\, x_2^6 + 585.9375\, x_2^5 - 468.75\, x_2^4 + 125\, x_2^3$$

(cf. Hintermüller and Kopacka [2009]) and

$$p(x) = \begin{cases} p_1(Q^T x), & \text{in } \Omega_1 \\ 0, & \text{otherwise,} \end{cases}$$

where $Q \in \mathbb{R}^{2\times 2}$ is a rotation matrix given by

$$Q = \begin{pmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{pmatrix} \tag{7.2}$$

and p_1 is defined by $p_1(x_1, x_2) = (-200(x_1 - 0.8)^2 + 0.5)(-200(x_2 - 0.9)^2 + 0.5)$. Moreover, Ω_1 is a square with midpoint $(0.8, 0.9)$ and edge length 0.1, which is rotated by Q around its midpoint such that its boundary is not captured by the mesh. Note that $\bar{y} \in C^2(\bar{\Omega})$, whereas p has a kink located on $\partial\Omega_1$ so that $p \notin H^2(\Omega)$. Thus the distributional Laplace of p contains a line measure, which we identify with the multiplier μ, i.e.

$$\langle \mu,\, v\rangle_{H^{-1}(\Omega),H^1(\Omega)} = \int_{\partial\Omega_1} n_1 \cdot \left(\nabla \bar{p}|_{\Omega_1} - \nabla \bar{p}|_{\Omega\setminus\Omega_1}\right) v\, ds, \quad v \in H^1(\Omega).$$

Herein, n_1 denotes the outward unit normal on $\partial\Omega_1$. Observe that $\mu \in H^{-1}(\Omega)$ such that μ provides the generic regularity according to the strong stationarity system (4.11). Using the regular parts of $\triangle p$ and (4.11c), we define the desired state by

$$y_d(x) = \begin{cases} \bar{y}(x) + \triangle p_1(Q^T x), & \text{in } \Omega_1 \\ \bar{y}(x), & \text{otherwise.} \end{cases}$$

Note that y_d possesses a discontinuity that is not covered by the meshes. To evaluate the corresponding integrals involving y_d with sufficient accuracy, we therefore introduce additional meshes, which are locally refined around the discontinuity of y_d. Let us point out that these meshes are only used for the numerical integration of the data y_d and not for the rest of the computation so that our error analysis is not affected.

The slack variable $\bar{\xi}$ is set to

$$\bar{\xi}(x_1, x_2) = \begin{cases} y_1(x_1 - 0.5) \cdot y_2(x_2), & \text{in } (0.5, 1) \times (0, 0.8) \\ 0, & \text{otherwise} \end{cases} \tag{7.3}$$

with y_1 and y_2 as defined above. Note that the complementarity system in (4.11b) is satisfied with this setting. Observe moreover that the biactive set is given by

$$\{x \in \Omega : \bar{y}(x) = \bar{\xi}(x) = 0\} = [0 : 1] \times [0.8 : 1]$$

and thus has measure greater zero. Hence the control-to-state mapping is not Gâteaux-differentiable in this local optimum (cf. Kunisch and Wachsmuth [2009]). Furthermore, the optimal control is obtained by (4.11a), i.e. $\bar{u} = -\triangle\bar{y} - \bar{\xi}$. To fulfill the gradient equation (7.1), we finally set $u_d = \bar{u} + \frac{1}{\nu}\, p$. With these settings, one easily verifies that (4.11d) is satisfied so that the strong stationarity conditions are fulfilled. In addition, it is easily seen that the second-order sufficient conditions (5.3) and (5.4) hold true. Hence $(\bar{y}, \bar{u})$ indeed represent a locally optimal solution that satisfies the quadratic growth condition. Consequently, Corollary 6.7 predicts a convergence rate of $1 - \varepsilon$ for this example.

Figs. 7.1-7.4 show the numerical solution for $h/\sqrt{2} = 0.005$. We point out that the discrete multiplier μ_h shown in Fig. 7.4 indeed appears to be an approximation of the line measure located on $\partial\Omega_1$.

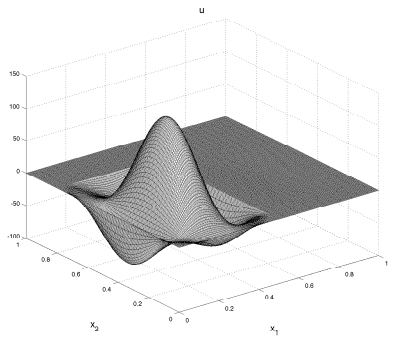

Figure 7.1: Example 1: optimal control $\bar{u}_h$ for $h/\sqrt{2} = 0.005$.

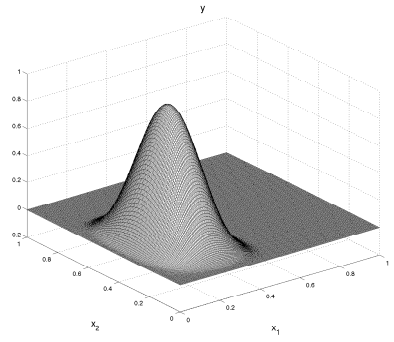

Figure 7.2: Example 1: optimal state $\bar{y}_h$ for $h/\sqrt{2} = 0.005$.

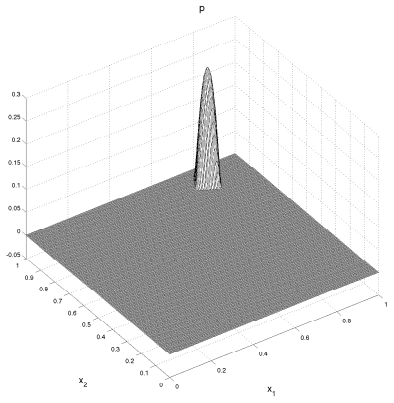

Figure 7.3: Example 1: adjoint state p_h for $h/\sqrt{2} = 0.005$.

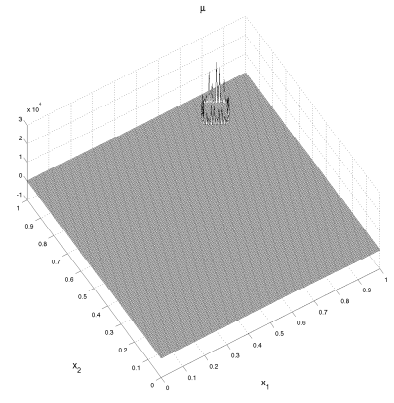

Figure 7.4: Example 1: multiplier μ_h for $h/\sqrt{2} = 0.005$.

To verify our theoretical order of convergence, we compute the experimental order of convergence. In case of the control u, it is given by

$$EOC_2(u) := \frac{\log(e_2(u,h_1)) - \log(e_2(u,h_2))}{\log(h_1) - \log(h_2)},$$

where h_1 and h_2 denote two consecutive mesh sizes and e_2 refers to the approximation of relative error in the L^2-norm, i.e.

$$e_2(u,h) := \frac{\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)}}{\|\bar{u}\|_{L^2(\Omega)}}$$

In addition $e_{1,2}$ denotes the approximative relative error in the H^1-norm, i.e.

$$e_{1,2}(y,h) := \frac{\|\bar{y} - \bar{y}_h\|_{H^1(\Omega)}}{\|\bar{y}\|_{H^1(\Omega)}}$$

and $EOC_{1,2}(y)$ is computed with $e_{1,2}(y,h)$ instead of $e_2(u,h)$. Table 7.1 presents the experimental orders of convergence and the relative errors for this example. We observe that $EOC_{1,2}(y)$ is approximately one, while $EOC_2(u)$ is slightly larger

$h/\sqrt{2}$	$e_2(u)$	$e_{1,2}(y)$	$EOC_2(u)$	$EOC_{1,2}(y)$
1/50	1.9436e-01	8.4567e-02	-	-
1/100	9.2354e-02	4.2389e-02	1.0735	0.9964
1/150	5.6347e-02	2.8273e-02	1.2186	0.9985
1/200	4.0105e-02	2.1208e-02	1.1819	0.9994
1/250	2.9831e-02	1.6968e-02	1.3263	0.9997
1/300	2.3991e-02	1.4140e-02	1.1948	0.9998
1/350	1.9629e-02	1.2121e-02	1.3021	0.9998
1/400	1.6622e-02	1.0606e-02	1.2453	0.9999

Table 7.1: Relative errors and experimental order of convergence in the first example.

than one. Nevertheless, the example already indicates that the convergence rates for optimal control of the obstacle problem are significantly lower compared to optimal control of the Poisson problem. This is also confirmed by the next test case whose numerical results are in agreement with the theoretical predictions.

7.2. Example 2. The second exact solution is constructed similarly to the first one in Section 7.1. While control $\bar{u}$, state $\bar{y}$, and slack variable $\bar{\xi}$ are the same as in the first example, we now choose

$$p(x) = \begin{cases} p_1(Q^T x) \cdot p_2(Q^T x), & \text{in } \Omega_2 \\ q_1(Q^T x) \cdot p_2(Q^T x), & \text{in } \Omega_3 \\ p_1(Q^T x) \cdot q_2(Q^T x), & \text{in } \Omega_4 \\ q_1(Q^T x) \cdot q_2(Q^T x), & \text{in } \Omega_5 \\ 0, & \text{otherwise} \end{cases}$$

with

$$\begin{aligned} p_1(x_1,x_2) &= -0.5[0.1(60(x_1-0.9)+1)^3 + 0.2(60(x_1-0.9))^2] + 0.5 \\ q_1(x_1,x_2) &= -0.5[0.1(-60(x_1-0.9)+1)^3 + 0.2(-60(x_1-0.9))^2] + 0.5 \\ p_2(x_1,x_2) &= -0.5[0.1(60(x_2-0.9)+1)^3 + 0.2(60(x_2-0.9))^2] + 0.5 \\ q_1(x_1,x_2) &= -0.5[0.1(-60(x_2-0.9)+1)^3 + 0.2(-60(x_2-0.9))^2] + 0.5 \end{aligned}$$

for the adjoint state. Herein, $\bigcup_{i=2,...,5} \Omega_i$ is the square with midpoint $(0.9, 0.9)$ and edge length 0.01, which is rotated by the rotation matrix Q, given in (7.2) around its midpoint. Each Ω_i, $i = 2, ..., 5$, then represents a quarter of this domain. Again we have $p \notin H^2(\Omega)$ so that the distributional Laplace of p is no proper function. As in the first case the irregular parts of $\triangle p$ are identified with the multiplier μ which is thus again a line measure located on $\bar{\Omega}_i \cap \bar{\Omega}_j$, $i, j = 2, ..., 5$, $i \neq j$. The desired state y_d is then again defined based on the regular parts of $\triangle p$ and (4.11c). Moreover, as before, the desired control u_d is chosen such that (7.1) is fulfilled. It is easily checked that the strong stationarity conditions as well as the second-order sufficient conditions are also satisfied in this example. Thus we again have a locally optimal solution fulfilling the quadratic growth condition and hence, Corollary 6.7 applies.

In Fig. 7.5-7.6 the numerical solution for the adjoint state and for the multiplier can be seen for $h/\sqrt{2} = 0.005$. Again the discrete multiplier appears to be an

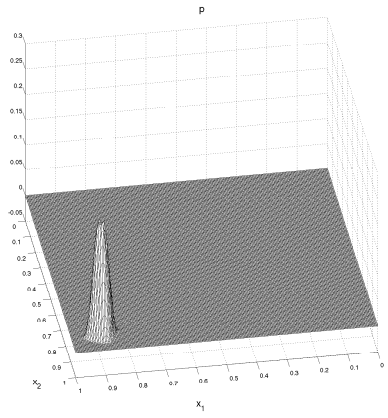

Figure 7.5: Example 2: adjoint state p_h for $h/\sqrt{2} = 0.005$.

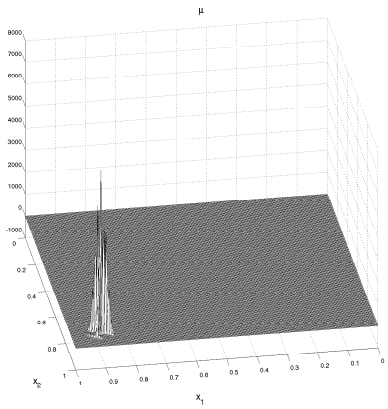

Figure 7.6: Example 2: multiplier μ_h for $h/\sqrt{2} = 0.005$.

approximation of the multiplier of the infinite dimensional problem, which is only a line measure as depicted above. The experimental order of convergence can be found in Tab. 7.2. In this case both, $EOC_2(u)$ and $EOC_{1,2}(y)$, amount approximately one and hence coincide with the theoretical results. Thus the numerical findings in this example agree with the theoretical predictions.

A L^∞-error estimates for the obstacle problem

In this section, we establish the L^2- and L^∞-error estimates for the obstacle problem that are used in Section 6. The proof is along the lines of a technique introduced by Nitsche [1977] which is based on the discrete maximum principle. The main differences to the analysis presented in Nitsche [1977] is that we do not require $W^{2,\infty}$-regularity and that we allow the discrete solution to hit the obstacle also at isolated points, edges, and facets, respectivley. Moreover, we use the technique to verify an L^2-estimate that employs higher interior regularity of the solutions.

We consider an obstacle problem of the form (2.2), i.e.

$$a(y, v - y) \geq \langle g,\, v - y \rangle \quad \forall v \in K,\ y \in K \tag{A.1}$$

$h/\sqrt{2}$	$e_2(u)$	$e_{1,2}(y)$	$EOC_2(u)$	$EOC_{1,2}(y)$
1/50	2.1912e-01	8.4567e-02	-	-
1/100	1.0578e-01	4.2390e-02	1.0506	0.9964
1/150	6.8792e-02	2.8273e-02	1.0613	0.9988
1/200	5.1199e-02	2.1208e-02	1.0267	0.9994
1/250	4.0729e-02	1.6968e-02	1.0253	0.9997
1/300	3.3822e-02	1.4140e-02	1.0192	0.9998
1/350	2.8879e-02	1.2121e-02	1.0249	0.9998
1/400	2.5253e-02	1.0606e-02	1.0047	0.9999

Table 7.2: Relative errors and experimental order of convergence in the second example.

with $K = \{v \in H_0^1(\Omega) : v(x) \geq \psi(x) \quad \text{a.e. in } \Omega\}$ and given right hand side $g \in L^2(\Omega)$. Throughout this section, we suppose the following

Assumption A.1. *The inhomogeneity, the solution of* (A.1)*, and the bound* ψ *satisfy* $g \in L^2(\Omega)$*,* $y \in H^2(\Omega)$*, and* $\psi \in H^2(\Omega)$*.*

As mentioned in Section 4, by standard arguments, one can reformulate (A.1) by a complementarity system provided e.g. that $y \in H^2(\Omega)$ which is ensured by Assumption A.1. Thus there is a slack variable $\xi \in L^2(\Omega)$ such that

$$-\Delta y = g + \xi \text{ in } \Omega, \quad y = 0 \text{ on } \Gamma \tag{A.2a}$$

$$\xi(x) \geq 0, \quad y(x) \geq \psi(x), \quad \xi(x)\,(y(x) - \psi(x)) = 0 \quad \text{a.e. in } \Omega. \tag{A.2b}$$

Recall now the discretization from Section 3. If $\mathcal{T}_h$ is a given mesh, we denote the standard linear FE space by V_h and set $V_h^0 = V_h \cap H_0^1(\Omega)$. Moreover, the nodes of $\mathcal{T}_h$ in the interior of the domain are denoted by $\{x_1, ..., x_n\}$, while $\{x_{n+1}, ..., x_{n+m}\}$ are the nodes on $\Gamma = \partial\Omega$. It will be convenient to define the index sets

$$\mathbb{G} := \{1, \dots, n\} \quad \text{and} \quad \mathbb{B} := \{n+1, ..., n+m\}.$$

Let us again denote by $\{\varphi_1, ..., \varphi_{n+m}\}$ the nodal basis of V_h. As already done in Section 3, we associate to every function $v_h \in V_h$ the coefficient vector $\boldsymbol{v} \in \mathbb{R}^{n+m}$ w.r.t. the nodal basis, i.e. $v_h(x) = \sum_{i=1}^{n+m} v_i \varphi_i(x)$ (and analogously for V_h^0). The discrete counterpart to (A.2) reads: find $y_h \in V_h^0$, $\boldsymbol{\xi} \in \mathbb{R}^n$ so that

$$a(y_h, \varphi_i) = \int_\Omega g\,\varphi_i dx + \xi_i \quad \forall\, i \in \mathbb{G} \tag{A.3a}$$

$$\boldsymbol{\xi} \geq 0, \quad \boldsymbol{y} \geq \boldsymbol{\psi}, \quad \boldsymbol{\xi}^T(\boldsymbol{y} - \boldsymbol{\psi}) = 0, \tag{A.3b}$$

cf. (4.1). Herein, as before, we set $\boldsymbol{\psi} = \big(\psi(x_i)\big)_{i=1}^n$. By Lemma 3.2 there exists a unique solution $(y_h, \boldsymbol{\xi}) \in V_h^0 \times \mathbb{R}^n$ of (A.3).

For the derivation of the error estimate, we split the error via the triangle inequality into

$$\|y - y_h\|_{L^2(\Omega)} \leq \|y - Y_h\|_{L^2(\Omega)} + \|Y_h - y_h\|_{L^2(\Omega)}, \tag{A.4}$$

where $Y_h \in V_h^0$ denotes the Ritz-projection, i.e. the unique solution to

$$a(Y_h, v_h) = a(y, v_h) \quad \forall\, v_h \in V_h^0.$$

Concerning the Ritz-projection, the standard Aubin-Nitsche trick for elliptic PDEs implies

$$\|y - Y_h\|_{L^2(\Omega)} \leq c\,h^2\,\|y\|_{H^2(\Omega)} \tag{A.5}$$

The estimation of the second error $\|Y_h - y_h\|_{L^2(\Omega)}$ will be done in several steps. To be more precise, we estimate the differences

$$Y_i - y_i \quad \text{and} \quad y_i - Y_i$$

$i = 1, ..., n + m$, separately, where, as above, y_i and Y_i denote the components of the coefficient vectors to y_h and Y_h. Let us first consider $Y_i - y_i$. We start by defining the active set

$$\mathcal{A} := \{x \in \Omega : y(x) = \psi(x)\} \tag{A.6}$$

and the following index sets

$$\begin{aligned} \mathbb{A}_1 &:= \{i \in \mathbb{G} : \omega_i \cap \mathcal{A} \neq \emptyset\}, \quad & \mathbb{I}_1 &:= \mathbb{G} \setminus \mathbb{A}_1 \\ \mathbb{A}_2 &:= \{i \in \mathbb{G} : y_i = \psi_i(x_i)\}, \quad & \mathbb{I}_2 &:= \mathbb{G} \setminus \mathbb{A}_2, \end{aligned} \tag{A.7}$$

where $\omega_i = \operatorname{supp}(\varphi_i)$, $i = 1, ..., n + m$ denotes the support of φ_i. Furthermore, we set

$$\mathcal{A}_{h,1} := \bigcup_{i \in \mathbb{A}_1} \omega_i \quad \text{and} \quad \mathcal{A}_{h,2} := \bigcup_{i \in \mathbb{A}_2} \omega_i. \tag{A.8}$$

Lemma A.2. *Suppose that the meshes $\mathcal{T}_h$ are such that the discrete maximum principle in Assumption 6.1 is fulfilled. Then, under Assumption A.1, there holds for all $i = 1, ..., n + m$*

$$Y_i - y_i \leq \|Y_h - y\|_{L^\infty(\mathcal{A}_{h,1})} + \max_{i \in \mathbb{A}_1} |y(x_i) - \psi(x_i)|.$$

Proof. Our proof is devided into three parts showing the assertion on the different sets $\mathbb{A}_1$, $\mathbb{I}_1$, and $\mathbb{B}$.

1st case, $i \in \mathbb{B}$:
Since $y_h \in V_h^0$ and $Y_h \in V_h^0$ it follows immediately $y_i = Y_i = 0$ and thus $Y_i - y_i = 0$.

2nd case, $i \in \mathbb{A}_1$:
Because of $y_i \geq \psi(x_i)$, it holds

$$Y_i - y_i = Y_i - y(x_i) + y(x_i) - y_i \leq \|Y_h - y\|_{L^\infty(\mathcal{A}_{h,1})} + y(x_i) - \psi(x_i), \tag{A.9}$$

which already implies the assertion on $\mathbb{A}_1$.

3rd case, $i \in \mathbb{I}_1$:
Here, we employ the discrete maximum principle from Assumption 6.1 with $\mathbb{J} = \mathbb{I}_1$. Therefore we have to verify condition (6.2), i.e. $a(Y_h - y_h, \varphi_i) \leq 0$ for all $i \in \mathbb{I}_1$. Due to $i \notin \mathbb{A}_1$ we have $y(x) > \psi(x)$ in ω_i giving in turn $\xi(x) = 0$ a.e. in ω_i by (A.2b). Thus (A.2a) implies $-\Delta y = g$ a.e. in ω_i which yields together with (A.3a)

$$\begin{aligned} 0 = \int_{\omega_i} (-\Delta y - g)\varphi_i \, dx &= -\int_\Omega \Delta y \, \varphi_i dx - \int_\Omega g \, \varphi_i \, dx \\ &= a(y, \varphi_i) - a(y_h, \varphi_i) + \xi_i \geq a(Y_h - y_h, \varphi_i), \end{aligned}$$

where we used $\boldsymbol{\xi} \geq 0$ and the definition of the Ritz-projection in the last step. Note that no boundary integral appears in the integration by parts in the above estimate since $\varphi_i \in V_h^0$ because of $i \in \mathbb{I}_1 \subset \mathbb{G}$, i.e. x_i is an interior node. Now, we can apply the discrete maximum principle (6.3) to obtain

$$Y_i - y_i \leq \max\big\{0, \max_{i \in \bar{\mathbb{I}}_1 \setminus \mathbb{I}_1} (Y_i - y_i)\big\}. \tag{A.10}$$

(Note that $\mathbb{I}_1 = \underline{\mathbb{I}}_1$.) Furthermore, in view of $\bar{\mathbb{I}}_1 \setminus \mathbb{I}_1 \subset \{\mathbb{G} \cup \mathbb{B}\} \setminus \mathbb{I}_1 = \mathbb{B} \cup \mathbb{A}_1$, we find

$$\max_{i \in \bar{\mathbb{I}}_1 \setminus \mathbb{I}_1} (Y_i - y_i) \leq \max_{i \in \mathbb{B} \cup \mathbb{A}_1} (Y_i - y_i) \leq \max_{i \in \mathbb{B}} (Y_i - y_i) + \max_{i \in \mathbb{A}_1} (Y_i - y_i).$$

Thus, together with (A.10), (A.9), and $Y_i = y_i = 0$ for $i \in \mathbb{B}$, this yields the assertion on $\mathbb{I}_1$.

Since $\mathbb{A}_1 \cup \mathbb{I}_1 \cup \mathbb{B} = \{1, ..., n+m\}$ the result is proven. □

Lemma A.3. *Suppose that the Assumptions A.1 and 6.1 are fulfilled. Then for every $i \in 1, ..., n+m$ the estimate*

$$y_i - Y_i \leq \|Y_h - y\|_{L^\infty(\mathcal{A}_{h,2})}$$

is valid.

Proof. The proof is almost the same as the one of Lemma A.2. Nevertheless we sketch the arguments for convenience of the reader. On $\mathbb{B}$ we again have $y_i = Y_i = 0$ so there is nothing to show here. If $i \in \mathbb{A}_2$, then

$$y_i - Y_i = \psi(x_i) - Y_i = \psi(x_i) - y(x_i) + y(x_i) - Y_i \leq \|y - Y_h\|_{L^\infty(\mathcal{A}_{h,2})} \tag{A.11}$$

because of $\psi(x_i) - y(x_i) \leq 0$.

It remains to study the case $i \in \mathbb{I}_2$. To apply again the discrete maximum principle (6.3), the condition $a(y_h - Y_h, \varphi_i) \leq 0$ is to be shown. In view of $i \in \mathbb{I}_2$, i.e. $y_i \neq \psi(x_i)$ and thus $\xi_i = 0$ by complementarity, (A.3a) gives

$$a(y_h, \varphi_i) = \int_\Omega f\varphi_i dx + \xi_i = \int_\Omega f\varphi_i dx \tag{A.12}$$

Furthermore, we find for the Ritz projection

$$a(Y_h, \varphi_i) = a(y, \varphi_i) = \int_\Omega f\varphi_i dx + \int_\Omega \underbrace{\xi}_{\geq 0} \underbrace{\varphi_i}_{\geq 0} dx \geq \int_\Omega f\varphi_i dx. \tag{A.13}$$

Combining (A.12) and (A.13) yields $a(y_h - Y_h, \varphi_i) \leq 0$ for all $i \in \mathbb{I}_2$ so that the discrete maximum principle (6.3) is applicable. Thus we obtain

$$y_i - Y_i \leq \max\left\{0, \max_{i \in \bar{\mathbb{I}}_2 \setminus \underline{\mathbb{I}}_2} (y_i - Y_i)\right\}$$

Moreover, by construction it holds $\underline{\mathbb{I}}_2 = \mathbb{I}_2$ and $\bar{\mathbb{I}}_2 \setminus \mathbb{I}_2 \subset \mathbb{I} \cup \mathbb{B} \setminus \mathbb{I}_2 = \mathbb{B} \cup \mathbb{A}_2$. Hence we arrive at

$$\max_{i \in \bar{\mathbb{I}}_2 \setminus \underline{\mathbb{I}}_2} (y_i - Y_i) \leq \max_{i \in \mathbb{B}} (y_i - Y_i) + \max_{i \in \mathbb{A}_2} (y_i - Y_i). \tag{A.14}$$

Together with the result on $\mathbb{A}_2$ this implies the estimate on $\mathbb{I}_2$. The assertion then again follows from $\mathbb{A}_2 \cup \mathbb{I}_2 \cup \mathbb{B} = \{1, ..., n+m\}$. □

The next results is an immediate consequence of the above lemmata.

Proposition A.4. *Let the Assumptions A.1 and 6.1 be fulfilled. Then*

$$\|y_h - Y_h\|_{L^\infty(\Omega)} \leq \|y - Y_h\|_{L^\infty(\mathcal{A}_{h,1} \cup \mathcal{A}_{h,2})} + \max_{i \in \mathbb{A}_1} |y(x_i) - \psi(x_i)|.$$

holds true.

Proof. For every $x \in \Omega$, we find

$$|y_h(x) - Y_h(x)| = |\sum_{i=1}^{n+m} (y_i - Y_i)\varphi_i(x)| \leq \max_{i \in \mathbb{G} \cup \mathbb{B}} |y_i - Y_i| \sum_{i=1}^{n+m} \varphi_i(x).$$

Due to $\sum_{i=1}^{n+m} \varphi_i(x) = 1$ for all $x \in \Omega$, Lemma A.2 and A.3 imply the assertion. □

In view of the above result, we need an estimate for $y - \psi$ on $\mathbb{A}_1$, which is provided by the following

Lemma A.5. *Let $i \in \mathbb{A}_1$ be arbitrary and assume that $y, \psi \in C^{k,\alpha}(\mathcal{A}_{h,1})$ with $k = 0, 1$ and $\mathcal{A}_{h,1}$ as defined in* (A.8)*. Then*

$$0 \le y(x_i) - \psi(x_i) \le c\, h^{k+\alpha} \left(\|y\|_{C^{k,\alpha}(\mathcal{A}_{h,1})} + \|\psi\|_{C^{k,\alpha}(\mathcal{A}_{h,1})} \right) \tag{A.15}$$

with and a constant $c > 0$ independent of i and h.

Proof. Since $i \in \mathbb{A}_1$, there is an $\tilde{x} \in \omega_i \cap \mathcal{A}$. Hence $|x_i - \tilde{x}| \le c\, h$ and $y(\tilde{x}) = \psi(\tilde{x})$ so that

$$\begin{aligned} y(x_i) - \psi(x_i) &= y(x_i) - y(\tilde{x}) + \psi(\tilde{x}) - \psi(x_i) \\ &\le \left(\|y\|_{C^{0,\alpha}(\mathcal{A}_{h,1})} + \|\psi\|_{C^{0,\alpha}(\mathcal{A}_{h,1})} \right) |x_i - \tilde{x}|^\alpha, \end{aligned}$$

which gives the assertion for $k = 0$. Note that $y(x) \ge \psi(x)$ holds due to $y \in K$. To prove the statement for $k = 1$, observe that, by construction, $\tilde{x}$ solves the following minimization problem:

$$(y - \psi)(\tilde{x}) = \min_{x \in \bar{\Omega}} \big((y - \psi)(x) \big). \tag{A.16}$$

Furthermore, by definition of $\mathcal{A}$, we have $\tilde{x} \notin \partial\Omega$ and hence the necessary optimality condition for (A.16) is given by

$$\nabla (y - \psi)(\tilde{x}) = 0$$

(notice that $y, \psi \in C^1(\Omega)$ by Assumption A.1). Thus the mean value theorem gives

$$\begin{aligned} y(x_i) - \psi(x_i) &= (y - \psi)(\tilde{x}) + \int_0^1 \nabla (y - \psi)(\tilde{x} + \delta(x_i - \tilde{x}))^T (x_i - \tilde{x}) d\delta \\ &= \int_0^1 \left[\nabla (y - \psi)(\tilde{x} + \delta(x_i - \tilde{x})) - \nabla (y - \psi)(\tilde{x}) \right]^T (x_i - \tilde{x}) d\delta \\ &\le c\, h^{1+\alpha} \left(\|y\|_{C^{1,\alpha}(\mathcal{A}_{h,1})} + \|\psi\|_{C^{1,\alpha}(\mathcal{A}_{h,1})} \right), \end{aligned}$$

which is just the assertion for $k = 1$. □

Theorem A.6. *Suppose that the discrete maximum principle from Assumption 6.1 is satisfied. Assume further that, in addition to Assumption A.1, the solution y of* (3.6) *and the bound ψ fulfill $y, \psi \in W^{2,p}(\Omega)$ with some $p > d$, where d denotes the spatial dimension. Then there holds*

$$\|y - y_h\|_{L^2(\Omega)} \le C\, h^{2-d/p} \, |\log h| \left(\|y\|_{W^{2,p}(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)} \right)$$

with some constant $C > 0$ independent of h and p.

Proof. In view of (A.4), (A.5), and Proposition A.4, we have

$$\begin{aligned} \|y - y_h\|_{L^2(\Omega)} \le c \Big(& h^2 \|y\|_{H^2(\Omega)} + \|y - Y_h\|_{L^\infty(\mathcal{A}_{h,1} \cup \mathcal{A}_{h,2})} \\ & + \max_{i \in \mathbb{A}_1} |y(x_i) - \psi(x_i)| \Big). \end{aligned} \tag{A.17}$$

For the second addend, well known L^∞-error estimates for the Poisson equation, cf. e.g. Schatz [1998] and the references therein, imply

$$\|y - Y_h\|_{L^\infty(\Omega)} \le c\, h \, |\log h| \, \|y - I_h y\|_{W^{1,\infty}(\Omega)} \le c\, h^{2-d/p} \, |\log h| \, \|y\|_{W^{2,p}(\Omega)}, \tag{A.18}$$

where $I_h : C(\bar{\Omega}) \to V_h$ denotes the Lagrange interpolation operator. The last inequality in the above estimate follows from standard interpolation error estimates, see e.g. Brenner and Scott [1994]. For the last addend in (A.17), we find by means of Lemma A.5

$$\max_{i \in \mathbb{A}_1} |y(x_i) - \psi(x_i)| \le c\, h^{2-d/p} \left(\|y\|_{W^{2,p}(\Omega)} + \|\psi\|_{W^{2,p}(\Omega)} \right), \tag{A.19}$$

where we used the Sobolev embedding $W^{2,p}(\Omega) \hookrightarrow C^{1,\alpha}(\Omega)$ for $\alpha = 1 - d/p$. Inserting (A.18) and (A.19) in (A.17) gives the desired estimate. □

Remark A.7. *The above proof immediately shows that the same order of convergence is also obtained for the L^∞-error $\|y - y_h\|_{L^\infty(\Omega)}$. However, since the error analysis of the optimal control problem only requires an estimate for the L^2-error, the statement of Theorem A.6 is sufficient here.*

The next Theorem exploits the fact that the L^∞-estimates are only needed on $\mathcal{A}_{h,1} \cup \mathcal{A}_{h,2}$.

Theorem A.8. *Assume that the discrete maximum principle from Assumption 6.1 holds. Let us further suppose that $\psi|_\Gamma < 0$. Then there exist $h_0 > 0$ and $\Omega' \subset\subset \Omega$ so that*

$$\mathcal{A}_{h,1} \cup \mathcal{A}_{h,2} \subset\subset \Omega'$$

for all $0 < h \le h_0$. Moreover, if in addition to Assumption A.1 one has $y, \psi \in W^{2,p}(\Omega')$ with $p > d$, then the estimate

$$\|y - y_h\|_{L^2(\Omega)} \le C\, h^{2-d/p}\, |\log h| \left(\|y\|_{W^{2,p}(\Omega')} + \|\psi\|_{W^{2,p}(\Omega')} + \|y\|_{H^2(\Omega)} \right)$$

is valid for all $0 < h \le h_0$.

Proof. By Assumption A.1 and Sobolev embeddings we know that $y, \psi \in C^{0,\alpha}(\Omega)$ with $\alpha = 2 - d/2 \ge 1/2$ for $d = 2, 3$. Hence, ψ is continuous up to the boundary Γ such that $\delta := \max_{x\in\Gamma} \psi(x)$ exists and is negative by assumption. Since $y|_\Gamma = 0$, we find

$$\begin{aligned} y(x) - \psi(x) &\ge y(\xi) - \psi(\xi) + y(x) - y(\xi) + \psi(\xi) - \psi(x) \\ &\ge |\delta| - \big(\|y\|_{C^{0,1/2}(\Omega)} + \|\psi\|_{C^{0,1/2}(\Omega)} \big) |x - \xi|^{1/2} \\ &\ge |\delta| - c \left(\|y\|_{H^2(\Omega)} + \|\psi\|_{H^2(\Omega)} \right) |x - \xi|^{1/2}. \end{aligned}$$

for every $x \in \Omega$ and every $\xi \in \Gamma$. Therefore, if

$$\operatorname{dist}(x, \Gamma) \le \frac{1}{4} \left(\frac{|\delta|}{c \left(\|y\|_{H^2(\Omega)} + \|\psi\|_{H^2(\Omega)} \right)} \right)^2 =: d_\Gamma > 0,$$

then

$$y(x) \ge \psi(x) + |\delta|/2. \tag{A.20}$$

This implies that $\operatorname{dist}(\mathcal{A}, \Gamma) \ge d_\Gamma$ by the definition of $\mathcal{A}$ in (A.6). Consequently, if $h_0 > 0$ is chosen sufficiently small, then (A.7) and (A.8) yield $\operatorname{dist}(\mathcal{A}_{h,1}, \Gamma) \ge \mathrm{d}_\Gamma/2$ for all $h < h_0$, since $\operatorname{diam}(\omega_i) \le 2h$ for all $i \in \{1, ..., n + m\}$. To verify the first assertion of the theorem, it remains to show that $\operatorname{dist}(\mathcal{A}_{h,2}, \Gamma) \ge d_\Gamma/2$ for sufficiently small h. To this end, observe that H^2-regularity of y together with standard interpolation and inverse estimates imply for the Ritz-projection

$$\|y - Y_h\|_{L^\infty(\Omega)} \le c\, h^{2-d/2}\, \|y\|_{H^2(\Omega)},$$

cf. Brenner and Scott [1994]. Then arguing similarly to the proof of Theorem A.6, one obtains

$$\begin{aligned} \|y - y_h\|_{L^\infty(\Omega)} &\le \|y - Y_h\|_{L^\infty(\Omega)} + \|y_h - Y_h\|_{L^\infty(\Omega)} \\ &\le c\, h^{2-d/2} \left(\|y\|_{H^2(\Omega)} + \|\psi\|_{H^2(\Omega)} \right), \end{aligned} \tag{A.21}$$

where we again employed Proposition A.4 and Lemma A.5 together with the Sobolev embedding $H^2 \hookrightarrow C^{0,\alpha}$, $\alpha = 2 - d/2$. Now assume that there is an index $\ell \in \mathbb{A}_2$ so that $\operatorname{dist}(x_\ell, \Gamma) < \mathrm{d}_\Gamma$. Then, in view of (A.20) and (A.21), we obtain

$$y_\ell - \psi(x_\ell) = y_h(x_\ell) - y(x_\ell) + y(x_\ell) - \psi(x_\ell) \ge \frac{|\delta|}{2} - \|y - y_h\|_{L^\infty(\Omega)} \ge \frac{|\delta|}{4},$$

for all $0 < h \le h_0$ provided that $h_0 > 0$ is chosen sufficiently small. This contradicts $\ell \in \mathbb{A}_2$ by definition of $\mathbb{A}_2$ in (A.7). Therefore $\operatorname{dist}(x_i, \Gamma) \ge d_\Gamma$ for all $i \in \mathbb{A}_2$ and, as above, $\operatorname{diam}(\omega_i) \le 2h$ for all $i \in \{1, ..., n + m\}$ implies $\operatorname{dist}(\mathcal{A}_{h,2}, \Gamma) \ge d_\Gamma/2$.

Thus there exists a set $\Omega' \subset\subset \Omega$ that strictly contains $\mathcal{A}_{h,1}$ and $\mathcal{A}_{h,2}$ as claimed. Without loss of generality we may assume that Ω' is a mesh domain, i.e. it is a union of elements of $\mathcal{T}_h$.

If further $y \in W^{2,p}(\Omega')$ with some $p > d$, then interior maximum norm estimates according to [Schatz and Wahlbin, 1977, Thm. 5.1] yield the existence of a constant $c > 0$ such that

$$\begin{aligned} \|y - Y_h\|_{L^\infty(\mathcal{A}_{h_1} \cup \mathcal{A}_{h_2})} &\le c \left(|\log h| \, \|y - I_h y\|_{L^\infty(\Omega')} + \|y - y_h\|_{L^2(\Omega)} \right) \\ &\le c \left(h^{2-d/p} \, |\log h| \, \|y\|_{W^{2,p}(\Omega')} + h^2 \, \|y\|_{H^2(\Omega)} \right), \end{aligned} \tag{A.22}$$

for all $0 < h \le h_0$, provided that $h_0 > 0$ sufficiently small. Herein we used (A.5) and standard interpolation error estimates for the last inequality. Moreover, Lemma A.5 and Sobolev embeddings give

$$\begin{aligned} \max_{i \in \mathbb{A}_1} |y(x_i) - \psi(x_i)| &\le c \, h^{2-d/p} \left(\|y\|_{C^{1,1-d/p}(\mathcal{A}_{h,1})} + \|\psi\|_{C^{1,1-d/p}(\mathcal{A}_{h,1})} \right) \\ &\le c \, h^{2-d/p} \left(\|y\|_{W^{2,p}(\Omega')} + \|\psi\|_{W^{2,p}(\Omega')} \right). \end{aligned} \tag{A.23}$$

Finally inserting (A.22) and (A.23) in (A.17) implies the second assertion of the theorem. □

B A discrete maximum principle

This section is devoted to the derivation of a necessary condition for the discrete maximum principle from Assumption 6.1. The arguments are fairly standard. Let us first recall our particular form of the discrete maximum principle. Given an index set $\mathbb{J} \subset \mathbb{G} \cup \mathbb{B}$ we set

$$\overline{\mathbb{J}} := \{ i \in \mathbb{G} \cup \mathbb{B} : \ \exists \, j \in \mathbb{J} \text{ with } x_i \in \omega_j \} \quad \text{and} \quad \underline{\mathbb{J}} := \{ j \in \mathbb{J} : j \in \mathbb{G} \}.$$

We say that the discrete maximum principle is fulfilled if, for all $v_h \in V_h^0$,

$$a(v_h, \varphi_i) \le 0 \quad \forall \, i \in \underline{\mathbb{J}} \tag{B.1}$$

implies

$$v_j \le \max\{0, m\} \quad \forall \, j \in \overline{\mathbb{J}}, \text{ with } m = \max_{i \in \overline{\mathbb{J}} \setminus \underline{\mathbb{J}}} v_i. \tag{B.2}$$

Lemma B.1. *Assume that the triangulation $\mathcal{T}_h$ is such that*

- *all secondary diagonal entries of the stiffness matrix are non-positive, i.e.*

$$a(\varphi_i, \varphi_j) \le 0 \quad \forall \, i, j \in \mathbb{G} \cup \mathbb{B}, \ i \ne j, \tag{B.3}$$

- *the stiffness matrix is weakly diagonally dominant, i.e.*

$$a(\varphi_i, \varphi_i) \ge - \sum_{j \ne i} a(\varphi_i, \varphi_j) \quad \forall \, i \in \mathbb{G}, \tag{B.4}$$

- *all principal minors of the stiffness matrix are non-zero, i.e.*

$$a(\varphi_i, \varphi_j)_{i,j \in \underline{\mathbb{J}}} \text{ is invertible for all } \mathbb{J} \subset \mathbb{G} \cup \mathbb{B}. \tag{B.5}$$

Then the discrete maximum principle holds, i.e. (B.1) *implies* (B.2) *for all index sets $\mathbb{J} \subset \mathbb{G} \cup \mathbb{B}$.*

Proof. Suppose that $v_h \in V_h^0$ satisfies (B.1) so that $b_i := a(v_h, \varphi_i) \le 0$ for all $i \in \underline{\mathbb{J}}$. Now let $\epsilon > 0$ be given and define a vector $\boldsymbol{v}^\epsilon \in \mathbb{R}^{|\underline{\mathbb{J}}|}$ by

$$\sum_{j \in \underline{\mathbb{J}}} a(\varphi_i, \varphi_j) v_j^\epsilon = b_i - \epsilon - \sum_{j \in \overline{\mathbb{J}} \setminus \underline{\mathbb{J}}} a(\varphi_i, \varphi_j) v_j, \quad i \in \underline{\mathbb{J}}$$

Note that $\boldsymbol{v}^\epsilon$ is well defined due to condition (B.5). Moreover, since $\sum_{j\in\overline{\mathbb{J}}} a(\varphi_i,\varphi_j)v_j = a(v_h,\varphi_i) = b_i$, we have by construction $v_j^\epsilon \to v_j$ for $\epsilon \searrow 0$ for all $j \in \underline{\mathbb{J}}$. Next we set

$$\tilde{v}_j^\epsilon := \begin{cases} v_j^\epsilon, & j \in \underline{\mathbb{J}} \\ v_j, & j \in \overline{\mathbb{J}} \setminus \underline{\mathbb{J}}, \end{cases}$$

and consequently

$$\tilde{v}_j^\epsilon \to v_j \quad \text{for } \epsilon \searrow 0 \ \forall\, j \in \overline{\mathbb{J}}. \tag{B.6}$$

Note that $\tilde{\boldsymbol{v}}^\epsilon \in \mathbb{R}^{|\overline{\mathbb{J}}|}$ satisfies

$$\sum_{j\in\overline{\mathbb{J}}} a(\varphi_i,\varphi_j)\tilde{v}_j^\epsilon = \sum_{j\in\underline{\mathbb{J}}} a(\varphi_i,\varphi_j)v_j^\epsilon + \sum_{j\in\overline{\mathbb{J}}\setminus\underline{\mathbb{J}}} a(\varphi_i,\varphi_j)v_j = b_i - \epsilon \le -\epsilon \tag{B.7}$$

for all $i \in \underline{\mathbb{J}}$, where we used (B.1) for the last inequality. We first show the assertion for $\tilde{\boldsymbol{v}}^\epsilon$ and (B.6) then implies the statement of the lemma. To this end, assume that there is an index $k \in \underline{\mathbb{J}}$ with

$$\tilde{v}_k^\epsilon = \max_{i\in\overline{\mathbb{J}}} \tilde{v}_i^\epsilon \quad \text{and} \quad \tilde{v}_k^\epsilon > 0. \tag{B.8}$$

Then it follows that

$$\begin{aligned} a(\varphi_k,\varphi_k)\tilde{v}_k^\epsilon &= \sum_{j\in\overline{\mathbb{J}}} a(\varphi_k,\varphi_j)\tilde{v}_j^\epsilon - \sum_{j\in\overline{\mathbb{J}},\, j\neq k} a(\varphi_k,\varphi_j)\tilde{v}_j^\epsilon \\ &< \Big(- \sum_{j\in\overline{\mathbb{J}},\, j\neq k} a(\varphi_k,\varphi_j)\Big)\tilde{v}_k^\epsilon \le a(\varphi_k,\varphi_k)\tilde{v}_k^\epsilon, \end{aligned}$$

where we used (B.7), (B.3), and (B.8) for the first inequality and (B.4) and (B.8) for the second estimate. Hence, we have a contradiction so that such an index k cannot exist. Due to $\tilde{v}_j^\epsilon = v_j$ for $j \in \overline{\mathbb{J}} \setminus \underline{\mathbb{J}}$, we therefore arrive at

$$\tilde{v}_j^\epsilon \le \max\big\{0, \max_{i\in\overline{\mathbb{J}}\setminus\underline{\mathbb{J}}} v_i\big\} \quad \forall\, j \in \overline{\mathbb{J}},$$

Thanks to (B.6), taking the limit $\epsilon \searrow 0$ in the above inequality finally gives the assertion. □

Remark B.2. *It is well known that the conditions* (B.3)–(B.5) *are for instance fulfilled, if the stiffness matrix* $a(\varphi_i,\varphi_j)_{i,i\in\mathbb{G}}$ *is a weakly diagonally dominant M-matrix, see e.g. Fiedler [1986].*

References

N. Arada, E. Casas, and F. Tröltzsch. Error estimates for a semilinear elliptic optimal control problem. *Computational Optimization and Applications*, 23:201–229, 2002.

S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods.* Springer, New York, 1994.

E. Casas and F. Tröltzsch. Error estimates for the finite-element approximation of a semilinear elliptic control problem. *Control and Cybernetics*, 31:695–712, 2002.

F. H. Clarke. A new approach to lagrange multipliers. *Mathematics of Operations Research*, 1(2):165–174, 1976.

K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state constrained elliptic control problem. *SIAM Journal on Numerical Analysis*, 45:1937–1953, 2007.

R. Falk. Approximation of a class of optimal control problems with order of convergence estimates. *Journal of Mathematical Analysis and Applications*, 44:28–47, 1973.

R. S. Falk. Error estimates for the approximation of a class of variational inequalities. *Mathematics of Computation*, 28(128):963–971, 1974.

M. Fiedler. *Special Matrices and Their Applications in Numerical Mathematics*. Martinus Nijhoff, Dordrecht, 1986.

H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*. Akademie-Verlag, Berlin, 1974.

T. Geveci. On the approximation of the solution of an optimal control problem governed by an elliptic equation. *R.A.I.R.O. Analyse numérique/Numericl Analysis*, 13(4):313–328, 1979.

P. Grisvard. *Singularities in Boundary Value Problems*. Masson, Paris, 1992.

M. Hintermüller. Inverse coefficient problems for variational inequalities: Optimality conditions and numerical realization. *ESAIM Mathematical Modelling and Numerical Analysis*, 35(1):129–152, 2001.

M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM Journal on Optimization*, 20:868–902, 2009.

M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Computational Optimization and Applications*, 20: 45–61, 2005.

K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Applied Mathematics and Optimization*, 41:343–364, 2000.

Jiří Jarušek, Jiří Outrata, and Jana Stará. On optimality conditions in control of elliptic variational inequalities. *Set-Valued Analysis*, 2010. To appear.

D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. Academic Press, New York, 1980.

K. Kunisch and D. Wachsmuth. Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities. Technical Report 2009-04, Johann Radon Institute for Computational and Applied Mathematics (RICAM), 2009. Submitted.

C. Meyer. Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints. *Control and Cybernetics*, 37(1):51–85, 2008.

F. Mignot. Contrôle dans les inéquations variationelles elliptiques. *Journal of Functional Analysis*, 22(2):130–185, 1976.

F. Mignot and J.-P. Puel. Optimal control in some variational inequalities. *SIAM Journal on Control and Optimization*, 22(3):466–476, 1984.

F. Natterer. Optimale L^2-Konvergenz Finiter Elemente bei Variationsungleichungen. *Bonner Mathematische Schriften*, 89:1–12, 1976.

J. Nitsche. L^∞-convergence of finite element approximations. In *Mathematical Aspects of Finite Element Methods*, volume 606 of *Lecture Notes in Mathematics*, pages 261–274, Berlin, 1977. Springer.

A. Rösch. Error estimates for parabolic optimal control problems with control constraints. *Zeitschrift für Analysis und ihre Anwendungen*, 23(2):353–376, 2004.

A. H. Schatz. Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids. i: Global estimates. *Mathematics of Computation*, 67:877–899, 1998.

A. H. Schatz and L. B. Wahlbin. Interior maximum norm estimates for finite element methods. *Mathematics of Computation*, 31(138):414–442, 1977.

Holger Scheel and Stefan Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.

Technische Universität Dortmund, Fakultät für Mathematik, Lehrstuhl X, Vogelpothsweg 87, 44227 Dortmund, Germany

E-mail address: Christian.Meyer@math.tu-dortmund.de

Technische Universität Dortmund, Fakultät für Mathematik, Lehrstuhl X, Vogelpothsweg 87, 44227 Dortmund, Germany

E-mail address: Oliver.Thoma@math.tu-dortmund.de