ERROR ESTIMATES FOR THE FEM APPROXIMATION OF STATE–CONSTRAINED ELLIPTIC OPTIMAL CONTROL PROBLEMS WITH SPARSITY IN A FINITE–DIMENSIONAL CONTROL SPACE*

PEDRO MERINO† AND ALEXANDER NENJER †

Abstract. In this work, we derive an a–priori error estimate of order $h^2|\log(h)|$ for the finite element approximation of an sparse optimal control problem governed by an elliptic equation, which is controlled in a finite dimensional space. Furthermore, box–constrains on the control are considered and finitely many pointwise state–constrains are imposed in specific points in the domain. With this choice for the control space, the achieved order of approximation for the optimal control is optimal, in the sense that the order of the error for the optimal control is of the same order of the approximation for the state equation.

Key words. Optimal control, elliptic partial differential equations, error estimates, finite element method, sparsity.

AMS subject classifications. 80M10, 41A25, 49J20, 49K20.

1. Introduction.

It is well known that sparse optimal controls have very attractive properties for their practical implementation c.f. [17]. When the decision variables are functions, sparse optimal control problems produce solutions with small supports on the domain, which can be viewed as a localized action of the optimal control c.f. [32]. However, when the control is a vector, sparsity takes the form of many null components in the solution vector. This is interesting since a simpler solution is not only easier to understand and implement, but it can also be used as an strategy to detect and select the most relevant quantities taking part in the control process, aiming to save scarce resources.

In this paper, we consider finite dimensional sparse optimal control problems and its approximation by the standard finite element method. We study the estimation of the convergence order (with respect to the size of the mesh h) of a state constrained linear–quadratic optimal control problem. Here, the control space is finite dimensional; therefore, in order to promote sparsity in the solution an 1–norm penalization term is included in the cost functional. The choice of finite–dimensional control space as well as pointwise state–constraints are motivated by technological requirements since, in practice, we often choose from a set of finite quantities in order to control certain processes in real applications. For instance, this is the case of digital technology, where finitely many quantities are used rather than functions.

In turn, pointwise constraints on the state are also important for applications where it is critical that the associated state remains constrained on specific points on the domain. Several applications in medicine have these kind of situations. For example, in the treatment of cancer with hypertermia or cryosurgery thechniques, state–constraints prevent damage on healthy tissue and organs exposed to extreme temperatures, see for example [22] and [31].

*Submitted to the editors DATE.

Funding: This research has been partially supported by project PIJ-15-26 granted by Escuela Politécnica Nacional. Also, we acknowledge partial support of MATHAmSud project SOCDE "Sparse Optimal Control of Differential Equations".

†Department of Mathematics, *Ladrón de Guevara E11-253, Escuela Politécnica Nacional (Quito–Ecuador)*, Research center on mathematical modeling MODEMAT, Quito–Ecuador (pedro.merino@epn.edu.ec, hernan.alex_nenjer@hotmail.com).

As mentioned in [7, 8], there are mainly two ways for inducing sparsity on the solutions: by considering a sparsity inducting term or, alternatively, by choosing an appropriate control space; for instance, the space of regular Borel measures. In recent years a considerable amount of literature has been published on sparse optimal control problems and many authors have studied several questions arising in its analysis, numerical approximation and numerical methods. See for example [7, 8, 9, 32] for a complete review of the theoretical results for sparse optimal control of elliptic and parabolic equations.

The question of error estimates by the finite element method for sparse optimal control problems is rather a recent subject of research. In the recent contributions [6] and [7], the linear and semilinear elliptic cases have been considered and an error of order h was derived. The case where the control space consists of the regular Borel measures was studied in [5]. There, an the error of order h^κ was deduced, where $\kappa = 1$ and $\kappa = 1/2$ for two and three dimensions respectively.

From the finite–dimensional perspective, error estimates for optimal control problems, where the control space is finite dimensional, have also been considered in [25]. However, there sparsity terms were not considered.

In this article, we derive an error of convergence of order $h^2|\log(h)|$, which is optimal in the sense that that the order of the error for the optimal control is of the same order of the approximation of the state equation. This is not highly surprising since the control variables are vectors and they do not need to be approximated. The only source of error comes from the approximation of the state equation. However, unlike the case of functional controls, the derivation of the error estimates follows a different analysis. Even though the results obtained by [7] can be applied to the finite–dimensional case, they do not take into account the finite–dimensional nature of the control space. In this regard, our results improve the existing estimations by taking advance of this particularity.

We consider an analogous problem as in [24] but the novelty consists in including the non–differentiable penalization by means of the 1–norm. The lack of differentiability involves new numerical challenges in the approximation and its numerical solution. At our best knowledge, this is the first contribution of sparse optimal control in a finite–dimensional control space. Our strategy for deriving error estimates consists in circumventing the non differentiability of the optimization problem by formulating an equivalent differentiable optimal control problem, obtained by splitting the control variable in its positive and negative parts. This splitting procedure is known in mathematical programming in finite dimensions but, due to the duplication of the number of variables, it is not useful for numerical algorithms. On the other hand, in context of optimal control, where a state equation has to be approximated by finite elements, this trick helps to analyze the numerical approximation. Indeed, having reformulated the original problem we proceed as in [25], by transforming the original optimal control problem into an optimization problem in finite dimensions for which we carry out a similar stability analysis and, subsequently to establish optimal error estimates. Our stability analysis developed for the equivalent differentiable problem differs from [25] in that we exploit convexity, and the so called *strong regularity* is weakened by the concept of *regularity*, allowing us to only require a Slater type condition as constraint qualification.

In addition, under further considerations we show that these results also cover the case when the Tikhonov regularization is not present, which an additional contribution of this paper.

Our paper is organized as follows: in the first section we introduce the optimal

control problem and the associated terminology and definitions. Then, in Section 2 we derive optimality conditions characterizing sparse solutions. In the third section we describe the numerical approximation by the finite element method of the optimal control problem and state the main result of this paper in Theorem 3.3. The fourth section focuses on an equivalent differentiable formulation of our problem which is the bridge to apply the stability analysis. The fifth section is devoted to the derivation of the error estimate and the sixth and final section, numerical tests confirming the theory are presented.

2. The sparse optimal control problem and optimality conditions. Let $\Omega \subset \mathbb{R}^2$ be an open bounded convex set with Lipschitz boundary Γ and $M \in \mathbb{N}\backslash\{0\}$. We are interested in the finite–element approximation of the following linear–quadratic sparse optimal control problem in $\mathbb{R}^M$:

$$(\mathbf{P_M}) \quad \begin{cases} \min\limits_{(y,u)} J(y,u) = \dfrac{1}{2}\displaystyle\int_\Omega (y-y_d)^2\,dx + \dfrac{\alpha}{2}||u||_2^2 + \beta||u||_1 \\ \text{subject to:} \\ Ay(x) = \displaystyle\sum_{i=1}^{M} u_i e_i(x), \quad \text{in } \Omega, \\ y(x) = 0, \qquad\qquad \text{on } \Gamma, \\ y(x_j) \le b_j, \qquad\qquad \forall j = 1,\dots,\ell, \\ u \in \mathcal{U}_{ad}, \end{cases}$$

In this problem, we assume that $y_d \in L^2(\Omega)$ is a desired state, $\alpha > 0$, $\beta > 0$, as well as fixed basis functions e_i, for $i = 1, 2, ..., M$, belong to $C^{0,\gamma}(\Omega)$, for some $0 < \gamma < 1$. The admissible set of controls is defined by

$$\mathcal{U}_{ad} = \{u \in \mathbb{R}^M : u_a \le u \le u_b\},$$

with $u_a, u_b \in \mathbb{R}^M$ and $u_a < u_b$. For vectors, the symbols $\le$ and $<$ are understood component–wise.

The differential equation is based upon the uniform ellipticity and symmetry of the differential operator A, defined by:

$$Ay(x) := -\sum_{i,j=1}^{2} \partial_j(a_{ij}(x)\partial_i y(x)) + a_0(x)y(x),$$

with coefficients $a_{ij} \in C^{1+\delta}(\Omega)$, $0 < \delta < 1$, and $a_0 \in C^{0,\gamma}(\Omega)$ such that $a_0(x) \ge 0$ in Ω.

The constraints on the state, whose bound is given by the vector $b \in \mathbb{R}^\ell$, are imposed in $\ell \in \mathbb{N}$ a priori selected points $x_i \in \Omega_0$, $i = 1,\dots,\ell$, where Ω_0 is a subdomain of Ω.

Let us summarize some fundamental results on elliptic partial differential equations which will be applied to the state and adjoint equations associated to our optimal control problem $(\mathbf{P_M})$. In what follows, we will use the symbols $(\cdot,\cdot)$ and $\|\cdot\|$ to denote the inner product and its corresponding norm in $L^2(\Omega)$.

Let f be a given function in $L^2(\Omega)$. We say that $y \in H_0^1(\Omega)$ is a weak solution of the equation

$$\begin{cases} Ay(x) = f(x), & \text{in } \Omega, \\ y(x) = 0, & \text{on } \Gamma, \end{cases} \tag{2.1}$$

if y fulfills the variational equation:

$$\sum_{i,j=1}^{2}(a_{ij}\partial_i y, \partial_j \phi) + (a_0 y, \phi) = (f, \phi),$$

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

PROPOSITION 2.1. *For each function $f \in L^2(\Omega)$, there exists a unique weak solution $y \in H_0^1(\Omega) \cap H^2(\Omega)$ of (2.1) and the mapping $G : L^2(\Omega) \to H_0^1(\Omega)$ defined by*

$$G(f) = y, \quad \forall f \in L^2(\Omega),$$

is linear and continuous. Moreover, if $f \in C^{0,\gamma}(\Omega)$, then $y \in C^{2,\gamma}(\Omega)$.

Proof. For the existence and uniqueness we refer to [16] [Theorem 3, p. 301] and for the regularity result see [18],[19], since Ω is convex. ∎

PROPOSITION 2.2. *For each $u \in \mathbb{R}^M$, the state equation has a unique weak solution $y_u \in H_0^1(\Omega) \cap H^2(\Omega) \cap C^{2,\gamma}(\Omega)$, that is y_u satisfies*

$$\sum_{i,j=1}^{2}(a_{ij}\partial_i y_u, \partial_j \phi) + (a_0 y_u, \phi) = (\sum_{i=1}^{M} u_i e_i, \phi), \tag{2.2}$$

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

For the analysis of the existence of an optimal solution, we reformulate the optimal control problem ($\mathbf{P_M}$) as an optimization problem in terms of u only. Let y_i be the weak solution of (2.1) associated to the right–hand side $f = e_i$. By utilizing the superposition principle, we define the control-to-state mapping S, which associates to each vector $u \in \mathbb{R}^M$ the corresponding state y_u defined by

$$\begin{array}{rccl} S: & \mathbb{R}^M & \to & C(\bar{\Omega}) \\ & u & \mapsto & y_u = Su = \displaystyle\sum_{i=1}^{M} u_i y_i. \end{array} \tag{2.3}$$

Clearly, by the Proposition 2.2, Su is the solution to the state equation for $u \in \mathbb{R}^M$.

Remark 2.3. The mapping S is linear and continuous.

Now, by replacing $y = Su$ in the cost functional J we introduce the reduced optimal control problem:

$$(\mathcal{P}) \qquad \begin{cases} \min\limits_{u \in \mathcal{U}_{ad}} f(u) = \dfrac{1}{2}\displaystyle\int_\Omega \left(\sum_{i=1}^{M} u_i y_i - y_d\right)^2 dx + \dfrac{\alpha}{2}||u||_2^2 + \beta||u||_1 \\ \text{subject to:} \\ \displaystyle\sum_{i=1}^{M} u_i y_i(x_j) \le b_j, \quad \forall j = 1, \ldots, \ell. \end{cases}$$

We denote the feasible set of controls by

$$\mathcal{U}_{feas} = \left\{ u \in \mathcal{U}_{ad} : \sum_{i=1}^{M} u_i y_i(x_j) \le b_j, \ \forall j = 1, \ldots, \ell \right\}. \tag{2.4}$$

Due to the presence of the state constraints, depending on the selection of b the last set has no reason to be non empty. Moreover, since we are looking for sparse solutions, control constraints should contain 0 as feasible vector. Thus, we make the following hypothesis.

Hypothesis 2.4. The upper and lower bounds u_b and u_a satisfy $u_a < 0 < u_b$ pointwise. Moreover, we assume that $\mathcal{U}_{feas}$ is not empty.

THEOREM 2.5. *Let the Tikhonov parameter α be positive, then there exists a unique solution for the problem $(\mathcal{P})$.*

Proof. It is clear from the definition of f that it is a continuous and strictly convex function because the Tikhonov term $\frac{\alpha}{2}\|u\|_2^2$. In this case, the existence of a unique solution is obtained even if $\mathcal{U}_{ad}$ is unbounded since $\mathcal{U}_{feas}$ remains convex and closed. ∎

Remark 2.6. Let $\bar{u}$ be the optimal control of $(\mathcal{P})$, then its associated state is given by

$$\bar{y} = \sum_{i=1}^{M} \bar{u}_i y_i.$$

We will refer to $\bar{y}$ as the *optimal state*, and to $(\bar{y}, \bar{u})$ as the *optimal pair*. Next, we are going to derive the necessary and sufficient conditions for problem $(\mathcal{P})$. For convenience, in the subsequent formulation the description of inequality constraints for problem $(\mathcal{P})$, will be represented using the function $g : \mathbb{R}^M \to \mathbb{R}^{\ell+2M}$ defined by

$$\begin{aligned} g_j(u) &= \sum_{i=1}^{M} u_i y_i(x_j) - b_j, && \text{for } j = 1, \dots, \ell, \\ g_{\ell+i}(u) &= u_{a,i} - u_i, && \text{for } i = 1, \dots, M, \\ g_{\ell+M+i}(u) &= u_i - u_{b,i}, && \text{for } i = 1, \dots, M. \end{aligned} \tag{2.5}$$

With this definition, we are able to formulate our optimal control problem as an optimization problem in finite dimensions as follows:

$$(\mathcal{P}') \qquad \begin{cases} \min\limits_{u \in \mathbb{R}^M} f(u) = \dfrac{1}{2} \displaystyle\int_\Omega \left(\sum_{i=1}^{M} u_i y_i - y_d \right)^2 dx + \dfrac{\alpha}{2} ||u||_2^2 + \beta ||u||_1 \\ \text{subject to:} \\ g(u) \leq 0. \end{cases}$$

Hypothesis 2.7 (Slater condition). There exist a control vector $u^\circ \in \operatorname{int} \mathcal{U}_{ad}$, such that

$$\sum_{i=1}^{M} u_i^\circ y_i(x_j) < b_j, \quad \forall j = 1, \dots, \ell. \tag{2.6}$$

Remark 2.8. In the absence of equality constraints, we shall notice that Hypothesis 2.7 implies Mangasarian–Fromovitz constraint qualification.

The following proposition establishes first order optimality conditions for $(\mathbf{P_M})$. The following conditions are obtained by applying Theorem 2.2 in [2].

THEOREM 2.9. *Let $\bar{u} \in \mathcal{U}_{feas}$ be the optimal solution for the problem $(\mathbf{P_M})$ and $\bar{y}$ be the optimal state. Then, under Hypothesis 2.7 there exists $(\nu, \mu_a, \mu_b) \in \mathbb{R}^\ell \times \mathbb{R}^M \times$*

$\mathbb{R}^M$ *such that for all* $i = 1, \dots, M$*, it follows that*

$$\int_\Omega (\bar{y} - y_d) y_i \, dx + \alpha \bar{u}_i + \sum_{j=1}^{\ell} \nu_j y_i(x_j) + \mu_{bi} - \mu_{ai} = -\beta, \qquad \text{if } \bar{u}_i > 0, \tag{2.7a}$$

$$\int_\Omega (\bar{y} - y_d) y_i \, dx + \alpha \bar{u}_i + \sum_{j=1}^{\ell} \nu_j y_i(x_j) + \mu_{bi} - \mu_{ai} = \beta, \qquad \text{if } \bar{u}_i < 0, \tag{2.7b}$$

$$\left| \int_\Omega (\bar{y} - y_d) y_i \, dx + \sum_{j=1}^{\ell} \nu_j y_i(x_j) + \mu_{bi} - \mu_{ai} \right| \leq \beta, \qquad \text{if } \bar{u}_i = 0, \tag{2.7c}$$

$$\nu \geq 0, \quad \text{and} \quad \nu_j(\bar{y}(x_j) - b_j) = 0, \tag{2.8a}$$

$$\mu_a \geq 0, \quad \text{and} \quad \mu_{ai}(u_{a,i} - \bar{u}_i) = 0, \tag{2.8b}$$

$$\mu_b \geq 0, \quad \text{and} \quad \mu_{bi}(\bar{u}_i - u_{b,i}) = 0. \tag{2.8c}$$

Proof. By composition, $f := J(S\,\cdot, \cdot)$ is a Lipschitz function. Moreover, g is a continuously differentiable affine function. By Hypothesis 2, the condition (i) of Theorem 2 in [2] is satisfied, therefore we conclude the existence of $(\nu, \mu_a, \mu_b) \in \mathbb{R}^\ell \times \mathbb{R}^M \times \mathbb{R}^M$ satisfying:

$$\sum_{j=1}^{\ell} \nu_j (v_j - g_j(\bar{u})) \leq 0, \qquad \forall v \in \mathbb{R}^\ell, \text{ with } v \leq 0 \tag{2.9a}$$

$$\sum_{i=1}^{M} \mu_{ai} (u_i - g_{\ell+i}(\bar{u})) \leq 0, \qquad \forall u \in \mathbb{R}^M, \text{ with } u \leq 0 \tag{2.9b}$$

$$\sum_{i=1}^{M} \mu_{bi} (u_i - g_{\ell+M+i}(\bar{u})) \leq 0, \qquad \forall u \in \mathbb{R}^M, \text{ with } u \leq 0. \tag{2.9c}$$

and

$$0 \in \partial f(\bar{u}) + (\nu, \mu_a, \mu_b)^\top \nabla g(\bar{u}) \tag{2.10}$$

Since the previous inequalities (2.9) fulfill for any $v \leq 0$ and $u \leq 0$ respectively, by appropriately testing v and u, we deduce that $(\nu, \mu_a, \mu_b) \geq 0$ and we also get the relations:

$$\nu_j g_j(\bar{u}) = \nu_j \left(\sum_{i=1}^{M} \bar{u}_i y_i(x_j) - b_j \right) = 0, \qquad \forall j = 1, \dots, \ell,$$

$$\mu_{ai} g_{i+\ell}(\bar{u}) = \mu_{ai}(u_{ai} - \bar{u}_i) = 0, \qquad \forall i = 1, \dots, M$$

$$\mu_{bi} g_{i+\ell+M}(\bar{u}) = \mu_{bi}(\bar{u}_i - u_{bi}) = 0, \qquad \forall i = 1, \dots, M.$$

On the other hand, from (2.10) and since f is the sum of a differentiable term and the 1–norm term, the usual rules of subdifferential calculus (see [12]) imply

$$0 \in [(\bar{y} - y_d, y_i)]_{i=1}^M + \alpha \bar{u} + \beta \partial \| \cdot \|_1(\bar{u}) + \left[\sum_{j=1}^{\ell} \nu_j y_i(x_j) \right]_{i=1}^M - \mu_a + \mu_b,$$

which is equivalent to:

$$[(\bar{y} - y_d, y_i)]_{i=1}^M + \alpha\bar{u} + \left[\sum_{j=1}^{\ell} \nu_j y_i(x_j)\right]_{i=1}^M - \mu_a + \mu_b \in -\beta\partial\|\cdot\|_1(\bar{u}).$$

Finally, (2.7) is a consecuence the subdifferential of the 1–norm:

$$\partial\|\cdot\|_1(\bar{u}) = \left\{ z \in \mathbb{R}^M : z_i \in \left\{ \begin{array}{ll} \{\text{sign}(\bar{u}_i)\}, & \text{if } \bar{u}_i \neq 0, \\ [-1,1], & \text{if } \bar{u}_i = 0. \end{array} \right\}. \right.$$

∎

3. Approximation by the finite element method. In the previous section our problem was transformed into a finite–dimensional programming problem. However, its numerical solution still requires the approximation of the state equation. Let us describe the numerical approximation of the state equation by finite elements. The associated definitions and classical results of the approximation can be found in [10] and [26].

3.1. Discretization of the state equation . We consider a family of meshes $(\mathcal{T}_h)_{h>0}$ of $\bar{\Omega}$, which consist of triangles $T \in \mathcal{T}_h$ such that the following conditions are satisfied:

- $\bigcup_{T\in\mathcal{T}_h} T = \bar{\Omega}$ and
- For two triangular elements T_i and T_j, $i \neq j$ they share a vertex, a side or are disjoints.

For each triangle $T \in \mathcal{T}_h$, we denote $\rho(T)$ the diameter of T, and $\sigma(T)$ the diameter of the largest ball contained in T. The mesh size h is defined by

$$h = \max_{T\in\mathcal{T}_h} \rho(T).$$

In addition, we impose in addition the following regularity assumption on the grid:

Hypothesis 3.1. There exist two positive constants ρ y σ such that

$$\frac{\rho(T)}{\sigma(T)} \leq \sigma \quad \text{y} \quad \frac{h}{\rho(T)} \leq \rho, \qquad \forall T \in \mathcal{T}_h,$$

for all $h > 0$.

Associated with the triangulation T_h, we define the set of piecewise linear and continuous functions:

$$Y_h = \{y_h \in C(\bar{\Omega}) : y_h|_T \in P_1(T),\ \forall T \in \mathcal{T}_h,\ y_h = 0 \text{ on } \Gamma\},$$

where $P_1(T)$ denotes the set of affine real-valued functions defined on T. Additionally, for each $i = 1, \ldots, M$, we define the discrete state y_i^h, as the unique function of Y_h that satisfies

$$\sum_{j,k=1}^{2} (a_{jk}\partial_j y_i^h, \partial_k \phi_h) + (a_0 y_i^h, \phi_h) = (e_i, \phi_h), \tag{3.1}$$

for all $\phi_h \in Y_h$. The the error of the approximation of the solution of (2.1) by the solution of equation (3.1) can be estimated in the several norms, summarized in the next result.

PROPOSITION 3.2. *There exists a constant $c > 0$ independent of h such that*

$$||y_i - y_i^h|| \leq ch^2, \tag{3.2a}$$

$$||y_i - y_i^h||_{L^\infty(\Omega)} \leq ch, \tag{3.2b}$$

$$||y_i - y_i^h||_{L^\infty(\Omega_0)} \leq ch^2|\log h|, \tag{3.2c}$$

for $i = 1, \ldots, M$.

For the proof of (3.2a) and (3.2c) see [24, Proposition 3.3]. The estimate (3.2b) can be found in [25, Theorem 3.1].

The *discrete state* associated to $u \in \mathcal{U}_{ad}$ is denoted by y_u^h and it is defined as the unique element y_u^h of Y_h that satisfies the equation

$$\sum_{j,k=1}^{2} (a_{jk}\partial_j y_u^h, \partial_k \phi_h) + (a_0 y_u^h, \phi_h) = \left(\sum_{i=1}^{M} u_i e_i, \phi_h\right), \quad \forall \phi \in Y_h. \tag{3.3}$$

Analogously, using the superposition principle the discrete state y_u^h associated to $u \in \mathcal{U}_{ad}$ can be expressed as:

$$y_u^h = \sum_{i=1}^{M} u_i y_i^h. \tag{3.4}$$

Observe that by the triangular inequality it follows that the Proposition 3.2 holds for y_u and y_u^h. Furthermore, we can also define the *discrete control–to-state operator* $S^h : \mathbb{R}^M \to Y_h$ by

$$u \mapsto \sum_{i=1}^{M} u_i y_i^h. \tag{3.5}$$

3.2. Discretization of the optimal control problem . We are in place to define the following discrete counterpart of problem $(\mathbf{P_M})$ by replacing the discrete state y_u^h defined in (3.4) in $(\mathbf{P_M})$ instead of y. Therefore, the discrete optimal control problem reads:

$$(\mathbf{P_M^h}) \qquad \begin{cases} \min\limits_{(y^h,u)} J(y^h, u) := \dfrac{1}{2}\displaystyle\int_\Omega (y^h - y_d)^2\, dx + \dfrac{\alpha}{2}||u||_2^2 + \beta||u||_1 \\ \text{subject to:} \\ \text{discrete equation (3.3)}, \\ \quad y^h(x_j) \leq b_j, \quad \forall j = 1, \ldots, \ell, \\ \quad u \in \mathcal{U}_{ad}. \end{cases}$$

The existence of a unique optimal control $\bar{u}^h$ for $(\mathbf{P_M^h})$ is obtained by the same finite–dimensional arguments as in the case of problem $(\mathbf{P_M})$. The main result of this paper is given in the following theorem.

THEOREM 3.3. *Let $\bar{u}$ and $\bar{u}^h$ be the optimal controls of problems $(\mathbf{P_M})$ and $(\mathbf{P_M^h})$ respectively. Then, the estimate*

$$||\bar{u} - \bar{u}^h|| \leq Ch^2|\log h|, \tag{3.6}$$

holds for some constant $C > 0$, independent of mesh parameter h.

We claim that this estimate is optimal in the sense that it has the same order of error than for the state equation given in Proposition 3.2. These kind of estimates are important for a–priori knowledge of how fast we can improve the expected quality of the solution when the number of elements in the discretization is increased. Moreover, error estimates can be used to design strategies for the reduction of the dimensionality of the optimal control problem, see [23].

The error estimate (3.6) is analogous to the estimate obtained in [24] for the smooth case. However, its derivation requires to cope with the non differentiability of the 1–norm. In order to derive the estimate (3.6) we proceed to reformulate problem $(\mathcal{P})$ as a equivalent differentiable problem by splitting the control variable in its positive and negative parts. Then, a stability analysis is carried out which allows us to establish the desired estimate. We reserve the proof of this theorem for Section 5.

4. Reformulation of $(\mathbf{P_M})$ as a smooth problem. In order to reformulate our problem, we introduce the following notation. The positive and negative parts of a vector $u \in \mathbb{R}^M$ are denoted by the vectors $u^+ \in \mathbb{R}^M$ and $u^- \in \mathbb{R}^M$, respectively and whose components are defined by

$$u_i^+ = \begin{cases} u_i & \text{if } u_i \geq 0, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad u_i^- = \begin{cases} -u_i & \text{if } u_i \leq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

We recall following properties of the positive and negative parts of $u \in \mathbb{R}^M$:

$$u = u^+ - u^-, \tag{4.2a}$$

$$u^{+\top} u^- = 0, \tag{4.2b}$$

$$||u||_1 = \sum_{i=1}^{M} (u_i^+ + u_i^-), \tag{4.2c}$$

$$||u||_2^2 = \sum_{i=1}^{M} \left[u_i^{+\,2} + u_i^{-\,2} \right]. \tag{4.2d}$$

Using relations (4.2) we reformulate problem $(\mathbf{P_M})$ as a smooth problem, however the number of variables is duplicated. In the following we will use bold characters to denote variables associated with $(\mathbf{P_M})$. The reformulated smooth optimal control problem reads:

$$(\mathbf{P_{2M}}) \begin{cases} \min\limits_{(y,\mathbf{u})} \mathbf{J}(y,\mathbf{u}) = \dfrac{1}{2}\displaystyle\int_\Omega (y - y_d)^2\, dx + \dfrac{\alpha}{2}||\mathbf{u}||_2^2 + \beta \sum_{i=1}^{2M} \mathbf{u}_i \\ \text{subject to:} \\ \begin{aligned} Ay(x) &= \sum_{i=1}^{M} (\mathbf{u}_i - \mathbf{u}_{i+M}) e_i(x), && \text{in } \Omega, \\ y(x) &= 0, && \text{on } \Gamma, \\ y(x_j) &\leq b_j, && \forall j = 1, \ldots, \ell, \\ [\mathbf{u}_i - \mathbf{u}_{i+M}]_{i=1}^{M} &\in \mathcal{U}_{ad}, \\ \mathbf{u}_i &\geq 0, && \forall i = 1, \ldots, 2M. \end{aligned} \end{cases}$$

Note that problem $(\mathbf{P_{2M}})$ has a unique solution if $\alpha > 0$. Let us denote its solution by $\bar{\mathbf{u}}$. Before we proof the equivalence of problems $(\mathbf{P_M})$ and $(\mathbf{P_{2M}})$ and the relation

of the positive and negative parts of a vector, the following orthogonality result is expected from the decomposition (4.2a) and its proof can be found in [33].

LEMMA 4.1. *If* $\bar{\mathbf{u}}$ *is the optimal control of* $(\mathbf{P_{2M}})$ *then* $\bar{\mathbf{u}}_i\bar{\mathbf{u}}_{i+M} = 0$ *for all* $i = 1, \dots, M$.

The equivalence between problems $(\mathbf{P_M})$ and $(\mathbf{P_{2M}})$ is stated in the following lemma.

LEMMA 4.2. *Let* $\bar{u}$ *and* $\bar{\mathbf{u}}$ *be the solutions of the problems* $(\mathbf{P_M})$ *and* $(\mathbf{P_{2M}})$ *respectively, then*

$$\left[\bar{\mathbf{u}}_i\right]_{i=1}^{M} = \bar{u}^+ \quad \textit{and} \quad \left[\bar{\mathbf{u}}_{i+M}\right]_{i=1}^{M} = \bar{u}^-,$$

i.e. $\bar{u} = \left[\bar{\mathbf{u}}_i\right]_{i=1}^{M} - \left[\bar{\mathbf{u}}_{i+M}\right]_{i=1}^{M}$.

Proof. Let $\bar{\mathbf{u}}$ be the solution for problem $(\mathbf{P_{2M}})$ and $\bar{u}$ the solution of $(\mathbf{P_M})$ with $\bar{u}^+$ and $\bar{u}^-$ their positive and negative parts respectively. Then, let us define the vectors:

$$\bar{\mathbf{w}} = \begin{bmatrix} \bar{u}^+ \\ \bar{u}^- \end{bmatrix} \in \mathbb{R}^{2M}, \text{ and } \bar{w} = [\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i+M}]_{i=1}^{M} \in \mathbb{R}^{M}.$$

Since $[\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_{i+M}]_{i=1}^{M} = \bar{u}^+ - \bar{u}^- = \bar{u} \in \mathcal{U}_{ad}$ and

$$\sum_{i=1}^{M}(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_{i+M})y_i(x_j) - b_j = \sum_{i=1}^{M}\bar{u}_i y_i(x_j) - b_j \leq 0, \quad \forall j = 1, \dots, \ell,$$

it follows that $\bar{\mathbf{w}}$ is and admisible control for $(\mathbf{P_{2M}})$ with associated state $y_{\bar{\mathbf{w}}} = S[\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_{i+M}]_{i=1}^{M}$. Moreover, taking into account (4.2b) we have that

$$\begin{aligned}
\mathbf{J}(y_{\bar{\mathbf{w}}}, \bar{\mathbf{w}}) &= \frac{1}{2}\left|\left|y_{\bar{\mathbf{w}}} - y_d\right|\right|^2 + \frac{\alpha}{2}||\bar{\mathbf{w}}||_2^2 + \beta\sum_{i=1}^{2M}\bar{\mathbf{w}}_i \\
&= \frac{1}{2}\left|\left|\sum_{i=1}^{M}(\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_{i+M})y_i - y_d\right|\right|^2 + \frac{\alpha}{2}\sum_{i=1}^{2M}\bar{\mathbf{w}}_i^2 + \beta\sum_{i=1}^{2M}\bar{\mathbf{w}}_i \\
&= \frac{1}{2}\left|\left|\sum_{i=1}^{M}\bar{u}_i y_i - y_d\right|\right|^2 + \frac{\alpha}{2}\sum_{i=1}^{M}(\bar{u}_i^+ + \bar{u}_i^-)^2 + \beta\sum_{i=1}^{M}\bar{u}_i^+ + \bar{u}_i^- \\
&= \frac{1}{2}\left|\left|\bar{y} - y_d\right|\right|^2 + \frac{\alpha}{2}\sum_{i=1}^{M}(\bar{u}_i)^2 + \beta\sum_{i=1}^{M}|\bar{u}_i| = J(\bar{y}, \bar{u}).
\end{aligned} \tag{4.3}$$

Analogously, $\bar{w}$ is admissible for problem $(\mathbf{P_M})$ and $J(y_{\bar{w}}, \bar{w}) = \mathbf{J}(y_{\bar{\mathbf{u}}}, \bar{\mathbf{u}})$. Then, by (4.3) and the optimality of $(\bar{y}, \bar{u})$ we have that

$$\mathbf{J}(y_{\bar{\mathbf{w}}}, \bar{\mathbf{w}}) = J(\bar{y}, \bar{u}) \leq J(y_{\bar{w}}, \bar{w}) = \mathbf{J}(y_{\bar{\mathbf{u}}}, \bar{\mathbf{u}}).$$

Therefore, uniqueness of the solution of the optimal control problem $(\mathbf{P_{2M}})$ allows us to conclude that $\bar{\mathbf{w}} = \bar{\mathbf{u}}$. Similarly, we deduce that $\bar{u} = \bar{w}$, therefore the proof is complete. ∎

Let us rewrite problem $(\mathbf{P_{2M}})$ as a finite dimensional programming problem. First, by replacing the control–to–state operator S, consider the reduced cost function

$$\mathbf{f}(\mathbf{u}) = \frac{1}{2}\int_{\Omega}(S([\mathbf{u}_i - \mathbf{u}_{i+M}]_{i=1}^{M}) - y_d)^2\,dx + \frac{\alpha}{2}||\mathbf{u}||_2^2 + \beta\sum_{i=1}^{2M}\mathbf{u}_i. \tag{4.4}$$

For convenience, we introduce the symmetric and positive semi–definite matrix $\mathcal{Y} \in \mathbb{R}^{2M\times 2M}$ which has the form

$$\mathcal{Y} = \begin{bmatrix} Y & -Y \\ -Y & Y \end{bmatrix}, \tag{4.5}$$

where the entries of $Y \in \mathbb{R}^{M\times M}$ are given by

$$Y_{ij} = \int_\Omega y_i(x) y_j(x)\, dx. \tag{4.6}$$

Moreover,the vector $\mathbf{b} \in \mathbb{R}^{2M}$ is defined by its components $\mathbf{b}_i = \int_\Omega y_i y_d\, dx$ and $\mathbf{b}_{i+M} = -\int_\Omega y_i y_d dx$, for $i = 1, \ldots, M$. In this way, we can rewrite

$$\mathbf{f}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top(\mathcal{Y} + \alpha I)\mathbf{u} + \mathbf{b}^\top\mathbf{u} + \beta\sum_{i=1}^{2M}\mathbf{u}_i + c, \tag{4.7}$$

with corresponding constant c term depending on y_d. In addition, we collect all inequality constraints by defining the constraint function $\mathbf{g} : \mathbb{R}^{2M} \to \mathbb{R}^{\ell+4M}$ as follows:

$$\begin{aligned}
\mathbf{g}_j(\mathbf{u}) &= \sum_{i=1}^{M}(\mathbf{u}_i - \mathbf{u}_{i+M})y_i(x_j) - b_j \\
&= [\mathcal{Z}\mathbf{u} - b]_j, && \text{for } \ j = 1, \ldots, \ell, && \text{(4.8a)}\\
\mathbf{g}_{\ell+i}(\mathbf{u}) &= u_{a,i} - \mathbf{u}_i + \mathbf{u}_{i+M}, && \text{for } \ i = 1, \ldots, M, && \text{(4.8b)}\\
\mathbf{g}_{i+\ell+M}(\mathbf{u}) &= \mathbf{u}_i - \mathbf{u}_{i+M} - u_{b,i}, && \text{for } \ i = 1, \ldots, M, && \text{(4.8c)}\\
\mathbf{g}_{i+\ell+2M}(\mathbf{u}) &= -\mathbf{u}_i, && \text{for } \ i = 1, \ldots, 2M. && \text{(4.8d)}
\end{aligned}$$

Here, the matrix $\mathcal{Z} \in \mathbb{R}^{\ell\times 2M}$ is such that $\mathcal{Z}_{ji} = y_i(x_j)$ for and $\mathcal{Z}_{ji+M} = -y_i(x_j)$ $i = 1, \ldots, M$ and $j = 1, \ldots, \ell$. Hence, using the above definitions we can write problem ($\mathbf{P_{2M}}$) in the compact form: we can write problem ($\mathbf{P_{2M}}$) in the compact form:

$$\begin{cases} \min\limits_{\mathbf{u}\in\mathbb{R}^{2M}} \mathbf{f}(\mathbf{u}) \\ \text{subject to:} \\ \mathbf{g}(\mathbf{u}) \le 0. \end{cases} \tag{4.9}$$

The *Lagrangean* associated to problem (4.9) is given by:

$$\begin{aligned}
\mathcal{L}(\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) =& \mathbf{f}(\mathbf{u}) + \sum_{j=i}^{\ell}\boldsymbol{\nu}_j\mathbf{g}_j(\mathbf{u}) + \sum_{i=1}^{M}\boldsymbol{\mu}_{a\,i}\mathbf{g}_{i+\ell}(\mathbf{u}) + \sum_{i=1}^{M}\boldsymbol{\mu}_{b\,i}\mathbf{g}_{i+\ell+M}(\mathbf{u}) \\
&+ \sum_{i=1}^{2M}\boldsymbol{\eta}_i\mathbf{g}_{i+\ell+2M}(\mathbf{u}).
\end{aligned} \tag{4.10}$$

In the following lemma we establish *first–order necessary optimality conditions* c.f. [3].

LEMMA 4.3. *Under Hypothesis 2.7, it follows that the constraints given by* $\mathbf{g}$ *are regular in the sense of [27] and [28]. That is,*

$$0 \in \operatorname{int}\{\mathbf{g}(\bar{\mathbf{u}}) + \mathbf{g}'(\bar{\mathbf{u}})(\mathbb{R}^{2M} - \bar{\mathbf{u}}) + \mathbb{R}_+^{4M+\ell}\}. \tag{4.11}$$

Proof. Since $\mathbf{g}$ is defined in terms of affine functions c.f. (4.8), then there are a matrix $G \in \mathbb{R}^{\ell+4M\times 2M}$ and a vector $d \in \mathbb{R}^{\ell+4M}$ such that the constraint $\mathbf{g}(\mathbf{u}) \leq 0$ is equivalently expressed as $G\mathbf{u} \leq d$ for all $\mathbf{u} \in \mathbb{R}^{2M}$. In this case, condition (4.11) is equivalent to

$$d \in \operatorname{int}\{G(\mathbb{R}^{2M}) + \mathbb{R}_+^{\ell+4M}\}, \tag{4.12}$$

which is verified with the help of $\mathbf{u}^\circ$. Indeed, by its definition we have that $0 > \mathbf{g}(\mathbf{u}^\circ) = G\mathbf{u}^\circ - d$. This fact implies that $d \in \operatorname{int}\{\mathbb{R}_+^{\ell+4M}\} + G\mathbf{u}^\circ \subset \operatorname{int}\{G(\mathbb{R}^{2M}) + \mathbb{R}_+^{\ell+4M}\}$ ∎

THEOREM 4.4. *Let* $\bar{\mathbf{u}} \in \mathbb{R}^{2M}$ *the solution of problem (4.9) with associated optimal state* $\bar{\mathbf{y}} = \sum_{i=1}^{M}(\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i+M})y_i$. *Then, there exists* $(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}}) \in \mathbb{R}_+^{\ell} \times \mathbb{R}_+^{M} \times \mathbb{R}_+^{M} \times \mathbb{R}_+^{2M} \cup \{0\}$ *such that the conditions below are satisfied,*

$$(\bar{\mathbf{y}} - y_d, y_i) + \alpha\bar{\mathbf{u}}_i \quad +\beta + \sum_{j=1}^{\ell} \boldsymbol{\nu}_j y_i(x_j) + \boldsymbol{\mu}_{bi} - \boldsymbol{\mu}_{ai} - \boldsymbol{\eta}_i = 0, \tag{4.13a}$$

$$-(\bar{\mathbf{y}} - y_d, y_i) + \alpha\bar{\mathbf{u}}_{i+M} + \beta - \sum_{j=1}^{\ell} \boldsymbol{\nu}_j y_i(x_j) - \boldsymbol{\mu}_{bi} + \boldsymbol{\mu}_{ai} - \boldsymbol{\eta}_{i+M} \ = 0. \tag{4.13b}$$

for $i = 1, \ldots, M$, *holds, together with the following conditions:*

$$\bar{\mathbf{u}} \geq 0, \tag{4.14a}$$

$$u_{ai} \leq \bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i+M} \leq u_{bi}, \quad \forall i = 1, \ldots, 2M. \tag{4.14b}$$

$$\boldsymbol{\nu}_j\,(\bar{\mathbf{y}}(x_j) - b_j) = 0, \quad \forall j = 1, \ldots, \ell, \tag{4.14c}$$

$$\boldsymbol{\mu}_{a_i}(u_{a,i} - \bar{\mathbf{u}}_i + \bar{\mathbf{u}}_{i+M}) = 0, \quad \forall i = 1, \ldots, M, \tag{4.14d}$$

$$\boldsymbol{\mu}_{b_i}(\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i+M} - u_{b,i}) = 0, \quad \forall i = 1, \ldots, M, \tag{4.14e}$$

$$\boldsymbol{\eta}_i\bar{\mathbf{u}}_i = 0, \quad \forall i = 1, \ldots, 2M. \tag{4.14f}$$

Proof. Hypothesis 2.7 and Lemma 4.2 imply that $\bar{\mathbf{u}}$ is a regular point for problem $(\mathbf{P_{2M}})$. Therefore, the result follows by applying the standard theory of constrained optimization c.f. [3]. ∎

Similarly, we consider the reduced discretized counterpart of problem $(\mathbf{P_{2M}})$ by replacing the state equation by the finite element approximation (3.3). The problem is formulated as follows:

$$(\mathbf{P}_{\mathbf{2M}}^{\mathbf{h}}) \quad \begin{cases} \min\limits_{(y^h,\mathbf{u})} \mathbf{J}(y^h, \mathbf{u}) = \dfrac{1}{2}\displaystyle\int_\Omega (y^h - y_d)^2\,dx + \dfrac{\alpha}{2}||\mathbf{u}||_2^2 + \beta\sum_{i=1}^{2M}\mathbf{u}_i \\ \text{subject to:} \\ \text{discrete equation (3.3) with } u_i = \mathbf{u}_i - \mathbf{u}_{i+M},\ i = 1, \ldots, M. \\ \qquad y^h(x_j) \ \leq b_j, \qquad \forall j = 1, \ldots, \ell, \\ \ [\mathbf{u}_i - \mathbf{u}_{i+M}]_{i=1}^{M} \ \in \mathcal{U}_{ad}, \\ \qquad \mathbf{u}_i \ \geq 0, \qquad \forall i = 1, \ldots, 2M. \end{cases}$$

Observe that $\bar{\mathbf{u}}^h$ is a unique solution for $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$. This fact can be deduced by the same arguments as in the non discretized case. Analogously to Lemma 4.2 for problem $(\mathbf{P_{2M}})$, it is also true that $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$ is equivalent to problem $(\mathbf{P}^{\mathbf{h}}_{\mathbf{M}})$ and this equivalence is expressed by the relations:

$$\left[\bar{\mathbf{u}}^h_i\right]_{i=1}^M = \bar{u}^{h+} \quad \text{and} \quad \left[\bar{\mathbf{u}}^h_{i+M}\right]_{i=1}^M = \bar{u}^{h-} \tag{4.15}$$

$$\text{i.e., } \bar{u}^h = \left[\bar{\mathbf{u}}^h_i\right]_{i=1}^M - \left[\bar{\mathbf{u}}^h_{i+M}\right]_{i=1}^M. \tag{4.16}$$

Because of the equivalence of problems $(\mathbf{P_M})$ and $(\mathbf{P_{2M}})$ (also between $(\mathbf{P}^{\mathbf{h}}_{\mathbf{M}})$ and $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$) and the relation of the positive and negative parts of a vector, the following orthogonality result is expected and its proof can be found in [33].

LEMMA 4.5. *If* $\bar{\mathbf{u}}$ *(respectively* $\bar{\mathbf{u}}^h$*) is the optimal control of* $(\mathbf{P_{2M}})$ *(respectively* $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$*) then* $\bar{\mathbf{u}}_i\bar{\mathbf{u}}_{i+M} = 0$ *(respectively* $\bar{\mathbf{u}}^h_i\bar{\mathbf{u}}^h_{i+M} = 0$*) for all* $i = 1, \dots, M$.

Moreover, we denote by $\mathbf{g^h}$ and $\mathcal{Y}^h$ the discrete analogues of $\mathbf{g}$ and $\mathcal{Y}$ defined in (4.8) and in (4.5), respectively, which are obtained by replacing y_i by y_i^h correspondingly. $\mathbf{b}^h$ is defined analogously. The discrete objective function $\mathbf{f^h}$ is also defined in this manner,

$$\mathbf{f^h}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top(\mathcal{Y}^h + \alpha I)\mathbf{u} + {\mathbf{b}^h}^\top\mathbf{u} + \beta\sum_{i=1}^{2M}\mathbf{u}_i + c^h, \tag{4.17}$$

where c^h is an appropriate constant depending on y_d and h. Now we are able to write problem $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$ compactly, as the following finite–dimensional optimization problem.

$$\begin{cases} \min\limits_{\mathbf{u}\in\mathbb{R}^{2M}} \mathbf{f^h}(\mathbf{u}) \\ \text{subject to:} \\ \mathbf{g^h}(\mathbf{u}) \leq 0. \end{cases} \tag{4.18}$$

Remark 4.6. From (3.2b) it follows that $|y_i(x_j) - y_i^h(x_j)| \leq ch^2|\log(h)|$ for all $i = 1, \dots, M$ and all $j = 1, \dots, \ell$. This fact, together with Hypothesis 2.7 imply that u° is also a slater point for $(\mathbf{P}^{\mathbf{h}}_{\mathbf{M}})$. Therefore, $\bar{\mathbf{u}}^h$ is a regular point for $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$ if h is sufficiently small. Applying the Lagrange principle again, we obtain a similar optimality system as in Theorem 4.4. The *discrete optimality system* for the discrete optimal pair $(\bar{\mathbf{u}}^h, \bar{y}^h)$ and the corresponding associated multipliers $(\bar{\boldsymbol{\nu}}^h, \bar{\boldsymbol{\mu}}^h_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}^h_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}}^h) \in \mathbb{R}^\ell_+ \times \mathbb{R}^M_+ \times \mathbb{R}^M_+ \times \mathbb{R}^{2M}_+ \cup \{0\}$, is given below.

$$(\bar{\mathbf{y}}^h - y_d, y_i^h) + \alpha\bar{\mathbf{u}}^h_i \quad +\beta + \sum_{j=1}^{\ell}\boldsymbol{\nu}^h_j y_i^h(x_j) + \boldsymbol{\mu}^h_{b\,i} - \boldsymbol{\mu}^h_{a\,i} - \boldsymbol{\eta}^h_i = 0, \tag{4.19a}$$

$$-(\bar{\mathbf{y}}^h - y_d, y_i^h) + \alpha\bar{\mathbf{u}}^h_{i+M} + \beta - \sum_{j=1}^{\ell}\boldsymbol{\nu}^h_j y_i^h(x_j) - \boldsymbol{\mu}^h_{b\,i} + \boldsymbol{\mu}^h_{a\,i} - \boldsymbol{\eta}^h_{i+M} \quad = 0. \tag{4.19b}$$

for $i = 1, \dots, M,$. Moreover, the following conditions are satisfied:

$$\bar{\mathbf{u}}^h \geq 0, \tag{4.20a}$$

$$u_{ai} \leq \bar{\mathbf{u}}_i^h - \bar{\mathbf{u}}_{i+M}^h \leq u_{bi}, \quad \forall i = 1, \dots, 2M. \tag{4.20b}$$

$$\boldsymbol{\nu}_j^h \left(\bar{\mathbf{y}}^h(x_j) - b_j \right) = 0, \quad \forall j = 1, \dots, \ell, \tag{4.20c}$$

$$\boldsymbol{\mu}_{a\,i}^h (u_{a,i} - \bar{\mathbf{u}}_i^h + \bar{\mathbf{u}}_{i+M}^h) = 0, \quad \forall i = 1, \dots, M, \tag{4.20d}$$

$$\boldsymbol{\mu}_{b\,i}^h (\bar{\mathbf{u}}_i^h - \bar{\mathbf{u}}_{i+M}^h - u_{b,i}) = 0, \quad \forall i = 1, \dots, M, \tag{4.20e}$$

$$\boldsymbol{\eta}_i^h \bar{\mathbf{u}}_i^h = 0, \quad \forall i = 1, \dots, 2M. \tag{4.20f}$$

5. Derivation of the error of convergence. This section is devoted to the derivation of the order of convergence for the finite element approximation ($\mathbf{P_{2M}^h}$) of ($\mathbf{P_{2M}}$) which is the intermediate step in order to derive an estimation of the error for the solution of ($\mathbf{P_M^h}$). Therefore, we develop the necessary tool for the error estimation by regarding problem (4.18) as a perturbation of (4.9).

Continuing with our analysis, we define the cone $K = \mathbb{R}_+^\ell \times \mathbb{R}_+^M \times \mathbb{R}_+^M \times \mathbb{R}_+^{2M} \cup \{0\}$. Then, we are interested in the stability analysis, with respect to h, of the problem

$$\begin{cases} \min\limits_{\mathbf{u}} \mathbf{f}(\mathbf{u}) \\ \mathbf{g}(\mathbf{u}) \leq_K 0, \end{cases} \quad \text{and its perturbation} \quad \begin{cases} \min\limits_{\mathbf{u}} \mathbf{f^h}(\mathbf{u}) \\ \mathbf{g^h}(\mathbf{u}) \leq_K 0. \end{cases}$$

For the comparison of the solutions of the preceding problems we first obtain the following estimation of the objective and the constraint functions.

LEMMA 5.1. *There exist $C > 0$, independent of h, such that for h small enough it follows that*

$$\begin{aligned} &|\mathbf{f}(\mathbf{u}) - \mathbf{f^h}(\mathbf{w})| + ||\nabla \mathbf{f}(\mathbf{u}) - \nabla \mathbf{f^h}(\mathbf{w})|| + ||\nabla^2 \mathbf{f}(\mathbf{u}) - \nabla^2 \mathbf{f^h}(\mathbf{w})|| \\ &\quad + \sum_{j=1}^{\ell} \left(|\mathbf{g}_j(\mathbf{u}) - \mathbf{g^h}_j(\mathbf{w})| + ||\nabla \mathbf{g}_j(\mathbf{u}) - \nabla \mathbf{g^h}_j(\mathbf{w})|| \right) \leq C(||\mathbf{u} - \mathbf{w}|| + h^2 |\log h|) \end{aligned}$$

for all $\mathbf{u}, \mathbf{w} \in \mathbb{R}^{2M}$.

Proof. The proof of this Lemma can be found in [25], and follows by taking into account the estimate (3.2c) within the Taylor expansion of $\mathbf{f}$. The estimates for $\mathbf{g}$ are direct since $\mathbf{g}$ is affine linear with respect to $\mathbf{u}$. ∎

LEMMA 5.2. *There exists a constant $C > 0$, independent of h, such that for h small enough, we have the estimate*

$$\|\mathcal{Y} - \mathcal{Y}^h\| \leq Ch^2, \tag{5.1}$$

in the Frobenius norm. Also the estimate,

$$\|\mathbf{b} - \mathbf{b}^h\| \leq Ch^2. \tag{5.2}$$

holds.

Proof. In order to estimate the Frobenius norm of $\mathcal{Y} - \mathcal{Y}^h$, we consider the dif-

ference of their entries. By applying Cauchy–Schwarz inequality we get:

$$\begin{aligned}|Y_{ij} - Y_{ij}^h| &= \left| \int_\Omega y_i(x) y_j(x)\, dx - \int_\Omega y_i^h(x) y_j^h(x)\, dx \right| \\ &\le \left| \int_\Omega (y_i - y_i^h) y_j dx \right| + \left| \int_\Omega (y_j^h - y_j) y_i^h dx \right| \\ &\le \|y_i - y_i^h\| \|y_j\| + \|y_j^h - y_j\| \|y_i^h\|,\end{aligned}$$

for all $i, j = 1, \dots, M$. Therefore, recalling that $\|y_i^h\|$ is bounded for $i = 1, \dots, M$, and in view of (3.2a), we conclude that there exist a constant $C > 0$ such that

$$|Y_{ij} - Y_{ij}^h| \le C h^2.$$

From this estimate we compute (5.1). The estimate (5.2) follows by analogous arguments. ∎

Among other properties of the approximated problem ($\mathbf{P^h_{2M}}$), boundedness of its multipliers is an important feature in the derivation of error estimates. We show this property in the following Lemma.

LEMMA 5.3. *For sufficiently small h, the corresponding multipliers $\boldsymbol{\nu}^h, \boldsymbol{\mu}_a^h, \boldsymbol{\mu}_b^h, \boldsymbol{\eta}^h$ associated with optimality conditions for problem ($\mathbf{P^h_{2M}}$) are uniformly bounded.*

Proof. Let us multiply equations (4.19) by $\mathbf{u}_i^\circ - \bar{\mathbf{u}}_i^h$. By adding them up we obtain:

$$\begin{aligned}0 =& \nabla \mathbf{f^h}(\bar{\mathbf{u}}^h)^\top (\mathbf{u}^\circ - \bar{\mathbf{u}}^h) + \sum_{i=1}^{M} \sum_{j=1}^{\ell} \boldsymbol{\nu}_j^h (\mathbf{u}_i^\circ - \bar{\mathbf{u}}_i^h - \mathbf{u}_{i+M}^\circ + \bar{\mathbf{u}}_{i+M}^h) y_i^h(x_j) \\ &+ \sum_{i=1}^{M} \boldsymbol{\mu}_{b\,i}^h (\mathbf{u}_i^\circ - \bar{\mathbf{u}}_i^h - \mathbf{u}_{i+M}^\circ + \bar{\mathbf{u}}_{i+M}^h) - \sum_{i=1}^{M} \boldsymbol{\mu}_{a\,i}^h (\mathbf{u}_i^\circ - \bar{\mathbf{u}}_i^h - \mathbf{u}_{i+M}^\circ + \bar{\mathbf{u}}_{i+M}^h) \\ &- \sum_{i=1}^{2M} \boldsymbol{\eta}_i^h (\mathbf{u}_i^\circ - \bar{\mathbf{u}}_i^h) \\ =& \nabla \mathbf{f^h}(\bar{\mathbf{u}}^h)^\top (\mathbf{u}^\circ - \bar{\mathbf{u}}^h) + \sum_{j=1}^{\ell} \boldsymbol{\nu}_j^h (\mathbf{y}^{\circ\, h}(x_j) - b_j) \\ &+ \sum_{i=1}^{M} \boldsymbol{\mu}_{b\,i}^h (\mathbf{u}_i^\circ - \mathbf{u}_{i+M}^\circ - u_{bi}) - \sum_{i=1}^{M} \boldsymbol{\mu}_{a\,i}^h (\mathbf{u}_i^\circ - \mathbf{u}_{i+M}^\circ - u_{ai}) - \sum_{i=1}^{2M} \boldsymbol{\eta}_i^h \mathbf{u}_i^\circ,\end{aligned}$$

which results in the relation

$$\begin{aligned}&\sum_{j=1}^{\ell} \boldsymbol{\nu}_j^h (b_j - \mathbf{y}^{\circ\, h}(x_j)) + \sum_{i=1}^{M} \boldsymbol{\mu}_{b\,i}^h (u_{bi} - \mathbf{u}_i^\circ + \mathbf{u}_{i+M}^\circ) \\ &\qquad + \sum_{i=1}^{M} \boldsymbol{\mu}_{a\,i}^h (\mathbf{u}_i^\circ - \mathbf{u}_{i+M}^\circ - u_{ai}) + \sum_{i=1}^{2M} \boldsymbol{\eta}_i^h \mathbf{u}_i^\circ = \nabla \mathbf{f^h}(\bar{\mathbf{u}}^h)^\top (\mathbf{u}^\circ - \bar{\mathbf{u}}^h).\end{aligned}$$

Notice that $\min_j \{b_j - \mathbf{y}^{\circ\, h}(x_j)\} > \min_j \{b_j - \mathbf{y}^\circ(x_j)\} - c h^2 |\log(h)| > \gamma_1 > 0$ for some positive γ_1 and h sufficiently small. Moreover, taking into account that $\gamma_2 := \min_i (u_{bi} - \mathbf{u}_i^\circ + \mathbf{u}_{i+M}^\circ) > 0$, $\gamma_3 := \min_i (\mathbf{u}_i^\circ - \mathbf{u}_{i+M}^\circ - u_{ai}) > 0$ and that $|\boldsymbol{\eta}^h| \le \beta$,

together with the nonnegativity of the multipliers, it follows that

$$\gamma_1 \sum_{j=1}^{\ell} |\boldsymbol{\nu}_j^h| + \gamma_2 \sum_{i=1}^{M} |\boldsymbol{\mu}_{b\,i}^h| + \gamma_3 \sum_{i=1}^{M} |\boldsymbol{\mu}_{a\,i}^h| \leq \nabla \mathbf{f}^{\mathbf{h}}(\bar{\mathbf{u}}^h)^\top (\mathbf{u}^\circ - \bar{\mathbf{u}}^h).$$

The right–hand side in the last inequality is bounded because of the boundedness of $\bar{\mathbf{u}}^h$ and its associated state. Thus, the last inequality just implies that the multipliers are bounded. ∎

Remark 5.4. The optimality systems (4.18) and (4.9) are obtained by the standard Lagrange principle, i.e. the corresponding optimal vectors satisfy:

$$\begin{cases} \nabla_u \mathcal{L}(\bar{\mathbf{u}}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) = 0, \\ \mathbf{g}(\bar{\mathbf{u}}) \leq_K 0, \\ (\boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta})^\top \mathbf{g}(\bar{\mathbf{u}}) = 0, \end{cases} \quad \text{and} \quad \begin{cases} \nabla_u \mathcal{L}^h(\bar{\mathbf{u}}^h, \boldsymbol{\nu}^h, \boldsymbol{\mu}_a^h, \boldsymbol{\mu}_b^h, \boldsymbol{\eta}^h) = 0, \\ \mathbf{g}^{\mathbf{h}}(\bar{\mathbf{u}}^h) \leq_K 0, \\ (\boldsymbol{\nu}^h, \boldsymbol{\mu}_a^h, \boldsymbol{\mu}_b^h, \boldsymbol{\eta}^h)^\top \mathbf{g}^{\mathbf{h}}(\bar{\mathbf{u}}^h) = 0. \end{cases}$$

By introducing the *normal cone* of K:

$$\partial \psi_K(x) = \begin{cases} z \in \mathbb{R}^\ell \times \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^{2M} \text{ with } z^T(w-x) \leq 0 \quad \forall w \in K, & \text{if } x \in K, \\ \emptyset, & \text{if } x \notin K, \end{cases}$$

these optimality systems can be rewritten as the generalized equations. Indeed, following [29] we have:

$$0 \in \begin{bmatrix} \nabla_u \mathcal{L}(\bar{\mathbf{u}}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) \\ -\mathbf{g}(\bar{\mathbf{u}}) \end{bmatrix} + \begin{bmatrix} \{0\} \\ \partial \psi_K(\boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) \end{bmatrix} \tag{5.3}$$

and

$$0 \in \begin{bmatrix} \nabla_u \mathcal{L}^h(\bar{\mathbf{u}}^h, \boldsymbol{\nu}^h, \boldsymbol{\mu}_a{}^h, \boldsymbol{\mu}_b{}^h, \boldsymbol{\eta}^h) \\ -\mathbf{g}^{\mathbf{h}}(\bar{\mathbf{u}}^h) \end{bmatrix} + \begin{bmatrix} \{0\} \\ \partial \psi_K(\boldsymbol{\nu}^h, \boldsymbol{\mu}_a{}^h, \boldsymbol{\mu}_b{}^h, \boldsymbol{\eta}^h) \end{bmatrix} \tag{5.4}$$

respectively. Moreover, according to [25][Section 5.3] the solutions of (5.3) and (5.4) can be viewed as solutions of a perturbed problem. Therefore, we consider the perturbed problem

$$0 \in \begin{bmatrix} \nabla_u \mathcal{L}(\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) \\ -\mathbf{g}(\mathbf{u}) \end{bmatrix} + \begin{bmatrix} \{0\} \\ \partial \psi_K(\boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) \end{bmatrix} - \delta. \tag{5.5}$$

It is clear that we can express (5.3) in terms of (5.5) by taking $(\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) = (\bar{\mathbf{u}}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}})$ and $\delta = 0$. For (5.4) we take $(\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) = (\bar{\mathbf{u}}^h, \bar{\boldsymbol{\nu}}^h, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}^h, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}^h, \bar{\boldsymbol{\eta}}^h)$ with associated perturbation $\delta(h)$ defined by:

$$\delta(h) = \begin{cases} 0 \quad \text{if } h = 0 \\ \begin{bmatrix} \nabla_u \mathcal{L}(\bar{\mathbf{u}}^h, \bar{\boldsymbol{\nu}}^h, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}^h, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}^h, \bar{\boldsymbol{\eta}}^h) - \nabla_u \mathcal{L}^h(\bar{\mathbf{u}}^h, \bar{\boldsymbol{\nu}}^h, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}^h, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}^h, \bar{\boldsymbol{\eta}}^h) \\ \mathbf{g}(\bar{\mathbf{u}}^h) - \mathbf{g}^{\mathbf{h}}(\bar{\mathbf{u}}^h) \end{bmatrix}, \text{ if } h \neq 0. \end{cases} \tag{5.6}$$

At this point, we have all the required elements to proof the error estimate for the approximation of the optimal control. Observe that by Lemma 4.3 we do not need strong regularity as in [25].

Proof of Theorem 3.3 .

Non optimal estimate. Consider the slater point defined in (2.6). In view of the state constraints, we have that $\bar{\mathbf{u}}$ does not need to be feasible for $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$ neither $\bar{\mathbf{u}}^h$ need to be feasible for $(\mathbf{P_{2M}})$. Therefore, we define the intermediate vectors

$$\begin{aligned}\tilde{\mathbf{u}} &= (1-t)\bar{\mathbf{u}}^h + t\mathbf{u}^\circ, \qquad \text{for } t \in (0,1), \text{and}\\ \hat{\mathbf{u}} &= (1-t)\bar{\mathbf{u}} + t\mathbf{u}^\circ, \qquad \text{for } t \in (0,1),\end{aligned}$$

where $\mathbf{u}^\circ$ is defined by $\mathbf{u}^\circ_i = u_i^{\circ +}$ and $\mathbf{u}^\circ_{i+M} = u_i^{\circ -}$, for $i = 1, \dots, M$. By convexity, $\tilde{\mathbf{u}}$ is feasible for $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$ and so does $\hat{\mathbf{u}}$ for problem $(\mathbf{P_{2M}})$.

By defining $\varrho := \max_{j=1,\dots,\ell} \sum_{i=1}^{M} u_i^\circ y_i(x_j) < b_j$, then $\varrho - b_j < 0$. Moreover, since $\bar{\mathbf{u}}^h$ satisfies the state constraints for the discrete problem and also satisfies the estimate (3.2b) then, for h sufficiently small, $\tilde{\mathbf{u}}$ fulfills

$$\begin{aligned}\sum_{i=1}^{M}(\tilde{\mathbf{u}}_i - \tilde{\mathbf{u}}_{i+M})y_i(x_j) &= (1-t)\sum_{i=1}^{M}(\bar{\mathbf{u}}^h_i - \bar{\mathbf{u}}^h_{i+M})y_i(x_j) + t\sum_{i=1}^{M}(\mathbf{u}^\circ_i - \mathbf{u}^\circ_{i+M})y_i(x_j)\\ &< (1-t)\sum_{i=1}^{M}(\bar{\mathbf{u}}^h_i - \bar{\mathbf{u}}^h_{i+M})(y_i(x_j) - y^h_i(x_j)) + (1-t)b_j + t\varrho\\ &< ch^2|\log(h)|\sum_{i=1}^{M}|\bar{\mathbf{u}}^h_i - \bar{\mathbf{u}}^h_{i+M}| + t(\varrho - b_j) + b_j\\ &< ch^2|\log(h)|C_{ad} + t(\varrho - b_j) + b_j,\end{aligned}$$

where C_{ad} is a positive bound for the admissible set U_{ad} not depending on h. Here, we realize that the right–hand side is less than b_j by selecting

$$t = \frac{-c}{\varrho - b_j}h^2|\log(h)|C_{ad} > 0. \tag{5.7}$$

Therefore, $\tilde{\mathbf{u}}$ is feasible for problems $(\mathbf{P_{2M}})$ and $(\mathbf{P}^{\mathbf{h}}_{\mathbf{2M}})$. The same conclusion is obtained for the control for $\hat{\mathbf{u}}$ with the same arguments. In view of feasibility of $\tilde{\mathbf{u}}$ and $\hat{\mathbf{u}}$, and the optimality of $\bar{\mathbf{u}}$ and $\bar{\mathbf{u}}^h$, we have

$$\nabla\mathbf{f}(\bar{\mathbf{u}})^\top(\tilde{\mathbf{u}} - \bar{\mathbf{u}}) \geq 0 \quad \text{and} \quad \nabla\mathbf{f}^{\mathbf{h}}(\bar{\mathbf{u}}^h)^\top(\hat{\mathbf{u}} - \bar{\mathbf{u}}^h) \geq 0 \tag{5.8}$$

By adding both inequalities in (5.8), and by replacing the corresponding gradients of $\mathbf{f}$ and $\mathbf{f}^{\mathbf{h}}$ as well as the definition of $\hat{\mathbf{u}}$ and $\tilde{\mathbf{u}}$, we get that

$$\begin{aligned}0 \leq\ & \bar{\mathbf{u}}^\top(\mathcal{Y} + \alpha I)(t(\mathbf{u}^\circ - \bar{\mathbf{u}}) + (1-t)(\bar{\mathbf{u}}^h - \bar{\mathbf{u}})) + \bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h + \alpha I)(t(\mathbf{u}^\circ - \bar{\mathbf{u}}^h) + (1-t)(\bar{\mathbf{u}} - \bar{\mathbf{u}}^h))\\ &+\mathbf{b}^\top\big(t(\mathbf{u}^\circ - \bar{\mathbf{u}}) + (1-t)(\bar{\mathbf{u}}^h - \bar{\mathbf{u}})\big) + \mathbf{b}^{h\top}(t(\mathbf{u}^\circ - \bar{\mathbf{u}}^h) + (1-t)(\bar{\mathbf{u}} - \bar{\mathbf{u}}^h))\\ &+\beta\sum_{i=1}^{2M}(t(\mathbf{u}^\circ_i - \bar{\mathbf{u}}_i) + (1-t)(\bar{\mathbf{u}}^h_i - \bar{\mathbf{u}}_i)) + \beta\sum_{i=1}^{2M}(t(\mathbf{u}^\circ_i - \bar{\mathbf{u}}^h_i) + (1-t)(\bar{\mathbf{u}}_i - \bar{\mathbf{u}}^h_i))\\ =\ & t\Big[\bar{\mathbf{u}}^\top(\mathcal{Y} + \alpha I)(\mathbf{u}^\circ - \bar{\mathbf{u}}) + \bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h + \alpha I)(\mathbf{u}^\circ - \bar{\mathbf{u}}^h) + \mathbf{b}^\top(\mathbf{u}^\circ - \bar{\mathbf{u}}) + \mathbf{b}^{h\top}(\mathbf{u}^\circ - \bar{\mathbf{u}}^h)\\ &+\beta\sum_{i=1}^{2M}(\mathbf{u}^\circ_i - \bar{\mathbf{u}}_i) + \beta\sum_{i=1}^{2M}(\mathbf{u}^\circ_i - \bar{\mathbf{u}}^h_i)\Big]\\ &+(1-t)\Big[\bar{\mathbf{u}}^\top(\mathcal{Y} + \alpha I)(\bar{\mathbf{u}}^h - \bar{\mathbf{u}}) + \bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h + \alpha I)(\bar{\mathbf{u}} - \bar{\mathbf{u}}^h) + (\mathbf{b} - \mathbf{b}^h)(\bar{\mathbf{u}}^h - \bar{\mathbf{u}})\Big].\end{aligned}$$

By rearranging terms, the last inequality is equivalent to

$$
\begin{aligned}
&(1-t)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})^\top(\mathcal{Y}+\alpha I)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})\\
&\leq t\Big[\bar{\mathbf{u}}^\top(\mathcal{Y}+\alpha I)(\mathbf{u}^\circ-\bar{\mathbf{u}})+\bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h+\alpha I)(\mathbf{u}^\circ-\bar{\mathbf{u}}^h)+\mathbf{b}^\top(\mathbf{u}^\circ-\bar{\mathbf{u}})+{\mathbf{b}^h}^\top(\mathbf{u}^\circ-\bar{\mathbf{u}}^h)\\
&+\beta\sum_{i=1}^{2M}(\mathbf{u}_i^\circ-\bar{\mathbf{u}}_i)+(\mathbf{u}_i^\circ-\bar{\mathbf{u}}_i^h)\Big]+(1-t)\Big[\bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h-\mathcal{Y})(\bar{\mathbf{u}}-\bar{\mathbf{u}}^h)+(\mathbf{b}-\mathbf{b}^h)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})\Big]\\
&=t\Big[\nabla\mathbf{f}(\bar{\mathbf{u}})^\top(\mathbf{u}^\circ-\bar{\mathbf{u}})-\nabla\mathbf{f}^{\mathbf{h}}(\bar{\mathbf{u}}^h)^\top(\bar{\mathbf{u}}^h-\mathbf{u}^\circ)\Big]\\
&+(1-t)\Big[\bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h-\mathcal{Y})(\bar{\mathbf{u}}-\bar{\mathbf{u}}^h)+(\mathbf{b}-\mathbf{b}^h)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})\Big].
\end{aligned}
$$

Now, by applying Lemma 5.1, we estimate the right–hand side as follows

$$
\begin{aligned}
&(1-t)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})^\top(\mathcal{Y}+\alpha I)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})\\
&\leq t\Big[\nabla\mathbf{f}(\bar{\mathbf{u}})^\top(\mathbf{u}^\circ-\bar{\mathbf{u}})-\nabla\mathbf{f}^{\mathbf{h}}(\bar{\mathbf{u}}^h)^\top(\bar{\mathbf{u}}^h-\mathbf{u}^\circ)\Big]+(1-t)\Big[\bar{\mathbf{u}}^{h\top}(\mathcal{Y}^h-\mathcal{Y})(\bar{\mathbf{u}}-\bar{\mathbf{u}}^h)\\
&\qquad+(\mathbf{b}-\mathbf{b}^h)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})\Big]\\
&\leq t\,C(\|2\mathbf{u}^\circ-\bar{\mathbf{u}}-\bar{\mathbf{u}}^h\|+h^2|\log(h)|)+(1-t)\big(\|\bar{\mathbf{u}}^h\|\|\mathcal{Y}^h-\mathcal{Y}\|+\|\mathbf{b}-\mathbf{b}^h\|\big)\|\bar{\mathbf{u}}^h-\bar{\mathbf{u}}\|.
\end{aligned}
$$

From Lemma 5.2, one gets

$$
\begin{aligned}
&(1-t)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})^\top(\mathcal{Y}+\alpha I)(\bar{\mathbf{u}}^h-\bar{\mathbf{u}})\\
&\qquad\leq t(\tilde{C}_{ad}+h^2|\log(h)|)+\big(\|\bar{\mathbf{u}}^h\|\|\mathcal{Y}^h-\mathcal{Y}\|+\|\mathbf{b}-\mathbf{b}^h\|\big)\|\bar{\mathbf{u}}^h-\bar{\mathbf{u}}\|\\
&\qquad\leq t(\tilde{C}_{ad}+h^2|\log(h)|)+\tilde{C}_{ad}h^2,
\end{aligned}
$$

where $\tilde{C}_{ad}>0$ is a generic constant independent of h. Therefore, thanks to Lemma 5.3, the estimate (5.7) and the positive definiteness of the matrix $\mathcal{Y}+\alpha I$ we can conclude the existence of a positive constant C, such that

$$
\|\bar{\mathbf{u}}^h-\bar{\mathbf{u}}\|\leq C\sqrt{h^2|\log(h)|}\to 0,\quad \text{whenever } h\to 0. \tag{5.9}
$$

Also, in view of (5.9), we notice that $\bar{\mathbf{y}}^h\to\bar{\mathbf{y}}$ as $h\to 0$.

Improved estimate. From our discussion in Remark 5.4 we know that optimality conditions can be represented as perturbed generalized equation depending on the perturbation $\delta(h)$ given in (5.6). We do not apply directly the results in [25] since there, the LICQ constraint qualifications and strong regularity were required. However, we follow their general ideas.

Using the fact that the constraints are regular at $\bar{\mathbf{u}}$ (in the sense given in Lemma 4.3) and the properties of the objective and constraint functions we are able to satisfy the stability result given in [30][Theorem 4.2]

Let us begin by considering $\delta(h)$ defined by (5.6) for $h\neq 0$. It is clear that Lemma 5.1 implies that the second component of $\delta(h)$ obeys $\|\mathbf{g}^{\mathbf{h}}(\bar{\mathbf{u}}^h)-\mathbf{g}(\bar{\mathbf{u}}^h)\|\leq Ch^2|\log(h)|$. On the other hand, a direct computation of the first components gives

$$
\begin{aligned}
&\|\nabla_{\mathbf{u}}\mathcal{L}^h(\bar{\mathbf{u}}^h,\bar{\boldsymbol{\nu}}^h,\bar{\boldsymbol{\mu}}_{\boldsymbol{a}}^h,\bar{\boldsymbol{\mu}}_{\boldsymbol{b}}^h,\bar{\boldsymbol{\eta}}^h)-\nabla_{\mathbf{u}}\mathcal{L}(\bar{\mathbf{u}}^h,\bar{\boldsymbol{\nu}}^h,\bar{\boldsymbol{\mu}}_{\boldsymbol{a}}^h,\bar{\boldsymbol{\mu}}_{\boldsymbol{b}}^h,\bar{\boldsymbol{\eta}}^h)\|\\
&\leq\|\nabla\mathbf{f}^{\mathbf{h}}(\bar{\mathbf{u}}^h)-\nabla\mathbf{f}(\bar{\mathbf{u}}^h)\|+\|\sum_{j=1}^{\ell}\bar{\boldsymbol{\nu}}_j^h y_i^h(x_j)-\sum_{j=1}^{\ell}\bar{\boldsymbol{\nu}}_j^h y_i(x_j)\|\leq Ch^2|\log(h)|,
\end{aligned}
$$

where the last bound follow from Lemma 5.1 and estimate (3.2c). In particular, this implies that the perturbation $\delta(h)$ is of order $h^2|\log(h)|$.

As stablished in Lemma (4.3), our constraints, considered in the polihedral cone $\mathbb{R}^{\ell+4M}_+$, are regular. Moreover, the second order sufficient conditions are satisfied in view of the positive definiteness of the matrix $\mathcal{L}''(\bar{\mathbf{u}}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}}) = \mathbf{f}''(\bar{\mathbf{u}}) = \mathcal{Y} + \alpha I$. Hence, we fulfill the hypothesis of Theorem 4.2 in [30] for the function $\mathcal{F} : \mathbb{R}^{2M} \times \mathbb{R}^{\ell+4M} \times (\mathcal{U}_{feas} \times \mathbb{R}^{\ell} \times \mathbb{R}^{M} \times \mathbb{R}^{M} \times \mathbb{R}^{2M}) \to \mathbb{R}^{2M} \times \mathbb{R}^{\ell} \times \mathbb{R}^{M} \times \mathbb{R}^{M} \times \mathbb{R}^{2M}$ given by

$$\mathcal{F}(\delta, (\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta})) = \begin{bmatrix} \nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_a, \boldsymbol{\mu}_b, \boldsymbol{\eta}) \\ -\mathbf{g}(\mathbf{u}) \end{bmatrix} - \delta. \tag{5.10}$$

Using the following definitions from [30]:

$$\begin{aligned} \boldsymbol{\Lambda}(\mathbf{u}, \delta) &:= \{(\boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta}) : (\mathbf{u}, (\boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta}), \delta) \text{ satisfying (5.5)}\}, \\ \boldsymbol{\Lambda}_0 &:= \boldsymbol{\Lambda}(\bar{\mathbf{u}}, 0), \\ \mathcal{SP}(\delta) &:= \{\mathbf{u} \in U : (\mathbf{u}, (\boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta}), \delta) \text{ satisfying (5.5), for some } (\boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta})\}, \end{aligned}$$

Here U is a neighborhood of $\bar{\mathbf{u}}$. We have as a consequence of [30][Theorem 4.2] that there exists a positive constant c, such that the inequality

$$\begin{aligned} \operatorname{dist}[(\mathbf{u}, (\boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta})), \bar{\mathbf{u}} \times \boldsymbol{\Lambda}_0] &\leq c\|\mathcal{F}(\delta, (\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta})) - \mathcal{F}(0, (\mathbf{u}, \boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta}))\| \\ &= c\|\delta\| \end{aligned} \tag{5.11}$$

is satisfied for δ sufficiently close to 0 and for all $(\mathbf{u}, (\boldsymbol{\nu}, \boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\eta})) \in \mathcal{SP}(\delta) \times \boldsymbol{\Lambda}(\mathbf{u}, \delta)$.

Since $\delta(h) \to 0$ and $\bar{\mathbf{u}}^h \in U$ for $h \to 0$, and $(\bar{\mathbf{u}}^h, (\boldsymbol{\nu}^h, \boldsymbol{\mu}^h_{\boldsymbol{a}}, \boldsymbol{\mu}^h_{\boldsymbol{b}}, \boldsymbol{\eta}^h), \delta(h))$ satisfies (5.5), in particular (5.11) implies that

$$\|\bar{\mathbf{u}}^h - \bar{\mathbf{u}}\| \leq ch^2|\log(h)|.$$

Finally, by Lemma 4.2, we know that $\bar{u} = [\bar{\mathbf{u}}_i]_{i=1}^M - [\bar{\mathbf{u}}_{i+M}]_{i=1}^M$ and $\bar{u}^h = [\bar{\mathbf{u}}^h_i]_{i=1}^M - [\bar{\mathbf{u}}^h_{i+M}]_{i=1}^M$. Hence, the estimate

$$\begin{aligned} ||\bar{u} - \bar{u}^h|| &= || \left(\left[\bar{\mathbf{u}}_i\right]_{i=1}^M - \left[\bar{\mathbf{u}}_{i+M}\right]_{i=1}^M \right) - \left(\left[\bar{\mathbf{u}}^h_i\right]_{i=1}^M - \left[\bar{\mathbf{u}}^h_{i+M}\right]_{i=1}^M \right) || \\ &\leq C||\bar{\mathbf{u}} - \bar{\mathbf{u}}^h|| \\ &\leq Ch^2|\log h|, \end{aligned}$$

holds for $h > 0$ small enough.

6. Comments on the case $\alpha = 0$. We have discussed so far the case when the Tikhonov regularization term is present in the cost function. Here, we briefly discuss the case $\alpha = 0$. We will denote this problem by $(\mathbf{P^0_M})$. Several results are analogous to the previous section, therefore, their proofs will not be given. If $\alpha = 0$, the lack of strict convexity of the cost function prevents us to fulfill second order sufficient conditions which are needed in our analysis. Therefore, we make the following assumption.

Hypothesis 6.1. We assume that the matrix Y is positive definite.

Remark 6.2. Note that this assumption can be satisfied for instance if the set of states $\{y_i \in H^1_0(\Omega) : i = 1, \dots, M\}$ is orthonormal. For example, we may choose finitely many normalized eigenfunctions of the Laplace operator (i.e. $A = -\Delta$).

We also notice that this assumption implies the uniqueness of the solution for problem $(\mathbf{P^0_M})$.

First order necessary optimality conditions for $(\mathbf{P^0_M})$ are obtained analogously.

THEOREM 6.3. *Under assumption 2.7, there exists* $(\nu, \mu_a, \mu_b) \in \mathbb{R}^\ell_+ \times \mathbb{R}^M_+ \times \mathbb{R}^M_+ \cup \{0\}$ *such that for all* $i = 1, \dots, M$, *it follows that*

$$\int_\Omega (\bar{y} - y_d) y_i \, dx + \sum_{j=1}^{\ell} \nu_j y_i(x_j) + \mu_{b_i} - \mu_{a_i} = -\beta, \qquad \text{if } \bar{u}_i > 0, \tag{6.1a}$$

$$\int_\Omega (\bar{y} - y_d) y_i \, dx + \sum_{j=1}^{\ell} \nu_j y_i(x_j) + \mu_{b_i} - \mu_{a_i} = \beta, \qquad \text{if } \bar{u}_i < 0, \tag{6.1b}$$

$$\left| \int_\Omega (\bar{y} - y_d) y_i \, dx + \sum_{j=1}^{\ell} \nu_j y_i(x_j) + \mu_{b_i} - \mu_{a_i} \right| \leq \beta, \qquad \text{if } \bar{u}_i = 0, \tag{6.1c}$$

moreover, the following conditions are satisfied:

$$\nu \geq 0, \quad \text{and} \quad \nu_j(\bar{y}(x_j) - b_j) = 0, \tag{6.2a}$$

$$\mu_a \geq 0, \quad \text{and} \quad \mu_{a_i}(u_{a,i} - \bar{u}_i) = 0, \tag{6.2b}$$

$$\mu_b \geq 0, \quad \text{and} \quad \mu_{b_i}(\bar{u}_i - u_{b,i}) = 0. \tag{6.2c}$$

Again, by using (4.2) we reformulate problem for the case $\alpha = 0$ in $\mathbb{R}^{2M}$ as follows:

$$(\mathbf{P^0_{2M}}) \begin{cases} \min\limits_{(y,\mathbf{u})} \mathbf{J}(y, \mathbf{u}) = \dfrac{1}{2} \displaystyle\int_\Omega (y - y_d)^2 \, dx + \beta \sum_{i=1}^{2M} \mathbf{u}_i \\ \text{subject to:} \\ \quad Ay(x) = \displaystyle\sum_{i=1}^{M} (\mathbf{u}_i - \mathbf{u}_{i+M}) e_i(x), \quad \text{in } \Omega, \\ \quad y(x) = 0, \quad \text{on } \Gamma, \\ \quad y(x_j) \leq b_j, \quad \forall j = 1, \dots, \ell, \\ \quad [\mathbf{u}_i - \mathbf{u}_{i+M}]_{i=1}^{M} \in \mathcal{U}_{ad}, \\ \quad \mathbf{u}_i \geq 0, \quad \forall i = 1, \dots, 2M. \end{cases}$$

In turn, we deduce the corresponding first order optimality conditions in the same fashion as in Theorem 4.4.

THEOREM 6.4. *Under Assumption 2.7, there exists* $(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}}) \in \mathbb{R}^\ell_+ \times \mathbb{R}^M_+ \times \mathbb{R}^M_+ \times \mathbb{R}^{2M}_+ \cup \{0\}$ *such that the conditions below are satisfied*

$$(\bar{\mathbf{y}} - y_d, y_i) + \beta + \sum_{j=1}^{\ell} \boldsymbol{\nu}_j y_i(x_j) + \boldsymbol{\mu}_{b_i} - \boldsymbol{\mu}_{a_i} - \boldsymbol{\eta}_i = 0, \tag{6.3a}$$

$$-(\bar{\mathbf{y}} - y_d, y_i) + \beta - \sum_{j=1}^{\ell} \boldsymbol{\nu}_j y_i(x_j) - \boldsymbol{\mu}_{b_i} + \boldsymbol{\mu}_{a_i} - \boldsymbol{\eta}_{i+M} = 0. \tag{6.3b}$$

for $i = 1, \dots, M$, holds, together with the following slackness conditions

$$\bar{\mathbf{u}} \geq 0, \tag{6.4a}$$

$$u_{ai} \leq \bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i+M} \leq u_{bi}, \quad \forall i = 1, \dots, 2M. \tag{6.4b}$$

$$\boldsymbol{\nu}_j \left(\bar{\mathbf{y}}(x_j) - b_j\right) = 0, \quad \forall j = 1, \dots, \ell, \tag{6.4c}$$

$$\boldsymbol{\mu}_{a\,i}(u_{a,i} - \bar{\mathbf{u}}_i + \bar{\mathbf{u}}_{i+M}) = 0, \quad \forall i = 1, \dots, M, \tag{6.4d}$$

$$\boldsymbol{\mu}_{b\,i}(\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_{i+M} - u_{b,i}) = 0, \quad \forall i = 1, \dots, M, \tag{6.4e}$$

$$\boldsymbol{\eta}_i \bar{\mathbf{u}}_i = 0, \quad \forall i = 1, \dots, 2M. \tag{6.4f}$$

Following the same procedure of discretization using the finite element method we formulate the associated discretized problem, using the superscript h to indicate the corresponding discretized variables. Similar to Theorem 4.4, we conclude that there exists $(\bar{\boldsymbol{\nu}}^h, \bar{\boldsymbol{\mu}}^h_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}^h_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}}^h) \in \mathbb{R}^\ell_+ \times \mathbb{R}^M_+ \times \mathbb{R}^M_+ \times \mathbb{R}^{2M}_+ \cup \{0\}$, such that

$$(\bar{\mathbf{y}}^h - y_d, y_i^h) + \beta + \sum_{j=1}^{\ell} \boldsymbol{\nu}_j^h y_i^h(x_j) + \boldsymbol{\mu}^h_{b\,i} - \boldsymbol{\mu}^h_{a\,i} - \boldsymbol{\eta}^h_i = 0,$$

$$-(\bar{\mathbf{y}}^h - y_d, y_i^h) + \beta - \sum_{j=1}^{\ell} \boldsymbol{\nu}_j^h y_i^h(x_j) - \boldsymbol{\mu}^h_{b\,i} + \boldsymbol{\mu}^h_{a\,i} - \boldsymbol{\eta}^h_{i+M} = 0.$$

for $i = 1, \dots, M$, together with the following conditions

$$\bar{\mathbf{u}}^h \geq 0,$$

$$u_{ai} \leq \bar{\mathbf{u}}^h_i - \bar{\mathbf{u}}^h_{i+M} \leq u_{bi}, \quad \forall i = 1, \dots, 2M.$$

$$\boldsymbol{\nu}_j^h \left(\bar{\mathbf{y}}^h(x_j) - b_j\right) = 0, \quad \forall j = 1, \dots, \ell,$$

$$\boldsymbol{\mu}^h_{a\,i}(u_{a,i} - \bar{\mathbf{u}}^h_i + \bar{\mathbf{u}}^h_{i+M}) = 0, \quad \forall i = 1, \dots, M,$$

$$\boldsymbol{\mu}^h_{b\,i}(\bar{\mathbf{u}}^h_i - \bar{\mathbf{u}}^h_{i+M} - u_{b,i}) = 0, \quad \forall i = 1, \dots, M,$$

$$\boldsymbol{\eta}^h_i \bar{\mathbf{u}}^h_i = 0, \quad \forall i = 1, \dots, 2M.$$

As mentioned earlier, second order sufficient conditions must be satisfied at the optimum. These conditions are given in [29][Definition 2.1] which, after the computation of the tangent cone at $\mathbf{g}(\bar{\mathbf{u}})$, are formulated as follows:

Let $\bar{\mathbf{u}} \in \mathbb{R}^{2M}$ a feasible point for $(\mathbf{P^0_{2M}})$ and $(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}}) \in \mathbb{R}^\ell_+ \times \mathbb{R}^M_+ \times \mathbb{R}^M_+ \times \mathbb{R}^{2M}_+ \cup \{0\}$ satisfying the first order necessary conditions (6.1) and (6.2). The *second–order sufficient conditions* are satisfied at $\bar{\mathbf{u}}$ with multipliers $(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}})$ if for each $\mathbf{v} \in \mathbb{R}^{2M}$ such that

$$\nabla \mathbf{g}_j(\bar{\mathbf{u}})^\top \mathbf{v} \leq 0 \quad \text{for } j \in \{1, \dots, \ell + 4M : \mathbf{g}_j(\bar{\mathbf{u}}) = 0\} \quad \text{and} \quad \nabla \mathbf{f}(\bar{\mathbf{u}})^\top \mathbf{v} = 0, \tag{6.7}$$

the quadratic growth condition:

$$\mathbf{v}^\top \mathcal{L}''(\bar{\mathbf{u}}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}})\mathbf{v} \geq c\|\mathbf{v}\|^2, \tag{6.8}$$

holds, for some constant $c > 0$.

LEMMA 6.5. *Under Hypothesis 6.1 the unique solution $\bar{\mathbf{u}}$ for problem $(\mathbf{P^0_M})$ satisfies the second–order sufficient conditions given by (6.7) and (6.8).*

Proof. Let $\mathbf{v} \in \mathbb{R}^{2M}$ be a vector satisfying condition (6.7). By using the first–order optimality conditions (6.3), we have that

$$\nabla \mathbf{f}(\bar{\mathbf{u}})^\top \mathbf{v} + \sum_{j=1}^{\ell} \bar{\boldsymbol{\nu}}_j y_{\mathbf{v}}(x_j) + \sum_{i=1}^{M} \bar{\boldsymbol{\mu}}_{\boldsymbol{b}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) - \sum_{i=1}^{M} \bar{\boldsymbol{\mu}}_{\boldsymbol{a}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) - \sum_{i=1}^{2M} \bar{\boldsymbol{\eta}}_i \mathbf{v}_i = 0,$$

where $y_{\mathbf{v}} = \sum_{i=1}^{M} (\mathbf{v}_i - \mathbf{v}_{i+M}) y_i$. Moreover, by taking into account (6.7), we replace the identity $\nabla \mathbf{f}(\bar{\mathbf{u}})^\top \mathbf{v} = 0$ in the previous equation, which leads to

$$\sum_{j=1}^{\ell} \bar{\boldsymbol{\nu}}_j y_{\mathbf{v}}(x_j) + \sum_{i=1}^{M} \bar{\boldsymbol{\mu}}_{\boldsymbol{b}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) - \sum_{i=1}^{M} \bar{\boldsymbol{\mu}}_{\boldsymbol{a}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) - \sum_{i=1}^{2M} \bar{\boldsymbol{\eta}}_i \mathbf{v}_i = 0. \tag{6.9}$$

From the complementarity slackness conditions (6.4c)-(6.4f) we know that if $\mathbf{g}_j(\bar{\mathbf{u}}) \neq 0$ for some $j \in \{1, \ldots, \ell + 4M\}$, thus the corresponding multiplier vanishes. Hence, we have that the sums on expression (6.9) are meaningful only on those j's for which $\mathbf{g}_j(\bar{\mathbf{u}}) = 0$. In view of this observation, we have that (6.9) becomes

$$\begin{aligned} \sum_{\substack{j \in \{1,\ldots,\ell\} \\ \mathbf{g}_j(\bar{\mathbf{u}})=0}} \bar{\boldsymbol{\nu}}_j y_{\mathbf{v}}(x_j) + \sum_{\substack{i \in \{1,\ldots,M\} \\ \mathbf{g}_{\ell+i}(\bar{\mathbf{u}})=0}} \bar{\boldsymbol{\mu}}_{\boldsymbol{b}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) - \sum_{\substack{i \in \{1,\ldots,M\} \\ \mathbf{g}_{\ell+M+i}(\bar{\mathbf{u}})=0}} \bar{\boldsymbol{\mu}}_{\boldsymbol{a}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) \\ - \sum_{\substack{i \in \{1,\ldots,2M\} \\ \bar{\mathbf{u}}_i=0}} \bar{\boldsymbol{\eta}}_i \mathbf{v}_i = 0, \end{aligned} \tag{6.10}$$

On the other hand, since the Lagrange multipliers are nonnegative and (6.7) holds, we have

$$\begin{aligned} \bar{\boldsymbol{\nu}}_j y_{\mathbf{v}}(x_j) \leq 0, \text{ for all } j \in \{j \in \{1, \ldots, \ell\} : \mathbf{g}_j(\bar{\mathbf{u}}) = 0\}, \\ \bar{\boldsymbol{\mu}}_{\boldsymbol{b}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) \leq 0, \text{ for all } i \in \{i \in \{1, \ldots, M\} : \mathbf{g}_{\ell+i}(\bar{\mathbf{u}}) = 0\}, \\ -\bar{\boldsymbol{\mu}}_{\boldsymbol{b}_i}(\mathbf{v}_i - \mathbf{v}_{i+M}) \leq 0, \text{ for all } i \in \{i \in \{1, \ldots, M\} : \mathbf{g}_{\ell+M+i}(\bar{\mathbf{u}}) = 0\}, \\ -\bar{\boldsymbol{\eta}}_i \mathbf{v}_i \leq 0, \text{ for all } i \in \{i \in \{1, \ldots, 2M\} : \bar{\mathbf{u}}_i = 0\}. \end{aligned}$$

From these relations, we conclude that all terms in (6.10) vanish. In particular we get

$$\bar{\boldsymbol{\eta}}_i \mathbf{v}_i = 0, \text{ for all } i \in \{i \in \{1, \ldots, 2M\} : \bar{\mathbf{u}}_i = 0\}. \tag{6.11}$$

The first order optimality conditions (6.1) and (6.2) imply that $0 < \beta = \frac{\bar{\boldsymbol{\eta}}_i + \bar{\boldsymbol{\eta}}_{i+M}}{2}$ for all $i = 1, \ldots, M$. We are going to show that $\mathbf{v}_i \mathbf{v}_{i+M} = 0$ for all $i = 1, \ldots, M$. Indeed, using the fact that $\bar{\mathbf{u}}_i \bar{\mathbf{u}}_{i+M} = 0$ we analyze the following cases:

(i) If $\bar{\mathbf{u}}_i = 0$ and $\bar{\mathbf{u}}_{i+M} > 0$, from the complementarity slackness conditions we have $\bar{\boldsymbol{\eta}}_{i+M} \bar{\mathbf{u}}_{i+M} = 0$, therefore $\bar{\boldsymbol{\eta}}_{i+M} = 0$ which implies $\bar{\boldsymbol{\eta}}_i > 0$, by (6.11) it follows that $\mathbf{v}_i = 0$.

(ii) If $\bar{\mathbf{u}}_i > 0$ and $\bar{\mathbf{u}}_{i+M} = 0$ we get that $\mathbf{v}_{i+M} = 0$. The arguments are analogous to the case (i).

(iii) If $\bar{\mathbf{u}}_i = 0$ and $\bar{\mathbf{u}}_{i+M} = 0$, again, from the complementarity slackness conditions we have $\bar{\boldsymbol{\eta}}_i \bar{\mathbf{u}}_i = 0$ and $\bar{\boldsymbol{\eta}}_{i+M} \bar{\mathbf{u}}_{i+M} = 0$, and since $\bar{\boldsymbol{\eta}}_i + \bar{\boldsymbol{\eta}}_{i+M} > 0$ we get $\bar{\boldsymbol{\eta}}_i > 0$ or $\bar{\boldsymbol{\eta}}_{i+M} > 0$ implying that $\mathbf{v}_i = 0$ or $\mathbf{v}_{i+M} = 0$. Thus $\mathbf{v}_i \mathbf{v}_{i+M} = 0$.

We have shown that $\mathbf{v}_i\mathbf{v}_{i+M} = 0$ for all $i = 1, \dots, M$, therefore

$$\begin{aligned}\mathbf{v}^\top \mathcal{L}''(\bar{\mathbf{u}}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{a}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{b}}, \bar{\boldsymbol{\eta}})\mathbf{v} &= \mathbf{v}^\top \mathcal{Y}\mathbf{v} \\ &= \left([\mathbf{v}_j - \mathbf{v}_{j+M}]_{j=1}^M\right)^\top Y [\mathbf{v}_i - \mathbf{v}_{i+M}]_{i=1}^M \\ &= \left([\mathbf{v}_j]_{j=1}^M\right)^\top Y [\mathbf{v}_i]_{i=1}^M + ([\mathbf{v}_{j+M}]_{j=1}^M)^\top Y [\mathbf{v}_{i+M}]_{i=1}^M \\ &\quad - 2\sum_{i=1}^M Y_{ii}\mathbf{v}_i\mathbf{v}_{i+M} \geq c\|\mathbf{v}\|^2,\end{aligned}$$

for a constant $c > 0$. ∎

Finally, having satisfied the second order conditions in Lemma 6.5, we follow the same analysis as in Theorem 3.3, obtaining the following error estimate.

THEOREM 6.6. *Let $\bar{u}$ be the optimal control of problems $(\mathbf{P_M^0})$ and $\bar{u}_h$ the solution of the corresponding discretized problem. Then, the estimate*

$$||\bar{u} - \bar{u}^h|| \leq Ch^2|\log h|, \tag{6.12}$$

holds for some constant $C > 0$, independent of mesh parameter h.

7. Numerical experimentation. In this section we focus on the numerical confirmation of our theoretical estimates by means of examples implemented in Matlab. In our experiments, we observe that the order of the error corresponds approximately to our prediction $h^2 \approx h^2|\log h|$. The $|\log(h)|$–term is hard to detect in our computations because the size of the mesh is small.

In order to compute the numerical solution in our examples, we transform our problem in a finite–dimensional problem, after the discretization of the state equation and the variables represented in the finite–element space. The cost function is also transformed to a function defined in finite–dimensional spaces using the mid–point rule to approximate the integrals. The contribution of the integration error is known to be of order h^2. Therefore, it should not affect to the observation of the error estimates of the optimal variables. The associated optimization problems are nonsmooth since a term with 1–norm is present. In oder to solve these problems numerically we apply the *second order orthant–wise method* developed in [14], which is able to compute solutions with exact null components where the solution is sparse.

The measures of convergence rate with respect to h were calculated by using decreasing mesh sizes at different values of h. As in [25] we use a very thin mesh in order to compute a high precision solution subsequently considered as the "exact" solution, with which we will compare the approximated solutions for every mesh. The experimental order of convergence is measure with the help of the following formula:

$$EOC = \frac{\log(||\bar{u}^{h_1} - \bar{u}_h^*||) - \log(||\bar{u}_h^* - \bar{u}^{h_2}||)}{\log(h_1) - \log(h_2)}, \tag{7.1}$$

for two consecutive mesh sizes h_1 y h_2, and u_h^* the approximate the reference solution of the problem.

Example 1.. In our first example we consider $\mathbb{R}^8$ as control space and we chose 5 different points of the domain where the state constraints are imposed.

(E1)
$$\begin{cases} \min\limits_{u\in\mathcal{U}_{ad}\subset\mathbb{R}^8} J(y,u) = \dfrac{1}{2}\displaystyle\int_\Omega (y-y_d)^2\,dx + \dfrac{1}{80}\|u\|_2^2 + 5\|u\|_1 \\ \text{subject to} \\ -\Delta y(x) + y(x) = \displaystyle\sum_{i=1}^{8} u_i e_i(x), \quad \text{in } \Omega = (0,1)^2, \\ \qquad\qquad y(x) = 0, \qquad\qquad \text{on } \Gamma, \\ \qquad\qquad y(x_j) \le -10, \qquad\qquad \forall j = 1,2,3,4,5; \end{cases}$$

The state constraints are given in the following points of the domain: $x_1 = \binom{0.08}{0.4}$, $x_2 = \binom{0.4}{0.4}$, $x_3 = \binom{0.84}{0.12}$, $x_4 = \binom{0.12}{0.44}$, $x_5 = \binom{0.2}{0.24}$ and the functions e_i are defined by

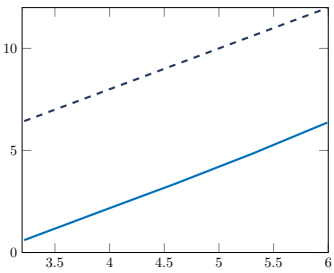
$$\begin{aligned} e_1(x) &= x_1 + x_2, & e_2(x) &= 8\pi^2 \sin(2\pi x_1)\sin(2\pi x_2), \\ e_3(x) &= x_1 - x_2, & e_4(x) &= \cos^3(\pi x_1) \\ e_5(x) &= 4\pi^2 \cos(2\pi(x_1+x_2)), & e_6(x) &= 4\pi^2 \cos(2\pi(x_1-x_2)), \\ e_7(x) &= x_1^2 + x_2^2, & e_8(x) &= (x_1^2-1)(x_2^2-1)(x_1^2+x_2^2). \end{aligned}$$

Additionally, the set of admissible controls is determined by the constants $u_a = -500$ and $u_b = 500$. The chosen desired state is $y_d(x) = -2\sin(2\pi x_1)\sin(2\pi x_2) + 5$.

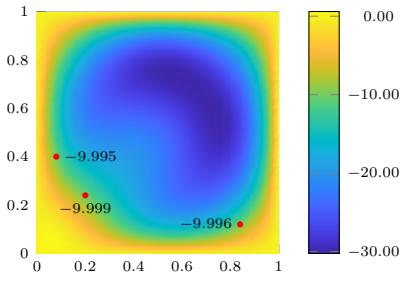
The exact solution of this problem is not known. Therefore, we compute a reference solution $\bar{u}_h^* = \bar{u}^h$ for $h = 0.00125$. The order of error computed with (7.1) gives the following results:

h	$\|\|\bar{u}^h - \bar{u}_h^*\|\|$	EOC
0.04	0.54515935	-
0.02	0.13564030	2.0069
0.01	0.03410037	1.9919
0.005	0.00802171	2.0878
0.0025	0.00171625	2.2247

Table 1: Experimental order of error for Example 1.



(a) Example 1: $-\log(h)$ versus $-\log(\|\bar{u}^h - \bar{u}_h^*\|)$ (solid line) compared with $-2\log(h)$ (dashed line).



(b) $\bar{y}^h$ computed at $h = 0.02$, showing the active state constraints (in red dots)

We notice that the error estimate is approximately h^2. Figure 1a, shows similar slopes of the numerical error compared with h^2 in logarithmic scale.

The approximated optimal control for **(E1)** with $h = 0.04$, is given by

$$\bar{u}^h = \begin{pmatrix} -455.9386 \\ 0 \\ 0 \\ 0 \\ 9.7017 \\ 13.1094 \\ -160.0708 \\ 0 \end{pmatrix};$$

were sparsity property is satisfied in several of its entries.

In the Figure 1b, we observe that the approximated optimal state $\bar{y}^h$ for $h = 0.04$ satisfies (approximately) the state constraints at the prescribed points of the domain.

$$\left[\bar{y}^h(x_j)\right]_{j=1}^{5} = \begin{pmatrix} -9.9948 \\ -19.7645 \\ -9.9964 \\ -14.5492 \\ -9.9994 \end{pmatrix},$$

i.e. in the points x_1, x_3 and x_5 the constraints are active.

Example 2. Here, we are interested in solving the optimal control problem without the Tikhonov regularization ($\alpha = 0$). We also consider five disjoint points on the domain where state constraints must be satisfied. Our problem reads

$$\textbf{(E2)} \qquad \begin{cases} \min\limits_{u \in \mathcal{U}_{ad} \subset \mathbb{R}^5} J(y,u) = \dfrac{1}{2}\|y - y_d\|^2 + 2\|u\|_1 \\ \text{subject to} \\ -\Delta y(x) = \sum\limits_{i=1}^{5} u_i e_i(x), & \text{in } \Omega = (0,1)^2, \\ y(x) = 0, & \text{on } \Gamma, \\ y(x_j) \leq b_j, & \forall j = 1,2,3,4,5; \end{cases}$$

with $x_1 = \binom{0.2}{0.5}$, $x_2 = \binom{0.7}{0.5}$, $x_3 = \binom{0.8}{0.5}$, $x_4 = \binom{0.2}{0.1}$, $x_5 = \binom{0.5}{0.1}$ and $b = (-8, -10, -12, -12, 9)^\top$ Functions e_i are defined by

$$e_1(x) = \sin(\pi(4x_1 + 5x_2)), \qquad e_2(x) = \sin(\pi(2x_1 + 3x_2)),$$
$$e_3(x) = -\cos(\pi(-3x_1 + 17x_2)), \qquad e_4(x) = \cos(\pi(x_1 + x_2)),$$
$$e_5(x) = -\sin(\pi(x_1 - x_2)). \qquad z$$

On set of admissible controls is given by

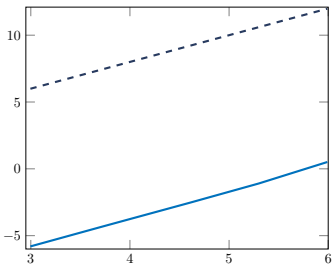
$$\mathcal{U}_{ad} = \{u \in \mathbb{R}^8 : -5000 \leq u_i \leq 5000, \quad \forall i = 1, 2, \ldots, 8\},$$

and the desired state is $y_d(x) = -2\sin(2\pi x_1)\cos(2\pi x_2) + 5$.

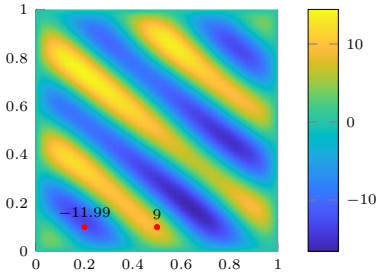
We proceed as in the first example since exact solution of this problem is also not known. The reference solution $\bar{u}_h^* = \bar{u}^h$ is computed at $h = 0.00125$. The experimental order of convergence can be observed in Table 2.

h	$\|\|\bar{u}^h - \bar{u}_h^*\|\|$	EOC
0.05	332.1845	-
0.025	80.5232	2.0445
0.020	51.2134	2.028
0.010	12.6040	2.0226
0.005	3.0017	2.07
0.0025	0.6	2.3227

Table 2: Experimental order of error for Example 2.



(a) Example 2: $-\log(h)$ versus $-\log(||\bar{u}^h - \bar{u}_h^*||)$ (solid line) compared with $-2\log(h)$ (dashed line).



(b) $\bar{y}^h$ computed at $h = 0.04$, showing the active state constraints (in red dots)

Note that the error estimator is close to 2. Therefore, the order of the error is approximately h^2. Figure 2a, shows the numerical error compared with the function h^2 in logarithmic scale. The optimal control calculated for $h = 0.02$, is given by:

$$\bar{u}^h = \begin{pmatrix} 4657.8 \\ 0 \\ 0 \\ 29 \\ 150.3 \end{pmatrix}.$$

Sparsity is satisfied in components $\bar{u}_2^h$ and $\bar{u}_3^h$. In Figure 2a the approximated optimal state $\bar{y}^h$ for $h = 0.02$ is depicted. According to the definition of the bound b, the values of the state at the constrained points are given by

$$\left[\bar{y}^h(x_j)\right]_{j=1}^{5} = \begin{pmatrix} -8 \\ -12.7414 \\ -13.116 \\ -11.992 \\ 9.000 \end{pmatrix},$$

were activity at the points x_4 and x_5 is observed.

REFERENCES

[1] Bartle, R.G. *The elements of integration and Lebesgue Measure*. EU : Wiley Classics Library (1995).

[2] Bonnans, J., Casas, E. . Contrôle de systèmes elliptiques similinéaires comportant des contraintes sur l'état. *(French. English summary) [Control of semilinear elliptic systems with state constraints] Nonlinear partial differential equations and their applications. Collège de France seminar, VIII* (1988), 69-86.

[3] Bonnans, J. F., Gilbert, J. C., Lemaréchal C., Sagastizábal C. A. *Numerical Optimization.* Paris, Grenoble, Rio de Janeiro, Springer (2006).

[4] Brezis, H. *Functional analysis, sobolev spaces and partial differential equations.* EU: Springer (2010).

[5] Casas, E., Clason, C., Kunisch, K. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim., 50* (2012), 1735–1752.

[6] Casas, E., Herzog, R., Wachsmuth, G. Approximation of sparse controls in semilinear equations by piecewise linear functions. *Numer. Math. 122* (2012), no. 4, 645–669.

[7] Casas, E., Herzog, R., Wachsmuth, G. Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional. *SIAM J. Optim., 22* (2012), 795–820.

[8] Casas, E., Kunisch, K. Optimal control of semilinear elliptic equations in measure spaces. *SIAM J. Control Optim., 52* (2014), 339-364.

[9] Casas, E., Tröltzsch, F. Second-order and stability analysis for state-constrained elliptic optimal control problems with sparse controls. *SIAM J. Control Optim., 52* (2014), 1010–1033.

[10] Ciarlet, P. *The finite element method for elliptic problems.* Amsterdam, NL: North-Holland (1978).

[11] Clarke, F. *Functional Analysis, Calculus of Variations and Optimal Control.* FR: Springer (2013).

[12] Clarke, F. H. *Optimization and Nonsmooth Analysis.* EU: John Wiley & Sons (1983).

[13] De Los Reyes, J. C. *Numerical PDE-Constrained Optimization.* EC: Springer (2015).

[14] De Los Reyes, J. C., Loayza, E., Merino, P. Second-order orthant-based methods with enriched Hessian information for sparse ℓ_1-optimization. *Comput. Optim. Appl., 67* (2017), 225-258.

[15] De Los Reyes, J. C., Merino, P., Rehberg, J., Tröltzsch, F. Optimality conditions for state-constrained PDE control problems with time–dependent controls. *Control Cybernet, 37* (2008), 5–38.

[16] Evans, L. *Partial Differential Equations.* EU: AMS (1998).

[17] Fu, H., Ng, M.K., Nikolova, M., Barlow, J.L. Efficient minimization methods of mixed $\ell_2 - \ell_1$ and $\ell_1 - \ell_1$ norms for image restoration. *SIAM J. Sci. Comput., 27* (2006), 1881–1902.

[18] Gilbarg, D & Trudinger N. *Elliptic Partial Differential Equations of Second Order.* DE: Springer (1998).

[19] Grisvard, P. *Elliptic Problems in Nonsmooth Domains.* Philadelphia, EU: Pitman (1985).

[20] Henrion, R. On constraint qualifications. *Optimization Theory and Applications, 72*, (1992). 187-197.

[21] Kunisch, K., Wachsmuth, D. On time optimal control of the wave equation and its numerical realization as parametric optimization problem. *SIAM J. Control Optim., 51* (2013), 1232-1262.

[22] Kok, H., Wust, P., Stauffer, P., Bardati, F., van Rhoon, G., & Crezee, J. Current state of the art of regional hyperthermia treatment planning: a review. *Radiation Oncology*, 10, 196. (2015). http://doi.org/10.1186/s13014-015-0503-8

[23] Merino, P.Neitzel I., Tröltzsch F. An adaptive numerical method for semi-infinite elliptic control problems based on error estimates. *Optimization Methods and Software, 30* (2015), x492–515.

[24] Merino, P., Neitzel I., Tröltzsch F. On linear-quadratic elliptic control problems of semi-infinite type. *Applicable Analysis, 90* (2011), 1047–1074.

[25] Merino, P., Tröltzsch, F., Vexler, B. Error Estimates for the Finite Element Approximation of a Semilinear Elliptic Control Problem with State Constraints and Finite Dimensional Control Space. *Mathematical Modelling and Numerical Analysis, 44* (2010), 167–188.

[26] Rannacher, R., Vexler, B. A priori error estimates for the finite element discretization of elliptic parameter identification problems with pointwise measurements. *SIAM J. Control Optim., 44* (2005), 1844 –1863.

[27] Robinson, S. M. Stability Theory for Systems of Inequalities, part I: Linear Systems*. *SIAM J. Numer. Anal., 12* (1975), 754–769.

[28] Robinson, S. M. Stability Theory for Systems of Inequalities, part II: Differentiable Nonlinear Systems*. *SIAM J. Numer. Anal., 13* (1976), 497–513.

[29] Robinson, S. M. Strongly Regular Generalized Equations. *Mathematics of Operations Research, 5* (1980), 43–62.

[30] Robinson, S. M. Generalized Equations and Their Solutions, Part II Applications to Nonlinear Programming. *Mathematical Programming Study, 19* (1982), 200–221.

[31] Roubíček, T. Optimal control of a Stefan problem with state-space constraints. *Numerische Mathematik, 50* (1987), 723–744.
[32] Stadler, G. Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices. *Comput. Optim. Appl., 44* (2009), 159–181.
[33] Tianhong, H. *Lasso and General ℓ_1 Regularized Regression under Linear Equality and Inequality Constraints*. Purdue University, EU: PhD thesis (2011).
[34] Tröltzsch, F. *Optimal Control of Partial Differential Equations Theory, Methods and Applications*. EU: AMS (2010).