FUNDAÇÃO GETÚLIO VARGAS

MASTER THESIS

On the Shooting Algorithm for Partially Affine Control Problems

Author:
João Miguel MACHADO

Supervisor:
María Soledad ARONNA

A thesis submitted in fulfillment of the requirements for the degree of M.Sc.

July 21, 2020

roll the dice

if you're goin to try, go all the
way.
otherwise, don't even start.

if you're going to try, go all the
way.
this could mean losing girlfriends,
wives, relatives, jobs and
maybe your mind.

go all the way.
it could mean not eating for 3 or
4 days.
it could mean freezing on a
park bench.
it could mean jail,
mockery,
isolation.
isolation is the gift,
all the others are a test of your,
endurance, of
how much you really want to
do it.
and you'll do it
despite rejection and the
worst odds
and it will be better than
anything else
you can imagine.

go all the way,
there is no other feeling like
that.
you will be alone with the
gods
and the nights will flame with
fire.

do it, do it, do it.
do it.

all the way
all the way.

you will ride life straight to
perfect laughter, it's
the only good fight
there is.

Charles Bukowski

“ América latina no quiere ni tiene por qué ser un alfil sin albedrío, ni tiene nada de quimérico que sus designios de independencia y originalidad se conviertan en una aspiración occidental. No obstante, los progresos de la navegación que han reducido tantas distancias entre nuestras Américas y Europa, parecen haber aumentado en cambio nuestra distancia cultural. ¿Por qué la originalidad que se nos admite sin reservas en la literatura se nos niega con toda clase de suspicacias en nuestras tentativas tan difíciles de cambio social? ¿Por qué pensar que la justicia social que los europeos de avanzada tratan de imponer en sus países no puede ser también un objetivo latinoamericano con métodos distintos en condiciones diferentes? No: la violencia y el dolor desmesurados de nuestra historia son el resultado de injusticias seculares y amarguras sin cuento, y no una confabulación urdida a 3 mil leguas de nuestra casa. Pero muchos dirigentes y pensadores europeos lo han creído, con el infantilismo de los abuelos que olvidaron las locuras fructíferas de su juventud, como si no fuera posible otro destino que vivir a merced de los dos grandes dueños del mundo. Este es, amigos, el tamaño de nuestra soledad.
$(\cdots)$
Ante esta realidad sobrecogedora que a través de todo el tiempo humano debió de parecer una utopía, los inventores de fábulas que todo lo creemos nos sentimos con el derecho de creer que todavía no es demasiado tarde para emprender la creación de la utopía contraria. Una nueva y arrasadora utopía de la vida, donde nadie pueda decidir por otros hasta la forma de morir, donde de veras sea cierto el amor y sea posible la felicidad, y donde las estirpes condenadas a cien años de soledad tengan por fin y para siempre una segunda oportunidad sobre la tierra.”

Gabriel García Márquez - Extracted from his Nobel prize speech

FUNDAÇÃO GETÚLIO VARGAS

Abstract

EMAp - Escola de Matemática Aplicada

M.Sc.

On the Shooting Algorithm for Partially Affine Control Problems

by João Miguel MACHADO

In this thesis we propose a shooting algorithm for partially affine optimal control problems, this is, systems in which the controls appear both linearly and nonlinearly in the dynamics. Since the shooting system generally has more equations than unknowns, the algorithm relies on the Gauss-Newton method. As a consequence, the convergence is locally quadratic provided that the derivative of the shooting function is Lipschitz continuous at the optimal solution. We provide a proof of the convergence for the proposed algorithm using recently developed second order conditions for weak optimality of partially affine problems. We illustrate the applicability of the algorithm by solving an optimal treatment-vaccination epidemiological problem.

Acknowledgements

Happy families are all alike; every unhappy family is unhappy in its own way. Those were the first words from Lev Tolstoy's famous novel *Anna Karenina*, after finishing this work it seems that the inverse holds for academic research. There are only so many ways where one can do poor research, but most times they might end up resembling one another. What I have learned during the preparation of this thesis is that quality academic work is an expression of one's individuality. It is the continuous process to learn to walk by one's own feet and to think with one's own thoughts. However, as in infancy we start copying our parents actions before learning to walk by ourselves, we can not think originally without the example of others that came before.

It is only fair that I start acknowledging my parents for the stepping stones that led me here to this moment of realization. Indeed, I am quite sure that my dreams for a career in academia come from my mother, who taught me to value knowledge for knowledge's sake. From my father I got the habit of reading and thinking in a somewhat compulsive manner. Both of those traits have shaped my goals and my values in such a way that I am not sure what path in life I would have taken otherwise, but I suspect that any other path would not have given me as much satisfaction.

I can not express enough my gratitude for my advisor Maria Soledad Aronna. If I have any hope to become a successful mathematician one day I owe a lot of it to her guidance and for her example as a professional and researcher. Her supervision has not only given me the tools to develop my (still evolving) own personal way of thinking about mathematics, but also has reassured my desire to pursue this dream.

I have also met many people from EMAp over the years that have helped me through this journey. There are too many names to mention, but among all the incredibly talented people I have met, I thank my colleges Felipe Antunes, Guilherme Hossaka, Laura Sant'Anna, Gabriel Jardim, Brenda Prallon, Daniel Carletti, Davi Barreira, Pablo Aguiar and Pedro Castilho for all their support, friendship and the many mathematical discussions we shared. My deepest thanks also goes for all the professors from EMAp, for their contribution to this school and the flourishing environment they all have built, which I hope to have contributed in a constructive way. And finally my thanks to all the administrative staff from our school, without whom we could not rest easy and keep on doing mathematics.

To all my friends out of mathematics and academia, I thank the many moments we have shared and for indulging my excitement about mathematics and science in general, even when they were not as excited as me. On the contrary, I always found in them the encouragement I needed to keep on pushing and following my dreams. For any one wishing to pursue her/his dreams in life, I wish them the luck I had in finding such good companions.

Contents

Introduction

Optimization has been a central branch of applied mathematics for many years. The most general problem can be stated as follows

$$\begin{array}{ll} \text{minimize} & \phi(x) \\ \text{subject to} & x \in K. \end{array}$$

Where $\phi : X \to \mathbb{R}$ is a function often called the *cost function* defined over a Banach space X, or over some subset of such space, and $K \subset X$ is the *feasible set*, the set of points perceived as viable in some sense, by the agent willing to perform such optimization.

As examples of famous optimization problems, we can mention the classical *least squares regression* commonly used in statistics to find optimal estimators, or the *Markowitz problem* for portfolio optimization in finance. Even thought many real problems can be accurately described in such form, notice that this approach to choose an optimal strategy does not take into account the underlying dynamics of the process of interest. The area of mathematics that has had the most success in modeling the dynamical evolution of processes in time is *Differential Equations*.

It was only a matter of time that mathematicians would start to wonder if this power to describe the evolution of complex systems could be used to control their behavior with some prescribed goal in mind. Thus came the birth of *Control Theory* addressing questions such as *controllability*; what conditions should a system satisfy in order for one to be able to drag its states from a given initial condition to another arbitrary state? How can we formalize the notion of stability? Given a naturally unstable system, when one is able to act on it in order to make it stable?

There is a vast literature covering such questions, but given that one is able to control a system, how can it be controlled in an optimal manner? These are the questions that the field of *Optimal Control* is concerned with. Having such considerations in mind, an optimal control problem is not so different than a classical optimization problem. The only difference is that now we must figure out how to deal with differential equations as constraints for our optimization. In fact, the way to deal with them is not so different than what is done in the classical theory of Lagrange multipliers to solve constrained optimization problems.

The problems addressed in this thesis

In this work we propose and study the convergence of a shooting algorithm for the numerical solution of optimal control problems governed by equations of the form

$$\dot{x}(t) = f_0(x(t), u(t)) + \sum_{i=1}^{m} v_i(t) f_i(x(t), u(t)), \quad \text{a.e. on } [0, T]. \tag{1}$$

Note that when $m = 0$ then a nonlinear control system arises and when the f_i's do not depend on u, for all $i = 0, \ldots, m$, then the resulting system is (totally) control-affine. In this thesis, however, we are particularly interested in the case where both

types of control appear. This study is motivated by many models that appear in practice and the system is then *partially control-affine*. Among them we can cite the followings: the Goddard's problem proposed in [23] and analyzed in Bonnans *et al.* [36], some models for rocket motion studied in Lawden [34], Bell and Jacobson [9], Goh [24, 25], Oberle [44], Azimov [6] and Hull [32], an optimal hydrothermal electricity production problem investigated in Bortolossi *et al.* [14], a problem of atmospheric flight considered by Oberle in [42], and an optimal production process in Cho *et al.* [17] and Maurer *et al.* [38].

For optimal control problems subject to the dynamics (1), with endpoint cost and constraints, and with control constraints, we propose a shooting algorithm and show that its local convergence is guaranteed if second order sufficient optimality conditions proved in Aronna [3] hold. These second order conditions are written in terms of the second derivative of the Lagrangian function associated to the optimal control problem and are an extension of the conditions proved in Dmitruk [18] for control-affine systems. It is worth mentioning that these conditions rely on Goh transform [26]. More details, references and timeline for second order conditions for partially control-affine and control-affine problems can be found in Aronna [3] and Aronna *et al.* [4], respectively.

Shooting-like methods applied to the numerical solution of partially control-affine problems can be found in the literature. See, for instance, Oberle [43, 42] and Oberle-Taubert [45], where a generalization of the algorithm proposed by Maurer [37] for (totally) affine systems is given. These works present practical implementations of shooting-like algorithms, but they do not deal with the problem of its convergence through optimality conditions.

The organization of this thesis

This theses is organized as follows.

- Chapter I focuses on introducing the main concepts from optimal control that will be discussed along the thesis. In Section I.1 we start with motivating examples that lead to the general formulation of partially affine problems. Section I.2 is dedicated to formalize the previous discussion, giving the general problem that will be addressed and presenting a general form of the Pontryagin's Maximum Principle. We chose the chapter in Section I.3 with a general discussion some of the numerical methods found in the literature.

- Chapter II is dedicated to the optimality conditions that will be used to prove the convergence of our shooting algorithm. Section II.1 starts with the PMP in the control unconstrained case and contextualizes the assumptions we make throughout the thesis. In Section II.2 we specify the different approaches found in the literature to obtain the optimal controls as a function of the states and costates, yielding a differential algebraic system of equations as a consequence of Pontryagin's Maximum Principle. Section II.3 continues the discussion around optimality conditions, exploring second order necessary and sufficient conditions that are used in the proof of convergence for our algorithm.

- In Chapter III we formally introduce the algorithm in Section III.1, and prove its convergence in the control unconstrained case in Section III.2. Afterwards, the convergence results are extended to the controls constrained case in Section III.3.

- The goal of Chapter IV is to give a detailed discussion of the implementation of our algorithm. In Section IV.1 we discuss how symbolic computations were employed to automate the analytic calculations necessary before the more computationally demanding steps. Afterwards, in Section IV.2 we discuss some less trivial examples, one where we are able to verify sufficient conditions of optimality and later apply our numerical scheme and another example coming from the epidemiology literature.

Notations. Throughout the text we shall omit the arguments of some function h whenever the context is clear, *e.g.* the time dependence is frequently omitted. When it is a function of time and some other variable $h = h(t, x)$, the time derivative is frequently referred as $\dot{h}$. For other variables partial derivatives are referred as $D_x h$ or even h_x. The same convention is adopted for higher derivatives. By $\mathbb{R}^k$ we denote the k-dimensional real space, the space of k-dimensional column vectors with the usual euclidean norm; and by $\mathbb{R}^{k,*}$ its dual space, consisting of k-dimensional row vectors and $\mathbb{B}$ denotes the corresponding unitary ball centered at zero. By $L^p([0,T];\mathbb{R}^k)$ we mean the Lebesgue space with domain being the interval $[0,T]$ and taking values in $\mathbb{R}^k$; while $W^{q,s}([0,T];\mathbb{R}^k)$ denotes the Sobolev spaces.

I

Optimal Control: the Pontryagin's Maximum Principle and numerics

I.1 Motivating Examples

Optimal control of *ordinary differential equations* (ODE) is no more than optimization problems such that the constraints include ODEs. Even thought we can not solve the ODE analytically, we can still obtain enough information concerning the optimal solution if we learn to deal with these new types of constraints.

In order to do this, we introduce the *costates*, or *adjoint variables*, $p(\cdot)$. This newly introduced variables assume the role of multipliers corresponding to the ODE constraints. Considering a general control system of the form

$$\dot{x}(t) = f(x(t), u(t)), \quad u(t) \in U_{ad},$$

these new costate variables satisfy the *adjoint dynamics* given by

$$-\dot{p} = p \cdot D_x f(x, u). \tag{I.1}$$

The advantage of introducing these variables is that we can derive necessary conditions of optimality that will enable us to find expressions for the optimal controls in terms of the states and costates. In this Section we will not discuss these conditions formally, this is left for Section I.2. Our goal here is to give a collection of examples and discuss the different kinds of behavior that optimal solutions can present. All of the examples discussed on this thesis can be found on the link.

We start with the simplest case, where the controls appear non linearly and without constraints.

Example 1.

$$\begin{array}{ll} \textit{minimize} & \frac{1}{2}\int_0^{12} \left((x(t)-10)^2 + u(t)^2\right) dt \\ \textit{subject to} & \dot{x}(t) = u(t) - 5\sin(t),\ 0 \le t \le 12, \\ & x(0) = 5. \end{array} \tag{I.2}$$

Defining the adjoint variable p satisfying the variational system $-\dot{p} = p \cdot D_x f(x, u)$, we will see that it is easy to verify that the optimal control for this problem satisfies the simple analytical expression.

$$u^* = -p.$$

The optimal solution can be found on Figure I.1.

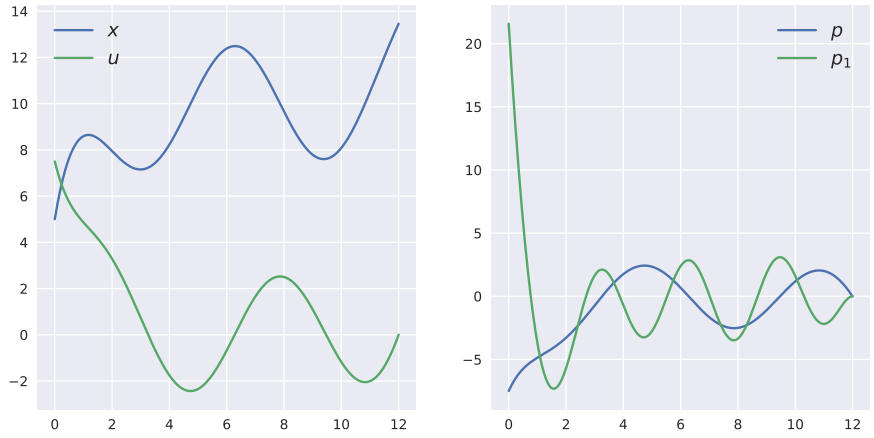

FIGURE I.1: Optimal trajectories for Example (1).

Indeed this type of systems, with controls appearing quadratically in the dynamics or in the cost function, usually are the easiest to solve. In fact it is fairly common to obtain such analytic expression for problems in this category.

In the sequel, we investigate a problem where the controls appear linearly.

Example 2.

$$\begin{aligned} &\textit{minimize} \quad r \cdot N(T) + \frac{1}{2}\int_0^T (q \cdot N(t) + su(t))\, dt \\ &\textit{subject to} \quad \begin{pmatrix} \dot{N}_1(t) \\ \dot{N}_2(t) \end{pmatrix} = \begin{pmatrix} -a_1 & 2a_2 \\ a_1 & -a_2 \end{pmatrix} \begin{pmatrix} N_1(t) \\ N_2(t) \end{pmatrix} + u(t) \begin{pmatrix} -2a_2 N_2(t) \\ 0 \end{pmatrix}, \\ &\quad N_1(0) = N_1^*, \quad N_2(0) = N2^*, \\ &\quad u(t) \in [0, u_{\max}]. \end{aligned} \tag{I.3}$$

This examples was proposed in [49] to model the effects of drug in a 2-compartment model for the growth of cancerous cells. The quantities N_2 represent the portion of the cancerous cell population that are in the reproduction stage, through *mitosis*, each give birth to two active cells that enter compartment N_1. A drug, whose dosage is represent by the control u, only affects the population of active cells N_1. The problem in controlling this system is the limit of the total amount of drug that the patient can be exposed to during the course of her/his treatment. For this reason, we introduce a maximum dosage constraint $u \leq u_{\max}$, but also include an integral term depending on the control in the cost function. This is meant to minimize long term damage to the exposition to the drug and the cost associated to the treatment.

As in example 1, defining the adjoint variables is key to characterize the optimal controls. In this example, the adjoint system becomes

$$-\dot{p} = p \cdot \begin{pmatrix} -a_1 & 2a_2 \\ a_1 & -a_2 \end{pmatrix} + q, \quad p(T) = r,$$

and the optimal controls assume the form

$$u^* = \begin{cases} u_{\max}, & s - 2a_2 p_1 < 0, \\ 0, & s - 2a_2 p_1 > 0, \end{cases}$$

what we call *bang-bang controls*, since they switch from their maximum and minimum values. This switching behavior can be observed in Figure I.2. Note that we have not specified what values the controls assume when $s - 2a_2p_1 = 0$. In fact, in [50] it was proven that this model only presents bang-bang solutions, hence this case does not play a role here.

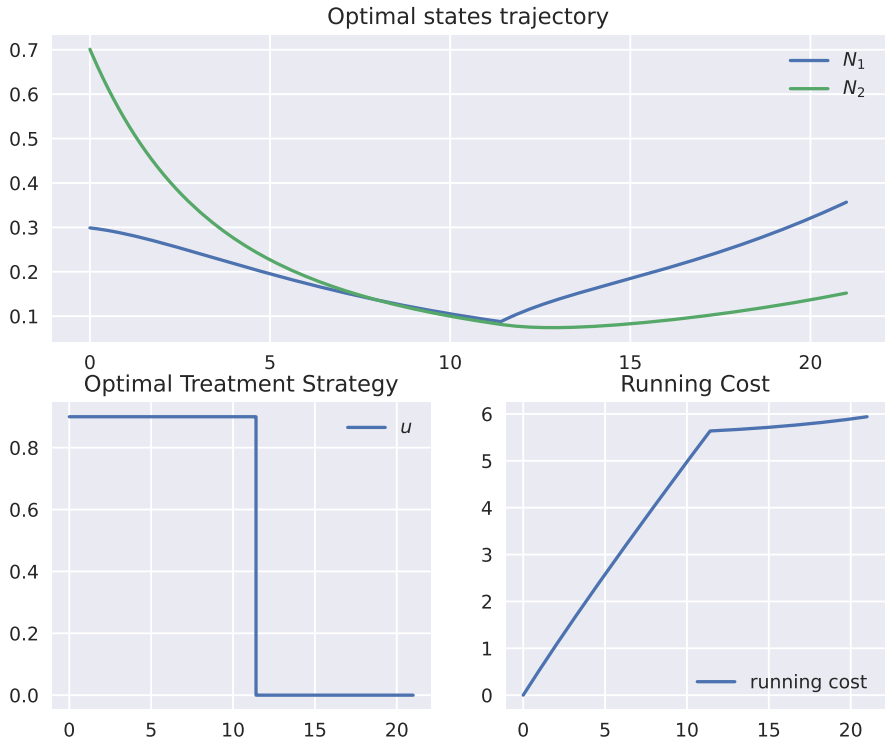

FIGURE I.2: Optimal trajectories for Example (2).

This however is not the case in our next example, known as the *singular linear quadratic regulator*, see [39].

Example 3.

$$\begin{aligned} \textit{minimize} \quad & \frac{1}{2}\int_0^5 \left(x_1(t)^2 + x_2(t)^2\right) dt \\ \textit{subject to} \quad & \dot{x}_1(t) = x_2(t), \quad x_1(0) = 0, \\ & \dot{x}_2(t) = v(t), \quad x_2(0) = 1, \\ & v(t) \in [-1, 1]. \end{aligned} \tag{I.4}$$

This example is a degenerate case of the *linear quadratic regulator* (LQR), common in the engineering literature. The difference is in the running cost function, in the LQR there should appear a term deepening on v^2, or more generally, a quadratic form v^TQv, where $Q \in \mathbb{R}^{m\times m}$ is non singular. This time, the optimal controls assume the form

$$v^*(t) = \begin{cases} 1, & p_2(t) < 0, \\ x_1(t), & p_2(t) = 0, \\ -1, & p_2(t) > 0. \end{cases}$$

When $p_2 = 0$ in a time interval with positive measure, we say that the controls present a *singular arc*. The computations leading to this characterization are less trivial when comparing to the bang-bang case. Moreover, is not easy to determine a priori when such *bang-singular structure* appears. In Figure I.3 we see an example of this behavior.

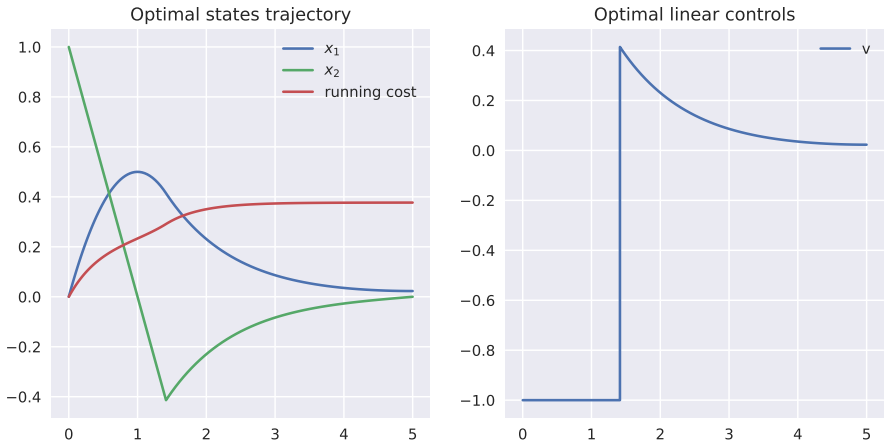

FIGURE I.3: Optimal trajectories for Example (3).

Our final example is the case when we have the presence of controls appearing both linearly and non linearly in the dynamics, what we call the *partially-affine case*.

Example 4.

$$\begin{aligned} \textit{minimize} \quad & -2x_2(2) + \int_0^2 \left(x_1(t)^2 + x_2(t)^2 + u(t)^2 + 10x_2(t)v(t)\right) dt \\ \textit{subject to} \quad & \dot{x}_1(t) = x_2(t) + u(t), \quad x_1(0) = 0, \\ & \dot{x}_2(t) = v(t), \quad x_2(0) = 0, \\ & v(t) \in [0, 0.5], \quad x_1(2) = 1. \end{aligned} \tag{I.5}$$

In Figure I.4, the behavior of the controls is similar to what we have observed in the previous examples; the nonlinear controls appear to be smooth, while the linear controls present the bang-singular structure observed in Example 3.

I.2 The Pontryagin's Maximum Principle

After the bestiary of motivating examples we have given in Section I.1, we intend to formulate a framework which encompasses all the previous examples. Consider the following general problem

$$\begin{aligned} \text{minimize} \quad & \phi(x(0), x(T)) && \text{[cost function]} \\ \text{subject to:} \quad & \dot{x}(t) = f_0(x(t), u(t)) + \sum_{i=1}^{m} v_i(t) f_i(x(t), u(t)) && \text{[dynamics]} \\ & \eta_j(x(0), x(T)) = 0, \quad j = 1, \dots, d_\eta, && \text{[equality end point constraints]} \\ & \phi_i(x(0), x(T)) \leq 0, \quad i = 1, \dots, d_\phi, && \text{[inequality end point constraints]} \\ & u(t) \in U_{ad}, && \text{[nonlinear control constraints]} \\ & a_i \leq v_i(t) \leq b_i. && \text{[linear control constraints]} \end{aligned} \tag{OC}$$

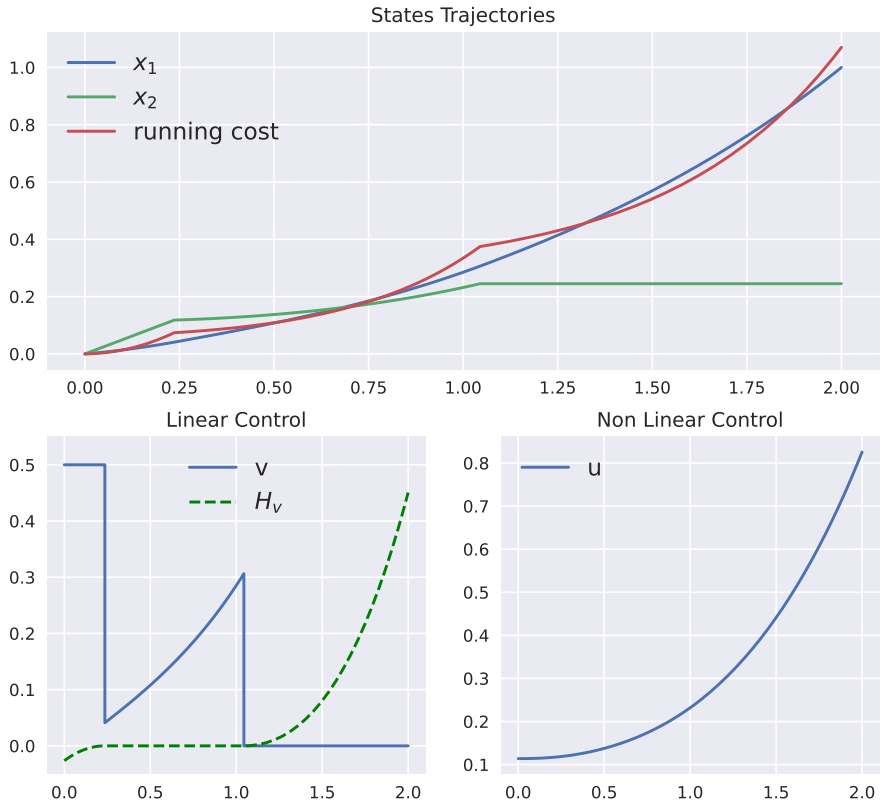

FIGURE I.4: Optimal trajectories for Example (4).

The main point of interest that we have stressed out in the examples given was concerning the type of dependence the controls presented. Notice however that problem (OC) generalizes all the such examples into this same model, that we will call *partially affine*.

If the problem at hand has only controls appearing nonlinearly, we take the vector fields $f_i = 0$; remaining only $f_0(x, u)$. On the other hand, if all the controls appear linearly, it will correspond to the case where the vector fields $f_0, f_1, \ldots, f_m$ depend only on x.

The main motivation to make this distinction is in the techniques required to find analytical expressions for the optimal controls. This will be addressed in Section II.2. For the time being, for the sake of clarity of exposition, we shall assume enough regularity for all the data functions as we conduct our computations. In Chapter II we shall formalize all the assumptions made throughout this thesis.

There are two main approaches for solving (OC), the *Pontryagin's Maximum Principle* (PMP) and Bellman's *Dynamic Programming Principle*. The former is a first order necessary condition for the optimality of a feasible state-control tuple $(\hat{x}, \hat{u}, \hat{v})$ that resembles the Lagrange multiplier rule from continuous optimization. The latter is a technique to derive a Hamilton-Jacobi equation, that is a partial differential equation to be satisfied by the *value function* associated to the problem. We will not detail this second approach any further, but the interested reader can check the monograph

[7]. In this work we focus on techniques derived from the PMP. For a proof of the Pontrygin's Principle we refer the reader to the original work from Pontryagin [48] or the more recent monographs [49, 53].

To state the PMP we will need to define some important functionals.

(i) The *non-minimized Hamiltonian function*

$$H(x,u,v,p) := p \cdot \left(f_0(x,u) + \sum_{i=1}^{m} v_i f_i,(x,u) \right); \tag{I.6}$$

(ii) The *end-point lagrangian*

$$\ell(x_0,x_T,\alpha_0,\alpha,\beta) := \alpha_0\phi(x_0,x_T) + \sum_{i=1}^{d_\phi} \alpha_i\phi_i(x_0,x_T) + \sum_{j=1}^{d_\eta} \beta_j\eta_j(x_0,x_T); \tag{I.7}$$

as well as specify the notion of optimality that will be used.

Definition I.2.1 (Weak minimum)**.** A feasible trajectory $\hat{w} = (\hat{x},\hat{u},\hat{v}) \in \mathcal{W}$ is said to be a *weak minimum* of problem (OC) if, for some real $\varepsilon > 0$, it is optimal in the set of feasible trajectories $w = (x,u,v)$ that satisfy

$$\|x-\hat{x}\|_\infty + \|u-\hat{u}\|_\infty + \|v-\hat{v}\|_\infty < \varepsilon.$$

For the reminder of this thesis we shall fix a nominal feasible trajectory $\hat{w} = (\hat{x},\hat{u},\hat{v})$ for which optimality conditions will be given and, whenever the arguments of a function are omitted, we mean that it is evaluated at such trajectory. In the sequel we state the PMP.

Theorem I.2.1. *Assume that f and ϕ are $\mathcal{C}^1$ functions, the set of admissible controls U_{ad} is a closed subset of $\mathbb{R}^l$ and let $(\hat{x},\hat{u},\hat{v})$ be a weak local minimizer for problem* (OC)*. Then there exists a function $p : [0,T] \to \mathbb{R}^{n,*}$, multipliers $\alpha \in \mathbb{R}_+^{d_\eta}$, $\beta \in \mathbb{R}^{d_\eta}$ and $\alpha_0 \in \{0,1\}$ satisfying the conditions:*

(i) the ***nontriviality condition***

$$(\alpha_0,\alpha,\beta,p(\cdot)) \neq 0; \tag{I.8}$$

(ii) the ***complementarity condition***

$$\alpha_i\phi_i(\hat{x}(0),\hat{x}(T)) = 0, \quad \text{for } i = 1\ldots,d_\phi; \tag{I.9}$$

(iii) the ***transversality conditions***

$$(-p(0),p(T)) = \nabla_{x_0,x_T}\ell(\hat{x}(0),\hat{x}(T),\alpha_0,\alpha,\beta); \tag{I.10}$$

(iv) the ***costate dynamics***

$$-\dot{p} = D_x H(\hat{x},\hat{u},\hat{v},p), \quad \text{a.e. on } [0,T]; \tag{I.11}$$

(v) *the* ***minimization of the Hamiltonian*** *condition*

$$(\hat{u}(t), \hat{v}(t)) = \underset{\substack{u \in U_{ad} \\ v_i \in [a_i, b_i]}}{\operatorname{argmin}} \; H(\hat{x}(t), u, v, p(t)), \quad \text{a.e. on } [0, T]. \tag{I.12}$$

A tuple $(\hat{x}, \hat{u}, \hat{v}, p, \alpha_0, \alpha, \beta)$ that satisfies the conditions (I.8)-(I.12) from the (PMP), is called an *extremal*. In addition, when $\alpha_0 > 0$, the extremal is called *normal*. The case when $\alpha_0 = 0$, the extremal being called *abnormal*, is pathological because the minimization of the cost becomes irrelevant, since in this case it gives no contribution to the transversality conditions. In this work we shall consider only problems that admit normal extremals.

Now let us comment on each of the conditions specified on Theorem I.2.1.

1. **Non triviality:** Notice that conditions $(ii) - (iv)$ are trivially satisfied if we allow the tuple $(\alpha_0, \alpha, \beta, p(\cdot))$ to be null. Therefore, the nontriviality condition states that there is at least one more set of multipliers that satisfy the optimality conditions, other than the trivial solution. Furthermore, for normal extremals, we can always assume $\alpha_0 = 1$ once we normalize the set of multipliers.

2. **Complementarity:** The complementarity condition plays a similar role as in the KKT theory for continuous optimization. The indexes i such that $\phi_i(x(0), x(T)) = 0$ are called active, that is, the inequality constraints that actively influence in the local optimality of a trajectory. The inactive indexes, the ones such that $\phi_i(x(0), x(T)) < 0$, do not influence in the local optimality since we can find a neighborhood of the optimal trajectory $\hat{x}$ where these constraints remain inactive, hence the local analysis remains unchanged, up to a shrinkage of the neighborhood we state that $\hat{x}$ is optimal. A direct consequence then is that for an inactive index i, one must have $\alpha_i = 0$.

3. **Minimization of the Hamiltonian:** The minimization of the Hamiltonian condition is our primary resource to compute the optimal controls. One viable strategy is to try to solve the minimization in (I.12) with the variables x, p as parameters. If we are able to obtain the minimizing controls for this problem analytically, we get a representation of the controls as a function of the states and costates. When the nonlinear controls are unconstrained, *i.e.* $U_{ad} = \mathbb{R}^l$, or the set U_{ad} is open, the minimization w.r.t. u becomes the *stationarity of the Hamiltonian*

   $$H_u(\hat{x}(t), \hat{u}(t), \hat{v}(t), p(t)) = 0,$$

   provided that the Hamiltonian is differentiable. Hence obtaining the nonlinear controls reduces to solving a nonlinear system of equations, under appropriate hypotheses that will be discussed in Section II.2. When this nonlinear dependence is quadratic, as in Examples 1 and 4, this representation is easy to obtain.

   The linear controls can be trickier to characterize. Notice that the Hamiltonian can be rewritten as

   $$H = p \cdot f_0 + \sum_{i=1}^{m} v_i H_{v_i},$$

   hence in order to minimize the Hamiltonian with respect to the control v_i, it suffices to minimize the quantity $v_i H_{v_i} = v_i p \cdot f_i$. When $H_{v_i} < 0$, the minimum is attained at the maximum value admissible for v_i; if $H_{v_i} > 0$, the minimum is

achieved at the minimum value admissible for v_i. So the optimal controls will assume the form

$$\hat{v}_i(t) = \begin{cases} b_i, & H_{v_i}(t) < 0, \\ ?, & H_{v_i}(t) = 0, \\ a_i, & H_{v_i}(t) > 0, \end{cases}$$

as we have observed in Example 2.

For this reason H_v is many times called the *switching function*, for it characterizes the times when the linear controls switch from their saturation values. However, the reader should have noticed that when the switching function becomes null, the controls are not specified by this method. We will discuss in Section III.3 that when $H_{v_i} = 0$ inside an interval of positive measure, the control v_i is said to be *singular*, or to present a *singular arc*. The way to circumvent this issue is to take time derivatives of the switching function until the dependence on the controls becomes explicit, and then solve a linear system. More details on such procedures will be addressed in Section II.2.

4. **Costate dynamics:** We have discussed that the costate variables are viewed as the Lagrange multipliers concerning the ODE constraints. On the other hand, the costate dynamics are the reason the Hamiltonian function gains its name. Notice that the pair of state and costate variables satisfy *Hamiltonian dynamics*,

$$\dot{x} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial x},$$

very celebrated in the physics literature, [1]. This is also reminiscent of the roots of optimal control theory in the calculus of variations. The latter is to this date an extremely fruitful area of mathematics, but that was initially conceived as a tool for solving problems coming from classical mechanics, [52].

Take a general control system of the form

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

its corresponding linearized dynamics of the form

$$\dot{\bar{x}} = D_x f(x, u)\bar{x} + D_u f(x, u)\bar{u}$$

has the property of transporting tangent vectors of the reachable set of the control system, [15, 49] with a careful choice of control variations $\bar{u}$.

This property is fundamental in the proof of the PMP since we can generate tangent vectors at some arbitrary time τ inside the time horizon $[0, T]$ by means of specially crafted variations of the controls and transport them with the linearized dynamics into tangent vectors to the reachable set at the final time. In the original work of Pontryagin and his group, these were given the name of *needle variations*, [48, 15]. Many other types of special variations came afterwords, see *e.g.* [49] and the references therein.

5. **Transversality:** Perhaps the transvesality conditions are the most subtle ones. To keep the discussion simpler, we will address the case with constraints only on the final time. Once we have constructed the cone of tangent vectors to the reachable set, that we will call Γ, checking the local optimility of a trajectory $\hat{x}$ reduces to checking if any of those tanget directions, $v \in \Gamma$, is such that

$\hat{x}(T) + \varepsilon v$ does not attains a smaller cost and keeps feasibility, in the sense of end-point constraints.

This is equivalent to saying that no direction $v \in \Gamma$ is in the tangent cone to the set

$$\left\{ x \in \mathbb{R}^n : \begin{array}{ll} \phi(x) \leq \phi(\hat{x}(T)), & \\ \phi_i(x) \leq 0, & \text{for } i = 1, \cdots, d_\phi \\ \eta_j(x) = 0, & \text{for } j = 1, \cdots, d_\eta \end{array} \right\},$$

which could be called the set of *profitable directions*. Under sufficient regularity conditions, or *qualification conditions* in the terminology of continuous optimization, the tangent cone at the point $\hat{x}(T)$ can be written as

$$T = \left\{ v \in \mathbb{R}^n : \begin{array}{ll} \nabla\phi(\hat{x}(T)) \cdot v \leq 0, & \\ \nabla\phi_i(\hat{x}(T)) \cdot v \leq 0, & \text{for } i = 1, \cdots, d_\phi \\ \nabla\eta_j(\hat{x}(T)) \cdot v = 0, & \text{for } j = 1, \cdots, d_\eta \end{array} \right\}.$$

This way, the essential element of the PMP becomes a theorem of separation of convex cones, in the sense that we need to separate the cones Γ and T with a hyperplane. Or equivalently, to finding a vector $p(T)$, which depends on the final time by our construction, that satisfies

$$p(T) \cdot v \leq 0, \text{ for all } v \in \Gamma, \tag{I.13}$$

$$p(T) \cdot v \geq 0, \text{ for all } v \in T. \tag{I.14}$$

Meaning, $p(T)$ is in the normal cone to the set of profitable directions.[1] Under the same qualification conditions, one can prove that the normal cone assumes the form

$$N = \left\{ \alpha_0 \nabla\phi(\hat{x}(T)) + \sum_{i=1}^{d_\phi} \alpha_i \nabla\phi_i(\hat{x}(T)) + \sum_{j=1}^{d_\eta} \beta_j \nabla\eta_j(\hat{x}(T)) : \begin{array}{c} \alpha_0 \geq 0, \\ \alpha \in \mathbb{R}_+^{d_\phi}, \\ \beta \in \mathbb{R}^{d_\eta} \end{array} \right\},$$

recovering the transversality conditions as stated.

To achieve the separation of the reachable set itself from the set of feasible directions, we need a cone with special properties that carries the information about the frontier of the reachable set, distinguishing its interior from the boundary points, points that will admit a supporting hyperplane. This is the motivation behind the definition of the *Boltyansky's approximating cone*, see the original work [48] or the discussion in [49]. After this construction, the PMP follows from classical separation theorems for convex sets.

Finally we point out that the transversality conditions state that the final costate value is a separating hyperplane for the tangent cone to the set of profitable directions and the approximating cone to the reachable set. This separation need not be unique, hence in general we can expect a set of multipliers instead of a single one.

[1] The normal cone is defined as the polar cone of the tangent cone, that is $N = T^\triangle$, where the polar of a set K is defined as

$$K^\triangle := \{v : v \cdot y \geq 0, \text{ for all } y \in K.\}$$

I.3 Numerical Methods

I.3.1 First Direct Approach

The problems in the form of (OC) can be regarded as optimization problems in functional spaces. It is possible to extend the classical KKT theory to functional spaces, however for the development suitable numerical algorithms it is necessary to discretize the domain over which our functions are defined. Hence, a first approach to solving these problems numerically would be to propose a discretization for the ODE and solve the finite dimensional *non linear program* (NLP) associated to it. This is called a *Direct Method* since it follows the philosophy of *first discretize, then optimize.*

We discretize our ODE with a general, possibly implicit, s-stage Runge-Kutta (RK) method as done in [28, 11]. For standard references on the RK methods we refer to [29, 54]. The associated NLP becomes

$$\begin{aligned}
\underset{x^0,(x_k),(x_{ki}),(u_{ki})}{\text{minimize}} \quad & \phi(x^0, x_N) \\
\text{subject to:} \quad & x_{k+1} = x_k + h_k \sum_{i=1}^{s} b_i f\left(x_{ki}, u_{ki}\right), \qquad k = 0, \dots, N-1 \\
& x_{ki} = x_k + h_k \sum_{i=1}^{s} a_{ij} f\left(x_{kj}, u_{kj}\right), \qquad \begin{matrix} k = 0, \dots, N-1 \\ i = 1, \dots, s \end{matrix} \\
& x_0 = x^0, \\
& \eta(x^0, x_N) = 0.
\end{aligned} \tag{NLP}$$

Here we are considering autonomous dynamics, the discretization for non autonomous follows in the exact same way and only adds complexity to our notation. The variables x_{ki} and u_{ki} correspond to the values of states and controls at times $t_k + c_i h_k$.

It will facilitate our calculations to introduce the quantities $K_{ki} = f\left(x_{ki}, u_{ki}\right)$. Noting that $K_{ki} = f\left(x_k + h_k \sum_{j=1}^{s} a_{ij} K_{kj}, u_{ki}\right)$, the previous system can be rewritten as

$$\begin{aligned}
\underset{x^0,(x_k),(K_{ki}),(u_{ki})}{\text{minimize}} \quad & \phi(x^0, x_N) \\
\text{subject to:} \quad & x_{k+1} = x_k + h_k \sum_{i=1}^{s} b_i K_{ki}, \qquad k = 0, \dots, N-1 \\
& K_{ki} = f\left(x_k + h_k \sum_{i=1}^{s} a_{ij} K_{kj}, u_{kj}\right), \qquad \begin{matrix} k = 0, \dots, N-1 \\ i = 1, \dots, s \end{matrix} \\
& x_0 = x^0, \\
& \eta(x^0, x_N) = 0.
\end{aligned} \tag{I.15}$$

Hence we are in a standard optimization setting and assuming some qualification conditions upon the constraints we have that local minima of problem (I.15) shall satisfy the KKT conditions [10, 41]. We define the *lagrangian function* for this problem as follows

$$\begin{aligned}
\mathcal{L}_{NLP} = \; & \beta^0 \phi(x_0, x_N) + p_0 \cdot (x_0 - x^0) + \sum_{k=0}^{N-1} \left\{ p_{k+1} \cdot \left(h_k \sum_{i=1}^{s} b_i K_{ki} + x_k - x_{k+1} \right) + \right. \\
& \left. \sum_{i=1}^{s} \psi_{ki} \cdot \left(f(x_k + h_k \sum_{\ell=1}^{s} a_{i\ell} K_{k\ell}, u_{ki}) - K_{ki} \right) \right\} + \beta \cdot \eta(x_0, x_N),
\end{aligned} \tag{I.16}$$

where the p_{k+1} are the multipliers referring to the equality constraints coming from the RK update for the discretized state x_{k+1}; ψ_{ki} refers to the equality constraints referent to the definition of K_{ki} and β is the multiplier associated to the end point constraints of the original problem. Under qualification assumptions, the KKT system that describes a minima for the discretized problem assumes the form

$$\frac{\partial \mathcal{L}_{NLP}}{\partial K_{ki}} = -\psi_{ki} + \sum_{j=1}^{s} a_{ji}\psi_{kj} \cdot D_x f\left(x_k + h_k + \sum_{\ell=1}^{s} a_{i\ell}K_{k\ell}, u_{kj}\right) + h_k b_i p_{k+1} = 0 \quad \text{(I.17)}$$

$$\frac{\partial \mathcal{L}_{NLP}}{\partial x_k} = \sum_{j=1}^{s} \psi_{kj} \cdot D_x f\left(x_k + h_k + \sum_{\ell=1}^{s} a_{i\ell}K_{k\ell}, u_{kj}\right) + p_{k+1} - p_k = 0 \quad \text{(I.18)}$$

$$\frac{\partial \mathcal{L}_{NLP}}{\partial u_{ki}} = 0 \quad \text{(I.19)}$$

Solving equation (I.17) for p_{k+1}, substituting into equation (I.18) and, provided that all the coefficients b_i are strictly positive, defining the quantities $p_{ki} = \psi_{ki}/b_i h_k$ we obtain the update scheme for thex multipliers p_k

$$\begin{aligned} p_{k+1} &= p_k - h_k \sum_{i=1}^{s} b_i p_{ki} \cdot D_x f(x_{ki}, u_{ki}), \\ p_{ki} &= p_k - h_k \sum_{j=1}^{s} \hat{a}_{ij} p_{kj} \cdot D_x f(x_{kj}, u_{kj}), \end{aligned} \qquad \text{for } k = 1, ..., N-1, \quad \text{(I.20)}$$

where the new coefficients are defined as $\hat{a}_{ij} := b_j - \frac{b_j}{b_i} a_{ji}$.

This update equation for the costates is nothing more than a Runge-Kutta discretization of the costate dynamics obtained from the PMP. Note however that the Runge-Kutta coefficients $(\hat{a}_{ij}, b_j)$ may not coincide with the coefficients used in the original Runge-Kutta discretization of the state dynamics.

Taking partial derivatives with respect to x_0, x_N, we obtain

$$\begin{aligned} p_N &= \beta^0 D_{x_T}\phi(x^0, x_N) + \beta \cdot D_{x_N}\eta(x_0, x_N), \\ -p_0 &= \beta^0 D_{x_0}\phi(x^0, x_N) + \beta \cdot D_{x_0}\eta(x_0, x_N), \end{aligned} \quad \text{(I.21)}$$

which is a discretized version of the transversality conditions of the PMP.

Finally, with (I.19), we obtain the discretized version of the stationarity of the Hamiltonian

$$H_u(x_{ki}, u_{ki}, p_{ki}) = 0, \quad \text{(I.22)}$$

note however that for simplicity of the exposition we have opted to avoid control constraints. Gathering all equations from this KKT system, we conclude that solving the discrete optimal control problem reduces to solving the following *Discretized*

Optimality System (DOS):

$$\begin{cases} x_{k+1} = x_k + h_k \sum_{i=1}^{s} b_i f\left(x_{ki}, u_{ki}\right), \\ x_{ki} = x_k + h_k \sum_{i=1}^{s} a_{ij} f\left(x_{kj}, u_{kj}\right), \\ p_{k+1} = p_k - h_k \sum_{i=1}^{s} b_i p_{ki} \cdot D_x f(x_{ki}, u_{ki}), \\ p_{ki} = p_k - h_k \sum_{j=1}^{s} \hat{a}_{ij} p_{kj} \cdot D_x f(x_{kj}, u_{kj}), \\ x_0 = x^0, \quad H_u(x_{ki}, u_{ki}, p_{ki}) = 0, \\ p_N = \beta^0 D_{x_T} \phi(x^0, x_N) + \beta \cdot D_{x_N} \eta(x_0, x_N), \\ -p_0 = \beta^0 D_{x_0} \phi(x^0, x_N) + \beta \cdot D_{x_0} \eta(x_0, x_N). \end{cases} \tag{DOS}$$

This analysis shows that in order to integrate the coupled dynamics of (x, p), one should be careful with the discretization method employed. Since the RK-coefficients for integrating the dynamics of x and p are different, it gives rise to the concept of *Partitioned Runge-Kutta* numerical integration methods, we refer to [30] for further details.

I.3.2 An Indirect Approach

We have explored the strategy of first optimize, then discretize. Now let us try the converge strategy. Given an optimal control problem, we use first order necessary conditions for optimality, *e.g.* the Pontryagin's Maximum Principle, to write a Two Point Boundary Value Problem (TPBVP) whose solution coincides, under certain conditions, with the solution for our original problem. To solve such (TPBVP) we will require some numerical integration method that will introduce a discretization scheme, hence the name *first optimize, then discretize*, or even *Indirect Methods*.

To exemplify this strategy, we recall our Example 1.

$$\begin{array}{ll} \text{minimize} & J(x,u) = \dfrac{1}{2} \displaystyle\int_0^{12} \left((x(t) - 10)^2 + u(t)^2 \right) \mathrm{d}t \\ \text{subject to} & \dot{x}(t) = u(t) - 5\sin(t), \ 0 \le t \le 12, \\ & x(0) = 5. \end{array} \tag{I.23}$$

Before going into further details of indirect methods, we need to rewrite this problem in a autonomous form and without the running cost, in order to apply the PMP as it was stated. This is easily done introducing the variables x_1, x_2, such that

$$\begin{aligned} \dot{x}_1 &= 1, \ x_1(0) = 0, \\ \dot{x}_2 &= \frac{1}{2}\left((x(t) - 10)^2 + u(t)^2 \right), \ x_2(0) = 0. \end{aligned}$$

This way, problem (I.23) becomes

$$\begin{aligned} &\text{minimize} \quad x_2(12) \\ &\text{subject to} \quad \begin{pmatrix} \dot{x} \\ \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} u - \sin(x_1) \\ 1 \\ \frac{1}{2}\left((x(t)-10)^2 + u(t)^2\right) \end{pmatrix}, \quad 0 \leq t \leq 12, \\ &\qquad\quad x(0) = 5, \; x_1(0) = x_2(0) = 0. \end{aligned} \tag{I.24}$$

The PMP for this new problem yields the following boundary value problem.

$$\begin{cases} \begin{pmatrix} \dot{x} \\ \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} u - \sin(x_1) \\ 1 \\ \frac{1}{2}\left((x(t)-10)^2 + u(t)^2\right) \end{pmatrix}, \\ \begin{pmatrix} \dot{p} \\ \dot{p}_1 \\ \dot{p}_2 \end{pmatrix} = \begin{pmatrix} p_2\,(-x+10) \\ 5p_1\cos(x_1) \\ 0 \end{pmatrix}, \\ u = -\dfrac{p}{p_2}, \quad \begin{matrix} (x(0), x_1(0), x_2(0)) & = (5,0,0) \\ (p(12), p_1(12), p_2(12)) & = (0,0,1) \end{matrix} \end{cases} \tag{I.25}$$

Some simplifications to problem (I.25) are in place. First, note that the variable p_2 has null dynamics therefore it remains constant, $p_2(t) = 1$ for all $0 \leq t \leq 12$, in virtue of the transversality condition $p_2(12) = 1$. Second, since we have constraints in the control, the minimization of the Hamiltonian condition becomes $H_u = 0$, also called the stationarity of the Hamiltonian. Hence we obtain $u = -p$ by means of a simple algebraic equation. Substituting this expression in the dynamics, our ODE depends only on (x, p).

Therefore, the difficulty os solving this problem lies in the boundary condition. Since the stationarity gives information only for the final values of the costates, we have no standard way to integrate this system directly. Fortunately, this problem is well known in the numerical analysis literature. One of the standard ways to solving it is by means of a *shooting algorithm*.

This class of algorithms are suitable to solving general *differential-algebraic systems of equations* (DAE), that is differential equations coupled with algebraic conditions. This is done defining a suitable function, taking the initial values $(p(0), p_1(0), p_2(0))$ as arguments, integrating the dynamics with these corresponding initial conditions, obtaining the values of $(x(12), x_1(12), x_2(12), p(12), p_1(12), p_2(12))$. Then, the shooting function is defined in such a way that the initial and final values of the states and costates satisfy the algebraic constraints.

For our toy problem (I.25), the initial conditions for the states will always be satisfied since they are fixed. The variable p_2 is a known constant, so we need not worry about it. Hence, a suitable shooting function can be defined as

$$(p(0), p_1(0)) \mapsto \begin{pmatrix} p(12) \\ p_1(12) \end{pmatrix}. \tag{I.26}$$

Now, solving (I.25) reduces to finding the roots of the map in (I.26). This can be done by a number of numerical methods, but we will postpone the details for Chapter IV, where we present a general form for the shooting function, suitable to solve a wider class of problems. The solution for problem (I.23) can be found in Figure I.1

I.3.3 Comparing Direct and Indirect Methods

Problem (I.23) is fairly simple when compared with other optimal control problems found in the literature and even to the scope of the problems that will be addressed in this thesis. The algorithm converged with a tolerance of 10^{-10} on the norm of the shooting function, even thought the first guess of the initial values of the costates was done very poorly, we just arbitrarily set them both to 1. However, this is not usually the case with shooting methods. Indeed, for more general problems, the shooting function becomes very sensitive to the initial conditions, in such a manner that the region of convergence can be very narrow.

This does not make direct methods objectively better. Naturally, in order to obtain a precise solution, one would need to refine the time discretization further and further, obtaining a humongous optimization problem, specially if the time horizon of the problem in question is large. Implementing an adaptive step size integration method is not trivial either, since the time discretization must be known before hand to formulate the optimization problem and adaptive step integrators compute the optimal step sizes online. In fact, this is an advantage of indirect methods over their counterparts. In the beginning of each iteration of a shooting algorithm, all the necessary information are the states and costates dynamics and the estimate for their initial conditions that satisfy the PMP. Therefore, the use of adaptive step integrators is very welcome, many times reducing considerably the computational complexity of each iteration.

For the aforementioned reasons, an approach that unifies the strengths of both methods is to obtain a rude estimate for the initial conditions using a direct method with relatively high error tolerance and refine the solution with a shooting scheme. The only concern that arises is if the operations of *discretize* and *optimize* commute in the sense that the final optimization problem tackled by the indirect method is the same as the problem solved by the direct method.

The choice of numerical integrators is crucial for such equivalence between direct and indirect methods. This choice is a challenging matter by its own, but the previous analysis from Section I.3.1 shows that when combined with questions of optimal control, it becomes even more subtle. Using the Partitioned Runge-Kutta scheme described makes the problems obtained by first discretizing and then optimizing and first optimizing and then discretizing equivalent, see Figure I.5.

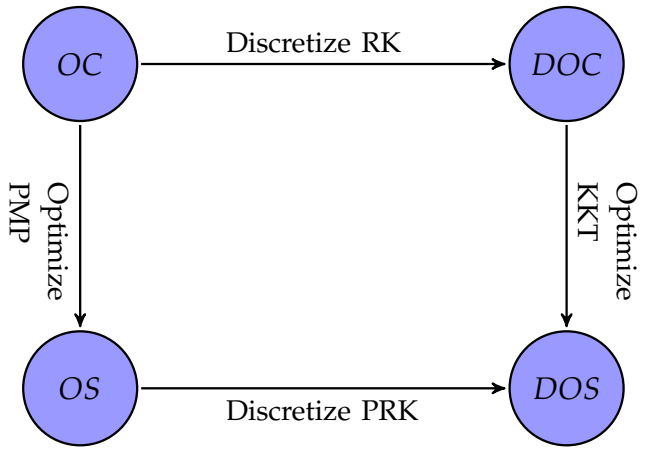

FIGURE I.5: Commutation between Discretization and Optimization steps.

So starting with a discretization for the state dynamics referent to a Butcher tableau (A, b, c), we construct a new tableau for the costate dynamics, denoted by $(\hat{A}, \hat{b}, c)$, where $\hat{b}_i = b_i, (\hat{A})_{ij} = b_j - \frac{b_j}{b_i} a_{ji}$.

Fortunately, some methods such as the Gauss collocation methods in Table I.1 are such that the tableau $(\hat{A}, \hat{b}, c)$ coincides with the original. This slightly simplifies the implementation work, since we do not need to distinguish between states and costates.

TABLE I.1: Gauss method of orders 2 (midpoint rule) and 4.

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac{1}{2} - \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4} - \frac{1}{6}\sqrt{3} \\ \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

II

Singular and Partially-Affine Problems

II.1 Problem Statement

In Section I.2 we have discussed the general case of the problem (OC) in a rather general form. In this section we introduce a simplified problem for which we will be able to prove the convergence of a shooting algorithm. The first simplification is in the controls, we will assume that the optimal solutions are contained in open subsets of the euclidean space. Regarding the end-point constraints, we drop the inequality components. The second simplification is less relevant to the scope we are working on, since in practice the only inequality constraints that are relevant are the active ones. Much like in the KKT theory for continuous optimization, we are only interested the indexes j for which the optimal state trajectory $\hat{x}$ satisfy $\phi_j(\hat{x}(0), \hat{x}(T)) = 0$. Hence using some strategy to obtain as estimate for the solution, *e.g.* a direct method as discussed in I.3, we can determine before hand which indexes are active and remove inactive constraints from the shooting scheme. We will come back to problems with control constraints in Section III.3.

Considering the function spaces $\mathcal{U} := L^\infty([0,T];\mathbb{R}^l)$, $\mathcal{V} := L^\infty([0,T];\mathbb{R}^m)$ and $\mathcal{X} := W^{1,\infty}([0,T];\mathbb{R}^n)$, we rewrite the optimal control problem (OC) as in the following form

$$\begin{aligned} &\text{minimize } \phi(x(0), x(T)) && \text{(II.1)}\\ &\text{subject to} \\ &\qquad \dot{x}(t) = f(x(t), u(t), v(t)), \ \text{ a.e. on } [0,T] && \text{(II.2)}\\ &\qquad \eta_j(x(0), x(T)) = 0, \ \text{ for } j = 1, \cdots, d_\eta. && \text{(II.3)} \end{aligned}$$

We let (OC) denote problem (II.1)-(II.3), where $\phi\colon \mathbb{R}^{2n} \to \mathbb{R}$, $\eta_j\colon \mathbb{R}^{2n} \to \mathbb{R}$, for $j = 1, \ldots, d_\eta$, and the dynamics $f : \mathbb{R}^{n+l+m} \to \mathbb{R}^n$ is of the form

$$f(x(t), u(t), v(t)) := f_0(x(t), u(t)) + \sum_{i=1}^{m} v_i(t) f_i(x(t), u(t)). \tag{II.4}$$

We make the following assumption for the aforementioned functions.

Assumption 1. *All data functions $f_0, f_1, \cdots, f_m$, η and ϕ have Lipschitz continuous second derivatives.*

A *feasible trajectory* is a tuple $w := (x, u, v) \in \mathcal{W} := \mathcal{X} \times \mathcal{U} \times \mathcal{V}$ that verifies the state dynamics (II.2) and initial-final constraints (II.3). The costates from the PMP will be elements from the space $\mathcal{X}_* := W^{1,\infty}([0,T];\mathbb{R}^{n,*})$. Given an element

$\lambda = (\beta, p) \in \mathbb{R}^{d_\eta,*} \times \mathcal{X}_*$, recall the definition of the *pre-Hamiltonian* from (I.6)

$$H[\lambda](w) := p(t) \cdot \left(f_0(x,u) + \sum_{i=1}^{m} v_i f_i(x,u) \right), \tag{II.5}$$

the *initial-final Lagrangian* from (I.7)

$$\ell[\lambda](x_0, x_T) := \phi(x_0, x_T) + \sum_{j=1}^{d_\eta} \beta_j \eta_j(x_0, x_T), \tag{II.6}$$

and finally, define the *Lagrangian* function $\mathcal{L}[\lambda] : \mathcal{W} \to \mathbb{R}$

$$\begin{aligned} \mathcal{L}[\lambda](w(\cdot)) := \ell[\lambda](x(0), x(T)) \\ + \int_0^T p(t) \cdot \left(f_0(x(t), u(t)) + \sum_{i=1}^{m} v_i(t) f_i(x(t), u(t)) - \dot{x}(t) \right) \mathrm{d}t. \end{aligned} \tag{II.7}$$

In the following analysis, we shall omit the dependence on time whenever convenient, in order to keep notation more readable.

Removing the control constraints, the minimization of the Hamiltonian conditions (I.12) assume the form of *stationarity of the Hamiltonian*. In our new setting, we use the following form of the PMP.

Theorem II.1.1 (Pontryagin's Maximum Principle)**.** *If $\hat{w} = (\hat{x}, \hat{u}, \hat{v})$ is a weak minimum of (OC), then there exists a multiplier $\lambda = (\beta, p) \in \mathbb{R}^{d_\eta,*} \times \mathcal{X}_*$, satisfying the costate dynamics:*

$$\dot{p} = -D_x H[\lambda](\hat{w}), \quad \text{a.e. on } [0,T]; \tag{II.8}$$

the transversality conditions:

$$\begin{aligned} p(0) &= -D_{x_0} \ell[\lambda](\hat{x}(0), \hat{x}(T)), \\ p(T) &= D_{x_T} \ell[\lambda](\hat{x}(0), \hat{x}(T)), \end{aligned} \tag{II.9}$$

and the stationarity of the Hamiltonian

$$D_u H[\lambda](\hat{w}) = 0 \text{ and } D_v H[\lambda](\hat{w}) = 0, \quad \text{a.e. on } [0,T]. \tag{II.10}$$

An element λ that satisfies the PMP for a trajectory $\hat{w} \in \mathcal{W}$ is called a *multiplier*. For a solution w, of (OC), we can, in general, expect a set of multipliers, instead of a single one. This is problematic for the *Shooting Algorithm* proposed later in this article, therefore we make the following assumption which guarantees the uniqueness of the Lagrange multipliers [48].

Assumption 2. *The derivative of the mapping*

$$\begin{aligned} \hat{\eta} \colon \mathbb{R}^n \times \mathcal{U} \times \mathcal{V} &\to \mathbb{R}^{d_\eta} \\ (x(0), u, v) &\mapsto \eta(x(0), x(T)) \end{aligned} \tag{II.11}$$

is onto. Here the vector x is the state of the system, given the control (u, v) and initial condition $x(0)$.

This hypothesis shall be assumed without declaration throughout the article. For now on, let $\hat{\lambda} := (\hat{\beta}, \hat{p})$ denote the unique multiplier.

II.2 The Differential-Algebraic System

The Pontryagin Maximum Principle implies that the optimal state $\hat{x}$ together with the multiplier $\hat{p}$ are solutions of a *differential-algebraic system of equations* (DAE) induced by equations (II.2), (II.3), (II.8), (II.9) and (II.10). The first step is to show that there exists a representation of the controls as a function of x and p. This can be achieved by using the stationarity of the Hamiltonian, or some of its consequences, along with a suitable strengthened version of the *Legendre-Clebsch conditions*. As a consequence, using this desired representation becomes equivalent to the algebraic conditions introduced by the stationarity condition, turning the DAE into a *two-point boundary value problem* (TPBVP).

In the following discussion we shall present a strategy to achieve this substitution of the controls that is a combination of the techniques in the particular cases when all the controls appear nonlinearly in the dynamics (*i.e.* for $m = 0$) and when the controls are *totally affine* (*i.e.* when $l = 0$). Next we give a brief review for each of these simpler cases and finally propose a technique for our mixed case.

We can trace a systematic approach that is common to all cases. It consists in using the stationarity of the Hamiltonian together with a suitable version of the Legendre-Clebsch condition to write a system of the form

$$\begin{aligned} \Theta(\xi, \alpha) &= 0, \\ D_\alpha \Theta(\xi, \alpha) &\succ 0. \end{aligned} \tag{IFTsys}$$

In this general system, α plays the role of the controls and ξ represents the tuple (x, p). After such a system with the form of (IFTsys) is assembled the *Implicit Function Theorem* (IFT) can be used to find a representation of α in terms of ξ.

Theorem II.2.1 (Implicit Function Theorem)**.** *Given Banach spaces X, Y and Z and a $\mathcal{C}^k$* [1] *mapping $\Theta\colon X \times Y \to Z$. Consider $(\bar{\xi}, \bar{\alpha}) \in X \times Y$ such that*

$$\Theta(\bar{\xi}, \bar{\alpha}) = 0 \ \text{ and } \ D_\alpha \Theta(\bar{\xi}, \bar{\alpha}) \text{ is invertible.} \tag{II.12}$$

Then, there exists an open neighborhood V of $\bar{\xi}$, and some real number $\gamma > 0$ and a $\mathcal{C}^k$ mapping $\psi\colon V \to Y$ such that

$$\Theta(\xi, \alpha) = 0, \quad \text{for } \xi \in V, \quad \|\alpha - \bar{\alpha}\|_Y \leq \gamma \tag{II.13}$$

holds if, and only if, $\alpha = \psi(\xi)$.

II.2.1 The totally nonlinear case

When we have no affine controls in the dynamics, to form a system such as (IFTsys) and satisfy the requirements of the IFT, the following assumption is necessary.

Assumption 3 (Strengthened Legendre-Clebsch condition)**.** *The second derivative of the Hamiltonian function, with respect to the control is positive definite,*

$$H_{uu}[\lambda](\hat{w}) \succ 0.$$

Under such assumption we obtain

$$\hat{u}(t) = U(\hat{x}(t), \hat{p}(t)). \tag{II.14}$$

[1] Here $\mathcal{C}^k$ in a Banach space is in the sense of Frechet differenciability.

Remark II.2.2. In fact, the second partial derivative of the Hamiltonian function with respect to the control being positive semi-definite is a necessary condition for optimality of a trajectory. Therefore, this assumption is a not too strong.

II.2.2 The totally affine case

When the controls appear linearly, we immediately face complications concerning the strengthened Legendre-Clebsch condition, since the affine dependency of the controls implies that the matrix H_{vv} is null and can not be positive definite. This way, the process for achieving a representation for the optimal control as

$$\hat{v}(t) = V(\hat{x}(t), \hat{p}(t)) \tag{II.15}$$

is not so trivial.

The solution for such problem is to turn our analysis to the time derivatives of H_v, which is usually referred as the *switching function*. In order to simplify the calculations involved in computing these derivatives, let us consider a general formula for the time derivative of a product $p \cdot F$, where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a vector field.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\hat{p} \cdot F(\hat{x})) &= \dot{\hat{p}} \cdot F(\hat{x}) + \hat{p} \cdot D_x F(\hat{x})\dot{\hat{x}} \\ &= \hat{p} \cdot \left(D_x F(\hat{x}) f_0(\hat{x}) - D_x f_0(\hat{x}) F(\hat{x}) + \sum_{i=1}^{m} \hat{v}_i \left(D_x F(\hat{x}) f_i(\hat{x}) - D_x f_i(\hat{x}) F(\hat{x}) \right) \right).\end{aligned} \tag{II.16}$$

It is advantageous to introduce the definition of *Lie brackets*. Given two differentiable vector fields $g, h : \mathbb{R}^n \to \mathbb{R}^n$ the *Lie bracket* between them is defined as

$$[g, h] := D_x h(x) g(x) - D_x g(x) h(x). \tag{II.17}$$

We use the same notation for functions depending on u and v as well, nevertheless, the derivatives are always taken w.r.t. x. Using the newly introduced notation, Equation (II.16) assumes the form

$$\frac{\mathrm{d}}{\mathrm{d}t}(\hat{p} \cdot F(\hat{x})) = \hat{p} \cdot [f_0, F] + \sum_{i=1}^{m} \hat{v}_j \hat{p} \cdot [f_i, F]. \tag{II.18}$$

We can obtain the first time derivative of H_{v_i} by choosing $F = f_i$ so that equation (II.18) becomes

$$\dot{H}_{v_j}[\hat{\lambda}](\hat{w}) = \hat{p} \cdot [f_0, f_j] + \sum_{i=1}^{m} \hat{v}_j \hat{p} \cdot [f_i, f_j]. \tag{II.19}$$

At this point, we are still not read to retrieve v since, in fact, (II.18) does not depend explicitly on the linear controls. This is a consequence of the following proposition, which is a corollary of second order necessary conditions for optimality when the set of multipliers is a singleton, as it is our case by Assumption 2.

Proposition II.2.3 (Goh conditions). *Assume that $\hat{w}(\cdot)$ is a weak minimum having an unique multiplier. Then the following identities hold*

$$\hat{p} \cdot [f_i, f_j](\cdot) \equiv 0, \quad \text{for } i, j = 1, \cdots, m.$$

Proposition II.2.3 was proposed and proved by Goh. A generalization that applies to the framework of the current paper was given by Aronna in [3, Cor. 5.2] (see also [5] and [21]). It is a direct consequence that the time derivative of the switching function (II.19) can now be expressed as

$$\dot{H}_{v_i}[\hat{\lambda}](\hat{w}) = \hat{p} \cdot [f_0, f_i], \tag{II.20}$$

showing explicitly the independence on the controls. Our hope is to turn to the second time derivative, which can be achieved once again with Equation (II.18), this time choosing $F = [f_0, f_i]$. We obtain

$$\ddot{H}_{v_i}[\hat{\lambda}](\hat{w}) = \hat{p} \cdot [f_0, [f_0, f_i]] + \sum_{j=1}^{m} \hat{v}_j \hat{p} \cdot \left[f_j, [f_0, f_i]\right]. \tag{II.21}$$

Remark II.2.4. One could easily be misled into thinking that the control shall appear explicitly in equation (II.21), as was the case with the first derivative. This is not necessarily the case, in fact, as was proved by *Kelly et al.* [33], for each control index i, the order M_i of the first derivative of the switching function which presents the control explicitly is even.

Or equivalently, for every $k \in \mathbb{N}$, the $2k$-th time derivative of the switching function is such that

$$\frac{\partial}{\partial v_i} \frac{\mathrm{d}^{2k-1}}{\mathrm{d}t^{2k-1}} H_v[\hat{\lambda}](\hat{w}) \equiv 0,$$

assuming the value zero when computed along the optimal trajectories.

The way to proceed is taking as many derivatives as necessary in order to force the explicit appearance of the linear controls. For simplicity, we shall consider that such task is achieved with only two derivatives. From the stationarity of the Hamiltonian for the linear controls, we have $\ddot{H}_v[\lambda](\hat{w}) = 0$. As a consequence, the controls can be retrieved by means of the IFT, once we assume the following *strengthened generalized Legendre-Clebsch condition*.

Assumption 4 (Strengthened Generalized Legendre-Clebsch condition).

$$-\frac{\partial \ddot{H}_v}{\partial v}[\hat{\lambda}](\hat{w}) \succ 0.$$

Once again, as in the totally nonlinear case, the positive semi-definiteness of the matrix in Assumption 4 is a necessary condition for the optimality of a feasible trajectory. Finally the controls can be retrieved assembling a system such as (IFTsys) with $\Theta = -\ddot{H}_v[\hat{\lambda}](\hat{w})$. In the sequence we substitute their expressions in the state and costate dynamics, again obtaining a TPBVP.

II.2.3 The partially-affine case

In our case of interest, we keep the same strategy as previously, attempting to form a system such as (IFTsys) in order to derive a feedback representation. The main difference now will be the appearance of terms depending on $\dot{\hat{u}}$ and $\dot{\hat{v}}$, which can be overcome with further applications of the IFT. Here, the conventional Legendre-Clebsch condition assumes the form

$$\begin{pmatrix} H_{uu}[\hat{\lambda}](\hat{w}) & H_{uv}[\hat{\lambda}](\hat{w}) \\ \\ H_{vu}[\hat{\lambda}](\hat{w}) & H_{vv}[\hat{\lambda}](\hat{w}) \end{pmatrix} \succeq 0. \tag{LC}$$

A proof of this can be found in Aronna [3, Corollary 1]. Furthermore, since $H_{vv}[\lambda](\hat{w}) = 0$, condition (LC) holds if, and only if,

$$H_{uu}[\lambda](\hat{w}) \succeq 0 \text{ and } H_{uv}[\lambda](\hat{w}) \equiv 0. \tag{II.22}$$

In order to find analogous representations to (II.14) and (II.15), we proceed in a similar manner. First, note that in the context of mixed controls, Equation (II.18) assumes the form

$$\frac{\mathrm{d}}{\mathrm{d}t}(\hat{p} \cdot F(\hat{x}, \hat{u})) = \hat{p} \cdot [f_0, F] + \sum_{i=1}^{m} \hat{v}_i \hat{p} \cdot [f_i, F] + \hat{p} \cdot D_u F \dot{\hat{u}}. \tag{II.23}$$

Once again, we obtain $\dot{H}_{v_i}$ by choosing $F = f_i$. Hence, recalling that $H_{vu} \equiv 0$ from (II.22) and the Goh conditions from Proposition II.2.3, we obtain the same expression as in the totally affine case

$$\dot{H}_{v_j}[\hat{\lambda}](\hat{w}) = \hat{p} \cdot [f_0, f_j]. \tag{II.24}$$

Differentiating the identify (II.23) once more, this time we chose $F = [f_0, f_i]$, and obtain

$$\ddot{H}_{v_i} = \hat{p} \cdot [f_0, [f_0, f_i]] + \sum_{j=1}^{m} \hat{v}_j \hat{p} \cdot \big[f_j, [f_0, f_i]\big] + \hat{p} \cdot D_u [f_0, f_i] \dot{\hat{u}}. \tag{II.25}$$

The difference from the previously discussed case is in the appearance of the term depending on $\dot{\hat{u}}$, which initially disables us from applying the IFT. To circumvent this, we use the stationarity condition for the nonlinear controls, $H_u[\lambda](\hat{w}) = 0$, to find a representation of $\dot{\hat{u}}$ in terms of the desired variables, *i.e.* x, p, u and v. Differentiating the stationarity condition formally with respect to time, and assuming enough regularity, yields

$$\dot{H}_u[\lambda](\hat{w}) = H_{ux}[\lambda](\hat{w})\dot{\hat{x}} + H_{up}[\lambda](\hat{w})\dot{\hat{p}} + H_{uu}[\lambda](\hat{w})\dot{\hat{u}} = 0. \tag{II.26}$$

where the term $H_{uv}[\hat{\lambda}]\dot{v}$ vanishes because of (II.22). To formalize (II.26), we make the following assumption on the controls.

Assumption 5 (Regularity of the controls). *The nonlinear control $\hat{u}(\cdot)$ is continuously differentiable and the linear control $\hat{v}(\cdot)$ is continuous.*

This assumption is not restrictive since it follows from the IFT, once we assume the strengthened generalized Legendre-Clebsch condition, (SLC) below. We do not require differentiability for v, since the coefficient of $\dot{v}$ vanishes. Taking $\Theta = \dot{H}_u$, in virtue of (II.26), system (IFTsys) assumes the form

$$\begin{aligned} \dot{H}_u &= 0, \\ H_{uu} &\succ 0, \end{aligned} \tag{II.27}$$

yielding the following representation of $\dot{\hat{u}}$

$$\dot{\hat{u}} = \Gamma(\hat{u}, \hat{v}, \hat{x}, \hat{p}), \tag{II.28}$$

where Γ is a $\mathcal{C}^1$ function.

Equation (II.28) shows that the dependence on $\dot{\hat{u}}$ can be dropped from (II.25). Therefore we are in position to formulate a system that can be used to achieve our

desired representation. Consider the mapping

$$(w,\lambda) \mapsto \begin{pmatrix} H_u[\lambda](w) \\ \\ -\ddot{H}_v[\lambda](w) \end{pmatrix}, \tag{II.29}$$

whose jacobian w.r.t. (u,v) at the extremal $(\hat{w},\hat{\lambda})$ is

$$\mathcal{J} := \begin{pmatrix} H_{uu}[\hat{\lambda}](\hat{w}) & H_{uv}[\hat{\lambda}](\hat{w}) \\ \\ -\dfrac{\partial \ddot{H}_v}{\partial u}[\hat{\lambda}](\hat{w}) & -\dfrac{\partial \ddot{H}_v}{\partial v}[\hat{\lambda}](\hat{w}) \end{pmatrix}. \tag{II.30}$$

To apply the IFT and retrieve the controls, we assume the following *Strengthened Generalized Legendre-Clebsch condition*

$$H_{uu}[\hat{\lambda}](\hat{w}) \succ 0, \quad -\frac{\partial \ddot{H}_v}{\partial v}[\hat{\lambda}](\hat{w}) \succ 0. \tag{SLC}$$

The next theorem discusses that we can write the controls in feedback form and solving our optimal control problem implies the solution of a TPBVP, which is often called the optimality system.

Theorem II.2.5. *Assume that SLC holds. If $\hat{w}$ is a* weak minimum *with associated multiplier $\hat{\lambda}$, then the optimal controls $(\hat{u},\hat{v})$ admit the feedback form*

$$\hat{u} = U(\hat{x},\hat{p}), \quad \hat{v} = V(\hat{x},\hat{p}), \tag{II.31}$$

where U and V are smooth functions of the states and costates.

Furthermore, the extremal $(\hat{w},\hat{\lambda})$ satisfy the optimality system

$$\begin{cases} \dot{x} = f(x,U(x,p),V(x,p)), & \text{a.e. on } [0,T], \\ \dot{p} = -p \cdot D_x f(x,U(x,p),V(x,p)), & \text{a.e. on } [0,T], \\ \eta_j(x(0),x(T)) = 0, & \text{for } j = 1,\cdots,d_\eta, \\ p(0) = -D_{x_0}\ell[\lambda](x(0),x(T)), & \\ p(T) = D_{x_T}\ell[\lambda](x(0),x(T)), & \\ H_v(x(T),U(x(T),p(T))) = 0, & \\ \dot{H}_v(x(0),U(x(0),p(0))) = 0. & \end{cases} \tag{OS}$$

Proof. We start with the feedback representation (II.31). From our previous discussion, since $H_{uu} \succ 0$, we can remove the dependence of $\hat{u}$ from $\ddot{H}_v$. Note that

$$\begin{pmatrix} H_{uu} & 0 \\ -\dfrac{\partial \ddot{H}_v}{\partial u} & -\dfrac{\partial \ddot{H}_v}{\partial v} \end{pmatrix} = \begin{pmatrix} H_{uu} & 0 \\ 0 & -\dfrac{\partial \ddot{H}_v}{\partial v} \end{pmatrix} \begin{pmatrix} I & 0 \\ \dfrac{\partial \ddot{H}_v}{\partial v}^{-1}\dfrac{\partial \ddot{H}_v}{\partial u} & I \end{pmatrix}. \tag{II.32}$$

Recalling that $H_{uv} \equiv 0$, we conclude that matrix (II.30) is invertible, since the second matrix in (II.32) is invertible from (SLC) and the third in invertible by inspection.

Representation (II.31) follows from the IFT.

Moving on to (OS), note that it is derived from the PMP. However the feedback forms II.31 are equivalent to $H_u = \dot{H}_v = 0$. To obtain the stationarity of the Hamiltonian w.r.t. v, we include the boundary conditions $H_v(T) = \dot{H}_v(0) = 0$.

We could have chosen any pair of terminal points from $H_v(0), \dot{H}_v(0), H_v(T)$ and $\dot{H}_v(T)$, including at least one of order zero. This choice will simplify the presentation of the results that follow. □

II.2.4 Computing the Linear Controls

To solve (OS), we need explicit expressions for the controls. The nonlinear controls usually can be obtained from the stationarity of the Hamiltonian, we aim to provide a practical strategy to obtain the linear controls by writing the set of equations (II.25) as a linear system. Our concern is whether the term $\hat{u}$ introduces dependencies on the linear controls. We start assuming that the representation $\hat{u} = U(\hat{x}, \hat{p})$ was already obtained.

In the sequel we introduce the *Poisson bracket notation*. Given two functions g, h that depend on x, p, the *Poisson bracket* is given by

$$\{g, h\} := D_x g D_p h - D_p g D_x h = \sum_{i=1}^{n} \left(\frac{\partial g}{\partial x_i} \frac{\partial h}{\partial p_i} - \frac{\partial g}{\partial p_i} \frac{\partial h}{\partial x_i} \right) \tag{II.33}$$

The following proposition is a direct consequence of this definition.

Proposition II.2.6. *Let $F = F(x, p, t)$ be a C^1 function. Then the following identity holds*

$$\frac{d}{dt} F(x, p, t) = \{F, H\} + \frac{\partial F}{\partial t}, \tag{II.34}$$

provided that the variables (x, p) follow Hamiltonian dynamics, i.e.

$$\dot{x} = H_p, \quad -\dot{p} = H_x. \tag{II.35}$$

As a consequence of Proposition II.2.6, if the optimal control $\hat{u}$ admits a feedback representation $\hat{u} = U(x, p)$, then

$$\dot{\hat{u}} = \{U, H\} = \{U, p \cdot f_0\} + \sum_{j=1}^{m} \hat{v}_j \{U, p \cdot f_j\}. \tag{II.36}$$

By substituting (II.36) in equation (II.25), we obtain

$$\begin{gathered} \ddot{H}_{v_i} = \gamma_{i0} + \sum_{j=1}^{m} \hat{v}_j \gamma_{ij} = 0, \\ \text{where } \gamma_{ij} := \hat{p} \cdot \left([f_j, [f_0, f_i]] + D_u [f_0, f_i] \{U, \hat{p} \cdot f_j\} \right), \end{gathered} \tag{II.37}$$

for $i, j = 1, \ldots, m$.

II.3 Second Order Optimality Conditions

Aiming for a proof of convergence for the shooting algorithm, we shall make use of second order optimality conditions. The goal of our present work is not to make an extensive study, but to briefly review the results given in Aronna [3] which are

relevant to our case of interest. However, there are practical benefits in following the arguments in a more general framework, where the Lagrange multipliers are not unique. The results given in this more general setting, when taken in the particular case where the set of multipliers is a singleton, imply the well-known Legendre-Clebsch and Goh conditions stated in Proposition II.2.3.

The optimality conditions to be presented will be in terms of the quadratic form

$$\begin{aligned}\Omega[\lambda](\bar{w}) :=& D^2\ell[\lambda](\bar{x}(0),\bar{x}(T))^2 + \int_0^T \Big(\bar{x}^T H_{xx}\bar{x} \\ &+\bar{u}^T H_{uu}\bar{u} + 2\bar{x}^T H_{ux}\bar{u} + 2\bar{x}^T H_{vx}\bar{v} + 2\bar{v}^T H_{uv}\bar{u}\Big)\,\mathrm{d}t,\end{aligned} \tag{II.38}$$

or some transformed version of this functional. A well-known result around such quadratic form, obtained by means of a second order Taylor expansion, is that

$$D^2\mathcal{L}[\lambda](\bar{w})^2 = \Omega[\lambda](\bar{w}), \tag{II.39}$$

where the derivatives are evaluated in the nominal trajectories.

We are interested in a set of critical directions for which we can extract second order conditions of optimality. Such set of directions is called the *critical cone* and can be viewed as the set of directions coming from the linearization of the DAE given by Equations (II.2),(II.3).

First let us introduce the notion of linearization of a system. A general differential-algebraic control system can be written as

$$\begin{cases}\dot{\xi}(t) = \mathcal{F}(\xi(t),\alpha(t)),\\ \quad 0 = \mathcal{G}(\xi(t),\alpha(t)),\\ \quad 0 = \mathcal{I}(\xi(0),\xi(T)),\end{cases} \tag{II.40}$$

where $\mathcal{F} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n, \mathcal{G} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{d_\mathcal{G}}$ and $\mathcal{I} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{d_\mathcal{I}}$ are $\mathcal{C}^1$ functions. The functions $\xi(\cdot)$ and $\alpha(\cdot)$ represent the state and control input of this general system. Consider $\tilde{w} = (\tilde{\xi},\tilde{\alpha})$ a solution of problem (II.40), then the *linearized system* (II.40) at $\tilde{w}$ is given by the DAE

$$\begin{cases}\dot{\bar{\xi}}(t) = D_\xi\mathcal{F}(\tilde{w})\bar{\xi} + D_\alpha\mathcal{F}(\tilde{w})\bar{\alpha},\\ \quad 0 = D_\xi\mathcal{G}(\tilde{w})\bar{\xi} + D_\alpha\mathcal{G}(\tilde{w})\bar{\alpha},\\ \quad 0 = D_{\xi_0}\mathcal{I}\left(\tilde{\xi}(0),\tilde{\xi}(T)\right)\bar{\xi}(0) + D_{\xi_T}\mathcal{I}\left(\tilde{\xi}(0),\tilde{\xi}(T)\right)\bar{\xi}(T).\end{cases} \tag{II.41}$$

Following (II.41), the linearization of the system (II.2), (II.3) is given by

$$\dot{\bar{x}} = D_x f(w)\bar{x} + D_u f(w)\bar{u} + D_v f(w)\bar{v}, \tag{II.42}$$

$$0 = D\eta(\hat{x}(0),\hat{x}(T))(\bar{x}(0),\bar{x}(T)). \tag{II.43}$$

Notice that it does not contains a running constraint, the term represented by the function $\mathcal{G}$, since we did not include such type of constraints in problem OC. However this more general case will still be relevant to us in the linearization of the DAE obtained from the PMP, for this type of running algebraic conditions appears in the form of the stationarity of the Hamiltonian, (II.10). This way, the critical cone is defined as

$$\mathcal{C} := \{\bar{w} \in \mathcal{W} : \text{(II.42) and (II.43) hold}\}. \tag{II.44}$$

Since we are interested in the second variation of the trajectories, it will be of use to

consider perturbations of the controls and states in L^2, instead of only in L^∞, as we have formulated so far. We can continously extend the quadratic mapping defined above to the function space $\mathcal{W}_2 := \mathcal{X}_2 \times \mathcal{U}_2 \times \mathcal{V}_2$, where

$$\begin{array}{rcl} \mathcal{X}_2 & := & W^{1,2}([0,T];\mathbb{R}^n), \\ \mathcal{U}_2 & := & L^2([0,T];\mathbb{R}^l), \\ \mathcal{V}_2 & := & L^2([0,T];\mathbb{R}^m). \end{array} \tag{II.45}$$

Therefore we can also extend the critical cone defined in (II.44) to the L^2 function spaces in (II.45) giving

$$\begin{array}{c} \mathcal{C}_2 := \{\bar{w} \in \mathcal{W}_2 : \text{(II.42)} - \text{(II.43)} \text{ hold}\} \\ \mathcal{C} := \mathcal{C}_2 \cap \mathcal{W}, \end{array} \tag{II.46}$$

hence $\mathcal{C} \subset \mathcal{C}_2$ and the inclusion is dense, [18].

The difficulty in obtaining sufficient conditions is in the fact that the second variation w.r.t. the linear controls vanish. Hence Ω does not have a quadratic term in $\bar{v}$ so that coercivity conditions w.r.t. $(\bar{u}, \bar{v})$ cannot hold.

II.3.1 Second Order Necessary Conditions of Optimality

A classical second order necessary condition of optimality is stated in the following theorem. The reader is referred to Milyutin *et al.* [40] (or [3]) for a proof.

Theorem II.3.1 (Classical Second Order Necessary Conditions). *Suppose $\hat{w}$ is a weak minimum of problem (OC), and let Λ be the set of Lagrange multipliers. Then it is necessary that*

$$\max_{\lambda \in \Lambda} \Omega[\lambda](\bar{w}) \geq 0, \text{ for all } \bar{w} \in \mathcal{C}. \tag{II.47}$$

This result can also be strengthened to a wider set of variations as done in [3], which translates to the following theorem.

Theorem II.3.2. *Suppose $\hat{w}$ is a weak minimum of problem (OC), denote by Λ the set of admissible Lagrange multipliers. Then it is necessary that*

$$\max_{\lambda \in \Lambda} \Omega[\lambda](\bar{w}) \geq 0, \text{ for all } \bar{w} \in \mathcal{C}_2. \tag{II.48}$$

Remark II.3.3. As discussed in [3], for the case when the set of multipliers is not a singleton, the previous Theorems II.47 and II.48 can be extended by restricting the set of multipliers to a set with more information around the nominal trajectory.

The new second order necessary condition is given as

$$\max_{\lambda \in (\mathrm{co}\Lambda)^\#} \Omega[\lambda](\bar{w}) \geq 0, \text{ for all } \bar{w} \in \mathcal{C}_2, \tag{II.49}$$

where $\mathrm{co}\Lambda$ denotes the convex hull of Λ and the set $(\mathrm{co}\Lambda)^\#$ is defined as

$$(\mathrm{co}\Lambda)^\# = \{\lambda \in \mathrm{co}\Lambda : H_{uu}[\lambda] \succeq 0 \text{ and } H_{uv}[\lambda] \equiv 0, \text{ a.e. on } [0,T]\}. \tag{II.50}$$

This strengthening should not be treated merely as a technical result. When the set of multipliers is a singleton, the condition from (II.49) implies that the set $(\mathrm{co}\Lambda)^\#$ is nonempty and in fact coincides with Λ. As a consequence, the Legendre-Clebsch condition stated in (II.22) follows. Since we assumed the uniqueness of multipliers

from the beginning, we can omit the term H_{uv} from the quadratic form Ω along weak minima.

There still remains the necessity of extending such result to sufficient conditions. This is usually done by strengthening the nonnegativity in the necessary conditions into a coercivity condition. However, this may become troublesome since the second derivative of the Hamiltonian with respect to the linear controls is identically null, so that the second variation Ω can never be coercive with respect to the linear controls. In order to overcome this problem, the *Goh transform* is employed. The latter is a change of variables introduced by Goh in [26] and applied by him, and later by several authors to derive second order necessary and sufficient conditions. We define the Goh transformation as

$$\begin{aligned} \bar{y}(t) &:= \int_0^t \bar{v}(\tau)\mathrm{d}\tau, \\ \bar{\xi}(t) &:= \bar{x}(t) - f_v(t)\bar{y}(t), \end{aligned} \qquad \text{for } t \in [0,T]. \tag{II.51}$$

One can easily check that the dynamics of the new variable $\bar{\xi}(\cdot)$ is given by

$$\dot{\bar{\xi}} = f_x\bar{\xi} + f_u\bar{u} + B\bar{y}, \quad \bar{\xi}(0) = \bar{x}(0), \tag{II.52}$$

$$\text{where } B := f_x f_v - \frac{\mathrm{d}}{\mathrm{d}t} f_v, \tag{II.53}$$

and the matrix B is well defined since the nonlinear control u must be differentiable, as stated in Assumption 5.

We are interested in how the functional Ω and the critical cone are expressed in terms of the new variables $(\bar{\xi}(\cdot), \bar{u}(\cdot), \bar{y}(\cdot))$. Starting with the transformed cones, consider a critical direction $\bar{w} \in \mathcal{C}$, note that $\bar{x}(T) = \bar{\xi}(T) + f_v(T)\bar{y}(T)$ and $\bar{x}(0) = \bar{\xi}(0)$. Therefore, the initial and final values of the states are not sufficient to characterize a critical direction, we also define $\bar{h} := \bar{y}(T)$, which appears in the transformation of the quadratic functional by means of an integration by parts. Equation (III.8) can be rewritten as

$$D\eta_j(\hat{x}(0), \hat{x}(T)) \left(\bar{\xi}(0), \bar{\xi}(T) + f_v(T)\bar{h}\right) = 0, \ \text{ for } j = 1, \cdots, d_\eta, \tag{II.54}$$

so that the critical cones $\mathcal{C}_2$ and $\mathcal{C}$ are respectively mapped into the sets

$$\mathcal{P}_2 := \left\{(\bar{\xi}(\cdot), \bar{u}(\cdot), \bar{y}(\cdot), \bar{h}) \in \mathcal{W}_2 \times \mathbb{R}^m : \bar{y}(0) = 0, \bar{y}(T) = \bar{h}, \text{(II.52) and (II.54) hold.}\right\} \tag{II.55}$$

and

$$\mathcal{P} := \mathcal{P}_2 \cap \mathcal{W}_2 \times \mathbb{R}^m. \tag{II.56}$$

As for the second variation of the Lagrangian, the quadratic functional Ω can also be written in terms of the new variables $(\bar{\xi}(\cdot), \bar{u}(\cdot), \bar{y}(\cdot), \bar{h})$. For an element $\lambda \in (\mathrm{co}\Lambda)^\#$, that is a multiplier where the terms on $H_{vu}[\lambda]$ vanish, the transformed second variation assumes the form

$$\begin{aligned} \Omega_{\mathcal{P}}[\lambda](\bar{\xi}, \bar{u}, \bar{y}, \bar{h}) := g[\lambda](\bar{\xi}(0), \bar{\xi}(T), \bar{h}) + \int_0^T \Big(&\bar{\xi}^T H_{xx}[\lambda]\bar{\xi} + 2\bar{u}^T H_{ux}[\lambda]\bar{\xi} \\ &+ 2\bar{y}^T M[\lambda]\bar{\xi} + \bar{u}^T H_{uu}[\lambda]\bar{u} + 2\bar{y}^T E[\lambda]\bar{u} + \bar{y}^T R[\lambda]\bar{y} + 2\bar{v}^T G[\lambda]\bar{y}\Big)\,\mathrm{d}t, \end{aligned} \tag{II.57}$$

where

$$M := f_v^T H_{xx} - \dot{H}_{vx} - H_{vx} f_x, \quad E := f_v^T H_{ux}^T - H_{vx} f_u, \tag{II.58}$$

$$S := \tfrac{1}{2}\left(H_{vx} f_v + (H_{vx} f_v)^T\right), \quad G := \tfrac{1}{2}\left(H_{vx} f_v - (H_{vx} f_v)^T\right), \tag{II.59}$$

$$R := f_v^T H_{xx} f_v - (H_{vx} B + (H_{vx} B)^T) - \dot{S}, \tag{II.60}$$

$$g[\lambda](\bar{\xi}_0, \bar{\xi}_T S, h) := \ell''(\xi_0, \xi_T + f_v(T)h)^2 + h^T(2H_{vx}(T)\xi_T + S(T)h). \tag{II.61}$$

For every multiplier $\lambda \in (co\Lambda)^\#$, critical variation $(\bar{x}, \bar{u}, \bar{v})$ and its respective transformed variables $(\bar{\xi}, \bar{u}, \bar{y}, \bar{y}(T), \bar{h})$, one can relate the quadratic functionals Ω and $\Omega_{\mathcal{P}}$ through integration by parts, obtaining

$$\Omega[\lambda](\bar{x}, \bar{u}, \bar{v}) = \Omega_{\mathcal{P}}[\lambda](\bar{\xi}, \bar{u}, \bar{v}, \bar{y}, \bar{y}(T), \bar{h}). \tag{II.62}$$

As a consequence of equation (II.62), Theorem II.3.1 and the equivalence of the critical cones and the transformed cones, one gets the following inequality as a necessary condition for optimality of an optimal trajectory:

$$\max_{\lambda \in (\mathrm{co}\Lambda)^\#} \Omega_{\mathcal{P}}[\lambda](\bar{\xi}, \bar{u}, \bar{v}, \bar{y}, \bar{y}(T)) \geq 0, \ on\ \mathcal{P}. \tag{II.63}$$

However, this result is not so informative since the quadratic form $\Omega_{\mathcal{P}}$ has the term $2\bar{v}^T G[\lambda]\bar{y}$, which depends on both the original linear control and the variable $\bar{y}$ obtained after Goh's transform. In order to circumvent such issue, we restrict our set of multipliers to the subset where such term $G[\lambda]$ vanishes. Define the set

$$G(\mathrm{co}\Lambda)^\# := \{\lambda \in (\mathrm{co}\Lambda)^\# : G[\lambda] = 0\}. \tag{II.64}$$

The following Theorem gives a further strengthened necessary condition in terms of this new set of multipliers.

Theorem II.3.4. *If $\hat{w}(\cdot)$ is a weak minimum of problem (OC), then*

$$\max_{\lambda \in G(co\Lambda)^\#} \Omega_{\mathcal{P}}[\lambda](\bar{\xi}, \bar{u}, \bar{y}, \bar{y}(T)) \geq 0, \ on\ \mathcal{P}. \tag{II.65}$$

Remark II.3.5. Developing the expression of the matrix G, defined in (II.59), one obtains that each element G_{ij} is given by

$$G_{ij} = -p \cdot [f_i, f_j]. \tag{II.66}$$

With analogous reasoning from Remark II.3.3, whenever the set of multipliers Λ is a singleton, we conclude that $G(\mathrm{co}\Lambda)^\#$ is not empty. In fact it is equal to Λ and $G_{ij} \equiv 0$, or equivalently the matrix $H_{vx} f_v$ is symmetric. Hence under the Assumption 2, of uniqueness of multipliers, we recover Goh's conditions stated in Proposition II.2.3.

As done for Theorems II.3.1 and II.3.2, the new necessary conditions, stated in terms of the functional $\Omega_{\mathcal{P}}$, can also be extended to take variations in $\mathcal{W}_2$. As previously, we note that the unique multiplier is in $G(\mathrm{co}\Lambda)^\#$, so that the term G vanishes and shall be omitted, as the dependence of $\Omega_{\mathcal{P}}$ with the variations $\bar{v}$. The extended quadratic form is denoted as $\Omega_{\mathcal{P}_2}$.

$$\begin{aligned}\Omega_{\mathcal{P}_2}[\lambda](\bar{\xi}, \bar{u}, \bar{y}, \bar{h}) := g[\lambda](\bar{\xi}(0), \bar{\xi}(T), \bar{h}) + \int_0^T \Big(&\bar{\xi}^T H_{xx}[\lambda]\bar{\xi} + 2\bar{u}^T H_{ux}[\lambda]\bar{\xi} \\ &+ 2\bar{y}^T M[\lambda]\bar{\xi} + \bar{u}^T H_{uu}[\lambda]\bar{u} + 2\bar{y}^T E[\lambda]\bar{u} + \bar{y}^T R[\lambda]\bar{y}\Big)\, \mathrm{d}t.\end{aligned} \tag{II.67}$$

Finally we state a version of necessary conditions which can be strenghtened to achieve sufficient conditions through the coerciveness of the quadratic functional.

Theorem II.3.6. *If $\hat{w}(\cdot)$ is a weak minimum of problem (OC), then*

$$\max_{\lambda \in G(co\Lambda)^{\#}} \Omega_{\mathcal{P}_2}[\lambda](\bar{\xi}, \bar{u}, \bar{y}, \bar{y}(T)) \geq 0, \text{ on } \mathcal{P}_2. \tag{II.68}$$

II.3.2 Second Order Sufficient Conditions of Optimality

As discussed above, the sufficient conditions are obtained by strengthening the positivity of the quadratic functional $\Omega_{\mathcal{P}_2}$. We introduce the following γ-order, which shall be used to state the sufficient conditions

$$\gamma_{\mathcal{P}}(\bar{x}(0), \bar{u}, \bar{y}, \bar{h}) := |\bar{x}(0)|^2 + |\bar{h}|^2 + \int_0^T (|\bar{u}(t)|^2 + |\bar{y}(t)|^2) \mathrm{d}t, \tag{II.69}$$

which is defined in $\mathbb{R}^n \times \mathcal{U}_2 \times \mathcal{V}_2 \times \mathbb{R}^m$. We can also express it as a function o the original variations, as a function of $(\bar{x}(0), \bar{u}, \bar{v}) \in \mathbb{R}^n \times \mathcal{U}_2 \times \mathcal{V}_2$ by setting

$$\gamma(\bar{x}(0), \bar{u}, \bar{v}) := \gamma_{\mathcal{P}}(\bar{x}(0), \bar{u}, \bar{y}, \bar{h}),$$

where $\bar{y}$ is obtained from $\bar{v}$ through Goh's transform (II.51).

Definition II.3.1 (γ-growth). We say that a trajectory $\hat{w} = (\hat{x}, \hat{u}, \hat{v})$ satisfies the γ-growth condition in the weak sense if there exist $\varepsilon, \rho > 0$ such that

$$\phi(x(0), x(T)) \geq \phi(\hat{x}(0), \hat{x}(T)) + \rho\gamma(x(0) - \hat{x}(0), u - \bar{u}, v - \hat{v}), \tag{II.70}$$

for every feasible trajectory w that verifies $\|w - \hat{w}\|_\infty < \varepsilon$.

The following theorem was proved in [3], and previously proposed by Dmitruk in [18] in the totally affine setting.

Theorem II.3.7 (Sufficient condition for weak optimality). *If for some $\rho > 0$ the quadratic functional $\Omega_{\mathcal{P}_2}$ satisfies*

$$\Omega_{\mathcal{P}_2}(\bar{\xi}, \bar{u}, \bar{y}, \bar{y}(T)) \geq \rho\gamma_{\mathcal{P}}(\bar{x}(0), \bar{u}, \bar{y}, \bar{y}(T)), \quad \text{on } \mathcal{P}_2, \tag{II.71}$$

than $\hat{w}$ is a weak minimum satisfying the γ-growth in the weak sense.

And conversely, if $\hat{w}(\cdot)$ is a weak minimum satisfying γ-growth and admitting a unique normal multiplier, then equation (II.71) is satisfied for some $\rho > 0$.

Corollary II.3.8. *If a feasible trajectory $\hat{w}$ has a unique associated multiplier and satisfies the coercivity condition (II.71) then*

$$\begin{pmatrix} H_{uu} & E^T \\ E & R \end{pmatrix} \succeq \rho I, \quad \text{a.e. on } [0, T]. \tag{II.72}$$

III

The Shooting Algorithm: Formulation and Convergence

III.1 The Shooting Algorithm

A well known method for solving TPBVPs is the *shooting algorithm*. Given an initial guess for the states and costates, the method iteratively adjusts the initial values for states and costates to verify the boundary conditions. Such approach is known in the optimal control literature as an *indirect method*, since it follows the philosophy of *first optimize, then discretize*, *i.e.* first obtain the Optimality System (OS) from the PMP, then discretize the differential equations.

However, (OS) is parametrized by the multipliers β, each resulting in a solution for the costate equation. Therefore, our goal is to find the initial values for the states and costates, as well as the multipliers that satisfy (OS).

III.1.1 The shooting function

In order to formulate the desired numerical scheme we define the *shooting function* as follows.

Definition III.1.1 (shooting function). Let $\mathcal{S} : \mathbb{R}^n \times \mathbb{R}^{n,*} \times \mathbb{R}^{d_\eta} =: D(\mathcal{S}) \to \mathbb{R}^{d_\eta} \times \mathbb{R}^{2n+2m}$ be the *shooting function* given by

$$(x_0, p_0, \beta) =: \nu \mapsto \mathcal{S}(\nu) = \begin{pmatrix} \eta(x_0, x(T)) \\ p_0 + D_{x_0}\ell[\lambda](x_0, x(T)) \\ p(T) - D_{x_T}\ell[\lambda](x_0, x(T)) \\ H_v[\lambda]\left(x(T), U(x(T), p(T))\right) \\ \dot{H}_v[\lambda]\left(x(0), U(x(0), p(0))\right) \end{pmatrix}, \tag{III.1}$$

where (x, p) is the solution of the *Initial Value Problem* (IVP)

$$\begin{aligned} \dot{x} &= H_p[\lambda](x, U(x,p), V(x,p)), \quad x(0) = x_0, \\ \dot{p} &= -H_x[\lambda](x, U(x,p), V(x,p)), \quad p(0) = p_0. \end{aligned} \tag{III.2}$$

Solving the differential-algebraic system (OS) is equivalent to finding the roots of the shooting function. With this in mind, the shooting algorithm is reduced to

finding the roots of such function. Since the number of unknowns can be smaller than the number of equations in $\mathcal{S}(\hat{v}) = 0$, the Gauss-Newton method is a suitable approach. At each step the method updates the current approximation v_k by

$$v_{k+1} \leftarrow v_k + \Delta_k, \tag{III.3}$$

where the increment Δ_k is computed by solving the linear approximation of the least squares problem

$$\min_{\Delta \in D(\mathcal{S})} \left| \mathcal{S}(v_k) + \mathcal{S}'(v_k)\Delta \right|^2. \tag{III.4}$$

The solution of the linear regression problem (III.4) is known to be

$$\Delta_k = -\left(\mathcal{S}'(v_k)^T \mathcal{S}'(v_k) \right)^{-1} \mathcal{S}'(v_k)^T \mathcal{S}(v_k), \tag{III.5}$$

provided that the matrix $\mathcal{S}'(v_k)^T \mathcal{S}'(v_k)$ is non-singular.

One can prove that the Gauss-Newton method converges at least linearly, as long as the derivative $\mathcal{S}'(\hat{v})$ exists and is injective. If in addition it is also Lipschitz continuous, the method converges locally quadratically. The reader can check Appendix A, or the works by Fletcher [20], or alternatively Bonnans [13], for more information on the Gauss-Newton method.

III.1.2 Computation of the derivative of the shooting function

In this paragraph we aim at obtaining a differential system to be used afterwards to compute the derivative of the shooting function. Recall the general DAE system from (II.40)

$$\begin{cases} \dot{\xi}(t) = \mathcal{F}(\xi(t), \alpha(t)), \\ 0 = \mathcal{G}(\xi(t), \alpha(t)), \\ 0 = \mathcal{I}(\xi(0), \xi(T)), \end{cases}$$

and its respective linearization with respect to some solution $\tilde{w} = (\tilde{\xi}, \tilde{\alpha})$, equation (II.41),

$$\begin{cases} \dot{\bar{\xi}}(t) = D_\xi \mathcal{F}(\tilde{w})\bar{\xi} + D_\alpha \mathcal{F}(\tilde{w})\bar{\alpha}, \\ 0 = D_\xi \mathcal{G}(\tilde{w})\bar{\xi} + D_\alpha \mathcal{G}(\tilde{w})\bar{\alpha}, \\ 0 = D_{\xi_0} \mathcal{I}\left(\tilde{\xi}(0), \tilde{\xi}(T)\right) \bar{\xi}(0) + D_{\xi_T} \mathcal{I}\left(\tilde{\xi}(0), \tilde{\xi}(T)\right) \bar{\xi}(T). \end{cases}$$

In order to write the linearization of (OS), we consider the generalized state ξ to be (x, p), α to be (u, v) and $w = (\xi, \alpha)$. The linearized state and costate dynamics can be written in terms of the variables $(\bar{x}, \bar{u}, \bar{v}, \bar{p})$ as

$$\dot{\bar{x}} = D_x f(w)\bar{x} + D_u f(w)\bar{u} + D_v f(w)\bar{v}, \tag{III.6}$$

$$\dot{\bar{p}} = -\left(\bar{p} H_{xp} + \bar{x}^T H_{xx} + \bar{u}^T H_{ux} + \bar{v}^T H_{vx} \right). \tag{III.7}$$

The end point conditions are also easily linearized, giving

$$0 = D\eta(\hat{x}(0), \hat{x}(T))(\bar{x}(0), \bar{x}(T)), \tag{III.8}$$

$$\bar{p}(0) = -\left(\bar{x}^T(0)D^2_{x_0}\ell[\hat{\lambda}](\hat{w}) + \bar{x}^T(T)D^2_{x_0x_T}\ell[\hat{\lambda}](\hat{w}) + \sum_{j=1}^{d_\eta}\beta_j D_{x_0}\eta_j\right), \tag{III.9}$$

$$\bar{p}(T) = \left(\bar{x}^T(T)D^2_{x_T}\ell[\hat{\lambda}](\hat{w}) + \bar{x}^T(T)D^2_{x_0x_T}\ell[\hat{\lambda}](\hat{w}) + \sum_{j=1}^{d_\eta}\beta_j D_{x_T}\eta_j\right). \tag{III.10}$$

Finally, the algebraic conditions can be linearized using the following lemma.

Lemma III.1.1. *For some sufficiently smooth function $\mathcal{F}$ commutation of the operations of linearization and differentiation holds, that is*

$$\frac{d}{dt}\mathrm{Lin}\mathcal{F} = \mathrm{Lin}\frac{d}{dt}\mathcal{F}. \tag{III.11}$$

Therefore, it suffices for us to compute the linearization of the switching function and perform successive differentiations, in order to achieve explicit expressions for the last components of (III.1.1). The linearization gives

$$\mathrm{Lin}H_u = \bar{p}D_u f + \bar{x}^T H^T_{ux} + \bar{u}^T H_{uu} \tag{III.12}$$

$$\mathrm{Lin}\ddot{H}_v = \bar{p}D_v f + \bar{x}^T H^T_{vx} \tag{III.13}$$

$$\mathrm{Lin}H_v|_{t=T} = \bar{p}D_v f + \bar{x}^T H^T_{vx}\Big|_{t=T} \tag{III.14}$$

$$\mathrm{Lin}\dot{H}_v|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\left(\bar{p}D_v f + \bar{x}^T H^T_{vx}\right). \tag{III.15}$$

Once the linearized system is computed, we can evaluate the derivative of the shooting function. Such derivative, evaluated in the direction $\bar{v} := (\bar{x}_0, \bar{p}_0, \bar{\beta})$ is given by

$$\mathcal{S}'(\hat{v})\bar{v} := \begin{pmatrix} D\eta(\hat{x}(0), \hat{x}(T))(\bar{x}(0), \bar{x}(T)) \\ \bar{p}(0) + \left[\bar{x}^T(0)D^2_{x_0}\ell + \bar{x}^T(T)D^2_{x_0x_T}\ell + \sum_{j=1}^{d_\eta}\bar{\beta}_j D_{x_0}\eta_j\right] \\ \bar{p}(T) - \left[\bar{x}^T(T)D^2_{x_T}\ell + \bar{x}^T(T)D^2_{x_0x_T}\ell + \sum_{j=1}^{d_\eta}\bar{\beta}_j D_{x_T}\eta_j\right] \\ \bar{p}D_v f + \bar{x}^T H^T_{vx}\Big|_{t=T} \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\bar{p}D_v f + \bar{x}^T H^T_{vx}\right)\Big|_{t=0} \end{pmatrix}, \tag{III.16}$$

where the derivatives are taken along $\hat{w}$ and $\hat{\lambda}$. Note that for each iteration it is necessary to numerically integrate both the original system of states and costates and the associated variational dynamics (i.e. the linearized system).
Notation: The lineariazation (III.6)-(III.10), (III.12)-(III.15) is referred as (LS).

III.2 Convergence of the unconstrained case

Now we turn to the proof of convergence for the shooting scheme proposed. For this we shall use the quadratic functional defined in (II.67) to formulate an auxiliary linear quadratic system.

III.2.1 The auxiliary linear quadratic problem

Let (LQ) denote the optimal control problem defined by (III.17)-(III.20) below

$$\text{minimize } \Omega_{\mathcal{P}_2}(\bar{\xi}, \bar{u}, \bar{y}, \bar{h}) \tag{III.17}$$

subject to

$$\dot{\bar{\xi}} = f_x\bar{\xi} + f_u\bar{u} + B\bar{y}, \tag{III.18}$$

$$\dot{\bar{h}} = 0, \tag{III.19}$$

$$0 = D\eta_j(\hat{x}(0), \hat{x}(T))\left(\bar{\xi}(0), \bar{\xi}(T) + f_v(T)h\right), \tag{III.20}$$

where $\bar{u}$ and $\bar{y}$ denote the control variables, $\bar{\xi}$ and $\bar{h}$, the states of our auxiliary system. Note that the feasible trajectories of (LQ) are the critical directions from $\mathcal{P}_2$. Once the coercivity condition (II.71) is assumed the unique optimal solution is $(\bar{\xi}, \bar{u}, \bar{y}, h) = 0$. Our strategy will be to exploit (LQ)'s corresponding differential-algebraic system, obtained by applying the PMP, along with the sufficient condition for the original problem discussed in Theorem II.3.7.

Let $\bar{\chi}$ and $\bar{\chi}_h$ denote the costates associated with $\bar{\xi}$ and $\bar{h}$, respectively. The qualification condition for the original problem given in Assumption 2 easily translates into the same constraints qualification of problem (LQ). Therefore, the weak minimizer of the auxiliary problem, $(\bar{\xi}, \bar{u}, \bar{y}, \bar{h}) = 0$, also has a unique multiplier, which we shall refer as $\lambda^{LQ} := \left(\bar{\chi}, \bar{\chi}_h, \beta^{LQ}\right)$.

We proceed with the formulation of the auxiliary differential-algebraic system. Define the pre Hamiltonian for problem (LQ)

$$\begin{aligned}\mathcal{H}[\lambda^{LQ}](\bar{\xi}, \bar{u}, \bar{y}) := \bar{\chi}(f_x\bar{\xi} + f_u\bar{u} + B\bar{y}) + \tfrac{1}{2}\bar{\xi}^T H_{xx}\bar{\xi} \\ + \bar{u}^T H_{ux}\bar{\xi} + \bar{y}^T M\bar{\xi} + \tfrac{1}{2}\bar{u}^T H_{uu}\bar{u} + \bar{y}^T E\bar{u} + \tfrac{1}{2}\bar{y}^T R\bar{y},\end{aligned} \tag{III.21}$$

as for the endpoint Lagragian

$$\begin{aligned}\ell^{LQ}[\lambda^{LQ}](\bar{\xi}(0), \bar{\xi}(T), \bar{h}) := \tfrac{1}{2}g(\bar{\xi}(0), \bar{\xi}(T), \bar{h}) \\ + \sum_{j=1}^{d_\eta} \beta_j^{LQ} D\eta_j\left(\bar{\xi}(0), \bar{\xi}(T) + f_v(T)\bar{h}\right),\end{aligned} \tag{III.22}$$

where g was defined in (II.61).The costate dynamics becomes

$$-\dot{\bar{\chi}} = \frac{\partial \mathcal{H}}{\partial \bar{\xi}}[\lambda^{LQ}] = \bar{\chi}f_x + \bar{\xi}^T H_{xx} + \bar{u}^T H_{ux} + \bar{y}^T M, \tag{III.23}$$

with transversality conditions

$$\bar{\chi}(0) = -\bar{\xi}^T(0)D^2_{x_0^2}\ell + (\bar{\xi}(T) + f_v(T)\bar{h})^T D^2_{x_0x_T}\ell + \sum_{j=1}^{d_\eta} D_{x_0}\eta_j, \tag{III.24}$$

$$\bar{\chi}(T) = \bar{\xi}^T(T)D^2_{x_T^2}\ell + \bar{\xi}^T(0)D^2_{x_0x_T}\ell + \bar{h}^T H_{vx}(T) + \sum_{j=1}^{d_\eta} D_{x_T}\eta_j. \tag{III.25}$$

The costate variable $\bar{\chi}_h$ is trivial and vanishes identically since $\dot{\bar{\chi}}_h = 0$ and $\bar{\chi}_h(0) = 0$. Finally, the stationarity of the Hamiltonian gives

$$0 = \mathcal{H}_{\bar{u}} = \bar{\chi}f_u + \bar{\xi}^T H_{xu}^T + \bar{u}^T H_{uu} + \bar{y}^T E, \tag{III.26}$$

$$0 = \mathcal{H}_{\bar{y}} = \bar{\chi}B + \bar{\xi}^T M^T + \bar{u}^T E^T + \bar{y}^T R. \tag{III.27}$$

The set of equations (III.18)-(III.20), (III.23)-(III.25) and (III.26)-(III.27) will be referred as (LQS), the Linear Quadratic System. Notice that for this system, the Legendre-Clebsch conditions translate to

$$D^2_{(\bar{u},\bar{y})}\mathcal{H} = D^2_{(\bar{u},\bar{y})}\Omega_{\mathcal{P}_2} = \begin{pmatrix} H_{uu} & E^T \\ E & R \end{pmatrix} \succeq 0, \tag{III.28}$$

hence if we assume coercivity for the original problem, Corollary II.3.8 implies that solving (LQ) is equivalent to solve (LQS) as in Theorem II.2.5.

III.2.2 Linking the auxiliary problem with the optimality system

In this section we recall the linearization (LS) of the optimality system (OS) along the nominal trajectory $\hat{w} = (\hat{x}, \hat{u}, \hat{v})$. Define the mapping

$$(\bar{x}, \bar{u}, \bar{v}, \bar{p}, \beta) \mapsto \left(\bar{\xi}, \bar{u}, \bar{y}, \bar{h}, \bar{\chi}, \bar{\chi}_h, \beta^{LQ}\right) \tag{III.29}$$

throught the equations

$$\begin{aligned} \bar{y}(t) := \int_0^t \bar{v}(s)\mathrm{d}s, \quad \bar{\xi} := \bar{x} - f_v\bar{y}, \quad \bar{\chi} := \bar{p} + \bar{y}^T H_{vx}, \\ \bar{\chi}_h := 0, \qquad \bar{h} := \bar{y}(T), \qquad \beta^{LQ} = \beta. \end{aligned} \tag{III.30}$$

This Goh-type transformation is clearly one-to-one. We will show that it maps solutions of (LS) into solutions of (LQS) and use the coercivity condition to obtain information about (LS). We also obtain information on the solution of (OC), since coercivity is a sufficient condition for optimality in Theorem II.3.7. We start with the following lemma.

Lemma III.2.1. *If $\hat{w}$ is a weak solution of* (OC)*, the injective mapping $(\bar{x}, \bar{u}, \bar{v}, \bar{p}, \beta) \mapsto (\bar{\xi}, \bar{u}, \bar{y}, \bar{h}, \bar{\chi}, \bar{\chi}_h, \beta^{LQ})$ defined in* (III.30) *converts solutions of* (LS) *into solutions of* (LQS).

Proof. We must check that given a solution $(\bar{x}, \bar{u}, \bar{v}, \bar{p}, \beta)$ of (LS), the corresponding transformed variables $(\bar{\xi}, \bar{u}, \bar{y}, \bar{h}, \bar{\chi}, \bar{\chi}_h, \beta^{LQ})$ solve (LQS).

Starting with the state $\bar{\xi}$, we recall the dynamics of the linearized variable $\bar{x}$ given in equation (III.6) so that one has

$$\dot{\bar{\xi}} = \dot{\bar{x}} - \dot{f}_v\bar{y} - f_v\dot{\bar{y}} = f_x\bar{\xi} + f_u\bar{u} + B\bar{y},$$

retrieving the dynamics in (III.18). The initial conditions are trivially satisfied since $\bar{y}(0) = 0$. The dynamics for $\bar{h}$ are satisfied by the definition.

For the costate dynamics we recall the dynamics of the linearized costates from (III.7) and the definition of the matrix M in (II.58).

$$\begin{aligned} -\dot{\bar{\chi}} &= -\dot{\bar{p}} - \dot{\bar{y}}^T H_{vx} - \bar{y}^T\dot{H}_{vx} \\ &= \bar{p}H_{xp} + \bar{x}^T H_{xx} + \bar{u}^T H_{ux} + \bar{v}^T H_{vx} - \bar{v}^T H_{vx} - \bar{y}^T\dot{H}_{vx} \\ &= \underbrace{(\bar{p} + \bar{y}^T H_{vx})}_{=\bar{\chi}} f_x + \underbrace{(\bar{x} - f_v\bar{y})^T}_{=\bar{\xi}^T} H_{xx} + \bar{y}\underbrace{(f_v^T H_{xx} - \dot{H}_{vx} - H_{vx}f_x)}_{=M} \\ &= \bar{\chi}f_x + \bar{\xi}^T H_{xx} + \bar{y}^T M. \end{aligned}$$

Hence the dynamics of $\bar{\chi}$ match (III.23). From equation (III.30) we obtain $\bar{\chi}(0) = \bar{p}(0)$ and conclude (III.24). The final conditions are trickier, one substitutes expressions for $\bar{x}(T)$ and $\bar{p}(T)$ into equation (III.14) and conclude since

$$S = H_{vx} f_v = f_v^T H_{vx}^T,$$

which is a consequence of the Goh conditions (II.2.3), recovering the transversality condition for $\bar{\chi}(T)$.

Finally we must check the stationarity of the Hamiltonian for (LQS), equations (III.26) and (III.27). Starting from (III.12) and (III.30) we obtain

$$\begin{aligned} 0 &= (\bar{\chi} - \bar{y}^T H_{vx}) f_u + (\bar{\xi} + f_v \bar{y})^T H_{ux}^T + \bar{u}^T H_{uu} \\ &= \bar{\chi} f_u + \bar{\xi}^T H_{ux}^T + \bar{u}^T H_{uu} + \bar{y}^T (\underbrace{f_v^T H_{ux}^T - H_{vx} f_u}_{=E}), \\ &= \bar{\chi} f_u + \bar{\xi}^T H_{ux}^T + \bar{u}^T H_{uu} + \bar{y}^T E, \end{aligned}$$

the stationarity with respect to $\bar{u}$. On the other hand, the same substitutions applied to (III.13) yield

$$0 = \bar{\chi} f_v + \bar{\xi}^T H_{vx}^T.$$

Differentiating with respect to time and using the definitions of B in (II.53) and E in (II.58), we recover the stationarity (III.27) with respect to $\bar{y}$.

This shows that the tuple $(\bar{\xi}, \bar{u}, \bar{y}, \bar{h}, \bar{\chi}, \bar{\chi}_h, \beta^{LQ})$ is a solution of (LQS). □

III.2.3 Convergence of the shooting algorithm

We are in position to prove the convergence of the shooting algorithm. We will use the following result on the behavior of the Gauss-Newton algorithm. Recall the algorithm given in (III.3)-(III.5).

Proposition III.2.2 ([13, 20]). *If the matrix $\mathcal{S}'(\hat{v})$ is injective, then the Gauss-Newton algorithm,* (III.3)-(III.5)*, is locally convergent. If in addition $\mathcal{S}'(\hat{v})$ is Lipschitz continuous, the algorithm converges locally quadratically.*

The main result of this article is the theorem below that states a sufficient condition for the quadratic convergence of the shooting algorithm near a local optimal solution.

Theorem III.2.3. *Let $\hat{w}$ be a trajectory of problem (OC) that verifies condition* (II.71) *and the Legendre-Clebsch conditions* (SLC)*. Then the shooting algorithm is locally quadratically convergent.*

Proof. From Theorem II.3.7, the trajectory $\hat{w}$ is a local weak minimum for problem (OC). In addition, since (SLC) hold, we can use Theorem II.2.5 and solve (OS) to find $\hat{w}$. Therefore, consider some solution $(\bar{x}, \bar{u}, \bar{v}, \bar{p}, \beta)$ of (LS), and the associated transformed process $(\bar{\xi}, \bar{u}, \bar{y}, \bar{h}, \bar{\chi}, \bar{\chi}_h, \beta^{LQ})$ given by (III.30). The latter is a solution of (LQS) by Lemma III.2.1.

However, once we assume condition (II.71) the unique solution to (LQS) is the null trajectory, and since the transformation (III.30) is one-to-one, the solution to (LS) is also null. But from equation (III.16), the vectors $\bar{v}$ in the kernel of $\mathcal{S}'(\hat{v})$ are precisely the initial conditions for the solutions of (LS). We conclude that $\mathcal{S}'(\hat{v})$ is injective, in addition, it is Lipschitz continuous from Assumption 1. The proof follows from III.2.2. □

The proof of Theorem (III.2.3) is summarized in the following schematic in Figure III.1.

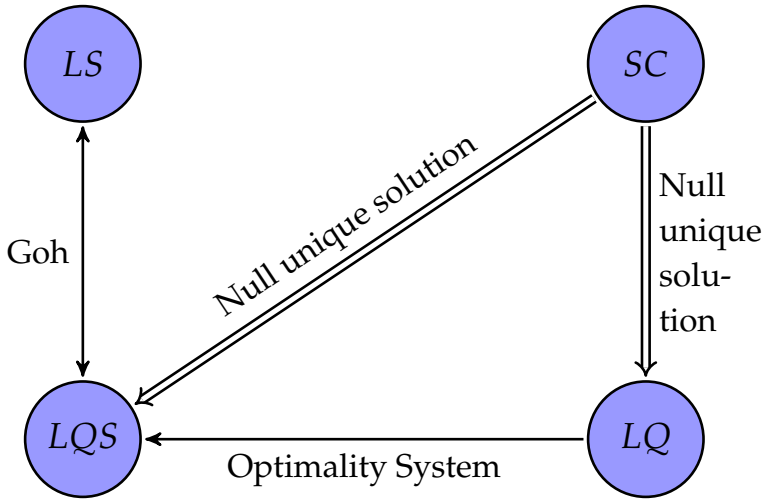

FIGURE III.1: Schematic proof of Theorem III.2.3. The sufficient conditions (SC) from Theorem II.3.7 imply that LQ and LQS have a null unique solution. From Lemma III.2.1, the solutions of LS and LQS are obtained one from the other through the Goh transform, hence LS admits only the null solution. Since these initial conditions for solutions of LS spam the kernel of $\mathcal{S}'(\hat{v})$, this kernel is trivial and $\mathcal{S}'(\hat{v})$ is injective.

III.3 Including control constraints

In this section we extend the proposed algorithm to problems where the controls are subject to bounds. We denote by (CP) the problem obtained by adding the following control constraints to (OC):

$$\begin{array}{ll} u(t) \in U, & \text{a.e. on } [0,T], \\ 0 \leq v_i(t) \leq 1, & \text{a.e. on } [0,T], \text{ for } i = 1, \cdots, m, \end{array} \tag{III.31}$$

where U is an open subset of $\mathbb{R}^l$. Consider also the following definition.

Definition III.3.1. The component $\hat{v}_i$ is said to have a *singular arc* in an interval I, whenever $0 < \hat{v}_i(t) < 1$ a.e. on I. On the other hand, a component $\hat{v}_i$ has an *upper bang arc* (resp. *lower bang arc*) on an interval I whenever $\hat{v}_i(t) = 1$(resp. $\hat{v}_i(t) = 0$) a.e. on this interval, and has a *lower bang arc* whenever $\hat{v}_i(t) = 0$ a.e. on such interval. If $\hat{v}_i$ has either an upper or a lower bang arc on I then we can say, shortly, that it has a *bang arc* on I.

Assumption 6. *We assume the following hypotheses on the optimal* $(\hat{u}, \hat{v})$.

(i) Each linear control $\hat{v}_i$*, with* $i = 1, \ldots, m$*, presents a* bang-singular structure*, i.e.* $\hat{v}_i$ *is a finite concatenation of* bang and singular arcs.

(ii) The bang-singular structure of $\hat{v}$ *induces a partition of the time interval* $[0, T]$*, that we write as*

$$\{0 := \hat{T}_0 < \hat{T}_1 < \hat{T}_2 < \cdots < \hat{T}_{N-1} < \hat{T}_N := T\}.$$

At each interval $\hat{I}_k := [\hat{T}_k, \hat{T}_{k+1}]$, *every component* $\hat{v}_i$ *is either bang or singular, and at* $\hat{T}_k$ *some control* $\hat{v}_i$ *switches from its current arc, and presents a discontinuity of first kind. Hence, defining the sets*

$$\begin{aligned} S_k &:= \{1 \leq i \leq m : \hat{v}_i \text{ is singular on } \hat{I}_k\}, \\ A_k &:= \{1 \leq i \leq m : \hat{v}_i = 0 \text{ a.e. on } \hat{I}_k\}, \\ B_k &:= \{1 \leq i \leq m : \hat{v}_i = 1 \text{ a.e. on } \hat{I}_k\}, \end{aligned}$$

there must exist some $\rho > 0$ *such that*

$$\rho < \hat{v}_i(t) < 1 - \rho, \quad \text{for all } i \in S_k, \text{ a.e. on } t \in \hat{I}_k. \tag{III.32}$$

In addition, we assume that the nonlinear control satisfies

$$\hat{u}([0,T]) + \rho\mathbb{B} \subset U. \tag{III.33}$$

(iii) For each $k = 1, \ldots, N$, *let* v_{S_k} *denote the vector with components* v_i *with* $i \in S_k$. *To obtain a feedback representation in a similar manner as done in Section* **??**, *we assume*

$$\begin{aligned} &\hat{u} \text{ is continuously differentiable in } [0,T], \\ &\hat{v}_{S_k} \text{ is continuous in } \hat{I}_k, \text{for } k = 1, \ldots, N. \end{aligned} \tag{III.34}$$

In addition, on each interval $\hat{I}_k$ *we assume the following form of the generalized strengthened Legendre-Clebsch conditions*

$$H_{uu}[\hat{\lambda}](\hat{w}) \succ 0, \quad -\frac{\partial \ddot{H}_{v_{S_k}}}{\partial v_{S_k}}[\hat{\lambda}](\hat{w}) \succ 0.$$

As a consequence of the minimization of the Hamiltonian given by the PMP, if a component v_i is singular in some interval I, then $H_{v_i}(t) = 0$ a.e. on I. Hence, as done in Section II.2, we can use the system

$$\begin{pmatrix} H_u[\hat{\lambda}](\hat{w}) \\ -\ddot{H}_{v_{S_k}}[\hat{\lambda}](\hat{w}) \end{pmatrix} = 0, \quad \text{a.e. on } \hat{I}_k, \tag{III.35}$$

along with item (iii) from Assumption 6 to write the controls $\hat{u}$ and $\hat{v}_{S_k}$ in feedback form, which we represent as

$$\hat{u} = U(\hat{x}, \hat{p}), \quad \hat{v}_{S_k} = V_{S_k}(\hat{x}, \hat{p}). \tag{III.36}$$

III.3.1 The transformed problem

Given a feasible control $(\hat{u}, \hat{v})$, we call *control structure* the configuration of bang and singular arcs of $\hat{v}$. In (CP), there may be feasible trajectories with a bang-singular structure different from the one of $(\hat{u}, \hat{v})$. However, if $(\hat{u}, \hat{v})$ is a local solution for (CP), it will also be a local solution for a problem with a fixed control structure. We use the knowledge of the optimal control structure to formulate a new unconstrained problem whose feasible controls correspond to controls of the original problem that have such fixed structure. This is achieved by a reparametrization from $[0,T]$ to the interval $[0,1]$ as described next.

In this new unconstrained problem, for each switching time we associate a state variable T_k having null dynamics, keeping the convention that $T_0 = 0$ and $T_N = T$. Such variables are initialized in the algorithm as a rough estimate of the optimal switching times, that will be iteratively tunned by the shooting scheme. For each interval $I_k := [T_k, T_{k+1}]$, we also associate a state variable x^k, that is the reparametrization of $x\big|_{I_k}$ to the interval $[0,1]$.

The control variables of the new problem are defined as follows. For each interval I_k of the partition we define a control variable $u^k \colon I_k \to \mathbb{R}^l$ that appears nonlinearly and an affine control $v^k \colon I_k \to \mathbb{R}^{|S_k|}$. This way each v^k has as many components as the number of singular components of $\hat{v}$ in $\hat{I}_k$. The bang components of v appear as constants and not as control variables, *i.e.* are fixed to 0 or 1.

The trajectories of the transformed problem have the form

$$W := \left(\left(x^k\right)_{k=1}^N, \left(u^k\right)_{k=1}^N, \left(v^k\right)_{k=1}^N, (T_k)_{k=0}^N \right), \tag{III.37}$$

and the transformed problem, denoted as (TP), is the following:

$$\begin{aligned}
&\min \phi(x^1(0), x^N(1)) \\
&\text{s.t. } \dot{x}^k = (T_k - T_{k-1}) \left(\sum_{i \in B_k \cup \{0\}} f_i(x^k, u^k) + \sum_{i \in S_k} v_i^k f_i(x^k, u^k) \right), \quad k = 1, \cdots, N, \\
&\quad \dot{T}_k = 0, \quad k = 1, \cdots, N-1, \\
&\quad \eta(x^1(0), x^N(T)) = 0, \\
&\quad x^k(1) = x^{k+1}(0), \quad k = 1, \cdots, N-1.
\end{aligned}$$

Note that given some admissible trajectory (x, u, v) of (CP), and its associated switching times (T_k), we can obtain a feasible trajectory for (TP) via the following transformation

$$\begin{aligned}
x^k(t) &:= x\left(T_{k-1} + (T_k - T_{k-1})t\right), \\
u^k(t) &:= u\left(T_{k-1} + (T_k - T_{k-1})t\right), \qquad \text{for } t \in [0,1]. \\
v^k(t) &:= v\left(T_{k-1} + (T_k - T_{k-1})t\right).
\end{aligned} \tag{III.38}$$

In fact, we prove below that we can derive the weak local optimality of a solution for (TP) from a solution for (CP), in a sense of optimality slightly weaker than L^∞. To do this, consider the definition of *Pontryagin minimum* [40].

Definition III.3.2. A feasible trajectory $\hat{w} \in \mathcal{W}$ is a Pontryagin minimum of (CP) if for any positive N, there exists some $\varepsilon_N > 0$ such that $\hat{w}$ is a minimum in the set of feasible trajectories $w = (x, u, v) \in \mathcal{W}$ satisfying

$$\|x - \hat{x}\|_\infty < \varepsilon_N, \ \|(u,v) - (\hat{u},\hat{v})\|_1 < \varepsilon_N, \ \|(u,v) - (\hat{u},\hat{v})\|_\infty < N.$$

Lemma III.3.1. *If $\hat{w}$ is a Pontryagin minimum of* (CP)*, then $\hat{W}$ obtained from $\hat{w}$ using transformation* (III.38) *is a weak minimum of* (TP).

Proof. Since $\hat{w}$ is a Pontryagin minimum of (CP), from Definition III.3.2, there exists $\varepsilon > 0$ such that

$$\|x - \hat{x}\|_\infty < \varepsilon, \ \|(u,v) - (\hat{u},\hat{v})\|_1 < \varepsilon, \ \|(u,v) - (\hat{u},\hat{v})\|_\infty < 1. \tag{III.39}$$

Let $\hat{W}$ be the transformation of $\hat{w}$ through (III.38). We now prove that $\hat{W}$ is weakly optimal for (TP). Hence we search appropriate $\bar{\delta}, \bar{\varepsilon}$ for which all feasible trajectories $W = \big((x^k), (u^k), (v^k), (T_k)\big)$ of (TP) that satisfy

$$\left|T_k - \hat{T}_k\right| < \bar{\delta}, \ \left\|(u^k, v^k) - (\hat{u}^k, \hat{v}^k)\right\|_\infty < \bar{\varepsilon}, \ \text{ for all } k = 1, \cdots, N \tag{III.40}$$

will be mapped into neighborhood of $\hat{w}$ where it is optimal. Such mapping of $W \mapsto w$ is done as follows

$$x(t) := x^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right), \quad u(t) := u^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right), \quad \text{for } t \in I_k, \tag{III.41}$$

$$v_i(t) := \begin{cases} 0, & \text{if } t \in I_k \text{ and } i \in A_k, \\ v_i^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right), & \text{if } t \in I_k \text{ and } i \in S_k, \\ 1, & \text{if } t \in I_k \text{ and } i \in B_k. \end{cases} \tag{III.42}$$

The dynamics (II.2) are clearly satisfied by (x, u, v) obtained from (III.41)-(III.42). The end point constraints in (II.3) are also easy to verify since $x(0) = x^1(0)$ and $x(T) = x^N(1)$ along with the feasibility of W.

The last step to check feasibility of w are the control constraints. For the nonlinear controls, note that since $\left\|u^k - \hat{u}^k\right\|_\infty < \bar{\varepsilon}$, we have that $\|u - \hat{u}\|_\infty < \bar{\varepsilon}$. Recalling ρ given in (III.32)-(III.33), if we choose $\bar{\varepsilon} < \rho$, then $u\left([0, T]\right) \subset U$. To discuss the feasibility of the linear controls, from equation (III.32), we can choose $\bar{\varepsilon}$ so that, whenever $t \in I_k$ and $i \in S_k$,

$$0 < \rho - \bar{\varepsilon} \leq v_i(t) \leq 1 - \rho + \bar{\varepsilon} < 1. \tag{III.43}$$

On the other hand, for $i \in A_k \cup B_k$, we know that $v_i(t) \in \{0, 1\}$ in view of (III.42), so that the control constraints are still satisfied. This concludes the proof of the feasibility of (x, u, v).

In the sequel, we find $\bar{\delta}$ and $\bar{\varepsilon}$ so that, if W satisfies (III.40), then the transformed w verifies (III.39) for the given ε. The analysis is analogous for both controls u and v, hence we will conduct the calculations component wise only for u. We have

$$\begin{aligned} \int_{I_k \cap \hat{I}_k} |u_i(t) - \hat{u}_i(t)| \mathrm{d}t \ &\leq \int_{I_k \cap \hat{I}_k} \left| u_i^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right) - \hat{u}_i^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right)\right| \mathrm{d}t \\ &+ \int_{I_k \cap \hat{I}_k} \left| \hat{u}_i^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right) - \hat{u}_i^k\left(\frac{t - \hat{T}_{k-1}}{\hat{T}_k - \hat{T}_{k-1}}\right)\right| \mathrm{d}t. \end{aligned} \tag{III.44}$$

The first integral in the r.h.s. of latter display is bounded by $\bar{\varepsilon}|I_k \cap \hat{I}_k|$ in view of (III.40). For the second term, recall that $\hat{u}$ is continuous on $[0, T]$ and so are the components of $\hat{u}^k$ over $\hat{I}_k$, so that they are uniformly continuous over these intervals. Therefore, for each $k = 1, \cdots, N$, we can find some $\bar{\delta}_k > 0$ such that, if $|T_k - \hat{T}_k| < \bar{\delta}_k$, then

$$\left| \hat{u}_i^k\left(\frac{t - T_{k-1}}{T_k - T_{k-1}}\right) - \hat{u}_i^k\left(\frac{t - \hat{T}_{k-1}}{\hat{T}_k - \hat{T}_{k-1}}\right)\right| < \bar{\varepsilon}$$

for every component of $\hat{u}^k$. Hence we only need to choose $\bar{\delta} := \min_{k=1,\cdots,N} \bar{\delta}_k$. We proved that

$$\int_{I_k \cap \hat{I}_k} |u_i(t) - \hat{u}_i(t)| \mathrm{d}t \leq 2\bar{\varepsilon}|I_k \cap \hat{I}_k|. \tag{III.45}$$

Next, we need to estimate the integral outside the intersection $I_k \cap \hat{I}_k$. We assume w.l.o.g. that $T_k < \hat{T}_k$ hence, in view of (III.40),

$$\int_{T_k}^{\hat{T}_k} |u_i(t) - \hat{u}_i(t)| \mathrm{d}t \leq \bar{\delta}\bar{\varepsilon}. \tag{III.46}$$

Adding up all the terms, we get from (III.45)-(III.46), that

$$\|u_i - \hat{u}_i\|_1 < \bar{\varepsilon}(2T + (N-1)\bar{\delta}).$$

An analogous estimate can be obtained for $\|v - \hat{v}\|_1$. Finally, taking into account all the control components m of the linear controls and l from the nonlinear controls, we get that, if

$$\bar{\varepsilon}(2T + (N-1)\bar{\delta}) < \frac{\varepsilon}{m+l},$$

then $\|u - \hat{u}\|_1 < \varepsilon$, as desired. □

III.3.2 The shooting algorithm for the transformed problem

In order to have a proper algorithm to solve control constrained problems, our final step is to define a proper shooting function and apply the shooting formulation described in Section III.1 to problem (TP).

We start by stating the PMP for this unconstrained problem (TP). Define the end-point Lagrangian

$$\tilde{\ell} := \phi(x^1(0), x^N(1)) + \sum_{j=1}^{d_\eta} \beta_j \eta_j(x^1(0), x^N(T)) + \sum_{k=1}^{N-1} \theta^k \left(x^k(1) - x^{k+1}(0) \right). \tag{III.47}$$

Note that each multiplier β_j is associated with the end point constraints that come from the original problem and each θ^k is associated with the additional constraints from (TP), included to guarantee the continuity of x. The pre-Hamiltonian of (TP) is given by

$$\begin{aligned} \tilde{H} &:= \sum_{k=1}^{N} (T_k - T_{k-1}) H^k, \\ \text{where } H^k &:= p^k \cdot \left(\sum_{i \in B_k \cup \{0\}} f_i(x^k, u^k) + \sum_{i \in S_k} v_i^k f_i(x^k, u^k) \right). \end{aligned} \tag{III.48}$$

Hence, from the PMP, the costates follow the dynamics

$$\dot{p}^k = -(T_k - T_{k-1}) D_{x^k} H^k, \tag{III.49}$$

with transversality conditions

$$p^1(0) = -D_{x_0^1}\phi - \sum_{j-1}^{d_\eta} \beta_j D_{x_0^1}\eta_j(x^1(0), x^N(T)) \tag{III.50}$$

$$\begin{aligned} p^k(1) &= \theta^k, \quad \text{for } k = 1, \cdots, N-1, \\ p^k(0) &= \theta^{k-1}, \quad \text{for } k = 2, \cdots, N, \end{aligned} \tag{III.51}$$

$$p^N(1) = D_{x_1^N}\phi + \sum_{j-1}^{d_\eta} \beta_j D_{x_1^N}\eta_j(x^1(0), x^N(T)). \tag{III.52}$$

Note that equation (III.51) can be replaced by

$$p^k(1) = p^{k+1}(0), \quad \text{for } k = 1, \cdots, N-1, \tag{III.53}$$

hence eliminating the multipliers θ^k. We must also address the costates p^{T_k} associated with the switching times, which satisfy

$$\dot{p}^{T_k} = -H^k + H^{k+1}, \quad p^{T_k}(0) = 0, \quad p^{T_k}(1) = 0, \quad \text{for } k = 1, \cdots, N-1. \tag{III.54}$$

Combining all conditions from (III.54), we obtain

$$\int_0^1 \left(H^{k+1} - H^k\right) \mathrm{d}t = p^{T_k}(0) - p^{T_k}(1) = 0. \tag{III.55}$$

Since the dynamics are autonomous the Hamiltonian is constant for the optimal trajectory and we equivalently express the conditions (III.55) for p^{T_k} as

$$H^k = H^{k+1}, \quad \text{for } k = 1, \cdots, N-1. \tag{III.56}$$

Now we are in position to adapt the shooting scheme for solving (TP). Following the steps from Section **??** we start finding the feedback form for the controls. It suffices to use the representation found in equation (III.36)

$$u^k = U\left(x^k, p^k\right) \quad v^k = V_{S_k}\left(x^k, p^k\right), \text{ for } k = 0, \cdots, N. \tag{III.57}$$

By Lemma III.3.1 such controls must also be feasible for (TP) and when the feedback arguments $\hat{x}^k$ and $\hat{p}^k$ correspond to the nominal trajectory, we obtain the optimal controls.

We must also define an appropriate shooting function that will express the stationarity of the Hamiltonian, the initial-final constraints and transversality conditions.

Stationarity with respect to the nonlinear controls is equivalent to the feedback representation for u given in equation (III.36). For the linear controls, the feedback form is equivalent to $\dot{H}_{v_{S_k}} = 0$. Hence we must also impose $H^k_{v_i^k}(0) = 0$ and $\dot{H}^k_{v_i^k}(0) = 0$ to ensure the stationarity $H_{v_k} = 0$.

Note that we can choose to include the constraints related the continuity of the states and costates or integrate each x^k and p^k using the final values of x^{k-1} and p^{k-1} as initial conditions. The clear advantage of the latter strategy is the smaller number of shooting variables, i.e. the initial conditions for states and costates at the switching times can be omitted. On the other hand, explicitly including these constraints makes the algorithm more stable numerically and favors parallelization for computational implementations, see [51].

The following is the shooting function associated to (TP) with the full set of shooting variables

$$\mathcal{S}\colon \mathbb{R}^{Nn}\times\mathbb{R}^{Nn,*}\times\mathbb{R}^{N-1}\times\mathbb{R}^{d_\eta}\to\mathbb{R}^{(N-1)n+d_\eta}\times\mathbb{R}^{(N+1)n+N-1+2\sum|S_k|,*}$$

$$\nu\mapsto\mathcal{S}(\nu):=\begin{pmatrix}\eta\left(x^1(0),x^N(1)\right)\\ \left(x^k(1)-x^{k+1}(0)\right)_{k=1,\cdots,N-1}\\ \left(p^k(1)-p^{k+1}(0)\right)_{k=1,\cdots,N-1}\\ p^1(0)+D_{x_0^1}\tilde{\ell}\left(x^1(0),x^N(1)\right)\\ p^N(1)-D_{x_1^N}\tilde{\ell}\left(x^1(0),x^N(1)\right)\\ \left(H^k(1)-H^{k+1}(0)\right)_{k=1,\cdots,N-1}\\ \left(p^i\cdot f_i\left(x^i,U^i\right)(0)\right)_{\substack{i\in S_k\\ k=1,\dots,N}}\\ \left(p^i\cdot[f_0,f_i]^x(0)\right)_{\substack{i\in S_k\\ k=1,\dots,N}}\end{pmatrix} \tag{III.58}$$

where we define the vector of shooting arguments as

$$\nu:=\left(\left(x^k(0)\right)_{k=1}^N,\left(p^k(0)\right)_{k=1}^N,(T_k)_{k=1}^{N-1},\beta\right). \tag{III.59}$$

We recall equation (II.20) that gives a concise analytical form for $\dot{H}_{v_i}$ and was used in the formulation of the shooting function.

Since the new problem (TP) falls in the same category of unconstrained problem (OC), we join Lemma III.3.1 and Theorem III.2.3 in the following.

Theorem III.3.2. *If $\hat{w}$ is a Pontryagin minimum of* (CP) *such that $\hat{W}$* (III.37) *satisfies the coercivity condition* (II.71) *for problem* (TP)*, then the shooting algorithm for* (TP) *is locally quadratically convergent.*

IV

Implementation and Examples

IV.1 The Algorithm

In this section we give all the details concerning the implementation of our shooting scheme. To keep the presentation clear, we will start with the unconstrained case, in the sequel we discuss how we have used symbolic computations to automate the work of assembling the transformed problem (TP), effectively reducing the original system to the unconstrained case. We recall the definition of the shooting function from (III.1) and the shooting arguments ν. Recall that ν has only initial conditions for x, p, on the other hand, to evaluate $\mathcal{S}(\nu)$ we require the values of states and costates at terminal time. Therefore we need a numerical scheme to integrate the Hamiltonian system on (x, p) encapsulated inside the mapping $\nu \mapsto \mathcal{S}(\nu)$. Hence we obtain the final values $(x(T), p(T))$ implicitly and finally use such values in the computation of $\mathcal{S}(\nu)$.

In order to make this dependence explicit, we define the function $\mathcal{T} = \mathcal{T}(x_0, x_T, p_0, p_T, \beta)$

$$\mathcal{T}(x_0, x_T, p_0, p_T, \beta) := \begin{pmatrix} \eta(x_0, x_T) \\ p_0 + D_{x_0}\ell[\lambda](x_0, x_T) \\ p_T - D_{x_T}\ell[\lambda](x_0, x_T) \\ H_v[\lambda](x_T, U(x_T, p_T)) \\ \dot{H}_v[\lambda](x_0, U(x_0, p_0)) \end{pmatrix}.$$

The difference between $\mathcal{S}$ and $\mathcal{T}$ is in their respective domains. Recall that $\mathcal{S}$ is defined over $\mathbb{R}^n \times \mathbb{R}^{n,*} \times \mathbb{R}^{d_\eta}$, on the other hand $\mathcal{T}$ is defined over $\mathbb{R}^{2n} \times \mathbb{R}^{2n,*} \times \mathbb{R}^{d_\eta}$. Furthermore, it is perfectly possible to evaluate $\mathcal{T}$ with arguments x_T, p_T which are not the result of the numerical integration of the initial conditions x_0, p_0. This is not the case with $\mathcal{S}$ since the choice of numerical integrator is "hardwired" into the evaluation of the shooting function. However, it is clear that given an argument $\nu = (x_0, p_0, \beta)$, the equality between $\mathcal{S}$ and $\mathcal{T}$ holds in the following sense,

$$\mathcal{S}(\nu) = \mathcal{T}(x_0, x(T, x_0), p_0, p(T, p_0), \beta).$$

The core of the algorithm consists in the Newton-like optimization loop, with the update scheme

$$\begin{aligned}\Delta_k &= \left(\mathcal{S}'(\nu_k)^T\mathcal{S}'(\nu_k)\right)^{-1}\mathcal{S}'(\nu_k)^T\mathcal{S}(\nu_k),\\ \nu_{k+1} &= \nu_k + \Delta_k.\end{aligned}$$

As we have discussed, due to the dependence of the shooting function on the integration scheme, the practical computation of its derivative becomes a difficult issue. Indeed, one cannot ignore the implicit dependence on the terminal values $(x(T), p(T))$ and their contribution in the derivatives with respect to the initial conditions. Fortunately, we can compute the derivatives of the terminal values with respect to the initial conditions by means of the variational equation. Since the states (x, p) evolve with the Hamiltonian system

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial H}{\partial p} \\ -\dfrac{\partial H}{\partial x} \end{pmatrix},$$

then we can compute derivatives with respect to the initial conditions, see [15, 31],

$$\Psi(t) = \begin{pmatrix} \dfrac{\partial x(t)}{\partial x_0} & \dfrac{\partial x(t)}{\partial p_0} \\ \dfrac{\partial p(t)}{\partial x_0} & \dfrac{\partial p(t)}{\partial p_0} \end{pmatrix}$$

with the following variational ODE

$$\dot{\Psi}(t) = \begin{pmatrix} \dfrac{\partial^2 H}{\partial p \partial x} & \dfrac{\partial^2 H}{\partial p \partial p} \\ \\ -\dfrac{\partial^2 H}{\partial x \partial x} & -\dfrac{\partial^2 H}{\partial x \partial p} \end{pmatrix} \Psi(t), \quad \Psi(0) = I_{2n\times 2n}. \tag{IV.1}$$

Note however, that in general, the second order derivatives of the Hamiltonian depend on the states and costates (x, p), therefore we need to integrate the variables x, p, Ψ simultaneously. Finally, once we have computed the final values $x(T), p(T), \Psi(T)$ we can compute the derivative of the shooting function as

$$\mathcal{S}'(\nu) = \begin{pmatrix} D_{x_0,p_0}\mathcal{T} + D_{x_T,p_T}\mathcal{T}\cdot\Psi(T) & \dfrac{\partial \mathcal{T}}{\partial \beta} \end{pmatrix}. \tag{IV.2}$$

The same analysis is easily done for the constrained case, we need only to pay some attention to the notation introduced by problem (TP). We unify the state and costate variables in the same vector $X \in \mathbb{R}^{nN} \times \mathbb{R}^{nN,*} \times \mathbb{R}^{N-1}$ where

$$X = \left((x^k)_{k=1}^N, (p^k)_{k=1}^N, (T_k)_{k=1}^{N-1}\right). \tag{IV.3}$$

After substituting the unconstrained controls from problem (TP) with their feedback representations, the dynamics of X and Ψ assume the following form

$$\begin{aligned}\dot{X} &= F(X), & X(0) &= X_0,\\ \dot{\Psi}_t &= D_x F(X)\Psi_t, & \Psi(0) &= I.\end{aligned} \tag{IV.4}$$

The following subsections discuss the different methods we have used to automate the task of assembling the appropriate system (IV.4) for problem (TP) and the variation of a Runge-Kutta method that exploits the particularities of this coupled system in the numerical integration. We end the section summarizing our algorithm and giving the final details of our implementation.

IV.1.1 Symbolic Computations and Assembling Problem TP

The first difficulty of our proposed algorithm is to assembly the associated transformed problem (TP). Even the simpler constrained problems often present many arcs, making the manual declaration of the required functions, as well as their jacobians, a tiring and time consuming task.

Therefore, the main goal of our implementation is to automate such processes, specially regarding the computations of the singular linear controls following equation (II.37). For this we have employed symbolic computations using a *Computer Algebra System* (CAS) to compute the quantities γ_{ij} from equation (II.37).

In the sequel we specify the steps for the case with a single singular control. In such case, the singular arc in feedback form becomes

$$V(x,p) = -\frac{\gamma_{10}}{\gamma_{11}}.$$

Even this simple expression can require lots of computations. Take for instance the example treated in Section IV.2.2, we dedicated Appendix B to the manual computations of the singular controls for this problem. In this example we have reached a satisfactory analytical expression, however using many tricks and identities coming from the optimality conditions.

These computations might be misleading since identities such as the Goh conditions

$$p \cdot [f_i, f_j] \equiv 0,$$

do not state that these equalities hold in the sense that this computation will be satisfied for any choice of x, p, they are only guaranteed to hold along the optimal trajectories. For this reason, it is not unlikely that some of these conserving quantities are the sum of components that do not vanish, and hence making the manual computations for the singular controls larger.

Other strategies to obtain the linear controls, for example solve the system of equations (II.37) numerically, also have some disadvantages. Even if this linear system can be solved with high accuracy, there is still the need to compute the jacobian of the dynamics. This would require some finite differences approximation for the partial derivatives, introducing another source of numerical residual to the algorithm, which should not be neglected. With symbolic computations we do not face this issue, since we always have the exact expressions, even if they are not simplified.

Once we have the singular controls, the dynamics are treated modularly, according to the type of arc we must deal with. Let $\omega \in \{\omega_+, \omega_-, \omega_s\}$ denote the possible types of arcs, where ω_+ and ω_- denote the upper and lower bang arcs and ω_s the singular arcs, we define the following parametrized vector fields

$$f^{\omega}(x,p) = \begin{cases} f_0\left(x, U(x,p)\right), & \text{if } \omega = \omega_-, \\ f_0\left(x, U(x,p)\right) + V(x,p) f_1\left(x, U(x,p)\right), & \text{if } \omega = \omega_s, \\ f_0\left(x, U(x,p)\right) + f_1\left(x, U(x,p)\right), & \text{if } \omega = \omega_+. \end{cases} \tag{IV.5}$$

This parametrized vector fields can be easily implemented using hashtables, mapping arcs to the appropriated dynamic, $\omega \mapsto f^{\omega}$. Hence, given a user input of a sequence of arcs $(\omega_k)_{k=1}^N$, we define $f^k = f^{\omega_k}(x^k, p^k)$, where x^k, p^k denote the states and costates, referent to arc k, introduced by problem (TP). Finally, we define the dynamics in (IV.4) as

$$F(X) = \begin{pmatrix} \left((T_k - T_{k-1})f^k\right)_{k=1}^N \\ \left((T_k - T_{k-1})p^k \cdot D_x f^k\right)_{k=1}^N \\ \mathbf{O}_{N-1\times 1} \end{pmatrix}. \tag{IV.6}$$

Afterwards, the computation of the jacobian DF is straight forward with symbolic differentiation.

IV.1.2 Integrating the Variational System

In this section we intend to analyse general Runge-Kutta schemes to integrate the variational system coupled with the states dynamics.

$$\begin{aligned} \dot{X} &= F(t, X), & X(0) &= X_0, \\ \dot{\Psi}_t &= D_x F(t, X)\Psi_t, & \Psi(0) &= I. \end{aligned} \tag{IV.7}$$

Here we have chosen to keep the dependence on time to make the exposition of the Runge-Kutta methods clearer. The original ODE can be integrated numerically independently of the second system, but the variational dynamics requires the values of the states at each time in order to evaluate the jacobian. We will assume that numerical approximations for the states $x(t_k) \approx x_k$ are available and describe an update formula for a general Implicit Runge Kutta method for the variational ODE. We will show that those update schemes can be written as the solution of linear systems that admit a parallelizable implementation.

The general s-stage Runge Kutta update scheme is given by the following system

$$\begin{aligned} k_i &= F\left(t_0 + c_i h, X_0 + h\sum_{\ell=1}^{s} a_{i\ell}k_\ell\right), \quad \text{for } i = 1, \dots, s \\ X_1 &= X_0 + h\sum_{i=1}^{s} b_i k_i, \end{aligned} \tag{IV.8}$$

also represented by a Butcher tableau (A, b, c), where $A \in \mathbb{R}^{s\times s}, b \in \mathbb{R}^{1\times s}, c \in \mathbb{R}^{s\times 1}$. Note that we can rewrite the previous system as a fixed point problem defining the quantities

$$\begin{aligned} g_i &:= X_0 + h\sum_{\ell=1}^{s} a_{i\ell}k_\ell \\ &= X_0 + h\sum_{\ell=1}^{s} a_{i\ell}\underbrace{F\left(t_0 + c_\ell h, X_0 + h\sum_{j=1}^{s} a_{\ell j}k_j\right)}_{=F(t_0+c_\ell h, g_\ell)} \end{aligned}$$

this way, system (IV.8) can be written in terms of the new variables g_i as

$$\begin{aligned} g_i &= X_0 + h\sum_{\ell=1}^{s} a_{i\ell}F(X_0 + c_\ell h, g_\ell), \quad \text{for } i = 1,\dots,s \\ X_1 &= X_0 + h\sum_{i=1}^{s} b_i F(X_0 + c_i h, g_i). \end{aligned} \tag{IV.9}$$

This new system is useful to provide conditions upon the maximum step size h that guarantees the existence of an unique solution of for the Runge Kutta scheme[1], see for instance [29, 30]. This is why we choose to present this intermediate transformation of the original problem, but to facilitate our computations, we introduce the new variables $Z_i := g_i - X_0$. Obtaining such quantities reduces to solving the following nonlinear system of equations

$$Z_i = h\sum_{\ell=1}^{s} a_{i\ell}F(X_0 + c_i h, X_0 + Z_\ell), \text{ for } i = 1,\dots,s. \tag{IV.10}$$

In matrix notation, we obtain the following system for $Z := (Z_1^T,\dots,Z_s^T)^T$,

$$Z := \begin{pmatrix} Z_1 \\ \vdots \\ Z_s \end{pmatrix} = h(A \otimes I_{n\times n}) \begin{pmatrix} F(t_0 + c_i h, X_0 + Z_1) \\ \vdots \\ F(t_0 + c_i h, X_0 + Z_s) \end{pmatrix}, \tag{IV.11}$$

where we have employed the *Kronecker product* notation

$$A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}, \tag{IV.12}$$

This formulation is valid for any ODE, and indeed is useful in a Newton scheme to solve the nonlinear fixed point problem of the variables g_i. For our purposes, we exploit the fact that the dynamics of the variational system can be computed by means of matrix multiplications to formulate a linear system for Z. For this we introduce the notation

$$F_i := DF(t_0 + c_i h, X(t_0 + c_i h)), \quad \text{for } i = 1,\dots,s$$

so that equation (IV.11) becomes the linear system

$$Z = h(A \otimes I_{n\times n}) \underbrace{\operatorname{diag}(F_1,\dots,F_s)}_{=:F_{sn\times sn}} \begin{pmatrix} \Psi_0 + Z_1 \\ \vdots \\ \Psi_0 + Z_s \end{pmatrix}. \tag{IV.13}$$

Solving for Z, we obtain

$$Z = (I_{sn\times sn} - h(A \otimes I_{n\times n})F_{sn\times sn})^{-1} h(A \otimes I_{n\times n})F_{sn\times sn} \begin{pmatrix} \Psi_0 \\ \vdots \\ \Psi_0 \end{pmatrix}. \tag{IV.14}$$

[1] The proof follows by means of a fixed point argument, controlling h in order to obtain a contraction over $g := (g_1,\dots,g_s)$ and finalizing using Banach's fixed point theorem. One can also prove that the increment operator is smooth as a consequence of the implicit function theorem.

We are still left with the issue of computing the update Ψ_1. The difficulty lies in the fact that the update equation (IV.8) is stated in terms of the quantities k_ℓ, instead of Z_i. This can be overcame with the following trick

$$\Psi_1 = \Psi_0 + \sum_{i=1}^{s} d_i Z_i, \quad \text{where } d = b^T A^{-1}. \tag{IV.15}$$

This is easy to prove noting that

$$\begin{aligned}\sum_{i=1}^{s} d_i Z_i = \left(b^T A^{-1} \otimes I\right) Z &= \left(b^T A^{-1} \otimes I\right) h(A \otimes I) \begin{pmatrix} F_1 \\ \vdots \\ F_s \end{pmatrix} \\ &= h\left(b^T A^{-1} \cdot A\right) \otimes (I \cdot I) \begin{pmatrix} F_1 \\ \vdots \\ F_s \end{pmatrix} \\ &= \sum_{i=1}^{s} b_i F_i = \Psi_1 - \Psi_0.\end{aligned}$$

Some small optimizations are in place when we look at the matrix $(A \otimes I_{n\times n})F_{sn\times sn}$, if fact $A \otimes I_{n\times n}$ has a nice sparse structure, even when A has arbitrary entries, and we obtain

$$(A \otimes I_{n\times n})F_{sn\times sn} = \begin{pmatrix} a_{11}F_1 & \cdots & a_{1s}F_s \\ \vdots & \ddots & \vdots \\ a_{s1}F_1 & \cdots & a_{ss}F_s \end{pmatrix}, \tag{IV.16}$$

saving some matrix multiplications in the computation of Z. This economy in each iteration makes a considerable difference when the dimension of problem (TP) grows.

Another alternative to the update scheme presented here would be to turn the matrix ODE (IV.7) and turn it into a big vector valued problem, of size $N(N+1)$ where N is the number of states in X. However, the approach we have presented takes greater advantage of the nonlinear system solvers that need to be implemented inside a RK algorithm to solve (IV.9). If those were to be applied to the vector valued variant, the nonlinear solver would take into account the variables coming from the variational ODE as well. The issue lies in the fact that although the RK methods for this part of the ODE can be computed directly as linear systems, this information is lost in the internal Newton loop necessary to solve the system (IV.9). This makes each iteration of the nonlinear solver much more costly, having to perform multiplications with matrices of size $N(N+1)$, instead of N. We still need to solve the linear system (IV.13), however this is done only once on each time step.

In practice we have observed that the vector valued approach to solve (IV.7) takes more steps with the adaptive step size algorithms, but one can expect more precision since the adaptive step size also takes the variational system into account. This lack of precision in the approach here presented could lead to slightly more imprecise estimates for the derivative of the Shooting function, further investigations are required in this direction.

Finally, to summarize our scheme, given an estimate X_k for the states at time t_k, we use an adaptive-step algorithm to find the optimal step h_k and the estimate X_{k+1} for the states at time $t_{k+1} = t_k + h_k$

$$X_k \mapsto (X_{k+1}, h_k),$$

where this mapping is defined by the specific RK scheme chosen. Afterwards, we need to compute the quantities $F_1, \ldots, F_s$. This is done using a fixed step size iteration of the same RK method to compute estimates for $X(t_k + c_i h_k)$. Notice that this step is parallelizable, since we can use X_k as initial condition for all of those computations. Afterwards it is just a matter of making the s function evaluations

$$F_i = DF(t_k + c_i h_k, X(t_k + c_i h_k)), \qquad \text{for } i = 1, \ldots, s.$$

This discussion is summarized in Algorithm 1.

Algorithm 1: Variational Integrator

Data: $t_0, T, (A, b, c), X_0$
Result: Estimatives for $X(T), \Psi(T)$
Compute $d = b^T A^{-1}$;
$t_k, X_k, \Psi_k \leftarrow t_0, X_0, \Psi_0$;
while $t_k < T$ **do**
  Compute X_{k+1} with $RK(A, b, c)$ using initial condition X_k;
  for $i = 1 \ldots, s$ **do**
    Compute $X(t_k + c_i h_k)$ with fixed step $c_i h_k$ and initial condition X_k;
    $F_i \leftarrow D_x F(t_k + c_i h_k, X(t_k + c_i h_k))$;
    Obtain $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_s \end{pmatrix}$ with equation (IV.14);
    Update $\Psi_{k+1} = \Psi_k + \sum_{i=1}^{s} d_i Z_i$;
  end
  Avance time step $t_{k+1} \leftarrow t_k + h_k$;
end

IV.1.3 Summing Up the Algorithm

Now we can gather all the steps discussed previously. The implementation was done in Python [46, 47]. Apart from the standard tools, we have chosen `symengine`, [16] as our CAS to perform the symbolic computations. The alternative was `sympy`, Python's standard package for symbolic computations. The problem is that the latter is implemented in pure Python, making the computations much slower than the former, which is implemented in C++.

For the numerical integration, we have used the `pygsl` package [27], a wrapper to the well-established General Scientific Library of the C programming language. We have used the Gauss collocation methods from Table I.1 since they coincide with their corresponding Partitioned Runge-Kutta methods, as discussed in Section I.3.3.

All of the examples discussed on this thesis can be found on the link.

Algorithm 2: Shooting Algorithm

Data: Inicial guess for ν, tolerance ε, arcs $(\omega_k)_{k=1}^N$
Result:
Obtain $(\omega_k)_{k=1}^N \mapsto F$ symbolically with (IV.6);
Symbolically differentiate F to get DF;
Obtain $\mathcal{S}$;
while $\|\mathcal{S}(\nu)\| > \varepsilon$ **do**
  $X_0 \leftarrow \nu(1 : 2nN + N - 1)$;
  $\beta \leftarrow \nu(-d_\beta : end)$;
  Integrate the system

$$\begin{aligned} \dot{X} &= F(X), & X(0) &= X_0, \\ \dot{\Psi} &= D_X F(X) \cdot \Psi, & \Psi(0) &= I_{2nN \times 2nN} \end{aligned}$$

  using Algorithm 1, obtain $(X(1), \Psi(1))$;
  Compute $\mathcal{S}(\nu) = \mathcal{T}(X_0, X(1), \beta)$;
  Compute $\mathcal{S}'(\nu) = \begin{pmatrix} \dfrac{\partial \mathcal{T}}{\partial X_0} + \dfrac{\partial \mathcal{T}}{\partial X_1} \cdot \Psi(1) & \dfrac{\partial \mathcal{T}}{\partial \beta} \end{pmatrix}$;
  $\Delta \leftarrow -\left(\mathcal{S}'(\nu)^T \mathcal{S}'(\nu)\right)^{-1} \mathcal{S}'(\nu)^T \mathcal{S}(\nu)$;
  $\nu \leftarrow \nu + \Delta$;
end

IV.2 Examples

IV.2.1 Degenerate Linear Quadratic Problem

In this section we check the sufficient optimality conditions for a toy problem. We consider the following partially affine example, inspired by the examples in [19, 2].

$$\begin{aligned} \text{minimize} \quad & -2x_2(2) + \int_0^2 \left(x_1^2 + x_2^2 + u^2 + 10x_2 v\right) \mathrm{d}t \\ \text{subject to} \quad & \dot{x}_1 = x_2 + u, \\ & \dot{x}_2 = v, \\ & 0 \le v(t) \le 0.5, \quad \text{a.e. on } [0, T] \\ & x_1(T) = 1 \\ & x_1(0) = x_2(0) = 0. \end{aligned} \tag{IV.17}$$

We start by obtaining an estimate for the optimal control structure. This was done by using the BOCOP package [12], where we obtained that the optimal solution presents a *Bang, Singular, Bang* structure. This way, the transformed problem (TP),

in the Mayer form, becomes

$$\begin{array}{ll}\text{minimize} & x_{3,3}(1) - 2x_{3,2}(1) \\ \text{subject to} & \dot{x}_{1,1} = (T_1 - T_0)\,(x_{1,2} + u_1)\,, \\ & \dot{x}_{1,2} = (T_1 - T_0)\,0.5, \\ & \dot{x}_{1,3} = (T_1 - T_0)\left(x_{1,1}^2 + x_{1,2}^2 + u_1^2 + 5x_{1,2}\right), \\ & \dot{x}_{2,1} = (T_2 - T_1)\,(x_{2,2} + u_2)\,, \\ & \dot{x}_{2,2} = (T_2 - T_1)\,(v)\,, \\ & \dot{x}_{2,3} = (T_2 - T_1)\left(x_{2,1}^2 + x_{2,2}^2 + u_2^2 + 10x_{2,2}v\right), \\ & \dot{x}_{3,1} = (2 - T_2)\,(x_{3,2} + u_3)\,, \\ & \dot{x}_{3,2} = 0, \\ & \dot{x}_{3,3} = (2 - T_2)\left(x_{3,1}^2 + x_{3,2}^2 + u_3^2\right), \\ & \dot{T}_i = 0, \;\; i = 1,2 \\ & x_{1,3}(1) = 1 \\ & x_{1,1}(0) = x_{1,2}(0) = x_{1,3}(0) = 0, \\ & x_{i+1,j}(0) = x_{i,j}(1), \;\; i,j = 1,\cdots,3. \end{array} \tag{IV.18}$$

Here the state variable $x_{i,j}$ indicates arc i and state index j. We have changed the superscript notation for the arc index to avoid confusion with exponents. We make a slight abuse of notation defining the variables:

$$X = \left((x_{i,j})_{i,j=1,\cdots,3}\right), (T_i)_{i=1,2}\,,\; U = (u_i)_{i=1,\cdots,3}\,,\; V = v \tag{IV.19}$$

Moving on to computing $\tilde{\Omega}_{\mathcal{P}_2}$, we first note that we only need to evaluate it in the cone of critical directions. Consider $(\tilde{\xi}, \tilde{u}, \tilde{y})$ in $\mathcal{P}_2$, the Goh transform of the linearized dynamics, satisfying (II.52). First note that we can partition $\tilde{\xi}$ as $\tilde{\xi} = \left((\xi_k)_{k=1}^3, \xi_T\right)$, where the terms ξ_T are the transformed variables corresponding to the switching times, and therefore are constant and equal to zero, since each T_k has null dynamics. The same happens for the transformed variation $\tilde{\xi}_{2,2}$. Therefore, writing each term of the quadratic form $\tilde{\Omega}_{\mathcal{P}_2}$, we obtain

$$\tilde{\xi}^T \tilde{H}_{xx} \tilde{\xi} = \sum_{k=1}^{3} 2(T_k - T_{k-1})\left(\xi_{k,1}^2 + \xi_{k,2}^2\right),\; 2\tilde{u}^T \tilde{H}_{ux} \xi = 0 \tag{IV.20}$$

$$2\tilde{y}^T \tilde{M} \tilde{\xi} = 4(T_2 - T_1)^2 \tilde{y} \tilde{\xi}_{2,2},\; \tilde{u}^T \tilde{H}_{uu} \tilde{u} = \sum_{k=1}^{3} 2(T_k - T_{k-1}) u_k^2 \tag{IV.21}$$

$$2\tilde{y}^T \tilde{E} \tilde{u} = 0,\; \tilde{y}^T R \tilde{y} = (T_2 - T_1)^3 2\tilde{y}^2,\; g\left(\tilde{\xi}(0), \tilde{\xi}(1), h\right) = 10\left(2h\tilde{\xi}_{3,2}(1) + h^2\right) \tag{IV.22}$$

Finally, recalling that $\tilde{\xi}_{3,2} = 0$, $\tilde{\Omega}_{\mathcal{P}_2}$ assumes the form

$$\tilde{\Omega}_{\mathcal{P}_2} = \int_0^1 \left(\sum_{k=1}^{3} 2(T_k - T_{k-1})\left(\xi_{k,1}^2 + \xi_{k,2}^2 + u_k^2\right) + 4(T_2 - T_1)^2 \tilde{y} \tilde{\xi}_{2,2} + 2(T_2 - T_1)^3 \tilde{y}^2 \right) \mathrm{d}t + 10h^2.$$

Completing squares for the cross term $\tilde{y}\tilde{\xi}_{2,2}$, we can find some positive constant C, depending on the switching times, such that

$$\tilde{\Omega}_{\mathcal{P}_2} \geq 10h^2 + C \int_0^1 \left(\tilde{y}^2 + \sum_{k=1}^{3} u_k^2 \right) \mathrm{d}t. \tag{IV.23}$$

The coercivity is proven, so that we have verified the sufficient condition from Theorem II.3.7. Figure IV.1 shows a comparison the solutions of our shooting algorithm and the one obtained from BOCOP. BOCOPS's solution already shows a good approximation of the singular control, however it has poor performance around the switching times. Another interesting numerical phenomenon is the fact that, generally in direct methods, the control variables have a tendency to have a slower convergence than the states and costates variables [28]. This can be observed in the comparison graph of the nonlinear controls. Since the shooting algorithm uses the analytical expression of the optimal controls, we can expect more accurate results.

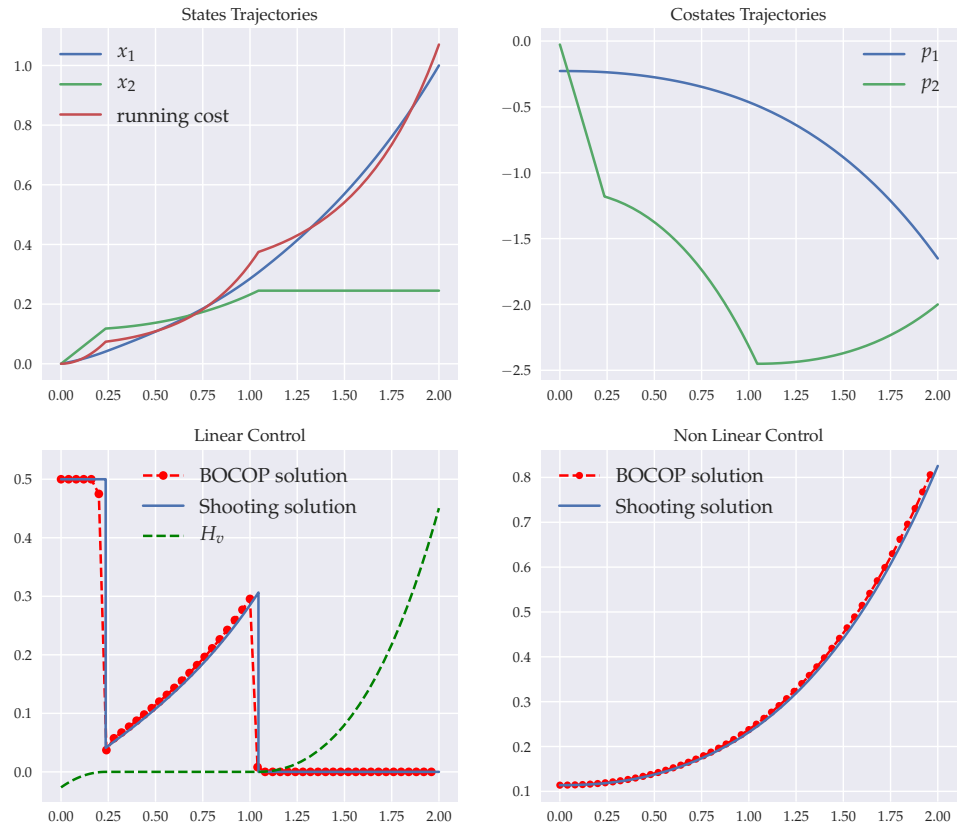

FIGURE IV.1: Optimal trajectories and controls for problem (IV.17)

IV.2.2 Optimal Control of an SIRS Epidemiological Model

In this section we follow [22, 35] where the authors discuss problems regarding the optimal control of various SIR models used to describe the spread of an epidemic in some demographic population. The control is performed through a term v modeling vaccination of susceptible individuals S, leading them to the recovered class R before they enter the infected class I, and the treatment of infected individuals is represented by a second control variable u. We consider the variation known as

SIRS model, which takes into account the effect of temporary immunity of recovered individuals, gradually reintroducing them into the susceptible class. This brief discussion is encapsulated in (IV.24) below:

$$\begin{aligned} \dot{N} &= F(N) - \delta I - \mu N \\ \dot{S} &= F(N) - \beta \tfrac{IS}{N} - vS + \omega(N - S - I) - \mu S \\ \dot{I} &= \beta \tfrac{IS}{N} - (\gamma + \delta + u)I - \mu I, \end{aligned} \tag{IV.24}$$

In equation (IV.24), instead of using the traditional states S, I, R, we introduce the total number of individuals in the population, $N = S + I + R$. Working with such variables simplifies the computations, as done in [22, 35]. The function $F : [0, K) \to \mathbb{R}_+$ is the population growth function. We assume the logistic growth model, *i.e.* $F(N) = \alpha N(1 - N/K)$ and that all newborn individuals enter the susceptible class.

Our goal is to minimize the amount of ill individuals with the lowest cost of vaccination and treatment over a time window, hence we choose the following cost function

$$\int_0^T \left(B_1 I(t) + B_2 v(t) + B_3 u^2(t)\right) \mathrm{d}t \tag{IV.25}$$

The choice of terms $B_1 I$ and $B_3 u^2$ follows [22], the integral of the former is proportional to the total amount of deaths due to the disease, while the former is chosen to model the difficulty of public health agents to implement treatment strategies to a wide portion of the population. Vaccination on the other hand is more easily implemented and hence appears linearly in the cost as in [35]. The linear dependence on the vaccination might result in bang-bang optimal controls as in [8], however the parameters values in Table IV.1 were chosen to favor the appearance of singular arcs among realistic parameters given in [22].

Parameter	**Biological Meaning**	**Values**
N_0	initial total population	5000 humans
S_0	initial susceptible population	4500 humans
I_0	initial infected population	499 humans
α	population growth rate	4.10^{-5} / day
K	carrying capacity	5000
μ	natural death rate of population	10^{-5} / day
β	incidence rate	0.5 / day
ω	waning rate	0.01 / day
γ	infection time	0.1 / day
δ	death rate due to disease	0.1 / day
B_1	cost per death	1
B_2	cost per vaccination	50
B_3	cost per treatment	1000
$v_{\max}$	maximum vaccination rate	0.25
T	period of analysis	100 days

TABLE IV.1: Biologically feasible parameters.

Now we introduce the optimal control problem in Mayer form

$$\begin{array}{ll} \text{minimize} & C(T) \\ \text{subject to} & \dot{N} = F(N) - \delta I - \mu N, \\ & \dot{S} = F(N) - \beta\frac{IS}{N} - vS + \omega(N - S - I) - \mu S, \\ & \dot{I} = \beta\frac{IS}{N} - (\gamma + \delta + u)I - \mu I, \\ & \dot{C} = B_1 I + B_2 v + B_3 u^2, \\ & 0 \leq v(t) \leq v_{\max}, \quad \text{a.e. on } [0, T] \\ & 0 \leq u(t), \quad \text{a.e. on } [0, T] \\ & N(0) = N_0, \quad S(0) = S_0, \quad I(0) = I_0, \quad C(0) = 0. \end{array} \tag{IV.26}$$

We will show that the restriction of non negativity on the nonlinear controls is redundant, since a "negative treatment" is never optimal. This is intuitive since negative treatment values would introduce more infected individuals to the overall population, and therefore increase the cost.

Proposition IV.2.1. *An optimal treatment strategy* $(\hat{u}, \hat{v})$ *for problem* (IV.26) *will never present negative values for* $\hat{u}$.

Proof. Suppose an optimal solution $(\hat{u}, \hat{v})$ is such that $\hat{u}(t)$ presents negative values in a measurable set. Define a new control strategy, where $\hat{v}$ remains unchanged and exchange $\hat{u}$ by $\tilde{u} := \max\{\bar{u}, 0\}$. The cost associated with $\hat{v}$ is unaffected, the term depending on the treatment, $\int_0^T u^2(s)\mathrm{d}s$, is clearly less expensive for u_0 and it remains to be checked the influence on the cost associated with the amount of infected individuals of this strategy.

With this in mind, let (N, S, I) and $(\tilde{N}, \tilde{S}, \tilde{I})$ be the solutions for (IV.24) with the control strategies $(\hat{u}, \hat{v})$ and $(\tilde{u}, \hat{v})$, respectively. To conclude our argument, it suffices to show that the quantity $z := \tilde{I} - I$ is non positive. Note that

$$\dot{z} = \dot{\tilde{I}} - \dot{I} = \beta\tilde{I}\frac{\tilde{S}}{\tilde{N}} - \beta I\frac{S}{N} - (\gamma + \delta + \mu)(\tilde{I} - I) + \hat{u}I\chi_{\{\hat{u}<0\}}.$$

Hence, we can define a continuous function $c(t)$, depending on $\tilde{S}, \tilde{N}, S, N$, such that

$$\dot{z} \leq \left(\beta c(t) - \left(\gamma + \delta + \mu + \hat{u}\xi_{\{\hat{u}>0\}}\right)\right) z + \hat{u}I\chi_{\{\hat{u}<0\}}.$$

Setting $a(t) := \beta c(t) - \left(\gamma + \delta + \mu + \hat{u}\chi_{\{\hat{u}>0\}}\right)$ and $b(t) := \hat{u}I\chi_{\{\hat{u}<0\}}$, by Gronwall's lemma, we have that

$$z(t) \leq z(0)\exp\left(\int_0^t a(s)\mathrm{d}s\right) + \int_0^t b(s)\exp\left(\int_0^s a(\sigma)\mathrm{d}\sigma\right)\mathrm{d}s.$$

By definition, $z(0) = 0$ and $b \leq 0$, thus $z \leq 0$, this is $\tilde{I} \leq I$. □

With the aid of the previous Proposition IV.2.1, our control problem (IV.26) satisfies all assumptions from Section III.3, since the constraints $u \geq 0$ can be removed, and we can apply our algorithm. The singular vaccination strategies are obtained using the expression for $\ddot{H}_v$ derived in (II.37). The complete analytical computation can be found in appendix B, however, our computational implementation relies on SymEngine - a *Computer Algebra System* (CAS), see [16] - to automate this laborious task and other computations necessary to formulate our algorithm.

We employed the shooting algorithm proposed in the present work to solve the optimal control problem (IV.26). As before, we used the BOCOP software [12] to get

an estimate of the shooting parameters and initialize our algorithm. The results are shown in Figures IV.2 and IV.3.

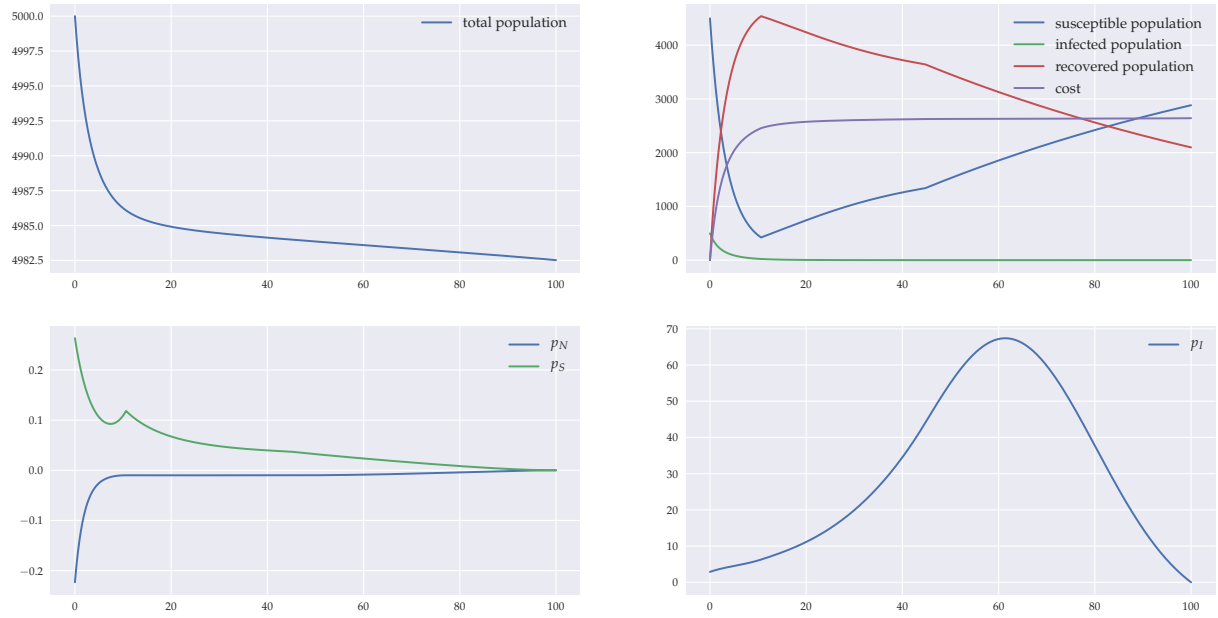

FIGURE IV.2: Optimal trajectories for problem (IV.26).

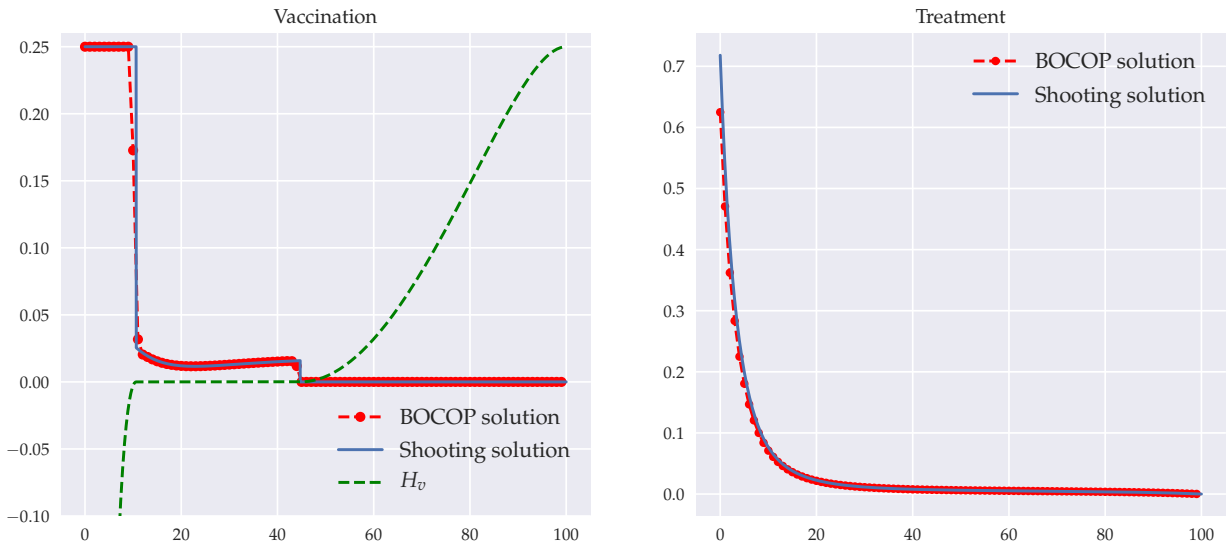

FIGURE IV.3: Optimal controls for problem (IV.26).

V

Conclusion

In this work we have studied the shooting algorithm for partially affine control problems, problems in which some of the control components appear linearly in the Hamiltonian and others do not. Our analysis was focused on the theoretical convergence of the algorithm and in its practical implementation.

Concerning the convergence, the difficulties were twofold: first the usual strategies to prove the convergence of such algorithms in the literature consist in assume the coercivity of the second variation. However, this variation with respect to the linear controls vanishes, making the usual second order optimality conditions non informative. This has been overcome with the use of the Goh transform that enables the derivation of more informative second order conditions. The second difficulty was in obtaining convergence results for control constrained problems. The derivation of second order sufficient conditions for such problems is still an open question, hence an analogous approach to the proof of convergence given in the unconstrained case becomes unfeasible. The method used in the present work to overcome this issue was to assemble a transformed unconstrained problem, whose solution is equivalent to the solution of a prescribed constrained one.

Regarding our implementation, the same transformed problem is the main source of difficulties. From one side, it is desirable that the programming interface to the end user does not requires much more than the modeling of the system in question and the criteria for optimality, *e.g.* cost function and constraints. From the other side, as we have discussed, tackling a constrained problem directly presents many complications for the shooting algorithm. Hence, the main goal of our implementation was to automate the step of assembling the transformed problem. This and the lengthy calculations of singular controls was achieved with symbolic computations. Our implementation was thoroughly tested in the various examples presented throughout the text.

A

On the Gauss-Newton Method

In this appendix we further discuss the formulation and convergence properties of the Gauss-Newton method. Even thought we have chosen a notation that is consistent with our purposes of finding the roots of the shooting function, the analysis here is applicable to any problem were the Gauss-Newton method is suitable. For simplicity we allow the abuse of notation that $\mathcal{S} : \mathbb{R}^n \to \mathbb{R}^m$. The reader is advised not to mistake n and m with the dimensions of the states and nonlinear controls.

We remind that our primal objective is to solve

$$\mathcal{S}(\hat{v}) = 0. \tag{A.1}$$

Such problem is equivalent to solving the optimization problem

$$\min_{v \in D(\mathcal{S})} |\mathcal{S}(v)|^2 . \tag{A.2}$$

First order conditions of optimality for this convex problem make (A.1) and (A.2) equivalent. In turn we make updates in the approximations of the solutions as $v_{k+1} = v_k + \Delta_k$ where Δ_k is the solution of the least squares problem

$$\min_{\Delta \in D(\mathcal{S})} \left|\mathcal{S}(v_k) + \mathcal{S}'(v_k)\Delta\right|^2 . \tag{A.3}$$

That is an interactive approach to solve (A.2) done by means of a first order Taylor approximations at each step. While Δ_k can be expressed as the solution of

$$S'(v_k)^T S'(v_k)\Delta_k + S'(v_k)^T S(v_k) = 0, \tag{A.4}$$

and explicitly written provided that $S'(v_k)^T S'(v_k)$ is invertible, we shall still have some error at each approximation that comes from the truncation of the residual terms from the Taylor expansion. Since there are various estimates of such error depending on the regularity of the objective function, we can also have different speeds of convergence depending on such regularity. This discussion is formalized in the following theorem.

Theorem A.0.1 (Gauss-Newton convergence). *Let $\hat{v}$ be a solution of problem (A.1), such that the matrix $\mathcal{S}'(\hat{v})^T\mathcal{S}'(\hat{v})$ is nonsingular. Consider also the sequence $(v_k)_{k\in\mathbb{N}}$ defined as in (A.3), such that v_0 is chosen sufficiently close to the solution so that $\mathcal{S}'(v_0)^T\mathcal{S}'(v_0)$ is also non singular. Then the Gauss Newton method converges at least linearly. Moreover, in case the function $S' : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ is Lipschitz continuous, quadratic convergence holds.*

Proof. We define $\epsilon_k := \nu_k - \hat{\nu}$, the error of each approximation of the solution using the Gauss Newton method. Consider the following first order Taylor expansion of the objective function

$$\mathcal{S}(\nu_k + h) = \mathcal{S}(\nu_k) + \mathcal{S}'(\nu_k) \cdot h + r(h), \tag{A.5}$$

where r is the residual of the Taylor expansion. We choose $h = -\epsilon_k$ so that the left side of (A.5) vanishes, since it is evaluated at the solution. We obtain

$$\mathcal{S}(\nu_k) - \mathcal{S}'(\nu_k) \cdot \epsilon_k + r(-\epsilon_k) = 0. \tag{A.6}$$

Multiplying both sides by $\mathcal{S}'(\nu_k)^T$ and using the definition of Δ_k in (A.4), we can write

$$\epsilon_k + \Delta_k = \left(\mathcal{S}'(\nu_k)^T \mathcal{S}'(\nu_k)\right)^{-1} r(-\epsilon_k), \tag{A.7}$$

since we assumed the sequence to be close enough to $\hat{\nu}$ so that, for each element of the sequence, the inverse is well-defined and bounded. We call the reader's attention to the fact that $\epsilon_k + \Delta_k = \epsilon_{k+1}$ and turn to the analysis of the term $r(\epsilon_k)$. If we consider the derivative $\mathcal{S}'$ to be Lipschitz continuous, then it is useful to write the residual term as

$$r(-\epsilon_k) = \int_0^1 \left[\mathcal{S}'(\hat{\nu} + t\epsilon_k) - \mathcal{S}'(\hat{\nu})\right] \cdot \epsilon_k \mathrm{d}t. \tag{A.8}$$

It suffices to add (A.8) to (A.5) in order to check this characterization for the residual. This leads to the following estimate

$$|r(-\epsilon_k)| \leq \left| \int_0^1 \left(\mathcal{S}'(\nu_k + t\epsilon_k) - \mathcal{S}'(\hat{\nu} + t\epsilon_k)\right) \mathrm{d}t \right| |\epsilon_k| = \mathcal{O}\left(|\epsilon_k|^2\right). \tag{A.9}$$

The last equality is valid given that $\mathcal{S}'(\cdot)$ is Lipschitz. Therefore, we have $|\epsilon_{k+1}| = \mathcal{O}\left(|\epsilon_k|^2\right)$ and there exists some constant $c > 0$ that satisfies

$$|\epsilon_{k+1}| \leq c|\epsilon_k|^2. \tag{A.10}$$

If the initial approximation is chosen close enough so that $|\epsilon_0| = |\nu_0 - \hat{\nu}| < \dfrac{\varepsilon}{c}$, where $0 < \varepsilon < 1$, then we can easily show with finite induction that

$$|\epsilon_k| \leq \frac{\varepsilon^{2k}}{c}, \tag{A.11}$$

proving that the sequence $(\nu_k)_{k \in \mathbb{N}}$ converges quadratically to $\hat{\nu}$.

When we drop the assumption on $\mathcal{S}'$ being Lipschitz continuous, through the same analysis of the Taylor expansion of the objective function, the residual is of the form $r(|\epsilon_k|) = o(|\epsilon_k|)$ and we write

$$|\epsilon_{k+1}| = |\epsilon_k| \rho(|\epsilon_k|), \tag{A.12}$$

where the function $\rho(|\epsilon_k|)$ approaches zero whenever $|\epsilon_k| \to 0$. So, given some $0 < \varepsilon < 1$, there exists some $\delta > 0$ such that when $|\epsilon_0| < \delta$, we have that $|\epsilon_1| < \varepsilon|\epsilon_0|$.

Proceeding inductively, we check that $|\epsilon_k| < \varepsilon|\epsilon_{k-1}|$, since $\varepsilon < 1$, and our induction hypothesis holds $|\epsilon_k| < |\epsilon_0| < \delta$, implying that $\rho(|\epsilon_k|) < \epsilon$. With (A.12), this gives the estimate

$$|\epsilon_k| < \varepsilon^k |\epsilon_0|, \tag{A.13}$$

proving the linear convergence. □

Remark A.0.2. In the Chapter III the assumptions of this theorem, for the convergence of the Gauss-Newton method are implied since the derivative of the shooting function is one-to-one. This implies that

$$\mathcal{S}'(\nu_k)^T \mathcal{S}'(\nu_k)\bar{\nu} = 0$$

if, and only if, $\bar{\nu}$ is the null vector and hence $\mathcal{S}'(\nu_k)^T \mathcal{S}'(\nu_k)$ is nonsingular.

B

Computations of Singular Arcs for Optimal Control of SIRS System

In this appendix we expose the computations of the singular vaccination strategies from Section IV.2.2 in full detail. To shorten notation, we define the state vector $x := (N, S, I, C)^T$ and rewrite the dynamics as

$$\dot{x} \quad = f_0(x,u) + v f_1(x) \tag{B.1}$$

where

$$f_0(x,u) = \begin{pmatrix} F(N) - \delta I - \mu N \\ F(N) - \beta \frac{IS}{N} + \omega R - \mu S \\ \beta \frac{IS}{N} - (\delta + \gamma + u + \mu) I \\ B_1 I + B_3 u^2 \end{pmatrix}, \quad f_1(x) = \begin{pmatrix} 0 \\ -S \\ 0 \\ B_2 \end{pmatrix} \tag{B.2}$$

Following the arguments from Section II.2, the singular arcs for the linear controls, *i.e.* vaccination, satisfy the following expression

$$\gamma_{01} + v_{\text{sing}} \gamma_{11} = 0. \tag{B.3}$$

Let us compute the quantities γ_{01} and γ_{11}, as defined in (II.37). Initially note that

$$Df_0 = \begin{pmatrix} F'(N) & 0 & -\delta & 0 \\ F'(N) + \omega + \beta \frac{SI}{N^2} & -\omega - \beta \frac{I}{N} & -\omega - \beta \frac{S}{N} & 0 \\ -\beta \frac{SI}{N^2} & \frac{I}{N} & \beta \frac{S}{N} - (\delta + \gamma + u + \mu) & 0 \\ 0 & 0 & B_1 & 0 \end{pmatrix},$$

$$Df_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

In order to compute γ_{01} and γ_{11}, we start with the Lie bracket $[f_0, f_1]$:

$$[f_0, f_1] = Df_1 f_0 - Df_0 f_1 = \begin{pmatrix} 0 \\ -\left(F(N) + \omega(N - I)\right) \\ \beta \frac{SI}{N} \\ 0 \end{pmatrix}.$$

Notice that $[f_0, f_1]$ does not depend on the nonlinear control u, hence the expressions for γ_{01} and γ_{11} become $p \cdot [f_0, [f_0, f_1]]$ and $p \cdot [f_1, [f_0, f_1]]$, respectively. Computing

$p \cdot [f_1, [f_0, f_1]]$ we obtain

$$
\begin{aligned}
[f_1, [f_0, f_1]] &= D[f_0, f_1]f_1 - Df_1[f_0, f_1] \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 \\ -F'(N) - \omega & 0 & \omega & 0 \\ -\beta\frac{SI}{N^2} & \beta\frac{I}{N} & \beta\frac{S}{N} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ -S \\ 0 \\ B_2 \end{pmatrix} \\
&\quad - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ -\left(F(N) + \omega(N - I)\right) \\ \beta\frac{SI}{N} \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ 0 \\ -\beta\frac{SI}{N} \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ F(N) + \omega(N - I) \\ 0 \\ 0 \end{pmatrix} = -2\begin{pmatrix} 0 \\ 0 \\ \beta\frac{SI}{N} \\ 0 \end{pmatrix} + [f_0, f_1].
\end{aligned}
$$

Using the Goh conditions (II.2.3) and the fact that $\dot{H}_v = 0$, we obtain

$$\gamma_{11} = p \cdot [f_1, [f_0, f_1]] = -2\beta\frac{SIp_I}{N}. \tag{B.4}$$

Moving on to $[f_0, [f_0, f_1]]$, after some algebraic simplifications, we have

$$
\begin{aligned}
&[f_0, [f_0, f_1]] = D[f_0, f_1]f_0 - Df_0[f_0, f_1] \\
&= \begin{pmatrix} 0 & 0 & 0 & 0 \\ -F'(N) - \omega & 0 & \omega & 0 \\ -\beta\frac{SI}{N^2} & \beta\frac{I}{N} & \beta\frac{S}{N} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} F(N) - \delta I - \mu N \\ F(N) - \beta\frac{IS}{N} + \omega R - \mu S \\ \beta\frac{IS}{N} - (\delta + \gamma + u + \mu)I \\ B_1 I + B_3\tau^2 \end{pmatrix} \\
&- \begin{pmatrix} F'(N) & 0 & -\delta & 0 \\ F'(N) + \omega + \beta\frac{SI}{N^2} & -\omega - \beta\frac{I}{N} & -\omega - \beta\frac{S}{N} & 0 \\ -\beta\frac{SI}{N^2} & \frac{I}{N} & \beta\frac{S}{N} - (\delta + \gamma + u + \mu) & 0 \\ 0 & 0 & B_1 & 0 \end{pmatrix} [f_0, f_1] \\
&= \beta\frac{SI}{N}w_1 + w_2 + \left(\beta\frac{I}{N} + \omega\right)[f_1, [f_0, f_1]],
\end{aligned}
$$

where the vectors w_1 and w_2 are given by

$$
w_1 := \begin{pmatrix} \delta \\ 2\omega + \beta\frac{S}{N} \\ -(F(N) - \delta I)/N + 2\beta\frac{I}{N}\frac{p_I}{p_S} \\ -B_1 \end{pmatrix},
$$

$$
w_2 := \begin{pmatrix} 0 \\ -(F'(N) + \omega)(F(N) - \delta I) - \omega I(\delta + \gamma + u) \\ 0 \\ 0 \end{pmatrix}.
$$

Notice the appearance of the term $[f_1, [f_0, f_1]]$, simplifies the final expression of the singular controls since this term cancels out with the denominator γ_{11}. Hence the expression for the singular control becomes

$$v_{\text{sing}} = -\frac{\gamma_{01}}{\gamma_{11}} = -\left(\omega + \beta\frac{I}{N}\right) + \frac{p}{2p_I} \cdot \left(w_1 + \frac{N}{\beta SI}w_2\right). \tag{B.5}$$

Bibliography

[1] V.I. Arnol'd. *Mathematical methods of classical mechanics*. Vol. 60. Springer Science & Business Media, 2013.

[2] M.S. Aronna. "Convergence of the shooting algorithm for singular optimal control problems". In: *Proceedings of the IEEE European Control Conference (ECC)*. 2013, pp. 215–220.

[3] M.S. Aronna. "Second order necessary and sufficient optimality conditions for singular solutions of partially-affine control problems". In: *Discrete Contin. Dyn. Syst. - S* 11.6 (2018), pp. 1179–1199.

[4] M.S. Aronna et al. "Quadratic order conditions for bang-singular extremals". In: *Numer. Algebra, Control Optim., AIMS Journal, special issue dedicated to Professor Helmut Maurer on the occasion of his 65th birthday* 2.3 (2012), pp. 511–546.

[5] M.S. Aronna et al. "Quadratic order conditions for bang-singular extremals". In: *Numer. Algebra, Control Optim., AIMS Journal, special issue dedicated to Professor Helmut Maurer on the occasion of his 65th birthday* 2.3 (2012), pp. 511–546.

[6] D.M. Azimov. "Active sections of rocket trajectories. A survey of research". In: *Avtomat. i Telemekh.* 11 (2005), pp. 14–34.

[7] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Springer Science & Business Media, 2008.

[8] H. Behncke. "Optimal control of deterministic epidemics". In: *Optimal control applications and methods* 21.6 (2000), pp. 269–285.

[9] D.J. Bell and D.H. Jacobson. *Singular Optimal Control Problems*. Academic Press, 1975.

[10] Şİ. Birbil, J.B. Frenk, and G.J. Still. "An elementary proof of the Fritz-John and Karush–Kuhn–Tucker conditions in nonlinear programming". In: *European journal of operational research* 180.1 (2007), pp. 479–484.

[11] J.F. Bonnans and J. Laurent-Varin. "Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control". In: *Numerische Mathematik* 103.1 (2006), pp. 1–10.

[12] J.F. Bonnans, P. Martinon, and V. Grélard. "Bocop-A collection of examples". In: (2012).

[13] J.F. Bonnans et al. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[14] H.J. Bortolossi, M.V. Pereira, and C. Tomei. "Optimal hydrothermal scheduling with variable production coefficient". In: *Math. Methods Oper. Res.* 55.1 (2002), pp. 11–36.

[15] A. Bressan and B. Piccoli. *Introduction to the mathematical theory of control*. Vol. 1. American institute of mathematical sciences Springfield, 2007.

[16] O Certık et al. *SymEngine: A fast symbolic manipulation library, 2020*.

[17] D.I. Cho, P.L. Abad, and M. Parlar. "Optimal production and maintenance decisions when a system experience age-dependent deterioration". In: *Optimal Control Appl. Methods* 14.3 (1993), pp. 153–167.

[18] A.V. Dmitruk. "Quadratic conditions for a weak minimum for singular regimes in optimal control problems". In: *Soviet Math. Doklady* 18.2 (1977).

[19] A.V. Dmitruk and K.K. Shishov. "Analysis of a quadratic functional with a partly degenerate legendre condition". In: *Moscow University Computational Mathematics and Cybernetics* 34.2 (2010), pp. 56–65.

[20] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[21] H. Frankowska and D. Tonon. "Pointwise second-order necessary optimality conditions for the Mayer problem with control constraints". In: *SIAM Journal on Control and Optimization* 51.5 (2013), pp. 3814–3843.

[22] H. Gaff and E. Schaefer. "Optimal control applied to vaccination and treatment strategies for various epidemiological models". In: *Math. Biosci. Eng* 6.3 (2009), pp. 469–492.

[23] R.H. Goddard. *A Method of Reaching Extreme Altitudes*. Vol. 71(2). Smithsonian Miscellaneous Collections. City of Washington: Smithsonian institution, 1919.

[24] B.S. Goh. "Necessary Conditions for the Singular Extremals in the Calculus of Variations". PhD thesis. University of Canterbury, 1966.

[25] B.S. Goh. "Optimal singular rocket and aircraft trajectories". In: *2008 Chinese Control and Decision Conference*. IEEE. 2008, pp. 1531–1536.

[26] B.S. Goh. "The second variation for the singular Bolza problem". In: *J. SIAM Control* 4.2 (1966), pp. 309–325.

[27] B. Gough. *GNU scientific library reference manual*. Network Theory Ltd., 2009.

[28] W.W. Hager. "Runge-Kutta methods in optimal control and the transformed adjoint system". In: *Numerische Mathematik* 87.2 (2000), pp. 247–282.

[29] E. Hairer. *SP N rsett, and G. Wanner. Solving Ordinary Differential Equations I. Nonsti Problems*. Springer Verlag, 1987.

[30] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Vol. 31. Springer Science & Business Media, 2006.

[31] M.W. Hirsch, S. Smale, and R.L. Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2012.

[32] D.G. Hull. "Optimal guidance for quasi-planar lunar ascent". In: *J. Optim. Theory Appl.* 151.2 (2011), pp. 353–372.

[33] H.J. Kelley, R.E. Kopp, and H. G. Moyer. "[Mathematics in Science and Engineering] Topics in Optimization Volume 31 —— 3 Singular Extremals". In: vol. 31. Elsevier, 1967, pp. 63–101.

[34] D.F. Lawden. *Optimal trajectories for space navigation*. Butterworths, London, 1963.

[35] U. Ledzewicz and H. Schättler. "On optimal singular controls for a general SIR-model with vaccination and treatment". In: *Discrete and continuous dynamical systems* 2 (2011), pp. 981–990.

[36] P. Martinon et al. "Numerical study of optimal trajectories with singular arcs for an Ariane 5 launcher". In: *Journal of Guidance, Control, and Dynamics* 32.1 (2009), pp. 51–55.

[37] H. Maurer. "Numerical solution of singular control problems using multiple shooting techniques". In: *J. Optim. Theory Appl.* 18.2 (1976), pp. 235–257.

[38] H. Maurer, J.-H. Kim, and G. Vossen. "On A State-Constrained Control Problem in Optimal Production and Maintenance". In: *Optimal Control and Dynamic Games: Applications in Finance, Management Science and Economics*. Ed. by C. Deissenberg and R.F. Hartl. Springer, 2005, pp. 289–308.

[39] V. Mehrmann. "Existence, uniqueness, and stability of solutions to singular linear quadratic optimal control problems". In: *Linear Algebra and Its Applications* 121 (1989), pp. 291–331.

[40] A.A. Milyutin and N.P. Osmolovskii. *Calculus of variations and optimal control*. Vol. 180. Translations of Mathematical Monographs. Translated from the Russian manuscript by Dimitrii Chibisov. American Mathematical Society, Providence, RI, 1998, pp. xii+372. ISBN: 0-8218-0753-6.

[41] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[42] H.J. Oberle. "Numerical computation of singular control functions in trajectory optimization problems". In: *J. Guidance Control Dynam.* 13.1 (1990), pp. 153–159.

[43] H.J. Oberle. "Numerische Behandlung singulärer Steuerungen mit der Mehrzielmethode am Beispiel der Klimatisierung von Sonnenhäusern". In: *PhD thesis. Technische Universität München* (1977).

[44] H.J. Oberle. "On the numerical computation of minimum-fuel, Earth-Mars transfer". In: *J. Optim. Theory Appl.* 22.3 (1977), pp. 447–453.

[45] H.J. Oberle and K. Taubert. "Existence and multiple solutions of the minimum-fuel orbit transfer problem". In: *J. Optim. Theory Appl.* 95.2 (1997), pp. 243–262.

[46] T.E. Oliphant. "Python for scientific computing". In: *Computing in Science & Engineering* 9.3 (2007), pp. 10–20.

[47] F. Perez, B.E. Granger, and J.D. Hunter. "Python: an ecosystem for scientific computing". In: *Computing in Science & Engineering* 13.2 (2010), pp. 13–21.

[48] L.S. Pontryagin et al. "The mathematical theory of optimal processes (International series of monographs in pure and applied mathematics)". In: *Interscience, New York* (1962).

[49] H. Schättler and U. Ledzewicz. *Geometric optimal control: theory, methods and examples*. Vol. 38. Springer Science & Business Media, 2012.

[50] H. Schättler and U. Ledzewicz. *Optimal control for mathematical models of cancer therapies*. Vol. 42. Springer, 2015.

[51] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. Vol. 12. Springer Science & Business Media, 2013.

[52] H.J. Sussmann and J.C. Willems. "300 years of optimal control: from the brachystochrone to the maximum principle". In: *IEEE Control Systems Magazine* 17.3 (1997), pp. 32–44.

[53] R. Vinter. *Optimal control*. Springer Science & Business Media, 2010.

[54] G. Wanner and E. Hairer. *Solving ordinary differential equations II*. Springer Berlin Heidelberg, 1996.