INAUGURAL-DISSERTATION

submitted

to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of

Ruprecht-Karls-University

Heidelberg

for the Degree of Doctor of Natural Sciences

Put forward by
Bao Duy, Tran, M.Sc.
from Vietnam

Day of the oral exam:

Numerical Methods For Optimal Control Of Switched Systems

Supervisor: Prof. Dr. Ekaterina Kostina

Zusammenfassung

Die Herausforderungen in realen Anwendungen, etwa beim Betrieb von Systemen mit Lastschwankungen sowie bei Anfahr- und Abschaltvorgängen, stellen komplexe mathematische Problemstellungen dar. Diese Komplexität resultiert aus ausgeprägten Nichtlinearitäten (insbesondere in transienten Phasen), gemischt-ganzzahligen Entscheidungsvariablen und Stellgrößen (z. B. zur Kopplung einzelner Komponenten), zustandsabhängigen Diskontinuitäten (hervorgerufen durch Phasenübergänge oder Regler) sowie einer hohen Systemdimension. Während in der Industrie häufig auf Entkopplungsstrategien und rezeptbasierte Regelungen zurückgegriffen wird, erweisen sich diese Ansätze für derart komplexe und eng gekoppelte Systeme als unzureichend, was den Bedarf an innovativen nichtlinearen Optimierungsmethoden deutlich macht. Für Prozesse unter Unsicherheit sind statische Open-Loop-Stellstrategien nicht geeignet; bevorzugt werden optimale rückgekoppelte Stellgesetze, die auf geschätzten Zuständen basieren. Der derzeit meistverwendete Ansatz für allgemeine nichtlineare optimale Regelungsprobleme mit Zustands- und Stellgrößenbeschränkungen ist die Nichtlineare Modellprädiktive Regelung (NMPC). Das Grundprinzip besteht darin, den aktuellen Zustand aus Messdaten über einen endlichen „bewegten“ Zeithorizont der Vergangenheit zu schätzen und die Stellgrößen über einen „bewegten“ Zeithorizont im Open-Loop zu optimieren. Der erste Steuerimpuls wird anschließend über ein Abtastintervall angewandt, während bereits die nächste Reoptimierung durchgeführt wird.

Diese Dissertation entwickelt numerische Methoden zur Berechnung von Open-Loop- und Feedback-Reglern in bestimmten Klassen gemischt-ganzzahliger optimaler Steuerungsprobleme mit geschalteten gewöhnlichen Differentialgleichungen (SwOCP). Diese Probleme finden wichtige Anwendung in der Charakterisierung der komplexen Eigenschaften von Trockenreibungsproblemen. Wir folgen der Filippov-Regel, nach der das SwOCP in ein optimales Steuerungsproblem mit gemischt-ganzzahligen Steuerfunktionen und speziellen gemischten Steuer-Zustands-Beschränkungen umformuliert wird. Wir untersuchen die relaxierte Formulierung dieses optimalen Steuerungsproblems und leiten notwendige Optimalitätsbedingungen aus dem Pontryagin-Maximumprinzip (PMP) ab, wobei die Regularitätseigenschaft der gemischten Beschränkungen sorgfältig berücksichtigt wird. Numerische Methoden für das relaxierte Problem, basierend auf dem Multiple-Shooting-Ansatz und einem geeigneten „Rundungsschema“ zur Behandlung impliziter Schaltvorgänge, werden untersucht. Um optimale Feedback-Regelgesetze zu berechnen, verallgemeinern wir den „NMPC“-Ansatz auf die allgemeine SwOCP-Klasse. Wir entwickeln einen direkten Ansatz zur Ableitung von Feedback-Regelgesetzen. Es basiert auf dem PMP-Ansatz zur Berechnung von „Nachbar-Feedback“-Reglern, um die explizite Umschaltung von ganzzahligen Reglern zu ermitteln. Die numerischen Methoden werden anhand von Benchmark-Problemen mithilfe der MUSCOD-II-Tool-Software mit PGPLOT oder MATLAB veranschaulicht.

Abstract

The challenges in real-life applications, like e.g., managing systems with load fluctuations, start-up, and shut-down, represent complex mathematical problems. This complexity stems from strong nonlinearities (especially in transients), mixed-integer decision variables and controls (e.g., for coupling components), state-dependent discontinuities (from phase transitions or controllers), and the large system dimension. While industry often relies on decoupling and recipe-based controls, these prove insufficient for such intricate, coupled systems, highlighting a need for innovative nonlinear optimization methods. For processes under uncertainty, static open-loop controls are inadequate; optimal feedback control laws, dependent on estimated states, are preferred. The presently most popular approach for general nonlinear optimal control problems with state and control constraints is Nonlinear Model Predictive Control (NMPC). The main idea is to estimate the present state from measured data on a finite "moving" time horizon of the past and to optimize the control on a "moving" time horizon in an open-loop. The first instant of the control is then applied during a sampling time interval, during which the next re-optimization is computed.

This dissertation develops numerical methods for computing open-loop and feedback controls in certain classes of mixed-integer optimal control problems with switched ODEs (SwOCP), which exhibit important applications to characterize the complex properties of dry friction problems. We follow Filippov's rule, according to which the SwOCP is reformulated to an optimal control problem with mixed integer controls and special mixed control-state constraints. We investigate the relaxed formulation of this optimal control problem and derive necessary optimality conditions from the Pontryagin maximum principle (PMP), where the regularity property of the mixed constraints is carefully considered. Numerical methods for the relaxed problem based on the multiple shooting approach and an appropriate "rounding scheme" to handle implicit switching are investigated. In order to compute optimal feedback control laws, we generalize the "NMPC" approach to the general SwOCP class above. We develop a direct approach to derive feedback control laws. It is based on the PMP approach to computing "neighbouring feedback" controls to find out the explicit switching of integer controls. The numerical methods are illustrated with benchmark problems via the MUSCOD-II tool software with PGPLOT or MATLAB.

Acknowledgments

I should thank a lot of people. I apologize if I forgot to mention someone.
First of all, I thank my supervisor Prof. Dr. Ekaterina Kostina, who allowed me to study in her research group at the Institute for Mathematics, Heidelberg University. She also carefully instructed and encouraged me very much during my Ph.D. journey. Without her support, I would not have had enough motivation to finish this study. I appreciate Prof. Dr. Gerhard Reinelt, for giving me valuable instructions in dealing with discrete optimization. Moreover, I also received influential comments from him on the direction for the second research stay of my project at CNR, Rome.
A special thank goes to Dr. Giovanni Rinaldi, Dr. Claudio Gentile, Diego Maria Pinto, Tiziano Bacci (CNR, Rome); Prof. Frank Vallentin, Jan Hendrik Rolfes, and Stefan Krupp (University of Cologne); and Dr. Kurt Majewski (Siemens AG, Munich). Their suggestions and constant kindness helped to make my secondments very nice experiences. I am grateful to all of my friends in the Numerical Optimization Group, and the Discrete & Combinatorial Optimization Group, e.g. Mrs. Catherine Proux-Wieland, Ms. Herta Fitzer, Andreas Sommer, Matthias Schlöder, Robert Scholz, Kaushal Kumar, Ihno Schrot and specially Sumet Khumphairan; and the MINOA EU network, e.g. Esteban Salgado, Daniel Brosch, and al., as well as my friends in the lunch-board game group, badminton group, e.g. Diego Costa, Sebastian Lackner, Dennis Aumiller, Achim Hildenbrandt, Andreas Splitz, and Chaiyod Kamthorncharoen, for helping me to balance quotidian life. We are truly a big international family where people are very friendly with each other. I learned a lot from them and working with them is a great pleasure.
Last, but absolutely not least, I would like to thank my wife, Quynh Nhu, my little daughter, Nhu An, and my family for always supporting me and being interested in what I am doing. I dedicate this dissertation to them.

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 764749, is gratefully acknowledged.

List of Acronyms

ACQ	Abadie Constraint Qualification
AD	Automatic Differentiation
BDF	Backward Differential Formula
BVP	Boundary-Value Problem
CIA	Combinatorial Integral Approximation
CHA	Competing Hamiltonian Approach
CQ	Constraint Qualification
DAE	Differential Algebraic Equation
DP	Disjunctive Programming
EFS	Externally Forced Switches
END	External Numerical Differentiation
FD	Finite Differences
FHA	Flat Hybrid Automata
GCQ	Guignard Constraint Qualification
GDP	Generalized Disjunctive Programming
IFS	Internally Forced Switches
IVP	Initial Value Problem
KKT	Karush–Kuhn–Tucker
LICQ	Linear Independence Constraint Qualification
LMP	Local Maximum Principle
LMPC	Linear Model Predictive Control
LP	Linear Program
MFCQ	Mangasarian–Fromowitz Constraint Qualification
MILP	Mixed Integer Linear Programming
MINLP	Mixed Integer Nonlinear Programming Problem
MIOCP	Mixed Integer Optimal Control Problem
MIQP	Mixed Integer Quadratic Program
MPBVP	Multi-Point Boundary-Value Problem

MPC	Model Predictive Control
MPVC	Mathematical Program with Vanishing Constraints
NFL	Neighboring Feedback Law
NLP	Nonlinear Programming Problem
NMPC	Nonlinear Model Predictive Control
OC	Outer Convexification
OCP	Optimal Control Problem
ODE	Ordinary Differential Equation
PMP	Pontryagin's Maximum Principle
POC	Partial Outer Convexification
PWA	Piecewise Affine
QP	Quadratic Program
QPVC	Convex Quadratic Programs with Vanishing Constraints
RC.SwP	Relaxed Convexified Switched Optimal Control Problem
RHC	Receding Horizon Control
RMC	Regulation of Mixed Constraints
SAR	Switching Aware Rounding
SOS	Special Ordered Set
SOS-1	Special Ordered Set of type 1
SOS-2	Special Ordered Set of type 2
SQP	Sequential Quadratic Programming
SUR	Sum-Up Rounding
SwOCP	Switched Optimal Control Problem
SOCP	Switched Optimal Control Problem without integer control
VC	Vanishing Constraint
VDE	Variational Differential Equation

Contents

Chapter 1

Introduction

Many phenomena occurring in industrial productions and plants can be described using mathematical expressions, such as differential equations. With the advances in computer technology and the invention of numerical methods, it became possible to accurately predict the efficiency of a new production and plant design or the effect of new control strategies. This offers a potent instrument to process engineers, who may want to evaluate the usability and benefits of their ideas in a theoretical way before realizing them. As a consequence, process engineering has become a very important discipline within chemical engineering over the past forty years, avoiding the necessity of expensive pilot plants and yielding profit increases by improving existing processes.

Complex switched systems are one of the most challenging topics in optimal controls with mixed state-control constraints. One particular class of these OCPs is Switched Optimal Control Problem (SwOCP), which are consisting of a switching law specifying binary or integer control variables at each time instant, i.e., controls that can only take values from a finite admissible set. There is plenty of the number of researches on this topic, which include both theoretical and computational numerical results. Recently, there has been a huge number of research to solve this problem, both direct and indirect methods.

1.1 Contributions

Several approaches to solving SwOCP have been investigated. Some instances of mechanical problems are used to illustrate our ideas. The main results and contributions of this dissertation are described as follows.

First Approach: LMP, Filippov's Rule with CQ

The Local Maximum Principle (LMP) is used as an indirect method for SwOCP in the extended version (relaxed-convexified) of Optimal Control Problem (OCP) with mixed state-control constraints. Some reformulations are used as start-of-the-art techniques to deduce the optimality conditions, where Filippov's rule is applied carefully, and the regularity of the mixed constraints is discussed.

Second Approach: Filippov's Rule and Feedback Algorithm

After reformulating SwOCP by Filippov's Rule and the relaxation, and analyzing the condensing block structure, a feedback algorithm is proposed to track switches. That results are confirmed by comparing with the exploiting of the active set method for vanishing constraints. In order to reduce the complication of the quadratic terms in the mixed state-control constraint, a new efficient reformulation for SwOCP is proposed to linearized these vanishing constraints.

Third Approach: Switching Point Algorithm

To treat the switches on each interval after multiple shooting method is applied, a Switching Point Algorithm is proposed by handling the discontinuities in ODE of SwOCP.

Numerical Studies

We show several numerical examples to perform effective approaches to solving SwOCP.

The first example concerns the New York Subway problem.

The second example considers the Flat Hybrid Automaton with DC electrical network.

The third and the fourth examples, respectively, deal with OCP with dry friction problems: material points on a straight line, and a mass point on a rough plane.

1.2 Dissertation Outline

This dissertation contributes four major parts, which consider the indirect approach for SwOCP, the direct approaches for SwOCP, determination of switches in SwOCP, and SwOCP with dry friction. The dissertation's structure is organized as follows.

The introduction is followed by Chapter 2, where we recall some needed elements of mathematical background and the states of the art to investigate SwOCP.

The first part deals to the indirect approach for SwOCP. In Chapter 3, to solve SwOCP effectively, a new reformulation for SwOCP with the irregular mixed constraints will be proposed, then they are treated by LMP. The optimality conditions here can be exploited to treat the integer controls in SwOCP, which similar to the concept of the Competing Hamiltonian algorithm. Furthermore, the convexification of the velocity set will be studied by using Filippov's rule.

The second part of the dissertation considers direct methods for solving SwOCP. In Chapter 4, we present a solution approach for SwOCP based on Filippov's rule reformulation, together with an expansion of the rounding scheme, which deals with a neighboring feedback law. Subsequently, a feedback algorithm is proposed after using the condensing procedure to explore the block structure of the QP subproblem. On the other hand, we consider the active set method for vanishing constraints to obtain the optimality conditions for comparing with the previous one.

The third part is Chapter 5 to determine switches, including a switching point algorithm. Therein, the derivative generation with variational differential equations is calculated.

In the fourth part of the dissertation, we consider some problems of SwOCP with dry friction in Chapter 6. The general framework with both indirect and direct approaches is to study the OCP with dry friction of a system of material points in a straight line and to investigate the optimal control of a point mass on a rough plane.

In this end, we conclude the dissertation with a summary and an outlook in Chapter 7.

In Appendix A, we collect all auxiliary results, which include the Competing Hamiltonian algorithm, the sliding regime for OCP, the LP in Maximum Principle, and some examples with numerical results.

Finally, Appendix B gathers some open problems regarding the Gröbner basis approach, and an idea about over-under estimating.

1.3 Computational Environment

All computational results and times presented in this dissertation have been obtained on a `64-bit Ubuntu 22.04.1 LTS` system powered by an `Intel Core i7-8700 CPU` @ `3.2GHz`×`12`, with `32 GB` main memory available; and all source code is written in `MATLAB R2021b` and `C++`.

Chapter 2

State of the Art

This chapter presents the nomenclature and terminology used throughout this dissertation by providing and mathematical background knowledge. Various aspects of the analysis, the implicit function theorem, and nonlinear optimization are mentioned. The most important parts here are the state of the art for SwOCP including FILIPPOV's theory, the outer reformulations, disjunctive programming, rounding schemes. For OCPs with LMP, the discussion includes feedback control, solution approaches via indirect methods as well as direct shooting methods with control discretization and sensitivity generation.

2.1 Mathematical Background

The section begins with some basic function definitions that needed to define SwOCP, and the concepts of positively-linearly independence and special ordered set are introduced. Subsequently, the Gaussian elimination algorithm is considered to prepare the direct approach for SwOCP, the implicit function theorem and the definition of functions of bounded variation are stated for the later work with the indirect approach for SwOCP.

2.1.1 Some Elements of Analysis

Definition 1. [130, Def. 4.13] Consider a map $f : D(f) \subseteq X \times Y \to Z$ by $(x, y) \mapsto f(x, y)$, where X, Y and Z are Banach spaces.
Let y be fixed and set $g(x) = f(x, y)$. If g has an derivative at x, then we define the *partial* derivative of f at (x, y) with respect to the variable x to be $f_x(x, y) = g'(x)$.
The derivative $f_y(x, y)$ is defined similarly.

Lemma 1 (**Basic Theorems of Differential Calculus - Partial Derivatives**). *Let $f : X \times Y \to Z$ is differentiable at (x^*, y^*), then the partial derivatives $f'_x(x^*, y^*)$ and $f'_y(x^*, y^*)$ exist at (x^*, y^*). Furthermore, it holds for all $x \in X$ and $y \in Y$ that*

$$f'(x^*, y^*)(x, y) = f'_x(x^*, y^*)(x) + f'_y(x^*, y^*)(y).$$

Proof. See [130, Prop. 4.14]. □

Definition 2 (**Monotone Function**). Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We call f *a monotone increasing function* if $f(t_1) \leq f(t_2)$ for any $t_1 < t_2$. The function f is called *a monotone*

decreasing function if $f(t_1) \geq f(t_2)$ for any $t_1 < t_2$. We call f *a monotone function* if it is either monotone increasing or monotone decreasing.

Theorem 1 (**Properties of Monotone Function**). *Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. Then $f(t^+)$ and $f(t^-)$ exist and are finite for all $t \in \mathbb{R}$. Moreover, for all $t \in \mathbb{R}$, it holds that*

(i) $f(t^-) \leq f(t) \leq f(t^+)$ if f is monotone increasing,

(ii) $f(t^-) \geq f(t) \geq f(t^+)$ if f is monotone decreasing.

The limits $f(\infty^-)$ and $f((-\infty)^+)$ also exist, but are not necessarily finite.

Proof. See [31]. □

Corollary 1. Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. Then

(i) $f(a^+) \leq f(b^-)$ if f is monotone increasing and $a, b \in \bar{\mathbb{R}}$ with $a < b$.

(ii) $f(a^+) \geq f(b^-)$ if f is monotone decreasing and $a, b \in \bar{\mathbb{R}}$ with $a < b$.

Proof. See [31]. □

The following result is about reveals differentiability properties of monotone function.

Theorem 2 (**Lebesgue**). *A monotone function $f : [a, b] \to \mathbb{R}$ has a finite derivative almost everywhere on $[a, b]$.*

Proof. See [80, Thm. 6]. □

According to Theorem 1 the values $f(t^-), f(t), f(t^+)$ all exist for any t, if f is a monotone function. Hence, the only discontinuities that a monotone function can have are jump discontinuities.

Definition 3 (**Jump Discontinuity**). A function $f : \mathbb{R} \to \mathbb{R}$ is said to have a *jump discontinuity* at t if the following conditions hold

(i) the value $f(t^-)$, $f(t)$ and $f(t^+)$ all exist and are finite,

(ii) $f(t^-)$, $f(t)$ and $f(t^+)$ are not all equal.

Theorem 3. *Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. The set of points at which f is discontinuous is either empty, finite, or countably infinite.*

Proof. See [31]. □

Definition 4 (**Jump of a Function**). Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. The *jump of function f* at $t \in \mathbb{R}$ is defined as

$$\Delta_f(t) := f(t^+) - f(t^-).$$

Definition 5 (**Càdlàg function**). [118, def. 2.10] A function $f : [t_0, t_f] \to \mathbb{R}^d$ is said to be *Càdlàg* if it is right-continuous with left limits, i.e., for each $t \in [t_0, t_f]$ the limits

$$\lim_{s \to t, s<t} f(s) \text{ and } \lim_{s \to t, s>t} f(s) \tag{2.1}$$

exist and $f(t) = \lim_{s \to t, s>t} f(s)$.

2.1.2 Positively-Linearly Independence and Special Ordered Set

In this subsection, two properties are defined: positively-linearly independence, and special ordered set of type one and two.

Definition 6. A system consisting of two tuples of vectors $p_1, \ldots, p_m$ and $q_1, \ldots, q_k$ in the space $\mathbb{R}^{n_r}$ is said to be *positively-linearly independent* if there does not exist a nontrivial tuple of multipliers $a_1, \ldots, a_m$, $b_1, \ldots, b_k$ with all $a_i \geq 0$, $i = 1, \ldots, m$, such that

$$\sum_{i=1}^{m} a_i p_i + \sum_{j=1}^{k} b_j q_j = 0.$$

Remark 1. The content "positively-linearly independent" holds an important role when applying to the regular characteristic of the mixed state-control constraints, see the work of Dubovitskii and Milyutin [48], or later by Dmitruk and Osmolovskii [45].

Definition 7. We say that the variables $(\omega_1, \ldots, \omega_2)$ fulfill the special ordered set type one property (SOS-1) if they satisfy

$$\sum_{i=1}^{n} \omega_i = 1, \quad \omega_i \in \{0, 1\}, \quad 1 \leq i \leq n.$$

If they fulfill

$$\sum_{i=1}^{n} \omega_i = 1, \quad \omega_i \in [0, 1], \quad 1 \leq i \leq n,$$

and at most two of the ω_i are nonzero and if so, they are consecutive, then $(\omega_1, \ldots, \omega_2)$ is said to have the SOS type two property (SOS-2).

Remark 2. SOS-1 restrictions will occur automatically after the convexifications. When nonlinear functions are approximated by piecewise linear functions, SOS-2 restrictions will typically occur.

2.1.3 Block Gaussian Elimination

Consider a system $Az = b$ where the matrix A is of dimension $m \times m$ with $m = pn$ and the vectors b and z are of dimension m, with $m, n, p \in \mathbb{N}$. The matrix and the vectors can be partitioned into:

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \ldots & A_{1,n} \\ A_{2,1} & A_{2,2} & \ldots & A_{2,n} \\ \vdots & \vdots & & \vdots \\ A_{n,1} & A_{n,2} & \ldots & A_{n,n} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

where the block $A_{i,j}$ are matrices of dimension $p \times p$ and b_i are vectors of dimension p. The block version of Gaussian elimination, cf. [66, 67], is written as below.

Algorithm 1 (**Gaussian Elimination**).

1. *Forward Elimination:*
   Loop on k from $k = 1, \ldots, n-1$
   - *if* $A_{k,k}$ *is singular, set singularity indicator, exit and calculate* $X = A_{k,k}^{-1}$
   - *Loop on i from* $k+1, \ldots, n$
     i. $T := A_{i,k}X$
     ii. *Loop on j from* $k+1, \ldots, n$
     $A_{i,j} := A_{i,j} - T * A_{k,j}$
     iii. *End Loop on j*
     iv. $b_i := b_i - T * b_k$
   - *End Loop on i*

   End Loop on k

2. *Back substitution:*
   Loop on i from $n, \ldots, 1$
   - $x_i := b_i$
   - *Loop on j from* $i+1, \ldots, n$
     $z_i := z_i - A_{i,j} * z_j$
   - *End Loop on j*
   - $z_i := A_{i,i} - Iz_i$

   End Loop on i

2.1.4 The Implicit Function Theorem

The implicit function theorem has various important roles in local convergence theory of optimization algorithms. In [97, Thm. A.1], Nocedal and Wright present an version of one with Lipschitz continuously, while in this dissertation, we follow the work of Zeidler, cf. [130, Sec. 4.7].
We want to solve the equation

$$F(x, y) = 0, \tag{2.2}$$

which has a given point solution, $F(x_0, y_0) = 0$, for y in a neighborhood of (x_0, y_0), i.e., we want to find a mapping $x \mapsto y(x)$ such that $y(x_0) = y_0$ and $F(x, y(x) = 0$ (See [130, Fig. 4.2]). The determinative condition for the existence of a unique solution is the following:

$$\textit{The inverse operator, } F_y(x_0, y_0)^{-1} : Z \to Y, \textit{ exists as a continuous linear operator.} \tag{2.3}$$

Since Y and Z are Banach spaces, this condition is equivalent to the following

$$\textit{The partial derivative } F_y(x_0, y_0) : Y \to Z \textit{ is bijective.} \tag{2.4}$$

The underlying concept is to rewrite equation (2.2) in the equivalent form

$$y - y_0 = (y - y_0) - F_y(x_0, y_0)^{-1} F(x, y). \tag{2.5}$$

If we write F as a classical power series, we obtain the form

$$F(x,y) = F(x_0,y_0) + a(x-x_0) + b(y-y_0) + \text{ higher-order terms},$$

now note that $F(x_0,y_0) = 0$ and $F_y(x_0,y_0) = b$. Thus, the initial equation $F(x,y) = 0$ for $x, y \in \mathbb{R}$ is equivalent to

$$y - y_0 = -b^{-1}a(x-x_0) + \text{ higher-order terms}.$$

This corresponds exactly to Eq. (2.5). The key condition (2.3) guarantees the existence of the inverse b^{-1}. This makes it clear that the right hand side of (2.5) is of first order with respect to the small parameter $(x - x_0)$, and of second order with respect to $(y - y_0)$.

Theorem 4 (**Implicit Function Theorem** of Hildebrant and Graves (1927))**.** [130, Thm. 4.B] *Suppose that:*

(i) the mapping $F : U(x_0,y_0) \subseteq X \times Y \to Z$ is defined on an open neighborhood $U(x_0,y_0)$, and $F(x_0,y_0) = 0$, where X, Y and Z are Banach spaces over $\mathbb{R}$ or $\mathbb{C}$,

(ii) F_y exists as a partial derivative on $U(x_0,y_0)$ and condition (2.4) holds,

(iii) F and F_y are continuous at (x_0,y_0).

Then the following are true:

(a) Existence and uniqueness. There exist positive numbers r_0 and r such that for every $x \in X$ satisfying $\|x - x_0\| \leq r_0$, there exists exactly one $y(x) \in Y$ for which $\|y(x) - y_0\| \leq r$ and $F(x,y) = 0$.

(b) Construction of the solution. The sequence $(y_n(x))$ of successive approximations, defined by $y_0(x) \equiv y_0$, and

$$y_{n+1}(x) = y_n(x) - F_y(x_0,y_0)^{-1}F(x,y_n(x)),$$

converges to the solution $y(x)$, as $n \to \infty$, for all points $x \in X$ satisfying $\|x - x_0\| \leq r_0$.

(c) Continuity. If F is continuous in a neighborhood of (x_0,y_0), then $y(\cdot)$ is continuous in a neighborhood of x_0.

(d) Continuous differentiability. If F is a C^m-map, $1 \leq m \leq \infty$, on a neighborhood of (x_0,y_0), then $y(\cdot)$ is also a C^m-map on a neighborhood of x_0.

Proof. See [130, Thm. 4.B, pp. 152]. □

Remark 3. In particular,

$$y'(x) = F_y(x,y(x))^{-1}F_x(x,y(x)) \tag{2.6}$$

for all x in a suitable open neighborhood of x_0.

2.1.5 Functions of Bounded Variation and Measure

Denote by $BV([t_0,t_f],\mathbb{R}^n)$ the space of functions $\lambda : [t_0,t_f] \to \mathbb{R}^n$ of bounded variation which have also values $\lambda(t_0-)$ and $\lambda(t_f+)$, independent of the values on segment $[t_0,t_f]$, cf. [47, sec. 3].
The *jump* of f at a point $t \in [t_0,t_f]$ is defined by the vector $[\lambda](t) := \lambda(t+) - \lambda(t-)$. In particular,

$$[\lambda](t_0) := \lambda(t_0+) - \lambda(t_0-), \quad [\lambda](t_f) := \lambda(t_f+) - \lambda(t_f-).$$

Any function $\lambda \in BV$ determines a Lebesgue-Stieltjes measure $d\lambda$ which satisfies: for any $[t_1,t_2] \subset [t_0,t_f]$,

$$\int_{[t_1,t_2]} d\lambda(t) = \lambda(t_2+) - \lambda(t_1-). \tag{2.7}$$

In particular, $\int_{[t_0,t_f]} d\lambda(t) = \lambda(t_f+) - \lambda(t_0-)$.
Let distinguish measures $d\lambda \in C^*$ from the functions of bounded variation $\lambda \in BV$ that define them. As is well known,

$$\|d\lambda\|_{C^*} = \int_{[t_0,t_f]} |d\lambda(t)|, \text{ and } \|\lambda\|_{BV} = |\lambda(t_0-) + \|d\lambda\|_{C^*}.$$

Furthermore, $\|\lambda\|_\infty = \max\{\operatorname*{ess\,sup}_{[t_0,t_f]}|\lambda(t)|,\ |\lambda(t_0-)|,\ |\lambda(t_f+)|\} \le \|\lambda\|_{BV}$, where "ess sup" stands for essential supremum (sometimes denoted by "vrai max"),

$$\operatorname*{ess\,sup}_{[t_0,t_f]}|\lambda| = \inf_{c\in\mathbb{R}} c, \text{ such that } \mu(\{t : |\lambda(t)| > c\}) = 0 \text{ (measure zero)}.$$

If a function $\lambda \in BV$ is absolutely continuous (hence $\lambda(t_0-) = \lambda(t_0)$ and $\lambda(t_f+) = \lambda(t_f)$), then the measure $d\lambda$ is also called *absolutely continuous*. In this case, there exists $\dot\lambda \in L_1[t_0,t_f]$ such that

$$d\lambda(t) = \dot\lambda(t)dt, \text{ and } \|\lambda\|_{BV} = |p(t_0)| + \int_{t_0}^{t_f} |\dot\lambda(t)|dt.$$

2.2 Nonlinear Optimization

This section introduces a nonlinear program in general form and discusses some constraint qualifications (CQs). The first and second order optimality conditions are also introduced.

Definition 8 (NLP). An optimization problem of the general form

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & g(x) = 0, \\ & h(x) \ge 0. \end{array} \tag{2.8}$$

with the objective function $f : \mathbb{R}^n \to \mathbb{R}$, equality constraints $g : \mathbb{R}^n \to \mathbb{R}^{n_g}$, and inequality constraints $h : \mathbb{R}^n \to \mathbb{R}^{n_h}$ is called a *Nonlinear Program* (therein $f, g, h \in \mathcal{C}^2$ w.r.t. x).

The feasible set of NLP (2.8) is defined as follows

$$\mathcal{F} \overset{def}{=} \{x \in \mathbb{R}^n \mid g(x) = 0, h(x) \geq 0\} \subseteq \mathbb{R}^n.$$

In optimization, the points of concern are the feasible points that minimize the objective function.

Definition 9 (**Global minimum**). A point $x^* \in \mathbb{R}^n$ is a *global minimizer* if and only if $x^* \in \mathcal{F}$ and $\forall x \in \mathcal{F} : f(x) \geq f(x^*)$. The value $f(x^*)$ is called the *global minimum*.

However, finding the global minimum is usually difficult, and most algorithms only allow us to obtain local minimizers and verify optimality locally.

Definition 10 (**Local minimum**). $x^* \in \mathbb{R}^n$ is a *local minimizer* if and only if $x^* \in \mathcal{F}$ and there exists a neighborhood U of x^* so that $\forall x \in \mathcal{F} \cap U : f(x) \geq f(x^*)$. The value $f(x^*)$ is called a *local minimum*.

To check if a candidate x^* is a local minimizer or not, we need to consider about the optimality conditions to describe the feasible set in the neighborhood of x^*. It means that not all inequality constraints need to be considered locally, but only the *active* ones.

Definition 11 (**Active constraint, active set**). Let $\bar{x} \in \mathbb{R}^n$ be a feasible point of problem (2.8). An inequality constraint $h_i(x) \geq 0, i \in \{1, \ldots, n\} \subset \mathbb{N}$, is called *active* at $\bar{x}$ if $h_i(\bar{x}) = 0$ holds. It is called inactive otherwise. Set of indices of all active constraints

$$\mathcal{A}(\bar{x}) \overset{def}{=} \{i \mid h_i(\bar{x}) = 0\} \subseteq \{1, \ldots, n_h\} \subset \mathbb{N}. \tag{2.9}$$

is called the *active set* associated with $\bar{x}$.

Remark 4. We often required that the set of active constraints to be linear independent.

Definition 12. The restriction of the inequality constraint function h onto the active inequality constraints is denoted by

$$\begin{aligned} h_{\mathcal{A}} : \mathbb{R}^n &\to \mathbb{R}^{|\mathcal{A}|} \\ x &\to h_{\mathcal{A}}(x). \end{aligned}$$

Definition 13. Let $\bar{x} \in \mathbb{R}^n$ be a feasible point of NLP (2.8). We state the definition of the tangent cone $\mathcal{T}(\bar{x}, \mathcal{F})$ of $\mathcal{F}$ in the point $\bar{x}$ as

$$\mathcal{T}(\bar{x}, \mathcal{F}) \overset{\text{def}}{=} \left\{ d \in \mathbb{R}^n \mid \exists \{x^k\} \subseteq \mathcal{F}, \{t^k\} \to 0^+ : x^k \to \bar{x}, \frac{1}{t^k}(x^k - \bar{x}) \to d \right\},$$

and the linearized cone $\mathcal{L}(\bar{x})$ of problem (2.8) in $\bar{x}$ as

$$\mathcal{L}(\bar{x}) \overset{\text{def}}{=} \left\{ d \in \mathbb{R}^n \mid \nabla G(\bar{x})^T d = 0, \nabla h_{\mathcal{A}}\left(\bar{x}\right)^T d \geq 0 \right\},$$

where $h_{\mathcal{A}} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{A}|}$ is defined in Def. 12.

Now we consider some constraint qualifications for problem (2.8) in $\bar{x}$. In general, a constraint qualification is a property of the feasible set represented by the constraint functions, which guarantees that the KKT conditions are in fact necessary optimality conditions. Three of the most common ones are considered, see Def. 14, Def. 15 and Def. 16, as follows.

Definition 14. (**Linear Independence Constraint Qualification, Regular Point**)
We say that *Linear Independence Constraint Qualification* (LICQ) holds for (2.8) in $\bar{x} \in \mathbb{R}^n$ if it holds that

$$rank \begin{pmatrix} \nabla g(\bar{x}) & \nabla h_{\mathcal{A}}(\bar{x}) \end{pmatrix}^T = n_g + n_{h_{\mathcal{A}}}. \tag{2.10}$$

And, $\bar{x}$ is referred to as a *regular point* of (2.8).

Denoting $\tilde{g}(x) := \begin{pmatrix} g(x) & h_{\mathcal{A}}(x) \end{pmatrix}^T$. Now we can see further meaning to the LICQ condition, i.e., LICQ is equivalent to full row rank of the Jacobian matrix $\nabla \tilde{g}(\bar{x})$.

Remark 5. LICQ holds at $\bar{x} \Leftrightarrow g(\bar{x}) = 0, h(\bar{x}) = 0$. It means that $\bar{x}$ lies on the border/boundary of the feasible set $\mathcal{F}$.

Definition 15. (**Mangasarian-Fromovitz Constraint Qualification**)
Mangasarian-Fromovitz Constraint Qualification (MFCQ) holds in $\bar{x}$ if the Jacobian $g_x(\bar{x})$ has full rank and there exists $d \in \mathbb{R}^n$ such that

$$\nabla g(\bar{x})^T d = 0, \text{ and } \nabla h_{\mathcal{A}}(\bar{x})^T d > 0. \tag{2.11}$$

Definition 16. (**Abadie Constraint Qualification**)
Abadie Constraint Qualification (ACQ) holds in $\bar{x}$ if

$$\mathcal{T}(\bar{x}, \mathcal{F}) = \mathcal{L}(\bar{x}), \tag{2.12}$$

where the definitions of the Bouligand tangent cone $\mathcal{T}(\bar{x}, \mathcal{F})$ of the set $\mathcal{F}$ in the point $\bar{x}$ and the linearized cone $\mathcal{L}(\bar{x})$ of problem (2.8) in $\bar{x}$ can be seen in Def. 13.

Remark 6. LICQ $\Rightarrow$ MFCQ $\Rightarrow$ ACQ, whereas the converse never holds, where counterexamples, e.g., can be found in [104, Appx. C]. Furthermore, readers can see in [69], where the MFCQ is not satisfied, but the ACQ holds under some assumptions.

2.2.1 Quadratic Optimization

The quadratic expansion of NLP (2.8) as the SQP reads

$$\begin{aligned} \min_{\Delta x \in \mathbb{R}^n} \quad & \tfrac{1}{2} \Delta x^T \tfrac{\partial^2 L(x,\lambda,\mu)}{\partial x^2} \Delta x + \nabla f(x) \Delta x \\ \text{s.t.} \quad & g(x) + \nabla g(x) \Delta x = 0, \\ & h(x) + \nabla h(x) \Delta x \geq 0. \end{aligned} \tag{2.13}$$

where $L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x)$.

2.2.2 First Order Optimality Conditions

An important question is if a feasible point $x^* \in \mathcal{F}$ satisfies necessary first order optimality conditions. If it satisfies these conditions, x^* is a candidate for a local minimizer. It it does not satisfy these condition, it cannot be a local minimizer. The first order condition, i.e., the KKT conditions, can only be formulated if a "constraint qualification" is satisfied.

Theorem 5 (**KKT conditions**, [74, 85])**.** *If x^* is a local minimizer of the NLP (2.8) and LICQ holds at x^* then there exist so called multiplier vectors $\lambda^* \in \mathbb{R}^{n_g}$ and $\mu^* \in \mathbb{R}^{n_h}$ with*

$$\nabla f(x^*) + \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^* = 0 \tag{2.14a}$$
$$g(x^*) = 0 \tag{2.14b}$$
$$h(x^*) \geq 0 \tag{2.14c}$$
$$\mu^* \geq 0 \tag{2.14d}$$
$$\mu_i^* h_i(x^*) = 0, \quad i = 1, \ldots, n_h. \tag{2.14e}$$

Proof. See [97, Sec. 12.4]. □

In the case of convex problems, the KKT conditions are not only *necessary* for a *local* minimizer, but also *sufficient* for a *global* minimizer. The Lagrangian function and the complementarity are considered as follows.

Definition 17 (**Lagrangian Function**)**.** We define the so called "Lagrangian function" to be

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x).$$

Here, $\lambda \in \mathbb{R}^{n_g}$ and $\mu \in \mathbb{R}^{n_h}$ are the so called "LAGRANGE multipliers" (or "dual variables"). Since the KKT conditions and the definition of the Lagrangian, we have (2.14a)$\Leftrightarrow \frac{\partial \mathcal{L}(x^*, \lambda^*, \mu^*)}{\partial x} = 0$.

Remark 7 (**Complementarity**)**.** The last three KKT condition (2.14c)-(2.14e) are called the *complementarity* conditions.
For each index i, if $h_i(x^*) = 0$ and $\mu_i^* = 0$ then this is called a weakly active constraint. On the other hand, an active constraint with $\mu_i^* > 0$ is called strictly active.

Definition 18. Consider a KKT point (x^*, λ^*, μ^*). We say that *strict complementarity* holds at this KKT point if and only if all active constraints are strictly active.

2.2.3 Second Order Optimality Conditions

Theorem 6 (**Second Order Optimality Conditions**)**.** *Let us regard a point x^* at which LICQ holds together with multipliers λ^*, μ^* so that the LICQ conditions (2.14a)-(2.14e) are satisfied and let strict complementarity hold. Regard a basis matrix $Z \in \mathbb{R}^{n \times (n - n_{\tilde{g}})}$ of the null space of $\nabla \tilde{g}(x^*) \in \mathbb{R}^{n_{\tilde{g}} \times n}$, i.e., Z has full column rank and $\nabla \tilde{g}(x^*) \in \mathbb{R}^{n_{\tilde{g}} \times n} Z = 0$. Then the following two statements hold:*

(a) If x^* is a local minimizer, then $Z^T \frac{\partial^2 \mathcal{L}(x^*, \lambda^*, \mu^*)}{\partial x^2} Z \succeq 0$.
(Second Order Necessary Condition)

(b) If $Z^T \frac{\partial^2 \mathcal{L}(x^*, \lambda^*, \mu^*)}{\partial x^2} Z \succ 0$, then x^* is a local minimizer.
This minimizer is unique in its neighborhood, i.e., a strict local minimizer, and stable against small differentiable perturbations of the problem data.
(Second Order Sufficient Condition)

Proof. See [97] on pages 332 and 333 for statements (a) and (b), respectively. □

The matrix $\frac{\partial^2 \mathcal{L}(x^*, \lambda^*, \mu^*)}{\partial x^2}$ is called the *Hessian of the Lagrangian*, while its projection on the null space of the Jacobian, $Z^T \frac{\partial^2 \mathcal{L}(x^*, \lambda^*, \mu^*)}{\partial x^2} Z$, is called the *reduced Hessian*.

2.3 Basic Theory of Optimal Control Problems

In this section, the basic results in optimal control that needed for solving SwOCP will be derived in the general nonlinear setting. We start by considering nonlinear control systems in continuous time

$$\dot{x}(t) = f(x(t), u(t)), \tag{2.15}$$

where control function $u(t)$ with values in $U \subset \mathbb{R}^m$, function $f : \mathbb{R}^n \times U \to \mathbb{R}^n$. Existence and uniqueness is then delivered by the well-known Theorem of Carathéodory as follows.

Theorem 7 (**Theorem of Carathéodory**,[116])**.** *Consider a control system with the following properties:*

(i) The space of control functions is given by

$$\mathcal{U} := \{u : \mathbb{R} \to U \mid u \text{ is measureable and bounded}\}.$$

(ii) The vector field $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ is continuous.

(iii) For any $R > 0$ there exists a constant $L_R > 0$ such that the condition

$$\|f(x_1, u) - f(x_2, u)\| \leq L_R \|x_1 - x_2\|$$

holds for all $x_1, x_2 \in \mathbb{R}^n$ and all $u \in U$ with $\|x_1\|, \|x_2\|, \|u\| \leq R$.

Then for any initial value $x_0 \in \mathbb{R}^n$, any initial time $t_0 \in \mathbb{R}$, and any control function $u \in \mathbb{U}$, there exists a maximal open interval I with $t_0 \in I$ and a unique absolutely continuous function $x(t)$, which solves the following integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f(x(\tau), u(\tau)) d\tau$$

for all $t \in I$.

Proof. The proof of Theorem 7 can be found in the book [116, Appendix C]. □

Definition 19. Denote the unique function $x(t)$ from Theorem 7 with $x_u(t; t_0, x_0)$ and call it the *solution* of (2.15) with *initial value* $x_0 \in \mathbb{R}^n$ and control function $u \in \mathcal{U}$.

Remark 8. In case $t_0 = 0$, we briefly write $x_u(t, x_0) = x_u(t; 0, x_0)$. Since $x_u(t, x_0)$ is absolutely continuous, it is differentiable w.r.t. t for almost all $t \in I$. In particular, Theorem 7 and the fundamental theorem of calculus imply that $x_u(t, x_0)$ satisfies (2.15) for almost all $t \in I$, i.e.,

$$\dot{x}(t, x_0, u) = f(x(t, x_0, u), u(t))$$

holds for almost all $t \in I$.

Now the OCP is defined in the following subsection.

2.3.1 Formulation of OCP

In the following definitions, we state the cost function $\varphi(\cdot,\cdot)$ in BOLZA type, which includes the MEYER term $m(\cdot)$ and the LAGRANGE term $\int_{t_0}^{t_f} l(\cdot,\cdot)dt$. The ODE and the constraints are also introduced.

Definition 20. For continuous *cost function* $l : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and $m : \mathbb{R}^n \to \mathbb{R}$, we define the *cost functional*

$$\varphi(x,u) := m(x(t_f)) + \int_{t_0}^{t_f} l(x(t),u(t))dt. \tag{2.16}$$

Then the OCP is given by the optimization problem

$$\text{minimize } \varphi(x,u) \text{ with respect to } u \in \mathcal{U} \text{ for each } x \in \mathbb{R}^n.$$

The function

$$V(x) := \inf_{u\in\mathcal{U}} \varphi(x,u)$$

is called the *optimal value function* of this OCP. A pair $(x^*,u^*) \in \mathbb{R}^n \times \mathcal{U}$ with $\varphi(x^*,u^*) = V(x^*)$ is called *optimal solution*, or rarely *optimal pair*.

Definition 21. The ordinary differential equation (ODE) with initial value of $x(\cdot)$ is defined

$$\dot{x}(t) = f(x(t),u(t)), \quad x(t_0) = x_0. \tag{2.17}$$

Definition 22. The point constraints are introduced as

$$r(x(t_0),x(t_f)) \geq 0. \tag{2.18}$$

For the path constraints, see in Section 2.4.

2.3.2 The Existence of Solution of OCP

We consider the optimal control problem (OCP) where the control functions belong to the class of bounded measurable functions. In the 1960s, FILIPPOV [55, Sec. I] or ROXIN [110] proved that there exists a solution of this kind of OCP in Subsection 2.3.1. More foundational investigations on the conditions of the control functions, i.e., piecewise smoothness control or continuous control, and also the general formulation of the OCP, are early discussed in [55, Sec. II,III,IV].

2.3.3 Solution Approaches

To solve an OCP, the function space approach, which considers to solve the OCP as an infinite dimensional optimization problem. The function space is also well-known as the classical *indirect approach*, or as *first optimize, then discretize approach*. For more investigations, readers can find in Chapter 3.
While on other side, a suitable discretization scheme can be applied to transform the OCP into a finite dimensional optimization problem. This discretization approach is also known as the so-called *direct approach*, which is based on the *first-discrete-then-optimize* paradigm. Detailed researches for the direct approach are done in Chapter 4 and Chapter 5.

Indirect methods

In indirect methods, the necessary or first-order optimality conditions are exploited to a nonlinear multipoint boundary value problem (BVP) that has to be tackled by a minimizer. The BVP is solved by using multiple shooting, cf. [15]. Exploitation of the maximum principle is usually not accomplished automatically by an algorithm and must be specified by user. The challenged tasks are to deal with the constraints, usually mixed state-control constraints which consist discontinuous state, discontinuous control, and may be jumps in the adjoint variables, cf. [14], and furthermore the constraints can be active or inactive. Especially mixed state-control constraints are commonly provided the transition from one stage to another stage, or from one arc type to another model of the right hand side function in the ODE. This transition is considered by the so-called *switching conditions* $\sigma(x(t_{sw}), u(t_{sw})) = 0$ where t_{sw} is a switching time.
The maximum principle is formulated in the weak minimum conditions in Chapter 3. The regularity of the constraint qualification is discussed, and then the local maximum principle is stated, respectively. Moreover, a general scheme for using Filippov's rule (or using the second time of convexification), cf. Subsection 3.1.3, and a neighboring feedback law to investigate the feedback control, cf. Subsection 2.6.8, are proposed.

Direct methods

Instead of formulating optimality conditions like the indirect methods, the direct approaches transcribe the original (finite or infinite dimensional) optimization problem into a finite dimensional NLP and then the resulting problem is solved effectively by the numerical methods such as interior point methods or SQP algorithm, cf. Section 4.1.4.
In Chapter 4, the resulting QP Subproblem comes out after Filippov's reformulation, relaxation and direct multiple shooting method, will be solved with SQP algorithm. Furthermore, to reduce the rambling behavior of the rounded control, the switching aware rounding algorithm is presented, cf. Section 4.2, and also the CIA is considered to track the average of a relaxed solution over a given rounding grid by a piecewise constant integer control and to minimize the integrity error.
While later, Chapter 5 will deals with the switching point algorithm for the resulted discretized multiple shooting QP Subproblem.

2.3.4 Maximum Principle

This subsection follows the work of Pontryagin et al., cf. [25, 106] to formulate the Maximum Principle for the OCP, where the ODE is defined in Eq.(2.17). For more general OCP, which includes path and point constraints, readers can see in the survey paper [63].

Definition 23. The *control theory Hamiltonian* is the function

$$H(x(t), \lambda(t), u(t)) := \lambda(t)^T f(x(t), u(t)) + l(x(t), u(t)), \qquad x, \lambda \in \mathbb{R}^n,\ u \in \mathcal{U}. \tag{2.19}$$

Theorem 8 (**Pontryagin Maximum Principle**). [25, 106] *Assume $(x^*(\cdot), u^*(\cdot))$ is optimal for OCP with ODE (2.17), cost function (2.16). Then there exists a function $\lambda^* : [t_0, t_f] \to \mathbb{R}^n$ such that*

$$\dot{x}^*(t) = \frac{\partial H(x^*(t), \lambda^*(t), u^*(t))}{\partial \lambda}, \tag{2.20}$$

$$\dot{\lambda}^*(t) = -\frac{\partial H(x^*(t), \lambda^*(t), u^*(t))}{\partial x}, \tag{2.21}$$

and

$$H(x^*(t), \lambda^*(t), u^*(t)) = \max_{u \in \mathcal{U}} H(x^*(t), \lambda^*(t), u), \quad t_0 \leq t \leq t_f. \tag{2.22}$$

Also,

$$H(x^*(t), \lambda^*(t), u^*(t)) \equiv 0, \quad t_0 \leq t \leq t_f.$$

Remark 9. Note that in the above theorem, Theorem 8, the end time point t_f is fixed. In the case that t_f is free, the terminal condition is added

$$\lambda^*(t_f) = \nabla m(x^*(t_f)), \tag{2.23}$$

and the mapping $t \mapsto H(x^*(t), \lambda^*(t), u^*(t))$ is constant.

One calls $x^*(\cdot)$ the *state trajectory* of the optimality controlled system and $\lambda^*(\cdot)$ the *costate*. The identities (2.21) are the *adjoint equations* and (2.22) the *maximization principle*. Notice that (2.20) and (2.21) resemble the structure of HAMILTON's equation. For a proof of the maximum principle and more references, see, e.g. [25, 106]. Ones also call (2.23) the *transversality condition* and will consider its significance later, see Section 2.6.

2.3.5 Local Maximum Principle

This subsection mostly follows the papers of DMITRUK [43], DMITRUK et. al. [46, 47]. Consider the OCP on a fixed interval of time $[t_0, t_f]$:

$$\begin{aligned}
\min \quad & \varphi\left(x(t), u(t)\right) && \text{(2.24a)}\\
\text{s.t.} \quad & \dot{x}(t) = f(x(t), u(t)),\ t \in [t_0, t_f], && \text{(2.24b)}\\
& g_i(x(t), u(t)) = 0,\ i = 1, \ldots, d(g),\ t \in [t_0, t_f], && \text{(2.24c)}\\
& G_j(x(t), u(t)) \leq 0,\ j = 1, \ldots, d(G),\ t \in [t_0, t_f],, && \text{(2.24d)}\\
& r(x(t_0), x(t_f)) \leq 0,\ t \in [t_0, t_f], && \text{(2.24e)}
\end{aligned}$$

where the functions $\varphi : \mathbb{R}^{n_x+n_u} \to \mathbb{R}$, $\varphi(x, u) := m(x(t_f)) + \int_{t_0}^{t_f} l(x(t), u(t))dt$, $f : \mathbb{R}^{n_x+n_u} \to \mathbb{R}^n$, $g_i : \mathbb{R}^{n_x+n_u} \to \mathbb{R}$, $i = 1, \ldots, d(g)$, and $G_j : \mathbb{R}^{n_x+n_u} \to \mathbb{R}$, $j = 1, \ldots, d(G)$, are continuously differentiable.
Conditions (2.24c), (2.24d) are called *mixed state-control constraints* or shortly *mixed constraints*. According to DUBOVITSKII and MILYUTIN, mixed constraints (2.24c-2.24d) are *regular* if for any point (x, u) satisfying these constraints, the gradients in control u

$$\frac{\partial g_i(x(t), u(t))}{\partial u},\ i = 1, \ldots, d(g), \quad \frac{\partial G_j(x(t), u(t))}{\partial u},\ j \in I(x(t), u(t)),$$

are *positive-linear independent* (see Def. 6), where $I(x(t), u(t)) := \{j \mid G_j(x(t), u(t)) = 0\}$ is the set of active indices for inequality mixed constraints $G \leq 0$ at the given point.

Remark 10. Note that here the state constraint (e.g., $\Phi(x(t)) \leq 0$, cf. [43]) cannot be considered as a special case of the mixed constraints due to regularity assumption.

Definition 24. [43, Def. 2] A pair $(x, u) \in \mathbb{R}^{n_x+x_u}$ is called a *phase point* (of the mixed constraint) if there exists $a \in \mathbb{R}^{d(G)}$, $a \geq 0$, and $b \in \mathbb{R}^{d(g)}$ such that $\sum a_j = 1$, and

$$a^T \frac{\partial G(x,u)}{\partial u} + b^T \frac{\partial g(x,u)}{\partial u} = 0, \quad a^T G(x,u) = 0,$$

i.e., the positive-linear independence fails to hold.
The corresponding vector $s = a^T \frac{\partial G(x,u)}{\partial x} + b^T \frac{\partial g(x,u)}{\partial x}$ is called a *phase jump*.

Remark 11. Note that the set of all phase points is determined only by the mixed state-control constraints and does not depend on the control system nor the endpoint of the OCP.

To state maximum principle, we start by denoting

$$y := (x, u), \quad \bar{n} := n_x + n_u,$$

and $\mathbb{R}^{n*}$ the space of row vectors of the dimension n. Then we introduce the simplex

$$\Delta = \{\gamma \in \mathbb{R}^{d(G)*} : \gamma \geq 0,\ |\gamma| = 1\},$$

where $|\gamma| := \sum_{i=1}^{d(G)} |\gamma_i|$ is the norm of an element γ in the space $\mathbb{R}^{d(G)*}$, and $d(G)$ is dimension of the mixed constraints G.
We also need the definition of *weak minimum*.

Definition 25. [45, Def] An admissible process $(x^*(t), u^*(t))$, $t \in [t_0, t_f]$ is a week minimum for (2.24) if there exists an $\epsilon > 0$ such that for any admissible process $(x(t), u(t))$, $t \in [t_0, t_f]$, satisfying the conditions

$$|x(t) - x^*(t)| \leq \epsilon, \quad |u(t) - u^*(t)| \leq \epsilon, \quad t \in [t_0, t_f],$$

the following inequality holds: $\varphi(x, u) \geq \varphi(x^*, u^*)$.

For any $y \in \mathbb{R}^{\bar{n}}$ satisfying the mixed constraints, i.e. $\mathcal{G}(y) \leq 0$, we define the set

$$\Lambda(y) = \left\{ \gamma \in \Delta :\ \gamma G(y) = 0,\ \gamma \frac{\partial G(y)}{\partial u} + b \frac{\partial g(y)}{\partial u} = 0 \right\}, \quad b \in \mathbb{R}^{d(g)},$$

and the set of phase points of the mixed constraints

$$\mathcal{N}(G) = \left\{ y \in \mathbb{R}^{\bar{n}} : \Lambda(y) \neq \varnothing \right\},$$

Clearly, $\mathcal{N}(G)$ is closed. We assume that $\mathcal{N}(G)$ is nonempty, otherwise the mixed constraints are regular.
Define the following set-valued mapping $y \in \mathbb{R}^{n_r} \rightrightarrows S(y) \in \mathbb{R}^{n*}$:

(i) if $y \in \mathcal{N}(G)$ then $S(y) = \left\{ s = \gamma \frac{\partial G(y)}{\partial x} + b \frac{\partial g(y)}{\partial x} : \gamma \in \Lambda(y) \right\}$,

(ii) if $y \notin \mathcal{N}(G)$ then $S(y) = \varnothing$.

For any nonempty set $M \subset \mathbb{R}^{\bar{n}}$ we define $S(M) = \bigcup_{y \in M} S(y)$.
Let $\hat{y} = (\hat{x}, \hat{u})$ be a given admissible process in problem (2.24) investigated for optimality. Denote for short $\hat{x}_{ini} = (\hat{x}(t_0), \hat{x}(t_f))$. Now we will formulate the conditions of the *local minimum principle* (LMP) for the process $\hat{y}$.
Recall that for the function $\hat{u}$ we introduced the set-valued mapping

$$\text{clm}\,(\hat{u})(t) = \{u \in \mathbb{R}^{n_u} : (t, u) \in \text{clm}\,(\hat{u})\},$$

where clm $(\hat{u})$ is the closure in measure of $\hat{u}$, and recall also that $(\hat{x}(t), \text{clm}\,(\hat{u})(t)) = \text{clm}\,(\hat{y})(t)$ for all $t \in [t_0, t_f]$.
Define a set

$$\mathcal{D} := \{t \in [t_0, t_f] : \text{clm}\,(\hat{y})(t) \cap \mathcal{N}(G) \neq \varnothing\}\,. \tag{2.25}$$

We see that $\mathcal{D}$ is a closed (possibly empty) subset in $[t_0, t_f]$, since the set clm $(\hat{y})$ is compact, and the set $\mathcal{N}(G)$ is closed. Let $\chi_{\mathcal{D}}$ be its characteristic function.

Lemma 2. *The case $\mathcal{D} = \varnothing$ means that the mixed state-control constraints are regular.*

Proof. It is easy to see that the case $\mathcal{D} = \varnothing$ means that the process $\hat{y}$ does not pass "closely" to the set of phase points $\mathcal{N}(G)$, i.e. $\exists \varepsilon > 0$ such that dist $(\hat{y}(t), \mathcal{N}(G)) \geq \text{const} > 0$ a.e. on $[t_0, t_f]$. In fact, the mixed constraints are regular.

□

For any $t \in \mathcal{D}$, consider the set conv $S(\text{clm}\,(\hat{y})(t))$, where "conv" stands for the convex hull.
Now, let us define the *Pontryagin function* and the *endpoint Lagrange function*

$$\mathcal{H}(\lambda, x, u) = \lambda^T F(x, u) + \delta^T l(x(t), u(t)), \quad L(\nu, \hat{x}_{ini}) = \nu^T r(x(t_0), x(t_f)), \tag{2.26}$$

where $\lambda \in \mathbb{R}^{n*}$ is a adjoint (costate) row-vector, $\delta \in \mathbb{R}^{n*}$ and $\nu \in \mathbb{R}^{(1+n_r)*}$ are Lagrange multipliers.
Then we introduce the so-called *augmented Pontryagin function*

$$\bar{\mathcal{H}}(\lambda, \mu, x, u) = \mathcal{H}(\lambda, x, u) + (\mu^G)^T G(x, u) + (\mu^g)^T g(x, u),$$

where $\mu \in \mathbb{R}^{d(G)*}$, $\mu^g \in \mathbb{R}^{d(g)*}$.
The conditions of LMP at the point $\hat{y}$ are as follows: there exist multipliers

$$\hat{\nu} \in \mathbb{R}^{(1+n_r)*}, \quad \hat{\lambda} \in BV\,([t_0, t_f]\mathbb{R}^{n*})\,, \tag{2.27}$$

$$\hat{\mu}^G \in L^1\left([t_0, t_f], \mathbb{R}^{d(G)*}\right), \quad \hat{\mu}^g \in L^1\left([t_0, t_f], \mathbb{R}^{d(g)*}\right), \quad d\hat{\eta} \in (C([t_0, t_f], \mathbb{R}))^*\,, \tag{2.28}$$

such that

$$\hat{\nu} \geq 0, \quad \hat{\nu} r(\hat{x}_{ini}) = 0,$$

$$\hat{\mu}^G \geq 0, \quad \hat{\mu}^G G(\hat{y}) = 0, \quad d\hat{\eta} \geq 0, \quad d\hat{\eta}\ \chi_{\mathcal{D}} = d\hat{\eta},$$

$$|\hat{\nu}| + \left\|\hat{\mu}^G\right\|_1 + \int_{[t_0, t_f]} d\hat{\eta} > 0,$$

and a $d\hat{\eta}$-measurable essentially bounded function $\hat{s} : [t_0, t_f] \to \mathbb{R}^{n*}$ such that

$$\hat{s}(t) \in \text{conv } S(\text{clm } (\hat{y})(t)) \quad \text{for almost all } t \text{ in } d\hat{\eta}\text{–measure}, \tag{2.29}$$

there hold the following adjoint equation in terms of measures

$$\begin{aligned} -d\hat{\lambda} &= \frac{\partial \hat{\mathcal{H}}(\hat{\lambda}, \hat{\mu}, \hat{y})}{\partial x} dt + \hat{s} d\hat{\eta}, \\ &= \frac{\partial \mathcal{H}(\hat{\lambda}, \hat{y})}{\partial x} dt + \hat{\mu}^G \frac{\partial G(\hat{y})}{\partial x} dt + \hat{\mu}^g \frac{\partial g(\hat{y})}{\partial x} dt + \hat{s} d\hat{\eta}, \end{aligned} \tag{2.30}$$

the tranversality conditions:

$$\hat{\lambda}(t_0-) = -L_{x_0}(\hat{\nu}, \hat{x}_{ini}), \quad \hat{\lambda}(t_f-) = -L_{x_f}(\hat{\nu}, \hat{x}_{ini}), \qquad x_0 = x(t_0),\ x_f = x(t_f),$$

and finally, the stationary condition w.r.t. the control:

$$\frac{\partial \bar{\mathcal{H}}(\hat{\lambda}(t), \hat{\mu}(t), \hat{y}(t))}{\partial u} = 0 \quad \text{a.e. in } [t_0, t_f], \tag{2.31}$$

where "a.e." means "almost everywhere with respect to the Lebesgue measure".
Note that the condition (2.30) can be understood in the following integral form: for almost all t

$$\hat{\lambda}(t) = \hat{\lambda}(t_0 - 0) + \int_{t_0}^{t} \frac{\partial \bar{\mathcal{H}}(\hat{\lambda}, \hat{\mu}, \hat{y})}{\partial x} d\tau + \int_{[t_0, t]} \hat{s}(\tau) d\hat{\eta}(\tau).$$

Theorem 9. [47, Thm. 3] *If $\hat{y} = (\hat{x}, \hat{u})$ is a weak local minimum in problem (2.24), then it satisfies the local minimum principle (2.27)-(2.31).*

Proof. See [47, Section. 6]. □

Remark 12. We consider the significance and application of LMP later, see Chapter 3 and Chapter 6. For more details on the general OCP with nonregular mixed constraints, see [47].

2.4 Direct Approach

In the recent decades, various approaches are investigated to address OCPs with discrete control variables, often known as Mixed-Integer Optimal Control Problems (MIOCPs). Direct approaches are widely employed to solve MIOCPs, see, for instance [22, 58].
In this section, the OCP is considered in the following formulation

$$\begin{aligned} \min_{x(\cdot), u(\cdot)} \quad & m(x(t_f)) + \textstyle\int_{t_0}^{t_f} l(x(t), u(t)) dt \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t), u(t)), \qquad t \in \mathcal{T} \stackrel{\text{def}}{=} [t_0, t_f]. \\ & 0_{n_r} \le r(x(t_0), x(t_f)), \\ & 0_{n_c} \le c(x(t), u(t)), \end{aligned} \tag{2.32}$$

in which we minimize a Bolza type objective function of a dynamic process $x(\cdot)$ defined on the horizon $\mathcal{T} \subset \mathbb{R}$ in terms of an ODE system with right hand side function $f(x(\cdot), u(\cdot))$. The precess is controlled by a control trajectory $u(\cdot)$ subject to minimization. The inequality point constraints $r(\cdot)$ and inequality path constraints $c(\cdot)$ must be satisfied. Moreover, all functions in (2.32) are assumed to be twice continuously differentiable.

2.4.1 Control Discretization

In this section, we introduce how to approximate the space of feasible control functions $u(\cdot)$ by a finite dimensional subspace. We begin by partitioning the control horizon $\mathcal{T}$ into N intervals

$$t_0 < t_1 < \ldots < t_N = t_f \tag{2.33}$$

such that the $\{t_n\}$ is called *shooting grid*. On each interval $[t_n, t_{n+1}]$, $0 \leq n \leq N-1$, we choose a vector of base functions $\theta_n(t, q_n) = \left[\theta_{n,1}(t, q_n^1), \ldots, \theta_{n,n_u}(t, q_n^{n_u})\right]$ where $q_n \stackrel{\text{def}}{=} \left[q_{n,1}^T, \ldots, q_{n,n_u}^T\right]^T$ and $\theta_{n,i} : [t_n, t_{n+1}] \times \mathbb{R}^{n_q^{ni}} \to \mathbb{R}$ for each i, $1 \leq i \leq n_u$, of the control $u(\cdot)$. Some popular choices for base functions w.r.t. the value of n_q^{ni} are as follows

- $n_q^{ni} = 1$, piecewise constant controls: $\theta_{i,n} = q_n^i$.
- $n_q^{ni} = 2$, piecewise linear controls: $\theta_{i,n} = \frac{t_{n+1}-t}{t_{n+1}-t_n} q_{n,1}^i + \frac{t-t_n}{t_{n+1}-t_n} q_{n,2}^i$.
- $n_q^{ni} = 4$, piecewise cubic spline controls: $\theta_{i,n} = \sum_{j=1}^{4} q_{n,j}^i \mu_j \left(\frac{t-t_n}{t_{n+1}-t_n}\right)^{j-1}$, with the appropriate spline function coefficients μ_j, $j = 1, 2, 3, 4$.

Various discretization types can be used for each of the n_u control trajectory components. Certain control discretization choices, such as piecewise linear controls, may need the discretized control trajectory to be continuous over the complete control horizon. To achieve this for the control trajectory component $u_i(\cdot)$, additional control continuity conditions can be added

$$\theta_{n,i}(t_{n+1}, q_n^i) - \theta_{n+1,i}(t_{n+1}, q_{n+1}^i) = 0,$$

for all points of the control discretization grid $\{t_n\}$, $n \in \{1, \ldots, N-1\}$.

This dissertation deals with direct single and direct multiple shooting methods, that is why we present both of them as follows.

2.4.2 Direct Single Shooting Method

In the first consideration of the *direct single shooting* method, cf. [65], the first parametrizes of the control function $u(\cdot)$ with techniques presented in Subsection 2.4.1. Therein $\theta(\cdot, q)$ stand for the control parametrization, where q denotes the parameter that will be determined by optimization.
We introduce the single shooting method for the easiest choice, i.e., piecewise constant controls. For the grid $t_0 < t_1 < \ldots < t_N = t_f$ we set parameters $q_n \in \mathbb{R}^{n_u}$, $n \in \{1, \ldots, N\}$. Then the control parametrization is defined as follows

$$\theta(t, q) \stackrel{\text{def}}{=} q_n, \quad \text{for } t \in [t_n, t_{n+1}), \quad 0 \leq n \leq N-1. \tag{2.34}$$

Therefore, the dimension of the parameter vector q is $N \cdot n_u$, where $q \stackrel{\text{def}}{=} [q_1^T, \ldots, q_N^T]^T$. For completeness, at the final time, the control is defined as $\theta(t_f, q) \stackrel{\text{def}}{=} q_N \stackrel{\text{def}}{=} q_{N-1}$.
In direct single shooting the states are obtained by a forward integration of the IVP, i.e.,

the states $x(\cdot)$ are considered as dependent variables of the controls $u(\cdot)$ respectively their parametrization $\theta(\cdot, q)$ together with the initial state s_0, as follows

$$\dot{x}(t) = f(x(t), \theta(t, q)), \qquad t \in \mathcal{T}, \tag{2.35}$$
$$x(t_0) = s_0. \tag{2.36}$$

The objective function, the control and path constraints are usually discretized and enforce only on the control discretization grid $\{t_n\}$.
This approach allows us to obtain a NLP with the unknowns $[s_0^T, q_0^T, \ldots, q_{N-1}^T]^T$, which can be solved by SQP algorithm, cf. [95, Sec. 3.6].
Instead of there are some advantages over other methods, such as the initialization of the NLP variables is restricted to the initial state s_0 and the control parameters q, the direct single shooting method still has some disadvantages: the potential infeasible of the numerical integration might break down during the integration process due to a very unstable set of differential equations or due to a singularity in time. It is prone to numerical instability especially for long time horizons, and it also struggles when applied to chaotic dynamics and stiff ODEs.

2.4.3 Direct Multiple Shooting Method

The *direct multiple shooting* method was developed by Bock and Plitt [22]. Then the direct multiple shooting code is implemented in details in `MUSCOD-II` by Leineweber [88]. In a direct multiple shooting method, the horizon interval $\mathcal{T}$ is split into N subintervals. The single shooting method is then applied to each subinterval independently. To guarantee state trajectory continuity, additional continuity constraints are included in the resulting NLP.

Control discretization

The discretized control together with the base functions are defined as the same way in Subsection 2.4.1 on the N subintervals.

State parameterization

A parameterization of the state trajectory $x(\cdot)$ is introduced on the shooting grid $\{t_n\}$ that implies N IVPs with initial values $s_i \in \mathbb{R}^{n_x}$ on the intervals $[t_n, t_{n+1}]$ of the horizon $\mathcal{T}$,

$$\dot{x}_n(t) = f(x_n(t), \theta_n(t, q_n)) \quad \forall t \in [t_n, t_{n+1}], \quad 0 \le n \le N-1, \tag{2.37a}$$
$$x_n(t) = s_n. \tag{2.37b}$$

To guarantee continuity of the resulted trajectory $x(\cdot)$ on the whole of the horizon $\mathcal{T}$, $N-1$ additional matching conditions are given as follows

$$x_n(t_{n+1}; t_n, s_n, q_n) - s_{n+1} = 0, \quad 0 \le n \le N-1, \tag{2.38}$$

where $x_n(t_{n+1}; t_n, s_n, q_n)$ stands for the final solution value $x(t_{n+1})$ obtained from the IVP (2.37) on $[t_n, t_{n+1}]$ when starting in the initial value $x(t_n) = s_n$ and applying the control trajectory $u(t) = \theta_n(t, q_n)$ on $[t_n, t_{n+1}]$. Thus the evaluation of the residual of constraint (2.38) needs the solution of an IVP by an appropriate numerical method, cf. [2, 49, 76].

Constraints discretization

The point constraint can be simply rewritten as $r(s_0, s_N) \geq 0_{n_r}$. The path constraint $c(\cdot)$ is discretized as follows

$$c_n(s_n, \theta_n(t_n, q_n)) \geq 0_{n_c}, \quad 0 \leq n \leq N. \tag{2.39}$$

This discretization increases the feasible set of the discretized OCP when comparing to the continuous one, and impacting the obtained optimal solution. In most real world problems, an optimal trajectory $(x^*(\cdot), u^*(\cdot)$ implied as a solution to the discretized problem exhibits only small violations of the path constraints $c(\cdot)$ in the interior of the shooting intervals if they are enforced on the shooting nodes. If large violations occur or strict feasibility on $\mathcal{T}$ is important, remaining violations can sometimes be successfully treated by choosing an adapted, perhaps tighter shooting grid t_n. An alternatively semi-infinite programming algorithm, in the interior of shooting intervals, for tracking of constraint violations is proposed in [107].

The nonlinear problem

By writing the MEYER term $m(t_N, s_N)$ as final term $l_N(t_N, s_N, q_N)$ of the objective function, the discretized OCP resulting from applying the direct multiple shooting method to problem (2.32) casts

$$\begin{aligned} \min_{s,q} \quad & \textstyle\sum_{n=0}^{N} l_n(s_n, q_n) \\ \text{s.t.} \quad & 0 = x_n(t_{n+1}; t_n, s_n, q_n) - s_{n+1}, \quad 0 \leq n \leq N-1, \\ & 0_{n_r} \leq r(s_0, s_N), \\ & 0_{n_c} \leq c(s_n, \theta(t_n, q_n)), \quad 0 \leq n \leq N. \end{aligned} \tag{2.40}$$

The SQP method for solving problem (2.40) are presented in details, cf. [95, Sec. 3.6] and [76, Chap. 3].

2.4.4 Derivative Generation

To analyze the sensitivity of derivative generation in forward mode, some needed definitions are presented during this section.

Implicitly defined discontinuities

Definition 26 (Parameter-dependent IVP with switches). Let $f : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \times \{-1, 0, 1\}^{n_\sigma} \to \mathbb{R}^{n_x}$, $p \in \mathbb{R}^{n_p}$, and $n_x, n_p, n_\sigma \in \mathbb{N}$. The parameter-dependent IVP with switches is defined as

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), p, \operatorname{sgn}(\sigma(t, x(t), p))), \quad t \in [t_0, t_f], \\ x(t_0) &= x_0, \end{aligned} \tag{2.41}$$

for $x \in \mathbb{R}^{n_x}$, $p \in \mathbb{R}^{n_p}$ and the switching function $\sigma : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_\sigma}$ with the components

$$\begin{aligned} \sigma_j : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} &\to \mathbb{R} \\ (t, x(t), p) &\to \sigma_j(t, x(t), p) \end{aligned}$$

where $j = 1, \ldots, n_\sigma$.

For the following definition we assume a problem with only one switching time t_{sw}.

Definition 27. 1. The left and right limits of the state vector at the switching time t_{sw} are defined as follows,

$$x_- := x_-(t_{sw}; t_0, x_0, p) = \lim_{\varepsilon \to 0} x(t_{sw} - \varepsilon), \tag{2.42}$$

$$x_+ := x_+(t_{sw}; t_0, x_0, p) = \lim_{\varepsilon \to 0} x(t_{sw} + \varepsilon), \tag{2.43}$$

respectively, with $\varepsilon > 0$.

2. f_- and f_+ are the right hand side of f in (t_{sw}, y_-, p) and (t_{sw}, y_+, p), respectively, i.e.,

$$f_- := f_-(t_{sw}, x_-, p) = \lim_{\varepsilon \to 0} f(t_{sw} - \varepsilon, x_-, p), \tag{2.44}$$

$$f_+ := f_+(t_{sw}, x_+, p) = \lim_{\varepsilon \to 0} f(t_{sw} + \varepsilon, x_+, p), \tag{2.45}$$

where $\varepsilon > 0$.

3. The jump vector δ of the right hand side f is defined as follows

$$\delta := \delta(t_{sw}, y_-, p) = f_+(t_0, x_+, p) - f_-(t_0, x_-, p) \tag{2.46}$$

Sensitivity analysis

Consider the IVP

$$\begin{aligned} &\dot{x}(t) = f(t, x(t), p), \quad t \in [t_0, t_f], \\ &x(t_0) = x_0, \end{aligned} \tag{2.47}$$

where $f : [t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$, $p \in \mathbb{R}^{n_p}$ and $n_x, n_p \in \mathbb{N}$.

Definition 28 (Sensitivities). The matrix $G_x(t; t_0, x_0, p) \in \mathbb{R}^{n_x \times n_x}$ denotes the sensitivity of the solution x at time point t w.r.t. the initial value $x_0 \in \mathbb{R}^{n_x}$ and the matrix $G_p(t; t_0, x_0, p) \in \mathbb{R}^{n_x \times n_p}$ analogously denotes the sensitivity of the solution x at time point t w.r.t. the parameter $p \in \mathbb{R}^{n_p}$,

$$G_x(t; t_0, x_0, p) := \frac{\partial x}{\partial x_0}(t; t_0, x_0, p), \tag{2.48}$$

$$G_p(t; t_0, x_0, p) := \frac{\partial y}{\partial p}(t; t_0, x_0, p). \tag{2.49}$$

To calculate sensitivities, we use the approximation by external numerical differentiation (END), and the variational differential equations (VDE).

External numerical differentiation

For the sensitivities w.r.t. x_0 that means calculating

$$\frac{\partial x}{\partial x_{0,i}}(t;t_0,x_0,p) \approx \frac{x(t;t_0,x_0+h_{x,i}\cdot e_i,p)-x(t;t_0,x_0-h_{x,i}\cdot e_i,p)}{2h_x},\ i=1,\dots,n_x, \quad (2.50)$$

where $h_x \in \mathbb{R}^{n_x}$ is the vector of step sizes for the finite differences (FD) w.r.t. the initial values and e_i is the i-th unit vector of the dimension n_x. In the similar way, the sensitivities w.r.t. p are calculated by perturbing each parameter, i.e.,

$$\frac{\partial x}{\partial p_i}(t;t_0,x_0,p) \approx \frac{x(t;t_0,x_0,p+h_{p,i}\cdot e_i)-x(t;t_0,x_0,p-h_{p,i}\cdot e_i)}{2h_p}, \quad i=1,\dots,n_p, \quad (2.51)$$

therein, $h_p \in \mathbb{R}^{n_p}$ is the vector of step sizes for the FD w.r.t. the initial values and e_i is the i-th unit vector of the dimension n_p.

Variational differential equations

Another possibility for calculating sensitivities is solving the VDE. To deduce the sensitivity matrix $G_x(t;t_0,x_0,p)$ the following system needs to be solved

$$\begin{aligned}&\frac{\partial G_x}{\partial t}(t;t_0,x_0,p) = \frac{\partial f}{\partial x}(t,x,p)\cdot G_x(t;t_0,x_0,p), \quad (2.52)\\ &G_x(t_0;t_0,x_0,p) = I_{n_x},\end{aligned}$$

where $I_{n_x} \in \mathbb{R}^{n_x\times n_x}$ is the identity matrix.
Analogously one obtains the sensitivity matrix $G_p(t;t_0,x_0,p)$ as the solution of the system as follows

$$\begin{aligned}&\frac{\partial G_p}{\partial t}(t;t_0,x_0,p) = \frac{\partial f}{\partial x}(t,x,p)\cdot G_p(t;t_0,x_0,p) + \frac{\partial f}{\partial p}(t,x,p), \quad (2.53)\\ &G_p(t_0;t_0,x_0,p) = \frac{\partial x_0}{\partial p}(p) = 0.\end{aligned}$$

2.5 Feedback Control

Optimization techniques play a fundamental role in real world processes and especially in the current industrial practice. In many applications, an OCP is solved *off-line* with full time horizon, or in an *open-loop*, i.e., the process operation is no longer tracked anymore and the obtained solution is applied without further feedback from the actual process. This leads open-loop controls to invalidate the previously optimal solution and are of limited applicability.
In practical applications, the actual system behavior is considered and the controller is constantly updated with the system state. Hence, there is a great interest in the so-called optimization-based *feedback control* or *closed-loop* approaches, where OCP is solved *on-line*.

This section introduces a powerful state-of-the-art feedback control approach via the principle of Model Predictive Control (MPC), cf. [41, 76] and [115, Chap. 4], which can be summarized as follows: one repeatedly off-line solves OCPs on a finite prediction horizon. At

sampling times t_j^s, where j stands for the sample index, we retrieve the current real process state x_j^s. Using x_j^s as initial state, we solve an OCP on a prediction horizon $[t_j^s, t_j^s + \Delta h]$, and obtain an optimal control $u_j^*(\cdot)$. The initial state conditions are chosen to generate a coupling of the real state and the state prediction. We only apply $u_j^*(\cdot)$ for the sampling time period Δt. At the subsequent sampling time $t_{j+1}^s = t_j^s + \Delta t$, we solve a new OCP with the updated initial state x_{k+1}^s on the horizon $[t_{j+1}^s, k_{j+1}^s + \Delta h]$, and apply the obtained optimal control $u_{j+1}^*(\cdot)$. Applied to OCP (2.32), at each sampling time t_j^s, the MPC feedback approach solves the following OCP

$$\begin{aligned} \min_{x(\cdot),u(\cdot)} \quad & m(x(t_j^s + \Delta h)) + \textstyle\int_{t_j^s}^{t_j^s+\Delta h} l(x(\tau), u(\tau))d\tau \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t), u(t)), \quad t \in \mathcal{T}, \\ & 0_{n_r} \leq r(x(t_j^s), x(t_j^s + \Delta h)), \\ & 0_{n_c} \leq c(x(t), u(t)), \quad t \in \mathcal{T}. \end{aligned} \tag{2.54}$$

MPC subject to OCPs with quadratic objective function and linear dynamic equations and inequality constraints are referred to as *Linear Model Predictive Control*. If the objective function is nonlinear but not quadratic, or nonlinear dynamic equations or inequality constraints, one calls it *Nonlinear Model Predictive Control* (NMPC).

2.6 Switched Optimal Control Problems

Switched Optimal Control Problems (SwOCPs) are a particular class of hybrid dynamic systems. There has been numerous research over the last few decades, and significant progress has been made in this topic, both theoretically and computationally, cf. [3, 60, 132].
Hybrid systems are dynamic systems that combine continuous and discrete event models, where the system switches between different models. Hybrid systems have applications in a variety of disciplines, including industrial process management, gas traffic control, power systems, gas and water networks. For a detailed survey on this field, readers can see in [132]. To discuss the necessary conditions for trajectories for hybrid systems, the existence of optimal control law is obtained based on dynamic programming, cf. [30], or by using the maximum principle, cf. [105, 117]. Then convex dynamic programming is employed to approximate the hybrid optimal control laws as well as the objective value's bounds, cf. [64]. In [18], switched systems are OCPs with state-dependent discontinuities with no jumps in the states, i.e., switched systems are represented by an indexed set of differential equations

$$\dot{x}(t) = F_{i(t)}(x(t), u(t)), \quad x(t_0) = x_0, \quad \mathcal{T} \stackrel{\text{def}}{=} [t_0, t_f],\ i : \mathcal{T} \to \{1, \ldots, M\}, \tag{2.55}$$

where the unified framework for SwOCP is presented with both implicit switches (internally forced switches(IFS)) and explicit switches (externally forced switches (EFS)).
Numerous literature dealing with IFS problems concentrates on piecewise affine (PWA), cf. [10, 71, 109]. A PWA system, cf. [18],

$$x(t+1) = A_i x(t) + B_i u(t) + f_i, \text{ if } \begin{bmatrix} x(t) & u(t) \end{bmatrix}^T \in \mathcal{X}_i,$$

where $\mathcal{X}_i \stackrel{\text{def}}{=} \left\{ \begin{bmatrix} x(t) & u(t) \end{bmatrix}^T \mid G_i x + H_i u \leq K_i \right\}$, divides the state space into polyhedral regions and assigns with each region its own linear difference equation. It can be expanded by

the following mode independent constraints

$$Ix(t) + Ju(t) \leq L.$$

For solving PWA constrained optimization problems, there are two types of solution approaches are proposed. First method is the mixed logical dynamic, cf [11], while the remain one is the dynamic programming strategies in combination with multi-parametric program solver are applied, cf. [28].
Plenty of algorithms for solving EFS are proposed, such as bi-level hierarchical algorithm, cf. [90, 92], direct single shooting method, cf. [9], or gradient projection and constrained NEWTON's method, cf. [132].

2.6.1 Filippov's Theory

Reminder that an optimal control may not exist if the right hand side function $f(\cdot)$ in the ODE of OCPs is not convex. For the counter example, readers can see in [55, Problem (14), Sec. V]. In the case of the general form of SwOCP, sliding regimes arise from the nonconvexity of $f(\cdot)$. The theory of FILIPPOV gives a generalized definition of the solution of switched systems in the sense that the definition holds for a larger class of differential equations, cf. [56]. FILIPPOV's rule describes three basic forms of dynamics that would occur on the switching manifold: sewing (sticking to a surface), sliding (motion constrained along the manifold), and escaping (leaving the surface abruptly). Solution in the FILIPPOV's rule is continuous in time, where jump conditions are not considered.
We consider FILIPPOV theory for the general case, therein a natural idea to extend the classic solution concept is to replace the right hand side $f(\cdot)$ with a set-valued function $F(\cdot)$ such that $f(\cdot)$ and $F(\cdot)$ are identical at points where $f(\cdot)$ is continuous in x. There is a suitable choice for $F(\cdot)$ required at points for which $f(\cdot)$ is discontinuous in x. Then the differential equation is replaced by the *differential inclusion*

$$\dot{x}(t) \in F(t, x(t)).$$

At points of discontinuity, $F(\cdot)$ is defined by means of the *generalized differential*, cf. [36]. The *generalized derivative* of a function $x : \mathbb{R} \to \mathbb{R}^{n_x}$ at t is defined as any value $\dot{x}_\alpha(t)$, cf. [37], which can be implies by means of a convex combination of its left and right derivatives as follows

$$\dot{x}_\alpha(t) = \alpha \cdot \dot{x}_+(t) + (1 - \alpha) \cdot \dot{x}_-(t), \quad 0 \leq \alpha \leq 1.$$

The values $\dot{x}_+(t)$ and $\dot{x}_-(t)$ are respective determined as $f_+(t, x(t))$ and $f_-(t, x(t))$. Denote $\partial x(t)$ the set of all the generalized differential of $x(\cdot)$ at t, i.e., it is the convex hull of the derivative extremes

$$\begin{aligned} \partial x(t) &= \operatorname{conv}\{\dot{x}_+(t), \dot{x}_-(t)\} \\ &= \{\dot{x}_\alpha(t) \in \mathbb{R}^n : \dot{x}_\alpha(t) = \alpha \cdot \dot{x}_+(t) + (1 - \alpha) \cdot \dot{x}_-(t), \alpha \in [0, 1]\}, \end{aligned} \tag{2.56}$$

where $\operatorname{conv} \mathcal{A}$ stands for the smallest closed convex set containing $\mathcal{A}$. The set-valued sign function is then defined as the generalized differential of $|x|$

$$\operatorname{sgn}(x) \stackrel{\text{def}}{=} \partial|x| = \begin{cases} \{-1\}, & \text{if } x < 0, \\ [-1, 1], & \text{if } x = 0, \\ \{1\}, & \text{if } x > 0. \end{cases}$$

The idea of replacing a switched ODE with a differential inclusion is transferred from one dimension to dimension n. The space $\mathbb{R}^n$ is split into two subspaces $\mathcal{S}_+$ and $\mathcal{S}_-$ by a hyper-surface $\mathcal{S}$ such that $\mathbb{R}^n = \mathcal{S}_- \cup \mathcal{S} \cup \mathcal{S}_+$, where the hyper-surface $\mathcal{S}$ is implicitly defined by the switching function $\sigma : \mathbb{R}^n \to \mathbb{R}$ as

$$\mathcal{S} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \sigma(x) = 0\}, \tag{2.57}$$

and the subspaces $\mathcal{S}_+$ and $\mathcal{S}_-$ as

$$\mathcal{S}_+ \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \sigma(x) > 0\}, \quad \mathcal{S}_- \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \sigma(x) < 0\}.$$

Consider the nonlinear system with discontinuous right hand side

$$\dot{x}(t) = f(t, x(t)) \stackrel{\text{def}}{=} \begin{cases} f_+(t, x(t)), & \text{if } x(t) \in \mathcal{S}_+, \\ f_-(t, x(t)), & \text{if } x(t) \in \mathcal{S}_-, \end{cases} \quad t \in \mathcal{T} \setminus \mathcal{S}, \quad x(t_{sw}) = x_{sw}. \tag{2.58}$$

We assume that $f(\cdot)$ fulfills all assumptions from [120, Thm. 2.2] in $\mathbb{R}^n \setminus \mathcal{S}$ such that the solution $x(\cdot)$ within $\mathcal{S}_+$ and $\mathcal{S}_-$ exists and unique. Furthermore, we assume that the smooth functions f_+ and f_- are extended uniquely to smooth functions on $S_+ \cup \mathcal{S}$ and $S_- \cup \mathcal{S}$, respectively.
Recall problem (2.58) where $f(\cdot)$ is not defined for t with $x(t) \in \mathcal{S}$. This allows for some freedom in extending the vector field on $\mathcal{S}$. To accomplish this, we study the set-valued extension $F : \mathcal{T} \times \mathbb{R}^{n_x} \to \mathbb{R}^n$ of $f(\cdot)$ for $x_\sigma \in \mathcal{S}$, represented as

$$F(t, x_\sigma) \stackrel{\text{def}}{=} \text{conv}\{y \in \mathbb{R}^n : y = \lim_{x \to x_\sigma} f(t, x),\ x \in \mathbb{R} \setminus \mathcal{S}\}. \tag{2.59}$$

Note that all the limits exist due to the assumptions on $f(\cdot)$. The convexification of the switched IVP (2.58) into the convex differential inclusion

$$\dot{x}(t) \in F(t, x(t)) \stackrel{\text{def}}{=} \begin{cases} f_+(t, x(t)), & \text{if } x(t) \in \mathcal{S}_+, \\ \text{conv}\{f_+(t, x(t)), f_-(t, x(t))\}, & \text{if } x(t) \in \mathcal{S}, \quad x(t_{sw}) = x_{sw}, \\ f_-(t, x(t)), & \text{if } x(t) \in \mathcal{S}_-, \end{cases} \tag{2.60}$$

where the convex set from (2.59) can be expressed on $\mathcal{S}$ by combinations of f_+ and f_- and

$$\text{conv}\{f_+, f_-\} = \{f \in \mathbb{R}^n : f = \alpha \cdot f_+ + (1 - \alpha) \cdot f_-,\ \alpha \in [0, 1]\} \tag{2.61}$$

is well-known as Filippov's convex method. To understand the IVP (2.58) as a mathematical model of a physical system, it's important to consider a solution notion that guarantees the existence of solutions. Thus, the choice of the set valued extension $F(\cdot)$ of $f(\cdot)$ should be appropriate in the sense that the existence of a solution is guaranteed. The notion of upper semi-continuity of set-valued functions, cf. [95, Def. 1.17], ensures the existence of solutions of a differential inclusion. Combining this condition with Cathathéodory solutions, the existence of differential inclusion trajectories, that are absolutely continuous, which is guaranteed by the following results.

Theorem 10 (**Existence of Differential Inclusion Solution**). [95, Thm. 1.18] *Let $F(\cdot)$ be a set valued function. Assuming $F(\cdot)$ to be a upper semi-continuous and $F(t, x)$ to be*

closed, convex, and bounded for all $t \in \mathbb{R}$ and $x \in \mathbb{R}$, then for each $x_{sw} \in \mathbb{R}^n$ there exists a $\tau > 0$ and an absolutely continuous function $x(\cdot)$ defined on $[t_{sw}, t_{sw} + \tau]$, which is a solution of the IVP

$$\dot{x}(t) \in F(t, x(t)), \quad x(t_{sw}) = x_{sw}.$$

Proof. See [7]. □

A *Filippov's rule* solution for an implicitly switched system of type (2.58) can be defined by combining of Filippov's convex method and the result from Theorem 10.

Definition 29 (**Solution in the Filippov's rule**). [95, Def. 1.19] An absolute continuous function $x : [t_{sw}, t_{sw} + \tau] \to \mathbb{R}^n$ is called a solution of IVP (2.58) in the Filippov's rule if for almost all $t \in [t_{sw}, t_{sw} + \tau]$ it holds that

$$\dot{x}(t) \in F(t, x(t)),$$

where $F(t, x(t))$ is defined as in (2.60).

Definition 30 (**Another sense of Filippov's solution**). [49, Def. 5.1] The function $x(t)$, $t \in [t_0, t_f]$ is called the solution of the differential equation $\dot{x} = f(x(t))$, if the following conditions are met:

- x is absolutely continuous,
- for almost all $t \in [t_0, t_f]$ and any $\delta > 0$, the vector $\dot{x} = \frac{dx}{dt}$ belongs to the smallest closed convex set that contains all values $f(\cdot)$ in a δ-neighborhood of $x(t)$:

$$\dot{x}(t) \in \bigcap_{\delta > 0} \bigcap_{\mu(N)=0} \overline{conv}\left(f(U(x(t), \delta) \setminus N, \cdot)\right).$$

Here, μ denotes the Lebesgue measure.

Remark 13. In the domain where $x(\cdot)$ is smooth, i.e., $x(t) \in \mathcal{S}_+ \cup \mathcal{S}_-$, the equality $f(t, x(t)) = F(t, x(t))$ holds true. If $x(\cdot)$ slides along a switching boundary, i.e., $x(t) \in \mathcal{S}$, then $\dot{x}(t) \in F(t, x(t))$. However, when the solution $x(\cdot)$ leaves from the switching manifold $\mathcal{S}$ or enters to $\mathcal{S}$, the state derivative $\dot{x}(t_\sigma)$ is not defined, at time instances t_σ. Therein, a solution trajectory $x(\cdot)$ leaves or enters $\mathcal{S}$ if for any $\varepsilon > 0$ there exists a $t_* \in t_\sigma + \mathcal{U}_\varepsilon(0) \setminus \{0\}$ such that $x(t_*) \notin \mathcal{S}$ and $x(t_\sigma) \in \mathcal{S}$.

Remark 14. Theorem 10 ensures the existence of a solution on $[t_{sw}, t_{sw} + \tau]$ with $\tau > 0$. To obtain existence over the whole horizon, one needs further assumptions: let $f(t, x)$ be linearly bounded for $x \notin \mathcal{S}$, i.e., there exists positive constants c_0 and c_1 such that for all $t \in [0, \infty)$ and $x \in \mathcal{S}_+ \cup \mathcal{S}_-$ it holds

$$\|f(t, x)\| \leq c_0 \|x\| + c_1.$$

Additionally, if $F(\cdot)$ is bounded at (t, x) for which F is set-valued, then a solution of IVP (2.60) exists on $[t_{sw}, \infty)$, cf. [37].
However, these assumptions are not sufficient to guarantee the uniqueness of a solution. Readers may refer to [95, Sec. 1.4], where the uniqueness of solutions is examined in several scenarios, such as transversal intersection mode, sliding mode, and higher order conditions.

2.6.2 Problem and Constraints Formulation of SwOCP

We start this subsection with a problem formulation of a general SwOCP involving a dynamic system with switches.

Definition 31. A SwOCP is a constrained optimization problem of the form

$$\begin{aligned} \min_{x(\cdot),u(\cdot),w(\cdot)} \quad & \varphi(x(t),u(t)) \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t),u(t),w(t),\operatorname{sgn}(\sigma(x(t)))), \\ & 0_{n_r} \leq r(x(t_0),x(t_f)), \\ & 0_{n_c} \leq c(x(t),u(t),w(t)), \\ & w(t) \in \Omega \subset \mathbb{R}^{n_w}, \end{aligned} \qquad t \in \mathcal{T}, \tag{2.62}$$

where a dynamic process $x : \mathcal{T} \to \mathbb{R}^{n_x}$ on the time horizon $\mathcal{T} \overset{\text{def}}{=} [t_0, t_f] \subset \mathbb{R}$ is determined. A solution $x(\cdot)$ is described by a system of ODEs, where $f : \mathcal{T} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_w} \times \{-1,0,1\}^{n_\sigma} \to \mathbb{R}^{n_x}$ is the right hand side function. This system is affected by a continuous-valued control function $u : \mathcal{T} \to \mathbb{R}^{n_u}$ as well as another discrete-valued control function $w : \mathcal{T} \to \Omega$, which includes only values from a finite set $\Omega \overset{\text{def}}{=} \{w_1, w_2, \ldots, w_{n_\omega}\} \subseteq \mathbb{R}^{n_w}$ with cardinality $|\Omega| < \infty$. Furthermore, the system is affected by an implicit switch determined by the sign structure of a switching function $\sigma : \mathbb{R}^{n_x} \to \mathbb{R}^{n_\sigma}$. The objective function $\varphi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$ is minimized. Moreover, mixed state-control (path) constraints $c(x(t),u(t),w(t)) \geq 0_{n_c}$ with $c : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_w} \to \mathbb{R}^{n_c}$ and point constraints $r(x(t_0),x(t_f)) \geq 0_{n_r}$ where $r : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_r}$ must be satisfied.

Note that $w(\cdot)$ has the special character when compared to $u(\cdot)$, and due to its special status, we will introduce a new term of this type of control functions in Chapter 3.

2.6.3 Consistent Switches

Main contributions of this dissertation are belonging to this type of switches, where the transversality assumption is hold true.

The sign structure of σ in the right hand side function f of the dynamic system in (2.62) leads to non-differentiability in the dynamics. This coincides with the concept of a switched system (2.55). The switch is an explicit switch, i.e., EFS, if σ is independent of x, otherwise it is an implicit switch, i.e., IFS. At a switching point $t_{sw} \in \mathcal{T}$, the left and right hand side limits of x are defined as

$$x_+(t_{sw}) \overset{\text{def}}{=} \lim_{t \searrow t_{sw}} x(t), \quad x_-(t_{sw}) \overset{\text{def}}{=} \lim_{t \nearrow t_{sw}} x(t).$$

Then the one-sides derivatives of $\sigma(\cdot)$ at $t_{sw} \in \mathcal{T}$ are defined as, cf. [18],

$$\mathcal{D}\sigma_+(t_{sw}) \overset{\text{def}}{=} \frac{d\sigma}{dt}(t_{sw}, x_+(t_{sw})), \quad \mathcal{D}\sigma_-(t_{sw}) \overset{\text{def}}{=} \frac{d\sigma}{dt}(t_{sw}, x_-(t_{sw})).$$

The *transversality assumption* holds true for a large class of switched dynamic systems, which allow solutions that cross the zero manifold $\mathcal{S}$, see Eq. (2.57), in either direction. As a result, a finite number of isolated switching events occurs on a finite time horizon. These solutions are typically referred to as "classical", and the switching behavior is often called "consistent".

Assumption 2.6.1 (**Transversality**). [18, Assumption 2.1] *Problem (2.62) satisfies the transversality assumption if* $\mathcal{D}\sigma_-(t_{sw}) \cdot \mathcal{D}\sigma_+(t_{sw}) > 0$ *for all* $t_{sw} \in \mathcal{T}$ *with* $\sigma(t_{sw}) = 0$.

If the transversality assumption holds, only a finite number of isolated points of the zero manifold $\mathcal{S}$ are part of a solution trajectory x.

2.6.4 Inconsistent Switches and Filippov Solutions

If the transversality assumption is violated, which means that $\mathcal{D}\sigma_-(t_{sw}) \cdot \mathcal{D}\sigma_+(t_{sw}) \leq 0$, we have to consider the additionally Filippov case of sliding on the zero manifold, i.e., a continuation on the manifold $\mathcal{S}$ in the Filippov's rule [56] can be found by replacing f in (2.62) by an appropriate combination

$$f_\alpha(x(t), u(t)) := \alpha(t) f(x(t), u(t), +1) + (1 - \alpha(t)) f(x(t), u(t), -1), \quad \alpha(t) \in (0, 1),$$

that satisfies $\sigma(x(t)) = 0$ for all $t \in [t_{sw}, t^+]$, where $t^+ > t_{sw}$ is a later time point with switching function derivatives that allow leaving the manifold. If one of derivatives $\mathcal{D}\sigma_-$ or $\mathcal{D}\sigma_+$ vanishes, the state trajectory tangentially leaves or enters the zero manifold $\mathcal{S}$. Since $\mathcal{D}\sigma_-$ and $\mathcal{D}\sigma_+$ are first order derivatives, higher-order derivatives of σ can be analyzed to study the type of continuation is open as a future research question.
In connection with switching functions, the phenomenon that the manifold cannot be pierced is called *inconsistent switching*. The switching is inconsistent in the type that changing the right hand side does not change the sign of the switching function. Section 5.1.3 will discuss more about directional fields and three-valued switching logic for the inconsistent switches with its general formulation of SwOCP.

2.6.5 Partial Outer Convexification and Relationship between SwOCP with Relaxed Problem

The idea of convexification and relaxation is similar to the concept of *generalized curves*, which was proposed by Young [127] to investigate existence questions in the domain of *calculus of variations*. The partial outer convexification (POC) *approach* has been studied in the context of OCPs by Sager [111], Sager et al., cf. [112, 113]. The term "partial" is due to the exclusive convexification of the only integer controls, not to the rest of OCP. To describe the POC approach the integer controls $w(\cdot)$ are lifted into a higher dimensional space by introducing binary controls $\omega_i : \mathcal{T} \to \{0, 1\}$, $i \in \{1, \ldots, n_\omega\}$. The value $\omega_i(t) = 1$ indicates mode i is active, otherwise not active ($\omega_i(t) = 0$) at instant time $t \in \mathcal{T}$. For better realization of the POC, we consider a SwOCP where ODE is considered without the sign function and the objective function is simply in Mayer type, as follows

$$\begin{aligned} \min_{x(\cdot),u(\cdot),w(\cdot)} \quad & m(x(t_f)) \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t), u(t), w(t)), \\ & 0_{n_r} \leq r(x(t_0), x(t_f)), \\ & 0_{n_c} \leq c(x(t), u(t), w(t)), \\ & w(t) \in \Omega \subset \mathbb{R}^{n_w}, \end{aligned} \qquad t \in \mathcal{T}, \tag{2.63}$$

while later, the general SwOCP (2.62) will be reformulated in Chapter 3. The ODE in (2.63) then reads as

$$\dot{x}(t) = \sum_{i=1}^{n_\omega} \omega_i \cdot f(x(t), u(t), w_i).$$

To guarantee that exactly one mode is active at any $t \in \mathcal{T}$, one additionally imposes the SOS-1 constraint $\sum_{i=1}^{n_\omega} \omega_i(t) = 1$.
Similarly, the path constraints of (2.63) is rewritten as follows

$$\sum_{i=1}^{n_\omega} \omega_i(t) \cdot c(x(t), u(t), w_i) \geq 0_{n_c}, \tag{2.64}$$

together with the additional constraint $\sum_{i=1}^{n_\omega} \omega_i(t) = 1$, $\omega_i(t) \in \{0, 1\}$.
Then (2.63) reads as the following form after POC reformulation

$$\begin{aligned} \min_{x(\cdot),u(\cdot),\omega(\cdot)} \quad & m(x(t_f)) \\ \text{s.t.} \quad & \dot{x}(t) = \textstyle\sum_{i=1}^{n_\omega} \omega_i \cdot f(x(t), u(t), w_i), \\ & 0_{n_r} \leq r(x(t_0), x(t_f)), \\ & 0_{n_c} \leq \textstyle\sum_{i=1}^{n_\omega} \omega_i(t) \cdot c(x(t), u(t), w_i), \\ & 1 = \textstyle\sum_{i=1}^{n_\omega} \omega_i(t),\ \omega(t) \in \{0, 1\}^{n_\omega}, \end{aligned} \qquad t \in \mathcal{T}, \tag{2.65}$$

where $\omega(\cdot) \stackrel{\text{def}}{=} [\omega_1(\cdot), \ldots, \omega_{n_\omega}(\cdot)]^T$. Problem (2.65) can be relaxed by construction as follows

$$\begin{aligned} \min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & m(x(t_f)) \\ \text{s.t.} \quad & \dot{x}(t) = \textstyle\sum_{i=1}^{n_\omega} \alpha_i \cdot f(x(t), u(t), w_i), \\ & 0_{n_r} \leq r(x(t_0), x(t_f)), \\ & 0_{n_c} \leq \textstyle\sum_{i=1}^{n_\omega} \alpha_i(t) \cdot c(x(t), u(t), w_i), \\ & 1 = \textstyle\sum_{i=1}^{n_\omega} \alpha_i(t),\ \alpha(t) \in [0, 1]^{n_\omega}, \end{aligned} \qquad t \in \mathcal{T}, \tag{2.66}$$

where analogously to $\omega(\cdot)$, the components of $\alpha(\cdot)$ are denoted by $\alpha_i(\cdot)$, $i \in \{1, \ldots, n_\omega\}$. The binary convexified OCP (2.65) is equivalent to OCP (2.62) in the type as follows:

Proposition 1. *The binary convexified OCP (2.65) has a solution if and only if the explicit switched OCP (2.63) has a solution. Let $(x_C^*, u_C^*, \omega_C^*)$ be a solution of (2.65). Then (x^*, u^*, w^*), with $x^* = x_C^*$, $u^* = u_C^*$, and $w^*(t) = \sum_{i=1}^{n_\omega} \omega_i(t) w_i$, is a solution of (2.63).*

Proof. See [89, Proposition 6.6]. □

Moreover, considering the relation between SwOCP and its relaxed problem, the following theorem has shown that: for a feasible point of the relaxed convexified OCP (2.66) there is an essentially feasible point of the binary convexified OCP (2.65) which has essentially the same objective function value.

Theorem 11. *Let $(\bar{x}, u, \bar{\alpha})$ be feasible for OCP (2.66) and suppose that $t \mapsto f(\bar{x}(t), u(t), w_i)$, $i \in \{1, \ldots, n_\omega\}$, are functions of type $W^{1,\infty}(\mathcal{T}, \mathbb{R}^{n_x})$. Let $\varepsilon > 0$.*
Then there are functions $x^\varepsilon \in W^{1,\infty}(\mathcal{T}, \mathbb{R}^{n_x})$ and $\omega^\varepsilon \in L^\infty(\mathcal{T}, \{0, 1\}^{n_\omega})$ such that

$$|m(x^\varepsilon(t_f)) - m(\bar{x}(t_f))| < \varepsilon$$

and

$$\begin{aligned}
&\dot{x}^{\varepsilon}(t) = \textstyle\sum_{i=1}^{n_\omega} \omega_i \cdot f(x^{\varepsilon}(t), u(t), w_i), \quad t \in \mathcal{T},\\
&0_{n_r} \leq r(x^{\varepsilon}(t_0), x^{\varepsilon}(t_f)),\\
&0_{n_c} \leq \omega_i^{\varepsilon}(t) \cdot c(x^{\varepsilon}(t), u(t), w_i), \quad 1 \leq i \leq n_\omega,\ t \in \mathcal{T},\\
&1 = \textstyle\sum_{i=1}^{n_\omega} \omega_i^{\varepsilon}(t), \quad t \in \mathcal{T}.
\end{aligned}$$

Proof. See [89, Theorem 6.7]. □

2.6.6 Generalized Disjunctive Programming

Disjunctive Programming (DP) is an approach for solving the OCPs, which includes both continuous and discrete controls, cf. [8]. DP models consist of logic disjunctions, algebraic constraints and logic propositions. A particular case of disjunctive programming is so-called General Disjunctive Programming (GDP), cf. [108]. A GDP problem is reached as follows

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n, \omega_{ik}\in\{0,1\}} \quad \psi(x) + \textstyle\sum_{k\in\mathcal{K}} c_k,\\
&\text{s.t.} \quad r(x) \leq 0,\\
&\qquad \bigoplus_{i\in\mathcal{D}_k} \begin{bmatrix} \omega_{ik} = 1 \\ g_{ik}(x) \leq 0 \\ c_k = \gamma_{ik} \end{bmatrix}, \quad k \in \mathcal{K} \stackrel{\text{def}}{=} \{1, \ldots, K\},\ \mathcal{D}_k \stackrel{\text{def}}{=} \{1, \ldots, D_k\},\\
&\qquad \Omega(\omega) = 1, \ \ x \in [x_l, x_u],
\end{aligned}$$

where the continuous variables $x \in \mathbb{R}^n$ in the bounds $[x_l, x_u]$ and binary variables $\omega \stackrel{\text{def}}{=} \{\omega_{ik}\}_{i,k}$, $\omega_{ik} \in \{0, 1\}$. The function $\psi : \mathbb{R}^n \to \mathbb{R}$ in the objective function and the global constraint function $g : \mathbb{R}^n \to \mathbb{R}^m$ are assumed to be sufficiently smooth, and $r(x)$ is the inequality sign include the case of equalities. K logical expressions must hold, and each of these expressions is composed of D_k terms which are connected by the EX-OR operator $\oplus$, indicating that exactly one of the boolean variables ω_{ik} must be defined to one. In the particular case for variable ω_{ik}, the associated constraint $g_{ik}(x) \leq 0$ and the objective weight c_k are enforced. The constraint $\Omega(\omega) = 1$ summarizes further constraints on the boolean variables ω_{ik}. All $\omega_{ik} = 0$ are ignored.
The GDP was recently used to reformulate the SwOCP by Meyer et al. [18], where they deal with explicit and implicit switches of OCPs.

2.6.7 Rounding Schemes

In this section, all available rounding schemes, which are used to return the integer value of control from the relaxed one, are shortly summarized. For the convergence of rounding schemes, i.e., solution quality, readers can see in [12, Prop. 3.1], [73, Prop. 4.8].
Suppose that the optimal solution of the relaxed convexified of the switched optimal control problem (RC.SwP) is (x^*, u^*, α^*). Hence the corresponding optimal solution of SwOCP with switched DAEs/ODEs is (x^*, u^*, w^*). Now we just take care for the term α^*, and looking for the relation between α^* and w^*.

i. If $\alpha_j^*(t) = 0$ or $\alpha_j^*(t) = 1$, then it is also the optimal solution of SwOCP.

ii. Otherwise, $\alpha_j^*(t) \in (0, 1)$, we must apply rounding strategies to imply the solution of SwOCP. Denote $\alpha^*(t) = \tilde{q}_i$, and $w^*(t) = q_i$, where $t \in [t_i, t_{i+1}] \subset [t_0, t_f]$, $i =$

$0, \ldots, n-1$. Then we could apply one of the suitable rounding strategies, cf. [111], as follows:

- Rounding strategy SR (standard rounding)

$$q_{j,i} = \begin{cases} 1 & \text{if } \tilde{q}_{j,i} \geq 0.5, \\ 0 & \text{else.} \end{cases}$$

- Rounding strategy SUR (sum-up rounding)

$$q_{j,i} = \begin{cases} 1 & \text{if } \sum_{k=0}^{i} \tilde{q}_{j,k} - \sum_{k=0}^{i-1} q_{j,k} \geq 1, \\ 0 & \text{else.} \end{cases} \tag{2.67}$$

- Rounding strategy SUR-0.5 (sum-up rounding with a different threshold)

$$q_{j,i} = \begin{cases} 1 & \text{if } \sum_{k=0}^{i} \tilde{q}_{j,k} - \sum_{k=0}^{i-1} q_{j,k} \geq 0.5, \\ 0 & \text{else.} \end{cases}$$

If the control function has to fulfill the (SOS-1) restriction (as it arises from a convexification, see its definition in Def. 7), the above SUR strategies are not enough. For these problems with the (SOS-1) property, we use one of the following strategies

- Rounding strategy SR-SOS-1 (standard)

$$q_{j,i} = \begin{cases} 1 & \text{if } \tilde{q}_{j,i} \geq \tilde{q}_{k,i}, \forall k \neq j, \text{and } j < k, \forall k : \tilde{q}_{j,i} = \tilde{q}_{k,i}, \\ 0 & \text{else.} \end{cases}$$

- Rounding strategy SUR-SOS-1 (sum-up rounding)

$$q_{j,i} = \begin{cases} 1 & \text{if } \hat{q}_{j,i} \geq \hat{q}_{k,i}, \forall k \neq j, \text{and } j < k, \forall k : \hat{q}_{j,i} = \hat{q}_{k,i}, \\ 0 & \text{else.} \end{cases}$$

  where $\hat{q}_{j,i} = \sum_{k=0}^{i} \tilde{q}_{j,k} - \sum_{k=0}^{i-1} q_{j,k}$.

- Direct SOS-1 Rounding (Definition 2.15, [76])

$$q_{j,i} = \begin{cases} 1 & \text{if } (\forall k : \int_{t_i}^{t_{i+1}} \tilde{q}_{ji} dt \geq \int_{t_i}^{t_{i+1}} \tilde{q}_{ki} dt) \\ & \wedge (\forall k : \int_{t_i}^{t_{i+1}} \tilde{q}_{ji} dt = \int_{t_i}^{t_{i+1}} \tilde{q}_{ki} dt : \ j < k), \quad 1 \leq j \leq n^w. \\ 0 & \text{otherwise.} \end{cases}$$

- SOS-1 - SUR Rounding

$$q_{j,i} = \begin{cases} 1 & \text{if } (\forall k : \hat{q}_{j,i} \geq \hat{q}_{k,j}) \wedge (\forall k, \hat{q}_{j,i} = \hat{q}_{k,j} : j < k) \\ 0 & \text{else} \end{cases} \quad 1 \leq j \leq n^w,$$

  therein, control $\hat{q}_{j,i}$ for the j step of sum-up rounding be defined as

$$\hat{q}_{j,i} := \int_{t_0}^{t_i} \beta_j^*(t) dt - \sum_{k=0}^{i-1} q_{j,k}.$$

- Vanishing Constraint SOS-Sum-Up Rounding (VC-SOS-SUR)

$$q_{j,i} = \left[j = \arg \max_{\substack{k \in [|V|] \\ \int_{t_i}^{t_{i+1}} \beta_k(t)dt > 0}} \int_{t_0}^{t_{i+1}} \beta_k^*(t)dt - \int_{t_0}^{t_i} q_k^{VC}(t)dt \right]$$

  where $V := \{v_1, \ldots, v_{|V|}\}$, $[|V|] := \{1, \ldots, |V|\} \subset \mathbb{N}$, $q^{VC}|(t_i, t_{i+1}) := (q_{j,i})_{j \in [|V|]}$.
- Next force rounding strategy for SOS-1-coupled controls, see [73, Alg. 4.1].
- A developed rounding scheme for computing integer feedback solutions: the neighboring feedback law for the switching aware rounding (see Subsection 4.2.2).

In conclusion, we could obtain that

$$w^*(t) = \begin{cases} \alpha^*(t) & \text{if } \alpha^*(t) \in \{0, 1\}, \\ RS(\alpha^*(t)) & \text{if } \alpha^*(t) \in (0, 1). \end{cases}$$

with RS denote for rounding strategies.

2.6.8 Neighboring Feedback Control

The neighboring feedback control was developed by Kramer-Eis, Bock, et. al., cf. [83,84]. In this approach, ones use the fact that the optimal controls $(u(t), w(t))$ maximize the Hamiltonian $\mathcal{H}(x(t), \lambda(t), u, w)$ pointwise a.e., from which leading to the relations $(u^*(x, \lambda), w^*(x, \lambda))$

$$(u^*(x, \lambda), w^*(x, \lambda)) = \arg \max_{u \in \mathcal{U}, w \in \mathcal{W}} \mathcal{H}(x(t), \lambda(t), u, w).$$

Exploiting the implicitly function theorem, see Section 2.1.4, on the MPBVP derived from the maximum principle along its solution, ones can express $\lambda(\hat{t})$ as a function of the initial value $x(\hat{t})$, from which the derivative $\Lambda(\hat{t}) := \frac{\partial \lambda(\hat{t})}{\partial x(\hat{t})}$ can be computed from the Jacobian of the MPBVP. Consequently, one obtains a neighboring feedback law of the following type

$$(u^{**}(x, \lambda), w^{**}(x, \lambda)) = \arg \max_{u \in \mathcal{U}, w \in \mathcal{W}} \mathcal{H}(\hat{x}, \lambda(\hat{t}) + \Lambda(\hat{t})(\hat{x} - x(\hat{t})), u, w),$$

where $\hat{x}$ is the estimated perturbed value of the state $x(t)$.
For more details and the application of this feedback control, readers can see in Section 4.2.2.

Chapter 3

Indirect Approach for Switched Optimal Control Problem: Maximum Principle

The classical indirect approach is based on the necessary conditions of optimality called Pontryagin's maximum principle of the 1960s, which had an enormous impact on solving engineering problems. Depending on the given optimal control problem, the optimality conditions lead to multi-point BVP. It consists of differential equations for the state and adjoint variables, an algebraic equation for the control, and boundary conditions for the states, the adjoint variables, and time.
For the Maximum Principle, one can read on the paper of Pesch & Bulirsch, see [103], about the historical creation and development. For a survey of OCP in the various forms of Pontryagin's maximum principle, the readers can see [63, 124]. Nowadays, the theory has been extended in many ways, for instance, see [17], [19], [43–48], and Kostina et. al., cf. [82]. The necessary optimality conditions for problems with mixed constraints, where the regularity assumption had been observed can been found in the work of, e.g., [4, 13, 38, 42, 44]. The GDP reformulation and LMP are exploited to obtain the optimality condition for SwOCP, cf. [121]. Then in [122], the mixed state-control constraints in regular and nonregular (or irregular) are carefully studied.

In this chapter, the Local Maximum Principle is first considered in Section 3.1 and Section 3.2 to derive the necessary optimality conditions when the mixed state-control constraints become nonregular (irregular). Therein, in Subsection 3.1.3 we consider Filippov's rule, which is employed as an innovative reformulation, and discuss how this reformulation affects the solution. The convexification of the velocity set is also investigated. Finally, numerical results with New York subway problem is mentioned in Section 3.3.

3.1 Maximum Principle for SOCP

This section deals with a simple SwOCP (SOCP), therein the controls are considered without integer control w. SOCP is reformulated by Filippov's rule, then the optimality condition is obtained by exploiting local maximum principle.

3.1.1 Reformulation for SOCP with Filippov's Solution

We start by considering a SOCP as follows

$$\begin{aligned} \min \quad & \varphi\left(x(t_f)\right) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} f_+\left(x(t)\right) + b(x)u(t), & \text{if } \sigma(x(t)) > 0, \\ f_-\left(x(t)\right) + b(x)u(t), & \text{if } \sigma(x(t)) < 0, \end{cases} \quad t \in [t_0, t_f], \\ & u(t) \in \mathcal{U}_0 \overset{\text{def}}{=} [-1, 1],\ x(t_0) = x_0,\ r(x(t_f)) \leq 0, \end{aligned} \tag{3.1}$$

where the process $x : \mathcal{T} \to \mathcal{X}$ is determined by a discontinuous dynamical system, affected by (piecewise) continuously-valued control function $u : \mathcal{T} \to \mathcal{U}_0 \in \mathbb{R}$ on a time horizon $\mathcal{T} := [t_0, t_f] \subset \mathbb{R}$, such that an objective function $\varphi : \mathcal{X} \to \mathbb{R}$ is minimized. The right hand side function is determined by the sign structure of the switching function $\sigma : \mathcal{X} \to \mathbb{R}$. Point constraints $r(x(t_f)) \leq 0$ with $r : \mathcal{X} \to \mathbb{R}^{n_r}$ must be satisfied.
The discontinuity in (3.1) may lead to the situation, where classical solution does not exist for $\sigma(x(t)) = 0$. Hence, we redefine solution of problem (3.1) according to the Filippov's rule, as follows

$$\dot{x}(t) = \begin{cases} f_+\left(x(t)\right) + b(x)u(t) =: rs_+(x,u), & \text{if } \sigma(x(t)) > 0, \\ f_-\left(x(t)\right) + b(x)u(t) =: rs_-(x,u), & \text{if } \sigma(x(t)) < 0, \\ \alpha(t) rs_+(x,u) + (1-\alpha(t))\, rs_-(x,u), & \text{if } \sigma(x(t)) = 0, \end{cases} \quad t \in \mathcal{T}, \tag{3.2}$$

where $\alpha(t) \in [0,1]$. The "if" formulation in (3.2) is written not in analytical form, so we propose to reformulate problem (3.1) with the redefinition (3.2) and additional mixed state-control constraints as a following problem

$$\min \quad \varphi\left(x(t_f)\right) \tag{3.3a}$$

$$\text{s.t.} \quad \dot{x}(t) = F(x(t), u(t), \alpha(t)),\ t \in \mathcal{T}, \tag{3.3b}$$

$$u(t) \in \mathcal{U}_0, t \in \mathcal{T},\ x(t_0) = x_0, r(x(t_f)) \leq 0, \tag{3.3c}$$

$$\alpha(t)\sigma(x(t)) \geq 0, (1-\alpha(t))\sigma(x(t)) \leq 0,\ t \in \mathcal{T}, \tag{3.3d}$$

$$\alpha(t) \in [0,1],\ t \in \mathcal{T}, \tag{3.3e}$$

where $F(x,u,\alpha) := \alpha(t) rs_+(x,u) + (1-\alpha(t))\, rs_-(x,u)$, $t \in \mathcal{T}$.
Let us discuss the additional constraints (3.3d).
If $\sigma(x(t)) > 0$, then we obtain $\alpha(t) \geq 1$, so a unique possible candidate is $\alpha(t) = 1$.
On the other hand, if $\sigma(x(t)) < 0$, then one implies $\alpha(t) \leq 0$, thus a unique possible candidate is $\alpha(t) = 0$.
For the remain case, i.e., $\sigma(x(t)) = 0$, the constraints (3.3d) hold true for all $\alpha(t) \in [0,1]$. These constraints become the vanishing constraints, leading to difficulties in formulating their optimality conditions.

We have some discussions about the regularity of mixed constraints (3.3d) (see Subsection 2.3.5). For convenience, denote $\mathcal{G}_1(\alpha, x) := -\alpha(t)\sigma(x(t))$, and $\mathcal{G}_2(\alpha, x) := (1-\alpha(t))\sigma(x(t))$. If $\sigma(x(t)) \neq 0$ then mixed constraints are regular (see Subsection 2.3.5). Indeed, if $\alpha(t) = 0$ then $\mathcal{G}_1 = 0$, $\mathcal{G}_2 = \sigma(x(t)) \neq 0$, and $\frac{\partial \mathcal{G}_1}{\partial(u,\alpha)} = \begin{pmatrix} 0 & -\sigma(x(t)) \end{pmatrix} \neq 0$; otherwise if $\alpha(t) = 1$ then $\mathcal{G}_1 = \sigma(x(t)) \neq 0$, $\mathcal{G}_2 = 0$, and $\frac{\partial \mathcal{G}_2}{\partial(u,\alpha)} = \begin{pmatrix} 0 & -\sigma(x(t)) \end{pmatrix} \neq 0$.
If $\sigma(x(t)) = 0$ then the constraints are nonregular, since $\mathcal{G}_1 = \mathcal{G}_1 = 0$, $\frac{\partial \mathcal{G}_2}{\partial(u,\alpha)} = \frac{\partial \mathcal{G}_2}{\partial(u,\alpha)} = 0$.

Lemma 3. *For SOCP* (3.3)*, the set of phase points of the mixed state-control constraints is determined by*

$$\mathcal{N}(\mathcal{G}) = \{(t, x, u, \alpha) \mid \alpha(t) \in [0,1], u(t) \in \mathcal{U}_0, t \in \mathcal{T} : \sigma(x(t)) = 0\}.$$

Proof. For problem (3.3) we have $d(\mathcal{G}) = 2$, $\mathcal{G} = \begin{pmatrix}\mathcal{G}_1 & \mathcal{G}_2\end{pmatrix}^T$, $y = (\alpha, x)$, and

$$\frac{\partial \mathcal{G}(y)}{\partial(u,\alpha)} = \begin{pmatrix}0 & -\sigma \\ 0 & -\sigma\end{pmatrix} = 0, \text{ if } \sigma(x(t)) = 0.$$

To determine the phase points we will find $a \in \Delta$ (see Subsection 2.3.5), $a = (a_1, a_2) \in \mathbb{R}^2$, $a \geq 0$ such that $a_1 + a_2 = 1$, and $a^T \mathcal{G}(y) = 0$, $a^T \frac{\partial \mathcal{G}(y)}{\partial(u,\alpha)} = 0$, i.e.,

$$\begin{cases} -a_1\alpha\sigma + a_2(1-\alpha)\sigma & = 0 \\ -a_1\sigma - a_2\sigma & = 0. \end{cases} \tag{3.4}$$

If $\sigma = 0$ then (3.4) is satisfied for all $a \geq 0$, $|a_1| + |a_2| = 1$.
If $\sigma \neq 0$ then the second equation of (3.4) is equivalent to $a_1 + a_2 = 0$, which contradict to $a \geq 0$, $|a_1| + |a_2| = 1$. This means that for $\sigma \neq 0$, mixed constraints are regular.
Therefore, for $\sigma = 0$ mixed constraints are irregular and phase points are

$$\mathcal{N}(\mathcal{G}) = \{(t, x, u, \alpha) \mid \alpha(t) \in [0,1], u(t) \in \mathcal{U}_0, t \in \mathcal{T} : \sigma(x(t)) = 0\}.$$

□

In the case $\sigma(x(t)) = 0$, the phase jump (see Def. 24) is defined as

$$s(t) = a^T \begin{pmatrix} -\alpha(t)\frac{\partial\sigma(x(t))}{\partial x} \\ (1-\alpha(t))\frac{\partial\sigma(x(t))}{\partial x}\end{pmatrix}, \quad a = (a_1, a_2) \in \mathbb{R}^2,\ a \geq 0,\ a_1 + a_2 = 1,\ t \in \mathcal{T}, \tag{3.5}$$

resulting in $s(t) = (a_2 - \alpha)\frac{\partial\sigma}{\partial x}$, $0 \leq a_2 \leq 1$, $0 \leq \alpha \leq 1$, $t \in \mathcal{T}$.

Remark 15. Let us discuss the phase jump (3.5) w.r.t. the values of α.
If $\alpha = 0$ then the phase jump is assumed to be $s = \frac{\partial\sigma}{\partial x}$.
If $\alpha = 1$ then the phase jump is $s = -\frac{\partial\sigma}{\partial x}$.
If $0 < \alpha < 1$ then the phase jump is assumed to be $s = b_s\frac{\partial\sigma}{\partial x}$, where $-1 \leq b_s \leq 1$.

To formulate the maximum principle, we exploit LMP (see Theorem 9) for the SOCP (3.3). We assume that the controls $u(t)$ and $\alpha(t)$ are piecewise continuous and we introduce the necessary notation as follows. Denote

$$\mathcal{V}(x) := \{\alpha \in [0,1] : \alpha\sigma \geq 0, (1-\alpha)\sigma \leq 0\}, \tag{3.6}$$
$$V(x) := \{(u, \alpha) : u \in \mathcal{U}_0,\ \alpha \in \mathcal{V}(x)\}. \tag{3.7}$$

3.1.2 Discussion on Optimality Conditions

Let $\overset{\circ}{\lambda}$ is a measure of function $\lambda(t)$, given the relations

$$\overset{\circ}{\lambda}([t_1, t_2]) = \lambda(t_2) - \lambda(t_1), \quad t_2 > t_1,$$

where function $\lambda(t)$ is assumed to be a left-continuous function of bounded variation.
The maximum principle (see Subsection 2.3.5) can be formulated as follows. Let $x^*(\cdot)$, $u^*(\cdot), \alpha^*(t)$ be optimal trajectory and (piecewise continuous) controls. Then there exist

- a vector $v \geq 0$ and a number v_0,
- a function of bounded variation $\lambda(t)$, $t \in \mathcal{T} = [t_0, t_f]$,
- a function $\vartheta(\cdot) \in L_1$, $\vartheta(t) \geq 0$ if $t \in \{T_* : \alpha(t) = 0\}$, $\vartheta(t) \leq 0$ if $t \in \{T_* : \alpha(t) = 1\}$, and $\vartheta(t) = 0$ with $t \in \mathcal{T} \setminus T_*$, where $T_* := \{t \in \mathcal{T} \mid \sigma(x^*(t)) = 0\}$,
- the measure $\mu, d\mu \geq 0$, focused on the set T_*, function $\lambda(\cdot), \vartheta(\cdot)$ and measure μ are related by the relations:

$$\overset{\circ}{\lambda}{}^T(dt) = \begin{cases} -\lambda^T(t)\left(\alpha^*(t)\frac{\partial f_+(x^*(t))}{\partial x} + (1-\alpha^*(t))\frac{\partial f_-(x^*(t))}{\partial x} + \frac{\partial b(x^*(t))}{\partial x}u^*(t)\right)dt + \Sigma_{dt}^{d\mu},\ t \in T_* \\ -\lambda^T(t)\left(\frac{\partial f_-(x^*(t))}{\partial x} + \frac{\partial b(x^*(t))}{\partial x}u^*(t)\right),\ t \in T_-, \\ -\lambda^T(t)\left(\frac{\partial f_+(x^*(t))}{\partial x} + \frac{\partial b(x^*(t))}{\partial x}u^*(t)\right),\ t \in T_+, \end{cases} \tag{3.8}$$

where $\Sigma_{dt}^{d\mu} := \vartheta(t)\frac{\partial \sigma(x^*(t))}{\partial x}dt + s(t)d\mu$, $T_+ := \{t \in \mathcal{T} \mid \sigma(x^*(t)) > 0\}$, $T_- := \{t \in \mathcal{T} \mid \sigma(x^*(t)) < 0\}$, and

$$\lambda(t_f - 0) = -v^T\frac{\partial r(x^*(t_f))}{\partial x} - v_0\frac{\partial \varphi(x^*(t_f))}{\partial x} - \mu(t_f)s(t_f), \tag{3.9}$$

at each point $t \in [t_0, t_f)$ of the discontinuity of μ we have

$$\lambda(t+0) - \lambda(t-0) = \mu(t)s(t), \tag{3.10}$$

normalization conditions are fulfilled

$$v_0 + \|v\| + \|\mu\| + \int_{\mathcal{T}} \vartheta(t)dt = 1. \tag{3.11}$$

Here, the maximum condition is satisfied

$$\mathcal{H}(x^*(t), u^*(t), \alpha^*(t), \lambda(t)) = \max_{(u,\alpha)\in V(x^*(t))} \mathcal{H}(x^*(t), u(t), \alpha(t), \lambda(t)),\ \text{a.e. } t \in \mathcal{T}, \tag{3.12}$$

where $\mathcal{H}(x, u, \alpha, \lambda) = \lambda^T F(x, u, \alpha) = \lambda(t)^T\left(\alpha(t)f_+(x(t)) + (1-\alpha(t))f_-(x(t)) + b(x(t))u(t)\right)$.
The measure μ is involved in the formulation of this maximum principle. In general the measure μ decomposes into a sum:

$$\mu = \mu_a + \mu_s + \mu_\delta,$$

where

- μ_a is an absolutely continuous component (in measure dt),
- μ_s is a singular component (in measure dt),

- μ_δ is an atomic (jump) component (in measure dt).

Let us analyze the maximum conditions (3.12). Consider the set $V(x^*(t))$ (see Eq. (3.7)), conditions (3.12) take the form:

$$\begin{aligned}\mathcal{H}(x^*,u^*,\alpha^*,\lambda(t)) &= \max_{(u,\alpha)\in V(x^*)} \lambda^T(t)\left(\alpha f_+(x^*)+(1-\alpha)f_-(x^*)+b(x^*)u\right), \text{a.e. in } \mathcal{T},\\ \text{s.t.}\quad & |u(t)|\le 1,\\ & \alpha(t)\sigma(x^*(t))\ge 0,\ (1-\alpha(t))\sigma(x^*(t))\le 0,\quad 0\le\alpha(t)\le 1,\end{aligned} \tag{3.13}$$

Let us analyze the last conditions in more details as follows.
A. If $\sigma(x^*(t))>0$, then it follows that $\alpha=1$. Therefore, the conditions of the maximum principle (3.13) take the form

$$\mathcal{H}(x^*,u^*,\alpha^*,\lambda(t)) = \lambda^T(t)f_+(x^*(t)) + \max_{u(t)\in[-1,1]}\{\lambda^T(t)b(x^*)u(t)\},\quad t\in T_+.$$

B. If $\sigma(x^*(t))<0$ then $\alpha=0$. Similarly, the conditions of the maximum principle (3.13) get the form

$$\mathcal{H}(x^*,u^*,\alpha^*,\lambda(t)) = \lambda^T(t)f_-(x^*(t)) + \max_{u(t)\in[-1,1]}\{\lambda^T(t)b(x^*)u(t)\},\quad t\in T_-.$$

C. If $\sigma(x^*(t))=0$ then the maximum conditions (3.13) take the form

$$\mathcal{H}(x^*,u^*,\alpha^*,\lambda(t)) = \max_{\alpha(t)\in[0,1]}\lambda^T(t)(\alpha(t)f_+(x^*)+(1-\alpha(t))f_-(x^*)) + \max_{u(t)\in[-1,1]}\lambda^T(t)b(x^*)u(t), t\in T_*.$$

Let us summarize optimal conditions. We can rewrite relations (3.8),(3.10) in an equivalent form as the following conditions. As a results, ones obtain the following theorem.

Theorem 12 (Optimality conditions for SOCP (3.3))**.** *Let $u^*(t),\alpha^*(t)$ and $x^*(t)$, $t\in\mathcal{T}$ be the optimal control and trajectory of problem (3.3), where assuming control $u^*(t)$ is piecewise continuous. Then there exist $\lambda(t)$, bounded variation, such that the measure $\overset{\circ}{\lambda}$ is absolutely continuous w.r.t. the measure $d\mu+dt$; there exists a vector v, a number v_0, a function $\vartheta\in L_1$, and a measure μ such that condition* (3.11) *is fulfilled, and the adjoint system has the form*

$$\frac{d\lambda(t)}{dt}\overset{dt}{=}\begin{cases}-\lambda^T(t)\frac{\partial(f_+(x^*)+b(x^*)u^*)}{\partial x}, & t\in T_+,\\ -\lambda^T(t)\frac{\partial(f_-(x^*)+b(x^*)u^*)}{\partial x}, & t\in T_-,\\ -\lambda^T(t)\frac{\partial F(x^*,u^*,\alpha^*)}{\partial x}+\left(\vartheta(t)\pm\frac{d\mu_a}{dt}\right)\frac{\partial\sigma(x^*(t))}{\partial x}dt, & t\in T_*,\end{cases} \tag{3.14}$$

$$\frac{d\lambda}{d\mu_s}\overset{d\mu_s}{=}\pm\frac{\partial\sigma(x^*(t))}{\partial x},\quad t\in T_*, \tag{3.15}$$

$$\lambda(t+0)=\lambda(t-0)+(\mu(t-0)+\mu(t+0))\frac{\partial\sigma(x^*(t))}{\partial x},\quad t\in\mathcal{D}\cap T_*, \tag{3.16}$$

where $\vartheta(t)$ is a Lebesgue absolutely integrable function, with $\vartheta(t) = 0$ if $t \in \mathcal{T} \setminus T_*$, $\mathcal{D}$ is the set of jumps of measure μ_δ, $\overset{dt}{=}$ means equality holds a.e. in the measure dt, and

$$\lambda(t_f - 0) = -v^T \frac{\partial r(x^*(t_f)}{\partial x} - v_0 \frac{\partial \varphi(x^*(t_f)}{\partial x} - \mu(t_f)s(t_f),$$

and the optimal conditions take the following forms

$$\vartheta(t) \geq 0, \ \text{if } \alpha^*(t) = 0, \quad \vartheta(t) \leq 0, \ \text{if } \alpha^*(t) = 1, \quad t \in T_*, \tag{3.17}$$

$$\mu(t) \geq 0, \ \text{if } \alpha^*(t) = 0, \quad \mu(t) \leq 0, \ \text{if } \alpha^*(t) = 1, \quad t \in T_*, \tag{3.18}$$

$$\lambda^T(t) b(x^*(t)) u^*(t) = \max_{u \in \mathcal{U}_0} \{\lambda^T(t) b(x^*(t)) u\}, \quad t \in \mathcal{T}, \tag{3.19}$$

$$\lambda^T(t) \left(f_+(x^*(t)) - f_-(x^*(t))\right) \begin{cases} = 0, & \text{if } 0 < \alpha^*(t) < 1, \\ \leq 0, & \text{if } \alpha^*(t) = 0, \\ \geq 0, & \text{if } \alpha^*(t) = 1, \end{cases} \quad t \in T_*, \tag{3.20}$$

Furthermore, ones have

$$\begin{aligned} &\lambda^T(\tau_i - 0) F(x^*(\tau_i - 0), u^*(\tau_i - 0), \alpha^*(\tau_i - 0)) \\ &= \lambda^T(\tau_i + 0) F(x^*(\tau_i + 0), u^*(\tau_i + 0), \alpha^*(\tau_i + 0)), \ i = 1, \ldots, p, \end{aligned} \tag{3.21}$$

where τ_i, $i = 1, \ldots, p$, be the minimum number of points such that:

$$\begin{aligned} &0 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_p < \tau_{p+1} = t^*, \\ &int\ T_* = \bigcup_{i \in N^*} (\tau_i, \tau_{i+1}), \quad int\ T_+ = \bigcup_{i \in N^+} (\tau_i, \tau_{i+1}), \quad int\ T_- = \bigcup_{i \in N^-} (\tau_i, \tau_{i+1}), \\ &N^* \cup N^- \cup N^+ = \{0, 1, \ldots, p\}. \end{aligned} \tag{3.22}$$

Remark 16. There are no examples in literatures, see [133–135], where μ contains the singular component. In case μ does not contain the singular component, (3.15) disappears.

Example 3.1. [81, Example 1 ($c = 0$)] Consider the optimal control problem

$$\begin{aligned} \min_{x,u} \quad & \varphi(x(2)) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} f_+(x(t)) + bu(t), & \text{if } \sigma > 0, \\ f_-(x(t)) + bu(t), & \text{if } \sigma < 0, \end{cases} \\ & |u(t)| \leq 1, \ t \in [-0.5, 2], \\ & x(-0.5) = x_0, \ r(x(2)) = 0, \ x_0^T = (19/32, -37/16, -3/4), \end{aligned} \tag{3.23}$$

where $\varphi(x(t)) := x_1(t) - 2.5x_2(t)$, $r(x(t)) := x_3(t) - 1$, switching function $\sigma(x(t)) = -x_3(t)$, $t \in [-0.5, 2]$, and $f_+(\cdot) = \begin{pmatrix} x_2 & x_3 + 5 & 0.5 \end{pmatrix}^T$, $f_-(\cdot) = \begin{pmatrix} x_2 & x_3 & 0 \end{pmatrix}^T$, $b = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T$. Reformulate problem (3.23) by using Filippov's rule and relaxation, one obtains

$$\begin{aligned} \min_{x,u,\alpha} \quad & x_1(2) - 2.5x_2(2) \\ \text{s.t.} \quad & \dot{x}(t) = F(x(t), u(t), \alpha(t)), \ t \in [-0.5, 2], \\ & |u(t)| \leq 1, \ 0 \leq \alpha(t) \leq 1, \ t \in [-0.5, 2], \\ & \mathcal{G}_j(\alpha, x) \leq 0, \ j = 1, 2, \\ & x(-0.5) = x_0, \ x_3(2) = 1, \ x_0^T = (19/32, -37/16, -3/4), \end{aligned} \tag{3.24}$$

where $\mathcal{G}_1(\cdot) = \alpha(t)x_3(t)$, $\mathcal{G}_2(\cdot) = (\alpha(t)-1)x_3(t)$, $t \in [-0.5, 2]$, and

$$F(\cdot) = \alpha f_+(\cdot) + (1-\alpha)f_-(\cdot) + bu = \begin{pmatrix} x_2 & x_3 + 5\alpha & u + 0.5\alpha \end{pmatrix}^T.$$

It is easy to check that the mixed constraints $\mathcal{G}_j(\cdot)$, $j = 1, 2$, are irregular when $\sigma = 0$. Then the set of phase points of these mixed constraints is determined as

$$\mathcal{N}(\mathcal{G}) = \left\{(x, u, \alpha) \in \mathbb{R}^5 : x_3 = 0\right\}.$$

Consider the control $u^*(\cdot)$, $\alpha^*(\cdot)$:

$$u^*(t) = \begin{cases} 1, & \text{if } t \in [-0.5, 0], \\ -0.5, & \text{if } t \in [0, 1], \\ 1, & \text{if } t \in [1, 2], \end{cases} \qquad \alpha^*(t) = \begin{cases} 1, & \text{if } t \in [-0.5, 0], \\ 1, & \text{if } t \in [0, 1], \\ 0, & \text{if } t \in [1, 2]. \end{cases} \tag{3.25}$$

The corresponding trajectory is denoted by $x^*(t) = (x_1^*(t), x_2^*(t), x_3^*(t))$, $t \in [-0.5, 2]$. Then, using the initial conditions, we have

$$\dot{x}_3(t) = u + 0.5\alpha = \begin{cases} 1.5, & \text{if } t \in [-0.5, 0], \\ 0, & \text{if } t \in [0, 1], \\ 1, & \text{if } t \in [1, 2], \end{cases} \Leftrightarrow x_3^*(t) = \begin{cases} 1.5t, & \text{if } t \in [-0.5, 0], \\ 0, & \text{if } t \in [0, 1], \\ t - 1, & \text{if } t \in [1, 2], \end{cases}$$

here, $D = [0, 1]$ is a switching interval of x_3, see Fig 3.1.

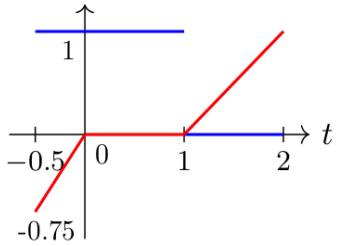


Figure 3.1: Trajectory x_3^* (red) switches at $t = 0$, $t = 1$; control α^* (blue) jumps at $t = 1$.

Let us show that the process (x^*, u^*, α^*) satisfies LMP in problem (3.23).
1. Consider the segment $t \in [1, 2]$. Boundary condition for solution to the adjoint system (see (3.9)) has the form

$$\lambda(2) = -v_0 \frac{\partial \varphi(x^*(2))}{\partial x} - v \frac{\partial r(x^*(2))}{\partial x} = (-v_0, 2.5v_0, -v)^T, \quad v_0 \geq 0, \tag{3.26}$$

here $s(2) = 0$, i.e., there are no phase jumps at $t = 2$.
With $v_0 = 1$, we have $\lambda(2) = (-1, 2.5, -v)^T$. For $t \in [1, 2]$, the adjoint system takes the form $\dot{\lambda}^T(t) = -(0, \lambda_1(t), \lambda_2(t))^T$. The adjoint solution with the boundary condition (3.26) is

$$\lambda_1(t) = -1, \ \lambda_2(t) = t + 0.5, \ \lambda_3(t) = -0.5(2-t)^2 + 2.5(2-t) - v, \ t \in [1, 2],$$
$$\lambda(1+0) = (-1, 1.5, 2-v)^T, \ \lambda(1-0) = \lambda(1+0) + \mu s(1) = (-1, 1.5, 2 - v + \mu^1)^T.$$

The optimality condition (see Theorem 12) and inequalities $|u^*(t)| < 1$, imply $\lambda_3(t) = 0$, $t \in [0,1]$. Hence, $\lambda_3(1-0) = 0$ holds true. Exploiting this equality and the equality $\lambda^T(1-0)\dot{x}(1-0) = \lambda^T(1+0)\dot{x}(1+0)$ (see [81, Eq.(20)]), ones get $\mu^1 = v - 2$, $v = -5.5$. Here there is a jump in λ.
2. Consider the segment $t \in [0,1]$. The adjoint system takes the form

$$\dot{\lambda}(t) = (0, -\lambda_1(t), 0)^T, \quad \lambda(1-0) = (-1, 1.5, 0)^T,$$

so, it has the solution

$$\lambda_1(t) = -1,\ \lambda_2(t) = t + 0.5,\ \lambda_3(t) = 0,\ t \in [0,1].$$

Consequently,

$$\lambda(0+0) = (-1, 0.5, 0)^T,\ \lambda(0-0) = \lambda(0+0) + \mu s(0) = (-1, 0.5, \mu^1)^T.$$

The condition (see [81, Eq.(20)]) $\lambda^T(0-0)\dot{x}(0-0) = \lambda^T(0+0)\dot{x}(0+0)$ implies $\mu^1 = 0$.
3. Consider the segment $t \in [-0.5, 0]$. Here the adjoint system takes the form $\dot{\lambda}^T(t) = -(0, \lambda_1(t), \lambda_2(t))^T$ with boundary condition $\lambda(0-0) = (-1, 0.5, 0)^T$, and hence it has the solution

$$\lambda_1(t) = -1,\ \lambda_2(t) = t + 0.5,\ \lambda_3(t) = -t^2/2 - t/2,\ t \in [-0.5, 0].$$

Now let us check the optimality conditions of Theorem 12. Here we have

$$T_* = \{t \in [-0.5, 2] \mid \sigma(x^*(t)) = 0\} = \{t \in [-0.5, 2] \mid -x_3^*(t) = 0\} = [0,1].$$

The normalization condition implies

$$\int_{-0.5}^{2} \vartheta(t)dt = 1 - v_0 - \|v\|\|\mu\| = -5.5\|\mu\| = 0,\ t \in T_*,$$

so $\vartheta(t) \le 0$ for all $t \in T_*$, i.e., condition (3.17) holds true.
Since $\mu^1(t) = 0 \le 0$, for all $t \in T_*$, hence condition (3.18) also holds true.
The left hand side and right hand side of condition (3.19) are respectively equal to

$$\begin{aligned} lhs &= \lambda^T F(x^*, u^*, \alpha^*) = (-1, t+0.5, \lambda_3)^T (x_2^*, x_3^* + 5\alpha^*, u^* + 0.5\alpha^*) \\ &= (-x_2^* + x_3^*(t+0.5)) + 5\alpha^*(t+0.5) + \lambda_3(u^* + 0.5\alpha^*), \qquad (3.27) \\ rhs &= \lambda^T f_-(x^*) + \max_{\alpha \in [0,1]} \{\lambda^T(\alpha(f_+(x^*) - f_-(x^*)))\} + \max_{u \in [-1,1]} \{\lambda^T b u\}, \\ &= (-x_2^* + x_3^*(t+0.5)) + \max_{\alpha \in [0,1]} \{\alpha(5(t+0.5) + 0.5\lambda_3)\} + \max_{u \in [-1,1]} \{\lambda_3 u\}. \qquad (3.28) \end{aligned}$$

By comparing (3.27) with (3.28), the optimality condition (3.19) holds true.
Moreover, condition (3.20) holds true, since

$$\lambda^*(t)^T(f_+(x^*(t)) - f_-(x^*(t))) = (-1, t+0.5, 0)^T(0, 5, 0.5) = 5(t+0.5) \ge 0, \forall t \in T_*.$$

Summing up, the local optimal control (3.25) satisfies the optimality conditions formulated in Theorem 12.

3.1.3 Discussion on Filippov's Rule

This subsection deals with SOCP's reformulation by using convexification, i.e., we apply Filippov's rule correctly. We consider the following OCP with discontinuous dynamics of nonlinear form on the right hand side of the ODE constraint

$$\begin{aligned} \min \quad & \varphi\left(x(t_f)\right) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} f_+\left(x(t)\right) + b_+(x(t))u(t), & \text{if } \sigma(x(t)) > 0 \\ f_-\left(x(t)\right) + b_-(x(t))u(t), & \text{if } \sigma(x(t)) < 0 \end{cases} \\ & u(t) \in \mathcal{U} \stackrel{\text{def}}{=} [u_l, u_u], t \in [t_0, t_f],\ x(t_0) = x_0,\ r(x(t_f)) \geq 0. \end{aligned} \tag{3.29}$$

Due to discontinuity, a classical solution of the ODE model may not exist. Then problem (3.29) must be reformulated according to Filippov's rule,

$$\begin{aligned} \min \quad & \varphi\left(x(t_f)\right) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} f_+\left(x(t)\right) + b_+(x(t))u(t) =: rhs_+, & \text{if } \sigma(x(t)) > 0 \\ f_-\left(x(t)\right) + b_-(x(t))u(t) =: rhs_-, & \text{if } \sigma(x(t)) < 0 \\ \alpha(t) rhs_+ + (1-\alpha(t))\, rhs_-, & \text{if } \sigma(x(t)) = 0 \end{cases} \\ & u(t) \in \mathcal{U}, t \in [t_0, t_f],\ x(t_0) = x_0,\ r(x(t_f)) \geq 0, \\ & \alpha(t) \in [0,1],\ t \in [t_0, t_f]. \end{aligned} \tag{3.30}$$

In problem (3.30) the velocity set

$$U(x) := \{v \in \mathbb{R}^n : v = \alpha\left(f_+(\cdot) + b_+(\cdot)u\right) + (1-\alpha)\left(f_-(\cdot) + b_-(\cdot)u\right), \alpha \in [0,1], u \in \mathcal{U}\}. \tag{3.31}$$

Remind that Filippov's rule works in the previous problem (3.1), but in problem (3.29) it does not work correctly since ones can show that problem 3.30 does not have solution, see Appendix A.1 for a numerical example with clear details and explanations. What is the reason? Is Filippov's rule incorrect? The answer is in the incorrect application of Filippov's rule due to the non-convexity of the set $U(x)$ – (3.31). Hence, correct applying Filippov's rule is necessary. In particular, we replace the set $U(x)$ by its "convex hull" $\operatorname{conv}(U(x))$.

Lemma 4. *Let the set $U(x)$ be defined by (3.31). Then $\operatorname{conv}(U(x)) = U^*(x)$, where*

$$U^*(x) = \{v \in \mathbb{R}^n : v = \alpha f_+ + (1-\alpha) f_- + \beta_1 b_+ + \beta_2 b_-, \alpha \in [0,1], \beta_1 \in T_1, \beta_2 \in T_2\}, \tag{3.32}$$

where $T_1 := [\alpha u_l, \alpha u_u]$, $T_2 := [(1-\alpha)u_l, (1-\alpha)u_u]$.

Proof. Let us first show that the set $U^*(x)$ is convex. Indeed, let $\bar{v} \in U^*(x)$ and $\tilde{v} \in U^*(x)$:

$$\begin{aligned} \bar{v} &= \bar{\alpha} f_+(x) + (1-\bar{\alpha}) f_-(x) + \bar{\beta}_1 b_+ + \bar{\beta}_2 b_-, \\ \tilde{v} &= \tilde{\alpha} f_+(x) + (1-\tilde{\alpha}) f_-(x) + \tilde{\beta}_1 b_+ + \tilde{\beta}_2 b_-, \end{aligned}$$

where $\bar{\alpha} \in [0,1]$, $\bar{\alpha} u_l \leq \bar{\beta}_1 \leq \bar{\alpha} u_u$, $(1-\bar{\alpha})u_l \leq \bar{\beta}_2 \leq (1-\bar{\alpha})u_u$, and $\tilde{\alpha} \in [0,1]$, $\tilde{\alpha} u_l \leq \tilde{\beta}_1 \leq \tilde{\alpha} u_u$, $(1-\tilde{\alpha})u_l \leq \tilde{\beta}_2 \leq (1-\tilde{\alpha})u_u$. For $\mu \in [0,1]$ consider the vector

$$\begin{aligned} v(\mu) &= \mu\bar{v} + (1-\mu)\tilde{v} \\ &= (\mu\bar{\alpha} + (1-\mu)\tilde{\alpha}) f_+(x) + (\mu(1-\bar{\alpha}) + (1-\mu)(1-\tilde{\alpha})) f_-(x) \\ &\qquad + (\mu\bar{\beta}_1 + (1-\mu)\tilde{\beta}_1) b_+ + (\mu\bar{\beta}_2 + (1-\mu)\tilde{\beta}_2) b_-. \end{aligned} \tag{3.33}$$

Define $\alpha(\mu) := \mu\bar{\alpha} + (1-\mu)\tilde{\alpha}$, $\beta_j(\mu) := \mu\bar{\beta}_j + (1-\mu)\tilde{\beta}_j$, $j = 1, 2$. Obviously, $\alpha(\mu) \in [0, 1]$,

$$\alpha(\mu)u_l = \mu\bar{\alpha}u_l + (1-\mu)\tilde{\alpha}u_l \le \beta_1(\mu) \le \mu\bar{\alpha}u_u + (1-\mu)\tilde{\alpha}u_u = \alpha(\mu)u_u,$$
$$\mu(1-\bar{\alpha})u_l + (1-\mu)(1-\tilde{\alpha})u_l \le \beta_2(\mu) \le \mu(1-\bar{\alpha})u_u + (1-\mu)(1-\tilde{\alpha})u_u.$$

These relations together with (3.33) imply

$$v(\mu) = \alpha(\mu)f_+(x) + (1-\alpha(\mu))f_-(x) + \beta_1(\mu)b_+ + \beta_2(\mu)b_- \in U^*(x).$$

Hence, $U^*(x)$ is convex.
Next, let us show that

$$U(x) \subset U^*(x). \tag{3.34}$$

Consider $\bar{v} \in U(x)$. Then for some $\bar{\alpha} \in [0, 1]$ and $\bar{u} \in [u_l, u_u]$ we have

$$\bar{v} = \bar{\alpha}(f_+(x) + b_+\bar{u}) + (1-\bar{\alpha})(f_-(x) + b_-\bar{u}).$$

Denote $\alpha := \bar{\alpha}$, $\beta_1 := \bar{\alpha}\bar{u}$, $\beta_2 := (1-\bar{\alpha})\bar{u}$. By construction ones have

$$\alpha \in [0, 1],\ \alpha u_l \le \beta_1 \le \alpha u_u,\ (1-\alpha)u_l \le \beta_2 \le (1-\alpha)u_u,$$
$$\bar{v} = \bar{\alpha}(f_+(x) + b_+\bar{u}) + (1-\bar{\alpha})(f_-(x) + b_-\bar{u}) = \alpha f_+(x) + (1-\alpha)f_-(x) + \beta_1 b_+ + \beta_2 b_-.$$

Therefore, $\bar{v} \in U^*(x)$ and $U(x) \subset U^*(x)$.
Now we will prove that

$$U^*(x) \subset \mathrm{conv}(U(x)). \tag{3.35}$$

Denote $v_{(1)}(x) := f_+(x) + b_+(x)u_l$, $v_{(2)}(x) := f_+(x) + b_+(x)u_u$, $v_{(3)}(x) := f_-(x) + b_-(x)u_l$, $v_{(4)}(x) := f_-(x) + b_-(x)u_u$. Obviously, $v_{(j)}(x) \in U(x)$, $j = 1, \dots, 4$. Hence, the inclusion

$$\bar{U}^*(x) := \{v \in \mathbb{R}^n : v = \sum_{j=1}^{4} \mu_j v_{(j)}(x), \mu_j \ge 0,\ j = 1, \dots, 4,\ \sum_{j=1}^{4} \mu_j = 1\} \subset \mathrm{conv}(U(x))$$

holds true. We may rewrite the set $\bar{U}^*(x)$ as

$$\bar{U}^*(x) := \{v \in \mathbb{R}^n : v = (\mu_1+\mu_2)f_-(x) + (\mu_3+\mu_4)f_+(x) + (\mu_1u_l+\mu_2u_u)b_+ + (\mu_3u_l+\mu_4u_u)b_-\}. \tag{3.36}$$

Let $\tilde{v} \in U^*(x)$, i.e., there exists numbers $\tilde{\alpha}$, $\tilde{\beta}_1$, $\tilde{\beta}_2$, such that

$$\tilde{\alpha} \in [0, 1],\ \tilde{\alpha}u_l\beta_1 \le \tilde{\alpha}u_u,\ (1-\tilde{\alpha})u_l\beta_2 \le (1-\tilde{\alpha})u_u, \tag{3.37}$$
$$\tilde{v} = \tilde{\alpha}f_+(x) + (1-\tilde{\alpha})f_-(x) + \tilde{\beta}_1 b_+ + \tilde{\beta}_2 b_-. \tag{3.38}$$

Define the numbers $\tilde{\mu}_j$, $j = 1, \dots, 4$, as a solution of the following system

$$\tilde{\mu}_1 + \tilde{\mu}_2 = \tilde{\alpha},\ \ \tilde{\mu}_3 + \tilde{\mu}_4 = 1 - \tilde{\alpha},\quad \tilde{\mu}_1u_l + \tilde{\mu}_2u_u = \tilde{\beta}_1,\ \ \tilde{\mu}_3u_l + \tilde{\mu}_4u_u = \tilde{\beta}_2. \tag{3.39}$$

The system (3.39) has a solution

$$\tilde{\mu}_1 = \frac{\tilde{\alpha}u_u - \tilde{\beta}_1}{u_u - u_l},\ \ \tilde{\mu}_2 = \frac{\tilde{\beta}_1 - \tilde{\alpha}u_l}{u_u - u_l},\ \ \tilde{\mu}_3 = \frac{(1-\tilde{\alpha})u_u - \tilde{\beta}_2}{u_u - u_l},\ \ \tilde{\mu}_4 = \frac{\tilde{\beta}_2 - (1-\tilde{\alpha})u_l}{u_u - u_l}. \tag{3.40}$$

Taking into account (3.37) it is easy to show that this solution (3.40) satisfies the relations

$$\tilde{\mu}_j \geq 0, \;\; j = 1, \ldots, 4, \; \sum_{j=1}^{4} \tilde{\mu}_j = 1. \tag{3.41}$$

Combining (3.36), (3.38)-(3.41), one obtains

$$\begin{aligned} \tilde{v} &= \tilde{\alpha} f_+(x) + (1-\tilde{\alpha}) f_-(x) + \tilde{\beta}_1 b_+ + \tilde{\beta}_2 b_- \\ &= (\tilde{\mu}_1 + \tilde{\mu}_2) f_-(x) + (\tilde{\mu}_3 + \tilde{\mu}_4) f_+(x) + (\tilde{\mu}_1 u_l + \tilde{\mu}_2 u_u) b_+ + (\tilde{\mu}_3 u_l + \tilde{\mu}_4 u_u) b_-, \end{aligned}$$

hence $\tilde{v} \in \operatorname{conv}(U(x))$, i.e., the inclusion (3.35) holds true. The convexity of $U^*(x)$ and the inclusion (3.34) imply that $\operatorname{conv}(U(x)) = U^*(x)$.

□

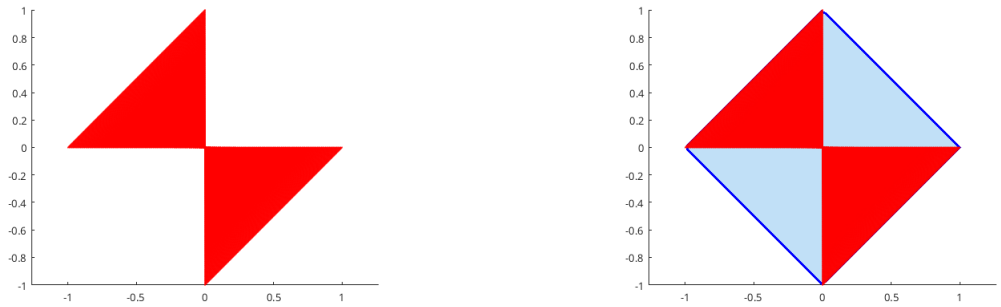


Figure 3.2: For $\mathcal{U} = [-1, 1]$, $b_+ = \begin{pmatrix} -1 & 0 \end{pmatrix}^T$, $b_- = \begin{pmatrix} 0 & 1 \end{pmatrix}^T$, $f_+ = f_- = 0$, the left figure describes the set $U(x)$, see Eq. (3.31), while the right one illustrates the correct application of Filippov's rule, where the polygon (both blue & red) is the set $\operatorname{conv}(U(x))$, see Eq. (3.32).

By applying Filippov's rule correctly for problem (3.30), we obtain the following problem

$$\begin{array}{ll} \min & \varphi\left(x(t_f)\right) \\ \text{s.t.} & \dot{x}(t) = F(x(t), \alpha(t), \beta_1(t), \beta_2(t)), \\ & x(t_0) = x_0, r(x(t_f)) \geq 0, \\ & \alpha(t)\sigma(x(t)) \geq 0, (1-\alpha(t))\sigma(x(t)) \leq 0, \\ & \alpha(t) u_l \leq \beta_1(t) \leq \alpha(t) u_u, \; (1-\alpha(t)) u_l \leq \beta_2(t) \leq (1-\alpha(t)) u_u, \\ & \alpha(t) \in [0, 1], \end{array} \quad t \in [t_0, t_f], \tag{3.42}$$

therein we use the notation, where $t \in [t_0, t_f]$,

$$F(\cdot) := \alpha(t) f_+(x(t)) + (1-\alpha(t)) f_-(x(t)) + \beta_1(t) b_+(x(t)) + \beta_2(t) b_-(x(t)). \tag{3.43}$$

Remark 17. In general, the discontinuity of the right hand side function may lead to the non-existence of the solution of the ODE, so we must carefully apply Filippov's rule, see the problem formulation in the sense of (3.44) with further relaxed reformulation.

3.1.4 Optimality Conditions for SOCP

Here we consider a more general SOCP as follows

$$\begin{aligned} \min_{x(\cdot),u(\cdot)} \quad & \varphi[x(\cdot),u(\cdot)] \\ \text{s.t.} \quad & \dot{x}(t) = f\left(x(t),u(t),\operatorname{sgn}\left(\sigma(x(t))\right)\right), \\ & 0 \le r\left(x(t_f)\right), \\ & t \in \mathcal{T} \stackrel{\text{def}}{=} [t_0,t_f]\,. \end{aligned} \tag{3.44}$$

therein, the dynamic process $x : \mathcal{T} \to \mathcal{X}$ is determined by a dynamical system with right hand side function $f : \mathcal{X} \times \mathcal{U} \times \{-1,0,1\} \to \mathbb{R}^{n_x}$, affected by continuously-valued bounded control function $u : \mathcal{T} \to \mathcal{U} := [u_l,u_u]^{n_u}$ on a time horizon $\mathcal{T} \subset \mathbb{R}$, such that an objective function $\varphi = m(x(t_f))$, is minimized. The right hand side function f is determined by the sign structure of the scalar switching function $\sigma : \mathcal{X} \to \mathbb{R}$. Point constraints $r\left(x(t_0),x(t_f)\right) \ge 0$ with $r : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n_r}$ must be satisfied.
Exploiting the definition of the sign of $\sigma(\cdot)$, we can rewrite (3.44) as follows:

$$\begin{aligned} \min_{x(\cdot),u(\cdot)} \quad & m(x(t_f)) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} f_+\left(x(t),u(t)\right), & \text{if } \sigma(x(t)) \ge 0, \\ f_-\left(x(t),u(t)\right), & \text{if } \sigma(x(t)) \le 0, \end{cases} \qquad t \in \mathcal{T}. \\ & 0 \le r\left(x(t_f)\right), \end{aligned} \tag{3.45}$$

Note that the case $\sigma(\hat{t},x(\hat{t})) = 0$, $\hat{t} \in \mathcal{T}$, is in principle contradictory. If $\hat{t}$ is an isolated zero of $\sigma(x(t))$, right-hand side is chosen to be either f_+ or f_-. If $\sigma(x(t))$ vanishes on an interval, the right-hand side has yet to be determined appropriately. How to use Filippov's rule in this problem?
By reformulating (3.45) with Filippov's rule, we obtain the equivalent problem

$$\begin{aligned} \min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & m(x(t_f)) \\ \text{s.t.} \quad & \dot{x}(t) = \bar{f}(x(t),u(t),\alpha(t)), \\ & 0 \le r\left(x(t_f)\right), \qquad t \in \mathcal{T}, \\ & 0 \le \alpha(t)\sigma(x(t)),\ (1-\alpha(t))\sigma(x(t)) \le 0, \\ & \alpha(t) \in [0,1], \end{aligned} \tag{3.46}$$

where $\bar{f}(\cdot) := \alpha(t) f_+\left(x(t),u(t)\right) + (1-\alpha(t)) f_-(x(t),u(t))$.
The velocity field in (3.46)

$$U(x) := \{v \in \mathbb{R}^{n_x} : v = \bar{f}(x,u,\alpha), \alpha \in [0,1], u \in \mathcal{U}\} \tag{3.47}$$

may be nonconvex, see Subsection 3.1.3. In order to to get an idea how to exploit Filippov's rule correctly, let use consider an instance as follows.

Example 3.2. Consider the following problem

$$\begin{aligned} \min_{x,u} \quad & \varphi\left(x(t_f)\right) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} b_+u(t), & \text{if } \sigma(x(t)) > 0, \\ b_-u(t), & \text{if } \sigma(x(t)) < 0, \end{cases} \qquad t \in \mathcal{T} := [t_0,t_f]. \\ & u(t) \in \mathcal{U} = [u_l,u_u],\ x(t_0) = x_0,\ r(x(t_f)) \ge 0, \end{aligned} \tag{3.48}$$

Let $u_j \in \text{conv}(\mathcal{U})$, $j = 1, 2$, we have

$$\begin{aligned}\alpha b_+ u_1 + (1-\alpha) b_- u_2 &= \alpha b_+(\gamma u_l + (1-\gamma)u_u) + (1-\alpha) b_-(\gamma u_l + (1-\gamma)u_u) \\ &= \gamma_1 b_+ u_l + \gamma_2 b_+ u_u + \gamma_3 b_- u_l + \gamma_4 b_- u_u,\end{aligned}$$

where $\alpha \in [0,1]$, $\gamma \in [0,1]$, and $\gamma_1 := \alpha\gamma$, $\gamma_2 := \alpha(1-\gamma)$, $\gamma_3 := (1-\alpha)\gamma$, $\gamma_4 := (1-\alpha)(1-\gamma)$. Thus $(\alpha b_+ u_1 + (1-\alpha) b_- u_2) \in \text{conv}(\mathcal{U})$ since $\sum_{j=1}^4 \gamma_j = 1$, $\gamma_j \geq 0$.
The velocity field $U(x) = \{\alpha b_+ u + (1-\alpha) b_- \mid \alpha \in [0,1], u \in [u_l, u_u]\}$ may be non-convex, see also Fig. 3.2, resulting that problem (3.48) does not have solution (see Appendix A.1 for an example). Hence, we relax the ODE in problem (3.48) as the following formulation

$$\dot{x}(t) = \begin{cases} b_+(\gamma_1 u_l + (1-\gamma_1)u_u), & \text{if } \sigma(x(t)) > 0, \\ b_-(\gamma_2 u_l + (1-\gamma_2)u_u), & \text{if } \sigma(x(t)) < 0, \end{cases} \quad t \in \mathcal{T}, \tag{3.49}$$

where $\gamma_j \in [0,1]$, $j = 1, 2$, are new controls, and apply Filippov's rule to this ODE leads to

$$\dot{x} = \alpha\gamma_1 b_+ u_l + \alpha(1-\gamma_1) b_+ u_u + (1-\alpha)\gamma_2 b_- u_l + (1-\alpha)(1-\gamma_2) b_- u_u.$$

That is why, instead of (3.47), we use further relaxation, i.e., we replace $U(x)$ by its convex hull

$$\begin{aligned}\tilde{U}(x) := \{v \in \mathbb{R}^{n_x} : v = \sum_{j=1}^{\tilde{n}} (\alpha\gamma_j^+ f_+(x,u_j) + (1-\alpha)\gamma_j^- f_-(x,u_j)),\ \gamma_j^+ \geq 0, \\ \sum_{j=1}^{\tilde{n}} \gamma_j^+ = 1, \gamma_j^- \geq 0, \sum_{j=1}^{\tilde{n}} \gamma_j^- = 1, \alpha \in [0,1], u_j \in \mathcal{U}\}.\end{aligned} \tag{3.50}$$

We here assume that we know such presentation $u_j/\tilde{n}$, $\tilde{n} < \infty$, however it is difficult to construct it. As a result, problem (3.46) is rewritten in the formulation as follows

$$\begin{aligned}\min_{x(\cdot),u(\cdot),\alpha(\cdot),\gamma(\cdot)} \quad & m(x(t_f)) \\ \text{s.t.} \quad & \dot{x}(t) = F(x(t), u(t), \alpha(t), \gamma(t)), \\ & 0 \leq r\left(x(t_f)\right), \\ & 0 \leq \alpha(t)\sigma(x(t)),\ (1-\alpha(t))\sigma(x(t)) \leq 0, \qquad t \in \mathcal{T}. \\ & u_j(t) \in \mathcal{U},\ \ j = 1, 2, \ldots, \tilde{n}, \\ & \textstyle\sum_{j=1}^{\tilde{n}} \gamma_j(t)^+ = 1,\ \gamma_j^+(t) \geq 0,\ j = 1, 2, \ldots, \tilde{n}, \\ & \textstyle\sum_{j=1}^{\tilde{n}} \gamma_j^-(t) = 1,\ \gamma_j^-(t) \geq 0,\ j = 1, 2, \ldots, \tilde{n},\end{aligned} \tag{3.51}$$

where $F(\cdot) := \sum_{j=1}^{\tilde{n}} (\alpha(t)\gamma_j^+(t) f_+(x(t), u_j(t)) + (1-\alpha(t))\gamma_j^-(t) f_-(x(t), u_j(t)))$.

To state the optimality condition for problem (3.51), we consider the following:
Denote $\tilde{u}_j^+ := \alpha\gamma_j^+ \in \mathbb{R}$, $\tilde{u}_j^- := (1-\alpha)\gamma_j^- \in \mathbb{R}$, $j = 1, \ldots, \tilde{n}$. The velocity field of ODE in problem (3.51) is convex. Analyze the maximum condition (3.13), which are now equivalent

to the following LP problem

$$\begin{aligned}
&\max_{\alpha,\tilde{u}_j^+,\tilde{u}_j^-}\{\sum_j^{\tilde{n}}\tilde{u}_j^+a_j^+ + \sum_j^{\tilde{n}}\tilde{u}_j^-a_j^-\}\\
&\text{s.t. } 0\le\alpha\le 1,\ 0\le\tilde{u}_j^+\le\alpha,\ 0\le\tilde{u}_j^-\le 1-\alpha,\ j=1,\dots,\tilde{n},\\
&\quad\sum_{j=1}^{\tilde{n}}\tilde{u}_j^+=\alpha,\ \sum_{j=1}^{\tilde{n}}\tilde{u}_j^-=1-\alpha,
\end{aligned}\tag{3.52}$$

where $a_j^+ := \lambda^T(t)f_+(x,u_j)$, $a_j^- := \lambda^T(t)f_-(x,u_j)$, $j=1,\dots,\tilde{n}$.
For the case $\alpha=1$, the maximum condition (3.52) implies $\tilde{u}_j^-=0$, $j=1,\dots,\tilde{n}$, and

$$\tilde{u}_{k_1}^+=\alpha,\ \tilde{u}_j^+=0,\ j\neq k_1,\ j,k_1\in\{1,\dots,\tilde{n}\},\ \text{where } k_1:=\arg\max_j\{\lambda^T(t)f_+(x,u_j)\}.\tag{3.53}$$

For $\alpha=0$, from the maximum condition (3.52) one gets $\tilde{u}_j^+=0$, $j=1,\dots,\tilde{n}$, and

$$\tilde{u}_{k_2}^-=1-\alpha,\ \tilde{u}_j^-=0,\ j\neq k_2,\ j,k_2\in\{1,\dots,\tilde{n}\},\text{where } k_2:=\arg\max_j\{\lambda^T(t)f_-(x,u_j)\}.\tag{3.54}$$

For the remain case $0<\alpha<1$, the maximum condition (3.52) implies

$$\begin{aligned}
&\tilde{u}_{k_1}^+=\alpha,\ \tilde{u}_j^+=0,\ j\neq k_1,\quad j,k_1\in\{1,\dots,\tilde{n}\},\\
&\tilde{u}_{k_2}^-=1-\alpha,\ \tilde{u}_j^-=0,\ j\neq k_2,\quad j,k_2\in\{1,\dots,\tilde{n}\},
\end{aligned}\tag{3.55}$$

where k_1 and k_2 are given in (3.53) and (3.54), respectively.

Remark 18. From the optimality conditions of problem (3.51), we get the following rule component for control γ:

$$\begin{aligned}
&\gamma_{k_1}^+=1,\ \gamma_j^+=0,\ j\neq k_1,\ j,k_1\in\{1,\dots,\tilde{n}\}, && \text{if } 0<\alpha(t)\le 1,\\
&\gamma_j^+=0,\ j\in\{1,\dots,\tilde{n}\}, && \text{if } \alpha(t)=0,
\end{aligned}\tag{3.56}$$

$$\begin{aligned}
&\gamma_{k_2}^-=1,\ \gamma_j^-=0,\ j\neq k_2,\ j,k_2\in\{1,\dots,\tilde{n}\}, && \text{if } 0\le\alpha(t)<1,\\
&\gamma_j^-=0,\ j\in\{1,\dots,\tilde{n}\}, && \text{if } \alpha(t)=1,
\end{aligned}\tag{3.57}$$

where k_1, k_2 are respectively determined by (3.53), (3.54).

Lemma 5. *Let $(\{u_j^*\}_{j=1}^{\tilde{n}},\alpha^*,\gamma^*)$ be a control solution of problem* (3.51)*, then the corresponding control of problem* (3.46) *is determined by (u^0,α^*), where $u^0=\sum_{j=1}^{\tilde{n}}({\gamma_j^+}^*+{\gamma_j^-}^*)u_j^*$.*

Proof. Similar to the proof of Proposition 1. □

Lemma 6. *For any absolutely continuous solution $x^*(t)$, $t\in\mathcal{T}$, of problem* (3.46)*, there exists a sequence of absolutely continuous solutions $x^k(t)$, $t\in\mathcal{T}$, of problem* (3.51)*, $k=1,2,\dots$, such that $\lim_{k\to\infty}\max_{t\in\mathcal{T}}\|x^*(t)-x^k(t)\|=0$.*
The optimal trajectory $x^0(t)$, $t\in\mathcal{T}$, of the original problem (3.44) *satisfies ODE of problem* (3.51)*, and ones have,*

$$\lim_{k\to\infty}\max_{t\in\mathcal{T}}\|x^0(t)-x^k(t)\|=0.$$

Proof. The proof is done by exploiting the construction in [82, Section 3]. □

Remark 19. $\tilde{U}(x)$ is generally have no closed form, even numerically very difficult to compute. However if control enters linearly in right hand side, and $\mathcal{U}$ is convex polytope, then $\tilde{U}(x)$ can be described explicitly. Assume

$$f_+(x,u) = A_+(x) + b_+(x)u, \quad f_-(x,u) = A_+(x) + b_-(x)u, \tag{3.58}$$

and $\mathcal{U} = [u_l, u_u]^{n_u}$. Then u_j, $j = 1, \ldots, \tilde{n}$, are vertices of $\mathcal{U}$, and

$$\begin{aligned}
\operatorname{conv}(\alpha f_+ + (1-\alpha) f_-) &= \sum_{j=1}^{\tilde{n}} (\alpha \gamma_j^+ f_+(x,u_j) + (1-\alpha)\gamma_j^- f_-(x,u_j)) \\
&= \sum_{j=1}^{\tilde{n}} (\alpha \gamma_j^+ (A_+(x) + b_+(x)u_j) + (1-\alpha)\gamma_j^- (A_-(x) + b_-(x)u_j)) \\
&= \alpha(A_+(x) - A_-(x)) + (b_+(x)a_+ + b_-(x)a_-)\,,
\end{aligned}$$

with controls $\alpha \in [0,1]$, $a_+ := \alpha \sum_{j=1}^{\tilde{n}} \gamma_j^+ u_j \in \alpha\mathcal{U}$, and $a_- := (1-\alpha)\sum_{j=1}^{\tilde{n}} \gamma_j^- u_j \in (1-\alpha)\mathcal{U}$. Furthermore, from the optimality conditions for problem (3.52), one gets

$$\lambda^T(A_+ - A_-) + \max_{u\in\mathcal{U}}\{\lambda^T b_+ u\} - \max_{u\in\mathcal{U}}\{\lambda^T b_- u\} \begin{cases} \geq 0, & \text{if } \alpha = 1, \\ \leq 0, & \text{if } \alpha = 0, \\ = 0, & \text{if } \alpha \in (0,1). \end{cases} \tag{3.59}$$

3.2 Optimality conditions for SwOCP

In this section, we propose a solution approach for SwOCP (with integer controls) exploiting Filippov's rule reformulation, relaxation and LMP.

3.2.1 Reformulation

Consider a general SwOCP as follows

$$\begin{aligned}
\min_{x(\cdot),u(\cdot),w(\cdot)} \quad & \varphi[x(\cdot),u(\cdot),w(\cdot)] \\
\text{s.t.} \quad & \dot{x}(t) = f\left(x(t),u(t),w(t),\operatorname{sgn}\left(\sigma(x(t))\right)\right), \\
& 0 \leq r\left(x(t_0),x(t_f)\right), \\
& w(t) \in \mathcal{W} \stackrel{\text{def}}{=} \{w_1,\ldots,w_{n_w}\}, t \in \mathcal{T} \stackrel{\text{def}}{=} [t_0,t_f]\,.
\end{aligned} \tag{3.60}$$

therein, the dynamic process $x : \mathcal{T} \to \mathcal{X}$ is determined by a dynamical system with right hand side function $f : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \times \{-1,0,1\} \to \mathbb{R}^{n_x}$, affected by continuously-valued bounded control function $u : \mathcal{T} \to \mathcal{U}$ (which is assumed entering linearly into the right hand side of the ODE of (3.60)) and a discretely-valued control function $w : \mathcal{T} \to \mathcal{W}$ on a time horizon $\mathcal{T} \subset \mathbb{R}$, such that an objective function $\varphi = m(x(t_f)) + \int_{t_0}^{t_f} l(x(t),u(t))dt$, is minimized. The right hand side function f is determined by the sign structure of the scalar switching

function $\sigma : \mathcal{X} \to \mathbb{R}$. Point constraints $r : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n_r}$ must be satisfied.
Since the definition of the sign of $\sigma(\cdot)$, we can rewrite (3.60) as follows:

$$\begin{aligned} \min_{x(\cdot),u(\cdot),w(\cdot)} \quad & m(x(t_f)) + \textstyle\int_{t_0}^{t_f} l(x(t),u(t))dt \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} f_+\left(x(t),u(t),w(t)\right), & \text{if } \sigma(x(t)) \geq 0, \\ f_-\left(x(t),u(t),w(t)\right), & \text{if } \sigma(x(t)) \leq 0, \end{cases} \quad t \in \mathcal{T}. \\ & 0 \leq r\left(x(t_0), x(t_f)\right), \\ & w(t) \in \mathcal{W}, \end{aligned} \tag{3.61}$$

Remind that in the case $\sigma(\hat{t}, x(\hat{t})) = 0$, $\hat{t} \in \mathcal{T}$, if $\hat{t}$ is an isolated zero of $\sigma(x(t))$, right-hand side is chosen to be either f_+ or f_-, and if $\sigma(x(t))$ vanishes on an interval, the right-hand side has yet to be determined appropriately.
By exploiting the POC, the ODE of problem (3.61) is rewritten as follows

$$\dot{x}(t) = \begin{cases} \sum_{w\in\mathcal{W}} \beta_w(t) f_+\left(x(t),u(t),w\right), & \text{if } \sigma(x(t)) \geq 0, \\ \sum_{w\in\mathcal{W}} \beta_w(t) f_-\left(x(t),u(t),w\right), & \text{if } \sigma(x(t)) \leq 0, \end{cases}, \quad t \in \mathcal{T}, \tag{3.62}$$

where $\sum_{w\in\mathcal{W}} \beta_w(t) = 1$, $\beta_w(t) \in \{0,1\}$.
We exploit further relaxations for (3.62), i.e. we use Filippov's rule correctly, where the integer controls β_w, $w \in \mathcal{W}$, are treated by the following relaxed formulation

$$\dot{x}(t) = \sum_{w\in\mathcal{W}} \alpha_w(t) \sum_{j=1}^{\tilde{n}} (\alpha_\sigma(t)\alpha_j^+(t) f_+\left(x(t),u_j(t),w\right) + (1-\alpha_\sigma(t))\alpha_j^-(t) f_-\left(x(t),u_j(t),w\right)),$$

$$\sum_{j=1}^{\tilde{n}} \alpha_j^+(t) = 1,\ \sum_{j=1}^{\tilde{n}} \alpha_j^-(t) = 1,\ \alpha_j^+(t) \geq 0,\ \alpha_j^-(t) \geq 0,\ j = 1,2,\ldots,\tilde{n},$$

$$\sum_{w\in\mathcal{W}} \alpha_w(t) = 1,\ \alpha_w : \mathcal{T} \to [0,1],\ w \in \mathcal{W}, \quad u_j(t) \in \mathcal{U},\ j = 1,2,\ldots,\tilde{n},$$

where $\tilde{n} < \infty$. Then, SwOCP (3.61) is reformulated as follows

$$\begin{aligned} \min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & m\left(x(t_f)\right) + \textstyle\int_{t_0}^{t_f} l(x(t),u(t))dt \\ \text{s.t.} \quad & \dot{x}(t) = F(x(t),u(t),\alpha(t)), \\ & 0 \leq r\left(x(t_0), x(t_f)\right), \\ & 0 \leq \alpha_\sigma(t)\sigma(x(t)),\ (1-\alpha_\sigma(t))\sigma(x(t)) \leq 0, \\ & 0 \leq \alpha_\sigma(t) \leq 1,\ \textstyle\sum_{w\in\mathcal{W}} \alpha_w(t) = 1,\ 0 \leq \alpha_w(t) \leq 1,\ w \in \mathcal{W}, \\ & \textstyle\sum_{j=1}^{\tilde{n}} \alpha_j^l(t) = 1,\ \alpha_j^l(t) \geq 0,\ l = + \vee -,\ j = 1,2,\ldots,\tilde{n}, \end{aligned} \quad t \in \mathcal{T}, \tag{3.63}$$

where $u_j(t) \in \mathcal{U}$, $j = 1,2,\ldots,\tilde{n}$, and

$$F(\cdot) := \sum_{w\in\mathcal{W}} \alpha_w(t) \sum_{j=1}^{\tilde{n}} (\alpha_\sigma(t)\alpha_j^+(t) f_+(x(t),u_j(t),w) + (1-\alpha_\sigma(t))\alpha_j^-(t) f_-(x(t),u_j(t),w)). \tag{3.64}$$

Remark 20.
- Note that the convexity of velocity field of (3.63) is guaranteed similarly to the previous sections, see Subsection 3.1.3 and Subsection 3.1.4.
- Here we distinguish the simple case with a scalar switching function σ. Ones can generate the general cases by re-noting $\sigma_1 =: \sigma$, so $\sigma_j = 0$ for $j = 2, \dots, n$, where n is the number of switching functions. Readers can find more details about this issue in Chapter 5 (see Section 5.1) and Chapter 6 (see Section 6.2 and Section 6.3).

3.2.2 Local Maximum Principle

Denote

$$\bar{\mathcal{V}}(x) := \{\alpha \in [0,1]^{2\tilde{n}+n_w+1} : \alpha_\sigma \sigma \geq 0, (1-\alpha_\sigma)\sigma \leq 0, \sum_{w\in\mathcal{W}} \alpha_w = 1, \sum_{j=1}^{\tilde{n}} \alpha_j^l = 1\}, \quad (3.65)$$

$$\bar{V}(x) := \left\{(u,\alpha) : u \in \mathcal{U},\ \alpha \in \bar{\mathcal{V}}(x)\right\}, \quad (3.66)$$

where $l = + \vee -$, and $\tilde{n} < \infty$. Analyzing similarly as in Subsection 3.1.2, ones can write down the LMP for problem (3.63) as follows.

Theorem 13 (Optimality conditions for SwOCP (3.63))**.** *Let $u^*(t), \alpha^*(t)$ and $x^*(t)$, $t \in \mathcal{T}$ be the optimal control and trajectory of problem (3.63), where assuming control $u^*(t)$ is piecewise continuous and F is given by* (3.64)*. Then there exists a vector δ, vectors v_{t_f}, v_{t_0}, a number v_0, a function $\vartheta \in L_1$, and a measure μ such that normalization conditions*

$$v_0 + \|v_{t_0}\| + \left\|v_{t_f}\right\| + \|\delta\| + \|\mu\| + \int_{\mathcal{T}} \vartheta(t)dt = 1,$$

are fulfilled, and the adjoint system has the form

$$\frac{d\lambda(t)}{dt} \overset{dt}{=} \begin{cases} -\lambda^T(t) \sum\limits_{w\in\mathcal{W}} \alpha_w^* \sum\limits_{j=1}^{\tilde{n}} \alpha_j^{+*} \frac{\partial f_+(x^*,u_j^*,w^*)}{\partial x} + \delta^T \sum\limits_{j=1}^{\tilde{n}} \alpha_j^{+*} \frac{\partial l(x^*,u_j^*)}{\partial x}, & t \in T_+, \\ -\lambda^T(t) \sum\limits_{w\in\mathcal{W}} \alpha_w^* \sum\limits_{j=1}^{\tilde{n}} \alpha_j^{-*} \frac{\partial f_-(x^*,u_j^*,w^*)}{\partial x} + \delta^T \sum\limits_{j=1}^{\tilde{n}} \alpha_j^{-*} \frac{\partial l(x^*,u_j^*)}{\partial x}, & t \in T_-, \\ -\lambda^T(t) \frac{\partial F(x^*,u^*,\alpha^*)}{\partial x} + \delta^T \frac{\partial l(x^*,u^*)}{\partial x} + \left(\vartheta(t) \pm \frac{d\mu_a}{dt}\right) \frac{\partial \sigma(x^*(t))}{\partial x} dt, & t \in T_*, \end{cases}$$

$$\frac{d\lambda}{d\mu_s} \overset{d\mu_s}{=} \pm \frac{\partial \sigma(x^*(t))}{\partial x}, \quad t \in T_*, \quad (3.67)$$

$$\lambda(t+0) = \lambda(t-0) + (\mu(t-0) + \mu(t+0)) \frac{\partial \sigma(x^*(t))}{\partial x}, \quad t \in \mathcal{D} \cap T_*,$$

where $T_+ := \{t \in \mathcal{T} \mid \sigma(x^*(t)) > 0\}$, $T_- := \{t \in \mathcal{T} \mid \sigma(x^*(t)) < 0\}$, $T_* := \{t \in \mathcal{T} \mid \sigma(x^*(t)) = 0\}$, $\vartheta(t)$ is a Lebesgue absolutely integrable function, $\mathcal{D}$ is the set of jumps of measure μ_δ, $\overset{dt}{=}$ means equality holds a.e. in the measure dt, and

$$\lambda(t_f - 0) = -v_{t_f}^T \frac{\partial r(\cdot, x^*(t_f))}{\partial x} - v_0 \frac{\partial m(x^*(t_f))}{\partial x} - \mu(t_f) \frac{\partial \sigma(x^*(t_f))}{\partial x},$$
$$\lambda(t_0 + 0) = -v_{t_0}^T \frac{\partial r(x^*(t_0), \cdot)}{\partial x} - \mu(t_0) \frac{\partial \sigma(x^*(t_0))}{\partial x},$$

and the optimal conditions take the following forms

$$\vartheta(t) \geq 0, \ \textit{if } \alpha_\sigma^*(t) = 0, \quad \vartheta(t) \leq 0, \ \textit{if } \alpha_\sigma^*(t) = 1, \quad t \in T_*, \ j = 1, \dots, \tilde{n}, \tag{3.68}$$

$$\mu(t) \geq 0, \ \textit{if } \alpha_\sigma^*(t) = 0, \quad \mu(t) \leq 0, \ \textit{if } \alpha_\sigma^*(t) = 1, \quad t \in T_*, \ j = 1, \dots, \tilde{n}, \tag{3.69}$$

$$\lambda^T(t) \sum_{w \in \mathcal{W}} \alpha_w^* \sum_{j=1}^{\tilde{n}} (\alpha_\sigma^* \alpha_j^{+*} f_+ \left(x^*, u_j^*(t), w^*\right) + (1 - \alpha_\sigma^*) \alpha_j^{-*} f_- \left(x^*, u_j^*(t), w^*\right)) \tag{3.70}$$

$$= \max_{(u,\alpha) \in \bar{V}} \{\lambda^T(t) \sum_{w \in \mathcal{W}} \alpha_w \sum_{j=1}^{\tilde{n}} (\alpha_\sigma \alpha_j^+ f_+ \left(x^*, u_j, w^*\right) + (1 - \alpha_\sigma) \alpha_j^- f_- \left(x^*, u_j, w^*\right))\}, t \in \mathcal{T},$$

$$\begin{aligned}
&\alpha^{+*}_{k_1} = 1, \ \alpha^{+*}_j = 0, \ j \neq k_1, \ j, k_1 \in \{1, \dots, \tilde{n}\}, && \textit{if } 0 < \alpha_\sigma(t) \leq 1, \\
&\alpha^{+*}_j = 0, \ j \in \{1, \dots, \tilde{n}\}, && \textit{if } \alpha_\sigma(t) = 0, \\
&\alpha^{-*}_{k_2} = 1, \ \alpha^{-*} = 0, \ j \neq k_2, \ j, k_2 \in \{1, \dots, \tilde{n}\}, && \textit{if } 0 \leq \alpha_\sigma(t) < 1, \\
&\alpha^{-*}_j = 0, \ j \in \{1, \dots, \tilde{n}\}, && \textit{if } \alpha_\sigma(t) = 1,
\end{aligned} \tag{3.71}$$

$$\lambda^T(t) \sum_{j=1}^{\tilde{n}} (\alpha_j^{+*}(t) f_+(*j) - \alpha_j^{-*}(t) f_-(*j)) \begin{cases} = 0, & \textit{if } 0 < \alpha_\sigma^*(t) < 1, \\ \leq 0, & \textit{if } \alpha_\sigma^*(t) = 0, \\ \geq 0, & \textit{if } \alpha_\sigma^*(t) = 1, \end{cases} \quad t \in T_*, \tag{3.72}$$

where $k_1 = \arg\max_j \{\lambda^T(t) f_+(x, u_j, w)\}$, $k_2 = \arg\max_j \{\lambda^T(t) f_-(x, u_j, w)\}$, *and* $f_+(*j) := f_+(x^*(t), u_j^*(t), w^*)$, $f_-(*j) := f_-(x^*(t), u_j^*(t), w^*)$.
Furthermore, ones have

$$\begin{aligned}
&\lambda^T(\tau_i - 0) F(x^*(\tau_i - 0), u^*(\tau_i - 0), \alpha^*(\tau_i - 0)) \\
&= \lambda^T(\tau_i + 0) F(x^*(\tau_i + 0), u^*(\tau_i + 0), \alpha^*(\tau_i + 0)), \ i = 1, \dots, p,
\end{aligned}$$

where τ_i, $i = 1, \dots, p$, *be the minimum number of points satisfy condition* (3.22).

Remark 21. Note that we can rewrite the right hand side of the optimality condition (3.70) as follows

$$\max_{\alpha_w} \{ \sum_{w \in \mathcal{W}} \alpha_w (\max_{\alpha_\sigma, \alpha_j^+, \alpha_j^-} \lambda^T(t) \sum_{j=1}^{\tilde{n}} (\alpha_\sigma \alpha_j^+ f_+ \left(x^*, u_j, w^*\right) + (1 - \alpha_\sigma) \alpha_j^- f_- \left(x^*, u_j, w^*\right)))\}$$

which returns the maximum value of the augmented Hamiltonian $\tilde{H}$ for each w. This is similar to the idea of Competing Hamiltonian Approach, see Appendix A.3.

3.3 Numerical Examples with LMP for Subway Problem

We consider a problem about Subway Optimization which goes back to work of [17, 19–21] for the city of New York. The aim is to minimize the energy used for a subway ride from one station to another, taking into account boundary conditions and a restriction on the time. The following contribution is based on our conference paper [121].
One can formulate this problem as constrained optimal control problem of the following form

$$\min_{x,w,T} \int_0^T L(x(t), w(t))dt \tag{3.73}$$

subject to an ODE system

$$\dot{x}(t) = f(x(t), w(t)), \quad t \in [t_0, T] \tag{3.74}$$

path constraints

$$0 \leq g(x(t)), \quad t \in [t_0, T] \tag{3.75}$$

interior point constraints

$$0 \leq r^{\text{ieq}}(x(t_0), x(t_1), \ldots, x(T), T),\ t_i \in [t_0, T], \quad r^{\text{eq}}(x(t_0), x(t_1), \ldots, x(T), T) = 0, \tag{3.76}$$

and binary admissibility of $w(\cdot)$

$$w(t) \in \{1, 2, 3, 4\}. \tag{3.77}$$

The terminal time T denotes the time of arrival of a subway train in the next station. The states $x_0(\cdot)$ and $x_1(\cdot)$ describe distance from starting point and velocity of the train, respectively. The train can be operated in one of four different models

$$w(t) = \begin{cases} 1 & \text{Series} \\ 2 & \text{Parallel} \\ 3 & \text{Coasting} \\ 4 & \text{Braking.} \end{cases} \tag{3.78}$$

that influences the acceleration and the deceleration of the train and therewith the energy consumption, which is to be minimized and given by the Lagrange term

$$L(x(t), 1) = \begin{cases} ep_1 & \text{for } x_1(t) \leq v_1 \\ ep_2 & \text{for } v_1 < x_1(t) \leq v_2 \\ e \sum_{i=0}^{5} c_i(1) \left(\frac{1}{10}\gamma\, x_1(t)\right)^{-i} & \text{for } x_1(t) > v_2 \end{cases} \tag{3.79a}$$

$$L(x(t), 2) = \begin{cases} 0 & \text{for } x_1(t) \leq v_2 \\ ep_2 & \text{for } v_2 < x_1(t) \leq v_3 \\ e \sum_{i=0}^{5} c_i(2) \left(\frac{1}{10}\gamma\, x_1(t)\right)^{-i} & \text{for } x_1(t) > v_3 \end{cases} \tag{3.79b}$$

$$L(x(t), 3) = 0, \tag{3.79c}$$

$$L(x(t), 4) = 0. \tag{3.79d}$$

In the considered problem, the right hand side function $f(\cdot)$ is dependent on the model $w(\cdot)$ and on the state variable velocity $x_1(\cdot)$, but not on distance. For all $t \in [0, T]$, we have

$$\dot{x}_0(t) = x_1(t) \tag{3.80}$$

For operation in series mode, $w(t) = 1$, we have

$$\dot{x}_1(t) = f_1(x,1) = \begin{cases} f_1^{1A}(x) & \text{for } x_1(t) \le v_1 \\ f_1^{1B}(x) & \text{for } v_1 < x_1(t) \le v_2 \\ f_1^{1C}(x) & \text{for } x_1(t) > v_2 \end{cases} \tag{3.81}$$

where $f_1^{1A} = \frac{gea_1}{W_{eff}}$, $f_1^{1B} = \frac{gea_2}{W_{eff}}$, $f_1^{1C} = \frac{g(eT(x_1(t),1)-R(x_1(t)))}{W_{eff}}$.
By using Filippov's rule, we can rewrite (3.81) as follows

$$\dot{x}_1(t) = \sum_j \beta_{1j}(t) f_1^{1j}\left(x(t)\right), \ \sum_j \beta_{2j}(t) = 1, \beta_{2j}(t) \in [0,1], j = A, B, C, \tag{3.82}$$

with the additional mixed state-control constraints

$$\beta_{1A}\mathcal{G}_{1A} \le 0, \ \ \beta_{1B}\mathcal{G}_{1A} \ge -\varepsilon, \ \beta_{1B}\mathcal{G}_{1B} \le 0, \ \ \beta_{1C}\mathcal{G}_{1B} \ge -\varepsilon, \tag{3.83}$$

where $\mathcal{G}_{1A}(t, x(t)) := x_1(t) - v_1$, $\mathcal{G}_{1B}(t, x(t)) := x_1(t) - v_2$, and $\varepsilon > 0$ sufficient small, which is needed to ensure the regularity of the mixed constraints.
For operation in parallel, $w(t) = 2$, we have

$$\dot{x}_1(t) = f_1(x,2) = \begin{cases} f_1^{2A}(x) & \text{for } x_1(t) \le v_2 \\ f_1^{2B}(x) & \text{for } v_2 < x_1(t) \le v_3 \\ f_1^{2C}(x) & \text{for } x_1(t) > v_3 \end{cases} \tag{3.84}$$

where $f_1^{2A} = 0$, $f_1^{2B} = \frac{gea_3}{W_{eff}}$, $f_1^{2C} = \frac{g(eT(x_1(t),2)-R(x_1(t)))}{W_{eff}}$.
By using Filippov's rule, and since $f_1^{2A} = 0$, we can similarly rewrite (3.84) as follows

$$\dot{x}_1(t) = \sum_j \beta_{2j}(t) f_1^{2j}\left(x(t)\right), \ \sum_j \beta_{2j}(t) = 1, \ \beta_{2j}(t) \in [0,1], \ j = B, C, \tag{3.85}$$

with the additional mixed constraints

$$\beta_{2B}\mathcal{G}_{2B} \ge -\varepsilon, \quad \beta_{2B}\mathcal{G}_{1B} \le 0, \quad \beta_{2C}\mathcal{G}_{2B} \ge -\varepsilon, \tag{3.86}$$

where $\mathcal{G}_{2B}(t, x(t)) := x_1(t) - v_3$, and $\varepsilon > 0$ sufficient small.
For coasting, $w(t) = 3$, we have

$$\dot{x}_1(t) = f_1(x,3) = -\frac{gR(x_1(t))}{W_{eff}} - C,$$

and for braking, $w(t) = 4$, we have

$$\dot{x}_1(t) = f_1(x,4) = -u(t), \quad u(t) \in [u_{\min}, u_{\max}].$$

The braking deceleration $u(t)$ can be varied between some given natural force $u_{\min}$ as present in coasting and a given limit $u_{\max}$ representing a maximum braking consistent with passenger comfort. It can be shown easily that for the problem at hand only maximal braking can be

optimal, thus, without loss of generality, we fix $u(t)$ to $u_{\max}$.
The occurring forces are given by

$$R(x_1(t)) = c(n_{wag})a\gamma^2 x_1(t)^2 + \frac{bW}{2000}\gamma x_1(t) + \frac{1.3}{2000}W + 116,$$
$$T(x_1(t),1) = \sum_{i=0}^{5} b_i(1)\left(\frac{1}{10}\gamma x_1(t) - 0.3\right)^{-i},$$
$$T(x_1(t),2) = \sum_{i=0}^{5} b_i(2)\left(\frac{1}{10}\gamma x_1(t) - 1\right)^{-i}.$$

Path constraints for subway trains typically are velocity limits

$$x_1(t) \le v_{\max}, \quad t \in [t_{ca}, t_{ce}]$$

where the time interval may be implicitly characterized to cover a certain section of the track. The interior point equality constraints $r^{\text{eq}}(\cdot)$ are the initial and terminal constraints on the state trajectory and constraints to characterize intermediate stops at stations of a line

$$x(0) = (0,0)^T, \quad x(t_i) = (S_i, 0)^T \; t_i \in [0,T], \quad x(T) = (S,0).$$

The interior point inequality constraints $r^{\text{ieq}}(\cdot)$ include a maximal driving time $T^{\max}$ to get from $x(0) = (0,0)^T$ to $x(T) = (S,0)^T$,

$$T \le T^{\max}. \tag{3.87}$$

In the equations above the parameters e (percentage of working motors – a peculiarity of the New York subway), p_1, p_2, p_3, $b_i(w), c_i(w)$, γ, g, a_1, a_2, a_3, W_{eff}, C, c, n_{wag}, b, W, $u_{\max}, T^{\max}$, v_1, v_2, and v_3 are fixed. Values for these parameters are given in the appendix of [111, Appendix C] and in the following description of the applications treated in this section.

The previous approach, see [17, 19–21] and [111] was used to treat several station-to-station rides for different station spacings, weight, travel time, etc. We use their ideas, but with different formula of the solution approach. Here we show results for a subway problem with 10 wagons ($n_{wag} = 10$), a medium loaded train ($W = 78000$ lbs), for a local run ($S = 2112$ ft), a transit time $T^{\max} = 65$ s that is about 20% longer than the fastest possible and with all engines working ($e = 1.0$).

By reformulating the problem with Filippov's rule, we obtain the equivalent one

$$\begin{aligned} &\min_{x,w,\alpha} \quad \int_0^T L(x(t),w)dt \\ &\text{s.t. } \dot{x}_0(t) = x_1(t), \quad \dot{x}_1(t) = \sum_{i=1}^{4} \alpha_i(t) f_1(x(t),i), \quad t \in [0,T], \\ &\qquad \textstyle\sum_{i=1}^{4} \alpha_i(t) = 1, \; \alpha(t) \in [0,1]^4, \; t \in [0,T], \quad x(0) = (0,0)^T, x(T) = (S,0)^T, \end{aligned} \tag{3.88}$$

where $w \in \mathcal{W} = \{1,2,3,4\}$, $S = 2112$ ft, and $T \le T^{\max} = 65$, and

$$L(x(t),w) := \alpha_1(t)L(x(t),1) + \alpha_2(t)L(x(t),2), \tag{3.89}$$

where $L(x(t),1) = \beta_{1A}ep_1 + \beta_{1B}ep_2 + \beta_{1C}e\sum_{i=0}^{5} c_i(1)(\frac{1}{10}\gamma x_1(t))^{-i}$, $L(x(t),2) = \beta_{2B}ep_2 + \beta_{2C}e\sum_{i=0}^{5} c_i(1)(\frac{1}{10}\gamma x_1(t))^{-i}$.

Exploiting the reformulation of $f_1(x,w)$ from (3.82) to (3.85), we finally can rewrite (3.88) in more details, which includes mixed state-control constraints, as follows

$$\begin{aligned}
&\min_{x,w,\alpha,\beta} \quad \int_0^T L(x(t),w)dt \\
&\text{s.t. } \dot{x}_0(t) = x_1(t), \\
&\quad \dot{x}_1(t) = \alpha_1(t)\textstyle\sum_j \beta_{1j}(t) f_1^{1j}(x(t),u_{\max},1) - \alpha_3(t)\left(\frac{gR(x_1(t))}{W_{eff}} + C\right) \\
&\qquad +\alpha_2(t)\textstyle\sum_k \beta_{2k}(t) f_1^{2k}(x(t),u_{\max},2) - \alpha_4(t)u_{\max}, \\
&\quad \beta_{1A}\mathcal{G}_{1A} \le 0,\ \beta_{1B}\mathcal{G}_{1A} \ge -\varepsilon,\ \beta_{1B}\mathcal{G}_{1B} \le 0,\ \beta_{1C}\mathcal{G}_{1B} \ge -\varepsilon, \\
&\quad \beta_{2B}\mathcal{G}_{2B} \ge -\varepsilon,\ \beta_{2B}\mathcal{G}_{1B} \le 0,\ \beta_{2C}\mathcal{G}_{2B} \ge -\varepsilon, \\
&\quad x(0) = (0,0)^T,\ x(T) = (S,0)^T, \\
&\quad \alpha \in [0,1]^4,\ \textstyle\sum_{i=1}^4 \alpha_i(t) = 1,\ \forall t \in [0,T], \\
&\quad \textstyle\sum_j \beta_{1j}(t) = 1, \quad \beta_{1j}(t) \in [0,1],\ j = A,B,C, \\
&\quad \textstyle\sum_k \beta_{2k}(t) = 1, \quad \beta_{2k}(t) \in [0,1],\ k = B,C,
\end{aligned} \tag{3.90}$$

where $\mathcal{G}_{1A}(t,x(t)) = x_1(t) - v_1$, $\mathcal{G}_{1B}(t,x(t)) = x_1(t) - v_2$, $\mathcal{G}_{2B}(t,x(t)) = x_1(t) - v_3$, and $L(x(t),w)$ is given by (3.89).
We will solve (3.90) by using Theorem 9. We start by assuming that $(x^*(t),w^*(t),\alpha^*(t),\beta^*(t))$ is a weak minimum of (3.90) and denote

$$\begin{aligned}
&c_1 := \beta_{1A}\mathcal{G}_{1A} = \beta_{1A}(x_1 - v_1) \le 0, && c_2 := \beta_{1B}\mathcal{G}_{1B} = \beta_{1B}(x_1 - v_2) \le 0, \\
&c_3 := \beta_{2B}\mathcal{G}_{1B} = \beta_{2B}(x_1 - v_2) \le 0, && c_4 := -\beta_{1B}\mathcal{G}_{1A} - \varepsilon \le 0, \\
&c_5 := -\beta_{1C}\mathcal{G}_{1B} - \varepsilon \le 0, && c_6 := -\beta_{2B}\mathcal{G}_{2B} - \varepsilon \le 0, \\
&c_7 = -\beta_{2C}\mathcal{G}_{2B} - \varepsilon = -\beta_{2C}(x_1 - v_3) - \varepsilon \le 0.
\end{aligned}$$

Since the set of phase points

$$\mathcal{N} = \{(x,w,\alpha,\beta) \mid x_1 - v_j = 0, \forall j = 1,2,3\} = \emptyset,$$

the mixed constraints are regular, i.e., the assumption [45, RMC] is satisfied.
We define some needed functions as below.

(i) Augmented Pontryagin function (extended Pontryagin function)

$$\bar{\mathcal{H}}(x,w,\alpha,\beta) = \lambda(t)^T F_0(x(t),\alpha(t),\beta(t)) + \delta L(x(t),w) + \sum_{j=1}^{7} \mu_j c_j,$$

where $\mu_j \ge 0$, $j = 1,\dots,7$, and $F_0(x,\alpha,\beta)$ is the right-hand-side of ODE of (3.90).

(ii) Endpoint Lagrange function

$$l_L(x_0,x_T) = \nu^T \begin{pmatrix} x_0(0) & x_0(T) - S \\ x_1(0) & x_1(T) \end{pmatrix}$$

Subsequently, from the Theorem (9), there exists a tuple of multipliers (λ,δ,μ,ν) satisfying the properties $\lambda : [t_0,t_f] \to I\mathbb{R}^n$ is a Lipschitz continuous function, $\mu_j : [t_0,t_f] \to \mathbb{R}_+$, $j = 1,\dots,7$, are measurable bounded functions, $\delta > 0$, and $\nu > 0$ is a vector; and such that the conditions of the local minimum principle (9) hold true.

(a) the non-negativity conditions: $\nu \geq 0$, $\delta \geq 0$, $\mu_j \geq 0$, $j = 1, \ldots, 7$,

(b) the nontrivial condition: $|\nu| + |\delta| + \sum_{j=1}^{7} \int_{t_0}^{t_f} \mu_j(t)dt > 0$,

(c) the complementary slackness conditions: $\nu^T \begin{pmatrix} x_0(0) & x_0(T) - S \\ x_1(0) & x_1(T) \end{pmatrix} = 0$,

(d) the point-wise complementary slackness conditions: $\mu_i(t)c_i(x^*(t), \beta^*(\cdot)) = 0$ a.e. on $[0, T]$, $i = 1, \ldots, 7$,

(e) the adjoint equation $\dot{\lambda}(t) = -\frac{\partial \bar{\mathcal{H}}}{\partial x}(x^*(t), w^*(t), \alpha^*(t), \beta^*(t))$,

(f) the transversality conditions: $\lambda(0) = -\frac{\partial L_L}{\partial x_0}(x^*(0), x^*(T))$, $\lambda(T) = \frac{\partial L_L}{\partial x_T}(x^*(0), x^*(T))$,

(g) the stationarity condition of the extended PONTRYAGIN function w.r.t. the control

$$\frac{\partial \bar{\mathcal{H}}}{\partial w}(x^*, w^*, \alpha^*, \beta^*) = 0, \ \frac{\partial \bar{\mathcal{H}}}{\partial \alpha}(x^*, w^*, \alpha^*, \beta^*) = 0, \ \frac{\partial \bar{\mathcal{H}}}{\partial \beta}(x^*, w^*, \alpha^*, \beta^*) = 0 \text{ a.e. on } [0, T].$$

Before using the above (a) - (g) conditions, we calculate some needed derivatives.

$$\frac{\partial L(x(t), k)}{\partial x_1} = \begin{cases} 0 & x_1 \leq v_{k+1} \\ e \sum_{i=1}^{5} \frac{-i\gamma}{10} c_i(k) \left(\frac{1}{10}\gamma x_1(t)\right)^{-i-1} & x_1 > v_{k+1} \end{cases}, \quad k = 1, 2, \tag{3.91}$$

$$\frac{\partial T(x_1(t), 1)}{\partial x_1} = \sum_{i=1}^{5} \frac{-i\gamma}{10} b_i(1) \left(\frac{1}{10}\gamma x_1(t) - 0.3\right)^{-i-1}, \tag{3.92}$$

$$\frac{\partial T(x_1(t), 2)}{\partial x_1} = \sum_{i=1}^{5} \frac{-i\gamma}{10} b_i(2) \left(\frac{1}{10}\gamma x_1(t) - 1\right)^{-i-1}, \tag{3.93}$$

$$\frac{\partial R(x_1(t))}{\partial x_1} = 2c(n_{wag})a\gamma^2 x_1(t) + \frac{bW}{2000}\gamma, \tag{3.94}$$

and

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha} = \begin{pmatrix} \lambda_1(t) \sum_j \beta_{1j}(t) f_1^{1j} + \mu_1^g(w - 1) \\ \lambda_1(t) \sum_k \beta_{1k}(t) f_1^{1k} + \mu_2^g(w - 2) \\ -\lambda_1(t) \left(\frac{gR(x_1(t))}{W_{eff}} + C\right) + \mu_3^g(w - 3) \\ -\lambda_1(t) u_{\max} + \mu_4^g(w - 4) \end{pmatrix}, \ \frac{\partial \bar{\mathcal{H}}}{\partial \beta} = \begin{pmatrix} \lambda_1(t)\alpha_1(t) f_1^{1A} + \mu_1 (x_1 - v_1) \\ \lambda_1(t)\alpha_1(t) f_1^{1B} + \mu_2 (x_1 - v_2) \\ \lambda_1(t)\alpha_1(t) f_1^{1C} \\ \lambda_1(t)\alpha_2(t) f_1^{2B} + \mu_3 (x_1 - v_2) \\ \lambda_1(t)\alpha_2(t) f_1^{2C} \end{pmatrix}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial x} = \begin{pmatrix} 0 \\ \lambda_0 + \lambda_1 \left(\alpha_1 \beta_{1C} \frac{\partial f_1^{1C}}{\partial x_1} + \alpha_2 \beta_{2C} \frac{\partial f_1^{2C}}{\partial x_1} - \alpha_3 \frac{g}{W_{eff}} \frac{\partial R(x_1)}{\partial x_1}\right) + \delta \frac{\partial L(x, w)}{\partial x_1} \end{pmatrix}$$

where $\frac{\partial R(x_1(t))}{\partial x_1}$ is calculated in (3.94), and

$$\frac{\partial f_1^{1C}}{\partial x_1} = \frac{g}{W_{eff}} \left(\frac{e\gamma}{10} \sum_{i=0}^{5} -i b_i(1) \left(\frac{\gamma}{10} x_1 - 0.3\right)^{-i-1} - 2c(n_{wag})a\gamma^2 x_1 - \frac{bW\gamma}{2000}\right),$$

$$\frac{\partial f_1^{2C}}{\partial x_1} = \frac{g}{W_{eff}} \left(\frac{e\gamma}{10} \sum_{i=1}^{5} -i b_i(2) \left(\frac{\gamma}{10} x_1 - 1\right)^{-i-1} - 2c(n_{wag})a\gamma^2 x_1 - \frac{bW\gamma}{2000}\right),$$

$$\frac{\partial L(x(t),k)}{\partial x_1} = \beta_{kC}\frac{e\gamma}{10}\sum_{i=1}^{5} -ic_i(k)\left(\frac{\gamma}{10}x_1(t)\right)^{-i-1}, \quad k=1,2.$$

The adjoint equation helps us imply

$$\lambda_0(t) = constant, \tag{3.95}$$

$$\dot{\lambda}_1(t) = -\lambda_1(t)\sum(\lambda,\cdot) - \delta\frac{\partial L(x(t),k)}{\partial x_1} - \lambda_0(t), \tag{3.96}$$

where $\sum(\lambda,\cdot) := \alpha_1(t)\beta_{1C}\frac{\partial f_1^{1C}}{\partial x_1} + \alpha_2(t)\beta_{2C}\frac{\partial f_1^{2C}}{\partial x_1} - \alpha_3(t)\frac{g}{W_{eff}}\frac{\partial R(x_1(t))}{\partial x_1}$.
Denote $\bar{\mathcal{H}} := \bar{\mathcal{H}}(x(t),w(t),\alpha(t),\beta(t))$, the maximality condition gets

$$\bar{\mathcal{H}} = \min_{0\le\alpha,\beta\le 1}\{\lambda(t)^T F_0(x(t),\alpha,\beta) + \delta L(x,w) + \sum_{j=1}^{4}\mu_j^g g_j(\alpha,w) + \sum_{j=1}^{7}\mu_j c_j(x_1(t),\beta)\}. \tag{3.97}$$

The minimum value of (3.97) is directly dependent on the ranges which the value of $x_1(\cdot)$ is belonged. We obtain several cases as follows.
i. If $x_1(t) \le v_1$ then $\beta_{1A} = \beta_{2B} = 1$, since $f_1^{1A} < f_1^{1B}$. Hence we have $\beta_{1B} = \beta_{1C} = \beta_{2C} = 0$. Together with $0 < f_1^{1A} < f_1^{2B}$, we imply $\alpha_1 = 1$, and $\alpha_2 = 0$. Therefore, we lastly obtain $\alpha_3 = \alpha_4 = 0$.
In conclusion, one has $\alpha = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}^T$, and $\beta_1 = \begin{pmatrix}1 & 0 & 0\end{pmatrix}^T$, $\beta_2 = \begin{pmatrix}1 & 0\end{pmatrix}$, where $x_1(t) \le v_1$.
ii. If $v_1 < x_1(t) \le v_2$ then $\beta_{1A} = 0$, and $\beta_{1B} = \beta_{2B} = 1$. Hence $\beta_{1C} = \beta_{2C} = 0$. Together with $0 < f_1^{1B} < f_1^{2B}$, again, we obtain $\alpha_1 = 1$, and $\alpha_2 = 0$. Therefore, $\alpha_3 = \alpha_4 = 0$.
In conclusion, $\alpha = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}^T$, $\beta_1 = \begin{pmatrix}0 & 1 & 0\end{pmatrix}^T$, $\beta_2 = \begin{pmatrix}1 & 0\end{pmatrix}^T$, where $v_1 < x_1(t) \le v_2$.
iii. If $x_1 > v_2$ then $\beta_{1A} = \beta_{1B} = \beta_{2B} = 0$. Hence $\beta_1 = \begin{pmatrix}0 & 0 & 1\end{pmatrix}^T$, and $\beta_2 = \begin{pmatrix}0 & 1\end{pmatrix}^T$. By comparing $T(x_1(t),1)$ with $T(x_1(t),2)$, i.e., f_1^{1C} with f_1^{2C}, we can imply 5 following cases by using four "switch points" of f_1^{1C}, f_1^{2C}.

1. If $v_2 < x_1 \le 8.6572$ then $0 < f_1^{1C} < f_1^{2C}$. Hence $\alpha_1 = 1$, $\alpha_2 = 0$, then $\alpha_3 = \alpha_4 = 0$.
   Therefore $\alpha = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}^T$, where $v_2 < x_1 \le 8.6572$.

2. If $8.6572 < x_1 \le 25.6452$ then $f_1^{1C} > 0$, $f_1^{2C} < 0$. Hence $\alpha_1 = 0$, $\alpha_2 = 1$, so $\alpha_3 = \alpha_4 = 0$. Thus $\alpha = \begin{pmatrix}0 & 1 & 0 & 0\end{pmatrix}^T$, where $8.6572 < x_1 \le 25.6452$.

3. If $25.6452 < x_1 \le 26.8579$ then $f_1^{1C} < 0$, and $f_1^{2C} > 0$. Hence, $\alpha_1 = 1$, and $\alpha_2 = 0$, so $\alpha_3 = \alpha_4 = 0$. Therefore $\alpha = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}^T$, where $25.6452 < x_1 \le 26.8579$.

4. If $26.8579 > x_1 \ge 23.5201$ then $f_1^{1C} > 0$, $f_1^{2C} > 0$. Hence $\alpha_1 = \alpha_2 = 0$. Since $u_{\max} > \frac{gR(x_1(t))}{W_{eff}}$, where $x_1 \in [23.5201, 26.8579)$, we obtain $\alpha_3 = 1$, and $\alpha_4 = 0$.
   Therefore $\alpha = \begin{pmatrix}0 & 0 & 1 & 0\end{pmatrix}^T$, where $26.8579 > x_1 \ge 23.5201$.

5. If $23.5201 > x_1$ then $f_1^{1C} > 0$, $f_1^{2C} > 0$. Hence $\alpha_1 = \alpha_2 = 0$. Since $u_{\max} < \frac{gR(x_1(t))}{W_{eff}}$, where $x_1 \in (0, 23.5201)$, we imply $\alpha_3 = 0$, and $\alpha_4 = 1$.
   Therefore $\alpha = \begin{pmatrix}0 & 0 & 0 & 1\end{pmatrix}^T$, where $23.5201 > x_1$.

Finally, we obtain the optimal controls with the switched points, see Tab. 3.1, which are confirmed by solving this problem by our direct approach, see Subsection 4.4.1.

Table 3.1: Summary of result of controls corresponding to the velocity $x_1(t)$.

$x_1(t)$ [mph]	α	w	β_1	β_2
0.0	$(1,0,0,0)$	1	$(1,0,0)$	$(1,0)$
$v_1 = 0.979474$	$(1,0,0,0)$	1	$(1,0,0)$	$(1,0)$
$v_2 = 6.73211$	$(1,0,0,0)$	1	$(0,1,0)$	$(1,0)$
8.6572	$(1,0,0,0)$	1	$(0,0,1)$	$(0,1)$
25.6452	$(0,1,0,0)$	2	$(0,0,1)$	$(0,1)$
26.8579	$(1,0,0,0)$	1	$(0,0,1)$	$(0,1)$
23.5201	$(0,0,1,0)$	3	$(0,0,1)$	$(0,1)$
0.0	$(0,0,0,1)$	4	$(0,0,1)$	$(0,1)$

Chapter 4

Direct Approaches for Switched Optimal Control Problem: Reformulations and Rounding Solutions

There exists plenty of direct methods for SwOCP. Approaches deal with implicit switched systems can use direct multiple shooting method (see [22, 75]), and switching time instant with variational formulations (see [126]). Some algorithms consider systems with explicit switches based on problem-specific continuous reformulations of discrete valued controls (see [59]), rounding heuristics (see [119]), direct collocation (see [95]).

This chapter investigates several direct approaches for solving Switched Optimal Control Problems (SwOCP). Section 4.1 introduces a novel solution approach for SwOCP, drawing upon a state of the art technique from Filippov's rule. Subsection 4.1.1 provides a detailed examination of this proposed solution. Subsequently, Subsection 4.1.5 explores the condensing procedure for the block structure of the quadratic programming subproblem. This leads to the development of a feedback algorithm for tracking integer control, proposed in Subsection 4.1.6. For comparison purposes, the active set method is leveraged in Subsection 4.1.7. Subsection 4.2 presents a switching-aware rounding algorithm, with Subsection 4.2.2 further expanding this scheme by proposing to exploit a neighboring feedback law. Subsection 4.3 outlines another algorithmic approach for SwOCP, termed Combinatorial Integral Approximation (CIA). The chapter concludes with applications of the proposed approaches to real-world problems: the New York subway problem in Subsection 4.4.1 and the Flat Hybrid Automaton (FHA) problem in Subsection 4.4.2.

4.1 Reformulation

4.1.1 Reformulation

By exploiting Filippov's rule reformulation and relaxation, see Subsection 3.2.1, SwOCP (3.60) is reached as the following formulation

$$\begin{aligned}
\min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & m\left(x(t_f)\right) + \int_{t_0}^{t_f} l(x(t),u(t)) = \varphi[\cdot] \\
\text{s.t.} \quad & \dot{x}(t) = F(x(t),u(t),\alpha_\sigma(t),\alpha_w(t)), \\
& 0 \le r\left(x(t_0),x(t_f)\right), \\
& 0 \le \alpha_\sigma(t)\sigma(x(t)),\ (1-\alpha_\sigma(t))\sigma(x(t)) \le 0, \\
& 0 \le \alpha_\sigma(t) \le 1,\ \textstyle\sum_{w\in\mathcal{W}} \alpha_w(t) = 1,\ 0 \le \alpha_w(t) \le 1, \\
& w \in \mathcal{W}, t \in \mathcal{T} = [t_0,t_f],
\end{aligned} \tag{4.1}$$

therein, we assume that the velocity field of $F(\cdot)$ is convex (otherwise one has to use further relaxation, see Subsection 3.2.1), where

$$F(x(t),u(t),\alpha(t)) = \sum_{w\in\mathcal{W}} \alpha_w(t)\left(\alpha_\sigma(t) f_+\left(x(t),u(t),w\right) + (1-\alpha_\sigma(t)) f_-\left(x(t),u(t),w\right)\right).$$

In this section, the general direct approach for solving SwOCP (4.1) will be presented by employing the direct multiple shooting method, where the original continuous optimal control problem is reformulated as a NLP which is then solved by an iterative solution procedure, a specially tailored *sequential quadratic programming (SQP) algorithm*.

4.1.2 Direct Multiple Shooting Method

Controls discretization

We introduce a discretization of the control trajectories $u(\cdot)$ and $\alpha(\cdot)$ by defining a shooting grid

$$t_0 < t_1 < \ldots < t_{m-1} < t_m = t_f, \quad m \in \mathbb{N}, m \ge 1.$$

On each interval $[t_i,t_{i+1}]$, $i = 0,\ldots,m-1$, of the shooting grid we introduce control parameters q_i^u, $q_i^{\alpha_\sigma}$ and $q_i^{\alpha_w}$ together with associated control base functions $\theta_i^u : \mathcal{T}\times\mathbb{R}^{n_i^{q^u}} \to \mathbb{R}^{n_i^u}$, $\theta_i^{\alpha_\sigma} : \mathcal{T}\times[0,1] \to [0,1]$, and $\theta_i^{\alpha_w} : \mathcal{T}\times[0,1]^{n_i^{q^{\alpha_w}}} \to [0,1]^{n_i^{\alpha_w}}$,

$$\begin{aligned}
& u(t) := \theta_{i,l}(t,q_{i,l}^u),\ 1 \le l \le n_i^{q^u}, \\
& \alpha_\sigma(t) := \theta_i(t,q_i^{\alpha_\sigma}), \qquad\qquad t \in [t_i,t_{i+1}] \subseteq \mathcal{T}, 0 \le i \le m-1. \\
& \alpha_w(t) := \theta_{i,l}(t,q_{i,l}^{\alpha_w}),\ 1 \le l \le n_i^{q^{\alpha_w}},
\end{aligned} \tag{4.2}$$

In conclusion, we could denote

$$\theta_{i,l}(t,q_{i,l}) := [\theta_{i,l}(t,q_{i,l}^u) \quad \theta_i(t,q_i^{\alpha_\sigma}) \quad \theta_{i,l}(t,q_{i,l}^{\alpha_w})]. \tag{4.3}$$

If for instance piecewise constant approximations are used for all control functions, we simply have $\theta_{i,l}(t,q_{i,l}) = \left(q_{i,l}^u,\ q_i^{\alpha_\sigma},\ q_{i,l}^{\alpha_w}\right)$ for $t \in [t_i,t_{i+1}]$. Since $\alpha(t) \in [0,1]^{n_w+1}$, and remember that here $\alpha^T = \left(\alpha_\sigma,\ \alpha_w^T\right)$, we obtain a bounded range

$$0 \le q_i^{\alpha_\sigma} \le 1, \qquad \sum_{w\in\mathcal{W}} q_{i,l}^{\alpha_w} = 1,\ 0 \le q_{i,l}^{\alpha_w} \le 1,\ 1 \le l \le n_i^{q^{\alpha_w}}. \tag{4.4}$$

State discretization

In addition to the control parameters, we introduce state vectors $s_i \in \mathbb{R}^{n_x}$ in all shooting nodes serving as initial values for m IVPs

$$\dot{x}_i(t) = F(x_i(t), q_i), \quad t \in [t_i, t_{i+1}] \subseteq \mathcal{T},\ 0 \leq i \leq m-1, \tag{4.5a}$$
$$x_i(t) = s_i, \qquad\qquad 0 \leq i \leq m-1. \tag{4.5b}$$

where $F(\cdot)$ is the right hand side of ODE in (4.1), and

$$q_i = (q_i^u,\ q_i^{\alpha_\sigma},\ q_i^{\alpha_w})\,. \tag{4.6}$$

Continuity of the solution is ensured by introduction of additional matching conditions

$$s_{i+1} = x_i(t_{i+1}; t_i, s_i, q_i), \quad 0 \leq i \leq m-1, \tag{4.7}$$

where $x_i(t_{i+1}; t_i, s_i, q_i)$ denotes the evaluation of $x_i(\cdot)$ at the final time t_{i+1} of shooting interval i, and depending on the start time t_i, initial value s_i, and control parameters q_i on that interval.

Constraint discretization

The point constraints can be rewritten as follows

$$0 \leq r(s_0, s_m),$$

while the additional constraints are reached as belows

$$0 \leq q_i^{\alpha_\sigma}\sigma(s_i), \quad (1 - q_i^{\alpha_\sigma})\sigma(s_i) \leq 0, \quad 0 \leq i \leq m-1. \tag{4.8}$$

Objective

By rewriting the Mayer term $m(s_m)$ as the final term $l_m(s_m, q_m)$, a formulation of the objective function with respect to the shooting grid structure is found

$$\varphi = \sum_{i=0}^{m} l_i(s_i, q_i). \tag{4.9}$$

Multiple shooting discretized for SwOCP

In conclusion, the discretized multiple shooting of SwOCP can be cast as a nonlinear problem

$$\begin{aligned}
\min_{y} \quad & \textstyle\sum_{i=0}^{m} l_i(s_i, q_i) \\
\text{s.t.} \quad & 0 = x_i(t_{i+1}; t_i, y_i) - s_{i+1},\ 0 \leq i \leq m-1, \\
& 0 \leq r\,(s_0, s_m)\,, \\
& 0 \leq q_i^{\alpha_\sigma}\sigma(s_i),\ 0 \leq i \leq m-1, \\
& 0 \leq (q_i^{\alpha_\sigma} - 1)\sigma(s_i),\ 0 \leq i \leq m-1, \\
& 0 \leq q_i^{\alpha_\sigma} \leq 1,\ 0 \leq i \leq m-1, \\
& 0 \leq q_i^{\alpha_w} \leq 1,\ \textstyle\sum_{w \in \mathcal{W}} q_i^{\alpha_w} = 1,\ 0 \leq i \leq m-1,\ w \in \mathcal{W}.
\end{aligned} \tag{4.10}$$

The constraints $0 \le q_i^{\alpha_\sigma}\sigma(s_i)$, $(1-q_i^{\alpha_\sigma})\sigma(s_i) \le 0$ (in (4.10)) are vanishing constraints and may create the problems with numerical methods. To overcome this trouble, we suggest to reformulate these constraints as follows, where the function $\sigma(s_i)$ is bounded on interval $[t_i, t_{i+1}]$, i.e., $l_\sigma(t_i) \le \sigma(s_i) \le u_\sigma(t_i)$, in short we write $l_\sigma \le \sigma(s_i) \le u_\sigma$. Later we investigate these constraints directly in Subsection 4.1.7.
We start by introducing a new variable $q_i^z := q_i^{\alpha_\sigma}\sigma(s_i)$, $i = 0, \ldots, m-1$. Add the following constraints

$$q_i^z \le u_\sigma q_i^{\alpha_\sigma}, \quad q_i^z \ge l_\sigma q_i^{\alpha_\sigma}, \qquad i = 0, \ldots, m-1, \tag{4.11}$$
$$q_i^z \le \sigma(s_i) - l_\sigma(1-q_i^{\alpha_\sigma}), \quad q_i^z \ge \sigma(s_i) - u_\sigma(1-q_i^{\alpha_\sigma}), \qquad i = 0, \ldots, m-1. \tag{4.12}$$

Consider the case $q_i^{\alpha_\sigma} = 0$. Then $q_i^z = 0$. Inequalities (4.11) force $q_i^z = 0$, while inequalities (4.12) say $\sigma(s_i) - u_\sigma \le q_i^z \le \sigma(s_i) - l_\sigma$, and $q_i^z = 0$ satisfies those inequalities.
The case $q_i^{\alpha_\sigma} \in (0,1)$ implies $0 < q_i^z < \sigma(s_i)$, and those constraints (4.11-4.12) are satisfied.
For the remain case $q_i^{\alpha_\sigma} = 1$, we have $q_i^z = \sigma(s_i)$. Inequalities (4.11) imply $l_\sigma \le q_i^z \le u_\sigma$, which is satisfied by $q_i^z = \sigma(s_i)$. Moreover, inequalities (4.12) force $q_i^z = \sigma(s_i)$ as desired.

By adding the above variable q_i^z, (4.10) is rewritten as the following nonlinear problem

$$\begin{aligned}
\min_{y} \quad & \textstyle\sum_{i=0}^{m} l_i(s_i, q_i) \\
\text{s.t.} \quad & 0 = x_i(t_{i+1}; t_i, y_i) - s_{i+1},\ 0 \le i \le m-1, \\
& 0 \le r\left(s_0, s_m\right), \\
& 0 \le q_i^z,\ 0 \le i \le m-1, \\
& 0 \le q_i^z - \sigma(s_i),\ 0 \le i \le m-1, \\
& 0 \le q_i^{\alpha_\sigma} \le 1,\ 0 \le i \le m-1, \\
& 0 \le q_i^{\alpha_w} \le 1,\ \textstyle\sum_{w\in\mathcal{W}} q_i^{\alpha_w} = 1,\ 0 \le i \le m-1,\ w \in \mathcal{W}, \\
& l_\sigma q_i^{\alpha_\sigma} \le q_i^z \le u_\sigma q_i^{\alpha_\sigma},\ 0 \le i \le m-1, \\
& u_\sigma(q_i^{\alpha_\sigma} - 1) \le q_i^z - \sigma(s_i) \le l_\sigma(q_i^{\alpha_\sigma} - 1),\ 0 \le i \le m-1,
\end{aligned} \tag{4.13}$$

where

$$y := (s_0, q_0, \ldots, s_{m-1}, q_{m-1}, s_m),\ y_i := (s_i, q_i),\ 0 \le i \le m-1,\ y_m := s_m, \tag{4.14}$$

$l_\sigma \le \sigma(s_i) \le u_\sigma$, and note that q_i is given by (4.6), $i = 0, \ldots, m-1$.

For $i = 1, \ldots, m-1$, we investigate some following instances:
1st case: $\sigma(s_i) > 0$. Then the additional constraints of (4.13) imply $q_i^{\alpha_\sigma} = 1$, and so $q_i^z = \sigma(s_i)$.
2nd case: $\sigma(s_i) < 0$. Then again, the additional constraints of (4.13) imply $q_i^{\alpha_\sigma} = 0$, and hence $q_i^z = 0$.
3rd case: $\sigma(s_i) = 0$. Then the additional constraints of (4.13) are true for all $q_i^{\alpha_\sigma}$ satisfy

$$0 \le q_i^{\alpha_\sigma} \le 1,$$

the additional constraints become the vanishing constraints. Therefore we propose a relaxed

technique to problem (4.13). As a result, one obtains the relaxed problem as follows

$$\begin{aligned}
\min_{y} \quad & \textstyle\sum_{i=0}^{m} l_i(s_i, q_i) \\
\text{s.t.} \quad & 0 = x_i(t_{i+1}; t_i, y_i) - s_{i+1},\ 0 \le i \le m-1, \\
& 0 \le r\left(s_0, s_m\right), \\
& 0 \le q_i^z,\ 0 \le i \le m-1, \\
& 0 \le q_i^z - \sigma(s_i),\ 0 \le i \le m-1, \\
& 0 \le q_i^{\alpha_\sigma} \le 1,\ 0 \le i \le m-1, \\
& 0 \le q_i^{\alpha_w} \le 1,\ \textstyle\sum_{w \in \mathcal{W}} q_i^{\alpha_w} = 1,\ 0 \le i \le m-1,\ w \in \mathcal{W}, \\
& l_\sigma q_i^{\alpha_\sigma} \le q_i^z \le u_\sigma q_i^{\alpha_\sigma},\ 0 \le i \le m-1, \\
& u_\sigma(q_i^{\alpha_\sigma} - 1) \le q_i^z - \sigma(s_i) \le l_\sigma(q_i^{\alpha_\sigma} - 1),\ 0 \le i \le m-1,
\end{aligned} \tag{4.15}$$

where y, $y_i, 0 \le i \le m-1$, are satisfied (4.14), and $l_\sigma \le \sigma(s_i) \le u_\sigma$, and $q_i = (q_i^u,\ q_i^{\alpha_\sigma},\ q_i^{\alpha_w},\ q_i^z)$, $i = 0, \ldots, m-1$.

4.1.3 Constraint Qualification

Lemma 7. *Consider NLP (4.15) with conditions* (4.14)*. Then LICQ and MFCQ are not satisfied if* $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$ *for some* $i = 0, \ldots, m-1$.

Proof. First, we can rewrite (4.13) as the Nonlinear Programming as follows

$$\begin{aligned}
\min_{y(.)} \quad & \varphi\left(y\right) \\
\text{s.t.} \quad & G(y) = 0, \\
& H(y) \ge 0,
\end{aligned} \tag{4.16}$$

where $\varphi(y) = \sum_{i=0}^{m} l_i(s_i, q_i)$, $G(y) = 0$ casts for all equality constraints from the matching conditions in (4.13), $H(y) \ge 0$ means all inequality constraints of (4.13).
The set of all feasible points of NLP (4.16) is denoted by

$$\mathcal{F} \overset{\text{def}}{=} \left\{ y = (s_0, q_0, \ldots, s_{m-1}, q_{m-1}, s_m) \mid G(y) = 0, H(y) \ge 0 \right\},$$

where $s_i \in \mathbb{R}^{n_i^x}$, $q_i = (q_i^u,\ q_i^{\alpha_\sigma},\ q_i^{\alpha_w},\ q_i^z)$, $q_i^u \in \mathbb{R}^{n_i^u}$, $q_i^{\alpha_\sigma} \in [0,1]$, $q_i^{\alpha_w} \in [0,1]^{n_i^{\alpha_w}}$, $-1 \le q_i^z \le 1$, for $i = 0, \ldots, m-1$.
Let $\bar{y}$ be a feasible point of NLP (4.16). For $i = 0, \ldots, m-1$, with $x = x(t_{i+1}; t_i, y_i)$, consider some cases as follows:
For $\sigma(s_i) > 0$, i.e., $q_i^{\alpha_\sigma} = 1$ and $q_i^z = \sigma(s_i)$, then only these constraints $q_i^z - \sigma(s_i) \ge 0$, $u_\sigma(q_i^{\alpha_\sigma} - 1) \le q_i^z - \sigma(s_i) \le l_\sigma(q_i^{\alpha_\sigma} - 1)$ are active, the remain constraints are inactive, and the Jacobian matrix of (4.16) includes rows

$$\begin{pmatrix}
-\frac{\partial \sigma(s_i)}{\partial s_i} & 0 & 0 & 0 & 1 \\
\frac{\partial \sigma(s_i)}{\partial s_i} & 0 & l_\sigma & 0 & -1 \\
-\frac{\partial \sigma(s_i)}{\partial s_i} & 0 & -u_\sigma & 0 & 1
\end{pmatrix}. \tag{4.17}$$

For $\sigma(s_i) < 0$, i.e., $q_i^{\alpha_\sigma} = 0$ and $q_i^z = 0$, then similarly, the Jacobian matrix of (4.16) includes rows

$$\begin{pmatrix}
0 & 0 & 0 & 0 & 1 \\
0 & 0 & u_\sigma & 0 & -1 \\
0 & 0 & -l_\sigma & 0 & 1
\end{pmatrix}. \tag{4.18}$$

For remain case $\sigma(s_i) = 0$, ones have $0 \leq q_i^{\alpha_\sigma} \leq 1$ and $q_i^z = 0$, the Jacobian matrix of (4.16) includes rows

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ -\frac{\partial \sigma(s_i)}{\partial s_i} & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.19}$$

From (4.17-4.19), we can conclude that this Jacobian matrix has the full rank, if $\frac{\partial \sigma(s_i)}{\partial s_i} \neq 0$, $i = 0, \ldots, m-1$. But if $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$ for some $i = 0, \ldots, m-1$, which will lead to two rows of vector are dependent, then the Jacobian matrix does not have the full rank.
Thus LICQ is satisfied if $\frac{\partial \sigma(s_i)}{\partial s_i} \neq 0$, $i = 0, \ldots, m-1$.
Moreover, suppose that there exits a vector d such that

$$\begin{cases} G_y(\bar{y})^T d = 0 \\ (H_{\mathcal{A}})_y (\bar{y})^T d > 0 \end{cases} \tag{4.20}$$

then it will lead to $d = (0 \ldots 0)^T$ where $\frac{\partial \sigma(s_i)}{\partial s_i} \neq 0$ for $i = 0, \ldots, m-1$. But if $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$ for some $i = 0, \ldots, m-1$, then which will lead to the constraints $q_i^z \geq 0$ and $q_i^z - \sigma(s_i) \geq 0$ are active at the same time, i.e., no vector d satisfy (4.20). Thus we can conclude that MFCQ is not satisfied if $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$ for some $i = 0, \ldots, m-1$. □

Remark 22. *From the results of Lem. 7 and Remark 6, we conclude that ACQ is satisfied if $\frac{\partial \sigma(s_i)}{\partial s_i} \neq 0$, for $i = 0, \ldots, m-1$.*

Lemma 8. *Consider NLP (4.15) with conditions (4.14) must be satisfied. Then ACQ is satisfied for all $\sigma(s_i)$, $i = 0, \ldots, m-1$.*

Proof. Since the result of of Remark 22, we only need to check ACQ for the case $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$, $i = 0, \ldots, m-1$. Let $\bar{y}$ be a feasible point of (4.16). We first require the definition of the tangent cone $\mathcal{T}(\bar{y}, \mathcal{F})$ of $\mathcal{F}$ in the point $\bar{y}$ and the linearized cone $\mathcal{L}(\bar{y})$ of problem (4.16) in $\bar{y}$, see Def. 13, where $H_{\mathcal{A}} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{A}|}$ is the restriction of the inequality constraint function H onto the active inequality constraints.
Since $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$ for $i = 0, \ldots, m-1$, then the constraints $q_i^z \geq 0$ and $q_i^z - \sigma(s_i) \geq 0$ are active at the same time, i.e., there are not exist any vector d satisfy (4.20), which leads to $\mathcal{L}(\bar{y}) = \emptyset$. Note that the inclusion $\mathcal{T}(\bar{y}, \mathcal{F}) \subseteq \mathcal{L}(\bar{y})$ always holds and that $\mathcal{T}(\bar{y}, \mathcal{F})$ is always closed, while $\mathcal{L}(\bar{y})$ is polyhedral and thus closed and convex. Therefore we have $\mathcal{T}(\bar{y}, \mathcal{F}) = \mathcal{L}(\bar{y}) = \emptyset$, i.e., the ACQ is satisfied. □

Remark 23. In numerical applications (see Sections 4.4, 6.2 and 6.3), one way to overcome the situation, in which LICQ is not satisfy where $\frac{\partial \sigma(s_i)}{\partial s_i} = 0$, $i = 0, \ldots, m-1$, is to relax the constraints $0 \leq q_i^z$, $0 \leq q_i^z - \sigma(s_i)$, $0 \leq i \leq m-1$, as follows

$$-\varepsilon \leq q_i^z, \ 0 \leq i \leq m-1, \tag{4.21}$$
$$-\varepsilon \leq q_i^z - \sigma(s_i), \ 0 \leq i \leq m-1, \tag{4.22}$$

with $\varepsilon > 0$ small enough. Then for NLP (4.15) with updated constraints (4.21-4.22), LICQ is satisfy for all $\sigma(s_i)$, $i = 0, \ldots, m-1$. Similarly to Lemma 7, one can easily check that active constraints satisfy LICQ.

4.1.4 Quadratic Programming Subproblem and SQP Algorithm

The SQP algorithm deals with the NLP problem where all functions are explicitly or implicitly defined as functions of the multiple shooting variables only. The numerical ODE solution on the multiple shooting intervals is performed in an underlying evaluation module and has to be carried out with sufficiently high integration tolerance.
Starting with an initial guess y^0 provided by the user, the SQP algorithm iterates

$$y^{k+1} = y^k + \delta^k \Delta y^k,$$

with step directions Δy^k (and relaxation factors $\delta^k \in (0, 1]$), until a pre-specified convergence criterion is satisfied.
At the k-th SQP iteration with multiple shooting variables y^k, the algorithm evaluates the NLP functions and their derivatives with respect to y. In this way, linearizations of the originally nonlinear NLP functions are obtained that are used to build a quadratic programming QP subproblem. Moreover, an approximation H^k of the Hessian matrix of the Lagrangian function is calculated. The QP subproblem solved at the k-th SQP iteration can be written as:

$$\begin{aligned}
\min_{\Delta y} \quad & \nabla_y \left(\textstyle\sum_{i=0}^{m} l_i(s_i^k, q_i^k)\right)^T \Delta y + \tfrac{1}{2}\Delta y^T H^k \Delta y \\
\text{s.t.} \quad & 0 = x_i^k - s_{i+1}^k + \left(\nabla_{y_i} x_i^k - \nabla_{y_i} s_{i+1}^k\right)^T \Delta y_i, \\
& 0 \le r(s_0^k, s_m^k) + \nabla_y r(s_0^k, s_m^k)^T \Delta y, \\
& 0 \le (q_i^z)^k - \sigma(s_i^k) + \left(\nabla_{y_i}\left((q_i^z)^k - \sigma(s_i^k)\right)\right)^T \Delta y_i, \\
& 0 \le u_\sigma(q_i^{\alpha_\sigma})^k - (q_i^z)^k + \left(\nabla_{y_i}\left(u_\sigma(q_i^{\alpha_\sigma})^k - (q_i^z)^k\right)\right)^T \Delta y_i, \qquad 0 \le i \le m-1, \\
& 0 \le (q_i^z)^k - l_\sigma(q_i^{\alpha_\sigma})^k + \left(\nabla_{y_i}\left((q_i^z)^k - l_\sigma(q_i^{\alpha_\sigma})^k\right)\right)^T \Delta y_i, \\
& 0 \le \sigma(s_i^k) + l_\sigma(q_i^{\alpha_\sigma})^k - (q_i^z)^k - l_\sigma + \left(\nabla_{y_i}\left(\sigma(s_i^k) + l_\sigma(q_i^{\alpha_\sigma})^k - (q_i^z)^k\right)\right)^T \Delta y_i, \\
& 0 \le (q_i^z)^k - \sigma(s_i^k) - u_\sigma(q_i^{\alpha_\sigma})^k + u_\sigma + \left(\nabla_{y_i}\left((q_i^z)^k - \sigma(s_i^k) - u_\sigma(q_i^{\alpha_\sigma})^k\right)\right)^T \Delta y_i,
\end{aligned} \tag{4.23}$$

where the bounds are satisfied $-q_i^{\alpha_\sigma} \le \Delta q_i^{\alpha_\sigma} \le 1 - q_i^{\alpha_\sigma}$, $-q_i^{\alpha_w} \le \Delta q_i^{\alpha_w} \le 1 - q_i^{\alpha_w}$ for $i = 0, \ldots, m-1$, and $\sum_{w \in \mathcal{W}} \Delta q_i^{\alpha_w} = 0$, $i = 0, \ldots, m-1$, while $\Delta y_m = \Delta s_m$; therein $l_\sigma \le \sigma(s_i) \le u_\sigma$, $0 \le q_i^{\alpha_\sigma} \le 1$, $0 \le q_i^{\alpha_w} \le 1$ for $i = 0, \ldots, m-1$, and $\sum_{w \in \mathcal{W}} q_i^{\alpha_w} = 1$ for $i = 0, \ldots, m-1$, and $q_i^z \ge 0$, $i = 0, \ldots, m-1$; with $x_i = x_i(t_{i+1}; t_i, y_i)$, $\Delta y_i := (\Delta s_i, \Delta q_i^u, \Delta q_i^{\alpha_\sigma}, \Delta q_i^{\alpha_w}, \Delta q_i^z)$, $i = 0, \ldots, m-1$, and

$$\Delta y = (\Delta s_0, \Delta q_0 \ldots, \Delta s_{m-1}, \Delta q_{m-1}, \Delta s_m),$$

where $\Delta q_i = (\Delta q_i^u, \Delta q_i^{\alpha_\sigma}, \Delta q_i^{\alpha_w}, \Delta q_i^z)$, $i = 0, \ldots, m-1$, and Ω is $\mathbb{R}^n$ or a suitably chosen box in $\mathbb{R}^n$ (that contains $\Delta y^k = 0$).
The QP subproblem is then solved and results in a direction Δy^k that helps to determine the next iterate $y^{k+1} = y^k + \delta^k \Delta y^k$. Different line search strategies are implemented that determine the relaxation factor δ^k.
For the new values of the multiple shooting variables all NLP functions and derivatives are again evaluated, a new Hessian matrix approximation H^{k+1} is provided and a new QP subproblem is solved for the next SQP iteration.
The iteration stops when the solution reaches accuracy, and is measured by the KKT-tolerance. It indicates how many digits the objective value is expected to be correct.

4.1.5 Condensing: Block Structure of QP Subproblem

We start with vectors h_i denoting the matching conditions residuals

$$h_i(y_i, s_{i+1}) := x_i(t_{i+1}; t_i, y_i) - s_{i+1}.$$

The matrices H_i denote the Hessians (or a suitable approximations) of the NLP's Lagrangian, and the vectors g_i denotes the gradients of the NLP's objective function. Matrices X_i, R_i, C_i and D_i denote linearizations of the constraint functions and the additional constraint functions obtained in y_i,

$$\begin{aligned} H_i &:\approx \frac{\partial^2 L(y_i, \eta_i)}{\partial y_i^2}, \quad & g_i &:= \frac{\partial l_i(s_i, q_i)}{\partial y_i}, \quad & X_i &:= \frac{\partial x_i(t_{i+1}; t_i, y_i)}{\partial y_i}, \quad & R_i &:= \frac{\partial r(s_0, s_m)}{\partial s_i}, \\ C_i &:= \frac{\partial \sigma(s_i)}{\partial y_i}, & D_i &:= \frac{\partial q_i^{\alpha_\sigma}}{\partial y_i}, & E_i &:= \frac{\partial q_i^z}{\partial y_i}, \end{aligned}$$

where $\eta_i^T = (\eta_{1,i}^T, \eta_2^T, \eta_{3,i}, \ldots, \eta_{12,i}, \eta_{13,i})$, and

$$\begin{aligned} L(y_i, \eta_i) := \varphi(y_i) &- \eta_{1,i}^T h_i(y_i, s_{i+1}) - \eta_{3,i} q_i^z - \eta_{4,i} \left(q_i^z - \sigma(s_i)\right) \\ &- \eta_{5,i} \left(u_\sigma q_i^{\alpha_\sigma} - q_i^z\right) - \eta_{6,i} \left(q_i^z - l_\sigma q_i^{\alpha_\sigma}\right) - \eta_{7,i} \left(\sigma(s_i) + l_\sigma q_i^{\alpha_\sigma} - q_i^z - l_\sigma\right) \\ &- \eta_{8,i} \left(q_i^z - \sigma(s_i) - u_\sigma q_i^{\alpha_\sigma} + u_\sigma\right) - \eta_{9,i} \left(1 - q_i^{\alpha_\sigma}\right) - \eta_{10,i} q_i^{\alpha_\sigma} \\ &- \eta_{11,i} \left(1 - q_i^{\alpha_w}\right) - \eta_{12,i} q_i^{\alpha_w} - \eta_{13,i} \left(q_i^{\alpha_w} - 1\right), \quad i = 0, \ldots, m-1, \end{aligned} \tag{4.24}$$

therein, $L(y_m, \eta_m) = \varphi(y_m) - \eta_2^T r(s_0, s_m)$, with $\eta_m = \eta_2$.
Then, with the abbreviations (4.27-4.28), we can rewrite (4.23) as follows

$$\begin{aligned} \min_{\Delta y} \quad & \textstyle\sum_{i=0}^m \left(g_i^T \Delta y_i + \frac{1}{2} \Delta y_i^T H_i \Delta y_i\right) \\ \text{s.t.} \quad & 0 = h_i(y_i, s_{i+1}) + X_i \Delta y_i - \Delta s_{i+1}, \\ & 0 \le q_i^z - \sigma(s_i) - C_i^s \Delta s_i + E_i^{q^z} \Delta q_i^z, \\ & 0 \le u_\sigma q_i^{\alpha_\sigma} - q_i^z + u_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} - E_i^{q^z} \Delta q_i^z, \\ & 0 \le q_i^z - l_\sigma q_i^{\alpha_\sigma} + E_i^{q^z} \Delta q_i^z - l_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma}, \qquad 0 \le i \le m-1, \\ & 0 \le \sigma(s_i) + l_\sigma q_i^{\alpha_\sigma} - q_i^z - l_\sigma + C_i^s \Delta s_i + l_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} - E_i^{q^z} \Delta q_i^z, \\ & 0 \le u_\sigma - \sigma(s_i) - u_\sigma q_i^{\alpha_\sigma} + q_i^z - C_i^s \Delta s_i - u_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + E_i^{q^z} \Delta q_i^z, \\ & 0 \le r(s_0, s_m) + R_0^T \Delta s_0 + R_m^T \Delta s_m, \\ & 0 \le q_i^z + \Delta q_i^z, \\ & -q_i^{\alpha_\sigma} \le \Delta q_i^{\alpha_\sigma} \le 1 - q_i^{\alpha_\sigma}, \\ & -q_i^{\alpha_w} \le \Delta q_i^{\alpha_w} \le 1 - q_i^{\alpha_w}, \ \textstyle\sum_{w \in \mathcal{W}} \Delta q_i^{\alpha_w} = 0, \end{aligned} \tag{4.25}$$

where $\Delta y_m = \Delta s_m$. Remember that $l_\sigma \le \sigma(s_i) \le u_\sigma$, $0 \le q_i^{\alpha_\sigma} \le 1$, $0 \le q_i^{\alpha_w} \le 1$ and $\sum_{w \in \mathcal{W}} q_i^{\alpha_w} = 1$, and $q_i^z \ge 0$, $q_i^z - \sigma(s_i) \ge 0$ for $i = 0, \ldots, m-1$.
To exploited the "block structure" in QP (4.25), we start with the linearized matching conditions

$$\Delta s_{i+1} = X_i^s \Delta s_i + X_i^{q^u} \Delta q_i^u + X_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + X_i^{q^{\alpha_w}} \Delta q_i^{\alpha_w} + X_i^{q^z} \Delta q_i^z + h_i, \tag{4.26}$$

with the abbreviations, for $i = 0, \dots, m-1$,

$$X_i^s = \frac{\partial x_i(t_{i+1}; t_i, y_i)}{\partial s_i} = \frac{\partial x_i(t_{i+1}; t_i, s_i, q_i^u, q_i^{\alpha_\sigma}, q_i^{\alpha_w}, q_i^z)}{\partial s_i}, \tag{4.27a}$$

$$X_i^{q^u} = \frac{\partial x_i(t_{i+1}; t_i, s_i, y_i)}{\partial q_i^u}, \qquad X_i^{q^{\alpha_\sigma}} = \frac{\partial x_i(t_{i+1}; t_i, s_i, y_i)}{\partial q_i^{\alpha_\sigma}}, \tag{4.27b}$$

$$X_i^{q^{\alpha_w}} = \frac{\partial x_i(t_{i+1}; t_i, s_i, y_i)}{\partial q_i^{\alpha_w}}, \qquad X_i^{q^z} = \frac{\partial x_i(t_{i+1}; t_i, s_i, y_i)}{\partial z_i}, \tag{4.27c}$$

$$R_0 = \frac{\partial r(s_0, s_m)}{\partial s_0}, \quad h_i = h_i(y_i, s_{i+1}), \tag{4.27d}$$

$$C_i^s = \frac{\partial \sigma(s_i)}{\partial s_i}, \qquad D_i^{q^{\alpha_\sigma}} = \frac{\partial q_i^{\alpha_\sigma}}{\partial q_i^{\alpha_\sigma}} = 1, \quad E_i^{q^z} = \frac{\partial q_i^z}{\partial z_i} = 1. \tag{4.27e}$$

where $y_i = (s_i, q_i^u, q_i^{\alpha_\sigma}, q_i^{\alpha_w}, q_i^z)$, and for $i = m$ (cf. [22, Eq. (28)]),

$$R_m = \frac{\partial r(s_0, s_m)}{\partial s_m}. \tag{4.28}$$

The block structure in QP (4.25) is exploited in a condensing step that transforms the QP into a related, considerably smaller, and densely populated one. Here we briefly review this *condensing* algorithm due to [22] and great detail in [86], and we adapt this algorithm to our problem. Therein, we also use the Block Gaussian Elimination algorithm, e.g. see Alg. 1.
We start by reordering the constraint matrix of (4.25) from the single shooting values $\Delta u = (\Delta s_0, \Delta q_0^u, \Delta q_0^{\alpha_\sigma}, \Delta q_0^{\alpha_w}, \Delta q_0^z \dots \Delta q_{m-1}^u, \Delta q_{m-1}^{\alpha_\sigma}, \Delta q_{m-1}^{\alpha_w}, \Delta q_{m-1}^z)$ to separate the additionally introduced values $\Delta v = (\Delta s_1, \dots, \Delta s_m)$, as dense matrix in page 69, where the blanks are mentioned zero-blocks.

Since the matching condition (4.26), we use the negative identity matrix blocks as pivots to eliminate the multiple shooting values $(\Delta s_1, \dots, \Delta s_m)$ from this system by the usual Gaussian method for triangular matrices.
The dense constraint matrix $\begin{pmatrix} \bar{X} & -I \\ \bar{R} & 0 \end{pmatrix}$ is obtained from the elimination procedure, see in page 70.
Then we deduce the transformed QP in terms of Δv and Δu as belows

$$\begin{aligned} \min_{\Delta v, \Delta u} \quad & g^T \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix}^T H \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} \\ \text{s.t.} \quad & 0 = \bar{h} + \bar{X} \Delta u - I \Delta v \\ & 0 \leq \bar{r} + \bar{R} \Delta u \end{aligned} \tag{4.29}$$

with appropriate right hand side vectors $\bar{h}$ and $\bar{r}$ obtained by applying the Gaussian elimination steps to h and r, respectively, and therein

$$H = \begin{pmatrix} \bar{H}_{11} & \bar{H}_{12} \\ \bar{H}_{12}^T & \bar{H}_{22} \end{pmatrix}, \quad g = \begin{pmatrix} \bar{g}_1 \\ \bar{g}_2 \end{pmatrix}.$$

$$
\left(\begin{array}{ccccccccc|cccc}
X_0^s & X_0^{q^u} & X_0^{q^{\alpha_\sigma}} & X_0^{q^{\alpha_w}} & X_0^{q^z} & & & & & -I & & & \\
 & & & X_1^{q^u} & X_1^{q^{\alpha_\sigma}} & X_1^{q^{\alpha_w}} & X_1^{q^z} & & & X_1^s & -I & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & & & \ddots & \ddots & \\
 & & & & & X_{m-1}^{q^u} & X_{m-1}^{q^{\alpha_\sigma}} & X_{m-1}^{q^{\alpha_w}} & X_{m-1}^{q^z} & & & X_{m-1}^s & -I \\
\hline
-C_0^s & & & & E_0^{q^z} & & & & & & & & \\
 & & & & & & E_1^{q^z} & & & -C_1^s & & & \\
 & & & & & & & \ddots & & & \ddots & & \\
 & & & & & & & & E_{m-1}^{q^z} & & & -C_{m-1}^s & \\
 & & u_\sigma D_0^{q^{\alpha_\sigma}} & & -E_0^{q^z} & & & & & & & & \\
 & & & & u_\sigma D_1^{q^{\alpha_\sigma}} & & -E_1^{q^z} & & & & & & \\
 & & & & & \ddots & & \ddots & & & & & \\
 & & & & & & u_\sigma D_{m-1}^{q^{\alpha_\sigma}} & & -E_{m-1}^{q^z} & & & & \\
 & & -l_\sigma D_0^{q^{\alpha_\sigma}} & & E_0^{q^z} & & & & & & & & \\
 & & & & -l_\sigma D_1^{q^{\alpha_\sigma}} & & E_1^{q^z} & & & & & & \\
 & & & & & \ddots & & \ddots & & & & & \\
 & & & & & & -l_\sigma D_{m-1}^{q^{\alpha_\sigma}} & & E_{m-1}^{q^z} & & & & \\
C_0^s & & l_\sigma D_0^{q^{\alpha_\sigma}} & & -E_0^{q^z} & & & & & & & & \\
 & & & & l_\sigma D_1^{q^{\alpha_\sigma}} & & -E_1^{q^z} & & & C_1^s & & & \\
 & & & & & & & \ddots & & & \ddots & & \\
 & & & & & & l_\sigma D_{m-1}^{q^{\alpha_\sigma}} & & -E_{m-1}^{q^z} & & & C_{m-1}^s & \\
-C_0^s & & -u_\sigma D_0^{q^{\alpha_\sigma}} & & E_0^{q^z} & & & & & & & & \\
 & & & & -u_\sigma D_1^{q^{\alpha_\sigma}} & & E_1^{q^z} & & & -C_1^s & & & \\
 & & & & & & & \ddots & & & \ddots & & \\
 & & & & & & -u_\sigma D_{m-1}^{q^{\alpha_\sigma}} & & E_{m-1}^{q^z} & & & -C_{m-1}^s & \\
R_0 & & & & & & & & & & & & R_m
\end{array}\right)
$$

By eliminating Δv, since $\Delta v = \bar{h} + \bar{X}\Delta u$, system (4.29) is rewritten as a final *condensed* QP

$$
\begin{array}{ll}
\min\limits_{\Delta u} & \hat{g}^T \Delta u + \frac{1}{2}\Delta u^T \hat{H} \Delta u \\
\text{s.t.} & 0 \leq \bar{r} + \bar{R}\Delta u
\end{array}
\tag{4.30}
$$

with the following dense Hessian matrix and gradient obtained from substitution of Δv in the objective of (4.29)

$$
\begin{aligned}
\hat{H} &= \bar{H}_{11} + \bar{H}_{12}\bar{X} + \bar{X}^T\bar{H}_{12}^T + \bar{X}^T\bar{H}_{22}\bar{X}, \\
\hat{g} &= \bar{g}_1 + \bar{X}^T\bar{g}_2 + \bar{H}_{12}^T\bar{h} + \bar{X}^T\bar{H}_{22}\bar{h}.
\end{aligned}
\tag{4.31}
$$

After the solution to the QP subproblem is obtained, an appropriate rounding procedure will be employed to return the "integer" controls. We are interested in the feedback controls, that's why we will investigate a "Feedback Algorithm" in the following subsection.

$$
\left(\begin{array}{c|cccccccccccccc|c}
 & X_0^s & X_0^u & X_0^{u,w} & X_0^{q,w} & X_0^{q,s} & X_1^u & X_1^{q,w} & X_1^{q,s} & & & & & & -I \\
 & X_1^s X_0^s & X_1^s X_0^u & X_1^s X_0^{u,w} & X_1^s X_0^{q,w} & X_1^s X_0^{q,s} & X_2^u X_1^u & X_2^{q,w} X_1^u & X_2^{q,s} X_1^u & & & & & & -I \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & & \ddots \\
 & \prod_0^{m-1} & \prod_1^{m-1} X_0^u & \prod_1^{m-1} X_0^{u,w} & \prod_1^{m-1} X_0^{q,w} & \prod_1^{m-1} X_0^{q,s} & \prod_2^{m-1} X_1^u & \prod_2^{m-1} X_1^{q,w} & \prod_2^{m-1} X_1^{q,s} & \cdots & X_{m-1}^u & X_{m-1}^{q,w} & X_{m-1}^{q,s} & & -I \\
\hline
 & -C_0^s & & & & E_0^{q,s} & & & E_1^{q,s} & & & & & & \\
 & -C_1^s X_0^s & -C_1^s X_0^u & -C_1^s X_0^{u,w} & -C_1^s X_0^{q,w} & -C_1^s X_0^{q,s} & & & & & & & & & \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & & \ddots & & & & & \\
 & -C_{m-1}^s \prod_0^{m-2} & -C_{m-1}^s \prod_1^{m-2} X_0^u & -C_{m-1}^s \prod_1^{m-2} X_0^{u,w} & -C_{m-1}^s \prod_1^{m-2} X_0^{q,w} & -C_{m-1}^s \prod_1^{m-2} X_0^{q,s} & -C_{m-1}^s \prod_2^{m-2} X_1^{q,s} & -C_{m-1}^s \prod_2^{m-2} X_1^{q,w} & -C_{m-1}^s \prod_2^{m-2} X_1^{q,s} & \cdots & & & & E_{m-1}^{q,s} & \\
 & & & u_\sigma D_0^{q,w} & & -E_0^{q,s} & & & & & & & & & \\
 & & & & & & & & & \ddots & & u_\sigma D_{m-1}^{q,w} & & -E_{m-1}^{q,s} & \\
 & & & -l_\sigma D_0^{q,w} & & E_0^{q,s} & & & & \ddots & & & & & \\
 & & & & & & & & & & & -l_\sigma D_{m-1}^{q,w} & & E_{m-1}^{q,s} & \\
 & C_0^s & & l_\sigma D_0^{q,w} & & -E_0^{q,s} & & & -E_1^{q,s} & & & & & & \\
 & C_1^s X_0^s & C_1^s X_0^u & C_1^s X_0^{u,w} & C_1^s X_0^{q,w} & C_1^s X_0^{q,s} & l_\sigma D_1^{q,w} & & & & & & & & \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & & & & & \\
 & C_{m-1}^s \prod_0^{m-2} & C_{m-1}^s \prod_1^{m-2} X_0^u & C_{m-1}^s \prod_1^{m-2} X_0^{u,w} & C_{m-1}^s \prod_1^{m-2} X_0^{q,w} & C_{m-1}^s \prod_1^{m-2} X_0^{q,s} & C_{m-1}^s \prod_2^{m-2} X_1^{q,s} & C_{m-1}^s \prod_2^{m-2} X_1^{q,w} & C_{m-1}^s \prod_2^{m-2} X_1^{q,s} & \cdots & & l_\sigma D_{m-1}^{q,w} & & -E_{m-1}^{q,s} & \\
 & -C_0^s & & -u_\sigma D_0^{q,w} & & E_0^{q,s} & & & & & & & & & \\
 & -C_1^s X_0^s & -C_1^s X_0^u & -C_1^s X_0^{u,w} & -C_1^s X_0^{q,w} & -C_1^s X_0^{q,s} & -u_\sigma D_1^{q,w} & & E_1^{q,s} & & & & & & \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & & & & & \\
 & -C_{m-1}^s \prod_0^{m-2} & -C_{m-1}^s \prod_1^{m-2} X_0^u & -C_{m-1}^s \prod_1^{m-2} X_0^{u,w} & -C_{m-1}^s \prod_1^{m-2} X_0^{q,w} & -C_{m-1}^s \prod_1^{m-2} X_0^{q,s} & -C_{m-1}^s \prod_2^{m-2} X_1^{q,s} & -C_{m-1}^s \prod_2^{m-2} X_1^{q,w} & -C_{m-1}^s \prod_2^{m-2} X_1^{q,s} & \cdots & & -u_\sigma D_{m-1}^{q,w} & & E_{m-1}^{q,s} & \\
\hline
 & R_0 + R_m \prod_0^{m-1} & R_m \prod_1^{m-1} X_0^u & R_m \prod_1^{m-1} X_0^{u,w} & R_m \prod_1^{m-1} X_0^{q,w} & R_m \prod_1^{m-1} X_0^{q,s} & R_m \prod_2^{m-1} X_1^u & R_m \prod_2^{m-1} X_1^{q,w} & R_m \prod_2^{m-1} X_1^{q,s} & \cdots & R_m X_{m-1}^u & R_m X_{m-1}^{q,w} & R_m X_{m-1}^{q,s} & R_m X_{m-1}^{q,s} & \\
\end{array}\right)
$$

where sensitivity matrix products $\prod_j^k := \prod_{l=j}^k X_l^s, 0 \leq j \leq k \leq m-1$, and $\prod_j^k := I$ for $j > k$.

4.1.6 Feedback Algorithm: Block Structure of QP Subproblem

We consider a simple (special) case that the QP subproblem (4.23) is considered using only one interval $I_i = [t_i, t_{i+1}]$, and the objective function is just in Mayer type, i.e., $\varphi(y) = \varphi(s_{i+1})$. We also assume that $l_\sigma < u_\sigma$, where $\sigma(s_i) \in [l_\sigma, u_\sigma]$. Then we can rewrite (4.25) as

$$
\begin{aligned}
\min_{\Delta y} \quad & \nabla_y \varphi(s_{i+1})^T \Delta y + \tfrac{1}{2} \Delta y^T \nabla_y^2 L(y_i, \eta) \Delta y \\
\text{s.t.} \quad & 0 = h_i(y_i) + X_i \Delta y_i - \Delta s_{i+1}, \\
& 0 \le q_i^z - \sigma(s_i) - C_i^s \Delta s_i + E_i^{q^z} \Delta q_i^z, \\
& 0 \le u_\sigma q_i^{\alpha_\sigma} - q_i^z + u_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} - E_i^{q^z} \Delta q_i^z, \\
& 0 \le q_i^z - l_\sigma q_i^{\alpha_\sigma} - l_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + E_i^{q^z} \Delta q_i^z, \\
& 0 \le \sigma(s_i) + l_\sigma q_i^{\alpha_\sigma} - q_i^z - l_\sigma + C_i^s \Delta s_i + l_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} - E_i^{q^z} \Delta q_i^z, \\
& 0 \le u_\sigma - \sigma(s_i) - u_\sigma q_i^{\alpha_\sigma} + q_i^z - C_i^s \Delta s_i - u_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + E_i^{q^z} \Delta q_i^z, \\
& 0 \le r(s_i, s_{i+1}) + R_i^T \Delta s_i + R_{i+1}^T \Delta s_{i+1}, \\
& 0 \le q_i^z + \Delta q_i^z, \\
& -q_i^{\alpha_\sigma} \le \Delta q_i^{\alpha_\sigma} \le 1 - q_i^{\alpha_\sigma}, \\
& -q_i^{\alpha_w} \le \Delta q_i^{\alpha_w} \le 1 - q_i^{\alpha_w}, \ \ \textstyle\sum_{w \in \mathcal{W}} \Delta q_i^{\alpha_w} = 0,
\end{aligned}
\tag{4.32}
$$

where $l_\sigma \le \sigma(s_i) \le u_\sigma$, $0 \le q_i^{\alpha_\sigma} \le 1$, $0 \le q_i^{\alpha_w} \le 1$, $\sum_{w \in \mathcal{W}} q_i^{\alpha_w} = 1$, and $q_i^z \ge 0$, $q_i^z - \sigma(s_i) \ge 0$, and

$$\Delta y = (\Delta s_i, \Delta q_i^u, \Delta q_i^{\alpha_\sigma}, \Delta q_i^{\alpha_w}, \Delta q_i^z, \Delta s_{i+1}).$$

Hereafter, $\Delta u = (\Delta s_i, \Delta q_i^u, \Delta q_i^{\alpha_\sigma}, \Delta q_i^{\alpha_w}, \Delta q_i^z)$, $\Delta v = \Delta s_{i+1}$, R_i, R_{i+1} and other derivatives are denoted as in (4.27), the dense constraint matrix in pp. 70 is

$$
\left(\begin{array}{ccccc|c}
X_i^s & X_i^{q^u} & X_i^{q^{\alpha_\sigma}} & X_i^{q^{\alpha_w}} & X_i^{q^z} & -I \\
\hline
-C_i^s & & & & E_i^{q^z} & \\
 & & u_\sigma D_i^{q^{\alpha_\sigma}} & & -E_i^{q^z} & \\
 & & -l_\sigma D_i^{q^{\alpha_\sigma}} & & E_i^{q^z} & \\
C_i^s & & l_\sigma D_i^{q^{\alpha_\sigma}} & & -E_i^{q^z} & \\
-C_i^s & & -u_\sigma D_i^{q^{\alpha_\sigma}} & & E_i^{q^z} & \\
R_i + R_{i+1} X_i^s & R_{i+1} X_i^{q^u} & R_{i+1} X_i^{q^{\alpha_\sigma}} & R_{i+1} X_i^{q^{\alpha_w}} & R_{i+1} X_i^{q^z} &
\end{array}\right)
$$

where the blanks mean zero blocks. The QP (4.32) reads

$$
\begin{aligned}
\min_{\Delta y} \quad & g_i^T \Delta y + \tfrac{1}{2} \Delta y^T H_i \Delta y \\
\text{s.t.} \quad & -\tilde{r} \le \tilde{R} \Delta y,
\end{aligned}
\tag{4.33}
$$

therein,

$$
g_i^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \frac{\partial \varphi(s_{i+1})}{\partial s_{i+1}} \end{pmatrix}, \tag{4.34}
$$

$$H_i = \begin{pmatrix} \frac{\partial^2 L(y_i,\eta)}{\partial s_i^2} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^u \partial s_i} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^{\alpha_\sigma} \partial s_i} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^{\alpha_w} \partial s_i} & 0 & 0 \\ \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^u \partial s_i} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{(\partial q_i^u)^2} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^u \partial q_i^{\alpha_\sigma}} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^u \partial q_i^{\alpha_w}} & 0 & 0 \\ \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^{\alpha_\sigma} \partial s_i} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^u \partial q_i^{\alpha_\sigma}} & 0 & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^{\alpha_\sigma} \partial q_i^{\alpha_w}} & 0 & 0 \\ \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^{\alpha_w} \partial s_i} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^u \partial q^{\alpha_w}} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{\partial q_i^{\alpha_w} \partial q_i^{\alpha_\sigma}} & \eta_{1,i}^T \frac{\partial x_i(\cdot)}{(\partial q_i^{\alpha_w})^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\partial^2 L(y_i,\eta)}{\partial s_{i+1}^2} \end{pmatrix} \tag{4.35}$$

with $x_i(\cdot) := x_i(t_{i+1}; t_i, y_i)$, $\frac{\partial^2 L(y_i,\eta)}{\partial s_{i+1}^2} = \frac{\partial^2 \varphi(s_{i+1})}{\partial s_{i+1}^2} + \eta_2^T \frac{\partial^2 r(s_i,s_{i+1})}{\partial s_{i+1}^2}$, $\frac{\partial^2 L(y_i,\eta)}{\partial s_i^2} = \frac{\partial^2 x_i(\cdot)}{\partial s_i^2} + (\eta_{4,i} - \eta_{7,i} + \eta_{8,i}) \frac{\partial^2 \sigma(s_i)}{\partial s_i^2}$, and

$$\tilde{R}\Delta y = \begin{pmatrix} -C_i^s \Delta s_i + E_i^{q^z} \Delta q_i^z \\ u_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} - E_i^{q^z} \Delta q_0^z \\ -l_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + E_i^{q^z} \Delta q_i^z \\ C_i^s \Delta s_i + l_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} - E_i^{q^z} \Delta q_i^z \\ -C_i^s \Delta s_i - u_\sigma D_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + E_i^{q^z} \Delta q_i^z \\ \left(\tilde{R}\Delta y\right)_6 \end{pmatrix}, \ \tilde{r} = \begin{pmatrix} q_i^z - \sigma(s_i) \\ u_\sigma q_i^{\alpha_\sigma} - q_i^z \\ -l_\sigma q_i^{\alpha_\sigma} + q_i^z \\ \sigma(s_i) + l_\sigma q_i^{\alpha_\sigma} - q_i^z - l_\sigma \\ u_\sigma - \sigma(s_i) - u_\sigma q_i^{\alpha_\sigma} + q_i^z \\ r(s_i, s_{i+1}) \end{pmatrix}$$

where $\left(\tilde{R}\Delta y\right)_6 := R_i + R_{i+1} X_i^s \Delta s_i + R_{i+1} X_i^{q^u} \Delta q_i^u + R_{i+1} X_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + R_{i+1} X_i^{q^{\alpha_w}} \Delta q_i^{\alpha_w} + R_{i+1} X_i^{q^z} \Delta q_i^z$.

To imply the solution of (4.33), we employ the KKT conditions. The necessary conditions read:

N1. Stationarity:

$$\begin{aligned} &\nabla_{\Delta y} \left[\left(g_i^T \Delta y + \frac{1}{2} \Delta y^T H_i \Delta y \right) - \mu^T \left(\tilde{R}\Delta y + \tilde{r} \right) \right] = 0 \\ &\Leftrightarrow H_i \Delta y = \tilde{R}^T \mu - g_i, \end{aligned} \tag{4.36}$$

where μ denotes the Lagrange multipliers in QP (4.33).

N2. Primal feasibility:

$$\tilde{R}\Delta y + \tilde{r} \geq 0. \tag{4.37}$$

N3. Dual feasibility:

$$\mu \geq 0. \tag{4.38}$$

N4. Complementary:

$$\mu^T (\tilde{R}\Delta y + \tilde{r}) = 0. \tag{4.39}$$

Since Lemma 7, we only consider the case that $\frac{\partial\sigma(s_i)}{\partial s_i} \neq 0$, i.e., $C_i^s \neq 0$. Then, from the conditions (N1-N4) and remember that $D_i^{q^{\alpha_\sigma}} = E_i^{q^z} = 1$, one obtains the following constraints

$$
\begin{aligned}
&- C_i^s \Delta s_i + \Delta q_i^z \geq -q_i^z + \sigma(s_i), && (4.40)\\
&u_\sigma \Delta q_i^{\alpha_\sigma} - \Delta q_i^z \geq -u_\sigma q_i^{\alpha_\sigma} + q_i^z, && (4.41)\\
&l_\sigma \Delta q_i^{\alpha_\sigma} - \Delta q_i^z \geq -l_\sigma q_i^{\alpha_\sigma} + q_i^z, && (4.42)\\
&C_i^s \Delta s_i + l_\sigma \Delta q_i^{\alpha_\sigma} - \Delta q_i^z \geq l_\sigma - \sigma(s_i) - l_\sigma q_i^{\alpha_\sigma} + q_i^z, && (4.43)\\
&- C_i^s \Delta s_i - u_\sigma \Delta q_i^{\alpha_\sigma} + \Delta q_i^z \geq \sigma(s_i) + u_\sigma q_i^{\alpha_\sigma} - q_i^z - u_\sigma, && (4.44)\\
&R_{i+1} \Delta s_{i+1} \geq r(s_i, s_{i+1}) - R_i \Delta s_i, && (4.45)\\
&\Delta s_{i+1} = X_i^s \Delta s_i + X_i^{q^u} \Delta q_i^u + X_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + X_i^{q^{\alpha_w}} \Delta q_i^{\alpha_w} + X_i^{q^z} \Delta q_i^z, && (4.46)\\
&\Delta q_i^z \geq -q_i^z. && (4.47)
\end{aligned}
$$

Remark 24. At each t_i, $i \in \{0, \ldots, m-1\}$, one can directly study the derivative of the switched function w.r.t the discretized state via the increment of the initial shooting state and the switched function's value, i.e.,

$$
\begin{aligned}
&\frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i \leq -\sigma(s_i), \quad \text{if } \sigma(s_i) \leq 0, && (4.48)\\
&\frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i > -\sigma(s_i), \quad \text{if } \sigma(s_i) > 0. && (4.49)
\end{aligned}
$$

Lemma 9. *Consider NLP* (4.15) *and corresponding QP* (4.32) *using one interval* $[t_i, t_{i+1}]$. *The following equalities are hold true:*

$$
\begin{aligned}
&q_i^{\alpha_\sigma} = 1,\ q_i^z = \sigma(s_i),\ \Delta q_i^{\alpha_\sigma} = 0,\ \Delta q_i^z = \frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i, && \text{if } \sigma(s_i) > 0, && (4.50)\\
&\Delta q_i^{\alpha_\sigma} = g_q(\Delta s_i, \Delta s_{i+1}),\ q_i^z = \Delta q_i^z = 0, && \text{if } \sigma(s_i) = 0, && (4.51)\\
&q_i^{\alpha_\sigma} = 0,\ q_i^z = 0,\ \Delta q_i^{\alpha_\sigma} = \Delta q_i^z = 0, && \text{if } \sigma(s_i) < 0. && (4.52)
\end{aligned}
$$

Proof. For $\sigma(s_i) > 0$, one has $q_i^{\alpha_\sigma} = 1$, $q_i^z = \sigma(s_i)$, and constraints $q_i^z - \sigma(s_i) \geq 0$, $u_\sigma(q_i^{\alpha_\sigma} - 1) \leq q_i^z - \sigma(s_i) \leq l_\sigma(q_i^{\alpha_\sigma} - 1)$ of the NLP are active, the remain constraints are inactive. By linearizing these active constraints, one obtains $\Delta q_i^{\alpha_\sigma} = 0$ and $\Delta q_i^z = q_i^\alpha \frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i = \frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i$. Proving equalities (4.50) is done.
Similarly, proving equalities (4.52) is completed by considering the case $\sigma(s_i) < 0$, therein $\Delta q_i^z = q_i^\alpha \frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i = 0$.
For the case $\sigma(s_i) = 0$, one has $q_i^z = 0$, only constraints $q_i^z \geq 0$ and $q_i^z - \sigma(s_i) \geq 0$ are active. Linearization these active constraints yields $\Delta q_i^z = 0$ and $\frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i = -\sigma(s_i) = 0$. The the necessary conditions (N1-N4) take form as follows

$$
\begin{aligned}
&H_i \Delta y - \tilde{R}_{\text{active}} \mu_{\text{active}} = g_i,\\
&\Delta s_{i+1} = X_i^s \Delta s_i + X_i^{q^u} \Delta q_i^u + X_i^{q^{\alpha_\sigma}} \Delta q_i^{\alpha_\sigma} + X_i^{q^{\alpha_w}} \Delta q_i^{\alpha_w} + X_i^{q^z} \Delta q_i^z + h_i,\\
&R_{i_{\text{active}}} \Delta s_i + R_{i+1_{\text{active}}} \Delta s_{i+1} = r(s_i, s_{i+1})_{\text{active}},\\
&\frac{\partial\sigma(s_i)}{\partial s_i} \Delta s_i = 0.
\end{aligned} \qquad (4.53)
$$

Solving (4.53) one can obtain $q_i^{\alpha_\sigma}$ and $\Delta q_i^{\alpha_\sigma}$, i.e., $\Delta q_i^{\alpha_\sigma} = g_q(\Delta s_i, \Delta s_{i+1})$. Proving the equalities (4.51) is done. □

Remark 25. The results in Lemma 9 show that there are no round-off procedures are necessary to recover integer-valued control when $\sigma(s_i) \neq 0$. On the other hand, sliding mode for control α is happened when $\sigma(s_i) = 0$.

Remark 26. The reformulations in Chapter 3, see Subsection 3.1.4 and 3.2.1, return poor numerical results due to the quadratic terms in the mixed constraints when applying for our direct approach. Therefore, we have suggested the linearized reformulation (4.11-4.12).

4.1.7 An Active Set Method for SwOCP

Constraint Qualifications

We consider a discretized SwOCP with vanishing constraints in the following form

$$\begin{aligned} \min_{s_i,q_i} \quad & \varphi(\cdot) \\ \text{s.t.} \quad & 0 = x_i(t_{i+1}; t_i, s_i, q_i) - s_{i+1}, \quad i = 0, \dots, m-1, \\ & 0 \leq r(s_0, s_m), \\ & \sigma(s_i) q_i \geq 0, \quad i = 0, \dots, m-1, \\ & \sigma(s_i)(q_i - 1) \geq 0, \quad i = 0, \dots, m-1, \\ & 0 \leq q_i \leq 1, \quad i = 0, \dots, m-1, \end{aligned} \tag{4.54}$$

where $\sigma(\cdot)$ is the switching function. We then analyze the vanishing constraints by dropping matching conditions and terminal condition from problem (4.54), i.e., we consider a simple discretized problem with vanishing constraints as follows,

$$\begin{aligned} \min_{s_i,q_i} \quad & \varphi(\cdot) \\ \text{s.t.} \quad & \sigma(s_i) q_i \geq 0, \quad i = 0, \dots, m-1, \\ & \sigma(s_i)(q_i - 1) \geq 0, \quad i = 0, \dots, m-1, \\ & 0 \leq q_i \leq 1, \quad i = 0, \dots, m-1. \end{aligned} \tag{4.55}$$

For a feasible point $(\bar{s}_j, \bar{q}_j) \in \mathbb{R}^{n_x} \times [0,1]$, $j \in \{0, \dots, m-1\} =: \mathcal{J}$, we define the active sets

$$\mathcal{A}_1(\bar{s}, \bar{q}) := \{j \in \mathcal{J} \mid \sigma(\bar{s}_j)\bar{q}_j = 0\}, \tag{4.56}$$

$$\mathcal{A}_2(\bar{s}, \bar{q}) := \{j \in \mathcal{J} \mid \sigma(\bar{s}_j)(\bar{q}_j - 1) = 0\}, \tag{4.57}$$

We then introduce the index sets

$$\begin{aligned} \mathcal{I}_{+1} &:= \mathcal{I}_{+1}(\bar{s}, \bar{q}) = \{j \in \mathcal{J} \mid \sigma(\bar{s}_j) > 0, \bar{q}_j = 1\}, \\ \mathcal{I}_{01} &:= \mathcal{I}_{01}(\bar{s}, \bar{q}) = \{j \in \mathcal{J} \mid \sigma(\bar{s}_j) = 0, \bar{q}_j = 1\}, \\ \mathcal{I}_{00} &:= \mathcal{I}_{00}(\bar{s}, \bar{q}) = \{j \in \mathcal{J} \mid \sigma(\bar{s}_j) = 0, \bar{q}_j = 0\}, \\ \mathcal{I}_{-0} &:= \mathcal{I}_{-0}(\bar{s}, \bar{q}) = \{j \in \mathcal{J} \mid \sigma(\bar{s}_j) < 0, \bar{q}_j = 0\}, \\ \mathcal{I}_{0+} &:= \mathcal{I}_{0+}(\bar{s}, \bar{q}) = \{j \in \mathcal{J} \mid \sigma(\bar{s}_j) = 0, 0 < \bar{q}_j < 1\}, \end{aligned} \tag{4.58}$$

which partition the set of active constraints according to signs of $\sigma(\cdot, \bar{s}_j)$ and $\bar{q}_j$,

$$\mathcal{A}_1(\bar{s}, \bar{q}) = \mathcal{I}_{01} \cup \mathcal{I}_{00} \cup \mathcal{I}_{-0} \cup \mathcal{I}_{0+}, \tag{4.59}$$

$$\mathcal{A}_1^C(\bar{s}, \bar{q}) := \mathcal{J} \setminus \mathcal{A}_1(\bar{s}, \bar{q}) = \mathcal{I}_{+1}, \tag{4.60}$$

$$\mathcal{A}_2(\bar{s}, \bar{q}) = \mathcal{I}_{+1} \cup \mathcal{I}_{01} \cup \mathcal{I}_{00} \cup \mathcal{I}_{0+}, \tag{4.61}$$

$$\mathcal{A}_2^C(\bar{s}, \bar{q}) := \mathcal{J} \setminus \mathcal{A}_2(\bar{s}, \bar{q}) = \mathcal{I}_{-0}. \tag{4.62}$$

Remark 27. If $\mathcal{I}_{00} = \emptyset$, $\mathcal{I}_{01} = \emptyset$ or $\mathcal{I}_{0+} = \emptyset$ then in the neighborhood of $(\bar{s}, \bar{q})$ problem (4.54) is a standard NLP including those constraints $\sigma(s_j) \leq 0$ for $j \in \mathcal{A}_2^C = \mathcal{I}_{-0}$, $\sigma(s_j) \geq 0$ for $j \in \mathcal{A}_1^C = \mathcal{I}_{+1}$, or $\sigma(s_j) = 0$ for $j \in \mathcal{I}_{0+}$ respectively. This condition refers to *Lower Level Strict Complementarity Condition*(LLSCC).
Otherwise, if $\mathcal{I}_{00} \cup \mathcal{I}_{01} \cup \mathcal{I}_{0+} \neq \emptyset$, i.e., LLSCC does not hold, then in a neighborhood of $(\bar{s}, \bar{q})$ the feasible set has combinatorial structure. Both LICQ and MFCQ are violated, which causes significant difficulties to KKT based descent methods.

Modified Stationarity Concept

In view of the practical difficulties (such as unbounded dual variables, ill-conditioned constraint Jacobian, cycling and stalling of active set methods, suboptimal and infeasible steps), a modified concept of optimality under a possibly weaker constraint qualification is desirable. This CQ should ensure that stationarity points of (4.54) are indeed KKT points to hold the concept of iterating towards KKT based optimality.
A Regularity Assumption To achieve this goal, we introduce the regularity assumption of MPVC-LICQ, cf. [1].

Definition 32. We say that MPVC-LICQ holds for a feasible point $(\bar{s}_j, \bar{q}_j) \in \mathbb{R}^{n_x} \times [0, 1]$, $j \in \mathcal{J}$, if

$$\begin{aligned} &\begin{pmatrix} \frac{\partial \sigma(\bar{s}_j)}{\partial s_j} & 1 \end{pmatrix}^T, && j \in \mathcal{I}_{0+}, \\ &\begin{pmatrix} \frac{\partial \sigma(\bar{s}_j)}{\partial s_j} & 1 \end{pmatrix}^T, \begin{pmatrix} 0 & 1 \end{pmatrix}^T, && j \in \mathcal{I}_{01} \cup \mathcal{I}_{00}, \end{aligned} \tag{4.63}$$

are linearly independent, i.e., the MPVC-LICQ holds if $\frac{\partial \sigma(\bar{s}_j)}{\partial s_j} \neq 0$, $j \in \mathcal{I}_{01} \cup \mathcal{I}_{00} \cup \mathcal{I}_{0+}$.

Strong Stationarity Conditions Under MPVC-LICQ, a KKT-like necessary condition for local optimality of a candidate point $(\bar{s}_j, \bar{q}_j)$, $j \in \mathcal{J}$, of problem (4.54) can be given. It is based on the so-called MPVC-Lagrangian $\mathcal{L}(s, q, \lambda, \mu_1, \mu_2, \mu_r)$ of problem (4.54),

$$\mathcal{L}(s_j, q_j, \lambda, \mu_1, \mu_2, \mu_r) := \varphi(\cdot) - \lambda^T(x_i - s_{i+1}) - \mu_1^T \sigma(s_j) q_j - \mu_2^T \sigma(s_j)(q_j - 1) - \mu_r r, \tag{4.64}$$

where $j \in \mathcal{J}$, and $\lambda, \mu_1, \mu_2 \in \mathbb{R}^{n_x+1}$, $\mu_r \in \mathbb{R}^{n_r}$ are referred to as MPVC multipliers. The notion of strong stationarity for MPVC has been defined in [68] as follows:

Definition 33. A feasible point $(\bar{s}_j, \bar{q}_j) \in \mathbb{R}^{n_x} \times [0, 1]$, $j \in \mathcal{J}$, is called MPVC *strongly stationary* if there exist MPVC multiplier $\lambda, \mu_1, \mu_2 \in \mathbb{R}^{n_x+1}$, $\mu_r \in \mathbb{R}^{n_r}$ such as that it holds that

$$\begin{aligned} &\mathcal{L}_{s_j}(\bar{s}_j, \bar{q}_j, \lambda, \mu_1, \mu_2, \mu_r) = 0,\ \mathcal{L}_{q_j}(\bar{s}_j, \bar{q}_j, \lambda, \mu_1, \mu_2, \mu_r) = 0, \quad \mu_r \geq 0, \\ &\mu_{1,j} \geq 0,\ j \in \mathcal{I}_{01} \cup \mathcal{I}_{00} \cup \mathcal{I}_{-0} \cup \mathcal{I}_{0+}, \quad \mu_{1,j} = 0,\ j \in \mathcal{I}_{+1}, \\ &\mu_{2,j} \geq 0,\ j \in \mathcal{I}_{+1} \cup \mathcal{I}_{01} \cup \mathcal{I}_{00} \cup \mathcal{I}_{0+}, \quad \mu_{2,j} = 0,\ j \in \mathcal{I}_{-0}. \end{aligned} \tag{4.65}$$

In [1] it has been shown that under MPVC-LICQ strong stationarity (4.65) for MPVC is equivalent to KKT stationarity for problem (4.54). The following stronger result is due to [69] and can be found in [72, 77].

Theorem 14. [77, thm 1] *Let $(\bar{s}_j, \bar{q}_j) \in \mathbb{R}^{n_x} \times [0,1]$, $j \in \mathcal{J}$, satisfy MPVC-LICQ. If $(\bar{s}, \bar{q})$ is a locally optimal point of problem (4.54) then $(\bar{s}, \bar{q})$ is an MPVC strongly stationary point. The associated MPVC multipliers $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_r)$ are unique.*

Convex Quadratic Programs with Vanishing Constraints

In the proposed SQP framework for vanishing constraint problems, the subproblems resulting from a local quadratic model of the MPVC-Lagrangian are convex quadratic programs with affine linear vanishing constraints, see [77, Eq. (12a-b)], as follows,

$$\begin{aligned} &\left(\frac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j + \sigma(s_j)\right)(q_j + \Delta q_j) \geq 0, \quad j \in \mathcal{J}, \\ &\left(\frac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j + \sigma(s_j)\right)(q_j + \Delta q_j - 1) \geq 0, \quad j \in \mathcal{J}, \end{aligned} \tag{4.66}$$

Problem (4.54) is then leading in the following QPVC

$$\begin{aligned} \min_{\Delta s, \Delta q} \quad & \tfrac{1}{2}(\Delta\eta)^T H \Delta\eta + \Delta\eta^T b \\ \text{s.t.} \quad & \tfrac{\partial x_j}{\partial s_j}\Delta s_j + \tfrac{\partial x_j}{\partial q_j}\Delta q_j + x_j(t_{j+1}; t_j, s_j, q_j) - s_{j+1} = 0, \quad j \in \mathcal{J}, \\ & \tfrac{\partial r}{\partial s_0}\Delta s_0 + r(s_0, s_m) \geq 0, \quad \tfrac{\partial r}{\partial s_m}\Delta s_m + r(s_0, s_m) \geq 0, \\ & \left(\tfrac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j + \sigma(s_j)\right)(q_j + \Delta q_j) \geq 0, \quad j \in \mathcal{J}, \\ & \left(\tfrac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j + \sigma(s_j)\right)(q_j + \Delta q_j - 1) \geq 0, \quad j \in \mathcal{J}, \\ & 0 \leq q_j + \Delta q_j \leq 1, \quad j \in \mathcal{J}, \end{aligned} \tag{4.67}$$

where $\Delta\eta := \begin{pmatrix} \Delta s & \Delta q \end{pmatrix}^T$.

Convex Quadratic Programs on Subsets with Partitioning QPVC Subproblems

In the neighborhood of a feasible point $\Delta\bar{\eta}_j = (\Delta\bar{s}_j, \Delta\bar{q}_j) \in \mathbb{R}^{n_x} \times [0,1]$, $j \in \mathcal{J}$, of the QPVC (4.67) we consider the following convex QP with smaller but convex feasible set

$$\begin{aligned} \min_{\Delta s, \Delta q} \quad & \tfrac{1}{2}(\Delta\eta)^T H \Delta\eta + (\Delta\eta)^T b \\ \text{s.t.} \quad & \tfrac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j \geq -\sigma(s_j), \quad j \in \mathcal{I}_{01} \cup \mathcal{I}_{+1}, \\ & -\tfrac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j \geq \sigma(s_j), \quad j \in \mathcal{I}_{00} \cup \mathcal{I}_{-0}, \\ & \tfrac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j = -\sigma(s_j), \quad j \in \mathcal{I}_{0+}, \\ & 0 \leq q_j + \Delta q_j \leq 1, \quad j \in \mathcal{J}, \\ & \Delta q_j = 0, \quad j \in \mathcal{I}_{01} \cup \mathcal{I}_{+1} \cup \mathcal{I}_{00} \cup \mathcal{I}_{-0}, \end{aligned} \tag{4.68}$$

where we assume problem (4.68) has a positive definite Hessian $\frac{\partial^2 \varphi}{\partial \eta^2} =: H \in \mathbb{R}^{(n_x+1)\times(n_x+1)}$ of the MPVC Lagrangian, $\frac{\partial \varphi}{\partial \eta} =: b \in \mathbb{R}^{n_x+1}$ denotes the gradient vector.
Based on KKT optimality for every solution $\Delta\eta^* = (\Delta s^*, \Delta q^*)$ of problem (4.68) there exists

a unique vector of MPVC multipliers $\lambda_0, \lambda_1, \mu_0^s, \mu_1^s, \mu \in \mathbb{R}^{n_x+1}$ such that the following system of optimality conditions for subproblem (4.68) is satisfied,

$$\begin{aligned}
0 &= H\begin{pmatrix}\Delta s^*\\ \Delta q^*\end{pmatrix} + \begin{pmatrix}\frac{\partial \varphi}{\partial s_j}\\ \frac{\partial \varphi}{\partial q_j}\end{pmatrix} + \begin{pmatrix}((\lambda_1^*)^T + (\mu^*)^T)\left(\frac{\partial^2 \sigma}{\partial s_j^2}\Delta s_j + \frac{\partial \sigma}{\partial s_j}\right)\\ \mu_0^{s^*} - \mu_1^{s^*}\end{pmatrix}, && j \in \mathcal{J}\setminus\mathcal{I}_{0+},\\
0 &\le \frac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j^* + \sigma(s_j), && j \in \mathcal{I}_{01}\cup\mathcal{I}_{+1},\\
0 &\le -\frac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j^* - \sigma(s_j), && j \in \mathcal{I}_{00}\cup\mathcal{I}_{-0},\\
0 &= \Delta q_j^*, && j \in \mathcal{J}\setminus\mathcal{I}_{0+}, \quad (4.69)\\
0 &= \left(\frac{\partial \sigma(s_j)}{\partial s_j}\Delta s_j^* + \sigma(s_j)\right)\mu_j^*, \ \ \mu_j^* \ge 0, && j \in \mathcal{J}\setminus\mathcal{I}_{0+},\\
0 &= (q_j^* + \Delta q_j^*)\mu_{0,j}^{s^*}, \ \ \mu_{0,j}^{s^*} \ge 0, && j \in \mathcal{J},\\
0 &= (1 - q_j^* - \Delta q_j^*)\mu_{1,j}^{s^*}, \ \ \mu_{1,j}^{s^*} \ge 0, && j \in \mathcal{J},
\end{aligned}$$

where the Lagrangian function of problem (4.68) is $\mathcal{L}_j := \frac{1}{2}(\Delta\eta)^T H\Delta\eta + (\Delta\eta)^T b + \lambda_0^T\Delta q_j + \lambda_1^T\left(\frac{\partial\sigma(s_j)}{\partial s_j}\Delta s_j + \sigma(s_j)\right) + \mu^T\left(\frac{\partial\sigma(s_j)}{\partial s_j}\Delta s_j + \sigma(s_j)\right) + (\mu_0^s)^T(q_j + \Delta q_j) + (\mu_1^s)^T(1 - q_j - \Delta q_j)$. If we let $\mu_j = 0$ for $j \in \mathcal{I}_{0+}$ then those vanishing constraints that have vanished in problem (4.68). Since H is positive definite the solution $\eta^* = (s^*, q^*)$ is unique and it is a global solution of (4.68). We obtain similar result as in [77, thm 4] as follows,

Theorem 15. [77, thm 4] *Let $(s^*, q^*, \lambda_0^*, \lambda_1^*, \mu_0^{s^*}, \mu_1^{s^*}, \mu^*)$ be a KKT point of the subset QP associated with $\mathcal{I}_{0+}$. Then this is MPVC strongly stationary if and only if $\mu_j^* = 0$ for all $j \in \mathcal{I}_{00}\cup\mathcal{I}_{01}$.*

Proof. Similar to the proof of [77, thm 4] by replacing appropriate index sets. □

Remark 28. Similar optimality conditions are obtained by the active set method (see Equations (4.69)) and the feedback algorithm (see Lemma 9 and Remark 24).

4.2 A Switching Aware Rounding Algorithm

We consider OCP (3.61) as follows

$$\begin{aligned}
&\min_{x(\cdot),u(\cdot),w(\cdot)} \quad m(x(t_f)) + \textstyle\int_{t_0}^{t_f} l(x(t),u(t))dt\\
&\text{s.t.} \qquad \dot{x}(t) = \begin{cases} f_+\left(x(t),u(t),w(t)\right), & \text{if } \sigma(x(t)) \ge 0,\\ f_-\left(x(t),u(t),w(t)\right), & \text{if } \sigma(x(t)) \le 0,\end{cases} \qquad t \in \mathcal{T}.\\
&\qquad\quad 0 \le r\left(x(t_0), x(t_f)\right),\\
&\qquad\quad w(t) \in \mathcal{W},
\end{aligned}$$

This problem can be solved by alternative approach as follows:

1. Reformulate the problem by uses the partial outer convexification.
2. Solve a continuous relaxation of the MIOCPs.

3. Compute a rounding on some discretization grid to obtain a discrete-valued control trajectory from the continuously-valued one.

The first and second steps allow us to reformulate the switched problems as problem (4.1) and solve them effectively. Then, one can use an appropriate rounding scheme (see Subsection 2.6.7), or the family of CIA algorithm (see [114]), to return the integer values. For the last step, inspired by [12, Alg. III], we will develop a rounding scheme, namely Switching Aware Rounding (SAR).

4.2.1 Switching Aware Rounding

We start with the following proposition.

Proposition 2. [12, Prop. 2.1] *Let $M := \{-1, 0, +1\}$ and $f_i : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous for all $i \in M$. Let $\alpha \in L^\infty((t_0, t_f), \mathbb{R}^M)$ be given and $(\beta^{(h)})_h \subset L^\infty((t_0, t_f), \mathbb{R}^M)$ satisfy the convergence property*

$$\sup_{t\in(t_0,t_f)} \left\| \int_{t_0}^{t} (\alpha(s) - \beta^{(h)}(s))ds \right\|_\infty \to 0, \tag{4.70}$$

for $h \to 0$. Then,

$$(x, u)^{(h)} \to (x, u),$$

if (x, u) denotes the solution of the IVP in (4.1) and the $(x, u)^{(h)}$ denote the solutions of the IVPs in (3.61) *with ODE (3.62).*

Definition 34. [12, Def. 2.3] Let α satisfy the last two constraints of (4.1). Let $t_0 < \ldots < t_N$ be a grid discretization (t_0, t_f) with $0 < t_k - t_{k-1} \leq h$ for all $k \in \{1, \ldots, N\}$. Then, the binary-valued step function

$$\beta : [t_0, t_f] \to \{0, 1\}^M$$

$$\beta_i(t) := \begin{cases} 1, & \text{if } i = i^*(k), \\ 0, & \text{else,} \end{cases} \qquad \text{for all } t \in [t_{k-1}, t_k).$$

is constructed iteratively for $1 \leq k \leq N$ by the following rule to determine the rounding index $i^*(k)$ for the interval $[t_{k-1}, t_k)$:

$$i^*(k) := \arg\max_{i\in M} \{\gamma_{k,i}, \},$$
$$\gamma_{k,i} := \int_{t_0}^{t_k} \alpha_i(t)dt - \int_{t_0}^{t_{k-1}} \beta_i(t)dt. \tag{4.71}$$

Algorithm (4.71) satisfies (4.70), which is stated in [12, Prop. 2.4].

Instead of minimizing the left side of (4.70), we can rewrite the left hand side as a constraint into an optimization problem.

A. Preparations

Let $t_0 < \ldots < t_N = t_f$ be a grid discretization (t_0, t_f) with maximum grid coarseness

$h := \max_{1\leq k\leq N}(t_k - t_{k-1})$ and let α satisfy the two mixed constraints of problem (4.1). We introduce the following variables and quantities

$$\alpha_k := \frac{1}{h_k}\int_{t_{k-1}}^{t_k} \alpha(t)dt \in [0,1]^M, \qquad h_k := t_k - t_{k-1},$$
$$\beta_k \in \{0,1\}^M, \qquad \epsilon_k \in \{0,1\}^M, \qquad \xi_k \in \{0,1\}^M,$$

for $k \in \{1,\dots,N\}$. Here, α_k denotes the value of α averaged over the k-th interval, β_k *is desired output of the rounding indicate which realization* i, $i \in M$, *of the derivative states is switched on in which interval*, $\epsilon_{k,i}$ will indicate a *switch on* of the i-th derivative state from interval $k-1$ to k and ξ_k *switch off* of the i-th derivative state. Then, we can reconstruct the function β from the β_k as $\beta = \sum_{k\in M}\chi_{[t_{k-1},t_k)}\beta_k$, where χ_A denotes the characteristic function for the set A.

B. The ILP for rounding

Now, we can state the switch aware rounding heuristic in the ILP *Switching Aware Rounding Problem* (4.72) as follow,

$$\begin{aligned}
&\min_{\beta_{k,i},\epsilon_{k,i},\xi_{k,i}} \quad \textstyle\sum_{i\in M} c_i\beta_{1,i} + \sum_{i\in M} d_i\beta_{N,i}\\
&\text{s.t. } \textstyle\sum_{i=-1}^{+1}\beta_{k,i} = 1, \forall k \in \{1,\dots,N\}\\
&\quad -Kh \leq \textstyle\sum_{l=1}^{h} h_l(\alpha_{l,i} - \beta_{l,i}) \leq Kh, \forall k \in \{1,\dots,N\}, i \in M,\\
&\quad \beta_{k+1,i} - \beta_{k,i} \leq \epsilon_{k,i}, \forall k \in \{1,\dots,N-1\}, i \in M,\\
&\quad \beta_{k,i} - \beta_{k+1,i} \leq \xi_{k,i}, \forall k \in \{1,\dots,N-1\}, i \in M,\\
&\quad \beta_{k,i}, \epsilon_{k,i}, \xi_{k,i} \in \{0,1\}, \forall i,k,\ M = \{-1,0,+1\}.
\end{aligned} \tag{4.72}$$

The following proposition guarantees that the convergence of the corresponding state vector sequences with the above-summarized theory.

Proposition 3. [12, Prop. 3.1] *Let* $K \geq 1$. *Let* $\alpha \in L^\infty((t_0,t_f),\mathbb{R}^M)$ *satisfy the last two constraints of (4.1). Let* $t_0 < \ldots < t_N = t_f$ *be a grid discretization* (t_0,t_f) *with* $h := \max_{1\leq k\leq N}(t_k - t_{k-1})$. *Then, (4.72) has a solution. Consider the function* $\beta^{(h)} := \sum_{k\in M}\chi_{[t_{k-1},t_k)}\beta_k^{(h)}$ *with the* $\beta_{k,i}^{(h)}$ *solving (4.72). Then,*

$$\sup_{t\in(t_0,t_f)}\left\|\int_{t_0}^{t}(\alpha(s)-\beta(s))ds\right\|_\infty \leq Kh.$$

In particular, (4.70) holds true.

C. Interpretation of (SARP)

Usually, the maximal frequency for switching is subject to some physical constraints, which will determine h. Thus, from the setup of (4.72), it is clear that the parameter governing the trade-off is K.

Note that for high values of K and N, we expect (4.72) to become prohibitively hard to compute as we assume it can be reduced to a weakly NP-hard problem.

Remark 29. After the rounding procedure, we get the integer controls, which help us to track exactly when the switches occur.

4.2.2 An Expansion of Rounding Scheme: Neighboring Feedback Law for the Switching Aware Rounding

In this subsection, the neighboring feedback law will be combined with the switch aware rounding heuristic to propose a new effective rounding scheme.
We start by exploiting Theorem 13, we obtain the following system

$$\begin{aligned}
&\dot{x}(t) = \alpha(t) f_+(x(t), u(t)) + (1-\alpha(t)) f_-(x(t), u(t)),\\
&\dot{\lambda}^T(t) = -\lambda^T(t)\left(\alpha(t)\frac{\partial f_+(x,u)}{\partial x} + (1-\alpha(t))\frac{\partial f_-(x,u)}{\partial x}\right) + \theta(t)\frac{\partial g(x)}{\partial x},\ t \in T_*,
\end{aligned} \tag{4.73}$$

with initial and end constraints

$$\begin{aligned}
&\lambda(t_f) = -v_{t_f}^T \frac{\partial r(\cdot, x(t_f))}{\partial x} - v_0 \frac{\partial m(x(t_f))}{\partial x},\ \ \lambda(t_0) = -v_{t_0}^T \frac{\partial r(x(t_0), \cdot)}{\partial x},\\
&r(x(t_0), x(t_f)) \geq 0,
\end{aligned} \tag{4.74}$$

and jump conditions

$$\lambda(t+0) = \lambda(t-0) + \vartheta(t)\frac{\partial \sigma(x^*(t))}{\partial x}, \quad t \in T_*, \tag{4.75}$$

therein we assume that there are no measure included in system (4.73), where $T_* = \{t \in [t_0, t_f] \mid \sigma(x(t)) = 0\}$, and $\theta(t)\frac{\partial g(x)}{\partial x} = \delta^T \frac{\partial l(x,u)}{\partial x} + \vartheta(t)\frac{\partial \sigma(x)}{\partial x}$. Then LMP gives us

$$\mathcal{H}(\hat{x}, \hat{u}, \hat{\alpha}, \lambda) = \max_{\alpha \in [0,1], u \in \mathcal{U}} \{\alpha \lambda^T (f_+(x,u) - f_-(x,u)) + \lambda^T f_-(x,u)\}, \tag{4.76}$$

which yields

$$(\hat{u}(t), \hat{\alpha}(t)) = \arg \max_{u \in \mathcal{U}, \alpha \in [0,1]} \mathcal{H}(\hat{x}, u, \alpha, \hat{\lambda}) = (\hat{u}(\hat{x}, \hat{\lambda}), \hat{\alpha}(\hat{x}, \hat{\lambda})), \tag{4.77}$$

where $\mathcal{H}(\cdot) = \lambda^T(\alpha f_+(x,u) + (1-\alpha) f_-(x,u))$. The transition points $\hat{t}_i$ are determined by switching functions, cf. [83],

$$\sigma_i(x(t), \lambda(t)) = 0. \tag{4.78}$$

The differential equations (4.73) together with switching condition (4.78) for the right hand side, initial and end conditions (4.74) for the state variable (x, λ), and jump conditions (4.75) yeild a MPBVP as follows

$$\begin{aligned}
&\dot{z}(t) = \mathcal{F}(z(t), \text{sgn } \sigma_i(z(t))),\\
&\mathcal{R}(z(t_0), \ldots, z(t_f)) = 0,
\end{aligned} \tag{4.79}$$

where $\mathcal{F} := (\partial/\partial x, -\partial/\partial \lambda)\mathcal{H}$, and $z := (x, \lambda)$.
Let $t_0 = \hat{t}_0 < \ldots < \hat{t}_m = t_f$ be a grid discretization, where $0 < \alpha(t) < 1$, $t \in (\hat{t}_0, \hat{t}_n)$, and on each subinterval $(\hat{t}_i, \hat{t}_{i+1})$, $i = 0, \ldots, m-1$, MPBVP (4.79) are solved by the multiple shooting technique to obtain the trajectory $z(\hat{t}_i)$. Denote

$$s_i = (s_i^x, s_i^\lambda) = (x(\hat{t}_i), \lambda(\hat{t}_i)),\ S := (s_0, \ldots, s_m), \tag{4.80}$$

and the typical multiple shooting equation

$$M(S) := \begin{pmatrix} \mathcal{R}(s_0, z(\hat{t}_1; S), \ldots, z(\hat{t}_f; S)) \\ z(\hat{t}_i; S) - s_i \ \ (i = 1, \ldots, m) \end{pmatrix} = \begin{pmatrix} h_0 \\ h_i \end{pmatrix} = 0. \tag{4.81}$$

At each iteration the following subproblem has to be solved

$$\begin{pmatrix} \mathcal{R}_0^s & \mathcal{R}_1^s & \ldots & \ldots & \mathcal{R}_{m-1}^s & \mathcal{R}_m^s \\ G_0 & -I & & & & \\ & G_1 & -I & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & G_{m-1} & -I \end{pmatrix} \begin{pmatrix} \Delta s_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \Delta s_m \end{pmatrix} = - \begin{pmatrix} h_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ h_m \end{pmatrix} \tag{4.82}$$

where the blanks mean zero blocks, and

$$\mathcal{R}_i^s = \frac{\partial \mathcal{R}}{\partial s_i}, \quad G_i = \frac{\partial z(\hat{t}_{i+1}; S)}{\partial s_i}, \ i = 0, \ldots, m-1. \tag{4.83}$$

Block-Gaussian elimination (4.82) is reduced to the condensed system, cf. [83, Eq. (1.20)],

$$\begin{aligned} &Z_0 \Delta s_0 =: \begin{pmatrix} I & 0 \\ Z_x^0 & Z_\lambda^0 \end{pmatrix} \Delta s_0 = -(u_0, \alpha_0), \\ &\Delta s_{i+1} = G_i \Delta s_i + h_{i+1}, \quad i = 0, \ldots, m-1, \end{aligned} \tag{4.84}$$

where $Z_0, (u_0, \alpha_0)$ are recursively determined by

$$\begin{aligned} &Z_m := \mathcal{R}_m, \ Z_i =: \begin{pmatrix} 0 & 0 \\ Z_x^i & Z_\lambda^i \end{pmatrix} = \mathcal{R}_i + Z_{i+1} G_i, \\ &(u_m, \alpha_m) := h_0, \ (u_i, \alpha_i) = (u_{i+1}, \alpha_{i+1}) + Z_{i+1} h_{i+1}, \end{aligned} \qquad i = m-1, \ldots, 0. \tag{4.85}$$

As a result of the iteration, a nominal trajectory $\hat{x}(t), \hat{\lambda}(t)$, and thus a nominal control $(\hat{u}(t), \hat{\alpha}(t)) = (\tilde{u}(\hat{x}(t), \hat{\lambda}(t)), \tilde{\alpha}(\hat{x}(t), \hat{\lambda}(t)))$ is obtained.
From [83, Section 2], we can suppose that $\hat{u}$ and $\hat{\alpha}$ can be embedded into piecewise $\mathcal{C}^1$ feedback controls u^{**} and α^{**}, respectively, which exists in the neighborhood of the nominal solution

$$\hat{u}(t) = u^{**}(\hat{x}(t), \hat{\lambda}(t)), \quad \hat{\alpha}(t) = \alpha^{**}(\hat{x}(t), \hat{\lambda}(t)). \tag{4.86}$$

Expanding the Hamiltonian of (4.76) with respect to states, controls, and adjoint variables, and maximizing this expansion implies an feedback control law, as follows

$$(u^{**}, \alpha^{**}) = \arg \min_{u \in \mathcal{U}, \alpha \in [0,1]} \mathcal{H}(\hat{x} + \delta x, u, \alpha, \hat{\lambda} + \Lambda \delta x) \tag{4.87}$$

therein, $\delta x := x - \hat{x}$, and Λ is the feedback matrix, which includes matrices at node $\hat{t}_i$,

$$\Lambda(\hat{t}_i) = -(Z_\lambda^i)^{-1} Z_x^i, \ \ i = 0, \ldots, m,$$

where Z_λ^i and Z_x^i, $i = 0, \ldots, m$, are given by Eq. (4.85).

Now, by incorporating with the previous Section 4.2, we use the feedback control (4.87) instead of the control α during the procedure of the (SAR) heuristic.

Definition 35. Let α^{**} satisfy the above explanation, i.e., (4.87). Let $\hat{t}_0 < \ldots < \hat{t}_n$ be a grid discretization $(\hat{t}_0, \hat{t}_f)$ with $0 < \hat{t}_k - \hat{t}_{k-1} \leq h$ for all $k \in \{1, \ldots, n\}$.
Then, the binary-valued step function

$$\beta^{**} : [\hat{t}_0, \hat{t}_f] \to \{0,1\}^M$$

$$\beta_i^{**}(\hat{t}) := \begin{cases} 1, & \text{if } i = i^{**}(k) \\ 0, & \text{else} \end{cases} \qquad \text{for all } \hat{t} \in [\hat{t}_{k-1}, \hat{t}_k)$$

is constructed iteratively for $1 \leq k \leq n$ by the following rule to determine the rounding index $i^{**}(k)$ for the interval $[\hat{t}_{k-1}, \hat{t}_k)$:

$$i^{**}(k) := \arg\max_{i \in M} \{\gamma_{k,i}, \}$$

$$\gamma_{k,i} := \int_{\hat{t}_0}^{\hat{t}_k} \alpha_i^{**}(\hat{t}) d\hat{t} - \int_{\hat{t}_0}^{\hat{t}_{k-1}} \beta_i^{**}(\hat{t}) d\hat{t}. \tag{SUR-SAR}$$

A. Preparations
Let $\hat{t}_0 < \ldots < \hat{t}_n = \hat{t}_f$ be a grid discretization $(\hat{t}_0, \hat{t}_f)$ with maximum grid coarseness $\hat{h} := \max_{1 \leq k \leq n}(\hat{t}_k - \hat{t}_{k-1})$ and let α^{**} be the neighboring feedback controls. We introduce the following variables and quantities

$$\alpha_k^{**} := \frac{1}{\hat{h}_k} \int_{\hat{t}_{k-1}}^{\hat{t}_k} \alpha^{**}(\hat{t}) d\hat{t} \in [0,1]^M, \qquad \beta_k^{**} \in \{0,1\}^M,$$
$$\hat{h}_k := \hat{t}_k - \hat{t}_{k-1}, \qquad \epsilon_k^{**} \in \{0,1\}^M, \qquad \xi_k^{**} \in \{0,1\}^M,$$

for $k \in \{1, \ldots, n\}$. Therein, α_k^{**} denotes the value of α^{**} averaged over the k-th interval, β_k^{**} *is desired output of the rounding indicate which realization* i, $i \in M$, *of the derivative states is switched on in which interval*, $\epsilon_{k,i}^{**}$ will indicate a *switch on* of the i-th derivative state from interval $k-1$ to k and ξ_k^{**} *switch off* of the i-th derivative state. Then, we can reconstruct the function β^{**} from the β_k^{**} as $\beta^{**} = \sum_{k \in M} \chi_{[\hat{t}_{k-1}, \hat{t}_k)} \beta_k^{**}$.

B. The ILP for rounding
Now, we can state the switch aware rounding heuristic for the neighboring feedback controls in the ILP *Switching Aware Rounding Problem* (N-SARP) as follow,

$$\begin{aligned} \min_{\beta_{k,i}^{**}, \epsilon_{k,i}^{**}, \xi_{k,i}^{**}} \quad & \textstyle\sum_{i \in M} c_i \beta_{1,i}^{**} + \sum_{i \in M} d_i \beta_{n,i}^{**} \\ \text{s.t. } & \textstyle\sum_{i=-1}^{+1} \beta_{k,i} = 1, \forall k \in \{1, \ldots, n\}, \\ & -Kh \leq \textstyle\sum_{l=1}^{h} h_l (\alpha_{l,i}^{**} - \beta_{l,i}^{**}) \leq Kh, \forall k \in \{1, \ldots, n\}, i \in M, \\ & \beta_{k+1,i}^{**} - \beta_{k,i}^{**} \leq \epsilon_{k,i}^{**}, \forall k \in \{1, \ldots, n-1\}, i \in M, \\ & \beta_{k,i}^{**} - \beta_{k+1,i}^{**} \leq \xi_{k,i}^{**}, \forall k \in \{1, \ldots, n-1\}, i \in M, \\ & \beta_{k,i}^{**}, \epsilon_{k,i}^{**}, \xi_{k,i}^{**} \in \{0,1\}, \forall i, k, \ M = \{-1, 0, +1\}. \end{aligned} \tag{4.88}$$

Finally, we obtain

$$\beta^{**(h)} = \sum_{k \in M} \chi_{[\hat{t}_{k-1}, \hat{t}_k)} \beta_k^{**(h)},$$

with the $\beta_{k,i}^{**(h)}$ solving (4.88), and the Prop. 3 and the convergence property (4.70) hold true, w.r.t. β^{**} and α^{**}.

In conclusion, we can summarize the main steps as the following algorithm.

Algorithm 2. *NFL-SAR*
***Input**: Controls α, u, state x, adjoint λ.*

for** each $(\hat{t}_i, \hat{t}_{i+1})$, $i = 0, \ldots, m-1$, **do

1. *(NFL). Compute feedback control (u^{**}, α^{**}) by using Eq. (4.87).*
2. *(SAR heuristic). Return $\beta_k^{**(h)}$ from solving N-SAR problem (4.88) by exploiting $\{\alpha^{**}\}_{i=1}^m$ in the preparation step.*

***Output**: Switching aware neighboring feedback controls $\beta^{**(h)} = \sum_{k\in M} \chi_{[\hat{t}_{k-1},\hat{t}_k)} \beta_k^{**(h)}$.*

For more details in the application, readers can see in Subsection 4.4.1.

4.3 An Advanced Algorithm Approach for SwOCP

Instead of the direct method for the reformulation of SwOCP by using GDP, relaxation and the rounding scheme, we consider another approach that is based on a decomposition of MINLP into a NLP and MILP, namely *Combinatorial Integral Approximation* (CIA). See [114] for the general idea and [131] for the latest extension and its application.
Recall, we can rewrite SwOCP by GDP and relaxation as follows:

$$\begin{aligned} \min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & m\left(x(t_f)\right) + \int_{t_0}^{t_f} l(x(t),u(t)) =: \varphi(\cdot) \\ \text{s.t.} \quad & \dot{x}(t) = F(x(t),u(t),\alpha_\sigma(t),\alpha_w(t)),\ w \in \mathcal{W}, \\ & 0 \le r\left(x(t_0),x(t_f)\right), \\ & 0 \le \alpha_\sigma(t)\sigma(x(t)) + \varepsilon,\ (1-\alpha_\sigma(t))\sigma(x(t)) - \varepsilon \le 0, \\ & \alpha_\sigma(t) \in [0,1],\ \textstyle\sum_{w\in\mathcal{W}} \alpha_w(t) = 1, \alpha_w(t) \in [0,1], \end{aligned} \qquad t \in \mathcal{T} = [t_0, t_f], \tag{4.89}$$

where $F(\cdot)$ is given in problem (4.1) and $\varepsilon > 0$. To approximate control functions by working with MILPs, we map between function space and $[0,1]^{2^{n_w}\times M}$ using a time grid $\mathcal{G} := \{t_0 < \ldots < t_M = t_f\}$ with $\Delta_j = t_{j+1} - t_j$ for $j = 0 \ldots M-1$.
The mappings are defined as follows:

$$\theta_{\alpha_w} : [0,1]^{2^{n_w}\times M} \to L^\infty(\mathcal{T}, [0,1]^{2^{n_w}}), \quad \alpha_w = \theta_{\alpha_w}(b),$$

$$\theta_{\alpha_\sigma} : [0,1]^{1\times M} \to L^\infty(\mathcal{T}, [0,1]), \quad \alpha_\sigma = \theta_{\alpha_\sigma}(a),$$

using piecewise constant functions, respectively,

$$\alpha_{w,i}(t) := b_{i,j} \quad i \in 1, \ldots, 2^{n_w}, t \in [t_j, t_{j+1}), j = 0 \ldots M-1, t_j \in \mathcal{G},$$

$$\alpha_{\sigma,i}(t) := a_{i,j} \quad i \in 1, 2, t \in [t_j, t_{j+1}), j = 0 \ldots M-1, t_j \in \mathcal{G},$$

The mappings in reverse direction, respectively,

$$\theta_{\alpha_w}^{-1} : L^\infty(\mathcal{T}, [0,1]^{2^{n_w}}) \to [0,1]^{2^{n_w}\times M}, \quad b = \theta_{\alpha_w}^{-1}(\alpha_w),$$

$$\theta_{\alpha_\sigma}^{-1} : L^\infty(\mathcal{T}, [0,1]) \to [0,1]^{1\times M}, \quad a = \theta_{\alpha_\sigma}^{-1}(\alpha_\sigma),$$

are defined by extracting integrals on the grid $\mathcal{G}$, respectively,

$$b_{i,j} := \frac{1}{\Delta_j} \int_{t_j}^{t_{j+1}} \alpha_{w,i}(\tau) d\tau, \quad i \in 1, \ldots, 2^{n_w}, j = 0 \ldots M-1, t_j \in \mathcal{G},$$

$$a_{i,j} := \frac{1}{\Delta_j} \int_{t_j}^{t_{j+1}} \alpha_{\sigma,i}(\tau) d\tau, \quad i \in 1,2, j = 0 \dots M-1, t_j \in \mathcal{G}.$$

In the following algorithm, RC.SwOCP denotes for the relaxed convexified reformulation of SwOCP, and C.SwOCP denotes for the convexified reformulation of SwOCP.

Algorithm 3. [114, Alg. 1] ***Decomposition of (RC.SwOCP)-(MIOCP)***
Input*: (MIOCP) instance, grid $\mathcal{G}$, algorithmic choices in sets S^{CIA} and S^{REC}.*

1. *Solve (RC.SwOCP)* $\rightarrow \varphi_{rel}, x, u, \alpha_w, \alpha_\sigma, a = \theta^{-1}_{\alpha_\sigma}(\alpha_\sigma), b = \theta^{-1}_{\alpha_w}(\alpha_w)$

2. ***for*** $milp \in S^{CIA}$ ***do***

2a. *Solve milp for data a, b with MILP solver* $\rightarrow w^{milp}$

2b. *Evaluate (C.SwOCP) with fixed* $\omega^{milp} := \theta_{\alpha_w}(w^{milp}) \rightarrow \varphi_{milp}, x, u$

2c. ***end***

3. ***for*** $rec \in S^{REC}$ ***do***

3a. *Create w^{rec} using w^{milp}, φ_{milp} from all milp$\in S^{CIA}$*

3b. *Evaluate (C.SwOCP) with fixed* $\omega^{rec} := \theta_{\alpha_w}(w^{rec}) \rightarrow \varphi_{rec}, x, u$

3c. ***end***

4. *Set* $\varphi^* = \min \left\{ \min_{milp \in S^{CIA}} \varphi_{milp}, \min_{rec \in S^{REC}} \varphi_{rec} \right\}$.

Output*: φ^*, x^*, u^*, w^*, and lower bound φ_{rel}.*

We use Algorithm Decomposition of (RC.SwOCP)-(MIOCP) to approximate the solution of (SwOCP) with a priori bounds. The state and relaxed control trajectories are obtained in Line 1. We approximate the relaxed control with binary ones by solving different MILPs in Line 2, and Line 2a. Their corresponding state trajectories, continuous control, and objective values are evaluated in Line 2b. In Line 3, and Line 3a, the binary controls (in several recombination heuristics) are used to create new candidate binary controls, which are computed in Line 3b. Finally, the solution is selected in Line 4.
The MILP formulations of combinatorial integral approximation type for S^{CIA} and recombination heuristics S^{REC} are appropriately selected from their definition sets. See more details in Section 3 and Section 4 in [131].

4.4 Applications

This section deals with two numerical instances, namely the New York subway problem, and the Flat Hybrid Automaton.

4.4.1 New York Subway Problem

We consider a problem about Subway Optimization goes back to work of [17] and [19] for the city of New York, which is already described in section 3.3, with details from Eq. (3.73) till Eq. (3.87). Our approach was used to treat several station-to-station rides for different station spacings, weight, travel time, etc. Here we show results for a subway problem with 10 wagons ($n_{wag} = 10$), a medium loaded train ($W = 78000$ lbs), for a local run ($S = 2112$ ft), a transit time $T^{\max} = 65$ s that is about 20% longer than the fastest possible and with all engines working ($e = 1.0$).

We transform the problem with the discrete-valued function $w(\cdot)$ to a convexified one with a four-dimensional control function $\alpha \in [0,1]^4$ and $\sum_{i=1}^4 \alpha_i(t) = 1$ for all $t \in [0,T]$. Therefore we can write the right hand side function $\tilde{f}$ and and the LAGRANGE term $\tilde{L}$ as

$$\tilde{f}_1(x,\alpha) = \sum_{i=1}^{4} \alpha_i(t) f_1(x,i),$$

and respectively as

$$\tilde{L}(x,\alpha) = \sum_{i=1}^{4} \alpha_i(t) L(x,i),$$

Now we can reformulate the problem as problem (3.88), as following

$$\begin{array}{ll} \min\limits_{x(\cdot),\alpha(\cdot)} & \int_0^T \tilde{L}(x,\alpha)dt \\ \text{s.t.} & \dot{x}_0(t) = x_1(t), \\ & \dot{x}_1(t) = \tilde{f}_1(x,\alpha), \\ & x(0) = (0,0)^T,\ x(T) = (S,0)^T, \\ & \alpha \in [0,1]^4,\ \sum_{i=1}^4 \alpha_i(t) = 1\ \forall t \in [0,T], \end{array} \tag{4.90}$$

with $S = 2112$ ft and $T \le T^{\max} = 65$ s.

Table 4.1: Optimal solution, where S, P, C, and B are denoted for Series, Parallel, Coasting and Braking, respectively. The column α is presented the relaxed controls (which are obtained by `MUSCOD-II`), and the column $\hat{\alpha}$ is described the resulting integer ones from the rounding scheme, while the last column is resulted the neighboring feedback controls α^{**}.

Time t	Mode	f_1	$x_0(t)$ [ft]	$x_1(t)$ [mph]/[ft/s]	α	$\hat{\alpha}$	α^{**}
0.0	S	f_1^{1A}	0.0	0.0	$(1,0,0,0)$	$(1,0,0,0)$	$(1,0,0,0)$
0.6317	S	f_1^{1B}	0.4537	0.979474/1.43656	$(1,0,0,0)$	$(1,0,0,0)$	$(1,0,0,0)$
2.8522	S	f_1^{1C}	11.6480	6.73211/9.87375	$(1,0,0,0)$	$(1,0,0,0)$	$(1,0,0,0)$
3.6434	P	f_1^{2B}	24.4836	8.6572/12.6972	$(0,1,0,0)$	$(0,1,0,0)$	$(0,1,0,0)$
5.5999	P	f_1^{2C}	52.1713	17.0273/24.9734	$(0,1,0,0)$	$(0,1,0,0)$	$(0,1,0,0)$
12.607	S	f_1^{1C}	277.711	25.6452/37.6129	$(0.5,0.5,0,0)$	$(1,0,0,0)$	$(1,0,0,0)$
45.7827	C	$f_1(3)$	1556.5	26.8579/39.3915	$(0.8,0,0.2,0)$	$(0,0,1,0)$	$(0,0,1,0)$
46.8938	C	$f_1(3)$	1600	26.5306/38.9115	$(0,0,1,0)$	$(0,0,1,0)$	$(0,0,1,0)$
57.16	B	$f_1(4)$	1976.78	23.5201/34.4961	$(0,0,0.65,0.35)$	$(0,0,0,1)$	$(0,0,0,1)$
65	-	–	2112	0.0/0.0	$(0,0,0,1)$	$(0,0,0,1)$	$(0,0,0,1)$

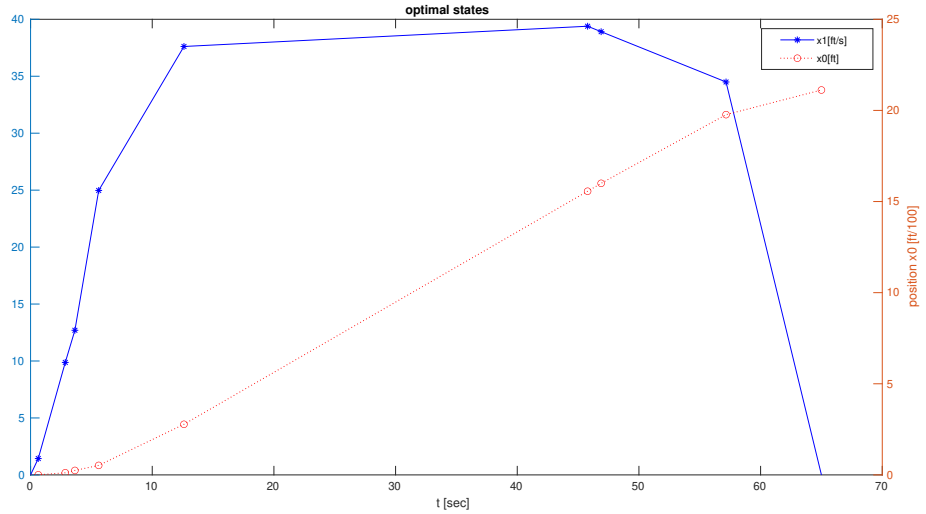


Figure 4.1: Optimal states (position $\cdots$ and velocity —).

We first operate in series until $\hat{t}_1 = 3.6434$, then we operate in parallel model until $\hat{t}_2 = 12.607$, then again in series until $\hat{t}_3 = 45.7827$; at $\hat{t}_4 = 57.16$ we stop coasting and brake until $T_{\max} = 65$, see Fig. 4.1 and Fig. 4.2. All numerical results are summarized as in Table 4.1. In other words, we finally determine the switches.

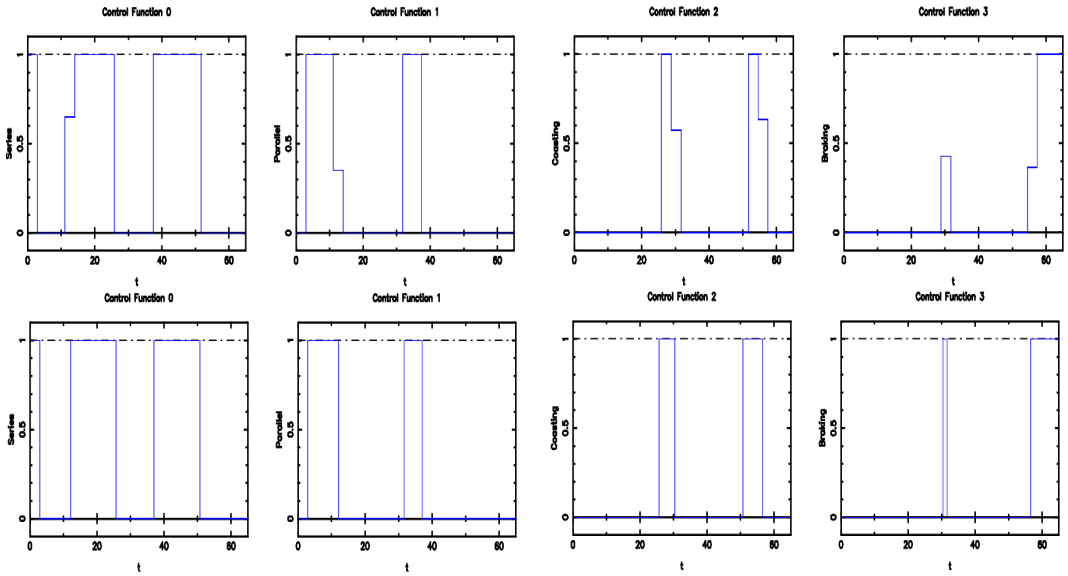


Figure 4.2: The upper solution is optimal for the relaxed problem, while the lowest row shows the optimal integer controls.

Using "the neighboring feedback law" for SAR via the implicit function theorem, see Subsection 4.2.2, we get the neighboring feedback law at time point $\hat{t}_0$ for control α^{**} by exploiting Eq. (4.87). By the similar way, we also obtain the neighboring feedback controls at time points $\hat{t}_j$, where $j = 2, \ldots, 10$.
A neighboring feedback control, see Tab. 4.1, is then given by

$$\{\alpha^{**}(\hat{x}_j)\}_{j=1}^{10}.$$

4.4.2 The Flat Hybrid Automaton

This approach can be applied to a new model class of hybrid system, which have been introduced by Kleinert and Hagenmeyer in 2019, namely, Flat Hybrid Automaton (FHA), see [79]. This subsection reports on some results obtained by using Filippov's rule and relaxation to solve FHA with computational results in special cases of a DC electrical network example.

Algorithm Approach for The Dynamic Optimization Problem of FHA

Consider a Flat Hybrid Automata FHA $= \{A^{fl}, C^{fl}\}$ and two states $(d_0, z_0), (d_{t_f}, z_{t_f})$, we want to find a path $P = \{e_{\varsigma 1}, e_{\varsigma 2}, \ldots, e_{\varsigma n}\}$ defined through the sequence of flat output $Z^* = \{z^*_{d_{\xi 1}}(t), z^*_{d_{\xi 2}}(t), \ldots, z^*_{d_{\xi n+1}}(t)\} \in Z_{d_{\xi i}}$ and the discrete inputs $v^*(t)$ that yields

$$\begin{aligned}
&\min_{\{z_{d_{\xi i}}(t)\}, v(t)} \quad J(.) = \sum_{i=1}^{n} \alpha(e_{\varsigma i}, z_{d_{\xi i}}(t), v(t)) \\
&\text{s.t.} \quad \alpha(e_{\varsigma i}, z_{d_{\xi i}}(t), v(t)) = \int_{t_i}^{t_{i+1}} L_i(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)))d\tau + \beta(e_{\varsigma i}) + \gamma(v(t)), \\
&\quad t_i \in t^*, \forall i \in [2, n-1], \\
&\quad 0 \le t_i < t_{i+1}, t_1 = t_0, \\
&\quad (d(t_1), z(t_1)) = (d_0, z_0), \\
&\quad (d(t_n), z(t_n)) = (d(t_f), z(t_f)), \\
&\quad 0 \le c(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau))), i \in [1, \ldots, n],
\end{aligned} \tag{4.91}$$

where $t^* = t', t'', \ldots$ state switching times. We refer [128] for more details.
The goal is to solve (4.91) by exploiting Filippov's rule to rewrite this problem to relaxed convexified one, together with the arising of the additional mixed state-control constraints, then the resulted problem can be solved by using an appropriate numerical method.
Since the input $v(t) \in \{0,1\}^{nv}$ of flat discrete subsystem A^{fl}, and the discrete-state transition $e_{\varsigma i} : d_{\varsigma i} \to d'_{\varsigma i}$, by reformulating (4.91) with Filippov's rule, POC and relaxation, we obtain the equivalent problem

$$\begin{aligned}
&\min_{\{z_{d_{\xi i}}(t)\}, \theta^v(t), \theta^{e_{\varsigma i}}(t)} \quad J(.) = \sum_{i=1}^{n} \alpha(e_{\varsigma i}, z_{d_{\xi i}}(t), v(t)) \\
&\text{s.t.} \quad \alpha(e_{\varsigma i}, z_{d_{\xi i}}(t), v(t)) = \textstyle\sum_{j=1}^{2^{nd_{\varsigma i}}} \bar{\theta}_j^{e_{\varsigma i}}(t) \sum_{k=1}^{2^{nv}} h_l(\cdot)\bar{\theta}_k^v(t),\ l \in \{++, +0, 0+, 00\}, \\
&\quad t_i \in t^*, \forall i \in [2, n-1], \\
&\quad 0 \le t_i < t_{i+1}, t_1 = t_0, \\
&\quad (d(t_1), z(t_1)) = (d_0, z_0), \quad (d(t_n), z(t_n)) = (d(t_f), z(t_f)), \\
&\quad 0 \le c(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau))), i \in [1, \ldots, n], \\
&\quad \textstyle\sum_{j=1}^{2^{nd_{\varsigma i}}} \bar{\theta}_j^{e_{\varsigma i}}(t) = 1, \bar{\theta}^{e_{\varsigma i}}(t) \in [0,1]^{nd_{\varsigma i}}, \quad \sum_{k=1}^{2^{nv}} \bar{\theta}_k^v(t) = 1, \bar{\theta}^v(t) \in [0,1]^{nv},
\end{aligned} \tag{4.92}$$

where

$$h_{++}(\cdot) := \int_{t_i}^{t_{i+1}} L_i(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)))d\tau + \beta(1) + \gamma(1),$$

$$h_{+0}(\cdot) := \int_{t_i}^{t_{i+1}} L_i(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)))d\tau + \beta(1) + \gamma(0),$$

$$h_{0+}(\cdot) := \int_{t_i}^{t_{i+1}} L_i(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)))d\tau + \beta(0) + \gamma(1),$$

$$h_{00}(\cdot) := \int_{t_i}^{t_{i+1}} L_i(\Phi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)), \Psi_{d_{\xi i}}(z_{d_{\xi i}}(\tau)))d\tau + \beta(0) + \gamma(0),$$

with the additional mixed constraints:

$$\begin{aligned}
&(\bar{\theta}_j^{e_{\varsigma i}}(t) - 1)(v(t) - 1) = 0, \text{ and, } (\bar{\theta}_k^v(t) - 1)(e_{\varsigma i} - 1) = 0,\\
&(\bar{\theta}_j^{e_{\varsigma i}}(t) - 1)(v(t) - 1) = 0, \text{ and, } \bar{\theta}_k^v(t)e_{\varsigma i} = 0,\\
&\bar{\theta}_j^{e_{\varsigma i}}(t)v(t) = 0, \text{ and, } (\bar{\theta}_k^v(t) - 1)(e_{\varsigma i} - 1) = 0,\\
&\bar{\theta}_j^{e_{\varsigma i}}(t)v(t) = 0, \text{ and, } \bar{\theta}_k^v(t)e_{\varsigma i} = 0,
\end{aligned} \qquad \begin{array}{c} i = \overline{1, n},\\ j = \overline{1, 2^{nd_{\varsigma i}}},\\ k = \overline{1, 2^{nv}}. \end{array}$$

Problem (4.92) is the relaxed convexified formulation of the dynamic optimization problem of FHA (4.91). We can solve (4.92) by direct multiple shooting methods.
Suppose that the optimal solution of (4.92) is $(\{z^*_{d_{\xi i}}\}, \{\bar{\theta}^v\}^*, \{\bar{\theta}^{e_{\varsigma i}}\}^*)$. Hence the corresponding optimal solution of (4.91) is $(\{z^*_{d_{\xi i}}\}, v^*)$. We are interested in the relation between $\{\bar{\theta}^v\}^*$ and v^*. Applying an appropriate rounding strategy (denoted by RS, see Section 2.6.7), we obtain

$$v^*(t) = \begin{cases} \bar{\theta}^{v*}(t) & \text{if } \bar{\theta}^{v*}(t) \in \{0, 1\},\\ RS(\bar{\theta}^{v*}(t)) & \text{if } \bar{\theta}^{v*}(t) \in (0, 1). \end{cases}$$

Remark 30. In comparison with [128, Alg. 1], instead of the computation of all possible paths P_j through the FHA connecting d_0 and d_f without visiting any node twice, our approach has just need to use the convexified combination of the choices of POC to find the optimal path with the minimal cost function value.

DC Electrical Network Example

We refer to [128, Sec. 5] and use their model and parameters for testing our approach on FHA. The following parameters are used in the calculation:

$$R = 5, C = 0.8, L = 7, R_{L1} = 2, R_{L2} = 3, v_0 = 6, i_0 = 2.5$$

We choose the initial and the final states

$$v_{L1}(t_0) = 0.5, v_{L1}(t_f) = 12, i_{L2}(t_0) = 0.1, i_{L2}(t_f) = 4, d_{t_0} = d_1, d_{t_f} = d_4.$$

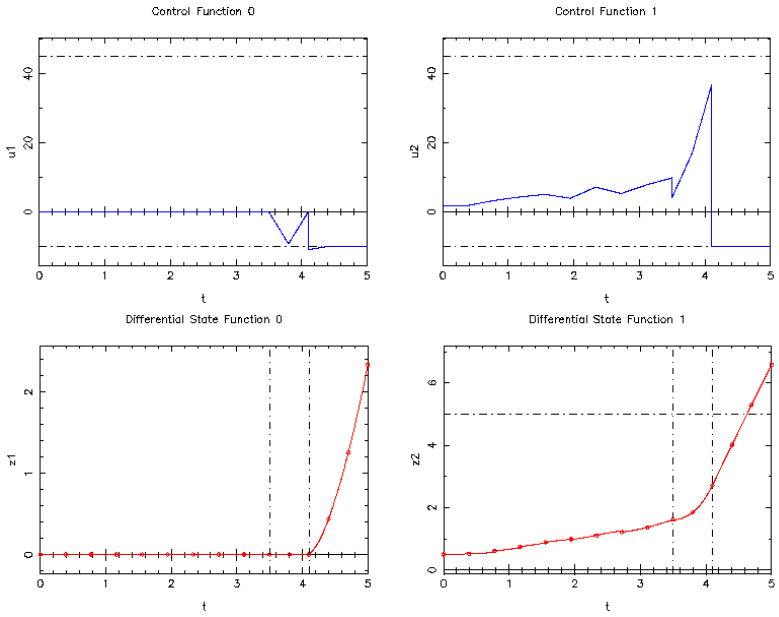


Figure 4.3: Optimal flat inputs u and outputs z. Blue lines u_1, u_2, red lines z_1, z_2, vertical black dashed lines show switching times.

Remark 31. Table 4.2 and Fig. 4.3, which are obtained by solving the DC Electrical Network in FHA, show that our approach could has wide applications to solve complex problems.

Table 4.2: Results on Electrical DC Network.

Optimal outputs	$(12, 4)$
Optimal path	$\{e_1, e_6\}$
Optimal discrete state	$\{d_1, d_2, d_4\}$
Convergence	achieved
Number of switches	2
Transition cost	30

Chapter 5

Determination of Switches in SwOCP

In order to handle the optimal control in hybrid systems, switched systems have been investigated by simplifying the details of the discrete behavior to switching patterns from a certain class with discontinuity in vector fields, cf. [126, 132]. The main tasks here are determining switches in SwOCP, where the switches come from the discontinuous of the right hand side of the ODE and the integer values of the controls. These challenged topics have been considered in the work of [18, 95] and [121–123], where the detailed methods are fully discussed in Chapter 3 and Chapter 4.

In applications, mechanical systems frequently have a large number of discontinuous transitions. Examples include force curves derived from discontinuous approximations of characteristic curves, hysteresis, friction, impacts, and controllers, cf. [87, 93, 96]. Furthermore, discontinuities can occur when working with implicit systems since the non-singularity of certain matrices, which is required to define the index, is not provided at individual points. Any change in the degrees of freedom of a system causes a discontinuity. In [40], an introductory tutorial on discontinuous dynamical systems with the notions of their solution as well as available tools to study their gradient information are presented. Since a certain minimum order of differentiability is required for consistency and convergence claims and the order and step size control of numerical integration methods, these points cannot simply intersect. In addition, an accuracy-controlled calculation of sensitivity matrices is required to use the integration methods in an optimization environment. This is only possible by explicitly considering the discontinuities in OCP.

This chapter is presented as follows. A general description of SwOCP with discontinuous differential equations as differential equations with switching conditions is given in Subsection 5.1.1. Subsequently, in Subsection 5.1.2, SwOCP is treated with a switching point algorithm. Next, a generalized three-valued switching logic is also considered in Subsection 5.1.3 with the general problem of inconsistent switching stated. The sensitivities are calculated and analyzed in the forward mode in Section 5.2. The chapter ends with Section 5.3, where brief comments on other approaches for tracking switches are considered.

5.1 A Discontinuous Dynamics-Based Approach to Handle Switches to SwOCP

The main idea of this section is inspired from [49, Chapter 5].

5.1.1 SwOCP with Switching Conditions in ODEs

The states s_i, $i = 0, \dots, m-1$ appearing on the right-hand side of the differential equation can exhibit discontinuous or non-differentiable behavior. In the following, we consider the OCP of the form of (4.13), which results after multiple shooting discretized taken,

$$\begin{aligned} \min_{y(\cdot)} \quad & \textstyle\sum_{i=1}^{m} l_i(s_i, q_i) \\ \text{s.t.} \quad & 0 = x_i(t_{i+1}; t_i, y_i) - s_{i+1}, \ i = 0, 1, \dots, m-1, \\ & 0 \le r(s_0, s_m). \end{aligned} \tag{5.1}$$

with piecewise smooth right hand side x_i of ODE, and $y_i := (s_i, q_i)$.

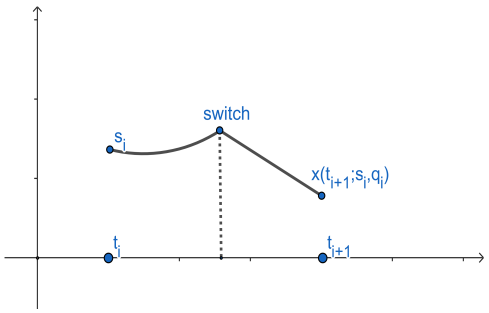


Figure 5.1: Switch occurs in the interval $[t_i, t_{i+1}]$.

The conditions for transitions between the regions where the functions are smooth are i.a. known and can be described as zeros of switching functions σ. Let σ be a vector-valued, state-dependent function

$$\sigma = \sigma(y(t)) = (\sigma_1, \dots, \sigma_{nsw})^T,$$

where nsw is "number of switches". For example, there is a switch in $[t_i, t_{i+1}]$, see Fig. 5.1. Then, problem (5.1) can then be written as a SwOCP

$$\begin{aligned} \min_{y(\cdot)} \quad & \textstyle\sum_{i=1}^{m} l_i(s_i, q_i) \\ \text{s.t.} \quad & 0 = x_i(t_{i+1}; t_i, y_i, \operatorname{sgn}(\sigma(y_i))) - s_{i+1}, \ i = 0, 1, \dots, m-1 \\ & 0 \le r(s_0, s_m). \end{aligned} \tag{5.2}$$

The expression $\operatorname{sgn}\sigma$ is to be understood component-wise, i.e., x_i depends on a combination of the signs of the components of σ, where $i = 0, 1, \dots, m-1$. For a fixed $\operatorname{sgn}\sigma$ then be sufficiently smooth:

$$x : \mathbb{R}^{n+1} \times \Omega \to \mathbb{R}^n, \quad x \in \mathcal{C}^l(\mathbb{R}^{n+1} \times \Omega), \quad \Omega = \{-1, 0, +1\}^{nsw},$$

where l is sufficiently large. Discontinuities only occur at zero points of a switching function.

In addition, the solution variables themselves may show jumps, for example, velocities in impact processes are discontinuous. The time when such a discontinuity occurs is implicitly provided as the zero of a switching function

$$\sigma(y_i(\hat{t}-)) = 0, \quad i = 0, 1, \ldots, m-1.$$

The right-hand limit $y_i(\hat{t}+)$ is a function of the left-hand limit $y_i(\hat{t}-)$

$$y_i(\hat{t}+) = s(\hat{t}, y_i(\hat{t}-)), \quad i = 0, 1, \ldots, m-1.$$

Difficulties in Discontinuous System Numerical Integration

In practice, discontinuous systems are frequently handled without the use of switch point search. This frequently causes in integrator failure, which manifests as method order collapse, step size control failure, and incorrect results:
Conventional integration methods with automatic step sizes and order control estimate the local error after each integration step and, by comparing it with a given error limit, determine whether it is accepted or rejected and with which step size and order the integration should be continued. If there is a point of discontinuity within an integration step, this usually leads to a large local error and thus to a drastic step size reduction. As a result, the step size often becomes very small. Then several steps with an increasing step size are usually carried out until there is a renewed attempt at the transition via the point of discontinuity. This can be repeated several times so that the effort is immense.
A second problem lies in the way the local error is estimated. The estimated local error (as well as the process coefficients for linear multi-step processes) are calculated using formulas that assume the continuity of trajectory and its derivatives. The error estimate is invalid if this requirement is not met. As a result, no remark is made about whether the solution remains inside the set tolerance limit after a step. This also applies to k consecutive steps in a k-step procedure.
If one also wishes to apply integration methods in conjunction with modern optimization methods, which require the efficient and accurate generation of sensitivity matrices, then localization of impact points by applying impact functions is unavoidable.
The challenges associated with directly integrating differential equations with discontinuous right-hand sides can often be avoided by using integrators with fixed step sizes or low order. The disadvantage of the first approach is that it does not allow for error checking. For a given accuracy, both methods have the disadvantage of requiring a great deal of effort.
The equally frequently used smoothing approaches make the system artificially stiff and often lead to very large discontinuous higher derivatives.
Alternatively, methods have been developed that control the point of discontinuity as the point of discontinuity increases significantly. In [57] a transition increment is then determined that keeps the local error below the required tolerance.
However, all approaches that do not or not explicitly localize discontinuities remain unsatisfactory and require a large number of heuristics. For this reason, a different approach is chosen here, the explicit localization of discontinuities as the zero point of switching functions.

5.1.2 A Switching Point Algorithm for Handling The Discontinuities in ODE

Switching point algorithm basically consist of the following main steps:

1. Determine a discontinuity by checking the sign of the switching functions
2. Localization of the discontinuity as the zero point of a switching function: switching point search
3. Integrated process of "continuous" system
4. Switch to the "new" right side
5. Discretization adjustment.

Based on the above core steps, we can propose a numerical algorithm approach for solving SwOCP with discontinuities.
Inputs: Objective function, ODE system, point constraints, and path constraints.

Algorithm 4 (**Switching Point Algorithm**)**.**
1. Initialize the problem by setting up the objective function, the ODE system, and the starting and ending point constraints.
2. Set up a tolerance level for detecting switches in the system.
3. Use a numerical method (e.g. forward integration method (`ode45 Matlab`), and/or backward differential formula (BDF) with Secant method) to integrate the system until a switch is detected.
4. Use a switching function to locate the exact point of the switch.
5. Integrate the system again from the switch point until the next switch is detected.
6. Repeat steps 4-5 until the final time is reached.
7. Adjust the discretization of the solution to ensure accuracy.
8. Output the solution with the exact time of the switches, and the objective's value.

Outputs: Solution with the exact time of the switches, and objective function value.
These main steps are described in more details as follows, where we assume that there are multiple switches in a shooting interval $[t_i, t_{i+1}]$.

Determining a Discontinuity

For each shooting interval $[t_i, t_{i+1}]$, multiple switches occur if $\sigma(x(t), q_i) = 0$ at τ_j, $\tau_{j+1}, \ldots,$ $\tau_{j+n_{sw}}$. We denote the initial mode k_i in $[t_i, t_{i+1}]$ by $k_i := \operatorname{sgn}(\sigma(s_i, q_i))$, and the subsequent modes are determined by the sign of σ after each switch. For each switch

$$\sigma(x(\tau_l), q_i) = 0, \quad l = j, j+1, \ldots, j+n_{sw},$$

with a sign change indicating a mode transition, e.g., from f_{k_l} to $f_{k_{l+1}}$.

Remark 32. Multiple zeros of σ require repeated detection within the same interval, which can be numerically challenging due to the sensitivity of σ and the adaptive steps of `ode45`. To overcome this situation, we use a numerical event detection mechanism (via `ode45`) to determine each zero of σ and ensure the switching function is smooth enough for reliable detection.

Searching Switching Points

We start by solving the ODE in each $[t_i, t_{i+1}]$:

$$\dot{x}(t) = f_{k_i}(t, x(t), q_i), \quad x(t_i) = s_i,$$

where $k_i = \text{sgn}(\sigma(s_i, q_i))$, and monitor $\sigma(x(t), q_i)$. When a switching point is detected at τ_j

$$\sigma(x(\tau_j), q_i) = 0.$$

the integration pauses, the mode updates, and integrated process resumes. For multiple switches, this process repeats for each τ_l, $l = j, \ldots, j + n_{sw}$,

1. Integrate from t_i to τ_j with mode k_i.
2. At τ_j, compute $x(\tau_j)$, update mode to $k_{i+1} = \text{sgn}(\sigma(x(\tau_j^+), q_i)))$, and continue integration from τ_j to τ_{j+1}.
3. Repeat for $\tau_{j+1}, \ldots, \tau_{j+n_{sw}}$, until reach t_{i+1}.

The state at t_{i+1} is $x(t_{i+1}) = x_i(t_{i+1}; t_i, s_i, q_i, \text{sgn}(\sigma(s_i, q_i)))$ obtained by piecewise integration over subsegments $[t_i, \tau_j]$, $[\tau_j, \tau_{j+1}], \ldots, [\tau_{j+n_{sw}}, t_{i+1}]$.

Remark 33. Since `ode45` may miss closely spaced switches if time step is too large or if σ change rapidly, we can set tight tolerances in `ode45` (e.g., 10^{-8} and define a robust function $\sigma(x(t), q_i) = 0$ with termination at each zero crossing. If switches are very close, we consider a finer shooting grid by increasing number of shooting nodes or a post-processing step to refine switch detection.

Remark 34 (**Other methods for switching point search**). In [49, sec 5.3.2], BDF with Secant method (or inverse interpolation) is used to search the switching points in dealing with Safeguard techniques.
In [33, 91] a NEWTON method is used to search for the switching point, in which a suitable start value for the iteration is also determined as the zero of a Hermite polynomial. Similar approaches can be found in [34, 35, 52, 62]. All of these methods require a large number of right-hand side evaluations because they do not have a continuous solution representation.
In [23, 50] a continuous solution representation is used for the first time. In [23] an Adams method is used for this purpose, in [50] a continuous solution representation is determined with the help of a 3rd-order Hermite polynomial, and the discretization is a 3rd/4th RUNGE-KUTTA pair order. ENRIGHT et al. [51] use a p-th order RUNGE-KUTTA method and corresponding local interpolation for localization. The switching point search is carried out with the help of a halving strategy until a termination criterion is met.
In [32, 33], a continuous representation of the switching functions themselves is obtained by setting up additional differential equations for the switching functions and using an integration method with a continuous solution representation. However, this procedure is impractical for many switching functions, since the differential equation system is very large and the additional effort is immense. Here the continuous solution representation of σ proposed above, which is obtained by substituting the continuous solution representation of y into σ, offers considerable advantages since it is more computationally essentially without additional effort.

Integration of Continuous System

Within each $[\tau_{l-1}, \tau_l]$, the system is continuous under mode f_{k_l}:

$$x(t) = x(\tau_{l-1}) + \int_{\tau_{l-1}}^{t} f_{k_l}(s, x(s), q_i)ds, \quad t \in [\tau_{l-1}, \tau_l].$$

The state at t_{i+1} is computed as follows

$$x_i(t_{i+1}) = x(\tau_{j+n_{sw}}) + \int_{\tau_{j+n_{sw}}}^{t_{i+1}} f_{k_i}(s, x(s), q_i)ds,$$

where $x(\tau_{j+n_{sw}})$ is obtained recursively through

$$x(\tau_l) = x(\tau_{l-1}) + \int_{\tau_{l-1}}^{\tau_l} f_{k_l}(s, x(s), q_i)ds, \quad l = j, j+1, \ldots, j+n_{sw}.$$

The continuity at the shooting node is ensured by the matching condition

$$x_i(t_{i+1}; t_i, s_i, q_i, \mathrm{sgn}(\sigma(s_i, q_i))) = s_{i+1}. \tag{5.3}$$

Remark 35. Multiple switches rise computational complexity, as each subsegment requires separate integration. Hence ones should exploit `ode45`'s event detection to pause and resume integration at each switching point τ_l, storing intermediate states $x(\tau_l)$.

Switching to the "New" Right Side

At each τ_l, the mode is updated as

$$k_l = \mathrm{sgn}(\sigma(x(\tau_l^+), q_i)),$$

and continue integration

$$\dot{x}(t) = f_{k_l}(t, x(t), q_i), \quad x(\tau_l^+) = x(\tau_l^-),$$

For $n_{sw}+1$ switches, the sequence of modes is $k_l, k_{l+1}, \ldots, k_{l+n_{sw}+1}$, with state continuity at each τ_l.

Remark 36. Since rapid mode changes may cause numerical instability in `ode45`, ones must ensure the dynamics f_{k_l} are continuous and use high-precision detection to accurately determine each τ_l.

Discretization Adjustment

Multiple switches within $[t_i, t_{i+1}]$ amplify discretization errors due to `ode45`'s adaptive steps. To this end, we adjust by:

- Solve $\sigma(x(\tau_j), q_i) = 0$ with high precision (e.g., tight tolerances in `ode45`).
- Verify state continuity: $x(\tau_j^-) = x(\tau_j^+)$ at each switch.
- Refine the solution near τ_j by integrating over $[\tau_j - \varepsilon, \tau_j + \varepsilon]$ with a finer appropriate grid, where $\varepsilon > 0$.

- If many switches occur, adjust the shooting node t_i to include τ_l as nodes, splitting $[t_i, t_{i+1}]$ into subintervals $[t_i, \tau_j]$, $[\tau_j, \tau_{j+1}], \ldots, [\tau_{j+n_{sw}}, t_{i+1}]$ with modified constraints.
- Increase the number of shooting interval or adaptively redistribute nodes to concentrate around regions with frequent switches, ensuring the matching condition (5.3).

5.1.3 Inconsistent Switching with Switching Logic

When treating differential equations with shifting conditions where the right-hand side f shows real jumps, e.g. when modeling with the help of Coulomb friction, the shifting process can become inconsistent. Beyond these discontinuities, there is no solution to the differential equation in the classical sense (e.g. [16, 56]). Treatment with the usual two-valued switching logic is impossible.
A solution in the classical sense is understood here as "the solution satisfies the differential equation almost everywhere". This corresponds to the approach in the last sections, in which the differential equation was only not fulfilled in the switching points. If the differential equation is discontinuous along a manifold $\sigma = 0$; it only has a solution in the classical sense if the solution allows for the manifold, i.e., breaking through is not possible, a solution in the classical sense is no longer defined. A generalized solution concept according to Filippov, see Section 2.6, provides a remedy here.
In the following, the problem of inconsistent switching is analyzed and a continuation of the solution is constructed based on a generalized solution according to Filippov [56]. The automatic handling using a *generalized three-valued switching logic* is described.
In mechanics, inconsistent switching often occurs when modeling Coulomb friction phenomena, which is described in Section 6.1.

Directional Fields with Inconsistent Switching

In the following, the problem with inconsistent switching is in the form of ordinary differential equations

$$\begin{aligned} \min_{y(\cdot)} \quad & \varphi\left(y(t)\right) \\ \text{s.t.} \quad & \dot{y} = f(t, y, \operatorname{sgn}\sigma) = \begin{cases} f(t, y, +1) = f_+(t, y) & \text{for } \sigma > 0 \\ f(t, y, -1) = f_-(t, y) & \text{for } \sigma < 0 \end{cases} \quad t \in \mathcal{T}, \\ & y(t_0) = y_0, \end{aligned} \tag{5.4}$$

explained with the switching function $\sigma : \mathbb{R}^{n+1} \to \mathbb{R}$.
The domain of definition of f can be divided into three areas:

$$\mathcal{S}_+ = \{(t, y) \mid \sigma(t, y) > 0\}, \quad \mathcal{S}_- = \{(t, y) \mid \sigma(t, y) < 0\}, \quad \mathcal{S} = \{(t, y) \mid \sigma(t, y) = 0\},$$

where $\mathcal{S}$ denotes the discontinuity surface.
$D\sigma_+, D\sigma_-$ denote the *auxiliary switches*:

$$\begin{aligned} D\sigma_+ &:= \frac{\partial \sigma}{\partial y} f_+ + \frac{\partial \sigma}{\partial t} \\ D\sigma_- &:= \frac{\partial \sigma}{\partial y} f_- + \frac{\partial \sigma}{\partial t}. \end{aligned}$$

$D\sigma_+, D\sigma_-$ can be interpreted as *directional derivatives* of the switching function in the direction f_+ or f_-: For a given $y(t)$ we define $\sigma^y(t) := \sigma(t, y(t))$, and hence

$$\frac{d\sigma^y}{dt} = \frac{\partial \sigma}{\partial y}\dot{y} + \frac{\partial \sigma}{\partial t}.$$

If ones insert $\dot{y} = f(t, y, +1)$ or $\dot{y} = f(t, y, -1)$, ones get $D\sigma_+$ or $D\sigma_-$.
Now the question is whether the switch can be pierced. Let $\sigma^y(\hat{t}) = 0$. The signs of $D\sigma_+$ and $D\sigma_-$ are decisive for answering this question (combinations with $D\sigma = 0$ are not considered, in this case further differentiations are necessary): In cases 1 and 2, the solution can be

Table 5.1: Four cases of directional fields.

Case	$D\sigma_+$	$D\sigma_-$	Exit switch
1	> 0	> 0	possible after $\sigma > 0$ (consistent)
2	< 0	< 0	possible after $\sigma < 0$ (consistent)
3	> 0	< 0	in both directions (bifurcation)
4	< 0	> 0	not possible (inconsistent)

continued in the classical sense and the numerical treatment with classical switching logic is possible, see Tab. 5.1 and Fig. 5.2.

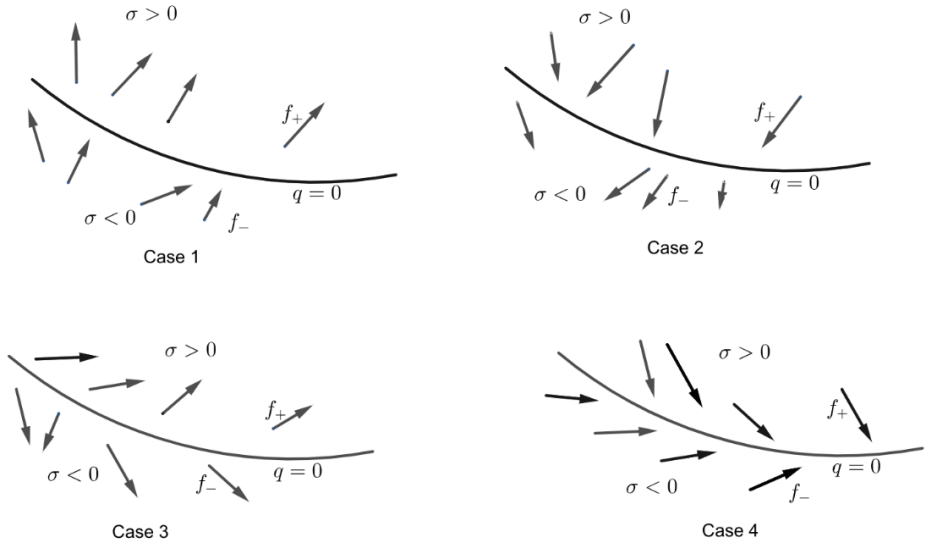


Figure 5.2: Four cases of different directional fields.

In case 4, inconsistent switching occurs. A change of sign on the right-hand side does not mean that the switching function σ can change its sign: "the solution gets stuck in the manifold", see Fig. 5.2, the differential equation has no solution in the classical sense.
Numerical treatment with a classic two-valued switching logic leads to oscillations around the switch: any attempt to change the right side results in a sign change of the switching function.

Example 5.1. Consider a switched problem

$$\begin{aligned} \min_{y(\cdot)} \quad & y(T) \\ \text{s.t.} \quad & \dot{y} = 1 - 2\,\mathrm{sgn}(y), \\ & y(0) = y_0 > 0, \\ & t \in [0, T]. \end{aligned}$$

It holds $\sigma = y$: The solution of this ODE system for $t < y_0$ is $y(t) = y_0 - t$. At time $t = y_0$, $y(t) = 0$ holds. Since the directional field points to the manifold $\mathcal{S} = \{y \mid y = 0\}$ from both sides, the solution cannot leave the manifold again, so for $t \geq y_0$ we have $y(t) = 0$. However, the differential equation no longer fulfills this function.
In conclusion, we have the solution for the given switched problem

$$y(T) = \begin{cases} y_0 - t, & \text{if } t < y_0, \\ 0, & \text{otherwise,} \end{cases}$$

and the optimal objective value is $\min\{0, y_0 - T\}$, where $y_0 > 0$.

To deal with these phenomena, Filippov [56] had extended the solution concept for ordinary differential equations by allowing set-valued right sides. This allows the solution to continue beyond such critical points.

Definition 36. [49, Def. 5.1] The function $y(t)$, $t \in [t_0, t_f]$ is called the solution of the differential equation $\dot{y}(t) = f(t, y(t))$, if the following conditions are met:

- y is absolutely continuous,
- for almost all $t \in [t_0, t_f]$ and any $\delta > 0$, the vector $\dot{y} = \frac{dy}{dt}$ belongs to the smallest closed convex set that contains all values $f(\cdot)$ in a δ-neighborhood of $y(t)$:

$$\dot{y}(t) \in \bigcap_{\delta > 0} \bigcap_{\mu(N)=0} \overline{conv}\,(f(U(y(t), \delta) \setminus N, \cdot)) = \tilde{f}(t, y).$$

Here, μ denotes the Lebesgue measure.

Remark 37. For continuous functions f the set $\tilde{f}$ consists only of the point $f(t, y(t))$ and this notion of solution agrees with the classical one.
The requirement "absolutely continuous" corresponds to the requirement for the existence of a generalized derivative: every absolutely continuous function x can be written as an indefinite integral over a summable function ϕ:

$$x(t) = x(a) + \int_a^t \phi(s) ds.$$

On the basis of this solution concept, Filippov [56] was able to show the existence, continuity, uniqueness, and continuous dependency of the solution from initial values and the right-hand side under some additional assumptions.

We explain this solution concept as follows. Let $\sigma(y) = 0$, $\sigma(y) \in \mathbb{R}$, and

$$f_+(t,y) := f(t,y,+1), \quad f_-(t,y) := f(t,y,-1),$$

Then the set $\tilde{f}(t,y)$ consists of the vectors whose endpoint lies on the line connecting f_+ and f_-:

$$\tilde{f}(t,y) = \{\alpha f_+ + (1-\alpha) f_-, \alpha \in [0,1]\}$$

so

$$\dot{y} = \alpha f_+ + (1-\alpha) f_-. \tag{5.5}$$

Now the question of choosing a suitable element from the convex hull arises, i.e., the question of choosing α. The parameter α is chosen randomly in the numerical implementation, convergence of the EULER method can then be shown. The method given in [16] makes more sense in the type that it uses the fact that the solution cannot leave the manifold for the choice of α. This procedure is described below.
Since the directional field of the differential equation is directed in such a way that the solution cannot leave the switch

$$\sigma(t, y(t)) = 0 \tag{5.6}$$

and thus it follows by differentiation

$$\sigma_y \dot{y} + \sigma_t = 0. \tag{5.7}$$

If we insert (5.5) into (5.7), we can obtain $\alpha = -\frac{D\sigma_-}{D\sigma_+ - D\sigma_-}$. From this one gets

$$\dot{y} = \frac{D\sigma_+ f_- - D\sigma_- f_+}{D\sigma_+ - D\sigma_-} \tag{5.8}$$

as a differential equation as long as the consistency conditions

$$D\sigma_+ < 0 \quad \text{and} \quad D\sigma_- > 0 \tag{5.9}$$

are fulfilled.
The solution can leave the $\mathcal{S}$ manifold again if $\alpha = 0$ or $\alpha = 1$. This corresponds to a sign change of one of the auxiliary switching functions $D\sigma_+$ or $D\sigma_-$ (see case 4 in Fig. 5.2).

Treatment of Inconsistent Switching by Three-Value Switching Logic

The treatment of inconsistent switching makes it necessary to expand the classic two-value switching logic ($s < 0, s > 0$) to three-value logic ($s < 0, s > 0, s = 0$). Here s stands for the sign of σ: $s = \operatorname{sgn} \sigma$, and $s = 0$ is assumed if $D\sigma_+ < 0$, $D\sigma_- > 0$ applies. This results in the switching logic shown in Fig. 5.3.

5.2 Sensitivity Analysis of Derivative Generation in Forward Mode

In this section we generate the derivative calculation by using the variational differential equations in forward mode. For the practical packages with the implementation in `Matlab`, readers can see on the work of SÖMMER et al., cf. [70]. On the other hand, for the backward differentiation formulas, readers can see in [2].

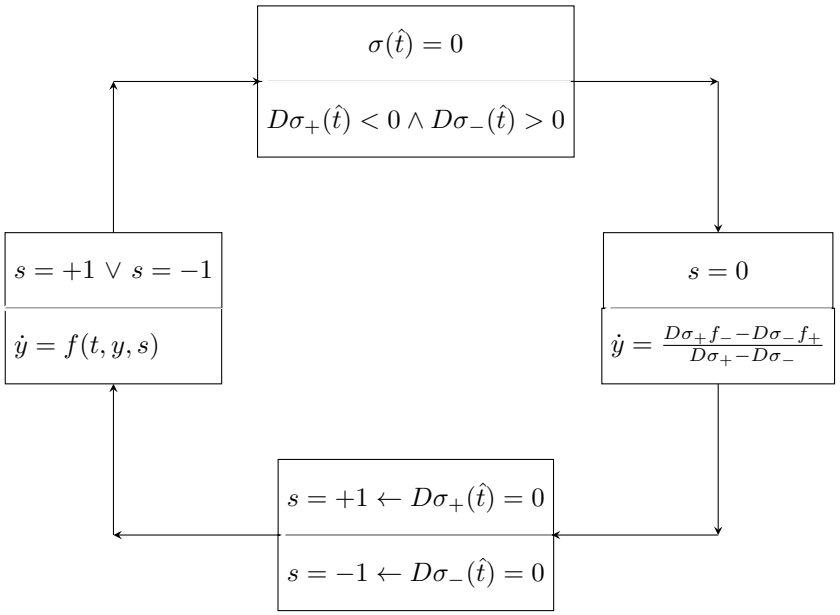


Figure 5.3: The treatment of inconsistent switching by a three-value switching logic.

5.2.1 Sensitivity Updates

By using END, the sensitivities of a switched IVP, where the right hand side has implicit discontinuities, can be obtained. However, in general, solving the VDEs (2.52) and (2.53) leads to wrong results, cf. [75, pp. 43]. One can overcome this problem by employing updates, whenever a switch occurs.
For the rest of this section, we assume that there is only one switch $t_{sw} \in (t_0, t_f)$.

Definition 37. Consider the IVP (2.41) with a single switch at t_{sw}. With $\varepsilon > 0$, at t_{sw} we define

$$G_x^-(t_{sw}; t_0, x_0, p) := \lim_{\varepsilon \to 0} G_x(t_{sw} - \varepsilon; t_0, x_0, p), \tag{5.10}$$

$$G_p^-(t_{sw}; t_0, x_0, p) := \lim_{\varepsilon \to 0} G_p(t_{sw} - \varepsilon; t_0, x_0, p), \tag{5.11}$$

and analogously

$$G_x^+(t_{sw}; t_0, x_0, p) := \lim_{\varepsilon \to 0} G_x(t_{sw} + \varepsilon; t_0, x_0, p), \tag{5.12}$$

$$G_p^+(t_{sw}; t_0, x_0, p) := \lim_{\varepsilon \to 0} G_p(t_{sw} + \varepsilon; t_0, x_0, p), \tag{5.13}$$

We consider the case with discontinuity only in the right hand side, i.e., there are switches, but no jumps. Then, cf. [75, sec. 4.4.3], one can derive the formulas for the sensitivities calculation using updates and for $t \in [t_0, t_f]$ it holds

$$G_x(t; t_0, x_0, p) = G_x(t; t_{sw}, x_+, p) U_x G_x^-(t_{sw}; t_0, x_0, p), \tag{5.14}$$

$$G_p(t; t_0, x_0, p) = G_x(t; t_{sw}, x_+, p) \left(U_x G_p^-(t_{sw}; t_0, x_0, p) + U_p \right) + G_p(t; t_{sw}, x_+, p) \tag{5.15}$$

therein $G_x^-(t_{sw}; t_0, x_0, p)$ and $G_p^-(t_{sw}; t_0, x_0, p)$ are defined according to (5.10) and (5.11), respectively, while x_+ is defined in (2.43).
U_x and U_p are the update matrices and can be calculated as follows

$$U_x = I_{n_x} + \delta \frac{\frac{\partial \sigma_j}{\partial x_-}}{\frac{d\sigma_j}{dt_s}}, \tag{5.16}$$

$$U_p = \delta \frac{\frac{\partial \sigma_j}{\partial p}}{\frac{d\sigma_j}{dt_s}}, \tag{5.17}$$

therein $I_{n_x} \in \mathbb{R}^{n_x \times n_x}$ is the identity matrix, δ and x_- are defined as in (2.46) and in (2.42), respectively. σ_j is the component of the switching function which zero-crossing caused the switch at t_{sw}, $\frac{\partial \sigma_j}{\partial x_-}$ and $\frac{d\sigma_j}{dt_s}$ denote the derivatives of σ_j w.r.t. x and w.r.t. t evaluated at x_- and t_{sw}, respectively, while $\frac{\partial \sigma_j}{\partial p}$ denotes the derivatives of σ_j w.r.t. p.
Since a switching point is a source of discontinuity, care must be taken when determining the sensitivities at $t = t_{sw}$. A possibility to define the sensitivities at the switching point is by using the updates. For more details, the following lemma will describe the relations between G_x^+ and G_x^-, G_p^+ and G_p^- at t_{sw} with initial values t_0, x_0.

Lemma 10. *The following equalities are hold true:*

$$G_x^+(t; t_0, x_0, p) = U_x G_x^-(t_{sw}; t_0, x_0, p), \tag{5.18}$$

$$G_p^+(t; t_0, x_0, p) = U_x G_p^-(t_{sw}; t_0, x_0, p) + U_p. \tag{5.19}$$

Proof. Using the definition 5 to x at the switching time $t = t_{sw}$, one gets $G_x^+(t_{sw}; t_{sw}, x_+, p) = I_{n_x}$ and $G_p^+(t_{sw}; t_{sw}, x_+, p) = 0_{n_x \times n_p}$. The proof is done by substituting t by t_{sw} into (5.14-5.15). □

5.2.2 Extension to Finitely Many Switches

Assumed there are n_{sw} switches on $[t_0, t_f]$ that occur at the time points $t_{sw}^{(i)} \in (t_0, t_f)$, $i = 1, \ldots, n_{sw}$. Moreover, we set $t_{sw}^{(0)} := t_0$, and $t_{sw}^{(n_{sw}+1)} := t_f$. For the sensitivities at $t \in (t_{sw}^{(i)}, t_{sw}^{(i+1)})$, $i = 0, 1, \ldots, n_{sw}$ or $t = t_f$ the following formulas hold:

$$G_x(t; t_0, x_0, p) = G_x(t; t_{sw}^{(i)}, x_+^{(i)}, p) \prod_{j=1}^{i} U_x^{(j)} G_x^-(t_{sw}^{(j)}; t_{sw}^{(j-1)}, x_+^{(j-1)}, p) \tag{5.20}$$

and

$$G_p(t; t_0, x_0, p) = G_x(t; t_{sw}^{(i)}, x_+^{(i)}, p) \left(U_x^{(i)} G_p^-(t_{sw}^{(i)}; t_0, x_0, p) + U_p^{(i)} \right) + G_p(t; t_{sw}^{(i)}, x_+^{(i)}, p). \tag{5.21}$$

The matrix $G_p^-(t_{sw}^{(i)}; t_0, x_0, p)$ in (5.21) is determined by using the formula (5.21) again for $t = t_{sw}^{(i)}$ with the starting matrix $G_p^-(t_{sw}^{(1)}; t_0, x_0, p)$ is calculated according to the VDE (2.53). That means the following equation holds for $i = 2, \ldots, n_s$,

$$G_p^-(t_{sw}^{(i)}; t_0, x_0, p) = G_x^-(t_{sw}^{(i)}; t_{sw}^{(i-1)}, x_+^{(i-1)}, p) \left(U_x^{(i-1)} G_p^-(t_{sw}^{(i-1)}; t_0, x_0, p) + U_p^{(i)} \right) + G_p^-(t_{sw}^{(i)}; t_{sw}^{(i-1)}, x_+^{(i-1)}, p). \tag{5.22}$$

The update matrices $U_x^{(i)}$ and $U_p^{(i)}$ are calculated for each switch $t_{sw}^{(i)}$ from (5.14) and (5.15), respectively, and $x_+^{(i)}$ is defined the same as y_+ in (2.43) for every i with $t_{sw} = t_{sw}^{(i)}$.
To calculate the sensitivities at the time point of a switch, i.e., for $t = t_{sw}^{(i)}$, the formulas (5.20) and (5.21) can be generalized as follows

$$G_x^+(t; t_0, x_0, p) = \prod_{j=1}^{i} U_x^{(j)} G_x^-(t_{sw}^{(j)}; t_{sw}^{(j-1)}, x_+^{(j-1)}, p), \quad i = 0, 1, \ldots, n_{sw}, \tag{5.23}$$

$$G_p^+(t; t_0, x_0, p) = U_x^{(i)} G_p^-(t_{sw}^{(i)}; t_0, x_0, p) + U_p^{(i)}, \quad i = 0, 1, \ldots, n_{sw}, \tag{5.24}$$

where $G_p^-(t_{sw}^{(i)}; t_0, x_0, p)$, $i = 2, 3, \ldots, n_{sw}$, are calculated as in (5.22).

5.3 Other Approaches for Treating Switches

Other approaches to handling switches in SwOCP are shortly considered as follows. Gröbner basis approach to return switching strategy (0 to 1 or 1 to 0; 1 to -1 or -1 to 1). See more details in Appendix B.2 with general heuristic approach and some illustrated examples. Moreover, switches in cost functions, cf. [12], and jumps, see [78].

Chapter 6

Switched Optimal Control Problems with Dry Friction

Over the past decades, the dry (Coulomb) friction has been considered a mechanical testing problem for discontinuous dynamical systems or SwOCPs. Previously, in the 1970s, Carver [32] studied a simple friction example of the discontinuity motion of a sliding object with mass and frictional resistance, where the externally applied switches. Next, in the 1990s, the suspension with Coulomb friction was numerically studied in the discontinuous dynamical systems, cf. [49, Chap. 6]. Later, dry friction was considered on the optimal control system of material points in a straight line, cf. [54]. Nowadays, many mathematicians investigate switched OCPs with friction, e.g. see [18] and [95, Chap. 15] for three different friction models.
This chapter deals with the applicability of our approaches to SwOCP on the benchmark problem with friction, where the results are presented both analytically with LMP and numerically with `MUSCOD-II`. Section 6.1 considers the dry friction, therein an idea from discontinuous dynamics is employed. Subsequently, in Section 6.2, the optimal control of a point mass on a rough plane will be considered. Last, Section 6.3 considers the general framework to solve the OCP with dry friction of a system of material points in a straight line, therein our solution approach is presented based on the correct application of Filippov's rule and LMP.

6.1 Dry Friction with Filippov's Rule

Friction is a complex phenomenon that significantly impacts mechanical systems. While much is known about friction in specific situations, a universal understanding remains elusive, making it challenging to anticipate and manage. It's rarely absent in natural or business processes, and its presence often limits performance. To mitigate its negative effects, model-based friction compensation is widely used, which requires an accurate friction model to apply counteracting forces. This need has led to the development of numerous mathematical models to describe friction's influence on machine behavior, as it's an unavoidable force in the feedback control of moving systems, cf. [6].

We will focus on a specific group of frictional phenomena: dry friction between material

points in a straight line, and dry friction acts between the rough plane and the mass point.

Dry friction (Coulomb friction) systems, as well as general differential equations with switches or jumps on the right side, can lead to locations where a classical solution does not exist. FILIPPOV's generalized solution notion for differential equations provides a solution in this case. FILIPPOV's concept of solution provides a physically relevant answer for the special situation of dry friction.

6.1.1 A General Framework: Discontinuous Dynamics's Idea

Dry friction always opposes the relative motion and is proportional to the normal force of contact, regardless of the area of impact. Dry friction, see Fig. 6.1, is a type of friction that is only affected by the direction of the velocity and not by the magnitude of the velocity. It is modeled as a static map between velocity and friction force that depends on the sign of the velocity,

$$F = F_\mu \operatorname{sgn}(v).$$

The frictional force between two bodies is assumed to be proportional to the normal force N on the area between the bodies at the point of contact. The proportionality factor is the material- and speed-dependent coefficient of friction μ, which was assumed to be constant in the original model:

$$|F_\mu| = |\mu||N|.$$

For zero velocities the above dry friction depends upon the sign function definition.

Figure 6.1: Dry friction.

The frictional force is tangential to the friction surface and is opposed to the direction of movement of the body:

$$F_\mu = -\mu|N|c\operatorname{sgn}(c^T\dot{p}),$$

where c is a unit vector orthogonal to the friction surface, $c^T\dot{p}$ describes the tangential velocity along the sliding surface, the coefficient of friction μ is continuous, $\mu > 0$ applies. If the slip surface is modeled by the equation $\tilde{g}(p) = 0$, then $N = \tilde{G}^T\lambda$, $\tilde{G} = \frac{\partial \tilde{g}}{\partial p}$ applies, since $\tilde{G}^T\lambda$ is even describes the coercive force. This gives you the equation of motion

$$M\ddot{p} = f + G^T\lambda - \mu|N|c\operatorname{sgn}(c^T\dot{p}),$$
$$g(p) = 0,$$

where M is the mass. Here, inconsistent switching for the switching function $\sigma = c^T\dot{p}$ occurs because of

$$D\sigma_+ = c^T M^{-1}(f + G^T\lambda - \mu|N|c), \tag{6.1a}$$
$$D\sigma_- = c^T M^{-1}(f + G^T\lambda + \mu|N|c), \tag{6.1b}$$

when the forces in the direction of c are smaller than the frictional force

$$D\sigma_+ \cdot D\sigma_- < 0 \Leftrightarrow c^T M^{-1}(f + G^T\lambda) < \mu|N|c^T M^{-1}c,$$

where $D\sigma_+, D\sigma_-$ are directional derivatives of the switching function in the direction f_+,f_-, respectively, see Subsection 5.1.3. Filippov's generalized solution concept leads to

$$\dot{y} = \alpha f_+(t,y) + (1-\alpha)f_-(t,y), \tag{6.2a}$$
$$0 = \sigma(t,y), \tag{6.2b}$$

on the system

$$M\ddot{p} = f + G^T\lambda + (1-2\alpha)\mu|N|c, \quad \alpha \in [0,1], \tag{6.3a}$$
$$g(p) = 0, \tag{6.3b}$$
$$c^T\dot{p} = 0. \tag{6.3c}$$

From the claim $\dot{\sigma} = c^T\ddot{p}$, the system is obtained for the sticking phase

$$M\ddot{p} = f + G^T\lambda - \frac{c^T M^{-1}(f + G^T\lambda)}{c^T M^{-1}c}c, \tag{6.4a}$$
$$g(p) = 0. \tag{6.4b}$$

Because of $\sigma = c^T\dot{p} = 0$, there is no motion in the direction of c, i.e., the body is stuck. Alternatively, the system equations during the sticking phase can also be obtained by adding the equation $\sigma = c^T\dot{p} = 0$ and with an additional Lagrange multiplier λ_R:

$$M\ddot{p} = f + G^T\lambda + c\lambda_R, \tag{6.5a}$$
$$g(p) = 0, \tag{6.5b}$$
$$c^T\dot{p} = 0. \tag{6.5c}$$

Differentiation of (6.5c) yields $\lambda_R = -\frac{c^T M^{-1}(f+G^T\lambda)}{c^T M^{-1}c}$, so after elimination of λ_R, one gets (6.4) again. This equivalence is summarized in the following lemma.

Lemma 11. *For systems of form*

$$M\ddot{p} = f + G^T\lambda - \mu|N|c\,\mathrm{sgn}(c^T\dot{p})$$
$$g(p) = 0,$$

the treatment with Filippov*'s solution concept and the additional condition $c^T\dot{p} = 0$ as well as the treatment by inserting additional* Lagrange *multipliers lead to the same result.*

Two-mass oscillator

To illustrate this idea, consider a two-mass oscillator with dry friction between the bodies:

$$m_1\ddot{p}_x = f_1 - \mu\,\mathrm{sgn}(\dot{p}_x - \dot{p}_y),$$
$$m_2\ddot{p}_y = f_2 + \mu\,\mathrm{sgn}(\dot{p}_x - \dot{p}_y).$$

The switching function is $\sigma = \dot{p}_x - \dot{p}_y$. Here $c = \begin{pmatrix}1 & -1\end{pmatrix}^T$ applies. The shifting process is inconsistent when the external forces are not large enough to overcome the stiction:

$$D\sigma_+ \cdot D\sigma_- < 0 \Leftrightarrow \left|\frac{f_1}{m_1} - \frac{f_2}{m_2}\right| < \mu\left(\frac{1}{m_1} + \frac{1}{m_2}\right).$$

The differential equation for the sticking phase is obtained from system (6.4)

$$(m_1 + m_2)\ddot{p}_x = f_1 + f_2,$$
$$(m_1 + m_2)\ddot{p}_y = f_1 + f_2,$$

i.e., both bodies move together under the influence of the total force.

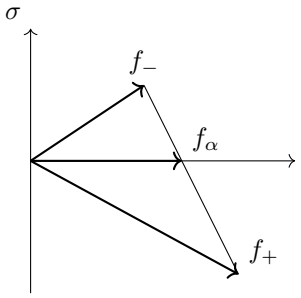
Remark 38 (Extension to the nonlinear case)**.** In the general case where the right-hand side of the differential equation depends non-linearly on $\mathrm{sgn}(\sigma)$, Lemma 11 no longer applies. In this case, the treatment with Filippov's concept of solution and with the help of additional Lagrange multipliers yield different results. They also differ from the treatment method proposed in [91]. This is shown below.
The system

$$\dot{y} = f(y, \mathrm{sgn}\,\sigma(y)) = \begin{cases} f_+(y) & \text{for } \sigma > 0 \\ f_-(y) & \text{for } \sigma < 0 \end{cases}, \quad f_+ \neq f_-,$$

be considered in the inconsistent case.

Case 1: Treatment with Filippov's concept of solution

$$\dot{y} = \alpha f_+ + (1-\alpha) f_- =: f_\alpha,\ \alpha \in [0,1],$$
$$0 = \sigma(y).$$



With that one gets

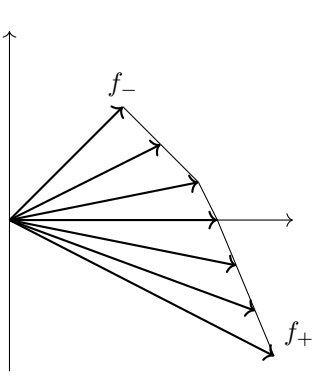
$$\dot{y} = \frac{D\sigma_+ f_- - D\sigma_- f_+}{D\sigma_+ - D\sigma_-}.$$

Case 2: Convexification within f

In [91] values from $[-1, 1]$ are allowed for the sign $s = \operatorname{sgn}(\sigma(y))$, in contrast to Filippov the convexification is carried out within f:

$$\dot{y} = f(y, s)$$
$$0 = \sigma(y).$$

If $\sigma \neq 0$, then $\dot{\sigma} = \sigma(y, s) = 0$ can be resolved to s. (Conditions for $s \in [-1, 1]$ can be found in [91].)



Case 3: Treatment with additional Lagrangian multipliers

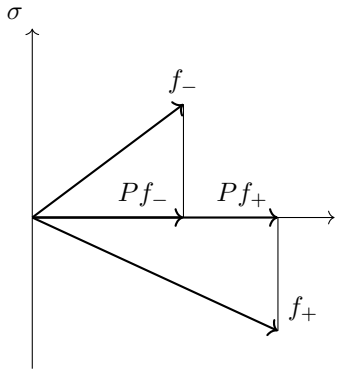
The new right-hand side is thus the projection of f onto $\sigma = 0$. Note that in the figure,

The treatment with additional Lagrangian multipliers μ gives nonsensical, ambiguous results here: Solve the system

$$\dot{y} = f(y, \operatorname{sgn}(\sigma)) + \frac{\partial \sigma}{\partial y}^T \mu$$
$$0 = \sigma(y),$$

after $\mu = -(\frac{\partial \sigma}{\partial y}(\frac{\partial \sigma}{\partial y})^T)^{-1} \frac{\partial \sigma}{\partial y} f$, this results after insertion

$$\dot{y} = \left(I - \left(\frac{\partial \sigma}{\partial y} \right)^T \left(\frac{\partial \sigma}{\partial y} \left(\frac{\partial \sigma}{\partial y} \right)^T \right)^{-1} \frac{\partial \sigma}{\partial y} \right) f.$$



Pf_+ and Pf_- denote for the respectively projection of f_+ and f_- onto $\sigma = 0$.

6.2 Optimal Control of a Point Mass on a Rough Plane

In this section, we consider solution approaches for optimal control of a point mass on a rough plane, both analytically and numerically. The model in this section (see Subsection 6.2.1) is proposed by Prof. N. Bolotnik in our private communication, see [24].

6.2.1 Mathematical Model of a Mechanical System

Consider a mass point that moves in a rigid inclined plane Π under the action of a control force $\mathbf{F}$, applied to this point. Coulomb's dry friction acts between the plane and the mass point. The contact between the mass point and the plane is represented by a unilateral constraint that prevents the mass point from passing through the plane but does not prevent

from separating from it. This fact is reflected in that the normal reaction force can be directed only to one side with respect to the plane. Introduce two right-handed rectangular coordinate systems: $OXYZ$ and $Oxyz$. The point O lies in the plane Π, the X-axis is horizontal, the Z axis points vertically upward, the x-axis coincides with the X-axis, and the coordinate plane xy lies in the plane Π. In this case, the angle γ is formed by the axes Y and y in the inclination angle of the plane. We assume that $\gamma \in [0, \pi/2)$ and that the point mass can physically move in the half-space $z \geq 0$.
Derive the equations of motion of the mass point in the plane Π. Let m denote the mass of the mass point; x and y the coordinates of this point in the plane Π; f_x, f_y, and f_z the components of the control force $\mathbf{F}$ in the coordinate system $Oxyz$; k the coefficient of the dry friction between the plane Π and the mass point; g the magnitude of the acceleration due to gravity.
Coulomb's friction force $\mathbf{R}$ acted on the mass point by the plane Π is defined as follows:

$$\mathbf{R} = \begin{cases} -kN\frac{\mathbf{v}}{\|\mathbf{v}\|}, & \mathbf{v} \neq 0, \\ -\mathbf{\Phi}, & \mathbf{v} = 0, \quad \|\mathbf{\Phi}\| \leq kN, \\ -kN\frac{\mathbf{\Phi}}{\|\mathbf{\Phi}\|}, & \mathbf{v} = 0, \quad \|\mathbf{\Phi}\| > kN, \end{cases} \tag{6.6}$$

where $\mathbf{v}$ is the velocity of the motion of the mass point in the plane Π, N the magnitude of the normal reaction of the plane on the mass point, $\mathbf{\Phi}$ the projection onto the plane Π of the resultant of the impressed forces applied to the mass point. By impressed forces, we understand all forces applied to the mass point, apart from the forces of interaction of this point with the plane, i.e., apart from the friction force and the normal reaction force. If $\mathbf{\Phi}_a$ is the resultant of the impressed forces applied to the point mass, then the force $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi} = \mathbf{\Phi}_a - (\mathbf{\Phi}_a, \mathbf{e}_z)\mathbf{e}_z, \tag{6.7}$$

where $\mathbf{e}_z$ is the unit vector of the z-axis of the coordinate system $Oxyz$.
In the case under consideration, the impressed forces are the control force $\mathbf{F}$ and the gravity force $m\mathbf{g}$, where $\mathbf{g}$ is the vector of the acceleration due to gravity. In the coordinate system $Oxyz$, we have

$$\mathbf{\Phi}_a = \mathbf{F} + m\mathbf{g} = \begin{bmatrix} f_x & f_y & f_z \end{bmatrix}^T + \begin{bmatrix} 0 & -mg\sin\gamma & -mg\cos\gamma \end{bmatrix}^T, \tag{6.8}$$

$$\mathbf{\Phi} = \begin{bmatrix} f_x & f_y - mg\sin\gamma & 0 \end{bmatrix}^T. \tag{6.9}$$

The components of the friction force in the plane Π are defined as follows:

$$\mathbf{R}_x = \begin{cases} -kN\frac{v_x}{\|\mathbf{v}\|}, & \text{if } \|\mathbf{v}\| \neq 0, \\ -f_x, & \text{if } \|\mathbf{v}\| = 0, \ \|\mathbf{\Phi}\| \leq kN, \\ -kN\frac{f_x}{\|\mathbf{\Phi}\|}, & \text{if } \|\mathbf{v}\| = 0, \ \|\mathbf{\Phi}\| > kN, \end{cases} \tag{6.10}$$

$$\mathbf{R}_y = \begin{cases} -kN\frac{v_y}{\|\mathbf{v}\|}, & \text{if } \|\mathbf{v}\| \neq 0, \\ -f_y + mg\sin\gamma, & \text{if } \|\mathbf{v}\| = 0, \|\mathbf{\Phi}\| \leq kN, \\ -kN\frac{f_y - mg\sin\gamma}{\|\mathbf{\Phi}\|}, & \text{if } \|\mathbf{v}\| = 0, \|\mathbf{\Phi}\| > kN, \end{cases} \tag{6.11}$$

where

$$\|\mathbf{\Phi}\| = \sqrt{f_x^2 + (f_y - mg\sin\gamma)^2}, \qquad \|\mathbf{v}\| = \sqrt{v_x^2 + v_y^2} = \sqrt{\dot{x}^2 + \dot{y}^2}. \tag{6.12}$$

In the vector form, the equations of motion of the mechanical system under consideration are given by

$$\dot{\mathbf{r}} = \mathbf{v}, \quad m\dot{\mathbf{v}} = \mathbf{F} + m\mathbf{g} + \mathbf{N} + \mathbf{R}, \tag{6.13}$$

where $\mathbf{r}$ is the position vector of the point mass relative to the point O and $\mathbf{N}$ is the vector of the normal reaction of the plane. In the coordinate system $Oxyz$, these vectors are represented as follows:

$$\mathbf{r} = \begin{bmatrix} x & y & z \end{bmatrix}^T, \quad \mathbf{N} = \begin{bmatrix} 0 & 0 & N \end{bmatrix}^T. \tag{6.14}$$

In the coordinate form, the equations of (6.13) become

$$\begin{aligned} m\ddot{x} &= f_x + R_x, \\ m\ddot{y} &= f_y - mg\sin\gamma + R_y, \\ 0 &= f_z - mg\cos\gamma + N. \end{aligned} \tag{6.15}$$

The last equation takes into account the fact that the mass point is constrained to move in the plane Π and, therefore, $z \equiv 0$.
Solve the last equation of (6.15) for N to obtain

$$N = -f_z + mg\cos\gamma. \tag{6.16}$$

Since the mass point is kept in the half-space $z \geq 0$ and the constraint is unilateral, the normal reaction force cannot be oriented in the negative direction of the z-axis; hence, we have

$$N \geq 0, \tag{6.17}$$

Inequality (6.17) and expression (6.16) imply the upper bound for the z-component of the control force:

$$f_z \leq mg\cos\gamma. \tag{6.18}$$

If this inequality violates, the mass point will separate from the plane Π, which is not allowed.

6.2.2 Optimal Control Problem

For the system

$$\begin{aligned} m\ddot{x} &= f_x + R_x, \\ m\ddot{y} &= f_y - mg\sin\gamma + R_y, \end{aligned} \tag{6.19}$$

where R_x and R_y are defined by expression (6.10) and (6.11) for $N = mg\cos\gamma - f_z$, find a control force $\mathbf{F}$ with the components f_x, f_y, and f_z that satisfies the constraints

$$f_x^2 + f_y^2 + f_z^2 \leq U^2, \tag{6.20}$$

$$f_z \leq mg\cos\gamma, \tag{6.21}$$

and transfers the system in a minimal time T from a given initial state

$$x(0) = x_0, \quad y(0) = y_0, \quad \dot{x}(0) = \dot{x}_0, \quad \dot{y}(0) = \dot{y}_0 \tag{6.22}$$

to a given terminal state

$$x(T) = x_T, \quad y(T) = y_T, \quad \dot{x}(T) = \dot{x}_T, \quad \dot{y}(T) = \dot{y}_T. \tag{6.23}$$

Inequality (6.20) constrains the control force in the absolute value. Inequality (6.21) requires that the z-component of this force does not exceed the magnitude of the projection of the gravity force onto the normal to the underlying plane; otherwise, it is impossible to provide a non-negative value for the normal reaction N.

Now, we can rewrite the system (6.19-6.23) as the following OCP

$$\begin{aligned} \min_{\mathbf{r}(\cdot),\mathbf{v}(\cdot),\mathbf{F}(\cdot)} \quad & T \\ \text{s.t.} \quad & \dot{x} = v_x, \quad \dot{y} = v_y, \\ & m\dot{v}_x = f_x + R_x, \quad m\dot{v}_y = f_y - mg\sin\gamma + R_y, \\ & f_x^2 + f_y^2 + f_z^2 \le U^2, \quad f_z \le mg\cos\gamma, \\ & x(0) = x_0, y(0) = y_0, \quad v_x(0) = v_{x_0}, v_y(0) = v_{y_0}, \\ & x(T) = x_T, y(T) = y_T, \quad v_x(T) = v_{x_T}, v_y(T) = v_{y_T}, \end{aligned} \tag{6.24}$$

where $\mathbf{r}$, R_x and R_y are defined by expression (6.14), (6.10) and (6.11), respectively, for $N = mg\cos\gamma - f_z$, the control force $\mathbf{F} = \begin{bmatrix} f_x & f_y & f_z \end{bmatrix}^T$, and the velocity $\mathbf{v} = \begin{bmatrix} v_x & v_y & 0 \end{bmatrix}^T$.

6.2.3 Reformulation

We will solve (6.24) by using Filippov's rule and the local minimum principle. We can rewrite ODEs in problem (6.24), i.e., $m\dot{v}_x = f_x + R_x$, and $m\dot{v}_y = f_y - mg\sin\gamma + R_y$, as follows

$$\dot{\mathbf{v}}(t) = \begin{cases} h_+(\cdot), & \text{if } \sigma_1 > 0, \\ h_0(\cdot), & \text{if } \sigma_1 = 0,\ \sigma_2 \le 0, \\ h_-(\cdot), & \text{if } \sigma_1 = 0,\ \sigma_2 > 0, \end{cases} \tag{6.25}$$

where switching functions $\sigma_1 := \|\mathbf{v}\| \ge 0$ (since (6.12)), $\sigma_2 := \|\mathbf{\Phi}\| - kN$; and

$$\begin{aligned} h_+(\cdot) &= \frac{1}{m}\begin{bmatrix} f_x - kN\frac{v_x}{\|\mathbf{v}\|} \\ f_y - mg\sin\gamma - kN\frac{v_y}{\|\mathbf{v}\|} \\ f_z - mg\cos\gamma + N \end{bmatrix} = \frac{1}{m}\begin{bmatrix} f_x - kN\frac{v_x}{\|\mathbf{v}\|} \\ f_y - mg\sin\gamma - kN\frac{v_y}{\|\mathbf{v}\|} \\ 0 \end{bmatrix}, \\ h_0(\cdot) &= \frac{1}{m}\begin{bmatrix} 0 & 0 & f_z - mg\cos\gamma + N \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T, \\ h_-(\cdot) &= \frac{1}{m}\begin{bmatrix} f_x - kN\frac{f_x}{\|\mathbf{\Phi}\|} \\ f_y - mg\sin\gamma - kN\frac{f_y - mg\sin\gamma}{\|\mathbf{\Phi}\|} \\ 0 \end{bmatrix}, \end{aligned}$$

First, (6.25) is rewritten by Filippov's rule in the following relaxed reformulation

$$\dot{\mathbf{v}}(t) = \sum_{j\in\mathcal{J}} \alpha_j(t) h_j\left(\mathbf{v}(t), \mathbf{F}(t)\right), \quad \sum_{j\in\mathcal{J}} \alpha_j(t) = 1, \alpha_j(t) \in [0,1],\ j \in \mathcal{J},$$

together with additional mixed constraints $\mathcal{G}(\alpha, \mathbf{v}, \mathbf{F}) \leq 0$, $g(\alpha, \mathbf{v}) = 0$, where

$$\begin{aligned} &\mathcal{G}_1(\cdot) := -\alpha_+(t)\sigma_1, \\ &g_1(\cdot) := \alpha_0(t)\sigma_1, \qquad \mathcal{G}_2(\cdot) := \alpha_0(t)\sigma_2, \\ &g_2(\cdot) := \alpha_-(t)\sigma_1, \qquad \mathcal{G}_3(\cdot) := -\alpha_-(t)\sigma_2. \end{aligned} \tag{6.26}$$

Some following cases are considered:
1st case: $\sigma_1 > 0$ (i.e., $\|\mathbf{v}\| > 0$) . Then the relaxed-additional constraints imply $\alpha_0(t) = \alpha_-(t) = 0$, so a unique possible solution is $\alpha_-(t) = \alpha_0(t) = 0$, and $\alpha_+(t) = 1$. Here $\mathcal{G}_2$ and $\mathcal{G}_3$ are active constraints, $\mathcal{G}_1$ is inactive one.
2nd case: $\sigma_1 = 0$, $\sigma_2 \leq 0$ (i.e., $\|\mathbf{v}\| = 0$, $\|\mathbf{\Phi}\| - kN \leq 0$). Then, from the relaxed-additional constraints, one obtains $\alpha_-(t) = 0$, and these constraints are satisfied for $\alpha_0(t) \in [0, 1]$, $\alpha_+(t) \in [0, 1]$ such that $\alpha_0(t) + \alpha_+(t) = 1$. Here $\mathcal{G}_j$, $j = 1, 3$, are active constraints, while $\mathcal{G}_2$ is inactive constraint if $\|\mathbf{\Phi}\| - kN \neq 0$ and $\alpha_0 \neq 0$.
3rd case: $\sigma_1 = 0$, $\sigma_2 > 0$ (i.e., $\|\mathbf{v}\| = 0$, $\|\mathbf{\Phi}\| - kN > 0$). Then, these constraints imply $\alpha_0(t) = 0$, and these constraints are satisfied for $\alpha_+(t) \in [0, 1]$, $\alpha_-(t) \in [0, 1]$ such that $\alpha_+(t) + \alpha_-(t) = 1$. Here $\mathcal{G}_j$, $j = 1, 2$, are active constraints, and $\mathcal{G}_3$ is inactive one if $\|\mathbf{\Phi}\| - kN \neq 0$ and $\alpha_- \neq 0$.
The phase points set is determined by

$$\mathcal{N}(\mathcal{G}) = \{(\mathbf{r}, \mathbf{v}, \mathbf{F}, \alpha) \mid \sigma_1 = 0, \sigma_2 = 0\} = \{(\mathbf{r}, \mathbf{v}, \mathbf{F}, \alpha) : \|\mathbf{v}\| = 0, \|\mathbf{\Phi}\| = kN\} \neq \emptyset.$$

(For e.g., $v_x = v_y = 0$, $f_x = 0$, $f_y = f_z$, $\cos\gamma = \sin\gamma$ and $k = 1$)
The corresponding phase jump is then determined by

$$s(t) = -a_1\alpha_+(t)\frac{\partial\sigma_1}{\partial(\mathbf{r}, \mathbf{v})} + (b_1\alpha_0(t) + b_2\alpha_-(t))\frac{\partial\sigma_1}{\partial(\mathbf{r}, \mathbf{v})}, \quad a_1 \geq 0, b_1, b_2 \in \mathbb{R},$$

where $\sigma_1 = \|\mathbf{v}\| = \sqrt{v_x^2 + v_y^2} = \sqrt{\dot{x}^2 + \dot{y}^2}$, $\mathbf{r} = \begin{bmatrix} x & y & 0 \end{bmatrix}^T$, and $\mathbf{v} = \begin{bmatrix} v_x & v_y & 0 \end{bmatrix}^T$.
Therefore, problem (6.24) is reformulated as follows

$$\begin{aligned} \min_{\mathbf{r}(\cdot), \mathbf{v}(\cdot), \mathbf{F}(\cdot), \alpha(\cdot)} \quad & T \\ \text{s.t.} \quad & \dot{\mathbf{r}}(t) = \mathbf{v}(t), \quad \dot{\mathbf{v}}(t) = h(\mathbf{v}, \mathbf{F}), \\ & \mathcal{G}_j(\alpha, \mathbf{v}, \mathbf{F}) \leq 0,\ j = 1, 2, 3, \quad g_i(\alpha, \mathbf{v}) = 0,\ i = 1, 2, \\ & \mathcal{G}_j^f(\mathbf{F}) \leq 0,\ j = 1, 2, \\ & x(0) = x_0, y(0) = y_0,\ x(0) = v_{x_0}, v_y(0) = v_{y_0}, \\ & x(T) = x_T, y(T) = y_T,\ v_x(T) = v_{x_T}, v_y(T) = v_{y_T}, \\ & \textstyle\sum_{j \in \mathcal{J}} \alpha_j(t) = 1, \alpha_j(t) \in [0, 1],\ j \in \mathcal{J}, \end{aligned} \tag{6.27}$$

where $\mathcal{G}_1^f(\mathbf{F}) := f_x^2 + f_y^2 + f_z^2 - U^2$, $\mathcal{G}_2^f(\mathbf{F}) := f_z - mg\cos\gamma$, and $h(\cdot)$ is defined by

$$h(\mathbf{v}, \mathbf{F}) := \sum_{j \in \mathcal{J}} \alpha_j(t) h_j(\mathbf{v}, \mathbf{F}). \tag{6.28}$$

with $\mathcal{G}_j$, $j \in \{1, 2, 3\}$, are active constraints, depending on $\|\mathbf{v}\|$.

6.2.4 A Solution Approach with LMP

Let us define the augmented PONTRYAGIN function and the endpoint Lagrange function for problem (6.27),

$$\bar{\mathcal{H}} = \mathcal{H} + \sum_{j=1}^{3} \mu_j(t)\mathcal{G}_j(\alpha, \mathbf{v}, \mathbf{F}) + \sum_{j=1}^{2} \mu_j^f(t)\mathcal{G}_j^f(\mathbf{F}) + \sum_{i=1}^{2} \mu_i^g g_i(\alpha, \mathbf{v}),$$
$$L(\nu, ini) = \nu^T m(ini),$$

where adjoint $\lambda(t) \in \mathbb{R}^{6*}$, integrable functions $0 \leq \mu(t) \in \mathbb{R}^{3*}$, $\mu^g(t) \in \mathbb{R}^{2*}$, Lagrange multipliers $\nu \in \mathbb{R}^{3*}$, $\mathcal{H} := \lambda^T(t)\begin{pmatrix}\mathbf{v}(t) & h(\mathbf{v}, \mathbf{F})\end{pmatrix}^T$, mixed constrains $\mathcal{G}$, g are given by (6.26), $ini := (x_0, y_0, v_{x_0}, v_{y_0}, x_T, y_T, v_{x_T}, v_{y_T})$, and

$$m(ini) := \begin{pmatrix} x(0) & y(0) & 0 \\ x(T) & y(T) & 0 \\ v_x(0) & v_y(0) & 0 \\ v_x(T) & v_y(T) & 0 \end{pmatrix}^T.$$

Supposed that $(\hat{\mathbf{r}}, \hat{\mathbf{v}}, \hat{\mathbf{F}}, \hat{\alpha})$ is a weak local minimum in problem (6.27), then it satisfies the LMP in Theorem 9, i.e., there exists multipliers: $\hat{\nu} \in \mathbb{R}^{3*}$, $\hat{\lambda} \in BV\left([0,T], \mathbb{R}^{6*}\right)$, $\hat{\mu} \in L^1\left([0,T], \mathbb{R}^{3*}\right)$, $\hat{\mu}^f \in L^1\left([0,T], \mathbb{R}^{2*}\right)$, $\hat{\mu}^g \in L^1\left([0,T], \mathbb{R}^{2*}\right)$, $d\hat{\eta} \in (C([0,T], \mathbb{R}))^*$, such that

$$\hat{\nu} \geq 0, \quad \hat{\nu}^T m(r_{ini}) = 0,$$
$$\hat{\mu} \geq 0, \quad \hat{\mu}\mathcal{G}(\hat{\alpha}, \hat{\mathbf{v}}, \hat{\mathbf{F}}) = 0, \quad d\hat{\eta} \geq 0, \quad d\hat{\eta}\,\chi_{\mathcal{D}} = d\hat{\eta},$$
$$|\hat{\nu}| + \|\hat{\mu}\|_1 + \int_{[0,T]} d\hat{\eta} > 0,$$

where $\mathcal{D} := \left\{t \in [0,T] : \text{clm}\,(\hat{\mathbf{r}}, \hat{\mathbf{v}}, \hat{\mathbf{F}}, \hat{\alpha})(t) \cap \mathcal{N}(\mathcal{G}) \neq \varnothing\right\}$, and a $d\hat{\eta}$-measurable essentially bounded function $\hat{s} : [0,T] \to \mathbb{R}^{6*}$ such that

$$\hat{s}(t) \in \text{conv}\ S(\text{clm}\,((\hat{\mathbf{r}}, \hat{\mathbf{v}}), (\hat{\alpha}, \hat{\mathbf{F}}))(t)) \quad \text{for almost all } t \text{ in } d\hat{\eta}\text{–measure},$$

there hold the following adjoint equation in terms of measure

$$-d\hat{\lambda} = \hat{\lambda}^T \frac{\partial \mathcal{H}}{\partial(\mathbf{r}, \mathbf{v})} dt + \hat{\mu}^T \frac{\partial \mathcal{G}}{\partial(\mathbf{r}, \mathbf{v})} dt + (\hat{\mu}^g)^T \frac{\partial g}{\partial(\mathbf{r}, \mathbf{v})} dt + \hat{s} d\hat{\eta}, \tag{6.29}$$

the tranversality conditions:

$$\hat{\lambda}(0-) = -\frac{\partial L(\hat{\nu}, ini)}{\partial(\mathbf{r}_0, \mathbf{v}_0)}, \quad \hat{\lambda}(T-) = -\frac{\partial L(\hat{\nu}, ini)}{\partial(\mathbf{r}_T, \mathbf{v}_T)}, \tag{6.30}$$

where $\mathbf{r}_0 = (x(0), y(0))$, $\mathbf{r}_T = (x(T), y(T))$, $\mathbf{v}_0 = (v_x(0), v_y(0))$, $\mathbf{v}_T = (v_x(T), v_y(T))$, and the stationary condition w.r.t. the control:

$$\mathcal{H}(\hat{\mathbf{v}}, \hat{\mathbf{F}}, \hat{\alpha}, \lambda) = \max_{\|\mathbf{F}\| \leq U, 0 \leq \alpha \leq 1} \mathcal{H}(\hat{\mathbf{v}}, \mathbf{F}, \alpha, \lambda) \tag{6.31}$$

To simplify the problem, consider the particular case where the particle is constrained to move on a line of maximum inclination and the control force acts in the vertical plane that passes through the line of motion of the particle, where the initial and terminal positions of the particle lie on a common line of maximum inclination and if, in addition, the initial and terminal velocities are parallel to this line, i.e., one has $\gamma \to \frac{\pi}{2}$, so

$$\sin\gamma = 1,\ \cos\gamma = 0, \quad \hat{f}_z = 0. \tag{6.32}$$

Subsequent, the stationary condition (6.31) implies

$$\begin{aligned}\mathcal{H}(\hat{\mathbf{v}}, \hat{\mathbf{F}}, \hat{\alpha}, \lambda) &= \lambda_1 v_x + \lambda_2 v_y + \max_{\|\mathbf{F}\|\le U, 0\le\alpha\le 1}\{\lambda_4(\alpha_+ + \alpha_-)\frac{f_x}{m} + \lambda_5(\alpha_+ + \alpha_-)\frac{f_y - mg}{m}\}\\ &= \lambda_1 v_x + \lambda_2 v_y + \max_{\|\mathbf{F}\|\le U, 0\le\alpha\le 1}\{f_x\frac{\lambda_4(1-\alpha_0)}{m} + f_y\frac{\lambda_5(1-\alpha_0)}{m}\} + \max_{0\le\alpha\le 1}\{\lambda_5 g(1-\alpha_0)\}.\end{aligned}$$

Here we obtain

$$\max_{0\le\alpha\le 1}\{\lambda_5 g(1-\alpha_0)\} = \begin{cases}\lambda_5 g, & \text{if } \lambda_5 > 0,\\ 0, & \text{if } \lambda_5 \le 0,\end{cases}, \quad \text{where } \hat{\alpha}_0 = \begin{cases}0, & \text{if } \lambda_5 > 0,\\ 1, & \text{if } \lambda_5 \le 0,\end{cases} \tag{6.33}$$

$$\max_{\|\mathbf{F}\|\le U, 0\le\alpha\le 1}\{f_x\frac{\lambda_4(1-\alpha_0)}{m} + f_y\frac{\lambda_5(1-\alpha_0)}{m}\} = \begin{cases}0, & \text{if } \lambda_4 = \lambda_5 = 0,\\ |U|(\lambda_4^2+\lambda_5^2)^{1/2}, & \text{otherwise,}\end{cases}$$

therein,

$$\hat{f}_x = |U|\frac{|\lambda_4|}{(\lambda_4^2+\lambda_5^2)^{1/2}},\ \hat{f}_y = |U|\frac{|\lambda_4|}{(\lambda_4^2+\lambda_5^2)^{1/2}},\ \hat{\alpha}_0 = 0, \quad \text{if } \lambda_4^2 + \lambda_5^2 > 0. \tag{6.34}$$

Consider the case $\sigma_1 = 0$, $\sigma_2 \le 0$ (i.e., $\|\hat{\mathbf{v}}\| = 0$, $\|\mathbf{\Phi}\| - kN \le 0$). Here ones have $\alpha_- = 0$, $\mathcal{G}_j$, $j = 1, 2, 3$, are active constraints with the phase points set $\mathcal{N}(\mathcal{G}) \ne \emptyset$, and the corresponding phase jump is

$$\hat{s}(t) = (b_1\hat{\alpha}_0 - a_1\hat{\alpha}_+)\frac{\partial\sigma_1}{\partial(\mathbf{r},\mathbf{v})} = \hat{\Sigma}\begin{pmatrix}\frac{\partial\|\hat{\mathbf{v}}\|}{\partial x} & \frac{\partial\|\hat{\mathbf{v}}\|}{\partial y} & 0 & \frac{\partial\|\hat{\mathbf{v}}\|}{\partial v_x} & \frac{\partial\|\hat{\mathbf{v}}\|}{\partial v_y} & 0\end{pmatrix}^T, \quad b_1 \in \mathbb{R},$$

where $\alpha_0 + \alpha_+ = 1$, $a_1 \ge 0$, and $\hat{\Sigma} := b_1\hat{\alpha}_0 - a_1\hat{\alpha}_+$. The adjoint equation (6.29) implies

$$\begin{aligned}
-d\hat{\lambda}_1 &= \hat{\lambda}_1\frac{\partial\hat{v}_x}{\partial x}dt + (\hat{\mu}_1^g\hat{\alpha}_0 - \hat{\mu}_1\hat{\alpha}_+)\frac{\partial\|\hat{\mathbf{v}}\|}{\partial x}dt + \hat{\Sigma}\frac{\partial\|\hat{\mathbf{v}}\|}{\partial x}d\hat{\eta},\\
-d\hat{\lambda}_2 &= \hat{\lambda}_2\frac{\partial\hat{v}_y}{\partial x}dt + (\hat{\mu}_1^g\hat{\alpha}_0 - \hat{\mu}_1\hat{\alpha}_+)\frac{\partial\|\hat{\mathbf{v}}\|}{\partial y}dt + \hat{\Sigma}\frac{\partial\|\hat{\mathbf{v}}\|}{\partial y}d\hat{\eta},\\
-d\hat{\lambda}_3 &= 0,\\
-d\hat{\lambda}_4 &= \hat{\lambda}_1 dt + (\hat{\mu}_1^g\hat{\alpha}_0 - \hat{\mu}_1\hat{\alpha}_+)\frac{\partial\|\hat{\mathbf{v}}\|}{\partial v_x}dt + \hat{\Sigma}\frac{\partial\|\hat{\mathbf{v}}\|}{\partial v_x}d\hat{\eta},\\
-d\hat{\lambda}_5 &= \hat{\lambda}_2 dt + (\hat{\mu}_1^g\hat{\alpha}_0 - \hat{\mu}_1\hat{\alpha}_+)\frac{\partial\|\hat{\mathbf{v}}\|}{\partial v_y}dt + \hat{\Sigma}\frac{\partial\|\hat{\mathbf{v}}\|}{\partial v_x}d\hat{\eta},\\
-d\hat{\lambda}_6 &= 0.
\end{aligned} \tag{6.35}$$

The tranversality conditions (6.30) lead to $\hat{\lambda}(0-) = 0$, and $\hat{\lambda}(T-) = 0$. Thus, exploiting the 3rd and the 6th equation of (6.35), i.e., $-d\hat{\lambda}_3 = 0$, and $-d\hat{\lambda}_6 = 0$, ones get $\hat{\lambda}_3(t) = 0$, and $\hat{\lambda}_6(t) = 0$, respectively. The remain equations of (6.35) get

$$\hat{\lambda}_1(t) = \int_0^t (\hat{\mu}_1^g \hat{\alpha}_0 - \hat{\mu}_1 \hat{\alpha}_+) \frac{\partial \|\hat{\mathbf{v}}\|}{\partial x} dt + (b_1 \hat{\alpha}_0 - a_1 \hat{\alpha}_+) \int_{[0,t]} \frac{\partial \|\hat{\mathbf{v}}\|}{\partial x}(\tau) d\hat{\eta}(\tau), \tag{6.36}$$

$$\hat{\lambda}_2(t) = \int_0^t (\hat{\mu}_1^g \hat{\alpha}_0 - \hat{\mu}_1 \hat{\alpha}_+) \frac{\partial \|\hat{\mathbf{v}}\|}{\partial y} dt + (b_1 \hat{\alpha}_0 - a_1 \hat{\alpha}_+) \int_{[0,t]} \frac{\partial \|\hat{\mathbf{v}}\|}{\partial y}(\tau) d\hat{\eta}(\tau), \tag{6.37}$$

$$\hat{\lambda}_4(t) = \int_0^t (\hat{\lambda}_1 + (\hat{\mu}_1^g \hat{\alpha}_0 - \hat{\mu}_1 \hat{\alpha}_+) \frac{\partial \|\hat{\mathbf{v}}\|}{\partial v_x}) dt + (b_1 \hat{\alpha}_0 - a_1 \hat{\alpha}_+) \int_{[0,t]} \frac{\partial \|\hat{\mathbf{v}}\|}{\partial v_x}(\tau) d\hat{\eta}(\tau), \tag{6.38}$$

$$\hat{\lambda}_5(t) = \int_0^t (\hat{\lambda}_2 + (\hat{\mu}_1^g \hat{\alpha}_0 - \hat{\mu}_1 \hat{\alpha}_+) \frac{\partial \|\hat{\mathbf{v}}\|}{\partial v_y}) dt + (b_1 \hat{\alpha}_0 - a_1 \hat{\alpha}_+) \int_{[0,t]} \frac{\partial \|\hat{\mathbf{v}}\|}{\partial v_x}(\tau) d\hat{\eta}(\tau). \tag{6.39}$$

Now we consider the case $\sigma_1 > 0$ (i.e., $\|\hat{\mathbf{v}}\| > 0$); the remain case, i.e., $\sigma_1 = 0$, $\sigma_2 > 0$, can be similarly analyzed. In this case, ones have $\hat{\alpha} = (1, 0, 0)^T$, $\mathcal{G}_1 = -\sigma_1$, $\mathcal{G}_j = 0$, $j = 2, 3$, $g_i = 0$, $i = 1, 2$, and $\mathcal{N}(\mathcal{G}) = \emptyset$, hence there are no phase point and the phase jump.
Since $\mathcal{G}_1$ is inactive constraint and the condition $\hat{\mu}\mathcal{G}(\cdot) = 0$, we obtain $\hat{\mu}_1 = 0$.
The adjoint equation (6.29) implies

$$-d\hat{\lambda}_1 = \hat{\lambda}_1 \frac{\partial \hat{v}_x}{\partial x} dt,\ -d\hat{\lambda}_2 = \hat{\lambda}_2 \frac{\partial \hat{v}_y}{\partial x} dt, d\hat{\lambda}_3 = d\hat{\lambda}_6 = 0,\ -d\hat{\lambda}_4 = \hat{\lambda}_1 dt,\ -d\hat{\lambda}_5 = \hat{\lambda}_2 dt. \tag{6.40}$$

Similar to the previous case, exploiting the 3rd equation of (6.40), i.e., $-d\hat{\lambda}_3 = 0$, and $-d\hat{\lambda}_6 = 0$, we obtain $\hat{\lambda}_3(t) = 0$, and $\hat{\lambda}_6(t) = 0$, respectively. The remain equations of (6.40) imply

$$\ln(\hat{\lambda}_1) = \int_0^t \frac{\partial \hat{v}_x}{\partial x} dt,\ \ln(\hat{\lambda}_2) = \int_0^t \frac{\partial \hat{v}_y}{\partial x} dt, \quad \hat{\lambda}_4 = \int_0^t \hat{\lambda}_1 dt,\ \hat{\lambda}_5 = \int_0^t \hat{\lambda}_2 dt. \tag{6.41}$$

6.2.5 Numerical Solution

Table 6.1: Parameters of the mechanical model with dry friction: A Mass Point on A Rough Plane, therein "–" means no unit, and $i = \overline{1, n}$.

Physical quantity	Identifier	Value	Unit
Number of points	n	3	–
Free fall acceleration	g	9.8	m/s/s
Identical point's mass	m	1	kg
Friction coefficient	k	0.5	–
Friction/control force	F_i/f_i		N (1 kg·m/s^2)
Angle $\angle yOY$ (see 6.2.1)	γ	$\pi/3$, $\pi/2$	rad
Upper force	U	10	N

For the system (6.19) with parameters are setting in Tab. 6.1, where R_x and R_y are defined by (6.10) and (6.11) (more details acan be seen in the previous subsections 6.2.1 and 6.2.2) for $N = mg\cos\ \gamma - f_z$, by solving the corresponding SwOCP (6.27) numerically, we

obtain a control force $\hat{\mathbf{F}}$ (see Fig. 6.2 and Fig. 6.3) which satisfies the constraints (6.20-6.21), and transfers the system (6.19) in a minimal time $\hat{T}$(s) from a given initial state

$$x(0) = x_0 = 10,\ y(0) = y_0 = 1, \quad \dot{x}(0) = \dot{x}_0 = 10,\ \dot{y}(0) = \dot{y}_0 = 5, \tag{6.42}$$

to a given terminal state

$$x(T) = x_T = 90,\ y(T) = y_T = 9, \quad \dot{x}(T) = \dot{x}_T = 0,\ \dot{y}(T) = \dot{y}_T = 0, \tag{6.43}$$

Note that here we exploit the mixed constraints (6.26) in the relaxed formula as $\mathcal{G}_j(\cdot, \varepsilon) \leq 0$, $j = 1, \ldots, 7$, i.e.,

$$\begin{aligned}
&\mathcal{G}_1(\cdot) \to \mathcal{G}_1(\cdot, \varepsilon) := -\alpha_+(t)\sigma_1 - \varepsilon,\\
&\mathcal{G}_2(\cdot) \to \mathcal{G}_2(\cdot) := \alpha_0(t)\sigma_2,\\
&\mathcal{G}_3(\cdot) \to \mathcal{G}_3(\cdot) := -\alpha_-(t)\sigma_2 - \varepsilon,\\
&g_1(\cdot) \to \mathcal{G}_4(\cdot, \varepsilon) := \alpha_0(t)\sigma_1 - \varepsilon \text{ and } \mathcal{G}_5(\cdot, \varepsilon) := -\alpha_0(t)\sigma_1 - \varepsilon,\\
&g_2(\cdot) \to \mathcal{G}_6(\cdot) := \alpha_-(t)\sigma_1 - \varepsilon \text{ and } \mathcal{G}_7(\cdot) := -\alpha_-(t)\sigma_1 - \varepsilon.
\end{aligned} \tag{6.44}$$

where $\varepsilon > 0$ small enough.
For more details, readers can see at `https://github.com/DuyTranHD/OCPswitched`.

6.3 Optimal Control of Material Points System in a Straight Line with Dry Friction

This section discusses analytical and numerical solution approaches for optimal control of a system of material points in a straight line with dry friction, which are based on our conference paper [121].

6.3.1 Optimal Control Problem

We consider a optimal control of system of material points in a straight line with dry friction, see [54], which consists $n \geq 3$ material points. The masses of these points are taken to be identical ($m_i = m, i = \overline{1,n}$) between the straight line and the points the dry Coulomb friction force acts. The forces that interact between neighboring points are assumed to control variables. Assume that x_i is the coordinate of the i-th point along the line, v_i and f_i are algebraic projections of the velocity of the i-point and the control force (acting on the $(i+1)$-th point from the i-th point) on the straight line, and F_i is an algebraic projection of the friction force acting on the i-th point, see Fig. 6.4.
The system of points' motion equations take the form

$$\begin{aligned}
&\dot{x}_i = v_i,\\
&m\dot{v}_i = f_{i-1} - f_i + F_i,
\end{aligned} \quad i = \overline{1,n}. \tag{6.45}$$

Hereafter we use the extension of the definition

$$f_0 = f_n = 0. \tag{6.46}$$

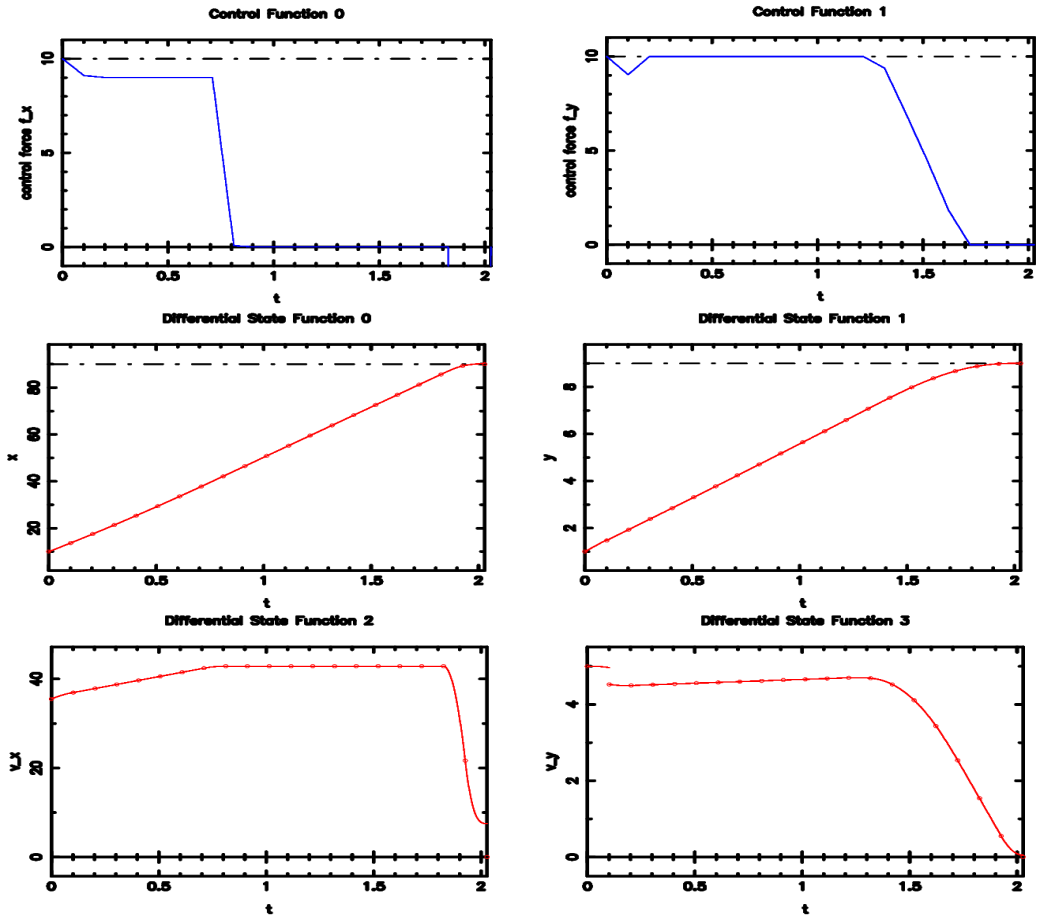


Figure 6.2: For the special case $\gamma = \pi/2$: control force $\hat{F}$ ($\hat{f}_x$-top left, $\hat{f}_y$ top right) drives trajectories positions x (middle left), y(middle right), with velocity v_x (bottom left), v_y (bottom right) in minimal time $\hat{T} = 2.02745$(s), where control $\hat{\alpha} = (1, 0, 0)$.

The friction forces are determined by the relations

$$F_i = \begin{cases} -kmg\,\mathrm{sgn}(v_i), & \text{if } v_i \neq 0, \\ -f_{i-1} + f_i, & \text{if } v_i = 0 \;\&\; |f_{i-1} - f_i| \leq kmg, \\ -kmg\,\mathrm{sgn}(f_{i-1} - f_i), & \text{if } v_i = 0 \;\&\; |f_{i-1} - f_i| > kmg, \end{cases} \tag{6.47}$$

where k is the coefficient of the friction between the points of the system and the plane; and g is the free fall acceleration.
Assume that at initial instant all points of the system are at rest and are located at one point of the straight line, without losing generality, we assume that this point is the coordinate axis's origin point.

$$x_i(0) = 0, \quad v_i(0) = 0, \quad i = \overline{1, n}. \tag{6.48}$$

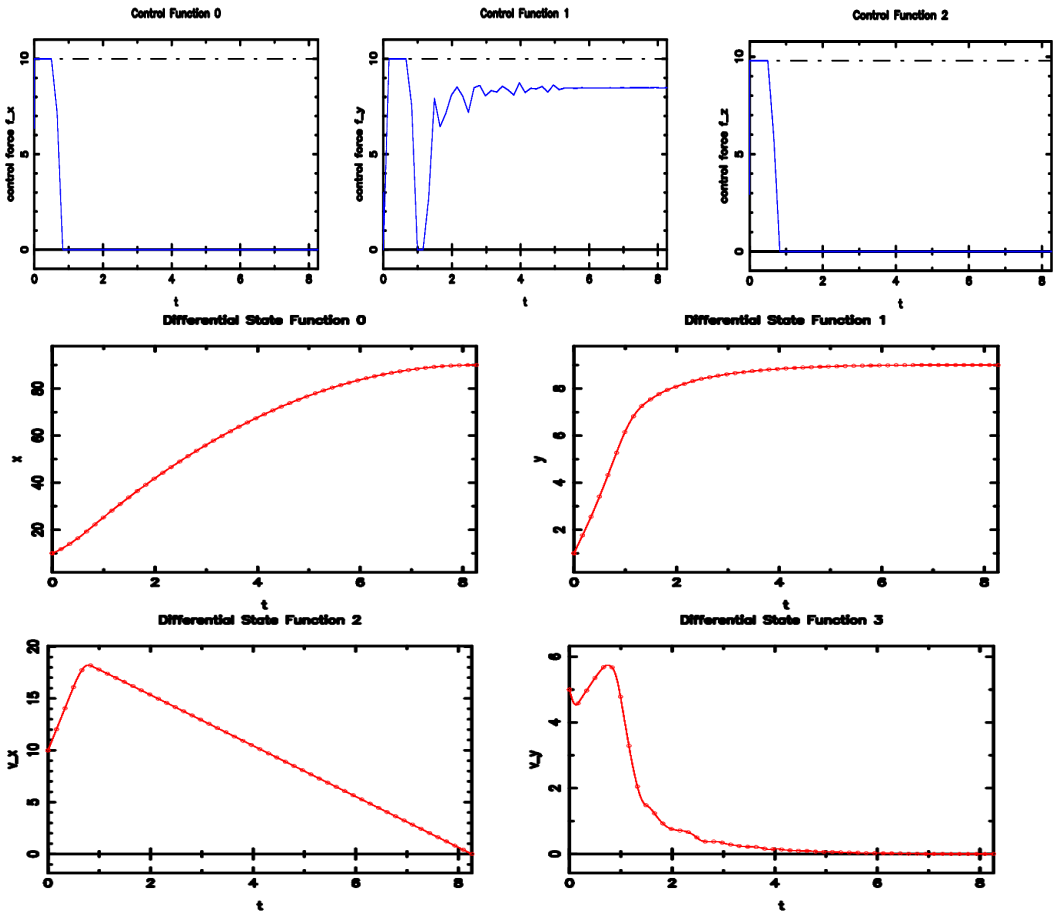


Figure 6.3: For the case $\gamma = \pi/3$: control force $\hat{\mathbf{F}}$ ($\hat{f}_x$-top left, $\hat{f}_y$-top middle, $\hat{f}_z$ top right) drives trajectories positions x (middle left), y (middle right), with velocity v_x(bottom left), v_y (bottom right) in minimal time $\hat{T} = 8.26367$(s), where control $\hat{\alpha} = (1, 0, 0)$.

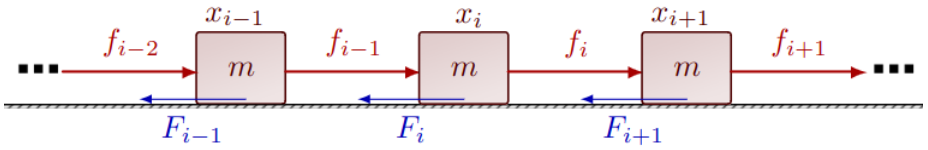


Figure 6.4: System of material points in a straight line with dry friction.

Let us fix the time of the system's motion ($t \in [0, T]$). We consider motions of the system that bring all of its points at the final instant T to the same position in the straight line with the zero velocity

$$x_i(T) = x_1(T), \quad i = \overline{2, n}, \quad v_i(T) = 0, i = \overline{1, n}. \tag{6.49}$$

Denote by x and v the coordinate of center of system's mass and the velocity of this center:

$$x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad v = \frac{1}{n}\sum_{i=1}^{n} v_i. \tag{6.50}$$

The goal is to find the motion of a system of point that moves the system from state (6.48) to state (6.49), where the motion is controlled by relations (6.45) and (6.47) under unbounded control forces f_i, and maximizes the displacement of the system

$$x(T) \to \max. \tag{6.51}$$

We can rewrite the above problem as in the below formulation of the optimal control problem

$$\begin{aligned}
&\max_{x(\cdot),v(\cdot),f(\cdot)} \quad x\,(T)\\
&\text{s.t. } \dot{x}_i(t) = v_i(t),\ i = \overline{1,n},\\
&\qquad \dot{v}_i(t) = \begin{cases} \frac{f_{i-1}-f_i}{m} - kg\,\mathrm{sgn}(v_i), & \text{if } v_i \neq 0,\\ 0, & \text{if } v_i = 0 \ \&\ |f_{i-1}-f_i| \leq kmg,\ i = \overline{1,n},\\ \frac{f_{i-1}-f_i}{m} - kg\,\mathrm{sgn}(f_{i-1}-f_i), & \text{if } v_i = 0 \ \&\ |f_{i-1}-f_i| > kmg, \end{cases}\\
&\qquad x_i(0) = 0, v_i(0) = 0,\ i = \overline{1,n},\\
&\qquad x_i(T) = x_1(T),\ i = \overline{2,n},\ v_i(T) = 0,\ i = \overline{1,n},
\end{aligned} \tag{6.52}$$

where (6.46) and (6.50) are satisfied.

6.3.2 Reformulation

We will solve (6.52) by exploiting Filippov's rule to rewrite this problem to a relaxed convexified one, together with the arising of the additional mixed state-control constraints, and the LMP is employed to obtain the solution.
By reformulating ODEs of problem (6.52) in Filippov's rule, we obtain the equivalent ones

$$\dot{v}_i(t) = \sum_{j\in\mathcal{J}} \alpha_{j,i}(t) h_j\left(f_i(t), f_{i-1}(t), v_i(t), i\right), \ \sum_{j\in\mathcal{J}} \alpha_{j,i}(t) = 1, \alpha_{j,i}(t) \in [0,1],\ j \in \mathcal{J}, i = \overline{1,n},$$

with additional mixed constraints $g_i(v,\alpha) = 0$, $i = 1,2,3$, $\mathcal{G}_j(v,\alpha,f) \leq 0$, $j = 1,\ldots,6$,

$\mathcal{G}_1 := -\alpha_{1,i}(t)v_i(t)$,
$\mathcal{G}_2 := \alpha_{2,i}(t)v_i(t)$,
$g_1 := \alpha_{3,i}(t)v_i(t)$, $\mathcal{G}_3 := \alpha_{3,i}(t)\left(f_{i-1} - f_i - kmg\right)$, $\mathcal{G}_4 := -\alpha_{3,i}(t)\left(f_{i-1} - f_i + kmg\right)$, $i = \overline{1,n}$,
$g_2 := \alpha_{4,i}(t)v_i(t)$, $\mathcal{G}_5 := -\alpha_{4,i}(t)\left(f_{i-1} - f_i - kmg\right)$,
$g_3 := \alpha_{5,i}(t)v_i(t)$, $\mathcal{G}_6 := \alpha_{5,i}(t)\left(f_{i-1} - f_i + kmg\right)$,

where switching functions $\sigma_{1,i} := v_i$, $\sigma_{2,i} := f_{i-1} - f_i - kmg$, $\sigma_{3,i} := f_{i-1} - f_i + kmg$, $i = \overline{1,n}$; and

$$h_1(\cdot,i) = \frac{f_{i-1}-f_i}{m} - kg, \ \ h_2(\cdot,i) = \frac{f_{i-1}-f_i}{m} + kg, \ \ h_3(\cdot,i) = 0, \quad i = \overline{1,n},$$

$$h_4(\cdot,i)=\frac{f_{i-1}-f_i}{m}-kg,\ \ h_5(\cdot,i)=\frac{f_{i-1}-f_i}{m}+kg,\quad i=\overline{1,n}.$$

Denoting $u_i:=(f_{i-1}\ f_i\ \alpha^T)=(f_{i-1}\ f_i\ \alpha_{1,i}\ \alpha_{2,i}\ \alpha_{3,i}\ \alpha_{4,i}\ \alpha_{5,i})$, $i=\overline{1,n}$, and

$$h(f,v,i):=\sum_{j\in\mathcal{J}}\alpha_{j,i}(t)h_j\left(f_i(t),f_{i-1}(t),v_i(t),i\right),\qquad i=\overline{1,n},\tag{6.53}$$

For $i=1,\ldots,n$, the phase points set is determined by

$$\begin{aligned}\mathcal{N}(\mathcal{G})&=\{(x,v,f,\alpha)\mid\sigma_{j,i}=0,j=1,2,3\}\\&=\{(x,v,f,\alpha)\mid v_i=0,\ f_{i-1}-f_i-kmg=0,\ f_{i-1}-f_i+kmg=0\}\\&=\emptyset,\end{aligned}$$

which means that the mixed constraints are regular constraints.
For $i=1,\ldots,n$, it is not difficult to show that the assumption [45, RMC] is satisfied. Some following cases are considered.
If $v_i<0$ then $\alpha_{1,i}=\alpha_{3,i}=\alpha_{4,i}=\alpha_{5,i}=0$, thus $\alpha_{2,i}=1$, the active set $\mathcal{I}=\{1,3,4,5\}$.
If $v_i=0$ then $0\le\alpha_{j,i}\le1$, $j\in\mathcal{J}$, and $\mathcal{I}\supseteq\{1,2\}$. If $\alpha_{1,i}=1$ or $\alpha_{2,i}=1$ then $\mathcal{I}=\mathcal{J}$.
If $v_i>0$ then $\alpha_{2,i}=\alpha_{3,i}=\alpha_{4,i}=\alpha_{5,i}=0$, thus $\alpha_{1,i}=1$ and $\mathcal{I}=\{2,3,4,5\}$.
Thus, problem (6.52) is reformulated as the following relaxed one

$$\begin{aligned}\min_{x(\cdot),v(\cdot),f(\cdot),\alpha(\cdot)}\quad&-x(T)\\\text{s.t.}\quad&\dot{x}_i(t)=v_i(t),\\&\dot{v}_i(t)=h(f,v,i),\\&\mathcal{G}_j(\cdot,i)\le0,\ j=1,\ldots,6,\quad g_l(\cdot,i)=0,\ l=1,2,3,\qquad i=\overline{1,n},\\&x_i(0)=0,v_i(0)=0,\\&x_i(T)=x_1(T),\ i=\overline{2,n},\ v_i(T)=0,\\&\textstyle\sum_{j\in\mathcal{J}}\alpha_{j,i}(t)=1,\alpha_{j,i}(t)\in[0,1],\ j\in\mathcal{J},i\end{aligned}\tag{6.54}$$

therein $h(\cdot)$ is defined as in (6.53), and equations (6.46-6.50) are required, with $\mathcal{G}_j(\cdot,i)$ is an active constraint, $j\in\mathcal{I}$, where the index set $\mathcal{I}$ depend on the value of v_i, $i=\overline{1,n}$.

6.3.3 A Solution Approach with LMP

We then can study (6.52) by solving (6.54), and the equivalent optimal solution can be implied. By using LMP, we can derive the necessary condition of problem (6.54).
We define some needed functions as follows.

(i) Augmented Pontryagin function

$$\bar{\mathcal{H}}(x,v,f,\alpha)=\lambda(t)^TH_0(x(t),v(t),f(t),\alpha(t))+\sum_{j=1}^{6}\mu_j\mathcal{G}_j(\alpha,f,v,i)+\sum_{l=1}^{3}\mu_l^g g(\alpha,v,i),$$

where $\mu_j(t)\ge0$, $j=1,\ldots,6$, $\mu_l^g(t)\in\mathbb{R}$, $l=1,2,3$, and

$$H_0(\cdot)=\begin{pmatrix}v_1(t) & \ldots & v_n(t) & h(f,v,1) & \ldots & h(f,v,n)\end{pmatrix}^T.$$

(iii) Endpoint LAGRANGE function

$$L_L(x_0, v_0, x_T, v_T) = \nu^T \begin{pmatrix} x_1(0) & \dots & x_n(0) & 0 & \dots & x_n(T) - x_1(T) \\ v_1(0) & \dots & v_n(0) & v_1(T) & \dots & v_n(T) \end{pmatrix}^T$$

Then, from the Theorem "local minimum principle", [45, Thm. 1], there exists a tuple of multipliers $(\lambda, \mu_1, \dots, \mu_6, \nu)$ satisfying the below properties: $\lambda : [0,T] \to I\!R^{2n}$ is a Lipschitz continuous function, $\mu_j : [0,T] \to I\!R_+$, $j = 1, \dots, 6$, $\mu_l^g : [0,T] \to I\!R$, $j = 1,2,3$, are measurable bounded functions, $\nu > 0$ is a vector; and such that the conditions (a) - (g) of LMP hold true.

(a) the nonnegativity conditions

$$\nu \geq 0, \quad \mu_i \geq 0, \quad i = 1, \dots, 6, \tag{6.55}$$

(b) the non-triviality condition

$$|\nu| + \int_0^T \sum_{j=1}^{6} \mu_j(t) dt > 0, \tag{6.56}$$

(c) the complementary slackness conditions: $L_L(x_0, v_0, x_T, v_T) = 0$,

(d) the pointwise complementary slackness conditions

$$\mu_j(t) \mathcal{G}_j(\cdot) = 0 \text{ a.e. on } [0,T], \ j = 1, \dots, 6, \tag{6.57}$$

(e) the adjoint equation

$$-\dot{\lambda}(t) = \left(\frac{\partial \bar{\mathcal{H}}}{\partial x}(x^*(t), v^*(t), f^*(t), \alpha^*(t)) \quad \frac{\partial \bar{\mathcal{H}}}{\partial v}(x^*(t), v^*(t), f^*(t), \alpha^*(t)) \right)^T \tag{6.58}$$

(f) the transversality conditions

$$\lambda(0) = -\left(\frac{\partial L_L}{\partial x_0}(x^*(0), x^*(T)) \quad \frac{\partial L_L}{\partial v_0}(x^*(0), x^*(T)) \right)^T \tag{6.59}$$

$$\lambda(T) = \left(\frac{\partial L_L}{\partial x_T}(x^*(0), x^*(T)) \quad \frac{\partial L_L}{\partial v_T}(x^*(0), x^*(T)) \right)^T \tag{6.60}$$

(g) the stationarity condition of the extended PONTRYAGIN function w.r.t. the control

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha}(x^*(t), v^*(t), f^*(t), \alpha^*(t)) = 0 \text{ a.e. on } [0,T], \tag{6.61}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial f}(x^*(t), v^*(t), f^*(t), \alpha^*(t)) = 0 \text{ a.e. on } [0,T]. \tag{6.62}$$

With $i = 1, 2, \ldots, n$, we have

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha_{1,i}} = \lambda_{n+i}\left(\frac{f_{i-1} - f_i}{m} - kg\right) - \mu_1 v_i, \tag{6.63a}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha_{2,i}} = \lambda_{n+i}\left(\frac{f_{i-1} - f_i}{m} + kg\right) + \mu_2 v_i, \tag{6.63b}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha_{3,i}} = \mu_1^g v_i - \mu_3(f_{i-1} - f_i - kmg) + \mu_4(f_{i-1} - f_i + kmg), \tag{6.63c}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha_{4,i}} = \lambda_{n+i}\left(\frac{f_{i-1} - f_i}{m} - kg\right) + \mu_2^g v_i - \mu_2^g v_i - \mu_5(f_{i-1} - f_i - kmg), \tag{6.63d}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha_{5,i}} = \lambda_{n+i}\left(\frac{f_{i-1} - f_i}{m} + kg\right) + \mu_3^g v_i + \mu_6(f_{i-1} - f_i + kmg), \tag{6.63e}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial f_i} = \lambda_{n+i}\frac{\alpha_{3,i} - 1}{m} + \mu_3\alpha_{3,i} - \mu_4\alpha_{3,i} + \mu_5\alpha_{4,i} - \mu_6\alpha_{5,i}, \tag{6.63f}$$

$$\frac{\partial \bar{\mathcal{H}}}{\partial f_{i-1}} = \lambda_{n+i}\frac{1 - \alpha_{3,i}}{m} - \mu_3\alpha_{3,i} + \mu_4\alpha_{3,i} - \mu_5\alpha_{4,i} + \mu_6\alpha_{5,i} = -\frac{\partial \bar{\mathcal{H}}}{\partial f_i}. \tag{6.63g}$$

Then, using the stationarity conditions (6.61-6.62), we obtain

$$\mu_1 v_i = \lambda_{n+i}\left(\frac{f_{i-1} - f_i}{m} - kg\right), \tag{6.64a}$$

$$\mu_2 v_i = -\lambda_{n+i}\left(\frac{f_{i-1} - f_i}{m} + kg\right) = -\mu_1 v_i - 2\lambda_{n+i}kg, \tag{6.64b}$$

$$0 = \mu_1^g v_i - \mu_3(f_{i-1} - f_i - kmg) + \mu_4(f_{i-1} - f_i + kmg), \tag{6.64c}$$

$$0 = -(f_{i-1} - f_i - kmg)\left(\frac{\lambda_{n+i}}{m} - \mu_5\right), \tag{6.64d}$$

$$0 = -(f_{i-1} - f_i + kmg)\left(\frac{\lambda_{n+i}}{m} + \mu_6\right), \tag{6.64e}$$

$$0 = \lambda_{n+i}\frac{1 - \alpha_{3,i}}{m} - \mu_3\alpha_{3,i} + \mu_4\alpha_{3,i} - \mu_5\alpha_{4,i} + \mu_6\alpha_{5,i}. \tag{6.64f}$$

From the pointwise complementary conditions (6.57) and the fact that $\mathcal{G}_j(\cdot)$ is an inactive constraint, $j \notin \mathcal{I}$, one obtains $\mu_j(t) = 0$, $j \notin \mathcal{I}$. Note that here $f_0 = f_n = 0$, and $n \geq 3$.
Case: $v_i = 0$, $i = \overline{1, n}$, i.e., $\mathcal{I} = \{1, 2, 3, 5\}$, $\alpha_i = (0, 0, 0, 1, 0)$, where we choose $\alpha_{4,i} = 1$.
The conditions (6.64c-6.64e) help us obtain

$$|f_{i-1} - f_i| = kmg, \tag{6.65}$$

therein we assume that $\mu_3 \neq 0$ and $\lambda_{n+i} \neq 0$, for $i = 1, \ldots, n$. Then we obtain

$$\dot{v}_i(t) = 0 \Leftrightarrow v_i(t) = 0, \qquad t \in [0, t_n] \setminus [t_{i-1}, t_i]. \tag{6.66}$$

Case: $v_i \neq 0$, $i = 1, 2, \ldots, n$. We consider the cases $v_i > 0$ and $v_i < 0$. From (6.64a-6.64b), and remember that $v_i(T) = 0$, $i = \overline{1,n}$, we can imply

$$\dot{v}_i(t) = \frac{f_{i-1} - f_i}{m} - kg\,\mathrm{sgn}(v_i(t)) = \begin{cases} -kg & \text{if } v_i > 0, \\ (n-2)kg & \text{if } v_i < 0 \end{cases}$$

$$\Leftrightarrow v_i(t) = \begin{cases} kg(T-t) & \text{if } t \in [t_n, T], \\ (n-2)kgt & \text{if } t \in [t_{i-1}, t_i], \end{cases} \tag{6.67}$$

where the stopping time of the ith point is determined by $t_i = \sqrt{i}t_1$, $i = \overline{1,n}$, and then the time of the speed-up of center of mass can be used to obtain

$$t_i = \frac{\sqrt{ni}}{2(n-1)}T, \qquad i = \overline{0,n}. \tag{6.68}$$

Finally, from (6.66), and (6.67), we deduce that the motion the system of points is determined by the following relations

$$v_i(t) = \begin{cases} (n-2)kgt, & \text{if } t \in [t_{i-1}, t_i], \\ 0, & \text{if } t \in [0, t_n] \setminus [t_{i-1}, t_i], \\ kg(T-t), & \text{if } t \in [t_n, T]. \end{cases} \qquad i = \overline{1,n}, \tag{6.69}$$

where t_i, $i = \overline{0,n}$, is defined by (6.68).

Remark 39. We obtain the same result by comparing with [54, Eq. 2.16] for $n \geq 3$.

6.3.4 Numerical Solution

We consider a system consisting of $n = 3$ material points which lie along a horizontal straight line. More descriptions in details can see in the previous subsections of Section 6.3. In this section, we use the following parameters, see Tab. 6.2, in order to find the motion of a system of point to maximizes the displacement of the system (6.51), in the case where this motion is governed by (6.45) and (6.47), takes the system from state (6.48) to (6.49) with $T = 10$(s), and $x_i(T) = x_1(T) = 10$(m), $i = 2, 3$, and a staring control forces $f_1 = 2$(N), $f_2 = 4$(N).

We solve problem (6.54) with above parameters and remembering that $f_0 = f_3 = 0$(N).

Table 6.2: Parameters of the mechanical model with dry friction, therein "–" means no unit, and $i = \overline{1,n}$.

Physical quantity	Identifier	Value	Unit
Number of points	n	3	–
Free fall acceleration	g	9.8	m/s/s
Identical point's mass	m	1	kg
Friction coefficient	k	0.5	–
Friction/control force	F_i/f_i		N

We obtain the maximal displacement is 94.514(m) by `MUSCOD-II`, see Figure 6.5. For more details, readers can see at `https://github.com/DuyTranHD/OCPswitched`. Note that here we also use further relaxed ε for mixed constraints of problem (6.54), similarly to Subsection 6.2.5.

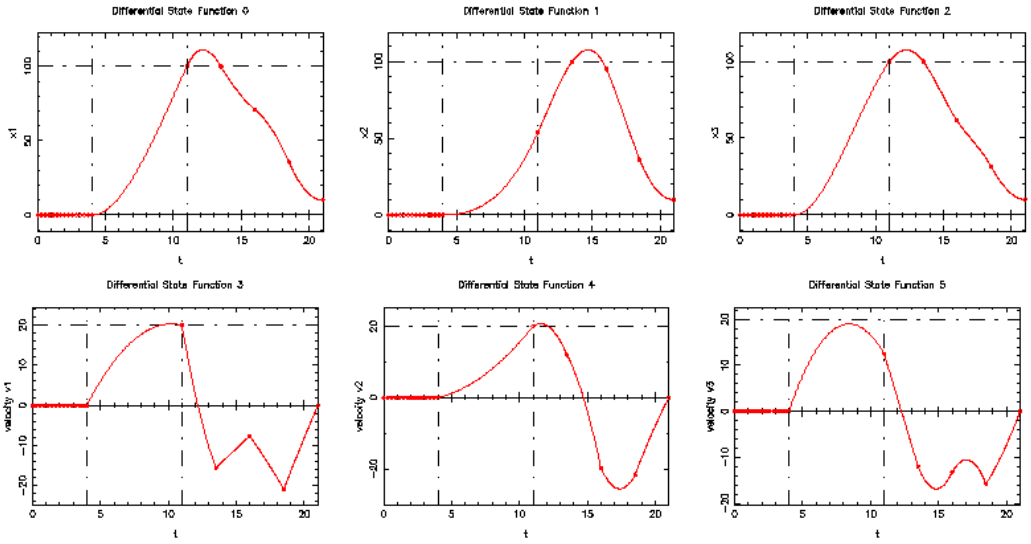


Figure 6.5: Positions of three points: x_1 (top left), x_2 (top middle), and x_3(top right). Velocity of three points: v_1 (bottom left), v_2 (bottom middle), and v_3(bottom right).

Chapter 7

Overview and Outlook

7.1 Overview

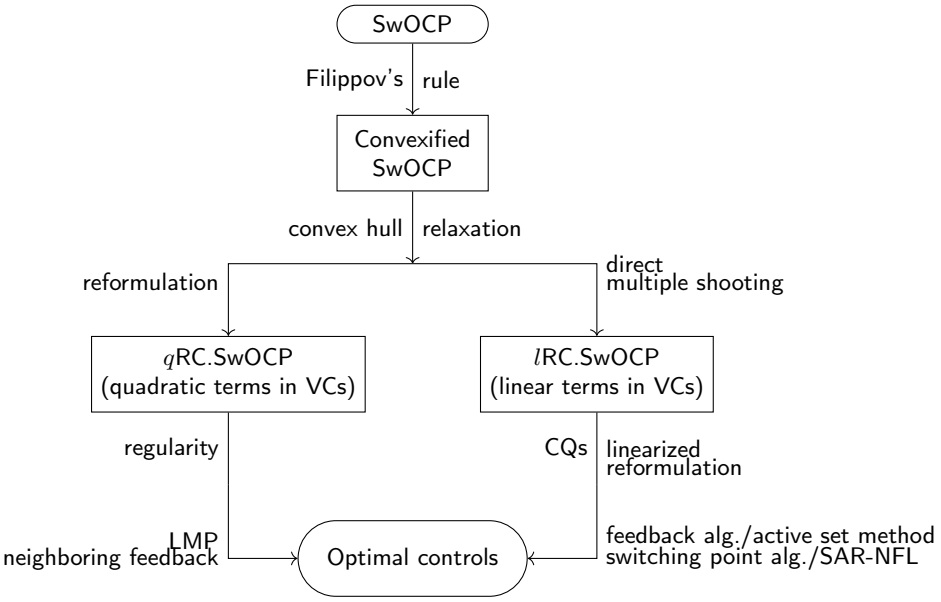


Figure 7.1: Dissertation overview diagram.

The general framework to solve SwOCPs is presented, as well as such problems that appear in real-life applications, like e.g., the subway problem, the flat hybrid automation with a DC electrical network, and dry friction problems. Through specially tailored Filippov's rule and relaxation, SwOCP is reformulated, and the resulting problem is then solved in two approaches.

As an indirect approach, the Maximum Principle takes the role of solving the SwOCP efficiently. From the local maximum principle, the optimality conditions are obtained by exploiting relaxed reformulations; therein, Filippov's rule is exploited carefully with further

convex hull relaxation, which allows us to treat the switches (i.e., the integer controls) exactly without using the rounding schemes. Numerical applications are considered in the subway problem and some instances of dry friction: a system of material points in a straight line with dry friction, and a point mass on a rough plane.

The direct approach with the direct multiple shooting method is proposed as the second approach that gives rise to Nonlinear Programming reformulations solved through a specially tailored SQP algorithm, wherein the regularity of the constraint qualifications is checked carefully. By employing the block structure of the quadratic programming subproblem in some special cases, one shows that the switches can be treated from the derivative with respect to the discretized state trajectory, which is confirmed by the active set method for vanishing constraints. Furthermore, a neighboring feedback law is proposed to obtain feedback controls. Subsequently, the CIA is considered as another approach to reformulating the SwOCP. As a state-of-the-art direct approach to treating switching points effectively, a switching point algorithm is proposed, which comes from discontinuous dynamics. Some applications are considered to illustrate the efficiency of the approach and include the subway problem, the flat hybrid automaton with a DC electrical network, and the dry friction problems.

Some auxiliary results, and open problems, such as the Gröbner basis approach, a concept about over-under estimating, and the Competing Hamiltonian algorithm, are considered in the Appendices.

7.2 Outlook

In future work for the feedback algorithm and condensing procedure with block structure for the QP subproblem, OCPs with vanishing constraints will be investigated for the general case. Furthermore, by considering the general case (not just the scalar function) of the switching function, and the path constraints will be taken in SwOCP, the general results will be established. Therefore, one can treat the switches directly to the derivatives of the point constraints and path constraints.

The case when the control u enters nonlinearly into the right hand side function of ODE of SwOCP (or SOCP), remains an open question for further investigation. Thus, future research will cover this problem by considering the reformulation and relaxation of Filippov's rule in some special structures of SwOCP (or SOCP).

The new reformulation for the mixed state-control constraints of SwOCP may be proposed, which is based on the analytical regular property of these constraints. One can apply it for other types of OCP, and some respective numerical tests for dry friction problems will also be considered.

Instead of using FD to generate derivatives, automatic differentiation (AD) can be used to improve the approximation. Moreover, internal numerical differentiation (IND) can help to freeze adaptive components. In future studies, we may apply this concept to numerical examples and compare the findings to those in this dissertation to demonstrate that the quality of our approach will be slightly better with AD.

Some open problems to work on are considered in Appendix B, including the Gröbner Basis Approach for solving SwOCP, and an idea about over-under estimating switches. This general heuristic approach will be developed to be more efficient and widely adopted.

Appendix A

Auxiliary Results

This appendix collects all auxiliary theories, and auxiliary numerical results.

A.1 An Example of Incorrect OCP

We start with an example as follows

$$\begin{array}{ll} \min & \|x(t_f)\|_2^2 \\ \text{s.t.} & \dot{x}(t) = \begin{cases} b_+ u(t), & \text{if } d^T x(t) > 0; \\ b_- u(t), & \text{if } d^T x(t) < 0; \end{cases} \\ & |u(t)| \le 1, t \in [0, t_f],\ x(0) = x_0, \end{array} \tag{A.1}$$

where $t_f = 3$, $d = \begin{pmatrix} 1 & -1 \end{pmatrix}^T$, $b_+ = \begin{pmatrix} -1 & 0 \end{pmatrix}^T$, $b_- = \begin{pmatrix} 0 & 1 \end{pmatrix}^T$, $x_0 = \begin{pmatrix} 2 & 1 \end{pmatrix}^T$, $x \in \mathbb{R}^2$. Problem (A.1) with a discontinuous right hand side under the solution of the ODE, we hence will consider the solution according to Filippov. Then we get a corresponding problem

$$\begin{array}{ll} \min & \|x(t_f)\|_2^2 \\ \text{s.t.} & \dot{x}(t) = \begin{cases} b_+ u(t), & \text{if } d^T x(t) > 0; \\ b_- u(t), & \text{if } d^T x(t) < 0; \\ (\alpha(t) b_- + (1 - \alpha(t)) b_+) u(t), & \text{if } d^T x(t) = 0;\ \ \alpha(t) \in [0,1], \end{cases} \\ & |u(t)| \le 1,\ x(0) = x_0, \\ & t \in [0, t_f]. \end{array} \tag{A.2}$$

Problem (A.2) can be rewritten in the equivalent form

$$\begin{array}{ll} \min & \|x(t_f)\|_2^2 \\ \text{s.t.} & \dot{x}(t) = (\alpha(t) b_- + (1 - \alpha(t)) b_+) u(t),\ t \in [0, t_f],\ x(0) = x_0, \\ & |u(t)| \le 1, \alpha(t) \in [0,1], \alpha(t) d^T x(t) \le 0, (1 - \alpha(t)) d^T x(t) \ge 0,\ t \in [0, t_f]. \end{array} \tag{A.3}$$

Note that in problem (A.3), it is easy to show that the set of admissible velocity

$$U := \{ v \in \mathbb{R}^2 : v = (\alpha b_- + (1 - \alpha) b_+)\, u, \alpha \in [0,1], |u| \le 1 \} \tag{A.4}$$

is non-convex. Therefore, problem (A.3) may not have a solution or be incorrect which we will show now. Together with the problem (A.3) we consider the following problem, where the set U is replaced by its convex hull $\mathrm{conv}(U)$:

$$\begin{array}{ll} \min & \|x(t_f)\|_2^2 \\ \text{s.t.} & \dot{x}(t) = \gamma_1(t)(b_- + b_+) + \gamma_2(t)(b_- - b_+) - b_-,\ x(0) = x_0, \\ & \gamma_j(t) \in [0,1],\ j = 1,2, \\ & (\gamma_1(t) + \gamma_2(t) - 1)^2 d^T x(t) \le 0, (\gamma_2(t) - \gamma_1(t))^2 d^T x(t) \ge 0, \end{array} \quad t \in [0, t_f]. \qquad \text{(A.5)}$$

In the problem (A.5) the control

$$\gamma_1^0(t) = 1,\ \gamma_2^0(t) = 0,\ t \in [0,1]; \qquad \gamma_1^0(t) = 0.5,\ \gamma_2^0(t) = 0,\ t \in [1,3],$$

together with the corresponding trajectory $x^0(t)$, $t \in [0,3]$,

$$x_1^0(t) = 2 - t, x_2^0(t) = 1, t \in [0,1];\ x_1^0(t) = 1 - (t-1)/2, x_2^0(t) = 1 - (t-1)/2, t \in [1,3],$$

satisfy the relations

$$d^T x^0(t) = 1 - t \ge 0,\ t \in [0,1]; \quad d^T x^0(t) = 0,\ t \in [1,3]; \quad x^0(t_f) = 0.$$

Hence the control $\gamma^0(t)$, $t \in [0,3]$, is feasible and optimal in the problem (A.5).

Let us come back to the problem (A.3). It is easy to show that the cost functional value in the problem (A.3) and consequently in the problem (A.2) is greater than zero for each feasible control. For example, the control

$$u^*(t) = 1, t \in [0,1],\ u^*(t) = 0, t \in [1,3], \quad \alpha^*(t) = 0, t \in [0,3],$$

is feasible, its cost functional has value zero, hence it is optimal in the problem (A.3).

To show that the problem (A.3) is incorrect, we construct a control

$$u^{(\varepsilon)}(t),\ \alpha^{(\varepsilon)}(t),\ t \in [0,3]$$

such that for any $\varepsilon > 0$ it satisfies the following constraints

$$|u^{(\varepsilon)}(t)| \le 1,\ \alpha^{(\varepsilon)}(t) \in [0,1], t \in [0,3]$$

the corresponding trajectory $x^{(\varepsilon)}(t)$, $t \in [0,3]$, satisfies the mixed constraints in the problem (A.3) with the accuracy ε and the cost functional is equal to zero. For this purpose, we choose an integer parameter $M > 0$ and define time points at the interval $[0,3]$ as follows:

$$\tau_j = 1 + jh = 1 + j/M,\ j = 0, 1, \ldots, 2M,\ h = 1/M.$$

The controls $\alpha^M(t)$, $u^M(t)$, $t \in [0,3]$, are constructed by the following rule

$$\begin{array}{l} \alpha^M(t) = 0,\ u^M(t) = 1,\ t \in [0,1], \\ \alpha^M(t) = 1,\ u^M(t) = -1,\ t \in [\tau_{2j}, \tau_{2j+1}), \\ \alpha^M(t) = 0,\ u^M(t) = 1,\ t \in [\tau_{2j+1}, \tau_{2(j+1)}),\ j = 0, 1, \ldots, M-1. \end{array}$$

The corresponding trajectory $x^M(t)$, $t \in [0,3]$, in the problem (A.3) is

$$
\begin{aligned}
&x_1^M(t) = 2 - t,\ x_2^M(t) = 1,\ t \in [0,1],\\
&x_1^M(t) = x_1^M(\tau_{2j}),\ x_2^M(t) = x_1^M(\tau_{2j}) - (t - \tau_{2j}),\ t \in [\tau_{2j}, \tau_{2j+1}),\\
&x_1^M(t) = x_1^M(\tau_{2j}) - (t - \tau_{2j+1}),\ x_2^M(t) = x_2^M(\tau_{2j+1}),\ t \in [\tau_{2j+1}, \tau_{2(j+1)}), j = 0, 1, \ldots, M-1.
\end{aligned}
$$

Hence the following equalities hold true

$$
\begin{aligned}
&x_1^M(\tau_{2j}) = x_2^M(\tau_{2j}),\ j = 0, 1, \ldots, M,\\
&d^T x^M(t) = x_1^M(t) - x_2^M(t) = 1 - t,\ t \in [0,1],\\
&d^T x^M(t) = x_1^M(t) - x_2^M(t) = t - \tau_{2j},\ t \in [\tau_{2j}, \tau_{2j+1}),\\
&d^T x^M(t) = x_1^M(t) - x_2^M(t) = h - (t - \tau_{2j+1}),\ t \in [\tau_{2j+1}, \tau_{2(j+1)}),\ j = 0, 1, \ldots, M-1,
\end{aligned}
$$

which means that

$$
0 \le d^T x^M(t),\ t \in [0,1];\ 0 \le d^T x^M(t) \le h = 1/M,\ t \in [1,3].
$$

Thus the control $\alpha^M(t)$, $u^M(t)$, $t \in [0,3]$, and the corresponding $x^M(t)$, $t \in [0,3]$, in the problem (A.3) satisfy mixed constraints with the accuracy $h = 1/M$ and the cost functional is equal to zero. For $M \to \infty$ the trajectory $x^M(t)$, $t \in [0,3]$, converges to the optimal trajectory $x^0(t)$, $t \in [0,3]$, of the problem (A.5), however the control $\alpha^M(t)$, $u^M(t)$, $t \in [0,3]$, does not have a "resonable" limit. This shows that the problem (A.3) is incorrect.

A.2 Numerical Example with FILIPPOV's Solution

This section will continue to discuss problem (A.1). As a result, we get the problem

$$
\begin{array}{ll}
\min & \|x(t_f)\|_2^2\\
\text{s.t.} & \dot{x}(t) = (\beta_2(t) b_- + \beta_1(t) b_+) u(t),\ t \in [0, t_f],\\
& x(0) = x_0,\\
& \alpha(t) \in [0,1], \alpha(t) d^T x(t) \le 0, (1 - \alpha(t)) d^T x(t) \ge 0,\ t \in [0, t_f],\\
& |\beta_1(t)| \le \alpha(t), |\beta_2(t)| \le (1 - \alpha(t)),\ t \in [0, t_f].
\end{array}
\tag{A.6}
$$

We will solve (A.3) and (A.6) numerically by using `MUSCOD-II`. We start by writing them in numerical-details form

$$
\begin{array}{ll}
\min & \|x(3)\|_2^2\\
\text{s.t.} & \dot{x}(t) = \begin{pmatrix} 1 - \alpha(t) & \alpha(t) \end{pmatrix}^T u(t),\ t \in [0,3],\\
& x(0) = x_0 = \begin{pmatrix} 2 & 1 \end{pmatrix}^T,\ |u(t)| \le 1, \alpha(t) \in [0,1],\ t \in [0,3],\\
& \alpha(t)(x_1(t) - x_2(t)) \le 0, (1 - \alpha(t))(x_1(t) - x_2(t)) \ge 0,\ t \in [0,3],
\end{array}
\tag{A.7}
$$

$$
\begin{array}{ll}
\min & \|x(3)\|_2^2\\
\text{s.t.} & \dot{x}(t) = \begin{pmatrix} -\beta_1(t) & \beta_2(t) \end{pmatrix}^T u(t),\ t \in [0,3],\\
& x(0) = x_0 = \begin{pmatrix} 2 & 1 \end{pmatrix}^T,\ |u(t)| \le 1, \alpha(t) \in [0,1],\ t \in [0,3],\\
& \alpha(t)(x_1(t) - x_2(t)) \le 0, (1 - \alpha(t))(x_1(t) - x_2(t)) \ge 0,\ t \in [0,3],\\
& |\beta_1(t)| \le \alpha(t), |\beta_2(t)| \le (1 - \alpha(t)),\ t \in [0,3].
\end{array}
\tag{A.8}
$$

Code, data, and result files (see `https://github.com/DuyTranHD/OCPswitched`) are saved under names: simple-test and simple-test-2, which are corresponding with (A.7) and (A.8), respectively. Then one obtains the optimal solution, see Tab. A.1, which shows that it has the strong significant (in the computation time and the number of SQP iterations) when dealing with multiple shooting.

Table A.1: Comparison between incorrect and correct application of Filippov's rule. Therein, control $(u, \alpha, \beta_1, \beta_2)$, state trajectory (x_1, x_2), and computing time is counted in second, and # means infeasible solution. In the multiple shooting method, 50 shooting intervals are hired.

Content	One time of convexification	Two times of convexification
Multiple shooting	29 SQP iterations	6 SQP iterations
Objective	$8.027E+00$	$1.678E-11$
Control	#	$(-9.523E-13, 3.686E-17, 1.037E-08, -8.620E-10)$
State trajectory	#	$(-2.514E-06, 3.235E-06)$
Convergence	error	convergence achieved
Computing time	5.286	0.436

A.3 Competing Hamiltonian Approach

In this section, we consider the OCP with integer controls in the formulation as follows

$$\begin{aligned} \min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & m\left(x(t_f)\right) + \int_{t_0}^{t_f} l(x(t),u(t)) \\ \text{s.t.} \quad & \dot{x}(t) = F(x(t),u(t),\alpha_w(t)), \\ & 0 \le r\left(x(t_0),x(t_f)\right), \\ & \textstyle\sum_{w\in\mathcal{W}} \alpha_w(t) = 1,\ 0 \le \alpha_w(t) \le 1, \\ & u \in \mathcal{U}, \quad w \in \mathcal{W}, t \in \mathcal{T} = [t_0, t_f], \end{aligned} \tag{A.9}$$

where $F(x(t),u(t),\alpha(t)) := \sum_{w\in\mathcal{W}} \alpha_w(t)\left(f_+\left(x(t),u(t),w\right) + f_-\left(x(t),u(t),w\right)\right)$.
We firstly would like to write down the maximum principle for problem (A.9).
The Hamilton function (augmented-extended Pontryagin function) of SwOCP (A.9) is

$$\bar{\mathcal{H}}(x,u,\alpha,\lambda) = \lambda^T F(x,u,\alpha) + \delta^T l(x,u),$$

$$\Psi(t_f,x,\rho) = m(x(t_f)) - \rho\, r(x(t_0),x(t_f)),$$

where $\alpha = [\alpha_\sigma \ \ \alpha_w]$. As the principal approach, the LMP reads

$$\begin{aligned} &\dot{x}(t) = F(x(t),u(t),\alpha(t)), \\ &\dot{\lambda}(t) = -\frac{\partial \bar{\mathcal{H}}^T}{\partial x} = -\frac{\partial F^T(x(t),u,\alpha)}{\partial x}\lambda(t) - \delta\frac{\partial l^T(x(t),u)}{\partial x}, \\ &x(t_0) = x_0, \end{aligned}$$

with the terminal condition of the costates

$$\lambda(t_f) = -\frac{\partial \Psi^T(t_f,x(t),\rho)}{\partial x} = -\frac{\partial m^T(x(t_f))}{\partial x} + \rho\frac{\partial r(x(t_0),x(t_f))}{\partial x},$$

if t_f is free, we have in addition

$$\bar{\mathcal{H}}|_{t=t_f} = \frac{\partial \Psi(t_f, x(t_f), \rho)}{\partial t_f} = \frac{\partial m(x(t_f))}{\partial t_f} - \rho \frac{\partial r(x(t_0), x(t_f))}{\partial t_f}.$$

The controls u and α must be determined as $u^*(x(t), \lambda(t))$ and $\alpha^*(x(t), \lambda(t))$ such that the Hamiltonian is maximized everywhere

$$\{u^*, \alpha^*\} = \arg \max_{u \in \mathcal{U}, \alpha \in [0,1]^{2^{n_w}}} \{\bar{\mathcal{H}}(x(t), u, \alpha, \lambda(t))\} \quad \text{a.e. on } [t_0, t_f]$$

This means that for every $t \in [t_0, t_f]$ a finite optimal control problem must be solved.

To track the integer controls w, we use the so-called *Competing Hamiltonians* approach, which has to our knowledge first been successfully applied to the optimization of operation of subway trains with discrete acceleration stages in New York by Bock and Longman, see [20, 21]. We recall that the discrete controls $w \in \mathcal{W} = \{w_1, \ldots, w_{n_w}\}$, and consider continuous controls $u \in \mathcal{U}, \alpha \in [0,1]^{2^{n_w}}$, and $F = F(x, u, \alpha, w)$. We will treat the controls by introducing $\mathcal{W} = \{w_1, \ldots, w_{n_w}\}$.
The Competing Hamiltonians algorithm is described as follows.

1. For all $w \in \mathcal{W}$ determine

   $$\{u^*, \alpha^*\} = \arg \max_{u \in \mathcal{U}, \alpha \in [0,1]^{2^{n_w}}} \{\bar{\mathcal{H}}(x(t), u, \alpha, \lambda(t))\}$$

   and solve the problem for the continuous controls. Then setting

   $$h_w := \bar{\mathcal{H}}(x, u^*, \alpha_w^*, \lambda), \quad w = w_1, \ldots, w_{n_w},$$

   therein, h_w are the Competing Hamiltonians.

2. Solve the Maximum Principle for the discrete control w

   $$w^*(x, \lambda) = \arg \max_{w \in \mathcal{W}} \{h_w(x, \lambda, w)\}.$$

   Then the switches can be treated by the difference between the function with maximum values and the function with values closest to the maximum function, so-called *switching functions*

   - define $Q_j := h_{\hat{w}} - h_j \geq 0$, with $\hat{w}$ is the current optimal value of w, and $j \in \mathcal{W}, j \neq \hat{w}$.
   - monitor if one $Q_{\hat{j}}$ has a zero, determine "switching point", change from $\hat{w}$ to $\hat{j}$, and redefine the Q_j.

Roots of this switching function indicate switched points from one optimal mode to the next. To illustrate some how a switch is treated, we explain as follows.

1. Suppose that w_1 is the current optimal value for w. We define switching functions $Q_j := h_{w_1} - h_j$, where $j \neq w_1$, $j \in \mathcal{W}$. See Fig. A.1 for $j = w_2$.

2. Check, if an equality $Q_j = 0$ is hold, then one determines a switching point j, and redefine Q_j with j is the current optimal value for w. For instance, we start with $j = w_2$, see Fig. A.2. If no, one moves to next step.

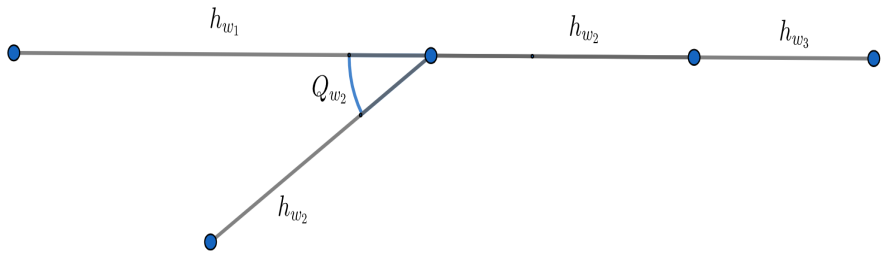


Figure A.1: Step 1 of the explanation.

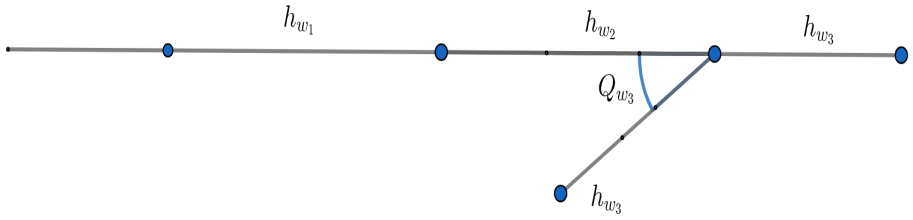


Figure A.2: Step 2 of the explanation, where $j = w_2$ is a switching point.

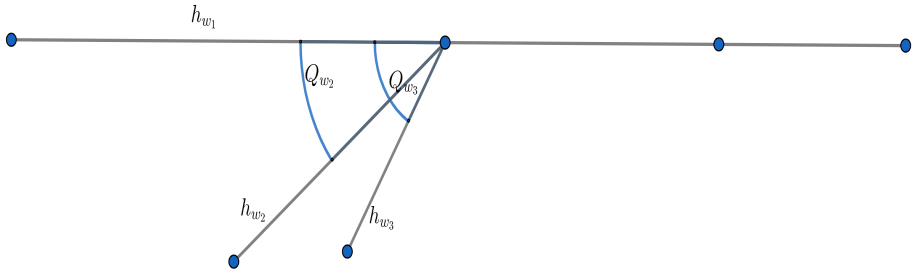


Figure A.3: Step 3 of the explanation.

3. One checks the switching condition for other point. For instance, we consider w_3 and check $Q_{w_3} = h_{w_1} - h_{w_3}$, see Fig. A.3.

4. Keep doing this until there are no more points left.

We obtain the following boundary value problem with switching function from the first order necessary conditions of optimality (the Maximum Principle)

$$\begin{aligned}
&\dot{x}(t) = F(x(t), u(t), \alpha_w),\\
&\dot{\lambda}(t) = -\frac{\partial \bar{\mathcal{H}}^T}{\partial x} = -\frac{\partial F^T(x(t), u, \alpha)}{\partial x}\lambda(t) - \delta\frac{\partial l^T(x(t), u)}{\partial x},\\
&x(t_0) = x_0,\\
&\lambda(t_f) = -\frac{\partial m^T(x(t_f))}{\partial x} + \rho\frac{\partial r(x(t_0), x(t_f))}{\partial x},\\
&\{u^*, \alpha^*\} = \arg\max_{u\in\mathcal{U}, \alpha\in[0,1]^{2^{n_w}}} \{\bar{\mathcal{H}}(x(t), u, \alpha, \lambda(t))\} \quad \text{a.e. on } [t_0, t_f]\\
&w^*(x, \lambda) = \arg\max_{w\in\mathcal{W}}\{h_w(x, \lambda, w)\},\\
&Q_j = h_{\hat{w}} - h_j = 0.
\end{aligned} \tag{A.10}$$

System (A.10) is solved with an advanced Multiple Shooting method, which is capable of treating such multi-point boundary value problems. The controls are indirectly determined by pointwise optimization of the Hamiltonian as functions of states and adjoints.

A.4 A Second-Order Sufficient Condition

This section covers a second-order sufficient condition for a Weak Local Minimum in an OCP with mixed state-control constraints. Main proposed result and the proof of the main theorem are considered in details.
Sufficient second-order conditions in optimal control are earlier investigated, see, for example, [26, 27, 94, 129]. The most general results on sufficient conditions of the second order in optimal control were published by Osmolovskii, see [98, 99].
Now we will formulate sufficient conditions of the second order for a weak local minimum. For more general, we consider OCP as follows:

$$\begin{aligned}
\min_{x(\cdot),u(\cdot),\alpha(\cdot)} \quad & J(x, u, \alpha) := m(x(0), x(T))\\
\text{s.t.} \quad & \dot{x}(t) = F(x(t), u(t), \alpha(t)) \text{ for a.a. } t \in [0, T],\\
& \mathcal{G}(x(t), \alpha(t), \varepsilon) \le 0 \text{ for a.a. } t \in [0, T],\\
& r(x(0), x(T)) \le 0.
\end{aligned} \tag{A.11}$$

where the mixed state-control constraints $\mathcal{G}(x(t), \alpha(t))$ is defined by

$$\mathcal{G}(x(t), \alpha(t), \varepsilon) = \begin{pmatrix} -\alpha(t)\sigma(x(t)) - \varepsilon \\ (1-\alpha(t))\sigma(x(t)) - \varepsilon \end{pmatrix}$$

with $\varepsilon > 0$ small enough. Denoting

$$q = (x(0), x(T)) = (x_0, x_T), \quad y = (x, u, \alpha), \quad \mathcal{Y} = \mathcal{X} \times \mathcal{U} \times [0, 1].$$

The local minimum here is a weak local minimum.
Let $\hat{y} = (\hat{x}, \hat{u}, \hat{\alpha}) \in \mathcal{Y}$ is an admissible of (A.11). Set $\hat{q} = (\hat{x}(0), \hat{x}(T))$.
We recall the Hamiltonian and the augmented Hamiltonian of (A.11)

$$\mathcal{H}(y, \lambda) = \lambda F(x, u, \alpha) = \lambda F(y), \quad \bar{\mathcal{H}}(y, \lambda, \mu) = \lambda F(y) + \mu\mathcal{G}(x, \alpha, \varepsilon).$$

Set $M_0 = \{t \in [0,T] : \mathcal{G}(\hat{x}(t), \hat{\alpha}(t)) = 0\}$. Define the critical cone K as follows:

$$K = \{\ y \in \mathcal{Y} : \dot{x}(t) = \nabla F(\hat{y}(t))y(t),\ \frac{\partial \mathcal{H}}{\partial u}(\hat{y}(t), \hat{\lambda}(t))u(t) = 0, \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}(t), \hat{\lambda}(t))\alpha(t) = 0,$$
$$\text{for a.a. } t \in [0,T];\ \frac{\partial \mathcal{G}}{\partial \alpha}(\hat{x}(t), \hat{\alpha}(t), \varepsilon)\alpha(t) \le 0 \text{ for a.a. } t \in M_0 \Big\}. \quad \text{(A.12)}$$

Since the classical definition of a critical cone, we see that K can be defined in the following equivalent way

$$K = \{\ y \in \mathcal{Y} : \nabla m(\hat{q})q \le 0, \dot{x}(t) = \nabla F(\hat{y}(t))y(t),\ \text{a.e. on } [0,T],$$
$$\frac{\partial \mathcal{G}}{\partial \alpha}(\hat{x}(t), \hat{\alpha}(t), \varepsilon)\alpha(t) \le 0 \text{ a.e. on } M_0 \Big\}.$$

Note that the condition $K = \{0\}$ is not sufficient for local minimality of $\hat{y}$. The counter example can be seen in [100, Example 2.1]. Now let us formulate the assumptions for (A.11).

Assumption A.4.1 (on the regularity of mixed constraints)**.** [45, Assumption RMC] *The mixed constraints $\mathcal{G}(x(t), \alpha(t), \varepsilon) \le 0$ are regular in the following sense: at any point (x, α) satisfying relations $\mathcal{G} \le 0$, the system of vectors $\frac{\partial \mathcal{G}}{\partial \alpha}(x, \alpha, \varepsilon)$ is positively-linearly independent.*

Assumption A.4.2. *The first order necessary optimality condition for a weak local minimum for $\hat{y} = (\hat{x}, \hat{u}, \hat{\alpha})$ is fulfilled: there exist $\hat{p}$ and $\hat{\lambda}$ such that*

$$(-\hat{p}(0), \hat{p}(1)) = \nabla m(\hat{q}), \quad \text{(A.13)}$$
$$-\dot{\hat{\lambda}}(t) = \frac{\partial \mathcal{H}}{\partial x}(\hat{y}(t), \hat{\lambda}(t)) = \hat{\lambda}(t)\frac{\partial F}{\partial x}(\hat{y}(t)) \text{ for a.a. } t \in [0,T], \quad \text{(A.14)}$$
$$\frac{\partial \bar{\mathcal{H}}}{\partial u}(\hat{y}(t), \hat{\lambda}(t), \hat{\mu}(t)) = \hat{\lambda}(t)\frac{\partial F}{\partial u}(\hat{y}(t)) = 0 \text{ for a.a. } t \in [0,T], \quad \text{(A.15)}$$
$$\frac{\partial \bar{\mathcal{H}}}{\partial \alpha}(\hat{y}(t), \hat{\lambda}(t), \hat{\mu}(t)) = \hat{\lambda}(t)\frac{\partial F}{\partial \alpha}(\hat{y}(t)) + \hat{\mu}(t)\frac{\partial \mathcal{G}}{\partial \alpha}(\hat{x}(t), \hat{\alpha}(t), \varepsilon) = 0 \text{ for a.a. } t \in [0,T], \quad \text{(A.16)}$$
$$\hat{\mu}(t) \ge 0 \text{ for a.a. } t \in [0,T], \quad \text{(A.17)}$$
$$\hat{\mu}(t)\mathcal{G}(\hat{x}(t), \hat{\alpha}(t), \varepsilon) = 0 \text{ for a.a. } t \in [0,T]. \quad \text{(A.18)}$$

Assumption A.4.3. *There exist $C > 0$ and $\epsilon > 0$ such that for a.a. $t \in sm(\epsilon)$ we have*

$$\mathcal{H}(\hat{x}(t), \hat{u}(t), \alpha, \hat{\lambda}(t)) - \mathcal{H}(\hat{x}(t), \hat{u}(t), \hat{\alpha}(t), \hat{\lambda}(t)) \ge C|\alpha - \hat{\alpha}(t)|^2$$
$$\mathcal{H}(\hat{x}(t), u, \hat{\alpha}(t), \hat{\lambda}(t)) - \mathcal{H}(\hat{x}(t), \hat{u}(t), \hat{\alpha}(t), \hat{\lambda}(t)) \ge C|u - \hat{u}(t)|^2$$
$$\text{whenever } |\alpha - \hat{\alpha}(t)| < \epsilon,\ |u - \hat{u}(t)| < \epsilon,\ \mathcal{G}(x, \alpha, \varepsilon) \le 0. \quad \text{(A.19)}$$

Therein this set of a small measure has the form

$$sm(\epsilon) = \{t \in [0,1] : 0 < |\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{x}(t), \hat{u}(t), \hat{\alpha}(t), \hat{\lambda}(t))| < \epsilon,\ 0 < |\frac{\partial \mathcal{H}}{\partial u}(\hat{x}(t), \hat{u}(t), \hat{\alpha}(t), \hat{\lambda}(t))| < \epsilon\}, \quad \text{(A.20)}$$

where $\epsilon > 0$ is arbitrarily small.
We introduce the *quadratic form*:

$$\Omega(y) := q^T \nabla^2 m(\hat{q}) q + \int_0^T y(t)^T \bar{\mathcal{H}}(\hat{y}(t), \hat{\lambda}(t), \mu(t)) y(t) dt, \tag{A.21}$$

where $q = (x(0), x(T))$.

Assumption A.4.4. *There exists $c_0 > 0$ such that*

$$\Omega(y) \geq c_0 \left(\|x\|_\infty^2 + \|u\|_2^2 + |\alpha|^2 \right) \quad \forall y \in K. \tag{A.22}$$

Proposition 4. *Assumption A.4.4 is equivalent to the following one: there exists $c_0 > 0$ such that*

$$\omega(y) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}(t), \hat{\lambda}(t)) v(u,t) + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}(t), \hat{\lambda}(t)) v(\alpha, t) \right) dt \geq c_0 (\|x\|_\infty^2 + \|u\|_2^2 + |\alpha|^2) \tag{A.23}$$

for all $y = (x, u, \alpha) \in K$ and for all $v(\cdot) \in L^\infty$ such that $v(u,t) \in T_U^{b(2)}(\hat{u}(t), u(t))$, $v(\alpha, t) \in T_A^{b(2)}(\hat{\alpha}(t), \alpha(t))$ a.e. on M_0, where

$$\omega(y) = \frac{1}{2} q^T \nabla^2 m(\hat{q}) q + \frac{1}{2} \int_0^T y(t)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}(t), \hat{\lambda}(t)) y(t) dt,$$

and

$$T_A^{b(2)}(\hat{\alpha}(t), \alpha(t)) = \{ v \in \mathbb{R}^n : \frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha}, \varepsilon) v + \frac{1}{2} \alpha^T \frac{\partial^2 \mathcal{H}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \alpha \leq 0 \}$$

is the second-order tangent to the set A for the pair $(\hat{\alpha}, \alpha) \in \mathbb{R}^{2n}$, see, for instance, [39].

A proof of the above proposition can be found in [100, pp. 156].
The main result of this subsection, the following theorem holds.

Theorem 16 (Sufficient second order condition)**.** *Let Assumption A.4.1 - A.4.4 be fulfilled. Then there exist $\delta > 0$ and $c > 0$ such that*

$$J(y) - J(\hat{y}) \geq c \left(\|x - \hat{x}\|^2 + \|u - \hat{u}\|^2 + |\alpha - \hat{\alpha}| \right) \tag{A.24}$$

for all admissible $y = (x, u, \alpha) \in \mathcal{Y}$ such that $\|y - \hat{y}\| < \delta$.

Proof. The majority of the following proof is based on and modified from the proof in [100, Sec. 3]. We omit the dependence on t for x, u, α, $\hat{x}$, $\hat{u}$, $\hat{\alpha}$, etc.
Step 1 For $y = (x, u, \alpha) \in \mathcal{Y}$ we set $\Delta y = y - \hat{y}$ and $\gamma(\Delta y) = \|\Delta x\|_\infty^2 + \|\Delta u\|_2^2 + |\Delta \alpha|^2$. Assume that condition (A.24) does not hold. Then, there is a sequence of admissible points $y_n \neq \hat{y}$ such that $\|y_n - \hat{y}\| \to 0$ and

$$\Delta_n J := J(y_n) - J(\hat{y}) \leq o(\gamma_n), \tag{A.25}$$

where $\gamma_n = \gamma(\Delta y_n) > 0$, and $\Delta y_n = (\Delta x_n, \Delta u_n, \Delta\alpha_n) = y_n - \hat{y}$. Set $\Delta_n F := F(y_n) - F(\hat{y})$. Since $\Delta\dot{x}_n = \Delta_n F$, we imply $\Delta_n J = \Delta_n J + \int_0^T \hat{\lambda}(\Delta_n F - \Delta\dot{x}_n)dt$.
Moreover, $\int_0^T \hat{\lambda}\Delta\dot{x}_n dt = \hat{\lambda}\Delta x_n \,|_0^T - \int_0^T \hat{\lambda}\Delta x_n dt = \nabla m(\hat{\lambda})\Delta q_n + \int_0^T \hat{\lambda}\frac{\partial F}{\partial x}(\hat{y})\Delta x_n dt$.
Therefore,

$$\begin{aligned}\Delta_n J &= \Delta_n m - \nabla m(\hat{\lambda})\Delta q_n + \int_0^T (\hat{\lambda}\Delta_n F - \hat{\lambda}\frac{\partial F}{\partial x}(\hat{y})\Delta x_n)dt \\ &= \Delta_n m - \nabla m(\hat{\lambda})\Delta q_n + \int_0^T (\Delta_n \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y},\hat{\lambda})\Delta x_n)dt, \end{aligned} \tag{A.26}$$

where $\Delta_n\mathcal{H} = \mathcal{H}(y_n,\hat{\lambda}) - \mathcal{H}(\hat{y},\hat{\lambda})$.
Step 2 We have

$$\begin{aligned}\Delta_n\mathcal{H} &:= \mathcal{H}(\hat{x}+\Delta x_n, \hat{u}+\Delta u_n, \hat{\alpha}+\Delta\alpha_n, \hat{\lambda}) - \mathcal{H}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda}) \\ &= \mathcal{H}(\hat{x}+\Delta x_n, \hat{u}+\Delta u_n, \hat{\alpha}+\Delta\alpha_n, \hat{\lambda}) - \mathcal{H}(\hat{x}, \hat{u}+\Delta u_n, \hat{\alpha}+\Delta\alpha_n, \hat{\lambda}) - \mathcal{H}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda}) \\ &\quad + \mathcal{H}(\hat{x}, \hat{u}+\Delta u_n, \hat{\alpha}+\Delta\alpha_n, \hat{\lambda}) - \mathcal{H}(\hat{x},\hat{u},\hat{\alpha}+\Delta\alpha_n,\hat{\lambda}) + \mathcal{H}(\hat{x},\hat{u},\hat{\alpha}+\Delta\alpha_n,\hat{\lambda}) \\ &= \frac{\partial\mathcal{H}}{\partial x}(\hat{x},\hat{u}+\Delta u_n,\hat{\alpha}+\Delta\alpha_n,\hat{\lambda})\Delta x_n + \Delta_{un}\mathcal{H} + \Delta_{\alpha n}\mathcal{H} + r_n,\end{aligned}$$

where $\|r_n\|_\infty = O(\gamma_n)$, $\Delta_{un}\mathcal{H} := \mathcal{H}(\hat{x},\hat{u}+\Delta u_n,\hat{\alpha}+\Delta\alpha_n,\hat{\lambda}) - \mathcal{H}(\hat{x},\hat{u},\hat{\alpha}+\Delta\alpha_n,\hat{\lambda})$, and $\Delta_{\alpha n}\mathcal{H} := \mathcal{H}(\hat{x},\hat{u},\hat{\alpha}+\Delta\alpha_n,\hat{\lambda}) - \mathcal{H}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})$. Let $\epsilon_n \searrow 0$. Set

$$sm(\epsilon_n) = \{t\in[0,1] : 0 < \left|\frac{\partial\mathcal{H}}{\partial\alpha}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\right| < \epsilon_n,\ 0 < \left|\frac{\partial\mathcal{H}}{\partial u}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\right| < \epsilon\}.$$

Clearly, $sm(\epsilon_n) \subset M_0$ and meas $sm(\epsilon_n) \to 0$ as $n\to\infty$. Since $\mathcal{G}(\cdot,\alpha_n) \le 0$ for all n, then, from Assumption A.4.3, we have $\Delta_{\alpha n}\mathcal{H} \ge C|\alpha_n|^2$, and $\Delta_{un}\mathcal{H} \ge C|u_n|^2$ for all sufficiently large n. Therefore,

$$\int_{sm(\epsilon_n)} \Delta_{\alpha n}\mathcal{H}\ dt \ge C\int_{sm(\epsilon_n)} |\alpha_n|^2 dt, \quad \int_{sm(\epsilon_n)} \Delta_{un}\mathcal{H}\ dt \ge C\int_{sm(\epsilon_n)} |u_n|^2 dt.$$

Consequently,

$$\begin{aligned}\int_{sm(\epsilon_n)} \left(\Delta_n\mathcal{H} - \frac{\partial\mathcal{H}}{\partial x}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\Delta x_n\right)dt &\ge \int_{sm(\epsilon_n)} \frac{\partial\mathcal{H}}{\partial x}(\hat{x},\hat{u}+\Delta u_n,\hat{\alpha}+\Delta\alpha_n,\hat{\lambda})\Delta x_n dt \\ &\quad - \int_{sm(\epsilon_n)} \frac{\partial\mathcal{H}}{\partial x}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\Delta x_n dt + C\int_{sm(\epsilon_n)} \left(|\alpha_n|^2 + |u_n|^2\right)dt + o(\gamma_n).\end{aligned}$$

Since

$$\begin{aligned}&\int_{sm(\epsilon_n)} |\Delta\alpha_n|\cdot|\Delta x_n| dt \le \|\Delta x_n\|_\infty \sqrt{\text{meas } sm(\epsilon_n)}\,|\Delta\alpha_n| = o(\gamma_n), \\ &\int_{sm(\epsilon_n)} |\Delta u_n|\cdot|\Delta x_n| dt \le \|\Delta x_n\|_\infty \sqrt{\text{meas } sm(\epsilon_n)}\,\|\Delta u_n\|_2 = o(\gamma_n),\end{aligned}$$

we imply $\int_{sm(\epsilon_n)} \left(\frac{\partial \mathcal{H}}{\partial x}(\hat{x}, \hat{u} + \Delta u_n, \hat{\alpha} + \Delta \alpha_n, \hat{\lambda}) - \frac{\partial \mathcal{H}}{\partial x}(\hat{x}, \hat{u}, \hat{\alpha}, \hat{\lambda}) \right) \Delta x_n = o(\gamma_n)$. Therefore,

$$\int_{sm(\epsilon_n)} \left(\Delta_n \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{x}, \hat{u}, \hat{\alpha}, \hat{\lambda}) \Delta x_n \right) dt \geq C \int_{sm(\epsilon_n)} \left(|\alpha_n|^2 + |u_n|^2 \right) dt + o(\gamma_n). \quad \text{(A.27)}$$

Step 3 Conditions (A.25)-(A.27) imply

$$\begin{aligned} o(\gamma_n) &\geq \Delta_n m - \nabla m(\hat{\lambda}) \Delta q_n + \int_0^T (\Delta_n \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt \\ &\geq \frac{1}{2} \Delta q_n^T \nabla^2 m(\hat{\lambda}) \Delta q_n + o(|\Delta q_n|^2) + \int_{[0,T] \setminus sm(\epsilon_n)} (\Delta_n \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt \\ &\quad + C \int_{sm(\epsilon_n)} \left(|\alpha_n|^2 + |u_n|^2 \right) dt + o(\gamma_n). \quad \text{(A.28)} \end{aligned}$$

We set $u_n' := \Delta u_n \chi_{sm(\epsilon_n)}$, $\Delta u_n^0 := \Delta u_n - u_n'$, $\Delta y_n^0 := (\Delta x_n, \Delta u_n^0, \Delta \alpha_n^0)$, $\alpha_n' := \Delta \alpha \chi_{sm(\epsilon_n)}$, $\Delta \alpha^0 := \Delta \alpha - \alpha_n'$, $\gamma_n^0 := \gamma(\Delta y_n^0)$, and $\gamma_n' := \int_0^T (|u_n'| + |\alpha_n'|) \, dt = \int_{sm(\epsilon_n)} \left(|\alpha_n|^2 + |u_n|^2 \right) dt$. Then we have $\gamma_n = \gamma_n^0 + \gamma_n'$. Further, set $\Delta_n^0 \mathcal{H} := \mathcal{H}(\hat{y} + \Delta y_n^0, \hat{\lambda}) - \mathcal{H}(\hat{y}, \hat{\lambda})$. Then

$$\begin{aligned} \int_{[0,T] \setminus sm(\epsilon_n)} (\Delta_n \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt &= \int_{[0,T] \setminus sm(\epsilon_n)} (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt \\ &= \int_0^T (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt - \int_{sm(\epsilon_n)} (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt. \end{aligned}$$

Obviously, we have $\int_{sm(\epsilon_n)} (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt = o(\gamma_n)$. Hence, we get

$$\int_{[0,T] \setminus sm(\epsilon_n)} (\Delta_n \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt = \int_0^T (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt + o(\gamma_n). \quad \text{(A.29)}$$

Note that $\frac{\partial \mathcal{H}}{\partial y}(\hat{y}, \hat{\lambda}) \Delta y_n^0 = \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n + \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0$. Therefore, relations (A.28) and (A.29) imply

$$\begin{aligned} o(\gamma_n) &\geq \frac{1}{2} \Delta q_n^T \nabla^2 m(\hat{\lambda}) \Delta q_n + \int_0^T (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt + C \gamma_n' \\ &= \frac{1}{2} \Delta q_n^T \nabla^2 m(\hat{\lambda}) \Delta q_n + \int_0^T (\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial x}(\hat{y}, \hat{\lambda}) \Delta x_n) dt \\ &\quad + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right) dt + C \gamma_n'. \end{aligned}$$

Since $\Delta_n^0 \mathcal{H} - \frac{\partial \mathcal{H}}{\partial y}(\hat{y}, \hat{\lambda}) \Delta y_n^0 = \frac{1}{2} \left(\Delta y_n^0 \right)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \Delta y_n^0 + o(|\Delta y_n^0|^2)$, we obtain from here that

$$\begin{aligned} o(\gamma_n) \geq \frac{1}{2} \Delta q_n^T \nabla^2 m(\hat{\lambda}) \Delta q_n &+ \int_0^T \left(\Delta y_n^0 \right)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \Delta y_n^0 dt \\ &+ \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right) dt + C \gamma_n', \end{aligned}$$

or, equivalently,

$$\omega(\Delta y_n^0) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right) dt + C\gamma_n' \leq o(\gamma_n), \tag{A.30}$$

where $\omega(y) = \frac{1}{2} q^T \nabla^2 m(\hat{q}) q + \frac{1}{2} \int_0^T y(t)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}(t), \hat{\lambda}(t)) y(t) dt$.
Step 4 Since $\omega(\Delta y_n^0) \leq O(\gamma_n^0) \leq O(\gamma_n)$, relation (A.30) implies

$$\int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right) dt \leq O(\gamma_n). \tag{A.31}$$

Further, condition $\mathcal{G}(\cdot, \hat{\alpha} + \Delta\alpha_n^0) \leq 0$ yields $\Delta_{\alpha n}^0 \mathcal{H} \geq C|\Delta\alpha_n|^2$, $\Delta_{un}^0 \mathcal{H} \geq C|\Delta u_n|^2$, and then

$$\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \geq O(|\Delta \alpha_n^0|^2), \quad \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 \geq O(|\Delta u_n^0|^2), \quad \text{a.e. on } M_0.$$

It follows that

$$\left(\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha^0 \right)^- \leq O(|\Delta \alpha_n^0|^2), \ \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 \right)^- \leq O(|\Delta u_n^0|^2) \text{ a.e. on } M_0, \tag{A.32}$$

where $a^- = \max\{-a, 0\}$, $a^+ = \max\{a, 0\}$, $a = a^+ - a^-$ for $a \in \mathbb{R}$.
Let us analysis conditions (A.31) and (A.32). We rewrite (A.31) in the following form

$$\int_0^T \left[\left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 \right)^+ + \left(\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right)^+ \right] dt$$
$$- \int_0^T \left[\left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 \right)^- + \left(\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right)^- \right] dt \leq O(\gamma_n). \tag{A.33}$$

Since, combine with (A.32), we imply $\int_0^T \left[\left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 \right)^+ + \left(\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right)^+ \right] dt \leq O(\gamma_n)$.
Consequently,

$$\int_0^T \left(\left| \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 \right| + \left| \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right| \right) dt \leq O(\gamma_n). \tag{A.34}$$

Step 5 Condition $\mathcal{G}(\cdot, \hat{\alpha} + \Delta\alpha_n^0) \leq 0$ yields

$$\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha}, \varepsilon) \Delta \alpha_n^0 + \frac{1}{2} (\Delta \alpha_n^0)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}, \varepsilon) \Delta \alpha_n^0 \leq o(|\Delta \alpha_n^0|^2) \text{ a.e. on } M_0, \tag{A.35}$$

By multiplying this equality with $\hat{\lambda} \geq 0$ and by taking into account $\hat{\mu} \frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha}) = -\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{p})$, we obtain

$$-\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 + \frac{1}{2} \hat{\lambda} (\Delta \alpha_n^0)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}, \varepsilon) \Delta \alpha_n^0 \leq o(|\Delta \alpha_n^0|^2) \text{ a.e. on } M_0, \tag{A.36}$$

with $-\int_0^T \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 dt + \int_0^T \frac{\hat{\mu}}{2} (\Delta \alpha_n^0)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}, \varepsilon) \Delta \alpha_n^0 dt \leq o(\gamma_n)$.
Upon putting this inequality to (A.30) and using $\bar{\mathcal{H}}(y, \lambda, \mu) = \lambda F(y) + \mu \mathcal{G}(x, \alpha, \varepsilon)$, we get

$$\Omega(\Delta \alpha_n^0) + C\gamma_n' \leq o(\gamma_n). \tag{A.37}$$

In the following $\mathcal{G}(\cdot,\alpha) := \mathcal{G}(x,\alpha,\varepsilon)$. Now, with $\gamma_n > 0$ for all n, we consider two possible cases:

$$(i)\ \liminf \frac{\gamma_n^0}{\gamma_n} = 0, \qquad (ii)\ \liminf \frac{\gamma_n^0}{\gamma_n} > 0.$$

Step 6 In case (i), there is a subsequence such that $\gamma_n^0/\gamma_n \to 0$ in this subsequence. Assume that the sequence itself satisfies this condition. Then, $\gamma_n^0 = o(\gamma_n)$. Since, obviously, $|\Omega(\Delta y_n^0)| \le O(\gamma_n^0)$, condition (A.37) implies

$$C\gamma_n' \le o(\gamma_n) + O(\gamma_n^0) = o_1(\gamma_n),$$

i.e., $\gamma_n' = o(\gamma_n)$. The latter contradicts the conditions $\gamma_n^0 = o(\gamma_n)$ and $\gamma_n^0 + \gamma_n' = \gamma_n > 0$.
Step 7 Case (ii) is the main case, where $\gamma_n = O(\gamma_n^0)$. Let us rewrite (A.30) in the form

$$\frac{\gamma_n^0}{\gamma_n} \cdot \frac{\omega(\Delta y_n^0) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y},\hat{\lambda})\Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y},\hat{\lambda})\Delta \alpha_n^0 \right) dt}{\gamma_n^0} + \frac{\gamma_n'}{\gamma_n} \cdot C \le o(1).$$

Thus, it follows

$$\min \left\{ \frac{\omega(\Delta y_n^0) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y},\hat{\lambda})\Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y},\hat{\lambda})\Delta \alpha_n^0 \right) dt}{\gamma_n^0}, C \right\} \le o(1).$$

Since $C > 0$, we obtain

$$\frac{\omega(\Delta y_n^0) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y},\hat{\lambda})\Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y},\hat{\lambda})\Delta \alpha_n^0 \right) dt}{\gamma_n^0} \le o(1),$$

or, equivalently,

$$\omega(\Delta y_n^0) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y},\hat{\lambda})\Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y},\hat{\lambda})\Delta \alpha_n^0 \right) dt \le o(\gamma_n^0). \tag{A.38}$$

Note that in general, Δy_n^0 does not belong to the critical cone K. We find a sequence $\delta y_n \in K$, which is "close" in some sense to Δy_n^0, and then we use condition (A.38) to analyze this condition by using Assumption (A.4.4).
Step 8 Set

$$M_+\left(\frac{\partial \mathcal{H}}{\partial u}\right) := \{t \in [0,T] : \left|\frac{\partial \mathcal{H}}{\partial u}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\right| > 0\},$$
$$M_+\left(\frac{\partial \mathcal{H}}{\partial \alpha}\right) := \{t \in [0,T] : \left|\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\right| > 0\},$$
$$M_+\left(\frac{\partial \mathcal{H}}{\partial u}, \epsilon_n\right) := \{t \in [0,T] : \left|\frac{\partial \mathcal{H}}{\partial u}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\right| \ge \epsilon_n\},$$
$$M_+\left(\frac{\partial \mathcal{H}}{\partial \alpha}, \epsilon_n\right) := \{t \in [0,T] : \left|\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda})\right| \ge \epsilon_n\},$$
$$M_0\left(\frac{\partial \mathcal{H}}{\partial u}\right) := \{t \in M_0 : \frac{\partial \mathcal{H}}{\partial u}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda}) = 0\},$$
$$M_0\left(\frac{\partial \mathcal{H}}{\partial \alpha}\right) := \{t \in M_0 : \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{x},\hat{u},\hat{\alpha},\hat{\lambda}) = 0\}.$$

Then

$$M_0 = M_0\left(\frac{\partial \mathcal{H}}{\partial u}\right) \cup M_+\left(\frac{\partial \mathcal{H}}{\partial u}\right) \cup M_0\left(\frac{\partial \mathcal{H}}{\partial \alpha}\right) \cup M_+\left(\frac{\partial \mathcal{H}}{\partial \alpha}\right)$$
$$= M_0\left(\frac{\partial \mathcal{H}}{\partial u}\right) \cup M_+\left(\frac{\partial \mathcal{H}}{\partial u}, \epsilon_n\right) \cup sm(\epsilon_n) \cup M_0\left(\frac{\partial \mathcal{H}}{\partial \alpha}\right) \cup M_+\left(\frac{\partial \mathcal{H}}{\partial \alpha}, \epsilon_n\right). \tag{A.39}$$

In view of (A.35), there exists $\tilde{\alpha}_{1n}$ and $\tilde{u}_{1n}$ such that

$$\tilde{\alpha}_{1n}\chi_{M_0(\frac{\partial \mathcal{H}}{\partial \alpha})} = \tilde{\alpha}_{1n}, \qquad \frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha}, \varepsilon)(\Delta\alpha_n^0 + \tilde{\alpha}_{1n})\chi_{M_0(\frac{\partial \mathcal{H}}{\partial \alpha})} \leq 0, \tag{A.40}$$
$$\tilde{u}_{1n}\chi_{M_0(\frac{\partial \mathcal{H}}{\partial u})} = \tilde{u}_{1n}, \qquad |\tilde{\alpha}_{1n}| \leq O(|\Delta\alpha_n^0|^2), \qquad |\tilde{u}_{1n}| \leq O(|\Delta u_n^0|^2). \tag{A.41}$$

hereinafter χ_M stands for the characteristic function of M, and therefore,

$$|\tilde{\alpha}_{1n}| \leq O(\gamma_n), \qquad |\tilde{\alpha}_{1n}| \leq O(|\Delta\alpha_n|^2) = o(1), \tag{A.42}$$
$$\|\tilde{u}_{1n}\|_1 \leq O(\gamma_n), \qquad \|\tilde{u}_{1n}\|_\infty \leq O(\|\Delta u_n\|_\infty^2) = o(1).$$

Moreover, one sets

$$\frac{\partial \mathcal{H}}{\partial u}^0 = \frac{\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda})}{\left|\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda})\right|},\ t \in M_+\left(\frac{\partial \mathcal{H}}{\partial u}\right), \quad \frac{\partial \mathcal{H}}{\partial \alpha}^0 = \frac{\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda})}{\left|\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda})\right|},\ t \in M_+\left(\frac{\partial \mathcal{H}}{\partial \alpha}\right).$$

There exists $\tilde{\alpha}_{2n}$ and $\tilde{u}_{2n}$ such that

$$\tilde{\alpha}_{2n}\chi_{M_+(\frac{\partial \mathcal{H}}{\partial \alpha}, \epsilon_n)} = \tilde{\alpha}_{2n}, \qquad \mathcal{H}(\hat{y}, \hat{\lambda})\left(\Delta\alpha_n^0 + \tilde{\alpha}_{2n}\right)\chi_{M_+(\frac{\partial \mathcal{H}}{\partial \alpha}, \epsilon_n)}, \tag{A.43}$$
$$\tilde{u}_{2n}\chi_{M_+(\frac{\partial \mathcal{H}}{\partial u}, \epsilon_n)} = \tilde{u}_{2n}, \qquad \mathcal{H}(\hat{y}, \hat{\lambda})\left(\Delta u_n^0 + \tilde{u}_{2n}\right)\chi_{M_+(\frac{\partial \mathcal{H}}{\partial u}, \epsilon_n)},$$

$$|\tilde{\alpha}_{2n}| \leq O\left(\left|\frac{\partial \mathcal{H}}{\partial \alpha}^0(\hat{y}, \hat{\lambda})\Delta\alpha_n^0\right|\right)\chi_{M_+(\frac{\partial \mathcal{H}}{\partial \alpha}, \epsilon_n)} \leq \frac{1}{\epsilon_n}O\left(\left|\frac{\partial \mathcal{H}}{\partial \alpha}^0(\hat{y}, \hat{\lambda})\Delta\alpha_n^0\right|\right)\chi_{M_+(\frac{\partial \mathcal{H}}{\partial \alpha}, \epsilon_n)},$$
$$|\tilde{u}_{2n}| \leq O\left(\left|\frac{\partial \mathcal{H}}{\partial u}^0(\hat{y}, \hat{\lambda})\Delta u_n^0\right|\right)\chi_{M_+(\frac{\partial \mathcal{H}}{\partial u}, \epsilon_n)} \leq \frac{1}{\epsilon_n}O\left(\left|\frac{\partial \mathcal{H}}{\partial u}^0(\hat{y}, \hat{\lambda})\Delta u_n^0\right|\right)\chi_{M_+(\frac{\partial \mathcal{H}}{\partial u}, \epsilon_n)}.$$

Consequently, $|\tilde{\alpha}_{2n}| \leq O(|\Delta\alpha_n|) = o(1)$, and $\|\tilde{u}_{2n}\|_\infty \leq O(\|\Delta u_n\|_\infty) = o(1)$.
Taking into account the estimation (A.34), we imply

$$|\tilde{\alpha}_{2n}| \leq \frac{1}{\epsilon_n}O(\gamma_n), \qquad \|\tilde{u}_{2n}\|_1 \leq \frac{1}{\epsilon_n}O(\gamma_n). \tag{A.44}$$

Choose $\epsilon_n > 0$ such that

$$\frac{\|\Delta y_n\|_\infty}{\epsilon_n} \to 0. \tag{A.45}$$

Then, one implies $\frac{1}{\epsilon_n}O(\gamma_n) = o(\sqrt{\gamma_n})$. Consequently,

$$|\tilde{\alpha}_{2n}| = o(\sqrt{\gamma_n}), \qquad \|\tilde{u}_{2n}\|_1 = o(\sqrt{\gamma_n}). \tag{A.46}$$

Set $\tilde{\alpha}_n = \tilde{\alpha}_{1n} + \tilde{\alpha}_{2n}$, and $\tilde{u}_n = \tilde{u}_{1n} + \tilde{u}_{2n}$. Then $|\tilde{\alpha}_n| \le O(|\Delta\alpha_n|) = o(1)$, $\|\tilde{u}_n\|_\infty \le O(\|\Delta u_n\|_\infty) = o(1)$, and

$$|\tilde{\alpha}_n| = o(\sqrt{\gamma_n}), \qquad |\tilde{\alpha}_n|^2 \le |\tilde{\alpha}_n|\,|\tilde{\alpha}_n| \le \frac{|\tilde{\alpha}_n|}{\epsilon_n} O(\gamma_n) = o(\gamma_n), \tag{A.47}$$

$$\|\tilde{u}_n\|_1 = o(\sqrt{u_n}), \qquad \|\tilde{u}_n\|_2^2 \le \|\tilde{u}_n\|_\infty \|\tilde{u}_n\|_1 \le \frac{\|\tilde{u}_n\|_\infty}{\epsilon_n} O(\gamma_n) = o(\gamma_n).$$

Moreover, since (A.39), (A.40), and (A.43), we have

$$\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha})(\Delta\alpha_n^0 + \tilde{\alpha}_n) \le 0, \qquad \frac{\partial \mathcal{G}}{\partial u}(\cdot, \hat{\alpha})(\Delta u_n^0 + \tilde{u}_n) \le 0 \qquad \text{a.e. on } M_0, \tag{A.48}$$

$$\frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda})(\Delta\alpha_n^0 + \tilde{\alpha}_n) = 0, \qquad \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda})(\Delta u_n^0 + \tilde{u}_n) = 0. \tag{A.49}$$

Set $\bar{\alpha}_n = -\alpha_n' + \tilde{\alpha}_n$, $\delta\alpha_n = \Delta\alpha_n + \bar{\alpha}_n = \Delta\alpha_n^0 + \tilde{\alpha}_n$, and $\bar{u}_n = -\alpha_n' + \tilde{u}_n$, $\delta u_n = \Delta u_n + \bar{u}_n = \Delta u_n^0 + \tilde{u}_n$. Then

$$\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha})\delta\alpha_n \le 0,\ \frac{\partial \mathcal{G}}{\partial u}(\cdot)\delta u_n \le 0 \ \text{ a.e. on } M_0, \quad \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda})\delta\alpha_n = 0,\ \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda})\delta u_n = 0. \tag{A.50}$$

Note that $|\alpha_n'| \le \sqrt{\operatorname{meas} sm(\epsilon_n)}\,|\alpha_n'| = o(|\alpha_n'|) = o(\sqrt{\gamma_n'}) = o(\sqrt{\gamma_n})$, and $\|u_n'\|_1 \le \sqrt{\operatorname{meas} sm(\epsilon_n)}\,\|u_n'\|_2 = o(\|u_n'\|_2) = o(\sqrt{\gamma_n'}) = o(\sqrt{\gamma_n})$. Therefore,

$$|\bar{\alpha}_n| = o(\sqrt{\gamma_n}), \qquad \|\bar{u}_n\|_1 = o(\sqrt{\gamma_n}). \tag{A.51}$$

Step 9 The equation $\Delta\dot{x}_n = \Delta_n F$ implies

$$\Delta\dot{x}_n = \frac{\partial F}{\partial x}(\hat{y})\Delta x_n + \frac{\partial F}{\partial u}(\hat{y})\Delta u_n + \frac{\partial F}{\partial \alpha}(\hat{y})\Delta\alpha_n + O(|\Delta y_n|^2). \tag{A.52}$$

There exists δx_n such that

$$\delta\dot{x}_n = \frac{\partial F}{\partial x}(\hat{y})\Delta x_n + \frac{\partial F}{\partial u}(\hat{y})\Delta u_n + \frac{\partial F}{\partial \alpha}(\hat{y})\Delta\alpha_n, \quad \delta x_n(0) = \Delta x_n(0). \tag{A.53}$$

Then, it follows from Eq. (A.52) and Eq. (A.53) that $\delta x_n = \Delta x_n + \bar{x}_n$, where $\bar{x}_n$ satisfies

$$\dot{\bar{x}} = \frac{\partial F}{\partial x}(\hat{y})\Delta x_n + \frac{\partial F}{\partial u}(\hat{y})\Delta u_n + \frac{\partial F}{\partial \alpha}(\hat{y})\Delta\alpha_n - O(|\Delta y_n|^2), \quad \bar{x}_n(0) = 0.$$

This implies the following estimation

$$\|\bar{x}_n\|_\infty \le O(\|\bar{u}_n\|_1) + O(|\bar{\alpha}_n|) + O(|\Delta y_n|_2^2) = o(\sqrt{\gamma_n}). \tag{A.54}$$

Set $\bar{y}_n = (\bar{x}_n, \bar{u}_n, \bar{\alpha}_n)$, and $\delta y_n = (\delta x_n, \delta u_n, \delta\alpha_n) := \Delta y_n^0 + \bar{y}_n$. Then, according to (A.50) and (A.53), we see that

$$\delta y_n \in K. \tag{A.55}$$

Step 10 Let us compare $\omega(\delta y_n)$ with $\omega(\Delta y_n^0)$. We have

$$\begin{aligned}(\delta y_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda})\delta y_n &= \left(\Delta y_n^0 + \bar{y}_n\right)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda})\left(\Delta y_n^0 + \bar{y}_n\right) \\ &= \left(\Delta y_n^0\right)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda})\delta y_n^0 + 2\bar{y}_n \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda})\Delta y_n^0 + (\bar{y}_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda})\bar{y}_n.\end{aligned}$$

Similarly,

$$(\delta q_n)^T \nabla^2 m(\hat{q}) \delta q_n = (\Delta q_n + \bar{q}_n)^T \nabla^2 m(\hat{q}) (\Delta q_n + \bar{q}_n)$$
$$= (\Delta q_n)^T \nabla^2 m(\hat{q}) \delta q_n + 2\bar{q}_n \nabla^2 m(\hat{q}) \Delta q_n + (\bar{q}_n)^T \nabla^2 m(\hat{q}) \bar{q}_n,$$

where $\delta q_n = (\delta x_n(0), \delta x_n(T))$, $\Delta q_n = (\Delta x_n(0), \Delta x_n(T))$, $\bar{q}_n = (\bar{x}_n(0), \bar{x}_n(T))$. Therefore,

$$\omega(\delta y_n) = \omega(\Delta y_n^0) + r_\omega(n),$$

where

$$r_\omega(n) = 2(\bar{q}_n)^T \nabla^2 m(\hat{q}) \Delta q_n + \bar{q}_n^T \nabla^2 m(\hat{q}) \bar{q}_n + \int_0^T (2(\bar{y}_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \Delta y_n^0 + \bar{y}_n^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \bar{y}_n) dt.$$

We will show that

$$|r_\omega(n)| = o(\gamma_n). \tag{A.56}$$

First, we have

$$\bar{y}_n \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \Delta y_n^0 = \bar{x}_n \frac{\partial^2 \mathcal{H}}{\partial x^2}(\hat{y}, \hat{\lambda}) \Delta x_n + \bar{x}_n \frac{\partial^2 \mathcal{H}}{\partial x \partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \bar{x}_n \frac{\partial^2 \mathcal{H}}{\partial x \partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0$$
$$+ \tilde{u}_n \frac{\partial^2 \mathcal{H}}{\partial u^2}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \tilde{u}_n \frac{\partial^2 \mathcal{H}}{\partial u \partial x}(\hat{y}, \hat{\lambda}) \Delta x_n + \tilde{u}_n \frac{\partial^2 \mathcal{H}}{\partial u \partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0$$
$$+ \tilde{\alpha}_n \frac{\partial^2 \mathcal{H}}{\partial \alpha^2}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 + \tilde{\alpha}_n \frac{\partial^2 \mathcal{H}}{\partial \alpha \partial x}(\hat{y}, \hat{\lambda}) \Delta x_n + \tilde{\alpha}_n \frac{\partial^2 \mathcal{H}}{\partial \alpha \partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0.$$

According to (A.54) and the first two estimations in (A.47) we get

$$(\|\Delta x_n\|_\infty + \|\Delta u_n\|_1 + |\Delta \alpha_n|) \|\bar{x}_n\|_\infty + (\|\Delta x_n\|_\infty + |\Delta \alpha_n|) \|\tilde{u}_n\|_1$$
$$+ |\tilde{\alpha}_n| (|\Delta \alpha_n| + \|\Delta x_n\|_\infty + \|\tilde{u}_n\|_1) = o(\gamma_n).$$

Let us estimate $\left\| |\Delta u_n^0| \cdot |\tilde{u}_n| \right\|_1$. Using the equality in (A.42), (A.44) and condition (A.45), we obtain

$$\int_0^T |\Delta u_n^0| \cdot |\tilde{u}_n| \, dt = \int_0^T |\Delta u_n^0| \cdot |\tilde{u}_{1n} + \tilde{u}_{2n}| \, dt \le \left\|\Delta u_n^0\right\|_\infty |\tilde{u}_{1n}|_1 + \left\|\Delta u_n^0\right\|_\infty |\tilde{u}_{2n}|_1$$
$$\le \left\|\Delta u_n^0\right\|_\infty O(\gamma_n) + \left\|\Delta u_n^0\right\|_\infty \frac{1}{\epsilon_n} O(\gamma_n) = o(\gamma_n). \tag{A.57}$$

Therefore, $\left\| (\bar{y}_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \Delta y_n^0 \right\| = o(\gamma_n)$.
Secondly, similarly way by using (A.54) and (A.47) we get

$$\left\| (\bar{y}_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \bar{y}_n \right\|_1 = o(\gamma_n).$$

Consequently, $\left| \int_0^T \left(2(\bar{y}_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \Delta y_n^0 + (\bar{y}_n)^T \frac{\partial^2 \mathcal{H}}{\partial y^2}(\hat{y}, \hat{\lambda}) \bar{y}_n \right) dt \right| = o(\gamma_n)$.
In addition,

$$\left| 2(\bar{q}_n)^T \nabla^2 m(\hat{q}) \Delta q_n + (\bar{q}_n)^T \nabla^2 m(\hat{q}) \bar{q}_n \right| \le c \left(\|\Delta x_n\|_\infty \|\bar{x}_n\|_\infty \right) + \|\bar{x}_n\|_\infty^2 = o(\gamma_n),$$

with some $c > 0$. This yields the estimation (A.56). Consequently,

$$\omega(\delta y_n) = \omega(\Delta y_n^0) + o(\gamma_n). \tag{A.58}$$

Step 11 Now let us compare $\gamma(\delta y_n)$ with $\gamma_n = \gamma(\Delta y_n)$. Similarly to Step 10, we obtain

$$\gamma(\delta y_n) = \gamma_n + o(\gamma_n). \tag{A.59}$$

Step 12 Finally, we consider the term $\int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta \alpha_n^0 \right) dt$ in the inequality (A.38). Let us use (A.35). Since

$$\begin{aligned}(\delta\alpha_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \delta\alpha_n &= (\Delta\alpha_n^0 + \tilde{\alpha}_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha})(\Delta\alpha_n^0 + \tilde{\alpha}_n) \\ &= (\Delta\alpha_n^0)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \Delta\alpha_n^0 + r_{\mathcal{G}(\alpha)}(n),\end{aligned}$$

where

$$r_{\mathcal{G}(\alpha)}(n) = 2(\tilde{\alpha}_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \Delta\alpha_n^0 + (\tilde{\alpha}_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \tilde{\alpha}_n \quad \text{and } \left\| r_{\mathcal{G}(\alpha)}(n) \right\|_1 = o(\gamma_n),$$

we obtain from (A.35) that

$$\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha}) \Delta\alpha_n^0 + \frac{1}{2}(\delta\alpha_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \delta\alpha_n \le o(|\Delta\alpha_n^0|^2) + r_{\mathcal{G}(\alpha)}(n) \quad \text{ a.e. on } M_0.$$

Due to Assumption A.4.1 there is a sequence $\{\tilde{\alpha}_{\mathcal{G}n}\ \tilde{u}_{\mathcal{G}n}\}$ such that

$$\begin{aligned}&\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha})(\Delta\alpha_n^0 + \tilde{\alpha}_{\mathcal{G}n}) + \frac{1}{2}(\delta\alpha_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \delta\alpha_n \le 0,\ |\tilde{\alpha}_{\mathcal{G}n}| \le o(\left|\Delta\alpha_n^0\right|^2) + c\left|r_{\mathcal{G}(\alpha)}(n)\right|, \\ &\frac{\partial \mathcal{G}}{\partial u}(\cdot, \hat{u})(\Delta u_n^0 + \tilde{u}_{\mathcal{G}n}) + \frac{1}{2}(\delta u_n)^T \frac{\partial^2 \mathcal{G}}{\partial u^2}(\cdot, \hat{u}) \delta u_n \le 0,\ |\tilde{u}_{\mathcal{G}n}| \le o(\left|\Delta u_n^0\right|^2) + c\left|r_{\mathcal{G}(u)}(n)\right|,\end{aligned}$$

with some $c > 0$. Set $\delta v_n(\alpha) = \Delta\alpha_n^0 + \tilde{\alpha}_{\mathcal{G}n}$, and $\delta v_n(u) = \Delta u_n^0 + \tilde{u}_{\mathcal{G}n}$. Then

$$\begin{aligned}&\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{\alpha}) \delta v_n(\alpha) + \frac{1}{2}(\delta\alpha_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{\alpha}) \delta\alpha_n \le 0, \qquad |\tilde{\alpha}_{\mathcal{G}n}| = o(\gamma_n), \\ &\frac{\partial \mathcal{G}}{\partial \alpha}(\cdot, \hat{u}) \delta v_n(u) + \frac{1}{2}(\delta u_n)^T \frac{\partial^2 \mathcal{G}}{\partial \alpha^2}(\cdot, \hat{u}) \delta u_n \le 0, \quad \|\tilde{u}_{\mathcal{G}n}\|_1 = o(\gamma_n), \quad \|\tilde{u}_{\mathcal{G}n}\|_\infty = o(1).\end{aligned}$$

Consequently,

$$\begin{aligned}\int_0^T \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \delta v_n(\alpha) dt &= \int_0^T \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \Delta\alpha_n^0 dt + o(\gamma_n), \\ \int_0^T \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \delta v_n(u) dt &= \int_0^T \frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \Delta u_n^0 dt + o(\gamma_n).\end{aligned} \tag{A.60}$$

Step 13 Conditions (A.38), (A.58), (A.59), and (A.60) imply

$$\omega(\delta y_n) + \int_0^T \left(\frac{\partial \mathcal{H}}{\partial u}(\hat{y}, \hat{\lambda}) \delta v_n(u) + \frac{\partial \mathcal{H}}{\partial \alpha}(\hat{y}, \hat{\lambda}) \delta v_n(\alpha) \right) dt \le o(\gamma(\delta y_n)). \tag{A.61}$$

Since $\delta y_n \in K$, and $\delta v_n(u) \in T_U^{b(2)}(\hat{u}, \delta u_n)$, $\delta v_n(\alpha) \in T_A^{b(2)}(\hat{\alpha}, \delta\alpha)$, condition (A.61) contradicts Assumption A.4.4 in the form (A.23). The proof is finished. □

A.5 Sliding Regime for OCP

An Approximation for Sliding Regimes to Switches

This subsection proposes an idea of "How to construct an approximation for 'sliding regimes' to 'switches' in general?"
We discretize the non-integer control values, i.e., the controls on the interval (t_l^*, t_u^*) with appropriate time's gird, then we approximate these discretized values by an appropriate rounding scheme (see Subsection 2.6.7) to construct the output integer control.
We start with a discretization of the controls α_1^* and α_2^* on the interval (t_l^*, t_u^*) by defining a shooting grid

$$t_l^* = t_0 < t_1 < \ldots < t_{s-1} < t_s = t_u^*.$$

On each interval $[t_i, t_{i+1}]$, $i = 0, 1, 2, \ldots, s-1$, of the shooting grid we make known control parameters $\tilde{q}_i^{\alpha_1}$, $q_i^{\alpha_1}$ and $\tilde{q}_i^{\alpha_2}$, $q_i^{\alpha_2}$, where $\alpha_{j,i}^* = \tilde{q}_i^{\alpha_j}$, $j = 1, 2$, and the switched controls $w_{j,i}^* = q_i^{\alpha_j}$, $j = 1, 2$. We then take into account the rounding strategy SUR (see Eq. (2.67)), with $j = 1, 2$,

$$q_i^{\alpha_j} = \begin{cases} 1 & \text{if } \sum\limits_{k=0}^{i} \tilde{q}_k^{\alpha_j} - \sum\limits_{k=0}^{i-1} q_k^{\alpha_j} \geq 1, \\ 0 & \text{else.} \end{cases}$$

Finally, we end the construction for "sliding regimes" to "switches" by collecting the values of the rounding ones

$$w_1^* = \{q_i^{\alpha_1}\}_{i=0}^{s-1}, \qquad w_2^* = \{q_i^{\alpha_2}\}_{i=0}^{s-1}.$$

Sliding Regime for OCP

We consider OCP under basic form as (A.62) without the integer-values control function, i.e.,

$$\begin{array}{ll} \min\limits_{x(\cdot),u(\cdot)} & \varphi\left(x(t), u(t)\right) \\ \text{s.t.} & \dot{x}(t) = f\left(x(t), u(t)\right) \\ & 0 \leq r\left(x(t_0), x(t_f)\right) \\ & u(t) \in \mathcal{U}, \end{array} \quad t \in \mathcal{T}. \tag{A.62}$$

The optimal sliding regime is characterized by the non-uniqueness of the maximum with respect to $u(t)$ of the Hamilton function

$$\mathcal{H}(x, \psi, u) = \psi^T f(x, u),$$

where ψ are adjoint variables. Under these conditions (in (A.62)), on the section $[\tau_s, \tau_{s+1}] \in [t_0, t_f]$, with $s = 0, 1, \ldots, k$, of the $(k+1)$-slide $(k > 1)$ with the maxima $u_0, \ldots, u_k$, (A.62)

splits and takes the form

$$\begin{aligned}
&\min_{x(\cdot),u(\cdot)} \quad \sum_{s=0}^{k} \alpha_s \varphi\left(x(t), u_s(t)\right) \\
&\text{s.t. } \dot{x}(t) = \sum_{s=0}^{k} \alpha_s f(x, u_s), \\
&\qquad 0 \le r\left(x(t_0), x(t_f)\right), \\
&\qquad t \in [\tau_s, \tau_{s+1}] \in \mathcal{T},\ s = 0, 1, \ldots, k, \\
&\qquad u_s(t) \in \mathcal{U}_m \overset{\text{def}}{=} \{u_0, \ldots, u_k\},\ s = 0, 1, \ldots, k, \\
&\qquad \sum_{s=0}^{k} \alpha_s = 1, \alpha_s \ge 0, s = 0, 1, \ldots, k.
\end{aligned} \tag{A.63}$$

The HAMILTON function for (A.63) as follows:

$$\mathcal{H}(x, \psi, \alpha, u) = \psi^T \left(\sum_{s=0}^{k} \alpha_s f(x(t), u_s) \right),$$

after excluding α_0 and regrouping the terms, the above function can be reduced to the form

$$\begin{aligned}
\mathcal{H}(x, \psi, \alpha, u) &= \psi^T \sum_{s=1}^{k} \left[f(x, u_0) - f(x, u_s) \right] \alpha_s \\
&= \sum_{s=1}^{k} \left(\mathcal{H}(x, \psi, u_0) - \mathcal{H}(x, \psi, u_s) \right) \alpha_s.
\end{aligned}$$

Since $\mathcal{H}(x, \psi, u_s) = \max\limits_{u \in \mathcal{U}_m} \mathcal{H}$, for $s = 0, 1, \ldots, k$, on the section $[\tau_0, \tau_s]$ the optimal sliding regime with $(k+1)$ maximum the coefficients at the $(k+1)$ independent linear controls $\alpha_0, \alpha_1, \ldots, \alpha_k$ of the HAMILTON function of the split problem (A.62) are equal to zero. An optimal sliding regime with a "slide" through $(k+1)$ maxima is an optimal singular regime with $(k+1)$ components for the split problem (A.63). The maximum possible value of $(k+1)$ advisable to take when researching sliding regimes is defined by the convexity's condition of the set values of the right hand side vector and the convexity from below of the greatest lower bound of the set values of the integrand of the split system obtained when the control vector (α_s, u_s), $s = 0, \ldots, k$, runs through the whole admissible domain of values.
Here, we give a definition of ψ, which are determined by an adjoint system of equations as follows:

$$\psi = -\sum_{s=0}^{k} \alpha_s \frac{\partial f(x, u_s)}{\partial x}.$$

Example A.1.

$$\min_{x,u} \quad \int_0^3 \left(x^2 - u^2\right) dt \tag{A.64}$$

$$\text{s.t. } \dot{x}(t) = u(t), \tag{A.65}$$

$$x(0) = 1,\ x(3) = 1, \tag{A.66}$$

$$|u(t)| \le 1. \tag{A.67}$$

To obtain a minimum of objective function (A.64), it is desirable for every t to have $|x(t)|$ as small as possible, and $|u(t)|$ as large as possible. From (A.65-A.67), an "ideal" trajectory is found

$$x(t) = \begin{cases} 1-t, & 0 \le t \le 1, \\ 0, & 1 < t < 2, \\ t-2, & 2 \le t \le 3. \end{cases} \tag{A.68}$$

Here, the boundary values of a control

$$u(t) = +1 \text{ or } u(t) = -1, \tag{A.69}$$

can obtain the absolute minimum of (A.64).
But, if $1 \le t \le 2$ then $u(t) \equiv 0$, (A.68) can not be created for any control function $u(t)$, which satisfies (A.69). However, it is possible to using control functions $u_n(t)$, where $1 < t < 2$ and $n \to \infty$, which realize more frequent switchings from 1 to -1 and vice versa

$$u_n(t) = \begin{cases} -1, & 0 \le t \le 1, \\ +1, & 1 + \frac{k}{n} < t \le 1 + \frac{2k+1}{2n},\ k = 0, \dots, n-1, \\ -1, & 1 + \frac{2k+1}{2n} < t \le 1 + \frac{k+1}{n}\ k = 0, \dots, n-1, \\ +1, & 2 < t \le 3 \end{cases} \tag{A.70}$$

where $n = 1, 2, \dots$, to create a minimizing sequence of controls $\{u_n(t)\}$ which satisfies (A.69) and a minimizing sequence of trajectories $\{x_n(t)\}$ converging towards the (A.68).
Each trajectory $x_n(t)$ differs from (A.68) only on the interval $(1, 2)$ on which, instead of being a precise path along the x-axis, it makes a "saw-toothed" path with n identical "teeth", positioned above the x-axis. The "teeth of the saw" become ever finer when $n \to \infty$, such that $\lim_{n\to\infty} x_n(t) = 0$, $1 < t < 2$. In this way, the minimizing sequence of trajectories $\{x_n(t)\}$ converges towards (A.68), but the minimizing sequence of control $\{u_n(t)\}$, when $n \to \infty$ and $1 < t < 2$, which realizes ever more frequent switchings from 1 to -1 and vice versa, does not have a limit in the class of measurable functions. This means that on $(1, 2)$, an optimal sliding regime occurs.
Using heuristic reasoning, it is possible to describe the obtained optimal sliding regime in the following way: *An optimal control at each point of the interval* $(1, 2)$ *"slide", i.e., skips from the value* $+1$ *to* -1 *and back, such that, for any interval of time, nevertheless small, the measure of the set of points* t *in which* $u = +1$ *is equal to the measure of the set of points* t *in which* $u = -1$*, which, by virtue of* (A.65)*, ensures a precise motion along the* x*-axis. The description given above of the character of change of an optimal control on part of a sliding regime is non-rigorous, for it does not satisfy the ordinary definition of a function.*

It is possible to give a rigorous definition of an optimal sliding regime if, along with the initial problem (A.64-A.67), an auxiliary "split"problem is introduced: To find a minimum of the functional

$$J(x, \alpha, u) = \int_0^3 \left(x^2 - \alpha_0 u_0^2 - \alpha_1 u_1^2\right) dt, \tag{A.71}$$

with the constraints

$$\dot{x} = \alpha_0 u_0 + \alpha_1 u_1, \tag{A.72}$$

$$x(0) = 1, \quad x(3) = 1, \tag{A.73}$$

$$|u_0| \le 1,\ |u_1| \le 1, \quad \alpha_0 + \alpha_1 = 1,\ \alpha_0, \alpha_1 \ge 0. \tag{A.74}$$

The split problem (A.71-A.74) differs from the initial one, i.e., (A.64-A.67), in that, instead of one control function $u(t)$, two independent control functions $u_0(t)$ and $u_1(t)$ are employed; the integrand and the function of the right-hand side of (A.65) of the initial problem are replaced by a linear convex combination of corresponding functions, taken with different controls $u_0(t)$ and $u_1(t)$ and with coefficients $\alpha_0(t)$, $\alpha_1(t)$, which are also considered as control functions.

Hence, in problem (A.71-A.74) there are four controls $u_0, u_1, \alpha_0, \alpha_1$. Insofar as α_0 and α_1 are related by the equality-type condition $\alpha_0 + \alpha_1 = 1$, it is possible to drop one of the controls α_0 or α_1 by expressing it through the other. However, for the convenience of subsequent analysis, it is better to leave both controls in an explicit form.

Unlike the initial problem, an optimal control for the split problem (A.71-A.74) exists. On a section of the optimal sliding regime of the initial problem, the optimal control of the split problem takes the form

$$\alpha_0(t) = \alpha_1(t) = \frac{1}{2}, \quad u_0(t) = -1,\ u_1(t) = +1, \qquad 1 < t < 2,$$

while on the sections of entry and exit:

$$\alpha_0(t) = 1, \quad u_0(t) = -1, \quad \alpha_1(t) = 0, \qquad u_1(t) \text{ arbitrary}, \quad 0 \le t \le 1,$$
$$\alpha_1(t) = 1, \quad u_1(t) = +1, \quad \alpha_0(t) = 0, \qquad u_0(t) \text{ arbitrary}, \quad 2 \le t \le 3.$$

On the section of an optimal sliding regime, the controls α_0 and α_1, going linearly into the right-hand side, and the integrand accept values within the admissible domain. This means that the optimal sliding regime of the initial problem (A.64-A.67) is an optimal singular regime, or optimal singular control, for the auxiliary split problem (A.71-A.74).

A.6 Linear Program in Maximum Principle

In the Maximum Principle (see Section 3.1.4), one needs to solve the linear program with variable bounds (or box constraints) as belows

$$\begin{array}{ll} \min\limits_{\alpha, \tilde{u}_j^+, \tilde{u}_j^-} \{-\sum_j^{\tilde{n}} \tilde{u}_j^+ a_j^+ - \sum_j^{\tilde{n}} \tilde{u}_j^- a_j^-\} & \\ \text{s.t.} \qquad 0 \le \alpha \le 1, \quad 0 \le \tilde{u}_j^+,\ 0 \le \tilde{u}_j^-,\ j = 1, \dots, \tilde{n}, & \\ \qquad \sum_{j=1}^{\tilde{n}} \tilde{u}_j^+ = \alpha,\ \sum_{j=1}^{\tilde{n}} \tilde{u}_j^- = 1 - \alpha, & \end{array} \tag{A.75}$$

where $a_j^+ = \lambda^T(t) f_+(x, u_j)$, $a_j^- = \lambda^T(t) f_-(x, u_j)$, $j = 1, \dots, \tilde{n}$, see problem (3.52).

The Lagrangian function of LP (A.75) is

$$L = -{a^+}^T \tilde{u}^+ - {a^-}^T \tilde{u}^- + \lambda^+ (\sum_{j=1}^{\tilde{n}} \tilde{u}_j^+ - \alpha) + \lambda^- (\sum_{j=1}^{\tilde{n}} \tilde{u}_j^- + \alpha - 1) + \mu_1 \alpha + \mu_2 (1 - \alpha) + \mu_3^T \tilde{u}^+ + \mu_4^T \tilde{u}^-.$$

The optimal solution of (A.75) are obtained by the following conditions

(i) $\sum_{j=1}^{\tilde{n}} \tilde{u}_j^+ = \alpha$, $\sum_{j=1}^{\tilde{n}} \tilde{u}_j^- = 1-\alpha$, $\alpha \in [0,1]$, $\tilde{u}_j^+ \in [0,\alpha]$, $\tilde{u}_j^- \in [0,1-\alpha]$, $j = 1,\dots,\tilde{n}$,

(ii) $\mu_1, \mu_2, \mu_{3,j}, \mu_{4,j} \geq 0$, $j = 1,\dots,\tilde{n}$,

(iii) $-\lambda^+ + \lambda^- + \mu_1 - \mu_2 = 0$,
$-a_j^+ + \lambda^+ + \mu_{3,j} = 0$, $-a_j^- + \lambda^- + \mu_{4,j} = 0$, $j = 1,\dots,\tilde{n}$,

(iv) $\mu_1 \alpha = 0$, $\mu_2(1-\alpha) = 0$,
$\mu_{3,j}\tilde{u}_j^+ = 0$, $\mu_{4,j}\tilde{u}_j^- = 0$, $j = 1,\dots,\tilde{n}$.

Conditions (iii)-(iv) help us obtain

$$\lambda^+ - \lambda^- = \mu_1 - \mu_2 \begin{cases} \leq 0 & \text{if } \alpha = 1, \mu_1 = 0, \\ \geq 0 & \text{if } \alpha = 0, \mu_2 = 0, \\ = 0 & \text{if } \alpha \in (0,1), \mu_1 = \mu_2 = 0, \end{cases} \tag{A.76}$$

$$\lambda^+ = a_j^+ - \mu_{3,j} \leq a_j^+, \forall j = 1,\dots,\tilde{n}, \tag{A.77}$$

$$\lambda^- = a_j^- - \mu_{4,j} \leq a_j^-, \forall j = 1,\dots,\tilde{n}. \tag{A.78}$$

If $\tilde{u}_j^+ > 0$ then from conditions (iv) one implies $\mu_{3,j} = 0$, $\forall j = 1,\dots,\tilde{n}$, hence from (A.77) one gets

$$\lambda^+ = a_j^+ = \lambda^T f_+(x, u_j),\ \forall j = 1,\dots,\tilde{n}.$$

If $\tilde{u}_j^+ = 0$ then from conditions (iv) one obtains $\mu_{3,j} \geq 0$, $\forall j = 1,\dots,\tilde{n}$, hence from (A.77) one implies

$$\lambda^+ = \min_{j \in \{1,\dots,\tilde{n}\}} \{a_j^+\} = \min_{j \in \{1,\dots,\tilde{n}\}} \{\lambda^T f_+(x, u_j)\}.$$

Analyzing similar for $\tilde{u}_j^-$, one gets

$$\lambda^- = \begin{cases} \lambda^T f_-(x, u_j) & \text{if } \tilde{u}_j^- > 0,\ \forall j = 1,\dots,\tilde{n}, \\ \min\limits_{j \in \{1,\dots,\tilde{n}\}} \{\lambda^T f_-(x, u_j)\} & \text{if } \tilde{u}_j^- = 0,\ \forall j = 1,\dots,\tilde{n}. \end{cases}$$

A.7 Example

Example A.2. [82, Eq. (4.1)] Consider a problem

$$\begin{aligned} \min_{x,u} \quad & F_0(x(T)) \\ \text{s.t.} \quad & \dot{x}(t) = \begin{cases} F_+(x(t), u(t)) & \text{if } r(x(t)) > 0, \\ F_-(x(t), u(t)) & \text{if } r(x(t)) < 0, \\ F_+(x(t), u(t)) \vee F_+(x(t), u(t)) & \text{if } r(x(t)) = 0, \end{cases} \\ & u(t) \in [-1,1], \quad t \in \mathcal{T} = [0,T], \quad x(0) = x_0, \quad h(x(T)) = 0, \end{aligned} \tag{A.79}$$

with the following data:

$$x \in \mathbb{R}^3, \quad F_\pm(x,u) = Ax + b^\pm u, \quad h(x) = x_1 - x_2 + 1, \quad x(0) = (2,1,0), \quad T = 4,$$

$$d = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad b^+ = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \quad b^- = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}. \tag{A.80}$$

We can write problem (A.79) in the following form

$$\begin{array}{ll}\min\limits_{x,u} & x_3(4)\\ \text{s.t.} & \dot{x}(t)=\begin{cases}(-u,0,x_1+x_2) & \text{if } x_1(t)-x_2(t)>0,\\ (0,u,x_1+x_2) & \text{if } x_1(t)-x_2(t)<0,\\ (-u,0,x_1+x_2)\vee(0,u,x_1+x_2) & \text{if } x_1(t)-x_2(t)=0,\end{cases}\\ & u(t)\in[-1,1],\quad t\in[0,4],\quad x(0)=(2,1,0),\quad x_1(4)-x_2(4)=-1.\end{array}\tag{A.81}$$

By using the Filippov's rule, in terms of differential inclusions, (A.81) is stated as follows:

$$\begin{array}{ll}\min\limits_{x} & x_3(4)\\ \text{s.t.} & \dot{x}(t)\in U(x(t)),\quad t\in[0,4],\\ & x(0)=(2,1,0),\quad x_1(4)-x_2(4)=-1.\end{array}\tag{A.82}$$

where the mapping $U(x)$, $x\in\mathbb{R}^3$, is defined by the relations

$$\begin{aligned}&U(x):=\{v\in\mathbb{R}^3: v=(-u,0,x_1+x_2), u\in[-1,1]\} \text{ if } x_1(t)-x_2(t)>0,\\ &U(x):=\{v\in\mathbb{R}^3: v=(0,u,x_1+x_2), u\in[-1,1]\} \text{ if } x_1(t)-x_2(t)<0,\\ &U(x):=\{v\in\mathbb{R}^3: v=\alpha(-u,0,x_1+x_2)+(1-\alpha)(0,u,x_1+x_2), u\in[-1,1]\}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } x_1(t)-x_2(t)=0.\end{aligned}\tag{A.83}$$

Next, since the function $r(x)$ is linear and $U(x)$ is defined as in (A.83), we can deduce that

$$\operatorname{conv} U(x)=\{v\in\mathbb{R}^3: v=f(x,\alpha,u_+,u_-),\alpha\in[0,1],|u_+|\le\alpha,|u_-|\le 1-\alpha\},$$

where

$$\begin{aligned}f(x,\alpha,u_+,u_-)&:=\alpha a^+(x)+(1-\alpha)a^-(x)+u_+b^+(x)+u_-b^-(x)\\ &=\Delta a(x)\alpha+a^-(x)-\begin{pmatrix}u_+ & 0 & 0\end{pmatrix}^T+\begin{pmatrix}0 & u_- & 0\end{pmatrix}^T,\end{aligned}\tag{A.84}$$

with $\Delta a(x):=a^+(x)-a^-(x)$.
The weakened problem is

$$\begin{array}{ll}\min\limits_{\alpha,u_+,u_-} & x_3(4)\\ \text{s.t.} & \dot{x}(t)=\begin{pmatrix}-u_+(t)\\ u_-(t)\\ x_1(t)+x_2(t)\end{pmatrix},\qquad t\in[0,4].\\ & x(0)=(2,1,0),\quad x_1(4)-x_2(4)=-1,\\ & |u_+(t)|\le\alpha(t),\quad |u_-(t)|\le 1-\alpha(t),\quad \alpha(t)\in[0,1],\\ & \alpha(t)\,(x_1(t)-x_2(t))\ge 0,\quad (1-\alpha(t))\,(x_1(t)-x_2(t))\le 0,\end{array}\tag{A.85}$$

By using the necessary nondegenerate optimality conditions stated in [82, Thm. 2] with the following data:

$$\begin{aligned}&y_0=1,\quad y=2,\quad \gamma^1=2,\quad \gamma_1=0,\quad S^0(t)=A,\quad t\in[0,4],\\ &\psi(t)=(t-4,t-4,-1),\quad t\in[0,3],\quad \psi(t)=(t-4-y,t-4+y,-1),\quad t\in[3,4],\end{aligned}$$

one can obtain the optimal control and the trajectory in the form

$$\begin{aligned}
&\alpha^0(t) = 1, \quad u_+^0(t) = 1, \quad u_-^0(t) = 0, \quad t \in [0,1),\\
&\alpha^0(t) = 0.5, \quad u_+^0(t) = 0.5, \quad u_-^0(t) = -0.5, \quad t \in [1,3),\\
&\alpha^0(t) = 0, \quad u_+^0(t) = 0, \quad u_-^0(t) = 1, \quad t \in [3,4],\\
&x_1^0(t) = -t + 2, \quad x_2^0(t) = 1, \quad t \in [0,1],\\
&x_1^0(t) = -t/2 + 1.5, \quad x_2^0(t) = -t/2 + 1.5, \quad t \in [1,3]\\
&x_1^0(t) = 0, \quad x_2^0(t) = t - 3, \quad t \in [3,4],\\
&x_3^0(t) = \int_0^t (x_1^0(\tau) + x_2^0(\tau))d\tau, \quad \tau \in [0,4],
\end{aligned} \tag{A.86}$$

and the optimal value of the objective function is $F_0(x^0(4)) = x_3^0(4) = 5$.
Now we will use the above optimal solution to construct a control in the original problem. By construction, $|u_+^0| \le 1$, $u_-^0 = 0$ if $t \in \bar{\mathcal{T}}_1 = \{t \in \mathcal{T} : \alpha^0(t) = 1\} = [0,1)$, $u_+^0 = 0$, $|u_-^0| \le 1$ if $t \in \bar{\mathcal{T}}_0 = \{t \in \mathcal{T} : \alpha^0(t) = 0\} = [3,4]$, and $|u_+^0| \le \alpha^0(t)$, $|u_-^0| \le 1 - \alpha^0(t)$ if $t \in \bar{\mathcal{T}}_* = \mathcal{T} \setminus (\bar{\mathcal{T}}_0 \cup \bar{\mathcal{T}}_1) = [1,3)$. We set

$$\begin{aligned}
&u_1(t) = u_+^0(t) = 1, \quad u_2(t) = 0, \quad t \in \bar{\mathcal{T}}_1 = [0,1),\\
&u_1(t) = 0, \quad u_2(t) = u_-^0(t) = 1, \quad t \in \bar{\mathcal{T}}_0 = [3,4],\\
&u_1(t) = \frac{u_+^0(t)}{\alpha^0(t)} = 1, \quad u_2(t) = \frac{u_-^0(t)}{1 - \alpha^0(t)} = -1, \quad t \in \bar{\mathcal{T}}_* = [1,3).
\end{aligned}$$

We get optimal control in relaxed problem. However this is not feasible in the original one. Exploiting the procedure as in Section 3.1.4 (cf. [82, Section 3]), one can construct an approximate solution to problem (A.79).

Appendix B

Some Open Problems

This appendix reprints some ideas and solution approaches for tracking switches in OCP.

B.1 An Idea about Over-Under Estimating

In this section we propose a concept about over-under estimating, in order to have a better approximation for switches.
Consider an appropriate time grid

$$G_m := \{t_0 < t_1 \ldots < t_m\},$$

and suppose that the controls can only switch in the discretization points t_i, $i = 0, 1, \ldots, m$. The basic control functions on this grid would be

$$b_j^0 := \begin{cases} 1, & \text{if } t \in [t_j, t_{j+1}) \\ 0, & \text{else} \end{cases} \qquad 0 \leq j \leq m-1,$$

with the binary controls then being $\omega_i := \sum_{j=0}^{m-1} b_j^0(t) q_{i,j}$, $1 \leq i \leq n_\omega$, where q is the solution of RC.SwP.
A switch in control ω_i at time step t_j is captured by the term

$$\sigma_{i,j} := |q_{i,j} - q_{i,j-1}|, \tag{B.1}$$

and the total number of switches could be limited by $\sigma_{\max}$ as $\sigma_{\max} \geq \frac{1}{2} \sum_{i=1}^{n_\omega} \sum_{j=1}^{m} \sigma_{i,j}$.
Based on the definition of the absolute value, from (B.1) we could use the tightest underestimating hyperplanes

$$\begin{aligned} \sigma_{i,j} &\geq q_{i,j} - q_{i,j-1}, \\ \sigma_{i,j} &\geq q_{i,j-1} - q_{i,j}. \end{aligned} \tag{B.2}$$

The biggest disadvantage of (B.2) is that it yields lots of switches when $q_{i,j} \simeq q_{i,j-1}$. Therefore, Kirches [76] proposed to use the tightest overestimating hyperplanes

$$\begin{aligned} \sigma_{i,j} &= q_{i,j} + q_{i,j-1}, \\ \sigma_{i,j} &= 2 - q_{i,j} - q_{i,j-1}, \end{aligned} \tag{B.3}$$

which are based on the inequalities

$$|q_{i,j} - q_{i,j-1}| \leq |q_{i,j} + q_{i,j-1}|\,,$$
$$|q_{i,j} - q_{i,j-1}| = 2\max\{q_{i,j}, q_{i,j-1}\} - |q_{i,j} + q_{i,j-1}| \leq 2 - (q_{i,j} + q_{i,j-1}).$$

Instead of using (B.2) or (B.3), we propose *over-under* estimating as follows

$$\begin{aligned}\sigma_{i,j} &= \sqrt{q_{i,j}^2 + q_{i,j-1}^2},\\ \sigma_{i,j} &= \frac{(q_{i,j} - q_{i,j-1})^2}{\sqrt{q_{i,j}^2 + q_{i,j-1}^2}},\end{aligned} \tag{B.4}$$

which are based on the inequalities $\frac{|A-B|^2}{\sqrt{|A-B|^2+C}} \leq |A-B| \leq \sqrt{|A-B|^2 + C}$, for the choices $A = q_{i,j}$, $B = q_{i,j-1}$, and $C = 2q_{i,j}q_{i,j-1}$.
Here it has to requires the switching variable to be equal to a convex combination of those

$$\sigma_{i,j} := \begin{cases} 0, & \text{if } q_{i,j}^2 + q_{i,j-1}^2 = 0,\\ \sigma_{i,j}^D, & \text{else,}\end{cases}$$

therein, $\sigma_{i,j}^D = \alpha_{i,j}\sqrt{q_{i,j}^2 + q_{i,j-1}^2} + (1-\alpha_{i,j})\frac{(q_{i,j}-q_{i,j-1})^2}{\sqrt{q_{i,j}^2+q_{i,j-1}^2}}$, whereas,

$$\alpha_{i,j} = \begin{cases} 1, & \text{if } q_{i,j}^2 + q_{i,j-1}^2 \leq 1,\\ 0, & \text{else.}\end{cases}$$

The following table, Tab. B.1, shows a brief comparison of these three above estimations.

Table B.1: Comparison for the three over-under estimating.

		Estimating switch $\sigma_{i,j}$		
$q_{i,j}$	$q_{i,j-1}$	(B.1-B.2)	(B.3)	(B.4)
0	0	$0, 0$	$0, 2$	0, error
1/4	0	$1/4, -1/4$	$1/4, 7/4$	$1/4, 1/4$
1/4	1/4	$0, 0$	$1/2, 3/2$	$\sqrt{2}/4, 0$
1/2	0	$1/2, -1/2$	$1/2, 3/2$	$1/2, 1/2$
1/2	1/4	$1/4, -1/4$	$3/4, 5/4$	$\sqrt{5}/4, 1/4\sqrt{5}$
1/2	1/2	$0, 0$	$1, 1$	$1/\sqrt{2}, 0$
3/4	1/2	$1/4, -1/4$	$5/4, 3/4$	$\sqrt{13}/4, 1/4\sqrt{13}$
1	0	$1, -1$	$1, 1$	$1, 1$
1	1	$0, 0$	$2, 0$	$\sqrt{2}, 0$
0	1	$-1, 1$	$1, 1$	$1, 1$

Remark 40. The results in Tab. B.1 show that, except the case where $q_{i,j}$ and $q_{i,j-1}$ are both equal to zero, our "over-under" (B.4) is the better estimator when compared with other ones, i.e., (B.1-B.2) and (B.3). Therefore, the best choice here is to deal with (B.3) for the zero case, while the remaining cases are applied with (B.4).

B.2 Gröbner Basic Approach

In this section, we will investigate another approach in optimal control, namely *Gröbner basis*, which is algebraic in nature. This approach was use to determine the switching surfaces, e.g. see [125], or to synthesize a feedback control based switching law that nearly produces time-optimal switching, see [101, 102]. Recently, the Gröbner basis is also employed to detect the biological switches, cf. [5].
The main idea of Gröbner basis approach is to test directly if a particular switching strategy is feasible. We will go through this heuristic via two following instances in Subsection B.2.2, while the general heuristic approach is proposed in Subsection B.2.1.

B.2.1 General Heuristic Approach

This section describes the core concept of the Gröbner basis approach through the input, the general heuristic approach, and the output.
Input: A OCP in the following form

$$\begin{array}{ll} \min\limits_{x(.),u(.)} & x(t_f) \\ \text{s.t.} & \dot{x}(t) = f(x(t),u(t)),\ t \in [t_0,t_f] =: \mathcal{T}, \\ & r(x(t_0),x(t_f)) \leq 0, \\ & |u(t)| \leq 1,\ t \in \mathcal{T}. \end{array} \tag{B.5}$$

General Heuristic Approach

(i) Using PMP to obtain optimal control $u^*(t) = u^*(\lambda(t))$, optimal state trajectory $x^*(t)$, and optimal co-state $\lambda^*(t)$.

(ii) Determining a maximum number of switches, assumed as d, which is directly depended on the degree of $\lambda^*(t)$.

(iii) Designing $t_1, \ldots, t_d$ the length of the successive intervals where $u^*(t)$ stays constant. The particular choice (among the only two possible ones)

$$u^*(t) = \begin{cases} -1, & \text{for } t_0 \leq t < t_1, \\ +1, & \text{for } t_1 \leq t < t_1 + t_2, \\ \vdots & \\ (-1)^d, & \text{for } t_1 + \ldots + t_{d-1} \leq t < t_1 + \ldots + t_{d-1} + t_d =: t_f. \end{cases}$$

(while the remain is $+1, -1, +1, \ldots$)

(iv) Calculating $x^*(t_f) = (x_1^*(t_f), \ldots, x_n^*(t_f))$ w.r.t. $u^*(t)$, which is described in step (iii).

(v) Computing suitable Gröbner bases (one possibility is using Sturm sequences)

(v.i) Setting

$$\begin{aligned} &z_1 := t_1,\ z_2 := t_2,\ \ldots\ z_d := t_d, \\ &a_1 := x_1(t_f),\ a_2 := x_2(t_f),\ \ldots,\ a_n := x_n(t_f). \end{aligned}$$

(v.ii) Solving the complex version of the switching problem

$$a_j = x_j(z_1, \ldots, z_d), \quad \text{for } j = 1, \ldots, n. \tag{B.6}$$

by using the `Macaulay` symbolic program [61]. The system (B.6) has a complex solutions, or real solutions.

(v.iii) Considering the special case there (B.6) has real nonnegative solutions. Using Sturm sequences to compute suitable Gröbner bases together with the following algorithm
Input of Alg. 5: Current state $(a_1, \ldots, a_n)$.

Algorithm 5. ***The Switching Algorithm****(with Gröbner basis)*
Case 1 (Check whether $z_1 = 0$) Testing the consistency of the system (B.6) within $z_1 = 0$. If consistent solve it, if $z_j \geq 0$, $j = 2, \ldots, d$ set $u = -1$, otherwise set $u = 1$. If the system (B.6, $z_1 = 0$) is not consistent, go to the next case.
Case 2 (Check whether $z_d = 0$) Testing the consistency of the system (B.6) within $z_d = 0$. If consistent solve it, if $z_j \geq 0$, $j = 1, \ldots, d-1$ set $u = -1$, otherwise set $u = 1$. If the system (B.6, $z_d = 0$) is not consistent, go to the next case.
Case 3 $cd(a_1, \ldots, a_n) = 0$ (Check whether $z_j = 0$, $j = 2, \ldots, d-1$) If $a_n < 0$ let $u = -1$, otherwise let $u = +1$.
Case 4 $cd(a_1, \ldots, a_n) \neq 0$ Computing the Sturm sequences, and obtain the number of solutions. Depending on this number, set $u = 1$ or $u = -1$.

where $cd(\cdot)$ is the condition comes from system (B.6) with some special cases of z_j, $j = 1, \ldots, d$.
Output of Alg. 5: Value of u, either 1 or -1.

Output: Switching strategy, $+1, -1, +1, \ldots$, or $-1, +1, -1, \ldots$.

Remark 41. An example to visualize the General Heuristic Approach can be seen in Example B.1. On the other hand, in Example B.2, we can apply this heuristic to obtain switching strategy between 0 and 1, when the control $0 \leq u(t) \leq 1$ instead of $-1 \leq u(t) \leq 1$.

B.2.2 Numerical Examples

This subsection consists two following instances to illustrating the general heuristic approach in Section B.2.1.

Example B.1. Consider the classical time-optimal control problem for a system consisting of a chain of integrators, cf. [125], where the linear system with saturated control input, and the objective drives the system from an initial condition $x(0)$ to a target $x(t_f)$ in minimum time t_f

$$\begin{array}{ll} \min\limits_{x(\cdot),u(\cdot)} & \int_0^{t_f} 1dt \\ \text{s.t.} & \dot{x}_1(t) = x_2(t), \\ & \dot{x}_2(t) = x_3(t), \\ & \dot{x}_3(t) = u(t), \quad |u(t)| \leq 1. \end{array} \tag{B.7}$$

We drop the superscript * to simplify the notation. The HAMILTON function of (B.7) is

$$\mathcal{H}(x,\lambda,u) = 1 + \lambda_1 x_2 + \lambda_2 x_3 + \lambda_3 u.$$

The maximality condition gets

$$\begin{aligned}\mathcal{H}(x(t),\lambda(t),u(t)) &= \min_{-1\le u\le 1}\{1 + \lambda_1(t)x_2(t) + \lambda_2(t)x_3(t) + \lambda_3(t)u\}\\ &= 1 + \lambda_1(t)x_2(t) + \lambda_2(t)x_3(t) - \lambda_3(t)\operatorname{sgn}(\lambda_3(t))\end{aligned}$$

while the optimal control is given by

$$u(t) = -\operatorname{sgn}(\lambda_3(t)) = \begin{cases}1 & \text{if } \lambda_3(t) < 0\\ -1 & \text{if } \lambda_3(t) > 0.\end{cases}$$

Hence the optimal state trajectories are

$$x_1(t) = \begin{cases}\frac{t^3}{6} + \frac{x_3(0)t^2}{2} + x_2(0)t + x_1(0) & \text{if } \lambda_3(t) < 0\\ -\frac{t^3}{6} + \frac{x_3(0)t^2}{2} + x_2(0)t + x_1(0) & \text{if } \lambda_3(t) > 0,\end{cases}$$

$$x_2(t) = \begin{cases}\frac{t^2}{2} + x_3(0)t + x_2(0) & \text{if } \lambda_3(t) < 0\\ -\frac{t^2}{2} + x_3(0)t + x_2(0) & \text{if } \lambda_3(t) > 0,\end{cases}$$

$$x_3(t) = \begin{cases}t + x_3(0) & \text{if } \lambda_3(t) < 0\\ -t + x_3(0) & \text{if } \lambda_3(t) > 0.\end{cases}$$

The adjoint equations are

$$\dot\lambda_1(t) = 0, \quad \dot\lambda_2(t) = -\lambda_1(t), \quad \dot\lambda_3(t) = -\lambda_2(t)$$

which together with the terminal condition $\lambda(t_f) = 0$ imply the optimal co-states

$$\lambda_1(t) = t - t_f,\ \lambda_2(t) = -\frac{t^2}{2} + t_f t - \frac{t_f^2}{2},\ \lambda_3(t) = \frac{t^3}{6} - t_f\frac{t^2}{2} + t_f^2\frac{t}{2} - \frac{t_f^3}{6}. \tag{B.8}$$

We consider a Gröbner basis approach. In this example, there are only two possible strategies where the input alternates between -1 and $+1$, taking the values $-1,+1,-1,\ldots$, or $+1,-1,+1,\ldots$, respectively. In each case, taking into account the maximal number of switching, one can easily derive an expression for the final value of the state, $x(t_f)$, as a function of the switching times.
From (B.8), it is well known that there are no singular intervals and that the control input switches at most three times. Designate by t_1, t_2 and t_3 the length of the successive intervals where $u(t)$ stays constant. Any set of initial and final conditions can be translated to having $x(0) = 0$ and a given value for $x(t_f)$, and this is the setting from here on. The particular choice (among the only two possible ones)

$$u(t) = \begin{cases}-1, & \text{for } 0 \le t < t_1\\ +1, & \text{for } t_1 \le t < t_1 + t_2\\ -1, & \text{for } t_1 + t_2 \le t < t_1 + t_2 + t_3 =: t_f\end{cases}$$

drives the chain of integrators for the origin to the final point $x(t_f)$ given by

$$\begin{aligned}
x_3(t_f) &= -t_1 + t_2 - t_3 \\
x_2(t_f) &= -\frac{t_1^2}{2} - t_1 t_2 + \frac{t_2^2}{2} + t_2 t_3 - \frac{t_3^2}{2} - t_3 t_1 \\
x_1(t_f) &= -\frac{t_1^3}{6} + \frac{t_2^3}{6} - \frac{t_3^3}{6} - \frac{t_1^2}{2} t_2 - \frac{t_1^2}{2} t_3 - \frac{t_2^2}{2} t_1 + \frac{t_2^2}{2} t_3 - \frac{t_3^2}{2} t_1 + \frac{t_3^2}{2} t_2 - t_1 t_2 t_3.
\end{aligned} \tag{B.9}$$

It turns out that the selection between alternating values $-1, +1, -1, \ldots$, or $+1, -1, +1, \ldots$ for the optimal input $u(t)$ depends on whether the equations in (B.9) have a solution for a specified final condition $x(t_f) = (x_1, x_2, x_3)^T$. We refer to the computational algebraic geometry and Gröbner bases, cf. [125, Sec. III], for the following calculation. We set

$$x := t_1, \quad y := t_2, \quad z := t_3, \quad a := x_3(t_f), \quad b := x_2(t_f), \quad c := x_1(t_f).$$

Then we solve the complex version of the switching problem, namely, the following.

Problem 1: Given is the system of equations

$$\begin{aligned}
a &= y - x - z \\
b &= \frac{y^2}{2} + yz - \frac{x^2}{2} - xy - \frac{z^2}{2} - zx \\
c &= \frac{y^3}{6} + \frac{y^2 z}{2} + \frac{z^2 y}{2} - \frac{x^3}{6} - \frac{z^3}{6} - \frac{x^2 y}{2} - \frac{x^2 z}{2} - \frac{y^2 x}{2} - \frac{z^2 x}{2}.
\end{aligned} \tag{B.10}$$

We use the Macaulay symbolic program to compute the complex solutions x, y, z of the above system. By a similar way as in [125, Subsec. A. Complex Solutions], we conclude that

i. the system does always have a complex solution;

ii. if $a^2 = -2b$, $a^3 = 6c$ and $a, b, c \in \mathbb{R}$, the system has real solutions;

iii. if $a^2 = -2b$, $a^3 = 6c$ and $0 \leq a, b, c \in \mathbb{R}$, the system has real nonnegative solutions.

We are interested in the case iii. Thus, as a second step, we will investigate the following.

Problem 2: Given are $a, b, c \in \mathbb{R}$. Does there exist a nonnegative solution vector (x, y, z) for the system (B.10) in the sense that $x \geq 0$, $y \geq 0$, $z \geq 0$? If there is a positive solution x, y, z, then the value of the optimal control u assumes the values $-1, +1, -1$ successively, and in particular, the present value for the optimal control is $u(0) = -1$. If no positive solution exists then the present value of the optimal control is $u(0) = +1$.

We will use *Sturm sequences* to compute suitable Gröbner bases together with an algorithm from real algebraic geometry. Sturm sequences are associated to polynomials as follows. Suppose $f(x)$ is a single variable polynomial with real coefficients. We define $p_0(x) = f(x)$, $p_1(x) = f'(x)$, and then recursively p_i by $p_i = q_{i-1} p_{i-1} - p_{i-2}$ for $i > 1$, where q_{i-1} represents the quotient and p_{i-2} the respective remainder each time. Therein, we demand that $deg(p_i) < deg(p_{i-1})$. So, p_i is up to sign the remainder of Euclidean division of p_{i-2} by p_{i-1}.

As a first step, we compute a Gröbner basis for the three polynomials in (B.10) under an elimination order with $x > y > z > c > b > a$. Note that switch of the variables y and z in

the ordering. One gets

$$0 = y^4 - 4y^2b - 2y^2a^2 + 4yc + 4yba + 4/3ya^3 - b^2 - ba^2 - 1/4a^4 \tag{B.11a}$$
$$0 = zb + 1/2za^2 - 1/2y^3 + 3/2yb + 3/4ya^2 - 2c - ba - 1/6a^3 \tag{B.11b}$$
$$0 = zy - 1/2y^2 + ya - 1/2b - 1/4a^2 \tag{B.11c}$$
$$0 = x + z - y + a. \tag{B.11d}$$

We next solve (B.11b) or (B.11c) for z

$$z = \frac{1/2y^3 - 3/2yb - 3/4ya^2 + 2c + ba + 1/6a^3}{b + 1/2a^2} \tag{B.12a}$$
$$z = \frac{y^2/2 - ya + 1/2b + 1/4a^2}{y} \tag{B.12b}$$

respectively. Herein, of course, assuming that $b + a^2/2$ and y are not zero.
One sees that $y = 0$ implies $2b + a^2 = 6c - a^3 = 0$. These relations simplify the system to

$$\begin{aligned} &0 = b + 1/2a^2, && 0 = c - 1/6a^3, && 0 = y^3, \\ &0 = zy - 1/2y^2 + ya, && 0 = y + z - y + a. \end{aligned} \tag{B.13}$$

This has the solutions $y = 0$, z arbitrary, $x = -a - z$. Since $y = 0$ is actually equivalent to $a^3 - 6c = a^2 + b = 0$, testing the latter conditions is sufficient to find out whether $y = 0$. In that case, nonnegative solutions will exist precisely when a is negative. This covers the case $y = 0$.
If $x = 0$, the system takes the form

$$\begin{aligned} 0 &= c^2 + 2cba + 2/3ca^3 - b^3 - 1/2b^2a^2 - 1/12ba^4 - 1/72a^6s \\ 0 &= yb + 1/2ya^2 - c - ba - 1/3a^3 \\ 0 &= yc - 1/6ya^3 + ca - b^2 + 1/12a^4 \\ 0 &= y^2 - b - 1/2a^2 \\ 0 &= z - y + a. \end{aligned} \tag{B.14}$$

Since a, b, c are known, it is not so difficult to check the consistency of this system, by solving each of three middle equations for y and testing the vanishing of the first. If consistency fails, we are not in the case $x = 0$. If the system is consistent, one needs to check whether the obtained solutions for y, z are nonnegative. If that is so, set $u = -1$ and otherwise $u = +1$, finishing the case $x = 0$.
In a similar fashion, one does get the case $z = 0$. If $z = 0$, one gets

$$\begin{aligned} 0 &= c^2 - 2cba + 2/3ca^3 + b^3 - 1/2b^2a^2 + 1/12ba^4 - 1/72a^6 \\ 0 &= yb - 1/12ya^2 - c + 1/6a^3 \\ 0 &= yc - 1/6ya^3 - 2ca + b^2 + 1/12a^4 \\ 0 &= y^2 - 2ya + b + 1/2a^2 \\ 0 &= x - y + a \end{aligned} \tag{B.15}$$

which is quite similar to the case $x = 0$. One first check whether the first relation between the parameters holds. Then one solves the next three equations for y and then solves the last relation for x. If the system is consistent we have $z = 0$. If x, y turn out to be nonnegative, set $u = -1$ and otherwise $u = +1$.
This rules out all cases of vanishing variables. In order to predict when strictly positive solution exist, we are reduced to the cases ($a^2/2 = -b$, $a^3/6 \neq c$) and ($a^2/2 \neq -b$).
We consider the first case ($a^2/2 = -b$, $a^3/6 \neq c$). Then, we have a Gröbner basis

$$-b - 1/2a^2y^3 + 4c - 2/3a^3z - y^2 - ax + z - y + a.$$

It becomes obvious that in order to have a nonnegative solution, we need

$$y^3 = 4(a^3/6 - c) \geq 0,\ z = (4(a^3/6 - c)^{1/3})/2 + a \geq 0,\ x = (4(a^3/6 - c)^{1/3})/2 \geq 0,$$

which simplifies to the two conditions $a^3/6 - c \geq 0$, $(4(a^3/6 - c)^{1/3})/2 + a \geq 0$. These conditions that can easily be checked for given a, b, c and determine existence of a nonnegative solution (x, y, z) of the system (B.10).
Now, let us move to the most general situation $a^2/2 + b \neq 0$. In particular, $y \neq 0$ then. [125, Theorem 2] asserts that the Sturm sequence $\{p_i(y)\}$ corresponding to

$$f(y) = y^4 - 4y^2(b + a^2/2) + 4y(ba + c + 1/3a^3) - b^2 - ba^2 - a^4/4$$

counts the zeros of this quartic. In particular, there will be positive solutions for just y if and only if $v(0) - v(\infty) > 0$ since zero is not a root of the quartic [note that $-b^2 - ba^2 - a^4/4 = -(b + a^2/2)^2$].
Now, from (B.11c)

$$z = \frac{y^2/2 - ya + 1/2b + 1/4a^2}{y}.$$

This means that for positive y, z is positive as long as $y^2/2 - ya + 1/2b + 1/4a^2 > 0$. This parabola has roots in $r_{1,2} = a \pm \sqrt{a^2/2 - b}$ where $r_1 \leq r_2$. Since the parabola has positive leading coefficient, $y, z > 0$ for $y \notin [r_1, r_2]$ if $a^2/2 - b > 0$, and $y, z > 0$ for all $y > 0$ if $a^2/2 - b < 0$.
Similarly

$$x = y - a - z = \frac{y^2/2 - 1/2b - 1/4a^2}{y}.$$

Let $r'_{1,2} = \pm\sqrt{1/2a^2 + b}$ with $r'_1 \leq r'_2$. Hence, $x, y > 0$ if and only if $0 < y \notin [r'_1, r'_2]$ if $a^2/2 > -b$, and $x, y > 0$ for all $y > 0$ if $a^2/2 < -b$.

We conclude that to have x, y, z all positive at the same time we need to satisfy the following conditions all at the same time:

$$\begin{aligned} y^4 - 4y^2(b + a^2/2) + 4y(ba + c + 1/3a^3) - b^2 - ba^2 - a^4/4 &= 0 \\ y \notin [r_1, r_2],\ &\text{or } r_i \notin \mathbb{R} \\ y \notin [r'_1, r'_2],\ &\text{or } r'_i \notin \mathbb{R} \\ y > 0 & \end{aligned}$$

which can be checked with Sturm sequences.

These results pave the way for the following algorithm, which has as input the current state (a, b, c) of the system and as output the recommended value for u for time optimal control, either -1 or 1. The origin is then approached by an iterated repetition of the algorithm.

Algorithm: *The Switching Algorithm* ([125, Alg. 3], Dynamical Steering of the System to the Origin): Suppose our system is in the state (a, b, c).

Case 1 (Check whether $x = 0$) Test the consistency of the system (B.14). If consistent solve it; if $y, z \geq 0$ set $u = -1$, otherwise set $u = 1$. If the system (B.14) is not consistent, go to the next case.

Case 2 (Check whether $z = 0$) Test the consistency of the system (B.15). If consistent solve it; if $x, y \geq 0$ set $u = -1$, otherwise set $u = 1$. If the system (B.15) is not consistent, go to the next case.

Case 3 $a^2 = -2b, a^3 = 6c$. (Check whether $y = 0$) If $a < 0$, let $u = -1$ for a s, at which point the system will have reached the origin. If $a \geq 0$, let $u = +1$ for a s.

Case 4 $a^2 = -2b, a^3 \neq 6c$, $x \neq 0, y \neq 0, z \neq 0$. If $a^3 - 6c < 0$ and $11a^3 < 6c$, let $u = -1$. Else, let $u = +1$.

Case 5 $a^2 \neq -2b$, $x \neq 0, y \neq 0, z \neq 0$. Set $r_1 = a - \sqrt{a^2/2 - b}$, $r_2 = a + \sqrt{a^2/2 - b}$, $r_2' = \sqrt{a^2/2 + b}$. Let $f(y) = y^4 - 4y^2(b + a^2/2) + 4y(ba + c + 1/3a^3) - b^2 - ba^2 - a^4/4$ and compute the corresponding Sturm sequence $\{p_i(y)\}_{i \leq 0}$. Let $I = (0, r_1) \cup (r_2, \infty)$ if $r_i \in \mathbb{R}$ and $I = (0, \infty)$ else. Let $I' = (r_2', \infty)$ if $r_2' \in \mathbb{R}$ and $I' = (0, \infty)$ else. Let $S = I \cap I'$.
Using the Sturm sequence compute the number of solutions of $f(y)$ in S. If this number is positive, set $u = 1$, otherwise set $u = -1$.

Example B.2. We consider the Fuller's problem (`https://mintoc.de/index.php/Fuller's_problem`)

$$\begin{array}{ll} \min\limits_{x(\cdot), w(\cdot)} & \int_0^1 x_1^2 dt \\ \text{s.t.} & \dot{x}_1(t) = x_2(t), \quad t \in [0,1] \text{ a.e.} \\ & \dot{x}_2(t) = 1 - 2w(t), \quad w(t) \in \{0,1\},\ t \in [0,1] \text{ a.e.} \\ & x(0) = (0.01, 0), \end{array} \tag{B.16}$$

Firstly, we can rewrite (B.16) as a relaxed problem as follow

$$\begin{array}{ll} \min\limits_{x(\cdot), \alpha(\cdot)} & \int_0^1 x_1^2 dt \\ \text{s.t.} & \dot{x}_1(t) = x_2(t), \quad t \in [0,1] \text{ a.e.} \\ & \dot{x}_2(t) = 1 - 2\alpha(t), \quad \alpha(t) \in [0,1],\ t \in [0,1] \text{ a.e.} \\ & x(0) = (0.01, 0), \end{array} \tag{B.17}$$

We drop the superscript * to simplify the notation. The HAMILTON function reads

$$\mathcal{H}(x, \lambda, \alpha) = x_1^2 + \lambda_1 x_2 + \lambda_2(1 - 2\alpha).$$

The maximality condition implies

$$\begin{aligned}\mathcal{H}(x(t),\lambda(t),\alpha(t)) &= \min_{0\le\alpha\le 1}\{x_1^2(t)+\lambda_1(t)x_2(t)+(1-2\alpha)\lambda_2(t)\}\\ &= x_1^2(t)+\lambda_1(t)x_2(t)-\operatorname{sgn}(\lambda_2(t))\lambda_2(t)\end{aligned}$$

while the optimal control is given by

$$\alpha(t)=\begin{cases}0 & \text{if } \lambda_2(t)<0\\ 1 & \text{if } \lambda_2(t)>0.\end{cases}$$

Hence the optimal state trajectories, with c_1, c_2, c_1', c_2' are the appropriate constants such that $x_1(0)=0.01$ and $x_2(0)=0$, are

$$x_1(t)=\begin{cases}\frac{t^2}{2}+c_1t+c_1' & \text{if } \lambda_2(t)<0\\ -\frac{t^2}{2}+c_2t+c_2' & \text{if } \lambda_2(t)>0\end{cases},\quad x_2(t)=\begin{cases}t+c_1 & \text{if } \lambda_2(t)<0\\ -t+c_2 & \text{if } \lambda_2(t)>0\end{cases},$$

and the adjoint equations are $\dot\lambda_1(t)=-2x_1(t)$, $\dot\lambda_2(t)=-\lambda_1(t)$, which together with the terminal condition $\lambda(t_f)=\lambda(1)=0$ imply the optimal co-states

$$\begin{aligned}\lambda_1(t)&=\begin{cases}-\frac{t^3}{3}-\frac{c_1t^2}{2}-c_1't+c_1'' & \text{if } \lambda_2(t)<0\\ \frac{t^3}{3}-\frac{c_2t^2}{2}-c_2't+c_2'' & \text{if } \lambda_2(t)>0,\end{cases}\\ \lambda_2(t)&=\begin{cases}\frac{t^4}{12}+\frac{c_1t^3}{6}+\frac{c_1't^2}{2}-c_1''t+c_1''' & \text{if } \lambda_2(t)<0\\ -\frac{t^4}{12}+\frac{c_2t^3}{6}+\frac{c_2't^2}{2}-c_2''t+c_2''' & \text{if } \lambda_2(t)>0.\end{cases}\end{aligned}\tag{B.18}$$

We consider a Gröbner basis approach. In this example, there are only two possible strategies where the input alternates between 1 and 0, taking the values $0,1,0,\ldots$, or $1,0,1,\ldots$, respectively. In each case, taking into account the maximal numbers of switching, one can not so difficult to derive an expression for the objective function value,

$$\int_0^1 x_1^2dt\approx\frac{1}{4}\left(\frac{x_1^2(0)}{2}+x_1^2(t_1)+x_1^2(t_1+t_2)+x_1^2(t_1+t_2+t_3)+x_1^2(\sum_{i=1}^4 t_i)+\frac{x_1^2(1)}{2}\right)$$

as a function of the switching times. From (B.18), we see that the control input switches at most four times. Designate by t_1, t_2, t_3, t_4 and t_5 the length of the successive intervals where $\alpha(t)$ stays constant. Any set of initial and final conditions can be translated to having $x(0)=(0.01,0)$ and a given value for $x(1)$, and this is the setting from here on. The particular choice (among the only two possible ones)

$$\alpha(t)=\begin{cases}1, & \text{for } 0\le t<t_1\\ 0, & \text{for } t_1\le t<t_1+t_2\\ 1, & \text{for } t_1+t_2\le t<t_1+t_2+t_3\\ 0, & \text{for } t_1+t_2+t_3\le t<t_1+t_2+t_3+t_4\\ 1, & \text{for } t_1+t_2+t_3+t_4\le t<t_1+t_2+t_3+t_4+t_5=1\end{cases}$$

drives the chain of integrators for the origin to the approximated points (exactly the points at the switches) and the final point $x(1)$ given by

$$\begin{aligned}
&x_2(1) = -t_1 + t_2 - t_3 + t_4 - t_5 = 2t_2 + 2t_4 - 1\\
&x_1(1) = -\frac{t_2^2}{2} - t_4^2 - 2t_1t_4 + 2t_2t_4 + 2t_3t_4 - t_1t_2 + 2t_2 + 2t_4 + 0.01 - \frac{1}{2}\\
&x_2(t_1 + t_2 + t_3 + t_4) = -t_1 + t_2 - t_3 + t_4\\
&x_1(t_1 + t_2 + t_3 + t_4) = -\frac{t_1^2}{2} + t_2^2 - \frac{t_3^2}{2} + \frac{t_4^2}{2} - t_1t_3 - t_1t_4 + t_2t_3 + t_2t_4 - t_3t_4 + 0.01\\
&x_2(t_1 + t_2 + t_3) = -t_1 + t_2 - t_3\\
&x_1(t_1 + t_2 + t_3) = -\frac{t_1^2}{2} - \frac{t_3^2}{2} - t_1t_3 + t_2t_3 + 0.01\\
&x_2(t_1 + t_2) = -t_1 + t_2\\
&x_1(t_1 + t_2) = -\frac{t_1^2}{2} + 0.01\\
&x_2(t_1) = -t_1\\
&x_1(t_1) = -\frac{t_1^2}{2} + 0.01
\end{aligned} \tag{B.19}$$

therein,

$$x_2(t) = \begin{cases} -t, & \text{for } t \in T_0\\ t - 2t_1, & \text{for } t \in T_1\\ -t + 2t_2, & \text{for } t \in T_2\\ t - 2t_1 - 2t_3, & \text{for } t \in T_3\\ -t + 2t_2 + 2t_4, & \text{for } t \in T_4 \end{cases}$$

$$x_1(t) = \begin{cases} -\frac{t^2}{2} + 0.01, & \text{for } t \in T_0\\ \frac{t^2}{2} - 2t_1t + t_1^2 + t_1t_2 - \frac{t_2^2}{2} + 0.01, & \text{for } t \in T_1\\ -\frac{t^2}{2} + 2t_2t - \frac{3t_2^2}{2} - t_1t_2 + 0.01, & \text{for } t \in T_2\\ \frac{t^2}{2} - 2(t_1 + t_3)t + (t_1 + t_2 + t_3)^2 - t_1t_2 - \frac{t_2^2}{2} + 0.01, & \text{for } t \in T_3\\ -\frac{t^2}{2} + 2(t_2 + t_4)t - t_4^2 - 2t_4(t_1 + t_2 + t_3) - t_1t_2 - \frac{t_2^2}{2} + 0.01, & \text{for } t \in T_4 \end{cases}$$

where $T_0 := [0, t_1)$, $T_1 := [t_1, t_1 + t_2)$, $T_2 := [t_1 + t_2, t_1 + t_2 + t_3)$, $T_3 := [t_1 + t_2 + t_3, t_1 + t_2 + t_3 + t_4)$, $T_4 := [t_1 + t_2 + t_3 + t_4, 1)$.
It turns out that the selection between alternating values $1, 0, 1, \ldots$, or $0, 1, 0, \ldots$ for the optimal input $\alpha(t)$ depends on whether the equations in (B.19) have a solution for a specified condition $\{x(t_1), x(t_1 + t_2), x(t_1 + t_2 + t_3), x(t_1 + t_2 + t_3 + t_4), x(1)\}$.
We refer to the computational algebraic geometry and Gröbner bases, cf. [125, Sec. III], for the following calculation. We set

$$w := t_1, \qquad x := t_2, \qquad y := t_3, \qquad z := t_4,$$

and $a := x_2(1)$, $b := x_1(1)$, $c := x_2(t_1+t_2+t_3+t_4)$, $d := x_1(t_1+t_2+t_3+t_4)$, $e := x_2(t_1+t_2+t_3)$, $f := x_1(t_1 + t_2 + t_3)$, $g := x_2(t_1 + t_2)$, $h := x_1(t_1 + t_2)$, $i := x_2(t_1)$, $j := x_2(t_1)$.
Then we solve the complex version of the following switching problem.

Problem 1: Given is the system of equations

$$
\begin{aligned}
a &= 2x + 2z - 1 \\
b &= -\frac{x^2}{2} - z^2 - 2wz + 2xz + 2yz - wx + 2x + 2z - 0.49 \\
c &= -w + x - y + z \\
d &= -\frac{w^2}{2} + x^2 - \frac{y^2}{2} + \frac{z^2}{2} - wy - wz + xy + xz - yz + 0.01 \\
e &= -w + x - y \\
f &= -\frac{w^2}{2} - \frac{y^2}{2} - wy + xy + 0.01 \\
g &= -w + x \\
h &= -\frac{w^2}{2} + 0.01 \\
i &= -w \\
j &= -\frac{w^2}{2} + 0.01.
\end{aligned}
\tag{B.20}
$$

We use the Macaulay symbolic program to compute the complex solutions w, x, y, z of the above system. By a similar way as in [125, Subsec. A. Complex Solutions] and Ex. B.1, we can conclude that

i. the system does always have a complex solution;

ii. if $h = j \leq 0.01$ and $a, b, c, d, e, f, g, h, i, j \in \mathbb{R}$, the system has real solutions;

iii. if $h = j < 0.01$ and $a, b, c, d, e, f, g, h, i, j \in \mathbb{R}$, the system has real nonnegative solutions.

We are interested in the third case. Hence, we will investigate the following question as a second step.

Problem 2: Given are $a, b, c, d, e, f, g, h, i, j \in \mathbb{R}$. Does there exist a nonnegative solution vector (w, x, y, z) for the system (B.20) in the sense that $w > 0$, $x \geq 0$, $y \geq 0$, $z \geq 0$? If there is a positive solution w, x, y, z, then the value of the optimal control α assumes the values $1, 0, 1, 0, 1$ successively, and in particular, the present value for the optimal control is $\alpha(0) = 1$. If no positive solution exists then the present value of the optimal control is $\alpha(0) = 0$.
We will use *Sturm sequences* similarly as the previous example to compute suitable Gröbner bases together with an algorithm from real algebraic geometry.
As a first step, we compute a Gröbner basis for the ten polynomials in (B.20) under an

elimination order with $w > x > y > z > j > i > h > g > f > e > d > c > b > a$. One gets

$$\begin{aligned}
b &= -\frac{x^2}{2} - z^2 - 2wz + 2xz + 2yz - wx + 2x + 2z - 0.49 \\
d &= -\frac{w^2}{2} + x^2 - \frac{y^2}{2} + \frac{z^2}{2} - wy - wz + xy + xz - yz + 0.01 \\
f &= -\frac{w^2}{2} - \frac{y^2}{2} - wy + xy + 0.01 \\
h &= -\frac{w^2}{2} + 0.01 \\
j &= -\frac{w^2}{2} + 0.01 \\
w &= -i \\
x &= g - i \\
y &= g - e \\
z &= \frac{a+1}{2} - g + i.
\end{aligned} \tag{B.21}$$

Remark 42. From examples B.1 and B.2, for solving these problems by using the Gröbner basis, we realize that the complexity rises a lot when we increase the number of variables or the degrees of the polynomial in the equations.

Bibliography

[1] W. Achtziger and C. Kanzow. Mathematical programs with vanishing constraints: optimality conditions and constraint qualifications, *Mathematical Programming Series A*, 114:69-99, 2008. 75, 76

[2] J. Albersmeyer. *Adjoint-based algorithms and numerical methods for sensitivity generation and optimization of large scale dynamic systems*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2010. 21, 99

[3] P. Antsaklis and X. Koutsoukos. *On hybrid control of complex systems: A survey*, In 3rd International Conference ADMP'98, Automation of Mixed Processes: Dynamic Hybrid Systems, March 1998, pp. 1-8. 25

[4] N. Arada, J. P. Raymond. Optimal control problems with mixed control-state constraints. *SIAM J. Control Optim.* 39(5):1391-1407, 2000. 35

[5] Y. Arkun. Detection of biological switches using the method of Gröbner bases, *BMC Bioinformatics*, 20:615, 2019. 152

[6] B. Armstrong-Hétlouvry, P. Dupont, and C. Canudas de Wit. A survey of models, analysis tools and compensation methods for the control of machines with friction, *Automatica*, 30(7):1083-1138, 1994. 103

[7] J. P. Aubin and A. Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Springer-Verlag, Berlin Heidelberg, 1984. ISBN 978-3-642-69514-8. 28

[8] E. Balas. Disjunctive Programming and a Hierarchy of Relaxations for Discrete Optimization Problems, *SIAM Journal on Algebraic and Discrete Methods*, 6(3):466-486, 1985. 32

[9] V. Bansal, V. Sakizlis, R. Ross, J.D. Perkins, and E.N. Pistikopoulos. New algorithms for mixed-integer dynamic optimization, *Computers & Chemical Engineering*, 27:647-668, 2003. 26

[10] A. Bemporad, G. Ferrari Trecate, and M. Morari. Observability and controllability of piecewise affine and hybrid systems, *IEEE Transactions on Automatic Control*. 45:1864-1876, 1999. 25

[11] A. Bemporad and M. Morari. Control of systems integrating logic, dynamics, and constraints, *Automatica* 35(3):407-427, 1999. 26

[12] F. Bestehorn, C. Hansknecht, C. Kirches and P. Manns. *A switching cost aware rounding method for relaxations of mixed-integer optimal control problems*, IEEE 58th Conference on Decision and Control (CDC), Nice, France, pp. 7134-7139, 2019. 32, 78, 79, 102

[13] A. Boccia, M. R. de Pinho, R. Vinter. Optimal control problems with mixed and pure state constraints. *SIAM J. Control Optim.* 54(6):3061-3083, 2016. 35

[14] H. G. Bock. Numerical Solution of Nonlinear Multipoint Boundary Value Problems with Applications to Optimal Control. *Zeitschrift für Angewandte Mathematik und Mechanik*, 58:407, 1978. 15

[15] H. G. Bock. *Numerische Berechnung Zustandsbeschränkter Optimaler Steuerungen Mit Der Mehrzielmethode*. Carl-Cranz-Gesellschaft, Heidelberg, 1978. 15

[16] H. G. Bock. *Randwertproblemmethoden zur Parameteridentifierung in Systemen nichtlinearer Differentialgleichungen*. PhD dissertation, Bonner mathematische Schriften 183, Universität Bonn, 1987. 96, 99

[17] H. G. Bock, A. Cadi, R. W. Longman, J. P. Schlöder. *Minimum Energy Time Tables for Subway Operation - And Hamiltonian Feedback to Return to Schedule*, In: Bock H., Phu H., Rannacher R., Schlöder J. (eds) Modeling, Simulation and Optimization of Complex Processes HPSC 2015. Springer, Cham. 2017. `https://doi.org/10.1007/978-3-319-67168-0_1`. 35, 52, 55, 85

[18] H. G. Bock, C. Kirches, A. Meyer & A. Potschka. Numerical solution of optimal control problems with explicit and implicit switches, *Optimization Methods and Software*, 33(3):450-474, 2018. 25, 29, 30, 32, 90, 103

[19] H. G. Bock, R. W. Longman. *Optimal Control of Velocity Profiles for Minimization of Energy Consumption in the New York Subway System*, In: Proceedings of the Second IFAC Workshop on Control Applications of Nonlinear Programming and Optimization, Oberpfaffenhofen, pp. 34-43. International Federation of Automatic Control, 1980. 35, 52, 55, 85

[20] H. G. Bock, R. W. Longman. *Computation of optimal controls on disjoint control sets for minimum energy subway operation*, in: Proceedings of the American Astronomical Society, Symposium on Engineering Science and Mechanics, Taiwan, 1982. 52, 55, 130

[21] H. G. Bock, R. W. Longman. Computation of optimal controls on disjoint control sets for minimum energy subway operation. *Adv. Astronaut. Sci.* 50:949-972, 1985. 52, 55, 130

[22] H. G. Bock, K. J. Plitt. *A Multiple Shooting Algorithm for Direct Solution of Optimal Control Problems*, In Proceedings of the 9th IFAC World Congress, pp. 242-247, Budapest. Pergamon Press, 1984. 19, 21, 60, 68

[23] H. G. Bock and J. P. Schlöder. Numerical solution of retarded differential equations with state dependent time lags. *Z. Angew. Math. Mech.*, 61:T269-T271, 1981. 94

[24] N. Bolotnik. Optimal control of a point mass on a rough plane, (problem statement) 2023. `https://github.com/DuyTranHD/OCP_PMRP` 107

[25] V. G. Boltyanski, R. V. Gamkrelidze, E. F. Mishchenko, and L. S. Pontryagin. The Maximum Principle in the Theory of Optimal Processes of Control. *IFAC Proceeding Volumes*, 1(1):464-469, 1960. 15, 16

[26] J. F. Bonnans, A. Hermant. Second-order analysis for optimal control problems with pure state constraints and mixed control-state constraints, *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 26(2):561-598, 2009. 132

[27] J. F. Bonnans, and A. Shapiro. *Perturbation Analysis of Optimal Control Problems*, Springer, New York 2000. 132

[28] F. Borrelli, M. Baotic, A. Bemporad, and M. Morari. *An efficient algorithm for computing the state feedback optimal control law for discrete time hybrid systems*, Proceedings of the 2003 American Control Conference, 2003(6):4717-4722, 2003. 26

[29] U. Boscain, B. Piccoli. *An Introduction to Optimal Control*, http://www.cmapx.polytechnique.fr/∼boscain/AUTOMATICS/introduction-to-optimal-control.pdf.

[30] M. S. Branicky, V. S. Borkar, and S. K. Mitter, A Unified Framework for Hybrid Control: Model and Optimal Control, *IEEE Transactions on Automatic Control* 43(1):31-45, 1998. 25

[31] B. V. Brunt and M. Carter. *The Lebesgue-Stieltjes Integral: A Practical Introduction*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2000. ISBN `978-0-387-95012-9`. 5

[32] M. B. Carver. Efficient integration over discontinuities in ordinary differential equation simulations. *Math. Comp. Simul.*, 20(3):190-196, 1978. 94, 103

[33] M. B. Carver and S. R. MacEwen. Numerical analysis of a system described by implicitly-defined ordinary differential equations containing numerous discontinuities. *Applied Math. Modeling*, 2:280-286, 1978. 94

[34] B. A. Chartres and R. S. Stepleman. Actual order of convergence of Runge-Kutta methods of differential equations with discontinuities. *SIAM J. Numer. Anal.*, 11:1193-1206, 1974. 94

[35] B. A. Chartres and R. S. Stepleman. Convergence of linear multistep methods for differential equations with discontinuities. *Numer. Math.*, 27:1-10, 1976. 94

[36] F. H. Clarke. *Optimization and Nonsmooth Analysis. Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 1990. ISBN 978-0-89871-256-8. 26

[37] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory*, volume 178. Springer Science & Business Media, 2008. 26, 28

[38] F. H. Clarke, M. R. de Pinho. Optimal control problems with mixed constraints. *SIAM Journal on Control and Optimization*. 48(7):4500-4524, 2010. 35

[39] R. Cominetti. Metric regularity, tangent sets, and second-order optimality conditions. *Applied Mathematics and Optimization* 21:265-287, 1990. 134

[40] J. Cortes. Discontinuous dynamical systems, in *IEEE Control Systems Magazine*, 28(3):36-73, 2008. 90

[41] M. Diehl. *Real-Time Optimization for Large Scale Nonlinear Processes*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2001. 24

[42] A. V. Dmitruk. Maximum principle for the general optimal control problem with phase and regular mixed constraints. *Comput Math Model.* 4:364-377, 1993. 35

[43] A. V. Dmitruk. On the development of Pontryagin's maximum principle in the works of A. Y. Dubovitskii and A. A. Milyutin. *Control and Cybernetics*, 38(4A):923-957, 2009. 16, 17, 35

[44] A. Dmitruk, N. P. Osmolovskii. Necessary Conditions for a Weak Minimum in Optimal Control Problems with Integral Equations Subject to State and Mixed Constraints. *SIAM J. on Control and Optimization*, 52:3437-3462, 2014. 35

[45] A. V. Dmitruk, N. P. Osmolovskii. Local minimum principle for optimization problems with different types of control systems subject to mixed state-control constraints, *International Journal for Computational Civil and Structural Engineering*, 14(2):57-64, 2018. 6, 17, 35, 56, 119, 120, 133

[46] A. Dmitruk, N. Osmolovskii. Local Minimum Principle for An Optimal Control Problems with A Nonregular Mixed Constraint. `SIAM J. Control Optim.` 60(4):1919-1941, 2022. 16, 35

[47] A. Dmitruk, N. P. Osmolovskii. Local Minimum Principle for Optimal Control Problems with Mixed Constraints: The Nonregular Case. *Appl Math Optim* 88(16), 2023. 9, 16, 19, 35

[48] A. Y. Dubovitskii, A. A. Milyutin. Theory of the maximum principle. On Sat. Methods of the theory of extremal problems in economics. M., Science 6-47 (in Russian), 1981. 6, 35

[49] E. Eich. *Mehrschrittverfahren zur numerischen Lösung von Bewegungsgleichungen technischer Mehrkörpersysteme mit Zwangsbedingungen und Unstetigkeiten*, PhD dissertation, Universität Augsburg, 1991. 21, 28, 91, 94, 98, 103

[50] D. Ellison. Efficient automatic integration of ODEs with discontinuities. *Math. Comput. Simul.*, 23(1):12-20, 1981. 94

[51] W. H. Enright, K. R. Jackson, S. P. Norsett and P. G. Thomson. Effective solution of discontinuous IVPs using Runge-Kutta formula pair with interpolants. *Appl. Math. Comp.*, 27:313-335, 1986. 94

[52] D. J. Evans and S. O. Fatunla. Accurate numerical determination of the intersection point of the solution of a differential equation with a given algebraic relation. *J. Inst. Math. Appl.*, 16(3):355-369, 1975. 94

[53] L. C. Evans. An Introduction to Mathematical Optimal Control Theory, Department of Mathematics University of California, Berkeley, 2013. `https://math.berkeley.edu/ evans/control.course.pdf`

[54] T. Yu. Figurina. Optimal Control of System of Material Points in a Straight Line with Dry Friction, *J. Comput. Syst. Sci. Int.* 54:671-677, 2015. 103, 115, 122

[55] A. F. Filippov. On certain questions in the theory of optimal control, *J. SIAM Control*, Ser. A, 1(1), 1962. 14, 26

[56] A. F. Filippov. Differential Equations with discontinuous right hand side. *AMS Transl.*, 42:199-231, 1964. 26, 30, 96, 98

[57] C. W. Gear and O. Osterby. Solving ordinary differential equations with discontinuities, *ACM Trans. Math. Softw.*, 10(1): 23-44, 1984. 92

[58] M. Gerdts. Solving mixed-integer optimal control problems by branch & bound: A case study from automobile test-driving with gear shift, *Optim. Control Appl. Methods*, 116:1-18, 2005. 19

[59] B. Gnaedig, J. Burgschweiger and M. C. Steinbach. Optimization models for operative planning in drinking water networks. *Optimization and Engineering*, 10(1), 43-73, 2009. 60

[60] R. Goebel, R. Sanfelice, and A.R. Teel. Hybrid dynamical systems, *IEEE Control Systems Magazine*, 29(2):28-93, 2009. 25

[61] D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at `https://math.uiuc.edu/Macaulay2/`. 153

[62] H. J. Halin. Integration of ordinary differential equations containing discontinuities. In *Proceeding of the Summer Computer Simulation Conference, La Jolla*, SCI Pres, La Jolla, California, S. 46-53, 1976. 94

[63] R. F. Hartl, and S. P. Sethi, and R. G. Vickson. A Survey of the Maximum Principles for Optimal Control Problems with State Constraints. *SIAM Review*, 37(2):181-218, 1995. 15, 35

[64] S. Hedlund and A. Rantzer, Convex Dynamic Programming for Hybrid Systems, *IEEE Transactions on Automatic Control* 47 (2002), no. 9, 1536-1540. 25

[65] G. Hicks and W. Ray. Approximation Methods for Optimal Control Synthesis. *The Canadian Journal of Chemical Engineering*, 49(4):522-528, 1971. 20

[66] N. J. Higham. Schreiber R. Demmel, J.W. Block lu factorization. *RIACS Technical Report*, no.92-03, 1992. 6

[67] N. J. Higham. Gaussian elimination. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(3):230-238, 2011. 6

[68] T. Hoheisel and C. Kanzow. First- and second-other optimality conditions for mathematical programs with vanishing constraints, *Applications of Mathematics*, 52:459-514, 2007. 75

[69] T. Hoheisel and C. Kanzow. On the Abadie and Guignard constraint qualifications for Mathematical Programmes with Vanishing Constraints. *Optimization*, 58(4): 431-448, 2009. 11, 76

[70] IFDIFF - A Matlab Toolkit for ODEs with State-Dependent Switches `https://github.com/andreassommer/ifdiff/blob/public/README.md`. 99

[71] J. I. Imura and A. Van Der Schaft. Characterization of well-posedness of piecewise-linear systems, *IEEE Transactions on Automatic Control*, 45(9):1600-1619, 2000. 25

[72] A.F. Izmailov and M.V. Solodov. Mathematical programs with vanishing constraints: Optimality conditions, sensitivity, and a relaxation method, *Journal of Optimization Theory and Applications*, 142:501-532, 2009. 76

[73] M. Jung. *Relaxations and Approximations for Mixed-Integer Optimal Control*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2014. 32, 34

[74] W. Karush. *Minima of Functions of Several Variables with Inequalities as Side Conditions*. Master's thesis, Department of Mathematics, University of Chicago, 1939. 12

[75] C. Kirches. *A numerical method for nonlinear robust optimal control with implicit discontinuities and an application to powertrain oscillations*. Diploma thesis, IWR, Ruprecht-Karls University Heidelberg, 2006. 60, 100

[76] C. Kirches. *Fast Numerical Methods for Mixed-Integer Nonlinear Model-Predictive Control*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2010. 21, 22, 24, 33, 150

[77] C Kirches, A Potschka, HG Bock, S Sager. *A parametric active set method for quadratic programs with vanishing constraints*, 2012. `https://www.mathopt.de/PUBLICATIONS/Kirches2011.pdf`. 76, 77

[78] C. Kirches, E. Kostina, A. Meyer, M. Schlöder. *Numerical Solution of Optimal Control Problems with Switches, Switching Costs and Jumps*, 2019. `https://optimization-online.org/2018/10/6888/`. 102

[79] T. Kleinert and V. Hagenmeyer. *Flat Hybrid Automata as a Class of Reachable Systems: Introductory Theory and Examples*, `arXiv preprint arXiv:1906.02790`, 2019. 87

[80] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover Books on Mathematics. Dover Publications, New York, 1975. ISBN `978-0-486-61226-3`. 5

[81] E. Kostina, O. Kostyukova, W. Schmidt. New Necessary Conditions for Optimal Control Problems in Discontinuous Dynamic Systems. 25th System Modeling and Optimization (CSMO), Sep 2011, Berlin, Germany. pp.122-135, `10.1007/978-3-642-36062-6_13. hal-01347530` 40, 42

[82] O. I. Kostyukova and E. A. Kostina. Necessary Conditions for Optimality in Problems of Optimal Control of Systems with Discontinuous Right-Hand Side, *Differential Equations*, 55(3):374-389, 2019. 35, 49, 147, 148, 149

[83] P. Krämer-Eis, and H. Bock. Numerical Treatment of State and Control Constraints in the Computation of Feedback Laws for Nonlinear Control Problems. In et al., P. D., editor, *Large Scale Scientific Computing*, pages 287-306. Birkhäuser, 1987. 34, 80, 81

[84] P. Krämer-Eis, H. Bock, R. Longman, and J. Schlöder. Numerical Determination of Optimal Feedback Control in Nonlinear Problems with State/Control Constraints. *Advances in the Astronautical Sciences*, 105:53-71, 2000. 34

[85] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1951. University of California Press. 12

[86] D. Leineweber, I. Bauer, H. G. Bock, and J. P. Schlöder. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part I: Theoretical aspects, *Computers & Chemical Engineering*, 27:157-166, 2003. 68

[87] R. I. Leine. *Bifurcations in discontinuous mechanical systems of the Fillippov-type*. Dissertation, Mechanical Engineering, Technische Universiteit Eindhoven, 2000. `https://doi.org/10.6100/IR533239`. 90

[88] D. Leineweber. *Efficient Reduced SQP Methods for the Optimization of Chemical Processes Described by Large Sparse DAE Models*. PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 1999. 21

[89] F. Lenders. *Numerical methods for mixed-integer optimal control with combinatorial constraints*, PhD Dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2017, `DOI: 10.11588/heidok.00024070`. 31, 32

[90] R. C. Loxton, K. L. Teo, and V. Rehbock. Optimal control problems with multiple characteristic time points in the objective and constraints, *Automatica* 44(11):2923-2929, 2008. 26

[91] R. Mannshardt. One-step methods of any order for ordinary differential equations with discontinuous right hand sides. Differentialgleichungen mit Sprungfunktionen. *Numer. Math.*, 31(2):131-152, 1978. 94, 106, 107

[92] R. B. Martin. Optimal control drug scheduling of cancer chemotherapy, *Automatica*, 28(6):1113-1123, 1992. 26

[93] R. Martinez, J. Alvarez. A controller for 2-DOF under actuated mechanical systems with discontinuous friction. *Nonlinear Dyn* 53:191-200, 2008. 90

[94] H. Maurer. First and second order sufficient optimality conditions in mathematical programming and optimal control. *Mathematical Programming Study* 14:163-177, 1981. 132

[95] A. Meyer. *Numerical solution of optimal control problems with explicit and implicit switches*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2020. 21, 22, 27, 28, 60, 90, 103

[96] J. A. Moreno. *Discontinuous integral control for mechanical systems*, 14th International Workshop on Variable Structure Systems (VSS), Nanjing, China, pp. 142-147, 2016. 90

[97] J. Nocedal, S. J. Wright. *Numerical Optimization*, in Springer Series in Operations Research and Financial Engineering, 2nd ed, Springer New York, NY, USA, 2006. 7, 12

[98] N. P. Osmolovskii. Sufficient quadratic conditions of extremum for discontinuous controls in optimal control problems with mixed constraints. *J. Math. Science* 173:1-106, 2011. 132

[99] N. P. Osmolovskii. Second-order optimality conditions for control problems with linearly independent gradients of control constraints. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(29):452-482, April 2012. 132

[100] N. P. Osmolovskii. A second-order sufficient condition for a weak local minimum in an optimal control problem with an inequality constraint*, *Control and Cybernetics*, 51(2):151-169, 2022. 133, 134

[101] D. U. Patil, D. Chakraborty. Computation of Time Optimal Feedback Control Using Gröbner Basis, *IEEE Transactions on automatic control*, 59(8):2271-2276, 2015. 152

[102] D. U. Patil, D. Chakraborty. Time Optimal Feedback Control using Chebyshev Polynomials and Gröbner Basis, *IFAC-PapersOnLine*, 48(14):50-55, 2015. 152

[103] H. J. Pesch, R. Bulirsch. The maximum principle, Bellman's equation, and Carathéodory's work. *J Optim Theory Appl* 80, 199-225, 1994. 35

[104] D. Peterson. A review of constraint qualifications in finite dimensional spaces, *SIAM Review*, 15(3):639-654, July 1973. 11

[105] B. Piccoli, *Necessary conditions for hybrid optimization*, Proceedings of the 38th IEEE Conference on Decision and Control, vol. 1, 1999, pp. 410-415 vol.1. 25

[106] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, E. F. Mishechenko. *The Mathematical Theory of Optimal Processes*. VIII + 360 S. New York/London, John Wiley & Sons, 1962. 15, 16

[107] A. Potschka, H. Bock, and J. Schlöder. A minima tracking variant of semi-infinite programming for the treatment of path constraints within direct solution of optimal control problems. *Optimization Methods and Software*, 24(2):237-252, 2009. 22

[108] R. Raman and I. E. Grossmann. Modelling and computational techniques for logic based integer programming, *Computers and Chemical Engineering*, 18(7):563-578, 1994. 32

[109] J. Roll. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37-50, 2004. 25

[110] E. Roxin. The existence of optimal controls. *Michigan Math Journal*, 9(2):109-119, 1962, DOI: 10.1307/mmj/1028998668. 14

[111] S. Sager. *Numerical methods for mixed-integer optimal control problems*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2006. 30, 33, 55

[112] S. Sager, H. G. Bock, and G. Reinelt. Direct methods with maximal lower bound for mixed-integer optimal control problems. *Mathematical Programming*, 118(1):109-149, 2007. 30

[113] S. Sager, H. G. Bock, and M. Diehl. The integer approximation error in mixed-integer optimal control. *Mathematical Programming*, 133(1-2):1-23, 2010. 30

[114] S. Sager, M. Jung, C. Kirches. Combinatorial integral approximation, *Mathematical Methods of Operational Research* 73(3):363-380, 2011. 78, 83, 84

[115] R. Scholz. *Model-based optimal feedback control for microgrids*, PhD dissertation, Ruprecht-Karls University Heidelberg, Heidelberg, 2022. 24

[116] E. D. Sontag. *Mathematical Control Theory*, Springer Verlag, New York, 2nd ed., 1998. 13

[117] H. J. Sussmann. *A maximum principle for hybrid optimal control problems*, Proceedings of the IEEE conference on decision and control, Pheonix, AR pp. 425-430, 1999. 25

[118] P. Tankov. *Financial Modelling with Jump Processes.* Chapman and Hall/CRC Financial Mathematics Series. CRC Press, 2003. 5

[119] S. Terwen, M. Back, and V. Krebs. Predictive powertrain control for heavy duty trucks. *IFAC Proceedings Volumes*, 37(22), 105-110, 2004. 60

[120] G. Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of Graduate Studies in Mathematics. American Mathematical Society, 2012. ISBN 978-0-8218-8328-0. 27

[121] B. D. Tran, E. Kostina. *On the Optimality Conditions of the Switched Optimal Control Problem: Disjunctive Programming Reformulation and Local Maximum Principle*, In: Proceeding of 2024 9th International Conference on Control and Optimization with Industrial Applications (COIA), IEEE Turkey Section, pp. 670-675, August 2024. ISBN 978-625-97879-3-0. 35, 52, 90, 115

[122] B. D. Tran, E. Kostina. *On The Local Maximum Principle for Switched Optimal Control Problems: Efficient Regular Reformulation for Mixed Constraints and On-line Neighboring Feedback Control* [in preparation], 2025. 35, 90

[123] B. D. Tran, E. Kostina. *On the switched optimal control problems: a new reformulation, constraint qualifications, and a direct solution approach* [in preparation], 2025. 90

[124] R. B. Vinter. *Optimal Control and Pontryagin's Maximum Principle*, In: Baillieul J., Samad T. (eds) Encyclopedia of Systems and Control, Springer, London, 2013. DOI 10.1007/978-1-4471-5102-9_200-1. 35

[125] U. Walther, T. T. Georgiou, and A. Tannenbaum. On the Computation of Switching Surfaces in Optimal Control: A Gröbner Basis Approach, *IEEE Transactions on automatic control*, 46(4), April 2001. 152, 153, 155, 157, 158, 160, 161

[126] X. Xu and P. Antsaklis. Optimal control of switched autonomous systems. In *Proceedings of the 41st IEEE Conference on Decision and Control*, IEEE, 2002. doi:10.1109/cdc.2002.1185065. 60, 90

[127] L. Young. *Generalized curves and the existence of an attained absolute minimum in the calculus of variations.* C.R. Soc. Sci. Lett. Varsovie, CL III, 30:212-234, 1937. 30

[128] F. Zahn, T. Kleinert, V. Hagenmeyer. On Optimal Control of Flat Hybrid Automata, *IFAC-PapersOnLine*, 53(2):6800-6805, 2020, 21st IFAC World Congress. 87, 88

[129] V. Zeidan. *Second-order conditions for optimal control problems with mixed state-control constraints*, Proceedings of 32nd IEEE Conference on Decision and Control, vol.4 pp. 3800-3805, 1993. doi: 10.1109/CDC.1993.325929. 132

[130] E. Zeidler. *Nonlinear Functional Analysis and Its Applications: Fixed Point Theorems.* Nonlinear Functional Analysis and Its Applications. Springer-Verlag, New York, 1986. ISBN 978-0-387-90914-1. 4, 7, 8

[131] C. Zeile, T. Weber and S. Sager. Combinatorial Integral Approximation Decompositions for Mixed-Integer Optimal Control, *Algorithms* 15:121, 2022. 83, 84

[132] F. Zhu and P. J. Antsaklis. Optimal control of hybrid switched systems: A brief survey, *Discrete Event Dynamic Systems*, 25(3):345-364, 2015. 25, 26, 90

[133] * A. P. Afanasiev, V. V. Dikusar, A. A. Milyutin, S. A. Chukanov. *Necessary condition in optimal control*. M. (in Russian), 1990. 40

[134] * Aseyev, S.M. "The Method of Smooth Approximations in the Theory of Optimality Conditions for Differential Inclusions." Bulletin of the Russian Academy of Sciences. Mathematical Series, 61(2b): 3-26, (in Russian) 1997. 40

[135] * V. V. Dikusar, A. A. Milyutin. "Qualitative and Numerical Methods in the Maximum Principle." Moscow, (in Russian) 1989. 40

Nomenclature

List of Symbols

$\stackrel{\text{def}}{=} / :=$	Defined to be equal
$\square$	End of a proof
$\cup$	Set-theoretic union("unified with")
$\cap$	Set-theoretic intersection("intersected with")
$\wedge$	Logical conjunction("AND")
$\oplus, \vee$	Logical exclusive/inclusive disjunction("EX-OR/OR")
$\supseteq, \supset$	Superset of a set("is a (proper) superset of")
$\subseteq, \subset$	Subset of a set("is a (proper) subset of")
$\in, \notin$	Set membership("is (not) an element of")
$\setminus$	Set difference
$\emptyset$	The empty set
$\forall$	Universal quantification("for all")
$\exists$	Existential quantification("exists")
$A_{i,\cdot}, A_{\cdot,j}$	i-th row/j-th column of matrix A, row/column vector
A^T	Transpose of matrix A
A^{-1}	Inverse of matrix A
x_i	i-th entry of vector x
f_i	i-th entry of vector-valued function f
$\lceil x \rceil$	Least integer greater than or equal to x
$\frac{\partial F}{\partial x}(x, y)$	Partial derivative of F at (x, y)
$\nabla F(x)$	Gradient of $F : \mathbb{R} \to \mathbb{R}$ at x

Black Board Symbols and Function Space

$\mathbb{N}$	Set of natural numbers excluding zero
$\mathbb{Z}$	Set of integer numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^n$	Space of n-vectors with elements from the set $\mathbb{R}$
$\mathbb{R}^{m\times n}$	Space of $m \times n$-matrices with elements from the set $\mathbb{R}$
$\mathcal{C}$	Space of continuous functions

Interval Symbols and Norm Symbols

$\mathcal{T}$	Time horizon $\mathcal{T} = [t_0, t_f] \subset \mathbb{R}$ for an ODE or OCP
t	Model or process time $t \in \mathcal{T}$
t_0, t_f	Initial/Final model or process time, start/end of time horizon $\mathcal{T}$
$\lvert \cdot \rvert$	Component-wise mapping of a real number to the absolute value
$\lVert \cdot \rVert$	The Euclidean norm of a matrix or vector

Sets

$\lvert \mathcal{X} \rvert$	Cardinality of a set $\mathcal{X}$
$\mathcal{A}(x)$	Active set at x (see Definition 8)
$\mathcal{U}$	Set of all continuous control functions
$\mathcal{X}$	Set of all differential state trajectories
$\mathrm{conv}(\mathcal{X})$	Convex hull of a set $\mathcal{X}$

Functions

$\mathcal{H}(\cdot)$	HAMILTON function
$\bar{\mathcal{H}}(\cdot)$	Augmented HAMILTON function
$\mathcal{L}(\cdot)$	LAGRANGE function
$\mathrm{sgn}(\cdot)$	Sign function
$\sigma(\cdot)$	Switching function
$\varphi(\cdot)$	BOLZA cost function
$l(\cdot)$	LAGRANGE cost function
$m(\cdot)$	MAYER cost function
$c(\cdot)$	Path constraint function

$r(\cdot)$	Endpoint constraint function
$f(\cdot)$	ODE system right hand side
$u(\cdot)$	Trajectory of continuous process controls
$w(\cdot)$	Trajectory of discrete process controls
$x(\cdot)$	Trajectory of ODE system states
$\alpha(\cdot), \gamma(\cdot)$	Trajectory of relaxed convex multipliers

Dimensions

n_c	Number of path constraints $c(\cdot)$
n_r	Number of boundary constraints $r(\cdot)$
n_u	Number of controls $u(\cdot)$
n_x	Number of differential states $x(\cdot)$

List of Figures

List of Tables