

# Higher-Order Numerical Scheme for Linear Quadratic Problems with Bang-Bang Controls

*Teresa Scarinci and Vladimir M. Veliov*

**Research Report 2017-06**

June 2017

ISSN 2521-313X

**Operations Research and Control Systems**  
Institute of Statistics and Mathematical Methods in Economics  
Vienna University of Technology

Research Unit ORCOS  
Wiedner Hauptstraße 8 / E105-4  
1040 Vienna, Austria  
E-mail: [orcos@tuwien.ac.at](mailto:orcos@tuwien.ac.at)

# Higher-Order Numerical Scheme for Linear Quadratic Problems with Bang-Bang Controls\*

T. Scarinci<sup>†</sup> and V.M. Veliov<sup>‡</sup>

## Abstract

This paper considers a linear-quadratic optimal control problem where the control function appears linearly and takes values in a hypercube. It is assumed that the optimal controls are of purely bang-bang type and that the switching function, associated with the problem, exhibits a suitable growth around its zeros. The authors introduce a scheme for the discretization of the problem that doubles the rate of convergence of the Euler's scheme. The proof of the accuracy estimate employs some recently obtained results concerning the stability of the optimal solutions with respect to disturbances.

*Keywords:* optimal control, numerical methods, bang-bang control, linear-quadratic optimal control problems, time-discretization methods

*MSC Classification:* 49M25, 65L99, 49J30, 49N10, 49J15

## 1 Introduction

Discretization schemes for optimal control problems have been largely investigated in the last sixty years (see, e.g., [8, 21, 9, 17, 6, 7], and the more recent paper [4] and the references therein). In the aforementioned papers and in most of the literature, the optimal controls are typically assumed to be sufficiently smooth (at least Lipschitz continuous) and results are usually based on second-order optimality conditions. On the other hand, whenever the control appears linearly in the system, the lack of coercivity typically leads to discontinuities of the optimal controls.

Recently, new second-order optimality conditions for systems that are linear with respect to the control have been developed. We refer to [22] and [3] for analysis of second-order necessary conditions for bang-bang and singular-bang controls, respectively. Results on the stability of solutions with respect to disturbances were also recently obtained, see [5, 14, 12, 25] and the bibliography therein. Based on these results, error estimates for the accuracy of the Euler discretization scheme applied to various classes of affine optimal control problems were obtained in [1, 2, 18, 26, 13, 27].

---

\*Both authors were supported by the Austrian Science Foundation (FWF) under Grant No P26640-N25. The first author was also supported by the Doctoral Programme *Vienna Graduate School on Computational Optimization*, funded by Austrian Science Foundation under Project No W1260-N35.

<sup>†</sup>Department of Statistic and Operation Research, University of Vienna, Austria, [teresa.scarinci@gmail.com](mailto:teresa.scarinci@gmail.com) (corresponding author)

<sup>‡</sup>Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria, [vladimir.veliov@tuwien.ac.at](mailto:vladimir.veliov@tuwien.ac.at)

The error estimates are at most of first order with respect to the discretization step, which is natural in view of the discontinuity of the optimal control. For the same reason, using higher order Runge-Kutta discretization schemes on a fixed grid does not help to improve the order of accuracy. Seemingly, the first paper that addresses the issue of accuracy of discrete approximations for affine problems is [30], where a higher order Runge-Kutta scheme is applied to a linear system, but the error estimate is of first order or less.

A new type of discretization scheme was recently presented in [23] for Mayer’s problems for linear systems. The idea behind this scheme goes back to [28, 15, 20] and is based on a truncated Volterra-Fliess-type expansion of the state and adjoint equations. The analysis of the convergence and of the error estimate makes use of the *strong Hölder metric sub-regularity* of the map associated with the Pontryagin maximum principle, proved in [25].

The goal of the present paper is to extend this discretization scheme and the pertaining error analysis to affine linear-quadratic problems. This extension is not a routine task, due to the appearance of the state function in the associated with the problem switching function, and of both the state and the control, in the adjoint equation.

More precisely, we consider the problem

$$J(x, u) := \frac{1}{2}x(T)^\top Qx(T) + q^\top x(T) + \int_0^T \left( \frac{1}{2}x(t)^\top W(t)x(t) + x(t)^\top S(t)u(t) \right) dt \longrightarrow \min \quad (1.1)$$

subject to

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0, \quad \text{for a.e. } t \in [0, T], \quad (1.2)$$

$$u(t) \in U := [-1, 1]^m \quad \text{for a.e. } t \in [0, T]. \quad (1.3)$$

Here  $[0, T]$  is a fixed time horizon, the state  $x$  is  $n$ -dimensional, the initial state  $x_0$  is given, the control  $u$  is  $m$ -dimensional,  $Q \in \mathbb{R}^{n \times n}$ ,  $q \in \mathbb{R}^n$ , and the matrix functions  $A, W : [0, T] \rightarrow \mathbb{R}^{n \times n}$  and  $S, B : [0, T] \rightarrow \mathbb{R}^{n \times m}$  are given data; the superscript “ $\top$ ” denotes transposition. Admissible controls are all measurable functions with values in the set  $U$  for a.e.  $t \in [0, T]$ . Linear terms are not included in the integrand in (1.1), since they can be shifted in a standard way into the differential equation (1.2).

The optimal controls in the problem (1.1)–(1.3) are typically concatenations of bang-bang and singular paths. In this paper, we assume that the optimal controls are of strictly bang-bang type with a finite number of switches, and the components of the switching function have a certain growth rate at their zeros, characterized by a number  $\kappa \geq 1$ . This number appears in the error estimate obtained in this paper for the proposed discretization scheme. “Generically”,  $\kappa = 1$ , and in this case the error estimate is of second order.

The paper is organized as follows. In Section 2 we recall some notations and formulate the assumptions. In Section 3 we introduce our discretization scheme and present the main result – the error estimate. Section 4 contains the proof. Section 5 presents an error estimate in case of inexact solution of the discretized problem. A numerical experiment confirming the theoretical findings is given in Section 6. Concluding remarks and further perspectives are discussed in Section 7.

## 2 Preliminaries

First of all we pose some assumptions, which are standard in the context of problem (1.1)–(1.3).

**Assumption (A1).** The matrix functions  $A(t)$ ,  $B(t)$ ,  $W(t)$  and  $S(t)$ ,  $t \in [0, T]$ , have Lipschitz continuous first derivatives,  $Q$  and  $W(t)$  are symmetric.

The set of all admissible controls will be denoted by  $\mathcal{U} \subset L^\infty$ . A pair  $(x, u)$  formed by an admissible control  $u$  and the corresponding solution  $x$  of (1.2) is referred to as an admissible process, and the set of all admissible processes is denoted by  $\mathcal{F}$ . The set  $\mathcal{F}$  will be considered as a subset of the space  $W_{x_0}^{1,1} \times L^1$ , where  $W_{x_0}^{1,1} = W_{x_0}^{1,1}(0, T)$  is the (affine) space of all absolutely continuous functions  $x : [0, T] \rightarrow \mathbb{R}^n$  with  $x(0) = x_0$ , and  $L^1 = L^1(0, T)$  has the usual meaning.<sup>1</sup>

Due to the compactness and the convexity of  $U$ , the set  $\mathcal{F}$  is compact with respect to the  $L^2$ -weak topology for  $u$  and the uniform norm for  $x$ . Thus a minimizer  $(\hat{x}, \hat{u})$  does exist in the space  $W_{x_0}^{1,1} \times L^1$  (in fact, also in  $W_{x_0}^{1,\infty} \times L^\infty$ ). The following assumption requires a sort of “directional convexity” of the objective functional  $J$  at  $(\hat{x}, \hat{u})$ .

**Assumption (A2):**

$$\frac{1}{2}z(T)^\top Qz(T) + \int_0^T \left( \frac{1}{2}z(t)^\top W(t)z(t) + z(t)^\top S(t)v(t) \right) dt \geq 0 \quad \text{for every } (z, v) \in \mathcal{F} - (\hat{x}, \hat{u}).$$

Let  $(\hat{x}, \hat{u})$  be an optimal process in problem (1.1)–(1.3). According to the Pontryagin (minimum) principle, there exists  $\hat{p} \in W^{1,\infty}$  such that  $(\hat{x}, \hat{u}, \hat{p})$  satisfies the following system of generalized equations: for a.e.  $t \in [0, T]$ ,

$$0 = \dot{x}(t) - A(t)x(t) - B(t)u(t), \quad x(0) = x_0, \quad (2.1)$$

$$0 = \dot{p}(t) + A(t)^\top p(t) + W(t)x(t) + S(t)u(t), \quad (2.2)$$

$$0 \in B(t)^\top p(t) + S(t)^\top x(t) + N_U(u(t)), \quad (2.3)$$

$$0 = p(T) - Qx(T) - q, \quad (2.4)$$

where  $N_U(u)$  is the normal cone to  $U$  at  $u$ :

$$N_U(u) = \begin{cases} \emptyset & \text{if } u \notin U, \\ \{l \in \mathbb{R}^m : \langle l, v - u \rangle \leq 0 \quad \forall v \in U\} & \text{if } u \in U. \end{cases}$$

System (2.1)–(2.4) can be shortly rewritten as

$$0 \in F(x, p, u), \quad (2.5)$$

where  $F$  is the set-valued map defined by

$$F(x, p, u) := \begin{pmatrix} \dot{x} - Ax - Bu \\ \dot{p} + A^\top p + Wx + Su \\ B^\top p + S^\top x + N_U(u) \\ p(T) - Qx(T) - q \end{pmatrix}.$$

The mapping  $F$  is considered as acting from the space  $\mathcal{X}$  to the space  $\mathcal{Y}$ , where

$$\mathcal{X} := W_{x_0}^{1,1} \times W^{1,1} \times L^1, \quad \mathcal{Y} := L^1 \times L^1 \times L^\infty \times \mathbb{R}^n.$$

---

<sup>1</sup> To avoid a confusion, we mention that the admissible controls, hence the derivative of the state function,  $\dot{x}$ , as well as the derivative of the adjoint function,  $\dot{p}$ , appearing below, belong to  $L^\infty$ . However, we use the  $L^1$ -norms of these derivatives in most of the considerations.

The norms in these spaces are defined as usual: for  $(x, p, u) \in \mathcal{X}$  and  $(\xi, \pi, \rho, \nu) \in \mathcal{Y}$ ,

$$\|x\|_{1,1} + \|p\|_{1,1} + \|u\|_1$$

and

$$\|(\xi, \pi, \rho, \nu)\| := \|\xi\|_1 + \|\pi\|_1 + \|\rho\|_\infty + |\nu|,$$

where  $\|x\|_{1,1}$  abbreviates  $\|x\|_{W^{1,1}}$  and  $\|\cdot\|_s$  is the norm in  $L^s$ ,  $s \in \{1, \infty\}$ . Notice that the normal cone  $N_{\mathcal{U}}(u)$  to the set  $\mathcal{U} \subset L^1$  at  $u(\cdot) \in \mathcal{U}$  has the point-wise representation  $\{\xi \in L^\infty : \xi(t) \in N_U(u(t)) \text{ a.e. on } [0, T]\}$ .

We also define the distance

$$d^\#(u_1, u_2) = \text{meas} \{t \in [0, T] : u_1(t) \neq u_2(t)\}$$

in the space  $L^\infty$ . We mention that  $\mathcal{U} \subset L^\infty$  is a complete metric space with respect to this metric ([11, Lemma 7.2]).

Observe that the inclusion (2.3) is equivalent to  $u(t) \in \underset{w \in U}{\text{Argmin}} \sigma(t)^\top w$ , where  $\sigma : [0, T] \rightarrow \mathbb{R}^m$  is the so-called *switching function*, defined for all  $t \in [0, T]$  as

$$\sigma(t) = B(t)^\top \hat{p}(t) + S(t)^\top \hat{x}(t).$$

Thus, for  $j = 1, \dots, m$ ,

$$\hat{u}^j(t) = \begin{cases} -1 & \text{if } \sigma^j(t) > 0, \\ 1 & \text{if } \sigma^j(t) < 0, \end{cases} \quad (2.6)$$

where  $\sigma^j$  and  $\hat{u}^j$  stay for the  $j$ -component of  $\sigma$  and  $\hat{u}$ , respectively.

**Assumption (B)** (Strict bang-bang property). There exist positive real numbers  $\kappa \geq 1$ ,  $m_0$  and  $\delta$  such that for every  $j = 1, \dots, m$ , and for every zero  $\tau \in [0, T]$  of  $\sigma^j$ , the inequality  $|\sigma^j(t)| \geq m_0 |t - \tau|^\kappa$  holds for all  $t \in [\tau - \delta, \tau + \delta] \cap [0, T]$ .

*Remark 2.1.* Clearly, Assumption (B) implies that each component  $\sigma^j$  has a finite number of zeros in  $[0, T]$  and then each component of  $\hat{u}$  is piecewise constant with values  $-1$  and  $1$ .

The following theorem plays a crucial role in the error analysis of the discretization scheme presented below. It is a modification (under weaker conditions and a different space setting) of [5, Theorem 8], and is proved in [24].

**Theorem 2.2.** *Let Assumption (A1) and (A2) be fulfilled. Let  $(\hat{x}, \hat{p}, \hat{u})$  be a solution of generalized equation (2.5) and let assumption (B) be fulfilled with some real number  $\kappa \geq 1$ . Then for any  $b > 0$  there exists  $c > 0$  such that for every  $y := (\xi, \pi, \rho, \nu) \in \mathcal{Y}$  with  $\|y\| \leq b$ , there exists a triple  $(x, p, u) \in \mathcal{X}$  solving  $y \in F(x, p, u)$  and every such triple satisfies the inequality*

$$\|x - \hat{x}\|_{1,1} + \|p - \hat{p}\|_{1,1} + \|u - \hat{u}\|_1 \leq c \|y\|^{1/\kappa}. \quad (2.7)$$

We mention that the above property of the mapping  $F$  and the reference point  $(\hat{x}, \hat{u}, \hat{p}) \in \mathcal{X}$  and  $0 \in \mathcal{Y}$  is a stronger version (non-local with respect to  $(x, p, u)$ ) of the so-called *metric sub-regularity*, [10].

### 3 Discretization Scheme

In this section we propose a discretization scheme for problem (1.1)–(1.3) which has a higher accuracy than the Euler scheme without a substantial increase of the numerical complexity of the discretized problem. We recall that the Euler method has already been profoundly investigated in the case where bang-bang controls appear (e.g. [1, 2, 18, 26, 13]). As mentioned in the introduction, in doing this we use an idea that originates from [28, 15] and was implemented in [23] in the case of Mayer’s problems. The approach uses second order truncated Volterra-Fliess series, as described in the next subsection.

#### 3.1 Truncated Volterra-Fliess series

Given a natural number  $N$ , denote  $h = T/N$ , and  $t_i = ih$  for  $i = 0, \dots, N$ . Let  $u$  be an admissible control on  $[t_i, t_{i+1}]$ . The solution  $x$  of (1.2) on the interval  $[t_i, t_{i+1}]$  can be represented as (see [23, Section 3])

$$\begin{aligned} x(t) = & \left[ I + (t - t_i)A + \frac{(t - t_i)^2}{2}(A^2 + A') \right] x(t_i) + (B + (t - t_i)AB) \int_{t_i}^t u(s) ds \\ & + (-AB + B') \int_{t_i}^t (s - t_i)u(s) ds + O(h^3), \end{aligned} \quad (3.1)$$

where for shortness we skip the fixed argument  $t_i$  in the appearing functions, that is,  $A := A(t_i)$ ,  $B := B(t_i)$ , etc. As usual, we denote by  $O(\varepsilon)$ ,  $\varepsilon > 0$ , any function that satisfies  $|O(\varepsilon)|/\varepsilon \leq c$ , where  $c$  is a “generic” constant, that is, depending only on the data of the problem (thus, independent of  $i$  and  $t$ , although  $O(h^3)$  may depend on  $i$  and  $t$  in the above context). The second-order expansion of the solution of the adjoint equation (2.2) differs from that in [23, Section 3] due to the presence of an integral term in the objective functional (1.1), therefore we derive it below. For all  $t \in [t_i, t_{i+1}]$ ,

$$p(t) = \int_t^{t_{i+1}} \left( A(s)^\top p(s) + W(s)x(s) + S(s)u(s) \right) ds + p(t_{i+1}). \quad (3.2)$$

Applying the first order Taylor expansion for  $A$ ,  $W$ ,  $S$  at  $t_i$ , the representations (3.1) of  $x(s)$  and (3.2) of  $p(s)$ , and skipping all third order terms we obtain that

$$\begin{aligned} p(t) = & \int_t^{t_{i+1}} (A + (s - t_i)A')^\top \left( p(t_{i+1}) + \int_s^{t_{i+1}} \left( A^\top p(t_{i+1}) + Wx(t_i) + Su(\zeta) \right) d\zeta \right) ds \\ & + \int_t^{t_{i+1}} (W + (s - t_i)W') \left( x(t_i) + (s - t_i)Ax(t_i) + B \int_{t_i}^s u(\zeta) d\zeta \right) ds \\ & + \int_t^{t_{i+1}} (Su(s) + (s - t_i)S'u(s)) ds + p(t_{i+1}) + O(h^3). \end{aligned}$$

Hence, we obtain the following truncated Volterra-Fliess expansion for the adjoint function:

$$\begin{aligned}
p(t) &= \left( I + (t_{i+1} - t)A^\top + \frac{h^2 - (t - t_i)^2}{2}A'^\top \right) p(t_{i+1}) \\
&+ \frac{(t_{i+1} - t)^2}{2}A^\top \left( A^\top p(t_{i+1}) + Wx(t_i) \right) + A^\top S \int_t^{t_{i+1}} \int_s^{t_{i+1}} u(\zeta) d\zeta ds \\
&+ (t_{i+1} - t)Wx(t_i) + \frac{h^2 - (t - t_i)^2}{2}(W' + WA)x(t_i) + WB \int_t^{t_{i+1}} \int_{t_i}^s u(\zeta) d\zeta ds \\
&+ S \int_t^{t_{i+1}} u(s) ds + S' \int_t^{t_{i+1}} (s - t_i)u(s) ds + O(h^3).
\end{aligned} \tag{3.3}$$

Now we shall derive a second order approximation to the integral term of the objective functional  $J(x, u)$  on  $[t_i, t_{i+1}]$  in term of the first and second order momentum of the controls.

Concerning the quadratic term in the  $x$ -variable we make use of the first order part of representation (3.1) and of the Taylor expansion at  $t_i$  for  $W$ . Remembering that  $W$  is a symmetric-matrix-valued function, we obtain that

$$\begin{aligned}
&\int_{t_i}^{t_{i+1}} x(t)^\top W(t)x(t) dt = \int_{t_i}^{t_{i+1}} \left( x(t_i) + (t - t_i)A(t_i)x(t_i) + B(t_i) \int_{t_i}^t u(s) ds \right)^\top \\
&\quad (W(t_i) + (t - t_i)W'(t_i)) \left( x(t_i) + (t - t_i)A(t_i)x(t_i) + B(t_i) \int_{t_i}^t u(\tau) d\tau \right) dt + O(h^3) \\
&= hx(t_i)^\top \left( W(t_i) + hW(t_i)A(t_i) + \frac{h}{2}W'(t_i) \right) x(t_i) + 2x(t_i)^\top W(t_i)B(t_i) \int_{t_i}^{t_{i+1}} \int_{t_i}^t u(s) ds dt + O(h^3).
\end{aligned}$$

Note that an easy calculation implies that

$$\int_{t_i}^{t_{i+1}} \int_{t_i}^t u(\tau) d\tau dt = h \int_{t_i}^{t_{i+1}} u(t) dt - \int_{t_i}^{t_{i+1}} (t - t_i)u(t) dt. \tag{3.4}$$

Now we consider the mixed term in the integral in (1.1). A calculation of the same fashion as the previous one yields

$$\begin{aligned}
&\int_{t_i}^{t_{i+1}} x(t)^\top S(t)u(t) dt \\
&= \int_{t_i}^{t_{i+1}} \left( x(t_i) + (t - t_i)A(t_i)x(t_i) + B(t_i) \int_{t_i}^t u(s) ds \right)^\top (S(t_i) + S'(t_i)(t - t_i)) u(t) dt + O(h^3) \\
&= x(t_i)^\top \left( S(t_i) \int_{t_i}^{t_{i+1}} u(t) dt + S'(t_i) \int_{t_i}^{t_{i+1}} (t - t_i)u(t) dt \right) + (A(t_i)x(t_i))^\top S(t_i) \int_{t_i}^{t_{i+1}} (t - t_i)u(t) dt \\
&\quad + \int_{t_i}^{t_{i+1}} \left( B(t_i) \int_{t_i}^t u(s) ds \right)^\top S(t_i)u(t) dt + O(h^3).
\end{aligned} \tag{3.5}$$

Let us focus on the last term:

$$\int_{t_i}^{t_{i+1}} \left( B(t_i) \int_{t_i}^t u(s) ds \right)^\top S(t_i)u(t) dt = \int_{t_i}^{t_{i+1}} \left( B(t_i) \int_{t_i}^t u(s) ds \right)^\top S(t_i) d \int_{t_i}^t u(s) ds.$$

Integrating by parts we obtain the relation

$$\int_{t_i}^{t_{i+1}} u(t)^\top \left( B(t_i)^\top S(t_i) + S(t_i)^\top B(t_i) \right) \int_{t_i}^t u(s) \, ds \, dt = \int_{t_i}^{t_{i+1}} u(s)^\top \, ds \, B(t_i)^\top S(t_i) \int_{t_i}^{t_{i+1}} u(s) \, ds.$$

Following [28], in order to obtain a second-order expansion expressed in term of the first and second-order momentum of  $u$ , we assume the following.

**Assumption (I)** The matrix  $B^\top(t)S(t)$  is symmetric for all  $t \in [0, T]$ .

Indeed, using Assumption (I) we obtain from the last exposed equality the expression

$$\int_{t_i}^{t_{i+1}} \left( B(t_i) \int_{t_i}^t u(s) \, ds \right)^\top S(t_i) u(t) \, dt = \frac{1}{2} \int_{t_i}^{t_{i+1}} u(s)^\top \, ds \, B(t_i)^\top S(t_i) \int_{t_i}^{t_{i+1}} u(s) \, ds,$$

which can be substituted in (3.5).

Notice that Assumption (I) is always fulfilled if  $m = 1$ . The above obtained second order approximations will be used in the next subsection to define an appropriate discrete-time approximation of problem (1.1)–(1.3).

### 3.2 The numerical scheme

First of all, observe that the representation (3.1) of  $x(t)$  for  $t \in [t_i, t_{i+1}]$  depends on the control  $u$  only through the integrals  $\int_{t_i}^{t_{i+1}} u(t) \, dt$  and  $\int_{t_i}^{t_{i+1}} (t - t_i)u(t) \, dt$ . The same applies to the approximations of the integral terms of the objective functional obtained in the last subsection. By changing the variable  $t = t_i + hs$ , this pair of integrals can be represented in the form  $hz_1$  and  $h^2z_2$ , respectively, where

$$z_1 = \int_0^1 \varphi(s) \, ds, \quad z_2 = \int_0^1 s\varphi(s) \, ds,$$

and  $\varphi(s) = u(t_i + hs)$  is a measurable function with values in  $[-1, 1]$ . By varying  $u$ , hence  $\varphi$ , in the set of all admissible controls on  $[0, T]$ , the couple  $(z_1, z_2) \in \mathbb{R}^{2m}$  generates a strictly-convex and compact set  $Z^m \subset \mathbb{R}^{2m}$ . Note that  $Z^m$  can be expressed as the Cartesian product  $\prod_1^m Z$ , where  $Z$  is the Aumann integral

$$Z := \int_0^1 \begin{pmatrix} 1 \\ s \end{pmatrix} [-1, 1] \, ds. \quad (3.6)$$

As pointed out in [23], it is a matter of standard calculation to represent the set  $Z$  in the more convenient way as

$$Z = \{(\alpha, \beta) : \alpha \in [-1, 1], \beta \in [\phi_1(\alpha), \phi_2(\alpha)]\},$$

where  $\phi_1(\alpha) := \frac{1}{4}(-1 + 2\alpha + \alpha^2)$  and  $\phi_2(\alpha) := \frac{1}{4}(1 + 2\alpha - \alpha^2)$ .

Following the hint provided by the representation (3.1), we introduce the notations

$$\begin{aligned} A_i &:= A(t_i) + \frac{h}{2}(A(t_i)^2 + A'(t_i)), \\ B_i &:= B(t_i) + hA(t_i)B(t_i), \quad C_i := -A(t_i)B(t_i) + B'(t_i), \end{aligned}$$

and replace the differential equation (1.2) with the discrete-time controlled dynamics

$$x_{i+1} = x_i + h(A_i x_i + B_i u_i + hC_i v_i), \quad i = 0, \dots, N-1, \quad x_0 \text{ given}, \quad (3.7)$$

$$(u_i, v_i) \in Z^m \quad i = 0, \dots, N-1. \quad (3.8)$$



Taking into account the approximations of the objective functional in the previous subsection, we introduce its discrete-time counterpart: for  $x = (x_0, \dots, x_N)$ ,  $u = (u_0, \dots, u_{N-1})$ ,  $v = (v_0, \dots, v_{N-1})$ ,

$$\begin{aligned} J^h(x, u, v) &:= \frac{1}{2}x_N^\top(Qx_N + q) + \frac{h}{2} \sum_{i=0}^{N-1} \left( x_i^\top W(t_i) (x_i + hA(t_i)x_i) + \frac{h}{2}x_i^\top W'(t_i)x_i \right) \\ &+ h \sum_{i=0}^{N-1} \left( hx_i^\top W(t_i)B(t_i)(u_i - v_i) + x_i^\top (S(t_i)u_i + hS'(t_i)v_i) + h(A(t_i)x_i)^\top S(t_i)v_i \right) \\ &+ \frac{h^2}{2} \sum_{i=0}^{N-1} \langle B(t_i)^\top S(t_i)u_i, u_i \rangle. \end{aligned} \quad (3.9)$$

We denote by  $(P^h)$  the discrete problem of minimizing (3.9) subject to (3.7)-(3.8). The Karush-Kuhn-Tucker theorem gives the following necessary conditions for optimality of  $(x_0, \dots, x_N)$ ,  $(w_0, \dots, w_{N-1})$ , with  $w_i := (u_i, v_i) \in Z^m$ : there is an (adjoint) sequence  $(p_0, \dots, p_N)$  such that

$$0 = -x_{i+1} + x_i + h(A_i x_i + B_i u_i + hC_i v_i), \quad (3.10)$$

$$\begin{aligned} 0 &= -p_i + \left( I + hA_i^\top \right) p_{i+1} + h \left( S(t_i)u_i + hS'(t_i)v_i + hA(t_i)^\top S(t_i)v_i \right) \\ &+ h \left( W(t_i) + \frac{h}{2}W(t_i)A(t_i) + \frac{h}{2}A(t_i)^\top W(t_i) + \frac{h}{2}W'(t_i) \right) x_i + h^2 W(t_i)B(t_i)(u_i - v_i), \end{aligned} \quad (3.11)$$

$$0 \in N_{Z^m}(w) + \begin{pmatrix} B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i)x_i + hB(t_i)^\top S(t_i)u_i \\ h(C_i^\top p_{i+1} - B(t_i)^\top W(t_i)x_i + (S(t_i)^\top A(t_i) + S'(t_i)^\top) x_i) \end{pmatrix}, \quad (3.12)$$

$$0 = -p_N + Q^\top x_N + q. \quad (3.13)$$

In order to obtain (3.12) we use again Assumption (I).

### 3.3 Construction of continuous-time controls and order of convergence

Let  $\{(x_i, u_i, v_i, p_i)\}$  be any solution of system (3.10)-(3.13). Based on the sequence  $\{(u_i, v_i)\}_{i=0}^{N-1}$  we shall construct a continuous-time admissible control  $u$  such that

$$\int_{t_i}^{t_{i+1}} u(s)ds = hu_i, \quad \int_{t_i}^{t_{i+1}} (s - t_i)u(s)ds = h^2v_i, \quad i = 0, \dots, N-1. \quad (3.14)$$

The construction is by idea similar to that in [23] with the essential difference that now  $u$  takes values only in the set  $\{-1, 1\}$  and the construction is simpler.

For  $(\alpha, \beta) \in Z$ , with  $\alpha \neq -1$  (that is,  $\alpha \in (-1, 1]$ ) and  $\beta \in [\phi_1(\alpha), \phi_2(\alpha)]$  define

$$\tau(\alpha, \beta) := \frac{1 + 2\beta}{2(1 + \alpha)} - \frac{1 + \alpha}{4}, \quad \theta(\alpha, \beta) := \frac{1 + 2\beta}{2(1 + \alpha)} + \frac{1 + \alpha}{4}.$$

For  $\alpha = -1$  we set  $\tau = \theta = 0$ . (Given that  $\beta \in [\phi_1(\alpha), \phi_2(\alpha)]$ , this is, in fact, an extension by continuity.) Clearly,  $\tau \leq \theta$ , while  $\tau \geq 0$  is implied by  $\beta \geq \phi_1(\alpha)$  and  $\theta \leq 1$  is implied by  $\beta \leq \phi_2(\alpha)$ . Then define the admissible control  $u$  component-wise as follows: for  $j = 1, \dots, m$  and  $i = 0, \dots, N-1$  set  $\tau_i^j = \tau(u_i^j, v_i^j)$ ,  $\theta_i^j = \theta(u_i^j, v_i^j)$ , and

$$u^j(t) := \begin{cases} -1 & \text{if } t \in [t_i, t_i + h\tau_i^j), \\ 1 & \text{if } t \in [t_i + h\tau_i^j, t_i + h\theta_i^j], \\ -1 & \text{if } t \in (t_i + h\theta_i^j, t_{i+1}]. \end{cases} \quad (3.15)$$

The functions  $\tau(\cdot)$  and  $\theta(\cdot)$  are defined in such a way that the relations (3.14) are fulfilled. To show this, it is enough to substitute the above defined  $u(\cdot)$  in (3.14) and calculate the integrals. We skip this trivial but cumbersome calculation.

We mention that in our framework the pairs  $(u_i^j, v_i^j)$ ,  $i = 0, \dots, N-1$ ,  $j = 1, \dots, M$ , typically belong to the boundary of the set  $Z$ . In such a case every component of the control  $u$  defined in (3.15) has at most one switching point per mesh interval  $[t_i, t_{i+1}]$  and we can distinguish the following possibilities:

- (i) if  $u_i^j = -1$  or  $u_i^j = 1$ , then  $u^j(t) = -1$ , respectively  $u^j(t) = 1$ , in  $[t_i, t_{i+1}]$ ;
- (ii-a) if  $u_i^j \in (-1, 1)$  and  $v_i^j = \phi_1(u_i^j)$  then  $\tau_i^j = 0$ ,  $\theta_i^j = (1 + u_i^j)/2$ , thus

$$u^j(t) := \begin{cases} 1 & \text{if } t \in [t_i, t_i + h\theta_i^j], \\ -1 & \text{if } t \in (t_i + h\theta_i^j, t_{i+1}]; \end{cases} \quad (3.16)$$

- (ii-b) if  $u_i^j \in (-1, 1)$  and  $v_i^j = \phi_2(u_i^j)$  then  $\tau_i^j = (1 - u_i^j)/2$ ,  $\theta_i^j = 1$ , thus

$$u^j(t) := \begin{cases} -1 & \text{if } t \in [t_i, t_i + h\tau_i^j], \\ 1 & \text{if } t \in (t_i + h\tau_i^j, t_{i+1}]. \end{cases}$$

A third possibility is that  $(u_i^j, v_i^j)$  happens to belong to the interior of  $Z$ :

- (iii) if  $u_i^j \in (-1, 1)$  and  $v_i^j \in (\phi_1(u_i^j), \phi_2(u_i^j))$  then formula (3.15) has to be used to define  $u(\cdot)$ .

In fact, formula (3.15) gives a unified description of all the above cases, where some of the three subintervals intervals in (3.15) degenerate in the (typical) cases (i) and (ii).

**Theorem 3.1.** *Let Assumption (A1), (A2) and (I) be fulfilled. Let  $(\hat{x}, \hat{u})$  be a solution of problem (1.1)–(1.3) for which Assumption (B) is fulfilled with some  $\kappa \geq 1$ . Let  $\hat{p}$  be the corresponding adjoint function (so that  $(\hat{x}, \hat{p}, \hat{u})$  satisfies the Pontryagin system (2.1)–(2.4)). Then for every natural number  $N$  system (3.10)–(3.13) has a solution  $\{(x_i, u_i, v_i, p_i)\}$ . Moreover, for the continuous embedding of  $(u_i, v_i)$  defined in (3.15), it holds that*

$$\max_{k=0, \dots, N} (|x_k - \hat{x}(t_k)| + |p_k - \hat{p}(t_k)|) + d^\#(u, \hat{u}) \leq c h^{2/\kappa}. \quad (3.17)$$

We mention that, for time-invariant problems without singular arcs, Assumption (B) is typically fulfilled with a number  $\kappa \in \{1, 2, 3, \dots\}$ , corresponding to the multiplicity of the zeros of the switching function. As argued in [23], the case  $\kappa = 1$  is in a certain sense “generic” and the error estimate (3.17) is of second order in this case. Also in the case  $\kappa > 1$  the order of accuracy is doubled in comparison with that proved in [26] for the Euler scheme. Utilization of higher order Runge-Kutta schemes on a fixed mesh could not improve the accuracy of the Euler scheme due to the discontinuity of the optimal control.

A solution  $\{(x_i, u_i, v_i, p_i)\}$  of system (3.10)–(3.13) can be obtained by any method for solving the discrete problem (3.7)–(3.9). The adjoint variables  $p_i$  do not need to be directly involved. For example, we use for numerical computations a version of the steepest descent method, where the adjoint equation is only indirectly involved for calculation of the derivative of the objective function (3.9) with respect to the variables  $(u_i, v_i)$ . In any case, the adjoint functions  $\hat{p}(\cdot)$  and  $\{p_i\}$  are well defined and the error estimate (3.17) is valid.

As we argue in Section 5, the solution  $\{(x_i, u_i, v_i, p_i)\}$  can be inexact, which leads to a modification of the error estimate as stated there.

## 4 Proof of Theorem 3.1

### 1. Preliminaries.

Let  $\{(x_i, u_i, v_i, p_i)\}$  be a solution of the discrete system (3.10)-(3.13) and let  $u(\cdot)$  be the continuous embedding of  $\{(u_i, v_i)\}$  defined in (3.15).

We embed the sequences  $\{x_i\}$  and  $\{p_i\}$  into the spaces  $W_{x_0}^{1,1}$  and  $W^{1,1}$  using the hint provided by the expansions developed in Section 3.1. Namely, for  $t \in [t_i, t_{i+1})$ , we define

$$\begin{aligned} x(t) := & \left( I + (t - t_i)A(t_i) + \frac{(t - t_i)^2}{2}(A(t_i)^2 + A'(t_i)) \right) x_i \\ & + (B(t_i) + (t - t_i)A(t_i)B(t_i)) \int_{t_i}^t u(s) ds + C_i \int_{t_i}^t (s - t_i)u(s) ds. \end{aligned} \quad (4.1)$$

and

$$\begin{aligned} p(t) := & \left[ I + (t_{i+1} - t)A(t_i)^\top + \frac{(t_{i+1} - t)^2}{2}A^2(t_i)^\top + \frac{h^2 - (t - t_i)^2}{2}A'(t_i)^\top \right] p_{i+1} \\ & + (t_{i+1} - t)W(t_i)x_i + \frac{h^2 - (t - t_i)^2}{2}(W'(t_i) + W(t_i)A(t_i))x_i \\ & + W(t_i)B(t_i) \int_t^{t_{i+1}} \int_{t_i}^s u(\tau) d\tau ds + \frac{(t_{i+1} - t)^2}{2}A(t_i)^\top W(t_i)x_i \\ & + A(t_i)^\top S(t_i) \int_t^{t_{i+1}} \int_s^{t_{i+1}} u(\zeta) d\zeta ds + S(t_i) \int_t^{t_{i+1}} u(s) ds + S'(t_i) \int_t^{t_{i+1}} (s - t_i)u(s) ds. \end{aligned} \quad (4.2)$$

We show below that  $p \in W^{1,1}$ . From (3.4) and (3.14) it follows that

$$\begin{aligned} \lim_{t \rightarrow t_i^+} p(t) = & \left[ I + hA(t_i)^\top + \frac{h^2}{2}A^2(t_i)^\top + \frac{h^2}{2}A'(t_i)^\top \right] p_{i+1} + hW(t_i)x_i + \frac{h^2}{2}(W'(t_i)x_i + W(t_i)A(t_i)x_i) \\ & + h^2W(t_i)B(t_i)(u_i - v_i) + \frac{h^2}{2}A(t_i)^\top W(t_i)x_i + hS(t_i)u_i + h^2(A(t_i)^\top S(t_i) + S'(t_i))v_i. \end{aligned} \quad (4.3)$$

Since

$$\left[ I + hA(t_i)^\top + \frac{h^2}{2}A^2(t_i)^\top + \frac{h^2}{2}A'(t_i)^\top \right] p_{i+1} = \left( I + hA_i^\top \right) p_{i+1},$$

the right-hand side of (4.3) is equal to the expression of  $p_i$  given by (3.11). Thus,  $p$  is continuous at  $t_i$ , and hence  $p \in W^{1,1}$ . The proof that  $x \in W_{x_0}^{1,1}$  is analogous and can be found in [23, Section 5].

By Theorem 2.2, for every  $b > 0$  there exist a number  $c$  such that if  $\|y\| \leq b$  then

$$\|x - \hat{x}\|_{1,1} + \|p - \hat{p}\|_{1,1} + \|u - \hat{u}\|_1 \leq c\|y\|^{1/\kappa}, \quad (4.4)$$

where  $y = (\xi, \pi, \nu, \rho)$  is the residual that  $(x, p, u)$  gives in (2.1)-(2.4), that is,  $y \in F(x, p, u)$ . Thus we have to estimate the norm of this residual.

The estimate  $\|\xi\|_1 \leq O(h^2)$  of the first residual is obtained in [23, Section 4], where the primal differential equation is the same as in the present paper. We shall analyze below the residual in the remaining equations (2.2)-(2.4).

**2. Residual in (2.2) and (2.4).**

First, we differentiate the expression in (4.2) for  $t \in [t_i, t_{i+1}]$ :

$$\begin{aligned}
-\dot{p}(t) &= \left[ A(t_i)^\top + (t_{i+1} - t)A^2(t_i)^\top + (t - t_i)A'(t_i)^\top \right] p_{i+1} + W(t_i)x_i \\
&\quad + (t - t_i) \left( W'(t_i) + W(t_i)A(t_i) \right) x_i + W(t_i)B(t_i) \int_{t_i}^t u(s) ds + (t_{i+1} - t)A(t_i)^\top W(t_i)x_i \\
&\quad + A(t_i)^\top S(t_i) \int_t^{t_{i+1}} u(\zeta) d\zeta + S(t_i)u(t) + S'(t_i)(t - t_i)u(t) \\
&= A(t)^\top \left( I + (t_{i+1} - t)A(t_i)^\top \right) p_{i+1} + W(t)x_i + (t - t_i)W(t)A(t_i)x_i + W(t)B(t_i) \int_{t_i}^t u(s) ds \\
&\quad + (t_{i+1} - t)A(t)^\top W(t_i)x_i + A(t)^\top S(t_i) \int_t^{t_{i+1}} u(\zeta) d\zeta + S(t)u(t) + O(t; h^2).
\end{aligned}$$

Here and below  $O(t; h^2) \leq Ch^2$  for some constant  $C > 0$  which is independent of  $t \in [0, T]$ . Using (4.1) and then (4.2) we obtain that

$$\begin{aligned}
-\dot{p}(t) &= A(t)^\top \left( I + (t_{i+1} - t)A(t_i)^\top \right) p_{i+1} + (t_{i+1} - t)A(t_i)^\top W(t_i)x_i + A(t)^\top S(t_i) \int_t^{t_{i+1}} u(\zeta) d\zeta \\
&\quad + W(t)x(t) + S(t)u(t) + O(t; h^2) \\
&= A(t)^\top p(t) + W(t)x(t) + S(t)u(t) + O(t; h^2).
\end{aligned}$$

Hence, we deduce that  $\|\pi\|_\infty \leq O(h^2)$ . Notice that the equation (3.12) gives that the residual in (2.4) is zero, that is,  $\nu = 0$ .

**3. Residual in (2.3).**

First of all, we derive a second order expansion of the term  $B^\top p + S^\top x$  appearing in (2.3). By (4.1), (4.2) and the Taylor expansion for  $B$  and  $r$  we have

$$\begin{aligned}
B^\top(t)p(t) + S(t)^\top x(t) &= \left( B(t_i) + (t - t_i)B'(t_i) \right)^\top \left( p_{i+1} + (t_{i+1} - t)A(t_i)^\top p_{i+1} \right. \\
&\quad \left. + (t_{i+1} - t)W(t_i)x_i + S(t_i) \int_t^{t_{i+1}} u(s) ds \right) \\
&\quad + \left( S(t_i) + S'(t_i)(t - t_i) \right)^\top \left( (I + (t - t_i)A(t_i)) x_i + B(t_i) \int_{t_i}^t u(s) ds \right) + O(t; h^2).
\end{aligned}$$

Then, by using the definition of  $B_i$  and  $C_i$  we obtain that

$$\begin{aligned}
B^\top(t)p(t) + S(t)^\top x(t) &= \\
&= (B_i + (t - t_i)C_i)^\top p_{i+1} + B(t_i)^\top \left( (t_{i+1} - t)W(t_i)x_i + S(t_i) \int_t^{t_{i+1}} u(s) ds \right) + S(t_i)^\top B(t_i) \int_{t_i}^t u(s) ds \\
&\quad + S(t_i)^\top (I + (t - t_i)A(t_i)) x_i + S'(t_i)^\top (t - t_i)x_i + O(t; h^2).
\end{aligned}$$

Since Assumption (I) means  $B^\top S = S^\top B$ ,

$$\begin{aligned}
B^\top(t)p(t) + S(t)^\top x(t) &= (B_i + (t - t_i)C_i)^\top p_{i+1} + B(t_i)^\top \left( (t_{i+1} - t)W(t_i)x_i + S(t_i) \int_t^{t_{i+1}} u(s) ds \right) \\
&\quad + S(t_i)^\top (I + (t - t_i)A(t_i)) x_i + S'(t_i)^\top (t - t_i)x_i + O(t; h^2). \tag{4.5}
\end{aligned}$$

Our goal is now to estimate the norm  $\|\cdot\|_\infty$  of the residual in (2.3). Since  $N_U(u) = \prod_{j=1, \dots, m} N_{[-1,1]}(u^j)$ , we analyze a single component  $j$  of (2.3). Moreover, (2.3) is a point-wise relation, therefore we consider it on an arbitrarily fixed interval  $[t_i, t_{i+1}]$ . We also mention that the set  $Z$  is the area surrounded by the two parabolas  $\beta = \phi_1(\alpha)$  and  $\beta = \phi_2(\alpha)$ , where  $\phi_1(\alpha) \leq \phi_2(\alpha)$ . Thus the normal cone to  $Z$  is easy to calculate and the following expression is provided in [23, Section 4]:

$$N_Z(\alpha, \beta) = \begin{cases} \emptyset & \text{if } (\alpha, \beta) \notin Z \\ \{\alpha(\lambda, \mu - \lambda)^\top : \mu \geq 0, \lambda \geq 0\} & \text{if } \alpha \in \{-1, 1\} \\ \{\mu(\zeta + \alpha, -2\zeta)^\top : \mu \geq 0\} & \text{if } \alpha \in (-1, 1) \wedge \beta \in \{\phi_1(\alpha), \phi_2(\alpha)\} \\ \{0\} & \text{if } \alpha \in (-1, 1) \wedge \beta \in (\phi_1(\alpha), \phi_2(\alpha)), \end{cases} \quad (4.6)$$

where  $\zeta = \text{sgn}(\alpha - 2\beta)$ .

We consider separately each of the cases (i), (ii) and (iii) for construction of continuous time control that appear in Section 3.3.

*Case (i):*  $u_i^j \in \{-1, 1\}$ . To be specific, let us assume that  $u_i^j = -1$ , hence  $v_i^j = \phi_1(-1) = \phi_2(-1) = -1/2$ . (The case  $u_i^j = 1$  is similar). The normal cone to  $Z$  at the point  $(-1, -1/2)^\top$  is (see the second line in (4.6))

$$N_Z((-1, -1/2)) = \left\{ \mu(0, -1)^\top + \lambda(-1, 1)^\top : \mu \geq 0, \lambda \geq 0 \right\}.$$

Then due to (3.12) for every  $j = 1, \dots, m$  there exist  $\mu \geq 0$  and  $\lambda \geq 0$  such that

$$\left( B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i)x_i + hB^\top(t_i)S(t_i)u_i \right)^j = \lambda, \quad (4.7)$$

$$h \left( C_i^\top p_{i+1} - B(t_i)^\top W(t_i)x_i + \left( S(t_i)^\top A(t_i) + S'(t_i)^\top \right) x_i \right)^j = (\mu - \lambda). \quad (4.8)$$

Observe that, for  $t \in [t_i, t_{i+1}]$ ,

$$\lambda + (\mu - \lambda) \frac{(t - t_i)}{h} = \mu \frac{(t - t_i)}{h} + \lambda \left( 1 - \frac{(t - t_i)}{h} \right), \quad (4.9)$$

thus the quantity above is non-negative for all  $t \in [t_i, t_i + h)$ . Thus, adding up (4.7) and (4.8), the latter multiplied by  $(t - t_i)/h$ , we obtain that

$$\begin{aligned} & \left[ (B_i + (t - t_i)C_i)^\top p_{i+1} + B(t_i)^\top ((t_{i+1} - t)W(t_i)x_i + S(t_i)u_i) \right. \\ & \left. + S(t_i)^\top (I + (t - t_i)A(t_i))x_i + S'(t_i)(t - t_i)x_i \right]^j \geq 0. \end{aligned}$$

By (3.14) the quantity above is identical to the  $j$ -th component of the right-hand side of (4.5), modulo  $O(t; h^2)$ . By the fact that  $u^j(t) = -1$  in case (i), we thus obtain

$$(B^\top p(t) + S(t)^\top x(t))^j + O(h^2) \in -N_{[-1,1]}(u^j(t)). \quad (4.10)$$

*Case (ii):*  $u_i^j \in (-1, 1)$ ,  $v_i^j \in \{\phi_1(u_i^j), \phi_2(u_i^j)\}$ . We consider the case  $v_i^j = \phi_1(u_i^j)$ ; the case  $v_i^j = \phi_2(u_i^j)$  can be treated similarly. The continuous-time control  $u^j(\cdot)$  is defined by (3.16), where the jump point  $\theta_i^j = (1 - u_i^j)/2$ . The normal cone to  $Z$  at  $(u_i^j, \phi(u_i^j))$  is (see the third line in (4.6))

$$N_Z\left((u_i^j, \phi(u_i^j))\right) = \left\{ \mu \left( 1 + u_i^j, -2 \right)^\top : \mu \geq 0 \right\}.$$

By (3.12), there exists  $\mu \geq 0$  such that

$$\left( B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i)x_i + hB(t_i)^\top S(t_i)u_i \right)^j = -\mu(1 + u_i^j), \quad (4.11)$$

$$h \left( C_i^\top p_{i+1} - B(t_i)^\top W(t_i)x_i + \left( S(t_i)^\top A(t_i) + S'(t_i) \right) x_i \right)^j = 2\mu. \quad (4.12)$$

Observe that, from the definition of  $\theta_i^j$ , it follows that the quantity

$$-\mu(1 + u_i^j) + 2\mu \frac{(t - t_i)}{h} \quad (4.13)$$

is non-positive whenever  $t \in [t_i, t_i + h\theta_i^j)$ , and non-negative whenever  $t \in [t_i + h\theta_i^j, t_{i+1}]$ . Thus, adding up (4.11) and (4.12), the latter multiplied by  $(t - t_i)/h$ , and using the definition of  $u^j$  we obtain that

$$\begin{aligned} & \left( B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i)x_i + hB(t_i)^\top S(t_i)u_i \right)^j \\ & + (t - t_i) \left( C_i^\top p_{i+1} - B(t_i)^\top W(t_i)x_i + \left( S(t_i)^\top A(t_i) + S'(t_i) \right) x_i \right)^j \in -N_{[-1,1]}(u^j(t)). \end{aligned}$$

By (3.14) and (4.5), the left-hand side term of the relation above is equal to  $(B^\top p(t) + S(t)^\top x(t) + r(t))^j + O(h^2)$ . This proves (4.10) in the case (ii).

*Case (iii)* By (3.12) and the fact that  $(u_k^j, v_k^j) \in \text{Int } Z$ , we have

$$\begin{aligned} & \left( B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i)x_i + hB(t_i)^\top S(t_i)u_i \right)^j = 0, \\ & h \left( C_i^\top p_{i+1} - B(t_i)^\top W(t_i)x_i + \left( S(t_i)^\top A(t_i) + S'(t_i) \right) x_i \right)^j = 0. \end{aligned}$$

Then,

$$\begin{aligned} 0 = & \left( B_i^\top p_{i+1} + S(t_i)^\top x_i + hB(t_i)^\top W(t_i)x_i + hB(t_i)^\top S(t_i)u_i \right. \\ & \left. + (t - t_i) \left( C_i^\top p_{i+1} - B(t_i)^\top W(t_i)x_i + \left( S(t_i)^\top A(t_i) + S'(t_i) \right) x_i \right) \right)^j. \end{aligned}$$

This and (4.5) yield (4.10) in the case (iii). We can finally conclude that  $\|\rho\|_\infty = O(h^2)$ .

Summarizing, we have obtained that  $\|y\| \leq c_1 h^2$ , where  $c_1$  is independent of  $N$ . Since  $c_1 h^2 \leq c_1 T^2 =: b$ , Theorem 2.2 implies existence of  $c$  such that for every natural  $N$

$$\|x - \hat{x}\|_{1,1} + \|p - \hat{p}\|_{1,\infty} + \|u - \hat{u}\|_1 \leq c h^{2/\kappa}.$$

We know that  $x(t_i) = x_i$  and  $p(t_i) = p_i$ , hence

$$\max_{i=0,\dots,N} (|x_i - \hat{x}(t_i)| + |p_i - \hat{p}(t_i)|) + \|u - \hat{u}\|_1 \leq c_2 h^{2/\kappa}.$$

Now we focus on the last term in the left-hand side. Since  $\hat{u}$  and  $u$  take only values  $\pm 1$ , as already pointed out for instance in [25, Section 4], we have that  $d^\#(u, \hat{u}) \leq c_3 \|u - \hat{u}\|_1$  for some  $c_3 \geq 0$ , and this concludes the proof.

## 5 Error estimate in case of inexact solution of the discrete problem

The estimation (3.17) in Theorem 3.1 is valid on the assumption that the discrete-time problem (3.7)–(3.9) is exactly solved. In the present section we incorporate in the error estimation possible inaccuracy in solving the discrete-time problem. The basic argument for that is identical with the one for Mayer’s problems, presented in [23, Section 5], therefore we only sketch it.

We assume that as a result of a numerical procedure for solving the mathematical programming problem (3.7)–(3.9) we have obtained an approximate solution  $(\{\tilde{x}_i\}, \{\tilde{p}_i\}, \{\tilde{w}_i\})$  of the first order optimality (Karush-Kuhn-Tucker) system (3.10)–(3.13). This means that the relations (3.10)–(3.13) are satisfied by the sequences  $(\{\tilde{x}_i\}, \{\tilde{p}_i\}, \{\tilde{w}_i\})$  with some residual  $(\xi, \pi, \rho, \nu) = (\{\xi_i\}_0^{N-1}, \{\pi_i\}_0^{N-1}, \{\rho_i\}_0^{N-1}, \nu)$ . We measure the size of the residual by the number

$$\varepsilon := \|\xi\|_{l_1} + \|\pi\|_{l_\infty} + \|\rho\|_{l_\infty} + |\nu| = h \sum_{i=0}^{N-1} |\xi_i| + \max_{i=0, \dots, N-1} |\pi_i| + \max_{i=0, \dots, N-1} |\rho_i| + |\nu|.$$

Using the approximate solution  $\{\tilde{w}_i\}$ , one can define an approximation,  $\tilde{u}(\cdot)$ , of the optimal control  $\hat{u}$  in the same way as described in Section 3.3. Then the estimation (3.17) in Theorem 3.1 takes the form

$$\max_{i=0, \dots, N} (|\tilde{x}_i - \hat{x}(t_i)| + |\tilde{p}_i - \hat{p}(t_i)|) + d^\#(\tilde{u} - \hat{u}) \leq c(\varepsilon + h^2)^{1/\kappa}. \quad (5.1)$$

The proof of this statement is not straightforward, but the argument is identical with that in [23, Section 5], therefore we do not repeat it here.

Clearly, in order to make the error arising in the proposed discretization scheme and the error in solving the discretized problem consistent, one has to solve the latter with accuracy  $\varepsilon$  proportional to  $h^2$ .

## 6 A numerical experiment

*Example 6.1.* Let us consider the following optimal control problem on the plane:

$$\min \left\{ -by(1) + \int_0^1 \frac{1}{2}(x(t))^2 dt \right\}$$

subject to

$$\begin{aligned} \dot{x} &= y, & x(0) &= a, \\ \dot{y} &= u, & y(0) &= 0, \end{aligned}$$

with control constraint  $u \in [-1, 1]$  and for  $a > 1/2$ ,  $b > 0$ .

Here (for appropriate values of  $a$  and  $b$ ) there is a unique optimal solution with a switch from  $u = -1$  to  $u = 1$  at time  $\tau$  which is a solution of the equation

$$-5\tau^4 + 24\tau^3 - (12a + 36)\tau^2 + (24a + 20)\tau + 24b - 12a - 3 = 0.$$

Moreover,  $\tau$  is a simple zero of the switching function, thus  $\kappa = 1$ . Taking, for example,  $a = 1$  and  $b = 0.1$ , the equation above becomes

$$-5\tau^4 + 24\tau^3 - 48\tau^2 + 44\tau - 12.6 = 0$$

and the single real solution of this equation in  $[0, 1]$  is  $\tau = 0.492487520$  with all digits being correct. We solved system (3.10)-(3.13) for various values of the discretization step  $h = T/N$  ( $N$  is a natural number) by using a version of the gradient projection method. The computation of the approximate solutions  $(\{\tilde{x}_i\}, \{\tilde{p}_i\}, \{\tilde{w}_i\})$  is done with accuracy (measured by the residual  $\varepsilon$  – see Section 5) higher than  $h^2$ , so that the theoretical error estimate (5.1) is  $ch^2$ . As before,  $(\hat{x}, \hat{p}, \hat{u})$  is the exact solution of the considered problem, and  $\tilde{u}$  is the continuous time control constructed as described in Section 3.3 for the computed  $\{\tilde{w}_i\} = \{(\tilde{u}_i, \tilde{v}_i)\}$  (note that  $\tilde{u}$  depends on size of the mesh  $N$ , therefore further we use the notation  $\tilde{u}^N$  instead of  $\tilde{u}$ ).

In Table 1 we report numerical results, focusing on the most critical error  $e_N = d^\#(\hat{u} - \tilde{u}^N)$ . We also calculate the value  $e_N/h^2$ , which according to (5.1) should be bounded. This is confirmed by the results.

Table 1: Here  $e_N = d^\#(\hat{u} - \tilde{u}^N)$  is the error of the numerically obtained control  $\tilde{u}^N$  for various values of  $N$ . The last line gives the values  $e_N/h^2$ .

$N$	10	20	30	40	60
$e_N$	$1.50 \cdot 10^{-3}$	$3.64 \cdot 10^{-4}$	$1.54 \cdot 10^{-4}$	$9.35 \cdot 10^{-5}$	$3.97 \cdot 10^{-5}$
$e_N/h^2$	0.150	0.146	0.139	0.150	0.143

## 7 Concluding remarks

In this paper we extend the analysis of the discretization scheme introduced in [23] for Mayer’s problems to embrace convex linear-quadratic problems, affine with respect to the control. The optimal controls are assumed to be purely bang-bang, with an additional assumption involving the associated “switching function”. Our discretization approach opts for solving a discrete-time optimization problem involving an additional control variable. This yields an order of convergence which doubles that of the Euler’s method. The price for that is that the discrete problem involves quadratic control constraints (instead of the box-type constraints in the original problem).

It is worth noting that the components of the optimal controls could be, in general, concatenations of bang-bang and singular arcs. This challenging case will be a subject of further investigation. It requires, among other things, a deeper analysis of the metric sub-regularity of the system of first order optimality conditions under perturbations (a step in this direction is made in [13]). Another challenging issue is to avoid Assumption (I). Assumption of this kind is present also in [29], as well as in [16], in a nonlinear context, where it requires the Lie brackets of the involved controlled vector fields to vanish. An idea in this direction is presented in [19], but it does not seem to be efficient for numerical implementation.

## References

- [1] W. Alt, R. Baier, M. Gerdt, F. Lempio. Error bounds for Euler approximations of linear-quadratic control problems with bang-bang solutions. *Numer. Algebra, Control and Optim.*, **2**(3), 547–570 (2012).
- [2] W. Alt, R. Baier, F. Lempio, M. Gerdt. Approximations of linear control problems with bang-bang solutions. *Optimization*, **62**(1), 9–32 (2013).



- [3] M.S. Aronna, J.F. Bonnans, A.V. Dmitruk and P.A. Lotito, Quadratic order conditions for bang-singular extremals, *Numer. Algebra, Control Optim.*, **2(3)**, 511–546 (2012).
- [4] J.F. Bonnans, A. Festa. Error estimates for the Euler discretization of an optimal control problem with first-order state constraints. *SIAM J. Numerical Analysis*, **55** (2), 445–471 (2017).
- [5] W. Alt, C. Schneider, M. Seydenschwanz. *Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions*. *Appl. Math. Comput.*, Vol. 287/288, 104–124 (2016).
- [6] A.L. Dontchev. An a priori estimate for discrete approximations in nonlinear optimal control, *SIAM J. Control Optim.*, **34**, 1315–1328 (1996).
- [7] A.L. Dontchev and W.W. Hager. Lipschitzian stability in nonlinear control and optimization. *SIAM J. Control Optim.*, **31**, 569–603 (1993).
- [8] A.L. Dontchev, W.W. Hager, K. Malanowski. Error bounds for Euler approximation of a state and control constrained optimal control problem. *Numer. Func. Anal. Opt.*, **21**, 653–682 (2000).
- [9] A.L. Dontchev, W.W. Hager, and V.M. Veliov. Second-order Runge-Kutta approximations in control constrained optimal control, *SIAM J. Numerical Anal.*, **38**, 202–226 (2000).
- [10] A.L. Dontchev, R.T. Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis. Second edition*. Springer, New York, 2014.
- [11] I. Ekeland. On the variational principle. *J. Math. Anal. and Appl.*, **47**:324–353 (1974).
- [12] U. Felgenhauer. On stability of bang-bang type controls. *SIAM J. of Control and Optimization*, **41**(6), 1843–1867 (2003).
- [13] U. Felgenhauer. Discretization of semilinear bang-singular-bang control problems. *Comput. Optim. Appl.*, DOI 10.1007/s10589-015-9800-2 (2015).
- [14] U. Felgenhauer, L. Poggolini, G. Stefani. Optimality and stability result for bang-bang optimal controls with simple and double switch behavior. *Control&Cybernetics*, **38**(4B), 1305–1325 (2009).
- [15] R. Ferretti. High-order approximations of linear control systems via Runge-Kutta schemes. *Computing*, **58**(4), 351–364 (1997).
- [16] L. Grüne and P.E. Kloeden. Higher order numerical schemes for affinely controlled nonlinear systems. *Numer. Math.*, **89**, 669–690 (2001).
- [17] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system, *Numerische Mathematik*, **87**:247–282 (2000).
- [18] J. Haunschmied, A. Pietrus, and V.M. Veliov. The Euler Method for Linear Control Systems Revisited. In *Large-Scale Scientific Computing*, I. Lirkov, S. Margenov, J. Wasniewski (Eds.), *Lecture Notes in Computer Science*, **8353**, 90–97, Springer (2014).

- [19] M.I. Krastanov and V.M. Veliov. High-order approximations of nonholonomic affine control systems. In *Large-Scale Scientific Computing*, I. Lirkov, S. Margenov, J. Wasniewski (Eds.), Lecture Notes in Computer Science, v. 5910, 294–301, Springer (2010).
- [20] F. Lempio and V.M. Veliov. Discrete approximations of differential inclusion. *Bayreuther Mathematische Schriften*, **54**, 149–232 (1998).
- [21] K. Malanowski. *Convergence of approximations vs. regularity of solutions for convex control-constrained optimal control problems*. Applied Mathematics and Optimization, **8**, 65-95 (1981).
- [22] N.P Osmolovskii and H Maurer. *Applications to regular and bang-bang control: second-order necessary and sufficient conditions in calculus of variations and optimal control*. Philadelphia: SIAM, 2012.
- [23] A. Pietrus, T. Scarinci, V. Veliov. High order discrete approximations to Mayer’s problems for linear system. Accepted for publication in SIAM J. Control and Opt.   
[http://orcos.tuwien.ac.at/fileadmin/t/orcos/Research\\_Reports/2016-04.pdf](http://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2016-04.pdf)
- [24] J. Preininger, T. Scarinci, V.M. Veliov. Metric regularity properties in bang-bang type linear-quadratic optimal control problems. Preprint.   
(See [http://orcos.tuwien.ac.at/fileadmin/t/orcos/Research\\_Reports/2017-04.pdf](http://orcos.tuwien.ac.at/fileadmin/t/orcos/Research_Reports/2017-04.pdf))
- [25] M. Quincampoix and V.M. Veliov. Metric regularity and stability of optimal control problems for linear systems. *SIAM J. Contr. Optim.*, **51**(5), 4118-4137 (2013).
- [26] M. Seydenschwanz. Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions. *Comput. Optim. Appl.*, **61**(3), 731–760 (2015).
- [27] C. Schneider, G. Wachsmuth. Regularization and discretization error estimates for optimal control of ODEs with group sparsity. *ESAIM: COCV* (To appear).
- [28] V.M. Veliov. *Approximations of differential inclusions by discrete inclusions*. IIASA Working Paper WP-89-017, 1989.
- [29] V.M. Veliov. On the time-discretization of control systems. *SIAM J. Control Optim.*, **35**(5), 1470–1486 (1997).
- [30] V.M. Veliov. Error analysis of discrete Approximation to bang-bang optimal control problems: the linear case. *Control&Cybernetics*, **34**(3), 967–982 (2005).