LINEAR AND DISCONTINUOUS APPROXIMATIONS FOR OPTIMAL CONTROL PROBLEMS

A. RÖSCH * AND R. SIMON †

Abstract. An optimal control problem for 2-d and 3-d elliptic equations is investigated with pointwise control constraints. This paper is concerned with discretization of the control by piecewise linear, but discontinuous functions. The state and the adjoint state are discretized by linear finite elements. The paper is focussed on similarities and differences to piecewise constant and piecwise linear (continuous) approximation of the controls. Approximation of order h in the L^∞-norm is proved in the main.

Keywords: Linear-quadratic optimal control problems, error estimates, elliptic equations, numerical approximation, control constraints.

AMS subject classification: 49K20, 49M25, 65N30

1. Introduction. The paper is concerned with the discretization of the elliptic optimal control problem

$$J(u) = F(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)} \tag{1.1}$$

subject to the state equations

$$\begin{aligned} Ay + a_0 y &= u && \text{in } \Omega \\ y &= 0 && \text{on } \Gamma \end{aligned} \tag{1.2}$$

and subject to the control constraints

$$a \le u(t,x) \le b \qquad \text{for a.a. } x \in \Omega, \tag{1.3}$$

where Ω is a bounded domain in $\mathbb{R}^N$ with $N = 2, 3$ and Γ is the boundary of Ω; A denotes a second order elliptic operator of the form

$$Ay(x) = -\sum_{i,j=1}^{N} D_i(a_{ij}(x) D_j y(x))$$

where D_i denotes the partial derivative with respect to x_i, and a and b are real numbers. Moreover, $\nu > 0$ is a fixed positive number. We denote the set of admissible controls by U_{ad}:

$$U_{ad} = \{u \in L^2(\Omega) : \ a \le u \le b \ \text{ a.e. in } \Omega\}.$$

We discuss here the full discretization of the control and the state equations by a finite element method. The asymptotic behavior of the discretized problem is studied.

Approximation properties of discretized optimal control problems are often investigated in the last years. First results were known for piecewise constant functions, we refer to Falk [7], Geveci [8], and Malanowski [12]. A renaissance of this topic was mainly initiated by the papers of Arada,

*Johann Radon Institute for Computational and Applied Mathematics (RICAM),Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria, arnd.roesch@oeaw.ac.at

†SFB 013 Johannes-Kepler-Universität Linz Altenbergerstraße 69, A-4040 Linz, Austria, rene.simon@sfb013.uni-linz.ac.at

Casas, and Tröltzsch [1] and Casas, Mateos, and Tröltzsch [5]. Error estimates of order h in the L^2-norm and in the L^∞-norm are established in these articles.

Piecewise linear control discretizations for elliptic optimal control problems are studied by Casas and Tröltzsch, see [6] and Casas [4]. These papers contains error estimates of order h and $o(h)$ in the L^2-norm for general cases. For more regular cases an approximation order of $h^{3/2}$ can be proved, see Rösch [17], [16]. An error estimate of order h in the L^∞-norm for an elliptic problem is proved by Meyer and Rösch [14]. However, this result is restricted to the space dimension 2. General error estimates in the L^∞-norm are unknown until now.

The L^2-estimates motivate the use of piecewise linear functions. The known L^∞-estimates favor piecewise constant functions. These facts were our motivation to shed light on the properties of piecewise linear, but discontinuous approximations for the control.

Let us remark, that new discretization concepts has been developed in recent years. The variational approach by Hinze [10] and the superconvergence approach of Meyer and Rösch [13] can achieve approximation order h^2.

The paper is organized as follows: In section 2 the discretizations are introduced and the main results are stated. Section 3 contains auxiliary results. The proofs of the approximation result is placed in section 4. The paper ends with numerical experiments shown in section 5.

2. Discretization and main result. Throughout this paper, Ω denotes a convex bounded open subset in $\mathbb{R}^2$ of class $C^{1,1}$. The coefficients a_{ij} of the operator A belong to $C^{0,1}(\bar\Omega)$ and satisfy the ellipticity condition

$$m_0|\xi|^2 \le \sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \qquad \forall (\xi,x)\in \mathbb{R}^N\times\bar\Omega, \quad m_0>0.$$

Moreover, we require $a_{ij}(x)=a_{ji}(x)$ and $y_d\in L^p(\Omega)$ for some $p>N$. For the function $a_0\in L^\infty(\Omega)$, we assume $a_0\ge 0$. Next, we recall a result from Grisvard [9] Th. 2.4.2.5.

LEMMA 2.1. *[9] For every $p>N$ and every function $g\in L^p(\Omega)$, the solution y of*

$$Ay + a_0 y = g \qquad \text{in } \Omega, \quad y|_\Gamma = 0,$$

belongs to $H_0^1(\Omega)\cap W^{2,p}(\Omega)$. Moreover, there exists a positive constant c, independent of a_0 such that

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

Next, we introduce the adjoint equation

$$\begin{aligned} Ap + a_0 p &= y - y_d && \text{in } \Omega \\ p &= 0 && \text{on } \Gamma \end{aligned} \tag{2.1}$$

Due to Lemma 2.1, the state equation and the adjoint equation admit unique solutions in $H_0^1(\Omega)\cap W^{2,p}(\Omega)$, if $y_d\in L^p(\Omega)$ for $p>N$. This space is embedded in $C^{0,1}(\bar\Omega)$.

We call the solution y of (1.2) for a control u associated state to u and write $y(u)$. In the same way, we call the solution p of (2.1) corresponding to $y(u)$ associated adjoint state to u and write $p(u)$.

Introducing the projection

$$\Pi_{[a,b]}(f(x)) = \max(a, \min(b, f(x))),$$

we can formulate the necessary and sufficient first-order optimality condition for (1.1)–(1.3).

LEMMA 2.2. *A necessary and sufficient condition for the optimality of a control $\bar{u}$ with corresponding state $\bar{y} = y(\bar{u})$ and adjoint state $\bar{p} = p(\bar{u})$, respectively, is that the equation*

$$\bar{u}(x) = \Pi_{[a,b]}(-\frac{1}{\nu}\bar{p}(x)) \tag{2.2}$$

holds.

Since the optimal control problem is strictly convex, we obtain the existence of a unique optimal solution. The optimality condition can be formulated as variational inequality

$$(\nu\bar{u} + \bar{p}, u - \bar{u})_U \geq 0 \qquad \text{for all } u \in U_{ad}$$

where $(\cdot,\cdot)_U$ denotes the natural inner product of $U = L^2(\Omega)$. A standard pointwise a.e. discussion of this variational inequality leads to the above formulated projection formula, see [12].

We are now able to introduce the discretized problem. We define a finite-element based approximation of the optimal control (1.1)–(1.3). To this aim, we consider a family of triangulations $(T_h)_{h>0}$ of $\bar{\Omega}$. With each element $T \in T_h$, we associate two parameters $\rho(T)$ and $\sigma(T)$, where $\rho(T)$ denotes the diameter of the set T and $\sigma(T)$ is the diameter of the largest ball contained in T. The mesh size of the grid is defined by $h = \max_{T\in T_h} \rho(T)$. We suppose that the following regularity assumptions are satisfied.

(A1) There exist two positive constants ρ and σ such that

$$\frac{\rho(T)}{\sigma(T)} \leq \sigma, \qquad \frac{h}{\rho(T)} \leq \rho$$

hold for all $T \in T_h$ and all $h > 0$.

(A2) Let us define $\bar{\Omega}_h = \bigcup_{T\in T_h} T$, and let Ω_h and Γ_h denote its interior and its boundary, respectively. We assume that $\bar{\Omega}_h$ is convex and that the vertices of T_h placed on the boundary of Γ_h are points of Γ. From [15], estimate (5.2.19), it is known that

$$|\Omega \setminus \Omega_h| \leq Ch^2,$$

where $|.|$ denotes the measure of the set. Next, we set

$$\begin{aligned} U_h &= \{u_h \in L^\infty(\Omega) :\ u_h \in \mathcal{P}_1 \text{ for all } T \in T_h,\ u_h = \Pi_{[a,b]}(0) \text{ on } \bar{\Omega} \setminus \Omega_h\}, \quad U_h^{ad} = U_h \cap U_{ad}, \\ V_h &= \{y_h \in C(\bar{\Omega}) :\ y_h \in \mathcal{P}_1 \text{ for all } T \in T_h, \text{ and } y_h = 0 \text{ on } \bar{\Omega} \setminus \Omega_h\}, \end{aligned}$$

where $\mathcal{P}_1$ is the space of polynomials of degree less or equal than 1. Let us short motivate this choice of U_h. The adjoint state p is Lipschitz continuous and has homogeneous boundary values. A continuous extension of the optimal control $\bar{u}$ to the boundary exists because of (2.2). Moreover it holds for every boundary point $\hat{x}$

$$\lim_{x\to\hat{x}} \bar{u}(x) = \Pi_{[a,b]}(0).$$

Therefore, we set $u_h = \Pi_{[a,b]}(0)$ on $\bar{\Omega} \setminus \Omega_h$.

For each $u_h \in U_h$, we denote by $y_h(u_h)$ the unique element of V_h that satisfies

$$a(y_h(u_h), v_h) = \int_\Omega u_h v_h \, dx \qquad \forall v_h \in V_h, \tag{2.3}$$

where $a : V_h \times V_h \to \mathbb{R}$ is the bilinear form defined by

$$a(y_h, v_h) = \int_\Omega \left(a_0(x) y_h(x) v_h(x) + \sum_{i,j=1}^{2} a_{ij}(x) D_i y_h(x) D_j v_h(x) \right) dx.$$

In other words, $y_h(u_h)$ is the approximated state associated with u_h. Because of $y_h = v_h = 0$ on $\bar{\Omega} \setminus \Omega_h$ the integrals over Ω can be replaced by integrals over Ω_h. The finite dimensional approximation of the optimal control problem is defined by

$$\inf J(u_h) = \frac{1}{2} \|y_h(u_h) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2} \|u_h\|^2_{L^2(\Omega)} \qquad u_h \in U_h^{ad}. \tag{2.4}$$

The adjoint equation is discretized in the same way

$$a(p_h(u_h), v_h) = \int_\Omega (y_h(u_h) - y_d) v_h \, dx \qquad \forall v_h \in V_h. \tag{2.5}$$

Now, we are able to state the main result.

THEOREM 2.3. *Let $\bar{u}$ and u_h be the optimal solution of (1.1) and (2.4), respectively. Then an estimate*

$$\|\bar{u} - u_h\|_{L^\infty(\Omega)} \le Ch \tag{2.6}$$

holds true with a positive constant C.

The proof of Theorem 2.3 is contained in section 4. Moreover, the constant C is specified in that section.

3. Auxiliary results. We start with an L_2-estimate corresponding to Theorem 2.3.

LEMMA 3.1. *Let $\bar{u}$ and u_h be the optimal solution of (1.1) and (2.4), respectively. Then an estimate*

$$\|\bar{u} - u_h\|_{L^2(\Omega)} \le C_2 h \tag{3.1}$$

holds true with a positive constant C_2.

Proof. This statement can easily proved by the arguments of Casas and Tröltzsch [6] or Casas [4]. Here, we sketch only the modifications of this proof for discontinuous piecewise linear functions. Let us define a function $v_h \in U_h^{ad}$ on an arbitrary triangle T_i by

$$v_h = \begin{cases} a & \text{if } \min_{x \in T_i} \bar{u}(x) = a \\ b & \text{if } \max_{x \in T_i} \bar{u}(x) = b \\ I_h(\bar{u}) & \text{else} \end{cases}$$

where $I_h(\bar{u})$ denotes the linear Interpolate of $\bar{u}$. This definition is correct for sufficient small h: Then, it can not happen $\min_{x \in T_i} \bar{u}(x) = a$ and $\max_{x \in T_i} \bar{u}(x) = b$ simultaneously on the same triangle T_i. Along the lines of [6], Lemma 2.1 it can be proved

$$\|u_h - \bar{u}\|_{L^2(\Omega)} \le c \|v_h - \bar{u}\|_{L^2(\Omega)}.$$

The inequality

$$\|v_h - \bar{u}\|_{L^2(\Omega)} \le ch$$

is easy to verify. □

COROLLARY 3.2. *Let us define a set K of triangles T_i containing active and inactive points:*

$$K := \bigcup \{T_i : \text{ there exist points } x_1, x_2 \text{ with } \bar{u}(x_1) \in (a,b),\ \bar{u}(x_2) \in \{a,b\}\}$$

If the set K has only a size of order h, i.e. $|K| \le ch$, then we have

$$\|u_h - \bar{u}\|_{L^2(\Omega)} \le ch^{3/2}.$$

Lemma 3.1 implies easily the following L^∞-estimate

$$\|\bar{p} - p(u_h)\|_{L^\infty(\Omega)} \le c\|\bar{p} - p(u_h)\|_{H^2(\Omega)} \le ch. \tag{3.2}$$

LEMMA 3.3. *The inequality*

$$\|\bar{p} - p_h(u_h)\|_{L^\infty(\Omega)} \le \kappa h \tag{3.3}$$

is valid with a positive constant κ.

Proof. First, we recall a L^∞-estimate for the finite element solution

$$\|p(u_h) - p_h(u_h)\|_{L^\infty(\Omega)} \le ch, \tag{3.4}$$

see Braess [3]. Moreover, we find

$$\|\bar{p} - p_h\|_{L^\infty(\Omega)} \le \|\bar{p} - p(u_h)\|_{L^\infty(\Omega)} + \|p(u_h) - p_h\|_{L^\infty(\Omega)} \le \kappa h.$$

using the inequalities (3.2) and (3.4). □

Next, we introduce a new notation for the piecewise linear functions. Let E_j be an arbitrary vertex of a triangle T_i. Then, we define a linear function $\tilde{e}_{i,j}$ on T_i by

$$\tilde{e}_{i,j}(E_k) = \delta_{jk},$$

where δ_{ij} is the Kronecker symbol and E_k is a vertex of T_i. Next, we introduce our basis function

$$e_{ij} = \begin{cases} \tilde{e}_{i,j} & \text{on } T_i, \\ 0 & \text{else} \end{cases}$$

Therefore, we can represent the functions u_h and $p_h(u_h)$ by

$$u_h(x) = \sum_{T_i} \sum_{j=1}^{3} u_{ij} e_{ij}(x)$$

$$(p_h(u_h))(x) = \sum_{T_i} \sum_{j=1}^{3} p_j e_{ij}(x)$$

with $u_{ij} = \lim_{x \to E_j, x \in T_i} u_h(x)$ and $p_j = (p_h(u_h))(E_j)$.

Let T_i be an arbitrary triangle and E_j an arbitrary vertex. We denote the set of all vertices of T_i excepted E_j by $N_i(E_j)$.

LEMMA 3.4. *For every triangle T_i and every indices k, j with $E_k \in N_i(E_j)$ it holds*

$$\frac{1}{\nu}|p_j - p_k| \leq Dh, \tag{3.5}$$

with

$$D = \frac{L + 2\kappa}{\nu}$$

where L denotes the Lipschitz constant of $\bar{p}$.

Proof. Because of Lemma 2.1, $\bar{p}$ belongs to $W_p^2(\Omega)$ for a certain $p > 2$. Therefore $\bar{p}$ is Lipschitz continuous and we have

$$|\bar{p}(E_j) - \bar{p}(E_k)| \leq Lh.$$

Combining this inequality with (3.3), we obtain

$$\begin{aligned} |p_j - p_k| &\leq |p_j - \bar{p}(E_j)| + |\bar{p}(E_j) - \bar{p}(E_k)| + |\bar{p}(E_k) - p_k| \\ &\leq \kappa h + Lh + \kappa h. \end{aligned}$$

Dividing by ν, the assertion is proved. □

Next, we recall a property concerning the mass matrix.

LEMMA 3.5. *For a fixed triangle T_i and arbitrary basis functions e_{ij}, e_{ik} $(j \neq k)$.*

$$(e_{ij}, e_{ij})_U = 2(e_{ij}, e_{ik})_U \tag{3.6}$$

is valid.

Proof. The element mass matrix of the reference element T_r for $N = 2$ is given by

$$M_r^2 = \left((e_{ij}, e_{ik})_U\right)_{j,k=1,2,3} = \frac{1}{24}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

and for $N = 3$ we have

$$M_r^3 = \left((e_{ij}, e_{ik})_U\right)_{j,k=1,2,3,4} = \frac{1}{120}\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}.$$

Clearly, the entries of this matrix have the desired property. The mass matrix of an arbitrary triangle T_s is given by

$$M_s = \frac{|T_s|}{|T_r|} M_r.$$

Multiplication with a scalar factor preserves this property. □

We note that in the case $N = 3$ the inequality

$$(e_{ij}, e_{ij})_U < \sum_{E_k \in N_i(E_j)} (e_{ij}, e_{ik})_U$$

is valid. Therefore, the proving technique of [14] can not be applied in the 3-d case. However, the new technique presented here can not be transferred to the case of piecewise linear and continuous controls.

Next, we want to investigate the following quantity

$$M := \max_{ij} \left| u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j) \right|. \tag{3.7}$$

In all what follows, the index ij denotes a fixed vertex E_j and a corresponding triangle T_i where this maximum is attained.

Equations (3.7) means that one of the following two cases occurs

$$\begin{array}{ll} (A) & M = u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j), \\ (B) & M = -(u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j)). \end{array}$$

LEMMA 3.6. *Assume $M > 0$. Then, the control $v_h = u_h - \varepsilon e_{ij}$ is admissible in the case (A) and $v_h = u_h + \varepsilon e_{ij}$ is admissible in the case (B) for a sufficiently small $\varepsilon > 0$. Moreover, the inequalities*

$$\begin{array}{rclcll} M & = & u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j) & \leq & u_{ij} + \frac{1}{\nu} p_j & \text{in the case (A)}, \\ M & = & -(u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j)) & \leq & -(u_{ij} + \frac{1}{\nu} p_j) & \text{in the case (B)} \end{array} \tag{3.8}$$

hold true.

Proof. We discuss only case (A). Since M is positive and $\Pi_{[a,b]}(-\frac{1}{\nu} p_j) \in [a, b]$ by definition, this implies

$$u_{ij} > a.$$

Consequently, there exists a $\varepsilon > 0$ such that

$$u_{ij} - \varepsilon > a.$$

This means, that the control $v_h = u_h - \varepsilon e_{ij}$ is admissible. From $M > 0$ and $u_{ij} \in [a, b]$ we obtain $\Pi_{[a,b]}(-\frac{1}{\nu} p_j) < b$. Therefore, we have $\Pi_{[a,b]}(-\frac{1}{\nu} p_j) \geq -\frac{1}{\nu} p_j$ and consequently

$$M = u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j) \leq u_{ij} + \frac{1}{\nu} p_j. \tag{3.9}$$

□

LEMMA 3.7. *Let $v_h = u_h - \varepsilon e_{lm}$ and $w_h = u_h + \varepsilon e_{rs}$ be admissible for certain indices l, m, r, s and $\varepsilon > 0$. Then the inequalities*

$$\begin{array}{rcl} u_{ml} + \frac{1}{\nu} p_l & \leq & \frac{1}{2} \sum_{E_k \in N_m(E_l)} -(u_{ik} + \frac{1}{\nu} p_k), \\ -(u_{rs} + \frac{1}{\nu} p_s) & \leq & \frac{1}{2} \sum_{E_k \in N_r(E_s)} (u_{ik} + \frac{1}{\nu} p_k) \end{array} \tag{3.10}$$

are valid.

Proof. We derive only the first inequaliy. We start with the optimality condition for u_h

$$(\nu u_h + p_h(u_h), v_h - u_h)_U \geq 0 \qquad \text{for all } v_h \in U_h^{ad}.$$

We test this inequality with $v_h = u_h - \varepsilon e_{ml}$

$$(\nu u_h + p_h(u_h), -\varepsilon e_{ml})_U \geq 0.$$

From this, we obtain

$$(\nu u_{ml} + p_l)(e_{ml}, e_{ml})_U \leq \sum_{E_k \in N_m(E_l)} -(\nu u_{mk} + p_k)(e_{ml}, e_{mk})_U.$$

Using (3.6), we find

$$(\nu u_{ml} + p_l)(e_{ml}, e_{ml})_U \leq \frac{1}{2}(e_{ml}, e_{ml})_U \sum_{E_k \in N_m(E_l)} -(\nu u_{mk} + p_k).$$

Division by $(e_{ml}, e_{ml})_U$ yields (3.10). □

LEMMA 3.8. *Let $M > 0$ and ij be an index where the maximum in (3.7) is attained. Then there exists an index m with $E_m \in N_i(E_j)$ such that*

$$\begin{array}{lll} \Pi_{[a,b]}(-\frac{1}{\nu}p_m)) & \leq & -\frac{1}{\nu}p_m \quad \textit{in the Case (A)}, \\ \Pi_{[a,b]}(-\frac{1}{\nu}p_m)) & \geq & -\frac{1}{\nu}p_m \quad \textit{in the Case (B)}. \end{array} \tag{3.11}$$

Proof. The discussion of Case (A) can be splitted in two partial cases:

Case 1 There exist an index l with $E_l \in N_i(E_j)$ with

$$\nu u_{il} + p_l > 0.$$

We can apply Lemma 3.7 for the index ij, since $v_h = u_h - \varepsilon e_{ij}$ is admissible (Lemma 3.6). Then, we can increase the right-hand side of (3.10) by omitting the term $-(\nu u_{il} + p_l)$

$$(\nu u_{ij} + p_j) < \frac{1}{2} \sum_{E_k \in N_i(E_j), E_k \neq E_l} -(\nu u_{ik} + p_k).$$

We continue by

$$(\nu u_{ij} + p_j) < \max_{E_k \in N_i(E_j)} -(\nu u_{ik} + p_k).$$

We denote an index, where this maximum is attained by m

$$-(\nu u_{im} + p_m) = \max_{E_k \in N_i(E_j)} -(\nu u_{ik} + p_k).$$

Combining with (3.9), we find

$$M = u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu}p_j) \leq u_{ij} + \frac{1}{\nu}p_j < -(u_{im} + \frac{1}{\nu}p_m). \tag{3.12}$$

By definition of M, we have

$$-(u_{im} - \Pi_{[a,b]}(-\frac{1}{\nu}p_m)) \leq M.$$

Hence, we obtain

$$\Pi_{[a,b]}(-\frac{1}{\nu}p_m)) \leq -\frac{1}{\nu}p_m.$$

Case 2 For all indices l with $E_l \in N_i(E_j)$ we have

$$\nu u_{il} + p_l \leq 0.$$

Using (3.10), we find

$$(\nu u_{ij} + p_j) \le \frac{3}{2} \max_{E_k \in N_i(E_j)} -(\nu u_{ik} + p_k). \tag{3.13}$$

Again, we denote an index, where this maximum is attained by m

$$-(\nu u_{im} + p_m) = \max_{E_k \in N_i(E_j)} -(\nu u_{ik} + p_k).$$

We will show that $u_{im} = b$ must hold. Assuming $u_{im} < b$, the control $v_h = u_h + \varepsilon u_{im}$ is admissible for sufficiently small ε. Therefore, we can apply Lemma 3.7 for the index im

$$-(u_{im} + \frac{1}{\nu} p_m) \le \frac{1}{2} \sum_{E_k \in N_i(E_m)} (u_{ik} + \frac{1}{\nu} p_k).$$

Now, the assertion of Case 2 implies

$$-(u_{im} + \frac{1}{\nu} p_m) \le \frac{1}{2} (u_{ij} + \frac{1}{\nu} p_j).$$

From this and (3.13) we get

$$u_{ij} + \frac{1}{\nu} p_j \le \frac{3}{4} (u_{ij} + \frac{1}{\nu} p_j)$$

or $u_{ij} + \frac{1}{\nu} p_j \le 0$. This is a contradiction to (3.9) and $M > 0$. Consequently, we have $u_{im} = b$.

The inequalities $M > 0$, (3.9), and (3.13) imply

$$0 < M \le -(\nu u_{im} + p_m) \quad \Rightarrow \quad -\frac{1}{\nu} p_m > b.$$

From this the assertion is easily obtained. □

LEMMA 3.9. *Assume that*

$$Dh < b - a$$

is valid. Then, the estimate

$$M = \max_{ij} |u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu} p_j)| < Dh \tag{3.14}$$

holds true.

Proof. For $M = 0$ the assertion is automatically true. Let us assume $M > 0$. Again, we discuss only the Case (A). Inequality (3.11) implies directly

$$b = \Pi_{[a,b]}(-\frac{1}{\nu} p_m)) < -\frac{1}{\nu} p_m. \tag{3.15}$$

From this and (3.5), we obtain

$$-\frac{1}{\nu} p_j > b - Dh.$$

By assumption, the value $b - Dh$ is greater than a. Hence

$$-\frac{1}{\nu} p_j > a \tag{3.16}$$

holds. From (A)

$$u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu}p_j) = M > 0$$

and $u_{ij} \leq b$ we obtain

$$\Pi_{[a,b]}(-\frac{1}{\nu}p_j) < b. \tag{3.17}$$

From (3.16) and (3.17) we get

$$-\frac{1}{\nu}p_j = \Pi_{[a,b]}(-\frac{1}{\nu}p_j)$$

that implies

$$u_{ij} + \frac{1}{\nu}p_j = u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu}p_j) = M.$$

Using $u_{ij} \leq b$ and $\frac{1}{\nu}p_j < -(b - Dh)$, we find

$$u_{ij} + \frac{1}{\nu}p_j < b - (b - Dh) = Dh.$$

Combining the last two inequalities, the assertion is proved. □

4. Proof of the main result. The proof of Theorem 2.3 is divided in two parts. In the next lemma we derive a corresponding estimate for the grid points. The estimate for arbitrary points is obtained in a second step.

LEMMA 4.1. *The estimate*

$$\max_{ij} |u_{ij} - \bar{u}(E_j)| \leq (D + \frac{\kappa}{\nu})h.$$

is valid.

Proof. Because of $u_h(x) \in [a, b]$, the assertion is automatically true for $Dh \geq b - a$. We assume now $Dh < b - a$. From Lemma 3.9, we know

$$\max_{i} |u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu}p_j)| \leq Dh$$

or in other notation

$$\max_{i} |u_{ij} - \Pi_{[a,b]}(-\frac{1}{\nu}p_h(E_j))| \leq Dh.$$

From (3.3)

$$\|\bar{p} - p_h\|_{L^\infty(\Omega)} \leq \kappa h,$$

and the Lipschitz continuity of the projection operator we deduce

$$\|\Pi_{[a,b]}(-\frac{1}{\nu}\bar{p}(E_j)) - \Pi_{[a,b]}(-\frac{1}{\nu}p_h(E_j))\|_{L^\infty(\Omega)} \leq \frac{\kappa}{\nu}h.$$

Using

$$\bar{u}(E_j) = \Pi_{[a,b]}(-\frac{1}{\nu}\bar{p}(E_j))$$

and the triangle inequality we end up with

$$\max_i |u_h(E_i) - \bar{u}(E_j)| \leq (D + \frac{\kappa}{\nu})h.$$

□

Now, we are able to proof Theorem 2.3.

Proof. For a non grid point $x \in T_i$ we find a convex linear combination of the vertices E_j of the corresponding triangle

$$x = \sum_{E_j \in T_i} \lambda_j E_j, \quad \sum_{E_j \in T_i} \lambda_j = 1.$$

Since u_h is linear on T_i, we get

$$\begin{aligned}
|u_h(x) - \bar{u}(x)| &= \left| \sum_{E_j \in T_i} \lambda_j u_{ij} - \bar{u}(x) \right| \\
&\leq \sum_{E_j \in T_i} \lambda_j |u_{ij} - \bar{u}(E_j)| + \sum_{E_j \in T_i} \lambda_j |\bar{u}(x) - \bar{u}(E_j)| \\
&\leq (D + \frac{\kappa}{\nu})h + \sum_{E_j \in T_i} \lambda_j |\bar{u}(x) - \bar{u}(E_j)| \\
&\leq (D + \frac{\kappa}{\nu})h + \frac{L}{\nu}h.
\end{aligned}$$

In the final inequality we used that $\bar{u}$ is Lipschitz continuous with constant L/ν. Summarizing all results, we obtain

$$\|\bar{u} - u_h\|_{L^\infty(\Omega_h)} \leq (D + \frac{\kappa + L}{\nu})h.$$

Therefore, the assertion is true for $x \in \Omega_h$ with

$$C = D + \frac{\kappa + L}{\nu}.$$

It remains the case $x \in \Omega \setminus \Omega_h$. By definition, we have $u_h = \Pi_{[a,b]}(0)$ on this part. From (2.2), we obtain easily $\bar{u} = \Pi_{[a,b]}(0)$ on Γ. Let $x \in \Omega \setminus \Omega_h$ be an arbitrary point. From [15], we know that

$$\min_{x_\Gamma \in \Gamma} |x - x_\Gamma| \leq c_\Gamma h^2$$

holds with a certain constant $c_\Gamma > 0$ independent of h. Therefore, we find for $x \in \Omega \setminus \Omega_h$

$$|u_h(x) - \bar{u}(x)| = |\Pi_{[a,b]}(0) - \bar{u}(x)| = |\bar{u}(x_\Gamma) - \bar{u}(x)| \leq \frac{c_\Gamma L}{\nu} h^2.$$

□

5. Numerical tests. Our approximation theory is tested for two examples where the exact solution of the undiscretized optimal control problem is known. These examples were originally introduced in [13].

In both cases, the Laplace operator $-\Delta$ was chosen for the elliptic operator A. The domain Ω is the unit square $(0,1) \times (0,1)$. We used uniform meshs, where the parameter N denotes the number

of intervalls in which the edges are divided. Hence, the quantities N and h are connceted by the formula $N \cdot h = \sqrt{2}$. Both optimization problems were solved numerically by a primal-dual active set strategy, see [2] and [11]. The discretization was already described in section 2: The state y and the adjoint state p were approximated by piecewise linear functions, whereas the control u is dicretized by piecewise linear, but discontinuous functions. For comparison we also used piecewise constant functions for the control u.

The first example is a homogeneous Dirichlet problem, which fulfills the assumptions mentioned at the beginning of section 2, except the boundary regularity. Although Γ is not of class $C^{1,1}$, the $W^{2,p}$-regularity of $\bar{p}$ (see Lemma 2.1) is obtained by a result of Grisvard [9] for convex polygonal domain. In the second example, a Neumann boundary problem is studied. In this case, the theoretical results does not exactly fit to the problem. However, in the case $\Omega_h = \Omega$, the theory can be easily adapted.

Example 1. In this example, the Laplace equation with homogeneous Dirichlet boundary conditions is investigated, i.e. $a_0 \equiv 0$ in (1.2). Thus, the state equation is given by

$$\begin{aligned} -\Delta y &= u && \text{in } \Omega \\ y &= 0 && \text{on } \Gamma. \end{aligned} \tag{5.1}$$

We define the optimal state by

$$\bar{y} = y_a - y_g$$

with an analytical part $y_a = \sin(\pi x_1)\sin(\pi x_2)$ and a less smooth function y_g, which is defined as the solution of

$$\begin{aligned} -\Delta y_g &= g && \text{in } \Omega \\ y_g &= 0 && \text{on } \Gamma. \end{aligned}$$

The function g is given by

$$g(x_1, x_2) = \begin{cases} u_f(x_1, x_2) - a & , \text{if } u_f(x_1, x_2) < a \\ 0 & , \text{if } u_f(x_1, x_2) \in [a, b] \\ u_f(x_1, x_2) - b & , \text{if } u_f(x_1, x_2) > b \end{cases}$$

with $u_f(x_1, x_2) = 2\pi^2 \sin(\pi x_1)\sin(\pi x_2)$. Due to the state equation (5.1), we obtain for the exact optimal control $\bar{u}$

$$\bar{u}(x_1, x_2) = \begin{cases} a & , \text{if } u_f(x_1, x_2) < a \\ u_f(x_1, x_2) & , \text{if } u_f(x_1, x_2) \in [a, b] \\ b & , \text{if } u_f(x_1, x_2) > b \end{cases} .$$

For the optimal adjoint state $\bar{p}$, we find

$$\bar{p}(x_1, x_2) = -2\pi^2 \nu \sin(\pi x_1)\sin(\pi x_2).$$

Due to the adjoint state equation, we finally get

$$y_d(x_1, x_2) = \bar{y} + \Delta\bar{p} = y_a - y_g + 4\pi^4 \nu \sin(\pi x_1)\sin(\pi x_2).$$

It can be easily shown, that these functions fulfill the necessary and sufficient first order optimality conditions. In the numerical tests, we chose $a = 3$, $b = 15$ and $\nu = 1$.

Figure 5.1 shows the approximation behavior of $||\bar{u} - u_h||_{L^\infty(\Omega)}$. In the figures, $\bar{u}$ is denoted by u_{opt}. The expected linear approximation behavior is observed, see Figure 5.1.

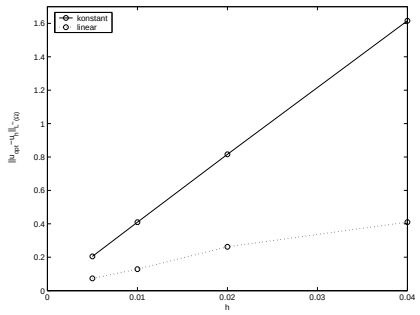

FIG. 5.1. $||\bar{u} - u_h||_{L^\infty(\Omega)}$

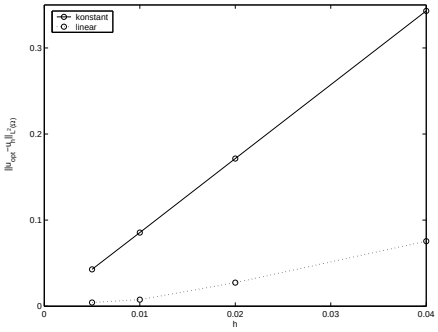

FIG. 5.2. $||\bar{u} - u_h||_{L^2(\Omega)}$

However, the error for the discontinuous linear ansatz functions is less than the error for piecewise constant functions as Table 5.2 shows. However, we should keep in mind that we have almost triplicate the number of degree of freedoms.

TABLE 5.1

	piecewise constant			piecewise linear (discontinuous)		
N	d.o.f.	$\|\|\bar{u} - u_h\|\|_{L^\infty}$	$N \cdot \|\|\bar{u} - u_h\|\|_{L^\infty}$	d.o.f.	$\|\|\bar{u} - u_h\|\|_{L^\infty}$	$N \cdot \|\|\bar{u} - u_h\|\|_{L^\infty}$
25	1058	1.6155	40.38	3456	1.0187	25.47
50	4608	0.8163	40.82	14406	0.5935	29.35
100	19208	0.4103	41.03	58806	0.2724	27.24
200	78408	0.2049	40.98	237606	0.0735	14.70

The situation will be even better in the L^2-norm, see. Figure 5.2. It is easy to verify, that the assumptions of Corollary 3.2 are fulfilled for this example. Hence, the approximation in the L^2-norm is of order $h^{3/2}$, which is confirmed by our numerical experiments. In contrast to this, the approximation in the L^2-norm is only of order h for piecewise constant functions, see Table 5.1. Here, we can achieve a better accuracy for linear discontinuous functions with about 18 percent of the degrees of freedom. Clearly, since the approximation order is larger for linear discontinuous function than for piecewise constant functions this relation will percentage decreases for finer discretizations.

TABLE 5.2

	piecewise constant			piecewise linear (discontinuous)		
N	d.o.f.	$\|\|\bar{u} - u_h\|\|_{L^2}$	$N \cdot \|\|\bar{u} - u_h\|\|_{L^2}$	d.o.f.	$\|\|\bar{u} - u_h\|\|_{L^2}$	$N^{3/2} \cdot \|\|\bar{u} - u_h\|\|_{L^2}$
25	1058	0.3431	8.58	3456	0.0755	9.44
50	4608	0.1712	8.56	14406	0.0274	9.69
100	19208	0.0856	8.56	58806	0.0076	11.10
200	78408	0.0428	8.56	237606	0.0043	7.60

Example 2. We consider here the problem

$$\begin{aligned} -\Delta y + cy &= u \qquad \text{in } \Omega \\ \partial_n y &= 0 \qquad \text{on } \Gamma \end{aligned} \tag{5.2}$$

where ∂_n denotes the normal derivative with respect to the outward normal vector.

The optimal state $\bar{y} = y_a - y_g$ is constructed with $y_a(x_1, x_2) = \cos(\pi x_1)\cos(\pi x_2)$. The function y_g is determined by the equation

$$\begin{aligned} -\Delta y_g + c y_g &= g \qquad \text{in } \Omega \\ \partial_n y_g &= 0 \qquad \text{on } \Gamma, \end{aligned}$$

with the inhomogenity

$$g(x_1, x_2) = \begin{cases} u_f(x_1, x_2) - a & , \text{if } u_f(x_1, x_2) < a \\ 0 & , \text{if } u_f(x_1, x_2) \in [a, b] \\ u_f(x_1, x_2) & , \text{if } u_f(x_1, x_2) > b \end{cases}$$

and $u_f(x_1, x_2) = (2\pi^2 + c)\cos(\pi x_1)\cos(\pi x_2)$. The optimal control $\bar{u}$ is given by the state equation (5.2)

$$\bar{u}(x_1, x_2) = \begin{cases} a & , \text{if } u_f(x_1, x_2) < a \\ u_f(x_1, x_2) & , \text{if } u_f(x_1, x_2) \in [a, b] \\ b & , \text{if } u_f(x_1, x_2) > b. \end{cases}$$

The optimal adjoint state is defined by

$$\bar{p}(x_1, x_2) = -(2\pi^2 + c)\nu \sin(\pi x_1)\sin(\pi x_2).$$

Moreover, the desired state y_d is chosen as

$$\begin{aligned} y_d(x_1, x_2) &= \bar{y} + \Delta\bar{p} - c\bar{p} \\ &= y_a - y_g + (4\pi^4\nu + 4\pi^2\nu c + \nu c^2)\sin(\pi x_1)\sin(\pi x_2). \end{aligned}$$

Again, it is easy to verify that these functions fulfill the necessary and sufficient first-order optimality conditions. In the numerical tests, we chose $a = -3$, $b = 15$ und $\nu = c = 1$.

Figure 5.3 and Figure 5.4 illustrate that the approximation behavior in the example with Neumann boundary conditions is similar to the example with Dirichlet boundary conditions.

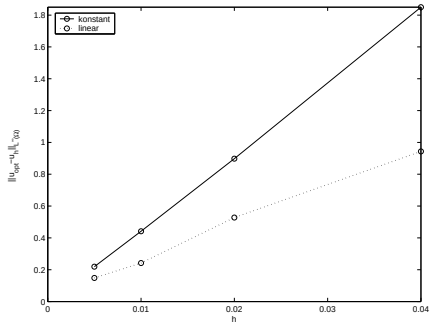

FIG. 5.3. $||\bar{u} - u_h||_{L^\infty(\Omega)}$

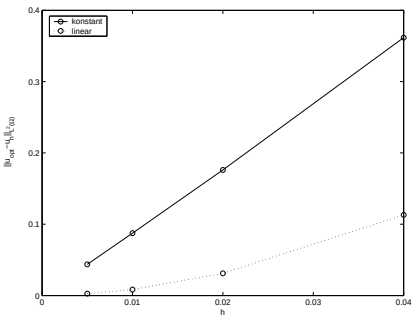

FIG. 5.4. $||\bar{u} - u_h||_{L^2(\Omega)}$

TABLE 5.3

	piecewise constant			piecewise linear (discontinuous)		
N	d.o.f.	$\|\|\bar{u}-u_h\|\|_{L^\infty}$	$N\cdot\|\|\bar{u}-u_h\|\|_{L^\infty}$	d.o.f.	$\|\|\bar{u}-u_h\|\|_{L^\infty}$	$N\cdot\|\|\bar{u}-u_h\|\|_{L^\infty}$
25	1250	1.8496	46.24	3750	0.9434	23.59
50	5000	0.8977	44.88	15000	0.5279	26.39
100	20000	0.4417	44.17	60000	0.2426	24.26
200	80000	0.2190	43.80	240000	0.1586	31.72

TABLE 5.4

	piecewise constant			piecewise linear		
N	d.o.f.	$\|\|\bar{u}-u_h\|\|_{L^2}$	$N\cdot\|\|\bar{u}-u_h\|\|_{L^2}$	d.o.f.	$\|\|\bar{u}-u_h\|\|_{L^2}$	$N^{3/2}\cdot\|\|\bar{u}-u_h\|\|_{L^2}$
25	1250	0.3617	9.04	3750	0.1131	14.14
50	5000	0.1761	8.81	15000	0.0311	10.99
100	20000	0.0874	8.74	60000	0.0084	8.40
200	80000	0.0437	8.74	240000	0.0027	7.64

Since we use discontinuous ansatz functions, it seems to be interesting to look at the maximal jump of the optimal control u_h in the gridpoints. If we double the number of degree of freedoms in one space-dimension, one can expect, that the maximal jump is halved. The following table shows this effect. Morover, these maximal jumps have similar values like the L^∞-error, see Table 5.5. The last value of the Dirichlet boundary example is an exceptional case. Calculations with finer grids show again the expected linear behavior.

TABLE 5.5

	Dirichlet		Neumann	
N	max. jump	$\|\|\bar{u}-u_h\|\|_{L^\infty}$	max. jump	$\|\|\bar{u}-u_h\|\|_{L^\infty}$
25	1.0073	1.0187	0.8507	0.9434
50	0.5894	0.5935	0.5643	0.5279
100	0.2719	0.2724	0.2377	0.2426
200	$2.9\cdot10^{-7}$	0.0735	0.1523	0.1586

The presented numerical results show that piecewise linear, but discontinuous controls can better perform than piecewise constants controls. This holds with respects to the number of degrees of freedom for the L^2-norm, too. Let us shortly discuss the different numerical effort. Although the number of degrees of freedom is triplicated for linear discontinuous controls, the numerical effort is not triplicated. The most expencive step in the computations is the numerical solution of the state equation and the adjoint equation. The control discretization influences only the right hand sight of the discretized state equation. Therefore, the numerical effort for linear discontinuous controls is not essential higher than the effort for piecewise constants controls.

REFERENCES

[1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for a semilinear elliptic optimal control problem*, Computional Optimization and Approximation, 23 (2002), pp. 201–229.

[2] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control and Optimization, 37 (1999), pp. 1176–1194.
[3] D. Braess, *Finite Elemente*, Springer-Verlag, Berlin Heidelberg, 1992.
[4] E. Casas, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*. submitted.
[5] E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Computational Optimization and Applications, (submitted).
[6] E. Casas and F. Tröltzsch, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems, V. B. et al, ed., Boston, 2003, Kluwer Academic Publishers, pp. 89–100.
[7] R. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
[8] T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, R.A.I.R.O. Analyse numeriqué, 13 (1979), pp. 313–328.
[9] P. Grisvard, *Elliptic problems in nonsmooth domains*, Pitman, Boston-London-Melbourne, 1985.
[10] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Computational Optimization and Applications, (Accepted for publication).
[11] K. Kunisch and A. Rösch, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM Journal Optimization, 13 (2002), pp. 321–334.
[12] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control–constrained optimal control problems*, Appl.Math.Opt., 8 (1981), pp. 69–95.
[13] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control and Optimization, 43 (2004), pp. 970–985.
[14] ———, *L^∞-estimates for approximated optimal control problems*, SIAM J. Control and Optimization, (submitted).
[15] P. Raviart and J. Thomas, *Introduction à l' Analyse Numérique des Équations aux Dérivées Partielles*, Masson, Paris, 1992.
[16] A. Rösch, *Error estimates for parabolic optimal control problems with control constraints*, ZAA, 23 (2004), pp. 353–376.
[17] ———, *Error estimates for linear-quadratic control problems with control constraints*, Optimization Methods and Software, (Accepted for publication).