Dedicated to
Professor Czesław Olech

Control and Cybernetics

vol. **34** (2005) No. 3

Error analysis of discrete approximations to bang-bang optimal control problems: the linear case[1]

by

Vladimir M. Veliov

Institute of Mathematical Methods in Economics,
Vienna University of Technology
Argentinierstrasse 8/119, A-1040 Vienna, Austria
and
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
1113 Sofia, Bulgaria
e-mail:vveliov@server.eos.tuwien.ac.at

Abstract: The paper presents an error estimate for Runge-Kutta direct discretizations of terminal optimal control problems for linear systems. The optimal control for such problems is typically discontinuous, and Lipschitz stability of the solution with respect to perturbations does not necessarily hold. The estimate (in terms of the optimal controls) is of first order if certain recently obtained sufficient conditions for structural stability hold, and of fractional order, otherwise. The main tool in the proof is the established relation between the local convexity index of the reachable set and the multiplicity of zeros of appropriate switching functions associated with the problem.

Keywords: linear control systems, discrete approximations, error estimates.

1. Introduction

In contrast to the rich literature on numerical methods for optimal control, error estimates are scarce, and are known only under conditions of coercivity and (certain) smoothness of the optimal solution (see e.g. Dontchev, Hager and Veliov, 2000 for a brief bibliographic account, and the contributions by Malanowski, Büskens and Maurer, 1998, and Dontchev and Hager, 2001, addressing the state

[1] This research was supported by the Austrian Science Foundation under contract No. P15618-G05.

constrained case). The coercivity is related to the Legendre-Clebsch second order sufficient condition, which implies also Lipschitz stability of the optimal solution with respect to perturbations (see Dontchev and Hager, 1993). Sufficient conditions for optimality in the case of non-coercive problems have been obtained only recently (Osmolovskii, 2000; Agrachev, Stefani and Zezza, 2002; Noble and Schättler, 2002; Felgenhauer, 2003, 2005; Maurer and Osmolovskii, 2004) and the research in this direction is still in progress.

In general, the high-order sufficient optimality conditions for non-coercive (bang-bang) problems do not imply Lipschitz stability of the optimal solution. The study of the sensitivity of the associated Hamiltonian (canonical) system is burdened by its state-discontinuity. These facts create difficulties for the sensitivity and the error analysis of bang-bang-type problems. To our knowledge, error estimates for direct discretization schemes are not available in the literature. In this paper we present such error estimates for Runge-Kutta discretization of terminal optimal control problems for linear systems. We use an indirect approach involving three main ingredients: (i) a sensitivity estimate for convex problems, depending on the convexity index of the objective function and the constraining set; (ii) estimation of the convexity index of the reachable set of a linear control system; (iii) estimation of the error in the reachable set caused by a time-discretization.

The (local) convexity index of a convex set at a given point on its boundary represents the (local) rate of deviation of the boundary from the tangential subspaces at the given point (see the next section for the strict definition). Sets with convexity index everywhere equal to two are known as R-convex sets. This notion was introduced by Pliś (1975) and is deeply studied by Frankowska and Olech (1980) in connection with the reachable set of linear control systems. A comprehensive analysis of the R-convexity is presented in Polovinkin (1996). As shown by Łojasiewicz (1979), however, for a state dimension bigger than two and for polyhedral control constraints, the reachable set generically fails to be R-convex. For this reason in the present paper we introduce the more general notion of local convexity index, and evaluate constructively the convexity index of the reachable set of linear systems, which turns out to be finite under rather general conditions. In this case the convexity index of the reachable set at a given boundary point is closely related with the multiplicity of the zeros of the switching function corresponding to this point, therefore – to the sensitivity of the system.

The estimation in (iii) is taken from Veliov (1997) and only adapted here to the particular consideration.

The discretization error estimate that we obtain (in terms of the optimal controls) for the Runge-Kutta discrete approximation is of first order in the case where the sufficient optimality condition from Felgenhauer (2003) holds, and of fractional order, in general. The estimation is sharp in several cases.

The paper is organized as follows: issues (i), (ii), and (iii) are presented in Sections 2, 3, and 4, respectively. Section 5 is devoted to the main result.

2. Sensitivity of the solution of σ-convex problems

In this section we obtain an estimation of the sensitivity of the solution of the problem

$$\min_{x \in R} g(x) \tag{1}$$

with respect to a perturbation in the constraining set $R \subset \mathbf{R}^n$. The principal assumption involves appropriate (in the context of the paper) convexity requirement that will be specified below.

We start with a list of (standard) notations that will be used later on:

$\mathbf{R}^n$ (resp. $\mathbf{R}^r$, $\mathbf{R}$) is the Euclidean space with the respective dimension;

$|\cdot|$ and $\langle \cdot, \cdot \rangle$ are the norm and the scalar product;

$\mathcal{B}$ is the unit ball in $\mathbf{R}^n$;

∂R is the boundary of the set $R \subset \mathbf{R}^n$;

$N_R(x)$ is the (external) normal cone to the convex closed set R at $x \in R$;

$N_R^1(x) = N_R(x) \cap \partial\mathcal{B}$ is the set of all unit normal vectors;

$H(X, Y)$ is the Hausdorff distance between two compact subsets $X, Y \subset \mathbf{R}^n$;

$\mathrm{meas}(\Delta)$ is the Lebesgue measure of $\Delta \subset \mathbf{R}$;

* means transposition.

Definition 2.1 A set $R \subset \mathbf{R}^n$ is locally σ-convex at the point $x \in R$ if there exists a constant $\gamma > 0$ and a neighborhood Z of x, such that for every $y \in R \cap Z$ the ball $0.5(x+y) + \gamma |x-y|^\sigma \mathcal{B}$ is contained in R.

For $\gamma = 0$ (and for a closed set R) this is merely the local star-shape property, which formally corresponds to the case $\sigma = +\infty$. If the above property is fulfilled at every point of the set with $\sigma = 2$ and with the same γ, then it is also called "strong convexity", or R-convexity (see Pliś, 1995; Frankowska and Olech, 1980; Polovinkin, 1996).

Remark 2.1 It is easy to see that it is enough to verify the requirement of Definition 1 for $y \in \partial R \cap Z$ only.

Definition 2.2 A function $g : R \mapsto \mathbf{R}$ (where R is a convex subset of $\mathbf{R}^n$) is locally κ-convex at $x \in R$ if there exists a constant $\rho > 0$ and a neighborhood Z of x, such that for every $y \in R \cap Z$ it holds that $g(0.5(x+y)) \leq 0.5(g(x) + g(y)) - \rho|x-y|^\kappa$.

We shall formally admit also $\kappa = +\infty$, in which case the last inequality is required with $\rho = 0$. For $\kappa = 2$ the above notion is introduced by Polyak (1983) and is called *strong convexity.*

Proposition 2.1 *Let $\hat{x}$ be the unique solution of (1). Assume that R is compact and locally σ-convex at $\hat{x}$ (with $\sigma \in [2, +\infty]$), and that $g : \mathbf{R}^n \mapsto \mathbf{R}$ is convex and* *(i) locally κ-convex at $\hat{x}$ (with $\kappa \in [2,+\infty]$); (ii) differentiable with a continuous derivative in a neighborhood of $\hat{x}$, satisfying $g'(\hat{x}) \neq 0$. Assume also that $s =$* $\min\{\kappa,\sigma\} < +\infty$.

Then there exist numbers $\varepsilon > 0$ and c such that for every compact set $\tilde{R} \subset \mathbf{R}^n$ with $H(\tilde{R},R) \leq \varepsilon$ and for every minimizer $\tilde{x}$ of g on $\tilde{R}$ it holds that

$$|\tilde{x}-\hat{x}| \leq c(H(\tilde{R},R))^{\frac{1}{s}}. \qquad (2)$$

Remark 2.2 As the proof shows, if $s = \kappa < +\infty$, then the claim remains true if just local Lipschitz continuity is required instead of (ii). Moreover, uniqueness of $\hat{x}$ follows from the rest of the assumptions if it is required, in addition, that R is (globally) star-shaped at $\hat{x}$.

Proof. Let $\tilde{x}$ be a minimizer of g on $\tilde{R}$. Let $y \in R$ be such that $|y-\tilde{x}| = \inf_{x\in R}|x-\tilde{x}|$. The mapping "set $R \longrightarrow$ set of solutions of (1)" is upper semicontinuous. Since $\hat{x}$ is the unique solution of (1), one can ensure that $\tilde{x}$ belongs to an arbitrarily given neighborhood of $\hat{x}$, by choosing $\varepsilon > 0$ sufficiently small. Then choosing an appropriate $\varepsilon > 0$ we may assume that the convex hull of $\hat{x}$, $\tilde{x}$ and y is contained in the neighborhood Z of $\hat{x}$ in which the local convexity and differentiability requirements are fulfilled.

We define

$$x = \frac{y+\hat{x}}{2} - \gamma\frac{g'(\hat{x})}{|g'(\hat{x})|}|y-\hat{x}|^{\sigma}.$$

(In connection with Remark 2.2, notice that if $s = \kappa$ one can take $\rho = 0$, so that g' does not appear in the considerations below and (ii) is of no use.) Obviously $x \in R$ due to the σ-convexity of R. Then we have (with $\bar{x}$ denoting an appropriate point on the segment $[\frac{y+\hat{x}}{2}, x]$) that

$$g(\hat{x}) \leq g(x) = g\left(\frac{y+\hat{x}}{2}\right) - \langle g'(\bar{x}), \gamma\frac{g'(\hat{x})}{|g'(\hat{x})|}|y-\hat{x}|^{\sigma}\rangle$$

$$\leq \frac{1}{2}g(y) + \frac{1}{2}g(\hat{x}) - \rho|y-\hat{x}|^{\kappa} - \gamma|g'(\hat{x})||y-\hat{x}|^{\sigma} + \gamma|g'(\bar{x}) - g'(\hat{x})||y-\hat{x}|^{\sigma}.$$

Since for an appropriate constant c_1

$$|\bar{x}-\hat{x}| \leq \frac{1}{2}|\hat{x}-y| + \gamma|y-\hat{x}|^{\sigma} \leq c_1|y-\hat{x}|,$$

we obtain that

$$\frac{1}{2}(g(y)-g(\hat{x})) \geq \rho|y-\hat{x}|^{\kappa} + \gamma|g'(\hat{x})||y-\hat{x}|^{\sigma} - \gamma c_1 L_{g'}|y-\hat{x}|^{\sigma+1},$$

where $L_{g'}$ is the Lipschitz constant of g' on Z. Choosing, if necessary, $\varepsilon > 0$ even smaller, we may ensure that $\gamma c_1 L_{g'}|y-\hat{x}| \leq 0.5\gamma|g'(\hat{x})|$. Hence,

$$\frac{1}{2}(g(y)-g(\hat{x})) \geq \rho|y-\hat{x}|^{\kappa} + \frac{1}{2}\gamma|g'(\hat{x})||y-\hat{x}|^{\sigma}.$$

Hence

$$c_2|y-\hat{x}|^s \le g(y)-g(\hat{x}),$$

where $c_2 = 2\rho + \gamma|g'(\hat{x})|$. Then we have

$$\begin{aligned} c_2|y-\hat{x}|^s &\le g(\tilde{x}) + L_g H(R,\tilde{R}) - \min_R g = \left(\min_{\tilde{R}} g - \min_R g\right) + L_g H(R,\tilde{R}) \\ &\le 2L_g H(R,\tilde{R}), \end{aligned}$$

where L_g is the Lipschitz constant of g in Z. Finally,

$$|\hat{x}-\tilde{x}| \le |\hat{x}-y| + |y-\tilde{x}| \le \left(\frac{2L_g}{c_2}\right)^{\frac{1}{s}} H(R,\tilde{R})^{\frac{1}{s}} + H(R,\tilde{R}),$$

which implies the claim of the lemma. ■

3. σ-convexity of the reachable set

In this section we give sufficient conditions for σ-convexity of the reachable set of the linear control system

$$\begin{aligned} &\dot{x} = A(t)x + B(t)u, \quad x(0) = x^0, \\ &u \in U, \end{aligned} \tag{3}$$

where $x \in \mathbf{R}^n$, $U \subset \mathbf{R}^r$. The reachable set of this system on $[0,T]$ will be denoted by R. That is, R consists of all end points $x(T)$ of trajectories of (3) corresponding to admissible controls $u(\cdot)$ (i.e. measurable functions with values in U).

Assumption (A1). For a natural number $\bar{\sigma} \ge 2$ the matrix $A : [0,T] \mapsto \mathbf{R}^{n\times n}$ is $(\bar{\sigma}-2)$-times differentiable, with Lipschitz continuous $(\bar{\sigma}-2)$-nd derivative; the matrix $B : [0,T] \mapsto \mathbf{R}^{n\times r}$ is $(\bar{\sigma}-1)$-times differentiable with Lipschitz continuous $(\bar{\sigma}-1)$-st derivative. The set U is a nondegenerate convex compact polyhedron in $\mathbf{R}^r$ $(r \ge 1)$, with finite number of vertices.

We denote by V the set of all vertices of U, and by E – the set of all edges[1] of U. Moreover, given $p \in \mathbf{R}^n$ we denote by $\lambda[p](\cdot)$ the backward solution of the adjoint equation

$$\dot{\lambda} = -A^*(t)\lambda, \quad \lambda(T) = p. \tag{4}$$

Following Pontryagin et al. (1962) we denote recursively

$$B_0(t) = B(t), \quad B_k(t) = -A(t)B_{k-1}(t) + B'_{k-1}(t), \ \ k = 1,\dots,\bar{\sigma}-1.$$

[1] The edges will be interpreted either as sets (segments) $[u,v]$, or as vectors $v-u$. In both cases $u,v \in V$, and $[u,v]$ is an extreme set of dimension one.

Assumption (A2). For every $t \in [0,T]$ and for every $e \in E$

$$\mathrm{rank}[B_0(t)e, \ldots, B_{\bar{\sigma}-1}(t)e] = n.$$

REMARK 3.1 Assumption (A2) is known as the *General Position Hypothesis* (GPH), Pontryagin et al. (1962). Thanks to the relation

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\langle \lambda[p](\cdot), B(\cdot)e\rangle = \langle \lambda[p](\cdot), B_k(\cdot)e\rangle, \quad k = 0, \ldots, \bar{\sigma}-1,$$

(GPH) implies that there is a natural number m such that for every $x \in \partial R$ there is a unique control steering x^0 to x on $[0,T]$, it is piece-wise constant with values in V, and has at most $m-1$ switching points[2]

For $l \in \mathbf{R}^n$ we define

$$V(l) = \{v \in V : \ \langle l, v\rangle = \max_{u \in U}\langle l, u\rangle\},$$

and

$$E(l) = \{[v,w] \in E : \ v, w \in V(l)\}.$$

That is, $E(l)$ consists of all "maximal" edges with respect to the direction l. Clearly, $E(l) = \emptyset$ if $V(l)$ is a singleton.

For $x \in \partial R$ we define the number $\sigma(x)$ as the minimal natural number $\sigma \in \{2, \ldots, \bar{\sigma}\}$ for which

$$\sum_{i=1}^{\sigma-1} |\langle \lambda[p](t), B_i(t)e\rangle| > 0 \ \ \forall t \in [0,T], \ \forall p \in N_R^1(x), \ \forall e \in E(B^*(t)\lambda[p](t)). \quad (5)$$

Assumptions (A1) and (A2), together with the fact that $\lambda[p](t) \neq 0$ for every $p \in N_R^1(x)$ and t, easily imply that the number $\sigma(x) \le \bar{\sigma}$ exists for every $x \in \partial R$.

PROPOSITION 3.1 *Assume (A1) and (A2). Then at every point $x \in \partial R$ the set R is locally $\sigma(x)$-convex.*

The above claim is related to, but essentially stronger, than that of Theorem 3.1 in Veliov (1987a), therefore we present the detailed proof. We still make use of a result from Veliov (1987a) (estimation (5) in the proof of Theorem 2.1), which can be reformulated in the following way.

For $k \ge 2$ and for an interval $[\tau_1, \tau_2]$, denote by $P_k([\tau_1,\tau_2]; L, \beta)$ the set of all $(k-1)$-times continuously differentiable functions $l(\cdot) : [\tau_1, \tau_2] \mapsto \mathbf{R}^n$ such that

(i) $l^{(k-1)}(\cdot)$ is Lipschitz continuous with a Lipschitz constant not bigger than L;

(ii) $\sup\{|l^{(k-1)}(t) : \ t \in [0,T]\} \le L$;

(iii) $\sum_{i=0}^{k-1} |l^{(i)}(t)| \ge \beta \ \ \forall t \in [\tau_1, \tau_2]$.

[2] GPH is not necessary for this strong bang-bang property even in the time-invariant case. A weaker condition, which is sufficient and necessary, is given in Veliov (1987b).

Lemma 3.1 *For every positive real numbers T, L, β and a natural number $k \geq 2$ there exists a constant $d = d(T, L, \beta, k)$ such that for every interval $[\tau_1, \tau_2] \subset [0, T]$, for every $l \in P_k([\tau_1, \tau_2]; L, \beta)$ and for every $\alpha > 0$ it holds that*

$$\text{meas}\{t \in [\tau_1, \tau_2] : \; |l(t)| \leq \alpha\} \leq d\alpha^{1/(k-1)}.$$

We shall use also the following corollary.

Corollary 3.1 *In the conditions of Lemma 3.1, for every $[\tau_1, \tau_2] \subset [0, T]$, for every $l \in P_k([\tau_1, \tau_2]; L, \beta)$, and for every measurable subset $\Delta \subset [\tau_1, \tau_2]$, there is*

$$\int_\Delta |l(t)| \, \mathrm{d}t \geq c \left(\text{meas}(\Delta)\right)^k,$$

where $c = 2^{-k} d^{-k+1}$ and d is the constant from Lemma 3.1.

Proof. Take Δ as in the formulation of the corollary. We shall apply Lemma 3.1 for

$$\alpha = (2d)^{-k+1} (\text{meas}(\Delta))^{k-1}.$$

It claims that for the set $\Delta_\alpha = \{t \in [\tau_1, \tau_2] : \; |l(t)| \leq \alpha\}$

$$\text{meas}(\Delta_\alpha) \leq d\alpha^{1/(k-1)} = 0.5 \, \text{meas}(\Delta).$$

We have

$$\begin{aligned} \int_\Delta |l(t)| \, \mathrm{d}t &\geq \alpha \int_{\Delta \setminus \Delta_\alpha} \mathrm{d}t \geq \alpha(\text{meas}(\Delta) - \text{meas}(\Delta_\alpha)) \\ &\geq 0.5 \, \alpha \, \text{meas}(\Delta) = c \left(\text{meas}(\Delta)\right)^k. \end{aligned} \qquad \blacksquare$$

Proof of Proposition 3.1. For $l \in \mathbf{R}^n$ denote by $\mathcal{E}(l)$ the set of all edges that are incident to an extremal vertex with respect to the direction l, that is,

$$\mathcal{E}(l) = \{e \in E : \; e = [v, w] \text{ with } v \in V(l)\}.$$

First we shall prove that there exist $\beta > 0$ and a neighborhood Z of x such that

$$\sum_{i=0}^{\sigma(x)-1} |\langle \lambda[\tilde{p}](t), B_i(t)e \rangle| \geq \beta \quad \forall \tilde{x} \in \partial R \cap Z, \; \forall t \in [0, T], \; \forall \tilde{p} \in N_R^1(\tilde{x}), \\ \forall e \in \mathcal{E}(B^*(t)\lambda[\tilde{p}](t)). \tag{6}$$

If this is not true, then there exist sequences $\partial R \ni x_k \longrightarrow x$, t_k, $p_k \in N_R^1(x_k)$, $e_k \in \mathcal{E}(B^*(t_k)\lambda[p_k](t_k))$ such that

$$\sum_{i=0}^{\sigma(x)-1} |\langle \lambda[p_k](t_k), B_i(t_k)e_k \rangle| \leq \frac{1}{k}.$$

Passing to subsequences we may assume that $t_k \longrightarrow t$, $p_k \longrightarrow p$, and $e_k = e$ (since E is a finite set). From the continuity of λ and B_i we obtain

$$\sum_{i=0}^{\sigma(x)-1} |\langle \lambda[p](t), B_i(t)e \rangle| = 0. \tag{7}$$

Clearly $p \in N_R^1(x)$ since the mapping $y \longrightarrow N_R^1(y)$ is upper semi-continuous. Moreover, $e = [v, w]$ for some $v \in V(B^*(t_k)\lambda[p_k](t_k))$ and $w \in V$, which implies $v \in V(B^*(t)\lambda[p](t))$ due to the upper semi-continuity of the mapping $l \longrightarrow V(l)$. Relation (7) implies that $\langle \lambda[p](t), B(t)(w - v) \rangle = 0$, hence $w \in V(B^*(t)\lambda[p](t))$. Therefore, $e \in E(B^*(t)\lambda[p](t))$. Then (7) contradicts the definition of $\sigma(x)$ in (5), which proves (6).

In order to prove the local $\sigma(x)$-convexity of R at x we fix an arbitrary $\tilde{x} \in R \cap Z$, $\tilde{x} \neq x$. According to Remark 2.1 we may assume that $\tilde{x} \in \partial R$. Denote

$$\bar{x} = \frac{\tilde{x} + x}{2}.$$

Let $\bar{y} \in \partial R$ be such that $|\bar{x} - \bar{y}| = \max\{\alpha : \; \bar{x} + \alpha\mathcal{B} \subset R\}$. Denote by $u(\cdot)$, $\tilde{u}(\cdot)$ and $\bar{v}(\cdot)$ admissible controls steering the initial state to x, $\tilde{x}$, and $\bar{y}$, respectively. Since all the three points belong to ∂R, according to Remark 3.1 the corresponding controls are uniquely determined piece-wise constant functions with values in V, and having at most $m - 1$ switching points. Clearly, the control $\bar{u} = 0.5(\tilde{u} + u)$ steers x^0 to $\bar{x}$. Since $\tilde{u} \neq u$, the function $\bar{u}$ does not take only values in V. Then assumption (A2) (see Remark 3.1) implies that $\bar{x}$ belongs to the interior of R. Hence, $\bar{y} \neq \bar{x}$, and

$$\bar{p} = \frac{\bar{y} - \bar{x}}{|\bar{y} - \bar{x}|} \in N_R^1(\bar{y})$$

due to the definition of $\bar{y}$. Denote $\zeta(t) = B^*(t)\lambda[\bar{p}](t)$, and let $\Phi(t, \tau)$ be the fundamental matrix solution of $\dot{x} = A(t)x$ normalized at $t = \tau$. From the (Pontryagin) maximum principle we obtain that

$$\langle \zeta(t), \bar{v}(t) \rangle = \max_{u \in U} \langle \zeta(t), u \rangle \tag{8}$$

for almost every $t \in [0, T]$. We obtain from (8) and the Cauchy formula

$$\begin{aligned}
\int_0^T |\langle \zeta(t), \bar{v}(t) - \bar{u}(t) \rangle| \,\mathrm{d}t &= \int_0^T \langle \zeta(t), \bar{v}(t) - \bar{u}(t) \rangle \,\mathrm{d}t \\
&= \int_0^T \langle B^*(t)\Phi^*(T, t)\bar{p}, \bar{v}(t) - \bar{u}(t) \rangle \,\mathrm{d}t \\
&= \int_0^T \langle \bar{p}, \Phi(T, t)B(t)(\bar{v}(t) - \bar{u}(t)) \rangle \,\mathrm{d}t = \left\langle \frac{\bar{y} - \bar{x}}{|\bar{y} - \bar{x}|}, \bar{y} - \bar{x} \right\rangle = |\bar{y} - \bar{x}|.
\end{aligned} \tag{9}$$

Let $W(t)$ be the set of all "best" adjacent vertices to $\bar{v}(t)$ with respect to the direction $\zeta(t)$. That is,

$$v \in W(t) \text{ if and only if } [\bar{v}(t), v] \in E, \text{ and } \langle \zeta(t), v \rangle \geq \langle \zeta(t), u \rangle \; \forall u \in V \setminus \{\bar{v}(t)\}.$$

The mapping W is nonempty compact-valued, and it is easy to check directly that it is measurable (in fact, W is upper semi-continuous, excluding the points of discontinuity of $\bar{v}(\cdot)$). Then it has a measurable selection $\tilde{v}$.

Denote

$$\Delta = \{t \in [0, T] : \; \tilde{u}(t) \neq u(t)\}.$$

Since $\bar{v}(\cdot)$ is piece-wise constant with $\bar{v}(t) \in V(\zeta(t))$, and has at most $m - 1$ switching points, there is an interval $[\tau_1, \tau_2] \subset [0, T]$, such that $\bar{v}(t) = \bar{v}$ is constant on $[\tau_1, \tau_2]$ and $\text{meas}(\Delta \cap [\tau_1, \tau_2]) \geq \text{meas}(\Delta)/m$. Since U has a finite number of vertices, M, there is a subset $\Delta_0 \subset \Delta \cap [\tau_1, \tau_2]$ such that $\tilde{v}(t) = \tilde{v}$ is constant on Δ_0, and $\text{meas}\, \Delta_0 \geq \text{meas}(\Delta)/Mm$.

By the definition of $\tilde{v}(t)$, for every $t \in \Delta$ at least one of the vertices $\tilde{u}(t)$ and $u(t)$ is different from $\tilde{v}(t)$, thus at least one of the inequalities

$$\langle \zeta(t), \tilde{v}(t) \rangle \geq \langle \zeta(t), u(t) \rangle, \quad \langle \zeta(t), \tilde{v}(t) \rangle \geq \langle \zeta(t), \tilde{u}(t) \rangle$$

is fulfilled. Taking into account also (8) we obtain for $t \in \Delta_0$

$$\begin{aligned} |\langle \zeta(t), \bar{v}(t) - \bar{u}(t) \rangle| &= \langle \zeta(t), \bar{v} - \bar{u}(t) \rangle = \langle \zeta(t), \bar{v} \rangle - \frac{1}{2} \langle \zeta(t), u(t) \rangle - \frac{1}{2} \langle \zeta(t), \tilde{u}(t) \rangle \\ &\geq \langle \zeta(t), \bar{v} \rangle - \frac{1}{2} \langle \zeta(t), \bar{v} \rangle - \frac{1}{2} \langle \zeta(t), \tilde{v} \rangle = \frac{1}{2} \langle \zeta(t), \bar{v} - \tilde{v} \rangle = \frac{1}{2} |\langle \zeta(t), \bar{v} - \tilde{v} \rangle|. \end{aligned}$$

Combining this with (9) we obtain

$$\begin{aligned} |\bar{y} - \bar{x}| &= \int_0^T |\langle \zeta(t), \bar{v}(t) - \bar{u}(t) \rangle| \, \mathrm{d}t \geq \int_{\Delta_0} |\langle \zeta(t), \bar{v}(t) - \bar{u}(t) \rangle| \, \mathrm{d}t \\ &\geq \frac{1}{2} \int_{\Delta_0} |\langle \zeta(t), \bar{v} - \tilde{v} \rangle| \, \mathrm{d}t. \end{aligned} \tag{10}$$

Denote $l(t) = \langle \zeta(t), \bar{v} - \tilde{v} \rangle$. According to Remark 3.1

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} l(t) = \langle \lambda[\bar{p}](t), B_k(t)(\bar{v} - \tilde{v}) \rangle.$$

For $t \in [\tau_1, \tau_2]$ we have $\bar{v} \in V(B^*(t)\lambda[\bar{p}](t))$ and $\tilde{v}$ is its adjacent vertex, hence $e = \bar{v} - \tilde{v} \in \mathcal{E}(B^*(t)\lambda[\bar{p}](t))$. Then (6) implies that

$$\sum_{i=0}^{\sigma(x)-1} \left| \frac{\mathrm{d}^k}{\mathrm{d}t^k} l(t) \right| \geq \beta > 0$$

on $[\tau_1, \tau_2]$. Then $l \in P_{\sigma(x)-1}([\tau_1, \tau_2]; L, \beta)$ for an appropriate L. From Corollary 3.1 we obtain that

$$\int_{\Delta_0} |l(t)| \, \mathrm{d}t \geq c \, (\mathrm{meas}\, \Delta_0)^{\sigma(x)} ,$$

where $c = c(T, L, \beta, \sigma(x))$ is as in Lemma 3.1. Combining this with (10) we obtain that

$$|\bar{y} - \bar{x}| \geq c \, (\mathrm{meas}\, \Delta_0)^{\sigma(x)} \geq c \left(\frac{\mathrm{meas}(\Delta)}{Mm} \right)^{\sigma(x)} = c_1 (\mathrm{meas}(\Delta))^{\sigma(x)}.$$

On the other hand, obviously

$$|x - \tilde{x}| \leq c_2 \|u - \tilde{u}\|_{L_1} \leq c_3 \, \mathrm{meas}(\Delta),$$

where c_2 and c_3 are appropriate constants. From the last two exposed inequalities we obtain the claim of the proposition. ■

4. The discrete-time problem

Consider the problem

$$\min g(x(1)), \tag{11}$$

$$\dot{x} = A(t)x + B(t)u, \quad x(0) = x^0, \tag{12}$$

$$u \in U, \tag{13}$$

where A, B, and U satisfy (A1), (A2) from the previous section. Let $(\hat{u}, \hat{x})$ be a solution of the above problem.

Assumption (A3): $g : \mathbf{R}^n \mapsto \mathbf{R}$ is differentiable with locally Lipschitz derivative, and is locally κ-convex at $\hat{x}(T)$ with $\kappa \in [2, +\infty)$; moreover, $g'(\hat{x}(T)) \neq 0$.

Under the above conditions $(\hat{u}, \hat{x})$ is the unique solution, and it has the bang-bang property described in Remark 3.1.

The discretized version of the above problem will be obtained in the following way. For a natural number N and $h = T/N$ we fix the uniform mesh $t_k = kh$, $k = 0, \ldots, N$. On every subinterval $[t_k, t_{k+1}]$, and for an arbitrary constant $u(t) \equiv u_k \in U$ we apply to (12) a given Runge-Kutta scheme of at least 3-rd order local consistency. This procedure formally defines a discrete-time control system of the form

$$x_{k+1} = A_N(k)x_k + B_N(k)u_k, \quad x_0 = x^0, \quad k = 0, \ldots, N-1, \tag{14}$$

$$u_k \in U, \tag{15}$$

where $A_N(k)$ and $B_N(k)$ are matrices of respective dimensions, which are explicitly defined if an explicit Runge-Kutta scheme is applied, or involve solving a linear system of equations, otherwise.

In general a (q-stage) Runge-Kutta scheme is defined by the so-called Butcher array

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1q} \\ \vdots & \vdots & & \vdots \\ c_q & a_{q1} & \dots & a_{qq} \\ \hline & b_1 & \dots & b_q \end{array}$$

(see Butcher, 1987). The following conditions implying 3-rd order local consistency will be assumed further.

Assumption (A4)

$$(i) \sum_{i=1}^{q} b_i = 1, \quad (ii) \sum_{i=1}^{q} \sum_{j=1}^{q} b_i a_{ij} = \frac{1}{2}, \quad (iii) \sum_{i=1}^{q} b_i c_i = \frac{1}{2},$$
$$(iv)\; b_i \geq 0, \quad c_i \in [0,1]. \tag{16}$$

For example, for the widely used (especially in optimal control context) Heun scheme, for which $b = (0.5, 0.5)$, $c = (0,1)$, $a_{21} = 1$, all other $a_{ij} = 0$, one has

$$\begin{aligned} A_N(k) &= I + 0.5h(A(t_k) + A(t_{k+1})) + 0.5h^2 A(t_{k+1})^2, \\ B_N(k) &= 0.5h(B(t_k) + B(t_{k+1})) + 0.5h^2 A(t_{k+1}) B(t_{k+1}). \end{aligned}$$

The following two lemmas are standard and we skip the proofs.

LEMMA 4.1 *Assume (A1) and (A4). Then there exist numbers N_0 and C, such that for every $N \geq N_0$ and for every $u_k \in U$, $k = 0, \dots, N-1$ it holds that*

$$|x_{k+1} - x(t_{k+1})| \leq Ch^2, \quad k = 0, \dots, N,$$

where $\{x_{k+1}\}$ is obtained from (14), and $x(t_{k+1})$ is the value at $t = t_{k+1}$ of the solution of (12) which corresponds to the piece-wise constant control defined by $\{u_k\}$.

LEMMA 4.2 *Assume (A1) and (A4). Then given a compact set $\Lambda \subset \mathbf{R}^n$, there exist numbers N_0 and C, such that for every $N \geq N_0$ and for every $\lambda_N \in \Lambda$ it holds that*

$$|\lambda_k - \lambda(t_k)| \leq Ch^2, \quad k = 1, \dots, N,$$

where the sequence $\{\lambda_k\}$ is generated by the discrete backward dynamics

$$\lambda_k = A_N^*(k) \lambda_{k+1},$$

and $\lambda(t_k)$ is the value at $t = t_k$ of the backward solution of adjoint equation (4) which starts from λ_N at time T.

The discretized version of problem (11)–(13) is

$$\min g(x_N), \quad \text{subject to (14) and (15).} \tag{17}$$

As an auxiliary step to our main result we consider the control system (12), (13) in a restrained set of admissible controls, $\mathcal{U}_N$, consisting of all function with values in U, which are constant on each subinterval (t_k, t_{k+1}). Denote by $\tilde{R}_N$ the reachable set of (12), (13) on $[0, T]$, in the set $\mathcal{U}_N$ of admissible controls. The following result is proven in Veliov (1997):

Lemma 4.3 *Under (A1) there exist numbers N_0 and C such that for every $N \geq N_0$*

$$H(R, \tilde{R}_N) \leq Ch^2.$$

Denote by R_N the reachable set of (14), (15) at $k = N$, that is, the set of all end points x_N of trajectories of (14) corresponding to arbitrary selections $u_k \in U$. Due to Lemma 4.1 we obtain in a standard way the following

Lemma 4.4 *Under (A1) and (A4) there exist numbers N_0 and C such that for every $N \geq N_0$*

$$H(\tilde{R}_N, R_N) \leq Ch^2.$$

The following proposition combines the last two lemmas:

Proposition 4.1 *Let (A1) and (A4) hold. Then there exist numbers N_0 and C such that for every $N \geq N_0$*

$$H(R, R_N) \leq Ch^2.$$

5. The error estimate

In this section we formulate and prove the main result, which estimates the difference between the solution $(\hat{x}(\cdot), \hat{u}(\cdot))$ of (11)–(13) and an arbitrary solution $(\hat{x}^N, \hat{u}^N) = ((\hat{x}_0, \dots, \hat{x}_N), (\hat{u}_0, \dots, \hat{u}_{N-1}))$ of the discrete-time problem (17). The proof combines propositions 2.1, 3.1, and 4.1, with the following auxiliary result. We shall abbreviate $\sigma = \sigma(\hat{x}(T))$, and $\lambda(t) = \lambda[-g'(\hat{x}(T))](t)$.

Lemma 5.1 *Assume (A1) and (A2). Then there exists a number c such that the following is true: For every interval $[\tau_1, \tau_2] \subset [0, T]$ in which $\hat{u}(t) = u$ is constant, and for every $\alpha > 0$ there exists a measurable set $\Delta_\alpha \subset [\tau_1, \tau_2]$ with*

$$\text{meas}(\Delta_\alpha) \leq c\alpha^{1/(\sigma-1)}, \tag{18}$$

such that each vector l satisfying

$$|B^*(\tau)\lambda(\tau) - l| < \alpha \tag{19}$$

for some $\tau \in [\tau_1, \tau_2] \setminus \Delta_\alpha$, verifies $V(l) = \{u\}$.

Proof. Let $\hat{u}(t) = u$ be constant on $[\tau_1, \tau_2]$. Since $-g'(\hat{x}(T)) \in N^1_R(\hat{x}(T))$, it follows from (6) that

$$\sum_{i=0}^{\sigma-1} |\langle \lambda(t), B_i(t)e\rangle| \geq \beta > 0 \quad \forall t \in [0,T],\ \forall e \in \mathcal{E}(B^*(t)\lambda(t)).$$

In particular, since $u \in V(B^*(t)\lambda(t))$ for $t \in [\tau_1, \tau_2]$, the above inequality holds for such t with any $e = u - v$, where v is an adjacent vertex to u. Thus, the function $\xi[v](\cdot) = \langle B^*(\cdot)\lambda(\cdot), u - v\rangle$ belongs to $P_\sigma([\tau_1, \tau_2]; L, \beta)$ (for an appropriate L) and we can apply Lemma 3.1. For the set

$$\Delta_\alpha(v) = \{t \in [\tau_1, \tau_2] :\ |\xi[v](t)| \leq \alpha \operatorname{diam}(U)\}$$

(where $\operatorname{diam}(U)$ is the diameter of the set U) we have

$$\operatorname{meas}(\Delta_\alpha(v)) \leq d(\alpha \operatorname{diam}(U))^{1/(\sigma-1)}.$$

Define $\Delta_\alpha = \cup \Delta_\alpha(v)$, where the union is taken over all vertices v for which $[u, v] \in E$. Clearly, Δ_α satisfies (18) with $c = jd(\operatorname{diam}(U))^{1/(\sigma-1)}$, where j is the number of vertices as above.

Take an arbitrary l as in the formulation of the lemma. Let $\tau \in [\tau_1, \tau_2] \setminus \Delta_\alpha$ be a time for which (19) holds, and let v be an arbitrary adjacent vertex to u. Since $\tau \notin \Delta_\alpha(v)$,

$$\begin{aligned} \langle l, u - v\rangle &> \langle B^*(\tau)\lambda(\tau), u - v\rangle - \alpha|u - v| \geq |\xi[l](t)| - \alpha|u - v| \\ &\geq \alpha \operatorname{diam}(U) - \alpha|u - v| \geq 0. \end{aligned}$$

Since the above strict inequality holds for every adjacent vertex to u, it holds also for every $v \in U \setminus \{u\}$, which means that $V(l) = \{u\}$. ■

In order to compare the solution $(\hat{x}(\cdot), \hat{u}(\cdot))$ of (11)–(13) with an arbitrary solution $(\hat{x}^N, \hat{u}^N)$ of the discrete-time problem (17) we introduce the piecewise constant extension $\hat{u}^N(\cdot)$ of $(\hat{u}_0, \ldots, \hat{u}_{N-1})$. As above we use the notation $\sigma = \sigma(\hat{x}(T))$, and also $s = \min\{\sigma, \kappa(\hat{x}(T))\}$, where $\kappa(\hat{x}(T))$ is the local convexity index of g at $\hat{x}(T)$.

THEOREM 5.1 *Assume (A1)–(A4). Then there exist numbers C and N_0 such that for every $N \geq N_0$ and for every optimal control $\hat{u}^N$ of (17) the following estimation holds:*

$$\operatorname{meas}\{t \in [0,T] :\ \hat{u}^N(t) \neq \hat{u}(t)\} \ \leq\ C\left(h^{\frac{1}{\sigma-1}} + L_{g'} h^{\frac{2}{s(\sigma-1)}}\right), \tag{20}$$

where $L_{g'}$ is the Lipschitz constant of g' at $\hat{x}(T)$.

Proof. Below $c_1,\ c_2, \ldots$ will be appropriate constants independent of N. As before, denote $\lambda(t) = \lambda[-g'(\hat{x}(T))](t)$.

The solution of (17) satisfies the maximum principle

$$\hat{u}_k \in V\left(\frac{1}{h}B_N^*(k)\lambda_{k+1}\right), \quad k = 0, \dots N-1, \tag{21}$$

where λ_k is the solution of

$$\lambda_k = A_N^*(k)\lambda_{k+1}, \quad k = N-1, \dots, 1, \quad \lambda_N = -g'(x_N).$$

Combining propositions 2.1, 3.1, and 4.1 we obtain that

$$|\hat{x}_N - \hat{x}(T)| \le c(H(R, R_N))^{\frac{1}{s}} \le c_1 h^{\frac{2}{s}}.$$

Hence,

$$|\lambda_N - \lambda(T)| \le L_{g'} c_1 h^{\frac{2}{s}}.$$

Due to Lemma 4.2 and the Lipschitz continuity of the solution of the adjoint equation (4) on the final condition, we obtain that

$$|\lambda_{k+1} - \lambda(t_{k+1})| \le c_2\left(h^2 + L_{g'} h^{\frac{2}{s}}\right).$$

Thanks to condition (16, i) for the scheme, we have $|\frac{1}{h}B_N(k) - B(t_{k+1})| \le c_3 h$, thus

$$\left|\frac{1}{h}B_N^*(k)\lambda_{k+1} - B^*(t_{k+1})\lambda(t_{k+1})\right| \le c_4\left(h + L_{g'} h^{\frac{2}{s}}\right),$$

Let $[\tau_1, \tau_2]$ be a maximal interval in which $\hat{u}(t)$ is constant and equals some u (hence, $u \in V(B^*(t)\lambda(t))$ on $[\tau_1, \tau_2]$). Since λ is Lipschitz continuous, for all k and $t \in [t_k, t_{k+1}]$

$$\left|\frac{1}{h}B_N^*(k)\lambda_{k+1} - B^*(t)\lambda(t)\right| \le c_5\left(h + L_{g'} h^{\frac{2}{s}}\right). \tag{22}$$

Now we shall apply Lemma 5.1 with

$$\alpha = 2c_5\left(h + L_{g'} h^{\frac{2}{s}}\right).$$

From (18) we obtain an estimation as in (20) for the set Δ_α from Lemma 5.1. Now we shall prove that for all k such that $[t_k, t_{k+1}]$ is contained in $[\tau_1, \tau_2]$ but is not contained in Δ_α, it holds that $u_k = u$. This implies the claim of the theorem, since $\hat{u}$ has only a finite number of jumps, that is, there is only a finite number of maximal intervals like $[\tau_1, \tau_2]$.

Take some k as in the last paragraph. We apply Lemma 5.1 with $l = \frac{1}{h}B_N^*(k)\lambda_{k+1}$. By the choice of k, the interval $[t_k, t_{k+1}]$ contains some $\tau \not\in \Delta_\alpha$. Due to (22) we have $|\lambda_k - \lambda(t)| < \alpha$. Then Lemma 5.1 implies $V(l) = \{u\}$, and (21) gives $u_k = u$. The theorem is proven. ■

If g is linear, then $L_{g'} = 0$ and the estimation becomes of order $h^{\frac{1}{\sigma-1}}$. Clearly, the more interesting case is that of $L_{g'} > 0$. Then the first term in the right-hand side of (20) can be skipped. The most important case is that of a (locally) strongly convex function g. In this case $s = \kappa(\hat{x}(T)) = 2$ and (20) reduces again to

$$\text{meas}\{t \in [0,T] : \; \hat{u}^N(t) \neq \hat{u}(t)\} \; \leq \; Ch^{\frac{1}{\sigma-1}}. \tag{23}$$

If $\sigma = 2$ (which is related to the conditions for structural stability of the optimal control, Felgenhauer, 2003) the estimation is of first order with respect to h, which obviously cannot be improved for the considered discretization. The estimation (23) seems to be sharp also for any value of σ, since in the case $\sigma > 2$ the sensitivity of the optimal control to perturbations is of non-Lipschitz (Hölder) type: an arbitrarily small perturbation in the data may create a bifurcation of "hidden" switching points of multiplicity $\sigma - 1$.

Clearly, the number $\sigma = \sigma(\hat{x}(T))$ is not known in advance. Condition (A2), however, can be verified and the number $\bar{\sigma}$ can be found, which can be used in (23) instead of σ. We mention also that according to the results in Veliov (1987a) the points x on ∂R for which $\sigma(x) > 2$ are situated on a $(n-2)$-dimensional continuous parametric manifold, and in this sense the case $\sigma(x) = 2$ is "typical".

We mention also that in the case $\sigma(x) = 2$ the result in Theorem 5.1 can be improved in a straightforward way: in addition to (23), the set $\{t \in [0,T] : \hat{u}^N(t) \neq \hat{u}(t)\}$ consists of finite number of intervals (this is clear in advance) and each interval contains a jump point of $\hat{u}$ or one of the points 0 and T. That is, the structure of $\hat{u}$ is preserved in $\hat{u}^N$.

Finally we mention that the results from sections 2 and 3, can be used to estimate the sensitivity of the optimal control with respect to "'regular" perturbations, hence to extend the results in Felgenhauer (2003) in two directions: (i) to obtain estimates of the sensitivity; (ii) to treat the case of multiple zeros of the switching function.

References

Agrachev, A.A., Stefani, G. and Zezza, P. (2002) Strong optimality for a bang-bang trajectory. *SIAM J. Control Optim.* **41** (4), 991–1014.

Butcher, J.C. (1987) *The numerical analysis of ordinary differential equations.* John Wiley and Sons.

Dontchev, A.L. and Hager, W.W. (1993) Lipschitzian stability in nonlinear control and optimization. *SIAM J. Control Optim.* **31** (3), 569–603.

Dontchev, A.L. and Hager, W.W. (2001) The Euler approximation in state constrained optimal control. *Math. Comp.* **70** (233), 173–203.

Dontchev, A.L., Hager, W.W. and Veliov, V.M. (2000) Second-order Runge-Kutta approximations in control constrained optimal control, *SIAM J. Numerical Anal.* **38** (1), 202–226.

Felgenhauer, U. (2003) On stability of bang-bang type controls. *SIAM J. Control Optim.* **41** (6), 1843–1867.

Felgenhauer, U. (2005) On the optimality of optimal bang-bang controls for linear and semilinear systems. *Control & Cybernetics*. To appear.

Frankowska, H. and Olech, Cz. (1980) *R*-convexity of the integral of set-valued functions. *Contributions to analysis and geometry*, Baltimore, Md., 1980, 117–129; Johns Hopkins Univ. Press, Baltimore, Md., 1981.

Łojasiewicz, St. Jr. (1979) Some properties of accessible sets in nonlinear control systems. *Ann. Polon. Math.*, **36**(2):123–137, .

Malanowski, K., Büskens, Ch. and Maurer, H. (1998) Convergence of approximations to nonlinear optimal control problems. In: A.V. Fiacco, ed., *Mathematical programming with data perturbations*, *Lecture Notes in Pure and Appl. Math.* **195**, Dekker, New York, 253–284.

Maurer, H. and Osmolovskii, N. (2004) Second order sufficient conditions for time-optimal bang-bang control. *SIAM J. Control Optim.* **42** (6), 2239–2263.

Noble, J. and Schättler, H. (2002) Sufficient conditions for relative minima of broken extremals in optimal control theory. *J. Math. Anal. Appl.* **269** (1), 98–128.

Osmolovskii, N.P. (1998) Second-order conditions for broken extremal. In: *Calculus of variations and optimal control*, Haifa, 1998, 198–216; Chapman & Hall/CRC Res. Notes Math. **411**, Boca Raton, FL, 2000.

Pliś, A. (1975) Accessible sets in control theory. *International Conference on Differential Equations*, Academic Press, 646–650.

Polovinkin, E. (1996) Strongly convex analysis. *Mat. Sb.* **187** (2), 103–130; translation in *Sb. Math.* **187** (2), 259–286.

Polyak, B.T. (1983) *Introduction to optimization* (in Russian). Nauka, Moscow.

Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V. and Mishchenko, E.F. (1962) *The mathematical theory of optimal processes.* John Wiley & Sons.

Veliov, V.M. (1987a) On the convexity of integrals of multivalued mappings: applications in control theory. *J. Optim. Theory Appl.* **54** (3), 541–563.

Veliov, V.M. (1987b) On the bang-bang principle for linear control systems. *Comptes Rendus de l'Academie Bulgare des Sciences* **4** (2), 31–33.

Veliov, V.M. (1997) On the time-discretization of control systems. *SIAM J. Control Optim.* **35** (5), 1470–1486.