Contents lists available at ScienceDirect

Results in Applied Mathematics

journal homepage: www.elsevier.com/locate/results-in-applied-mathematics

A cubic C^0 interior penalty method for elliptic distributed optimal control problems with pointwise state and control constraints[☆]

Susanne C. Brenner[a,*], Li-yeng Sung[a], Zhiyu Tan[b]

[a] Department of Mathematics and Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, United States of America

[b] Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, United States of America

ARTICLE INFO

Article history:
Received 30 March 2020
Received in revised form 9 May 2020
Accepted 12 May 2020
Available online xxxx

MSC:
49J40
49N10
65K10
65N15
65N30

Keywords:
Elliptic distributed optimal control problems
Pointwise state and control constraints
C^0 interior penalty methods
Cubic Hermite finite element

ABSTRACT

We design and analyze a cubic C^0 interior penalty method for linear–quadratic elliptic distributed optimal control problems with pointwise state and control constraints. Numerical results that corroborate the theoretical error estimates are also presented.

1. Introduction

Let $\Omega \subset \mathbb{R}^2$ be a bounded convex polygonal domain, β be a positive constant, and $y_d \in L_2(\Omega)$. The optimal control problem is to

$$\text{find} \quad (\bar{y}, \bar{u}) = \underset{(y,u)\in\mathbb{K}}{\operatorname{argmin}} \frac{1}{2}\Big(\|y - y_d\|^2_{L_2(\Omega)} + \beta\|u\|^2_{L_2(\Omega)}\Big), \tag{1.1}$$

where $(y, u) \in \mathbb{K} \subset H^1_0(\Omega) \times L_2(\Omega)$ if and only if

$$\int_\Omega \nabla y \cdot \nabla z \, dx = \int_\Omega uz \, dx \qquad \forall z \in H^1_0(\Omega), \tag{1.2}$$

[☆] This work was supported in part by the National Science Foundation, United States of America under Grant No. DMS-19-13035.

[*] Corresponding author.
E-mail addresses: brenner@math.lsu.edu (S.C. Brenner), sung@math.lsu.edu (L.-y. Sung), ztan@lsu.edu (Z. Tan).

https://doi.org/10.1016/j.rinam.2020.100119

and

$$\psi_1 \le y \le \psi_2 \qquad \text{a.e. in } \Omega, \tag{1.3}$$

$$\phi_1 \le u \le \phi_2 \qquad \text{a.e. in } \Omega. \tag{1.4}$$

We assume that (i) $\psi_1, \psi_2 \in W^{2,\infty}(\Omega) \cap H^3(\Omega)$, (ii) $\psi_1 < \psi_2$ on $\bar{\Omega}$, (iii) $\psi_1 < 0 < \psi_2$ on $\partial\Omega$, (iv) $\phi_1, \phi_2 \in W^{1,\infty}(\Omega)$ and (v) $\phi_1 < \phi_2$ on $\bar{\Omega}$.

Remark 1.1. Throughout the paper we follow standard notation for differential operators, function spaces and norms that can be found for example in [1–3].

There are three different approaches to solving the optimal control problem (1.1)–(1.4) by finite element methods. The first one is based on the first order optimality condition of the reduced minimization problem involving the control variable [4,5]. The second one is based on a regularization of the state constraints, such as the Lavrentiev regularization approach in [6,7] and the Moreau–Yosida regularization approach in [8,9]. The third is based on a reformulation of the optimal control problem as a fourth order variational inequality [10]. The goal of this paper is to design and analyze a cubic C^0 interior penalty method based on the third approach. We note that the idea of the reformulation was discussed in [11] and a nonconforming finite element method based on this idea was investigated in [12] for state constrained problems. Other finite element methods for state constrained problems based on this approach can be found in [13–20].

The cubic finite element method in this paper performs better than the Morley finite element method in [10]. Moreover, by taking advantage of the internal degree of freedom for the cubic element, the discrete problem becomes a quadratic programming problem with box constraints that can be solved efficiently by a primal–dual active set method [21–24]. Our interest in the cubic finite element method is also motivated by the observation (cf. [19,25]) that cubic adaptive finite element methods for fourth order variational inequalities can capture the free boundary of the coincidence/active set much more effectively than their quadratic counterparts.

The rest of the paper is organized as follows. We recall some relevant results for the continuous problem in Section 2 and introduce the finite element method in Section 3. Tools for the convergence analysis are presented in Section 4. We establish a preliminary estimate in Section 5 and derive error estimates in Section 6. Numerical results that illustrate the performance of our method are presented in Section 7, followed by some concluding remarks in Section 8. Technical results concerning the tools in Section 4 are provided in Appendices A and B.

We will use C (with or without subscript) to denote a generic positive constant independent of the mesh size. To avoid the proliferation of constants, we also use the notation $A \lesssim B$ (or $B \gtrsim A$) to represent the statement $A \le (\text{constant})B$, where the positive constant is independent of the mesh size. The notation $A \approx B$ is equivalent to $A \lesssim B$ and $B \lesssim A$.

2. The continuous problem

Since Ω is convex, the constraint (1.2) implies $y \in H^2(\Omega)$ by elliptic regularity [26–28]. Hence the optimal control problem (1.1)–(1.4) can be reformulated as follows:

$$\text{Find} \quad \bar{y} = \underset{y\in K}{\operatorname{argmin}} \frac{1}{2}\Big(\|y - y_d\|^2_{L_2(\Omega)} + \beta\|\Delta y\|^2_{L_2(\Omega)}\Big) = \underset{y\in K}{\operatorname{argmin}}\Big(\frac{1}{2}\mathcal{A}(y,y) - (y_d, y)\Big), \tag{2.1}$$

where

$$K = \{y \in H^2(\Omega) \cap H^1_0(\Omega) : \ \psi_1 \le y \le \psi_2 \text{ and } \phi_1 \le (-\Delta y) \le \phi_2 \text{ a.e. in } \Omega\}, \tag{2.2}$$

$$\mathcal{A}(y,z) = \beta a(y,z) + (y,z), \quad a(y,z) = \int_\Omega (\Delta y)(\Delta z)dx, \quad \text{and} \quad (y,z) = \int_\Omega yz\,dx. \tag{2.3}$$

Remark 2.1. Note that we have an alternative expression (cf. [29, Lemma 2.2.2])

$$a(y,z) = \sum_{i,j=1}^{2} \int_\Omega \left(\frac{\partial^2 y}{\partial x_i \partial x_j}\right)\left(\frac{\partial^2 z}{\partial x_i \partial x_j}\right) dx =: \int_\Omega D^2 y : D^2 z\,dx \qquad \forall\, y,z \in H^2(\Omega)\cap H^1_0(\Omega). \tag{2.4}$$

We assume the following Slater condition:

$$\begin{aligned}&\text{There exists } y \in H^2(\Omega)\cap H^1_0(\Omega) \text{ such that (i) } \psi_1 < y < \psi_2 \text{ in } \Omega, \text{ and (ii) } u = -\Delta y\\ &\text{satisfies the constraint (1.4).}\end{aligned} \tag{2.5}$$

Under the condition (2.5), the closed convex subset K of $H^2(\Omega)\cap H^1_0(\Omega)$ is nonempty. Therefore, in view of the coercivity of $\mathcal{A}(\cdot,\cdot)$, the convex optimization problem (2.1) has a unique solution $\bar{y} \in K$ (cf. [30,31]) characterized by the fourth order variational inequality

$$\mathcal{A}(\bar{y}, y - \bar{y}) - (y_d, y - \bar{y}) \ge 0, \quad \forall y \in K. \tag{2.6}$$

Moreover, we have the following generalized Karush–Kuhn–Tucker (KKT) conditions (cf. [24, Chapter 1, Theorem 1.6]):

$$\mathcal{A}(\bar{y},z)-(y_d,z)=\int_\Omega z\,d\mu+\int_\Omega \lambda(-\Delta z)dx \qquad \forall z\in H^2(\Omega)\cap H_0^1(\Omega), \tag{2.7}$$

where $\lambda\in L_2(\Omega)$ and $\mu\in\mathcal{M}(\Omega)$ (the space of regular Borel measures on Ω) satisfy

$$\lambda\ge 0 \quad \text{if } -\Delta\bar{y}=\phi_1, \tag{2.8}$$

$$\lambda\le 0 \quad \text{if } -\Delta\bar{y}=\phi_2, \tag{2.9}$$

$$\lambda=0 \quad \text{otherwise;} \tag{2.10}$$

$$\mu\ge 0 \quad \text{if } \bar{y}=\psi_1, \tag{2.11}$$

$$\mu\le 0 \quad \text{if } \bar{y}=\psi_2, \tag{2.12}$$

$$\mu=0 \quad \text{otherwise.} \tag{2.13}$$

The derivations of the following regularity results, which are based on [26–28,32–36], can be found in [10, Section 2]:

$$\lambda\in W^{1,s}(\Omega) \qquad \forall s\in[1,2), \tag{2.14}$$

$$\mu\in W^{-1,s}(\Omega) \qquad \forall s\in[1,2), \tag{2.15}$$

$$\bar{u}=-\Delta\bar{y}\in H^1(\Omega)\cap L_\infty(\Omega), \tag{2.16}$$

$$\bar{y}\in H^3_{loc}(\Omega)\cap W^{2,p}_{loc}(\Omega)\cap H^{2+\alpha}(\Omega) \qquad \forall p\in[1,\infty), \tag{2.17}$$

where $\alpha\in(0,1]$ is determined by the angles at the corners of Ω.

Under additional assumptions, the regularity of λ, μ, $\bar{u}$ and $\bar{y}$ can be improved.

In the case where $\mathrm{supp}\lambda\cap\mathrm{supp}\mu=\emptyset$, we have

$$\lambda\in H^1(\Omega), \tag{2.18}$$

$$\mu\in H^{-1}(\Omega), \tag{2.19}$$

$$\bar{y} \text{ belongs to } W^{2,\infty}(G) \text{ in a neighborhood of } \mathrm{supp}\mu. \tag{2.20}$$

In the case where $\phi_1\le 0\le\phi_2$, we have

$$\bar{u}=-\Delta\bar{y}\in H_0^1(\Omega). \tag{2.21}$$

3. The discrete problem

Let $\mathcal{T}_h$ be a quasi-uniform simplicial triangulation of Ω. We denote the diameter of T by h_T and $h\approx\max_{T\in\mathcal{T}_h}\mathrm{diam}\,T$ is a mesh parameter. The set of interior (resp., boundary) edges of $\mathcal{T}_h$ is denoted by $\mathcal{E}_h^i$ (resp., $\mathcal{E}_h^b$). The cubic Hermite finite element space $V_h\subset H^2(\Omega)\cap H_0^1(\Omega)$ (cf. [2,3]) consists of piecewise cubic polynomial functions that are continuous up to first order derivatives at the vertices of $\mathcal{T}_h$.

3.1. A modified cubic Hermite finite element

Let T be a triangle. The degrees of freedom (dofs) for the standard cubic Hermite finite element are given by $v(p_i)$ and $\nabla v(p_i)$ for $1\le i\le 3$ and $v(c)$, where p_1, p_2, p_3 are the vertices of T and c is the center of T. For handling the control constraint (1.4), it is convenient to modify the internal degree of freedom, which does not change the finite element space V_h.

Let φ_i be the barycentric coordinate associated with p_i and $\varphi_T=\varphi_1\varphi_2\varphi_3$ be the cubic bubble function on T.

Lemma 3.1. *Let p_1, p_2 and p_3 be the vertices of a triangle T. A cubic polynomial v is uniquely determined by $v(p_i)$ and $\nabla v(p_i)$ for $1\le i\le 3$ together with the integral*

$$\int_T(1+\gamma\varphi_T)(\Delta v)\,dx,$$

where γ is any nonnegative number.

Proof. It suffices to show that v is uniquely determined by the 10 dofs. Suppose $v(p_i)$ and $\nabla v(p_i)$ vanish for $1\le i\le 3$. Then v vanishes on ∂T and hence v is a multiple of φ_T. It remains to verify that $\int_T(1+\gamma\varphi_T)(\Delta\varphi_T)\,dx\ne 0$.

A direct calculation shows that the normal derivative of φ_T is <0 on ∂T except at the vertices p_1, p_2 and p_3. Therefore we have

$$\int_T(1+\gamma\varphi_T)(\Delta\varphi_T)dx=\int_{\partial T}(\partial\varphi_T/\partial n)ds-\gamma\int_T(\nabla\varphi_T)\cdot(\nabla\varphi_T)dx<0. \qquad \square$$

According to Lemma 3.1 with $\gamma = 0$, we have a modified cubic Hermite finite element by replacing the dof $v(c)$ with the dof $\int_T (\Delta v) dx$. We will use this element in the computations.

Remark 3.2. The case where $\gamma = h_T^2$ will play a useful role in the convergence analysis.

3.2. The $\mathbf{C^0}$ *interior penalty method*

In the C^0 interior penalty approach [37–39], the bilinear form $a(\cdot,\cdot)$ in (2.4) is replaced by the bilinear form $a_h(\cdot,\cdot)$ defined by

$$a_h(w,v) = \sum_{T\in\mathcal{T}_h} \int_T D^2 w : D^2 v\,dx + \sum_{e\in\mathcal{E}_h^i} \int_e \left(\left\{\!\!\left\{ \frac{\partial^2 w}{\partial n^2} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] + \left\{\!\!\left\{ \frac{\partial^2 v}{\partial n^2} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial w}{\partial n} \right]\!\!\right] \right) ds$$
$$+ \sum_{e\in\mathcal{E}_h^i} \frac{\sigma}{|e|} \int_e \left[\!\!\left[\frac{\partial w}{\partial n} \right]\!\!\right] \left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] ds, \tag{3.1}$$

where $|e|$ is the length of the edge e, $\sigma > 0$ is a penalty parameter, and the jumps and averages of the normal derivatives for piecewise H^2 functions are defined as follows.

Let $e \in \mathcal{E}_h^i$ be the common edge of $T_e^{\pm} \in \mathcal{T}_h$ and n_e be the unit normal of e pointing from T_e^- to T_e^+. We define on the edge e

$$\left\{\!\!\left\{ \frac{\partial^2 v}{\partial n^2} \right\}\!\!\right\} = \frac{1}{2}\left(\frac{\partial^2 v_+}{\partial n_e^2}\bigg|_e + \frac{\partial^2 v_-}{\partial n_e^2}\bigg|_e \right) \quad\text{and}\quad \left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] = \frac{\partial v_+}{\partial n_e}\bigg|_e - \frac{\partial v_-}{\partial n_e}\bigg|_e,$$

where $v_\pm = v\big|_{T_e^\pm}$.

For σ sufficiently large (cf. [39]), we have

$$a_h(z_h,z_h) \gtrsim \sum_{T\in\mathcal{T}_h} |z_h|_{H^2(T)}^2 + \sum_{e\in\mathcal{E}_h^i} \frac{1}{|e|} \|[\![\partial z_h/\partial n]\!]\|_{L_2(e)}^2 \qquad \forall\, z_h \in V_h. \tag{3.2}$$

Let I_h be the nodal interpolation operator for the P_1 finite element space (cf. [2,3]) associated with $\mathcal{T}_h$, and Q_h be the L_2 projection onto the space of piecewise constant functions defined by

$$(Q_h v)\big|_T = (1/|T|) \int_T v\,dx \qquad \forall\, v \in L_2(\Omega),\ T \in \mathcal{T}_h, \tag{3.3}$$

where $|T|$ is the area of the triangle. The discrete constraint set $K_h \subset V_h$ is given by

$$K_h = \{y_h \in V_h :\ I_h\psi_1 \le I_h y_h \le I_h \psi_2 \quad\text{and}\quad Q_h\phi_1 \le Q_h(-\Delta_h y_h) \le Q_h\phi_2\}, \tag{3.4}$$

where Δ_h is the piecewise defined Laplace operator.

Remark 3.3. According to the definition of K_h, the constraint (1.3) is imposed at the vertices of $\mathcal{T}_h$ and the constraint (1.4) is imposed on each $T \in \mathcal{T}_h$ in the mean-value sense. These constraints are box constraints for the modified Hermite element introduced in Section 3.1.

The discrete problem for (2.6) is to find $\bar{y}_h \in K_h$ such that

$$\mathcal{A}_h(\bar{y}_h, y_h - \bar{y}_h) - (y_d, y_h - \bar{y}_h) \ge 0 \qquad \forall\, y_h \in K_h, \tag{3.5}$$

where

$$\mathcal{A}_h(z_h, y_h) = \beta a_h(z_h, y_h) + (z_h, y_h) \qquad \forall y_h, z_h \in V_h. \tag{3.6}$$

We will use the following mesh-dependent norm $\|\cdot\|_h$ in the error analysis:

$$\|z\|_h^2 = \beta\Big(\sum_{T\in\mathcal{T}_h} |z|_{H^2(T)}^2 + \sum_{e\in\mathcal{E}_h^i} |e|^{-1} \big\|[\![\partial z/\partial n]\!]\big\|_{L_2(e)}^2 \Big) + \|z\|_{L_2(\Omega)}^2. \tag{3.7}$$

It follows from (3.2), (3.6) and (3.7) that

$$\mathcal{A}_h(y,z) \lesssim \|y\|_h \|z\|_h \qquad \forall\, y,z \in [H^2(\Omega)\cap H_0^1(\Omega)] + V_h, \tag{3.8}$$

and

$$\mathcal{A}_h(z_h,z_h) \gtrsim \|z_h\|_h^2 \qquad \forall\, z_h \in V_h, \tag{3.9}$$

provided that σ is sufficiently large, which is assumed to be the case from here on.

Note that

$$\|z\|_h^2=\beta|z|^2_{H^2(\Omega)}+\|z\|^2_{L_2(\Omega)}\qquad\forall\, z\in H^2(\Omega).\tag{3.10}$$

4. Tools for the convergence analysis

Interpolation and enriching operators with appropriate properties are the two main tools in the convergence analysis developed in [40] for optimal control problems with state constraints that was extended to problems with both state and control constraints in [10].

We will use the following notation in the construction of these operators.

- $\mathcal{V}^c$ is the set of the corners of Ω.
- $\mathcal{V}_h$ is the set of the vertices of $\mathcal{T}_h$.
- $\mathcal{V}_h^b$ is the subset of $\mathcal{V}_h$ consisting of the vertices that belong to $\partial\Omega$.

For each $p\in\mathcal{V}_h$, we assign an element $T_p\in\mathcal{T}_h$ such that

$$\begin{aligned}&\text{(i) } p \text{ is a vertex of } T_p,\ \text{(ii) if } p\in\mathcal{V}_h^b \text{ is the common endpoint of two edges in } \mathcal{E}_h^b,\text{ then}\\ &\qquad\text{one of these edges should be an edge of } T_p.\end{aligned}\tag{4.1}$$

4.1. Interpolation operators

Let $\Pi^L_{h,T}$ be the nodal interpolation operator on T for the cubic Lagrange element (cf. [2,3]). The interpolation operators $\Pi_h,\Pi_{h,\rho}:H^2(\Omega)\cap H^1_0(\Omega)\longrightarrow V_h$ are defined as follows:

$$(\Pi_h\zeta)(p)=\quad\zeta(p)\quad=\quad(\Pi_{h,\rho}\zeta)(p)\qquad\forall\, p\in\mathcal{V}_h,\tag{4.2}$$

$$\partial_{x_i}(\Pi_h\zeta)(p)=\partial_{x_i}(\Pi^L_{h,T_p}\zeta)(p)=\partial_{x_i}(\Pi_{h,\rho}\zeta)(p)\qquad\forall\, p\in\mathcal{V}_h\setminus\mathcal{V}^c\quad(i=1,2),\tag{4.3}$$

$$\partial_{x_i}(\Pi_h\zeta)(p)=\qquad 0\qquad=\partial_{x_i}(\Pi_{h,\rho}\zeta)(p)\qquad\forall\, p\in\mathcal{V}^c\quad(i=1,2),\tag{4.4}$$

and

$$\int_T\Delta(\Pi_h\zeta)dx=\int_T\Delta\zeta\,dx\qquad\forall\, T\in\mathcal{T}_h,\tag{4.5}$$

$$\int_T\rho\Delta(\Pi_{h,\rho}\zeta)dx=\int_T\rho\Delta\zeta\,dx\qquad\forall\, T\in\mathcal{T}_h,\tag{4.6}$$

where the weight function ρ is defined by

$$\rho_T=\rho\big|_T=1+h_T^2\varphi_T\qquad\forall\, T\in\mathcal{T}_h.\tag{4.7}$$

Remark 4.1. It follows from Lemma 3.1 that (4.5) and (4.6) are well-defined. Moreover the choice of T_p specified in (4.1) guarantees that $\Pi_h\zeta$ and $\Pi_{h,\rho}\zeta$ vanish on $\partial\Omega$ for $\zeta\in H^2(\Omega)\cap H^1_0(\Omega)$.

Note that (3.3) and (4.5) imply

$$Q_h(\Delta\zeta)=Q_h[\Delta_h(\Pi_h\zeta)]\qquad\qquad\forall\,\zeta\in H^2(\Omega)\cap H^1_0(\Omega),\tag{4.8}$$

and similarly, the relation (4.6) implies

$$Q_{h,\rho}(\Delta\zeta)=Q_{h,\rho}[\Delta_h(\Pi_{h,\rho}\zeta)]\qquad\qquad\forall\,\zeta\in H^2(\Omega)\cap H^1_0(\Omega),\tag{4.9}$$

where $Q_{h,\rho}$ is the L_2 projection onto the space of piecewise constant functions defined by

$$(Q_{h,\rho}v)\big|_T=\left(\int_T\rho_T v\,dx\right)\Big/\left(\int_T\rho_T\,dx\right)\qquad\forall\, v\in L_2(\Omega),\ T\in\mathcal{T}_h.\tag{4.10}$$

Since piecewise constant functions are invariant under $Q_{h,\rho}$, we have a standard interpolation error estimate

$$\|\eta-Q_{h,\rho}\eta\|_{L_2(\Omega)}\lesssim h^s|\eta|_{H^s(\Omega)}\qquad\forall\,\eta\in H^s(\Omega)\quad\text{and}\quad 0\le s\le 1\tag{4.11}$$

by the Bramble–Hilbert lemma [41,42].

It follows immediately from (3.4), (4.2) and (4.8) that

$$\Pi_h K\subset K_h.\tag{4.12}$$

In particular, the closed and convex discrete constraint set K_h is nonempty, which together with (3.9) implies (3.5) has a unique solution $\bar{y}_h\in K_h$.

We have the following interpolation error estimates:

$$\sum_{k=0}^{2} h_T^k |\zeta - \Pi_{h,\rho}\zeta|_{H^k(T)} \lesssim h_T^{2+s} |\zeta|_{H^{2+s}(S_T)} \qquad \forall\, T \in \mathcal{T}_h,\ 0 \le s \le 2, \tag{4.13}$$

where S_T is the union of the triangles of $\mathcal{T}_h$ that share a common vertex with T, and

$$\|\zeta - \Pi_h\zeta\|_h + \|\zeta - \Pi_{h,\rho}\zeta\|_h \le h^s |\zeta|_{H^{2+s}(\Omega)} \qquad \forall\, 0 \le s \le 2. \tag{4.14}$$

Moreover we have

$$\|\Pi_{h,\rho}\zeta\|_h \lesssim \|\zeta\|_{H^2(\Omega)} \qquad \forall\, \zeta \in H^2(\Omega) \cap H_0^1(\Omega). \tag{4.15}$$

The closely related operators Π_h and $\Pi_{h,\rho}$ satisfy the estimates

$$\sum_{k=0}^{2} h^k |\Pi_h\zeta - \Pi_{h,\rho}\zeta|_{H^k(\Omega)} \lesssim h^4 |\Pi_{h,\rho}\zeta - \zeta|_{H^2(\Omega)} \qquad \forall\, \zeta \in H^2(\Omega) \cap H_0^1(\Omega), \tag{4.16}$$

$$\|\Pi_h\zeta - \Pi_{h,\rho}\zeta\|_h \lesssim h^2 |\Pi_{h,\rho}\zeta - \zeta|_{H^2(\Omega)} \qquad \forall\, \zeta \in H^2(\Omega) \cap H_0^1(\Omega). \tag{4.17}$$

The derivations of (4.13)–(4.17) are given in Appendix A.

4.2. An enriching operator

Let $W_h \subset H^2(\Omega) \cap H_0^1(\Omega)$ be the Hsieh–Clough–Tocher (HCT) finite element space associated with $\mathcal{T}_h$ (cf. [43]). On each $T \in \mathcal{T}_h$, the space $\Sigma_{\mathrm{HCT}}(T)$ of shape functions consists of C^1 functions that are piecewise cubic with respect to the triangulation of T determined by the vertices of T and the center of T. A function $v \in \Sigma_{\mathrm{HCT}}(T)$ is uniquely determined by the values of v and ∇v at the three vertices of T together with the means of $\partial v/\partial n$ over the three edges.

We will need an enhanced HCT finite element for the construction of the enriching operator.

Lemma 4.2. *A function v in $\tilde{\Sigma}_{\mathrm{HCT}}(T) = \Sigma_{\mathrm{HCT}}(T) \oplus \langle \varphi_T^2 \rangle$ is uniquely determined by the values of v and ∇v at the vertices of T, the means of $\partial v/\partial n$ over the three edges of T and the value of $\int_T \rho_T \Delta v\, dx$, where ρ_T is given in (4.7).*

Proof. Suppose $v \in \tilde{\Sigma}_{\mathrm{HCT}}(T)$ vanishes up to its first order derivatives at the three vertices and the means of $\partial v/\partial n$ vanish on the three edges, then v is a multiple of φ_T^2 and it only remains to show that $\int_T \rho_T \Delta(\varphi_T^2) dx \neq 0$. Indeed we have

$$\int_T \rho_T \Delta(\varphi_T^2) dx = -h_T^2 \int_T (\nabla\varphi_T) \cdot \nabla(\varphi_T^2) dx = -2h_T^2 \int_T \varphi_T |\nabla\varphi_T|^2 dx < 0. \qquad \square$$

Remark 4.3. Note that $\int_T \Delta(\varphi_T^2) dx = 0$, which is why it is necessary to include a weight function in the constructions of the enhanced HCT finite element and the enriching operator $E_{h,\rho}$ (see below). This in turn necessitates the construction of the weighted interpolation operator $\Pi_{h,\rho}$ (cf. Section 4.1) that is compatible with $E_{h,\rho}$.

Let $\tilde{W}_h \subset H^2(\Omega) \cap H_0^1(\Omega)$ be obtained from W_h by enlarging the space of shape functions $\Sigma_{\mathrm{HCT}}(T)$ to $\tilde{\Sigma}_{\mathrm{HCT}}(T)$ on each $T \in \mathcal{T}_h$. The operator $E_{h,\rho} : V_h \longrightarrow \tilde{W}_h$ is defined as follows:

$$(E_{h,\rho}v)(p) = v(p) \qquad \forall\, p \in \mathcal{V}_h, \tag{4.18}$$

$$\nabla(E_{h,\rho}v)(p) = \nabla v(p) \qquad \forall\, p \in \mathcal{V}_h, \tag{4.19}$$

$$\int_e \frac{\partial(E_{h,\rho}v)}{\partial n_e} ds = \frac{1}{2}\left(\int_e \frac{\partial v_+}{\partial n_e} ds + \int_e \frac{\partial v_-}{\partial n_e} ds\right) \qquad \forall\, e \in \mathcal{E}_h^i, \tag{4.20}$$

where e is a common edge of $T_e^{\pm}$, n_e is a unit normal of e and $v_{\pm} = v\big|_{T_e^{\pm}}$,

$$\int_e \frac{\partial(E_{h,\rho}v)}{\partial n_e} ds = \int_e \frac{\partial v}{\partial n_e} ds \qquad \forall\, e \in \mathcal{E}_h^b, \tag{4.21}$$

where n_e is the unit normal of e pointing towards the outside of Ω,

$$\int_T \rho_T \Delta(E_{h,\rho}v) dx = \int_T \rho_T \Delta v\, dx \qquad \forall\, T \in \mathcal{T}_h. \tag{4.22}$$

Remark 4.4. Note that (4.22) is well-defined because of Lemmas 3.1 and 4.2. Since v and $E_{h,\rho}v$ are cubic polynomials on the edges of $\mathcal{T}_h$, they are identical on the edges by the conditions (4.18)–(4.19).

It follows from (4.10) and (4.22) that

$$Q_{h,\rho}(\Delta E_{h,\rho}v) = Q_{h,\rho}(\Delta_h v) \qquad \forall\, v \in V_h, \tag{4.23}$$

which together with (4.9) implies

$$Q_{h,\rho}[\Delta(E_{h,\rho}\Pi_{h,\rho}\zeta)] = Q_{h,\rho}[\Delta_h(\Pi_{h,\rho}\zeta)] = Q_{h,\rho}(\Delta\zeta) \qquad \forall\, \zeta \in H^2(\Omega)\cap H_0^1(\Omega). \tag{4.24}$$

The operator $E_{h,\rho}$ enjoys the following properties:

$$\sum_{k=0}^{2} h^{2k} \sum_{T\in\mathcal{T}_h} |v - E_{h,\rho}v|^2_{H^k(T)} \lesssim h^4 \sum_{e\in\mathcal{E}_h^i} \frac{1}{|e|}\,\|[\![\partial v/\partial n]\!]\|^2_{L_2(e)} \qquad \forall\, v \in V_h, \tag{4.25}$$

and for $\zeta \in H^{2+s}(\Omega)\cap H_0^1(\Omega)$ and $0 \le s \le 2$,

$$\sum_{k=0}^{2} h^k |\zeta - E_{h,\rho}\Pi_{h,\rho}\zeta|_{H^k(\Omega)} \lesssim h^{2+s}|\zeta|_{H^{2+s}(\Omega)} \tag{4.26}$$

$$\|\zeta - E_{h,\rho}\Pi_{h,\rho}\zeta\|_{W^{1,2/(1-\epsilon)}(\Omega)} \lesssim h^{1+s-\epsilon}|\zeta|_{H^{2+s}(\Omega)} \qquad \forall\, 0 \le \epsilon \le 1/2, \tag{4.27}$$

$$|\mathcal{A}_h(\Pi_{h,\rho}\zeta, v) - \mathcal{A}(\zeta, E_{h,\rho}v)| \lesssim h^s\|\zeta\|_{H^{2+s}(\Omega)}\|v\|_h \qquad \forall\, v \in V_h. \tag{4.28}$$

The derivations of (4.25)–(4.28) are given in Appendix B. Note that (3.7) and (4.25) imply in particular

$$|E_{h,\rho}v|_{H^2(\Omega)} \lesssim \|v\|_h \qquad \forall\, v \in V_h. \tag{4.29}$$

Remark 4.5. The estimate (4.25) indicates that the norms of $v - E_{h,\rho}v$ measure the distance between v and $H^2(\Omega)\cap H_0^1(\Omega)$. The estimates (4.26) and (4.27) state that $E_{h,\rho}\Pi_{h,\rho}$ behaves like a quasi-local interpolation operator. The estimate (4.28) means that $E_{h,\rho}$ is essentially the adjoint of $\Pi_{h,\rho}$ with respect to the bilinear forms $\mathcal{A}(\cdot,\cdot)$ and $\mathcal{A}_h(\cdot,\cdot)$.

5. A preliminary estimate

We will follow the approach in [10,40] and begin with a preliminary estimate that reduces the error analysis to the continuous level. From (2.17), (3.5), (3.9), (4.12) and (4.14), we have

$$\begin{aligned}
\|\bar{y} - \bar{y}_h\|_h^2 &\le 2\|\bar{y} - \Pi_h\bar{y}\|_h^2 + 2\|\Pi_h\bar{y} - \bar{y}_h\|_h^2\\
&\lesssim h^{2\alpha} + \mathcal{A}_h(\Pi_h\bar{y} - \bar{y}_h, \Pi_h\bar{y} - \bar{y}_h)\\
&\le h^{2\alpha} + \mathcal{A}_h(\Pi_h\bar{y}, \Pi_h\bar{y} - \bar{y}_h) - (y_d, \Pi_h\bar{y} - \bar{y}_h)\\
&= h^{2\alpha} + \big[\mathcal{A}_h(\Pi_{h,\rho}\bar{y}, \Pi_{h,\rho}\bar{y} - \bar{y}_h) - (y_d, \Pi_{h,\rho}\bar{y} - \bar{y}_h)\big]\\
&\qquad + \big[\mathcal{A}_h(\Pi_h\bar{y}, \Pi_h\bar{y} - \bar{y}_h) - \mathcal{A}_h(\Pi_{h,\rho}\bar{y}, \Pi_{h,\rho}\bar{y} - \bar{y}_h)\big] + (y_d, \Pi_{h,\rho}\bar{y} - \Pi_h\bar{y}),
\end{aligned} \tag{5.1}$$

and

$$(y_d, \Pi_{h,\rho}\bar{y} - \Pi_h\bar{y}) \lesssim \|y_d\|_{L_2(\Omega)} h^4 |\Pi_{h,\rho}\bar{y} - \bar{y}|_{H^2(\Omega)} \lesssim h^{4+\alpha} \lesssim h^{2\alpha}, \tag{5.2}$$

by (2.17), (4.13) and (4.16).

Furthermore, it follows from (2.17), (3.8), (4.14), (4.15) and (4.17) that

$$\begin{aligned}
&\mathcal{A}_h(\Pi_h\bar{y}, \Pi_h\bar{y} - \bar{y}_h) - \mathcal{A}_h(\Pi_{h,\rho}\bar{y}, \Pi_{h,\rho}\bar{y} - \bar{y}_h)\\
&\qquad = \mathcal{A}_h(\Pi_h\bar{y} - \Pi_{h,\rho}\bar{y}, \Pi_{h,\rho}\bar{y} - \bar{y}_h) + \mathcal{A}_h(\Pi_h\bar{y} - \Pi_{h,\rho}\bar{y}, \Pi_h\bar{y} - \Pi_{h,\rho}\bar{y})\\
&\qquad\qquad + \mathcal{A}_h(\Pi_{h,\rho}\bar{y}, \Pi_h\bar{y} - \Pi_{h,\rho}\bar{y})\\
&\qquad \lesssim \|\Pi_h\bar{y} - \Pi_{h,\rho}\bar{y}\|_h\big(\|\Pi_{h,\rho}\bar{y} - \bar{y}_h\|_h + \|\Pi_h\bar{y} - \Pi_{h,\rho}\bar{y}\|_h + \|\Pi_{h,\rho}\bar{y}\|_h\big)\\
&\qquad \lesssim h^{2+\alpha}\big(\|\Pi_{h,\rho}\bar{y} - \bar{y}\|_h + \|\bar{y} - \bar{y}_h\|_h + h^{2+\alpha} + \|\Pi_{h,\rho}\bar{y}\|_h\big)\\
&\qquad \lesssim h^{2+2\alpha} + h^{2+\alpha}\|\bar{y} - \bar{y}_h\|_h + h^{4+2\alpha} + h^{2+\alpha}\\
&\qquad \lesssim h^{2\alpha} + h^{\alpha}\|\bar{y} - \bar{y}_h\|_h.
\end{aligned} \tag{5.3}$$

Next we consider the second term on the right-hand side of (5.1). We can write

$$\begin{aligned}
\mathcal{A}_h(\Pi_{h,\rho}\bar{y}, \Pi_{h,\rho}\bar{y} - \bar{y}_h) - (y_d, \Pi_{h,\rho}\bar{y} - \bar{y}_h) &= \mathcal{A}\big(\bar{y}, E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big) - \big(y_d, E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big)\\
&\qquad + \big[\mathcal{A}_h(\Pi_{h,\rho}\bar{y}, \Pi_{h,\rho}\bar{y} - \bar{y}_h) - \mathcal{A}\big(\bar{y}, E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big)\big]\\
&\qquad\qquad - \big(y_d, (\Pi_{h,\rho}\bar{y} - \bar{y}_h) - E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big),
\end{aligned} \tag{5.4}$$

and we have

$$\begin{aligned}&\mathcal{A}_h(\Pi_{h,\rho}\bar{y},\Pi_{h,\rho}\bar{y}-\bar{y}_h)-\mathcal{A}\big(\bar{y},E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big)\\ &\quad\lesssim h^\alpha\|\bar{y}\|_{H^{2+\alpha}(\Omega)}\|\Pi_{h,\rho}\bar{y}-\bar{y}_h\|_h\\ &\quad\le h^\alpha\|\bar{y}\|_{H^{2+\alpha}(\Omega)}\big(\|\Pi_{h,\rho}\bar{y}-\bar{y}\|_h+\|\bar{y}-\bar{y}_h\|_h\big)\lesssim h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h\end{aligned}\tag{5.5}$$

by (2.17), (4.14) and (4.28),

$$-\big(y_d,(\Pi_{h,\rho}\bar{y}-\bar{y}_h)-E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big)\lesssim h^2\|\Pi_{h,\rho}\bar{y}-\bar{y}_h\|_h\le h^2\big(\|\Pi_{h,\rho}\bar{y}-\bar{y}\|_h+\|\bar{y}-\bar{y}_h\|_h\big)\lesssim h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h \tag{5.6}$$

by (2.17), (4.14) and (4.25).

Putting (5.1)–(5.6) together with Young's inequality, we arrive at the preliminary estimate

$$\|\bar{y}-\bar{y}_h\|_h^2\lesssim h^{2\alpha}+\big[\mathcal{A}\big(\bar{y},E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big)-\big(y_d,E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big)\big].\tag{5.7}$$

Note that the second term on the right-hand side of (5.7) only involves the bilinear form $\mathcal{A}(\cdot,\cdot)$, which is defined on the continuous level.

6. Error estimates

It follows from (2.7) that

$$\mathcal{A}\big(\bar{y},E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big)-\big(y_d,E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big)=\int_\Omega E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)d\mu+\int_\Omega\lambda\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx.\tag{6.1}$$

The first (resp., second) integral on the right-hand side of (6.1) measures the discretization error due to the state (resp., control) constraint.

6.1. Discretization errors due to state constraints

Using (2.11)–(2.13), (2.15) and (4.25)–(4.27), we can obtain the estimate

$$\int_\Omega E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)d\mu\le\begin{cases}C(h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h) & \text{if } \alpha<1\\ C_\epsilon(h^{2-\epsilon}+h^{1-\epsilon}\|\bar{y}-\bar{y}_h\|_h) & \text{if } \alpha=1\end{cases},\tag{6.2}$$

where ϵ is any number strictly greater than 0.

Under the additional assumption that supp$\lambda\cap$supp$\mu=\emptyset$, we can take ϵ to be 0 by exploiting (2.19)–(2.20) and the estimate becomes

$$\int_\Omega E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)d\mu\le C(h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h)\qquad\text{for } \alpha\le 1.\tag{6.3}$$

The derivation of (6.2) and (6.3) follows the same steps in [10, Section 4.2] by replacing the operators Π_h and E_h there with the operators $\Pi_{h,\rho}$ and $E_{h,\rho}$ from Section 4 of this paper. We omit the identical arguments.

6.2. Discretization errors due to control constraints

If the discrete control constraint in (3.4) is replaced by

$$Q_{h,\rho}\phi_1\le Q_{h,\rho}(-\Delta_h y_h)\le Q_{h,\rho}\phi_2,$$

then the estimate for the second integral on the right-hand side of (6.1) can again be obtained as in [10, Section 4.1]. Since we use Q_h in (3.4) for a simpler implementation of the discrete problem, it is necessary to modify the arguments as follows.

In view of the fact that $\phi_1<\phi_2$ on $\bar{\Omega}$, we can write

$$\int_\Omega\lambda\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx=\int_\Omega\lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx+\int_\Omega\lambda_2\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx,\tag{6.4}$$

where

$$\lambda_1=\begin{cases}\lambda & \text{if } -\Delta\bar{y}=\phi_1\\ 0 & \text{otherwise}\end{cases}\quad\text{and}\quad\lambda_2=\begin{cases}\lambda & \text{if } -\Delta\bar{y}=\phi_2\\ 0 & \text{otherwise}\end{cases}.\tag{6.5}$$

Note that $\lambda_1=\max(\lambda,0)$ by (2.8) and (2.10), and $\lambda_2=\min(\lambda,0)$ by (2.9) and (2.10). Therefore $\lambda_j\in W^{1,s}(\Omega)$ for any $1\le s<2$ by (2.14) and Lemma 7.6 in [44]. It then follows from a standard interpolation error estimate (cf. [2,3]) that, for $j=1,2$,

$$\|\lambda_j-Q_{h,\rho}\lambda_j\|_{L_2(\Omega)}\le C_\epsilon h^{1-\epsilon}\qquad\forall\,T\in\mathcal{T}_h\text{ and }\epsilon>0.\tag{6.6}$$

Remark 6.1. Under the assumption that supp$\lambda \cap$ supp$\mu = \emptyset$, we have $\lambda_j \in H^1(\Omega)$ by (2.18). Consequently we can take ϵ in (6.6) (and below) to be 0 and remove all dependence on ϵ.

We split the first integral on the right-hand side of (6.4) into

$$\begin{aligned}&\int_\Omega \lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx\\ &\qquad = \int_\Omega \rho\lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx + \int_\Omega(\rho-1)\lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx,\end{aligned} \tag{6.7}$$

and observe that (4.7), (4.14) and (4.29) imply

$$\begin{aligned}\int_\Omega(\rho-1)\lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx &\lesssim h^2\|\lambda_1\|_{L_2(\Omega)}\|\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\|_{L_2(\Omega)}\\ &\lesssim h^2|E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)|_{H^2(\Omega)}\\ &\lesssim h^2(\|\Pi_{h,\rho}\bar{y}-\bar{y}\|_h+\|\bar{y}-\bar{y}_h\|_h)\\ &\lesssim h^{2+\alpha}+h^2\|\bar{y}-\bar{y}_h\|_h \lesssim h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h.\end{aligned} \tag{6.8}$$

In view of (6.5), we can decompose the first integral on the right-hand side of (6.7) into the sum of four integrals:

$$\begin{aligned}&\int_\Omega \rho\lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx\\ &\qquad = \int_\Omega \rho\lambda_1\big[-\Delta(E_{h,\rho}\Pi_{h,\rho}\bar{y}-\bar{y})\big]dx + \int_\Omega \rho\lambda_1(\phi_1-Q_{h,\rho}\phi_1)dx\\ &\qquad\quad + \int_\Omega \rho\lambda_1 Q_{h,\rho}\big[\phi_1+\Delta(E_{h,\rho}\bar{y}_h)\big]dx + \int_\Omega \rho\lambda_1\big[\Delta E_{h,\rho}\bar{y}_h - Q_{h,\rho}(\Delta E_{h,\rho}\bar{y}_h)\big]dx.\end{aligned} \tag{6.9}$$

For $\epsilon > 0$, we have

$$\int_\Omega \rho\lambda_1[-\Delta(E_{h,\rho}\Pi_{h,\rho}\bar{y}-\bar{y})]dx = \int_\Omega \rho(\lambda_1-Q_{h,\rho}\lambda_1)[-\Delta(E_{h,\rho}\Pi_{h,\rho}\bar{y}-\bar{y})]dx \le C_\epsilon h^{1+\alpha-\epsilon} \tag{6.10}$$

by (2.17), (4.10), (4.24), (4.26) and (6.6), and

$$\int_\Omega \rho\lambda_1(\phi_1-Q_\rho\phi_1)dx = \int_\Omega \rho(\lambda_1-Q_{h,\rho}\lambda_1)(\phi_1-Q_{h,\rho}\phi_1)dx \le C_\epsilon h^{2-\epsilon} \le C_\epsilon h^{1+\alpha-\epsilon} \tag{6.11}$$

by (4.10), (4.11) and (6.6), because $\phi_1 \in W^{1,\infty}(\Omega)$ by assumption.

Since $\lambda_1 \ge 0$, the third integral on the right-hand side of (6.9) satisfies

$$\begin{aligned}\int_\Omega \rho\lambda_1 Q_{h,\rho}\big[\phi_1+\Delta(E_{h,\rho}\bar{y}_h)\big]dx &= \int_\Omega \rho\lambda_1 Q_{h,\rho}(\phi_1+\Delta_h\bar{y}_h)dx\\ &= \int_\Omega \rho(Q_{h,\rho}\lambda_1)(\phi_1+\Delta_h\bar{y}_h)dx\\ &= \int_\Omega(\rho-1)(Q_{h,\rho}\lambda_1)(\phi_1+\Delta\bar{y})dx + \int_\Omega(\rho-1)(Q_{h,\rho}\lambda_1)(\Delta_h\bar{y}_h-\Delta\bar{y})dx\\ &\qquad + \int_\Omega(Q_{h,\rho}\lambda_1)Q_h(\phi_1+\Delta_h\bar{y}_h)dx\\ &\le \int_\Omega(\rho-1)(Q_{h,\rho}\lambda_1)(\Delta_h\bar{y}_h-\Delta\bar{y})dx\\ &\lesssim h^2\|\lambda_1\|_{L_2(\Omega)}\|\bar{y}-\bar{y}_h\|_h \lesssim h^\alpha\|\bar{y}-\bar{y}_h\|_h\end{aligned} \tag{6.12}$$

by (1.4), the discrete control constraint in (3.4), (3.7), (4.7), (4.10) and (4.23).

Finally we split the last integral on the right-hand side of (6.9) into

$$\begin{aligned}\int_\Omega \rho\lambda_1\big[\Delta E_{h,\rho}\bar{y}_h - Q_{h,\rho}(\Delta E_{h,\rho}\bar{y}_h)\big]dx &= \int_\Omega \rho\lambda_1[\Delta(E_{h,\rho}\bar{y}_h-\bar{y}) - Q_{h,\rho}\Delta(E_{h,\rho}\bar{y}_h-\bar{y})]dx\\ &\qquad + \int_\Omega \rho\lambda_1(\Delta\bar{y}-Q_{h,\rho}\Delta\bar{y})dx.\end{aligned}$$

We have

$$\int_\Omega \rho\lambda_1(\Delta\bar{y}-Q_{h,\rho}\Delta\bar{y})dx = \int_\Omega \rho(\lambda_1-Q_{h,\rho}\lambda_1)(\Delta\bar{y}-Q_{h,\rho}\Delta\bar{y})dx \le C_\epsilon h^{1+\alpha-\epsilon}$$

by (2.17), (4.10), (4.11) and (6.6), and

$$\begin{aligned}\int_\Omega \rho\lambda_1[\Delta(E_{h,\rho}\bar{y}_h-\bar{y})-Q_{h,\rho}\Delta(E_{h,\rho}\bar{y}_h-\bar{y})]dx &= \int_\Omega \rho(\lambda_1-Q_{h,\rho}\lambda_1)[\Delta(E_{h,\rho}\bar{y}_h-\bar{y})]dx\\ &\lesssim \|\lambda_1-Q_{h,\rho}\lambda_1\|_{L_2(\Omega)}\|\Delta(E_{h,\rho}\bar{y}_h-\bar{y})\|_{L_2(\Omega)}\\ &\le C_\epsilon h^{1-\epsilon}\Big(|E_{h,\rho}(\bar{y}_h-\Pi_{h,\rho}\bar{y})|_{H^2(\Omega)}+|E_{h,\rho}\Pi_{h,\rho}\bar{y}-\bar{y}|_{H^2(\Omega)}\Big)\\ &\le C_\epsilon(h^{1-\epsilon}\|\Pi_{h,\rho}\bar{y}-\bar{y}_h\|_h+h^{1+\alpha-\epsilon})\le C_\epsilon(h^{1-\epsilon}\|\bar{y}-\bar{y}_h\|_h+h^{1+\alpha-\epsilon})\end{aligned}$$

by (2.17), (4.10), (4.14), (4.29) and (6.6). We conclude that

$$\int_\Omega \rho\lambda_1\big[(\Delta E_{h,\rho}\bar{y}_h)-Q_{h,\rho}(\Delta E_{h,\rho}\bar{y}_h)\big]dx\le C_\epsilon(h^{1-\epsilon}\|\bar{y}-\bar{y}_h\|_h+h^{1+\alpha-\epsilon}). \tag{6.13}$$

Putting (6.7)–(6.13) together, we find

$$\int_\Omega \lambda_1\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx\le C(h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h)+C_\epsilon\big(h^{1+\alpha-\epsilon}+h^{1-\epsilon}\|\bar{y}-\bar{y}_h\|_h\big). \tag{6.14}$$

Similarly we have

$$\int_\Omega \lambda_2\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx\le C(h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h)+C_\epsilon\big(h^{1+\alpha-\epsilon}+h^{1-\epsilon}\|\bar{y}-\bar{y}_h\|_h\big). \tag{6.15}$$

It follows from (6.4), (6.14) and (6.15) that

$$\int_\Omega \lambda\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx\le\begin{cases}C(h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h) & \text{if } \alpha<1\\ C_\epsilon(h^{2-\epsilon}+h^{1-\epsilon}\|\bar{y}-\bar{y}_h\|_h) & \text{if } \alpha=1\end{cases}, \tag{6.16}$$

where ϵ is any number strictly greater than 0.

As mentioned in Remark 6.1, under the assumption that $\mathrm{supp}\lambda\cap\mathrm{supp}\mu=\emptyset$ we can replace (6.16) by

$$\int_\Omega \lambda\big[-\Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y}-\bar{y}_h)\big]dx\le C(h^{2\alpha}+h^\alpha\|\bar{y}-\bar{y}_h\|_h)\qquad\text{for}\quad\alpha\le 1. \tag{6.17}$$

6.3. Convergence results

The following theorem is a direct consequence of (5.7), (6.2), (6.16) and Young's inequality.

Theorem 6.2. *We have*

$$\|\bar{y}-\bar{y}_h\|_h\le\begin{cases}Ch^\alpha & \text{if } \alpha<1\\ C_\epsilon h^{1-\epsilon} & \text{if } \alpha=1\end{cases},$$

where α is the index of elliptic regularity in (2.17) and ϵ is any number strictly greater than 0.

The following corollary is obtained by using the Poincaré–Friedrichs (resp., Sobolev) inequality for piecewise H^2 functions in [45] (resp., [46]).

Corollary 6.3. *We have*

$$\|\bar{y}-\bar{y}_h\|_{H^1(\Omega)}+\|\bar{y}-\bar{y}_h\|_{L_\infty(\Omega)}\le\begin{cases}Ch^\alpha & \text{if } \alpha<1\\ C_\epsilon h^{1-\epsilon} & \text{if } \alpha=1\end{cases},$$

where α is the index of elliptic regularity in (2.17) and ϵ is any number strictly greater than 0.

The optimal control $\bar{u}=-\Delta\bar{y}$ can be approximated by $\bar{u}_h=-\Delta_h\bar{y}$ and the following corollary is immediate.

Corollary 6.4. *We have*

$$\|\bar{u}-\bar{u}_h\|_{L_2(\Omega)}\le\begin{cases}Ch^\alpha & \text{if } \alpha<1\\ C_\epsilon h^{1-\epsilon} & \text{if } \alpha=1\end{cases},$$

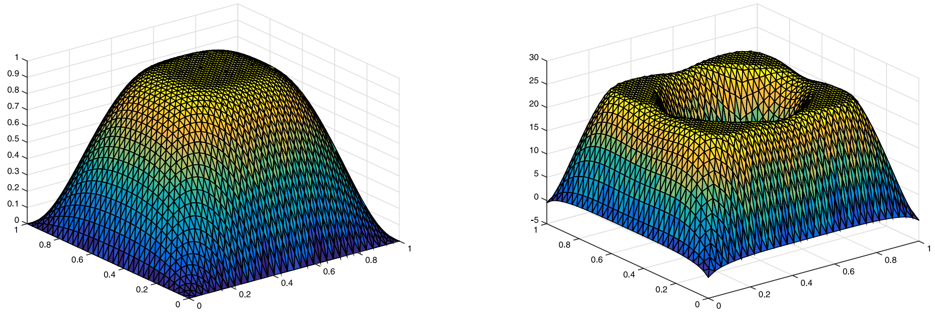
where α is the index of elliptic regularity in (2.17) and ϵ is any number strictly greater than 0.

In the case where $\mathrm{supp}\lambda\cap\mathrm{supp}\mu=\emptyset$, we can use (6.3) and (6.17) to improve these error estimates.

Theorem 6.5. *Under the assumption that* $\mathrm{supp}\lambda\cap\mathrm{supp}\mu=\emptyset$*, we have*

$$\|\bar{y}-\bar{y}_h\|_h+\|\bar{y}-\bar{y}_h\|_{H^1(\Omega)}+\|\bar{y}-\bar{y}_h\|_{L_\infty(\Omega)}+\|\bar{u}-\bar{u}_h\|_{L_2(\Omega)}\le Ch^\alpha,$$

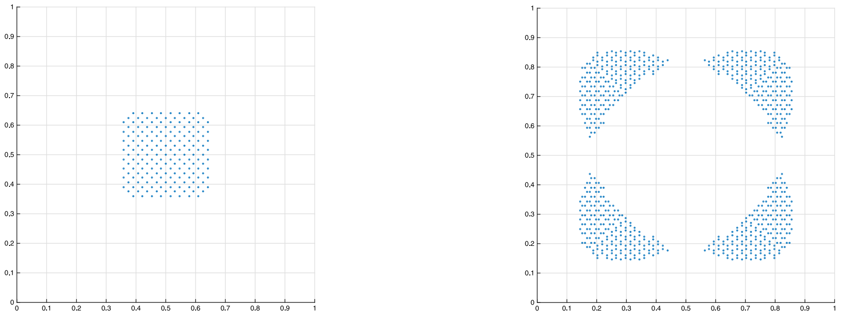
where $\alpha\le 1$ is the index of elliptic regularity in (2.17).



(a) $\bar{y}_h$

(b) $\bar{u}_h$

Fig. 7.1. Graphs of $\bar{y}_h$ and $\bar{u}_h$ from Example 7.1 with $h = 2^{-5}$.



(a) The active set for the upper bound of the state

(b) The active set for the upper bound of the control

Fig. 7.2. The discrete active sets from Example 7.1 with $h = 2^{-5}$.

Remark 6.6. Since $\|\cdot\|_{L_2(\Omega)}$, $|\cdot|_{H^1(\Omega)}$ and $\|\cdot\|_{L_\infty(\Omega)}$ are lower order norms, the error estimates in these norms are not expected to be sharp. This is confirmed by the numerical results in Section 7.

7. Numerical results

In the first two examples, which are taken from [10], we solve the optimal control problem defined by (1.1)–(1.4) (or equivalently the fourth order variational inequality defined by (2.2), (2.3) and (2.6)). Since the exact solutions for these examples are not available, the errors are estimated by comparing the solutions on consecutive levels. In the last two examples, we solve more general optimal control problems where the exact solutions are available, so that the errors can be computed directly. We take $\sigma = 10^6$ and use uniform meshes in all the computations. The discrete variational inequalities are solved by a primal–dual active set method [21–24].

The errors in the tables are defined as follows:

$$e_h = \|\bar{y} - \bar{y}_h\|_h, \quad e_{1,h} = |\bar{y} - \bar{y}_h|_{H^1(\Omega)}, \quad e_{\infty,h} = \max_{p \in \mathcal{V}_h} |\bar{y}(p) - \bar{y}_h(p)| \quad \text{and} \quad e_{0,h} = \|\bar{y} - \bar{y}_h\|_{L_2(\Omega)},$$

where $\mathcal{V}_h$ is the set of the vertices of $\mathcal{T}_h$.

Example 7.1. This is Example 5.1 in [10], where $\Omega = (0, 1)^2$,

$$\beta = 10^{-3}, \; y_d = 2, \; \psi_1 = -\infty, \; \psi_2 = 1, \; \phi_1 = -1 \text{ and } \phi_2 = 25.$$

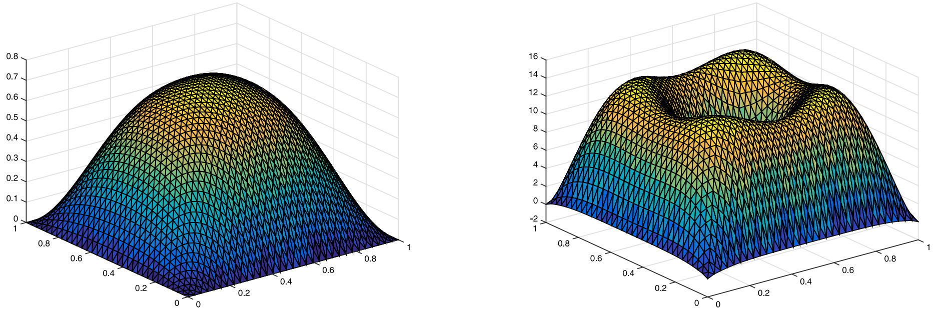
The Slater condition (2.5) is satisfied by $y = 0$.

The graphs of the discrete optimal state and optimal control are displayed in Fig. 7.1. The active sets for ψ_1 and ϕ_1 are empty, and the discrete active sets for ψ_2 and ϕ_2 are presented in Fig. 7.2. Both figures match the corresponding figures in [10].

Since $\operatorname{supp}\lambda \cap \operatorname{supp}\mu = \emptyset$, Theorem 6.5 predicts that the magnitude of e_h is $O(h^\alpha)$. Note also that $\phi_1 < 0 < \phi_2$ on $\partial\Omega$, and hence $\alpha = 1$ by (2.21) and elliptic regularity [26, Section 5.1]. From the numerical results in Table 7.1, we observe $O(h)$ convergence for e_h and better than $O(h)$ convergence for the lower order norms. The errors for $h = 2^{-5}$ are comparable to

Table 7.1
Estimated errors for Example 7.1.

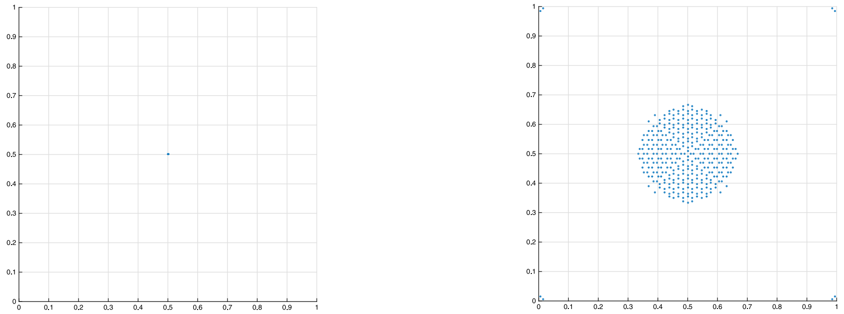
h	e_h	Order	$e_{1,h}$	Order	$e_{\infty,h}$	Order	$e_{0,h}$	Order
2^{-1}	2.9849e−1	–	6.618e−1	–	1.5976e−1	–	8.9645e−2	–
2^{-2}	8.3961e−2	1.77	8.9707e−2	2.90	1.5322e−2	3.38	7.4102e−3	3.60
2^{-3}	3.9797e−2	1.07	2.2393e−2	2.00	4.0938e−3	1.90	1.7041d−3	2.12
2^{-4}	1.9308e−2	1.04	5.2152e−3	2.10	5.2191e−4	2.97	1.6135e−4	3.40
2^{-5}	9.4405e−3	1.03	1.3166e−3	1.99	1.5938e−4	1.71	8.1776e−5	0.98



(a) $\bar{y}_h$

(b) $\bar{u}_h$

Fig. 7.3. Graphs of $\bar{y}_h$ and $\bar{u}_h$ from Example 7.2 with $h = 2^{-5}$.



(a) The active set for the upper bound of the state

(b) The active set for the lower bound of the control

Fig. 7.4. The discrete active sets from 7.2 with $h = 2^{-5}$.

the corresponding errors in [10, Example 5.1] for $h = 2^{-8}$. Therefore the method in this paper outperforms the method in [10] because the number of global dofs of the cubic finite element space is only (roughly) 25% more than that of the Morley finite element space.

Example 7.2. This is Example 5.2 in [10], where $\Omega = (0,1)^2$,

$$\beta = 10^{-3},\ y_d = 1,\ \psi_1 = -\infty,\ \psi_2 = 4(x_1 - x_1^2)(x_2 - x_2^2),\ \phi_2 = 100,$$

and

$$\phi_1 = \begin{cases} 8\exp\left(\dfrac{|x-(0.5,0.5)|^2}{|x-(0.5,0.5)|^2 - 0.25}\right) & \text{if } |x-(0.5,0.5)| \le 0.5 \\ 0 & \text{otherwise} \end{cases}.$$

The Slater condition (2.5) is satisfied by $y = 9(x_1 - x_1^2)(x_2 - x_2^2)$.

The graphs of the discrete optimal state and optimal control are presented in Fig. 7.3. The active set for ψ_1 and ϕ_2 are empty, and the discrete active sets for ψ_2 (the singleton $\{(0.5, 0.5)\}$) and ϕ_1 are displayed in Fig. 7.4. Both figures match the corresponding figures in [10].

Since $\phi_1 < 0 < \phi_2$ on $\partial\Omega$, the optimal state $\bar{y} \in H^3(\Omega)$ by (2.21) and elliptic regularity. Therefore the convergence rate for e_h predicted by Theorem 6.2 is $O(h^{1-\epsilon})$, which is observed in Table 7.2. The rate of convergence in the lower order

Table 7.2
Estimated errors for Example 7.2.

h	e_h	Order	$e_{1,h}$	Order	$e_{\infty,h}$	Order	$e_{0,h}$	order
2^{-1}	1.1139e−1	–	2.4540e−1	–	5.8482e−2	–	3.2612e−2	–
2^{-2}	3.4974e−2	1.67	4.3529e−2	2.50	9.8614e−3	2.57	4.80403−3	2.76
2^{-3}	1.6461e−2	1.09	7.8545e−3	2.47	6.4930e−4	3.92	3.3728e−4	3.83
2^{-4}	8.7523e−3	0.91	2.2021e−3	1.83	2.6570e−4	1.29	9.1899e−5	1.88
2^{-5}	4.5458e−3	0.95	5.5334e−4	1.99	6.6175e−5	2.01	2.2178e−5	2.05

norms are better than the results in Corollary 6.3. Again the errors for $h = 2^{-5}$ are comparable to the corresponding errors in [10, Example 5.2] for $h = 2^{-8}$.

In order to test our method on examples where the exact solutions are available, we consider the following more general optimal control problem: Given $f, u_d, y_d \in L_2(\Omega)$,

$$\text{find} \quad (\bar{y}, \bar{u}) = \underset{(y,u)\in\mathbb{K}}{\operatorname{argmin}} \frac{1}{2}\Big(\|y - y_d\|^2_{L_2(\Omega)} + \beta\|u - u_d\|^2_{L_2(\Omega)}\Big), \tag{7.1}$$

where $(y, u) \in \mathbb{K} \subset H^1_0(\Omega) \times L_2(\Omega)$ if and only if

$$\int_\Omega \nabla y \cdot \nabla z\,dx = \int_\Omega (u + f)z\,dx \qquad \forall z \in H^1_0(\Omega) \tag{7.2}$$

and the constraints (1.3)–(1.4) are satisfied.

By elliptic regularity, the problem (7.1) can be reformulated as follows:

$$\text{Find} \quad \bar{y} = \underset{y\in K}{\operatorname{argmin}} \frac{1}{2}\Big(\|y - y_d\|^2_{L_2(\Omega)} + \beta\|\Delta y + f + u_d\|^2_{L_2(\Omega)}\Big), \tag{7.3}$$

where K is given by (2.2). It follows from the classical theory that (7.3) has a unique solution provided K is nonempty. The corresponding fourth order variational inequality is to find $\bar{y} \in K$ such that

$$\mathcal{A}(\bar{y}, y - \bar{y}) - (y_d, y - \bar{y}) + \beta(f + u_d, \Delta(y - \bar{y})) \ge 0 \qquad \forall y \in K, \tag{7.4}$$

where $\mathcal{A}(\cdot,\cdot)$ is given by (2.3), and the KKT conditions are given by

$$\mathcal{A}(\bar{y}, z) - (y_d, z) + \beta(f + u_d, \Delta z) = \int_\Omega z\,d\mu + \int_\Omega \lambda(-\Delta z)dx \qquad \forall z \in H^2(\Omega) \cap H^1_0(\Omega) \tag{7.5}$$

together with (2.8)–(2.13). The finite element method for (7.4) is to find $\bar{y}_h \in K_h$ such that

$$\mathcal{A}_h(\bar{y}_h, y_h - \bar{y}_h) - (y_d, y_h - \bar{y}_h) + \beta(f + u_d, \Delta_h(y_h - \bar{y}_h)) \ge 0 \qquad \forall y_h \in K_h, \tag{7.6}$$

where $\mathcal{A}_h(\cdot,\cdot)$ is given by (3.6) and

$$K_h = \{y_h \in V_h : \ I_h\psi_1 \le I_h y_h \le I_h\psi_2 \quad \text{and} \quad Q_h\phi_1 \le Q_h(-\Delta_h y_h - f) \le Q_h\phi_2\}. \tag{7.7}$$

Under the condition that $f + u_d \in H^1_0(\Omega)$, we can rewrite (7.4) as

$$\mathcal{A}(\bar{y}, y - \bar{y}) - (y_d, y - \bar{y}) - \beta\int_\Omega \nabla(f + u_d)\cdot\nabla(y - \bar{y})dx \ge 0 \qquad \forall y \in K, \tag{7.8}$$

and replace (7.5) by

$$\mathcal{A}(\bar{y}, z) - (y_d, z) - \beta\int_\Omega \nabla(f + u_d)\cdot\nabla z\,dx = \int_\Omega z\,d\mu + \int_\Omega \lambda(-\Delta z)dx \qquad \forall z \in H^2(\Omega) \cap H^1_0(\Omega). \tag{7.9}$$

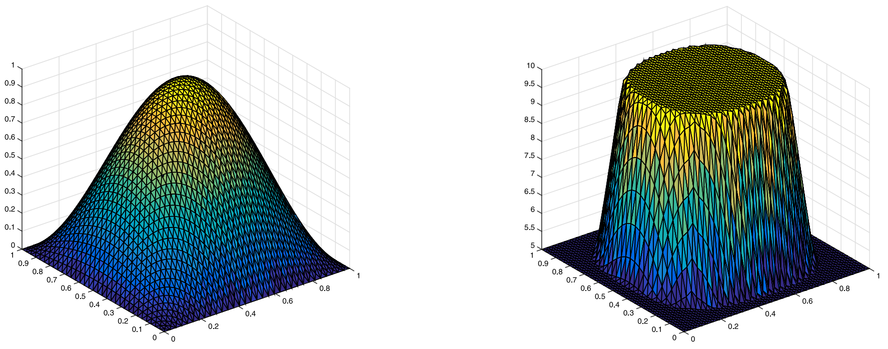
The finite element method for (7.8) is to find $\bar{y}_h \in K_h$ such that

$$\mathcal{A}_h(\bar{y}_h, y_h - \bar{y}_h) - (y_d, y_h - \bar{y}_h) - \beta\int_\Omega \nabla(f + u_d)\cdot\nabla(y_h - \bar{y}_h)dx \ge 0 \qquad \forall y_h \in K_h. \tag{7.10}$$

Example 7.3. Let $\Omega = (0, 1)^2$. We consider (7.1)/(7.8) with the data

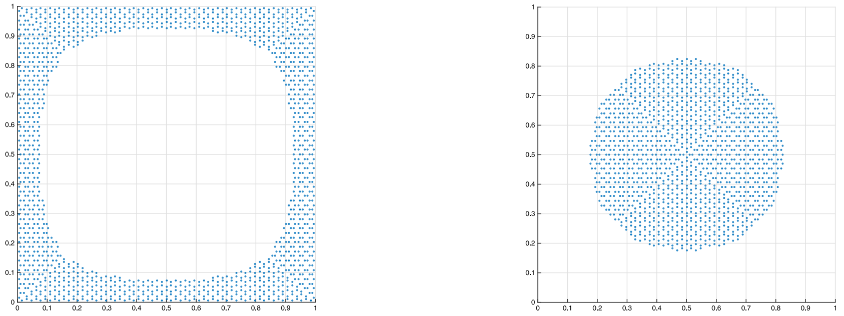
$$\begin{aligned}
&\beta = 1, \ \psi_1 = -\infty, \ \psi_2 = \infty, \phi_1 = 5, \ \phi_2 = 10,\\
&f = \min\{0, 2\pi^2\sin(\pi x_1)\sin(\pi x_2) - 5\} + \max\{0, 2\pi^2\sin(\pi x_1)\sin(\pi x_2) - 10\},\\
&y_d = \sin(\pi x_1)\sin(\pi x_2), \qquad u_d = 2\pi^2\sin(\pi x_1)\sin(\pi x_2) - f.
\end{aligned}$$

The exact solution is $\bar{y} = \sin(\pi x_1)\sin(\pi x_2)$ and $\bar{u} = -\Delta\bar{y} - f = u_d$.



(a) $\bar{y}_h$ (b) $\bar{u}_h$

Fig. 7.5. Graphs of $\bar{y}_h$ and $\bar{u}_h$ from Example 7.3 with $h = 2^{-5}$.



(a) The active set for the lower bound of the control (b) The active set for the upper bound of the control

Fig. 7.6. The discrete active sets from Example 7.3 with $h = 2^{-5}$.

Table 7.3
Errors for Example 7.3.

h	e_h	Order	$e_{1,h}$	Order	$e_{\infty,h}$	Order	$e_{0,h}$	Order
2^{-1}	8.6875e−1	–	4.7617e−2	–	1.4409e−3	–	4.8153e−3	–
2^{-2}	2.2394e−1	1.96	6.2988e−3	2.92	4.8527e−4	1.57	4.4907e−4	3.42
2^{-3}	5.6528e−2	1.99	7.8618e−4	3.00	5.7519e−5	3.08	3.7587e−5	3.58
2^{-4}	1.4166e−2	2.00	9.8563e−5	3.00	7.9290e−6	2.86	3.9788e−6	3.24
2^{-5}	3.5446e−3	2.00	1.5045e−5	2.71	3.9227e−6	1.02	1.7779e−6	1.16

This is an example with only control constraints so that μ does not appear in the KKT conditions. (A similar example can be found in [47].) Since $f + u_d = -\Delta\bar{y} \in H_0^1(\Omega)$, we use the formulation (7.10) in the computations.

The graphs of the discrete optimal state and optimal control are shown in Fig. 7.5, and the active sets for ϕ_1 and ϕ_2 are displayed in Fig. 7.6. They capture accurately their exact counterparts.

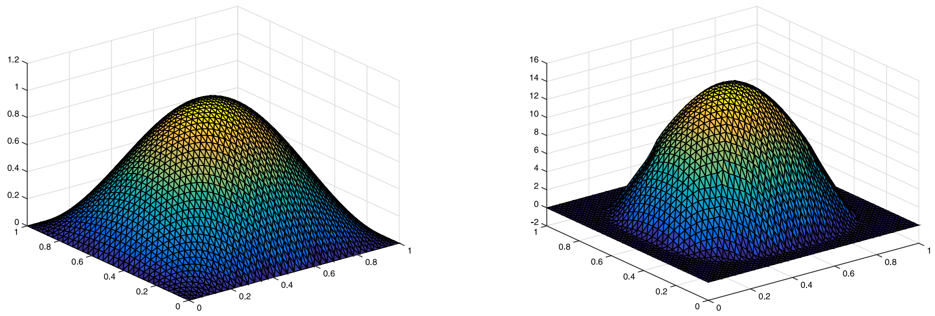
It is straightforward to check that (7.9) holds for $\lambda = 0$, and the analysis in Sections 5 and 6 can be extended to the variational inequality (7.8) because we can estimate $|\Pi_h\zeta - \Pi_{h,\rho}\zeta|_{H^1(\Omega)}$ (resp., $|v - E_h v|_{H^1(\Omega)}$ and $|\zeta - E_{h,\rho}\Pi_{h,\rho}\zeta|_{H^1(\Omega)}$) by (4.16) (resp., (4.25) and (4.26)). Therefore for this example Theorem 6.5 holds with $\alpha = 2$. The $O(h^2)$ convergence rate of e_h is observed in Table 7.3. The convergence rate for the lower norms are higher up to $h = 2^{-4}$, before round-off errors due to ill-conditioning take effect.

Example 7.4. This is the example in [6, Section 6], with $\Omega = (0, 1)^2$ and the exact solution is

$$\bar{y}(x) = \sin(\pi x_1)\sin(\pi x_2), \quad \bar{u} = \max(-\Delta\bar{y} - \kappa, 0),$$

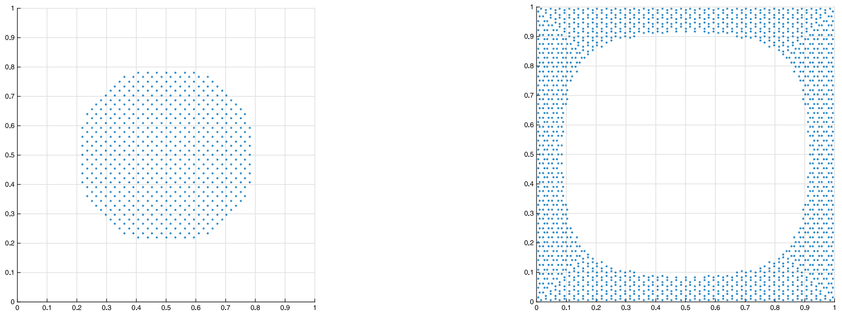
where $\kappa = 5$. The data for (7.1)/(7.4) are given by

$$\beta = 0.1,\ \phi_1 = 0,\ \phi_2 = 100,\ \psi_2 = 100,$$



(a) $\bar{y}_h$ (b) $\bar{u}_h$

Fig. 7.7. Graphs of $\bar{y}_h$ and $\bar{u}_h$ from Example 7.4 with $h = 2^{-5}$.



(a) The active set for the lower bound of the state (b) The active set for the lower bound of the control

Fig. 7.8. Discrete active sets from Example 7.4 with $h = 2^{-5}$.

$$\psi_1 = \begin{cases} \bar{y}(x) & \text{if } \bar{y}(x) \geq c \\ 2\bar{y}(x) - c & \text{if } \bar{y}(x) \leq c \end{cases}, \qquad f = -\Delta\bar{y} - \bar{u},$$

$$y_d(x) = \begin{cases} \bar{y}(x) - 1 & \text{if } \bar{y}(x) > c \\ \beta\Delta^2\bar{y}(x) + \bar{y}(x) & \text{if } \bar{y}(x) < c \end{cases}, \quad u_d(x) = \begin{cases} \bar{u}(x) - 2\pi^2 c & \text{if } \bar{y}(x) > c \\ -\kappa & \text{if } \bar{y}(x) < c \end{cases},$$

where $c = 0.6$. We use the formulation (7.6) in the computations.

The graphs of the discrete optimal state and optimal control are displayed in Fig. 7.7. The active sets for ψ_2 and ϕ_2 are empty and the discrete active sets for ψ_1 and ϕ_1 are presented in Fig. 7.8. Both figures capture their exact counterparts accurately.

It is straightforward to check that (7.5) holds with $\lambda \in H^1(\Omega)$ given by

$$\lambda = \begin{cases} -\Delta\bar{y} - \kappa & \text{if } -\Delta\bar{y} - \kappa \leq 0 \\ 0 & \text{otherwise} \end{cases},$$

and $\mu \in H^{-1}(\Omega)$ given by

$$\int_\Omega z\,d\mu = \int_{\bar{y}>c} z\,dx - 2\pi^2\beta \int_{\bar{y}=c} z\frac{\partial\bar{y}}{\partial n}\,ds,$$

where n is the unit outer normal on the boundary of the domain defined by $\bar{y} > c$ (cf. (a) of Fig. 7.8).

For this example we have $\tau = f + u_d \in H^1(\Omega)$ and the analysis in Sections 5–6 can be extended to the variational inequality (7.6) by using the estimate

$$\begin{aligned} &\big(\tau, \Delta_h(\Pi_h\bar{y} - \bar{y}_h)\big) - \big(\tau, \Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big) \\ &\quad = \big(\tau, \Delta_h(\Pi_h\bar{y} - \Pi_{h,\rho}\bar{y})\big) + \big((1-\rho)\tau, \Delta_h(\Pi_{h,\rho}\bar{y} - \bar{y}_h) - \Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big) \\ &\qquad + \big(\rho(\tau - Q_{h,\rho}\tau), \Delta_h(\Pi_{h,\rho}\bar{y} - \bar{y}_h) - \Delta E_{h,\rho}(\Pi_{h,\rho}\bar{y} - \bar{y}_h)\big) \\ &\quad \lesssim h^{2+\alpha} + (h^2 + h)\|\bar{y} - \bar{y}_h\|_h \end{aligned}$$

Table 7.4
Errors for Example 7.4.

h	e_h	Order	$e_{1,h}$	Order	$e_{\infty,h}$	Order	$e_{0,h}$	Order
2^{-1}	6.0834e−1	–	1.5616e−1	–	1.5392e−2	–	2.0150e−2	–
2^{-2}	2.7959e−1	1.12	5.4201e−2	1.53	1.1094e−2	0.47	5.1841e−3	1.96
2^{-3}	1.3699e−1	1.03	2.6974e−2	1.01	6.6380e−3	0.74	2.5055e−3	1.05
2^{-4}	8.4630e−2	0.70	1.5003e−2	0.85	3.2834e−3	1.02	1.3833e−3	0.86
2^{-5}	4.1034e−2	1.04	6.4865e−3	1.21	1.1477e−3	1.52	5.7424e−4	1.27

that follows from (2.17), (4.7), (4.11), (4.16), (4.23) and (4.29). Consequently we can apply Theorem 6.5 to this example with $\alpha = 1$. The $O(h)$ convergence of e_h is observed in Table 7.4, and the convergence rates in the lower order norms are higher.

8. Concluding remarks

We have developed a modified cubic Hermite finite element method for the optimal control problem defined by (1.1)–(1.4). By using the mean value of the Laplacian of a shape function as a dof of the modified cubic Hermite element, the resulting discrete variational inequality is a quadratic programming problem with box constraints that can be solved efficiently by a primal–dual active set method. This method performs better than the Morley finite element method in [10], and it is also less expensive than either a C^0 interior penalty method based on the cubic Lagrange element or a C^1 finite element method based on the Hsieh–Clough–Tocher element.

The cubic Hermite finite element method can be extended to three dimensional domains by adding a quartic bubble function on each tetrahedron. It can also be extended to the following problem where (1.1) is replaced by

$$(\bar{y}, \bar{u}) = \operatorname*{argmin}_{(y,u)\in\mathbb{K}} \frac{1}{2}\Big(\|y - y_d\|^2_{L_2(\Omega)} + \beta\|u\|^2_{L_2(\omega)}\Big)$$

for a subdomain ω of Ω, and the constraints (1.2) and (1.4) are replaced by

$$\int_\Omega \nabla y \cdot \nabla z\, dx = \int_\omega uz\, dx \qquad \forall z \in H^1_0(\Omega)$$

and $\phi_1 \le u \le \phi_2$ a.e. in ω.

These extensions and the adaptive version of the cubic finite element method are ongoing projects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Properties of the interpolation operators

Let p, a corner of Ω, be the common vertex of $e_{p,1}, e_{p,2} \in \mathcal{E}^b_h$, which are edges of the triangles $T_{p,1}, T_{p,2} \in \mathcal{T}_h$ (that may coincide). Observe that, for $\zeta \in H^2(\Omega) \cap H^1_0(\Omega)$, the definition (4.4) is equivalent to

$$\frac{\partial(\Pi_h\zeta)}{\partial t_{e_{p,i}}}(p) = \frac{\partial(\Pi^L_{h,T_{p,i}}\zeta)}{\partial t_{e_{p,i}}}(p) = \frac{\partial(\Pi_{h,\rho}\zeta)}{\partial t_{e_{p,i}}}(p) \qquad \text{for } i = 1, 2. \tag{A.1}$$

We can use (4.2), (4.3), (4.5), (4.6) and (A.1) to extend the definitions of Π_h and $\Pi_{h,\rho}$ to $H^2(\Omega)$, and the interpolation error estimate (4.13) (for the extended Π_h and $\Pi_{h,\rho}$) follows immediately from the Bramble–Hilbert lemma [41,42], since $\Pi_h\zeta = \zeta = \Pi_{h,\rho}\zeta$ if ζ is a cubic polynomial on S_T.

The estimate (4.14) follows from (3.7), (4.13) and the trace inequality with scaling:

$$|e|^{-1}\big\|[\![\partial z/\partial n]\!]\big\|^2_{L_2(e)} \lesssim \sum_{T\in\mathcal{T}_e}\big(h_T^{-2}|z|^2_{H^1(T)} + |z|^2_{H^2(T)}\big) \qquad \forall e \in \mathcal{E}^i_h, \tag{A.2}$$

where z is any piecewise H^2 function and $\mathcal{T}_e$ is the set of the two triangles in $\mathcal{T}_h$ that share e as a common edge.

Note that we also have

$$\|\zeta - \Pi_{h,\rho}\zeta\|_{L_2(T)} \lesssim h_T^2|\zeta|_{H^2(S_T)} \qquad \forall \zeta \in H^2(\Omega),\ T \in \mathcal{T}_h,$$

which implies

$$\|\zeta - \Pi_{h,\rho}\zeta\|_h \lesssim \|\zeta\|_{H^2(\Omega)} \qquad \forall \zeta \in H^2(\Omega) \tag{A.3}$$

by (3.7), (A.2) and standard inverse estimates [2,3]. The estimate (4.15) follows from (3.10), (A.3) and the triangle inequality.

Finally we turn to the estimates (4.16) and (4.17). Let $P_3(T)$ be the space of cubic polynomials on T. We have, by scaling,

$$\|v\|_{L_2(T)}^2 \approx \sum_{i=1}^{3}\Big(h_T^2[v(p_i)]^2 + h_T^4|(\nabla v)(p_i)|^2\Big) + h_T^2\Big(\int_T(\Delta v)dx\Big)^2 \qquad \forall\, v \in P_3(T),\ T \in \mathcal{T}_h, \tag{A.4}$$

and

$$\int_T \Delta(\Pi_h\zeta - \Pi_{h,\rho}\zeta)dx = \int_T(1-\rho_T)\Delta(\zeta - \Pi_{h,\rho}\zeta)dx \lesssim h_T^3\|\Delta(\zeta - \Pi_{h,\rho}\zeta)\|_{L_2(T)} \tag{A.5}$$

by (4.5)–(4.7).

The case of $k = 0$ in (4.16) follows immediately from (A.4) and (A.5). The rest of the estimates in (4.16) and the estimate (4.17) then follows from standard inverse estimates, (3.7) and (A.2).

Appendix B. Properties of the enriching operator

We have, by scaling,

$$\|v\|_{L_2(T)}^2 \approx \sum_{i=1}^{3}\Big(h_T^2[v(p_i)]^2 + h_T^4|(\nabla v)(p_i)|^2 + h_T^2\Big[\int_e(\partial v/\partial n)ds\Big]^2\Big) + h_T^2\Big(\int_T \rho_T(\Delta v)dx\Big)^2 \tag{B.1}$$

for all $v \in \tilde{\Sigma}_{\mathrm{HCT}}(T)$ and $T \in \mathcal{T}_h$.

Let $\mathcal{E}_T$ be the set of the edges of T that are interior to Ω. It follows from (4.18)–(4.22) and (B.1) that

$$\begin{aligned}\|v - E_{h,\rho}v\|_{L_2(T)}^2 &\approx \sum_{e\in\mathcal{E}_T} h_T^2\left[\int_e\left(\frac{\partial v_T}{\partial n_e} - \frac{\partial(E_{h,\rho}v)}{\partial n_e}\right)ds\right]^2\\ &= \frac{h_T^2}{4}\sum_{e\in\mathcal{E}_T}\left[\int_e [\![\partial v/\partial n_e]\!]\,ds\right]^2 \lesssim h_T^4\sum_{e\in\mathcal{E}_T}|e|^{-1}\|\,[\![\partial v/\partial n_e]\!]\,\|_{L_2(e)}^2.\end{aligned} \tag{B.2}$$

The estimate (4.25) follows from (B.2) and standard inverse estimates.

Note that we can use (4.18)–(4.22) to extend $E_{h,\rho}$ to an operator that maps the modified Hermite finite element space without any boundary condition to $H^2(\Omega)$. For the extended operators $\Pi_{h,\rho}$ and $E_{h,\rho}$, both (4.26) and (4.27) follow from the Bramble–Hilbert lemma since $E_{h,\rho}\Pi_{h,\rho}\zeta = \zeta$ on T if ζ is a cubic polynomial on $\tilde{S}_T$, the union of all the triangles that share at least a vertex with one of the triangles in S_T.

In order to establish (4.28), it suffices to show that

$$|a_h(\Pi_{h,\rho}\zeta, v) - a(\zeta, E_{h,\rho}v)| \lesssim h^s\|\zeta\|_{H^{2+s}(\Omega)}\|v\|_h \tag{B.3}$$

for all $\zeta \in H^{2+s}(\Omega)\cap H_0^1(\Omega)$ and $v \in V_h$. Indeed, it follows from (3.7), (4.13), (4.25) and (B.3) that

$$\begin{aligned}&|\mathcal{A}_h(\Pi_{h,\rho}\zeta, v) - \mathcal{A}(\zeta, E_{h,\rho}v)|\\ &\qquad = \big|\beta[a_h(\Pi_{h,\rho}\zeta, v) - a(\zeta, E_{h,\rho}v)] + (\Pi_{h,\rho}\zeta - \zeta, v) + (\zeta, v - E_{h,\rho}v)\big|\\ &\qquad \lesssim h^s\|\zeta\|_{H^{2+s}(\Omega)}\|v\|_h + \|\Pi_{h,\rho}\zeta - \zeta\|_{L_2(\Omega)}\|v\|_{L_2(\Omega)} + \|\zeta\|_{L_2(\Omega)}\|v - E_{h,\rho}v\|_{L_2(\Omega)} \lesssim h^s\|\zeta\|_{H^{2+s}(\Omega)}\|v\|_h\end{aligned}$$

for all $\zeta \in H^{2+s}(\Omega)\cap H_0^1(\Omega)$, $v \in V_h$ and $0 \le s \le 2$.

The proof of (B.3) proceeds as in [40, Appendix B], where a quadratic C^0 interior penalty method was treated. It begins with the formula

$$\begin{aligned}&a_h(\Pi_{h,\rho}\zeta, v) - a(\zeta, E_{h,\rho}v)\\ &\quad = \sum_{T\in\mathcal{T}_h}\left(\int_T D^2(\Pi_{h,\rho}\zeta) : D^2(v - E_{h,\rho}v)dx + \int_T D^2(\Pi_{h,\rho}\zeta - \zeta) : D^2(E_{h,\rho}v)dx\right)\\ &\qquad + \sum_{e\in\mathcal{E}_h^i}\int_e\left(\left\{\!\!\left\{\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(v - E_{h,\rho}v)}{\partial n}\right]\!\!\right] + \left\{\!\!\left\{\frac{\partial^2 v}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(\Pi_{h,\rho}\zeta - \zeta)}{\partial n}\right]\!\!\right]\right)ds\\ &\qquad\quad + \sum_{e\in\mathcal{E}_h^i}\frac{\sigma}{|e|}\int_e\left[\!\!\left[\frac{\partial(\Pi_{h,\rho}\zeta - \zeta)}{\partial n}\right]\!\!\right]\left[\!\!\left[\frac{\partial v}{\partial n}\right]\!\!\right]ds\end{aligned} \tag{B.4}$$

that follows from (3.1) and the fact that $[\![\partial(E_{h,\rho}v)/\partial n]\!] = 0 = [\![\partial\zeta/\partial n]\!]$ across $e \in \mathcal{E}_h^i$.

Using (4.13), (4.25), (4.29), (A.2), the Cauchy–Schwarz inequality and standard inverse estimates, we obtain

$$\Big|\sum_{T\in\mathcal{T}_h}\int_T D^2(\Pi_{h,\rho}\zeta-\zeta):D^2(E_{h,\rho}v)dx\Big|+\Big|\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2 v}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(\Pi_{h,\rho}\zeta-\zeta)}{\partial n}\right]\!\!\right]ds\Big| \\ +\Big|\sum_{e\in\mathcal{E}_h^i}\frac{\sigma}{|e|}\int_e\left[\!\!\left[\frac{\partial(\Pi_{h,\rho}\zeta-\zeta)}{\partial n}\right]\!\!\right]\left[\!\!\left[\frac{\partial v}{\partial n}\right]\!\!\right]ds\Big|\lesssim h^s\|\zeta\|_{H^{2+s}(\Omega)}\|v\|_h. \tag{B.5}$$

In view of Remark 4.4 and the integration by parts formula

$$\int_T D^2w:D^2v\,dx=\int_{\partial T}\left(\frac{\partial^2 w}{\partial n^2}\frac{\partial v}{\partial n}+\frac{\partial^2 w}{\partial n\partial t}\frac{\partial v}{\partial t}-\frac{\partial(\Delta w)}{\partial n}v\right)ds+\int_T(\Delta^2 w)v\,dx \tag{B.6}$$

that holds for $w\in H^4(T)$ and $v\in H^2(T)$, we have

$$\int_T D^2(\Pi_{h,\rho}\zeta):D^2(v-E_{h,\rho}v)dx=\int_{\partial T}\left(\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right)\left(\frac{\partial(v-E_{h,\rho}v)}{\partial n}\right)ds.$$

Therefore we can rewrite the sum of the two remaining terms on the right-hand side of (B.4) as

$$\sum_{T\in\mathcal{T}_h}\int_T D^2(\Pi_{h,\rho}\zeta):D^2(v-E_{h,\rho}v)dx+\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(v-E_{h,\rho}v)}{\partial n}\right]\!\!\right]ds \\ =\sum_{T\in\mathcal{T}_h}\int_{\partial T}\left(\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right)\left(\frac{\partial(v-E_{h,\rho}v)}{\partial n}\right)ds+\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(v-E_{h,\rho}v)}{\partial n}\right]\!\!\right]ds \\ =-\sum_{e\in\mathcal{E}_h^i}\int_e\left[\!\!\left[\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right]\!\!\right]\left\{\!\!\left\{\frac{\partial(v-E_{h,\rho}v)}{\partial n}\right\}\!\!\right\}ds. \tag{B.7}$$

Finally, the estimate

$$\left|\sum_{e\in\mathcal{E}_h^i}\int_e\left[\!\!\left[\frac{\partial^2(\Pi_{h,\rho}\zeta)}{\partial n^2}\right]\!\!\right]\left\{\!\!\left\{\frac{\partial(v-E_{h,\rho}v)}{\partial n}\right\}\!\!\right\}ds\right|\lesssim h^s\|\zeta\|_{H^{2+s}(\Omega)}\|v\|_h \tag{B.8}$$

can be derived by the same arguments in [40, Appendix B].

The estimate (B.3) follows from (B.4), (B.5), (B.7) and (B.8).

References

[1] Adams RA, Fournier JJF. Sobolev spaces. 2nd ed.. Amsterdam: Academic Press; 2003.

[2] Ciarlet PG. The finite element method for elliptic problems. Amsterdam: North-Holland; 1978.

[3] Brenner SC, Scott LR. The mathematical theory of finite element methods. 3rd ed.. New York: Springer-Verlag; 2008.

[4] Meyer C. Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints. Control Cybernet 2008;37:51–83.

[5] Rösch A, Wachsmuth D. A posteriori error estimates for optimal control problems with state and control constraints. Numer Math 2012;120:733–62.

[6] Cherednichenko S, Rösch A. Error estimates for the regularization of optimal control problems with pointwise control and state constraints. Z Anal Anwend 2008;27:195–212.

[7] Cherednichenko S, Rösch A. Error estimates for the discretization of elliptic control problems with pointwise control and state constraints. Comput Optim Appl 2009;44:27–55.

[8] Hintermüller M, Hinze M. Moreau-Yosida regularization in state constrained elliptic control problems: error estimates and parameter adjustment. SIAM J Numer Anal 2009;47:1666–83.

[9] Hintermüller M, Schiela A, Wollner W. The length of the primal–dual path in Moreau-Yosida-based path-following methods for state constrained optimal control. SIAM J Optim 2014;24:108–26.

[10] Brenner SC, Gudi T, Porwal K, Sung L-Y. A Morley finite element method for an elliptic distributed optimal control problem with pointwise state and control constraints. ESAIM:COCV 2018;24:1181–206.

[11] Pierre M, Sokołowski J. Differentiability of projection and applications. In: Control of partial differential equations and applications (Laredo, 1994). Lecture notes in pure and appl. math., vol. 174, New York: Dekker; 1996, p. 231–40.

[12] Liu W, Gong W, Yan N. A new finite element approximation of a state-constrained optimal control problem. J Comput Math 2009;27:97–114.

[13] Gong W, Yan N. A mixed finite element scheme for optimal control problems with pointwise state constraints. J Sci Comput 2011;46:182–203.

[14] Brenner SC, Sung L-Y, Zhang Y. A quadratic C^0 interior penalty method for an elliptic optimal control problem with state constraints. In: Feng OKX, Xing Y, editors. Recent developments in discontinuous Galerkin finite element methods for partial differential equations. The IMA volumes in mathematics and its applications, vol. 157, Cham-Heidelberg-New York-Dordrecht-London: Springer; 2013, p. 97–132, (2012 John H. Barrett Memorial Lectures).

[15] Brenner SC, Davis CB, Sung L-Y. A partition of unity method for a class of fourth order elliptic variational inequalities. Comput Methods Appl Mech Engrg 2014;276:612–26.

[16] Brenner SC, Sung L-Y, Zhang Y. Post-processing procedures for a quadratic C^0 interior penalty method for elliptic distributed optimal control problems with pointwise state constraints. Appl Numer Math 2015;95:99–117.

[17] Brenner SC, Oh M, Pollock S, Porwal K, Schedensack M, Sharma N. A C^0 interior penalty method for elliptic distributed optimal control problems in three dimensions with pointwise state constraints. In: Brenner S, editor. Topics in numerical partial differential equations and scientific computing. The IMA volumes in mathematics and its applications, vol. 160, Cham-Heidelberg-New York-Dordrecht-London: Springer; 2016, p. 1–22.

[18] Brenner SC, Gedicke J, Sung L-Y. C^0 interior penalty methods for an elliptic distributed optimal control problem on nonconvex polygonal domains with pointwise state constraints. SIAM J Numer Anal 2018;56:1758–85.

[19] Brenner SC, Sung L-Y, Zhang Y. C^0 interior penalty methods for an elliptic state-constrained optimal control problem with Neumann boundary condition. J Comput Appl Math 2019;350:212–32.

[20] Brenner SC, Oh M, Sung L-Y. P_1 Finite element methods for an elliptic state-constrained distributed optimal control problem with Neumann boundary conditions. RINAM 2020. http://dx.doi.org/10.1016/j.rinam.2019.100090, (published online 7 January 2020).

[21] Bergounioux M, Ito K, Kunisch K. Primal–dual strategy for constrained optimal control problems. SIAM J Control Optim 1999;37:1176–94, (electronic).

[22] Bergounioux M, Kunisch K. Primal–dual strategy for state-constrained optimal control problems. Comput Optim Appl 2002;22:193–224.

[23] Hintermüller M, Ito K, Kunisch K. The primal–dual active set strategy as a semismooth Newton method. SIAM J Optim 2003;13:865–88.

[24] Ito K, Kunisch K. Lagrange multiplier approach to variational problems and applications. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2008.

[25] Brenner SC, Gedicke J, Sung L-Y, Zhang Y. An a posteriori analysis of C^0 interior penalty methods for the obstacle problem of clamped Kirchhoff plates. SIAM J Numer Anal 2017;55:87–108.

[26] Grisvard P. Elliptic problems in non smooth domains. Boston: Pitman; 1985.

[27] Dauge M. Elliptic boundary value problems on corner domains. Lecture notes in mathematics, vol. 1341, Berlin-Heidelberg: Springer-Verlag; 1988.

[28] Maz'ya V, Rossmann J. Elliptic equations in polyhedral domains. Providence, RI: American Mathematical Society; 2010.

[29] Grisvard P. Singularities in boundary value problems. Paris: Masson; 1992.

[30] Ekeland I, Témam R. Convex analysis and variational problems. Classics in applied mathematics, Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM); 1999.

[31] Kinderlehrer D, Stampacchia G. An introduction to variational inequalities and their applications. Philadelphia: Society for Industrial and Applied Mathematics; 2000.

[32] Frehse J. On the regularity of the solution of the biharmonic variational inequality. Manuscripta Math 1973;9:91–103.

[33] Hörmander L. The analysis of linear partial differential operators. III. Berlin: Springer-Verlag; 1985.

[34] Casas E. L^2 estimates for the finite element method for the Dirichlet problem with singular data. Numer Math 1985;47:627–32.

[35] Evans LC. Partial differential equations. 2nd ed.. Providence, RI: American Mathematical Society; 2010.

[36] Casas E, Mateos M, Vexler B. New regularity results and improved error estimates for optimal control problems with state constraints. ESAIM Control Optim Calc Var 2014;20:803–22.

[37] Engel G, Garikipati K, Hughes TJR, Larson MG, Mazzei L, Taylor RL. Continuous/discontinuous finite element approximations of fourth order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity. Comput Methods Appl Mech Engrg 2002;191:3669–750.

[38] Brenner SC, Sung L-Y. C^0 Interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. J Sci Comput 2005;22/23:83–118.

[39] Brenner SC. C^0 Interior penalty methods. In: Blowey J, Jensen M, editors. Frontiers in numerical analysis-Durham 2010. Lecture notes in computational science and engineering, vol. 85, Berlin-Heidelberg: Springer-Verlag; 2012, p. 79–147.

[40] Brenner SC, Sung L-Y. A new convergence analysis of finite element methods for elliptic distributed optimal control problems with pointwise state constraints. SIAM J Control Optim 2017;55:2289–304.

[41] Bramble JH, Hilbert SR. Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation. SIAM J Numer Anal 1970;7:113–24.

[42] Dupont T, Scott R. Polynomial approximation of functions in Sobolev spaces. Math Comp 1980;34:441–63.

[43] Ciarlet PG. Sur l'élément de Clough et Tocher. RAIRO Anal Numér 1974;8:19–27.

[44] Gilbarg D, Trudinger NS. Elliptic partial differential equations of second order. Classics in mathematics, Berlin: Springer-Verlag; 2001.

[45] Brenner SC, Wang K, Zhao J. Poincaré-Friedrichs inequalities for piecewise H^2 functions. Numer Funct Anal Optim 2004;25:463–78.

[46] Brenner SC, Neilan M, Reiser A, Sung L-Y. A C^0 interior penalty method for a von Kármán plate. Numer Math 2017;135:803–32.

[47] Meyer C, Rösch A. Superconvergence properties of optimal control problems. SIAM J Control Optim 2004;43:970–85.