TECHNISCHE UNIVERSITÄT WIEN

Operations Research and Control Systems

SWM ORCOS

# High Order Discrete Approximations to Mayer's Problems for Linear Systems

*A. Pietrus, T. Scarinci and V.M. Veliov*

**Research Report 2016-04**

June, 2016

**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

# High Order Discrete Approximations to Mayer's Problems for Linear Systems*

A. Pietrus† T. Scarinci ‡ V.M. Veliov§

**Abstract**

The paper presents a discretization scheme for Mayer's type optimal control problems of linear systems. The scheme is based on second order Volterra-Fliess approximations, and on an augmentation of the control variable in a control set of higher dimension. Compared with the existing results, it has the advantage of providing a higher order accuracy without a substantial increase of computations. Error estimations (depending on the controllability index of the system at the solution) are proved by using a recent result about stability of the optimal solution with respect to disturbances. Numerical results are provided, which show the sharpness of the error estimations.

**Keywords**: optimal control, numerical methods, linear systems, error estimation.
**MSC Classification**: 49M25, 65L99.

## 1 Introduction

This paper presents a new discretization scheme and a related error analysis for the following optimal control problem:

$$\min g(x(T)) \tag{1}$$

subject to the linear control system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0, \tag{2}$$

$$u(t) \in U, \tag{3}$$

where $x \in \mathbf{R}^n$, $U = [-1, 1]^m$, $T$ and $x_0$ are fixed and considered as known. The set $\mathcal{U}$ of admissible controls consists of all measurable functions $u : [0, T] \to U$.

In most of the existing literature, the error analysis of discretization methods for ODE optimal control problems is based on certain coercivity properties of the Hamiltonian associated

*The second and the third authors are supported by the Austrian Science Foundation (FWF) under grant No P26640-N25. The preparation of this paper was done mainly during the visit of the third author at Université des Antilles, March, 2016.

†Laboratoire LAMIA, Dépt. de Mathématiques, Université des Antilles, Pointe-à-Pitre, Guadeloupe, apietrus@univ-ag.fr

‡Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria, teresa.scarinci@tuwien.ac.at,teresa.scarinci@gmail.com

§Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria, veliov@tuwien.ac.at

with the problem, and enough smoothness of the optimal control (see e.g. [7, 6, 13, 8, 5] among many). In contrast, the present paper contributes to the still developing area of numerical approximations of problems in which coercivity fails and the optimal controls are typically discontinuous.

Although the optimal control theory for linear systems was broadly developed in the middle of the last century, the investigation of regularity properties of the solutions of problem (1)–(3) and the error analysis of approximation schemes progressed substantially during the last 10–15 years. It is still a challenging issue due to the typical discontinuity of the optimal controls. Concerning the "regular" dependence on disturbances of the solutions of problems that are linear with respect to the control, we refer to [9, 20], and also to [10, 11] and the bibliography therein for extensions to more general affine control problems. The analysis in the present paper is based on [20], where the "regularity" is understood as the so-called "Hölder bi-metric regularity".

A widely used numerical approach to problem (1)–(3) involves solving a discretized problem (usually obtained using the Euler scheme) with the aim of approximately identifying the switching structure of the optimal control. Then, a low-dimensional optimization problem with the switching points used as variables may be solved to locate the switching points more precisely [16, 17, 18]. For this reason, it is important to ensure a high accuracy of the discrete approximation with low computational costs. The higher order approximation proposed in the present paper may make the second stage of the above approach even redundant.

In the next paragraphs we review the literature related to discrete approximations of problems that are linear with respect to the control.

A second order Runge-Kutta scheme (namely, the Heun scheme) is used in [24] to discretize problem (1)–(3), but the error estimate in that paper is not of a better order than the ones subsequently obtained for the Euler scheme. For the analysis of the convergence of the Euler scheme we refer, among others, to [1, 4, 21]. The expected first order error estimate is obtained under the condition that, roughly speaking, switching function associated with the optimal solution has only simple zeros. This condition is relaxed in [14], where the error estimate depends on the multiplicity of the zeros of the switching function, defined indirectly by the so-called *controllability index*, $\sigma$, which is a natural number (see the next section for a definition); the case of simple zeros of the switching function corresponds to $\sigma = 1$. The error estimate derived in [14] is $O(h^{1/\sigma})$, where $h$ is the discretization step. This estimate is sharp.

We mention that the result in [14] applies to Mayer's problem (1)–(3), but its extension to problems involving integral linear-quadratic objective functionals (linear in the control) is requires to overcome considerable difficulties and is done in [21] (see also the analysis of the implicit Euler scheme in [3] for $\sigma = 1$ and [2] where $\sigma$ is any natural number). To the authors' knowledge, the known estimation for the explicit Euler scheme in presence of a bilinear term involving the state and the control in the objective integrand is $O(h^{1/(\sigma+1)})$.

For extensions, concerning the application of Euler's scheme to affine control systems (linear with respect to the control variable), we refer to the recent paper [11].

In this paper we present a discretization scheme for problem (1)–(3), which is based on the Volterra-Fliess expansion of the solution of equation (2), rather than on Runge-Kutta schemes. The idea is based on [23] (see also [15, pp.203-206]) and the apparently independent similar result in [12]. Here we appropriately adapt and implement in the context of problem (1)–(3) this idea, which brings into consideration a discrete-time optimization problem involving additional control variables. The error analysis of this discretization scheme is quite involved

and essentially uses a result from [20]. We will show that the solution of the discrete problem can be used to constructively define an admissible control in problem (1)–(3) which approximates the optimal one with accuracy $O(h^{2/\sigma})$ (recall that $\sigma$ is the controllability index at the solution and $h$ is the discretization step). Thus, the order of accuracy doubles in comparison with the known estimates, while the computational effort is comparable with that for solving the discrete problem obtained by the Euler scheme. As explained later, the case $\sigma = 1$ is in a sense generic, thus our scheme has generically second order accuracy. We also stress that the error estimate is obtained in the metric in the space of control functions that seems most meaningful for the considered problem: the measure of the set on which the approximate control differs from the optimal one. Numerical experiments confirm theoretically obtained rate of convergence.

In addition, we analyze how the accuracy of our numerical scheme changes when the discretized problem is solved with a certain error $\varepsilon$. We will show that the overall error estimate involves both the discretization step $h$ and the solution error $\varepsilon$ and has the form $O\left((\varepsilon + h^2)^{1/\sigma}\right)$.

The paper is organized as follows. In the next section we formulate the assumptions and provide some preliminary results adapting [20]. In Section 3 we present the discretization scheme and formulate the main result – the error estimate. The proof is given in Section 4. Section 5 investigates the effect that inexact solving of the discretized problem has on the overall accuracy. In Section 6 we outline possible numerical implementations of the proposed discretization and present results obtained for test examples. Possible extensions are discussed in Section 7.

## 2 Preliminaries

In this section, some preliminary material is reviewed adapting results from [20]. We begin with some assumptions.

*Assumption (A1):* For some natural number $\bar{\sigma}$, the matrix functions $A : [0,T] \to \mathbf{R}^{n\times n}$ and $B : [0,T] \to \mathbf{R}^{n\times m}$ are $\bar{\sigma}$ times, respectively $\bar{\sigma}+1$ times, continuously differentiable. Moreover, $g : \mathbf{R}^n \to \mathbf{R}$ is convex and differentiable with a locally Lipschitz derivative.

On this assumption, the reachable set, $R$, of (2)–(3) is a convex and compact subset of $\mathbf{R}^n$, hence problem (1)–(3) has at least one solution $(\hat{x}, \hat{u})$.

Define the sequence of matrices

$$B_0(t) = B(t), \quad B_{i+1}(t) = -A(t)B_i(t) + \dot{B}_i(t), \quad i = 0, \dots, \bar{\sigma} - 1, \tag{4}$$

where the dot above a symbol denotes differentiation with respect to $t$.

*Assumption (A2):* $\operatorname{rank}\left[B_0^j(t), \dots, B_{\bar{\sigma}}^j(t)\right] = n$ for every $j = 1, \dots, m$ and every $t \in [0,T]$, where $B_i^j(t)$ is the $j$-th column of $B_i(t)$. Moreover, $\nabla g(x) \neq 0$ for every $x \in R$ ($\nabla g$ denotes the gradient of $g$).

The rank condition in the above assumption is the well-known *general position hypotheses* [19] (see Section 7 about a possible relaxation of (A2)). The second part of assumption (A2) makes the problem meaningful, since it rules out the occurrence of infinitely many solutions.

The Pontryagin maximum principle (here written in the form of "minimum" principle) claims that any optimal pair $(\hat{x}, \hat{u})$, together with a corresponding absolutely continuous function $\hat{p} : [0,T] \to \mathbf{R}^n$, satisfies the following (generalized) equations:

$$\begin{aligned} 0 &= \dot{x}(t) - A(t)\,x(t) - B(t)\,u(t), \quad x(0) = x_0, && (5)\\ 0 &= \dot{p}(t) + A^\top(t)\,p(t), && (6)\\ 0 &= p(T) - \nabla g(x(T)), && (7)\\ 0 &\in B^\top(t)\,p(t) + N_U(u(t)), && (8) \end{aligned}$$

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

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

(Notice that (8) is equivalent to $u(t) \in \operatorname*{Argmin}_{v\in U} \langle B^\top(t)\,p(t), v\rangle$.)

The following lemma is well-known.

**Lemma 1** *Let the matrix-valued functions $A$ and $B$ be measurable and essentially bounded, and let $g$ be differentiable and convex. Then $(\hat{x}, \hat{u})$ is a solution of problem (1)–(3) if and only if the triple $(\hat{x}, \hat{p}, \hat{u})$ (with an absolutely continuous $\hat{p}$) is a solution of system (5)–(8). If (A1) and (A2) hold, then the solution $(\hat{x}, \hat{u})$ of (1)–(3) is unique, hence so is the solution of (5)–(8). Moreover, $\hat{u}(t)$ is a vertex of $U$ for a.e. $t \in [0,T]$.*

Let $(\hat{x}, \hat{u})$ be a solution of problem (1)–(3).

**Definition 1** *The* controllability index *of the solution $(\hat{x}, \hat{u})$ of problem (1)–(3) is the minimal number $\sigma$ such that for every $t \in [0,T]$ and for every $j = 1, \dots, m$ at least one of the numbers $\left\langle \hat{p}(t), B_i^j(t) \right\rangle$, $i = 0, \dots, \sigma$, is not equal to zero. Here, $\hat{p}$ is the solution of the equations (6), (7) with $x(T) = \hat{x}(T)$.*

Assumptions (A1) and (A2) imply that the controllability index $\sigma \le \bar{\sigma}$ does exist.

**Remark 1** *Notice that, according to (7) and (6), $\hat{p}$ is uniquely determined by $\nabla g(\hat{x}(T))$, which belongs to $-N_R(\hat{x}(T))$. Moreover, under (A2) any solution $p$ of (6) with $p(T) \in -N_R(\hat{x}(T))$ produces the same controllability index $\sigma$ in the spirit of Definition 1. Thus $\sigma$ depends only on the location of $\hat{x}(T)$ on $\partial R$. The relation of $\sigma$ with the index of convexity of $R$ at $\hat{x}(T) \in \partial R$ is clarified in [22].*

The generalized equations (5)–(7) can be written in the form $0 \in F(x,p,u)$, where

$$F(x,p,u) := \begin{pmatrix} \dot{x} - A\,x - B\,u \\ \dot{p} + A^\top p \\ p(T) - \nabla g(x(T)) \\ B^\top p + N_{\mathcal{U}}(u) \end{pmatrix}. \qquad (9)$$

The set $N_{\mathcal{U}}(u)$ in (9) is defined point-wise as $\{\rho \in L^\infty : \rho(t) \in N_U(u(t)) \ \forall\, t \in [0,T]\}$. Notice that, strictly speaking, $N_{\mathcal{U}}(u)$ is not the normal cone to the convex set $\mathcal{U}$, since the latter is a subset of the dual space to $L^\infty$. Apparently, $N_{\mathcal{U}}(u)$ is only a subset of "true" normal cone.

Thus, under (A1) and (A2) the inclusion $0 \in F(x,p,u)$ is equivalent to our original problem (1)–(3). Namely, it has a unique solution $(\hat{x},\hat{p},\hat{u})$ and $(\hat{x},\hat{u})$ is the unique solution of problem (1)–(3).

The norms in $L^1(0,T)$ and $L^\infty(0,T)$ are denoted by $\|\cdot\|_1$ and $\|\cdot\|_\infty$, respectively. The notation $W^{1,s} = W^{1,s}([0,T];\mathbf{R}^n)$ (with $s=1$ or $s=\infty$) is used for the space of all absolutely continuous functions $x : [0,T] \to \mathbf{R}^n$ with the derivative $\dot{x}$ belonging to $L^s(0,T)$. The norm in this space is $\|x\|_{1,s} := \|x\|_\infty + \|\dot{x}\|_s$.

The set of admissible controls $\mathcal{U}$ is viewed as a subset of $L^\infty(0,T)$ equipped with the metric

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

This metric is shift-invariant and we use the shorthand notation $d^{\#}(u_1,u_2) = d^{\#}(u_1-u_2,0) =: d^{\#}(u_1-u_2)$. The triple $(x,p,u)$ is considered as an element of the (affine) space

$$\mathcal{X} = W^{1,1}_{x_0} \times W^{1,\infty} \times \mathcal{U},$$

where $W^{1,1}_{x_0} = \{x \in W^{1,1} :\ x(0) = x_0\}$.

Correspondingly, the image space of $F$ will be $\mathcal{Y} = L^1 \times L^\infty \times \mathbf{R}^n \times L^\infty$ with the norm

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

The following is a simplified version of [20, Theorem 2].

**Theorem 1** *Let assumptions (A1) and (A2) be fulfilled, let $(\hat{x},\hat{p},\hat{u})$ be a solution of the generalized equation $0 \in F(x,p,u)$ (with $F$ given in (9)) and let $\sigma$ be its controllability index. Then for every number $b > 0$ there exists a number $c$ such that for every $y = (\xi,\pi,\rho,\nu) \in \mathcal{Y}$ with $\|y\| \le b$ and for every solution $(x,p,u) \in \mathcal{X}$ of the inclusion $y \in F(x,p,u)$ it holds that*

$$\|x-\hat{x}\|_{1,1} + \|p-\hat{p}\|_{1,\infty} + \|u-\hat{u}\|_1 \ \le\ c\,\|y\|^{\frac{1}{\sigma}}. \tag{10}$$

In the proof of our main result we shall need the following known result, [22, Corollary 2.1]. The numbers $m$, $T$ and $\sigma$ are fixed as above.

**Lemma 2** *Let $L$ and $\gamma$ be positive reals, and let $P^\sigma(L,\gamma)$ be the set of all functions $l : [0,T] \to \mathbf{R}^m$ which are $\sigma$-times differentiable, the $\sigma$-th derivative $l^{(\sigma)}$ is Lipschitz continuous with Lipschitz constant $L$, each of the functions $l^{(0)} = l$, $l^{(1)}$, ..., $l^{(\sigma)}$ is uniformly bounded over $[0,T]$ by $L$, and $\sum_{i=0}^{\sigma} |l^{(i)}(t)| \ge \gamma$ for every $t \in [0,T]$. Then there exists a constant $d$ such that*

$$\int_\Delta |l(t)|\,\mathrm{d}t \ge d\,(\text{meas}\,\Delta)^{\sigma+1}$$

*for every $l \in P^\sigma(L,\gamma)$ and every measurable set $\Delta \subset [0,T]$.*

## 3 Approximation scheme and error estimate

The idea of the approximation scheme introduced below originates from [23] and is based on utilization of the Volterra-Fliess series for the solution of (2). Namely, if $x(\theta) = x_\theta$ and $u$ is an admissible control on $[\theta, \theta+h)$, where $\theta \geq 0$, $h > 0$, $\theta+h \leq T$, then the solution $x$ of (2) satisfies for $t \in [\theta, \theta + h]$

$$\begin{aligned} x(t) = x_\theta \quad & + \int_\theta^t [A(s)x_\theta + B(s)u(s)]\,\mathrm{d}s \\ & + \int_\theta^t \int_\theta^s [A(s)A(\tau)x_\theta + A(s)B(\tau)u(\tau)]\,\mathrm{d}\tau\,\mathrm{d}s + O(t;h^3), \end{aligned}$$

where hereafter $O(t;s)$ will denote a function (different at different places) such that $O(t;s)/s$ is bounded when $t, s \in [0,T]$. Inserting the first order Taylor expansion of $A$ and $B$ in the first integral, replacing the arguments of $A$ and $B$ with $\theta$ in the second integral, and changing the order of integration, we obtain that

$$\begin{aligned} x(t) = & \left[I + (t-\theta)A + \frac{(t-\theta)^2}{2}(A^2 + A')\right] x_\theta + (B + (t-\theta)AB)\int_\theta^t u(s)\,\mathrm{d}s \\ & + (-AB + B')\int_\theta^t (s-\theta)u(s)\,\mathrm{d}s + O(t;h^3), \end{aligned} \tag{11}$$

where $A = A(\theta)$, and similarly for $B$, $A'$ and $B'$. Substituting $s = \theta + ht$ and $z(t) = u(\theta + ht)$ in the integrals, we obtain the representation

$$\begin{aligned} x(\theta+h) = & \left[I + hA + \frac{h^2}{2}(A^2 + A')\right] x_\theta \\ & + h(B + hAB)w_1 + h^2(-AB + B')w_2 + O(t;h^3), \end{aligned} \tag{12}$$

where

$$w_1 = \int_0^1 z(s)\,\mathrm{d}s \quad w_2 = \int_0^1 sz(s)\,\mathrm{d}s.$$

Notice that when $u$ runs over the set of all admissible controls on $[\theta, \theta+h]$ the corresponding vectors $(w_1, w_2) \in \mathbf{R}^{2n}$ form a convex and compact set of the form $W^m := \prod_1^m W$ (meaning that each component of the pair of vectors $(w_1, w_2)$ belongs to $W$), where $W \subset \mathbf{R}^2$ is the Aumann integral

$$W := \int_0^1 \begin{pmatrix} 1 \\ s \end{pmatrix} [-1,1]\,\mathrm{d}s. \tag{13}$$

Thus, we obtain that the set of transitions from $x_\theta$ that the control system (2) defines on $[\theta, \theta+h]$ coincides, modulo $O(h^3)$, with the set of transitions defined by the discrete system (12) using the vectors in $W^m$ as control parameters. The approximation scheme below implements this observation, and the main goal of this paper is to prove that this implementation is advantageous in the context of optimal control.

Before presenting the discrete approximation scheme, we mention that it is a standard exercise to represent the above Aumann integral as

$$W = \left\{(\alpha, \beta) \in \mathbf{R}^2 : \alpha \in [-1,1],\ \beta \in [\varphi_1(\alpha), \varphi_2(\alpha)]\right\}, \tag{14}$$

where

$$\varphi_1(\alpha) := \frac{1}{4}(-1+2\alpha+\alpha^2), \quad \varphi_2(\alpha) := \frac{1}{4}(1+2\alpha-\alpha^2). \tag{15}$$

In fact, this will be implied by the proof of the theorem below, but for the need of the formulation of the discretization scheme and the error estimate one can take (14) as a definition of $W$.

The approximating discrete problem reads as follows: given $N$, $h := T/N$, $t_k := kh$, we consider

$$\min g(x_N) \tag{16}$$

subject to the discrete linear control system

$$\begin{aligned} x_{k+1} &= x_k + h(A_k x_k + B_k u_k + hC_k v_k), \quad x_0 \text{ – given}, \\ &\quad (u_k, v_k) \in W^m, \quad k = 0, \dots, N-1, \end{aligned} \tag{17}$$

where

$$\begin{aligned} A_k &= A(t_k) + \frac{h}{2}(A(t_k)^2 + A'(t_k)), \\ B_k &= B(t_k) + hA(t_k)B(t_k), \quad C_k = -A(t_k)B(t_k) + B'(t_k), \end{aligned}$$

The Karush-Kuhn-Tukker theorem gives the following necessary conditions (discrete maximum principle) for the optimality of $(x_0, \dots, x_N)$, $(w_0, \dots, w_{N-1})$, with $w_k := (u_k, v_k) \in W^r$: there is an (adjoint) sequence $(p_0, \dots p_N)$ such that

$$0 = -x_{k+1} + x_k + h(A_k x_k + B_k u_k + hC_k v_k), \quad k = 0, \dots, N-1, \tag{18}$$

$$0 = -p_k + p_{k+1} + hA_k^\top p_{k+1}, \quad k = N-1, \dots, 0, \tag{19}$$

$$0 = -p_N + \nabla g(x_N), \tag{20}$$

$$0 \in (B_k, hC_k)^\top p_{k+1} + N_{W^m}(w_k). \tag{21}$$

Now we consider an arbitrary triplet $(\{x_k\}, \{p_k\}, \{w_k\})$ that satisfies the above four equations. Next, we explain how we can define an appropriate "embedding" of the sequence $\{w_k\}$ into the set $\mathcal{U}$ of admissible controls.

*Construction of a continuous-time control.*
Define the mapping $\mathcal{F}_k : W \to L_\infty(t_k, t_k+1)$ in the following way. Take $(\alpha, \beta) \in W$.

(i) If $\alpha \in \{-1, 1\}$ define $\mathcal{F}_k(\alpha, \beta)(t) = \alpha$, for $t \in [t_k, t_{k+1})$.

(ii) If $\alpha \in (-1, 1)$ and $\beta \in \{\varphi_1(\alpha), \varphi_2(\alpha)\}$ define $\zeta = \operatorname{sgn}(\alpha - 2\beta)$, $\tau = (1+\zeta\alpha)/2$, and

$$\mathcal{F}_k(\alpha, \beta)(t) = \begin{cases} \zeta & \text{for } t \in [t_k, t_k + h\tau), \\ -\zeta & \text{for } t \in [t_k + h\tau, t_{k+1}). \end{cases}$$

(iii) If $\alpha \in (-1, 1)$ and $\beta \in (\varphi_1(\alpha), \varphi_2(\alpha))$, define $\mathcal{F}_k(\alpha, \beta)(t) = 0$ on $[t_k, t_{k+1})$.

Then define an admissible control $u$ of the original problem as

$$u^j(t) = \mathcal{F}_k(u_k^j, v_k^j)(t), \quad t \in [t_k, t_{k+1}), \;\; k = 0, \dots, N-1, \;\; j = 1, \dots, m, \tag{22}$$

where $u^j(t)$, $u_k^j$ and $v_k^j$ are the $j$-th components of $u(t)$, $u_k$ and $v_k$, respectively.

The next theorem is the main result of the paper, which provides an error estimate of the approximation scheme presented above.

**Theorem 2** *Let assumptions (A1), (A2) be fulfilled and let $(\hat{x}, \hat{u})$ be the unique solution of problem (1)–(3) and $\hat{p}$ be the corresponding adjoint function (so that $(\hat{x}, \hat{p}, \hat{u})$ satisfies the Pontryagin system (5)–(8)). Let $\sigma$ be the controllability index of this solution.*

*Then there exists a number $c$ such that for any natural number $N$ and the corresponding $h = T/N$ the following statement is true. For any triple $\{(x_k, p_k, (u_k, v_k))\}$ solving the discrete-time system (18)–(21), the function $u$ defined in (22) belongs to $\mathcal{U}$ and*

$$\max_{k=0,\dots,N} (|x_k - \hat{x}(t_k)| + |p_k - \hat{p}(t_k)|) + d^{\#}(u - \hat{u}) \le c\, h^{2/\sigma}. \tag{23}$$

The proof will be given in the next section. Below we make some comments. As mentioned in the introduction, the case $\sigma = 1$ is generic. More precisely, due to the assumption that $\nabla g(x) \neq 0$ for every $x \in R$, we have that $\hat{x}(T)$ belongs to the boundary, $\partial R$, of the convex and compact set $R$. In [22] it was proved for stationary systems that on the controllability assumption in (A2) the set of points in $\partial R$ (which is an $(n-1)$-dimensional continuous parametric manifold) for which the controllability index $\sigma$ (see Remark 1) is bigger than one forms an $(n-2)$-dimensional continuous manifold $\Gamma$ (it is even empty if $n \le 2$ and $A$ and $B$ are stationary). Thus the case $\sigma = 1$ is "typical". In this case the error estimation (23) is of second order. However, if $\hat{x}(T)$ happens to belong to $\Gamma$, the order of the estimation drops down to $2/\sigma$. In the next section it will be shown by numerical experiments that this error estimation is still sharp, at least for $\sigma = 1, 2, 3, 4$. We also mention that even if $\hat{x}(T) \in \partial R \setminus \Gamma$, the constant $c$ in (23) may become arbitrarily large if $\hat{x}(T)$ is sufficiently close to $\Gamma$. This motivates our goal to study also the "non-generic" case $\sigma > 1$. Notice that the number $\bar{\sigma}$ (which is calculable) in assumption (A2) provides an upper bound for $\sigma$, so that the estimation (23) has always a finite order. This order is doubled (from $1/\sigma$ to $2/\sigma$) compared with the approximations obtained by the Euler scheme (see [14, 21]).

# 4 Proof of Theorem 2

We begin with some preliminaries. First, it is easy to calculate an explicit representation of the normal cone $N_W(\alpha, \beta)$, namely,

$$N_W(\alpha,\beta) = \begin{cases} \emptyset & \text{if } (\alpha,\beta) \notin W \\ \{\alpha(\nu, \mu - \nu)^\top : \mu \ge 0, \nu \ge 0\} & \text{if } \alpha \in \{-1, 1\} \\ \{\mu\,(\zeta + \alpha, -2\zeta)^\top : \mu \ge 0\} & \text{if } \alpha \in (-1,1) \wedge \beta \in \{\varphi_1(\alpha), \varphi_2(\alpha)\} \\ \{0\} & \text{if } \alpha \in (-1,1) \wedge \beta \in (\varphi_1(\alpha), \varphi_2(\alpha)) \end{cases}$$

where $\zeta = \operatorname{sgn}(\alpha - 2\beta)$, as in case (ii) of the construction of a continuous time control.

Second, we shall prove that the construction of the control $u$ in the previous section implies in the cases (i) and (ii) the equalities

$$\int_{t_k}^{t_{k+1}} u(s)\,\mathrm{d}s = h u_k, \quad \int_{t_k}^{t_{k+1}} (s - t_k) u(s)\,\mathrm{d}s = h^2 v_k. \tag{24}$$

In the case (i) we have $u_k^j = \alpha \in \{-1, 1\}$, hence (according to (14) and (15)) $v_k^j = \alpha/2$, and $u^j(t) \equiv \alpha$ on $[t_k, t_{k+1})$. Then

$$\int_{t_k}^{t_{k+1}} u^j(s)\,\mathrm{d}s = \int_{t_k}^{t_{k+1}} \alpha\,\mathrm{d}s = h\alpha = h u_k^j, \qquad \int_{t_k}^{t_{k+1}} (s - t_k) u^j(s)\,\mathrm{d}s = \frac{h^2}{2}\alpha = h^2 v_k^j.$$

In the case (ii) we have $u_k^j = \alpha$, $v_k^j = \beta = \varphi_i(\alpha)$, where $i$ equals 1 or 2, depending on whether $\zeta = \operatorname{sgn}(\alpha - 2\beta)$ equals 1 or $-1$. Indeed, if $\beta = \varphi_1(\alpha)$, then

$$\zeta = \operatorname{sgn}(\alpha + 2\varphi_1(\alpha)) = \operatorname{sgn}(1 - \alpha^2) = 1.$$

if $\beta = \varphi_2(\alpha)$, then

$$\zeta = \operatorname{sgn}(\alpha + 2\varphi_2(\alpha)) = \operatorname{sgn}(-1 + \alpha^2) = -1.$$

In both cases we obtain from (15) that

$$v_k^j = \frac{1}{4}\left(\zeta + \frac{1}{2}\alpha + \zeta\alpha^2\right).$$

According to the definition of $u^j(t)$ on $[t_k, t_{k+1})$, we have

$$\begin{aligned}\int_{t_k}^{t_{k+1}} u^j(s)\,\mathrm{d}s \quad &= \int_{t_k}^{t_k+h\tau} \zeta\,\mathrm{d}s - \int_{t_k+h\tau}^{t_{k+1}} \zeta\,\mathrm{d}s = h\zeta(2\tau - 1) \\ &= h\zeta\left(2\frac{1+\zeta\alpha}{2} - 1\right) = h\alpha = hu_k^j, \end{aligned} \tag{25}$$

$$\begin{aligned}\int_{t_k}^{t_{k+1}} (s - t_k)u^j(s)\,\mathrm{d}s &= \int_{t_k}^{t_k+h\tau} (s - t_k)\zeta\,\mathrm{d}s - \int_{t_k+h\tau}^{t_{k+1}} (s - t_k)\zeta\,\mathrm{d}s \\ &= \frac{2\tau^2 - 1}{2}\zeta h^2 = \left[\left(\frac{1+\zeta\alpha}{2}\right)^2 - \frac{1}{2}\right]\zeta h^2 \\ &= \frac{1}{4}\left(\zeta + \frac{1}{2}\alpha + \zeta\alpha^2\right)h^2 = h^2 v_k^j. \end{aligned}$$

Now we shall modify the control $u$ in intervals $[t_k, t_{k+1})$ where at least one of its components is defined as in point (iii) of the construction of $u$. For each such $k$ and component $j$, we define

$$\tilde{u}^j(t) = \begin{cases} 0 & \text{for } t \in [t_k, t_k + h\theta), \\ \zeta & \text{for } t \in [t_k + h\theta, t_k + h\tau), \\ -\zeta & \text{for } t \in [t_k + h\tau, t_{k+1}), \end{cases} \tag{26}$$

where $\zeta = \operatorname{sgn}(u_k^j - 2v_k^j)$ and $0 < \theta < \tau < 1$ are chosen in such a way, that the equalities (24) are fulfilled for $\tilde{u}$. For intervals $[t_k, t_{k+1})$ for which all components of $u$ are defined as in points (i) and (ii) we set $\tilde{u}(t) = u(t)$. Existence and uniqueness of numbers $\theta$ and $\tau$ as above can be proven as follows. For a fixed $\theta \in [0, 1]$ we have

$$\begin{aligned}&\left\{\left(\int_\theta^1 u(s)\,\mathrm{d}s, \int_\theta^1 su(s)\,\mathrm{d}s\right)^\top : u - \text{measurable}, u(t) \in [-1, 1]\right\} \\ &= \int_\theta^1 \begin{pmatrix} 1 \\ s \end{pmatrix}[-1, 1]\,\mathrm{d}s =: W_\theta. \end{aligned} \tag{27}$$

By change of the variable $s$ with $t = (s - \theta)/(1 - \theta)$ we obtain the relation

$$W_\theta = \begin{pmatrix} 1 - \theta & 0 \\ 0 & (1 - \theta)^2 \end{pmatrix} W.$$

Thus the boundary of $W_\theta$ continuously shrinks from $\partial W$ to 0 when $\theta$ changes from 0 to 1. Since $(u_k^j, v_k^j) \in \operatorname{int} W$, there will be some $\theta > 0$ such that $(u_k^j, v_k^j) \in \partial W_\theta$. Then the choice of $\tau > \theta$ can be made exactly as in point (ii) of the construction of $u$, which results in (26).

The proof below is based on Theorem 1. Having embedded the sequence $\{(u_k, v_k)\}$ in the set $\mathcal{U}$ as $\tilde{u}$, we need to embed also the sequences $\{x_k\}$ and $\{p_k\}$ into the spaces $W^{1,1}_{x_0}$ and $W^{1,1}$, respectively. Using the hint given by the representation in (11) we define, for $t \in [t_k, t_{k+1})$,

$$\begin{aligned} x(t) \;:=\; & \left( I + (t-t_k)A(t_k) + \frac{(t-t_k)^2}{2}(A(t_k)^2 + A'(t_k)) \right) x_k \\ & + (B(t_k) + (t-t_k)A(t_k)B(t_k)) \int_{t_k}^t \tilde{u}(s)\,\mathrm{d}s + C_k \int_{t_k}^t (s-t_k)\tilde{u}(s)\,\mathrm{d}s. \end{aligned} \tag{28}$$

Similarly, we define

$$p(t) := \left[ I + (t_{k+1}-t)A(t_k)^\top + \frac{(t_{k+1}-t)^2}{2} A^2(t_k)^\top + \frac{h^2 - (t-t_k)^2}{2} A'(t_k)^\top \right] p_{k+1}. \tag{29}$$

We observe from (28), the definitions of $A_k$, $B_k$ and $C_k$, and (24) that

$$\begin{aligned} \lim_{t \to t_{k+1}} x(t) &= (I + hA_k)x_k + B_k \int_{t_k}^{t_{k+1}} u(s)\,\mathrm{d}s + C_k \int_{t_k}^{t_{k+1}} (s-t_k)u(s)\,\mathrm{d}s \\ &= (I + hA_k)x_k + hB_k u_k + h^2 C_k v_k = x_{k+1}. \end{aligned}$$

Thus $x$ is continuous at $t_{k+1}$, hence it is absolutely continuous. Since $x(0) = x_0$, we obtain that $x \in W^{1,1}_{x_0}$. Similarly we obtain that $p \in W^{1,1}$. Thus, $(x, p, \tilde{u}) \in \mathcal{X}$.

In order to apply Theorem 1, we shall estimate the residual $y = (\xi, \pi, \nu, \rho)$ that $(x, p, \tilde{u})$ gives in (5)–(8).

**1.** Residual in (5).
From (28) we have

$$\begin{aligned} \dot{x}(t) &= [A(t_k) + (t-t_k)(A(t_k)^2 + A'(t_k))]x_k + A(t_k)B(t_k) \int_{t_k}^t \tilde{u}(s)\,\mathrm{d}s \\ &\quad + [B(t_k) + (t-t_k)A(t_k)B(t_k)]\tilde{u}(t) + (t-t_k)C_k\tilde{u}(t) \\ &= [A(t) + (t-t_k)A(t_k)^2 + O(t;h^2)]x_k + A(t)B(t) \int_{t_k}^t \tilde{u}(s)\,\mathrm{d}s + O(t;h^2) \\ &\quad + [B(t_k) + (t-t_k)(A(t_k)B(t_k) + C_k)]\tilde{u}(t) \\ &= A(t) \left[ (I + (t-t_k)A(t_k))\, x_k + B(t) \int_{t_k}^t \tilde{u}(s)\,\mathrm{d}s \right] + [B(t_k) + (t-t_k)B'(t_k)]\tilde{u}(t) + O(t;h^2) \\ &= A(t)(x(t) + O(t;h^2)) + B(t)\tilde{u}(t) + O(t;h^2), \end{aligned}$$

where $O(t;h^2)/h^2$ (which may be different at different places) is uniformly bounded in $t \in [0,T]$ and $h \in [0,T]$. Thus, $\|\xi\|_1 = O(h^2)$.

**2.** Residual in (6) and (7).

Since $p^N(\cdot)$ interpolates the sequence $\{p_k\}$, we have $p^N(T) = p_N$. Due to (20) and $x^N(T) = x_N$, we have that $(x^N, p^N)$ satisfy (7) exactly, that is, $\nu = 0$.

To estimate the residual in (6), we differentiate the expression in (29). This gives

$$\begin{aligned} \dot{p}^N(t) &= -\left[A(t_k)^\top + (t_{k+1} - t)A^2(t_k)^\top + (t - t_k)A'(t_k)^\top\right] p_{k+1} \\ &= -A(t)^\top \left[I + (t_{k+1} - t)A(t_k)^\top\right] p_{k+1} + O(h^2) \\ &= -A(t)^\top p^N(t) + O(t; h^2). \end{aligned}$$

Thus, $\|\pi\|_\infty = O(h^2)$.

**3.** Residual in (8).

First of all we shall prove for every $k = 0, \ldots, N-1$ the inclusion

$$\left\langle p_{k+1}, B_k^j + (t - t_k)C_k^j \right\rangle \in -N_{[-1,1]}(\tilde{u}^j(t)), \quad t \in [t_k, t_{k+1}), \;\; j = 1, \ldots, m. \tag{30}$$

We consider separately the three cases in the definition of the mapping $\mathcal{F}_k$.

Consider first the case (i), where $u_k^j = \alpha \in \{-1, 1\}$. In this case $\tilde{u}(t) = u(t) = \alpha$ on $[t_k, t_{k+1})$. According to (21) and the representation of $N_W$, there exists $\mu \geq 0$ and $\nu \geq 0$ such that

$$\langle p_{k+1},\, B_k^j \rangle = \alpha\nu, \quad h\langle p_{k+1},\, C_k^j \rangle = \alpha(\mu - \nu).$$

Hence,

$$\begin{aligned} \langle p_{k+1},\, B_k^j \rangle + (t - t_k)\langle p_{k+1},\, C_k^j \rangle &= \alpha\left[\nu + \frac{\mu - \nu}{h}(t - t_k)\right] = \alpha\left[\mu\frac{t - t_k}{h} + \nu\left(1 - \frac{t - t_k}{h}\right)\right] \\ &\in N_U(\tilde{u}(t)), \end{aligned}$$

where the last inclusion holds since the expression in the brackets is non-negative. Thus (30) is fulfilled.

Now, consider the case (ii) in the definition of the mapping $\mathcal{F}_k$. Here $u_k^j = \alpha \in (-1, 1)$, $v_k^j = \beta \in \{\varphi_1(\alpha), \varphi_2(\alpha)\}$, and

$$\tilde{u}(t) = u(t) = \begin{cases} \zeta & \text{for } t \in [t_k, t_k + h\tau), \\ -\zeta & \text{for } t \in [t_k + h\tau, t_{k+1}), \end{cases}$$

where $\zeta = \operatorname{sgn}(\alpha - 2\beta)$, $\tau = (1 + \zeta\alpha)/2$. According to (21) and the representation of $N_W$, there exists $\mu \geq 0$ such that

$$\langle p_{k+1},\, B_k^j \rangle = -\mu(\zeta + \alpha), \quad h\langle p_{k+1},\, C_k^j \rangle = 2\mu\zeta.$$

Then

$$\langle p_{k+1},\, B_k^j \rangle + (t - t_k)\langle p_{k+1},\, C_k^j \rangle = \mu\left(-\zeta - \alpha + 2\zeta\frac{t - t_k}{h}\right).$$

Now, let $t \in [t_k, t_k + h\tau)$. Having in mind that $\tau = (1 + \zeta\alpha)/2$, we obtain that

$$2\zeta\frac{t - t_k}{h} \leq 2\tau = 1 + \zeta\alpha,$$

hence

$$\zeta\left(\langle p_{k+1},\, B_k^j\rangle + (t-t_k)\langle p_{k+1},\, C_k^j\rangle\right) \quad = \zeta\mu\left(-\zeta - \alpha 2\zeta\frac{t-t_k}{h}\right)$$
$$\leq \mu\left(-1-\alpha\zeta+1+\alpha\right) = 0.$$

Since for $t \in [t_k, t_k + h\tau)$ the definition of $u^j(\cdot)$ in (22) gives $u^j(t) = \zeta$, thus also $\tilde{u}^j(t) = \zeta$, the last inequality is equivalent to (30). For $t \in [t_k + h\tau, t_{k+1})$ we have $u^j(t) = -\zeta$ and (30) is obtained by a similar calculation as above.

Now, consider the case (iii), where $u_k^j \in (-1, 1)$ and $v_k^j \in (\varphi_1(u_k^j), \varphi_2(u_k^j))$. Then $N_W(u_k^j, v_k^j) = \{0\}$, hence

$$\langle p_{k+1},\, B_k^j\rangle = 0, \quad h\langle p_{k+1},\, C_k^j\rangle = 0. \tag{31}$$

This immediately implies (30).

Thus (30) is proved in all cases. Then taking into account that

$$N_U(u) = \Pi_{j=1}^m N_{[-1,1]}(u^j),$$

we obtain

$$(p_{k+1})^\top(B_k + (t-t_k)C_k) \in -N_{[-1,1]}(\tilde{u}(t)), \quad t \in [t_k, t_{k+1}), \tag{32}$$

On the other hand we represent for $t \in [t_k, t_{k+1})$

$$\begin{aligned}
p(t)^\top B(t) &= (p_{k+1})^\top[I + (t_{k+1}-t)A(t_k) + O(t;h^2)]\,[B(t_k) + (t-t_k)B'(t_k) + O(t;h^2)] \\
&= (p_{k+1})^\top(B(t_k) + (t_{k+1}-t)A(t_k)B(t_k) + (t-t_k)B'(t_k)) + O(t;h^2) \\
&= (p_{k+1})^\top(B(t_k) + hA(t_k)B(t_k) + (t_k-t)A(t_k)B(t_k) + (t-t_k)B'(t_k)) + O(t;h^2) \\
&= (p_{k+1})^\top(B_k + (t-t_k)C_k) + O(t;h^2). \qquad (33)
\end{aligned}$$

Combining the above equality with (32), we obtain that

$$B^\top(t)\,p(t) + O(t;h^2) \in -N_U(\tilde{u}(t)), \tag{34}$$

hence $\|\rho\|_\infty = O(h^2)$. 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 1 implies existence of $c$ such that for every natural $N$

$$\|x - \hat{x}\|_{1,1} + \|p - \hat{p}\|_{1,\infty} + \|\tilde{u} - \hat{u}\|_1 \;\leq\; c\,\|y\|^{1/\sigma}.$$

We know that $x(t_k) = x_k$ and $p(t_k) = p_k$, hence

$$\max_{k=0,\ldots,N}(|x_k - \hat{x}(t_k)| + |p_k - \hat{p}(t_k)|) + \|\tilde{u} - \hat{u}\|_1 \leq c_2\, h^{2/\sigma}. \tag{35}$$

Now we focus on the last term in the right-hand side. First, we have that $\hat{u}(t) \in \{-1,\, 1\}$ and $\tilde{u}(t) \in \{-1,\, 0,\, 1\}$. This easily implies

$$d^{\#}(\tilde{u} - \hat{u}) \leq \sqrt{Tm}\,\|\tilde{u} - \hat{u}\|_1. \tag{36}$$

Second, we notice that $\tilde{u}(t) \neq u(t)$ for some $t$ only if $t$ belongs to some interval $[t_k, t_{k+1}]$ where some of the components of $u$, say $u^j$, is constructed as in point (iii). In this case we have (31) and from (33) we obtain existence of a constant $c_3$ such that

$$|p(t)^\top B^j(t)| \leq c_3 h^2.$$

Denote $l(t) = p(t)^\top B^j(t)$, $\Delta = \{t \in [0,T] : |l(t)| \le c_2 h^2\}$. From the definition of $B_i$ and (6) we see that $l^{(i)}(t) = \langle p(t), B_i^j(t)\rangle$. Then (A2) implies the existence of $\gamma > 0$ such that $\sum_{i=0}^{\sigma} |l^{(i)}(t)| \ge \gamma$ for every $t \in [0,T]$, thus $l \in P^\sigma(L,\gamma)$ for an appropriate $L$. Using Lemma 2 we obtain

$$d\,(\operatorname{meas}\Delta)^{\sigma+1} \le \int_\Delta |l(t)|\,\mathrm{d}t \le c_3 h^2 \operatorname{meas}\Delta,$$

hence

$$\operatorname{meas}\Delta \le \left(\frac{c_2 h^2}{d}\right)^{1/\sigma} = c_4 h^{\frac{2}{\sigma}}.$$

Thus

$$d^\#(u-\hat u) \le d^\#(\tilde u - \hat u) + m c_4 h^{\frac{2}{\sigma}}.$$

Combining this with (35) and (36) we finish the proof.

# 5 Error estimate in case of inexact solutions of problem (18)–(21)

In Section 3 we assume that the discrete-time system (18)–(21) is exactly solved. Having the solution, we may obtain an approximation of the solution of the original problem for which the estimation in Theorem 2 holds. In practice, finding a solution of this system requires (excluding the case of a linear function $g$) application of an iterative procedure, resulting in an approximate solution $(\{\tilde x_k\},\ \{\tilde p_k\},\ \{\tilde w_k\})$. We measure the inexactness of this approximate solution by the residual $(\xi,\pi,\nu,\rho)$ that $(\{\tilde x_k\},\ \{\tilde p_k\},\ \{\tilde w_k\})$ produces in the left-hand side of (18)–(21). Here each of the components of $(\xi,\pi,\nu,\rho)$ has corresponding dimension; for example, $\xi = (\xi_0,\dots,\xi_{N-1}) \in \mathbf{R}^{N\times n}$, etc. Denote

$$\varepsilon := \|\xi\|_{l_1} + \|\pi\|_{l_\infty} + |\nu| + \|\rho\|_{l_\infty} = h\sum_{k=0}^{N-1} |\xi_k| \ + \max_{k=0,\dots,N-1} |\pi_k| \ + \ |\nu| \ + \max_{k=0,\dots,N-1} |\rho_k|.$$

Using the approximate solution $(\{\tilde x_k\},\ \{\tilde p_k\},\ \{\tilde w_k\})$ of system (18)–(21), one can define an approximation, $\tilde u(\cdot)$, of the optimal control $\hat u$ in the same way as described in the part "Construction of a continuous-time control" of Section 3. Having in mind the proof of Theorem 2, it is to expect that it remains true with $(\{x_k\},\ \{p_k\})$ replaced with $(\{\tilde x_k\},\ \{\tilde p_k\})$ and $u$ replaced with $\tilde u$, and with the following modification of the error estimation (23):

$$\max_{k=0,\dots,N} (|\tilde x_k - \hat x(t_k)| + |\tilde p_k - \hat p(t_k)|) + d^\#(\tilde u - \hat u) \le c\,(\varepsilon + h^2)^{1/\sigma}. \tag{37}$$

It is straightforward to prove this estimation if $\xi = \pi = 0$, since the relations (20) and (21) are pointwise. Thus the residual $\rho$ can be just added to the left-hand side of (34). Similarly for (20). If there are inaccuracies in (18) and (19), the situation becomes more complicated, since the embedding of $(\{\tilde x_k\},\ \{\tilde p_k\})$ into $W^{1,1}_{x_0} \times W^{1,1}$ so that the residual in (5) and (6) is of order $\varepsilon + h^2$ becomes problematic. However, this is not a principle trouble, as argued below.

Given the sequence $\{\tilde w_k\}$, one can recalculate the solution of (18)–(20) for $\tilde w_k$, obtaining new sequences $(\{\bar x_k\},\ \{\bar p_k\})$. Observe, that this calculation can be done exactly (neglecting round-off

computational errors). Then the triplet $(\{\bar{x}_k\}, \{\bar{p}_k\}, \{\tilde{w}_k\})$ satisfies relations (18)–(21) with a residual $(0, 0, 0, \bar{\rho})$, where

$$|\bar{\rho}_k| \le |\rho_k| + |(B_k, hC_k)^\top (\bar{p}_{k+1} - \tilde{p}_{k+1})| \le |\rho_k| + c_0 \, |\bar{p}_{k+1} - \tilde{p}_{k+1}|,$$

with an appropriate constant $c_0$. In a standard way one can estimate $|\bar{x}_k - \tilde{x}_k| \le c_1 \|\xi\|_{l_1}$, with some constant $c_1$. Then an estimation $|\bar{p}_N - \tilde{p}_N| \le c_2(|\nu| + \|\xi\|_{l_1})$ follows from (7), hence also $|\bar{p}_k - \tilde{p}_k| \le c_3(\|\pi\|_{l_\infty} + |\nu| + \|\xi\|_{l_1})$ for some constants $c_2$ and $c_3$.

Summarizing the above and using (37), now with $(\{\bar{x}_k\}, \{\bar{p}_k\}, \{\tilde{w}_k\})$ and residual $(0, 0, 0, \bar{\rho})$, we obtain that

$$\max_{k=0,\dots,N} (|\bar{x}_k - \hat{x}(t_k)| + |\bar{p}_k - \hat{p}(t_k)|) + d^{\#}(\tilde{u} - \hat{u}) \le c \, (\varepsilon + h^2)^{1/\sigma} \tag{38}$$

with an appropriate constant $c$. Thus, in order to keep the overall error of order $2/\sigma$, one has to solve the discrete system (18)–(21) with accuracy (in terms of residual) $h^2$.

## 6 Implementation of the approach and numerical experiments

In the implementation of the approximation scheme presented in Section 3 one needs to approximately solve the discrete-time problem (18)–(21) (see the previous section about error analysis). This can be done in many ways, out of which we mention the shutting method (where minimization of the residual in (20) is sought iteratively) and the direct approach, which is based on the observation that system (18)–(21) represents a necessary optimality condition for the discrete-time problem

$$\min g(x_N)$$

subject to equation (18) and the control constraints $w_k = (u_k, v_k) \in W^m$. This is a mathematical programming problem to which various algorithms can be applied. We implement the gradient projection method in the control space, since the gradient of the objective function can be easily calculated using the adjoint equation (19) with the end-point condition (20) at each iteration.

We mention that solving problem (18)–(21) is not substantially more complicated than solving the one obtained by the Euler discretization (cf. [4, 14]), especially if the matrices $A_k$, $B_k$ and $C_k$ are pre-calculated. Solving the variational inequality (21) with respect to $w_k$ is not problematic, since it splits to $m$ independent variational inequalities, each of them of the form $(\xi_1, \xi_2) \in N_W(w)$. The solution $w = (\alpha, \beta)$ is given by the simple formula

$$(\alpha, \beta) = \begin{cases} (-1, -1/2) & \text{if } \xi_1 \le 0 \text{ and } \xi_1 + \xi_2 \le 0, \\ (1, 1/2) & \text{if } \xi_1 > 0 \text{ and } \xi_1 + \xi_2 \ge 0, \\ (-1 - 2\xi_1/\xi_2, \varphi_1(\alpha)) & \text{if } \xi_1 > 0 \text{ and } \xi_1 + \xi_2 < 0, \\ (1 + 2\xi_1/\xi_2, \varphi_2(\alpha)) & \text{if } \xi_1 \le 0 \text{ and } \xi_1 + \xi_2 > 0. \end{cases}$$

The inexactness of the iterative procedure for solving system (18)–(21) influences the overall error estimate as described in Section 5. In order to focus on the main result of this paper, Theorem 2, in the examples below we consider linear functions $g$, where system (18)–(21) can be exactly solved (modulo round-off computational errors).

**Example 1** (Control of a harmonic oscillator)

Consider the following problem on the interval $[0, 3\pi]$:

$$\begin{array}{ll} \text{minimize} & x_2(3\pi) \\ \text{subject to} & \dot{x}_1(t) = x_2(t), \\ & \dot{x}_2(t) = -x_1(t) + u(t), \\ & x(0) = 0, \\ & u(t) \in [-1, 1]. \end{array}$$

The exact optimal control in this problem is known:

$$\hat{u}(t) = \begin{cases} 1 & \text{if } t \in [0, \pi/2) \cup (3\pi/2, 5\pi/2)\,, \\ -1 & \text{if } t \in (\pi/2, 3\pi/2) \cup (5\pi/2, 3\pi]\,. \end{cases}$$

Here assumption (A2) is fulfilled with $\sigma = 1$, therefore Theorem 2 claims accuracy estimation proportional to $h^2$. In the three rows of Table 1 we present for various values of $N$: (i) the absolute error of the numerically obtained control $u^N$, $e_N := d^\sharp(u^N - \hat{u})$; (ii) the ratio $e_N/h^2$, which is claimed to be bounded; (iii) the ration $e_N/e_{2N}$, which is expected to be around 4 (although this is not formally implied by Theorem 2). The numerical results completely support the theoretical prediction.

| $N$ | 50 | 100 | 200 | 400 | 800 | 1600 | 3200 | 10000 |
|---|---|---|---|---|---|---|---|---|
| $e_N$ | 0.0780 | 0.0204 | 0.0052 | 0.0013 | $3.26 \cdot 10^{-4}$ | $8.16 \cdot 10^{-5}$ | $2.04 \cdot 10^{-5}$ | $2.09 \cdot 10^{-6}$ |
| $\frac{e_N}{h^2}$ | 195.00 | 204.00 | 208.00 | 208.00 | 208.44 | 208.87 | 209.09 | 209.23 |
| $\frac{e_N}{e_{2N}}$ | 3.8235 | 3.9231 | 4.000 | 3.9914 | 3.992 | 3.996 | 3.998 | |

Table 1: Here $e_N$ is the error $e_N = d^\sharp(u^N - \hat{u})$ of the numerically obtained control $u^N$ in Example 1 for various values of $N$. The quantities $e_N/h^2$ and $e_N/e_{2N}$ are given, which, according to Theorem 2, are expected to be bounded and approximately equal to 4, respectively.

**Example 2** (A non-stationary harmonic oscillator)

The second example numerically checks Theorem 2 in the non-stationary case. Consider the following problem on the time interval $[0, 1]$:

$$\begin{array}{ll} \text{minimize} & x_1(1) \\ \text{subject to} & \dot{x}_1(t) = a\,t\,x_2(t) + tu(t), \\ & \dot{x}_2(t) = t\,u(t), \\ & x(0) = 0, \\ & u(t) \in [-1, 1], \end{array}$$

where $a = 32/(\pi^2 - 16)$. The constant $a$ is chosen in such a way that the optimal control in this problem is

$$\hat{u}(t) = \begin{cases} 1 & \text{if } t \in [0, \pi/4)\,, \\ -1 & \text{if } t \in (\pi/4, 1]\,. \end{cases}$$

Here assumption (A2) is also fulfilled with $\sigma = 1$. The structure of Table 2 is as in the previous example and the results are also consistent with Theorem 2.

| $N$ | 50 | 100 | 200 | 400 | 800 | 1600 | 3200 |
|---|---|---|---|---|---|---|---|
| $e_N$ | $3.3 \cdot 10^{-4}$ | $7.7 \cdot 10^{-5}$ | $1.8 \cdot 10^{-5}$ | $4.9 \cdot 10^{-6}$ | $1.3 \cdot 10^{-6}$ | $2.6 \cdot 10^{-7}$ | $8.2 \cdot 10^{-8}$ |
| $\frac{e_N}{h^2}$ | 0.8352 | 0.7713 | 0.7223 | 0.7896 | 0.8482 | 0.6703 | 0.8420 |
| $\frac{e_N}{e_{2N}}$ | 4.3316 | 4.3030 | 3.6596 | 3.7245 | 5.0403 | 5.0190 | 3.200 |

Table 2: The structure of this table for Example 2 is the same as that of Table 1.

**Example 3** (Cases with index $\sigma > 1$)

Here we present experiments with a family of problems with various controllability indexes $\sigma$, given in [21]. Below, the time-interval is $[0, 1]$, the dimension of the state is $n = \sigma + 1$ and the dynamics depends on parameters $s_j$:

$$\begin{array}{ll} \text{minimize} & x_1(1) \\ \text{subject to} & \dot{x}_j(t) = s_j x_{j+1}(t) + u(t), \quad j = 1, \dots, \sigma, \\ & \dot{x}_{\sigma+1}(t) = u(t), \\ & x(0) = 0, \\ & u(t) \in [-1, 1]. \end{array}$$

For any natural number $\sigma$ the values of the parameters $s_j$ are chosen in such a way that the solution is

$$\hat{u}(t) = \begin{cases} 1 & \text{if } t \in [0, 1/2)\,, \\ -1 & \text{if } t \in (1/2, 1]\,, \end{cases}$$

if $\sigma$ is odd, and $\hat{u}(t) \equiv -1$ if $\sigma$ is even. Moreover, the controllability index of the solution is $\sigma$. This is achieved by choosing $s_j := -2(\sigma - j + 1)$, $j = 1, \dots, \sigma$ (see [21]).

Our numerical experiments for $\sigma = 2$, 3, 4 are presented on Table 3. As asserted by Theorem 2, the values $e_N/h^{2/\sigma}$ are bounded (as above $e_N = d^\sharp(u^N - \hat{u})$).

| | $N$ | 50 | 100 | 200 | 400 | 800 | 1600 | 3200 |
|---|---|---|---|---|---|---|---|---|
| $\sigma = 2$ | $e_N$ | 0.0300 | 0.0150 | 0.0075 | 0.0038 | $6.250 \cdot 10^{-4}$ | $3.12 \cdot 10^{-4}$ | $1.56 \cdot 10^{-4}$ |
| | $\frac{e_N}{h^{2/2}}$ | 1.5000 | 1.5000 | 1.5000 | 1.5000 | 0.5000 | 0.5000 | 0.5000 |
| $\sigma = 3$ | $e_N$ | 0.0607 | 0.0396 | 0.0242 | 0.0150 | 0.0093 | 0.0058 | 0.0037 |
| | $\frac{e_N}{h^{2/3}}$ | 0.8244 | 0.8538 | 0.8265 | 0.8143 | 0.8010 | 0.7976 | 0.7988 |
| $\sigma = 4$ | $e_N$ | 0.0823 | 0.0525 | 0.0327 | 0.0210 | 0.0136 | 0.0086 | 0.0055 |
| | $\frac{e_N}{h^{2/4}}$ | 0.5818 | 0.5250 | 0.4625 | 0.4206 | 0.3836 | 0.3466 | 0.3119 |

Table 3: The errors $e_N := d^\sharp(u_N - \hat{u})$ and the ratios $e_N/h^{2/\sigma}$ for controllability indexes $\sigma = 2$, 3, 4 and various values of $N$.

# 7 Concluding remark

First, we point out that assumption (A2) may be too restrictive in the multi-control case, since it requires the "controllability" for each of the control components separately. This, however, is not necessary for the finiteness of the controllability index of a given triple $(x, u, p)$ satisfying the necessary optimality conditions (5)–(8), which is the property actually used in the proofs. We assume (A2) in order to utilize the results in [20]; extending the error estimate in the present paper for relaxed versions of (A2) would require further analysis in line with [20].

The applicability of our discretization scheme to problems with an integral term in the objective functional is a subject of future research. It will also need elaboration of results in [20].

In this paper, the feasible set $U$ is box-like: a product of intervals. However, the Aumann integral in (13) is constructively representable for some more general sets $U$ (in that case the integral is, in general, a non-decomposable subset of $\mathbf{R}^{2m}$). Our approach can be extended also to such cases, but this requires a proper definition of the "continuous-time control" in Section 3.

Formally, the discretization approach presented in this paper can be extended to affine (linear with respect to the control) problems, as in [23]. However, the theoretical ground in the spirit of [20] for establishing sharp error estimates is still missing in the non-linear case and its development is an important subject of further research.

**Acknowledgment:** The authors wish to thank Asen Dontchev for the valuable suggestions concerning the exposition.

## References

[1] W. Alt, R. Baier, M. Gerdts, F. Lempio. Error bounds for Euler approximations of linear-quadratic control problems with bang-bang solutions. *Numerical Algebra, Control and Optimization*, **2**(3) (2012), 547–570.

[2] W. Alt, C. Schneider, M. Seydenschwanz. Regularization and implicit euler discretization of linear-quadratic optimal control problems with bang-bang solutions. *Applied Mathematics and Computation*, 287-288C (2016), 104–124.

[3] W. Alt and M. Seydenschwanz. An implicit discretization scheme for linear-quadratic control problems with bang-bang solutions. *Optim. Methods & Software*, **29**(3) (2014), 535–560.

[4] W. Alt, R. Baier, F. Lempio, M. Gerdts. Approximations of linear control problems with bang-bang solutions. *Optimization*, **62**(1) (2013), 9–32.

[5] J.F. Bonnans and A. Festa. Error estimates for the Euler discretization of an optimal control problem with first-order state constraints. Inria Report, Dec. (2014).

[6] A.L. Dontchev. An a priori estimate for discrete approximations in nonlinear optimal control, *SIAM J. Control Optim.*, **34**(1996), 1315–1328.

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

[8] 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**(2000), 202–226.

[9] U. Felgenhauer. On stability of bang-bang type controls. *SIAM J. of Control and Optimization*, **41**(6) (2003), 1843–1867.

[10] 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) (2009), 1305–1325.

[11] U. Felgenhauer. Discretization of semilinear bang-singular-bang control problems. *Comput. Optim. Appl.*, DOI 10.1007/s10589-015-9800-2.

[12] R. Ferretti. High-order approximations of linear control systems via Runge-Kutta schemes. *Computing*, **58**(4)(1997), 351–364.

[13] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system, *Numerische Mathematik*, **87** (2000), 247-282.

[14] 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** (2014) 90–97, Springer.

[15] F. Lempio and V.M. Veliov. Discrete approximations of differential inclusion. *Bayreuther Mathematische Schriften*, **54**(1998), 149–232.

[16] N.P Osmolovskii and H Maurer. Equivalence of second order optimality conditions for bang-bang control problems. Part 1: Main results. *Control&Cybernetics*, **34** (2005), 927–950.

[17] N.P Osmolovskii and H Maurer. Equivalence of second order optimality conditions for bang-bang control problems. Part 2: Proofs, variational derivatives and representations. *Control&Cybernetics*, **36** (2007), 5–45.

[18] 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.

[19] L. S. Pontryagin, V. G. Boltyanskij, R. V. Gamkrelidze, E. F. Mishchenko, *The mathematical theory of optimal processes*, Fizmatgiz, Moscow, 1961 (Pergamon, Oxford, 1964).

[20] M. Quincampoix and V.M. Veliov. Metric regularity and stability of optimal control problems for linear systems. *SIAM J. Contr. Optim.*, **51**(5)(2013) 4118-4137.

[21] M. Seydenschwanz. Convergence results for the discrete regularization of linear-quadratic control problems with bang-bang solutions. *Comput. Optim. Appl.*, **61**(3) (2015) 731–760.

[22] V.M. Veliov. On the convexity of integrals of multivalued mappings: applications in control theory. *J. of Optimization Theory and Applications*, **54**(3) (1987), 541–563.

[23] V.M. Veliov. *Approximations of differential inclusions by discrete inclusions.* IIASA Working Paper WP-89-017, 1989.

[24] V.M. Veliov. Error analysis of discrete Approximation to bang-bang optimal control problems: the linear case. *Control&Cybernetics*, **34**(3) (2005), 967–982.