TECHNISCHE UNIVERSITÄT WIEN

Operations Research and Control Systems
SWM ORCOS

Metric sub-regularity in optimal control of affine problems with free end state

N.P. Osmolovskii, V.M. Veliov

Research Report 2019-04

March 2019

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

Metric sub-regularity in optimal control of affine problems with free end state*

N.P. Osmolovskii† V.M. Veliov‡

Abstract

The paper investigates the property of Strong Metric sub-Regularity (SMsR) of the mapping representing the first order optimality system for a Lagrange-type optimal control problem which is affine with respect to the control. The terminal time is fixed, the terminal state is free, and the control values are restricted in a convex compact set U. The SMsR property is associated with a reference solution of the optimality system and ensures that small additive perturbations in the system result in solutions which are at distance to the reference one, at most proportional to the size of the perturbations. A general sufficient condition for SMsR is obtained for appropriate space settings and then specialized in the case of a polyhedral set U and purely bang-bang reference control. Sufficient second-order optimality conditions are obtained as a by-product of the analysis. Finally, the obtained results are utilized for error analysis of the Euler discretization scheme applied to affine problems.

Key words: optimal control, affine control problems, bang-bang control, metric regularity, Pontryagin's maximum principle, second-order optimality conditions, Euler discretization

AMS subject classifications: 49K40, 49J53, 49J30, 49K15, 47J30

1 Introduction

The paper investigates the following Lagrange-type optimal control problem:

$$\min \Big\{ J(u) := \int_0^T g(t, x(t), u(t))\, \mathrm{d}t \Big\} \tag{1}$$

subject to

$$\dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x^0, \tag{2}$$

$$u(t) \in U, \quad t \in [0, T], \tag{3}$$

where the state x is a vector in $\mathbb{R}^n$, the control u has values $u(t)$ that belong to a given set U in $\mathbb{R}^m$ for almost every (a.e.) $t \in [0, T]$. The initial state x^0 and the final time $T > 0$ are fixed. The set of feasible control functions u, denoted in the sequel by $\mathcal{U}$, consists of all Lebesgue measurable

*This research is supported by the Austrian Science Foundation (FWF) under grant No P31400-N32. The second author also acknowledges the partial support of the Erwin Schrdinger International Institute (ESI), Vienna.

†Systems Research Institute, Polish Academy of Sciences, Warszawa, Poland, Nikolai.Osmolovskii@ibspan.waw.pl

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

and bounded functions $u : [0,T] \to U$. Accordingly, the state trajectories x, that are solutions of (2) for feasible controls, are Lipschitz continuous functions of time $t \in [0,T]$.

An important specific feature of the investigated problem is that it is assumed to be affine with respect to the control, that is, f and g have the following form:

$$f(t,x,u) = a(t,x) + B(t,x)u, \qquad g(t,x,u) = w(t,x) + \langle s(t,x), u\rangle, \tag{4}$$

with an $(n \times m)$-dimensional matrix function B, a scalar function w and an m-dimensional vector function s.

It is well known that the Pontryagin (local) maximum principle can be written in the form of a generalized equation

$$0 \in \mathcal{F}(y), \tag{5}$$

where $y = (x(\cdot), u(\cdot), p(\cdot))$ encapsulates the state function $x(\cdot)$, the control function $u(\cdot) \in \mathcal{U}$, and the adjoint (co-state) function $p(\cdot)$, and the inclusion $0 \in \mathcal{F}(y)$ represents the state equation, the co-state equation, the transversality condition, and the maximization condition in the maximum principle (the last being the inclusion of the derivative of the associated Hamiltonian with respect to the control in the normal cone to $\mathcal{U}$ at $u(\cdot)$). The detailed formulations will be given in the next two sections.

The main aim of the paper is to obtain sufficient conditions for Strong Metric sub-Regularity (SMsR) of the mapping $\mathcal{F}$. We remind this notion, following [6, p. 202]. Let $(\mathcal{Y}, d_{\mathcal{Y}})$ and $(\mathcal{Z}, d_{\mathcal{Z}})$ be two metric spaces. In any metric space, we denote by by $\mathbb{B}(q;\alpha)$ the ball with radius α centered at the point q.

Definition 1.1 *A set-valued mapping $\mathcal{F} : \mathcal{Y} \rightrightarrows \mathcal{Z}$ is* Strongly Metrically sub-Regular (SMsR) *at $\hat{y}$ for $\hat{z}$ if $\hat{z} \in \mathcal{F}(\hat{y})$ and there exist numbers $\alpha > 0$, $\beta > 0$ and c such that for any $z \in \mathbb{B}(\hat{z};\alpha)$ and for any solution $y \in \mathbb{B}(\hat{y};\beta)$ of the inclusion $z \in \mathcal{F}(y)$ it holds that $d_{\mathcal{Y}}(y,\hat{y}) \le c d_{\mathcal{Z}}(z,\hat{z})$.*

In the terms of the inverse mapping $\mathcal{F}^{-1}(z) := \{y \in \mathcal{Y} : z \in \mathcal{F}(y)\}$, the SMsR property reads as

$$\mathcal{F}^{-1}(z) \cap \mathbb{B}(\hat{y};\beta) \subset \mathbb{B}(\hat{y}; c d_{\mathcal{Z}}(z,\hat{z})) \quad \text{ for all } z \in \mathbb{B}(\hat{z};\alpha).$$

Notice that the mapping $z \rightrightarrows \mathcal{F}^{-1}(z) \cap \mathbb{B}(\hat{y};\beta)$ can be empty- or multi-valued, but its value at $\hat{z}$ is the singleton $\{\hat{y}\}$.

For the particular case of mapping $\mathcal{F}$ resulting from the optimality system for an optimal control problem as (1)–(3) there are various options for the choice of the spaces $\mathcal{Y}$ and $\mathcal{Z}$. For problems satisfying the so-called *coercivity condition*, introduced (to the best of our knowledge) in [10], a stronger property than SMsR has been proved in [5], where the L^∞-metric in $\mathcal{U}$ is used for the controls, and the metrics (norms) in the other components of $\mathcal{Y}$ and $\mathcal{Z}$ are defined correspondingly. However, this coercivity condition never holds for affine problems.

In order to cope with the regularity issue for affine problems, one has to use the L^1-metric in $\mathcal{U}$, and define appropriately the metric in the image space $\mathcal{Z}$ of the optimality mapping $\mathcal{F}$. This is done in Section 3, and a sufficient condition for SMsR is obtained in terms of positive definiteness of a linear-quadratic functional defined on the set of feasible variations of the control component, $\hat{u}$, of a reference solution $\hat{y}$ of inclusion (5). As shown in Section 2, this linear-quadratic functional represents the second-order variation of the objective functional, $J(u)$ in (1), which (despite of the non-differentiability in L^1) turns out to provide a second-order approximation of the cost $J(u)$ at a point $u \in \mathcal{U}$ in an L^1-neighborhood of $\hat{u}$. As a by-product, also in Section 2, we formulate simple second-order sufficient conditions for local minimum in the space L^1.

The obtained sufficient condition for SMsR is somewhat stronger than the second-order sufficient optimality condition in Section 2. In the same time, it is similar to (but weaker than) a condition introduced in [1] in the context of error estimates for the Euler discretization scheme. The condition is investigated in more details in Section 4, in the case of a polyhedral set U and purely bang-bang optimal control $\hat{u}$, where previous results from [2] and [13] are extended to the case of non-linear affine problems and general polyhedral sets U.

The SMsR of the optimality system is a key property for obtaining error estimates for discrete approximations to problem (1)–(3). In Section 5 we prove an error estimate of first order (with respect to the mesh size) for the Euler discretization. As explained in more details in Section 5, the result extends the ones in a sequence of previous publications (see [1] and the references therein).

2 Preliminary analysis

For the problem (1)–(3) with the affine specification (4) we make the following assumptions.

Assumption (A1). The set U is convex and compact, the functions $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ (having the form as in (4)) are two times differentiable in x, the second derivatives are continuous in x locally uniformly in t;[1] for every $x \in \mathbb{R}^n$ and $u \in U$ the functions f, g and their first and second derivatives in x are measurable and bounded in t.

Here and in the squeal, we use the following standard notations. The euclidean norm and the scalar product in $\mathbb{R}^n$ (the elements of which are regarded as column-vectors) are denoted by $|\cdot|$ and $\langle \cdot, \cdot \rangle$, respectively. The transpose of a matrix (or vector) E is denoted by $E^\top$. For a function $\psi : \mathbb{R}^p \to \mathbb{R}^r$ of the variable z we denote by $\psi_z(z)$ its derivative (Jacobian), represented by an $(r \times p)$-matrix. If $r = 1$, $\nabla_z \psi(z) = \psi_z(z)^\top$ denotes its gradient (a vector-column of dimension p). Also for $r = 1$, $\psi_{zz}(z)$ denotes the second derivative (Hessian), represented by a $(p \times p)$-matrix. For a function $\psi : \mathbb{R}^{p+q} \to \mathbb{R}$ of the variables (z, v), $\psi_{zv}(z, v)$ denotes its mixed second derivative, represented by a $(p \times q)$-matrix. The space $L^k = L^k([0, T], \mathbb{R}^r)$, with $k = 1, 2$ or $k = \infty$, consists of all (classes of equivalent) Lebesgue measurable r-dimensional vector-functions defined on the interval $[0, T]$, for which the standard norm $\|\cdot\|_k$ is finite. Often the specification $([0, T], \mathbb{R}^r)$ will be omitted in the notations. As usual, $W^{1,1} = W^{1,1}([0, T], \mathbb{R}^r)$ denotes the space of absolutely continuous functions $x : [0, T] \to \mathbb{R}^r$ for which the first derivative belongs to L^1. For convenience, the norm in $W^{1,1}$ is defined as $\|x\|_{1,1} := |x(0)| + \|\dot{x}\|_1$, so that $\|x\|_\infty \leq \|x\|_{1,1}$.

Define the Hamiltonian associated with problem (1)–(3) as usual:

$$H(t, x, u, p) := g(t, x, u) + \langle p, f(t, x, u) \rangle, \qquad p \in \mathbb{R}^n.$$

Although the feasible controls $u \in \mathcal{U}$ are bounded, we consider the control-trajectory pairs (x, u) as elements of the space $W^{1,1}([0, T], \mathbb{R}^n) \times L^1([0, T], \mathbb{R}^m)$.

The local form of the Pontryagin maximum (here minimum) principle for problem (1)-(3) can be represented by the following optimality system for (x, u) and an absolutely continuous (here

[1] Applied to f, for example, this means that for every bounded set $S \subset \mathbb{R}^n$ there exists a monotone increasing function (called *modulus of continuity*) $\omega : (0, +\infty) \to [0, +\infty)$ with $\lim_{s \to 0} \omega(s) = 0$, such that $|f(t, x', u) - f(t, x, u)| \leq \omega(|x' - x|)$ for every $t \in [0, T]$, $u \in U$ and x, $x' \in S$.

Lipschitz) function $p : [0,T] \to \mathbb{R}^n$: for a.e. $t \in [0,T]$

$$0 = -\dot{x}(t) + f(t, x(t), u(t)), \quad x(0) - x^0 = 0, \tag{6}$$

$$0 = \dot{p}(t) + \nabla_x H(t, x(t), u(t), p(t)), \tag{7}$$

$$0 = p(T), \tag{8}$$

$$0 \in \nabla_u H(t, x(t), u(t), p(t)) + N_U(u(t)), \tag{9}$$

where the normal cone $N_U(u)$ to the set U at $u \in \mathbb{R}^m$ is defined as

$$N_U(u) = \begin{cases} \{y \in \mathbb{R}^n \mid \langle y, v - u\rangle \le 0 \text{ for all } v \in U\} & \text{if } u \in U, \\ \emptyset & \text{otherwise.} \end{cases}$$

Let a reference solution $\hat{y} = (\hat{x}, \hat{u}, \hat{p}) \in W^{1,1} \times \mathcal{U} \times W^{1,1}$ of the optimality system (6)–(9) be fixed. To shorten the notations we skip arguments with "hat" in functions, shifting the "hat" on the top of the notation of the function, so that $\hat{f}(t) := f(t, \hat{x}(t), \hat{u}(t))$, $\hat{s}(t) = s(t, \hat{x}(t))$, $\hat{H}(t) := H(t, \hat{x}(t), \hat{u}(t), \hat{p}(t))$, $\hat{H}(t,u) := H(t, \hat{x}(t), u, \hat{p}(t))$, etc. Moreover, denote

$$\hat{A}(t) := f_x(t, \hat{x}(t), \hat{u}(t)), \quad \hat{B}(t) := f_u(t, \hat{x}(t), \hat{u}(t)) = B(t, \hat{x}(t)), \quad \hat{\sigma}(t) := \nabla_u \hat{H}(t) = \hat{B}(t)^\top \hat{p}(t) + \hat{s}(t).$$

Remark 2.1 Due to Assumption (A1), and since the solution $\hat{x}$ of (2) with $u = \hat{u}$ exists on $[0,T]$, there exist a number $r > 0$ and a convex compact set $\bar{S} \subset \mathbb{R}^n$ such that for every $u \in \mathcal{U}$ with $\|u - \hat{u}\|_1 \le r$ the solution x of (2) exists on $[0,T]$ and $\mathbb{B}(x(t); 1) \subset \bar{S}$ for all $t \in [0,T]$. By taking $\bar{S}$ sufficiently large we may also ensure that $\mathbb{B}(\hat{p}(t); 1) \subset \bar{S}$ for all $t \in [0,T]$. Using Assumption (A1), we denote by L a Lipschitz constant with respect to $x \in \bar{S}$ (uniformly with respect to $t \in [0,T]$, $u \in U$, $p \in \bar{S}$) of the functions f, g and H, their first derivatives in x, and H_{ux}, H_{up}. Further, we denote by M a bound of the functions f, f_u, f_x, f_{ux}, H_x, H_{xx}, H_{xu}, H_{xxu} and H_{xxp} for $(t, x, u, p) \in [0,T] \times \bar{S} \times U \times \bar{S}$. Finally, we denote by $\bar{\omega}$ the modulus of continuity of H_{xx}, uniformly with respect to $(t, u, p) \in [0,T] \times U \times \bar{S}$ (see Footnote 1).

According to this remark, for any $u \in \mathcal{U}$ with $\|u - \hat{u}\|_1 \le r$ the value of the objective functional

$$J(u) := \int_0^T g(t, x(t), u(t))\, \mathrm{d}t$$

is well defined. In the next proposition we obtain a sort of second order expansion of J in an L^1-neighborhood of $\hat{u}$ in the set $\mathcal{U}$ (although the functional J is, in general, not Fréchet directionally differentiable at $\hat{u}$).

For any measurable function $\delta u(t) \in U - \hat{u}(t)$ a.e. in $[0,T]$, we introduce the linearized version of equation (2):

$$\dot{\delta x}(t) = \hat{A}(t)\delta x(t) + \hat{B}(t)\delta u(t), \quad \delta x(0) = 0, \quad t \in [0,T]. \tag{10}$$

Adapting the usual definition to the affine case (see e.g. [11]) we introduce the following quadratic functional of $(\delta x, \delta u) \in W^{1,1} \times L^1$:

$$\Omega(\delta x, \delta u) := \int_0^T \left[\frac{1}{2} \langle \hat{H}_{xx}(t)\delta x(t), \delta x(t)\rangle + \langle \hat{H}_{ux}(t)\delta x(t), \delta u(t)\rangle \right] \mathrm{d}t. \tag{11}$$

Proposition 2.1 *Let Assumption (A1) be fulfilled. Then there exist constants $\bar{c}$ and c_f such that for every $u \in \mathcal{U}$ for which $\delta u = u - \hat{u}$ satisfies $\|\delta u\|_1 \le r$ (see Remark 2.1 for r) the solution x of (2) exists on $[0,T]$ and the following representation holds:*

$$J(u) = J(\hat{u}) + \int_0^T \langle \hat{\sigma}(t), \delta u(t)\rangle \,\mathrm{d}t + \Omega(\delta x, \delta u) + \hat{\gamma}(\delta u), \tag{12}$$

where δx is the solution of (10) and the number $\hat{\gamma}(\delta u)$ satisfies

$$|\hat{\gamma}(\delta u)| \le \bar{c}\,[\bar{\omega}(c_f\|\delta u\|_1) + \|\delta u\|_1]\;\|\delta u\|_1^2. \tag{13}$$

The numbers $\bar{c}$ and c_f depend on the reference solution $\hat{y} = (\hat{x}, \hat{u}, \hat{p})$ of the optimality system (6)-(9) and on the data of the problem (1)–(3), f, g, x^0, T, U, only through the constants L and M, the time horizon, T, and the modulus $\bar{\omega}$ (see Remark 2.1). Moreover, the fact that $\hat{y}$ satisfies inclusion (9) is not needed for the representation (12) with (13).

Proof. Let a triple $(\hat{x}, \hat{u}, \hat{p}) \in W^{1,1} \times \mathcal{U} \times W^{1,1}$ satisfy (6)-(8). Take an arbitrary function $u \in \mathcal{U}$ such that $\delta u(t) = u(t) - \hat{u}(t)$ satisfies $\|\delta u\|_1 \le r$ and let x be the solution of (2) on $[0,T]$. According to Remark 2.1 this solution exists and $x(t) \in \bar{S}$ for every $t \in [0,T]$.

Due to the relations

$$f(t,x(t),u(t)) - \hat{f}(t) = f(t,x(t),u(t)) - f(t,\hat{x}(t),u(t)) + \langle \hat{B}(t), u(t) - \hat{u}(t)\rangle,$$

$$|f(t,x(t),u(t)) - f(t,\hat{x}(t),u(t))| \le L|x(t) - \hat{x}(t)|, \quad \left|\langle \hat{B}(t), u(t) - \hat{u}(t)\rangle\right| \le M|u(t) - \hat{u}(t)|,$$

the Grönwall inequality implies the estimation

$$\|x - \hat{x}\|_C \le c_f \|u - \hat{u}\|_1 \quad \text{with} \;\; c_f = Me^{LT}. \tag{14}$$

Setting $\Delta x = x - \hat{x}$, using the definition of H, equations (6)-(8) and integrating by parts, we obtain the identity

$$\begin{aligned}
J(u) - J(\hat{u}) &= \int_0^T [g(t,x(t),u(t)) - \hat{g}(t)]\,\mathrm{d}t \\
&= \int_0^T [H(t,x(t),u(t),\hat{p}(t)) - \hat{H}(t)]\,\mathrm{d}t - \int_0^T \langle \hat{p}(t), f(t,x(t),u(t)) - \hat{f}(t)\rangle \,\mathrm{d}t \\
&= \int_0^T [H(t,x(t),u(t),\hat{p}(t)) - \hat{H}(t)]\,\mathrm{d}t - \int_0^T \langle \hat{p}(t), \dot{x}(t) - \dot{\hat{x}}(t)]\,\mathrm{d}t \\
&= \int_0^T [H(t,x(t),u(t),\hat{p}(t)) - \hat{H}(t)]\,\mathrm{d}t + \int_0^T \langle \dot{\hat{p}}(t), \Delta x(t)\rangle \,\mathrm{d}t \\
&= \int_0^T [H(t,x(t),u(t),\hat{p}(t)) - \hat{H}(t)]\,\mathrm{d}t - \int_0^T \langle \nabla_x \hat{H}(t), \Delta x(t)\rangle \,\mathrm{d}t. \qquad (15)
\end{aligned}$$

Using the Taylor formula and the equality $H_{uu} = 0$, we obtain that

$$\begin{aligned}
(16)\; & H(t,x(t),u(t),\hat{p}(t)) - \hat{H}(t) = \langle \nabla_x \hat{H}(t), \Delta x(t)\rangle + \langle \nabla_u \hat{H}(t), \delta u(t)\rangle \\
& + \frac{1}{2}\langle \hat{H}_{xx}(t)\Delta x(t), \Delta x(t)\rangle + \langle \hat{H}_{ux}(t)\Delta x(t), \delta u(t)\rangle + \frac{1}{2}\langle \tilde{W}(t)\Delta x(t), \Delta x(t)\rangle + \langle \tilde{S}(t)\Delta x(t), \delta u(t)\rangle,
\end{aligned}$$

where

$$|\tilde{W}(t)| \le \sup_{\tilde{x},\tilde{u}} |H_{xx}(t,\tilde{x},\tilde{u},\hat{p}(t)) - \hat{H}_{xx}(t)|, \qquad |\tilde{S}(t)| \le \sup_{\tilde{x},\tilde{u}} |H_{ux}(t,\tilde{x},\tilde{u},\hat{p}(t)) - \hat{H}_{ux}(t)|$$

and the supremum is over $\tilde{x} \in [x(t), \hat{x}(t)]$ and $\tilde{u} \in [u(t), \hat{u}(t)]$. Having in mind that H_{ux} is independent of u, Remark 2.1 and (14), and also using the equality

$$H_{xx}(t,\tilde{x},\tilde{u},\hat{p}(t)) - \hat{H}_{xx} = H_{xx}(t,\tilde{x},\tilde{u},\hat{p}(t)) - H_{xx}(t,\hat{x},\tilde{u},\hat{p}(t)) + H_{xx}(t,\hat{x},\tilde{u},\hat{p}(t)) - \hat{H}_{xx},$$

we obtain the estimations

$$\begin{aligned} |\tilde{W}(t)| &\le \bar{\omega}(|\Delta x(t)|) + M|\delta u(t)| \le \bar{\omega}(c_f\|\delta u\|_1) + M|\delta u(t)|, \\ |\tilde{S}(t)| &\le L|\Delta x(t)| \le L c_f \|\delta u\|_1. \end{aligned}$$

Combining (15), (16) and the last inequalities, we obtain that

$$J(u) - J(\hat{u}) = \int_0^T \hat{\sigma}(t)\delta u(t)\,\mathrm{d}t + \Omega(\Delta x, \delta u) + r_1(\delta u), \tag{17}$$

where

$$|r_1(\delta u)| \le \frac{c_f}{2}\Big(T c_f \bar{\omega}(c_f\|\delta u\|_1) + (c_f M + 2L)\|\delta u\|_1\Big)\|\delta u\|_1^2. \tag{18}$$

Let δx be the solution of the linear equation (10). Now we replace $(\Delta x, \delta u)$ with $(\delta x, \delta u)$ in the quadratic form Ω in (17). We have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(\Delta x(t) - \delta x(t)) &= f(t,x(t),u(t)) - f(t,\hat{x}(t),\hat{u}(t)) - \hat{A}(t)\delta x(t) - \hat{B}(t)\delta u(t) \\ &= \hat{f}_x(t)\Delta x(t) + \hat{f}_u(t)\delta u(t) + r_f(t) - \hat{A}(t)\delta x(t) - \hat{B}(t)\delta u(t) \\ &= \hat{A}(t)(\Delta x(t) - \delta x(t)) + r_f(t), \end{aligned}$$

where, due to the linearity of f in u,

$$r_f(t) = f(t,x(t),u(t)) - \hat{f}(t) - \hat{f}_x(t)\Delta x(t) - \hat{f}_u(t)\delta u(t)$$

$$= f(t,x(t),u(t)) - f(t,\hat{x}(t),u(t)) + f(t,\hat{x}(t),u(t)) - \hat{f}(t) - \hat{f}_x(t)\Delta x(t) - \hat{f}_u(t)\delta u(t)$$

$$= f(t,x(t),u(t)) - f(t,\hat{x}(t),u(t)) - \hat{f}_x(t)\Delta x(t) = (f_x(t,\hat{x}(t),u(t)) - \hat{f}_x(t))\Delta x(t) + \frac{1}{2}\zeta_f(t)|\Delta x(t)|^2$$

with $\|\zeta_f(t)\|_\infty \le L$, $|r_f(t)| \le M|\delta u(t)||\Delta x(t)| + \frac{1}{2}L|\Delta x(t)|^2$. Using the Grönwall inequality we obtain that

$$\begin{aligned} \|\Delta x - \delta x\|_C \le e^{MT}\|r_f\|_1 &\le e^{MT}\left(M\|\delta u\|_1\|\Delta x\|_C + \frac{1}{2}LT\|\Delta x\|_C^2\right) \\ &\le \frac{1}{2}e^{MT}c_f(2M + TLc_f)\,\|\delta u\|_1^2 =: d\|\delta u\|_1^2. \end{aligned} \tag{19}$$

Then we can estimate the difference

$$r_{\Delta\Omega}(\delta u) := \Omega(\Delta x, \delta u) - \Omega(\delta x, \delta u)$$

as follows:

$$|r_{\Delta\Omega}(\delta u)| \;\leq\; \frac{1}{2}\int_0^T \left[|\hat{H}_{xx}(t)|\,|\Delta x(t)+\delta x(t)|\,|\Delta x(t)-\delta x(t)| + 2|\hat{H}_{ux}(t)|\,|\Delta x(t)-\delta x(t)|\,|\delta u(t)|\right] \mathrm{d}t$$

$$\leq\; \frac{1}{2}MT\|\Delta x+\delta x\|_C\,\|\Delta x-\delta x\|_C + M\|\Delta x-\delta x\|_C\,\|\delta u\|_1. \tag{20}$$

In view of (10) we have $\|\delta x\|_C \leq e^{MT}M\|\delta u\|_1 =: d_1\|\delta u\|_1$. Using this estimate together with $\|\Delta x\|_C \leq c_f\|\delta u\|_1$ and (19) we obtain from (20) that

$$|r_{\Delta\Omega}(\delta u)| \leq \frac{1}{2}MT(c_f+d_1)d\|\delta u\|_1^3 + Md\|\delta u\|_1^3 =: d_2\|\delta u\|_1^3. \tag{21}$$

Combining this estimate with (17) we obtain the first claim of the proposition with

$$\hat{\gamma}(\delta u) = r_1(\delta u) + r_{\Delta\Omega}(\delta u).$$

Estimation (13) follows from this equality, (18) and (21) with

$$\bar{c} = \frac{1}{2}\max\left\{c_f^2T,\; Mc_f^2 + 2Lc_f + 2d_2\right\}.$$

The second claim of the proposition follows from the definition of the constants r, c_f, d, d_1 and d_2 above. In the proof we have assumed that $(\hat{x},\hat{u},\hat{p})$ satisfies (6)-(8) only, hence the last claim of the proposition. Q.E.D.

Notice that Ω in (11) is the usual "second variation" of the objective functional, adapted to the affine case. Here we point out another feature which is specific for affine problems. For a general (not necessarily affine) optimal control problem, the right-hand side of the estimate (13) contains the term $\|\delta u\|_2$ (possibly even $\|\delta u\|_\infty$, if the dependence of f and g on u is not linear-quadratic) instead of $\|\delta u\|_1$, which makes a substantial difference. The terms $\|\delta u\|_2$ and $\|\delta u\|_\infty$ do not appear in (13) because in the affine case the derivatives f_{uu} and H_{uu} vanish.

Remark 2.2 It will be useful to observe that in the case of a linear function f (that is, linear function a and independent of x function B) and a function g which is quadratic in x and bilinear in (x,u), the number $\hat{\gamma}$ in Proposition 2.1 equals zero. Indeed, as seen in the proof, in this case $r_1(\delta u)=0$ and $\Delta x=\delta x$, hence $r_{\Delta\Omega}(\delta u)=0$.

Denote by Γ the set of all pairs $(\delta x,\delta u)\in W^{1,1}\times L^1$ such that $\delta u(t)\in U-\hat{u}(t)$ a.e. in $[0,T]$ and δx is the solution of the linearized equation (10). Proposition 2.1 invokes the following assumption, which turns out to be a sufficient optimality condition for $(\hat{x},\hat{u})$.

Assumption (A2). There exists a constant $c_0>0$ such that

$$\int_0^T \langle\hat{\sigma}(t),\delta u(t)\rangle\,\mathrm{d}t + \Omega(\delta x,\delta u) \geq c_0\|\delta u\|_1^2 \quad \text{for all } (\delta x,\delta u)\in\Gamma.$$

Corollary 2.1 *Let Assumption (A1) be fulfilled and let $\hat{y}=(\hat{x},\hat{u},\hat{p})\in W^{1,1}\times\mathcal{U}\times W^{1,1}$ be a solution of the part (6)–(8) of the optimality system (6)–(9). Let, in addition, Assumption (A2) be fulfilled. Then $(\hat{x},\hat{u})$ is a strict strong local solution of problem (1)–(3). Consequently, inclusion (9) is also satisfied.*

We clarify that "strict strong local solution" has the following meaning: there is a number $\varepsilon > 0$ such that for every $u \in \mathcal{U}$ with $\|u - \hat{u}\|_1 \leq \varepsilon$ and $u \neq \hat{u}$ (in the sense of L^1) it holds that if the corresponding solution, x, of (2) exists on $[0, T]$, then $J(u) > J(\hat{u})$.

Proof of Corollary 2.1. Take $\varepsilon > 0$ so small that $\varepsilon \leq r$ (see Remark 2.1 for the number r) and

$$|\hat{\gamma}(\delta u)| \leq \frac{c_0}{2}\|\delta u\|_1^2$$

whenever $u \in \mathcal{U}$ and $\delta u = u - \hat{u}$ satisfies $\|\delta u\|_1 \leq \varepsilon$ (see (13)). Then the solution x exists on $[0, T]$ and, according to Proposition 2.1 and Assumption (A2), we have

$$\begin{aligned} J(u) - J(\hat{u}) &= \int_0^T \langle \hat{\sigma}(t), \delta u(t) \rangle \, \mathrm{d}t + \Omega(\delta x, \delta u) + \hat{\gamma}(\delta u) \geq c_0 \|\delta u\|_1^2 - |\hat{\gamma}(\delta u)| \\ &\geq \frac{c_0}{2}\|\delta u\|_1^2 > 0. \end{aligned}$$

Q.E.D.

Even more than claimed by Corollary 2.1, the last inequality in the proof shows *quadratic growth* of the cost functional J at the reference point $\hat{u}$ in the L^1-norm.

3 Strong metric sub-regularity of the optimality mapping

We begin this section with an abstract result about "stability" of the SMsR property, proved in the recent paper [4]. Let $(\mathcal{Y}, d_{\mathcal{Y}})$ and $(\mathcal{Z}, d_{\mathcal{Z}})$ be two metric spaces and let $\mathcal{F} : \mathcal{Y} \rightrightarrows \mathcal{Z}$ be a set-valued mapping. The numbers (α, β, c) in Definition 1.1 will be referred to as parameters of SMsR.

Proposition 3.1 *Let $\mathcal{Z}$ be a linear space and the metric $d_{\mathcal{Z}}$ be shift-invariant, that is $d_{\mathcal{Z}}(z, z') = d_{\mathcal{Z}}(z - z', 0)$. Let $\mathcal{F} : \mathcal{Y} \rightrightarrows \mathcal{Z}$ be SMsR at $\hat{y}$ for $\hat{z}$ with parameters (α, β, c). Let $\alpha' > 0$, $\beta' > 0$, c' and λ be any numbers such that*

$$\lambda c < 1, \qquad \beta' \leq \beta, \qquad \alpha' + \lambda\beta' \leq \alpha, \qquad c' \geq \frac{c}{1 - \lambda c}. \tag{22}$$

Then for any function $\varphi : \mathcal{Y} \to \mathcal{Z}$ that satisfies the conditions

$$\varphi(\hat{y}) = 0, \qquad d_{\mathcal{Z}}(\varphi(y), \varphi(\hat{y})) \leq \lambda d_{\mathcal{Y}}(y, \hat{y}) \quad \text{for every } \ y \in \mathbb{B}(\hat{y}; \beta'), \tag{23}$$

the mapping $\varphi + \mathcal{F}$ is SMsR at $\hat{y}$ for $\hat{z}$ with parameters (α', β', c').

We adapt the simple proof from [4], in order to explicitly formulate the conditions for the parameters α', β', c' and λ, which will be needed later.

Proof. Let the numbers $\alpha' > 0$, $\beta' > 0$, c' and λ, and the function φ be as in the formulation of the proposition. Obviously $\hat{z} \in \varphi(\hat{y}) + \mathcal{F}(\hat{y})$. Take an arbitrary $z \in \mathbb{B}(\hat{z}; \alpha')$ and let $y \in \mathbb{B}(\hat{y}; \beta')$ satisfy $z \in \varphi(y) + \mathcal{F}(y)$. Then

$$d_{\mathcal{Z}}(z - \varphi(y), \hat{z}) \leq d_{\mathcal{Z}}(z, \hat{z}) + d_{\mathcal{Z}}(\varphi(y), 0) \leq \alpha' + \lambda d_{\mathcal{Y}}(y, \hat{y}) \leq \alpha' + \lambda\beta' \leq \alpha,$$

and

$$d_{\mathcal{Y}}(y, \hat{y}) \leq \beta' \leq \beta.$$

Hence, using the SMsR property of $\mathcal{F}$ and conditions (23) for φ we obtain that

$$d_{\mathcal{Y}}(y,\hat{y}) \le c d_{\mathcal{Z}}(z-\varphi(y),\hat{z}) \le c d_{\mathcal{Z}}(z,\hat{z}) + c\lambda d_{\mathcal{Y}}(y,\hat{y})$$

which, in view of the first and the last inequalities in (22), implies $d_{\mathcal{Y}}(y,\hat{y}) \le c' d_{\mathcal{Z}}(z,\hat{z})$. Q.E.D.

Now we return to the affine problem (1)–(3), for which we specify the spaces

$$\mathcal{Y} := W_0^{1,1} \times \mathcal{U} \times W^{1,1}, \qquad \mathcal{Z} := L^1 \times L^1 \times \mathbb{R}^n \times L^\infty,$$

where $W_0^{1,1}$ is the affine space consisting of those $x \in W^{1,1}$ for which $x(0) = x^0$, and $\mathcal{U}$ is endowed with the L^1-metric (thanks to the compactness of U, the metric space $\mathcal{U}$ is complete). Correspondingly, the shift-invariant metrics in these spaces are defined as follows: for $y = (x,u,p) \in \mathcal{Y}$ and $z = (\xi,\pi,\nu,\rho) \in \mathcal{Z}$

$$d_{\mathcal{Y}}(y) := d_{\mathcal{Y}}(y,0) = \|x\|_{1,1} + \|u\|_1 + \|p\|_{1,1}, \qquad d_{\mathcal{Z}}(z) := d_{\mathcal{Z}}(z,0) = \|\xi\|_1 + \|\ \pi\|_1 + |\nu| + \|\rho\|_\infty.$$

The optimality system (6)–(9) can be recast as the generalized equation

$$0 \in \psi(y) + \Psi(y), \tag{24}$$

where $y = (x,u,p)$ and

$$\mathcal{Y} \ni y \mapsto \psi(y) := \begin{pmatrix} -\dot{x} + f(\cdot,x,u) \\ \dot{p} + \nabla_x H(\cdot,y) \\ p(T) \\ \nabla_u H(\cdot,y) \end{pmatrix} \in \mathcal{Z}, \qquad \mathcal{Y} \ni y \Rightarrow \Psi(y) := \begin{pmatrix} 0 \\ 0 \\ 0 \\ N_{\mathcal{U}}(u) \end{pmatrix} \subset \mathcal{Z}. \tag{25}$$

Here

$$N_{\mathcal{U}}(u) := \{v \in L^\infty :\ v(t) \in N_U(u(t)) \text{ for a.e. } t \in [0,T]\}$$

is the normal cone to the set $\mathcal{U}$ (considered as a subset of L^1) at $u \in \mathcal{U}$. For $u \notin \mathcal{U}$ the normal cone is empty.

We will prove the SMsR property of the mapping $\psi + \Psi$ in the optimality system (24) under a somewhat stronger requirement than Assumption (A2). Let $\hat{y} = (\hat{x},\hat{u},\hat{p}) \in \mathcal{Y}$ satisfy (24).

Assumption (A2'). There exists a constant $c_0 > 0$ such that

$$\int_0^T \langle \hat{\sigma}(t), \delta u(t)\rangle \,\mathrm{d}t + 2\,\Omega(\delta x, \delta u) \ge c_0 \|\delta u\|_1^2 \quad \text{for all } (\delta x, \delta u) \in \Gamma.$$

The difference with (A2) is the multiplier 2 of Ω. This assumption is stronger than (A2) because in view of (9), $\langle \hat{\sigma}(t)\delta u(t)\rangle \ge 0$ a.e. in $[0,T]$ for all $\delta u \in \mathcal{U} - \hat{u}$. Therefore (A2') implies:

$$2\int_0^T \langle \hat{\sigma}(t), \delta u(t)\rangle \,\mathrm{d}t + 2\,\Omega(\delta x, \delta u) \ge c_0 \|\delta u\|_1^2 \quad \text{for all } (\delta x, \delta u) \in \Gamma,$$

that is (A2) holds with the constant $c_0/2$ instead of c_0. Of course, (A2') and (A2) are equivalent if Ω is non-negative on Γ, but not in general.

Theorem 3.1 *Let assumptions (A1) and (A2') be fulfilled. Then the optimality mapping $\psi+\Psi$, associated with problem (1)-(3), is strongly metrically sub-regular at $\hat{y}=(\hat{x},\hat{u},\hat{p})$ for zero. Moreover, the parameters of SMsR can be chosen as depending on the data of the problem (1)–(3) only through the constants L, M and T, the modulus $\bar{\omega}$ (see Remark 2.1), and the constant c_0 in Assumption (A2').*

In the proof we will use Proposition 3.1, for which we need some preparation. Define the following linearized version of the mapping $\psi+\Psi$ (along the reference point $\hat{y}$): for $y\in\mathcal{Y}$

$$\mathcal{F}(y):=\begin{pmatrix} -\dot{x}+\hat{f}(\cdot)+\hat{A}(\cdot)(x-\hat{x})+\hat{B}(\cdot)(u-\hat{u}) \\ \dot{p}+\nabla_x\hat{H}(\cdot)+\hat{H}_{xy}(\cdot)(y-\hat{y}) \\ p(T) \\ \nabla_u\hat{H}(\cdot)+\hat{H}_{uy}(\cdot)(y-\hat{y})+N_{\mathcal{U}}(u) \end{pmatrix}\subset\mathcal{Z}, \tag{26}$$

Then we can represent $\psi+\Psi=\varphi+\mathcal{F}$, where

$$\varphi(y):=\begin{pmatrix}\varphi_1(y)\\ \varphi_2(y)\\ \varphi_3(y)\\ \varphi_4(y)\end{pmatrix}=\begin{pmatrix} f(\cdot,x,u)-\hat{f}(\cdot)-\hat{A}(\cdot)(x-\hat{x})-\hat{B}(\cdot)(u-\hat{u}) \\ \nabla_x H(\cdot,y)-\nabla_x\hat{H}(\cdot)-\hat{H}_{xy}(\cdot)(y-\hat{y}) \\ 0 \\ \nabla_u H(\cdot,y)-\nabla_u\hat{H}(\cdot)-\hat{H}_{uy}(\cdot)(y-\hat{y}) \end{pmatrix}\in\mathcal{Z},$$

Notice that $\hat{H}_{xp}=\hat{A}^\top$ and $\hat{H}_{uy}=(\hat{H}_{ux},0,\hat{H}_{up})=(\hat{H}_{ux},0,\hat{B}^\top)$. Now we can take advantage of Proposition 3.1.

Proposition 3.2 *Let Assumption (A1) be fulfilled. Then the optimality mapping $\psi+\Psi$ is SMsR at $\hat{y}$ for zero with parameters depending only on L, M, T and the modulus $\bar{\omega}$ (see Remark 2.1) if and only if the same is true for the mapping $\mathcal{F}$ in (26).*

Proof. Since $\psi+\Psi=\varphi+\mathcal{F}$, it suffices to establish a connection between the properties of SMsR for $\varphi+\mathcal{F}$ and for $\mathcal{F}$ using Proposition 3.1. Let $\mathcal{F}$ be SMsR at $\hat{y}$ for zero with parameters (α,β,c) depending only on L,M,T, and $\bar{\omega}$. Let the numbers α',β',c' and λ satisfy the inequalities (22), and in addition,

$$\beta'\le r,\quad 3L\beta'\Big(1+\frac{T}{2}\Big)\le\lambda,\quad 3\bar{\omega}(\beta')T+\frac{3}{2}M\beta'(3+T)+\frac{3}{2}L\beta'(1+T)\le\lambda,\quad 3(2L+M)\beta'\le\lambda. \tag{27}$$

Obviously this can be done in such a way that these numbers depend on L,M,T, and $\bar{\omega}$ only. Since $\varphi(\hat{y})=0$, in order to apply Proposition 3.1 we have to estimate $d_{\mathcal{Z}}(\varphi(y))$ by $d_{\mathcal{Y}}(y,\hat{y})$, where $d_{\mathcal{Y}}(y,\hat{y})\le\beta'$. In the estimations below we use, in addition to Remark 2.1, just the first order Taylor formula and the linearity with respect to u and p. We denote $\Delta x=x-\hat{x}$, $\Delta u=u-\hat{u}$, $\Delta p=p-\hat{p}$, which implies $\|\Delta x\|_{1,1}+\|\Delta u\|_1+\|\Delta p\|_{1,1}=d_{\mathcal{Y}}(y,\hat{y})$. Using the second inequality in (27) and skipping the dummy argument t of integration we obtain

$$\begin{aligned}\|\varphi_1(y)-\varphi_1(\hat{y})\|_1 &\le L\int_0^T\Big[|\Delta x||\Delta u|+\frac{1}{2}|\Delta x|^2\Big]\,\mathrm{d}t\le L\Big[\|\Delta u\|_1+\frac{1}{2}\|\Delta x\|_1\Big]\|\Delta x\|_\infty\\ &\le L\Big(\beta'+\frac{T}{2}\beta'\Big)\|\Delta x\|_{1,1}\le\frac{1}{3}\lambda\, d_{\mathcal{Y}}(y,\hat{y}).\end{aligned}$$

In the next estimation we denote $\Delta y = (\Delta x, \Delta u, \Delta p)$ and use the straightforward equality

$$\begin{aligned}
\nabla_x H(t, y(t)) &= \nabla_x \hat{H}(t) + \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} \nabla_x H(t, \hat{y}(t) + s\Delta y(t))\, \mathrm{d}s \\
&= \nabla_x \hat{H}(t) + \int_0^1 H_{xy}(t, \hat{y}(t) + s\Delta y(t)) \Delta y(t)\, \mathrm{d}s \\
&= \nabla_x \hat{H}(t) + \hat{H}_{xy}(t)\Delta y(t) + \int_0^1 \left[H_{xy}(t, \hat{y}(t) + s\Delta y(t)) - \hat{H}_{xy}(t) \right] \Delta y(t)\, \mathrm{d}s.
\end{aligned}$$

Moreover, since H is linear in u and p, we have

$$\begin{aligned}
& |H_{xx}(t, \hat{y}(t) + s\Delta y(t)) - \hat{H}_{xx}(t)| \\
\leq\ & |H_{xx}(t, \hat{y}(t) + s\Delta y(t)) - H_{xx}(t, \hat{x}(t), \hat{u}(t) + s\Delta u(t), \hat{p}(t) + s\Delta p(t))| \\
+\ & |H_{xx}(t, \hat{x}(t), \hat{u}(t) + s\Delta u(t), \hat{p}(t) + s\Delta p(t)) - \hat{H}_{xx}(t)| \\
\leq\ & \bar{\omega}(\|\Delta x\|_\infty) + Ms|\Delta u(t)| + Ms|\Delta p(t)| \quad \text{a.e. in} \quad (0,T), \\
& |H_{xu}(t, \hat{y}(t) + s\Delta y(t)) - \hat{H}_{xu}(t)| = |H_{xu}(t, \hat{x}(t) + s\Delta x(t), \hat{u}(t), \hat{p}(t) + s\Delta p(t)) - \hat{H}_{xu}| \\
\leq\ & |H_{xu}(t, \hat{x}(t) + s\Delta x(t), \hat{u}(t), \hat{p}(t) + s\Delta p(t)) - H_{xu}(t, \hat{x}(t), \hat{u}(t), \hat{p}(t) + s\Delta p(t))| \\
+\ & |H_{xu}(t, \hat{x}(t), \hat{u}(t), \hat{p}(t) + s\Delta p(t)) - \hat{H}_{xu}(t)| \leq Ls|\Delta x(t)| + Ms|\Delta p(t)| \quad \text{a.e. in} \quad (0,T),
\end{aligned}$$

and similarly,

$$|H_{xp}(t, \hat{y}(t) + s\Delta y(t)) - \hat{H}_{xp}(t)| \leq Ls|\Delta x(t)| + Ms|\Delta u(t)| \quad \text{a.e. in} \quad (0,T).$$

Then, using the above estimations, the third inequality in (27), and the inequality $d_{\mathcal{Y}}(y, \hat{y}) \leq \beta'$, we obtain that

$$\begin{aligned}
\|\varphi_2(y) - \varphi_2(\hat{y})\|_1 &\leq \int_0^T \int_0^1 \left| \left[H_{xy}(t, \hat{y}(t) + s\Delta y(t)) - \hat{H}_{xy}(t) \right] \Delta y(t) \right| \mathrm{d}s\, \mathrm{d}t \\
&\leq \int_0^T \left[\left(\bar{\omega}(\|\Delta x\|_\infty) + \frac{M}{2}|\Delta u| + \frac{M}{2}|\Delta p| \right) |\Delta x| \right. \\
&\quad \left. + \left(\frac{L}{2}|\Delta x| + \frac{M}{2}|\Delta p| \right) |\Delta u| + \left(\frac{L}{2}|\Delta x| + \frac{M}{2}|\Delta u| \right) |\Delta p| \right] \mathrm{d}t \\
&\leq \left(\bar{\omega}(\beta')T\|\Delta x\|_{1,1} + \frac{M}{2}\beta'\|\Delta u\|_1 + \frac{M}{2}T\beta'\|\Delta x\|_{1,1} \right) + M\beta'\|\Delta u\|_1 \\
&\quad + \left(\frac{L}{2}\beta'\|\Delta u\|_1 + \frac{L}{2}\beta' T\|\Delta p\|_{1,1} \right) \\
&\leq \left[\bar{\omega}(\beta')T + \frac{1}{2}M\beta'(3+T) + \frac{1}{2}L\beta'(1+T) \right] d_{\mathcal{Y}}(y, \hat{y}) \leq \frac{1}{3}\lambda\, d_{\mathcal{Y}}(y, \hat{y}).
\end{aligned}$$

Similarly (and shorter), using the linearity of H in u and p, and the last inequality in (27), we obtain that

$$\|\varphi_4(y) - \varphi_4(\hat{y})\|_\infty \leq \frac{1}{3}\,\lambda d_{\mathcal{Y}}(y, \hat{y}).$$

Hence, $d_{\mathcal{Z}}(\varphi(y), \varphi(\hat{y})) \leq \lambda\, d_{\mathcal{Y}}(y, \hat{y})$. Then the SMsR property of $\mathcal{F}$ follows from Proposition 3.1.

To prove the converse claim we just exchange the places of the non-linear and the linearized problem, so that now $\mathcal{F} = \psi + \Psi$ and the linearized mapping in (26) takes the form $\mathcal{F} - \varphi$, then apply again Proposition 3.1 Q.E.D.

It is worth mentioning (although it is not needed for the subsequent analysis) that the mapping $\mathcal{F}$ in (26) is just the optimality mapping for the following linear-quadratic affine problem (we skip the argument t of x, $\hat{x}$, u and $\hat{u}$):

$$\min \int_0^T \left[\frac{1}{2}\langle \hat{H}_{xx}(t)(x-\hat{x}), x-\hat{x}\rangle + \langle \hat{H}_{ux}(t)(x-\hat{x}), u-\hat{u}\rangle + \langle \nabla_u \hat{H}(t), u-\hat{u}\rangle\right] \mathrm{d}t \tag{28}$$

subject to

$$\dot{x} = \hat{A}(t)(x-\hat{x}) + \hat{B}(t)(u-\hat{u}) + \hat{f}(t), \quad x(0) = x^0, \tag{29}$$

$$u(t) \in U, \quad t \in [0,T]. \tag{30}$$

Notice that the switching function $\hat{\sigma}$ and the quadratic form Ω (see (11)) associated with the nonlinear problem (1)–(3) coincide with those associated with the linear-quadratic affine problem (28)–(30). Thus also assumptions (A2) and (A2') for these two problems are identical.

The proof of Theorem 3.1 follows.

ProofA. ccording to Proposition 3.2, it is enough to prove the claim of the theorem for the linearized mapping $\mathcal{F}$ in (26). The positive parameters α and β of SMsR of $\mathcal{F}$ will be arbitrary (that is, can be taken infinite) and c will be fixed later as depending only on L, M, T and the constant c_0 in Assumption (A2').

Take an arbitrary $z = (\xi, \pi, \nu, \rho) \in \mathcal{Z}$ with $d_{\mathcal{Z}}(z) \le \alpha$ and a solution $y = (x, u, p) \in \mathcal{Y}$ of the "perturbed" inclusion

$$z \in \mathcal{F}(y), \tag{31}$$

satisfying $\|u - \hat{u}\|_1 \le \beta$. In detail, inclusion (31) reads as

$$\dot{x}(t) = \hat{A}(t)(x(t)-\hat{x}(t)) + \hat{B}(t)(u(t)-\hat{u}(t)) + \hat{f}(t) - \xi(t), \quad x(0) = x^0, \tag{32}$$

$$\dot{p}(t) = -\nabla_x \hat{H}(t) - \hat{H}_{xy}(t)(y(t)-\hat{y}(t)) + \pi(t), \tag{33}$$

$$p(T) = \nu, \tag{34}$$

$$-N_U(u(t)) \ni \nabla_u \hat{H}(t) + \hat{H}_{uy}(t)(y(t)-\hat{y}(t)) - \rho(t), \tag{35}$$

where the differential equations (32) and (33) have to be fulfilled for a.e. $t \in [0,T]$.

We denote again $\Delta x = x - \hat{x}$, $\Delta u = u - \hat{u}$, $\Delta p = p - \hat{p}$, , $\Delta y = y - \hat{y}$ and we set $\Delta\dot{x} = \dot{x} - \dot{\hat{x}}$, $\Delta\dot{p} = \dot{p} - \dot{\hat{p}}$. Then

$$\Delta\dot{x}(t) = \hat{A}(t)\Delta x(t) + \hat{B}(t)\Delta u(t) - \xi(t), \quad \Delta x(0) = 0, \tag{36}$$

$$\Delta\dot{p}(t) = -\hat{H}_{xy}(t)\Delta y(t) + \pi(t), \quad \Delta p(T) = \nu, \tag{37}$$

$$-N_U(u(t)) \ni \hat{\sigma}(t) + \hat{H}_{uy}(t)\Delta y(t) - \rho(t) \tag{38}$$

Since

$$\begin{aligned}\langle \Delta x(T), \nu\rangle &= \langle \Delta x(T), \Delta p(T)\rangle - \langle \Delta x(0), \Delta p(0)\rangle = \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\langle \Delta x(t), \Delta p(t)\rangle \,\mathrm{d}t \\ &= \int_0^T \left[\langle \Delta\dot{x}(t), \Delta p(t)\rangle + \langle \Delta x(t), \Delta\dot{p}(t)\rangle\right] \mathrm{d}t,\end{aligned}$$

using (36) and (37) and the identity $\hat{H}_{xp}(t)^\top = \hat{A}(t)$ we obtain that

$$\begin{aligned}\langle \Delta x(T), \nu\rangle = \int_0^T \Big[&\langle \hat{B}(t)\Delta u(t), \Delta p(t)\rangle - \langle \xi(t), \Delta p(t)\rangle \\ &-\langle \hat{H}_{xx}(t)\Delta x(t), \Delta x(t)\rangle - \langle \hat{H}_{xu}(t)\Delta u(t), \Delta x(t)\rangle + \langle \pi(t), \Delta x(t)\rangle\Big]\,\mathrm{d}t.\end{aligned} \tag{39}$$

From (38) and the identity $\hat{H}_{up}(t) = \hat{B}(t)^\top$ we have

$$\begin{aligned} 0 &\le \langle \hat{\sigma}(t) + \hat{H}_{uy}(t)\Delta y(t) - \rho(t), \hat{u}(t) - u(t)\rangle \\ &= -\langle \hat{\sigma}(t), \Delta u(t)\rangle - \langle \hat{H}_{ux}(t)\Delta x(t), \Delta u(t)\rangle - \langle \hat{B}(t)^\top \Delta p(t), \Delta u(t)\rangle + \langle \rho(t), \Delta u(t)\rangle. \end{aligned}$$

Integrating this inequality and adding the result to (39), we obtain

$$\begin{aligned}\langle \Delta x(T), \nu\rangle \le \int_0^T \big[&- \langle \hat{\sigma}(t), \Delta u(t)\rangle - \langle \hat{H}_{xx}(t)\Delta x(t), \Delta x(t)\rangle - 2\langle \hat{H}_{ux}(t)\Delta x(t), \Delta u(t)\rangle \\ &-\langle \xi(t), \Delta p(t)\rangle + \langle \pi(t), \Delta x(t)\rangle + \langle \rho(t), \Delta u(t)\rangle\big]\,\mathrm{d}t.\end{aligned}$$

Hence,

$$\int_0^T \langle \hat{\sigma}(t), \Delta u(t)\rangle\,\mathrm{d}t + 2\Omega(\Delta x, \Delta u) \le |\Delta x(T)|\,|\nu| + \|\Delta p\|_\infty\,\|\xi\|_1 + \|\Delta x\|_\infty\,\|\pi\|_1 + \|\Delta u\|_1\,\|\rho\|_\infty. \tag{40}$$

Let us denote $\delta u := \Delta u \in \mathcal{U} - \hat{u}$, and let δx be the corresponding solution of the linearized equation (10). Then from (10) and (36), using the Cauchy formula we obtain that

$$\|\Delta x - \delta x\|_C \le e^{MT}\|\xi\|_1, \qquad \|\delta x\|_C \le e^{MT} M\|\delta u\|_1, \qquad \|\Delta x\|_C \le e^{MT}(M\|\delta u\|_1 + \|\xi\|_1).$$

Further c_1, $c_2, \dots$ will denote constants that depend only on the numbers M, L and T. Thus we can write

$$\varepsilon := \|\Delta x - \delta x\|_C \le c_1\|\xi\|_1 \le c_1 d_{\mathcal{Y}}(z), \qquad \|\delta x\|_C \le c_2\|\delta u\|_1, \qquad \|\Delta x\|_C \le c_3(\|\delta u\|_1 + d_{\mathcal{Y}}(z)), \tag{41}$$

and also in view of (37) (again due to the Cauchy formula, and using the above estimations for $\|\Delta x\|_C$),

$$\|\Delta p\|_C \le c_4(\|\delta u\|_1 + |\nu| + \|\pi\|_1 + \|\xi\|_1) \le c_4(\|\delta u\|_1 + d_{\mathcal{Z}}(z)). \tag{42}$$

Moreover, using (20) and (41) we easily get

$$|\Omega(\Delta x, \delta u) - \Omega(\delta x, \delta u)| \le c_5 d_{\mathcal{Y}}(z)(\|\delta u\|_1 + d_{\mathcal{Y}}(z)).$$

Using the obtained estimate, Assumption (A2') in (40), the last estimate in (41), and (42), we obtain

$$\begin{aligned} c_0\|\delta u\|_1^2 &\le \int_0^T \langle \hat{\sigma}(t), \delta u(t)\rangle\,\mathrm{d}t + 2\Omega(\delta x, \delta u) \\ &\le \int_0^T \langle \hat{\sigma}(t), \delta u(t)\rangle\,\mathrm{d}t + 2\Omega(\Delta x, \delta u) + 2c_5 d_{\mathcal{Y}}(z)(\|\delta u\|_1 + d_{\mathcal{Y}}(z)) \\ &\le d_{\mathcal{Z}}(z)\max\{|\Delta x(T)|,\, \|\Delta p\|_\infty,\, \|\Delta x\|_\infty,\, \|\Delta u\|_1\} + 2c_5 d_{\mathcal{Y}}(z)(\|\delta u\|_1 + d_{\mathcal{Y}}(z)) \\ &\le c_6 d_{\mathcal{Y}}(z)(\|\delta u\|_1 + d_{\mathcal{Y}}(z)). \end{aligned}$$

This inequality implies that

$$\|\delta u\|_1 \le \frac{c_6}{c_0}\,\frac{1+\sqrt{1+4c_0/c_6}}{2}\,d_{\mathcal{Y}}(z).$$

Then the needed inequality $d_{\mathcal{Y}}(y,\hat{y}) \le c d_{\mathcal{Z}}(z)$ follows from the last estimate, combined with the last estimate in (41) and (42). Q.E.D.

4 The case of a bang-bang optimal control

Now we give sufficient conditions under which Assumption (A2') is fulfilled. It applies to the case where the reference control $\hat{u}$ has a bang-bang structure. We remind that $\hat{y} = (\hat{x}, \hat{u}, \hat{p}) \in \mathcal{Y}$ is a fixed reference solution of the optimality system (24), and under (A2) $\hat{u}$ is a strict locally optimal solution of problem (1)-(3) (see Corollary 2.1).

Let U be a compact convex polyhedron. Using geometric (rather than analytic) terminology we denote by V the set of all vertices of U, and by E, the set of all unit vectors $e \in \mathbb{R}^m$ that are parallel to some edge of U. For every unit vector $e \in \mathbb{R}^m$ denote

$$\hat{\sigma}_e(t) = \langle \hat{\sigma}(t), e\rangle.$$

For a vertex $v \in V$, denote by $E(v)$ the set of all unit vectors $(v'-v)/|v'-v|$, where v' is any neighboring vertex to v (that is, the segment $[v,v']$ is an edge of U). Notice that $E = \cup_{v\in V} E(v)$.

We introduce the following two assumptions in a somewhat more general form than needed in the context of Theorem 3.1.

Assumption (B'). There exist numbers $\kappa \ge 1$, $\gamma_0 > 0$ and $\tau_0 > 0$ such that for every $e \in E$ and for every $s \in [0,T]$ for which $\hat{\sigma}_e(s) = 0$, it holds that

$$|\hat{\sigma}_e(t)| \ge \gamma_0 |t-s|^\kappa \quad \text{for every } \ t \in [s-\tau_0, s+\tau_0] \cap [0,T]. \tag{43}$$

Set $\hat{E}(s) := \underset{[0,T]\ni t\to s}{\text{Limsup}}\, E(\hat{u}(t))$, where Limsup denotes the Kuratowski upper limit.Whenever $\hat{u}$ is piece-wise constant, it holds that $\hat{E}(s) := E(\hat{u}(s))$ except at the jump points of $\hat{u}$. If s is a jump point of $\hat{u}$, then $\hat{E}(s) := E(\hat{u}(s-)) \cup E(\hat{u}(s+))$.

Assumption (B"). The function $\hat{u}$ is piecewise constant with values in the set of vertices, V, of U. Moreover, there exist numbers $\kappa \ge 1$, $\gamma_0 > 0$ and $\tau_0 > 0$ such that condition (43) is fulfilled for every $s \in [0,T]$ and $e \in \hat{E}(s)$, for which $\hat{\sigma}_e(s) = 0$.

Proposition 4.1 *Let assumption (A1) and at least one of the assumptions (B') and (B") be fulfilled. Then there exists a number $\mu > 0$ such that for every $u \in \mathcal{U}$ it holds that*

$$\int_0^T \langle \hat{\sigma}(t), u(t) - \hat{u}(t)\rangle \,\mathrm{d}t \ge \mu \|u - \hat{u}\|_1^{\kappa+1}. \tag{44}$$

The proof includes argumentation similar to those used in the case of a box-like set U, see [7, Lemma 3.3], [14], [16, Lemma 1.3], [13, Sect. 3]. However, the case of a general polyhedral set U requires additional analysis, therefore we present the proof below.

Proof. First we prove that (B') implies (B"). For this it is enough to prove that, under Assumption (B'), the control $\hat{u}$ is piecewise constant and takes values (after changing it on a set of measure zero) in V. Assumption (B') implies that for every $e \in E$, the function $\langle \hat{\sigma}(\cdot), e\rangle$ has at most $K_0 := T/\tau_0 + 1$ zeros in $[0, T]$. Then the set of all zeros of $\langle \hat{\sigma}(\cdot), e\rangle$ with $e \in E$ has at most KK_0 elements, where K is the number of edges of U. In every interval (s', s'') which does not contain any of these zeros, and for every $e \in E$, the function $\langle \hat{\sigma}(\cdot), e\rangle$ does not vanish in (s', s'') and since it is continuous, it has a constant sign in this interval. Then $\langle \hat{\sigma}(t), v\rangle$ has a unique minimizer $v \in U$, it is a vertex of U, and the vertex v is the same for every $t \in (s', s'')$. Then due to condition (9) in the Pontryagin principle, $\hat{u}(t) = v$ on (s', s'') modulo a set of zero measure.

Now, assume that Assumption (B") is fulfilled. For every vertex $v \in V$ and every $u \in U$ there exists a representation

$$u = v + \sum_{e \in E(v)} \lambda_e\, e \quad \text{with} \quad \lambda_e \geq 0. \tag{45}$$

Since $v \in V$, there exist $q \in \mathbb{R}^m$, $|q| = 1$, and a number $\varepsilon > 0$ such that

$$\langle q, e\rangle \geq \varepsilon \quad \text{for every} \;\; e \in E(v).$$

Then

$$|u - v| = |u - v|\,|q| \geq \langle q, u - v\rangle = \sum_{e \in E(v)} \lambda_e \langle q, e\rangle \geq \varepsilon \sum_{e \in E(v)} \lambda_e$$

Hence,

$$\sum_{e \in E(v)} \lambda_e \leq \frac{\operatorname{diam}(U)}{\varepsilon} =: d, \tag{46}$$

where $\operatorname{diam}(U) = \max\{|u' - u''| : u', u'' \in U\}$.

Let us fix an arbitrary $u \in \mathcal{U}$. From the representation (45) and (46) we have

$$u(t) - \hat{u}(t) \in G(t) := d \operatorname{co}\big\{E(\hat{u}(t)),\, 0\big\}.$$

The mapping G is closed-valued with values in $\mathbb{R}^m$ and Lebesgue measurable (we remind that $\hat{u}$ takes only finitely many values). Then Theorem 8.2.15 in [3] (Carathéodory representation) asserts that there exist measurable selections $e_0(\cdot), \dots e_m(\cdot)$ of $E(\hat{u}(\cdot))$ and measurable functions $\alpha_0(\cdot), \dots, \alpha_m(\cdot)$ such that

$$u(t) - \hat{u}(t) = d \sum_{i=0}^{m} \alpha_i(t) e_i(t), \quad \alpha_i(t) \geq 0, \quad \sum_{i=0}^{m} \alpha_i(t) \leq 1. \tag{47}$$

Denoting $\lambda_i(t) = d\alpha_i(t)$ we have

$$u(t) - \hat{u}(t) = \sum_{i=0}^{m} \lambda_i(t) e_i(t), \quad \lambda_i(t) \geq 0, \;\; e_i(t) \in E(\hat{u}(t)),$$

with measurable λ_i and e_i.

On the other hand,

$$|u(t) - \hat{u}(t)| \leq \sum_{i=0}^{m} \lambda_i(t) |e_i(t)| = \sum_{i=0}^{m} \lambda_i(t). \tag{48}$$

Now we consider the quantity

$$\Delta := \int_0^T \langle \hat{\sigma}(t), u(t) - \hat{u}(t) \rangle \, \mathrm{d}t = \int_0^T \left\langle \hat{\sigma}(t), \sum_{i=0}^{m} \lambda_i(t) e_i(t) \right\rangle \mathrm{d}t = \int_0^T \sum_{i=0}^{m} \lambda_i(t) \, |\hat{\sigma}_{e_i(t)}(t)| \, \mathrm{d}t,$$

where we use the fact that $\hat{\sigma}_e(t) \geq 0$ for every $e \in E(\hat{u}(t))$, in particular for $e = e_i(t)$. Denoting $\psi(t) := \min_{e \in E(\hat{u}(t))} |\hat{\sigma}_e(t)|$, and using the inclusion $e_i(t) \in E(\hat{u}(t))$ and (48), we obtain that

$$\Delta \geq \int_0^T \psi(t) \sum_{i=0}^{m} \lambda_i(t) \, \mathrm{d}t \geq \int_0^T \psi(t) \, |u(t) - \hat{u}(t)| \, \mathrm{d}t. \tag{49}$$

We shall show that there exists a number c_1 such that for every number $\beta > 0$, the Lebesgue measure of the set

$$\Psi_\beta = \{t \in [0, T] : \, \psi(t) \leq \beta\}$$

satisfies

$$\mathrm{meas}\,(\Psi_\beta) \leq c_1 \beta^{\frac{1}{\kappa}}. \tag{50}$$

Let us consider a maximal open interval (s', s'') in which $\hat{u}$ has a constant value $v \in V$, so that we have $E(\hat{u}(t)) = E(v)$ for all $t \in (s', s'')$. Let $e \in E(v)$ be arbitrarily fixed. Then $e \in \hat{E}(t)$ for every $t \in [s', s'']$.

Denote by Θ the union of all intervals $(s - \tau_0, s + \tau_0) \cap [s', s'']$, where s is a zero of $\hat{\sigma}_e$ in $[s', s'']$. Assumption (B") is applicable for any such s and the fixed e. It implies that the intervals $(s - \tau_0, s + \tau_0) \cap [s', s'']$ are disjoint and their number is at most $K_0 = T/(2\tau_0) + 1$. The set $[s', s''] \setminus \Theta$ is compact and $\hat{\sigma}_e(t) \neq 0$ for every $t \in [s', s''] \setminus \Theta$. Then there is a number $\beta_0 > 0$ such that $|\hat{\sigma}_e(t)| \geq \beta_0$ for every $t \in [s', s''] \setminus \Theta$. For $\beta \in (0, \beta_0)$ we have

$$\mathrm{meas}\,\{t \in [s', s''] : \, |\hat{\sigma}_e(t)| \leq \beta\} = \mathrm{meas}\,\{t \in \Theta : \, |\hat{\sigma}_e(t)| \leq \beta\}.$$

Using again Assumption (B") we obtain that for each of the intervals $(s - \tau_0, s + \tau_0) \cap [s', s'']$

$$\mathrm{meas}\,\{t \in (s - \tau_0, s + \tau_0) \cap [s', s''] : \, |\hat{\sigma}_e(t)| \leq \beta\} \leq 2 \left(\frac{\beta}{\gamma_0} \right)^{\frac{1}{\kappa}}.$$

Hence,

$$\mathrm{meas}\,\{t \in [s', s''] : \, |\hat{\sigma}_e(t)| \leq \beta\} \leq 2K_0 \left(\frac{\beta}{\gamma_0} \right)^{\frac{1}{\kappa}}.$$

For $\beta \geq \beta_0$ we have

$$\mathrm{meas}\,\{t \in [s', s''] : \, |\hat{\sigma}_e(t)| \leq \beta\} \leq T \leq T \left(\frac{\beta}{\beta_0} \right)^{\frac{1}{\kappa}}.$$

Thus for any $\beta > 0$ we have

$$\mathrm{meas}\,\{t \in [s', s''] : \, |\hat{\sigma}_e(t)| \leq \beta\} \leq c' \beta^{\frac{1}{\kappa}}, \qquad \text{where } c' = \max \left\{ \frac{2K_0}{(\gamma_0)^{\frac{1}{\kappa}}}, \, \frac{T}{(\beta_0)^{\frac{1}{\kappa}}} \right\}.$$

Since $e \in E(v)$ was arbitrarily chosen and the set $E(v)$ contains at most K elements,

$$\mathrm{meas}\,\{t \in [s', s''] : \, \psi(t) \leq \beta\} \leq K c' \beta^{\frac{1}{\kappa}}.$$

This implies (50) with $c_1 = (KK_0+1)Kc'$ (we remind that KK_0 is an upper bound for the number of jumps of $\hat{u}$).

From (49) and (50) we obtain that for every $\beta > 0$

$$\begin{aligned}\Delta &\geq \int_{[0,T]\setminus\Psi_\beta} \beta\,|u(t)-\hat{u}(t)|\,\mathrm{d}t \geq \int_0^T \beta\,|u(t)-\hat{u}(t)|\,\mathrm{d}t - \int_{\Psi_\beta} \beta\,|u(t)-\hat{u}(t)|\,\mathrm{d}t \\ &\geq \beta\|u-\hat{u}\|_1 - c_1\|u-\hat{u}\|_\infty \beta^{1+\frac{1}{\kappa}} \geq \beta\|u-\hat{u}\|_1 - c_1 d\beta^{1+\frac{1}{\kappa}},\end{aligned}$$

where d denotes the diameter of the set U. Using this inequality with $\beta = (2c_1 d)^{-\kappa}\|u-\hat{u}\|_1^\kappa$ and remembering the definition of Δ we obtain the inequality (44) with $\mu = 2^{-1}(2c_1 d)^{-\kappa}$. Q.E.D.

Corollary 4.1 *Let assumption (A1) and at least one of the assumptions (B') and (B") be fulfilled with $\kappa = 1$. Let, in addition, there exists a constant $\mu_0 \in (0, \mu)$ (where μ is the constant in Proposition 4.1) such that*

$$2\,\Omega(\delta x, \delta u) \geq -\mu_0\|\delta u\|_1^2 \quad \text{for all } (\delta x, \delta u) \in \Gamma. \tag{51}$$

Then the optimality mapping $\psi + \Psi$, associated with problem (1)-(3), is strongly metrically subregular at $\hat{y} = (\hat{x}, \hat{u}, \hat{p})$ for zero.

As shown in the proof of Proposition 4.1, Assumption (B') is stronger than (B"). However, it has the minor advantage that it explicitly involves only U and $\hat{\sigma}$, while (B") also explicitly involves the reference solution $\hat{u}$.

Remark 4.1 A similar assumption as (51) was introduced in [1] and was used for error analysis of the Euler discretization scheme applied to affine problems. The result of Proposition 4.1 was taken as an assumption there, but as mentioned in [1], this result was essentially known from e.g. [7, 16]. However, only the case of a box-like set U was investigated in these papers (which brings technical simplifications), and even in this case the assumptions made were somewhat stronger than our (B'). Namely, the switching functions σ_e were not allowed to have zeros at 0 and T. A stability result in a form close to Corollary 4.1 is presented in [2, Theorem 8] for linear-quadratic affine problems with box-like control constraints (more references about stability of solutions are also given there).

5 Error analysis of the Euler discretization

In this section we utilize the SMsR property of the optimality mapping, associated with the affine problem (1)-(3) with the specification (4), for error analysis of the Euler discretization scheme. The investigation of the convergence rate of Runge-Kutta discretization schemes applied to affine optimal control problems began with the paper [17] some three decades later than for problems satisfying a Legendre-type condition. A sequence of papers followed, most of them for problems with linear dynamics. We refer to the recent paper [1], where the dynamics is non-linear and more detailed bibliography is given. In all these papers (except [12, 15], where a different, not Runge-Kutta-type, discretization scheme is proposed) the convergence rate is of at most first order, and using higher order Runge-Kutta schemes on a uniform mesh cannot help to improve the convergence rate if the optimal control is discontinuous. Therefore, below we focus on the simplest scheme – the Euler one.

Let $t_0, \dots, t_N$ be the uniform mesh in $[0, T]$ with step $h = T/N$ (N is a natural number), that is, $t_i = iT/N$, $i = 0, \dots, N$. The discrete-time (mathematical programming) problem obtained by the Euler discretization, denoted further by $\mathcal{P}_h$, reads as

$$\min\ h \sum_{i=0}^{N-1} g(t_i, x_i, u_i)$$

subject to the constraints

$$x_{i+1} = x_i + hf(t_i, x_i, u_i), \quad i = 0, \dots, N-1, \qquad x_0 \text{ – given},$$

$$u_i \in U, \quad i = 0, \dots, N-1,$$

where $x = (x_0, \dots, x_N) \in \mathbb{R}^{n(N+1)}$ and $u = (u_0, \dots, u_{N-1}) \in \mathbb{R}^{mN}$ are the discrete-time state and control vectors. The local form of the discrete-time maximum (here minimum) principle (the Karush-Kuhn-Tacker conditions) for this problem claims that any locally optimal pair (x, u), satisfies, together with a co-state vector $p = (p_0, \dots, p_N) \in \mathbb{R}^{n(N+1)}$ the system

$$x_{i+1} = x_i + hf(t_i, x_i, u_i), \quad x_0 \text{ – given}, \tag{52}$$

$$p_i = p_{i+1} + h\nabla_x H(t_i, x_i, u_i, p_{i+1}), \quad p_N = 0, \tag{53}$$

$$0 \in \nabla_u H(t_i, x_i, u_i, p_{i+1}) + N_U(u_i), \tag{54}$$

where i runs between 0 and $N-1$.

Let (x^h, u^h) be a solution of problem $\mathcal{P}_h$ and let p^h be the corresponding co-state vector, so that $y^h := (x^h, u^h, p^h)$ satisfies system (52)–(54). In order to compare this solution with the reference solution $\hat{y} = (\hat{x}, \hat{u}, \hat{p})$ of the continuous-time problem we embed the sequence (x^h, u^h, p^h) into the space $W^{1,1} \times L^1 \times W^{1,1}$ defining

$$x_h(t) := x_i^h + \frac{t - t_i}{h}(x_{i+1}^h - x_i^h), \qquad u_h(t) := u_i^h, \qquad p_h(t) = p_i^h + \frac{t - t_i}{h}(p_{i+1}^h - p_i^h),$$

for $t \in [t_i, t_{i+1})$, $i = 0, \dots, N-1$. Denote $y_h := (x_h, u_h, p_h)$.

Assumption (C1). Assumption (A1) is fulfilled with the additional requirement that f and g and their first derivatives in x and u are Lipschitz continuous in t, uniformly with respect to (x, u) in any compact set.

Assumption (C2). The optimality mapping $\psi + \Psi$, associated with problem (1)-(3) (see (24) and (25)), is strongly metrically sub-regular at $\hat{y} = (\hat{x}, \hat{u}, \hat{p})$ for zero with some parameters $(\hat{\alpha}, \hat{\beta}, \hat{c})$.

Assumption (C3). For all sufficiently small $h > 0$, all components of x^h and p^h (thus also the values of x_h and p_h) belong to the set $\bar{S}$ defined in Remark 2.1. Moreover, $d_{\mathcal{Y}}(y_h, \hat{y}) \le \hat{\beta}$.

The first part of Assumption (C3) is technical. It means that the solution y^h of (52)–(54) is not far away from the reference solution $(\hat{x}, \hat{u}, \hat{p})$ of the continuous-time problem, at least for h small enough. It allowes us to use in the subsequent analysis the same constants L and M defined in Remark 2.1. The second part is crucial, at least because y_h may happen to be close to some other local solution of the continuous-time problem, and we have to eliminate this possibility by an assumption.

Theorem 5.1 *Let assumptions (C1)–(C3) be fulfilled. Then there exists a constant C such that the estimate*

$$\|x_h - \hat{x}\|_{1,1} + \|u_h - \hat{u}\|_1 + \|p_h - \hat{p}\|_{1,1} \le Ch$$

holds for all sufficiently small $h > 0$.

Proof. In order to make use of the SMsR property of the optimality mapping for problem (1)-(3), we have to estimate the residuals

$$\begin{aligned}
\xi(t) &:= \dot{x}_h(t) - f(t, x_h(t), u_h(t)),\\
\pi(t) &:= \dot{p}_h(t) + \nabla_x H(t, x_h(t), u_h(t), p_h(t)),\\
\rho(t) &:= \nabla_u H(t_i, x_i^h, u_i^h, p_i^h) - \nabla_u H(t, x_h(t), u_h(t), p_h(t)), \quad t \in [t_i, t_{i+1}), \quad i = 0, \dots, N-1.
\end{aligned}$$

The residual in the transversality condition (8) is $\nu = 0$. We mention that the inclusion (9) is satisfied by (x_h, u_h, p_h) with residual ρ because of the constancy of u^h in every interval $[t_i, t_{i+1})$, so that $N_U(u_h(t)) = N_U(u_i^h)$, $t \in [t_i, t_{i+1})$, $i = 0, \dots, N-1$.

We have

$$\begin{aligned}
\|\xi\|_1 &= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \left| \frac{1}{h}(x_{i+1}^h - x_i^h) - f\Big(t, x_i^h + \frac{t-t_i}{h}(x_{i+1}^h - x_i^h), u_i^h\Big) \right| \mathrm{d}t\\
&= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \left| f(t_i, x_i^h, u_i^h) - f\Big(t, x_i^h + \frac{t-t_i}{h}(x_{i+1}^h - x_i^h), u_i^h\Big) \right| \mathrm{d}t\\
&\le \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} L[(t-t_i) + \frac{t-t_i}{h}|x_{i+1}^h - x_i^h|] \,\mathrm{d}t \le \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} L(1+M)(t-t_i)\,\mathrm{d}t\\
&\le \frac{1}{2} TL(M+1)h.
\end{aligned}$$

The estimation for $\|\pi\|_1$ is similar. Moreover, for $t \in [t_i, t_{i+1})$ we have

$$|\rho(t)| \le L\Big[(t-t_i) + \frac{t-t_i}{h}|x_{i+1}^h - x_i^h| + \frac{t-t_i}{h}|p_{i+1}^h - p_i^h|\Big] \le L(1+2M)h,$$

thus $\|\rho\|_\infty \le L(2M+1)h$.

For all sufficiently small h we have $\|\xi\|_1 + \|\pi\|_1 + |\nu| + \|\rho\|_\infty \le \hat{\alpha}$. Then the claim of the theorem follows from the SMsR property of $\psi + \Psi$ and the above estimates of the norms $\|\xi\|_1$, $\|\pi\|_1$, $\|\rho\|_\infty$ via h, using the second part of Assumption (C3). Q.E.D.

The principle advantage of this theorem compared with [1] is that Theorem 4 in [1] claims a first order error estimate under particular sufficient conditions for SMsR (this notion is not used there), while our result is based only on the SMsR property, thus it holds under any sufficient conditions for this property. In particular, it holds under the conditions in Theorem 3.1 or under the more elaborated but more restrictive conditions in Corollary 4.1 (which are still weaker than those in [1]).

We mention that in practice one may only obtain an approximate solution of the discrete problem $\mathcal{P}_h$. Usually the numerical methods for solving this problem produce an approximate solution $\tilde{y}^h$ of the optimality system (52)–(54). In this case, the error estimate in Theorem 5.1 can

be modified by adding the additional residuals caused by the non-exactness in solving problem $\mathcal{P}_h$ to the residuals resulting from the discretization. More precisely, let the approximate solution $\tilde{y}^h$ satisfy (52)–(54) with residual $(\tilde{\xi}^h, \tilde{\pi}^h, \tilde{\nu}^h, \tilde{\rho}^h) = \left(\{\tilde{\xi}_i^h\}_0^{N-1}, \{\tilde{\pi}_i^h\}_0^{N-1}, \tilde{\nu}^h, \{\tilde{\rho}_i^h\}_0^{N-1}\right)$. We measure the size or this residual by the number

$$\varepsilon := h \sum_{i=0}^{N-1} (|\tilde{\xi}_i^h| + |\tilde{\pi}_i^h|) \; + \; |\tilde{\nu}^h| \; + \; \max_{i=0,\ldots,N-1} |\tilde{\rho}_i^h|.$$

Then, after embedding the sequence $\tilde{y}^h$ as $\tilde{y}_h$ into the space $W^{1,1} \times L^1 \times W^{1,1}$, as we did before for y^h, the estimate in Theorem 5.1 takes the form

$$\|\tilde{x}_h - \hat{x}\|_{1,1} + \|\tilde{u}_h - \hat{u}\|_1 + \|\tilde{p}_h - \hat{p}\|_{1,1} \leq C(h + \varepsilon).$$

References

[1] W. Alt, U. Felgenhauer and M. Seydenschwanz, *Euler discretization for a class of nonlinear optimal control problems with control appearing linearly*, Computational Optimization and Applications (2017), https://doi.org/10.1007/s10589-017-9969-7.

[2] W. Alt, C. Schneider and M. Seydenschwanz, *Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang-bang solutions*, Appl. Math. Comput., 287 (2016), pp. 104–124.

[3] J.-P. Aubin and H. Frankowska. *Set-valued Analysis*. Birkhäuser, Boston, Basel, Berlin, 1990.

[4] R. Cibulka, A.L. Dontchev, and A.Y. Kruger. *Strong metric subregularity of mappings in variational analysis and optimization*. Journal of Mathematical Analysis and Applications, 457 (2018), pp. 1247-1282.

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

[6] A.L. Dontchev, R.T. Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis. Second edition*. Springer, New York, 2014.

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

[8] U. Felgenhauer, *Discretization of semilinear bang-singular-bang control problems*, Computational Optimization and Applications, 64 (2016), pp. 295–326.

[9] U. Felgenhauer, *A Newton-type method and optimality test for problems with bang-singular-bang optimal control*. Pure and Applied Functional Analysis, 1 (2016), pp. 197–215.

[10] W. W. Hager. *Multiplier methods for nonlinear optimal control*. SIAM J. Numer. Anal., 27(1990), pp. 1061–1080.

[11] M.R. Hestenes. *Calculus of variations and optimal control theory*. John Wiley&Sons, 1966.

[12] A. Pietrus, T. Scarinci, and V.M. Veliov. *High order discrete approximations to Mayer's problems for linear systems*. SIAM J. Control Optim., 56 (2018), pp. 102–119.

[13] J. Preininger, T. Scarinci and V.M. Veliov, *Metric regularity properties in bang-bang type linear-quadratic optimal control problems*. Set-Valued and Variational Analysis, 27 (2019), pp. 381–404.

[14] M. Quincampoix and V. Veliov, *Metric Regularity and Stability of Optimal Control Problems for Linear Systems*, SIAM J. Control Optim, 51 (2013), pp. 4118–4137.

[15] T. Scarinci and V.M. Veliov, *Higher-order numerical schemes for linear quadratic problems with bang-bang controls*, Computational Optimization and Applications, 69 (2018), pp.403–422.

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

[17] V.M. Veliov, *Error analysis of discrete approximation to bang-bang optimal control problems: the linear case*, Control Cyberne., 34 (2005), pp. 967–982.