University of Trento

Dott. Marco Frego

NUMERICAL METHODS FOR OPTIMAL CONTROL PROBLEMS WITH APPLICATION TO AUTONOMOUS VEHICLES

Prof. Enrico Bertolazzi
Prof. Francesco Biral

2014

UNIVERSITY OF TRENTO

Numerical Methods for Optimal Control Problems with Application to Autonomous Vehicles

Ph. D. Head's Prof. Davide Bigoni

Final Examination 07 / 04 / 2014

Board of Examiners
Prof. Oreste Salvatore Bursi (Università degli Studi di Trento)
Prof. Dionisio P. Bernal (Northeastern University, Boston)
Prof. Michel Destrade (National University of Ireland Galway)
Dott. Andrea Giovanni Calogero (Università Milano-Bicocca)
Dott. Paola Falugi (Imperial College London)

SUMMARY - SOMMARIO

In the present PhD thesis an optimal problem suite is proposed as benchmark for the test of numerical solvers. The problems are divided in four categories, classic, singular, constrained and hard problems. Apart from the hard problems, where it is not possible to give the analytical solution but only some details, all other problems are supplied with the derivation of the solution. The exact solution allows a precise comparison of the performance of the considered software. All of the proposed problems were taken from published papers or books, but it turned out that an analytic exact solution was only rarely provided, thus a true and reliable comparison among numerical solvers could not be done before. A typical wrong conclusion when a solver obtains a lower value of the target functional with respect to other solvers is to claim it better than the others, but it is not recognized that it has only underestimated the true value.
In this thesis, a cutting edge application of optimal control to vehicles is showed: the optimization of the lap time in a race circuit track considering a number of realistic constraints. A new algorithm for path planning is completely described for the construction of a quasi G^2 fitting of the GPS data with a clothoid spline in terms of the G^1 Hermite interpolation problem. In particular the present algorithm is proved to work better than state of the art algorithms in terms of both efficiency and precision.

In questa tesi di dottorato di ricerca viene presentata una suite di problemi di controllo ottimo per effettuare un confronto tra software che li risolvono numericamente. I problemi sono stati divisi in quattro categorie, classici, singolari, vincolati e difficili. Tranne che per i problemi difficili, per i quali non è possibile trovare la soluzione analitica a parte qualche dettaglio, tutti gli altri sono stati corredati con la derivazione della soluzione esatta. La sua conoscenza permette di effettuare un confronto preciso sulla performance dei software testati. Tutti i problemi proposti sono stati raccolti da articoli pubblicati o da libri, tuttavia ne è emerso che solo raramente ne veniva presentata anche la soluzione esatta, dunque finora un confronto realistico e corretto non è ancora stato possibile. Una conclusione errata tipica è quella di considerare migliore un software che fornisce un valore del target minore di quello dato da altri, non riconoscendo che sta solamente sottostimando il valore corretto.
In questa tesi è presentata anche un'applicazione di punta del controllo ottimo applicato a veicoli: l'ottimizzazione del tempo minimo sul giro di un veicolo su un tracciato di gara considerando diversi vincoli realistici. Si descrive anche un nuovo algoritmo per il path planning che costruisce un fitting quasi G^2 con clotoidi[1] dei dati GPS sfruttando la soluzione del problema di Hermite di interpolazione G^1. In particolare il presente algoritmo è dimostrato essere migliore degli altri algoritmi stato dell'arte sia in termini di efficienza sia in termini di precisione.

[1] "quasi" means "almost".

UNIVERSITY OF TRENTO - Italy

Title of the Ph.D. Thesis:

Numerical Methods for Optimal Control Problems with Application to Autonomous Vehicles

Tutors:

prof. Enrico Bertolazzi: _____

prof. Francesco Biral: _____

Ph.D. candidate:

dott. Marco Frego: _____

If you have everything under control,
then you are not going fast enough.
M. Andretti

CONTENTS

1

INTRODUCTION AND SCOPE

1.1 STATE OF ART OF NUMERICAL METHODS FOR OCPS

The concept of dynamic optimization is the natural extension of the theory of static optimization. Some classic examples of static optimization problems are represented by the Linear Programming (LP), the Quadratic Programming (QP), the integer/mixed integer programming (MIP), and the most famous case of Nonlinear Programming (NLP). In all these problems, the unknowns variables are defined over the real or integer numbers $\mathbb{R}, \mathbb{Z}$. The theory of dynamic optimization looks instead at problems whose unknowns are real functions. The solution of this kinds of problems goes back to the origin of differential calculus and has become an independent branch of research, first in the Calculus of Variations, and nowadays, in the Optimal Control. The first results are due to Leonhard Euler and to the Bernoulli brothers, who gave the foundations of the calculus of variations. In the second half of the XIX century, other important names of Mathematics contributed to theorems of existence as Jacobi and Weierstrass. The passage from calculus of variations to optimal control, is attributed to the Russian mathematician Lev Pontryagin and to the American Richard Bellman in the Fifties of the last century. The first is the founder of the *indirect* methods based on variational observations, the second discovered the Dynamic Programming Principle of optimality (Dpp) which gave birth to Dynamic Programming (DP). Later, a new family of numerical methods for the solution of optimal control problems was introduced, it is the family of *direct* methods based on the direct transcription of the optimal control problem (Figure 1.1). Suppose to tackle the following optimal

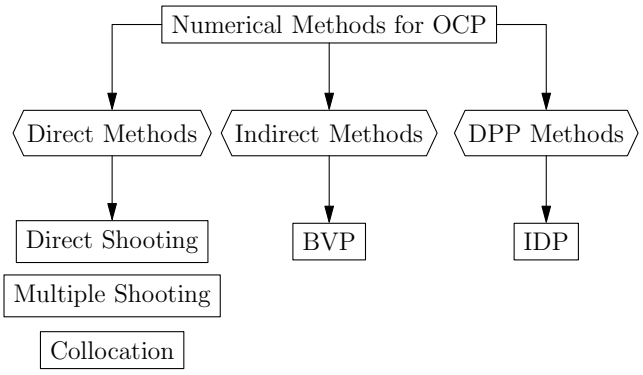


Figure 1.1: The main subdivision of numerical methods for optimal control problems.

control problem, consider the time interval $[a, b]$ and a finite sequence of transition points (corner points) $a < \tilde{s}_1 < \tilde{s}_2 < \cdots < \tilde{s}_{n_d} < b$ and a functional to be minimized

$$\begin{aligned}
\text{minimize: } & \mathcal{J} = \psi(\boldsymbol{x}(a), \boldsymbol{x}(b), \boldsymbol{p}) + \sum_{k=1}^{n_d} \psi_k(\boldsymbol{x}^+(\tilde{s}_k), \boldsymbol{p}) + \int_a^b \mathcal{F}(\boldsymbol{x}(s), \boldsymbol{u}(s), \boldsymbol{p}, s)\, \mathrm{d}s, \\
\text{ODE:} \quad & \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{u}, \boldsymbol{p}, s) = \mathbf{0}, \\
\text{BC:} \quad & \boldsymbol{b}(\boldsymbol{x}(a), \boldsymbol{x}(b), \boldsymbol{p}) = \mathbf{0},
\end{aligned}$$

at corner points there can be jump/transition conditions or various type of constraints

$$\begin{aligned}
\mathbf{S}(\boldsymbol{x}^-(\tilde{s}_k), \boldsymbol{x}^+(\tilde{s}_k), \tilde{s}_k) &= \mathbf{0}, \qquad k = 1, 2, \ldots, n_d, \\
\mathbf{P}_k(\boldsymbol{x}^+(\tilde{s}_k), \boldsymbol{p}, \tilde{s}_k) &= \mathbf{0}, \qquad k = 1, 2, \ldots, n_d
\end{aligned}$$

where

$$\boldsymbol{x}^+(\tilde{s}_k) = \lim_{h \to 0^+} \boldsymbol{x}(\tilde{s}_k + h), \qquad \boldsymbol{x}^-(\tilde{s}_k) = \lim_{h \to 0^+} \boldsymbol{x}(\tilde{s}_k - h).$$

$$\begin{aligned}
\mathbf{S}(\boldsymbol{x}^-(\tilde{s}_k), \boldsymbol{x}^+(\tilde{s}_k), \tilde{s}_k) &= \mathbf{0}, \qquad k = 1, 2, \ldots, n_d, \\
\mathbf{P}_k(\boldsymbol{x}^+(\tilde{s}_k), \boldsymbol{p}, \tilde{s}_k) &= \mathbf{0}, \qquad k = 1, 2, \ldots, n_d
\end{aligned}$$

1.1.1 *Indirect Methods*

The indirect methods are based on the classic theory of calculus of variations and on the famous Pontryagin's Maximum (Minimum) Principle (PMP). Starting from the necessary first order optimality conditions they obtain a two-point (in general a multi-point) boundary value problem. It is derived from the first variation of the Lagrangian function associated to the optimal control problem. An equivalent derivation is possible taking derivatives of the Hamiltonian function. The boundary conditions of this BV problem are given by the initial/final condition given by the problem itself, other are yielded from the transversal condition of the adjoint variables. Of course, by the intrinsic nature of the optimal control problems, a closed form analytical solution is seldom obtained, but the indirect methods can produce it. In presence of path constraint or inequalities it is difficult to apply the PMP to solve for an explicit formula for the control, this leads to state dependent switches. The claimed disadvantage of the indirect method is that the resulting BV problems are difficult to solve. This is not completely true, because today there are various techniques to solve systems of differential equations. It is also mandatory to analyse the computed solution, because it is only extremal but not necessary a minimum. This can be accomplished inspecting the problem (convexity, second variation, etc). The advantages are given by the underlying philosophy of "first optimize, then discretize": the boundary value problem has dimension $2 \times n_x$ where n_x is the number of state variables, therefore even large scale systems are feasible.

1.1.2 *Direct Methods*

A different approach to OCPs is given by the direct methods which follow the philosophy of "first discretize, then optimize" and are somehow the opposite of the indirect methods. Here the state and the control variables are approximated by polynomial interpolation, the target functional itself is approximated by a cost function. Hence the problem is discretized on a mesh, and the optimization variables become the unknowns of a general nonlinear programming problem. There are three main algorithms employed in the application of a direct method, the first is the shooting method (single and multiple) which results in small NLP problems; the second is the pseudospectral method (medium sized problem); the third is the collocation method, which is the most accurate at the price of a very large NLP. The main advantage of the direct methods is that NLPs are widely studied and

a plethora of state of art solution algorithms are available. Moreover it is easier to treat inequality constraints because they have their natural equivalent form in the associated NLP problem. The principal disadvantage is that direct methods produce only suboptimal or approximate solutions. Nowadays they are very popular because they are easy to understand and apply (no calculus of variations needed), they are also robust.

1.1.3 *Dynamic Programming*

The third family of numerical methods to solve an optimal control problem is given by algorithms that make use of the Hamilton-Jacobi-Bellman equation. The idea behind this algorithms is the Principle of Optimality, which states that any subarc of an optimal trajectory is also optimal. A grid $a = t_0 < \ldots < t_N = b$ is introduced over the time interval $[a, b]$, and by the principle of optimality, on each subinterval $[t_k, t_{k+1}]$ the restriction of the functional on that interval is optimized. The resulting partial differential equation is solved recursively backwards starting at the end of the time interval. Advantages of this method are that it searches the whole state space giving a global optimum, can have optimal feedback controls precomputed, admits some analytical solutions (for linear systems with quadratic cost), the so called *viscosity solutions* exist and are feasible for a quite general class of nonlinear problems. The main disadvantage of Dynamic Programming is that the resulting partial differential equation is in a high dimensional space and is in general non tractable. This is what Bellman called the “curse of dimensionality”.

1.2 SCOPE OF THE THESIS

The aim of the present PhD thesis is to propose a suite of optimal control problems together with the derivation of their analytical solution in order to compare the quality of the numerical results given by software numerical solvers. Those analytical solutions allow to understand the difficulties faced by the solvers on some families of problems and give the insight for the design of strategies that enhance the convergence of the numerical methods. The motivation of this study was the validation of the OCP solver XOptima, proposed by the Mechatronic Research Group of the University of Trento. The comparison was done with other three open source software, Acado [HFD11], Gpops [RBD$^+$10], Iclocs [FKvW10]. Acado is developed by the research group lead by M. Diehl at the University of Leuven; Gpops is the solver proposed by A. Rao and his group at the University of Florida, Gainesville and is used, among the others, by NASA; Iclocs is the software presented by F. Falugi of Imperial College London; XOptima is presented by E. Bertolazzi and F. Biral [BBDL03, BBDL05, BBDL07] and the focus of the thesis is to perform a deep test of its features, starting from the easiest classic problems to the well known hardest problems like the Hang Glider Problem [BNPS91], the third order singular problem proposed by Luus [Luu00] that exhibits the chattering phenomenon discovered by Fuller, the optimization of the minimum lap time for a high performance vehicle on a circuit track, the minimum time manoeuver for an underwater vehicle [CSMV04]. With respect to these problems, it is possible to derive the analytic exact solution only for the second one, for the hang glider it is only possible to compare the solution with two cases[1] found in literature (only [BNPS91] and [Bet01]), while for the minimum lap time, we can compare the results with the real laps performed by professional drivers and pilots. A relevant part of the thesis is devoted to the study of singular problems in sight of the analysis of the Fuller problem of third order and its numerical treatment from the Sixties until nowadays. It turns out that the formulation proposed by some authors since the Seventies can not exhibit the chattering phenomenon as claimed, this is shown in the section of the problem Luus n.4. In the present thesis it is recognized, for the first time, that the problem proposed in [FO77] is in facts the

[1] There are other numerical solutions for the Hang Glider problem on the user manuals of other software. They are not considered here because they are not reliable, there is lack of information or some conditions of the problem are violated.

third order Fuller problem with an important modification, and that the solution of Luus is not just a suboptimum, but is instead the true minimum. Finally, analytic and numeric details for a whole family of singular problems are given.
Another contribution was the theoretical study of the second variation (taking into account general initial and final conditions) of the functional to be minimized. The second variation leads to sufficient conditions of optimality that can be checked *a posteriori* to ensure the presence of the minimum point. It would be interesting, as a future work, to implement the derived second order conditions.
While solving OCPs, emerged a new idea on how to solve them with a different approach, which tries to collect the main advantages of the three families of methods described above. As starting point, it is desirable to split the single OCP in segments in order to solve more *smaller* problems (called *boxes*). This is the idea of the DPP and of the direct methods, what they do not have is the possibility to precompute the *global* control once and for the whole problem. The control is obtained via the Pontryagin's Maximum Principle or by solving explicitly the equation $\partial\mathcal{H}/\partial u = 0$ (where $\mathcal{H}$ is the Hamiltonian of the problem), when possible. Another benefit inherited by the indirect method is the additional knowledge of the adjoint variables (costate), that are not considered in the DPP or in some direct methods. They provide a richer differential structure that can be exploited in the solution of the problem.
The single boxes optimize many smaller optimal control problems, while the optimal control is fed

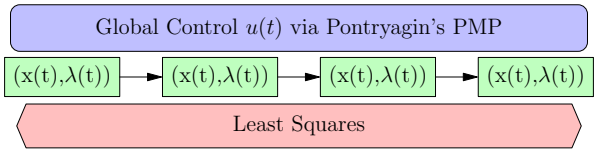


Figure 1.2: The three logical layers of the proposed method.

globally by the Pontryagin's Maximum Principle. The continuity of the functions and the satisfaction of the various constraints are left to the low-level least squares optimizer. In this way, all the knowledge available on the problem is used, everything will be contained in the boxes that will have a fast feedback. The algorithm works like many black boxes that solve a piece of the original OCP only on a small time interval. The boxes are connected in sequence imposing a nonlinear least square problem to provide continuity of the global solution. The aim is to have very efficient boxes that provide quick solutions for high speed computation of manoeuvers for vehicles. This new algorithm is not described here because it is still improving and under verification. We limit ourselves to report the numerical results obtained for some of the benchmark tests with the label "present method".
The third scope of this thesis is the application of the techniques described to the field of autonomous vehicles, XOptima was born to solve the optimization problems that arose while optimizing the models of vehicles and the related environment. In the field of intelligent vehicles the optimal control can be used to formulate and solve many interesting problems such as the motion planning and optimal manoeuver tracking in a receding horizon scheme [BBDL$^+$14]. The first problem (Optimal motion planning) finds the optimal way to drive a vehicle from a point A to point B along a strip of road. It turned out that the description of the road in curvilinear coordinates (i.e. arc length and curvature) is efficient and quite convenient to impose the path constraints. One common way to describe the road shape in curvilinear coordinates is using a clothoid spline which has some good properties, the most important is that the curvature varies linearly with the arc length, making a clothoid spline superior over other polynomial splines. In facts, it was soon recognized, when using polynomials, that the curvature at the extrema of the intervals of the subdivision was unacceptable. Clothoids are widely used in highways design and are herein applied for the description of the geometry of the road. The problem with this transcendent curve is that the numerical computation of its parameters is very unstable (see [BF14]). The quasi G^2 fitting algorithm permits a very smooth trajectory (from the point of view of the curvature). However,

from the practical point of view, the road shape can be derived from the GPS points in cartesian coordinates. The cloud of points is then clustered and fitted with a spline of cubic Bezier curves. From the Bezier spline the G^1 information is gathered and furnished to the clothoid fitting algorithm. The complete solution of the G^1 Hermite interpolation problem with clothoids and with the quasi G^2 interpolation with clothoids is exposed in Chapter 6 and can be found in [BF14]. An open source implementation in Matlab can be found in [BF13].
An example of this techniques is given in the OCP of minimum lap time of a vehicle on a race track: it combines an OCP with a complex dynamic system (both realistic and intrinsically unstable) with a path generated with the above results.

2

STATIC AND DYNAMIC OPTIMIZATION

The concept of *optimization* is nowadays universal. Among the various way we can perform an action, we are looking to the best way to do it, where the idea of "best" can change from situation to situation. In mathematics optimization makes sense if we can describe the object of our investigation with a model, that is, some equations or expressions. In sight of the *Optimal Control Problem* we start with some basic definitions that will help in understanding the successive topics. Once we have a model, depending on the situation, we can be interested in finding maximum or minimum points (or more in general, trajectories) that optimize the model. They are called *extremal* values but they do not need to exist. For example, on $\mathbb{R}$, the function $f(x) = x$ is unbounded and on the open interval $(-1, 1)$, although limited, does not have extremal points. On the other side, on the interval $[-1, 1]$, f assumes both maximum and minimum values. Another important remark arises noticing that it is not enough to restrict the image to a limited set. In facts, depending on the interval considered, a function can have only one extremal value or can assume it at more than one point. The previous examples show that neither compactness nor continuity can alone ensure the existence of extremal values. The presence of both these conditions leads to the theorem of Weierstrass, which can be weakened for the case of semi continuous functions.
It is clear that even with simple functions we need some necessary and sufficient conditions to ensure that f has a minimum or a maximum. The problem of finding the maximum of f is the same of the problem of minimum of $-f$, so we focus only on minimum problems, see Figure 2.1. Let us begin with the discussion on function classes from one to many real variables and then to functionals.

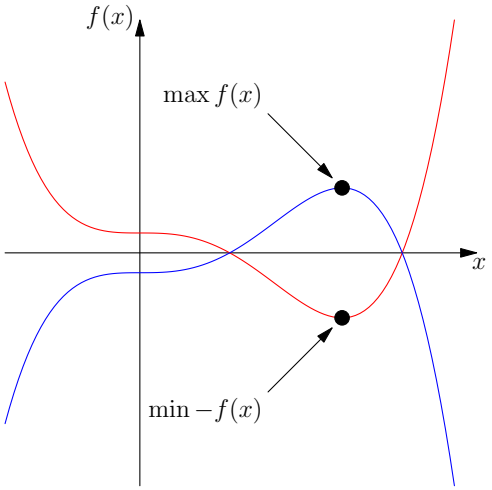


Figure 2.1: The maximum of f is the minimum of $-f$.

2.1 FUNCTIONS OF REAL VARIABLE

The space of work will be $\mathbb{R}^n$ or a subset $\Omega \subseteq \mathbb{R}^n$. We write $x \in \mathbb{R}$ for a real variable, and use bold for vectors or matrices, $\boldsymbol{x} \in \mathbb{R}^n$. The components of vector $\boldsymbol{x}$ are $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, that is, we consider column vectors with the only exception of the gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$, in that case $\nabla f(\boldsymbol{x}) = \left(\dfrac{\partial f}{\partial x_1}, \ldots, \dfrac{\partial f}{\partial x_n} \right)$ is a row vector. In general, the domain is described by some equalities $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ with $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^{n_h}$ and some inequalities $\boldsymbol{g}(\boldsymbol{x}) \geq \boldsymbol{0}$ with $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^{n_g}$. The set Ω is also called the *feasible set* and is defined as

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0},\, \boldsymbol{g}(\boldsymbol{x}) \geq \boldsymbol{0}\}.$$

Depending on the context, we can have two main classes of functions: smooth and non smooth functions. The first one, splits in continuous functions $C^0(\Omega, \mathbb{R})$, with continuous first $C^1(\Omega, \mathbb{R})$ or second $C^2(\Omega, \mathbb{R})$ derivatives. In particular cases we can have even higher derivatives or $C^\infty(\Omega, \mathbb{R})$ functions. We will give only a brief survey of the non smooth case, because in our application there will be some regularity. When dealing with Taylor's expansions, we adopt here the *small* o notation.

Definition 2.1 (small o)**.** *Assume $g(x) \neq 0$ for all $x \neq x_0$ in some interval containing x_0, the notation*

$$f(x) = \mathrm{o}(g(x)) \qquad \text{as } x \to x_0$$

means that

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0.$$

Some texts use the *big* $\mathcal{O}$ notation for the remainder, this means the following. For a function $g(x) \neq 0$ for all $x \neq x_0$ in some interval containing x_0, we say that $f(x) = \mathcal{O}(g(x))$ as $x \to x_0$ if and only if there exist a constant A such that $|f(x)| \leq A|g(x)|$ for all x in a neighbourhood of x_0.

2.1.1 *One Real Variable*

Definition 2.2. *The function f has a* local minimum *at point x_0 if for all $x \in (x_0 - \delta, x_0 + \delta)$ with $\delta > 0$ is $f(x) \geq f(x_0)$. The function f has a* global minimum *at point x_0 on an interval $[a, b]$ if $f(x) \geq f(x_0)$ for all $x \in [a, b]$.*

In this definition there are no differentiability nor continuity assumptions. An easy necessary condition for the smooth case is the following proposition.

Proposition 2.3 (Necessary condition)**.** *The necessary condition for a differentiable function $f(x)$ to have a local minimum at x_0 is*

$$f'(x_0) = 0.$$

An useful sufficient condition to ensure a minimum is given by the next proposition.

Proposition 2.4 (Sufficient condition)**.** *The sufficient condition for a twice differentiable function $f(x)$ to have a local minimum at x_0 is*

$$f'(x_0) = 0, \qquad f''(x_0) > 0.$$

From these easy examples we see that even for a smooth function in one real variable there is not a criterion for a local minimum both sufficient and necessary.
In some situations we do not have smooth functions, not even continuous functions, therefore we need a way to characterize their minima. Loosing for the moment the hypothesis of continuity, we observe that $f : I \to \mathbb{R}$, where $I \subset \mathbb{R}$, has a minimum at x_0 if

$$\inf_I f = f(x_0).$$

It follows that f limited is a necessary condition, but we need a *minimizing sequence* $\{x_n\} \subset I$ such that $x_n \to x$ with

$$\lim_{n\to\infty} f(x_n) = \inf_I f$$

such that there is a convergent subsequence to x_0. We also need a second property, f has to be *lower semi-continuous* at x_0 (see Figure 2.2), i.e. for every $\varepsilon > 0$ there exists a neighbourhood U of x_0 such that $f(x) \geq f(x_0) - \varepsilon$ for all $x \in U$, this can be written as

$$\liminf_{x\to x_0} f(x) \geq f(x_0).$$

Sequential compactness and lower semi-continuity should be both present to ensure the existence

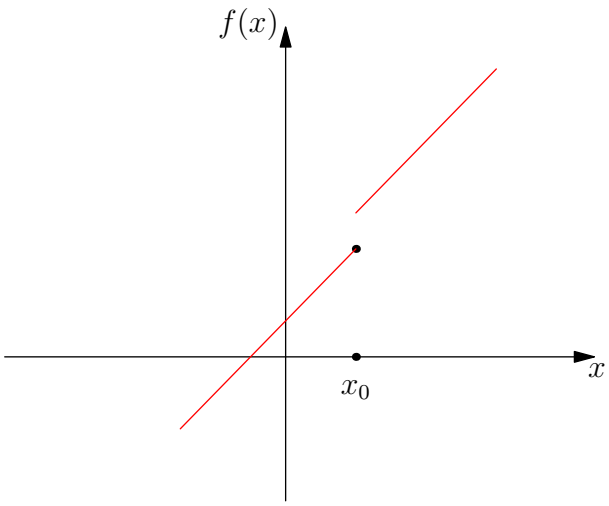


Figure 2.2: A lower semi-continuous function

of a minimum point. With this hypothesis, one can state the theorem of extreme values.

Theorem 2.5 (Weierstrass)**.** *If a function $f : [a, b] \to (-\infty, \infty]$ is lower semi-continuous in $[a, b]$ then f is bounded below and attains its minimum.*

Looking at a computational approach, these results only give existence of the minimum, but in order to find it, when possible, we try to solve $f'(x) = 0$ and check the nature of the stationary points. The *calculus of variations* arises as a generalization of these concepts, and applies to *functionals*. Before introducing functionals and optimal control, we discuss further the minimum problems of real functions.
From standard calculus, using the Taylor expansion of a continuously differentiable function $f(x)$, we have

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \mathrm{o}(\Delta x),$$

where $\mathrm{o}(\Delta x)$ is the Peano's remainder which means that

$$\lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x) - f'(x)\Delta x}{\Delta x} = 0.$$

This expansion extends to functions having m continuous derivatives.

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 + \cdots + \frac{i}{m!}f^{(m)}(x)\Delta x^m + \mathrm{o}(\Delta x^m).$$

If f is twice differentiable, the conditions of minimum can be retrieved as follows,

$$f(x + \Delta x) - f(x) = f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 + \mathrm{o}(\Delta x^2).$$

The right-hand side has the form of a quadratic in Δx, $a\Delta x^2 + b\Delta x + \mathrm{o}(\Delta x^2)$. If $b = f'(x) \neq 0$, for Δx small enough, the sign of $f(x + \Delta x) - f(x)$ is determined by that of $b\Delta x$. If $b > 0$ we have $f(x + \Delta x) - f(x) > 0$ and we arrive back to the necessary condition. Suppose now that $b = f'(x) = 0$, then the term $f''(x)\Delta x^2$ defines the value of the right-hand side when Δx is sufficiently small. So $f''(x) > 0$ is enough to ensure the presence of a minimum.

2.1.2 *Many Real Variables*

Now we extend some of the previous results and definitions to functions of n real variables. Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$.

Definition 2.6. *The function $f : \mathbb{R}^n \to \mathbb{R}$ has a global minimum in $\boldsymbol{x}_0$ if*

$$f(\boldsymbol{x}_0) \leq f(\boldsymbol{x}_0 + \boldsymbol{\Delta x})$$

holds for all nonzero $\Delta \boldsymbol{x} = (\Delta \boldsymbol{x}_1, \ldots, \Delta \boldsymbol{x}_n) \in \mathbb{R}^n$. Point $\boldsymbol{x}_0$ is a local minimum if there exists a radius $r > 0$ such that $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x}_0 + \boldsymbol{\Delta x})$ whenever $||\boldsymbol{\Delta x}|| < r$.

Proposition 2.7 (Necessary condition)**.** *The necessary condition for a differentiable function $f(\boldsymbol{x})$ to have a local minimum at $\boldsymbol{x}_0$ is*

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x})\bigg|_{\boldsymbol{x}=\boldsymbol{x}_0} = 0, \qquad i = 1, \ldots, n.$$

To express the sufficient second order condition, we need the extension of Taylor formula to n variables. Let $f(\boldsymbol{x})$ possess all continuous derivatives up to second order in some neighbourhood of a point $\boldsymbol{x}$ and suppose $\boldsymbol{x} + \Delta\boldsymbol{x}$ lies in this neighbourhood, then

$$\begin{aligned} f(\boldsymbol{x} + t\Delta\boldsymbol{x}) &= f(\boldsymbol{x}) + \left.\frac{\mathrm{d}f(\boldsymbol{x} + t\Delta\boldsymbol{x})}{\mathrm{d}t}\right|_{t=0} t + \frac{1}{2}\left.\frac{\mathrm{d}^2 f(\boldsymbol{x} + t\Delta\boldsymbol{x})}{\mathrm{d}t^2}\right|_{t=0} t^2 + o(t^2) \\ &= f(\boldsymbol{x}) + \sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x})}{\partial x_i}\Delta\boldsymbol{x}_i + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\Delta\boldsymbol{x}_i\Delta\boldsymbol{x}_j + \mathrm{o}(||\Delta\boldsymbol{x}||^2). \end{aligned}$$

Proposition 2.8 (Sufficient condition)**.** *The sufficient condition for a twice differentiable function $f(\boldsymbol{x})$ to have a local minimum at $\boldsymbol{x}_0$ is*

$$\left.\frac{\mathrm{d}^2 f(\boldsymbol{x}_0 + t\Delta\boldsymbol{x})}{\mathrm{d}t^2}\right|_{t=0} > 0$$

for $||\Delta\boldsymbol{x}||$ small enough. The associated quadratic form in the variables Δx_i is

$$\frac{1}{2}\begin{pmatrix}\Delta\boldsymbol{x}_1 & \Delta\boldsymbol{x}_2 & \cdots & \Delta\boldsymbol{x}_n\end{pmatrix}\begin{pmatrix}\frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n}\end{pmatrix}\begin{pmatrix}\Delta\boldsymbol{x}_1 \\ \Delta\boldsymbol{x}_2 \\ \vdots \\ \Delta\boldsymbol{x}_n\end{pmatrix} = \frac{1}{2}\Delta\boldsymbol{x}^T\boldsymbol{H}\Delta\boldsymbol{x}.$$

The matrix $\boldsymbol{H}(\boldsymbol{x})$ is the Hessian matrix[1]*. A sufficient condition for a local minimum is that $\boldsymbol{H}(\boldsymbol{x}_0)$ is Symmetric Positive Defined (SPD).*

2.1.2.1 *Constrained Optimization*

We need further theory when searching for minima on a constrained domain $\Omega \subset \mathbb{R}^n$.

Definition 2.9 (Active constraint)**.** *An inequality constraint $g_i(\boldsymbol{x}) \geq \mathbf{0}$ is called* active constraint *at $\boldsymbol{x}_0 \in \Omega$ if and only if $g_i(\boldsymbol{x}_0) = \mathbf{0}$, and otherwise* inactive*. The index set $\mathcal{A}(\boldsymbol{x}_0) \subset \{1, \ldots, n_g\}$ of active constraints is called* active set*.*

Definition 2.10 (Constraint qualification)**.** *The* linear independence constraint qualification *holds at $\boldsymbol{x}_0 \in \Omega$ if and only if all vectors (for the equalities $\boldsymbol{h}$) $\nabla h_j(\boldsymbol{x}_0)$ for $j = 1, \ldots n_h$ and $\nabla g_i(\boldsymbol{x}_0)$ for $i \in \mathcal{A}(\boldsymbol{x}_0)$ are linearly independent.*

We can now state the famous *Karush-Kuhn-Tucker optimality conditions* as first order necessary and second order sufficient conditions. We introduce here also the *Lagrangian* function. Consider the constrained minimization problem

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{x}) \\ &\text{subject to:} && h_i(\boldsymbol{x}) = 0 \qquad i = 1, 2, \ldots, m \end{aligned}$$

The solution algorithm prescribes to form the Lagrangian function

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x})$$

and to solve the nonlinear system $\nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathbf{0}^T$ with $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$. The for each solution point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ compute $\nabla\boldsymbol{h}(\boldsymbol{x}^\star)$ and check it is full rank, e.g. the rows are linearly independent. Compute

[1] $\boldsymbol{H}(\boldsymbol{x})$ is symmetric if f is smooth enough.

the matrix $\boldsymbol{K}$, the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$, i.e. $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{K} = \boldsymbol{0}$. Then project the reduced Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ in the kernel of the constraints $\boldsymbol{K}$:

$$\boldsymbol{H} = \boldsymbol{K}^T \nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)\boldsymbol{K},$$

A *necessary* condition of optimality is that $\boldsymbol{H}$ is semi positive definite, a *sufficient* condition is that $\boldsymbol{H}$ is positive defined, briefly written, the two conditions are respectively $\boldsymbol{H} \succeq \boldsymbol{0}$ and $\boldsymbol{H} \succ \boldsymbol{0}$. The next theorem proves those conditions.

Theorem 2.11 (Lagrange multipliers)**.** *Let $f \in C^2(\mathbb{R}^n, \mathbb{R})$ a map and $\boldsymbol{x}^\star$ a* local minimum *of $f(\boldsymbol{x})$ satisfying the constraints $\boldsymbol{h} \in C^2(\mathbb{R}^n, \mathbb{R}^m)$, i.e. $\boldsymbol{h}(\boldsymbol{x}^\star) = \boldsymbol{0}$. If $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ is* full rank*, then there exists m scalars λ_k such that*

$$\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) = \nabla f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}^\star) = \boldsymbol{0}^T \tag{A}$$

moreover, for all $\boldsymbol{z} \in \mathbb{R}^n$ which satisfy $\nabla \boldsymbol{h}(vX(s))\boldsymbol{z} = \boldsymbol{0}$ it follows

$$\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\boldsymbol{z} = \boldsymbol{z}^T \left(\nabla^2 f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla^2 h_k(\boldsymbol{x}^\star) \right) \boldsymbol{z} \geq 0 \tag{B}$$

in other words the matrix $\nabla^2_{\boldsymbol{x}}\big(f(\boldsymbol{x}^\star) - \boldsymbol{\lambda} \cdot \boldsymbol{h}(\boldsymbol{x}^\star)\big)$ is semi-SPD in the Kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$.

Proof. Let $\boldsymbol{x}^\star$ be a local minimum, then there exists $\varepsilon > 0$ such that

$$f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}), \qquad \text{for all } \boldsymbol{x} \in B \text{ with } \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \tag{2.1}$$

where $B = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \varepsilon\}$. Consider thus, the functions sequence

$$f_k(\boldsymbol{x}) = f(\boldsymbol{x}) + k \left\|\boldsymbol{h}(\boldsymbol{x})\right\|^2 + \alpha \left\|\boldsymbol{x} - \boldsymbol{x}^\star\right\|^2, \qquad \alpha > 0 \tag{2.2}$$

with the corresponding sequence of (unconstrained) local minima in B:

$$\boldsymbol{x}_k = \operatorname*{argmin}_{\boldsymbol{x} \in B} f_k(\boldsymbol{x}).$$

The sequence $\boldsymbol{x}_k$ is contained in the compact ball B and from compactness, there exists a converging sub-sequence $\boldsymbol{x}_{k_j} \to \bar{\boldsymbol{x}} \in B$. The rest of the proof is to verify that $\bar{\boldsymbol{x}} = X(s)$ and it is a minimum.

Step 1: $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \boldsymbol{0}$. Notice that the sequence $\boldsymbol{x}_k$ satisfy $f_k(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^\star)$, in fact

$$f_k(\boldsymbol{x}_k) \leq f_k(\boldsymbol{x}^\star) = f(\boldsymbol{x}^\star) + k \left\|\boldsymbol{h}(\boldsymbol{x}^\star)\right\|^2 + \alpha \left\|\boldsymbol{x}^\star - \boldsymbol{x}^\star\right\|^2 = f(\boldsymbol{x}^\star).$$

and by definition (2.2) we have

$$\begin{aligned} k_j \left\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\right\|^2 + \alpha \left\|\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\right\|^2 &\leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{k_j}) \\ &\leq f(\boldsymbol{x}^\star) - \min_{\boldsymbol{x} \in B} f(\boldsymbol{x}) = C < +\infty \end{aligned} \tag{2.3}$$

so that

$$\lim_{j \to \infty} \left\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\right\| = 0 \quad \Rightarrow \quad \left\|\boldsymbol{h}\left(\lim_{j \to \infty} \boldsymbol{x}_{k_j}\right)\right\| = \left\|\boldsymbol{h}\left(\bar{\boldsymbol{x}}\right)\right\| = 0 \quad \Rightarrow \quad \boldsymbol{h}(\bar{\boldsymbol{x}}) = \boldsymbol{0}.$$

Step 2: $\bar{\boldsymbol{x}} = \boldsymbol{x}^\star$. From (2.3)

$$\alpha \left\|\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\right\|^2 \leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{k_j}) - k_j \left\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\right\|^2 \leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{k_j})$$

and taking the limit

$$\alpha \left\|\lim_{j\to\infty} \boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\right\|^2 \leq \alpha \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^\star\right\|^2 \leq f(\boldsymbol{x}^\star) - \lim_{j\to\infty} f(\boldsymbol{x}_{k_j}) \leq f(\boldsymbol{x}^\star) - f(\bar{\boldsymbol{x}})$$

From $\|\boldsymbol{h}(\bar{\boldsymbol{x}})\| = 0$ it follows that from (2.1) that $f(\boldsymbol{x}^\star) \leq f(\bar{\boldsymbol{x}})$ and

$$\alpha \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^\star\right\|^2 \leq f(\boldsymbol{x}^\star) - f(\bar{\boldsymbol{x}}) \leq 0$$

and, thus $\bar{\boldsymbol{x}} = \boldsymbol{x}^\star$.

Step 3: Build multipliers. Because $\boldsymbol{x}_{k_j}$ are *unconstrained local minima* for $f_{k_j}(\boldsymbol{x})$, it follows that

$$\nabla f_{k_j}(\boldsymbol{x}_{k_j}) = \nabla f(\boldsymbol{x}_{k_j}) + k_j \nabla \left\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\right\|^2 + \alpha \nabla \left\|\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\right\|^2 = \boldsymbol{0}.$$

Recalling that

$$\begin{aligned} \nabla \left\|\boldsymbol{x}\right\|^2 &= \nabla(\boldsymbol{x}\cdot\boldsymbol{x}) = 2\boldsymbol{x}^T, \\ \nabla \left\|\boldsymbol{h}(\boldsymbol{x})\right\|^2 &= \nabla(\boldsymbol{h}(\boldsymbol{x})\cdot\boldsymbol{h}(\boldsymbol{x})) = 2\boldsymbol{h}(\boldsymbol{x})^T \nabla \boldsymbol{h}(\boldsymbol{x}), \end{aligned}$$

it follows (after transposition)

$$\nabla f(\boldsymbol{x}_{k_j})^T + 2k_j \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \boldsymbol{h}(\boldsymbol{x}_{k_j}) + 2\alpha(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star) = \boldsymbol{0}. \tag{2.4}$$

Left multiplying by $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})$

$$\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \left[\nabla f(\boldsymbol{x}_{k_j})^T + 2\alpha(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star)\right] + 2k_j \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \boldsymbol{h}(\boldsymbol{x}_{k_j}) = \boldsymbol{0}.$$

Now $\nabla \boldsymbol{h}(\boldsymbol{x}^\star) \in \mathbb{R}^{m\times n}$ is full rank for j large by continuity, $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})$ is full rank and thus the matrix $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \in \mathbb{R}^{m\times m}$ is nonsingular, thus

$$2k_j \boldsymbol{h}(\boldsymbol{x}_{k_j}) = -\left(\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right)^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \left[\nabla f(\boldsymbol{x}_{k_j})^T + 2\alpha(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star)\right]$$

taking the limit for $j \to \infty$

$$\lim_{j\to\infty} 2k_j \boldsymbol{h}(\boldsymbol{x}_{k_j}) = -\left(\nabla \boldsymbol{h}(\boldsymbol{x}^\star) \nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T\right)^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}^\star) \nabla f(\boldsymbol{x}^\star)^T = -\boldsymbol{\lambda} \tag{2.5}$$

and taking the limit of (2.4) with (2.5) we have $\nabla f(\boldsymbol{x}^\star)^T - \nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T \boldsymbol{\lambda} = \boldsymbol{0}$.

Step 4: Build a special sequence of $\boldsymbol{z}_j$. We need a sequence $\boldsymbol{z}_j \to \boldsymbol{z}$ such that $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z}_j = \boldsymbol{0}$ for all j. The sequence $\boldsymbol{z}_j$ is built as the projection of $\boldsymbol{z}$ into the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})$, i.e.

$$\boldsymbol{z}_j = \boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z},$$

in facts

$$\begin{aligned} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z}_j &= \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z} \\ &= \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \boldsymbol{z} = \boldsymbol{0} \end{aligned}$$

consider now the limit

$$\begin{aligned}\lim_{j\to\infty} \boldsymbol{z}_j &= \boldsymbol{z} - \lim_{j\to\infty} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z} \\ &= \boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{z}\end{aligned}$$

and thus, if $\boldsymbol{z}$ is in the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$, i.e. $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{z} = \boldsymbol{0}$ we have

$$\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z}_j = \boldsymbol{0} \qquad \text{with} \qquad \lim_{j\to\infty} \boldsymbol{z}_j = \boldsymbol{z}.$$

Step 5: Necessary conditions. Because $\boldsymbol{x}_{k_j}$ are *unconstrained local minima* for $f_{k_j}(\boldsymbol{x})$, it follows that matrices $\nabla^2 f_{k_j}(\boldsymbol{x}_{k_j})$ are semi positive defined, i.e.

$$\boldsymbol{z}^T \nabla^2 f_{k_j}(\boldsymbol{x}_{k_j})\boldsymbol{z} \geq 0, \qquad \forall \boldsymbol{z} \in \mathbb{R}^n$$

moreover

$$\begin{aligned}\nabla^2 f_{k_j}(\boldsymbol{x}_{k_j}) &= \nabla^2 f(\boldsymbol{x}_{k_j}) + k\nabla^2 \left\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\right\|^2 + 2\alpha\nabla(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star) \\ &= \nabla^2 f(\boldsymbol{x}_{k_j})^T + k\nabla^2 \sum_{i=1}^{m} h_i(\boldsymbol{x}_{k_j})^2 + 2\alpha \boldsymbol{I}\end{aligned} \tag{2.6}$$

using the identity

$$\nabla^2 h(\boldsymbol{x})^2 = \nabla(2h(\boldsymbol{x})\nabla h(\boldsymbol{x})^T) = 2\nabla h(\boldsymbol{x})^T \nabla h(\boldsymbol{x}) + 2h(\boldsymbol{x})\nabla^2 h(\boldsymbol{x})$$

in (2.6)

$$\nabla^2 f_{k_j}(\boldsymbol{x}_{k_j}) = \nabla^2 f(\boldsymbol{x}_{k_j}) + 2k_j \sum_{i=1}^{m} \nabla h_i(\boldsymbol{x}_{k_j})^T \nabla h_i(\boldsymbol{x}_{k_j}) + 2k_j \sum_{i=1}^{m} h_i(\boldsymbol{x}_{k_j})\nabla^2 h_i(\boldsymbol{x}_{k_j}) + 2\alpha \boldsymbol{I}.$$

Let $\boldsymbol{z} \in \mathbb{R}^n$, then $0 \leq \boldsymbol{z}^T \nabla^2 f_{k_j}(\boldsymbol{x}_{k_j})\boldsymbol{z}$, that is

$$0 \leq \boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}_{k_j})\boldsymbol{z} + \sum_{i=1}^{m} (2k_j h_i(\boldsymbol{x}_{k_j}))\boldsymbol{z}^T \nabla^2 h_i(\boldsymbol{x}_{k_j})\boldsymbol{z} + 2k_j \left\|\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z}\right\|^2 + 2\alpha \left\|\boldsymbol{z}\right\|^2.$$

The inequality is true for all $\boldsymbol{z} \in \mathbb{R}^n$ and thus for any $\boldsymbol{z}$ in the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$. Choosing $\boldsymbol{z}$ in the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$, from the previous step, the sequence $\boldsymbol{z}_j$ satisfies

$$0 \leq \boldsymbol{z}_j^T \nabla^2 f(\boldsymbol{x}_{k_j})\boldsymbol{z}_j + \sum_{i=1}^{m} (2k_j h_i(\boldsymbol{x}_{k_j}))\boldsymbol{z}_j^T \nabla^2 h_i(\boldsymbol{x}_{k_j})\boldsymbol{z}_j + 2\alpha \left\|\boldsymbol{z}_j\right\|^2$$

and taking the limit $j \to \infty$ with (2.5)

$$0 \leq \boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{z} + \sum_{i=1}^{m} \lambda_i \boldsymbol{z}^T \nabla^2 h_i(\boldsymbol{x}^\star)\boldsymbol{z} + 2\alpha \left\|\boldsymbol{z}\right\|^2.$$

The value of $\alpha > 0$ can be chosen arbitrarily, therefore

$$0 \leq \boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{z} - \sum_{i=1}^{m} \lambda_i \left[\boldsymbol{z}^T \nabla^2 h_i(\boldsymbol{x}^\star)\boldsymbol{z}\right]$$

which is the relation to be proved. □

It is possible to adapt theorem 2.11 for inequality constraints. Consider the NLP problem

$$\begin{aligned} \text{minimize:} \quad & f(\boldsymbol{x}) \\ \text{subject to:} \quad & h_i(\boldsymbol{x}) = 0 \qquad i = 1, 2, \ldots, m \\ & g_i(\boldsymbol{x}) \geq 0 \qquad i = 1, 2, \ldots, p \end{aligned}$$

introducing the *slack* variables e_i, $i = 1, 2, \ldots, p$ and $\boldsymbol{y}^T = (\boldsymbol{x}^T, \boldsymbol{e}^T)$ the new problem

$$\begin{aligned} \text{minimize:} \quad & \mathsf{f}(\boldsymbol{y}) = f(\boldsymbol{x}) \\ \text{subject to:} \quad & \mathsf{h}_i(\boldsymbol{y}) = h_i(\boldsymbol{x}) = 0 \qquad i = 1, 2, \ldots, m \\ & \mathsf{h}_{i+m}(\boldsymbol{y}) = g_i(\boldsymbol{x}) - e_i^2 = 0 \qquad i = 1, 2, \ldots, p \end{aligned}$$

with the Lagrangian function:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k \left(g_k(\boldsymbol{x}) - e_k^2 \right)$$

The first order condition becomes

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \nabla f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}^\star) - \sum_{k=1}^{p} \mu_k \nabla g_k(\boldsymbol{x}^\star) = \boldsymbol{0}^T, \\ \nabla_{\boldsymbol{e}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= 2(\mu_1 e_1, \ldots, \mu_p e_p) = \boldsymbol{0}^T, \\ h_k(\boldsymbol{x}^\star) &= 0, \\ g_k(\boldsymbol{x}^\star) &= e_k^2 \geq 0, \end{aligned}$$

and the second order condition becomes $\boldsymbol{z}^T \nabla^2_{(x,e)} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} \geq 0$ for $\boldsymbol{z}$ in the kernel of the matrix

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \boldsymbol{h}(\boldsymbol{x}^\star) & \boldsymbol{0} \\ \nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}^\star) & 2\,\mathrm{diag}(e_1, \ldots, e_p) \end{pmatrix} \tag{2.7}$$

where

$$\nabla^2_{(\boldsymbol{x},\boldsymbol{e})} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} = \begin{pmatrix} \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) & \boldsymbol{0} \\ \boldsymbol{0} & \nabla^2_{\boldsymbol{e}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \end{pmatrix} \tag{2.8}$$

and $\nabla_{\boldsymbol{x}} \nabla^T_{\boldsymbol{e}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{0}$, moreover

$$\begin{aligned} \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \nabla^2 f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla^2 h_k(\boldsymbol{x}^\star) - \sum_{k=1}^{p} \mu_k \nabla^2 g_k(\boldsymbol{x}^\star), \\ \nabla^2_{\boldsymbol{e}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= 2\,\mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_p). \end{aligned}$$

Notice that $\mu_k e_k = 0$ is equivalent of $\mu_k e_k^2 = 0$ and thus $\mu_k g_k(\boldsymbol{x}^\star) = 0$. So that when $g_k(\boldsymbol{x}^\star) > 0$ then $\mu_k = 0$. Up to a reordering, we split $\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{g}^{(1)}(\boldsymbol{x}) \\ \boldsymbol{g}^{(2)}(\boldsymbol{x}) \end{pmatrix}$ where

$$\begin{aligned} g_k(\boldsymbol{x}^\star) &= e_k^2 = 0, \qquad k = 1, 2, \ldots, r \\ g_k(\boldsymbol{x}^\star) &= e_k^2 > 0, \qquad k = r+1, r+2, \ldots, p \end{aligned}$$

and thus (2.7) becomes

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}^\star) & \mathbf{0} & \mathbf{0} \\ \nabla_{\boldsymbol{x}}\boldsymbol{g}^{(1)}(\boldsymbol{x}^\star) & \mathbf{0} & \mathbf{0} \\ \nabla_{\boldsymbol{x}}\boldsymbol{g}^{(2)}(\boldsymbol{x}^\star) & \mathbf{0} & \boldsymbol{E} \end{pmatrix}, \qquad 2\,\mathrm{diag}(e_{k+1},\ldots,e_p) = \boldsymbol{E}. \tag{2.9}$$

and

$$\nabla^2_{\boldsymbol{e}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu}) = \begin{pmatrix} \boldsymbol{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{M} = 2\,\mathrm{diag}(\mu_1,\mu_2,\ldots,\mu_r) \tag{2.10}$$

The group of constraints $\boldsymbol{g}^{(1)}(\boldsymbol{x}^\star)$ that are zeros are the active constraints. The kernel of (2.9) can be written as

$$\boldsymbol{\mathcal{K}} = \begin{pmatrix} \boldsymbol{K} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \\ -\boldsymbol{E}^{-1}\nabla_{\boldsymbol{x}}\boldsymbol{g}^{(2)}(\boldsymbol{x}^\star)\boldsymbol{K} & \mathbf{0} \end{pmatrix}, \tag{2.11}$$

where $\boldsymbol{K}$ is the kernel of the matrix

$$\begin{pmatrix} \nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}^\star) \\ \nabla_{\boldsymbol{x}}\boldsymbol{g}^{(1)}(\boldsymbol{x}^\star) \end{pmatrix}$$

thus $\boldsymbol{z}$ can be written as the scalar product $\boldsymbol{\mathcal{K}}\boldsymbol{d}$ and thus the second order necessary condition $\boldsymbol{z}^T\nabla^2_{(\boldsymbol{x},\boldsymbol{e})}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{z} \geq 0$ becomes

$$0 \leq \boldsymbol{d}^T\left[\boldsymbol{\mathcal{K}}^T\nabla^2_{(\boldsymbol{x},\boldsymbol{e})}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{\mathcal{K}}\right]\boldsymbol{d}, \qquad \boldsymbol{d}\in\mathbb{R}^s$$

and using (2.11) with (2.8) and (2.10)

$$\begin{aligned}\left[\boldsymbol{\mathcal{K}}^T\nabla^2_{(\boldsymbol{x},\boldsymbol{e})}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{\mathcal{K}}\right] &= \boldsymbol{\mathcal{K}}^T\begin{pmatrix} \nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}\boldsymbol{\mathcal{K}}, \\ &= \begin{pmatrix} \boldsymbol{K}^T\nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{K} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} \end{pmatrix}.\end{aligned}$$

Using the solution algorithm of the equality constrained problem, we have

- Necessary condition: the matrices

  $$\boldsymbol{K}^T\nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{K}, \qquad \text{and} \qquad \boldsymbol{M}$$

  must be semi positive defined. This implies that $\mu_k \geq 0$ for $k = 1,2,\ldots,p$

- Sufficient condition: the matrices

  $$\boldsymbol{K}^T\nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{e},\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{K}, \qquad \text{and} \qquad \boldsymbol{M}$$

  must be positive defined. This implies that $\mu_k > 0$ for the active constraints.

Consider the constrained minimization problem NLP

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{x}) && \\ &\text{subject to:} && h_i(\boldsymbol{x}) = 0 && i = 1,2,\ldots,m \\ & && g_i(\boldsymbol{x}) \geq 0 && i = 1,2,\ldots,p \end{aligned} \tag{2.12}$$

The solution algorithm requires the following steps

- Compute the Lagrangian function:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k g_k(\boldsymbol{x})$$

- Solve the nonlinear system

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \mathbf{0}^T \\ h_k(\boldsymbol{x}) &= 0 \qquad k = 1, 2, \ldots, m \\ \mu_k g_k(\boldsymbol{x}) &= 0 \qquad k = 1, 2, \ldots, p \end{aligned}$$

  and keep only the solutions with $\mu_k^\star \geq 0$ and $g_k(\boldsymbol{x}^\star) \geq 0$.

- For each solution point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ compute $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ with $\nabla g_k(\boldsymbol{x}^\star)$ where $g_k(\boldsymbol{x}^\star) = 0$ are the active constraints with $\mu_k > 0$ and check they are linearly independent.

- Compute matrix $\boldsymbol{K}$ the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ with $\nabla g_k(\boldsymbol{x}^\star)$ where $g_k(\boldsymbol{x}^\star) = 0$ are the active constraints with $\mu_k > 0$.

- Compute the reduced Hessian

$$\boldsymbol{H} = \boldsymbol{K}^T \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) \boldsymbol{K},$$

  - Necessary condition: $\boldsymbol{H}$ is semi-positive definite.
  - Sufficient condition: $\boldsymbol{H}$ is positive definite and $\mu_k > 0$ for all the active constraints.

The following theorem (see [Joh48]) give the necessary conditions for constrained minima. Notice that no condition on the constraints are necessary.

Theorem 2.12 (Fritz John)**.** *If the functions $f(\boldsymbol{x})$, $g_1(\boldsymbol{x})$,..., $g_p(\boldsymbol{x})$, are differentiable, then a necessary condition for $\boldsymbol{x}^\star$ to be a local minimum to problem:*

$$\begin{aligned} &\textit{minimize:} && f(\boldsymbol{x}) \\ &\textit{subject to:} && g_i(\boldsymbol{x}) \geq 0 \qquad i = 1, 2, \ldots, p \end{aligned}$$

is that there exist scalars $\mu_0^\star$, $\mu_1^\star$, $\mu_p^\star$, (not all zero) such that the following inequalities and equalities are satisfied:

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) &= \mathbf{0}^T \\ \mu_k^\star g_k(\boldsymbol{x}^\star) &= 0, \qquad k = 1, 2, \ldots, p; \\ \mu_k^\star &\geq 0, \qquad k = 0, 1, 2, \ldots, p; \end{aligned}$$

where

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k \, g_k(\boldsymbol{x})$$

In [KT51] Kuhn and Tucker showed that if a condition, called the *first order constraint qualification*, holds at $\boldsymbol{x}^\star$, $\boldsymbol{\lambda}^\star$ then λ_0 can be taken equal to 1.

Definition 2.13 (Constraints qualification LICQ)**.** *Let the unilateral and bilateral constraints be $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$, the point $X(s)$ is admissible if*

$$g_k(\boldsymbol{x}^\star) \geq 0, \qquad h_k(\boldsymbol{x}^\star) = 0.$$

The constraints $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$ are qualified *at $\boldsymbol{x}^\star$ if the point $\boldsymbol{x}^\star$ is admissible and the vectors*

$$\{\nabla g_k(\boldsymbol{x}^\star) \ : \ k \in \mathcal{A}(X(s))\} \bigcup \{\nabla h_1(\boldsymbol{x}^\star), \nabla h_2(\boldsymbol{x}^\star), \ldots, \nabla h_m(\boldsymbol{x}^\star)\}$$

are linearly independent*.*

Definition 2.14 (Constraint qualification (Mangasarian-Fromovitz))**.** *The constraints $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$ are* qualified *at $\boldsymbol{x}^\star$, if the point $\boldsymbol{x}^\star$ is admissible and it does not exists a linear combination*

$$\sum_{k \in \mathcal{A}(\boldsymbol{x}^\star)}^{m} \alpha_k \nabla g_k(\boldsymbol{x}^\star) + \sum_{k=1}^{m} \beta_k \nabla h_k(\boldsymbol{x}^\star) = \boldsymbol{0}$$

with $\alpha_k \geq 0$ for $k \in \mathcal{A}(\boldsymbol{x}^\star)$ and α_k with β_k not all 0. In other words, there is not a non trivial linear combination of the null vector such that $\alpha_k \geq 0$ for $k \in \mathcal{A}(\boldsymbol{x}^\star)$.

The next theorems are taken from [NW06].

Theorem 2.15 (First order necessary conditions)**.** *Let $f \in C^1(\mathbb{R}^n)$ and the constraints $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ and $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$. Suppose that $\boldsymbol{x}^\star$ is a local minimum of* (2.12) *and that the constraints qualification LICQ holds at $\boldsymbol{x}^\star$. Then there are Lagrange multiplier vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ such that the following conditions are satisfied at $(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu})$*

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) &= \boldsymbol{0}^T \\
h_k(\boldsymbol{x}^\star) &= 0, \qquad k = 1, 2, \ldots, m; \\
\mu_k^\star g_k(\boldsymbol{x}^\star) &= 0, \qquad k = 1, 2, \ldots, p; \\
\mu_k^\star &\geq 0, \qquad k = 1, 2, \ldots, p;
\end{aligned}$$

where

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k \, h_k(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k \, g_k(\boldsymbol{x})$$

Theorem 2.16 (Second order necessary conditions)**.** *Let $f \in C^2(\mathbb{R}^n)$ and the constraints $\boldsymbol{g} \in C^2(\mathbb{R}^n, \mathbb{R}^p)$ and $\boldsymbol{h} \in C^2(\mathbb{R}^n, \mathbb{R}^m)$. Let $\boldsymbol{x}^\star$ satisfying the First order necessary conditions, a* necessary *condition for $\boldsymbol{x}^\star$ be a* local minimum *is that the $m + p$ scalars (Lagrange Multiplier) of the first order necessary condition satisfy:*

$$\boldsymbol{d}^T \nabla_x^2 \mathcal{L}(X(s), \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \boldsymbol{d} \geq 0$$

for all $\boldsymbol{d}$ such that

$$\begin{aligned}
&\nabla h_k(\boldsymbol{x}^\star) \boldsymbol{d} = 0, && k = 1, 2, \ldots, m \\
&\nabla g_k(\boldsymbol{x}^\star) \boldsymbol{d} = 0, && \text{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \text{ and } \mu_k > 0 \\
&\nabla g_k(\boldsymbol{x}^\star) \boldsymbol{d} \geq 0, && \text{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \text{ and } \mu_k = 0
\end{aligned}$$

Remark 2.17. *The conditions*

$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \textit{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \textit{ and } \mu_k > 0$$
$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \qquad \textit{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \textit{ and } \mu_k = 0$$

restrict the space of direction to be considered. If changed with

$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \textit{if } k \in \mathcal{A}(\boldsymbol{x}^\star)$$

theorems 2.16 is still valid because the necessary condition is tested in a smaller set.

Theorem 2.18 (Second order sufficient conditions)**.** *Let $f \in C^2(\mathbb{R}^n)$ and the constraints $\boldsymbol{g} \in C^2(\mathbb{R}^n, \mathbb{R}^p)$ and $\boldsymbol{h} \in C^2(\mathbb{R}^n, \mathbb{R}^m)$. Let $\boldsymbol{x}^\star$ satisfy the First order necessary conditions, a* sufficient *condition for $\boldsymbol{x}^\star$ be a* local minimum *is that the $m + p$ scalars (Lagrange Multiplier) of the first order necessary condition satisfy:*

$$\boldsymbol{d}^T \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)\boldsymbol{d} > 0$$

for all $\boldsymbol{d} \neq \boldsymbol{0}$ such that

$$\nabla h_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad k = 1, 2, \ldots, m$$
$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \textit{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \textit{ and } \mu_k > 0$$
$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \qquad \textit{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \textit{ and } \mu_k = 0$$

Remark 2.19. *The condition*

$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \qquad \textit{if } k \in \mathcal{A}(\boldsymbol{x}^\star) \textit{ and } \mu_k = 0$$

restrict the space of direction to be considered. If omitted, theorem 2.18 is still valid because the sufficient condition is tested in a larger set.

2.2 FUNCTIONALS

The generalization of the concept of function, that is a special function in which the independent variable is a function itself, is called a *functional*. The object of the calculus of the variations is to find the functions that minimize a given functional. A classic example of a functional is the length of a curve. If we consider a curve in the $(x, y) \subset \mathbb{R}^2$ plane, i.e. a function in the form $y = y(x)$, the total length of the curve in the interval $[a, b]$ is the integral

$$J(y) = \int_a^b \sqrt{1 + y'(x)^2}\, \mathrm{d}x.$$

The general form of a functional in calculus of variations will depend not only on the value of the function $y(x)$ itself, but also on its derivative $y'(x)$,

$$J(y) = \int_a^b f(x, y, y')\, \mathrm{d}x := \int_a^b f[y]\, \mathrm{d}x. \tag{2.13}$$

An important point in seeking an extremum value of a functional, is to establish the class of functions we are dealing with. Different classes of functions have fundamental implications, even in the existence of the extremal values. It would be good to deal with smooth functions, or at least with continuous first derivative, but often in technical applications we encounter just piecewise-continuous functions. This becomes much more evident in the optimal control theory, when

discontinuities produce bang-bang controls.
We consider the functional J defined on a subset $\mathcal{D}$ of a linear space $\mathcal{Y}$. We have to put some care in the choice of the subset $\mathcal{D}$, because, for example $\mathcal{D} = \{y \in C[a,b] \,|\, y(a) = 0,\ y(b) = 1\}$ is not a linear space, while the subsets of vector valued functions with components in $C(\mathbb{R})$ are linear. When in equation (2.13), $f \in C([a,b] \times \mathbb{R}^2)$, then J is defined on $\mathcal{Y} = C^1[a,b]$, because for each function $y \in \mathcal{Y}$, $f(x, y, y') \in C[a,b]$. But when $f \in C([a,b] \times D)$, where $D \subset \mathbb{R}^2$, then J is defined only on a subset of $\mathcal{D} = \{y \in C^1[a,b] \,|\, (y, y') \in D \ \forall\ x \in [a,b]\}$. This shows that there are various situations when we try to optimize a functional J over a subset $\mathcal{D}$ of $\mathcal{Y}$. It is not strange that the natural domain of J can be larger than $\mathcal{D}$ and can be $\mathcal{Y}$ itself.
If $\mathcal{Y}$ is the vector space $\mathbb{R}^n$ it is routine to associate each vector to a real number, given by a norm. If $\mathcal{Y} = C[a,b]$ there are various choices for a norm, $||y||_M = \max_{x\in[a,b]} |y(x)|$ determines the *maximum* norm (see Figure 2.3, $||y||_1 = \int_a^b |y(x)|\,\mathrm{d}x$ is also common, the Euclidean norm $||y||_2$ is difficult to employ because it is nontrivial to apply. Once we have chosen a norm, we define continuity of a functional as follows.

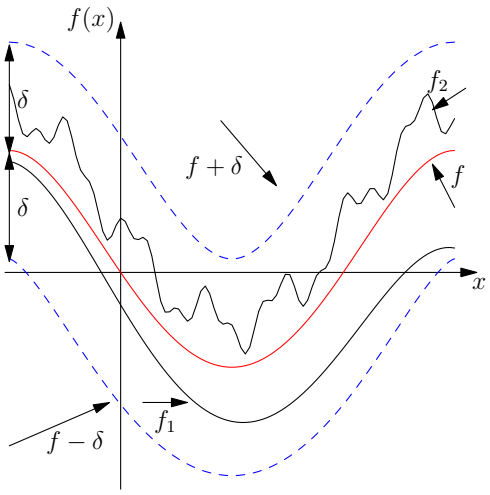


Figure 2.3: The graph of a family of functions $f_i \in \mathcal{Y} = C[a,b]$ uniformly bounded within δ with respect to the graph of f.

Definition 2.20 (continuity for functionals)**.** *In a normed linear space $\mathcal{Y}$, if $\mathcal{D} \subset \mathcal{Y}$, a functional $J : \mathcal{D} \to \mathbb{R}$ is continuous at $y_0 \in \mathcal{D}$ if and only if for each $\varepsilon > 0$ exists a $\delta > 0$ such that $|J(y) - J(y_0)| < \varepsilon$ for all $y \in \mathcal{D}$ with $||y - y_0|| < \delta$.*

In sight of the equivalent version of the theorem of Weierstrass for functionals, the next lemma yields a general result on continuity.

Lemma 2.21. *If K is a compact set in a normed linear space $\mathcal{Y}$, then a continuous functional $J : K \to \mathbb{R}$ is* uniformly *continuous on K, that is, for $\varepsilon > 0$ there exists $\delta > 0$ such that $y, y_0 \in K$ with $||y - y_0|| < \delta$ imply that $|J(y) - J(y_0)| < \varepsilon$.*

It is not a surprise that a continuous functional on a subset of a linear space need not to admit neither a maximum nor a minimum value on this subset, unless compactness is present.

Theorem 2.22 (Weierstrass for functionals)**.** *Let $J : K \to \mathbb{R}$ be a continuous functional on the compact set K, then J assumes both maximum and minimum values at points in K, in particular these values are finite.*

However, often the domains where we have to establish the presence of an extremal value are too large to be compact, hence other techniques are necessary: for example $C[0,1]$ with the

maximum norm is not compact, e.g. $J(y) = y(1)$ is unbounded. In facts the sequence given by $y_n(x) = nx$ for $x \in [0,1]$, for which $J(y_n) = n \to +\infty$ diverges.
As with in the static optimization problems, we are interested in the extremal (minimum) values of a functional J, they occur at $y_0 \in \mathcal{D}$ when

$$J(y_0) \leq J(y) \quad \forall y \in \mathcal{D}.$$

As said before, maximum points of J can be obtained from $-J(y_0) \leq -J(y)$, we can focus only on minimum points.

Proposition 2.23. *An element $y_0 \in \mathcal{D}$ minimizes (globally) J on $\mathcal{D}$ if and only if*

$$J(y_0 + v) - J(y_0) \geq 0 \quad \forall y_0 + v \in \mathcal{D}, \tag{2.14}$$

the equality holds if and only if $v = \mathbf{0}$. Moreover if $c_0, c \neq 0$ are constants, y_0 minimizes also $c^2 J + c_0$.

Example 2.24. *This proposition has interesting applications, consider the functional $J(y) = \int_a^b y'(x)^2 \,\mathrm{d}x$ on the set $\mathcal{D} = \{y \in C^1[a,b] \mid y(a) = 0,\ y(b) = 1\}$. It is clear that $J \geq 0$ and that $J(y) = 0$ for $y' = 0$, but $y = k$ with k constant is not an element of $\mathcal{D}$, therefore we should use equation* (2.14)*:*

$$\begin{aligned} J(y_0 + v) - J(y_0) &= \int_a^b (y_0'(x) + v'(x))^2 - y_0'(x)^2 \,\mathrm{d}x \\ &= \int_a^b v'(x)^2 \,\mathrm{d}x + 2\int_a^b y_0'(x)v'(x) \,\mathrm{d}x \\ &\geq 2\int_a^b y_0'(x)v'(x) \,\mathrm{d}x. \end{aligned}$$

Observing that $0 = y_0(a) = (y_0 + v)(a) = y_0(a) + v(a) = 0$, we have that $v(a) = 0$ and with the same argument $v(b) = 0$. If we try $y_0'(x) = k$ for a constant k, the last integral becomes

$$\int_a^b y_0'(x)v'(x) \,\mathrm{d}x = k\int_a^b v'(x) \,\mathrm{d}x = k(v(b) - v(a)) = 0 \quad \forall v.$$

This shows that equation (2.14) *is satisfied, and one can show that the minimizing function is $y_0(x) = \frac{x-a}{b-a}$. The equality is also satisfied because it is required that $v'(x)^2 = 0 \implies v(x) = \mathrm{const} = v(a) = 0$, hence $v = \mathbf{0}$. The second part of the proposition shows that y_0 minimizes uniquely also*

$$\tilde{J}(y) = 2\int_a^b y'(x)^2 - e^x \,\mathrm{d}x = 2\int_a^b y'(x)^2 \,\mathrm{d}x + 2\int_a^b e^x \,\mathrm{d}x = c^2 J(y) + c_0.$$

The minimization of functional constrained to the level set of a vector valued function $\boldsymbol{h} = \mathbf{0}$ reflects the technique of the Lagrange multiplier for static optimization. We transform the original functional in an augmented one without constraints.

Proposition 2.25. *If the functional J and the function $\boldsymbol{h} = (h_1, \ldots, h_N)^T$ are defined on $\mathcal{D}$, and for some constants $\lambda_1, \ldots, \lambda_N$ the function y_0 minimizes $J_1 = J + \lambda_1 h_1 + \ldots + \lambda_N h_N$ on $\mathcal{D}$, then y_0 minimizes J restricted to the set $H_{y_0} = \{y \in \mathcal{D} \mid h_j(y) = h_j(y_0),\ j = 1, 2, \ldots, N\}$.*

2.2.1 *Gâteaux Variations*

In order to further characterize the extremal values of a functional, we have to introduce the analogue of the partial derivatives for real valued functions. In general we will not have partial derivatives but only directional derivatives, which are called Gâteaux Variations.

Definition 2.26. *The Gâteaux variation of J at y in the direction v for $y, v \in \mathcal{Y}$ is*

$$\delta J(y; v) := \lim_{\varepsilon \to 0} \frac{J(y + \varepsilon v) - J(y)}{\varepsilon} = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} J(y + \varepsilon v)\bigg|_{\varepsilon=0},$$

assuming that the limit exists.

The existence of the limit relies on the definition of $J(y)$ and $J(y + \varepsilon v)$ for sufficently small ε, and on the existence of the ordinary derivative in ε. The variation need not to exist in any direction or it may exist only in some directions. It has the properties of linearity of the standard derivatives. When J is a real function, $\boldsymbol{y}, \boldsymbol{v}$ real vectors, then $\delta J(\boldsymbol{y}; \boldsymbol{v}) = \nabla J(\boldsymbol{y}) \cdot \boldsymbol{v}$ is just the directional derivative of J when $\boldsymbol{v}$ is a unit vector.

Definition 2.27. *In a normed linear space $\mathcal{Y}$, the Gâteaux variations $\delta J(y; v)$ of a real valued functional are said to be* weakly *continuous at $y_0 \in \mathcal{Y}$ if for each $v \in \mathcal{Y}$ we have that $\delta J(y; v) \to \delta J(y_0, v)$ as $y \to y_0$.*

Example 2.28. *Consider $\mathcal{Y} = C[a, b]$ and the functional, $J = \int_a^b y^2(x) + e^x \,\mathrm{d}x$ which is defined for all $y \in \mathcal{Y}$. For fixed $y, v \in \mathcal{Y}$ and $\varepsilon \neq 0$ we have that also $y + \varepsilon v \in \mathcal{Y}$ because it is a linear space, hence*

$$J(y + \varepsilon v) = \int_a^b (y + \varepsilon v)^2(x) + e^x \,\mathrm{d}x$$

is well defined. After some manipulations we have

$$\begin{aligned}
\frac{J(y + \varepsilon v) - J(y)}{\varepsilon} &= \frac{1}{\varepsilon} \int_a^b (y + \varepsilon v)^2(x) - y^2(x) \,\mathrm{d}x \\
&= \frac{1}{\varepsilon} \int_a^b y^2(x) + 2\varepsilon y(x)v(x) + \varepsilon^2 v^2(x) - y^2(x) \,\mathrm{d}x \\
&= 2 \int_a^b y(x)v(x) \,\mathrm{d}x + \varepsilon \int_a^b v^2(x) \,\mathrm{d}x.
\end{aligned}$$

When $\varepsilon \to 0$, the variation becomes

$$\delta J(y; v) = 2 \int_a^b y(x)v(x) \,\mathrm{d}x \qquad \forall y, v \in \mathcal{Y}.$$

The other way to compute $\delta J(y; v)$ is to make use of the explicit formula given in the definition of variation $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} J(y + \varepsilon v)\Big|_{\varepsilon=0}$ and compute

$$\begin{aligned}
J(y + \varepsilon v) &= \int_a^b (y + \varepsilon v)^2(x) + e^x \,\mathrm{d}x \\
&= \int_a^b y^2(x) + e^x \,\mathrm{d}x + 2\varepsilon \int_a^b y(x)v(x) + \varepsilon^2 \int_a^b v(x)^2 \,\mathrm{d}x.
\end{aligned}$$

For fixed y, v the derivative of the previous expression becomes

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} J(y + \varepsilon v) = 2 \int_a^b y(x)v(x) \,\mathrm{d}x + 2\varepsilon \int_a^b v^2(x) \,\mathrm{d}x,$$

thus the variation is

$$\delta J(y;v) = 2\int_a^b y(x)v(x)\,\mathrm{d}x.$$

Although the second method is technically easier, because of the familiarity with usual differentiation methods, it requires that $\mathrm{d}/\,\mathrm{d}\varepsilon J$ exist for small $\varepsilon \neq 0$ and that it be continuous at $\varepsilon = 0$. The first method requires only the existence of the derivative at $\varepsilon = 0$.
We collect in the next table some remarkable variations of functionals that are often encountered. Suppose f or $(x(t), y(t))$ are continuous functions over the appropriate domain.

$J(y)$	$\delta J(y;v)$
$\int_a^b f(x)\sqrt{1+y'(x)^2}\,\mathrm{d}x$	$\int_a^b \frac{f(x)y'(x)v'(x)}{\sqrt{1+y'(x)^2}}\,\mathrm{d}x$
$\int_a^b f[y(x)]\,\mathrm{d}x$	$\int_a^b f_y[y(x)]v(x) + f_{y'}[y(x)]v'(x)\,\mathrm{d}x$
$\int_a^b \sin y(x)\,\mathrm{d}x + y^2(b)$	$\int_a^b v(x)\cos y(x)\,\mathrm{d}x + 2y(b)v(b)$
$\int_0^1 x(t)y(t)\,\mathrm{d}t$	$\int_0^1 x(t)v'(t) + y'(t)u(t)\,\mathrm{d}t$
$\int_a^b f(x,\boldsymbol{y}(x),\boldsymbol{y}'(x))\,\mathrm{d}x$	$\int_a^b f_{\boldsymbol{y}}[\boldsymbol{y}(x)]\boldsymbol{v}(x) + f_{\boldsymbol{y}'}[\boldsymbol{y}(x)]\boldsymbol{v}'(x)\,\mathrm{d}x$
$\int_D \sqrt{1+u_x^2+u_y^2}\,\mathrm{d}A$	$\int_D \frac{u_xv_x+u_yv_y}{\sqrt{1+u_x^2+u_y^2}}\,\mathrm{d}A$
$\int_a^b f(x,y,y',y'')\,\mathrm{d}x$	$\int_a^b f_y[y]v(x) + f_{y'}[y]v'(x) + f_{y''}[y]v''(x)\,\mathrm{d}x$

We conclude this section with some properties of the Gâteaux variations.

Proposition 2.29. *If $\delta J(y;v)$ and $\delta J_1(y;v)$ both exist for $y, v \in \mathcal{Y}$ and supposing $f \in C^1(\mathbb{R})$, then*

- $\delta(JJ_1)(y;v) = \delta J(y;v)J_1(y;v) + J(y;v)\delta J_1(y;v)$,
- $\delta(J/J_1)(y;v) = \dfrac{\delta J(y;v)J_1(y;v) - J(y;v)\delta J_1(y;v)}{J_1(y;v)^2}$,
- $\delta(f(J))(y;v) = f'(J(y))\delta J(y;v)$.
- *If J is a linear functional on $\mathcal{Y}$, its variation is simply $\delta J(y;v) = J(v)$.*

2.2.2 *Convexity*

More about the existence of a minimum point can be said when there is the condition of convexity. We introduce the matter on sets and functions first, then we generalize it to functionals.

Definition 2.30 (Convex set)**.** *A set A is convex if the line segment between any two points in A lies in A, that is, for any $x_1, x_2 \in A$ and for any $\alpha \in [0,1]$,*

$$\alpha x_1 + (1-\alpha)x_2 \in A.$$

A point x of the form $x = \alpha_1 x_1 + \ldots + \alpha_n x_n$ where $\alpha_1 + \ldots + \alpha_n = 1$ with $\alpha_i \geq 0$ and $x_i \in A$ is called a convex combination of the points x_i. A set is convex if and only if it contains every convex combination of its points. The convex hull of a set A is the set of all convex combinations of points in A, formally

$$\mathrm{conv}A := \{\alpha_1 x_1 + \ldots + \alpha_n x_n \mid x_i \in A,\ \alpha_i \geq 0,\ \alpha_1 + \ldots + \alpha_n = 1,\ i = 1,\ldots,n\}.$$

By definition, the convex hull of a set is convex, see Figure 2.4. Some easy examples of convex

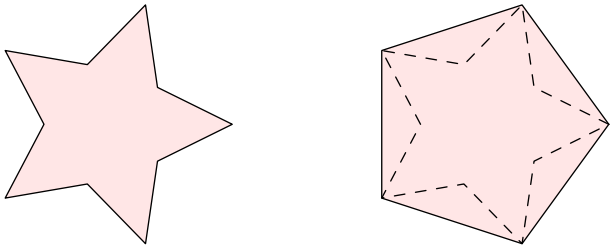


Figure 2.4: Left: the set A (non convex). Right: convex hull of A, $\operatorname{conv} A$ (convex).

sets of $\mathbb{R}^n$ are the empty set, every singleton, the whole space, a line or line segment, a linear subspace.
A hyperplane is a set of the form $\{\boldsymbol{x} \mid \boldsymbol{v}^T\boldsymbol{x} = \boldsymbol{b}\}$ for $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v} \neq \mathbf{0}$ and $\boldsymbol{b} \in \mathbb{R}$, i.e. is the solution of a linear equation among the components of $\boldsymbol{x}$. It can be thought as the set of points that are orthogonal to the vector $\boldsymbol{v}$ and $\boldsymbol{b}$ is the offset of the hyperplane from the origin. Finally, a hyperplane divides the space in two halfspaces, see Figure 2.5. Intersection of convex sets is still convex, as a

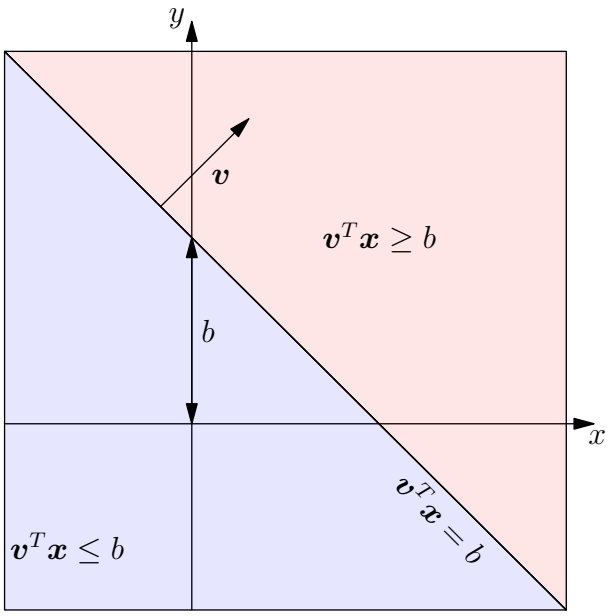


Figure 2.5: The hyperplane $\boldsymbol{v}^T\boldsymbol{x} = b$ cutting $\mathbb{R}^2$ in two halfspaces.

direct consequence, a polyhedron obtained as the intersection of halfspaces and hyperplanes is convex. Convexity is also preserved by isometries, that is, scaling and translation. The projection of a convex set onto some of its coordinates is convex.
A basic property that connects convex sets with hyperplanes is the theorem of Hahn-Banach, or separating hyperplane theorem.

Theorem 2.31. *If C and D are two convex sets that do not intersect, $C \cap D = \emptyset$, then there exists $\boldsymbol{v} \neq \mathbf{0}$ and b such that $\boldsymbol{v}^T\boldsymbol{x} \leq b$ for all $\boldsymbol{x} \in C$ and $\boldsymbol{v}^T\boldsymbol{x} \geq b$ for all $\boldsymbol{x} \in D$. The hyperplane $\boldsymbol{v}^T\boldsymbol{x} = b$ is called a separating hyperplane.*

If the two convex sets are disjoint, there is an hyperplane orthogonal to the shortest segment that connects two points of each set and bisecting it. In this case the inequality conditions are strict and it is called *strict* separation. Notice that in general disjoint sets need not to be strictly separable. An hyperplane that is tangent to a set C at one boundary point $x_0 \in \partial C$ is called a

supporting hyperplane. If a set is closed with nonempty interior and has a supporting hyperplane at every point in its boundary, then it is convex.
It is now natural to introduce convex functions: there are various definition that employ weaker or stronger conditions on the function. We choose a definition that does not require differentiability and is intuitive.

Definition 2.32 (convex function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if the domain of f is a convex set and if for all $\boldsymbol{x}, \boldsymbol{y}$ and for all $0 \leq \alpha \leq 1$ holds*

$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$$

which is called the Jensen *inequality. A function f is said to be concave if $-f$ is convex. If the Jensen inequality is strict, then f is called strictly convex.*

The inequality can be extended to any convex combination of points, if f is convex, for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in its domain, and for scalars $\alpha_1, \ldots, \alpha_n$ such that $\alpha_i \geq 0$ and $\alpha_1 + \ldots + \alpha_n = 1$, then

$$f(\alpha_1 \boldsymbol{x}_1 + \ldots + \alpha_n \boldsymbol{x}_n) \leq \alpha_1 f(\boldsymbol{x}_1) + \ldots + \alpha_n f(\boldsymbol{x}_n).$$

Remark 2.33. *From this inequality, many inequalities can be derived, for example the simple arithmetic-geometric mean, for $a, b \geq 0$, $\sqrt{ab} \leq (a+b)/2$; Hölder inequality, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and p, q dual norms, $||\boldsymbol{x} \cdot \boldsymbol{y}||_1 \leq ||\boldsymbol{x}||_p ||\boldsymbol{y}||_q$; the general arithmetic-geometric mean, for $a, b \geq 0$ and $0 \leq \alpha \leq 1$, $a^\alpha b^{1-\alpha} \leq \alpha a + (1-\alpha) b$.*

From a geometric point of view, a convex function has every chord from $f(\boldsymbol{x})$ to $f(\boldsymbol{y})$ above its graph. A useful property of convex functions is that they remain convex when restricted to any line intersecting their domains. Thus f is convex if and only if for all $\boldsymbol{v}$ the function $g(t) = f(\boldsymbol{x} + t\boldsymbol{v})$ is convex. This represents a practical test to check convexity, allowing to consider only the restriction to a line of a certain function.
When f is defined on a convex set and is also differentiable, then the condition of convexity can be stated with the inequality

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}).$$

The right hand side represents the Taylor expansion of first order, or, geometrically, the supporting hyperplane at $\boldsymbol{x}$. As before, if the inequality is strict, then we speak of strict convexity. For a concave function, the previous inequality changes the sign.
If f is twice differentiable, defined on an open convex domain, convexity is assured if

$$\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}.$$

The information given by the Hessian matrix is about the local curvature of f: positive eigenvalues are equivalent to positive curvature, hence they imply the presence of a minimum point. Strict convexity and concavity follow in the same way. We must point out that $\nabla^2 f \succ \mathbf{0}$ implies convexity, but the converse is not true, as the simple function $f(x) = x^4$ shows: it is strictly convex but has $f''(0) = 0$.

Remark 2.34. *The hypothesis that the domain of f is convex is important for first and second order conditions, for example $f(x) = \frac{1}{x^2}$ for $x \neq 0$ satisfies $f''(x) > 0$ but it is not convex.*

Some examples of elementary convex functions are the exponential e^x; the powers x^p for positive x and $p \geq 1$ or $p \leq 0$ (concave for $0 \leq p \leq 1$); the logarithm is concave. Some important convex functions are the norms of $\mathbb{R}^n$; the maximum function of $\mathbb{R}^n$ $\max\{x_1, \ldots, x_n\}$; squared functions over linear function as x^2/y on domain of kind $\{(x, y) \mid \ y > 0\}$; the Log-Sum-Exp function $\log(e^{x1} + \ldots + e^{x_n})$; the geometric mean $(\prod_{i=1}^n x_i)^{1/n}$.

From those simple functions we can build new convex functions via operations that preserve convexity: non negative weighted sums, for non negative weights $w_i \geq 0$, $f = w_1 f_1 + \ldots + w_n f_n$; similarly, a non negative weighted sum of concave functions is concave; if $f(x, y)$ is convex in x for each y and $w(y) \geq 0$, then $g(x) = \int w(y) f(x, y) dy$ is convex; the image of a convex set under a linear map is convex. The composition $g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ of a convex function f with an affine map $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is also convex. The pointwise maximum or supremum of two convex functions is convex, $f(x) = \max\{f_1(x), f_2(x)\}$; this property can be generalized to the maximum or supremum of n convex functions. An application of the last point is the distance to farthest point of a set (in any norm).
An important application is devoted to least squares. Let $\boldsymbol{x}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_n \in \mathbb{R}^m$, minimize the objective function

$$\sum_{i=1}^{n} w_i (\boldsymbol{v}_i^T \boldsymbol{x} - \boldsymbol{b}_i)^2$$

where the w_i are the weights that can be negative. Defining

$$g(\boldsymbol{w}) = \inf_x \sum_{i=1}^{n} w_i (\boldsymbol{v}_i^T \boldsymbol{x} - \boldsymbol{b}_i)^2,$$

g is the infimum of a family of linear functions of $\boldsymbol{w}$ and is a concave function of $\boldsymbol{w}$ (see [BV04]). Another remarkable example is the norm of a matrix (convexity follows from the supremum of linear functions).
We give now some general results of convexity when composing functions, we begin with the scalar case. Let $h, g : \mathbb{R} \to \mathbb{R}$, for twice differentiable g and h. Define $f(x) = (h \circ g)(x) = h(g(x))$. The f is convex if and only if $f''(x) \geq 0 \ \forall \ x$. This can be expanded as

$$f''(x) = h''(g(x)) g'(x)^2 + h'(g(x)) g''(x).$$

It can be proved the next result, which it turns out to hold in the general case of $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ (with the note that h should be monotone and extended-valued to $+\infty$ for points not in its domain):

$$\begin{array}{ccccc} h & & g & & f \\ \text{convex and nondecreasing} & & \text{convex} & \Longrightarrow & \text{convex} \\ \text{convex and nonincreasing} & & \text{concave} & \Longrightarrow & \text{convex} \\ \text{concave and nondecreasing} & & \text{concave} & \Longrightarrow & \text{concave} \\ \text{concave and nonincreasing} & & \text{convex} & \Longrightarrow & \text{concave} \end{array} \tag{2.15}$$

Example 2.35. *As an example consider $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$ and $h : [1, 2] \to \mathbb{R}$, $h(x) = 0$. In this case g is convex and h is convex nondecreasing. Define $f = h \circ g$, which has the domain $[-\sqrt{2}, -1] \cup [1, \sqrt{2}]$ and is $f(x) = 0$. Here f is not convex because the domain is not convex, the problem here is that h is not nondecreasing outside its domain, that is, the* extended valued h *should be nondecreasing, not just over its domain.*

We turn now to the general vector composition, let $h : \mathbb{R}^k \to \mathbb{R}$ and $g_i : \mathbb{R} \to \mathbb{R}$ where

$$f(x) = h(\boldsymbol{g}(x)) = h(g_1(x), \ldots, g_k(x)).$$

The formal expression of the condition of convexity can be expressed as

$$f''(x) = \boldsymbol{g}'(x)^T \nabla^2 h(\boldsymbol{g}(x)) \boldsymbol{g}'(x) + \nabla h(\boldsymbol{g}(x))^T \boldsymbol{g}''(x) > 0.$$

The conclusion of table (2.15) are still valid, but considering $\boldsymbol{g}$ convex/concave in each component. We extend the concept of convex function to functionals: the directional derivatives are substituted by the Gâteaux variations. In the case of $f \in C^1(\mathbb{R}^3)$, we have already seen that $\delta f(\boldsymbol{y}; \boldsymbol{v}) = \nabla f(\boldsymbol{y}) \cdot \boldsymbol{v}$, and convexity is provided by the condition

$$f(\boldsymbol{y} + \boldsymbol{v}) - f(\boldsymbol{y}) \geq \nabla f(\boldsymbol{y}) \cdot \boldsymbol{v} = \delta f(\boldsymbol{y}; \boldsymbol{v}).$$

Strict convexity is present if the previous relation in an equality if and only if $\boldsymbol{v} = \boldsymbol{0}$.

Definition 2.36 (convexity for functionals)**.** *A real valued functional J defined on $\mathcal{D}$ in a linear space $\mathcal{Y}$ is said to be convex on $\mathcal{D}$ provided that when y and $y + v \in \mathcal{D}$ then $\delta J(y; v)$ is defined and*

$$J(y + v) - J(v) \geq \delta J(y; v). \tag{2.16}$$

Definition 2.37 (strict convexity for functionals)**.** J *is* strictly *convex if* (2.16) *is an equality if and only if $\boldsymbol{v} = \boldsymbol{0}$.*

An useful property of convex functionals is that if J and J_1 are convex and $c \in \mathbb{R}$, $c > 0$, then $J + J_1$ and cJ are also convex.

Proposition 2.38. *If J is convex on $\mathcal{D}$, then each $y \in \mathcal{D}$ for which $\delta J(y; v) = 0$ minimizes J on $\mathcal{D}$. Moreover, if J is strictly convex, then the minimizer is unique.*

Example 2.39. *Consider $\mathcal{Y} = C[a, b]$ and the functional, $J = \int_a^b y^2(x) + e^x \, \mathrm{d}x$. J is strictly convex because $\delta J(y; v) = 2 \int_a^b y(x)v(x) \, \mathrm{d}x$ and*

$$J(y + \varepsilon v) - J(y) = 2 \int_a^b y(x)v(x) \, \mathrm{d}x + \int_a^b v^2(x) \, \mathrm{d}x \geq 2 \int_a^b y(x)v(x) \, \mathrm{d}x = \delta J(y; v).$$

The equality holds if and only if $\int_a^b v^2(x) \, \mathrm{d}x = 0$, that is $v = 0$. Therefore, for the previous proposition the function y such that $\delta J(y; v) = 0$ minimizes uniquely the functional.

It is an application of the definition to obtain that a linear functional is convex but not strictly convex.
The general case is a convex integral functional of the form

$$F(y) = \int_a^b f[y(x)] \, \mathrm{d}x := \int_a^b f(x, y(x), y'(x)) \, \mathrm{d}x,$$

which has the variation

$$\delta F(y; v) = \int_a^b f_y[y(x)]v(x) + f_{y'}[y(x)]v'(x) \, \mathrm{d}x.$$

Convexity implies that $J(y + v) - J(v) \geq \delta J(y; v)$, i.e.

$$\int_a^b f[y(x) + v(x)] - f[y(x)] \, \mathrm{d}x \geq \int_a^b f_y[y(x)]v(x) + f_{y'}[y(x)]v'(x) \, \mathrm{d}x, \tag{2.17}$$

this yields the pointwise relation

$$f[y(x) + v(x)] - f[y(x)] \geq f_y[y(x)]v(x) + f_{y'}[y(x)]v'(x).$$

This shows that f is convex when x is held fixed, a kind of partial convexity which is essential in the development of the theory and leads to the definition of strong convexity.

Definition 2.40 (convexity for integral functionals)**.** *Let x be fixed, $f(x, y, y')$ is said to be convex if f and its partial derivatives f_y and $f_{y'}$ are defined and continuous and satisfy the inequality*

$$f[y(x) + v(x)] - f[y(x)] \geq f_y[y(x)]v(x) + f_{y'}[y(x)]v'(x). \tag{2.18}$$

If the equality holds if and only if $v = 0$ or $v'(x) = 0$, then we speak of strong *convexity.*

It is clear that if f is convex by itself, then also fixing x yields a convex function. For the same reason, if f is strictly convex, then fixing x yields a strong convex function.

2.2.3 *The Two Equation of Euler-Lagrange*

Theorem 2.41. *Let D be a subset of $\mathbb{R}^2$, let a_1, b_1 be such that*

$$\mathcal{D} = \{y \in C^1[a, b] \quad | \quad y(a) = a_1,\ y(b) = b_1,\ (y(x), y'(x)) \in D\}.$$

If (fixing x) f is convex on $[a, b] \times D$ then

$$F(y) = \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x$$

is convex on $\mathcal{D}$. Moreover, strong convexity of f implies strict convexity of F. Each $y \in \mathcal{D}$ for which

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{y'}[y(x)] = f_y[y(x)], \tag{2.19}$$

on (a, b), minimizes F on $\mathcal{D}$ (uniquely if f is strongly convex). Equation (2.19) *is called the* first *Euler-Lagrange equation.*

Proof. A sketch, see [Tro96]. Integrating inequality (2.18), gives (2.17) or $F(y + v) - F(y) \geq \delta F(y; v)$, that is, F is convex. Each function that satisfies (2.19) allows to write

$$\delta F(y; v) = \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} \left(f_{y'}[y(x)]v(x)\right)\,\mathrm{d}x = f_{y'}[y(x)]v(x)\Big|_a^b = 0,$$

and by Proposition 2.38, y minimizes F. We remark that neither of the convexity implications of this theorem is reversible. □

The next example shows that strong convexity is weaker than strict convexity.

Example 2.42. *The functional $f(y, y') = y'(x)^2 + 4y(x)$ is strongly convex even if it is not strictly convex. First we check that F is not strictly convex, we have that*

$$\begin{aligned} F(y + v) - F(y) \ &= \int_0^1 2y'(x)v'(x) + v'(x)^2 + 4v(x)\,\mathrm{d}x \\ &\geq \delta F(y; v) = \int_0^1 2y'(x)v'(x) + 4v(x)\,\mathrm{d}x. \end{aligned}$$

Strict convexity requires that the previous inequality becomes an equality if and only if $v(x) = 0$, but $\int_0^1 v'(x)^2\,\mathrm{d}x = 0$ for each constant $v(x)$, so F is not strictly convex. Strong convexity implies that the inequality between integrands becomes an equality, this time allowing directly $v'(x) = 0$ hence strong convexity is present.
Suppose now to minimize

$$F(y) = \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x = \int_0^1 y'(x)^2 + 4y(x)\,\mathrm{d}x$$

on the set $\mathcal{D} = \{y \in C^1[0,1] \mid y(0) = 0,\ y(1) = 1,\ (y(x), y'(x)) \in \mathbb{R}^2\}$*. The set* $\mathcal{D}$ *implies strict convexity of* f*, because the only admissible variations* v *must satisfy* $v(0) = v(1) = 0$*, thus the only constant function that makes the inequality an equality is* $v(x) = 0$*, so in this case* f *is also strictly convex. The hypotheses of the previous theorem 2.41 are respected and* F *is minimized uniquely by a solution of the equation* (2.19) *for* $0 < x < 1$*. Such equation takes the form*

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{y'}[y(x)] = f_y[y(x)] \implies 2y''(x) = 4,$$

and has the general solution $y(x) = x^2 + \alpha x + \beta$*, for some constants* $\alpha, \beta \in \mathbb{R}$*. Solving the associated boundary value problem, it is easy to find* $a = b = 0$ *and* $y(x) = x^2$*.*

Depending on the explicit dependence of f from y or y', the Euler-Lagrange equation reduces to some special cases. If $f = f(x, y')$, that is when $f_y = 0$, then convexity is characterized by

$$f(x, y'(x) + v'(x)) - f(x, y'(x)) \geq f_{y'}(x, y'(x))v'(x),$$

and the Euler-Lagrange equation becomes

$$f_{y'}(x, y'(x)) = \text{const.} \tag{2.20}$$

If $f = f(y')$ only, the equation reduces to $f_{y'}(x) = \text{const}$ and the function $y(x) = m(x - a) + a_1$, for $m = \frac{b_1 - a_1}{b - a}$, minimizes $F(y) = \int_a^b f(y'(x))\,\mathrm{d}x$.
If x is fixed and $f = f(x, y)$ is convex on $[a, b] \times \mathbb{R}$, each $y \in C[a, b]$ that satisfies $f_y(x, y(x)) = \text{const}$ minimizes $F(y) = \int_a^b f(x, y(x))\,\mathrm{d}x$; if strong convexity is present, the minimizer is unique. If $f = f(y, y')$ the equation of Euler-Lagrange reduces to $f(y, y') - y'(x)f_{y'}(x) = \text{const}$.
When $f = f(x, y, y')$ is $C^1[a, b]$ and y is solution of the first Euler-Lagrange equation (2.19), the integration of the first equation yields

$$f_{y'}(x) = \int_a^x f_y(t)\,\mathrm{d}t + \text{const}.$$

When y is C^2, with the usual abuse of notation, we have

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x, y, y') = f_x(x, y, y') + f_y(x)y'(x) + f_{y'}y''(x) = f_x(x) + \frac{\mathrm{d}}{\mathrm{d}x}(y'(x)f_{y'}(x)),$$

in facts, by the chain rule, $\frac{\mathrm{d}}{\mathrm{d}x}(y'(x)f_{y'}(x)) = y''f_{y'} + y'\frac{\mathrm{d}}{\mathrm{d}x}f_{y'}$, but by the Euler-Lagrange equation (2.19), we can replace $\frac{\mathrm{d}}{\mathrm{d}x}f_{y'}$ with f_y, hence

$$\frac{\mathrm{d}}{\mathrm{d}x}(f(x, y, y') - y'(x)f_{y'}(x)) = f_x(x),$$

or, integrating the above expression,

$$f(x, y, y') - y'(x)f_{y'}(x) = \int_a^x f_x(t)\,\mathrm{d}t + \text{const}.$$

These properties lead to the second equation of Euler-Lagrange as exposed in the next proposition.

Proposition 2.43. *Let*

$$J(y) = \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x$$

and $\mathcal{D} = \{y \in C^1[a,b] \mid y(a) = a_1,\ y(b) = b_1\}$*. If* $f \in C^1[a,b] \times \mathbb{R}^2$ *and* $y \in \mathcal{D}$ *is a local extremal function for* J *on* $\mathcal{D}$*, then on* $[a,b]$*,* y *satisfies the* second *Euler-Lagrange equation*

$$f(x,y,y') - y'(x) f_{y'}(x) = \int_a^x f_x(t)\,\mathrm{d}t + c, \tag{2.21}$$

for some constant c*.*

When $f = f(y,y')$, a local extremal function must also satisfy the equation

$$\frac{\mathrm{d}}{\mathrm{d}x}(f(x,y,y') - y'(x) f_{y'}(x)) = 0,$$

without additional smoothness assumptions.
In order to apply these results, we need tools to check if a function is convex or not. One criterion resembles the condition for convexity of a function defined on $\mathbb{R}$, $f'' > 0$. We begin with the case of $f = f(x,y')$.

Proposition 2.44. *Let* $f = f(x,y')$ *and* $f_{y'y'}$ *be continuous on* $[a,b] \times I$*, and for each* $x \in [a,b]$*,* $f_{y'y'}(x,y') > 0$ *except possibly for a finite number of* y' *values, then, fixing* x*,* $f(x,y')$ *is strongly convex on* $[a,b] \times I$*. If for some* $x \in [a,b]$*,* $f_{y'y'}(x,y') = 0$*, then* $f_{y'}$ *is increasing along* y' *but not strictly, so* $f(x,y')$ *is only convex.*

Example 2.45. *Let* $g(x) > 0$ *be a continuous function on* $[a,b]$*,* $\alpha \neq 0$*, then the functional* $f(x,y') = g(x)\sqrt{\alpha^2 + y'(x)^2}$ *is strongly convex. The functional* $f(y') = -\sqrt{1 - y'(x)^2}$ *is also strongly convex on* $(-1,1)$*. Instead,* $f(x,y') = e^x y'(x)$ *is only convex and* $f(x,y') = x^2 - y'(x)^2$ *is never convex (but* $-f$ *is strongly convex).*

There is not such an easy criterion for functionals that depend explicitly on y. Often we can combine some elementary facts to obtain convexity of elaborated functionals: the sum of convex functionals is again convex; suppose to fix x, then for each $g(x) > 0$ the product $g(x)f(x,y,y')$ (for a convex $f(x,y,y')$) is convex (strong convexity is preserved); $g_1(x) + g_2(x)y(x) + g_3(x)y'(x)$ is only convex for continuous functions g_1, g_2, g_3; each convex function $f(x,y)$ or $f(x,y')$ is also convex when considered as $f(x,y,y')$ on an appropriate set.

Example 2.46. *The functional* $f(x,y,y') = -2\sin(x)y(x) + y'(x)^2$ *is strongly convex on* $\mathbb{R} \times \mathbb{R}^2$*, in fact it can be seen as the sum of the strongly convex function* $y'(x)^2$ *with the convex function* $-2\sin(x)y(x)$ *(recall that* x *is fixed). With the same argument,* $f(x,y,y') = -2\sin(x)y(x) + y'(x)^2 + x^2\sqrt{1 + y(x)^2}$ *is also strongly convex. A more involved strongly convex functional on* $\mathbb{R}^2$ *is* $f(x,y') = \sqrt{1 + y(x)^2 + y'(x)^2}$ *(it is even strictly convex). With this result,* $g(x)\sqrt{1 + y^2(x) + y'(x)^2}$ *is strongly convex too (for* $g(x) > 0$*).*

If $f(x,y,y')$ and f_{yy}, $f_{y'y'}$, $f_{yy'}$ are continuous on $[a,b] \times \mathbb{R}^2$, then f is convex if and only if the Hessian of $f[y(x)]$ is positive semidefinite. We can use these results on convexity to characterize the famous Lagrange Multiplier Theorem with convex constraints.

Theorem 2.47. *If* D *is a domain in* $\mathbb{R}^2$*, such that for some constants* λ_j*, for* $j = 1, \ldots, N$ *and fixed* x*,* $f(x,y,y')$ *and* $\lambda_j g_j(x,y,y')$ *are convex on* $[a,b] \times D$*, let*

$$\bar{f} = f + \sum_{j=1}^{N} \lambda_j g_j(x).$$

Then each solution y *of the differential equation of Euler-Lagrange*

$$\frac{\mathrm{d}}{\mathrm{d}x}\bar{f}_{y'}[y(x)] = \bar{f}_y[y(x)]$$

minimizes $F(y) = \int_a^b f[y(x)]\,\mathrm{d}x$ *on* (a,b) *under the constraining relations*

$$G_j(y) = \int_a^b g_j[y(x)]\,\mathrm{d}x.$$

If at least one of the $\lambda_j g_j$ *is strongly convex, then the minimizer is unique.*

The definitions and theorems so far exposed are useful to analyse the functionals of the calculus of variations and then of optimal control problems. The basic results are the Euler-Lagrange equation and the lemmas of Lagrange and du Bois-Reymond. They are important to determine necessary conditions for a minimizing function when convexity is not present. This section explores rigorously the properties, already encountered, that when a function h is constant, the integral $\int_a^b h(x)v'(x)\,\mathrm{d}x = 0$ for $v(a) = v(b) = 0$. We present herein the theorems that describe those integrals with the associated necessary conditions.

Lemma 2.48 (du Bois-Reymond)**.** *If* $h \in C[a,b]$ *and* $\int_a^b h(x)v'(x)\,\mathrm{d}x = 0$*, for all* $v \in \mathcal{D}_0 = \{v \in C^1[a,b] \quad | \quad v(a) = v(b) = 0\}$*, then* h *is constant on* $[a,b]$*.*

Proposition 2.49. *If* $g, h \in C[a,b]$ *and* $\int_a^b g(x)v(x) + h(x)v'(x)\,\mathrm{d}x = 0$*, for all* $v \in \mathcal{D}_0 = \{v \in C^1[a,b] \quad | \quad v(a) = v(b) = 0\}$*, then* $h \in C^1[A,b]$ *and* $h' = g$*. As a corollary, setting* $h = 0$ *yields* $g = 0$*.*

The generalization of this result is known as the Lemma of Lagrange.

Lemma 2.50 (Lagrange)**.** *If* $g \in C[a,b]$ *and for some* $m = 0, 1, 2, \ldots$ $\int_a^b g(x)v(x)\,\mathrm{d}x = 0$*, for all* $v \in \mathcal{D}_0 = \{v \in C^m[a,b] \quad | \quad v^{(k)}(a) = v^{(k)}(b) = 0,\ k = 1, \ldots m\}$*, then* $g = 0$ *on* $[a,b]$*.*

The generalization of the Lemma of du Bois-Reymond is given next.

Proposition 2.51. *If* $h \in C[a,b]$ *and for some* $m = 0, 1, 2, \ldots$ $\int_a^b h(x)v^{(m)}(x)\,\mathrm{d}x = 0$*, for all* $v \in \mathcal{D}_0 = \{v \in C^m[a,b] \mid v^{(k)}(a) = v^{(k)}(b) = 0,\ k = 1, \ldots m-1\}$*, then* h *is a polynomial of degree* $\deg h < m$ *on* $[a,b]$*.*

There are also the vector analogues of the previous theorems, and it is enough to consider the scalar version for each component.

Example 2.52. *Consider the characterization of the minimum values of the functional* $J(y) = \int_a^b f(x)\sqrt{1 + y'(x)^2}$ *for a continuous function* f*, on the domain* $\mathcal{D} = \{y \in C^1[a,b] \mid \ y(a) = a_1,\ y(b) = b_1\}$ *for given* $a_1, b_1 \in \mathbb{R}$*. We have already seen that an admissible variation is* $v \in \mathcal{D}_0 = \{v \in C^1[a,b] \mid v(a) = v(b) = 0\}$*. The necessary condition that* $y \in \mathcal{D}$ *is a local extremum is that* $\delta J(y; v) = 0$ *or,*

$$\delta J(y; v) = \int_a^b \frac{f(x)y'(x)v'(x)}{\sqrt{1 + y'(x)^2}}\,\mathrm{d}x = 0 \quad \forall\ v \in \mathcal{D}_0.$$

From Lemma (2.48) *of du Bois-Reymond, the necessary condition is satisfied by a function* y *for which*

$$\frac{f(x)y'(x)}{\sqrt{1 + y'(x)^2}} = k, \qquad k \in \mathbb{R}, \tag{2.22}$$

that is, after some manipulations,

$$y'(x)^2 = \frac{k^2}{f(x)^2 - k^2}.$$

Now we can observe that if f *vanishes at a single point, then from* (2.22) $k = 0$ *and this implies* $y = \mathrm{const}$ *which requires that* $y(a) = a_1 = y(b) = b_1$*, thus if* $a_1 \neq b_1$ *the problem has no solution.*

Therefore if $a_1 \neq b_1$, it is required that $f(x)^2 > k^2 > 0$, e.g. $f(x) > |k| > 0$ or $f(x) < -|k| < 0$. Considering the first case only (the second follows putting $-J$ instead of J), when $f(x) > 0$ the integrand is a strongly convex function, as we have already seen. Hence, from the special case of the Euler-Lagrange equation (2.20)*, the function y that solves* (2.22) *gives the only minimum value for J, provided that such y exists.*

2.2.4 *Fréchet Derivatives*

The Gâteaux variation in a normed linear space is the analogous of the directional derivative in $\mathbb{R}^n$, and like the directional derivatives, it can not provide a good local approximation of a function, except in each separate direction. As with usual functions, we require a stronger differentiability which is independent of the direction.

Definition 2.53 (Fréchet derivative)**.** *In a normed linear space $\mathcal{Y}$, a real valued functional J is said to be differentiable in the sense of Fréchet at $y_0 \in \mathcal{Y}$ provided that J is defined in a sphere $S(y_0)$ and there exists a continuous linear function $L : \mathcal{Y} \to \mathbb{R}$ for which*

$$J(y) = J(y_0) + L(y - y_0) + ||y - y_0||\mathrm{o}(||y - y_0||).$$

If J is Fréchet differentiable at y_0 then J has the Gâteaux variations $\delta J(y_0; v) = L(v)$ in each direction $v \in \mathcal{Y}$. The linear function L is uniquely determined and is denoted as $J'(y_0)$. The differentiability at y_0 implies the continuity of the functional at that point. As in $\mathbb{R}^n$, the converse is not true. To obtain a kind of viceversa, we need the additional hypothesis of uniformity.

Theorem 2.54. *In a normed linear space $\mathcal{Y}$, if a real valued functional J has at each $y \in S(y_0)$ the Gâteaux variations $\delta J(y; v)$ for all $v \in \mathcal{Y}$ and $\delta J(y; v)$ is linear in v; if for $y \to y_0$ the difference $|\delta J(y; u) - \delta J(y_0; v)| \to 0$ uniformly for $u \in \{u \in \mathcal{Y},\ ||u|| = 1\}$, then J is differentiable at y_0.*

Proposition 2.55. *When $f = f(x, y(x), y'(x))$ and $f_y, f_{y'} \in C([a, b] \times \mathbb{R}^2)$ then*

$$F(y) = \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x$$

is differentiable and has weakly continuous variations at each $y_0 \in \mathcal{Y} = C^1[a, b]$ with respect to the maximum norm $||y||_M$.

With the knowledge of the Fréchet derivative, we can extend the concept of separating hyperplane to functionals. The Fréchet derivative J' offers a good approximation of the functional J by the function,

$$T(y) = J(y_0) + J'(y_0)(y - y_0).$$

Roughly speaking, the graph of T is tangent to the graph of J at the point $(y_0, J(y_0)$. We can consider the level set of T at y_0 as $T_{y_0} = \{y \in \mathcal{Y} \mid T(y) = T(y_0)\} = \{y \in \mathcal{Y} \mid J'(y_0)(y - y_0) = 0\}$. The last equality follows from the fact that $T(y_0) = J(y_0)$. If we set $\mathcal{Y} = \mathbb{R}^3$ with the Euclidean norm, then all the apparatus reduces to $\delta J(\boldsymbol{y}_0; \boldsymbol{v}) = \nabla J(\boldsymbol{y}_0) \cdot \boldsymbol{v}$ and the linear function becomes $T(\boldsymbol{v}) = J'(\boldsymbol{y}_0) \cdot \boldsymbol{v} = \nabla J(\boldsymbol{y}_0) \cdot \boldsymbol{v}$. The tangent directions $\boldsymbol{v}$ are those which are orthogonal to the gradient $\nabla J(\boldsymbol{y}_0)$. If $\nabla J(\boldsymbol{y}_0) \neq \boldsymbol{0}$ then it is perpendicular to the plane $T(\boldsymbol{y}_0)$ through $\boldsymbol{y}_0$ determined by the tangent vectors, and therefore $\nabla J(\boldsymbol{y}_0)$ is normal to the level surface J_{y_0} through this point. The concept of level set permits to generalize the theorem of Lagrangian Multipliers. For example the constraint $\{y \in C[a, b] \mid y(a) = a_1, y(b) = b_1\}$ can be expressed as the intersection of the two level sets $G_1(y) = y(a)$ and $G_2(y) = y(b)$ respectively to level a_1 and b_1. The theorem is similar to Theorem 2.47.

Theorem 2.56 (Lagrange). *In a normed linear space $\mathcal{Y}$ let real valued functionals $J, G_1, \dots, G_n$ be defined in a neighborhood of y_0, a local extremal point for J constrained to the level sets $G_{y_0} = \{y \in \mathcal{Y} \mid G_i(y) = G_i(y_0),\, i = 1, \dots, n\}$, and have there weakly continuous Gâteaux variations. Then either:*

$$\det \begin{pmatrix} \delta G_1(y_0; v_1) & \delta G_1(y_0; v_2) & \dots & \delta G_1(y_0; v_n) \\ \delta G_2(y_0; v_1) & \delta G_2(y_0; v_2) & \dots & \delta G_2(y_0; v_n) \\ \vdots & & \ddots & \vdots \\ \delta G_n(y_0; v_1) & \delta G_n(y_0; v_2) & \dots & \delta G_n(y_0; v_n) \end{pmatrix} = 0,$$

for all $v_j \in \mathcal{Y}$ and $j = 1, \dots, n$; or there exist constants $\lambda_i \in \mathbb{R}$ for $i = 1, \dots, n$ such that

$$\delta J(y_0; v) = \sum_{i=1}^{n} \lambda_i \delta G_i(y_0; v) \qquad \forall\ v \in \mathcal{Y}.$$

The first condition implies that the constraints are locally linearly dependent, i.e. there exist constants μ_j such that $\sum \mu_j G_j(y) = 0$. Because the Gâteaux variations are linear, the previous relation yields $\sum \mu_j \delta G_j(y; v) = 0$ for each direction v and thus the determinant is zero. The second condition yields a linear dependence of $\delta J, \delta G_1, \dots, \delta G_n$, that is, all the functionals are differentiable in a direction simultaneously tangent to each level set G_{j,y_0}. In particular it must be tangential to the unconstrained J_{y_0}. The constraints on J that determine admissible directions v which are a linear subspace, can be considered for restricting the possible directions when applying the Lagrangian Multiplier theorem for the other constraints. This is shown in the next example.

Example 2.57. *Find the local extremals for the functional*

$$J(y) = \int_{-1}^{0} y'(x)^3 \,\mathrm{d}x$$

on the set $\mathcal{D} = \{y \in \mathcal{Y} = C^1[-1, 0] \mid y(-1) = 0,\ y(0) = \frac{2}{3}\}$, under the constraining relation

$$G(y) = \int_{-1}^{0} x y'(x) \,\mathrm{d}x = -\frac{4}{15}.$$

Instead of invoking the theorem with three constraints ($n = 3$), we observe that the fixed extrema imply that the directions form a linear subspace of $\mathcal{Y}$: in fact the admissible directions satisfy $\{v \in C^1[-1, 0] \mid v(-1) = v(0) = 0\}$. Hence we use the theorem in a restricted form, considering only

$$\delta J(y; v) = \int_{-1}^{0} 3y'(x)v'(x) \,\mathrm{d}x \qquad \textit{and} \qquad \delta G(y; v) = \int_{-1}^{0} x v'(x) \,\mathrm{d}x.$$

These variations are weakly continuous because of proposition 2.55, in facts the partial derivatives with respect to y' are continuous. The theorem of Lagrange 2.56 gives us two possibilities: either $\delta G(y; v) = \int_{-1}^{0} x v'(x) \,\mathrm{d}x = 0$ for all admissible directions or there exists λ such that $\delta(J + \lambda G)(y; v) = \int_{-1}^{0} (3y'(x)^2 + \lambda x) v'(x) \,\mathrm{d}x = 0$. Now, the first case yields that x should be a constant function over $[-1, 0]$ (lemma of du Bois-Reymond 2.48), and this is impossible. The second case implies that $3y'(x)^2 + \lambda x = c$ with $c \in \mathbb{R}$ constant. For the sake of simplicity replace λ with -3λ so that $y'(x)^2 = c + \lambda x \geq 0$, which gives two possibilities for y', $y'(x) = \pm\sqrt{c + \lambda x}$. The negative root does not satisfy the constraint $G(y) = -4/15$; it remains the positive root. First we notice that $\lambda \neq 0$ because $\lambda = 0$ implies $y'(x) = \sqrt{c}$, then $G(y) = -4/15 \implies \sqrt{c} = 8/15$, but then $y(x) = 8/15x + k$

for a constant $k \in \mathbb{R}$ is no more in $\mathcal{D}$, because this straight line does not match the boundary constraints, therefore $\lambda \neq 0$. Integrating $y'(x) = \sqrt{c + \lambda x}$ yields (for an integrating constant k),

$$y(x) = \frac{2}{3\lambda}(c + \lambda x)^{3/2} + k.$$

We can now impose the boundary conditions and find $k = -2/(3\lambda)(c - \lambda)^{3/2}$ and $\lambda = c^{3/2} - (c - \lambda)^{3/2}$. Next we have to match the requirement $G(y) = -4/15$ which gives $\lambda^2 = \frac{-5\lambda}{2}(c - \lambda)^{3/2} + c^{5/2} - (c - \lambda)^{5/2}$. We obtained a nonlinear system of three equations in the unknown c, k, λ. Its only feasible solution is $k = 0$, $\lambda = 1$, $c = 1$. Thus we have proved that $y(x) = \frac{2}{3}(x + 1)^{3/2}$ is the only possible extremal function. We can now employ convexity to show that it is not a local maximum with respect to the maximum norm $|| \cdot ||_M$. In facts the functional $f(x, y') = y'(x)^3 + \lambda x y'(x)$ is strongly convex on $[-1, 0] \times [0, \infty)$.

2.2.5 *Transversal Conditions*

In general, at the boundary we can have different conditions, that are called transversal conditions. An useful technique for handling such kind of constraints is to apply the Lagrange multipliers.
It is common that the upper extremum of the integral functional is free, this happens in minimum time problems. In this case the functional is of the form

$$J(y, t) = \int_a^t f(x, y(x), y'(x))\,\mathrm{d}x = \int_a^t f[y(x)]\,\mathrm{d}x$$

and is to be minimized on a set like (see Figure 2.6)

$$\mathcal{D}_\tau = \{y \in C^1[a, t] \mid y(a) = a_1,\ \tau(t, y(t)) = 0\},$$

where $\tau(t, y(t))$ is some kind of expression. We assume here that $\nabla\tau \neq \mathbf{0}$, but this condition is mild, because in most cases τ is a linear function. We need to perform the variation on the

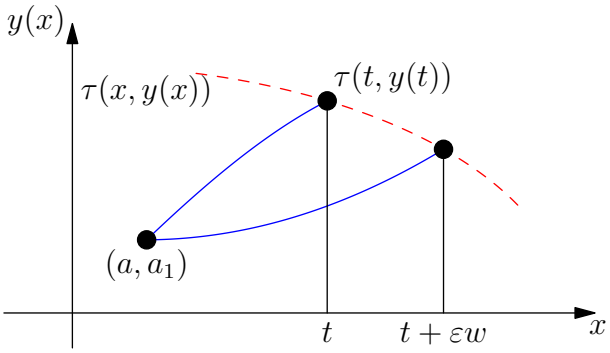


Figure 2.6: The transversal condition $\tau(x, y(x)) = 0$ for a free endpoint.

extended space $\mathcal{Y} = C^1[a, b] \times \mathbb{R}$ with the associated norm $||(y, t)|| = ||y||_M + |t|$. The variation of the functional $J(y, t)$ in the direction (v, w) becomes then

$$\begin{aligned}\delta J(y, t; v, w) &= f(t)w + \int_a^t f_y(x)v(x) + f_{y'}(x)v'(x)\,\mathrm{d}x \\ &= f(t)w + f_{y'}(x)v(x)\Big|_a^t.\end{aligned}$$

The endpoint constraint can be expressed as the zero level set of a function $G(y,t) = \tau(t, y(t)) = \tau[y(t)]$, so that its variation is

$$\begin{aligned}\delta G(y,t;v,w) &= \lim_{\varepsilon\to 0}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\tau(t+\varepsilon w,(y+\varepsilon v)(t+\varepsilon w))\\ &= \lim_{\varepsilon\to 0}\frac{\partial\tau}{\partial x}[y(t+\varepsilon w)]w + \frac{\partial\tau}{\partial y}[y(t+\varepsilon w)]\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(y(t+\varepsilon w)+\varepsilon v(t+\varepsilon w))\\ &= \tau_x[y(t)]w + \tau_y[y(t)]\lim_{\varepsilon\to 0}(y'(t+\varepsilon w)w + v(t+\varepsilon w) + \varepsilon v'(t+\varepsilon w)w)\\ &= \tau_x[y(t)]w + \tau_y[y(t)](y'(t)w + v(t)).\end{aligned}$$

Both $\delta J(y,t;v,w)$ and $\delta G(y,t;v,w)$ are weakly continuous, and we can apply the theorem of Lagrange multipliers to seek λ such that $\delta(J+\lambda G)(y,t;v,w) = 0$. Now the set of possible direction is $\mathcal{D}_0 = \{v \in C^1[a,t] \mid v(a) = v(t) = 0\}$. Explicitly, the previous requirement for $v(a) = v(t) = 0$ and w small is

$$(f(t) + \lambda(\tau_x[y(t)] + \tau_y[y(t)](y'(t)))\, w = 0.$$

In the same fashion, for $w = v(a) = 0$ and v small is

$$(f_{y'}(t) + \lambda\tau_y[y(t)])\, v = 0.$$

Combining and solving this two relations for λ (multiply the first equation by τ_y and isolate $\lambda\tau_y$ in the second), a local extremum y of J on $\mathcal{D}_\tau$ stationary on (a,t) satisfies the transversal condition

$$f(t)\tau_y[y(t)] = f_{y'}(t)(\tau_x[y(t)] + \tau_y[y(t)](y'(t)). \tag{2.23}$$

If the endpoint condition is fixed, i.e. $\tau(x,y) = b - x$, the previous equations reduce (because $\tau_y = 0$) to the condition $f_{y'}(b) = 0$; if $\tau(x,y) = y - b_1$ for an assigned b_1, the terminal value t of x is free and at (t,b_1) an extremal solution needs to meet $f(t) - y'(t)f_{y'}(t) = 0$. This situation is called free-horizon. If the value b_1 is also free at (t,b_1), the (free) end point condition is $f_{y'}(t) = 0$.

2.2.6 *Integral Constraints*

Other kinds of constraints that involve the whole integrating interval $[a,b]$ are the integral constraint. They appear very frequently in applications because they can be interpreted as equations of mechanics or physics. They are in general expressed by

$$G(y) = \int_a^b g(x, y(x), y'(x))\,\mathrm{d}x = \int_a^b g[y(x)]\,\mathrm{d}x.$$

There is a version of the theorem of Lagrange multipliers also for this case, and it is very similar to theorem 2.56.

Theorem 2.58 (Lagrange)**.** *In a normed linear space $\mathcal{Y}$ let real valued functionals $J, g_1, \dots, g_N$ be continuous with their y and y' partial derivatives. Let y be a local extremal function for*

$$J(y) = \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x$$

on the set

$$\mathcal{D} = \{y \in C^1[a,b] \mid y(a) = a_1,\ y(b) = b_1\},$$

with constraints

$$G_y = \{y \in C^1[a,b] \mid G_j(y) = \int_a^b g(x, y(x), y'(x))\,\mathrm{d}x,\ j = 1, 2, \ldots, N\}.$$

Then either:

$$\det \begin{pmatrix} \delta G_1(y_0; v_1) & \delta G_1(y_0; v_2) & \ldots & \delta G_1(y_0; v_N) \\ \delta G_2(y_0; v_1) & \delta G_2(y_0; v_2) & \ldots & \delta G_2(y_0; v_N) \\ \vdots & & \ddots & \vdots \\ \delta G_N(y_0; v_1) & \delta G_N(y_0; v_2) & \ldots & \delta G_N(y_0; v_N) \end{pmatrix} = 0,$$

for all $v_j \in \mathcal{D}_0 = \{v \in C^1[a,b] \mid v(a) = v(b) = 0,\ j = 1, \ldots, N\}$*; or there exist constants* $\lambda_i \in \mathbb{R}$ *for* $i = 1, \ldots, N$ *that make* y *stationary for the augmented functional* $\hat{f} = f + \sum_{i=1}^N \lambda_i g_i$*, that is,* y *is a solution on* (a, b) *of the equation*

$$\frac{\mathrm{d}}{\mathrm{d}x} \bar{f}_{y'}(x) = \bar{f}_y(x).$$

Clearly, if we replace $\mathcal{D}$ with $\mathcal{D}^b = \{y \in C^1[a,b] \mid y(a) = a_1\}$, then $\mathcal{D}_0$ becomes $\mathcal{D}_0^b = \{v \in C^1[a,b] \mid v(a) = 0\}$ with the additional requirement that $\hat{f}_{y'}(b) = 0$; if $\mathcal{D} = \{y \in C^1[a,b]\}$ then we must have $\hat{f}_{y'}(a) = \hat{f}_{y'}(b) = 0$. For a more general transversal condition such as $y(a) = a_1$ and $\tau(t, y(t)) = 0$ (with the standard assumptions on τ as in the section of transversal conditions), the general requirement is

$$\bar{f}(t)\tau_y[y(t)] = \bar{f}_{y'}(t)(\tau_x[y(t)] + \tau_y[y(t)]y'(t)$$

as in equation (2.23).

2.2.7 *Equality Constraints*

The method of Lagrangian multipliers can also be adapted to the case of equality constraints of the form

$$g[y(x)] = g(x, y(x), y'(x)) = 0 \qquad \forall x \in [a,b],$$

where $g \in C^1(D)$ for a domain $D \subset \mathbb{R}^{2d+1}$. It is enough to consider one constraint, the others can be added in the same fashion.

Theorem 2.59 (Lagrange)**.** *For* $f = f(x, y(x), y'(x))$ *and* $f_{x_j} \in C^1([a,b] \times \mathbb{R}^{2d})$*,* $j = 1, 2, \ldots, d$ *if* y_0 *is* C^2 *and it minimizes* $F(y) = \int_a^b f[y(x)]\,\mathrm{d}x$ *on* $\mathcal{D} = \{y \in \mathcal{Y} = (C^1[a,b])^d \mid y(a) = y_0(a),\ y(b) = y_0(b)\}$ *when it is subject to* $g[y(x)] = 0$*, where* g *is* C^2 *such that* $\nabla g[y_0(x)] \neq \mathbf{0}$*, then there exists* $\lambda \in C[a,b]$ *such that* y_0 *is stationary for the augmented function* $f + \lambda g$*.*

The extension to N constraints is straightforward by adding those constraint in the augmented functional with the appropriate multiplier, provided that the $N \times N$ Jacobian of the constraints is non vanishing along the trajectory.

2.2.8 *Extension to* C^1 *Piecewise Functions*

It is clear from the configuration of many classical examples, e.g. the minimal surfaces of revolution, that often we need optimal functions that exhibit corners. Those curves are called piecewise differentiable functions or C^1 piecewise functions. In the next we include this class of functions in the theory of calculus of variations and provide general necessary and sufficient properties of

minimal extremals. This situation occurs very frequently in applications and will be treated more extensively in the chapter of optimal control. In this section we introduce the basic theory of the conditions of Weierstrass-Erdmann and the condition of Legendre. In case of (strong) convexity, we give conditions to guarantee the minimality of the solution.

Definition 2.60 (C^1 piecewise functions)**.** *A function $y \in \hat{C}^1[a,b]$ is piecewise differentiable if there is a finite irreducible partition $a = c_0 < c_1 < \ldots < c_{N+1} = b$ such that y may be regarded as a function in $C^1[c_k, c_{k+1}]$ for each $k = 1, \ldots, N$. When present, the interior points c_k are called* corner points *of y.*

It is clear that such $y \in \hat{C}^1[a,b]$ is defined and continuously differentiable on $[a,b]$ except at corner points, where it has distinct limiting values. Let c be a corner point, then we denote with $y'(c)$ both values when the distinction is not important, otherwise, a good notation can be $y'(c^-)$ and $y'(c^+)$. We collect some facts on $\hat{C}^1[a,b]$ functions.

Proposition 2.61. *Le $y \in \hat{C}^1[a,b]$, then:*

- $y(x) = y(a) + \int_a^x y'(t)\,\mathrm{d}t$*, a form of the fundamental theorem of calculus.*
- *If* $\int_a^b y'(x)^2\,\mathrm{d}x = 0$ *then* $y' = 0$ *on* $[a,b]$.
- *If* $y' = 0$ *where defined, then* $y = \mathrm{const}$ *on* $[a,b]$.

A useful norm for the space $\hat{C}^1[a,b]$ can be $||y||_\infty = \max\{|y(x)| \mid x \in [a,b]\}$ because $\hat{C}^1[a,b] \subset C[a,b]$, that is, even if there is no control over the differentiability, it gives some information when the piecewise y is smoothed another function. It is called *strong* norm. Another choice for a norm is $||y|| := \max\{|y(x)| + |y'(x)| \mid x \in [a,b]\}$, which takes into account the differentiability of y and is called *weak* norm; or $||y||_1 := \int_a^b |y(x)| + |y'(x)|\,\mathrm{d}x$. The last choice permits to compare two functions which agree except in small neighbourhoods of their corner points to be close. These norms are not independent, and can be related by the next inequality:

$$A||y||_\infty \le ||y||_1 \le (b-a)||y||, \qquad A = \frac{b-a}{1+b-a}. \tag{2.24}$$

When a function $f(x,y,y')$ depends on $y \in \hat{C}^1[a,b]$ with simple discontinuities at corner points, then

$$F(y) = \int_a^b f(x,y,y')\,\mathrm{d}x = \int_a^b f[y(x)]\,\mathrm{d}x$$

is definite and finite, since the partition given by the corner points reduces the integral to a finite sum of integrals with all the good properties. But, in general, F is not continuous with respect to the norms $||F||_{\max}$ or $||F||_1$, F is continuous only with respect to the weak norm $||F||$ defined before.

Remark 2.62. *Notice that if $f \in C([a,b] \times \mathbb{R}^{2d})$, $f(x,y,y')$ and y_0 is an extremal point for F on $\mathcal{D} = \{y \in \mathcal{Y} \mid y(a) = A,\ y(b) = B\}$, then y_0 is also an extremal point for F on $\hat{\mathcal{D}} = \{\hat{y} \in \hat{\mathcal{Y}} \mid \hat{y}(a) = A,\ \hat{y}(b) = B\}$ with respect to the same norm. The characterization of* local C^1 *extremals given in the previous sections were with respect to an unspecified norm, but, as observed, weak local extremals need not be strong local extremals. However, in case they are global extremals, then the choice of the norm is indifferent, that is: if y is a global minimizer for F on $\mathcal{D}$, then it will be a global minimizer also for F on $\hat{\mathcal{D}}$. Moreover, the minima of convex functions minimize also over the corresponding class of piecewise C^1 functions.*

2.2.8.1 *The Weierstrass-Erdmann Conditions*

The previous remark 2.62 does not preclude a function from being extremized by a function which is only piecewise differentiable. A classical counterexample to this is given by the functional

$$F(y) = \int_{-1}^{1} y^2(x)(1 - y'(x))^2 \,\mathrm{d}x$$

defined on the set $\mathcal{D} = \{y \in \hat{C}^1[-1, 1] \mid y(-1) = 0, y(1) = 1\}$. The minimum is reached uniquely by the solution

$$y(x) = \begin{cases} 0 & -1 \leq x \leq 0, \\ x & 0 \leq x \leq 1. \end{cases}$$

There is clearly a corner point in $x = 0$. On the other hand such function $y(x)$ does not belong to the set of *continuously* differentiable functions.
When searching for necessary conditions that make $y \in \hat{C}^1$ a local extremal, we have to assume first that y is a weak local extremal, in fact each local extremal with respect to the $||y||_{\max}$ norm or the $||y||_1$ norm is automatically a weak local extremal (because of (2.24)). The definition of the variation of the functional

$$F(y) = \int_a^b y^2 f(x, y(x), y'(x)) \,\mathrm{d}x$$

has the same derivation but we have to take into account the corner points of both y and v, they must be split in a finite sum of integrals with continuous integrands and differentiate each under the integral sign. Then, after reassembly, we get again

$$\frac{\partial}{\partial \varepsilon} F(y + \varepsilon v) = \int_a^b f_y[(y + \varepsilon v)(x)]v(x) + f_{y'}[(y + \varepsilon v)(x)]v'(x) \,\mathrm{d}x,$$

and performing the limit for $\varepsilon \to 0$ we obtain the usual

$$\delta F(y; v) = \int_a^b f_y(x)v(x) + f_{y'}(x)v'(x) \,\mathrm{d}x$$

where the partial derivatives f_y and $f_{y'}$ are piecewise continuous on $[a, b]$. If y is a local extremal function, $\delta F(y; v) = 0$ must hold, and integrating by parts we have

$$\delta F(y; v) = \int_a^b \left(f_{y'} - \int_a^x f_y(t) \,\mathrm{d}t \right) v'(x) \,\mathrm{d}x = 0$$

and by the du Bois-Reymond lemma, the factor that multiplies $v'(x)$ should be zero, hence

$$f_{y'}(x) = \int_a^x f_y(t) \,\mathrm{d}t + k \quad \Longrightarrow \quad \frac{\mathrm{d}}{\mathrm{d}x} f_{y'}(x) = f_y(x)$$

except at each corner point c of y where the continuity of $f_{y'}(x) = \int_a^x f_y(t) \,\mathrm{d}t + k$ implies the *first* Weierstrass-Erdmann condition,

$$f_{y'}(c^-) = f_{y'}(c^+). \tag{2.25}$$

On each interval that excludes corner points, the local extremal function y must be C^1 and stationary. Moreover, at each corner c, the second derivative $f_{y'y'}(c, y(c), y'(c))$, if defined, must vanish for some values of y'. The other condition is derived starting from the second Euler-Lagrange equation,

$$f(x) - y'(x) f_{y'}(x) = \int_a^x f_x(t) \,\mathrm{d}t + k,$$

in facts we have that

$$\frac{\mathrm{d}}{\mathrm{d}x}(f - y'f_{y'})(x) = f_x(x) \qquad \forall x \in (a,b) - \{c_i\}.$$

At those corner points. holds the *second* Weierstrass-Erdmann condition,

$$(f - y'f_{y'})(c^-) = (f - y'f_{y'})(c^+). \tag{2.26}$$

We can rewrite the second condition (2.26) using the first condition (2.25) as follows:

$$f(c, y(c), y'(c^-)) - f(c, y(c), y'(c^+)) - \big(y'(c^-) - y'(c^+)\big)\, f_{y'}(c, y(c), y'(c^+)) = 0.$$

Now, because on corner points $y'(c^-) \neq y'(c^+)$, we have that $f(c, y(c), .)$ can not be strictly convex and this information can be useful to locate or preclude the presence of corner points. We summarize these results in the next theorem.

Theorem 2.63 (Weierstrass-Erdmann conditions)**.** *If a function $f(x, y, y') \in C^1([a,b] \times \mathbb{R}^2)$ and $y \in \mathcal{Y} = \hat{C}^1[a,b]$ provide a weak local extremal for*

$$F(y) = \int_a^b f[y(x)]\,\mathrm{d}x$$

on

$$\mathcal{D} = \{y \in \mathcal{Y} \mid y(a) = a_1,\ y(b) = b_1\},$$

then, except at its corner points, y is C^1 and satisfies the first and second Euler-Lagrange conditions (2.19) *and* (2.21)*. At each corner point c hold the two Weierstrass-Erdmann necessary conditions* (2.25) *and* (2.25)*:*

1. $f_{y'}(c^-) = f_{y'}(c^+)$,
2. $(f - y'f_{y'})(c^-) = (f - y'f_{y'})(c^+)$,
3. $\pm f(c, y(c), y'(.))$ *can not be strictly convex in y'.*

Example 2.64. *Fix x and $y(x)$, then $f(x, y(x), y'(x)) = (x^2 + y^2)\sqrt{1 + y'^2}$ is strictly convex except when $x^2 + y^2 = 0$, hence the associated local extremal y can have a corner point only for values of c such that $c = y(c) = 0$.*
Similarly, $f = (1 + y^2)y'^4$ is strictly convex in y' and therefore can not have extremals with corner points.

The theorem shows that the discontinuities of y' are permitted at corner points of a local extremal, but are limited to those which preserve the continuity of both $f_{y'}$ and $f - y'f_{y'}$, hence when $f_x \equiv 0$ the latter term is constant.

Example 2.65. *Consider the function $f(x, y, y') = y^2(1 - y')^2$, for which $f_{y'} = 2y^2(y' - 1)$. An extremal function y must be stationary on interval excluding corner points, at which both $f_{y'} = -2y^2(1 - y')$ and $f - y'f_{y'} = y^2(1 - y'^2)$ are continuous (the latter is constant because $f_x \equiv 0$). From the continuity of y, it follows that the first condition implies that y' is continuous except at corner points c such that $y(c) = 0$. Corner points can be only of this form in this example. Therefore, unless y vanishes at some point in $[a,b]$ it is not a local extremal. If it vanishes a single point c, then from the second condition, $y^2(1 - y'^2) \equiv 0$ so that for all $x \in [a,b]$ either $y(x) = 0$ or $y'(x) = 1$ or $y'(x) = -1$.*

We can also extend to piecewise C^1 functions the theorem 2.41 of uniqueness when f is convex, where any local extremum is the local minimum. The extremal y must be stationary on intervals excluding corner points, at which it must satisfy theorem 2.63.

Example 2.66. *Find the local extremal functions for*

$$F(y) = \int_0^2 y'(x)^2 \,\mathrm{d}x \qquad \text{on} \quad \mathcal{D} = \{y \in \hat{C}^1[0,2] \mid y(0) = y(2) = 1,\ y(1) = 0\}.$$

First we notice that $f = y'^2$ is strictly convex, and $\mathcal{D}_0 = \{v \in \hat{C}^1[0,2] \mid v(0) = v(1) = v(2) = 0\}$. Therefore, as usual, by convexity,

$$F(y+v) - F(y) \geq \delta F(y; v) = \int_0^2 2y'(x)v'(x) \,\mathrm{d}x.$$

The inequality is an equality when $v = 0$. A possible solution can be (see 2.24), y' constant, that is

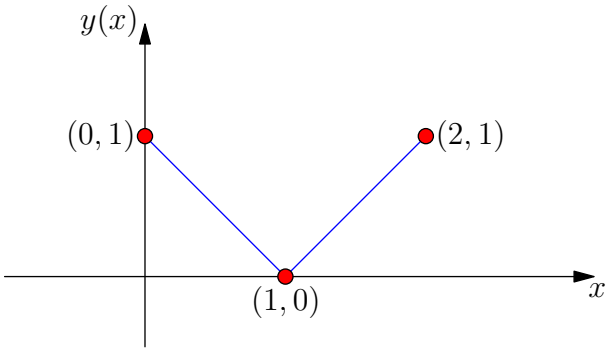


Figure 2.7: The plot of $y(x)$.

$$y'(x) = \begin{cases} c_1 & x \in [0,1) \\ c_2 & x \in (1,2]. \end{cases}$$

Thus, integrating the previous equation with the boundary condition given, yields (see Figure 2.7)

$$y(x) = \begin{cases} 1 - x & x \in [0,1) \\ x - 1 & x \in (1,2]. \end{cases}$$

That y is the only local extremal function for F that minimizes F uniquely.
Observe that y is clearly $\hat{C}^1$ but it does not satisfy the Weierstrass-Erdmann conditions, because the corner point in $x = 1$ is forced by the problem and is not natural.

Example 2.67. *Minimize the (strictly convex) distance function*

$$F(y) = \int_a^b \sqrt{1 + y'(x)^2} \,\mathrm{d}x \qquad \text{such that} \quad y(x) \leq x^2.$$

The domain is the set $\mathcal{D} = \{y \in \hat{C}^1[a,b] \mid y(a) = a_1,\ y(b) = b_1\}$. This time we have a constraint function $g(x, y(x)) = y(x) - x^2 = 0$, g is convex, so we consider the augmented problem

$$\tilde{F}(y) = \int_a^b f[y(x)] + \lambda(x) g[y(x)] \,\mathrm{d}x \qquad \lambda(x) \geq 0.$$

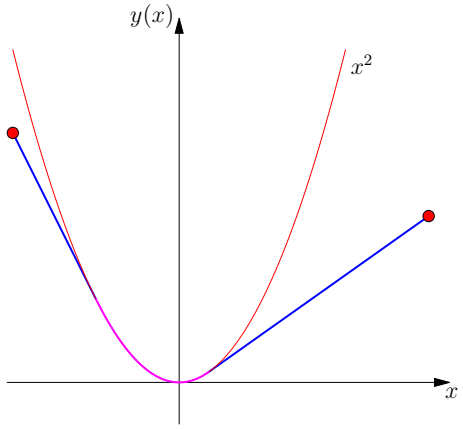


Figure 2.8: The plot of $y(x)$.

From the strong convexity of $\tilde{f}(x,y,y') = f(x,y,y') + \lambda(x)g(x,y)$ that y minimizes F uniquely under the inequality constraint. We search for intervals excluding corner points such that y is stationary for $\tilde{f}$, that is

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{y'}[y(x)] - f_y[y(x)] = \lambda(x) g_y[y(x)]. \tag{2.27}$$

This time, however, we admit intervals with $g \equiv 0$ and $\lambda \neq 0$ for a y not stationary for f. Moreover, since $\tilde{f}_{y'} = f_{y'}$, y has only the corner points permitted by f. It is clear from Figure 2.8 that for some configurations a portion of the minimizing curve lie on the parabola defined by g. We have that

$$f(x,y,y') = \sqrt{1+y'(x)^2} \implies \tilde{f}(x,y,y') = \sqrt{1+y'(x)^2} + \lambda(x)(y(x) - x^2).$$

For the part along the parabola, $y(x) = x^2$ and because of $g_y = 1$ and $f_y = 0$, from equation (2.27) *we can find $\lambda(x)$:*

$$\lambda(x) = \frac{\mathrm{d}}{\mathrm{d}x} \frac{y'(x)}{\sqrt{1+y'(x)^2}} = \frac{\mathrm{d}}{\mathrm{d}x} \left(\frac{2x}{\sqrt{1+4x^2}} \right) = \frac{2}{(1+4x^2)^{3/2}} \geq 0.$$

For the portions not on the parabola, $\lambda = 0$ hence y will be segment of the line tangential to the parabola at the point of contact. This analysis shows that this is the unique minimizer and that it has not corner points. This last fact can be seen also by noticing that $\tilde{f}_{y'y'} = f_{y'y'} > 0$ so y can not have corner points.

A useful generalization of these results is done substituting the scalar problem with vector valued extremals. The derivation is the same, we give only the comprehensive theorem.

Theorem 2.68 (Weierstrass-Erdmann for vector valued functions)**.** *For a domain $D \subset \mathbb{R}^{2n}$, let $f = f(x,Y,Y') \in C^1([a,b] \times D)$ and suppose that Y is a local extremal for*

$$F(Y) = \int_a^b f[Y(x)]\,\mathrm{d}x$$

on $\mathcal{D} = \{Y \in (\hat{C}^1[a,b])^n \mid Y(a) = A,\ Y(b) = B\}$. Then except at its corner points, Y is C^1 and satisfies the first and second Euler-Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{Y'}(x) = f_Y(x)$$

$$\frac{\mathrm{d}}{\mathrm{d}x}(f - Y' \cdot f_{Y'})(x) = f_x(x).$$

At each corner point c Y meets the Weierstrass-Erdmann conditions

1. $f_{Y'}(c^-) = f_{Y'}(c^+)$,
2. $(f - Y' \cdot f_{Y'})(c^-) = (f - Y' \cdot f_{Y'})(c^+)$,
3. $\pm f(c, Y(c), Y'(.))$ *can not be strictly convex in* Y'.

The last point of the Weierstrass-Erdmann conditions show that when present, second derivatives of f may give useful information about the location of corner points. In the next theorem of Hilbert, we show that at non corner points, a condition on the matrix $f_{Y'Y'}$ can guarantee higher differentiability of the extremal function.

Theorem 2.69 (Hilbert Differentiability Criterion)**.** *If $f_{Y'}$ is C^1 and $Y \in (C^1[a,b])^n$ is a solution of the integral equation*

$$f_{Y'}(x, Y(x), Y'(x)) = \int_a^x f_Y(t)\,\mathrm{d}t + K,$$

then Y is C^2 in a neighbourhood of each non corner point x_0 at which the matrix $f_{Y'Y'}$ is invertible.

When $n = 1$, the invertibility of matrix $f_{Y'Y'}$ reduces to the nonvanishing of the term $f_{y'y'}[y(x_0)]$.

Example 2.70. *For example the function $f(x, y, y') = e^{y(x)}\sqrt{1 + y'(x)^2}$ has extremals which are necessarily C^2. In facts, there are not corner points because $f_{y'y'} = e^y(x)/(1 + y'(x)^2)^{3/2} > 0$ never vanishes. Hence the Hilbert criterion is satisfied at all points.*

2.2.9 ***Necessary Conditions for Minima***

Characteristic of the Euler-Lagrange equations, is that they do not distinguish between maximal, minimal or saddle point behaviour neither globally nor locally. Therefore there are deep studies conducted by Legendre, Weierstrass and Jacobi starting from the first and the second derivatives of f.
To fix the ideas, we consider a functional to be minimized of kind

$$F(Y) = \int_a^b f[Y(x)]\,\mathrm{d}x = \int_a^b f(x, Y(x), Y'(x))\,\mathrm{d}x$$

locally on the set

$$\mathcal{D} = \{Y \in \mathcal{Y} = (\hat{C}^1[a,b])^n \mid Y(a) = A,\ Y(b) = B\}.$$

2.2.9.1 *The Weierstrass Condition*

Definition 2.71 (Weierstrass excess function)**.** *For a given function $f(x, y(x), y'(x))$, the* Weierstrass excess function *is defined as*

$$\mathcal{E}(x, y, y'; w) = f(x, y, w) - f(x, y, y') - (w - y')f_{y'}(x, y, y'). \tag{2.28}$$

We can notice that $f(x, y, y') + (w - y')f_{y'}(x, y, y')$ corresponds to the first order Taylor expansion of $f(x, y, w)$ interpreted as a function of w, around the point $w = y'$. This means that the Weierstrass excess function $\mathcal{E}(x, y, y'; w)$ measures the distance between the funtion f and its linear approximation around the point $w = y'$.

Theorem 2.72 (Weierstrass necessary condition)**.** *If $y(x)$ is a strong minimum, then*

$$\mathcal{E}(x, y, y'; w) \geq 0$$

for all non corner points and all $w \in \mathbb{R}$.

The geometric interpretation of this condition is that for each x, the graph of f (seen only as a function of y') lies above its tangent line at $y'(x)$, that is the function is locally convex.
It is interesting and more useful to reformulate this necessary conditions in terms of the Hamiltonian function, a theme that will be discussed in detail in the next chapter. This property of the Hamiltonian will lead to the Maximum (Minimum) Principle of Pontryagin. Supposing the variation problem is subject to a differential constraint of kind $y'(x) = g(x)$, introducing the associated multiplier $\lambda(x)$, the Hamiltonian becomes

$$\mathcal{H}(x, y, y', \lambda) = y'(x)\lambda(x) - f(x, y, y').$$

We can manipulate the Weierstrass excess function (2.28) in such a way to make the Hamiltonian appear:

$$\begin{aligned}\mathcal{E}(x, y, y'; w) &\equiv f(x, y, w) - f(x, y, y') - (w - y')f_{y'}(x, y, y') \\ &= [y' f_{y'}(x, y, y') - f(x, y, y')] - [w f_{y'}(x, y, y') - f(x, y, w)] \\ &= \mathcal{H}(x, y, y', \lambda) - \mathcal{H}(x, y, w, \lambda) \geq 0.\end{aligned}$$

2.2.9.2 *The Legendre Condition*

Now suppose to fix the (vectorial) variables x, $Y(x)$ and $Y'(x)$ and consider the excess function (2.28) only as a function depending on $\boldsymbol{w} \in \mathbb{R}^n$,

$$e(\boldsymbol{w}) = f(x, Y, \boldsymbol{w}) - f(x, Y, Y') - f_{Y'}(x, Y, Y')(\boldsymbol{w} - Y').$$

Both $e(\boldsymbol{w})$ and its gradient $e_{\boldsymbol{w}}(\boldsymbol{w}) = f_{Y'}(x, Y, \boldsymbol{w}) - f_{Y'}(x, Y, Y')$ vanish when $\boldsymbol{w} = Y'$. The second partial derivatives of $e(\boldsymbol{w})$ are given (when defined) by,

$$e_{w_i w_j}(Y') = f_{Y'_i Y'_j}(x, Y, Y') \qquad i, j = 1, 2, \ldots, n,$$

at the stationary point $\boldsymbol{w} = Y'$ of the excess e where $e(Y') = 0$.

Theorem 2.73 (Legendre necessary condition)**.** *If $f, f_Y, f_{Y'}$ are continuous on $[a, b] \times \mathbb{R}^{2n}$ and Y minimizes the functional F locally with respect to the strong norm $||.||_{\max}$, then Y satisfies the* Legendre condition

$$Q(x, \boldsymbol{v}) = \sum_{i,j=1}^{n} f_{Y'_i Y'_j}[Y(x)] v_i v_j \geq 0 \qquad \forall \boldsymbol{v} \in \mathbb{R}^n,$$

at each x at which the coefficient functions $f_{Y'_i Y'_j}$ are defined and continuous in the variable Y'. The condition is called strong Legendre condition *if the inequality is strict.*

Definition 2.74. *The function f is called* regular *if $Q(x, \boldsymbol{v}) > 0$ for all $x \in [a, b]$ and for all y and y'.*

The Weierstrass and the Legendre conditions are not equivalent, but if at some $x \in [a, b]$ the *strict* Legendre condition holds ($Q(x, \boldsymbol{v}) > 0$ for $\boldsymbol{v} \neq \boldsymbol{0}$), then for small $\boldsymbol{w}$ we have $\mathcal{E} > 0$. Moreover, the matrix $f_{Y'Y'}[Y(x)]$ is invertible and hence Y is C^2 in a neighbourhood of each non corner point.

Example 2.75 (Bolza's Problem)**.** *Consider $f(x, y, y') = f(y') = y'^2(y' + 1)^2$. Clearly $f_{y'} = 4y'^3 + 6y'^2 + 2y'$ and $f_{y'y'} = 2(6y'^2 + 6y' + 1)$. The linear function $y(x) = mx + q$ is stationary for f over $\mathcal{D}$ since $y'(x) = m$ is constant and*

$$\mathcal{D} = \{y \in C^1[a, b] \mid y(a) = a_1,\ y(b) = b_1\}.$$

In particular it is a computation to verify that $m = \frac{b_1 - a_1}{b - a}$ and $q = a_1 - ma = b_1 - mb$.
We have that $f_{y'y'} = 0$ when $m_\pm = -\frac{1}{2} \pm \frac{\sqrt{3}}{6} \approx -0.21, -0.78$. Therefore, for $m \leq m_-$ or $m \geq m_+$

the extremal y satisfies the Legendre condition $f_{y'y'} \geq 0$, and by strict convexity (e.g. Proposition 2.44) y provides the unique minimum for F.
If $-1 < m < 0$, then t can not give a strong local minimum for F over the piecewise *differentiable functions, because $F(y) = \int_a^b m^2(m+1)^2 \,\mathrm{d}x > 0$, while each strong norm neighbourhood of y contains a y for which $y' = 0$ or $y' = -1$ so that $F(y) = 0$. In this range, the Weierstrass condition is violated, indeed we have:*

$$\begin{aligned}\mathcal{E}(x, y, y'; w) &= f(w) - f(y') - f_{y'}(w - y') \\ &= f(w) - f(m) - f_{y'}(m)(w - m) \\ &= w^2(w+1)^2 - m^2(m+1)^2 - 2m(2m^2 + 3m + 1)(w - m) \\ &= (w - m)^2[(w + m + 1)^2 + 2m(m + 1)].\end{aligned}$$

This expression is negative for $-1 < m < 0$ when $w = -(m+1)$ and in particular for $-1 < m < m_-$ or $m_+ < m < 0$ provides a weak local minimum which is not a strong local minimum. With the same argument, $-1 < m < 0$ can not provide a strong local maximum.

2.2.10 *Sufficient Conditions for Minima*

We look now for sufficient conditions that characterize a minimum point when convexity is not present. In facts, convexity is a strong hypothesis that excludes many cases. An important feature, not exploited until now, is that, although arbitrary, the variations v and v' are not independent, but are connected by the relation $v' = \frac{\mathrm{d}v}{\mathrm{d}t}$. We turn back to the Legendre condition of the previous section.

Proposition 2.76. *Let f be C^2 and y extremal, if the strong Legendre condition holds, then y is C^2.*

Hence, when a problem is regular, any minimizing function is necessarily at least C^2. In general, it can be shown that if f is C^k with $k \geq 2$, then if the strong Legendre condition holds, an extremal function y is also C^k. We point out now that the Legendre condition (even the strong one) is only necessary but not sufficient to ensure the presence of a minimum. It can be derived starting from the second variation of the functional. Consider the usual functional

$$J(y) = \int_a^b f[y(x)] \,\mathrm{d}x = \int_a^b f(x, y(x), y'(x)) \,\mathrm{d}x,$$

the second variations is

$$\left.\frac{\mathrm{d}^2 J}{\mathrm{d}\varepsilon^2}\right|_{\varepsilon=0} = \int_a^b f_{yy}(x, y, y')v^2 + 2f_{yy'}(x, y, y')vv' + f_{y'y'}v'^2 \,\mathrm{d}x. \tag{2.29}$$

By integration by parts and because $v(a) = v(b) = 0$, we have

$$\int_a^b 2f_{yy'}vv' \,\mathrm{d}x = v^2 f_{yy'}\Big|_a^b - \int_a^b v^2 \frac{\mathrm{d}}{\mathrm{d}x} f_{yy'} \,\mathrm{d}x = -\int_a^b v^2 \frac{\mathrm{d}}{\mathrm{d}x} f_{yy'} \,\mathrm{d}x,$$

and thus the second variation simplifies to

$$\left.\frac{\mathrm{d}^2 J}{\mathrm{d}\varepsilon^2}\right|_{\varepsilon=0} = \int_a^b \left(f_{yy}(x, y, y') - \frac{\mathrm{d}}{\mathrm{d}x} f_{yy'}\right) v^2 + f_{y'y'}v'^2 \,\mathrm{d}x.$$

Because y is minimizing, the second variations has to be positive, and being v arbitrary, we must have

$$f_{yy}(x, y, y') - \frac{\mathrm{d}}{\mathrm{d}x} f_{yy'} \geq 0, \qquad f_{y'y'} \geq 0.$$

This is not a sufficient condition as we show in the next counterexample. Suppose the strong Legendre condition holds, and define

$$p(x) = f_{y'y'}, \qquad q(x) = f_{yy}(x, y, y') - \frac{\mathrm{d}}{\mathrm{d}x} f_{yy'}.$$

Consider any function $g(x)$ in $C^1[a, b]$. From $v(a) = v(b) = 0$ we have

$$0 = g(x)v(x)^2\Big|_a^b = \int_a^b \frac{\mathrm{d}}{\mathrm{d}x} g(x)v(x)^2 \,\mathrm{d}x = \int_a^b (g'v^2 + 2gvv')(x) \,\mathrm{d}x.$$

Substituting the last line in the second variation yields indeed,

$$\begin{aligned} \frac{\mathrm{d}^2 J}{\mathrm{d}\varepsilon^2}\Big|_{\varepsilon=0} &= \int_a^b (p(x)v'(x)^2 + q(x)v(x)^2 \,\mathrm{d}x \\ &= \int_a^b p(x)v'(x)^2 + 2gv(x)v'(x) + (q(x) + g'(x))v(x)^2 \,\mathrm{d}x \end{aligned}$$

The last quantity is a perfect square if and only if $g(x)^2 = p(x)(q(x) + g'(x))$. But in general this condition will not hold, therefore the quadratic form above can be not positive defined, thus y is not minimizing. What we need, is the so called *Jacobi condition*. Consider the integrand of equation (2.29),

$$\varphi(x, v, v') = f_{yy}v^2 + 2f_{yy'}vv' + f_{y'y'}v'^2 \implies \frac{\mathrm{d}^2 J}{\mathrm{d}\varepsilon^2}\Big|_{\varepsilon=0} = \int_a^b \varphi(x, v, v') \,\mathrm{d}x,$$

we set the so called *accessory* minimum problem:

$$\min_{v \in \mathring{C}^1[a,b]} \Phi(v) = \int_a^b \varphi(x, v, v') \,\mathrm{d}x \qquad \text{s.t.} \quad v(a) = v(b) = 0.$$

We study the accessory minimum problem to derive the Jacobi condition, that together with the strong Legendre condition is sufficient to characterize the presence of a minimum for the original problem. Suppose v is extremal for the accessory minimum problem, then the second variation of Φ must vanish. To see this, rewrite φ in the following way:

$$2\varphi(x, v, v') = \varphi_v(x, v, v')v + \varphi_{v'}(x, v, v')v'.$$

Being v extremal for Φ, it must satisfy the Euler-Lagrange equation,

$$\frac{\mathrm{d}}{\mathrm{d}x}\varphi_{v'} = \varphi_v, \tag{2.30}$$

hence the second variation of Φ becomes

$$\begin{aligned} 2\frac{\mathrm{d}^2\Phi}{\mathrm{d}\varepsilon^2}\Big|_{\varepsilon=0} &= 2\int_a^b \varphi(x, v, v') \,\mathrm{d}x \\ &= \int_a^b [\varphi_v(x, v, v')v(x) + \varphi_{v'}(x, v, v')v'(x)] \,\mathrm{d}x \\ &= \int_a^b \left[v(x)\frac{\mathrm{d}}{\mathrm{d}x}\varphi_{v'}(x, v, v') + v'(x)\varphi_{v'}(x, v, v')\right] \mathrm{d}x \\ &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}x}\left(\varphi_{v'}(x, v, v')v(x)\right) \mathrm{d}x \\ &= \varphi_{v'}(x, v, v')v(x)\Big|_a^b = 0. \end{aligned}$$

Notice that if f is C^4 and the strong Legendre condition holds for y, then φ is regular. The necessary condition (2.30) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}x}\varphi_{v'} = \varphi_v \implies \frac{\mathrm{d}}{\mathrm{d}x}\left(f_{yy}v + f_{y'y'}v'\right) = f_{yy'}v' + f_{yy}v. \tag{2.31}$$

Equation (2.31) is called the *Jacobi equation*.

Definition 2.77 (Conjugated point)**.** *A point $\xi \in (a, b]$ is conjugated to a if there exists a non null function $v_\xi : [a, \xi] \to \mathbb{R}$ such that $v_\xi \in \hat{C}^1$, $v_\xi(a) = v_\xi(\xi) = 0$ and v_ξ satisfies the Jacobi equation* (2.31)*.*

Definition 2.78 (Jacobi condition)**.** *The* Jacobi condition *holds if there are not conjugated points to a in (a, b). The* strong *Jacobi condition holds if there are not conjugated points to a in $(a, b]$.*

Theorem 2.79 (Jacobi necessary condition)**.** *Let y be C^3 and minimizing, f be C^4 and suppose that the strong Legendre condition holds. Then the Jacobi condition is satisfied.*

If we add the strong conditions, the previous theorem gives sufficient conditions for minima.

Theorem 2.80 (Jacobi sufficient condition)**.** *Let y be C^3 extremal, f be C^4 and suppose that the strong Legendre condition and the strong Jacobi condition hold. Then y gives a local minimum.*

Example 2.81. *Consider the problem*

$$\min \int_0^{3\pi/2} y'(x)^2 - y(x)^2 - 2y(x)\,\mathrm{d}x \qquad \text{s.t.} \quad y(0) = y\left(\frac{3\pi}{2}\right) = 0.$$

The Euler-Lagrange equation for this problem is $y'' + y = -1$, so the general integral is $y(x) = \alpha \sin x + \beta \cos x - 1$. The boundary conditions give $\alpha = 1$, $\beta = -1$. Thus a candidate to be an extremal is $y(x) = \sin x - \cos x - 1$. The Hessian of $f = y'^2 - y^2 - 2y$ with respect to y and y' is $(-2\ 0; 0\ 2)$ and is therefore not definite. The Legendre condition is clearly satisfied, because $f_{y'y'} = 2 > 0$ (hence also $Q(x, v) > 0$) for all $x \in [0, 3\pi/2]$. Function y is a minimizing candidate. We have to look at the Jacobi sufficient condition. We need to check if there is a conjugated point $\xi \in (0, 3\pi/2]$ to $x = 0$. The Jacobi equation for the accessory minimum, being $f_{yy} = -2$, $f_{yy'} = 0$, $f_{y'y'} = 2$, is:

$$\frac{\mathrm{d}}{\mathrm{d}x}2v'(x) = -2v(x) \implies v(x) = A\sin x + B\cos x.$$

The boundary conditions $v(0) = v(\xi) = 0$ give $v(x) = 0$ for $\xi \in (0, \pi) \cup (\pi, 3\pi/2)$, but $v(x) = B\cos x$ for all $B \in \mathbb{R}$, for $\xi = \pi$. Thus there is at least one non null solution to the accessory minimum problem that matches Jacobi equation, i.e. $\xi = \pi$ is a conjugated point to 0 and y can not be minimum.

3

OPTIMAL CONTROL

It turns out that scientists had been studying optimal control problems for quite a few years before they realized that it is part of the calculus of variations. The main issue was that they had to laid down assumptions of smoothness. In fact, OCPs are more general than the problems described so far. The Maximum (Minimum) Principle of Pontryagin applies to all problems that arise in calculus of variations, and gives equivalent results to those expected by the classical approach so far exposed. However, the two approaches differ and optimal control gives insights into problems that are less readily apparent in the calculus of variations. It also works for some classes of problems for which the calculus of variations is not useful, such as thse involving constraints on the derivatives of the unknown optimal function. This kind of constraints is very convenient for example when we have

to characterize increments that can not be negative. Let us see the connection that makes each problem of calculus of variation an optimal control problem, this will also show the generality of the formulation of an OCP. Consider a problem of kind

$$\min_{y\in C^1(a,b)} \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x, \tag{3.1}$$

with $y(a) = y_0$. If we fix $y(x)$, we can introduce a new variable $z(x)$ such that

$$z'(x) = f(x, y(x), y'(x)), \qquad z(a) = 0.$$

In this way, we have

$$z(b) = z(b) - z(a) = \int_a^b z'(x)\,\mathrm{d}x = \int_a^b f(x, y(x), y'(x))\,\mathrm{d}x,$$

and calling $u(x) = y'(x)$ the previous problems can be restated as a *terminal control* problem:

$$\min z(b) \qquad \text{s.t.} \quad y'(x) = u(x),\ z'(x) = f(x, y(x), u(x)), \tag{3.2}$$

with initial conditions $y(a) = y_0$ and $z(a) = 0$. We have to find the *control* $u(x)$ at which $z(b)$ attains its minimal value. This formulation does not involve the operation of integration, and it is well known that the solution of a Cauchy problem for a system of ordinary differential equations (ODE) is less computationally expensive than the solution of the corresponding integral equations formulation. Although problem (3.2) is equivalent to problem (3.1), it has extended the class of problems that we can treat, since it is not a problem of calculus of variations. If we consider (3.2) from a vectorial point of view, we can assume $\boldsymbol{x} \in \mathbb{R}^n$, $f(x, y, y') : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and the ODE from $\mathbb{R}^{2n+1}$ to $\mathbb{R}^m$. The problem is *nondegenerate* if $m < n$, otherwise the minimization is done over a discrete or an empty set of trajectories. For nondegenerate problems, the condition $m < n$ in the n-dimensional space of variables y', defines an $(n - m)$-dimensional manifold M, and in a neighbourhood of a point in M we can introduce coordinates $U \in \mathbb{R}^{n-m}$ and parametrize the manifold. In these coordinates, the system of ODE depending on (x, y, y') is equivalent to the system of coordinates (x, y, u). Therefore, classical variational problems can be seen as optimal control problems. A typical example is given by the brachistochrone, we will go through it later in detail.

3.1 THE PROBLEMS OF MAYER, LAGRANGE AND BOLZA

There are three types of optimal control problems, they differ apparently in the formulation of the functional to be optimized, but we will show that they are equivalent and that it is possible to convert each problem in the other two forms.

3.1.1 *The Problem of Mayer*

In the problem of Mayer, the functional is not an integral but a function M that depends in general from the dependent variable x and the final point of the x-domain. Often this is useful to be intended as a problem of optimizing the final *time*, such as in the time-optimal OCPs. The objective function is called *pay off* function and is constrained by a set of differential equations, in general ODE, but often we encounter also differential algebraic equations (DAE). The standard formulation is

$$\begin{aligned} \min_{\boldsymbol{u}\in U} J(\boldsymbol{u}) &= M(b, \boldsymbol{y}(b)) \\ \boldsymbol{y}'(x) &= g(x, \boldsymbol{y}(x), \boldsymbol{u}(x)) \qquad x \in (a, b) \\ \boldsymbol{y}(a) &= \boldsymbol{y}_0. \end{aligned} \tag{3.3}$$

The set U represents a general class of function available for the control, for example the control can be C^1 or $\hat{C}^1$, other possibilities are the piecewise constant functions. U can also contain limitations for the control, a classical case is $|u| \leq 1$. There are also more general formulations for the OCP, adding inequality state constraints or jump conditions. We will define more general problems in the next chapters. In some applications it is convenient to consider an *unbounded* domain for x, for example $x \in [a, \infty)$, or the final point b can be an unknown. We will analyse also these eventualities later.

3.1.2 *The Problem of Lagrange*

In the problem of Lagrange, we have the objective functional in (pure) integral form. The previous considerations for the Mayer problem still hold. The standard formulation is

$$\begin{aligned} \min_{\boldsymbol{u}\in U} J(\boldsymbol{u}) &= \int_a^b f(x, \boldsymbol{y}, \boldsymbol{u})\,\mathrm{d}x \\ \boldsymbol{y}'(x) &= g(x, \boldsymbol{y}(x), \boldsymbol{u}(x)) \qquad x \in (a, b) \\ \boldsymbol{y}(a) &= \boldsymbol{y}_0. \end{aligned} \tag{3.4}$$

3.1.3 *The Problem of Bolza*

The third standard form for an OCP is given by the formulation of Bolza, which consider a linear combination of the problems of Mayer and Lagrange.

$$\begin{aligned} \min_{\boldsymbol{u}\in U} J(\boldsymbol{u}) &= M(b, \boldsymbol{y}(b)) + \int_a^b f(x, \boldsymbol{y}, \boldsymbol{u})\,\mathrm{d}x \\ \boldsymbol{y}'(x) &= g(x, \boldsymbol{y}(x), \boldsymbol{u}(x)) \qquad x \in (a, b) \\ \boldsymbol{y}(a) &= \boldsymbol{y}_0. \end{aligned} \tag{3.5}$$

As before, the considerations done for the Mayer problem hold.

3.1.4 *Equivalence of the Three Problems*

Even if the Bolza problem looks more general than the other two, we show next that the three formulations are equivalent. It is clear that problems (3.3) and (3.4) are particular cases of (3.5), hence we have to show how (3.5) becomes (3.4) and how (3.4) becomes (3.3).

3.1.4.1 *From Bolza to Lagrange*

To do this conversion, add a new component to the vector $\boldsymbol{y} \in \mathbb{R}^n$, so that $y_{n+1}(x) = M(x, \boldsymbol{y}(x))$. According to this notation, the problem of Bolza becomes

$$\begin{aligned} \min_{\boldsymbol{u}\in U} \tilde{J}(\boldsymbol{u}) &= \int_a^b f(x, \boldsymbol{y}, \boldsymbol{u}) + y'_{n+1}(x)\,\mathrm{d}x \qquad x \in [a, b] \\ \begin{pmatrix} \boldsymbol{y}' \\ y'_{n+1} \end{pmatrix} &= \begin{pmatrix} g(x, \boldsymbol{y}, \boldsymbol{u}) \\ \frac{\mathrm{d}}{\mathrm{d}x} M(x, \boldsymbol{y}(x)) \end{pmatrix} \\ \begin{pmatrix} \boldsymbol{y}(a) \\ y_{n+1}(a)) \end{pmatrix} &= \begin{pmatrix} \boldsymbol{y}_0 \\ M(a, \boldsymbol{y}_0) \end{pmatrix}, \end{aligned}$$

which is a problem of Lagrange.

3.1.4.2 *From Lagrange to Mayer*

To transform (3.4) into a Mayer (3.3), consider a new variable y_{n+1} defined as $y'_{n+1}(x) = f(x, \boldsymbol{y}, \boldsymbol{u})$, with the initial condition $y_{n+1}(a) = 0$. Hence the problem of Lagrange becomes

$$\begin{aligned} \min_{\boldsymbol{u}\in U} \tilde{J}(\boldsymbol{u}) &= y_{n+1}(b) \qquad x \in [a, b] \\ (\boldsymbol{y}', y'_{n+1}) &= (g(x, \boldsymbol{y}, \boldsymbol{u}), f(x, \boldsymbol{y}, \boldsymbol{u})) \\ (\boldsymbol{y}(a), y_{n+1}(a)) &= (\boldsymbol{y}_0, 0), \end{aligned}$$

which is a problem of Mayer.
Finally we show how to pass from a Mayer to a Lagrange.

3.1.4.3 *From Mayer to Lagrange*

Consider a new variable y_{n+1} defined as $y'_{n+1}(x) = 0$ with the condition that $y_{n+1} = \frac{M(x,\boldsymbol{y}(b))}{b-a}$. The Mayer problem becomes then

$$\begin{aligned} \min_{\boldsymbol{u}\in U} \tilde{J}(\boldsymbol{u}) &= \int_a^b y_{n+1}(x)\,\mathrm{d}x \qquad x \in [a, b] \\ (\boldsymbol{y}', y'_{n+1}) &= (g(x, \boldsymbol{y}, \boldsymbol{u}), 0) \\ (\boldsymbol{y}(a), y_{n+1}(a)) &= \left(\boldsymbol{y}_0, \frac{M(x, \boldsymbol{y}(b))}{b-a}\right), \end{aligned}$$

which is a problem of Lagrange.

3.2 HAMILTONIAN FORMALISM

A fundamental tool in the solution of variational problems was suggested by Hamilton, it is the Legendre transform of a function f as a function of y' for fixed values of x and y is denoted by $\mathcal{H}(x, y, y')$, i.e. the Hamiltonian[1]:

$$\lambda(x) = f_{y'}(x, y(x), y'(x)), \qquad \mathcal{H}(x, y, v, \lambda) = \langle \lambda, v\rangle - f(x, y, v). \tag{3.6}$$

Taking the total derivative w.r.t. x of the Hamiltonian (3.6) with $v = y'$, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{H}(x, y, y') = \langle \lambda', y'\rangle + \left\langle \lambda, \frac{\mathrm{d}}{\mathrm{d}x}y'\right\rangle - f_x - \left\langle f_y, \frac{\mathrm{d}}{\mathrm{d}x}y\right\rangle - \left\langle f_{y'}, \frac{\mathrm{d}}{\mathrm{d}x}y'\right\rangle.$$

Notice that by the first of (3.6) and by the Euler-Lagrange equation, we can simplify

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{H}(x, y, y', \lambda) &= \langle \lambda', y'\rangle + \left\langle f_{y'}, \frac{\mathrm{d}}{\mathrm{d}x}y'\right\rangle - f_x - \left\langle f_y, \frac{\mathrm{d}}{\mathrm{d}x}y\right\rangle - \left\langle f_{y'}, \frac{\mathrm{d}}{\mathrm{d}x}y'\right\rangle \\ &= \left\langle \frac{\mathrm{d}}{\mathrm{d}x}f_{y'_i}, y'\right\rangle - f_x - \langle f_y, y'\rangle \\ &= \left\langle \frac{\mathrm{d}}{\mathrm{d}x}f_{y'_i} - f_y, y'\right\rangle - f_x \\ &= -f_x. \end{aligned}$$

[1] $\langle ., .\rangle$ is the standard scalar product.

This shows that the total derivative of the Hamiltonian is equal to the partial derivative of the Hamiltonian w.r.t. x, e.g.

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{H}(x,y,y',\lambda) &= \frac{\partial}{\partial x}\mathcal{H}(x,y,y',\lambda) \\ y'(x) &= \frac{\partial}{\partial \lambda}\mathcal{H}(x,y,y',\lambda) \\ -\lambda'(x) &= \frac{\partial}{\partial y}\mathcal{H}(x,y,y',\lambda) \\ 0 &= \frac{\partial}{\partial y'}\mathcal{H}(x,y,y',\lambda) \end{aligned} \tag{3.7}$$

These equations are called the *canonical system*. In particular, the first equation shows that for autonomous problems (i.e. when $\mathcal{H}$ does not depend explicitly on x) the Hamiltonian is constant along an optimal trajectory, because $f_x = 0$, moreover, if the final point $x = b$ is free, the Hamiltonian is zero.
The geometric meaning of the transform can be expressed as follows. We consider the function $y = h(x)$ and its graph $(x, y = h(x))$. From the first relation of (3.6) we have that the tangent to the function at a point x_0 has slope λ, from the second equation of (3.6) we have that $h^\star(\lambda)$ is the value by which the line $y = \lambda x$ should be lowered to become the tangent to the graph of function $h(x)$. Thus the function $h^\star(\lambda)$ defines a set of tangents of the function $y = h(x)$. In the vectorial case, $h^\star$ represents the value by which the plane $y = \langle \boldsymbol{\lambda}, \boldsymbol{x} \rangle$ should be lowered to become the tangent plane of $y = h(\boldsymbol{x})$.

Example 3.1. *Consider the function $y = h(x) = (x-1)^2 + 1$, the tangent at point $x = 3$ is given by the first order Taylor polynomial*

$$T_1(x) = h(x_0) + h'(x_0)(x - x_0) = 5 + 4(x-3) = 4x - 7.$$

Let us see it with the Legendre transform. We have $\lambda = h'(x) = 2x - 2$ and evaluated at x_0 gives $\lambda = 4$. The transform gives $h^\star(\lambda) = \lambda x - (x-1)^2 - 1$. At the required point its value is 7, so the line $y = \lambda x = 4x$ should be lowered by 7 to be the tangent of h at x_0. This gives exactly the expected solution.

3.3 THE FIRST VARIATION

Here it is convenient to change our notation and follow the literature: in variational problems, it is commonly used the variable x for the independent variable and $y(x)$ for the state variables, the variations are expressed as v. In optimal control theory, very often the dependent variable is the time and the state is $x(t)$, the variations involve many functions, so they are expressed with a δ followed by the corresponding variable, e.g. the variation of $\lambda(t)$ becomes $\delta\lambda$. Most authors follow this convention, so we restart stating a basic optimal control problem and deriving the "classic " canonical system again. We present a general problem of Bolza and we will investigate more involved problems, like Hestenes's problem later. We consider a time interval $[t_0, T]$ with a Mayer term M and a Lagrange term L, the functional to be minimized is

$$\begin{aligned} J(\boldsymbol{u}) &= M(t_0, \boldsymbol{x}(t_0), T, \boldsymbol{x}(T)) + \int_{t_0}^{T} L(t, \boldsymbol{x}(t), \boldsymbol{u}(t))\,\mathrm{d}t \qquad \text{s.t.} \\ \boldsymbol{x}'(t) &= \boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{u}) \\ \boldsymbol{B}(t_0, \boldsymbol{x}(t_0), T, \boldsymbol{x}(T)) &= \boldsymbol{0}. \end{aligned} \tag{3.8}$$

The *state* vector $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{u} \in \mathbb{R}^m$ denotes the control vector, M is the scalar Mayer term, L is the scalar Lagrange term, $\boldsymbol{f}$ is an ODE with values in $\mathbb{R}^n$, $\boldsymbol{B} \in \mathbb{R}^{p+1}$ is the vector of the boundary

conditions with $p \leq n$. To derive the canonical system consider the *Lagrangian* $\mathcal{L}$ of the problem, which is known also with the name of *augmented functional*

$$\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = M + \boldsymbol{\nu}^T \boldsymbol{B} + \int_{t_0}^{T} L + \boldsymbol{\lambda}^T (\boldsymbol{f} - \boldsymbol{x}') \, \mathrm{d}t.$$

The (first) variation of the Lagrangian gives the first order necessary conditions of optimality:

$$\delta\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \delta M + \delta\boldsymbol{\nu}^T \boldsymbol{B} + \boldsymbol{\nu}^T \delta\boldsymbol{B} + \delta \int_{t_0}^{T} L + \boldsymbol{\lambda}^T (\boldsymbol{f} - \boldsymbol{x}') \, \mathrm{d}t.$$

The resulting expressions for each variation depend on the characteristics of the problem, in particular on the presence of free or fixed boundary conditions, on the presence of constraints on the control $\boldsymbol{u}$. They become easily long, therefore we analyse them separately. For the Mayer term we have

$$\begin{aligned} \delta M \; = \; & \left[\frac{\partial M}{\partial \boldsymbol{x}(t_0)} \boldsymbol{x}'(t_0) + \frac{\partial M}{\partial t_0} \right] \delta t_0 + \frac{\partial M}{\partial \boldsymbol{x}(t_0)} \delta \boldsymbol{x}_0 + \\ & \left[\frac{\partial M}{\partial \boldsymbol{x}(T)} \boldsymbol{x}'(T) + \frac{\partial M}{\partial T} \right] \delta T + \frac{\partial M}{\partial \boldsymbol{x}(T)} \delta \boldsymbol{x}_T. \end{aligned}$$

With analogous computation we derive the variation of the boundary condition $\delta\boldsymbol{B}$,

$$\begin{aligned} \delta \boldsymbol{B} \; = \; & \left[\frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}(t_0)} \boldsymbol{x}'(t_0) + \frac{\partial \boldsymbol{B}}{\partial t_0} \right] \delta t_0 + \frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}(t_0)} \delta \boldsymbol{x}_0 + \\ & \left[\frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}(T)} \boldsymbol{x}'(T) + \frac{\partial \boldsymbol{B}}{\partial T} \right] \delta T + \frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}(T)} \delta \boldsymbol{x}_T. \end{aligned}$$

The term $\delta\boldsymbol{\nu}^T \boldsymbol{B}$ can not be further simplified, so we consider now the variation of integral,

$$\begin{aligned} & \delta \int_{t_0}^{T} \mathcal{H} - \boldsymbol{\lambda} \cdot \boldsymbol{x}' \, \mathrm{d}t = \\ & = \int_{t_0}^{T} \delta\mathcal{H} - \delta\boldsymbol{\lambda}^T \boldsymbol{x}' - \boldsymbol{\lambda}^T \delta\boldsymbol{x}' \, \mathrm{d}t + \left[\mathcal{H} - \boldsymbol{\lambda}^T \boldsymbol{x}' \right] \Big|_T \delta T - \left[\mathcal{H} - \boldsymbol{\lambda}^T \boldsymbol{x}' \right] \Big|_{t_0} \delta t_0. \end{aligned}$$

Again, we simplify the single variations. For the Hamiltonian we have

$$\delta\mathcal{H} = \frac{\partial \mathcal{H}}{\partial \boldsymbol{x}} \delta\boldsymbol{x} + \frac{\partial \mathcal{H}}{\partial \boldsymbol{\lambda}} \delta\boldsymbol{\lambda} + \frac{\partial \mathcal{H}}{\partial \boldsymbol{u}} \delta\boldsymbol{u}.$$

The term $-\boldsymbol{\lambda}^T \delta\boldsymbol{x}'$ can be reduced to first order terms by integration by parts,

$$\begin{aligned} \int_{t_0}^{T} -\boldsymbol{\lambda}^T \delta\boldsymbol{x}' \, \mathrm{d}t \; = \; & - \left[\boldsymbol{\lambda}^T \delta\boldsymbol{x} \right]_{t_0}^{T} + \int_{t_0}^{T} \boldsymbol{\lambda}'^T \delta\boldsymbol{x} \, \mathrm{d}t \\ & = \boldsymbol{\lambda}^T(t_0) \delta\boldsymbol{x}_0 - \boldsymbol{\lambda}^T(T) \delta\boldsymbol{x}_T + \int_{t_0}^{T} \boldsymbol{\lambda}'^T \delta\boldsymbol{x} \, \mathrm{d}t. \end{aligned} \tag{3.9}$$

The term $\delta\boldsymbol{\lambda}^T\boldsymbol{x}'$ can be rewritten as $\boldsymbol{x}'^T\delta\boldsymbol{\lambda}$, therefore, putting all the expansions together, the variation of the integral is

$$\begin{aligned}\delta\int_{t_0}^{T}\mathcal{H}-\boldsymbol{\lambda}^T\boldsymbol{x}'\,\mathrm{d}t \;=\; &\boldsymbol{\lambda}^T(t_0)\delta\boldsymbol{x}_0-\boldsymbol{\lambda}^T(T)\delta\boldsymbol{x}_T+\\ &\left[\mathcal{H}(T)-\boldsymbol{\lambda}^T(T)\boldsymbol{x}'(T)\right]\delta T-\left[\mathcal{H}(t_0)-\boldsymbol{\lambda}^T(t_0)\boldsymbol{x}'(t_0)\right]\delta t_0+\\ &\int_{t_0}^{T}\left[\frac{\partial\mathcal{H}}{\partial\boldsymbol{x}}+\boldsymbol{\lambda}'^T\right]\delta\boldsymbol{x}+\left[\frac{\partial\mathcal{H}}{\partial\boldsymbol{\lambda}}-\boldsymbol{x}'^T\right]\delta\boldsymbol{\lambda}+\frac{\partial\mathcal{H}}{\partial\boldsymbol{u}}\delta\boldsymbol{u}\,\mathrm{d}t\end{aligned}$$

Because the first variation should be zero to satisfy the first order necessary conditions and because the variations are independent, we can collect them to obtain the general form for first order conditions.

$$\begin{aligned}\delta\boldsymbol{\lambda}:\quad &\boldsymbol{x}'=\frac{\partial\mathcal{H}}{\partial\boldsymbol{\lambda}}=\boldsymbol{f}\\ \delta\boldsymbol{x}:\quad &\boldsymbol{\lambda}'=-\frac{\partial\mathcal{H}}{\partial\boldsymbol{x}}\\ \delta\boldsymbol{u}:\quad &\mathbf{0}=\frac{\partial\mathcal{H}}{\partial\boldsymbol{u}}\\ \delta\boldsymbol{x}_0:\quad &\boldsymbol{\lambda}(t_0)=-\frac{\partial M}{\partial\boldsymbol{x}(t_0)}^T-\frac{\partial\boldsymbol{B}}{\partial\boldsymbol{x}(t_0)}^T\boldsymbol{\nu}\\ \delta t_0:\quad &\mathcal{H}(t_0)=\frac{\partial M}{\partial t_0}+\boldsymbol{\nu}^T\frac{\partial\boldsymbol{B}}{\partial t_0}\\ \delta\boldsymbol{x}_T:\quad &\boldsymbol{\lambda}(T)=\frac{\partial M}{\partial\boldsymbol{x}(T)}^T+\frac{\partial\boldsymbol{B}}{\partial\boldsymbol{x}(T)}^T\boldsymbol{\nu}\\ \delta T:\quad &\mathcal{H}(T)=-\frac{\partial M}{\partial T}-\boldsymbol{\nu}^T\frac{\partial\boldsymbol{B}}{\partial T}\end{aligned}\tag{3.10}$$

Notice that the variations $\mathrm{d}\boldsymbol{x}_0$ and $\mathrm{d}t_0$ (and similarly $\mathrm{d}\boldsymbol{x}_T$ and $\mathrm{d}T$) are not independent, so they have to vanish together in the case of free boundary conditions.

3.4 THE SECOND VARIATION

In order to compute the second variation for problem (3.8), it is convenient to restate it in a more compact way.

$$\begin{aligned}J(\boldsymbol{u}) \;&=\; M(t_0,\boldsymbol{x}(t_0),T,\boldsymbol{x}(T))+\int_{t_0}^{T}\mathcal{H}-\boldsymbol{\lambda}\cdot\boldsymbol{x}'\,\mathrm{d}t\qquad\text{s.t.}\\ \boldsymbol{x}'(t) \;&=\; \boldsymbol{f}(t,\boldsymbol{x},\boldsymbol{u})\\ \boldsymbol{B}(t_0,\boldsymbol{x}(t_0),T,\boldsymbol{x}(T)) \;&=\; \mathbf{0}\end{aligned}$$

We compute again the first variation to perform another one and obtain the required second variation. This time we do not collect the independent variation $\delta\boldsymbol{x}_0$ and $\mathrm{d}t_0$ in $\mathrm{d}\boldsymbol{x}_0$ and similarly for the final point, and we work directly with $\delta\boldsymbol{x}_0$ and $\mathrm{d}t_0$. The first variation of the Lagrangian gives the first order necessary conditions of optimality:

$$\delta\mathcal{L}(t,\boldsymbol{x},\boldsymbol{u},\boldsymbol{\lambda},\boldsymbol{\nu})=\delta N+\delta\int_{t_0}^{T}\mathcal{H}-\boldsymbol{\lambda}\cdot\boldsymbol{x}'\,\mathrm{d}t.$$

where

$$N(t_0, \boldsymbol{x}(t_0), T, \boldsymbol{x}(T), \boldsymbol{\nu}) = M(t_0, \boldsymbol{x}(t_0), T, \boldsymbol{x}(T)) + \boldsymbol{\nu}^T \boldsymbol{B}(t_0, \boldsymbol{x}(t_0), T, \boldsymbol{x}(T)).$$

Adopting the more compact notation for the derivatives, we have:

$$\begin{aligned}\delta\mathcal{L} = \; & N_{t_0}\delta t_0 + N_{\boldsymbol{x}_0}\cdot\delta\boldsymbol{x}_0 + N_T\delta T + N_{\boldsymbol{x}_T}\cdot\delta\boldsymbol{x}_T + N_{\boldsymbol{\nu}}\cdot\delta\boldsymbol{\nu} \\ & +[\mathcal{H}(T) - \boldsymbol{\lambda}(T)\cdot\boldsymbol{x}'(T)]\delta T - [\mathcal{H}(t_0) - \boldsymbol{\lambda}(t_0)\cdot\boldsymbol{x}'(t_0)]\delta t_0 \\ & + \int_{t_0}^{T} \mathcal{H}_{\boldsymbol{x}}\cdot\delta\boldsymbol{x} + \mathcal{H}_{\boldsymbol{u}}\cdot\delta\boldsymbol{u} + \mathcal{H}_{\boldsymbol{\lambda}}\cdot\delta\boldsymbol{\lambda} - \delta\boldsymbol{\lambda}\cdot\boldsymbol{x}' - \boldsymbol{\lambda}\cdot\delta\boldsymbol{x}'\,\mathrm{d}t.\end{aligned} \tag{3.11}$$

We can integrate by parts the term $\boldsymbol{\lambda}\cdot\delta\boldsymbol{x}'$ inside the integral, as was done in (3.9), obtaining

$$\begin{aligned}\delta\mathcal{L} = \; & N_{t_0}\delta t_0 + N_{\boldsymbol{x}_0}\cdot(\delta\boldsymbol{x}_0 + \boldsymbol{x}'(t_0)\delta t_0) + N_T\delta T + N_{\boldsymbol{x}_T}\cdot(\delta\boldsymbol{x}_T + \boldsymbol{x}'(T)\delta T) + N_{\boldsymbol{\nu}}\cdot\delta\boldsymbol{\nu} \\ & +[\mathcal{H}(T) - \boldsymbol{\lambda}(T)\cdot\boldsymbol{x}'(T)]\delta T - [\mathcal{H}(t_0) - \boldsymbol{\lambda}(t_0)\cdot\boldsymbol{x}'(t_0)]\delta t_0 \\ & +\boldsymbol{\lambda}^T(t_0)\cdot\delta\boldsymbol{x}_0 - \boldsymbol{\lambda}^T(T)\cdot\delta\boldsymbol{x}_T \\ & + \int_{t_0}^{T} (\mathcal{H}_{\boldsymbol{x}} + \boldsymbol{\lambda}')\cdot\delta\boldsymbol{x} + \mathcal{H}_{\boldsymbol{u}}\cdot\delta\boldsymbol{u} + (\mathcal{H}_{\boldsymbol{\lambda}} - \boldsymbol{x}')\cdot\delta\boldsymbol{\lambda}\,\mathrm{d}t \\ = \; & \delta t_0\,(N_{t_0} + \boldsymbol{x}'(t_0)\cdot N_{\boldsymbol{x}_0} - \mathcal{H}(t_0) + \boldsymbol{\lambda}(t_0)\cdot\boldsymbol{x}'(t_0)) + \delta\boldsymbol{x}_0\cdot(N_{\boldsymbol{x}_0} + \boldsymbol{\lambda}(t_0)) \\ & +\delta T\,(N_T + \boldsymbol{x}'_T\cdot N_{\boldsymbol{x}_T} + \mathcal{H}(T) - \boldsymbol{\lambda}(T)\cdot\boldsymbol{x}'(T)) + \delta\boldsymbol{x}_T\cdot(N_{\boldsymbol{x}_T} - \boldsymbol{\lambda}(T)) \\ & + \int_{t_0}^{T} (\mathcal{H}_{\boldsymbol{x}} + \boldsymbol{\lambda}')\cdot\delta\boldsymbol{x} + \mathcal{H}_{\boldsymbol{u}}\cdot\delta\boldsymbol{u} + (\boldsymbol{f} - \boldsymbol{x}')\cdot\delta\boldsymbol{\lambda}\,\mathrm{d}t.\end{aligned}$$

The variation relative to $\boldsymbol{\nu}$ can be set to zero because $N_{\boldsymbol{\nu}} = \boldsymbol{B} = \boldsymbol{0}$. The next step is to denote (as in Hull [Hul03]) by Γ and Ω the coefficients of δt_0 and δT respectively, that is

$$\Gamma = N_{t_0} + \boldsymbol{x}'(t_0)\cdot N_{\boldsymbol{x}_0} - L(t_0), \qquad \Omega = N_T + \boldsymbol{x}'(T)\cdot N_{\boldsymbol{x}_T} + L(T).$$

Now we take advantage of this compact machinery to compute the second variation. With the above convention on the notation we write

$$\begin{aligned}\delta^2\mathcal{L} = \; & \delta t_0\,[\Gamma_{t_0}\delta t_0 + \Gamma_{\boldsymbol{x}_0}\cdot\delta\boldsymbol{x}_0 + \Gamma_T\delta T + \Gamma_{\boldsymbol{x}_T}\cdot\delta\boldsymbol{x}_T + \Gamma_{\boldsymbol{\nu}}\cdot\delta\boldsymbol{\nu}] \\ & +\delta\boldsymbol{x}_0\cdot[N_{\boldsymbol{x}_0 t_0}\delta t_0 + N_{\boldsymbol{x}_0\boldsymbol{x}_0}\delta\boldsymbol{x}_0 + N_{\boldsymbol{x}_0 T}\delta T + N_{\boldsymbol{x}_0\boldsymbol{x}_T}\delta\boldsymbol{x}_T + N_{\boldsymbol{x}_0\boldsymbol{\nu}}\delta\boldsymbol{\nu} + \delta\boldsymbol{\lambda}(t_0)] \\ & +\delta T\,[\Omega_{t_0}\delta t_0 + \Omega_{\boldsymbol{x}_0}\cdot\delta\boldsymbol{x}_0 + \Omega_T\delta T + \Omega_{\boldsymbol{x}_T}\cdot\delta\boldsymbol{x}_T + \Omega_{\boldsymbol{\nu}}\cdot\delta\boldsymbol{\nu}] \\ & +\delta\boldsymbol{x}_T\cdot[N_{\boldsymbol{x}_T t_0}\delta t_0 + N_{\boldsymbol{x}_T\boldsymbol{x}_0}\delta\boldsymbol{x}_0 + N_{\boldsymbol{x}_T T}\delta T + N_{\boldsymbol{x}_T\boldsymbol{x}_T}\delta\boldsymbol{x}_T + N_{\boldsymbol{x}_T\boldsymbol{\nu}}\delta\boldsymbol{\nu} - \delta\boldsymbol{\lambda}(T)] \\ & + \int_{t_0}^{T} \delta\boldsymbol{x}\cdot[\mathcal{H}_{\boldsymbol{x}\boldsymbol{x}}\delta\boldsymbol{x} + \mathcal{H}_{\boldsymbol{x}\boldsymbol{u}}\delta\boldsymbol{u} + \boldsymbol{f}_{\boldsymbol{x}}\delta\boldsymbol{\lambda} + \delta\boldsymbol{\lambda}'] \\ & \quad + \delta\boldsymbol{u}\cdot[\mathcal{H}_{\boldsymbol{x}\boldsymbol{u}}\delta\boldsymbol{x} + \mathcal{H}_{\boldsymbol{u}\boldsymbol{u}}\delta\boldsymbol{u} + \boldsymbol{f}_{\boldsymbol{u}}\delta\boldsymbol{\lambda}] + \delta\boldsymbol{\lambda}\cdot[\boldsymbol{f}_{\boldsymbol{x}}\delta\boldsymbol{x} + \boldsymbol{f}_{\boldsymbol{u}}\delta\boldsymbol{u} - \delta\boldsymbol{x}']\,\mathrm{d}t\end{aligned}$$

Here we made broad use of the first order necessary conditions, for example the variation of the extremals of the integral vanish. Here we integrate by parts the quantity $\delta\boldsymbol{x}\cdot\delta\boldsymbol{\lambda}'$, i.e. $\delta\boldsymbol{x}\cdot\delta\boldsymbol{\lambda}' = (\delta\boldsymbol{x}\cdot\delta\boldsymbol{\lambda})' - \delta\boldsymbol{x}'\cdot\delta\boldsymbol{\lambda}$; the variation of $\boldsymbol{x}'$ in the integral follows from the first order conditions, $\delta\boldsymbol{x}' = \boldsymbol{f}_{\boldsymbol{x}}\delta\boldsymbol{x} + \boldsymbol{f}_{\boldsymbol{u}}\delta\boldsymbol{u}$. Further simplifications can be done with the help of the following lemma.

Lemma 3.2. *There are the following simplifications in the coefficients of $\delta^2\mathcal{L}$:*

- $N_{\boldsymbol{x}_0 t_0} + N_{\boldsymbol{x}_0\boldsymbol{x}_0}\boldsymbol{x}'(t_0) + \boldsymbol{\lambda}'(t_0) = \Gamma_{\boldsymbol{x}_0}$.
- $N_{\boldsymbol{x}T} + N_{\boldsymbol{x}_T\boldsymbol{x}_T}\boldsymbol{x}'(T) - \boldsymbol{\lambda}'(T) = \Omega_{\boldsymbol{x}_T}$.

Proof. It is enough to collect the related terms making use of the first variation and the continuity of the solution. □

By rearranging and collecting terms, the second variation results in:

$$\begin{aligned}\delta^2\mathcal{L} = {} & \delta t_0^2[\Gamma_{t_0} + \boldsymbol{x}'(t_0)\cdot\Gamma_{\boldsymbol{x}_0}] + 2\delta t_0\delta\boldsymbol{x}_0\cdot\Gamma_{\boldsymbol{x}_0}\\ & +\delta T^2[\Omega_T + \boldsymbol{x}'(T)\cdot\Omega_{\boldsymbol{x}_T}] + 2\delta T\delta\boldsymbol{x}_T\cdot\Omega_{\boldsymbol{x}_T}\\ & +\delta\boldsymbol{x}_0\cdot N_{\boldsymbol{x}_0\boldsymbol{x}_0}\delta\boldsymbol{x}_0 + \delta\boldsymbol{x}_T\cdot N_{\boldsymbol{x}_T\boldsymbol{x}_T}\delta\boldsymbol{x}_T\\ & +\delta t_0\delta T[\Gamma_T + \boldsymbol{x}'(T)\cdot\Gamma_{\boldsymbol{x}_T} + \Omega_{t_0} + \boldsymbol{x}'(t_0)\cdot\Omega_{\boldsymbol{x}_0}]\\ & +\delta t_0\delta\boldsymbol{x}_T[\Gamma_{\boldsymbol{x}_T} + N_{\boldsymbol{x}_T t_0} + N_{\boldsymbol{x}_T\boldsymbol{x}_0}\boldsymbol{x}'(t_0)] + 2\delta\boldsymbol{x}_0\cdot N_{\boldsymbol{x}_0\boldsymbol{x}_T}\delta\boldsymbol{x}_T\\ & +\delta T\delta\boldsymbol{x}_0\cdot[N_{\boldsymbol{x}_0 T} + N_{\boldsymbol{x}_0\boldsymbol{x}_T}\boldsymbol{x}'(T) + \Omega_{\boldsymbol{x}_0}]\\ & +\delta\boldsymbol{\nu}\cdot[\Gamma_{\boldsymbol{\nu}} + N_{\boldsymbol{\nu}\boldsymbol{x}_0}\delta\boldsymbol{x}_0 + \Omega_{\boldsymbol{\nu}} + N_{\boldsymbol{\nu}\boldsymbol{x}_T}\delta\boldsymbol{x}_T]\\ & +\delta\boldsymbol{\lambda}(T)\cdot\delta\boldsymbol{x}_T - \delta\boldsymbol{\lambda}(t_0)\cdot\delta\boldsymbol{x}_0 + \delta\boldsymbol{\lambda}(t_0)\cdot\delta\boldsymbol{x}_0 - \delta\boldsymbol{\lambda}(T)\cdot\delta\boldsymbol{x}_T\\ & +\int_{t_0}^{T}\delta\boldsymbol{x}\cdot\mathcal{H}_{\boldsymbol{x}\boldsymbol{x}}\delta\boldsymbol{x} + 2\delta\boldsymbol{x}\cdot\mathcal{H}_{\boldsymbol{x}\boldsymbol{u}}\delta\boldsymbol{u} + \delta\boldsymbol{u}\cdot\mathcal{H}_{\boldsymbol{u}\boldsymbol{u}}\delta\boldsymbol{u}\,\mathrm{d}t\end{aligned}$$

The previous result can be summarized in matrix form as

$$\delta^2\mathcal{L} = \begin{pmatrix}\delta\boldsymbol{x}_0\\ \delta t_0\\ \delta\boldsymbol{x}_T\\ \delta T\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\alpha} & \boldsymbol{\beta}^T\\ \boldsymbol{\beta} & \boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\delta\boldsymbol{x}_0\\ \delta t_0\\ \delta\boldsymbol{x}_T\\ \delta T\end{pmatrix} + \int_{t_0}^{T}\begin{pmatrix}\delta\boldsymbol{x}\\ \delta\boldsymbol{u}\end{pmatrix}^T\begin{pmatrix}\mathcal{H}_{\boldsymbol{x}\boldsymbol{x}} & \mathcal{H}_{\boldsymbol{x}\boldsymbol{u}}\\ \mathcal{H}_{\boldsymbol{x}\boldsymbol{u}} & \mathcal{H}_{\boldsymbol{u}\boldsymbol{u}}\end{pmatrix}\begin{pmatrix}\delta\boldsymbol{x}\\ \delta\boldsymbol{u}\end{pmatrix}\mathrm{d}t \tag{3.12}$$

A general assumption can be made to reduce the number of elements and improve the elegance of the notation: we suppose that the cross derivatives of the initial and final point are zero so that the matrix $\boldsymbol{\beta} = \mathbf{0}$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ take the form (by using the previous lemma),

$$\boldsymbol{\alpha} = \begin{pmatrix}N_{\boldsymbol{x}_0\boldsymbol{x}_0} & \Gamma_{\boldsymbol{x}_0}^T\\ \Gamma_{\boldsymbol{x}_0} & \Gamma_{t_0} + \boldsymbol{x}'(t_0)\Gamma_{\boldsymbol{x}_0}\end{pmatrix}, \qquad \boldsymbol{\gamma} = \begin{pmatrix}N_{\boldsymbol{x}_T\boldsymbol{x}_T} & \Omega_{\boldsymbol{x}_t}^T\\ \Omega_{\boldsymbol{x}_t} & \Omega_T + \boldsymbol{x}'(T)\Omega_{\boldsymbol{x}_T}\end{pmatrix}.$$

3.5 SUFFICIENT CONDITIONS

The relations seen in (3.10) are only *necessary* conditions that an extremal function has to meet in order to be optimal. Since the first differential is zero, the total change of the functional is proportional to the second differential and must be nonnegative for all admissible controls $\boldsymbol{u}$ to be a minimum, that is $\delta^2\mathcal{L} \geq 0$. As one can expect, the condition $\delta^2\mathcal{L} \geq 0$ is only necessary. Depending on the problem, it is sometimes easy to directly check if $\delta^2\mathcal{L} > 0$, this is a sufficient condition. A sufficient, but not necessary condition for a minimum, is that all the quadratic forms that appear in $\delta^2\mathcal{L}$ be positive defined. To see that they are not necessary, consider the case of a time fixed optimal control problem, so that the part outside the integral in (3.12) reduces to $N_{\boldsymbol{x}_T\boldsymbol{x}_T}$. It is possible that the quadratic form in the integral be positive defined and hence positive, $N_{\boldsymbol{x}_T\boldsymbol{x}_T}$ be negative and a minimum still exist. Verifying those properties is not always possible or feasible, therefore we introduce some general sufficient conditions that are not problem dependant.
To characterize the presence of a minimizing function, some sufficient conditions have to hold. The first important case is in presence of convexity, it is the generalization of theorem 2.47 and is due to Mangasarian. After the convex case, we will consider the more general environment of the conjugated points theory of Jacobi.

3.5.1 *The Convex Case*

Theorem 3.3 (Mangasarian sufficient conditions)**.** *Consider the minimization problem* (3.8) *for a vectorial control $\boldsymbol{u} \in C[t_0, T]^{n_u}$ with fixed endpoints, with L and $\boldsymbol{f}$ with continuous first partial derivatives with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$, with L and $\boldsymbol{f}$ convex in $\boldsymbol{x}$ and $\boldsymbol{u}$, for all $(t, \boldsymbol{x}, \boldsymbol{u}) \in [t_0, T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that the extremal function given by $\boldsymbol{u}^\star$, $\boldsymbol{x}^\star$, $\boldsymbol{\lambda}^\star$ satisfies the Pontryagin Maximum (Minimum) Principle and $\boldsymbol{\lambda} \geq \mathbf{0}$, for all $t \in [t_0, T]$, then $\boldsymbol{u}^\star$ is a global minimizer for the problem. Moreover if L and $\boldsymbol{f}$ are strictly convex, then $\boldsymbol{u}^\star$ is a strict global minimizer.*

When $\boldsymbol{f}$ is linear in $\boldsymbol{x}$ and $\boldsymbol{u}$, the theorem holds without the sign restriction for the multiplier $\boldsymbol{\lambda}$.

Remark 3.4. *The hypotheses required by this theorem are of rather limited application, because in most problems the terminal cost or the integral cost or the differential equation are not convex. A weaker hypothesis, known as the Arrow condition, requires that the minimized Hamiltonian with respect to $\boldsymbol{u}$ be a convex function in $\boldsymbol{x}$.*

Another problem of this theorem is that the control is assumed to be continuous, but most of the optimal controls are only piecewise continuous, and in some cases, just integrable. Discontinuities in the control give rise to corner points in the state variable, we encountered them in the classic problems of calculus of variations. If $\boldsymbol{u} \in \hat{C}[t_0, T]^{n_u}$ then at a corner point $t_k \in (t_0, T)$ we have

$$\boldsymbol{x}(t_k^-) = \boldsymbol{x}(t_k^+) \qquad \boldsymbol{\lambda}(t_k^-) = \boldsymbol{\lambda}(t_k^+) \qquad \mathcal{H}(t_k^-) = \mathcal{H}(t_k^+).$$

It can be shown that these conditions are equivalent to the Weierstrass-Erdmann corner conditions of theorem 2.63.

3.5.2 *The General Case*

We introduced the concept of accessory minimum problem in the chapter devoted to calculus of variations. Herein we extend it to the more complex case of an optimal control problem, where the variations of several variables are involved. The tractation becomes quickly very involved because of the introduction of many new variables, therefore we prefer to consider only the case of a problem with fixed initial conditions and fixed final time, but with free endpoint, that is $\boldsymbol{B}(\boldsymbol{x}_T) = \mathbf{0}$. This is not a limitation because any other kind of problem can be converted to a fixed time. We begin from the first order differential necessary conditions furnished by the Pontryagin Maximum (Minimum) Principle (3.10). The variation of those conditions with control obtained via $\mathcal{H}_u = 0$, leads to

$$\begin{aligned} \delta\boldsymbol{x}' &= \boldsymbol{f}_{\boldsymbol{x}}\delta\boldsymbol{x} + \boldsymbol{f}_{\boldsymbol{u}}^T\delta\boldsymbol{u} \\ \delta\boldsymbol{\lambda}' &= -\mathcal{H}_{\boldsymbol{xx}}\delta\boldsymbol{x} - \mathcal{H}_{\boldsymbol{xu}}\delta\boldsymbol{u} - \boldsymbol{f}_{\boldsymbol{x}}^T\delta\boldsymbol{\lambda} \\ 0 &= \mathcal{H}_{\boldsymbol{xu}}\delta\boldsymbol{x} + \mathcal{H}_{\boldsymbol{uu}}\delta\boldsymbol{u} + \boldsymbol{f}_{\boldsymbol{u}}^T\delta\boldsymbol{\lambda}. \end{aligned} \tag{3.13}$$

With the assumption that the problem is not singular, that is $\mathcal{H}_{\boldsymbol{uu}} > 0$ (the strengthened Legendre-Clebsch condition holds), the variation of the candidate optimal control can be solved from the third equation above, yielding

$$\delta\boldsymbol{u} = -\mathcal{H}_{\boldsymbol{uu}}^{-1}\left(\mathcal{H}_{\boldsymbol{xu}}\delta\boldsymbol{x} + \boldsymbol{f}_{\boldsymbol{u}}^T\delta\boldsymbol{\lambda}\right).$$

The substitution of this value in the first two equation of (3.13) gives

$$\begin{aligned} \delta\boldsymbol{x}' &= \left(\boldsymbol{f}_{\boldsymbol{x}} - \boldsymbol{f}_{\boldsymbol{u}}\mathcal{H}_{\boldsymbol{uu}}^{-1}\mathcal{H}_{\boldsymbol{xu}}\right)\delta\boldsymbol{x} - \left(\boldsymbol{f}_{\boldsymbol{u}}\mathcal{H}_{\boldsymbol{uu}}^{-1}\boldsymbol{f}_{\boldsymbol{u}}^T\right)\delta\boldsymbol{\lambda} \\ &= \boldsymbol{A}\delta\boldsymbol{x} - \boldsymbol{D}\delta\boldsymbol{\lambda} \\ \delta\boldsymbol{\lambda}' &= -\left(\mathcal{H}_{\boldsymbol{xx}} - \mathcal{H}_{\boldsymbol{xu}}\mathcal{H}_{\boldsymbol{uu}}^{-1}\mathcal{H}_{\boldsymbol{xu}}\right)\delta\boldsymbol{x} - \left(\boldsymbol{f}_{\boldsymbol{x}} - \boldsymbol{f}_{\boldsymbol{u}}\mathcal{H}_{\boldsymbol{uu}}^{-1}\mathcal{H}_{\boldsymbol{xu}}\right)^T\delta\boldsymbol{\lambda} \\ &= -\boldsymbol{C}\delta\boldsymbol{x} - \boldsymbol{A}^T\delta\boldsymbol{\lambda}. \end{aligned} \tag{3.14}$$

The boundary conditions are taken from the differential of the initial and final conditions of (3.11). We keep the more compact notation of the section of the second variation. A procedure to solve this boundary value problem is the sweep method, in which the solution is assumed to have the form of the final conditions, that is

$$\begin{aligned} \delta\boldsymbol{\lambda} &= S(t)\delta\boldsymbol{x} + R(t)\delta\boldsymbol{\nu} \\ \delta\boldsymbol{B} &= P(t)\delta\boldsymbol{x} + Q(t)\delta\boldsymbol{\nu}. \end{aligned} \tag{3.15}$$

$$S(T) = N_{\boldsymbol{x}_T\boldsymbol{x}_T}, \qquad R(T) = \boldsymbol{B}^T_{\boldsymbol{x}_T}, \qquad P(T) = \boldsymbol{B}_{\boldsymbol{x}_T}, \qquad Q(T) = 0. \tag{3.16}$$

To obtain differential equations for S, R, P, Q, so that the system (3.14) is satisfied, we perform differentiation of (3.15), that leads to

$$\begin{aligned} \delta\boldsymbol{\lambda}' &= S'\delta\boldsymbol{x} + S\delta\boldsymbol{x}' + R'\delta\boldsymbol{\nu} \\ 0 &= P'\delta\boldsymbol{x} + P'\delta\boldsymbol{x}' + Q'\delta\boldsymbol{\nu}. \end{aligned}$$

Substitution of (3.14) gives

$$\begin{aligned} \left[S' - S\boldsymbol{D}S + S\boldsymbol{A} + \boldsymbol{A}^T S + vC\right]\delta\boldsymbol{x} + \left[R' + (\boldsymbol{A}^T - S\boldsymbol{D})R\right]\delta\boldsymbol{\nu} &= 0 \\ \left[P' + P(vA - vDS^T)\right]\delta\boldsymbol{x} + \left[Q' - P\boldsymbol{D}R\right]\delta\boldsymbol{\nu} &= 0. \end{aligned}$$

The resulting system of ODE is

$$\begin{aligned} S' &= -\boldsymbol{C} - \boldsymbol{A}^T S - S\boldsymbol{A} + S\boldsymbol{D}S \\ R' &= (S\boldsymbol{D} - \boldsymbol{A}^T)R \\ P' &= (S\boldsymbol{D} - \boldsymbol{A}^T)P \\ Q' &= R^T\boldsymbol{D}R. \end{aligned} \tag{3.17}$$

with boundary conditions given by (3.16). Hence, assuming it is possible to compute S, R, P, Q, we can restate equations (3.15). If also $Q(t_0)^{-1}$ exists, from the second equation of (3.15) it is possible to obtain $\delta\boldsymbol{\nu}$,

$$\delta\boldsymbol{\nu} = -Q(t_0)^{-1}R(t_0)^T\delta\boldsymbol{x}_0,$$

so that the other equation yields

$$\delta\boldsymbol{\lambda}(t_0) = \bar{S}(t_0)\delta\boldsymbol{x}_0, \qquad \bar{S} = S - RQ^{-1}R^T.$$

There are now three cases to be considered.

1. $\bar{S}$ is finite over $[t_0, T)$, thus a given value of $\delta\boldsymbol{x}_0$ induces a finite value for $\delta\boldsymbol{\lambda}(t_0)$. Therefore we can integrate the differential equations (3.14) for the neighbouring optimal paths $\delta\boldsymbol{x}$, $\delta\boldsymbol{\lambda}$ and synthesizing the control $\delta\boldsymbol{u}$. Taking the initial state perturbation $\delta\boldsymbol{x}_0$ to zero, leads to $\delta\boldsymbol{\lambda}(t_0) = 0$ and $\delta\boldsymbol{x} = \delta\boldsymbol{\lambda} = \delta\boldsymbol{u} = \boldsymbol{0}$, that is there are not admissible neighbouring optimal trajectories different from the optimal candidate, which is then a minimizing control.

2. $\bar{S}$ is infinite at t_0, that is, a finite value of $\delta\boldsymbol{x}_0$ induces an infinite value for $\delta\boldsymbol{\lambda}(t_0)$, this implies that there is no neighbouring optimal path. However, it is possible to choose $\delta\boldsymbol{x}_0$ such that it induces a finite $\delta\boldsymbol{\lambda}(t_0)$, moreover the resulting $\delta\boldsymbol{x}$, $\delta\boldsymbol{\lambda}$, $\delta\boldsymbol{u}$ are different from zero. This path is an admissible comparison trajectory and the time instant where $\bar{S}$ becomes infinite is called a *conjugate point*. In this situation the second variation vanishes and hence there is not a sufficient condition for the candidate path to be a minimum.

3. $\bar{S}$ is infinite at $t^\star > t_0$, i.e. the conjugate point is inside the interval (t_0, T) and an optimal comparison path can be established combining $\delta \boldsymbol{u} = \mathbf{0}$ for $t \in [t_0, t^\star]$ and then selecting the candidate optimal control. It can be shown that the associated second variation is negative and this proofs that the candidate solution is not a minimum.

In conclusion, for a candidate optimal solution to be a minimum, it is sufficient that there are not conjugate points in $[t_0, T)$.

Example 3.5. *Consider the optimal control problem of minimizing the functional*

$$\min \int_0^1 (u - x)^2 \, \mathrm{d}t, \qquad x' = u,$$

with $x(0) = 1$.
In this example $N = 0$ *and the Hamiltonian is* $\mathcal{H} = (u-x)^2 + \lambda u$. *Performing the first variation of the augmented functional leads to the first order necessary conditions of the theorem of Pontryagin,*

$$\begin{aligned} \lambda' &= 2(u - x) \\ x' &= u \\ 0 &= 2(u - x) + \lambda \\ \lambda(1) &= 0. \end{aligned}$$

It is a trite computation to solve for the control $u(t) = e^t$ *and the optimal state* $x(t) = e^t$ *with costate* $\lambda(t) = 0$. *To check if this solution is a minimum we first notice that the problem is non singular, in facts* $\mathcal{H}_{uu} = 2 > 0$, *then we set the accessory minimum problem* (3.14)*:*

$$f = u, \qquad f_x = 0, \qquad f_u = 1, \qquad \mathcal{H}_{uu} = 2, \qquad \mathcal{H}_{xu} = -2, \qquad \mathcal{H}_{xx} = 2.$$

With these information we have

$$\begin{aligned} A &= f_x - f_u \mathcal{H}_{uu}^{-1} \mathcal{H}_{xu} &&= 1 \\ D &= f_u \mathcal{H}_{uu}^{-1} f_u &&= \frac{1}{2} \\ C &= \mathcal{H}_{xx} - \mathcal{H}_{xu} \mathcal{H}_{uu}^{-1} \mathcal{H}_{xu} &&= 0. \end{aligned}$$

Because $N_{x_T x_T} = 0$ *and there are not final conditions, the functions* $R = P$ *and* Q *do not exist, thus the only differential equation to solve is the one for* S, *which is (from* (3.17)*),*

$$S' = -C - AS - SA + SDS = -0 - 1 \cdot S - S \cdot 1 + S \cdot \frac{1}{2} \cdot S = \frac{S^2}{2} - 2S.$$

The boundary conditions are given by (3.16) *and reduce in this case to* $S(T = 1) = 0$. *The solution of the differential equation is easily seen to be* $S(t) = 0$ *for* $t \in [0, 1]$. *This shows that* $\bar{S} = S$ *is finite everywhere and with the condition* $\mathcal{H}_{uu} > 0$ *is sufficient to ensure the presence of a minimum.*
In this case it was not possible to simply directly compute the second variation: without loss of generality we could consider the expression for $\delta^2 \mathcal{L}$ *given in* (3.12)*, where* $\delta t_0 = \delta \boldsymbol{x}_0 = \delta T = 0$. *There is only* $\delta \boldsymbol{x}_T$, *moreover,* $\Gamma = -L(0) = -(u(0) - x(0))^2 = 0$ *and* $\Omega = L(T) = (u(T) - x(T))^2 = 0$ *so that the matrices* $\boldsymbol{\alpha} = \boldsymbol{\beta} = \boldsymbol{\gamma} = \mathbf{0}$. *It remains only the integral part, but it is quick to check that the matrix in the integral is only semi positive defined:*

$$\delta^2 \mathcal{L} = \int_{t_0}^{T} (\delta \boldsymbol{x} \ \delta \boldsymbol{u}) \begin{pmatrix} \mathcal{H}_{\boldsymbol{xx}} & \mathcal{H}_{\boldsymbol{xu}} \\ \mathcal{H}_{\boldsymbol{xu}} & \mathcal{H}_{\boldsymbol{uu}} \end{pmatrix} \begin{pmatrix} \delta \boldsymbol{x} \\ \delta \boldsymbol{u} \end{pmatrix} \mathrm{d}t = \int_{t_0}^{T} (\delta \boldsymbol{x} \ \delta \boldsymbol{u}) \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \begin{pmatrix} \delta \boldsymbol{x} \\ \delta \boldsymbol{u} \end{pmatrix} \mathrm{d}t \geq 0,$$

in facts its eigenvalues are 0 and 4. Again we see that the Legendre-Clebsch condition $\mathcal{H}_{\boldsymbol{uu}} > 0$ *is only a necessary condition, but is not enough to ensure the presence of a minimum.*

3.6 INTERPRETATION OF THE MULTIPLIER

We give now an interpretation of the geometrical meaning of the Lagrangian multiplier $\boldsymbol{\lambda}$ associated to a certain equation as the sensitivity of the objective function to a change in that constraint. This subject is connected with the necessary condition derived from the PMP of Pontryagin (3.10). We rewrite the general Optimal Control Problem (3.8) in the formulation of a Lagrange problem. As shown before, this can be always done without loss of generality. For simplicity we restate it with a simple initial condition

$$\begin{aligned} J(\boldsymbol{u}) &= \int_{t_0}^{T} L(t, \boldsymbol{x}(t), \boldsymbol{u}(t))\,\mathrm{d}t \qquad \text{s.t.} \\ \boldsymbol{x}'(t) &= \boldsymbol{f}(t, \boldsymbol{x}, \boldsymbol{u}) \\ \boldsymbol{x}(t_0) &= \boldsymbol{x}_0. \end{aligned} \tag{3.18}$$

We also assume that t_0 and T are fixed extrema and L, $\boldsymbol{f}$ are continuous with continuous first partial derivatives with respect to $\boldsymbol{x}$ and $\boldsymbol{u}$. We suppose also that the optimal control $\boldsymbol{u}$ is unique with the adjoint variable $\boldsymbol{\lambda}$. Next we consider a perturbation by $\boldsymbol{\xi}$ of the initial value $\boldsymbol{x}_0$, that is the new initial state is $\boldsymbol{x}_0 + \boldsymbol{\xi}$. If the optimal control $\boldsymbol{v}(t, \boldsymbol{\xi})$ exists for the perturbed problem, we denote by $\boldsymbol{y}(t, \boldsymbol{\xi})$ the optimal trajectory. This means that

$$\boldsymbol{y}(t, \boldsymbol{\xi})' = f(t, \boldsymbol{y}(t, \boldsymbol{\xi}), \boldsymbol{v}(t, \boldsymbol{\xi}))$$

with $\boldsymbol{y}(t_0, \boldsymbol{\xi}) = \boldsymbol{x}_0 + \boldsymbol{\xi}$. It follows that $\boldsymbol{v}(t, \boldsymbol{0}) = \boldsymbol{u}(t)$ and $\boldsymbol{y}(t, \boldsymbol{0}) = \boldsymbol{x}(t)$. With this notation, consider the objective function for the perturbed problem, we have

$$\begin{aligned} J(\boldsymbol{v}, \boldsymbol{\xi}) &= \int_{t_0}^{T} L(t, \boldsymbol{y}(t), \boldsymbol{v}(t, \boldsymbol{\xi}))\,\mathrm{d}t \\ &= \int_{t_0}^{T} L(t, \boldsymbol{y}(t), \boldsymbol{v}(t, \boldsymbol{\xi})) + \boldsymbol{\lambda}^T f(t, \boldsymbol{y}(t, \boldsymbol{\xi}), \boldsymbol{v}(t, \boldsymbol{\xi}))\,\mathrm{d}t \end{aligned}$$

Taking the partial derivative of the previous expression with respect to $\boldsymbol{\xi}$ yields (omitting the obvious dependencies)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\xi}} J(\boldsymbol{v}, \boldsymbol{\xi}) &= \int_{t_0}^{T} \left(L_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{v}) + \boldsymbol{\lambda}^T f_{\boldsymbol{u}}(\boldsymbol{y}, \boldsymbol{v})\right)^T \boldsymbol{v}_{\boldsymbol{\xi}}(t, \boldsymbol{\xi})\,\mathrm{d}t \\ &\quad + \int_{t_0}^{T} \left(L_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{v}) + \boldsymbol{\lambda}^T f_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{v}) + \boldsymbol{\lambda}'\right)^T \boldsymbol{y}_{\boldsymbol{\xi}}(t, \boldsymbol{\xi})\,\mathrm{d}t \\ &\quad - \boldsymbol{\lambda}(T)^T \boldsymbol{y}_{\boldsymbol{\xi}}(T, \boldsymbol{\xi}) + \boldsymbol{\lambda}(t_0)^T \boldsymbol{y}_{\boldsymbol{\xi}}(t_0, \boldsymbol{\xi}). \end{aligned}$$

The limit for $\boldsymbol{\xi} \to \boldsymbol{0}$ gives

$$\begin{aligned} \lim_{\boldsymbol{\xi} \to \boldsymbol{0}} \frac{\partial}{\partial \boldsymbol{\xi}} J(\boldsymbol{v}, \boldsymbol{\xi}) &= \int_{t_0}^{T} \left(L_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{\lambda}^T f_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{u})\right)^T \boldsymbol{v}_{\boldsymbol{\xi}}(t, \boldsymbol{0})\,\mathrm{d}t \\ &\quad + \int_{t_0}^{T} \left(L_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{\lambda}^T f_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{\lambda}'\right)^T \boldsymbol{y}_{\boldsymbol{\xi}}(t, \boldsymbol{0})\,\mathrm{d}t \\ &\quad - \boldsymbol{\lambda}(T)^T \boldsymbol{y}_{\boldsymbol{\xi}}(T, \boldsymbol{0}) + \boldsymbol{\lambda}(t_0)^T \boldsymbol{y}_{\boldsymbol{\xi}}(t_0, \boldsymbol{0}) \\ &= \boldsymbol{\lambda}(t_0). \end{aligned}$$

In other words, the costate variable $\boldsymbol{\lambda}$ at the initial point can be interpreted as the sensitivity of the cost functional to a change in the initial condition $\boldsymbol{x}_0$. To understand the adjoint variables at a time instant inside the interval $[t_0, T]$, we need the principle of optimality.

Theorem 3.6 (Principle of Optimality)**.** *Let $\boldsymbol{u} \in \hat{C}[t_0, T]^{n_u}$ be an optimal control for problem* (3.18)*, and let $\boldsymbol{x}$ be the associated optimal trajectory. Then for any $t_1 \in [t_0, T]$, the restriction of $\boldsymbol{u}$ to the interval $t_1 \leq t \leq T$ is an optimal control for the same problem restricted to that time interval.*

The theorem shows how a sub arc of an optimal control restricted to a particular interval, is itself optimal for the restricted problem. We use this theorem to apply the argument used to interpret the multiplier at the initial time $\boldsymbol{\lambda}(vx_0)$, to the problem restricted to $t_1 \leq t \leq T$. If we consider the perturbed problem as above, we obtain that

$$\boldsymbol{\lambda}(t_1) = \lim_{\boldsymbol{\xi} \to \mathbf{0}} \frac{\partial}{\partial \boldsymbol{\xi}} J(\boldsymbol{v}, \boldsymbol{\xi}),$$

and because this relation is valid for any $t_1 \in [t_0, T]$, we can write

$$\boldsymbol{\lambda}(t) = \lim_{\boldsymbol{\xi} \to \mathbf{0}} \frac{\partial}{\partial \boldsymbol{\xi}} J(\boldsymbol{v}, \boldsymbol{\xi}).$$

In other words, if the problem is perturbed by a small quantity $\boldsymbol{\xi}$ at time t, and the corresponding optimal control is synthesized, the optimal cost J changes at the rate of $\boldsymbol{\lambda}(t)$. It is said that $\boldsymbol{\lambda}(t)$ is the *marginal valuation* in the OCP of the state variable at time t. We can further observe that the optimal cost remains the original if the perturbation happens at the terminal time T.

3.7 DIFFERENT INITIAL/FINAL CONDITIONS

Historically and in literature, optimal control problems were not always of the form proposed in (3.8), the classic problem that arises from calculus of variations, the brachistochrone, does not have fixed time extremals, but is instead a minimum time problem, that is, the final time T is to be minimized. Similarly there are problems with free initial point, or with infinite time horizon. We show in this section how to deal with such problems, and how any OCP can be restated as an autonomous fixed "time" problem. The independent variable t (the time) will no longer have physical meaning, and we point out that it is better to define the independent variable in a conveniently scaled way. For example, in XOptima, the OCPs are scaled and parametrized with independent variable $\zeta \in [0, 1]$.

3.7.1 *Free Initial Point*

The case of free initial conditions does not appear very often in non academic examples. In this situation we have to consider the two equations with respect to the variations of the initial point which are given by the Pontryagin first order necessary conditions of (3.10).

$$\begin{aligned} \delta \boldsymbol{x}_{t_0} : \qquad & \boldsymbol{\lambda}(t_0) = -\frac{\partial M}{\partial \boldsymbol{x}(t_0)}^T - \frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}(t_0)}^T \boldsymbol{\nu} \\ \delta t_0 : \qquad & \mathcal{H}(t_0) = \frac{\partial M}{\partial t_0} + \boldsymbol{\nu}^T \frac{\partial \boldsymbol{B}}{\partial t_0}. \end{aligned}$$

3.7.2 *Free Final Point*

The case of free final point occurs often in practical problems, we have to consider then the variations for δT and $\delta \boldsymbol{x}_T$ from (3.10).

$$\begin{aligned} \delta \boldsymbol{x}_T : \qquad & \boldsymbol{\lambda}(T) = \frac{\partial M}{\partial \boldsymbol{x}(T)}^T + \frac{\partial \boldsymbol{B}}{\partial \boldsymbol{x}(T)}^T \boldsymbol{\nu} \\ \delta T : \qquad & \mathcal{H}(T) = -\frac{\partial M}{\partial T} - \boldsymbol{\nu}^T \frac{\partial \boldsymbol{B}}{\partial T} \end{aligned}$$

This way to proceed requires the derivation of new second order conditions to ensure the presence of a minimum, therefore it is convenient to reparametrize the problem to fixed time.

3.7.3 *Infinite Horizon*

This situation can be intended as an optimal control problem with fixed final time, and instead of writing $\boldsymbol{x}(\infty)$ we have to consider the limit $\lim_{T\to\infty} \boldsymbol{x}(T)$. In general there should be enough conditions on the Lagrange term inside the integral, on the control and on the trajectory, in order to ensure the existence of the improper integral

$$\int_{t_0}^{\infty} L(t, \boldsymbol{x}, \boldsymbol{u})\,\mathrm{d}t.$$

The criticity is usually overcome by multiplying the integrand by the factor $e^{-\alpha t}$ for $\alpha > 0$.

3.7.4 *Autonomous Problems*

When the functions and the functionals involved in the optimal control problem do not depend explicitly on the time, or more correctly, when the independent variable does not occur explicitly in the problem, we speak of *autonomous problems*. An important property of the Hamiltonian in this case is that is constant along an optimal trajectory. This follows from the first equation of (3.7) and from the related considerations. Moreover, if the final point is free, then the Hamiltonian is equal to zero. Most of the literature considers only autonomous problems, this is because every non autonomous problem can be transformed in an autonomous by a change of variable. This can be done easily by enlarging the original problem by setting the independent variable, e.g. t, equal to a new state equation. This gives $\boldsymbol{x}_{n+1}(t) = t$, and the new differential equation contains $\boldsymbol{x}'_{n+1} = 1$ with initial value $\boldsymbol{x}_{n+1}(t_0) = t_0$.

3.7.5 *Minimum Time*

The problems in which the target to be minimized is the final time, are called *time optimal* problems. The functional to be minimized reduces (if $t_0 = 0$) to

$$\min T = \min \int_{t_0}^{T} 1\,\mathrm{d}t.$$

It it therefore transformed in a problem with free endpoint. Another way to treat this kind of problems is to reformulate them as a fixed time problem adding one more state variable, namely the final time T. To do that we have to change the independent variable t with a new monotone variable, which is called sometimes *dimensionless time* or *pseudotime*, by posing $\zeta = t/T$, in this way the differential becomes $T(\zeta)\,\mathrm{d}\zeta = \mathrm{d}t$ and an eventual differential equation $x'(t) = f(x, u, t)$ is rewritten as $x'(\zeta) = T(\zeta) f(x, u, \zeta)$. The differential equation associated to the new state $T(\zeta)$ is clearly $T'(\zeta) = 0$, because the final time is constant.

We conclude this section with an example that shows the application of these facts.

Example 3.7. *This example is from [Hul03], it is rather artificial but easy enough not to be excessively long in the tractation. It is a problem with free initial and final time, with a constraint on the states. In Hull [Hul03] it is solved applying specialized first order necessary conditions for a free initial time problem, then ad hoc second order conditions are also derived. We prefer instead to solve it after a conversion to a fixed time problem, in order to use the necessary and sufficient conditions for a minimum explained so far. The original problem statement requires to minimize the distance between a parabola and a line as in Figure 3.1. The problem is to find the control* $u(t)$

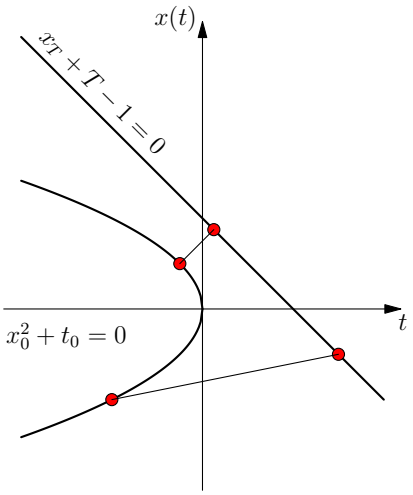


Figure 3.1: Graphical representation of the position of the parabola and the straight line.

that minimizes the target

$$\min J = \int_{t_0}^{T} \sqrt{1+u^2}\,\mathrm{d}t, \qquad x' = u,$$

with prescribed initial condition $x_0^2 + t_0 = 0$ *and final condition* $x_T + T - 1 = 0$.
First we do the change of variable that allows to pass from a functional with variable endpoints to a problem with fixed independent variable $\zeta \in [0,1]$. *This is done by letting* $\zeta = \frac{t-t_0}{T-t_0}$, *which implies* $(T-t_0)\,\mathrm{d}\zeta = \mathrm{d}t$ *and the two new differential states* $t_0(\zeta)$, $T(\zeta)$. *The corresponding differential equations are trivial,* $t_0' = T' = 0$. *The Hamiltonian for this problem is hence*

$$\mathcal{H} = (T-t_0)\sqrt{1+u^2} + \lambda_1(T-t_0)u + \lambda_2 \cdot 0 + \lambda_3 \cdot 0 = (T-t_0)[\sqrt{1+u^2} + \lambda_1 u].$$

We observe that the problem is autonomous so that $\mathcal{H} = \mathrm{const}$. *From the first order necessary conditions we obtain the differential problem*

$$\begin{aligned}
0 &= \frac{\partial \mathcal{H}}{\mathrm{d}u} = (T-t_0)\left(\frac{u}{\sqrt{1+u^2}} + \lambda_1\right)\\
\lambda_1' &= 0\\
\lambda_2' &= \sqrt{1+u^2} + \lambda_1 u = c_1\\
\lambda_3' &= -\lambda_2'.
\end{aligned}$$

From these equations we notice that λ_1 is constant, so that also the optimal control u is constant. Moreover, we have that $\lambda_2(\zeta) = c_1\zeta + c_2$ and $\lambda_3(\zeta) = -c_1\zeta + c_3$. Posing $N = \nu_1(x_0^2 + t_0) + \nu_2(x_T + T - 1)$, the natural boundary conditions for the differential equations are

$$\begin{aligned}
\nu_1 &= -\lambda_2(0) = -c_2 \\
0 &= -\lambda_3(0) = -c_3 \\
0 &= \lambda_1(1) = c_1 + c_2 \\
\nu_2 &= \lambda_3(1) = -c_1 + c_3 \\
0 &= \nu_1 + \int_0^1 \mathcal{H}_{t_0}\,\mathrm{d}\zeta = \nu_1 - \int_0^1 [\sqrt{1+u^2} + \lambda_1 u]\,\mathrm{d}\zeta \\
0 &= \nu_2 + \int_0^1 \mathcal{H}_T\,\mathrm{d}\zeta = \nu_2 + \int_0^1 [\sqrt{1+u^2} + \lambda_1 u]\,\mathrm{d}\zeta.
\end{aligned}$$

We can add also the equation for the state x, that is $x(\zeta) = (T - t_0)u\zeta + x_0$, which yields one more equation for the nonlinear system that we have to solve to obtain the constants for the BVP: for $\zeta = 1$ the final condition is $(T - t_0)u + x_0 = x_T$. We have thus 10 equations (6 boundary conditions, the initial and final conditions, the constraint of the final state, the equation for the control and the expression for c_1) but 11 unknowns, namely t_0, T, u, x_0, x_T, λ_1, c_1, c_2, c_3, ν_1, ν_2. Thus the solution is dependent from an unknown. Solving the NLP gives two solutions, one is not feasible, the second is

$$u = 1, \quad c_1 = \nu_1 = -\nu_2 = -c_2 = -\lambda_1 = \frac{\sqrt{2}}{2}, \quad c_3 = 0,$$
$$t_0 = -x_0^2, \quad T = \frac{1}{2}(-x_0^2 - x_0 + 1), \quad x_T = \frac{1}{2}(x_0^2 + x_0 + 1).$$

To find the value of x_0 we can not use the constant value of the Hamiltonian, because it is linearly dependent with the equations of the NLP. We minimize instead the value of the functional that can now be written in terms of x_0 only. This yields

$$\min J = \int_0^1 \frac{1}{2}(-x_0^2 - x_0 + 1) - (-x_0^2)\,\mathrm{d}\zeta,$$

and the result of this minimization is clearly $x_0 = \frac{1}{2}$, therefore the missing constant are $t_0 = -\frac{1}{4}$, $T = \frac{1}{8}$ and $x_T = \frac{7}{8}$. We check the sufficient conditions. The vectors $\boldsymbol{A}$ and vC are zero because $\frac{u}{\sqrt{1+u^2}} + \lambda_1 = 0$ for the optimal u and λ_1. The vector $\boldsymbol{D} = (T - t_0)u^2\sqrt{1+u^2} = \frac{3\sqrt{2}}{8}$, $\boldsymbol{f_x} = \sqrt{1+u^2}(0,-1,1)^T = \sqrt{2}(0,-1,1)^T$, $\mathcal{H}_{\boldsymbol{x}u} = \boldsymbol{0}$. The final conditions for S, R, T, Q are respectively 0, 1, 1 and 0, hence the differential system (3.17) *reduces to*

$$S' = S\boldsymbol{D}S, \qquad R' = S\boldsymbol{D}R, \qquad Q' = R\boldsymbol{D}R, \qquad S(1) = 0, R(1) = 1, Q(1) = 0.$$

In particular, the first equation has the only solution $S = 0$, which is finite for all $\zeta \in [0,1]$ and there are not conjugate points, thus the candidate optimal control is minimizing.

3.8 CONSTRAINED PROBLEMS

In this section we describe a number of constraint that appear in the formulation of optimal control problems. We present first the simplest cases and show how to reduce more complex constraints to a combination of simpler one. The necessary conditions rely on the first variation, in some cases we also give sufficient conditions of optimality based on the second variation.

3.8.1 *Initial or Final State Constraints*

For the first kind of simple constraints we consider those that restrict the initial or the final state to belong to a certain geometric variety V, that we assume regular enough and defined by

$$V = \{\boldsymbol{x} \in \mathbb{R}^n \mid \alpha_i(\boldsymbol{x}) = 0, i = 1, \dots, q \leq n\},$$

for functions α_i continuously differentiable with their gradient of full rank q for all $\boldsymbol{x} \in V$. We start by considering two regular varieties for the initial and the final state, that is

$$\begin{aligned}(t_0, \boldsymbol{x}(t_0)) \in V_0 &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \alpha_{0i}(\boldsymbol{x}) = 0, i = 1, \dots, q_0 \leq n\},\\ (T, \boldsymbol{x}(T)) \in V_T &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \alpha_{Ti}(\boldsymbol{x}) = 0, i = 1, \dots, q_T \leq n\}.\end{aligned}$$

A simple condition can be derived in the case of the problem of Lagrange (no Mayer term), that is the Hamiltonian at the initial and at the final point should be zero, there are also conditions on the multipliers. In particular if the time instants t_0 and T are free, the following conditions hold:

$$\begin{aligned}\mathcal{H}(t_0, \boldsymbol{x}(t_0), \boldsymbol{u}(t_0), \boldsymbol{\lambda}(t_0)) &= 0,\\ \mathcal{H}(T, \boldsymbol{x}(T), \boldsymbol{u}(T), \boldsymbol{\lambda}(T)) &= 0,\\ \boldsymbol{\lambda}(t_0) &= \sum_{i=1}^{q_0} \vartheta_{0i} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}} \alpha_{0i}(\boldsymbol{x}) \Big|_{\boldsymbol{x}=\boldsymbol{x}(t_0)},\\ \boldsymbol{\lambda}(T) &= \sum_{i=1}^{q_T} \vartheta_{Ti} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}} \alpha_{Ti}(\boldsymbol{x}) \Big|_{\boldsymbol{x}=\boldsymbol{x}(T)}.\end{aligned} \tag{3.19}$$

The numbers ϑ represent the real constant multipliers for the equations that define the varieties V_0 and V_T. The first two equations of (3.19) are called *transversality conditions* are valid only if the time is free, the last two equations of (3.19) are called *orthogonality conditions* because, from a geometric point of view, they can be interpreted as orthogonal vectors: the multiplier $\boldsymbol{\lambda}$ at the initial or final time is orthogonal to the hyperplane which is tangent to V_0 or V_T. When the initial or final time are specified, these conditions are trivially satisfied.
This type of constraint has been included in the standard tractation as the more general function $\boldsymbol{B}(t_0, \boldsymbol{x}_0, T, \boldsymbol{x}_T) = \boldsymbol{0}$ with associated multiplier $\boldsymbol{\nu}$. Another method for the treatment of these inequalities is in the next section.

3.8.2 *Integral Constraints*

One historically important constraint in the calculus of variations is the integral constraint. It arises in the *isoperimetric* problems. There can be two kinds of integral constraints, expressed as an equality or as an inequality. In the first case we have to handle an expression like

$$\int_{t_0}^{T} w(t, x(t), u(t))\,\mathrm{d}t = W_e,$$

where the function w (can be a vector) is continuous with its first derivative with respect to x and t, while W_e is a given constant (or a constant vector). It is possible to include this constraint as a standard ODE of a state extended problem by posing

$$\begin{aligned}z'(t) &= w(t, x(t), u(t))\\ z(t_0) &= 0,\end{aligned} \tag{3.20}$$

then we have to add the final condition $z(T) = W_e$. This can be done in the previous section.
To deal with an integral inequality, of kind

$$\int_{t_0}^{T} w(t, x(t), u(t))\,\mathrm{d}t \leq W_d,$$

where W_d is a known constant, the first step is to convert it to an equality and apply the previous procedure as in equation (3.20); the second step is to add the final condition $z(T) \leq W_d$. The inequality for the initial/final states can be treated as an inequality in a static NLP problem: written in canonical form as $\leq$, it is multiplied by a multiplier that should be *a posteriori* non negative. From a computational point of view it is useful to introduce an ulterior differential state k and set the following conditions. Replace the inequality $z_T \leq W_d$ with the equality $z_T - k = 0$ and add the condition $k'(t) = 0$ with constraint $k(t) \leq W_d$. In this way the variable is forced to be constant and less than W_d. Next we see how to take into account equalities or inequalities constraints over the time interval $[t_0, T]$.

3.8.3 *Equality Constraints*

The class of equality constraint can be divided in three categories, the constraint acts only on the control, only on the state, or on both control and state.

3.8.3.1 *Control Equality Constraints*

Consider the optimal control problem subject to the global equality constraint which is a vector of r components having the form

$$C(t, \boldsymbol{u}) = \mathbf{0}.$$

We assume that the control has m variables, therefore only $m - r$ are independent, assuming that the control is continuous. Each constraint is multiplied by a Lagrange multiplier $\boldsymbol{\mu}$, integrated over the time interval $[t_0, T]$ and added to the target functional. The new objective function becomes then

$$J = \int_{t_0}^{T} L + \boldsymbol{\mu}^T C\,\mathrm{d}t.$$

Taking the differential of J yields

$$\delta J = \int_{t_0}^{T} \left(\frac{\partial L}{\partial \boldsymbol{u}} + \boldsymbol{\mu}^T \frac{\partial C}{\partial \boldsymbol{u}} \right) \delta \boldsymbol{u} + \delta \boldsymbol{\mu}^T C\,\mathrm{d}t. \tag{3.21}$$

It can be readily noted that the term $\delta\boldsymbol{\mu}$ must vanish because its coefficient is $C(t, \boldsymbol{u})$ which is zero by hypothesis. Since the components of the variation of the control are now not all independent, it is not possible to just set the coefficient of $\delta\boldsymbol{u}$ equal to zero, r of the coefficients of $\boldsymbol{\mu}$ are chosen such that the r dependent coefficients of $\delta\boldsymbol{u}$ vanish, then, the remaining components can be considered independent and so their coefficient can be set equal to zero.
The second differential is obtained from equation (3.21) and is

$$\delta^2 J = \int_{t_0}^{T} \delta \boldsymbol{u}^T \left(\frac{\partial^2 L}{\partial \boldsymbol{u}^2} + \boldsymbol{\mu}_i \frac{\partial^2 C_i}{\partial \boldsymbol{u}^2} \right) \delta \boldsymbol{u} + 2\delta \boldsymbol{\mu}^T \frac{\partial C}{\partial \boldsymbol{u}} \delta \boldsymbol{u}\,\mathrm{d}t.$$

The components of $\delta\boldsymbol{u}$ must satisfy $C(t, \boldsymbol{u}) = \mathbf{0}$ and imply the first order condition $\dfrac{\partial C}{\partial \boldsymbol{u}} \delta \boldsymbol{u} = \mathbf{0}$: this fact causes the vanishing of the second term in the previous integral. Once the dependent

variations are obtained in terms of the independent and are simplified, a lower order quadratic form results, it must be non negative in order to obtain a minimizing control. A sufficient condition for a minimum is that the reduced quadratic is positive defined.

3.8.3.2 *State Equality Constraints*

Consider the optimal control problem subject to the global equality constraint which is a vector of s components having the form

$$S(t, \boldsymbol{x}) = \mathbf{0}.$$

We point out that this time the constraint does not depend on the control variable $\boldsymbol{u}$. This constraint reduces the number of independent control and affects the boundary conditions as well. Suppose the dynamical system is $\boldsymbol{x}' = \boldsymbol{f}$, taking the derivative of S yields

$$S' = S_t + S_{\boldsymbol{x}}\boldsymbol{x}' = S_t + S_{\boldsymbol{x}}\boldsymbol{f} = \mathbf{0}.$$

Now it is possible that one among S_T, $S_{\boldsymbol{x}}$ and $\boldsymbol{f}$ contains the control, but this is not mandatory. Hence we must assume that the control does not appear in S', thus the differentiation process is repeated with the substitution $\boldsymbol{x}' = \boldsymbol{f}$ until the q^{th} derivative introduces the control variable. In that case we speak of q^{th} order equality constraint, that can be stated as

$$S(t, \boldsymbol{x}) = \mathbf{0}, \quad S'(t, \boldsymbol{x}) = \mathbf{0}, \quad \ldots, \qquad S^{(q)}(t, \boldsymbol{x}, \boldsymbol{u}) = \mathbf{0}.$$

It is important to underline, that it is not enough to satisfy only the last equation, $S^{(q)}(t, \boldsymbol{x}, \boldsymbol{u}) = \mathbf{0}$, but it is necessary that the whole chain of $q+1$ equalities is satisfied at every point where additional constraints are present. If the interval of application of $S(t, \boldsymbol{x}) = \mathbf{0}$ is a subinterval $[a, b] \subset [t_0, T]$, three cases must be taken into account:

- $a = t_0$ and $b < T$. Then the initial conditions must satisfy all the chain of equalities:

$$S(t_0, \boldsymbol{x}) = \mathbf{0}, \quad S'(t_0, \boldsymbol{x}) = \mathbf{0}, \quad \ldots, \qquad S^{(q)}(t_0, \boldsymbol{x}, \boldsymbol{u}) = \mathbf{0}.$$

- $a > t_0$ and $b = T$. Then the final conditions must satisfy

$$S(T, \boldsymbol{x}) = \mathbf{0}, \quad S'(T, \boldsymbol{x}) = \mathbf{0}, \quad \ldots, \qquad S^{(q)}(T, \boldsymbol{x}, \boldsymbol{u}) = \mathbf{0}.$$

- $a > t_0$ and $b < T$. Then the point constraint need only to be satisfied at $t = a$.

3.8.3.3 *State and control equality constraints*

Consider the optimal control problem subject to the global equality constraint which is a vector of r components having the form

$$C(t, \boldsymbol{x}, \boldsymbol{u}) = \mathbf{0}.$$

Imposing such kind of constraint reduces the number of the independent controls from m to $m - r$. If all the constraint component depend on $\boldsymbol{u}$, then C can be added with a Lagrange multiplier to the Hamiltonian. If some of the components of C do not depend on $\boldsymbol{u}$ then they are state equality constraints, and can be handled as showed in the previous section by augmenting the Hamiltonian.

3.8.4 ***Inequality Constraints***

3.8.4.1 *Control Inequality Constraints*

Consider the fixed time optimal control problem subject to the scalar inequality constraint of the form

$$C(t,u) \leq 0.$$

One way to handle this constraint is to transform the inequality in an equality by means of a *slack variable* $\alpha(t)$ that is defined by

$$\bar{C}(t,u) = -\alpha(t)^2, \qquad C = \bar{C} + \alpha(t)^2 = 0.$$

Assuming the OCP is of Lagrange, the first order necessary conditions become

$$L_u + \mu \bar{C}_u = 0, \qquad 2\mu\alpha = 0.$$

The second equation allows the possibility of mixed arcs, that is arcs where $\mu = 0$ and the bound is not active ($\bar{C} < 0$) and parts where $\alpha = 0$ that are on the boundary $\bar{C} = 0$. In the first case the necessary conditions return the usual ones and the control is determined by $L_u = 0$, then from the equation $\bar{C} = -\alpha^2$ it is possible to recover α. In the second case, when $\alpha = 0$, the control is obtained by solving (the now equality) $\bar{C} = 0$, μ is obtained from $L_u + \mu \bar{C}_u = 0$.
To check if the control is a minimum, we look at the second differential of the functional, which is

$$\delta^2 J = \int_{t_0}^{T} \left(L_{uu} + \mu \bar{C}_{uu} \right) \delta u^2 + 2\mu\delta\alpha^2 \, \mathrm{d}t.$$

The two variations are not independent but are connected by the first order condition

$$\bar{C}_u \delta u + 2\alpha\delta\alpha = 0,$$

thus if $\bar{C}_u \neq 0$, the dependent variation δu can be eliminated from the second differential by posing

$$\delta u = -\frac{2\alpha\delta\alpha}{\bar{C}_u}.$$

Such substitution in the second differential leads to the necessary condition for a minimum,

$$\left(\frac{2\alpha}{\bar{C}_u} \right)^2 \left(L_{uu} + \mu \bar{C}_{uu} \right) + 2\mu \geq 0.$$

Where the bound is inactive ($\mu = 0$), the requirement that the second differential be nonnegative reduces to $L_{uu} \geq 0$; where the bound is active $\alpha = 0$ so that the second differential requires $\mu \geq 0$, which is a classical condition. Sufficient conditions are obtained by requiring the strict inequality of the previous relations.
A practical strategy is to first ignore the bound C and compute the control from $L_u = 0$, if this control satisfies $\bar{C} \leq 0$ we are done, otherwise the optimal control is the one that makes $\bar{C} = 0$. Next, if where the bound is active $\mu > 0$ and where it is inactive $L_{uu} > 0$, the control is a minimum.

3.8.4.2 *State Inequality Constraints*

A classic bound is the *path constraint*, which is an inequality that does not depend explicitly on the control input $\boldsymbol{u}$ but relies on the time (independent variable) and on the state $\boldsymbol{x}$. It is expressed as

$$S(t,\boldsymbol{x}) \leq \mathbf{0}.$$

To simplify the discussion we assume again that the case of one scalar control u and only one inequality, and we consider the case of the active bound, that is when $S(t,x)=0$. In this situation we resume the same strategy used in the case of state equalities and perform the process of differentiation as many times as the control appears in the equation, that is, after q differentiations we end up with

$$S^{(q)}(t,x,u)=0$$

and with a set of point conditions as in the previous section.
Suppose we have the active constraint on a subinterval $[a,b]\subset[t_0,T]$, then we already seen that the point conditions must hold for the entry point $t=a$. Thus we have

$$\begin{aligned}\theta_1 &= S^{(q-1)}(a,x(a))=0,\\ \theta_2 &= S^{(q-2)}(a,x(a))=0,\\ &\cdots\\ \theta_q &= S^{(0)}(a,x(a))=0.\end{aligned}$$

The augmented performance index becomes

$$\begin{aligned}J \;=\; & N(t_0,x_0,T,x_T,\boldsymbol{\nu},a,x(a),\boldsymbol{\xi})+\int_{t_0}^{a}\mathcal{H}(t,x,u,\lambda)-\lambda x'\,\mathrm{d}t\\ &+\int_a^b \mathcal{H}(t,x,u,\lambda,\mu)-\lambda x'\,\mathrm{d}t+\int_b^T \mathcal{H}(t,x,u,\lambda)-\lambda x'\,\mathrm{d}t,\end{aligned}$$

where

$$N=\boldsymbol{\nu}^T\boldsymbol{B}(t_0,x_0,T,x_T)+\boldsymbol{\xi}^T\boldsymbol{\theta},$$

$$\mathcal{H}=\begin{cases} L+\lambda f & t_0\le t\le a\\ L+\lambda f+\mu S^{(q)} & a\le t\le b\\ L+\lambda f & b\le t\le T\end{cases}$$

Performing the first variation the following conditions must hold:

$$\begin{gathered}x'=f,\qquad \lambda'=-\mathcal{H}_x^T,\qquad \mathcal{H}_u^T=0,\\ \mathcal{H}(a^+)=\mathcal{H}(a^-)+N(a),\qquad \lambda(a^+)=\lambda(a^-)-N^T(a),\\ \mathcal{H}(b^+)=\mathcal{H}(b^-),\qquad \lambda(b^+)=\lambda(b^-),\end{gathered}$$

together with the standard boundary conditions at the extremals $t=t_0$ and $t=T$. It is important to notice that a jump in the Hamiltonian and in the multiplier can occur only when the bound becomes active, but they are continuous when the bound becomes inactive. On the bound, the Hamiltonian must be a minimum with respect to the control obtained by $S^{(q)}=0$; off the bound, the Hamiltonian is minimized by the control that satisfies the Pontryagin Maximum Principle.

3.8.4.3 *State and Control Inequality Constraints*

The general form for a state and control inequality constraint is denoted by

$$C(t,\boldsymbol{x},\boldsymbol{u})\le\mathbf{0},$$

where C has r components, while the control has $m>r$. Again we consider only one inequality and one control, for multiple control and constraints the solution is similar but the procedure

becomes more involved and combinatoric. There are simple forms of C where a bounded control can be made unbounded by enlarging the control space with a slack variable. A typical example is the bound $u \geq k$ (for a constant k), that can be replaced by $u = k + \alpha^2$, similarly, the case of $k_1 \leq u \leq k_2$ can be simplified in

$$u = k_1 + (k_2 - k_1)\sin^2 \alpha.$$

There are many trick like this or like the penalty functions to obtain an unbounded control. The general approach of using the slack variables is similar to the one described in the previous sections, and requires to introduce

$$\bar{C}(t, x, u, \alpha) = C(t, x, u) + \alpha^2$$

The Hamiltonian is augmented with $\mu(\bar{C} + \alpha^2)$, and since this problem involves now only equalities, the conditions derived in the previous section are applicable:

$$\mathcal{H}_\alpha = 0 \implies 2\mu\alpha = 0.$$

Hence, either $\mu = 0$ for an off-boundary arc, or $\alpha = 0$ when the bound is active. In the first case the control is obtained by $\mathcal{H}_u = 0$, $\alpha(t)$ is solved from $\alpha^2 = -C$. In the second case the control is obtained from $C = 0$ and μ from $\mathcal{H}_u = 0$. The point where two subarcs join is called a corner point, since there are no conditions imposed on the location $t = c$ of the corner point, the conditions are just

$$\mathcal{H}(c^+) = \mathcal{H}(c^-), \qquad \lambda(c^+) = \lambda(c^-).$$

The second order conditions are given by

$$\mathcal{H}_{uu}\delta u^2 + 2\mu\delta\alpha^2 \geq 0, \qquad C_u\delta u + 2\alpha\delta\alpha = 0.$$

Off the boundary (i.e. $\mu = 0$) we require that $\mathcal{H}_{uu} \geq 0$. On the boundary $\alpha = 0$ implies $\delta u = 0$ which requires $C_u \neq 0$ so that $\mu \geq 0$.
The Legendre-Clebsch condition has the standard form $\mathcal{H}_{uu} \geq 0$ off the boundary, it requires that the Lagrange multiplier associated with the equality constraint be nonnegative on the boundary. This is often useful in determining the the subarcs contained in the minimal sequence.
For problems that are affine in the control where $\mathcal{H}_u \neq 0$ and the control is bounded as $k_1 \leq u \leq k_2$, the optimal control is bang-bang or singular. Since $\mathcal{H}_{uu} = 0$, the Weierstrass condition $\mathcal{H}(t, x, u, \lambda) - \mathcal{H}(t, x, u^\star, \lambda) > 0$, must be used to determine the control sequence.

3.8.5 *Jump Conditions*

An optimal control problem can often require different phases and the connection between each couple of arcs can be imposed by forcing the passage of trajectory for some “checkpoints”. They are formulated as *isolated* equality constraints of the form

$$w_i(t_i, x(t_i)) = 0, \qquad t_0 < t_1 < \ldots < t_i < \ldots < t_k < T,$$

for k time instants. The functions w_i are supposed continuous with their first derivative. The Hamiltonian and the multipliers λ may be discontinuous at the times t_i and if x, u, λ are optimal for the first order necessary conditions, then at each time instant t_i,

$$\lambda(t_i^-) = \lambda(t_i^+) + \frac{\partial w_i(t_i, x)}{\partial x}\mu_i, \qquad i = 1, \ldots, k.$$

Finally, because the time instants t_i themselves can be unknown, there is the additional relation

$$\mathcal{H}(t_i^-) = \mathcal{H}(t_i^+) - \frac{\partial w_i(t_i, x)}{\partial t}\mu_i, \qquad i = 1, \dots, k.$$

It is noticed that this way to handle the jump conditions is equivalent to the solution of a sequence of optimal control subproblems joined together by the isolated equality constraints, thus we can treat also isolated inequalities with the same technique of the previous section on the initial/final state constraints.

We end the chapter with an interesting example that shows typical behaviours of controls and constraints.

Example 3.8. *The performance index to be minimized is*

$$J = \int_0^T \frac{1}{2}u^2 \,\mathrm{d}t, \qquad x' = u, \quad x(0) = 0,$$

with the final state and time that belong to the manifold

$$V_T = \{(x, t) \in \mathbb{R}^2 \mid (x-2)^2 + (t-2)^2 - 1 = 0\}.$$

First we check that the final manifold is regular, we need to ensure that the gradient of $\alpha(t, x) = (x-2)^2 + (t-2)^2 - 1$ *has full rank. We have that*

$$\nabla\alpha(t, x) = (2t - 4,\ 2x - 4) \neq \mathbf{0} \iff (t, x)^T \neq (2, 2)^T$$

and it is easy to check that the vector $(2, 2)^T$ *does not belong to the manifold, hence* V_T *is regular. The Hamiltonian is* $\mathcal{H} = \frac{1}{2}u^2 + \lambda u$*. The first order necessary conditions give* $\lambda(t) = \lambda(0)$ *constant,* $u(t) = -\lambda(t) = -\lambda(0)$ *and* $x(t) = -\lambda(0)t + x_0$*. The initial conditions yields directly* $x_0 = 0$*. Next we need the second and the fourth of conditions* (3.19)*, together with the equation of the final manifold* $\alpha(t, x) = 0$*,*

$$\begin{aligned} 0 &= -\frac{1}{2}\lambda(0)^2 + 2\vartheta(T-2) \\ 0 &= \lambda(0) - 2\vartheta(x(T)-2) \\ 1 &= (x(T)-2)^2 + (T-2)^2. \end{aligned}$$

These are three equations in four unknowns: we can add the equation of the state x *to obtain the fourth relation necessary to solve the nonlinear system,* $x(T) = -\lambda(0)T$*. The system simplifies in*

$$\begin{aligned} 0 &= 3\lambda(0)^4 + 8\lambda(0)^3 + 16\lambda(0)^2 + 32\lambda(0) + 12 \\ T &= \frac{4 - 2\lambda(0)}{\lambda(0)^2 + 2} \\ \vartheta &= -\frac{\lambda(0)}{2(\lambda(0)T + 2)}, \end{aligned}$$

where the two real roots of the polynomial in $\lambda(0)$ *are approximated and lead to the following solutions:*

$$\begin{aligned} &\lambda(0) \approx -2.11, \qquad T \approx 1.27, \qquad x(T) \approx 2.69, \qquad J \approx 2.83 \\ &\lambda(0) \approx -0.46, \qquad T \approx 2.23, \qquad x(T) \approx 1.03, \qquad J \approx 0.24. \end{aligned}$$

Looking at Figure 3.2, we see that there is a third solution, marked as P_3*, which has the*

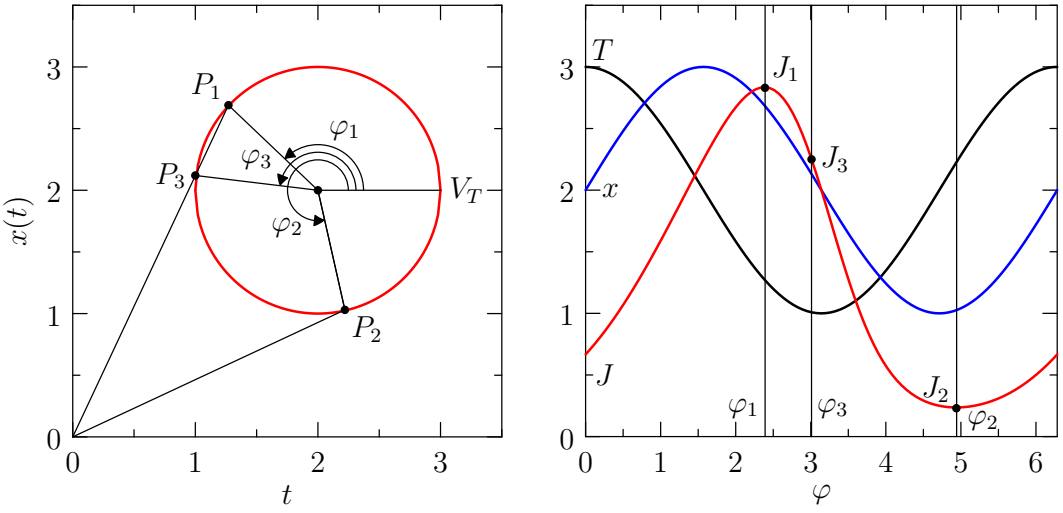


Figure 3.2: On the left the trajectories reaching the final manifold V_T, on the right the plot of the values of the final state $x(T)$, the final time T and the value of the performance index J as functions of the angle φ.

corresponding objective function $J_3 \approx 2.25$. We can notice that $J_1 > J_3$, but this fact is only a geometric question: Figure 3.2, on the right, shows that the two computed solutions (P_1 and P_2) are respectively the maximum and the minimum of the functional, while P_3 is not even a stationary point.

4

PROBLEMS AFFINE IN THE CONTROL

4.1 THE HAMILTONIAN AFFINE IN THE CONTROL

In this section we consider problems for which the Hamiltonian is an affine function of the control u. Despite the fact that, in general, dealing with linear problems should be an easier task, this is not the case. The complication arises from the fact that the necessary condition that the Hamiltonian is minimized with respect to the control along an optimal trajectory does not provide a well defined expression for the synthesis of the optimal control. Another difficulty connected with this family of problems is the presence of discontinuities in the control, which are difficult to locate, moreover the switching points can be infinite and can accumulate at a time instant. Because general results and theorems regarding existence and uniqueness are rather limited in this situation, additional relationships have to be introduced manipulating the other necessary conditions. We can say that problems with the Hamiltonian linear with respect to the control u constitute an almost independent research area in optimal control. Most of the theoretical results are obtained with a geometric approach and involve fiber bundles and symmetry groups instead of variational techniques. To keep notation as simple as possible, we consider here problems with a single control variable and fixed end point. The formulation of these problems is the following.

Definition 4.1. *Minimize the cost functional J subject to the dynamic $\boldsymbol{x}'$ with the control u constrained in $|u| \leq 1$ and the initial state $\boldsymbol{x}(0) = \boldsymbol{x}_0$, where*

$$\begin{aligned} J(u) &= \int_0^T f_0(\boldsymbol{x}(t)) + b_0(\boldsymbol{x}(t))u(t)\,\mathrm{d}t \\ \boldsymbol{x}'(t) &= \boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{b}(\boldsymbol{x}(t))u(t) \qquad \text{i.e.} \\ x_i'(t) &= f_i(\boldsymbol{x}(t)) + b_i(\boldsymbol{x}(t))u(t), \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{4.1}$$

$f_0(\boldsymbol{x}(t))$ and $b_0(\boldsymbol{x}(t))$ are scalar functions of the state $\boldsymbol{x}(t)$, $u(t)$ is the scalar control.

To further simplify the computations, we consider the scalar equation of the functional as an extra state variable posing

$$x_0'(t) = f_0(\boldsymbol{x}(t)) + b_0(\boldsymbol{x}(t))u(t), \qquad x_0(0) = 0,$$

(we have already seen that this is not a loss of generality), but the key point here is that the control appears only linearly in both dynamics and cost functional. We can write the Hamiltonian for this system, namely

$$\mathcal{H} = \sum_{i=0}^{n} f_i(\boldsymbol{x}(t))\lambda_i(t) + u(t)\sum_{i=0}^{n} b_i(\boldsymbol{x}(t))\lambda_i(t).$$

It is convenient to rename the two addends as

$$\begin{aligned} H_0(\boldsymbol{x}, \boldsymbol{\lambda}, t) &= \sum_{i=0}^{n} f_i(\boldsymbol{x}(t))\lambda_i(t), \\ H_1(\boldsymbol{x}, \boldsymbol{\lambda}, t) &= \sum_{i=0}^{n} b_i(\boldsymbol{x}(t))\lambda_i(t), \end{aligned}$$

so that the Hamiltonian becomes

$$\mathcal{H} = H_0(\boldsymbol{x}, \lambda, t) + u(t)H_1(\boldsymbol{x}, \lambda, t). \tag{4.2}$$

The adjoint variables of the costate are given by the usual relation

$$\lambda_i'(t) = -\frac{\partial \mathcal{H}}{\partial x_i(t)} = -\sum_{j=0}^{n} \lambda_j(t)\frac{\partial f_j(\boldsymbol{x}(t))}{\partial x_i(t)} - u(t)\sum_{j=0}^{n} \lambda_j(t)\frac{\partial b_j(\boldsymbol{x}(t))}{\partial x_i(t)}.$$

We state now the necessary conditions for problem (4.1).

Theorem 4.2. *If $u^\star(t)$ is an optimal control and if $\boldsymbol{x}^\star(t)$ is the corresponding optimal trajectory, then there are a costate $\lambda_i^\star(t)$ such that*

- *for $i = 1, \dots, n$*

$$\begin{aligned} x_i'^\star(t) &= f_i(\boldsymbol{x}^\star(t)) + b_i(\boldsymbol{x}^\star(t))u^\star(t), \\ \lambda_i'^\star(t) &= -\sum_{j=0}^{n} \lambda_j^\star(t)\frac{\partial f_j(\boldsymbol{x}^\star(t))}{\partial x_i^\star(t)} - u^\star(t)\sum_{j=0}^{n} \lambda_j^\star(t)\frac{\partial b_j(\boldsymbol{x}^\star(t))}{\partial x_i^\star(t)}; \end{aligned}$$

- *for all $t \in [0, T]$ and all $u(t)$ s.t. $|u(t)| \le 1$, holds*

$$H_0(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, t) + u^\star(t)H_1(\boldsymbol{x}^\star, \lambda^\star, t) \le H_0(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, t) + u(t)H_1(\boldsymbol{x}^\star, \lambda^\star, t);$$

- *If T is free, then for all $t \in [0, T]$*

$$\mathcal{H}^\star := H_0(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, t) + u^\star(t)H_1(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, t) = 0, \tag{4.3}$$

  while, if T is fixed, then

$$\mathcal{H}^\star := H_0(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, t) + u^\star(t)H_1(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, t) = c,$$

  for a real constant c.

It follows easily from equation (4.2) with the PMP, that the optimal control is given by

$$u^\star(t) = -\text{sign}\{H_1(\boldsymbol{x}^\star, \lambda^\star, t)\} = -\text{sign}\left\{\sum_{i=0}^{n} b_i(\boldsymbol{x}^\star(t))\lambda_i^\star(t)\right\}, \tag{4.4}$$

as long as the *function switching* $H_1(\boldsymbol{x}^\star, \lambda^\star, t)$ is not zero. If H_1 becomes zero, then the sign function is not defined. If H_1 is zero only in a point we speak of *bang-bang* controls, if H_1 is identically zero for an entire interval $t \in (t_1, t_2]$, the controls are called *singular*. In this case, the associated trajectory $\boldsymbol{x}(t)$ is called a *singular arc*.

4.2 BANG-BANG CONTROLS

From equation (4.4), it is clear the if the control is unbounded, the optimal control would be $u = \pm\infty$, hence it is more interesting to consider the case of a constrained control in a compact (convex) set. For a scalar control it is common in literature to consider the control $u \in [-1, 1]$. It is easy to adapt a different interval $[a, b]$ to fit into $[-1, 1]$ by a linear homotopy, if $x \in [a, b]$ and $u \in [-1, 1]$ then

$$u = \frac{2x}{b-a} - \frac{a+b}{b-a}, \qquad x = \frac{b-a}{2}u + \frac{a+b}{2}.$$

Again from equation (4.4) it is clear that the optimal control will varies on the boundary of the possible feasible control domain, the jumps from a border to the other are governed by the switching function $H_1(\boldsymbol{x}^\star, \lambda^\star, t)$. This behaviour reflects the linearity of the problem in the control and resembles what happens in a problem of linear programming, where the optimal point is on the border of the simplex. The name given to these kind of controls is *bang-bang*, because there are not transitions.

Example 4.3. *A classic example of bang-bang controls is given by the double integrator plant, where*

$$\min T = \int_0^T 1 \,\mathrm{d}t, \qquad x' = y,\ y' = u, \qquad |u| \le 1, \qquad x(T) = y(T) = 0,$$

for fixed initial states x_0 and y_0. The Hamiltonian for this system is given by

$$\mathcal{H} = 1 + \lambda_1 y + \lambda_2 u,$$

which gives the costate equations

$$\lambda_1' = 0, \qquad \lambda_2' = -\lambda_1 \implies \lambda_1(t) = \text{const}, \quad \lambda_2(t) = \lambda_1(T-t) + c.$$

The transversality conditions for the problem (recalling that $\lambda_2(T) = c$) and the equation for the control (4.4)*, give*

$$\lambda_2(T)u(T) = -1, \qquad u(t) = -\text{sign}\lambda_2(t).$$

This implies that either $c = 1$, $u(T) = -1$ or $c = -1$, $u(T) = 1$. Moreover, the linearity of the switching function $\lambda_2(t)$ ensures the there can be at most *one jump. Going back from the final instant making use of the boundary conditions $x(T) = y(T) = 0$ yields two possible trajectories:*

$$u = -1, \quad x = -\frac{(T-t)^2}{2}, \quad y = T - t, \quad x(y) = -\frac{y^2}{2}$$
$$u = +1, \quad x = +\frac{(t-T)^2}{2}, \quad y = t - T, \quad x(y) = +\frac{y^2}{2}.$$

The switching curve (in the state space) is made up of two pieces of parabola given by $x(y) = -\text{sign}(y)\frac{y^2}{2}$*. Above the curve the control takes the value* $u = -1$*, below the curve it takes the value* $u = +1$*. The situation is focused better in Figure 4.1. If the initial state is located on the switching*

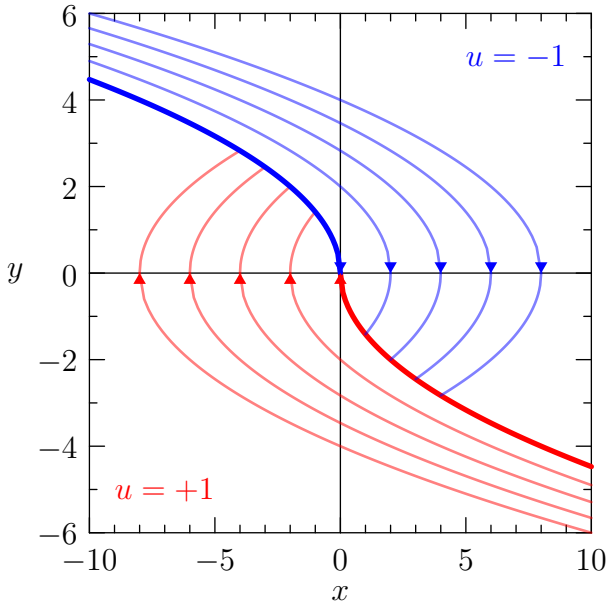


Figure 4.1: State space (x, y) for the double integrator plant, the thick curve is the switching function.

curve, then no jumps in the control are necessary to reach the target state in the origin. One switch is needed if the initial state is somewhere else in the plane. From an initial state, the optimal trajectory is a parabola of the form $x = \pm\frac{y^2}{2} + c$ *towards the curve* $x(y) = -\text{sign}(y)\frac{y^2}{2}$*, when this manifold is hit, the optimal trajectory follows the switching curve until the final state. The contour of constant final time* T *are given, above and below the switching curve, respectively by*

$$(y+T)^2 = 4\left(-x + \frac{T^2}{2}\right), \qquad (y-T)^2 = 4\left(x + \frac{T^2}{2}\right). \tag{4.5}$$

It is noticed that this contour has slope discontinuity when hitting the switching curve, and the vector of the multipliers

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}\right)$$

is orthogonal to the contours of constant T*, but the normal is not defined on the switching curve. To fix the ideas look at Figure 4.2, suppose* $x(0) = 2$ *and* $y(0) = 1$*: the initial point is above the switching surface, hence the optimal trajectory must follow a parabola with control* $u = -1$ *of kind* $x(y) = -\frac{y^2}{2} + k$ *towards the switching curve (*$x(y) = \frac{y^2}{2}$*). Imposing the initial condition in the previous equation, we get* $k = \frac{5}{2}$ *and the intersection with the switching curve occurs at time* $t^\star = \sqrt{5/2} + 1$ *at coordinates* $x = \frac{5}{4}$ *and* $y = \sqrt{5/2}$*. Then, on the switching curve, it takes* $T - t^\star = \sqrt{5/2}$ *to reach the origin, hence the total time required is* $T = 2\sqrt{5/2} + 1 \approx 4.16$*.*
We take another point on the contour of equal final time (4.5)*, for* $T = 2\sqrt{5/2}+1$*; a possible choice is* $x(0) = \frac{3}{4} - \frac{\sqrt{10}}{2}$ *and* $y(0) = 2$*. Making use of the relations:*

$$k = \frac{1}{2}y(0)^2 + x(0), \qquad t^\star = \sqrt{k} + y(0),$$

$$x(t^\star) = -\frac{1}{2}t^{\star 2} + y(0)t^\star + x(0), \qquad y(t^\star) = -t^\star + y0,$$

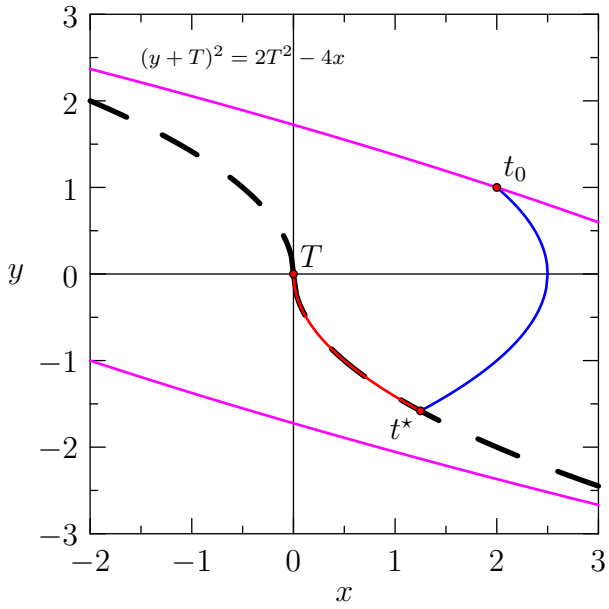


Figure 4.2: State space (x, y) for the double integrator plant, the thick dashed curve is the switching function, the initial point is $(2, 1)$ and in magenta the level set of isochrony.

the corresponding optimal trajectory has $u = -1$ *and* $x(y) = -\frac{y^2}{2} + k$ *until the switching curve, and it is satisfied by* $k = \frac{11}{4} - \frac{\sqrt{10}}{2} \approx 1.16$. *The intersection of the parabola with the switching curve occurs at* $\hat{t}^\star = \sqrt{k} + y(0) = \sqrt{\frac{11}{4} - \frac{\sqrt{10}}{2}} + 2 \approx 3.08$, $x(\hat{t}^\star) = \frac{11}{8} - \frac{\sqrt{10}}{4} \approx 0.58$ *and* $y(\hat{t}^\star) = \frac{1}{2} - \frac{\sqrt{10}}{2} \approx -1.08$. *The second part of the trajectory lies on the switching curve, where* $T - \hat{t}^\star = \sqrt{k}$ *thus* $T = \hat{t}^\star + \sqrt{k} = 2\sqrt{k} + y(0) = \sqrt{10} + 1$.
Finally, solving the multipliers, we have that with the last initial conditions,

$$\lambda_1 = -\frac{1}{\hat{t}^\star} = -\frac{2}{3 + \sqrt{10}} \approx -0.32, \qquad \lambda_2(t) = \lambda_1 t - 1.$$

Therefore the vector of the multipliers at time $t = 0$ *is given by* $\boldsymbol{\lambda} = (-\frac{2}{3+\sqrt{10}}, -1)^T$ *while solving the implicit function of the contour (equation* (4.5)*),* $(y + T)^2 = 4\left(-x + \frac{T^2}{2}\right)$*, gives locally* $y(x) = -T + \sqrt{2T^2 - 4x}$ *and has the derivative equal to* $\frac{-2}{\sqrt{2T^2-4x}}$. *Thus the tangent vector to the contour at the initial point is* $(1, -2/(3 + \sqrt{10}))^T$ *and it is easy to see that it is orthogonal to* $\boldsymbol{\lambda}$ *(Figure 4.3.*

4.3 SINGULAR CONTROLS

If problem (4.1) results singular, controls, costate and trajectory have the following property: there is at least one half-open interval $(t_1, t_2] \subset [0, T]$ such that

$$H_1 = \sum_{i=0}^{n} \lambda_i^\star(t) b_i(\boldsymbol{x}^\star(t)) = 0, \qquad t \in (t_1, t_2]. \tag{4.6}$$

As before, the existence of an extremal singular control, does not imply that the *optimal* control is singular: we need additional information as uniqueness to conclude its optimality. The function (4.6) is often called *switching function*.

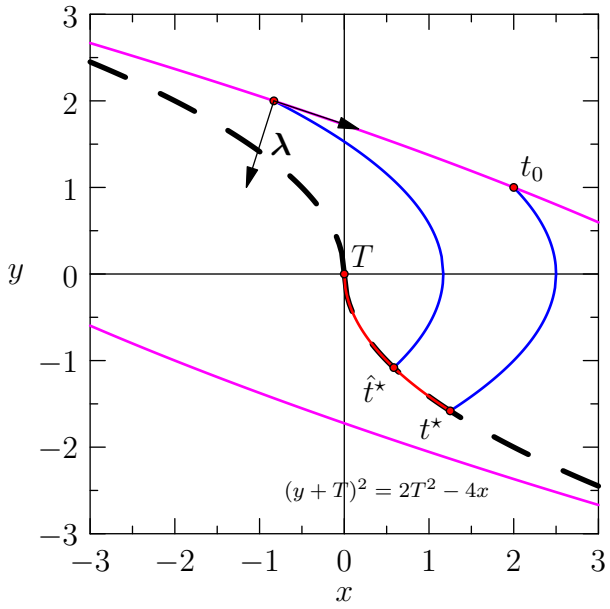


Figure 4.3: State space (x, y) for the double integrator plant, the thick dashed curve is the switching function, the arrows represent the tangent at the levelset of the isochrone manifold and the vector of the multipliers.

4.3.1 *Necessary Condition for Singular Controls*

Suppose we are in the free terminal time case, so that (4.3) holds, we test if it is possible to have a singular control as follows. First we need to assume that the switching function is zero in an unknown interval $(t_1, t_2]$, then, because of the free end time hypothesis, for $t \in (t_1, t_2]$,

$$\mathcal{H} = \sum_{i=0}^{n} f_i(\boldsymbol{x}(t))\lambda_i(t) + u(t)\sum_{i=0}^{n} b_i(\boldsymbol{x}(t))\lambda_i(t) = 0 \implies \sum_{i=0}^{n} f_i(\boldsymbol{x}(t))\lambda_i(t) = 0.$$

Moreover, for each $k \in \mathbb{N}$, we have that

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\sum_{i=0}^{n} \lambda_i(t) b_i(\boldsymbol{x}(t)) = 0, \qquad k = 1, 2, 3, \ldots, \quad t \in (t_1, t_2], \tag{4.7}$$

and similarly,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\sum_{i=0}^{n} \lambda_i(t) f_i(\boldsymbol{x}(t)) = 0, \qquad k = 1, 2, 3, \ldots, \quad t \in (t_1, t_2], \tag{4.8}$$

However, the canonical equations (we omit explicit time dependence) are

$$\begin{aligned} x_i' &= f_i + u b_i \\ \lambda_i' &= -\sum_{j=0}^{n} \lambda_j(t)\frac{\partial f_j}{\partial x_i} - u\sum_{j=0}^{n} \lambda_j \frac{\partial b_j}{\partial x_i}, \qquad i, j = 1, \ldots, n;\ k \in \mathbb{N}. \end{aligned}$$

Let $k = 1$ in (4.7), then by the chain rule we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=0}^{n} \lambda_i b_i = \sum_{i=0}^{n}\left(\lambda_i' b_i + \lambda_i \sum_{j=0}^{n} \frac{\partial b_j}{\partial x_i} x_j'\right) = 0.$$

Substituting the canonical equations in the previous one, yields

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\left(f_j\lambda_i\frac{\partial b_i}{\partial x_j}-b_i\lambda_j\frac{\partial f_j}{\partial x_i}\right)+u\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\lambda_i b_j\frac{\partial b_i}{\partial x_j}-\lambda_j b_i\frac{\partial b_j}{\partial x_i}\right)=0.$$

We notice that the coefficient of u is zero, and we conclude that

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j}-f_j\frac{\partial b_i}{\partial x_j}\right)=0. \tag{4.9}$$

Applying the same argument to equation (4.8), gives

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=0}^{n}\lambda_i f_i(\boldsymbol{x})=\sum_{i=0}^{n}\left(\lambda_i' f_i+\lambda_i\sum_{j=0}^{n}\frac{\partial f_i}{\partial x_j}x_j'\right)\\ &=\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\lambda_i f_i\frac{\partial f_i}{\partial x_j}-\lambda_j f_i\frac{\partial f_j}{\partial x_i}\right)+u\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\lambda_i b_j\frac{\partial f_i}{\partial x_j}-\lambda_j f_i\frac{\partial b_j}{\partial x_i}\right)\\ &=u\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j}-f_j\frac{\partial b_i}{\partial x_j}\right)=0.\end{aligned}$$

This last line implies that either $u=0$ or

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j}-f_j\frac{\partial b_i}{\partial x_j}\right)=0.$$

Noticing that the previous line is equal to equation (4.9), we can not conclude that $u=0$, and this is true for all $k\in\mathbb{N}$.
In conclusion, a necessary but not sufficient test for an extremal singular control is that in the interval $(t_1,t_2]$ the following relations are satisfied:

$$\begin{cases}\displaystyle\sum_{i=0}^{n}b_i\lambda_i=0,\\ \displaystyle\sum_{i=0}^{n}f_i\lambda_i=0,\\ \displaystyle\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j}-f_j\frac{\partial b_i}{\partial x_j}\right)=0.\end{cases}$$

We show now how to compute the explicit expression for the singular control, we return back to the equation of the switching function (4.7) and let $k=2,3,\ldots$. It is shown in [AF66][pag.487] that those derivatives for $k=2,3,\ldots$ require extensive manipulations but have all the following same structure:

$$\sum_{i=0}^{n}\lambda_i\psi_{ki}(\boldsymbol{x})+u\sum_{i=0}^{n}\lambda_i\phi_{ki}(\boldsymbol{x})=0.$$

If $\sum\lambda_i\phi_{ki}(\boldsymbol{x})=0$ we can perform successive derivatives of (4.7), in general there will be a finite value of $k=m$ for which $\sum\lambda_i\phi_{mi}(\boldsymbol{x})\neq 0$ and then the singular control u can be solved:

$$u=-\frac{\displaystyle\sum_{i=0}^{n}\lambda_i\psi_{mi}(\boldsymbol{x})}{\displaystyle\sum_{i=0}^{n}\lambda_i\phi_{mi}(\boldsymbol{x})}. \tag{4.10}$$

Up to now we have dealt with free final time, if the final time is fixed the we have to take in account the presence of the constant c in the Hamiltonian, but equations (4.7) and (4.8) still hold and hence also the conclusions.
In practical problems, we must check if $\sum \lambda_i \phi_{ki}(\boldsymbol{x}) = 0$ for all $k = 1, \dots, m-1$ and $\sum \lambda_i \phi_{mi}(\boldsymbol{x}) \neq 0$. If any of the relations are violated then this represent a violation of the necessary conditions and so singular controls can not occur, if these conditions are all met, then there may be a singular extremal control. The minimum value of differentiations needed for which we can express explicitly the control as in equation (4.10) for problems of kind (4.1) is always an even natural number. Therefore, let this number be $k = 2q$, then q is called the *order* of the singular arc. In applications coming from mechanical systems with linear controls, appear only singular controls of order 1 or 2. Particular cases lead to order 3, but there are also examples of higher order. A complete description of the cases $q = 1, 2$ is available in the works of [ZB94], there is not a complete knowledge of the third order arcs, but many facts are understood. Very little is known for higher orders, this is due to the complex geometrical construction behind their study. Here we will focus only on order 1 and 2, and we develop in the next section ad hoc theory.

4.3.1.1 *The Poisson Bracket*

We notice that the above notation for determining the order of singular arcs is somehow cumbersome. There is a geometric tool that can help in those computations, it is the *Poisson bracket*.

Definition 4.4 (Poisson bracket)**.** *The Poisson bracket of two functions $A(\boldsymbol{x}, \boldsymbol{\lambda})$ and $B(\boldsymbol{x}, \boldsymbol{\lambda})$, defined on the extended space of state and multipliers (for $\boldsymbol{x}, \boldsymbol{\lambda} \in \mathbb{R}^n$), is*

$$\{B, A\} := \sum_{i=1}^{n} \left(\frac{\partial B}{\partial \lambda_i} \frac{\partial A}{\partial x_i} - \frac{\partial B}{\partial x_i} \frac{\partial A}{\partial \lambda_i} \right).$$

The motivation comes from the following observation, if $A(\boldsymbol{x}, \boldsymbol{\lambda})$ is an arbitrary differentiable function, along the optimal trajectory we have

$$\frac{\mathrm{d}}{\mathrm{d}t} A(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \left(\frac{\partial A}{\partial \lambda_i} \frac{\mathrm{d}\lambda_i}{\mathrm{d}t} + \frac{\partial A}{\partial x_i} \frac{\mathrm{d}x_i}{\mathrm{d}t} \right) = \sum_{i=1}^{n} \left(-\frac{\partial A}{\partial \lambda_i} \frac{\partial \mathcal{H}}{\partial x_i} + \frac{\partial A}{\partial x_i} \frac{\partial \mathcal{H}}{\partial \lambda_i} \right) = \{\mathcal{H}, A\}.$$

There are two useful properties of the Poisson bracket, it is anticommutative and satisfies the Jacobi identity, that is, respectively:

$$\{B, A\} = -\{A, B\}, \qquad \{A, \{B, C\}\} + \{B, \{C, A\}\} + \{C, \{A, B\}\} = 0.$$

Thus turning back to the derivative of the switching function H_1 on a singular arc, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} H_1 = \{H_0 + uH_1, H_1\} = \{H_0, H_1\} + u\{H_1, H_1\} = \{H_0, H_1\} = 0.$$

Since the function $\frac{\mathrm{d}}{\mathrm{d}t} H_1$ does not depend on u, it gives no information on the optimal control. Differentiating it once more we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} H_1 = \frac{\mathrm{d}}{\mathrm{d}t} \{H_0, H_1\} = \{H_0, \{H_0, H_1\}\} + u\{H_1, \{H_0, H_1\}\} = 0.$$

Now, if $\{H_1, \{H_0, H_1\}\} \neq 0$ we are in the case of singular controls of order 1 and thus

$$u = -\frac{\{H_0, \{H_0, H_1\}\}}{\{H_1, \{H_0, H_1\}\}},$$

whereas if

$$\{H_1, \{H_0, H_1\}\} = 0, \tag{4.11}$$

we need to continue the differentiation process:

$$\frac{\mathrm{d}^3}{\mathrm{d}t^3} H_1 = \{H_0, \{H_0, \{H_0, H_1\}\}\} + u \{H_1, \{H_0, \{H_0, H_1\}\}\} = 0.$$

The last term can be simplified making use of the Jacobi identity,

$$\begin{aligned} \{H_1, \{H_0, \{H_0, H_1\}\}\} &= -\{H_0, \{\{H_0, H_1\}, H_1\}\} - \{\{H_0, H_1\}, \{H_1, H_0\}\} \\ &= -\{H_0, \{\{H_0, H_1\}, H_1\}\} \\ &= \{H_0, \{H_1, \{H_0, H_1\}\}\}. \end{aligned}$$

Because equation (4.11) holds,

$$\{H_0, \{H_1, \{H_0, H_1\}\}\} = \{H_0, 0\} = 0,$$

thus the third derivative does not contain the variable u and the process of differentiation continues as above. In case of order 2 we end with

$$\frac{\mathrm{d}^4}{\mathrm{d}t^4} H_1 = \{H_0, \{H_0, \{\{H_0, H_1\}, H_1\}\}\} + u \{H_1, \{H_0, \{\{H_0, H_1\}, H_1\}\}\} = 0$$

but in case of order 2, the coefficient of u is different from zero and the singular control is given by

$$u = -\frac{\{H_0, \{H_0, \{\{H_0, H_1\}, H_1\}\}\}}{\{H_1, \{H_0, \{\{H_0, H_1\}, H_1\}\}\}}.$$

The necessary conditions for an optimal control, is the extended Legendre-Clebsch condition, also known as the Kelley-Contensou condition,

$$(-1)^q \frac{\partial}{\partial u} \frac{\mathrm{d}^{2q}}{\mathrm{d}t^{2q}} H_1(\boldsymbol{x}, \boldsymbol{\lambda}) \leq 0.$$

When $q = 2$ this can be rewritten as

$$\{H_1, \{H_0, \{H_0, \{H_0, H_1\}\}\}\} \leq 0.$$

The constraint on the control, $|u| \leq 1$ implies the inequality

$$|\{H_0, \{H_0, \{\{H_0, H_1\}, H_1\}\}\}| \leq -\{H_1, \{H_0, \{\{H_0, H_1\}, H_1\}\}\}.$$

The conclusion of these equations is that if the singular arc is of order 2 (equation (4.11) must hold), it must lie in the manifold V defined in the extended space $(\boldsymbol{\lambda}, \boldsymbol{x})$ by the equations

$$H_1 = 0, \quad \{H_0, H_1\} = 0, \quad \{H_0, \{H_0, H_1\}\} = 0, \quad \{H_0, \{H_0, \{H_0, H_1\}\}\} = 0.$$

This fact is fundamental in the analysis of the optimal trajectory in the proximity of V because it offers a diffeomorphism that allows a change of coordinates in canonical form. The first four coordinates become

$$z_1 = H_1, \quad z_2 = \{H_0, H_1\}, \quad z_3 = \{H_0, \{H_0, H_1\}\}, \quad z_4 = \{H_0, \{H_0, \{H_0, H_1\}\}\}$$

and remembering the relation $\frac{\mathrm{d}}{\mathrm{d}t}A(\boldsymbol{x},\boldsymbol{\lambda}) = \{\mathcal{H}, A\}$ together with (4.11) and the Kelley condition,

$$z_1' = z_2, \quad z_2' = z_3, \quad z_3' = z_4, \quad z_4' = \frac{\mathrm{d}^4}{\mathrm{d}t^4}H_1 = a(\boldsymbol{\lambda},\boldsymbol{x}) + ub(\boldsymbol{\lambda},\boldsymbol{x}),$$

where, clearly,

$$a(\boldsymbol{\lambda},\boldsymbol{x}) = \{H_0, \{H_0, \{\{H_0, H_1\}, H_1\}\}\}, \qquad b(\boldsymbol{\lambda},\boldsymbol{x}) = \{H_1, \{H_0, \{\{H_0, H_1\}, H_1\}\}\}.$$

The other coordinates are $w_1, \ldots, w_{2n-4}$ and are chosen in a way that all the coordinates $(\boldsymbol{\lambda},\boldsymbol{x})$ can be expressed in terms of the $(\boldsymbol{z},\boldsymbol{w})$ and the determinant of the Jacobian of the change of coordinates is non zero. The transformed Hamiltonian system in the new coordinates $(\boldsymbol{z},\boldsymbol{w})$ is called *semi*canonical system and is studied to prove the optimality of singular arcs, but this topic is out of the scope. It is called semicanonical and not just canonical, because the choice for the variables $\boldsymbol{w}$ is not unique.

4.4 CHATTERING

It is very difficult to give general theorems for global properties of an arbitrary optimal control problem, because there are many different situations that occur. For example, the uniqueness of the solution requires Lipschitz continuity, but we can find trivial functions that are not Lipschitz. Another problem is that practical problems involve often the initial but also the final point, but most of the theorems on differential equations deal with the initial value problem (also known as the Cauchy problem) instead of the Boundary Value Problem (BVP). The motivation is that for the problem of Cauchy there are results of existence and uniqueness, while it is difficult to guarantee even the existence for a BVP: it is easy to find examples with infinite, only one or no solutions. A famous counterexample for the uniqueness of the solution is given by the scalar problem $x' = \sqrt{x}$ with initial condition $x(0) = 0$: the functions $x(t) = 0$ and $x(t) = \frac{t^2}{4}$ both satisfy the ODE. The motivation is that $f = \sqrt{x}$ is not Lipschitz at zero, where the Lipschitz constant grows to infinity. We have also to lose the concept of continuity of f because in many applications, the right hand side is discontinuous because, e.g. due to bang-bang controls. The concept of switching is very important and we dedicate a chapter on the case of OCP that are affine in the control. An important example is the chattering phenomenon, in literature is also known as Zeno behaviour. The concept is introduced considering a ball bouncing on the floor. If h is the heigh with respect to the floor and v is its velocity, assuming for simplicity unitary gravitational constant, the differential system that describes the motion of the ball is given by

$$h' = v, \qquad v' = -1.$$

But when the ball hits the floor, there is an instantaneous change in the velocity, $v(t) = -kv(t^-)$, with the coefficient k that models the elasticity of the impact and is therefore a number $k \in (0,1)$. The switching function for the jumps in the differential equation are given by the time instants where the ball has $h = 0$. Because the dynamic of the system is not different in the two configurations, we can integrate the differential equation and obtain

$$\begin{aligned} v(t) &= -(t - t_0) + v(t_0), \\ h(t) &= -\frac{(t-t_0)^2}{2} + v(t_0)(t - t_0) + h(t_0), \end{aligned}$$

where the initial conditions were $t_0 = 0$, $h(0) = 0$ and $v(0) = 1$. Before the first switch we have

$$\begin{aligned} v(t) &= -t + 1, \\ h(t) &= -\frac{t^2}{2} + t, \end{aligned}$$

and the switch instant is $t = 2$ because it gives $h(2) = 0$ with $v(2^-) = -1$ and $v(2) = k$. For $t > 2$ we have

$$\begin{aligned} v(t) &= -t + 2 + k, \\ h(t) &= -\frac{(t-2)^2}{2} + k(t-2). \end{aligned}$$

We can then compute the next switch for $t = 2 + 2k$ with velocity $v(2 + 2k) = k^2$, and by applying the same argument, the switching times for the progression $2, 2 + 2k, 2 + 2k + 2k^2, \ldots$ while the corresponding velocities form the sequence $k^2, k^3, k^4, \ldots$. The switching times constitute a geometric progression with a *finite* limit, an accumulation point, given by

$$\sum_{i=0}^{\infty} 2k^i = \frac{2}{1-k}.$$

This means that the ball does infinitely many bounces before this time, and this fact is called Zeno Behaviour. In reality this fact does not happens because of other physical constraints, and the ball will stop at rest after only a finite number of bounces. In optimal control problems this phenomenon has different names according to different authors. Nowadays all seem to agree with the term *chattering*, but some texts refers to the chatter when dealing to the *sliding mode* [Mar75]. Russian literature distinguishes between chattering and sliding mode, and refers to the first with *Fuller phenomenon* [Bor00].

4.4.1 *Sliding Mode*

The chattering related to sliding mode, means the relaxation of the control or the convexization of the maneuverability domain. A typical example is the situation when it is theoretically necessary to alternate at an infinite rate between two values of the control as in the astrodynamic problem of the Lawden's spiral. These solutions are characterized by an indetermination of the PMP and the control can not be determined directly. Consider the system

$$\min y(3), \qquad x' = u, \quad y' = x^2, \qquad x(0) = 1,\ x(3) = 1,\ y(0) = 0, \tag{4.12}$$

$u = \pm 1$, where the independent variable is the time $t \in [0, 3]$ and we want to minimize the final state $y(3)$. If the control were free to move in $u \in [-1, 1]$ we could have had the sequence of control $-1, 0, 1$ as showed in Figure 4.4. The optimal value of the target is $y(3) = \frac{2}{3}$. The solution

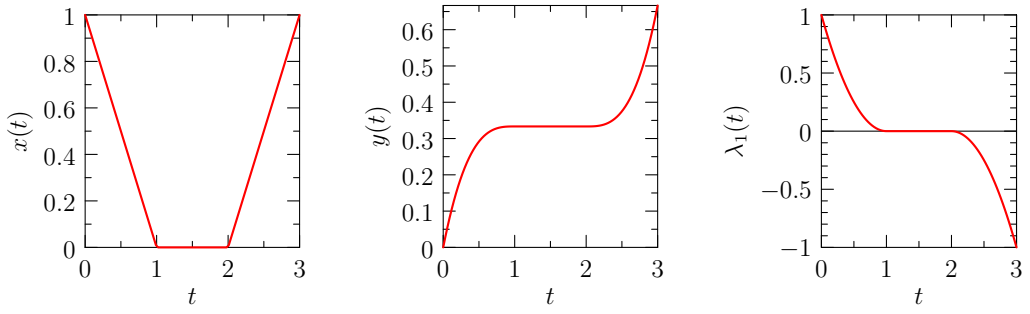


Figure 4.4: States and multiplier for the OCP (4.12) with control $u \in [0, 1]$.

for $t \in [1, 2]$ is singular, in fact the multiplier (see Figure 4.4) is vanishing on the interval $[1, 2]$. This solution is no more valid if the control is constrained to be $u = \pm 1$, as in the original problem, however, the value of $y(3)$ can be approximated as close as desired by solution similar to that of Figure 4.5. In proximity of the singular arc the control chatters between ± 1 and theoretically it

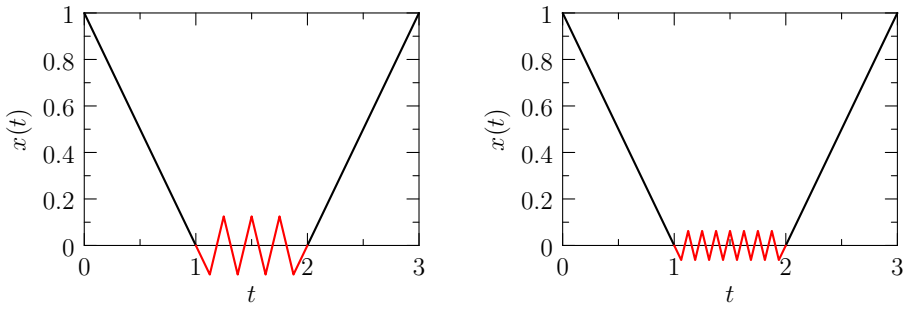


Figure 4.5: Suboptimal example states for the OCP (4.12) with control $u = \pm 1$.

must chatter infinite times. This fact is called sliding mode. It is understood that the difference of the two problems is only academic but in some problems, like the Marec Spiral [Mar73], it is physically impossible to relax the control and convexify the set ± 1 to $[-1, 1]$. So the question if it is always possible to relax the control set is not obvious as it could appear. For example, the OCP $x' = u$, $y' = x^2 + 2\sqrt{|y|}$ with the control $u = \pm 1$ can not be relaxed. Marchal gives three sufficient conditions for the feasibility of the relaxation. Briefly, they are

1. In any bounded subset of the state space $(t, \boldsymbol{x})$, the derivative of the state is bounded.
2. The control is a measurable function in a "good" domain (the description of the domain have a weak topology that the domains of interest in general posses).
3. The differential system is canonical in the sense of Pontryagin.

The problem with the above stated OCP is that it does not satisfy the third request, because of the presence of the term $\sqrt{|y|}$.
We explain now the motivation for the name "sliding", consider a system of two differential equations $\boldsymbol{x}' = \boldsymbol{f}$ where

$$\boldsymbol{x}' = \begin{cases} f_1(\boldsymbol{x}) & \text{if} \quad g(\boldsymbol{x}) > 0 \\ f_2(\boldsymbol{x}) & \text{if} \quad g(\boldsymbol{x}) < 0 \end{cases}$$

for f_i and g sufficiently differentiable. The switching manifold is assumed to be defined by $S = \{\boldsymbol{x} \mid g(\boldsymbol{x}) = 0\}$ and f_1, f_2 are one on each side of S; g is called the switching function. Supposing that there are no state jumps in the trajectory, when $\boldsymbol{x}$ reaches S, it crosses over to the other side. This is possible when both vectors f_1 and f_2 have the same direction with respect to S, and a solution is naturally obtained. A different situation is when f_1 and f_2 point both toward S, because in this case a solution can not be obtained. Let the point $\boldsymbol{x}_0$ be on the manifold (i.e. $g(\boldsymbol{x}_0) = 0$), then consider the two scalar products

$$\begin{aligned} v_1 &= \quad \langle \nabla g(\boldsymbol{x}_0), f_1(\boldsymbol{x}_0) \rangle \\ v_2 &= -\langle \nabla g(\boldsymbol{x}_0), f_2(\boldsymbol{x}_0) \rangle. \end{aligned}$$

The vector $\nabla g(\boldsymbol{x}_0)$ points toward the domain of f_1, then $v_1 < 0$ implies that the vector field for f_1 pushes against S; if $v_1 > 0$ the flow is pulling. The same argument holds for v_2 and hence the four cases:

- $v_1 > 0$ and $v_2 < 0$: the trajectory crosses S from $g < 0$ to $g > 0$.
- $v_1 < 0$ and $v_2 > 0$: the trajectory crosses S from $g > 0$ to $g < 0$.
- $v_1 > 0$ and $v_2 > 0$: the trajectory pulls on S from both sides and the solution is not unique, but this case in general does not occur.

- $v_1 < 0$ and $v_2 < 0$: the trajectory pushes on S from both sides and the solution is constrained on S as for example in Figure 4.5.

While the first two cases cause no difficulties and the third does not occur, the fourth is problematic and was studied for several years. Once on the manifold, the trajectory can not continue neither following $\boldsymbol{x}' = f_1$ nor $\boldsymbol{x}' = f_2$, however, we have seen that suboptimal solution of this kind exist (Figure 4.5). It became clear that the optimal solution was on the manifold itself, hence arises the necessity of solving a differential algebraic problem because of the presence of the term $g(\boldsymbol{x}) = 0$. The correct way of proceed is the one proposed by Filippov [HNW93], he suggested to search the vector field in S in the convex hull of f_1 and f_2, which is given by

$$f(\boldsymbol{x}, \lambda) = (1 - \lambda) f_1(\boldsymbol{x}) + \lambda f_2(\boldsymbol{x}). \tag{4.13}$$

The value of λ must be chosen such that the trajectory remains on S. This implies that we need to solve

$$\begin{aligned} \boldsymbol{x}' &= f(\boldsymbol{x}, \lambda) \\ 0 &= g(\boldsymbol{x}), \end{aligned}$$

which is a Differential Algebraic Equation (DAE) of index 2 [HNW96]. This can be done with DAE techniques by differentiating the constraint

$$0 = \nabla g(\boldsymbol{x}) \boldsymbol{x}' = \nabla g(\boldsymbol{x}) f(\boldsymbol{x}, \lambda), \tag{4.14}$$

and if it is possible to solve for λ in a form like $\lambda = G(\boldsymbol{x})$, then we have transformed the DAE in an ODE, namely

$$\boldsymbol{x}' = f(\boldsymbol{x}, G(\boldsymbol{x})).$$

The above equation can now be solved with standard ODE techniques. From the equation (4.13), the relation (4.14) can be written as

$$(1 - \lambda) v_1(\boldsymbol{x}) - \lambda v_2(\boldsymbol{x}) = 0 \implies \lambda = \frac{v_1(\boldsymbol{x})}{v_1(\boldsymbol{x}) + v_2(\boldsymbol{x})}.$$

With this solution, the only possible trajectory is constrained to slide along the manifold S. In the case of problem (4.12), the manifold g is given by $\lambda_1 = 0$ and the two f_i both point towards S. Therefore in the first region $\lambda_1 > 0$ and we follow the solution until it hits the manifold $\lambda_1 = 0$, here we have $v_1 < 0$ and $v_2 < 0$ so the solution remains inside the manifold until one of the two values of v_1, v_2 changes sign, this happens for $t = 2$ and after that point the solution follows the rule for $\lambda_1 < 0$. In applications the sliding mode is approximated by the *hysteresis switching* which is showed in Figure 4.5 and gives suboptimal solutions [Lib03].

4.4.2 *Fuller Phenomenon*

The proper description of the term chattering refers to the arcs that appear before and after singular arcs, as soon as the generalized Legendre-Clebsch condition (called also Kelly-Contensou test for singular extremals) requires four or more derivatives with respect to the independent variable. The optimal control has a countable infinite number of switching with an accumulation point at the beginning or at the end of a singular arc. This phenomenon was discovered in an innocent looking example prosed in the Sixties by Fuller, and called *Fuller phenomenon* after his name. It is not just an academic fact, because it appears in many optimal control problems that posses a linear control. In practical implementations it is not possible to realize those infinite sequence of switching and it is interesting to observe the difference between a suboptimal piecewise continuous

control in comparison with the chattering arc. The results are surprising and non intuitively, the common experience in optimization is found to be misleading. It is the hidden symmetry of Fuller's phenomenon that allows to obtain the optimal control synthesis, and this leads quickly away from calculus of variations directly to questions of higher geometry and group theory. The goal of this section is to make aware and recognise the presence of this phenomenon.
We consider herein the classic Fuller problem, and we present a closely related example, the Fuller-Marchal Problem, in the chapter of benchmarks. Consider a particle subjected to a force $u(t)$ and moving from an initial condition $(x(0) = x_0, x'(0) = x'_0)$ on a straight line without friction. The target functional is the minimization of the deviation of the particle from the origin $x = 0$, hence we can formulate the Fuller problem as

$$\min J = \int_0^T x^2(t)\,\mathrm{d}t, \qquad x' = y, \quad y' = u, \qquad |u| \leq 1, \tag{4.15}$$

with initial conditions fixed to $x(0) = x_0 = 2$, $y(0) = y_0 = -2$. It is clear that the particle must reach the origin $x = y = 0$ to minimize the functional, and this position is a singular manifold V. Experience would suggest to reach V as soon as possible, but it turns out that this is not true in the presence of a singular arc. We compare the Fuller problem with the following time optimal OCP, and show that they are not equivalent:

$$\begin{aligned} &\min T, \qquad x' = y, \quad y' = u, \qquad |u| \leq 1, \\ &x(0) = x_0,\ y(0) = y_0,\ x(T) = y(T) = 0. \end{aligned}$$

Without loss of generality, we can assume $x_0 > 0$ and $y_0 = 0$, then the time optimal problem has only one switch on $[0, T]$, in facts the force $u = -1$ pushes the particle toward the origin on the first time interval $[0, t^\star]$ and then switches to $u = 1$ to arrest the particle at the origin. This is the classical bang-bang solution.
The situation is completely different in the case of Fuller problem: the optimal strategy has infinitely many switches and consists of an infinite number of cycles (in the state space (x, y)) around the origin. The initial arc begins with $u = -1$ followed by a switch to $u = 1$ and so on. The instant t_1 of the first switch, causes the particle to reach a point $x(t_1) = -qx_0$ with $0 < q < 1$. This is repeated at successively points q^2x_0, $-q^3x_0$, q^4x_0 and so on. These points form an alternating convergent geometric sequence, the duration of these cycles forms a geometric sequence too, thus the entire process takes finite time to reach the origin, but not in the shortest time.
Chattering is closely related to the existence of singular extremals and to their order, it was proved by Robbins, Kelley, Kopp and Moyer that the order of the singular arc that permits to solve the control is even. They also proved that the concatenation of a piecewise smooth nonsingular arc with a singular arc of even order is *non optimal*. Usually, the singular manifold V is the most profitable point (or in general subset) of the state space and if singular solutions have second or higher order, chattering is necessary to enter V (and to escape from V). Consider the possibility of escaping from the manifold V, the optimal exit strategy is again chattering with the switches that accelerate to infinity in the reverse time. A graphical representation of this behaviour is shown in Figure 4.6. This escape can be imagined as a series of very fast pushes and pulls like vibrations which gives extremely small alternating deviations from V at the very beginning.
We turn back to Fuller problem (4.15) and show the symmetry properties of the homogeneity group that allows to find the optimal solution. The Hamiltonian for this problem is

$$\mathcal{H} = x^2 + \lambda_1 y + \lambda_2 u,$$

and the corresponding adjoint system is given by

$$\lambda'_1 = -2x, \qquad \lambda'_2 = -\lambda_1, \qquad u = -\mathrm{sign}(\lambda_2).$$

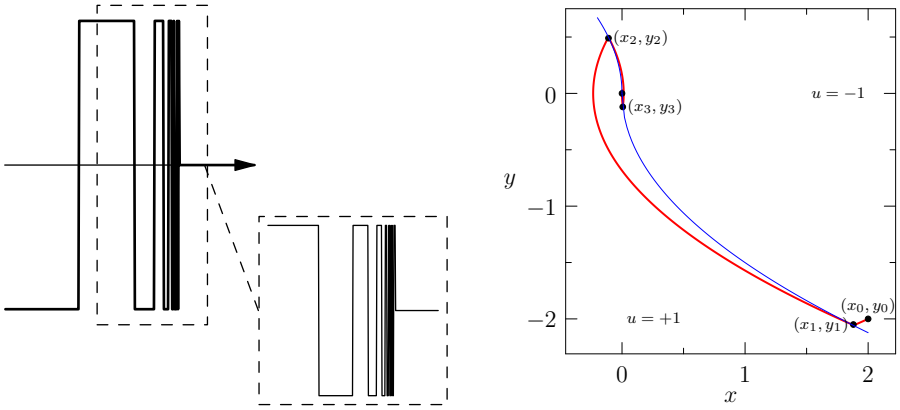


Figure 4.6: Optimal chattering control entrance to the singular arc and the corresponding trajectory in the state space (x, y). The red line is the optimal trajectory, the blue line represents the switching manifold.

To establish the singular control and find its order, it necessary to differentiate the switching function $H_1 = \lambda_2$. We do not need to apply the change of variable described in the previous section of the Poisson bracket, because the system is already semicanonical. In facts, we have

$$\begin{aligned} H_1' = \lambda_2' = -\lambda_1, \quad H_1'' = -\lambda_1' = 2x, \\ H_1''' = 2x' = 2y, \quad H_1'''' = 2y' = 2u = 0, \end{aligned} \tag{4.16}$$

hence the singular control is $u = 0$ and the order of the arc is 2 and we expect the chattering phenomenon. To show this, suppose that it is possible to reach the singular arc with a finite combination of bang-bang controls. Integrating the BVP (4.16) over the last bang-bang arc, that is between the time instants $[t_{n-1}, t_n]$, we have ($\Delta = t_n - t_{n-1}$):

$$y = u\Delta + a_2 = 0, \qquad x = \frac{u}{2}\Delta^2 + a_2\Delta + a_1 = 0,$$

$$\lambda_1 = -\frac{u}{3}\Delta^3 - a_2\Delta^2 - 2a_1\Delta + q_1 = 0, \quad \lambda_2 = \frac{u}{12}\Delta^4 + \frac{a_2}{3}\Delta^3 + a_1\Delta^2 - q_1\Delta + q_2 = 0$$

for suitable real constants a_1, a_2, q_1, q_2 and control $u = \pm 1$. As a first observation, we notice that $q_2 = 0$ because by hypothesis t_{n-1} is a switching instant and $\lambda_2(t_{n-1}) = 0$, then we can solve the resulting algebraic system and discover that the only solution is zero. We must then conclude that the number of bang-bang arcs to reach the singular manifold $\{x = y = \lambda_1 = \lambda_2 = 0\}$ is infinite. The next step is to prove that the total time required to reach the singular arc is finite, thus the switching times must accumulate. This is done in two parts, first we show that the Fuller problem possesses a symmetry group of dilatations, then we prove that the duration of the bang-bang arcs forms a converging geometrical progression.
Let $(\lambda_2(t), \lambda_1(t), x(t), y(t))$ be an admissible solution of the problem with control $u(t)$ and target J, we can see that for any $k > 0$ the tuple

$$(\lambda_{2,k}(t), \lambda_{1,k}(t), x_k(t), y_k(t)) = \left(k^4\lambda_2(kt), k^3\lambda_1(kt), k^2 x\,(kt)\,, ky\,(kt)\right)$$

is also an admissible solution of the system with control $u_k(t) = u(kt)$ and target $k^5 J$. Formally, this is a group of transformations (dilatations) $g_k : \mathbb{R}^4 \to \mathbb{R}^4$, $g_k(\lambda_2, \lambda_1, x, y) = (k^4\lambda_2, k^3\lambda_1, k^2 x, ky)$ for any $(\lambda_2, \lambda_1, x, y) \in \mathbb{R}^4$ for any $k > 0$. We can consider the parametric curve $k \to (x_i k^2, k y_i)$ and conclude that the switching instants (x_i, y_i) lie on the branches of the parabolas

$$x = -C\text{sign}(y)y^2, \qquad C = \frac{x_i}{y_i^2}, \quad i = 1, 2, 3, \ldots. \tag{4.17}$$

Another group that is acting on this problem is the group of the reflections, in facts if the tuple $(\lambda_2(t), \lambda_1(t), x(t), y(t))$ is a solution of the problem, then the tuple $(-\lambda_2(t), -\lambda_1(t), -x(t), -y(t))$ is also a solution of the problem. Those trajectories can be obtained from the previous by reflection with respect to the plane $y = 0$, $\lambda_1 = 0$.
With these properties we can impose the following system of equation in the interval between two successive switches t_{n-1} and t_n:

$$\begin{aligned} y(t_n) &= -ky(kt_{n-1}) \\ x(t_n) &= -k^2x(kt_{n-1}) \\ \lambda_1(t_n) &= -k^3\lambda_1(kt_{n-1}) \\ \lambda_2(t_n) &= -k^4\lambda_2(kt_{n-1}) = 0, \end{aligned}$$

the explicit system is

$$\begin{aligned} u\Delta + a_2 &= -ka_2 \\ \frac{u}{2}\Delta^2 + a_2\Delta + a_1 &= -k^2a_1 \\ -\frac{u}{6}\Delta^3 - a_2\Delta^2 - 2a_1\Delta + q1 &= -k^3q_1 \\ \frac{u}{24}\Delta^4 + \frac{a_2}{3}\Delta^3 + a_1\Delta^2 - q_1\Delta &= 0. \end{aligned}$$

The solution and substitution of the first three equation yields respectively

$$\Delta = -\frac{a_2(k+1)}{u}, \quad a_1 = -\frac{a_2^2(k^2-1)}{2u(k^2+1)}, \quad q_1 = -\frac{a_2^3(k^3-2k^2-2k+1)(k+1)}{3u^2(k^2+1)(k^2-k+1)},$$

finally, the last equation is

$$\frac{a_2^4(k+1)^2(k^2-1)(k^4-3k^3-4k^2-3k+1)}{24u^3(k^2+1)(k^2-k+1)} = 0,$$

where after the simplification of the trivial factors $k = \pm 1$ and factoring out $a_2 \neq 0$, reduces to

$$k^4 - 3k^3 - 4k^2 - 3k + 1 = 0.$$

If z is a root of this equation, then also $1/z$ is a root; moreover the equation evaluated in 0 and 1 gives respectively 1 and -8, therefore there is a root $0 < z < 1$. There are now several ways to continue, a standard algebraic substitution from Galois theory is $\sigma = k + \frac{1}{k}$, another possibility that has deeper connections with the problem is $k = \sqrt{\frac{1-2C}{1+2C}}$. The first case brings to

$$\sigma^2 - 3\sigma - 6 = 0 \implies \sigma = \frac{3 \pm \sqrt{33}}{2} \approx 4.37, \ -1.37$$

but the negative root is irrelevant, because it give rise to complex solutions for k, the positive root gives

$$\begin{aligned} k_1 &= \frac{3 + \sqrt{33} - \sqrt{26} + 6\sqrt{33}}{4} \approx 0.242121374 \\ k_2 &= \frac{3 + \sqrt{33} + \sqrt{26} + 6\sqrt{33}}{4} \approx 4.130159950. \end{aligned}$$

The second method yields the biquadratic equation

$$C^4 + \frac{C^2}{12} - \frac{1}{18} = 0, \tag{4.18}$$

which has two complex roots, and

$$\begin{aligned} C_1 &= \frac{\sqrt{6\sqrt{33}-6}}{12} \approx 0.4446235601 \\ C_2 &= -\frac{\sqrt{6\sqrt{33}-6}}{12} \approx -0.4446235601. \end{aligned}$$

We look for the location of the first switch (x_1, y_1, t_1) on the parabola $x = Cy^2$ with C the root of equation (4.18). Consider the initial point of the trajectory $(x_0, y_0) = (2, -2)$, it follows that the initial control is $u = -1$ until the switching curve is reached. The corresponding trajectory is given by

$$x(y) = \frac{1}{2u}y^2 + \left(x_0 - \frac{y_0^2}{2u}\right), \qquad t = \frac{y - y_0}{u}$$

and the intersection with the parabola $x = Cy^2$ gives

$$\begin{aligned} y_1 &= -\frac{\sqrt{(1+2C)(y_0^2 + 2x_0)}}{1+2C} && \approx -2.057787900 \\ x_1 &= Cy_1^2 && \approx 1.882754481 \\ t_1 &= y_0 - y_1 && \approx 0.057787900. \end{aligned} \tag{4.19}$$

With the same argument but with $u = 1$ we can integrate the differential system with initial point (x_1, y_1, t_1) to find the second switch. The result is

$$\begin{aligned} y_2 &= -\frac{\sqrt{(1+2C)(y_1^2 - 2x_1)}}{1+2C} && \approx 0.4982344331 \\ x_2 &= -Cy_2^2 && \approx -0.1103722634 \\ t_2 &= y_2 - y_1 + y_0 - y_1 && \approx 2.613810233. \end{aligned} \tag{4.20}$$

With the above expression we can prove that the switching points are in geometrical progression of ratio $k = k_1$ and the duration of the arcs is also a geometrical progression. Consider the expression for y_2, a simple computation shows that

$$|y_2| = \frac{\sqrt{(1+2C)(1-2C)y_1^2}}{1+2C} = |y_1|\sqrt{\frac{1-2C}{1+2C}} = k|y_1| \implies \frac{|y_2|}{|y_1|} = k.$$

For the x variable we have

$$|x_2| = Cy_2^2 = Ck^2y_1^2 = k^2|x_1|$$

To check that the duration of the arcs is also in geometric progression we need to compare two entire durations, thus we need to compute also the instant $t_3 = -y_3 + 2y_2 - 2y_1 + y_0 \approx 3.232677872$, then the ratio of two successive time intervals is

$$\frac{t_3 - t_2}{t_2 - t_1} = \frac{-y_3 + y_2}{y_2 - y_1} = k.$$

It is now possible to compute the total time T to reach the origin, set $\Delta = t_2 - t_1$, then

$$T = t_1 + \sum_{i=1}^{\infty}(t_{i+1} - t_i) = t_1 + \Delta + k\Delta + k^2\Delta + \ldots = t_1 + \frac{\Delta}{1-k} \approx 3.430389060.$$

Moreover, the n^{th} switch is located (if we start with the first control $u = -1$, otherwise the signs reverse) at

$$\begin{aligned} t_n &= t_1 + \Delta\frac{k^{n-1}-1}{k-1} \\ x_n &= (-1)^{n-1}k^{2n-2}x_1 \\ y_n &= (-1)^{n-1}k^{n-1}y_1. \end{aligned}$$

We conclude the exposition with the analysis of the target. As before, we study the integral for the initial partial arc for $t \in [0, t_1]$ and then we consider two entire arcs among the switching points t_1, t_2. In the first interval we have

$$I_0 = \int_0^{t_1} x(t)^2\,\mathrm{d}t = \int_0^{t_1}\left(\frac{-t}{2} + y_0 t + x_0\right)^2\,\mathrm{d}t = \frac{(t_1(2y_0-1)+2x_0)^3}{24y_0-12} - \frac{2x_0^3}{6y_0-3},$$

in particular $I_0 \approx 0.2148564335$. The integral over the interval $[t_1, t_2]$ has a simple but long analytic expression that here is omitted, it can be approximated to $I_1 \approx 1.29622064$, therefore the target can be evaluated with the relation

$$J = I_0 + \frac{I_1}{1-k^5} \approx 1.515228194.$$

As an appendix of this problem, we report a brief table (4.21) of the first six switching points.

t_i	x_i	y_i
0.057787900	1.8827544810	−2.057787900
2.613810233	−0.1103722637	0.498234433
3.232677872	0.0064703266	−0.120633205
3.382518955	−0.0003793084	0.029207877
3.418798684	0.0000222361	−0.007071851
3.427582782	−0.0000013035	0.001712246

(4.21)

5

BENCHMARKS ON A PROBLEM SUITE

5.1 CLASSIC PROBLEMS

5.1.1 *The Brachistochrone*

One of the first and most famous problems in the calculus of variations is the problem of the brachistochrone, proposed originally by Bernoulli in 1696. There are various ways to consider this problem, the most important deal with calculus of variations and optimal control theory, involving ordinary or differential algebraic equations (ODE and DAE), the law of conservation of energy.
The statement asks to find the path of minimum time that joins two points, when only the gravity force is active. Supposing motion in two dimensions, from a starting point A to a fixed end point B: the question is to find the shape of the rigid path on which a particle subject only to gravity force travels from A to B in minimum time.
A good coordinate system has the origin in the starting point A and the vertical axis y directed upward, i.e. with the opposite direction of the gravity force, the horizontal axis x orthogonal to y

such that the x coordinate of B is positive. Because the optimal path γ must be a continuous curve and can not have loops, it can be assumed without loss of generality that γ can be represented as a function $y(x)$. In order to simplify signs, a classical assumption is to consider $y > 0$ so that the minus of the orientation of g and the minus of the negative height of y_B vanishes. Applying the conservation of energy, it is clear that if m is the mass of the particle, g the acceleration of gravity, v the velocity of the particle,

$$mgy = \frac{1}{2}mv^2 \quad \Longrightarrow \quad v(y(x)) = \sqrt{2gy(x)}.$$

In particular, this shows that the solution is independent of the mass m. The distance L travelled by the particle when it approaches the end point B is given by the line integral

$$L(y, x) = \oint_\gamma \mathrm{d}s = \int_0^{x_B} \sqrt{1 + y'(x)^2}\,\mathrm{d}x.$$

The time used to travel the path is given by

$$T(y, x) = \int_0^{x_B} \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2gy(x)}}\,\mathrm{d}x. \tag{5.1}$$

Therefore the solution of the problem consists in minimizing the functional (5.1) subject to the following constraints,

$$y(0) = 0, \qquad y(x_B) = -y_B. \tag{5.2}$$

5.1.1.1 *Solution with Calculus of Variations*

The approach of the calculus of variations to solve the brachistochrone problem is making use of the Euler–Lagrange equation, and because the functional $T(y, x) = \int F(y, x)\,\mathrm{d}x$ has in facts no explicit dependence on x, the Euler–Lagrange equation reduces to Beltrami's identity [Por07]

$$\frac{\partial F}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial y'} = 0 \qquad \Longrightarrow \qquad F - y'\frac{\partial F}{\partial y'} = c, \tag{5.3}$$

where c is a real constant that can be determined using the information of the constraint (5.2). In the case of the brachistochrone F can be simplified in

$$\sqrt{2g}T = \int_0^{x_B} \sqrt{\frac{1 + y'(x)^2}{y(x)}}\,\mathrm{d}x \qquad \Longrightarrow \qquad F(y, x) = \sqrt{\frac{1 + y'(x)^2}{y(x)}},$$

therefore the Euler–Lagrange equation (5.3) becomes, skipping the dependence on x,

$$\sqrt{\frac{1 + y'^2}{y}} - \frac{y'^2}{\sqrt{y(1 + y'^2)}} = c \qquad \Longrightarrow \qquad y(1 + y'^2) = \frac{1}{c^2} = C > 0.$$

The latter is an autonomous nonlinear differential equation solvable with the separated form technique, that is

$$\int_{y_A}^{y_B} \sqrt{\frac{y}{C - y}}\,\mathrm{d}y = \int_{x_A}^{x_B} 1\,\mathrm{d}x.$$

A primitive of this integral can be computed via the substitution $y = C\sin^2\frac{\phi}{2}$ with differential $\mathrm{d}y = \sin\frac{\phi}{2}\cos\frac{\phi}{2}\,\mathrm{d}\phi$, thus

$$\begin{aligned}\int\sqrt{\frac{y}{C-y}}\,\mathrm{d}y &= C\int\sqrt{\frac{\sin^2\frac{\phi}{2}}{1-\sin^2\frac{\phi}{2}}}\sin\frac{\phi}{2}\cos\frac{\phi}{2}\,\mathrm{d}\phi\\ &= C\int\sin^2\frac{\phi}{2}\,\mathrm{d}\phi\\ &= \frac{C}{2}(\phi-\sin\phi)+k\end{aligned}$$

where k is the integrating constant.
Evaluating the integral in the extrema given by the constraints condition (5.2), permits to determine the two constants C and k. The fact that the starting point A coincides with the origin implies that $k=0$, the constant C depends on the coordinates of $B=(x_B,y_B)$. In particular

$$x_B = \frac{C}{2}(\phi_B-\sin\phi_B),\qquad y_B = -\frac{C}{2}(1-\cos\phi_B). \tag{5.4}$$

In conclusion the brachistochrone has parametric equation

$$x(\phi) = \frac{C}{2}(\phi-\sin\phi),\qquad y(\phi) = -\frac{C}{2}(1-\cos\phi) \tag{5.5}$$

which is the equation of a cycloid curve.
The geometric proof given by Bernoulli suffers of spurious solutions [SW01], with the variational form it is difficult to prove the existence of the brachistochrone. Applying optimal control theory, the existence follows directly from the theorem of Ascoli–Arzelá.

5.1.1.2 *The Brachistochrone as an Optimal Control Problem*

This example can be treated as an optimal control problem, the control is the angle ϑ of descent of the particle. Splitting the velocity v in its two components along the xy axes, the problem can be states as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v\sin\vartheta,\qquad \frac{\mathrm{d}y}{\mathrm{d}t} = -v\cos\vartheta,\qquad \frac{\mathrm{d}v}{\mathrm{d}t} = g\cos\vartheta \tag{5.6}$$

with border conditions, if T is the time used to travel from A to B,

$$\begin{aligned}&x(0)=0,\quad y(0)=0,\quad v(0)=0,\\ &x(T)=x_B,\ y(T)=y_B,\ v(T)=v_T,\end{aligned} \tag{5.7}$$

for a free final velocity v_T. The functional to be minimized is T.

5.1.1.3 *A Numerical Example*

Consider the brachistochrone problem of a particle that is left free in the origin, and travels to $B=(10,-3)$. In the variational formulation constant C must be retrieved. This is done using equation (5.4). Solving the nonlinear system for C and ϕ_B yields to

$$\phi_B \approx 4.17,\qquad C\approx 3.97.$$

The solution is plotted in figure 5.1. The same result, treated as an optimal control problem as

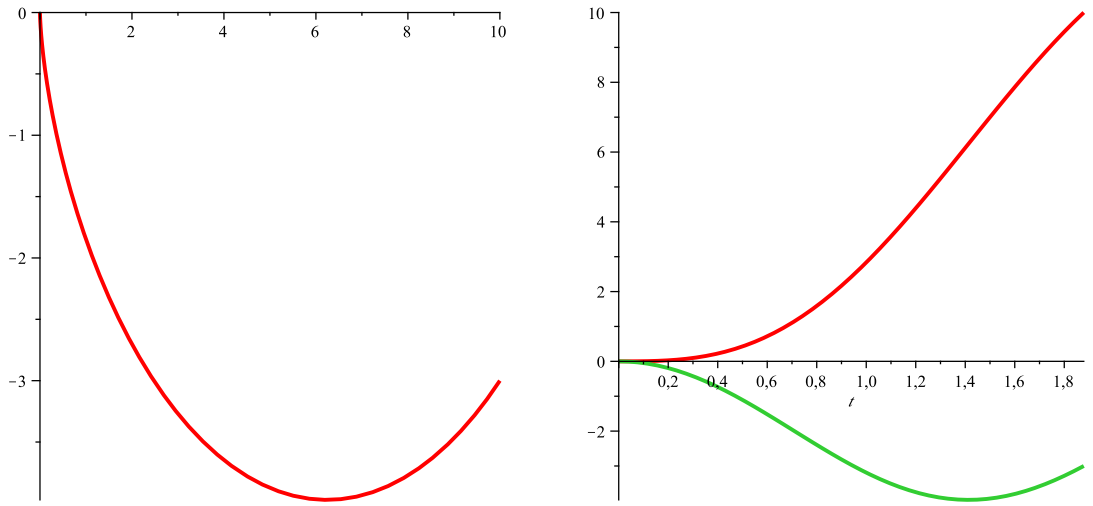


Figure 5.1: Variational solution to the brachistochrone problem: the cycloid. On the left the trajectory of the particle from the origin to B, on the right the two components of the solution, x in red and y in green.

stated in (5.6) with boundary conditions (5.7), is difficult to solve analitically, in facts, for example, the Hamiltonian function is

$$\mathcal{H} = 1 + \lambda v \sin\vartheta - \mu v \cos\vartheta + \xi g \cos\vartheta.$$

From the Maximum Principle of Pontryagin (PMP), the associated boundary value problem is

$$\lambda' = -\frac{\partial \mathcal{H}}{\partial x} = 0, \qquad \mu' = -\frac{\partial \mathcal{H}}{\partial y} = 0, \qquad \xi' = -\frac{\partial \mathcal{H}}{\partial v} = \lambda \sin\vartheta - \mu \cos\vartheta$$

for the multipliers, and

$$0 = \frac{\partial \mathcal{H}}{\partial \vartheta} = \lambda v \cos\vartheta + \mu v \sin\vartheta - \xi g \sin\vartheta.$$

To bypass the analytical solution, a way is to use the answer given by the variational formulation to deduce the solution of the OCP. This passage can not be done directly because equations (5.5) are function of the space angle ϕ and not of time t. To find a relation between them one can use the time funtional

$$T(x,y) = \int_0^{\phi_B} \frac{\sqrt{x'(\phi)^2 + y'(\phi)^2}}{\sqrt{2gy(\phi)}}\,\mathrm{d}\phi.$$

A short computation shows that

$$T(x,y) = \int_0^{\phi_B} \sqrt{\frac{C}{2g}}\,\mathrm{d}\phi = \phi_B \sqrt{\frac{C}{2g}} = \phi_B k \approx 1.87,$$

this yields to the property of tautochrone (isochrone) curve of the cycloid. Now the parametrization of time is simply $t = \phi k$, for $t \in [0, \phi_b k]$

From this point of view, the control $\vartheta(t)$ should satisfy

$$\vartheta(t) = \arctan\left(\frac{y'(t)}{x'(t)}\right) + \frac{\pi}{2} = \arctan\left(-\frac{\sin t/k}{1-\cos t/k}\right) + \frac{\pi}{2},$$

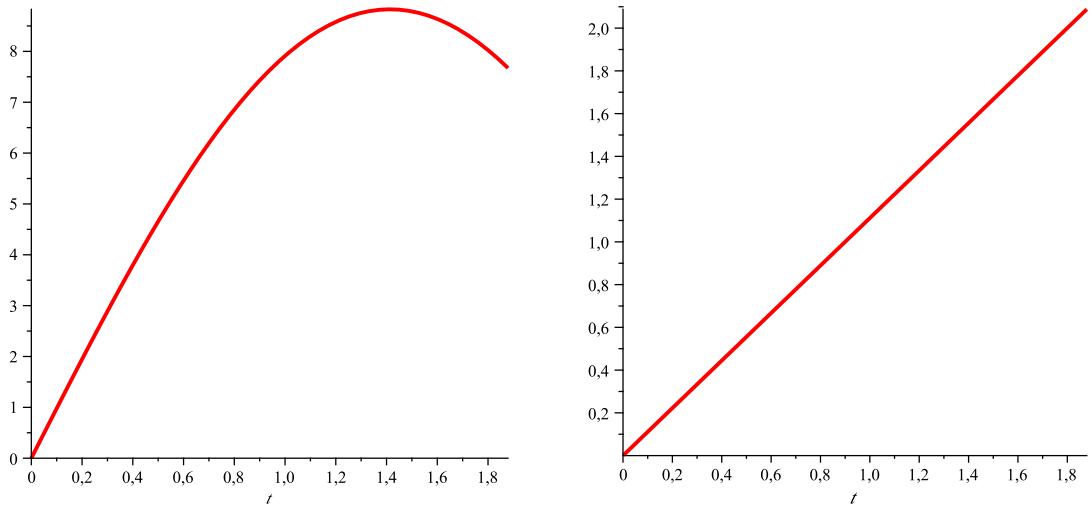


Figure 5.2: On the left the plot of the velocity $v(t)$, on the right the control variable $\vartheta(t)$

so that

$$\sin\vartheta = \frac{1-\cos t/k}{\sqrt{2}\sqrt{1-\cos t/k}}, \qquad \cos\vartheta = \frac{\sin t/k}{\sqrt{2}\sqrt{1-\cos t/k}}.$$

thus the velocity becomes

$$v(t) = \frac{\sqrt{2}C}{2k}\sqrt{1-\cos t/k}.$$

The plot of the velocity $v(t)$ and the control $\vartheta(t)$ is shown in figure 5.2. The differential equations for the space components become

$$x'(t) = v(t)\sin\vartheta = \frac{C}{2}(1-\cos t/k), \qquad y'(t) = -v(t)\cos\vartheta = -\frac{C}{2}(\sin t/k).$$

To compute the multipliers λ, μ and ξ it is enough to use the Hamiltonian function: substituting the expression for xi it remains a single function in one unknown λ, which gives $\lambda = -k/C$, therefore at the end the three multipliers are,

$$\lambda = -\frac{k}{C} \approx -0.11$$

$$\mu = -\frac{\lambda\sin T/k}{1-\cos T/k} \approx -0.06$$

$$\xi(t) = \frac{\sqrt{2}C}{2gk}\frac{\lambda\sin t/k + \mu(1-\cos t/k)}{\sqrt{1-\cos t/k}}.$$

The plot of the costate (multipliers) is shown in figure 5.3 Using the ACADO toolkit to solve the same problem yields to the following solution.The plots of this functions are showed in figure 5.4. We report in table 5.1 the results we collected and computed.

5.1.2 *Single Integrator Plant*

The single integrator plant problem, is a classic one and was proposed originally by Goh and Teo in [GT88], but also by Luus [Luu91]. Several results and comparison we collected here were

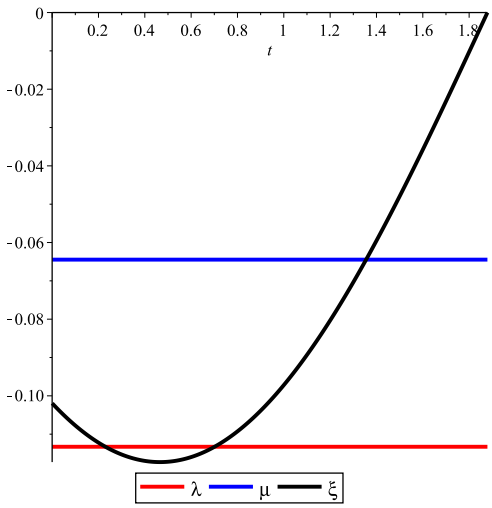


Figure 5.3: The three multipliers, λ in red, μ in blue and ξ in black.

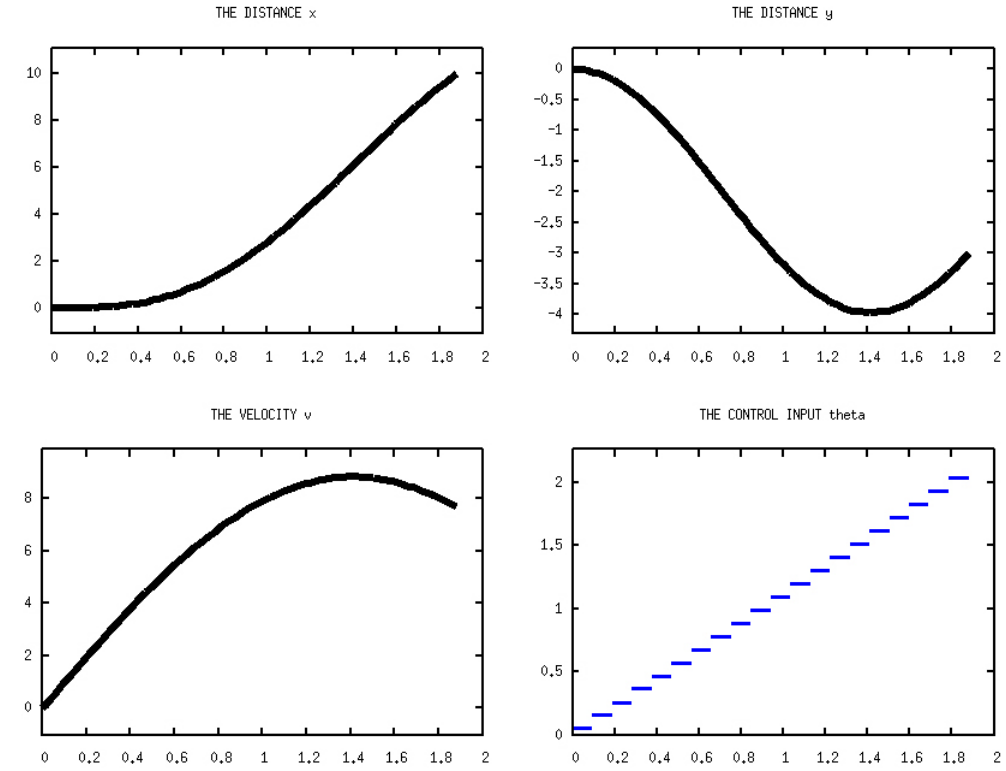


Figure 5.4: Solution given by ACADO with a coarse mesh.

discussed in the paper by Dadebo and McAuley in [DM95a].
The problem can be described by the following differential equations:

$$\min x_2(T), \qquad x_1'(t) = u(t), \quad x_2'(t) = x_1(t)^2 + u(t)^2,$$

Table 5.1: Summary of the results for the Brachistochrone, in the first column the name of the algorithm, in the second column the value of the target, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	1.8789403296413785	0
XOptima	1.8789410460182487	3.8 E-07
ICLOCS	1.8789488420535776	4.5 E-07
Gpops	1.8789403296291913	-6.4 E-12
ACADO	1.8789488680010755	4.5 E-06
PROPT	1.8789403291138431	-2.8 E-10

with the final time set to $T = 1$ and the initial conditions $x_1(0) = 1$ and $x_2(0) = 0$. The problem is further subject to the terminal constraint $x(1) = 1$. We report in table 5.2 the results we collected and computed. The optimal target is $x_2(T) = x_2(1) = \frac{2(e-1)}{e+1} \approx 0.92$.

Table 5.2: Summary of the results for the single integrator plant, in the first column the name of the algorithm, in the second column the value of the target, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	0.9242343145200195	0
Present method	0.9242875107734982	5.7 E-05
XOptima	0.9242343800573381	7.1 E-08
ICLOCS	0.9242346022361485	3.1 E-07
Gpops	0.9242343145186011	-1.5 E-12
ACADO	0.9242360964829455	1.9 E-06
Goh and Teo [GT88]	0.92518	1.0 E-03
Luus [Luu91]	0.92441	1.9 E-04
Dadebo and McAuley [DM95a]	0.92428	4.9 E-05

5.2 SINGULAR PROBLEMS

5.2.1 *Dubins Car*

The simple car model has three degrees of freedom, the car can be imagined as a rigid body that moves in a plane. The back wheels can only slide and that is why parallel parking is challenging. If all wheels could be rotated together, parking would be a trivial task. The position of a car can be identified with the triple $(x, y, \theta) \in \mathbb{R}^2 \times \mathbb{S}$, where x and y are the principal directions and θ is the angle of the car with the x axis. From the geometry of the model (simplifying some constant to unity), the following system of differential equations can be retrieved:

$$\begin{aligned} x' &= \cos(\theta), \\ y' &= \sin(\theta), \\ \theta' &= u. \end{aligned}$$

The problem is to drive in minimum time the car from an assigned position to the origin, The control u is designed to be in the interval $[-2, 2]$. Hence the optimal control problem is stated as:

$$\min T = \min \int_0^T 1 \, \mathrm{d}t \qquad s.t. \qquad |u| \leq 2 \qquad \text{with}$$

$$\begin{aligned} x' &= \cos(\theta), & x(0) &= 4, & x(T) &= 0, \\ y' &= \sin(\theta), & y(0) &= 0, & y(T) &= 0, \\ \theta' &= u, & \theta(0) &= \frac{\pi}{2}. \end{aligned}$$

The control u appears linearly, so we expect a singular arc. The Hamiltonian for this problem is

$$\mathcal{H} = 1 + \lambda_1 \cos(\theta) + \lambda_2 \sin(\theta) + \lambda_3 u.$$

From the PMP, $u = \arg\min \mathcal{H}$, therefore we can write

$$u = \begin{cases} 2 & \text{if} \quad \lambda_3 < 0, \\ ? & \text{if} \quad \lambda_3 = 0, \\ -2 & \text{if} \quad \lambda_3 > 0. \end{cases}$$

The equation of the costate are derived from the Hamiltonian,

$$\begin{aligned} -\lambda_1' &= \frac{\partial \mathcal{H}}{\partial x} = 0, \\ -\lambda_2' &= \frac{\partial \mathcal{H}}{\partial y} = 0, \\ -\lambda_3' &= \frac{\partial \mathcal{H}}{\partial \theta} = -\lambda_1 \sin\theta + \lambda_2 \cos\theta. \end{aligned}$$

From the previous equations, the multipliers λ_1 and λ_2 are real constants. Performing further differentiation on λ_3', we have that in the singular arc $\lambda_3'' = 0$, thus,

$$\lambda_3'' = \lambda_1 \theta' \cos\theta + \lambda_2 \theta' \sin\theta = \lambda_1 u \cos\theta + \lambda_2 u \sin\theta = 0,$$

that is, in the singular arc, $u(\lambda_1 \cos\theta + \lambda_2 \sin\theta) = 0 \implies u = 0$, i.e. the control is zero. With this information on u we reconstruct the singular arc, because from $\theta' = u = 0 \implies \theta(t) = K$ constant.

This forces $x' = \cos K$ and $y' = \sin K$, which integrated give the singular arc $x(t) = t\cos K + K_1$ and $y(t) = t\sin K + K_2$ for two real constants K_2, K_3.
We analyse now the non singular arc, there $\theta(t) = -2t\text{sign}(\lambda_3) + k_3$. Using the initial condition $\theta(0) = \frac{\pi}{2}$ we solve $k_3 = \frac{\pi}{2}$. Calling $m = -2\text{sign}(\lambda_3)$, the associated arc becomes:

$$\begin{aligned} x'(t) &= \cos(mt + k_3) \implies x(t) = \frac{1}{m}\sin(mt + \pi/2) + k_1, \\ y'(t) &= \sin(mt + k_3) \implies y(t) = -\frac{1}{m}\cos(mt + \pi/2) + k_2. \end{aligned}$$

Now we assume that there is only a singular arc, in the interval $(t_A, T]$, where t_A is the unknown switching instant. This assumption is suggested by the fact that the final condition of θ is free, so the associated multiplier λ_3 is zero for $t = T$. Therefore the arc is non singular in the interval $[0, t_A]$, studying the initial conditions we can guess $\lambda_3 < 0$ in that interval. This implies that $k_1 = 4 - 1/2 = 7/2$ and $k_2 = -\cos(\pi/2) = 0$, and the associated trajectory is

$$x(t) = \frac{1}{2}\cos(2t) + \frac{7}{2}, \qquad y(t) = \frac{1}{2}\sin(2t), \qquad \theta(t) = 2t + \frac{\pi}{2}.$$

It is now possible to join the first arc with the second one, because the trajectory is a continuous function, for $t = t_A$,

$$\begin{aligned} x(t_A) &= t_A\cos K + K_1 = \frac{1}{2}\cos(2t_A) + \frac{7}{2}, \\ y(t_A) &= t_A\sin K + K_2 = \frac{1}{2}\sin(2t_A), \\ \theta(t_A) &= K = 2t_A + \frac{\pi}{2}. \end{aligned}$$

We can also impose the final conditions

$$\begin{aligned} x(T) &= T\cos K + K_1 = 0, \\ y(T) &= T\sin K + K_2 = 0. \end{aligned}$$

This is a non linear system of five equations in the five unknowns K, K_1, K_2, t_A, T, we solve it next. It is a quick manipulation to simplify the dependence of K, K_1, K_2 and we end up with two equations in two unknowns, T and t_A:

$$\begin{aligned} -T\sin(2t_A) + t_A\sin(2t_A) + \frac{1}{2}\cos(2t_A) + \frac{7}{2} = 0, \\ T\cos(2t_A) - t_A\cos(2t_A) + \frac{1}{2}\sin(2t_A) = 0. \end{aligned}$$

Multiplying the first equation by $\cos(2t_A)$ and adding the second multiplied by $\sin(2t_A)$, then multiplying the first equation by $\sin(2t_A)$ and adding the second multiplied by $\cos(2t_A)$ yields

$$\begin{aligned} \frac{1}{2} + \frac{7}{2}\cos(2t_A) = 0, \\ -T + t_A + \frac{7}{2}\sin(2t_A) = 0. \end{aligned}$$

From this couple of expressions it is easy to solve the switching the instant t_A and the final (minimum) time T, and, consequently,the three constants K, K_1, K_2 :

$$\begin{aligned}
t_A &= \frac{\pi}{2} - \frac{1}{2}\arccos\frac{1}{7} && \approx 0.857071948,\\
T &= \frac{\pi}{2} - \frac{1}{2}\arccos\frac{1}{7} + 2\sqrt{3} && \approx 4.321173564,\\
K &= \frac{3\pi}{2} - \arccos\frac{1}{7} && \approx 3.284940223,\\
K_1 &= \frac{2\sqrt{3}}{7}\left(\pi - \arccos\frac{1}{7} + 4\sqrt{3}\right) && \approx 4.276852666,\\
K_2 &= \frac{\pi}{14} - \frac{1}{14}\arccos\frac{1}{7} + \frac{2\sqrt{3}}{7} && \approx 0.6173105091.
\end{aligned}$$

These constants permit to completely solve the state of the system at any time $t \in [0, T]$. It remains to specify the costate. We already saw that λ_1 and λ_2 are constant, for λ_3 we have the differential equation given by $-\partial\mathcal{H}/\partial\theta = 0$ for the singular arc, therefore we need other two equations in order to set up a non linear system in the three unknowns $\lambda_1, \lambda_2, \lambda_3$. One equation can be the Hamiltonian itself, which is autonomous and hence equal to zero. The third equation can be the expression of the multiplier in the interval $[0, t_A]$, $\lambda_3(t) = -\lambda_1/2\cos(2t + k_3) - \lambda_2/2\sin(2t + k_3) + \lambda_3(0)$. We introduce here a fourth unknown, $\lambda_3(0)$, but we do not need another equation. The non linear system, for $t = t_A$, is:

$$\begin{aligned}
\mathcal{H} &= 1 + \lambda_1\cos(\theta) + \lambda_2\sin(\theta) + \lambda_3 u = 0,\\
\lambda_3'(t_A) &= -\frac{\partial\mathcal{H}}{\partial\theta} = \lambda_1\sin\theta - \lambda_2\cos\theta = 0,\\
\lambda_3(t_A) &= -\lambda_1/2\cos(2t_A + k_3) - \lambda_2/2\sin(2t_A + k_3) + \lambda_3(0).
\end{aligned}$$

The solution of the system gives the following expressions,

$$\begin{aligned}
\lambda_1 &= \frac{\cos(k_3)}{\sin(2t_A)\cos k_3 - \sin K\cos(2t_A)} = \frac{4}{7}\sqrt{3} \approx 0.9897433188,\\
\lambda_2 &= \frac{\sin(k_3)}{\sin(2t_A)\cos k_3 - \sin K\cos(2t_A)} = \frac{1}{7} \qquad \approx 0.1428571429,\\
\lambda_3(0) &= -\frac{1}{2}.
\end{aligned}$$

In Figure 5.5 the plots for state, costate, control and trajectory. The numerical results 5.3 are quite surprising: we were able to make Gpops, Iclocs and Acado converge only with great effort imposing some extra bound on the states, while XOptima gives readily a good solution with a poor guess. The control solved by Iclocs has a very oscillating damped singular arc. In particular the path constraint for the angle had to be relaxed to the range $[-10, 10]$. On the contrary, Acado converged only with a very strict bound on the final time, e.g. $T \in [3.1, 4.8]$, while in Iclocs/Gpops was enough $[0.1, 100]$ and in XOptima only setting a penalty for the time to be positive.

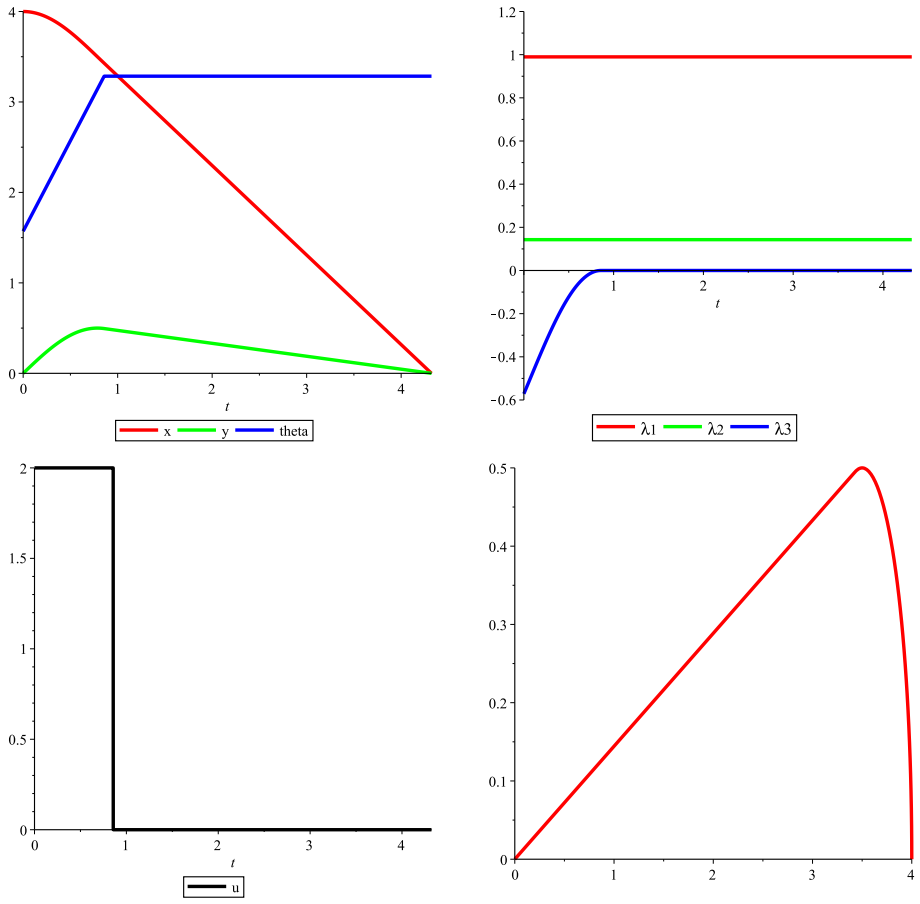


Figure 5.5: Variational solution to the Dubins car problem. On the top left the states , on the right the costates, below the control and the trajectory.

Table 5.3: Summary of the results for problem of the Dubins car, in the first column the the name of the algorithm, in the second column the value of the target, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	4.32117356298788557	0
XOptima	4.32117216744031918	-3.2 E-07
ICLOCS	4.3212508202939119	1.7 E-05
Gpops	4.3211747200514896	2.6 E-07
ACADO	4.1817537736144796	-3.2 E-02

5.2.2 *An OCP with Singular Controls*

When the Hamiltonian $\mathcal{H}$ is linear in the control u, the solution involves discontinuities in the optimal control. If the switching function associated to the control is not sustained over an interval of time, that is, the coefficient of u in $\mathcal{H}$ is equal to zero only for isolated instants, the control is bang-bang. A bang-bang control assumes always the extreme values of the control set. If, instead, the coefficient of u in $\mathcal{H}$ is equal to zero over an interval of time, the control is singular. The choice of u must be obtained from other information than the Pontryagin maximum principle. The times when the optimal control switches from a state to another or to singular controls, are called switch times and are sometimes difficult to find.
The next example shows a singular control. The problem asks to minimize the functional

$$\begin{aligned} &\min \int_{-1}^{1} (x - 1 + t^2)^2 \, \mathrm{d}t, && \text{subject to the dynamic:} \\ &x' = u, && \text{with control bounded by:} \\ &|u| \leq 1. \end{aligned}$$

The Hamiltonian of the problem is

$$\mathcal{H} = (x - 1 + t^2)^2 + \lambda u + p_\varepsilon(u),$$

where $p_\varepsilon(u)$ is a penalty function introduced to handle the constrained controls in the interval $[-1, 1]$. An example of such a function is $p_\varepsilon(u) = -\varepsilon \ln(\cos \frac{\pi}{2} u)$. When $\varepsilon \to 0$ the function is close to zero in $[-1, 1]$ and grows to infinity outside that interval. It is clear that the Pontryagin maximum principle is of no use in this case. The stationarity condition for the optimal control problem,

$$\frac{\partial \mathcal{H}}{\partial u} = \lambda + \frac{\partial p_\varepsilon(u)}{\partial u} = 0$$

implies that

$$\lambda(t) = -\varepsilon \frac{\pi}{2} \tan \frac{\pi}{2} u(t) \implies u(t) = \lim_{\varepsilon \to 0} -\frac{2}{\pi} \arctan\left(\frac{2\pi\lambda(t)}{\varepsilon}\right) = -\operatorname{sign}(\lambda(t)).$$

Hence, when $\lambda \neq 0$ the control is $u = \pm 1$, when $\lambda = 0$ the control is singular.
The state is not specified at the boundary, so there are the transversality conditions $\lambda(-1) = \lambda(1) = 0$. Differentiating the Hamiltonian with respect to x yields the information on the multiplier λ:

$$\lambda' = -\frac{\partial \mathcal{H}}{\partial x} = -2(x - 1 + t^2),$$

therefore, by integration, the expression of $\lambda(t)$ is

$$\lambda(t) = -2 \int_{-1}^{t} (x(s) - 1 + s^2) \, \mathrm{d}s + k \tag{5.8}$$

for a constant $k \in \mathbb{R}$ that can be determined by the initial value $\lambda(-1) = 0$, i.e.,

$$0 = \lambda(-1) = -2 \int_{-1}^{-1} (x(s) - 1 + s^2) \, \mathrm{d}s + k \implies k = 0.$$

Using $k = 0$ in equation (5.8) allows to write

$$0 = \lambda(1) = -2 \int_{-1}^{-1} (x(s) - 1 + s^2) \, \mathrm{d}s \implies \int_{-1}^{-1} (x(s) - 1 + s^2) \, \mathrm{d}s = 0.$$

For t in the time interval $[t_A, t_B] \subset [-1, 1]$ in which $\lambda(t) = 0$,

$$\begin{aligned} 0 = \lambda(t) &= -2\left(\int_{-1}^{t_A} (x(s) - 1 + s^2)\,\mathrm{d}s + \int_{t_A}^{t} (x(s) - 1 + s^2)\,\mathrm{d}s\right) \\ &= \lambda(t_A) - 2\int_{t_A}^{t} (x(s) - 1 + s^2)\,\mathrm{d}s, \end{aligned} \tag{5.9}$$

hence, differentiating the right hand side of the last line,

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left(-2\int_{t_A}^{t} (x(s) - 1 + s^2)\,\mathrm{d}s\right) \implies x(t) - 1 + t^2 = 0, \quad t \in [t_A, t_B].$$

The state x when the control is singular is thus $x(t) = 1 - t^2$ and from the dynamic of the problem, $u(t) = -2t$ for $t \in [t_A, t_B]$. It is important to point out that for $|t| > 1/2$ the control $u(t) = -2t$ does not satisfy the bounds of the hypothesis, the problem is thus to find the two switching times t_A and t_B such that $x' = u$ and $|u| \leq 1$. For $t \in [-1, t_A)$ the control is $u(t) = 1$ and for $t \in (t_B, 1]$ the control is $u(t) = -1$, the corresponding state is then respectively $x(t) = t + a$ and $x(t) = -t + b$. From equation (5.9) and the fact that for $t \in [t_A, t_B]$, $x(t) = 1 - t^2$, it follows that $\lambda(t_A) = 0$, i.e,

$$\begin{aligned} 0 = \lambda(t_A) &= -2\int_{-1}^{t_A} (x(s) - 1 + s^2)\,\mathrm{d}s \\ &= -2\int_{-1}^{t_A} (s + a - 1 + s^2)\,\mathrm{d}s \\ &= -2\left[\frac{1}{3}(t_A + 1)\left(t_A^2 + \frac{1}{2}t_A - \frac{7}{2} + 3a\right)\right]. \end{aligned}$$

This expression is equal to zero for

$$a = -\frac{1}{3}t_A^2 - \frac{1}{6}t_A + \frac{7}{6} \implies x(t) = t - \frac{1}{3}t_A^2 - \frac{1}{6}t_A + \frac{7}{6}. \tag{5.10}$$

Now, because x is continuous, in the switching point must hold $t_A + a = 1 - t_A^2$. From that equation it is possible to solve $t_A = -\frac{1}{4}$ and $a = \frac{19}{16}$.
With similar consideration it is possible to solve b and t_B: in fact,

$$0 = \lambda(1) = \lambda(t_B) - 2\int_{t_B}^{1} x(s) - 1 + s^2\,\mathrm{d}s \implies \int_{t_B}^{1} x(s) - 1 + s^2\,\mathrm{d}s = 0.$$

Using the expression $x(t) = -t + b$ for $t \in (t_B, 1]$, the previous integral yields

$$0 = \int_{t_B}^{1} -s + b - 1 + s^2\,\mathrm{d}s = -\frac{1}{3}(t_B - 1)\left(t_B^2 - \frac{1}{2}t_B - \frac{7}{2} + 3b\right),$$

which gives the same result of equation (5.10),

$$b = -\frac{1}{3}t_B^2 - \frac{1}{6}t_B + \frac{7}{6} \implies x(t) = -t - \frac{1}{3}t_B^2 - \frac{1}{6}t_B + \frac{7}{6}.$$

Again, for the continuity of x, $-t_B + b = 1 - t_B^2$, implies that $t_B = \frac{1}{4}$ and $b = \frac{19}{16}$. This also shows the symmetry of the solution. With this optimal control u and state x the minimum of the functional is

$$\begin{aligned} \int_{-1}^{1} (x - 1 + t^2)^2\,\mathrm{d}t &= \int_{-1}^{-1/4} \left(t + \frac{9}{16} - 1 + t^2\right)^2 \mathrm{d}t + \int_{1/4}^{1} \left(-t + \frac{9}{16} - 1 + t^2\right)^2 \mathrm{d}t \\ &= \frac{9}{1280} = 0.00703125. \end{aligned}$$

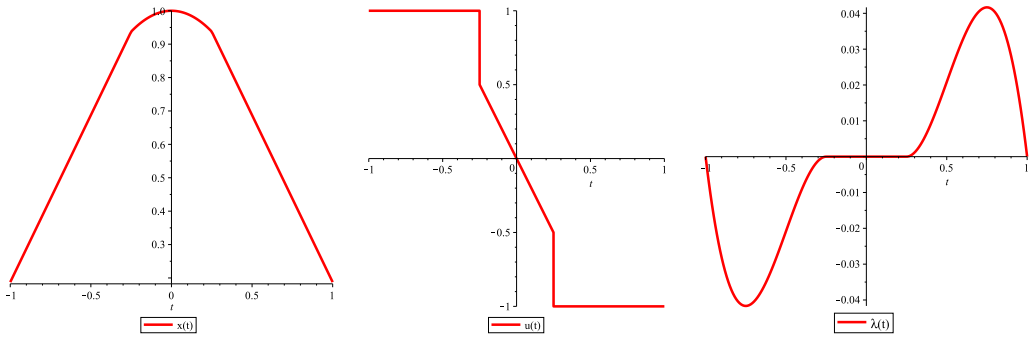


Figure 5.6: Variational solution to the singular arc problem. On the left the trajectory of the state $x(t)$, in the middle the control $u(t)$ and on the right the costate $\lambda(t)$.

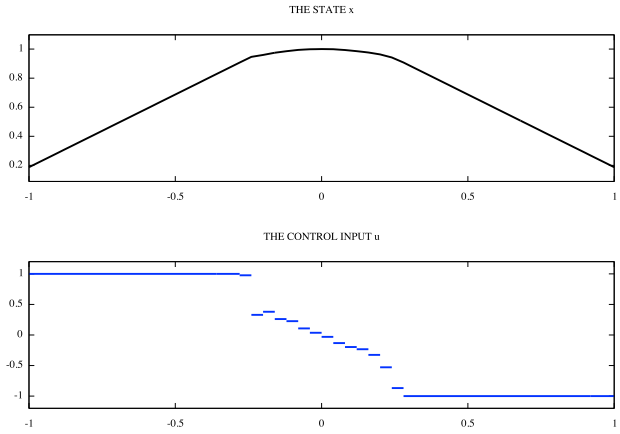


Figure 5.7: Numerical solution given by ACADO.

Collecting the three phases the analytical solution of the optimal control problem is,

$$x(t) = \begin{cases} t+a & \text{for} \quad t \in [-1, t_A) & = & [-1, -\frac{1}{4}) \\ 1-t^2 & \text{for} \quad t \in [t_A, t_B] & = & [-\frac{1}{4}, \frac{1}{4}], \\ -t+b & \text{for} \quad t \in (t_B, 1] & = & (\frac{1}{4}, 1] \end{cases} \qquad a = b = \frac{19}{16},$$

$$u(t) = \begin{cases} 1 & \text{for} \quad t \in [-1, t_A) & = & [-1, -\frac{1}{4}) \\ -2t & \text{for} \quad t \in [t_A, t_B] & = & [-\frac{1}{4}, \frac{1}{4}] \\ -1 & \text{for} \quad t \in (t_B, 1] & = & (\frac{1}{4}, 1] \end{cases}$$

$$\lambda(t) = \begin{cases} -\frac{2}{3}t^3 - t^2 - \frac{3}{8}t - \frac{1}{24} & \text{for} \quad t \in [-1, t_A) & = & [-1, -\frac{1}{4}) \\ 0 & \text{for} \quad t \in [t_A, t_B] & = & [-\frac{1}{4}, \frac{1}{4}], \\ -\frac{2}{3}t^3 + t^2 - \frac{3}{8}t + \frac{1}{24} & \text{for} \quad t \in (t_B, 1] & = & (\frac{1}{4}, 1]. \end{cases}$$

A graphical representation of the state x, the control u and the costate λ is shown in Figure 5.6. In Figure 5.7 the numerical solution obtained with ACADO.

5.2.3 *Luus n.1*

This is the first of four unconstrained singular optimal control problems proposed and analysed by Luus in [Luu00], the results are compared by him to the results obtained by other researchers in

Table 5.4: Summary of the results for problem of the Singular control problem, in the first column the the name of the algorithm, in the second column the value of the target, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	0.0070312500000000000	0
XOptima	0.0070312525135557231	3.5 E-07
ICLOCS	0.0070317890699292378	7.6 E-05
Gpops	0.0070292409711438441	-2.8 E-04
ACADO	0.0070371147114343157	8.3 E-04

their published papers.
Luus introduces the problem as simple-looking, pointing out that its triviality is only apparent from a numerical perspective [CH93]. As a matter of fact, the problem can be easily solved by hand analytically, because it is defined by two equations and the functional:

$$\begin{aligned} &\min x_2(T), && \text{subject to the dynamic:} \\ &x_1(t)' = u(t), && \\ &x_2(t)' = \frac{1}{2}x_1(t)^2, && \text{with control bounded by:} \\ &|u| \leq 1. && \end{aligned} \tag{5.11}$$

The given boundary conditions are $x_1(0) = 1$ and $x_2(0) = 0$, so that $\lambda_1(T) = 0$ and $\lambda_2(T) = 1$, where T is the final time set to $T = 2$. The Hamiltonian for this problem is

$$\mathcal{H} = \lambda_1 u + \frac{1}{2}\lambda_2 x_1^2.$$

The equation of the costate are derived from the Hamiltonian,

$$\begin{aligned} -\lambda_1' &= \frac{\partial \mathcal{H}}{\partial x_1} = -\lambda_2 x_1, \\ -\lambda_2' &= \frac{\partial \mathcal{H}}{\partial x_2} = 0, \end{aligned}$$

Therefore we have $\lambda_2(t) = 1$ constant. Supposing there is only one switching point, we can guess that the singular arc is the second, hence deriving the multiplier λ_1 in the singular part we obtain the singular control. From $\lambda_1' = -x_1$ we get $\lambda_1'' = -x_1' = -u = 0$, thus in the singular arc the control is zero. Now we can integrate the dynamical system to find the state and the costate. The result is

$$\begin{aligned} x_1(t) &= K, \\ x_2(t) &= \frac{1}{2}K^2 t + K_2, \\ \lambda_1(t) &= -Kt + 2K, \\ \lambda_2(t) &= 1, \end{aligned}$$

for some constants K, K_2 to be determined, and for $t \in [t_A, T]$, where t_A is the unknown switching time.

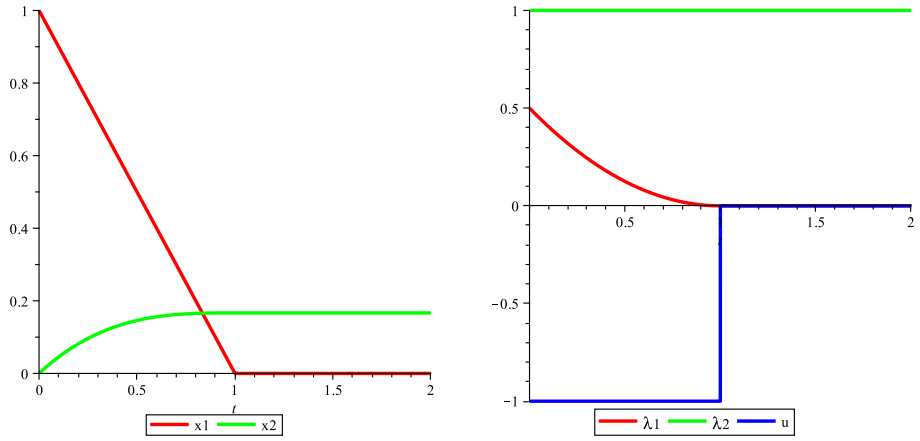


Figure 5.8: Variational solution to problem (5.11). On the left the states, on the right the costates and the control.

In the first arc the control is derived by Pontryagin maximum principle, and is constant $u(t) = -1$ for $t \in [0, t_A)$. Then using the initial conditions the state is solvable, and it results:

$$\begin{aligned} x_1(t) &= -t + 1, \\ x_2(t) &= \frac{1}{6}t^3 - \frac{1}{2}t^2 + \frac{1}{2}t, \\ \lambda_1(t) &= \frac{1}{2}t^2 - t + \lambda_0, \\ \lambda_2(t) &= 1, \end{aligned}$$

for $t \in [0, t_A)$. Using the continuity of the states at the switching point one can set a non linear system in the unknowns K, K_2, λ_0, t_A. That system has two solutions, one which doe not have physical meaning, the other gives $t_A = 1, K = 0, K_2 = 1/6$ and $\lambda_0 = 1/2$. The corresponding $x_2(T) = 1/6$. The plots of the system is shown in Figure 5.8. We report in table 5.5 the results we collected from several authors and summarized by Luus.

5.2.4 *Luus n.2*

The second test case proposed by Luus in [Luu00] is taken from [JGL70] and studied also by [FO77] and [DM95b]. It is formulated as

$$\begin{aligned} &\min x_3(T), && \text{subject to the dynamic:} \\ &x_1(t)' = x_2(t) \\ &x_2(t)' = u(t) \\ &x_3(t)' = x_1(t)^2, && \text{with control bounded by:} \\ &|u| \leq 1. \end{aligned}$$

The given boundary conditions are $x_1(0) = 0$, $x_2(0) = 1$ and $x_3(0) = 0$, T is the final time set to $T = 5$. It is easy to check that the system admits a singular control $u = 0$ of order two, and a reformulation of the problem makes it equivalent to a slightly modified Fuller problem. We do this

Table 5.5: Summary of the results for problem (5.11), in the first column the article with the first author or the name of the algorithm, in the second column the value of the target, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	0.166666666666666667	0
XOptima	0.16666241760617506	-2.5 E-06
ICLOCS	0.16650010203149829	-9.9 E-04
Gpops	0.16668416351901405	1.0 E-04
ACADO	0.16666666672273089	3.3 E-10
PROPT	0.166665695130345510	5.8 E-06
Luus [Luu00]	0.1666667	2.0 E-07
Jacobson [JGL70]	0.1717	3.0 E-02
Chen [CH93]	0.1683	9.7 E-03

transformation to make use of the theory developed in the chapter on singular controls, thus we restate the problem as

$$\begin{aligned} &\min \int_0^5 x^2 \,\mathrm{d}t, && \text{subject to the dynamic:} \\ &x(t)' = y(t) \\ &y(t)' = u(t), && \text{with control bounded by:} \\ &|u| \leq 1. \end{aligned}$$

This is a variant of the Fuller problem (4.15), and because the final conditions are free, we have only a chattering trajectory that *only enters* the singular manifold, and stays there until the final time. The question is if there is enough time to reach the origin, and to ensure that we need to compute the accumulation point (known as *Fuller Point*). The starting point $(x(0), y(0)) = (0, 1)$ is above the switching curve (4.17), i.e. $x = Cy^2$, thus the initial control is $u = -1$. The constant C and k derived for the Fuller problem remain the same, because they are independent of the initial point. We use formulas (4.19) and (4.20) for computing the first two switching points. The results are

$$\begin{aligned} x_1 &= \frac{C}{1+2C} \approx 0.235344310 & x_2 &= \frac{C(2C-1)}{(1+2C)^2} \approx -0.013796532 \\ y_1 &= -\frac{1}{\sqrt{1+2C}} \approx -0.727537889 & y_2 &= \frac{k}{\sqrt{1+2C}} \approx 0.1761524730 \\ t_1 &= 1+\frac{1}{\sqrt{1+2C}} \approx 1.727537889 & t_2 &= 1+\frac{k+2}{\sqrt{1+2C}} \approx 2.631228251 \end{aligned}$$

Setting $\Delta = t_2 - t_1$ we obtain

$$\Delta = \frac{k+1}{\sqrt{1+2C}} \approx 0.9036903627,$$

therefore, from the formula for the total time to reach the origin, we have

$$T = t_1 + \frac{\Delta}{1-k} = 1 - \frac{2}{(k-1)\sqrt{1+2C}} \approx 2.919932465 < 5.$$

The value of $T < 5$ shows that the chattering arcs accumulate before the final time $t = 5$, hence the optimal trajectory stays singular at the origin for $t \geq T$.
We conclude the analysis computing the optimal target value. The integral over the first time interval, e.g. from zero until the first switch is, for $\alpha = 1 + (1+2C)^{-1/2}$,

$$I_0 = \int_0^{t_1} x(t)^2 \, \mathrm{d}t = \int_0^{t_1} \left(\frac{-t}{2} + y_0 t + x_0 \right)^2 \mathrm{d}t = \frac{\alpha^5}{20} - \frac{\alpha^4}{4} + \frac{\alpha^3}{3} \approx 0.2612271922.$$

The integral over the interval $[t_1, t_2]$ has a simple but long analytic expression that here is omitted, it can be approximated to $I_1 \approx 0.007160675$, therefore the target can be evaluated with the relation

$$J = I_0 + \frac{I_1}{1-k^5} \approx 0.2683938305689113.$$

We computed also some suboptimal trajectories, we report here the values obtained with three

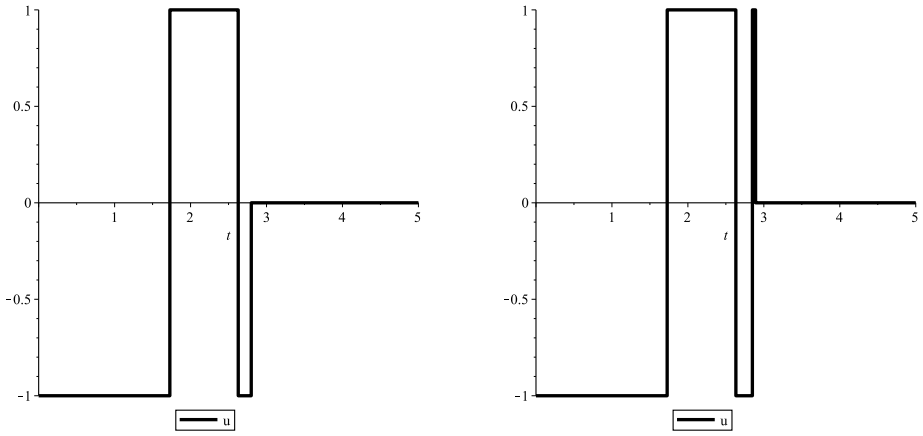


Figure 5.9: Suboptimal non chattering controls with 3 and 4 switches for the problem Luus n.2

and four switches (see Figure 5.9). To solve these problems, we made use of some techniques of computational algebra like the Gröbner Bases, because the resulting NLP polynomial system was composed by more than 30 equations. Denoting with J_3 and J_4 the respective target values, we found

$$J_3 \approx 0.2683941501, \qquad J_4 \approx 0.2683938764.$$

It is clear that those suboptimal solutions are very close to the real optimal value J in fact the error is respectively less than 10^{-5} and 10^{-6}.
We collected here all the numerical values found in literature for this problem, we extended them with the values computed with ACADO, Gpops, ICLOCS, Xoptima and the method proposed in this thesis. The comparison is done with respect to the exact value obtained from the semi-analytical solution. They are summarized in Table 5.6.

Table 5.6: Summary of the results for problem Luus n.2, in the first column the article with the first author or the name of the algorithm, in the second column the value of the target $x_3(T)$, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	0.2683938305689113	0
XOptima	0.268391015569164393	-1.0 E-05
ICLOCS	0.26727731172422453	-4.1E-03
Gpops	0.26840134117823489	2.7 E-05
ACADO	0.26839863859636331	1.7 E-05
PROPT	0.2683360594785408	-2.1 E-04
Luus [Luu00]	0.2683938	-1.13 E-07
Jacobson [JGL70]	0.2771	3.2 E-02
Flaherty [FO77]	0.269	2.2 E-03
Dadebo [DM95b]	0.269	2.2 E-03

5.2.5 ***Luus n.3***

This example is taken from the handbook of PROPT, a commercial software for solving OCPs. It is an example discussed by Luus and other author in several papers, but the exact solution is not given. We give here the semi-analytical solution to it, so we can do a precise comparison among the numerical solutions given by the software proposed for the benchmark. We follow the nomenclature adopted in the handbook of PROPT, where the problem is referred as Singular n3, although it is referred also as double integrator plant and harmonic oscillator.
We have chosen this example because it is studied and considered by many authors, indeed we have found numerical results in bibliography for some software, and although it possesses a semi-analytical solution, it is not trivial. The solution does not appear in papers, so we have derived it in the thesis: it admits an explicit closed form as a combination of cubic polynomials and simple exponential functions. Unfortunately, the coefficients of the equations depend on a linear combination of the only admissible root of a nonlinear function, which can not be solved by radicals or elementary functions. Hence the global solution is analytic except for one constant, the switching time of the control, that has to be computed numerically at arbitrary precision. Because in literature only single precision values are reported, and the software give double precision values, we computed the exact solution to 20 digits.
We compare the results obtained with the principal open source optimal control softwares, we choose one for each family of methods: ACADO is a software based on the multiple shooting algorithm, Gpops and ICLOCS employ the pseudospectral techniques of direct methods, Xoptima uses indirect variational methods, and the method presented in this PhD thesis. Moreover we show the results obtained by Luus with the Iterative Dynamic Programming (IDP), the results available on the handbook of PROPT, the results given by the authors that have been published on papers. We found the contribution of Jacobson, Gershwin, Stanley, Lele [JGL70], Flaherty, O'Malley [FO77], Dadebo, Mcauley [DM95b], Luus [Luu00].

5.2.5.1 *Problem Statement and Solution*

The optimal control problem Singular n3 has the following formulation.

$$
\begin{aligned}
&\min x_3(T), && \text{subject to the dynamic:}\\
&x_1(t)' = x_2(t) &&\\
&x_2(t)' = u(t) && \qquad (5.12)\\
&x_3(t)' = x_1(t)^2 + x_2(t)^2, && \text{with control bounded by:}\\
&|u| \leq 1. &&
\end{aligned}
$$

The given boundary conditions are $x_1(0) = 0$, $x_2(0) = 1$ and $x_3(0) = 0$, T is the final time set to $T = 5$.

5.2.5.2 *Preliminary Considerations*

Before to attack the problem from an analytical point of view, it is convenient to check if it admits a solution, and if that solution is unique and a minimum. At a first glance, problem (5.12) looks nonlinear, because of the nonlinear differential equation for x_3. However, it is possible to rewrite the formulation in order to avoid the nonlinearity. This is done easily by converting the Mayer problem into a Lagrange problem, in fact we can exploit the fact that $x_3(0) = 0$, therefore the target can be written

$$x_3(T) = x_3(T) - x_3(0) = \int_0^T x_3'(t)\,\mathrm{d}t = \int_0^T x_1(t)^2 + x_2(t)^2\,\mathrm{d}t.$$

With this formulation, we can apply standard theorems of existence of the optimal control.
Moreover, from the convexity of the problem, the solution is also unique and is a minimum. To see this, suppose that $\boldsymbol{x} = (x_1, x_2, x_3)^T$ is an optimal solution of the original problem and $\boldsymbol{y} \neq \boldsymbol{x}$ is another solution such that $y_3 \geq x_3$, let us see that in facts $y_3 > x_3$. We can build a family $\boldsymbol{z}$ of intermediate solutions by posing, for $0 < \lambda < 1$, $\boldsymbol{z}(t) = \lambda \boldsymbol{x}(t) + (1-\lambda)\boldsymbol{y}(t)$,

$$\begin{aligned} z_1(t) &= \lambda x_1(t) + (1-\lambda) y_1(t) \\ z_2(t) &= \lambda x_2(t) + (1-\lambda) y_2(t) \\ z_3(t) &= \lambda x_3(t) + (1-\lambda) y_3(t). \end{aligned}$$

It is possible to rewrite $z_3'(t)$ as

$$\begin{aligned} z_3' &= z_1^2 + z_2^2 \\ &= \lambda x_3' + (1-\lambda) y_3' \\ &= [\lambda x_1 + (1-\lambda) y_1]^2 + [\lambda x_2 + (1-\lambda) y_2]^2 \\ &= \lambda x_3' + (1-\lambda) y_3' - \lambda(1-\lambda)[(x_1 - y_1)^2 + (x_2 - y_2)^2], \end{aligned}$$

then, integrating both hand sides yields,

$$z_3(T) = \lambda x_3(T) + (1-\lambda) y_3(T) - \lambda(1-\lambda) \int_0^T [(x_1 - y_1)^2 + (x_2 - y_2)^2] \, \mathrm{d}t$$

and, since by hypothesis, $\boldsymbol{x}$ is the optimal result, $z_3(T) \geq x_3(T)$, we have that

$$y_3(T) \geq x_3(T) + \lambda \int_0^T [(x_1 - y_1)^2 + (x_2 - y_2)^2] \, \mathrm{d}t > x_3(T).$$

Thus the problem has one and only one optimal solution, which is the only one given by the generalization of the PMP.

5.2.5.3 *Semi-analytical Solution*

We have seen in the previous section that the problem admits unique solution, so we compute it using standard variational techniques. It have been shown in several papers, that the optimal control consists in a bang-bang arc followed by a (non trivial) singular arc, in particular there is not the chattering phenomenon because the order of the singular arc is 1. The optimal control starts at -1 until a switching time t_A, then it becomes singular and can not be synthesized with the theorem of Pontryagin.
The Hamiltonian for this problem is

$$\mathcal{H} = \lambda_1 x_2 + \lambda_2 u + \lambda_3 (x_1^2 + x_2^2). \tag{5.13}$$

The equation of the costate are derived from the Hamiltonian, so that $\lambda_1(T) = \lambda_2(T) = 0$ and $\lambda_3(T) = 1$,

$$\begin{aligned} \lambda_1' &= -\frac{\partial \mathcal{H}}{\partial x_1} = -2\lambda_3 x_1, \\ \lambda_2' &= -\frac{\partial \mathcal{H}}{\partial x_2} = -\lambda_1 - 2\lambda_3 x_2, \\ \lambda_3' &= -\frac{\partial \mathcal{H}}{\partial x_3} = 0. \end{aligned}$$

Therefore we have $\lambda_3(t) = 1$ constant. Since there is only one switching point, and because the singular arc is the second, we derive the multipliers and the control in the two configurations.
In the first segment the control is bang-bang and $u = -1$, therefore $x_2'(t) = -1$ and $x_2(t) = -t + \alpha_2$. The constant α_2 is obtained from $x_2(0) = 1$, that is $\alpha_2 = 1$. From the other differential equation, $x_1(t) = -\frac{1}{2}t^2 + \alpha_2 t + \alpha_1 = -\frac{1}{2}t^2 + t$. Similarly, we can conclude that $\alpha_3 = 0$ and $x_3'(t) = x_1(t)^2 + x_2(t)^2 = t^2 - 2t + 1 + \frac{1}{4}t^4 + t^2 - t^3$, hence before the first switching t_A:

$$\begin{aligned} x_1(t) &= -\frac{1}{2}t^2 + t \\ x_2(t) &= -t + 1 \\ x_3(t) &= \frac{1}{20}t^5 - \frac{1}{4}t^4 + \frac{2}{3}t^3 - t^2 + t \qquad 0 \le t \le t_A. \end{aligned}$$

The corresponding multipliers are, for some constants ℓ_1, ℓ_2,

$$\begin{aligned} \lambda_1(t) &= \frac{1}{3}t^3 - t^2 + \ell_1 \\ \lambda_2(t) &= -\frac{1}{12}t^4 + \frac{1}{3}t^3 - t^2 - (2 + \ell_1)t + \ell_2 \\ \lambda_3(t) &= 1 \qquad 0 \le t \le t_A. \end{aligned}$$

Now we have to write the singular part of the optimal control, first we consider the Hamiltonian (5.13) in canonical form $\mathcal{H} = h_0(t) + h_1(t)u(t)$, where the *switching function* is $h_1(t) = \lambda_2(t)$. During the singular tract, $h_1(t) = 0$, so taking its derivatives we obtain:

$$\begin{aligned} \lambda_2(t) &= 0 \\ \lambda_2(t)' &= -\lambda_1(t) - 2x_2(t) \\ \lambda_2(t)'' &= -\lambda_1(t)' - 2x_2(t)' = 2x_1(t) - 2u(t). \end{aligned}$$

From the last equation we can solve the singular control, which for $t_A \le t \le 5$ is equal to $u(t) = x_1(t)$. It is now possible to solve the differential system and obtain explicit expression for the states and the control, as well as the multipliers. We have,

$$\begin{aligned} x_1(t) &= -\frac{1}{4}(t_A^2 - 2)e^{t - t_A} - \frac{1}{4}(t_A^2 - 4t_A + 2)e^{-t + t_A} \\ x_2(t) &= -\frac{1}{4}(t_A^2 - 2)e^{t - t_A} + \frac{1}{4}(t_A^2 - 4t_A + 2)e^{-t + t_A} \\ x_3(t) &= \frac{1}{16}(t_A^2 - 2)^2 e^{2(t - t_A)} + \frac{1}{16}(t_A^2 - 4t_A + 2)^2 e^{2(-t + t_A)} + \\ &\quad + \frac{1}{20}t_A^5 - \frac{1}{4}t_A^4 + \frac{1}{6}t_A^3 + \frac{1}{2}t_A^2 \qquad t_A \le t \le 5. \end{aligned}$$

The multipliers are

$$\begin{aligned} \lambda_1(t) &= \frac{1}{6}(3t_A^2 - 6)e^{t - t_A} + \frac{1}{6}(-3t_A^2 + 12t_A - 6)e^{-t + t_A} + \\ &\quad + \frac{1}{3}t_A^3 - t_A^2 - 2t_A + 2 + \ell_1 \\ \lambda_2(t) &= -\left(\frac{1}{3}t_A^3 - t_A^2 - 2t_A + 2 + \ell_1\right)t + \frac{1}{4}t_A^4 - \frac{2}{3}t_A^3 - t_A^2 + \ell_2 = 0 \\ \lambda_3(t) &= 1 \qquad t_A \le t \le 5. \end{aligned}$$

To determine the three constants t_A, ℓ_1, ℓ_2 we have to impose a nonlinear problem by imposing the end conditions $\lambda_1(T) = 0, \lambda_2(T) = 0, \lambda_2(t_A) = 0$. This system is linear in ℓ_1, ℓ_2, but nonlinear in t_A, namely we have

$$\begin{aligned}
\ell_1 &= -\frac{1}{3}t_A^3 + t_A^2 + 2t_A - 2 \approx 1.8843466929567696441 \\
\ell_2 &= -\frac{1}{4}t_A^4 + \frac{2}{3}t_A^3 + t_A^2 + 2t_A \approx 2.8838203516249728328 \\
t_A &\implies \frac{1}{2}(t_A^2 - 2)e^{-t_A+5} - \frac{1}{2}(t_A^2 - 4t_A + 2)e^{t_A-5} = 0 \\
t_A &\approx 1.4137640876300641592.
\end{aligned}$$

As showed in figure 5.10, the last nonlinear equation has only one real root in the interval $[0, 5]$, hence there is no ambiguity in selecting the correct root.

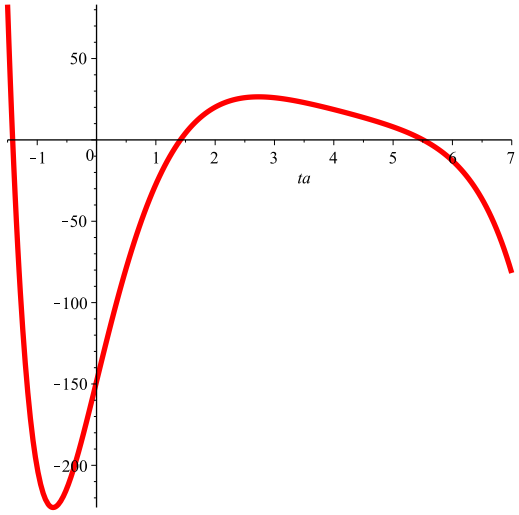


Figure 5.10: The nonlinear function for t_A possesses only one real root in the interval $[0, 5]$

5.2.5.4 *Numerical Results and Comparison*

We collected here all the numerical values found in literature for this problem, we extended them with the values computed with ACADO, Gpops, ICLOCS, Xoptima and the method proposed in this thesis. The comparison is done with respect to the exact value obtained from the semi-analytical solution. They are summarized in Table 5.7.

In Figure 5.12 there is the plot of the error in logarithmic scale.

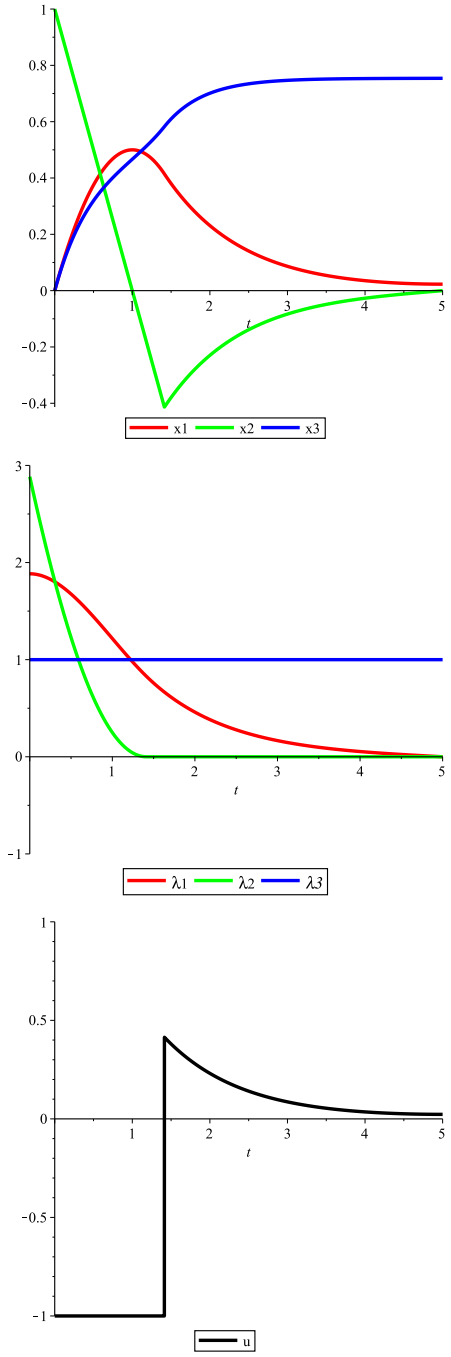


Figure 5.11: Variational solution to Singular Problem n3. From the top: states, costates and control.

Table 5.7: Summary of the results for problem Singular n3, in the first column the article with the first author or the name of the algorithm, in the second column the value of the target $x_3(T)$, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	0.75398386057588920820	0
Present method	0.753990154	8.35E-6
XOptima	0.75398389193771053751	4.16E-8
ICLOCS	0.75158391763498122	3.18E-4
Gpops	0.75398439909761550	7.14E-7
ACADO	0.75398395894495096	1.30E-7
PROPT	0.75399456159009870	1.41E-5
Luus [Luu00]	0.7539839	5.22E-8
Jacobson [JGL70]	0.828514	9.88E-2
Flaherty [FO77]	0.758	5.32E-3
Dadebo [DM95b]	0.754016	4.26E-5

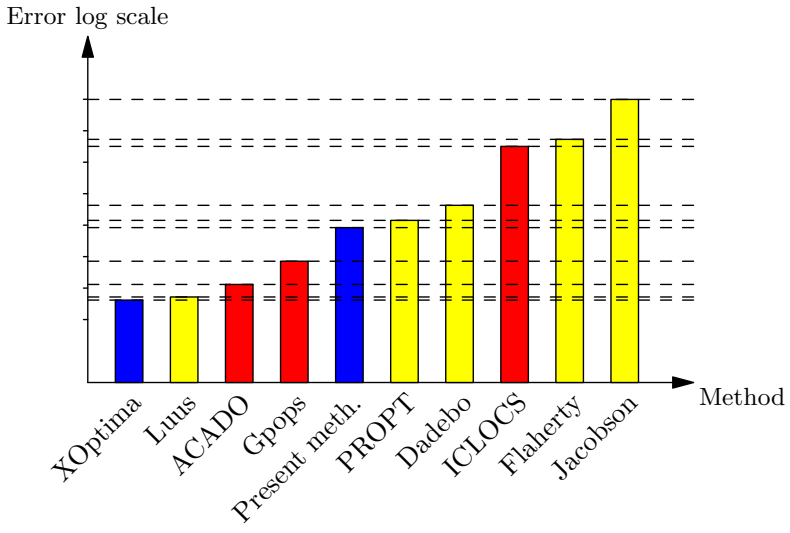


Figure 5.12: The bar plot for the errors reported in table 5.7. In blue the methods presented in present thesis, in yellow the values presented in the cited papers, in red the values computed by us.

5.2.6 ***Fuller-Marchal***

This problem is a variant of the Fuller problem we discussed in the chapter of OCP linear in the control. It was proposed and analysed in Marchal [Mar73], and shows the case of chattering entrance and escape of the control from a singular arc. The original version has the following statement,

$$\min J = \int_0^8 x(t)^2 \, \mathrm{d}t, \qquad x' = y, \quad y' = u, \qquad |u| \leq 1,$$

with boundary conditions given by $x(0) = x(8) = 2$ and $y(0) = -2$, $y(8) = 2$. It can be understood as two Fuller problems (see Figure 5.13), with the second in the reversed time. Because of this symmetry we can use all the computations done for the Fuller problem with some care: the singular arc $u = 0$ begins for $t = T_1 \approx 3.43$ and the control stays singular until $t = T_2 \approx 8 - T_1$. As a consequence of this symmetry the target value is doubled and the optimal target is $J \approx 3.030456$. These values are in accord with the paper of Marchal and the results our section on chattering control. The corresponding optimal trajectories can be obtained by mirroring along the axis $t = 4$ the trajectories of the Fuller problem (with $u = 0$ after the point of accumulation of the switching points).

5.2.6.1 *Numerical Results and Comparison*

We compare the results given by XOptima, Gpops, Iclocs and Acado .

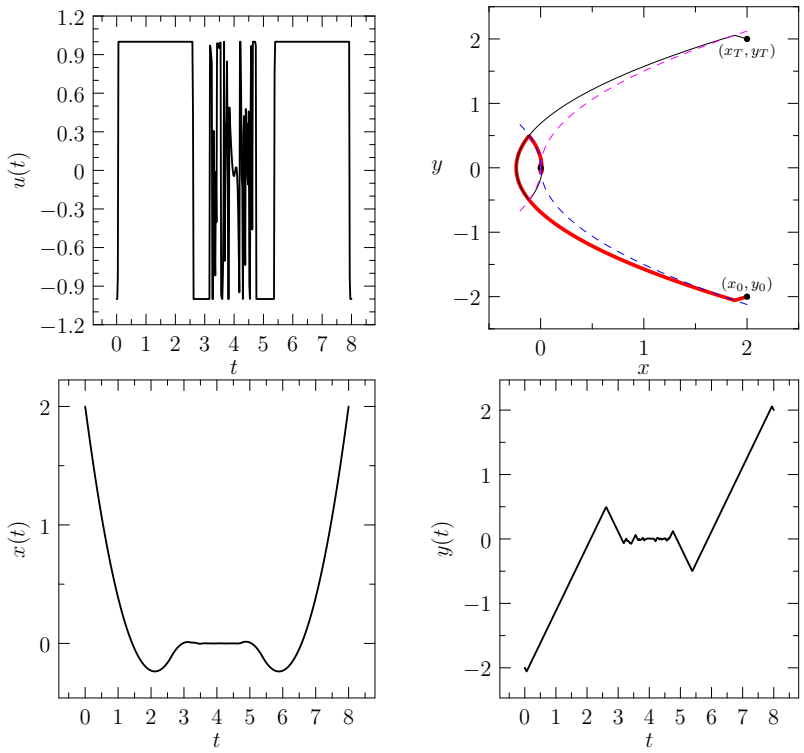


Figure 5.13: The plots of the results obtained with Acado, from the top: the control, the trajectory in the state space, the states $x(t)$ and $y(t)$.

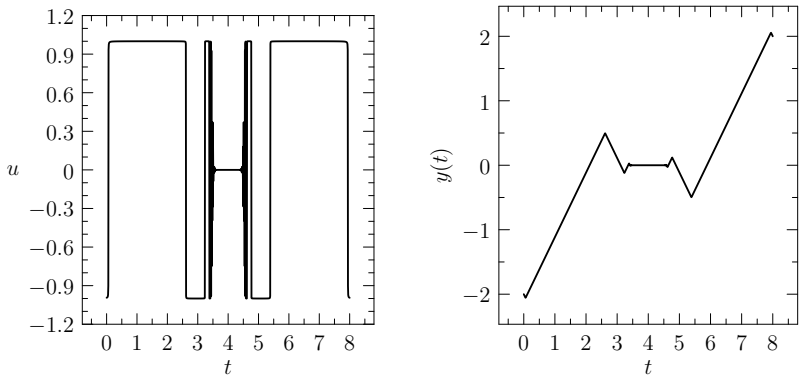


Figure 5.14: The plots of the control u and the velocity y obtained with Xoptima.

Table 5.8: Summary of the results for problem Fuller-Marchal, in the first column the article with the first author or the name of the algorithm, in the second column the value of the target $x_3(T)$, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	3.0304563877738555	0
XOptima	3.0305812059484234	4.1E-5
ICLOCS	3.0304696190904252	4.3E-6
Gpops	3.0304906866820898	1.1E-5
ACADO	3.0305914050027605	4.4E-5
Marchal [Mar73]	3.03046	1.1E-6

5.2.7 *Economic Growth*

We consider here the resource allocation problem proposed in [ZB94, Bor00] reduced to the formalization

$$\min T \qquad \text{s.t.} \qquad x_1' = u_1 x_1 x_2, \quad x_2' = u_2 x_1 x_2$$

with boundary conditions

$$x_1(0) = x_{10} = 1, \qquad x_2(0) = x_{20} = 2, \qquad (\boldsymbol{x}(T), \boldsymbol{y}(T)) \in M,$$

where M is a prescribed smooth manifold:

$$M := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 x_2 = c = 10\}$$

The problem is further constrained with

$$u_1 \geq 0, \quad u_2 \geq 0, \quad u_1 + u_2 = 1, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

First we notice that the control $\boldsymbol{u} = (u_1, u_2)$ can be reduced to a scalar variable posing $u_2 = 1 - u_1$, hence the $u_1 := u \in [0, 1]$. We can the form the Hamiltonian of the problem,

$$\begin{aligned} \mathcal{H} &= 1 + \lambda_1 u x_1 x_2 + \lambda_2 (1 - u) x_1 x_2 \\ &= 1 + \lambda_2 x_1 x_2 + u x_1 x_2 (\lambda_1 - \lambda_2) \\ &= H_0 + u H_1 = 0, \end{aligned}$$

the adjoint equations become

$$\begin{aligned} \lambda_1' &= -x_2(u(\lambda_1 - \lambda_2) + \lambda_2) \\ \lambda_2' &= -x_1(u(\lambda_1 - \lambda_2) + \lambda_2) \\ u &= -\operatorname{sign}(H_1), \quad H_1 \neq 0. \end{aligned}$$

If the switching function H_1 vanishes, the singular control must be determined taking the poisson bracket of H_1. We have that the singular control has essential order 1, in facts, after simplifying from H_1 the term $x_1 x_2 \neq 0$,

$$\begin{aligned} H_1 &= \lambda_1 - \lambda_2 \\ H_1' &= \lambda_2 (x_1 - x_2) \\ H_1'' &= -\lambda_2 x_1^2 + x_1 x_2 (\lambda_1 + \lambda_2) u. \end{aligned}$$

Thus the singular control is given by

$$u = \frac{\lambda_2 x_1}{x_2(\lambda_1 + \lambda_2)} = \frac{-\lambda_2(x_1 - x_2) + \lambda_2 x_2}{x_2(\lambda_1 + \lambda_2)} = \frac{\lambda_2}{\lambda_1 + \lambda_2},$$

whereas the last equality can be written also as $u = 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2}$, therefore adding the two quantities yields

$$2u = 1 - \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} = 1 \implies u = \frac{1}{2}.$$

In the case of a singular arc the differential system (with $u = \frac{1}{2}$) takes the form of

$$\begin{aligned} x_1' &= \frac{1}{2}x_1x_2 \qquad & x_2' &= \frac{1}{2}x_1x_2 \\ \lambda_1' &= -\lambda_2 x_2 \qquad & \lambda_2' &= -\lambda_2 x_1. \end{aligned} \tag{5.14}$$

The optimal synthesis has the control $u = 1$ for $t \in [0, t_0]$ and $u = \frac{1}{2}$ for $t \in [t_0, T]$, so on the first arc the integration of the system gives

$$\begin{aligned} x_1(t) &= x_{10}e^{x_{20}t} \qquad & x_{10} &= x_1(0) = 1 \\ x_2(t) &= x_{20} \qquad & x_{20} &= x_2(0) = 2 \\ \lambda_1(t) &= \lambda_{10}e^{-x_{20}t} \qquad & \lambda_{10} &= \lambda_1(0) \\ \lambda_2(t) &= -x_{10}\lambda_{10}t + \lambda_{20} \qquad & \lambda_{20} &= \lambda_2(0). \end{aligned} \tag{5.15}$$

On the singular arc, the integration of the differential system (5.14) is a little more tricky, first we notice that $x_1' = x_2'$, thus we can argue that $x_2(t) = x_1(t) + \Delta$ for a constant Δ and $t \in [t_0, T]$. We obtain Δ imposing continuity of the states x_1, x_2 at the junction point t_0, therefore

$$\Delta = x_2(t_0^-) - x_1(t_0^-) = x_{20} - x_{10}e^{x_{20}t_0} = 2 - e^{2t_0}.$$

We solve the differential equation of x_1, which is $x_1' = \frac{1}{2}x_1(x_1 + \Delta)$, with the method of separation of variables, and obtain

$$\begin{aligned} x_1(t) &= \frac{\Delta(x_{20} - \Delta)e^{-\frac{1}{2}\Delta t_0}}{x_{20}e^{-\frac{1}{2}\Delta t} - (x_{20} - \Delta)e^{-\frac{1}{2}\Delta t_0}} \\ x_2(t) &= x_1(t) + \Delta. \end{aligned}$$

Remark 5.1. *We notice that this solution requires $\Delta \neq 0$. The case with $\Delta = 0$ is simpler and yields*

$$x_1(t) = x_2(t) = \frac{2x_{20}}{2 + x_{20}t_0 - x_{20}t}.$$

We continue assuming $\Delta \neq 0$, evaluating the Hamiltonian in $t = t_0^-$ we can determine λ_{10},

$$\mathcal{H}(t_0^-) = 1 + \lambda_1(t_0^-)x_1(t_0^-)x_2(t_0^-) = 1 + \lambda_{10}e^{-x_{20}t_0}x_{10}e^{x_{20}t_0}x_{20} = 0,$$

thus $\lambda_{10} = -\frac{1}{x_{10}x_{20}} = -\frac{1}{2}$.
At $t = t_0^-$, the first condition of singularity must hold, i.e. $H_1 = x_1x_2(\lambda_1 - \lambda_2) = 0$, substituting the known values we get a relation for λ_{20}:

$$x_1(t_0^-)x_2(t_0^-)(\lambda_1(t_0^-) - \lambda_2(t_0^-)) = x_{10}x_{20}e^{x_{20}t_0}(\lambda_{10}e^{-x_{20}t_0} + \lambda_{10}x_{10}t_0 - \lambda_{20}) = 0,$$

that is, $\lambda_{20} = \lambda_{10}(e^{-x_{20}t_0} + x_{10}t_0) = -\frac{1}{2}(e^{-2t_0} + t_0)$.
We need now to introduce the singular part of the problem, it is possbile to solve the differential equation for λ_2 and then impose the second singularity condition $H_1' = 0$, the result of the computation requires the previous assumption of $\Delta = 0$, and is

$$H_1' = \lambda_2(x_1 - x_2) = \Delta(\lambda_{10}x_{10}t_0 - \lambda_{20}) = 0 \implies \lambda_{10}x_{10}t_0 - \lambda_{20} = 0. \tag{5.16}$$

We make use of the previously retrieved information for $\lambda_{10}, \lambda_{20}$ to simplify the above expression, but we arrive to an absurd:

$$\lambda_{10}x_{10}t_0 - \lambda_{20} = \frac{e^{-x_{20}t_0}}{x_{10}x_{20}} \neq 0,$$

and it is not possible to find a real t_0 that satisfies that equation, hence the assumption $\Delta \neq 0$ is wrong and we must restart from Remark 5.1 with $\Delta = 0$. The above relations for $\lambda_{10}, \lambda_{20}$ are not affected by this change, only the differential equation for λ_2 must be recomputed with the new expression for x_1, x_2 given in the Remark. We have that for $t \in [t_0, T]$

$$\lambda_2(t) = \frac{1}{4}\lambda_2(t_0^-)[x_{20}t - t_0 x_{20} - 2]^2, \qquad \lambda_2(t_0^-) = -x_{10}\lambda_{10}t_0 + \lambda_{20},$$

but equation (5.16) is trivially satisfied, so we use the value of the Hamiltonian in $t = t_0^+$

$$\mathcal{H}(t_0^+) = \lambda_2(t_0^+)x_1(t_0^+)x_2(t_0^+) = \left(\frac{t_0}{x_{20}} - \frac{x_{10}t_0 + e^{-x_{20}t_0}}{x_{10}x_{20}}\right)x_{20}^2 + 1 = 0,$$

which gives $t_0 = -\frac{\ln(x_{10}/x_{20})}{x_{20}} = -\frac{\ln 1/2}{2}$.
Finally, when we touch the manifold M we must impose $x_1(T)x_2(T) = c = 10$, that is

$$\frac{4x_{20}^2}{2 - (T - t_0)x_{20}} = c \implies T = \frac{cx_{20}t_0 + 2c \pm 2x_{20}\sqrt{c}}{cx_{20}},$$

The two values of T are both positive and equal $T = 1 + \frac{\ln 2}{2} \pm \frac{\sqrt{10}}{5}$, we take the smallest root $T \approx 0.7141180583$.
In conclusion, the optimal trajectory for this economic growth model is made up of two arcs, the first characterized by $u = 1$ for $t \in [0, t_0]$ and described by equations (5.15), the second is a singular arc with $u = \frac{1}{2}$ and for $t \in [t_0, T]$,

$$\begin{aligned} x_1(t) = x_2(t) &= \frac{2x_{20}}{2 + x_{20}t_0 - x_{20}t} \\ \lambda_1(t) = \lambda_2(t) &= -\frac{1}{4}(\lambda_{20} - x_{10}\lambda_{10}t_0)[x_{20}t - t_0x_{20} - 2]^2. \end{aligned}$$

The various constants are

$$\begin{aligned} \lambda_{10} &= -\frac{1}{x_{10}x_{20}} && = -\frac{1}{2} && \\ \lambda_{20} &= \lambda_{10}(e^{-x_{20}t_0} + x_{10}t_0) && = -\frac{1}{4}(1 + \ln 2) && \approx -0.4232867952 \\ t_0 &= -\frac{\ln(x_{10}/x_{20})}{x_{20}} && = \frac{\ln 2}{2} && \approx 0.3465735903 \\ T &= \frac{cx_{20}t_0 + 2c - 2x_{20}\sqrt{c}}{cx_{20}} && = 1 + \frac{\ln 2}{2} - \frac{\sqrt{10}}{5} && \approx 0.7141180583 \end{aligned}$$

The graph of the trajectory is showed in Figure 5.15.

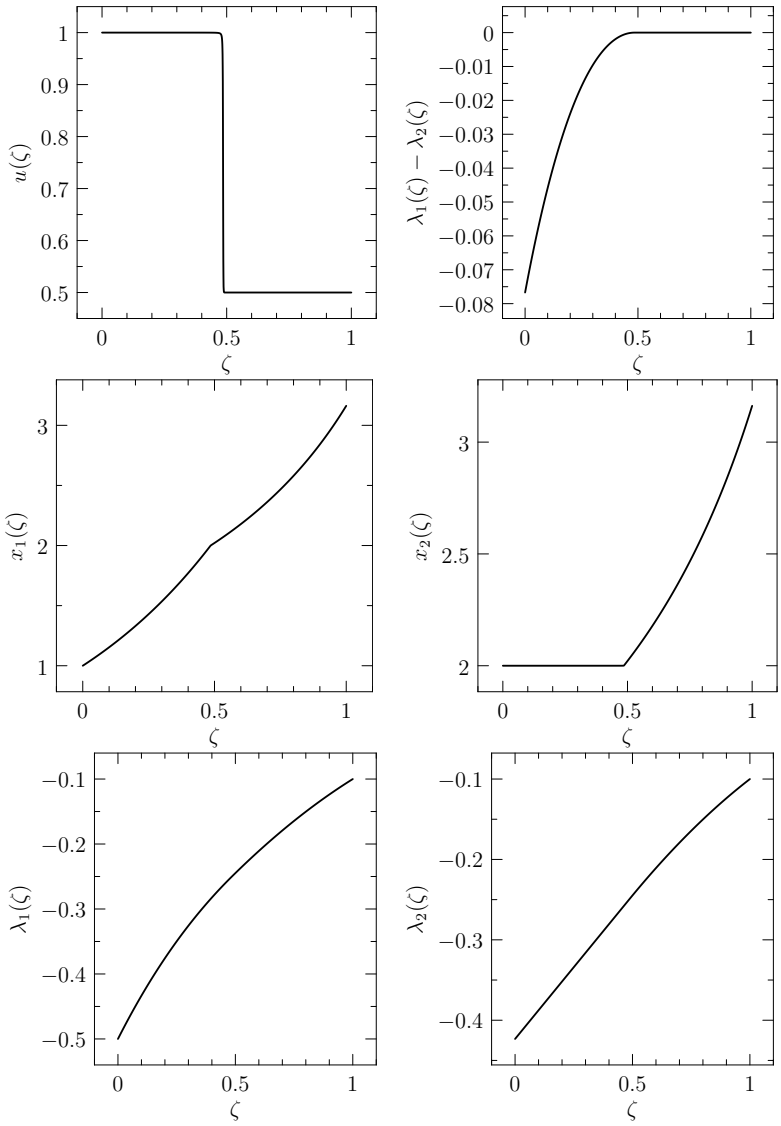


Figure 5.15: Results for the Resource Allocation problem. From the top: the control u and the switching function $\lambda_1 - \lambda_2$; the positions x_1 and x_2; the multipliers λ_1 and λ_2.

Table 5.9: Summary of the results for problem Economic growth, in the first column the name of the algorithm, in the second column the value of the target T, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	0.71411805824629678	0
XOptima	0.71411800841994765	-6.9E-8
ICLOCS	0.71405437822992757	8.9E-5
Gpops	NC	
ACADO	0.71411806970055869	-1.6E-8

5.3 CONSTRAINED PROBLEMS

5.3.1 *Constrained Car*

In this section we analyse the model of a one dimensional vehicle as described in [BBDL$^+$14]. The problem describes the longitudinal dynamic of the car that has a limited amount of braking force and acceleration. The control variable is the jerk and we will try two kinds of control, a linear one on the jerk, a quadratic on the jerk. The first give rise to singular and bang-bang arcs, the second one has a smooth control. The optimal control problem can be stated as

$$\min \int_0^T 1 + wJ^2 \, \mathrm{d}t$$

where w is set to zero in the first case, and $w = 0.1$ in the second case, subject to the following dynamic and path constraints:

$$\begin{aligned} s' &= v \\ v' &= a(p) \\ p' &= J \end{aligned}$$

with boundary conditions:

$$\begin{aligned} s(0) &= 0, & s(T) &= 200 \\ v(0) &= 10, & v(T) &= 10 \\ p(0) &= 0, & p(T) &= \text{free}. \end{aligned}$$

The path constraint is

$$a_{\min} - a(p) \leq 0, \qquad a(p) = \begin{cases} a_1 p & p \geq 0, \\ a_2 p & p < 0. \end{cases}$$

where the constants are $a_{\min} = -5$, $a_1 = 3$ and $a_2 = 10$. The state p is constrained in $[-1, 1]$ and the control $J \in [-1, 1]$.

5.3.1.1 *Jerk as linear control*

When setting $w = 0$ the problem becomes linear in the control and we expect a bang bang solution. The analytical solution requires to solve a square non linear system in 29 equations. We have 5 arcs with 4 switches, at instants:

$$\begin{aligned} t_A &= 1.0000000000000000000 & t_B &= 5.8209379730184301176 \\ t_C &= 6.8209379730184301176 & t_D &= 7.3209379730184301176. \end{aligned}$$

The final time is $T = 10.563500756829488188$.

5.3.1.2 *Jerk as quadratic control*

When we set $w \neq 0$, the control has a (small) quadratic component that alter the bang bang nature of the problem, this results in a smoother control. In practise, the solution of this problem is close to the previous solution with rounded corners at the switching times. However, the new problem has a higher number of arcs, so the size of the resulting NLP increases. In facts we end up with a nonlinear system of 39 equations, after simplifying the trivial substitutions and the easiest

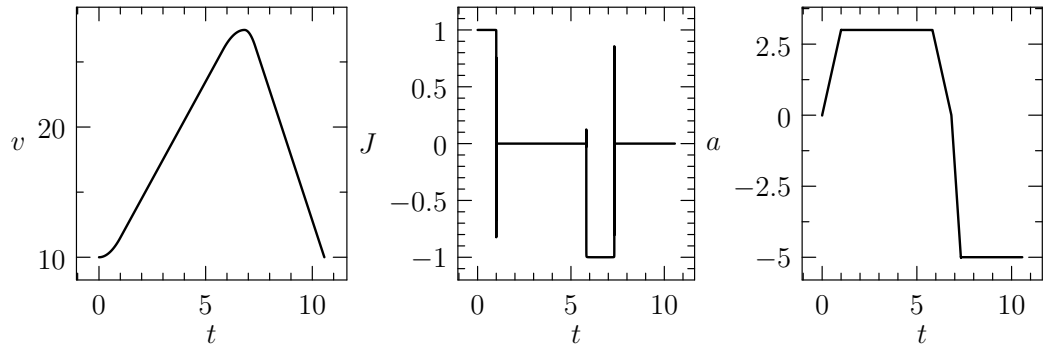


Figure 5.16: The numerical result for the case with $w = 0$ obtained with XOptima, from the left, the state v, the control J and the acceleration $a(p(t))$. The minimum time obtained is 10.56491.

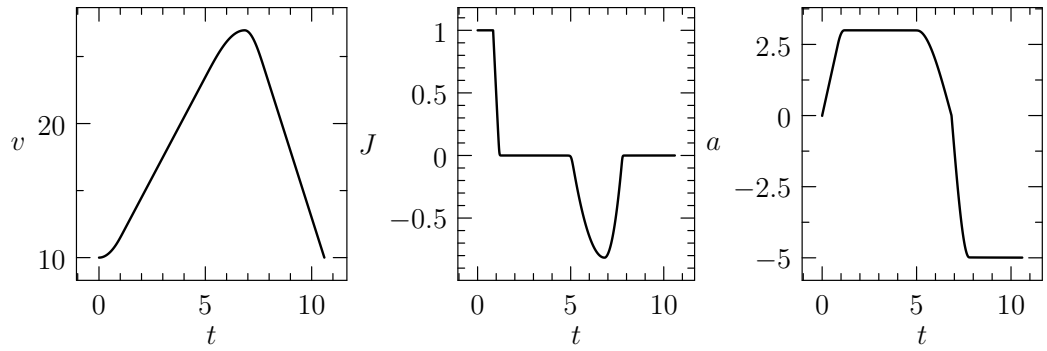


Figure 5.17: The numerical result for the case with $w = 0.1$ obtained with XOptima, from the left, the state v, the control J and the acceleration $a(p(t))$. The minimum time obtained is 10.59642.

equations, we could reduce the NLP to a polynomial system of 8 equations in 8 unknowns. This NLP consisted in polynomials over $\mathbb{Q}$ of degree 2,3,4,5 and could be solved numerically with high difficulty, because there were dozens of spurious solutions that did not satisfy the constraints of the original OCP. The time instants where the different arcs join are given by

$$t_A = 0.83562085921040736732395 \qquad t_B = 1.16754114382857409594695$$
$$t_C = 5.03270079920302976591860 \qquad t_D = 6.84697148865936836721862$$
$$t_E = 7.78184612489142465651121 \qquad T = 10.59102376306397182309365.$$

After the solution of the NLP, the analytic integration of the target and the evaluation at $t = T$ gave the exact value of the objective function, namely 10.78370428467098987654696.

5.3.2 *A Singular Constrained Problem*

The following problem involves a linear control which is bounded, and a constraint on both state and control. The formulation ([Calb, Cala]) is

$$\begin{aligned} &\min \int_0^3 (t-4)u\,\mathrm{d}t, && \text{subject to the dynamic:}\\ &x' = u,\ x(0) = 0,\ x(3) = 3 && \text{constrained by:}\\ &0 \le u \le 2,\\ &g(t,x,u) = x - t - u + 1 \le 0. \end{aligned}$$

The Hamiltonian of the problem is

$$\mathcal{H} = (t-4)u + \lambda u + \mu(x - t - u + 1) = (t - 4 + \lambda - \mu)u + \mu(x - t + 1).$$

For $t = 0$ we see that the constraint g is satisfied only if $g(0) = 1 - u(0) \le 0$, that is, the initial control must be $u(0) \ge 1$. From the Hamiltonian, we derive the equations needed to form the boundary value problem:

$$\frac{\partial \mathcal{H}}{\partial u} = t - 4 + \lambda - \mu = 0, \qquad \lambda' = -\frac{\partial \mathcal{H}}{\partial x} = -\mu, \qquad \mu \ge 0.$$

The multiplier of the constraint μ should be non negative when the bound is active (because then $g = 0$), and $\mu = 0$ when the bound is not sharp. Suppose that the constraint is inactive, thus $\mu = 0$, the boundary value problem gives $\lambda = c$ for a constant c from the equation $\frac{\partial \mathcal{H}}{\partial x}$, but gives $\lambda = 4 - t$ from the equation $\frac{\partial \mathcal{H}}{\partial u}$. Therefore there can not be an arc where the constraint is inactive, hence on $[0, 3]$ the constraint is sharp. Combining the two equations $\lambda' = -\mu$ and $t - 4 + \lambda - \mu = 0$ we obtain the differential equation $t - 4 + \lambda + \lambda' = 0$. Its solution is $\lambda(t) = \alpha e^{-t} - t + 5$ for an unknown constant α. From the multiplier we can get $\mu = -\lambda' = \alpha e^{-t} + 1$ which is non negative on $[0, 3]$ for $\alpha \ge e^3 \approx 20.08553692$. The control is given differentiating $g = 0$ with the combination of $x' = u$, that is $x' - 1 - u' = u - u' - 1 = 0$. The initial value of the control $u(0) = 1$ allows to solve the differential equation yielding $u(t) = 1$. Now from $g = 0$ we obtain $x(t) = t$. The presence of a minimum is given by the application of the Weierstrass condition $\mathcal{H}(u) - \mathcal{H}(u^\star) > 0$, where $u^\star$ is the optimal control and $u > 1$ is another admissible control. Observing that $\mu g = 0$ the Weierstrass condition must be checked only for $(t - 4 + \lambda)(u - u^\star) > 0$. Notice that $(t - 4 + \lambda)$ is strictly positive for $\alpha \ge e^3$ and $u > u^\star = 1$, see Figure 5.18. The value of the target is $-\frac{15}{2}$. Various things happens with this example: Gpops, although converges to the correct solution $u = 1$, $x = t$, but gives a completely wrong target (exactly zero); Iclocs and XOptima are influenced by the contribution of the bound; Acado converges to machine precision without iterations. The numerical results are in table 5.10.

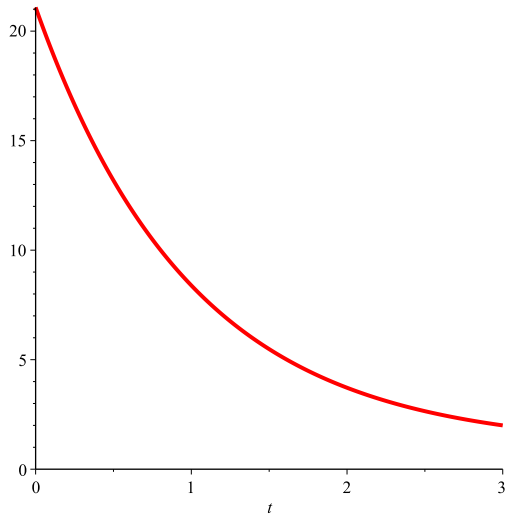


Figure 5.18: The plot of $t - 4 + \lambda$.

Table 5.10: Summary of the results for Singular constrained problem, in the first column the name of the algorithm, in the second column the value of the target T, in the third column the ratio with the exact value.

Method/Author	Reported value	Error
Exact value	-7.5000000000000000	0
XOptima	-7.49992004398472911	1.0E-05
ICLOCS	-7.5013166543897905	-1.7E-04
Gpops	NC	
ACADO	-7.5000000000000666	-8.8E-15

5.4 HARD PROBLEMS

5.4.1 *Hang Glider*

This problem was posed by Bulirsch et al [BNPS91] but we consider here the slightly modified version proposed by J.T.Betts in [Bet01]. It requires to maximize the maximum distance that a hang glider can travel in presence of a thermal updraft. The difficulty in solving this problem is the sensitivity to the accuracy of the mesh. Both references exploit a combination of direct and indirect methods with some *ad hoc* tricks in order to obtain convergence of the solver and the solution. The formulation and the constants defined in [Bet01] for the hang glider problem are the following. The dynamical system is

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}x(t) &= v_x(t) \\
\frac{\mathrm{d}}{\mathrm{d}t}y(t) &= v_y(t) \\
\frac{\mathrm{d}}{\mathrm{d}t}v_x(t) &= \frac{1}{m}(-L\sin\eta - D\cos\eta) \\
\frac{\mathrm{d}}{\mathrm{d}t}v_y(t) &= \frac{1}{m}(L\cos\eta - D\sin\eta - W).
\end{aligned}$$

The polar drag is $C_D(C_L) = C_0 + kC_L^2$, and the expressions are defined as

$$\begin{aligned}
D &= \frac{1}{2}C_D\rho S v_r^2, & L &= \frac{1}{2}C_L\rho S v_r^2, \\
X &= \left(\frac{x}{R} - 2.5\right)^2, & u_a(x) &= u_M(1-X)e^{-X}, \\
V_y &= v_y - u_a(x), & v_r &= \sqrt{v_x^2 + V_y^2}, \\
\sin\eta &= \frac{V_y}{v_r}, & \cos\eta &= \frac{v_x}{v_r}.
\end{aligned}$$

The constants are

$$\begin{aligned}
u_M &= 2.5, & m &= 100\ [kg], \\
R &= 100, & S &= 14\ [m^2], \\
C_0 &= 0.034, & \rho &= 1.13\ [kg/m^3], \\
k &= 0.069662, & g &= 9.80665\ [m/s^2],
\end{aligned}$$

finally $W = mg$ and the control is the lift coefficient C_L which is bounded in $0 \le C_L \le 1.4$. The boundary conditions for the problem are

$$\begin{aligned}
x(0) &= 0, & x(T) &: \text{ free}, \\
y(0) &= 1000, & y(T) &= 900, \\
v_x(0) &= 13.2275675, & v_x(T) &= 13.2275675, \\
v_y(0) &= -1.28750052, & v_y(T) &= -1.28750052.
\end{aligned}$$

Notice that also the final time T is free.

We first tried (with XOptima) the pure formulation of Betts without introducing tricks, but we could not achieve (good) convergence to a valid solution. Instead of performing simplifications of the

model, we found out that a new parametrization of the problem in the spatial coordinate, permitted to quickly solve the problem in few iterations and little time also on a coarse mesh. Next we give the result of the transform of the problem from the time dependence to the spatial variable. The first step is to change from t to x the independent variable, this is done via the condition $x(t(x)) = x$ so that we can obtain $t'(x)$ from the equation $v_x(t(x))t'(x) = 1$, whereas for a function $f(t)$ we have

$$\frac{\mathrm{d}f}{\mathrm{d}x}(t(x)) = f'(t(x))t'(x) = \frac{f'(t(x))}{v_x(x)}.$$

The second step is the change of variable $x = \zeta\ell(\zeta)$ for the new independent variable $\zeta \in [0, 1]$ and the maximum range $\ell(\zeta)$ which is constant. Hence, with this choice $\ell'(\zeta) = 0$ and

$$\frac{\mathrm{d}f}{\mathrm{d}\zeta}(x(\zeta)) = f'(x(\zeta)x'(\zeta) = f'(x(\zeta))\ell(\zeta).$$

The optimal control problem takes the new form of

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\zeta}t(\zeta) &= \frac{\ell(\zeta)}{v_x(\zeta)} \\
\frac{\mathrm{d}}{\mathrm{d}\zeta}y(\zeta) &= \frac{\ell(\zeta)v_y(\zeta)}{v_x(\zeta)} \\
\frac{\mathrm{d}}{\mathrm{d}\zeta}v_x(\zeta) &= \frac{\ell(\zeta)\rho S}{2v_x(\zeta)m}v_r(\zeta)\left(-C_D v_x(\zeta) - C_L V_y(\zeta)\right) \\
\frac{\mathrm{d}}{\mathrm{d}\zeta}v_y(\zeta) &= \frac{\ell(\zeta)\rho S}{2v_x(\zeta)m}v_r(\zeta)\left(-C_D V_y(\zeta) + C_L v_x(\zeta)\right) - \frac{\ell(\zeta)g}{v_x(\zeta)} \\
\frac{\mathrm{d}}{\mathrm{d}\zeta}\ell(\zeta) &= 0,
\end{aligned}$$

where

$$v_r(\zeta) = \sqrt{v_x(\zeta)^2 + V_y(\zeta)^2}, \qquad V_y(\zeta) = (v_y(\zeta) - u_a(\zeta\ell(\zeta)).$$

We started XOptima with a smooth penalty function on the control, that we made sharper at each iteration of the algorithm with a homotopy argument. We obtained a maximum value for $\ell = x(T) = 1248.02$ and a final time $T = 98.43$, while in [Bet01] is reported a value of 1248.031026. The plots for the control and the states are reported in Figure 5.19.

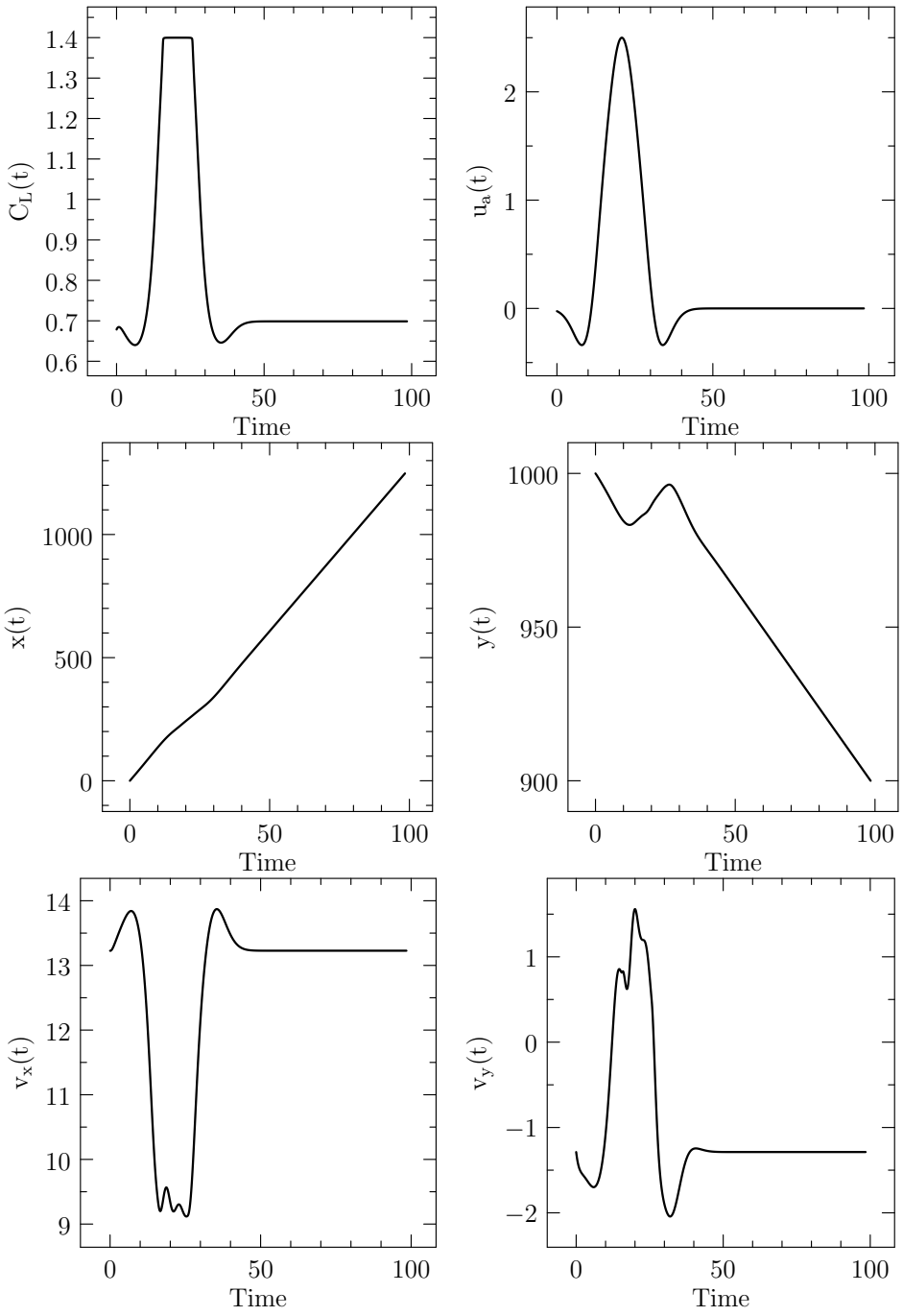


Figure 5.19: Results for the Hang Glider problem. From the top: the control C_L and the thermal drift u_a; the positions x and y; the velocities v_x and v_y.

5.4.2 *Luus 4*

This last example concludes the overview of the family of singular chattering controls proposed by Luus, the optimal control problem is a modification of the Fuller problem of order 3. A third order problem is hard because its geometric structure is very involved. The statement of the Fuller problem is the following.

$$\min J = \int_0^T x(t)^2 \, \mathrm{d}t, \qquad x' = y, \quad y' = z, \quad z' = u \qquad |u| \leq 1,$$

the final time T is free (with terminal condition the origin) or infinity. The initial point is set to $(x_0, y_0, z_0) = (1, 0, 0)$. The problem proposed by Luus sets the final time $T = 5$ and no terminal conditions. By some authors ([FO77]) it is claimed that the problem with $T = 5$ has an infinite number of switching times, but the situation is actually not so clear. While it is true that the Fuller problem of order 3 possesses a constant ratio solution, it possesses also other non chattering solutions. This variety makes it so difficult to give a precise answer. It is proved in [ZB94] that the Fuller problem admits trajectories that reach the origin without switches if the starting point is on the curve

$$\rho(t) = \left[\frac{u}{6}t^3, \quad \frac{u}{2}t^2, \quad ut\right], \quad u = \pm 1.$$

For points not on that curve, there is an optimal chattering control that switches on a switching surface contained in $\mathbb{R}^3$, but its equation is not given. Therefore, it is not clear if the optimal control for the problem of Luus 4 is chattering: the question is what is the final time to reach the origin for the Fuller problem? If it is smaller than 5, then the problem of Luus is chattering, if the final time is greater than 5, then the optimal control is just bang-bang.
We give here some partial results on the Fuller problem. Using the technique introduced for the Fuller problem of order 2, we can integrate the differential system exploiting its symmetry properties. The result of the integration for a constant control u is

$$\begin{aligned}
z &= \Delta u + a_3 \\
y &= \frac{1}{2}\Delta^2 u + a_3\Delta + a_2 \\
x &= \frac{1}{6}\Delta^3 u + \frac{1}{2}a_3\Delta^2 + a_2\Delta + a_1 \\
\lambda_1 &= -\frac{1}{12}\Delta^4 u - \frac{1}{2}a_3\Delta^3 - a_2\Delta^2 - 2a_1\Delta + q_1 \\
\lambda_2 &= \frac{1}{60}\Delta^5 u + \frac{1}{12}a_3\Delta^4 + \frac{1}{3}a_2\Delta^3 + a_1\Delta^2 - q_1\Delta + q_2 \\
\lambda_3 &= -\frac{1}{360}\Delta^6 u - \frac{1}{60}a_3\Delta^5 - \frac{1}{12}a_2\Delta^4 - \frac{1}{3}a_1\Delta^3 + \frac{1}{2}q_1\Delta^2 - q_2\Delta
\end{aligned}$$

Now we impose the symmetry relation

$$z = -ka_3, \quad y = -k^2 a_2, \quad x = -k^3 a_1, \quad \lambda_1 = -k^4 q_1, \quad \lambda_2 = -k^5 q_2, \quad \lambda_3 = 0,$$

and the solution of the resulting nonlinear system gives the following equation in k,

$$k^8 - 7k^7 - 2k^6 + 8k^5 + 17k^4 + 8k^3 - 2k^2 - 7k + 1 = 0. \tag{5.17}$$

Although it is a degree 8 polynomial, it is a reciprocal polynomial and it can be shown applying Galois Theory that its symmetry group is not the whole permutation group S_8, but the subgroup generated by $(1,4,2,3)(5,7,6,8)$, $(1,3,)(6,8)$, $(3,6)$. Moreover, if s_i are the 4 roots of $x^4 - 7x^3 - 6x^2 + 29x + 23$

the roots of polynomial (5.17) are obtained from the roots of $x^2 - s_i x + 1$. This shows that (5.17) is solvable by radicals, in practise the resulting expression is very complicated, so here we just give the numerical approximation of the 4 real roots, they are

$$k_i : \qquad 0.1414077939, 0.5757361184, 1.736906836, 7.071745993, \qquad i = 1 \dots 4.$$

It can be shown as in [Mar73], that the only admissible value is that of $k_2 = 0.5757\ldots = 1/1.7369\ldots$: the first value corresponds to a geometric progression toward the origin, the second value is the geometric progression that escapes from the origin. The uniqueness of the family of constant ratio solutions is in accord with the theorems and conjectures of [ZB94].
We try now to explicitly obtain the switching surface of these constant ratio solutions. Consider the Hamiltonian of the system,

$$\mathcal{H} = x^2 + \lambda_1 y + \lambda_2 z + \lambda_3 u = 0,$$

where $\mathcal{H} = 0$ because we are considering the free time case of the problem. We introduce the Bellman function $V(x, y, z)$ that associates to the initial point (x, y, z) the minimum of the functional to be minimized. The Bellman function for this problem has the property that

$$V(k^3 x, k^2 y, kz) = k^7 V(x, y, z).$$

Performing the partial derivative of this equation with respect to k we have

$$\begin{aligned} \frac{\partial V}{\partial k} &= 3k^2 V_x + 2ky V_y + k V_z = 7k^6 V, \\ V_x &= \lambda_1, \quad V_y = \lambda_2, \quad V_z = \lambda_3. \end{aligned}$$

Now we set the trivial value of $k = 1$ and end up with

$$3xV_x + 2yV_y + V_z - 7V = 0 \implies V_z = 7V - 3xV_x - 2yV_y.$$

With a change of variable coming from the symmetry properties, we can simplify V setting $\phi = \frac{x}{z^3}$, $\psi = \frac{y}{z^2}$ and $V = z^7 F(\phi, \psi)$. The derivatives of V are therefore

$$V_x = z^4 F_\phi = \lambda_1, \qquad V_y = z^5 F_\psi = \lambda_2$$

Hence we rewrite the Hamiltonian with respect to the new variables,

$$\begin{aligned} 0 = \mathcal{H}(\phi, \psi) &= \phi^2 z^6 + V_x \psi z^2 + V_y z + |V_z| \\ &= z^6 \left(\phi^2 + \psi F_\phi + F_\psi + |V_z|\right) \\ &= \phi^2 + \psi F_\phi + F_\psi + |V_z| \\ &= \phi^2 + \psi F_\phi + F_\psi + |7F - 3\phi F_\phi - 2\psi F_\psi| \end{aligned}$$

In order to solve this PDE, it is convenient to split the absolute value in two cases, F^- where $V_z = \lambda_3 > 0$, $u = -1$ and F^+ where $V_z = \lambda_3 < 0$, $u = +1$:

$$\begin{aligned} F^- &: \phi^2 + \psi F_\phi + F_\psi + 7F - 3\phi F_\phi - 2\psi F_\psi = 0, \\ F^+ &: \phi^2 + \psi F_\phi + F_\psi - 7F + 3\phi F_\phi + 2\psi F_\psi = 0. \end{aligned}$$

The two branches of the PDE are respectively

$$\begin{aligned} F^- &= -\phi^2 - \frac{1}{3}\psi^2 + \left(\psi - \frac{1}{4}\right)\phi + \frac{11}{60}\psi - \frac{11}{420} + (2\psi-1)^{(7/2)}G^-\left(\alpha^-\right) \\ F^+ &= \phi^2 + \frac{1}{3}\psi^2 + \left(\psi + \frac{1}{4}\right)\phi + \frac{11}{60}\psi + \frac{11}{420} + (2\psi+1)^{(7/2)}G^+\left(\alpha^+\right) \\ &\text{with} \qquad \alpha^- = \frac{3\phi - 3\psi + 1}{3(2\psi-1)^{(3/2)}}, \qquad \alpha^+ = \frac{3\phi + 3\psi + 1}{3(2\psi+1)^{(3/2)}} \end{aligned}$$

where G^- and G^+ are arbitrary functions coming from the resolution of the PDE. There is a relation between G^- and G^+ that can be found using the symmetry of the problem: since $V(x,y,z) = V(-x,-y,-z)$, we have that $F^-(\phi,\psi) = -F^+(\phi,-\psi)$ and $G^-(\alpha^-) = -G^+(\alpha^-)$, $G^-(\alpha^+) = -G^+(\alpha^+)$, thus $G^- = -G^+ = G$. If we return to the variables (x,y,z), we have

$$\begin{aligned} F^- &= -x^2z + xyz - \frac{1}{3}y^2z^3 - \frac{1}{4}xz^4 + \frac{11}{60}yz^5 - \frac{11}{420}z^7 + (2y-z^2)^{7/2}G(\alpha) \\ F^+ &= x^2z + xyz + \frac{1}{3}y^2z^3 + \frac{1}{4}xz^4 + \frac{11}{60}yz^5 + \frac{11}{420}z^7 - (2y+z^2)^{7/2}G(\alpha) \\ &\text{with} \qquad \alpha = \frac{z^3 - 3yz + 3x}{3(2y-z^2)^{3/2}}. \end{aligned}$$

To determine the switching manifold, we have to impose the continuity conditions

$$V^- = V^+, \qquad V_x^- = V_x^+, \qquad V_y^- = V_y^+.$$

The condition $V_z^- = V_z^+$ is not necessary because it is redundant and already contained in the previous relations. Performing the differentiations we have:

$$\begin{aligned} V_x^- &= z^4 F_\phi^-, \qquad F_\phi^- = -2\phi + \psi - \frac{1}{4} + (2\psi-1)^2 G'(\alpha) \\ V_x^+ &= z^4 F_\phi^+, \qquad F_\phi^+ = 2\phi + \psi + \frac{1}{4} - (2\psi+1)^2 G'(\alpha) \\ V_y^- &= z^5 F_\psi^-, \qquad F_\psi^- = \phi - \frac{2}{3}\psi + \frac{11}{60} + 7(2\psi-1)^{5/2}G(\alpha) - (2\psi-1)(3\phi-\psi)G'(\alpha) \\ V_y^+ &= z^5 F_\psi^+, \qquad F_\psi^+ = \phi + \frac{2}{3}\psi + \frac{11}{60} - 7(2\psi+1)^{5/2}G(\alpha) + (2\psi+1)(3\phi+\psi)G'(\alpha) \end{aligned}$$

Setting for simplicity $G(\alpha) = A$ and $G'(\alpha) = a$. The relation $V^- = V^+$ is equivalent to $F^-(\phi,\psi) = F^+(\phi,\psi)$, that is

$$2\phi^2 + \frac{2}{3}\psi^2 + \frac{1}{2}\phi + \frac{11}{210} - A\left((2\psi-1)^{7/2} + (2\psi+1)^{7/2}\right) = 0.$$

The relation $V_x^- = V_x^+$ becomes

$$4\phi + \frac{1}{2} - 2a(4\psi^2+1) = 0.$$

The relation $V_y^- = V_y^+$ becomes

$$\frac{4}{3}\psi - 7A\left((2\psi+1)^{5/2} + (2\psi-1)^{5/2}\right) + 2a\psi(6\phi+1) = 0.$$

A numerical simulation was still hard to obtain, but shows that the final time to reach the origin is around $T = 5.63 > 5$ so we are expecting that the problem of Flaherty [FO77] and [Luu00] does not have the chattering phenomenon. This is in agree with the numerical test performed with different OCP softwares, none of them exhibit chattering. We summarize in 5.11 the numerical results for the case of fixed final time to $T = 5$ as formulated in [FO77]. The plots of the corresponding controls are collected in Figure 5.20. The exact control has 5 switching points, trying to impose a

Table 5.11: Summary of the results for problem Singular n4, in the first column the article with the first author or the name of the algorithm, in the second column the value of the target, in the third column the ratio with the exact value, in the fourth the number of switches detected.

Method/Author	Reported value	Error	n. of switches
Exact value	1.2521117475984577	0	5
XOptima	1.2521117901796599	3.3E-8	5
ICLOCS	NC	1.0E+7	
Gpops	1.2521531043753287	3.3E-5	4
ACADO	1.2521241356056492	9.8E-6	4
PROPT	1.2523896453830434	2.2E-4	4
Luus [Luu00]	1.2521128	8.4E-7	5
Flaherty [FO77]	1.2521	9.3E-6	

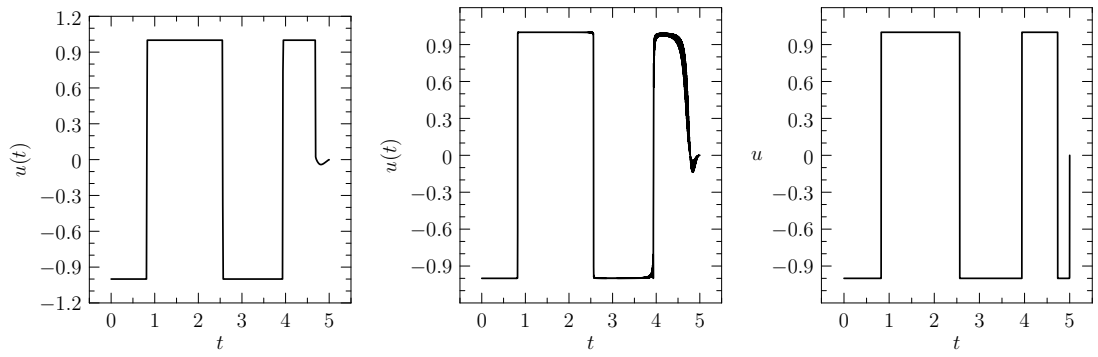


Figure 5.20: The plots of the results obtained with Acado, Gpops and XOptima.

big nonlinear system forcing 6 switches did not produce any solution. As pointed out in [FO77], the contribution of the last interval is negligible, we computed the exact value of the functional of the suboptimal solution with 4 switches and it is 1.2521120520842668, the difference with the value with 5 switches is around 10^{-7}.

We wanted to check the Fuller phenomenon, so we let the terminal time free. The solvers converged with great difficulty and the results were very different. We expect the value of the target to be lower than the target obtained in the fixed time problem, the problem is that even suboptimal solutions are very close to that value, as the case with 4 switches shows. Thus, in the free time case, Gpops converged to a suboptimum with higher value but with a sharper and more defined control, Acado converged far from the expected terminal time, Xoptima was almost close to the expected estimated terminal time (Table 5.12) . The resulting controls are in Figure 5.21. The estimated terminal time is the accumulation point (Fuller point) of the switching times, and can be obtained as a geometric progression. From the theory exposed so far, we know that the ratio of

Table 5.12: Summary of the results for problem Singular n4 in the free time case, in the first column the article with the first author or the name of the algorithm, in the second column the value of the target, in the third column the difference of the fixed time target and the free time target, in the fourth the terminal time T, in the fifth the number of switches detected.

Method/Author	Reported value	$J_5 - J_T$	T	n. of switches
XOptima	1.2521123265506712	5.7E-7	5.653	11
Gpops	1.2521258305303968	1.4E-5	4.794	5
ACADO	1.2521247813454326	1.3E-5	6.532	12

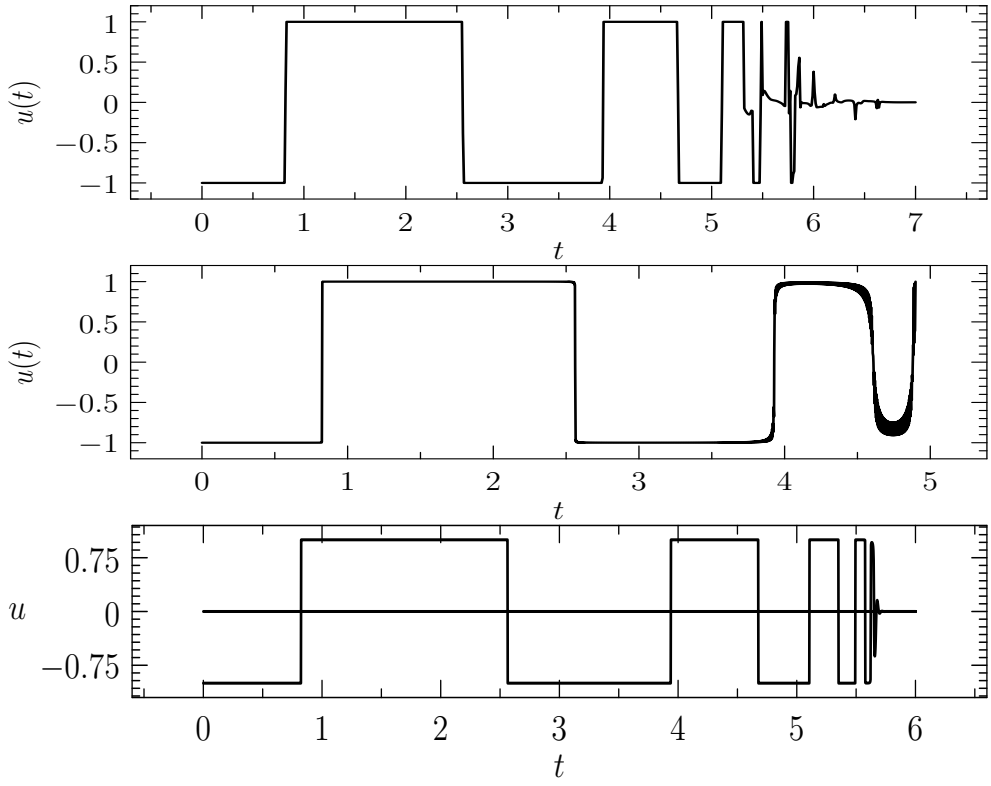


Figure 5.21: The plots of the results obtained with Acado, Gpops and XOptima.

successive time intervals is $k_2 = 0.5757\ldots$, so we can estimate T from the first switches. Once the trajectory is stabilized on the switching curve we take $\Delta = t_j - t_{j-1}$ and then $T = t_{j-1} + \frac{\Delta}{1-k_2}$. From the numerical result of XOptima we have the switching times and relative ratios reported in Table 5.13. The best deviation from k_2 is less than 0.01 corresponding to t_{10}, hence we have

$$T \approx t_9 + \frac{\Delta}{1-k_2} = t_9 + \frac{t_{10} - t_9}{1-k_2} = 5.634601330.$$

Table 5.13: The switching times obtained with XOptima for the problem Luus 4 with free terminal time.

i	t_i	$(t_{i+1} - t_i)/(t_i - t_{i-1})$
1	0.82427999792319	
2	2.56466959566023	0.7902
3	3.93992720492731	0.5357
4	4.67675043914512	0.5845
5	5.10747500039411	0.5685
6	5.35235393876917	0.5803
7	5.49447117979041	0.5846
8	5.57755510531052	0.5526
9	5.62346990625584	0.5714
10	5.64970693536746	0.0833
11	5.65189335446009	

5.4.3 *Underwater Vehicle*

The problem of the underwater vehicle is of minimum time and consists in driving the submarine from an initial configuration to a final configuration at rest. It is formulated in [CSMV04, Chy03] and has the peculiarity of having three controls that admit singular arcs and chattering arcs. The vehicle moves in the x, z plane and is actuated by thrusters while submerged. The equation of motion are

$$\begin{aligned} x' &= v_1 \cos\theta + v_3 \sin\theta \\ z' &= v_3 \cos\theta - v_1 \sin\theta \\ \theta' &= \Omega \\ v_1' &= -v_3\Omega\frac{m_3}{m_1} + \frac{u_1}{m_1} \\ v_3' &= v_1\Omega\frac{m_1}{m_3} + \frac{u_2}{m_3} \\ \Omega' &= v_1 v_3 \frac{m_3 - m_1}{I} + \frac{u_3}{I}, \end{aligned}$$

where the masses $m_1 = 13.2$ and $m_3 = 25.6$, the inertia $I = 0.12$ and the motors are limited to $|u_i| \leq 1$, $i = 1, 2, 3$. The boundary conditions used by the authors in their paper are $(0, 1, 0, 0, 0, 0)$ corresponding to $t = 0$ and $(2, 1, 0, 0, 0, 0)$ for $t = T$. A reasonable guess for the final minimum time is $T \approx 10$. The authors of [CSMV04] describe the intuitive bang-bang solution that involves the linear motion of the vehicle using only one thruster, then they give another "surprising" ([Chy03]) solution which is around 10% better. The optimal solution is a non intuitive sequence of chattering and singular arcs, so the problem can not be directly solved but some parameters need to be relaxed with a homotopy argument. Moreover the fineness of the grid is important to obtain a good convergence. In [CSMV04] the problem is solved via homotopy on I starting from a value of $I = 2$ to the desired value $I = 0.12$ with a grid of 1000 up to 10000 nodes. The software used was AMPL together with LOQO.

We were not able to solve the problem with Acado, Gpops and Iclocs, because they do not feature continuation/homotopy. We were successful with XOptima performing the continuation method on the inertia and on the relaxation parameter of the penalties on the three controls. We used a mesh of 10000 nodes but found no significant improvement with respect to the mesh with 2000 nodes. We obtained five different solutions that we show in the next figures. The first is the "intuitive" bang bang solution with only one active thruster that gives a final time of $T = 10.2761863103346265$, with switching time at half the period. We found then two other solution with time $T = 9.30138845993620045$, they are symmetric with respect to the controls u_2 and u_3, and this is not a surprise from the symmetry of the optimal manoeuver, Figure 5.22. Then there are other two (symmetric) solutions, that yields a slightly better time of $T = 9.17050944725101225$, Figure 5.23. In [CSMV04] it is reported the value of $T = 9.254699$ which corresponds to an intermediate solution between the four that we propose. This can be explained from the nature of the problem: the presence of the Fuller phenomenon causes the presence of many suboptimal solutions, as we have seen in the problems proposed in the previous section. Unfortunately, in [CSMV04] it is not reported the value of the first pure bang bang solution, so a comparison is not possible. We can only analyse the quality of the oscillations by having a look of the pictures presented in [CSMV04] with respect to Figures 5.22 and 5.23. We see that the oscillations are more defined towards ± 1 than the oscillations of the authors, and maybe this explains also better our lower value of the final time.

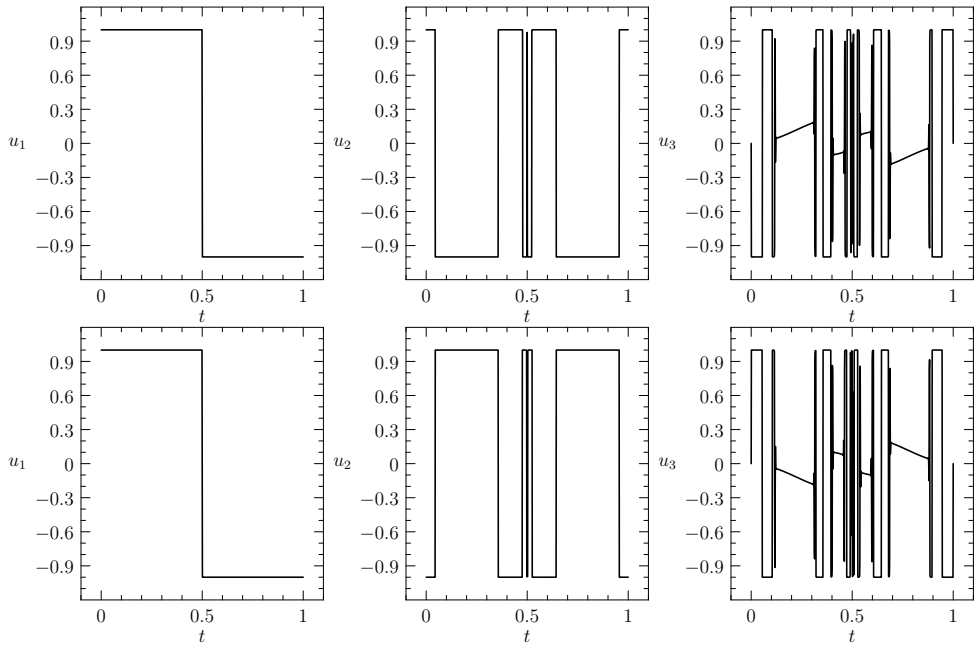


Figure 5.22: The plots of the results for $T = 9.30138845993620045$ with XOptima. The two solutions are symmetric with respect to u_2 and u_3. The authors of [CSMV04] converged to a solution of the first kind. We plot the pseudotime t/T so that it is normalized in $[0, 1]$.

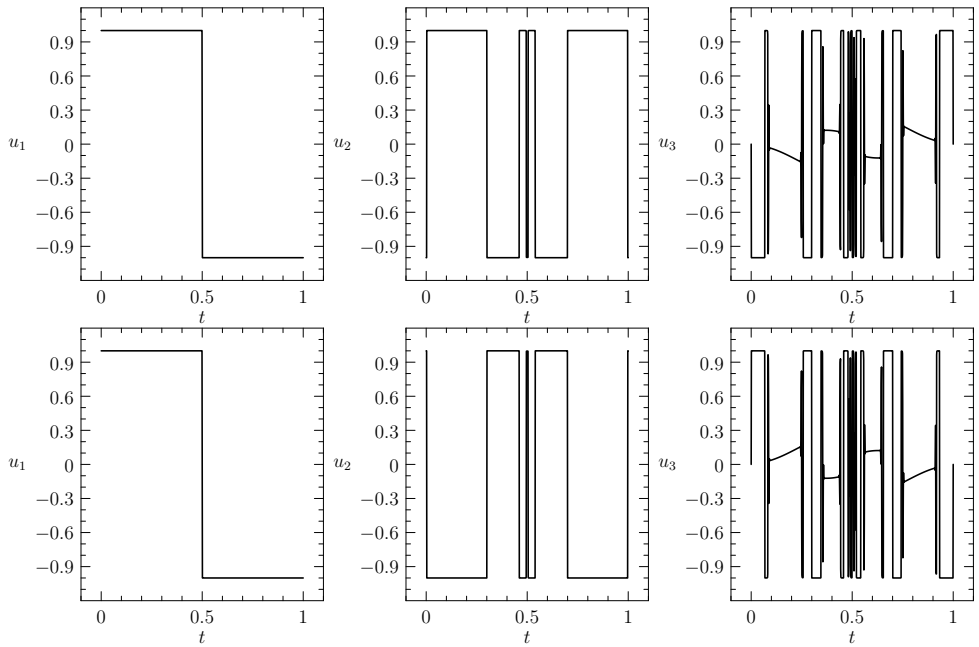


Figure 5.23: The plots of the results for $T = 9.17050944725101225$ with XOptima. We plot the pseudotime t/T so that it is normalized in $[0, 1]$.

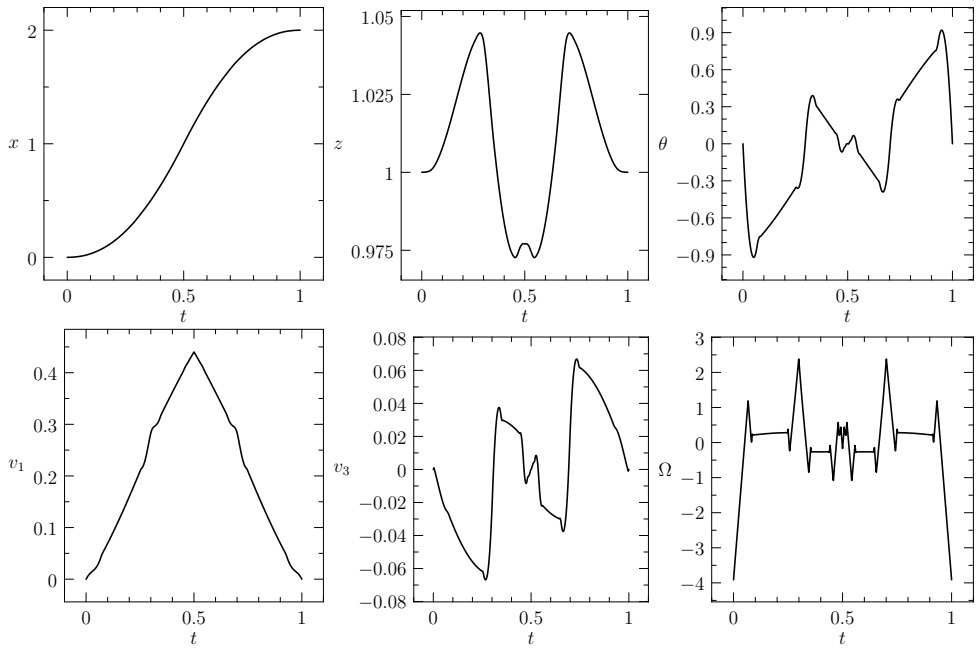


Figure 5.24: The plots of the states for $T = 9.170$ with XOptima (the second case). We plot the pseudotime t/T so that it is normalized in $[0, 1]$.

5.4.4 *Minimum Lap Time*

In [BBDL05] it is described the OCP of a realistic model of a competition motorcycle that have to perform the minimum lap time on a circuit track. One major problem is to simulate the driving skills of the pilot, that are different from the mathematical ideal model. This behaviour is evident if we take into account the ellipse of adherence of the vehicle: a professional driver does not use the whole ellipse, he stays away from the area of minimal longitudinal and lateral forces. This fact is modelled cutting away that area from the ellipse as is it shown below. A description of the model can be found in [TCS14] where there are also dynamic considerations and comments, the equations can be found in [BBDL03]. The road description is based on the article [BL13].
The problem of finding the minimum lap time of racing vehicle can be nicely formulated as an optimal control problem with cyclic condition at the boundaries. Cyclic conditions mean same initial and final states (but not fixed): their actual value would be calculated as part of the solution of the optimal control. The problem is challenging for many reasons:

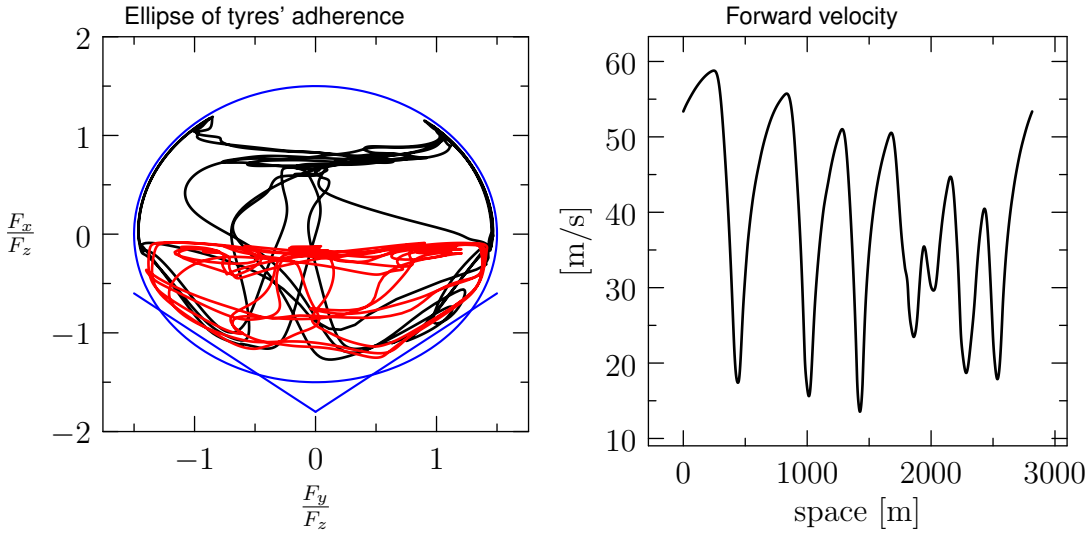


Figure 5.25: Usage of the ellipse of adherence for the front (in red) and the rear wheel (in black). On the right, optimal velocity on the circuit of Adria.

- The vehicle dynamics is quite complex and under some operative conditions is unstable.
- It is necessary to impose many path inequality constraints such as road borders, engine maps, etc.
- The resulting BVP problem is quite large and can reach up to 1 million equations for complex dynamic models and long circuits.
- It may happen that the vertical forces reach the zero value which means that the wheel detaches from the ground: therefore the nature of the dynamic system of equations changes and this must be either avoided or carefully handled.
- Ill conditioning due to some states variables having very large values of order of 10^4 and other (such as slip angles) of order 10^{-3}.
- The driver has more than one input to control the longitudinal dynamics, in particular the motorcyclist, which has to two independent controls to maximize the braking manoeuvre performance. This may lead to locally singular problems or singular Jacobians, due to the fact that a slight change in the control has the same little effect on some states yielding a flat descent direction. As an example during hard braking manoeuvres, the rider could

both decelerate with front and rear brakes with a slight traction force in order to increase the deceleration and additionally decelerate by fast steering the front tyre in order to exploit the barking component of the front tyre lateral force.

- Similarly, as for longitudinal dynamics, the lateral dynamics, i.e. finding the optimal trajectory, may pose some difficulties in the convergence.

All the points discussed above come into play in the example we propose in this section. We formulated a minimum lap time of a sport motorcycle on a real racing track (Adria, in Italy, and Spa-Francorchamps, in Belgium) that includes a first order approximation of the suspension effect, the engine map, non linear tyres with first order dynamics and constraints on pattern of longitudinal and lateral accelerations.
The approach adopted is the one discussed in [TCS14] using the software XOptima. The numerical problem dimensions and convergence performance are reported in the following table:

N. equations	=	91416
Iterations	=	253/300
Function Evaluations	=	1054
Jacobian Factorizations	=	253
Jacobian Inversions	=	1306
Tolerance	=	1.0 E-09
Last Residual	=	3.1 E-10
Elapsed Time	=	1 : 38 min
Processor	=	Intel Core i7 2.66GHz

To find the solution we did not used any special guess: a steady state longitudinal motion was used to initialize the states of the motorcycle model placed in the center line. All other variables (such as Lagrange multipliers) are set to zero. We used continuation to push the inequality constraints parameter to the limit.
Figure 5.29 shows the optimal trajectory that uses all the available road up to the limit allowed

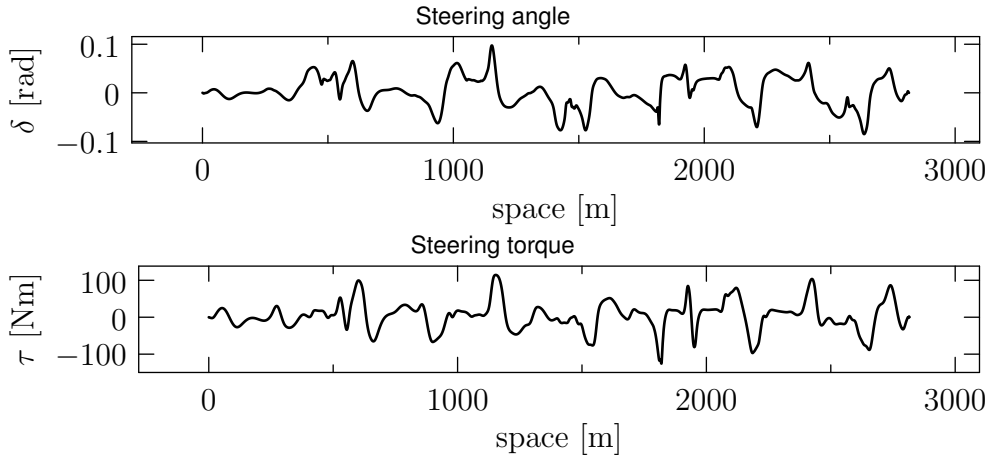


Figure 5.26: Optimal steering angle (top) and steering torque (bottom) on the circuit of Adria.

by the road constraint tolerance. Figure 5.25 shows the longitudinal velocity where one may appreciate that the initial and final velocity are the same due to the cyclic conditions. This is also

true for the lateral displacement for trajectory in Figure 5.25. The velocity highlights the hard braking manoeuvres and accelerations. Figure 5.26 shows the lateral dynamic control (i.e the steering torque) which is between the limit of a human rider and generates the steering angle. In Figure 5.27 the roll angle reaches the 60 degrees which is manifest of a manoeuvre pushed to the limit. One may also note that the maximum angle is kept for a larger period for corners on the right (positive roll angles) since the race track is run clockwise and most corners are on this side. Figure 5.28 is interesting since its shows the longitudinal and vertical forces. The vertical forces

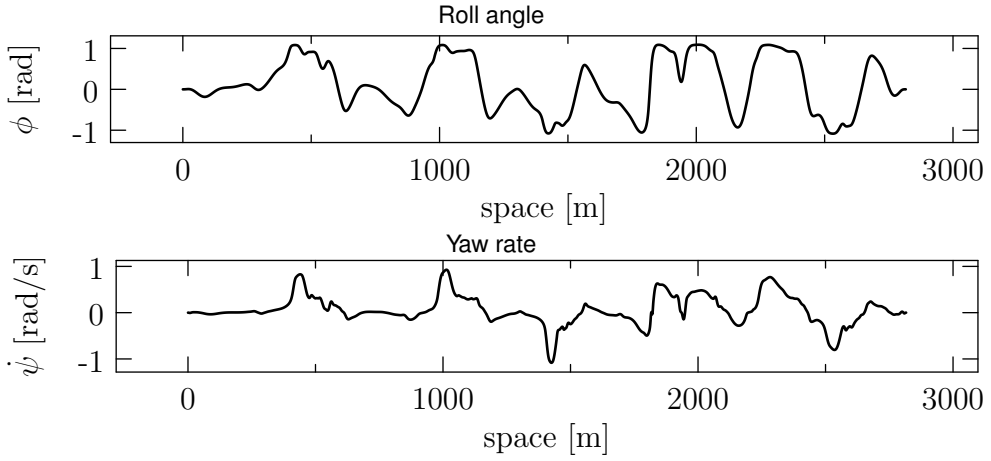


Figure 5.27: Optimal roll angle (top) and yaw rate (bottom) on the circuit of Adria.

are affected by a large load transfer due to longitudinal forces that almost generate the wheel detach from ground (zero vertical forces): for example, for traction phase at points $A_1, \ldots, A_4$, the front wheel reaches the minimum which is imposed by an inequality constraints (about $100N$). For braking phase is the rear wheel that is almost detached from ground at points $B_1, \ldots, B_4$, see Figure 5.28 and 5.29. The reader may also note that at $1800m$ there is a quick change of direction and the roll angle passes from -60 to $+60$ degrees. This generates a large centrifugal force in the vertical direction which tends to lift the motorcycle from ground (consistent reduction of both vertical forces). This is a peculiarity of motorcycles, due to the large roll angle they can reach a fast change of direction. At the same point, in the chart of the steering torque, one may see that a peak torque is reached, which means high level of rider effort. Another interesting comments is the use of both traction force and braking force (combined use of longitudinal controls) in order to maximize the deceleration. This means that the rider is always braking and pushing the vehicle when riding at the limits and that steady state conditions are never reached along the whole track. Finally, Figure 5.25 shows the ellipse of adherence which is the engagement of rear and front tyre, that is, the ratio between lateral and longitudinal force with respect to vertical forces. The rear and front tyre engagement stays inside the ellipse of the maximum tyre adherence and has a particular slightly triangular pattern for the braking phase. This is due to the inequality constraints that is imposed to limit the combination of longitudinal and lateral acceleration during braking to mimic the experimental data. This is called willingness envelope and represents the set of accelerations that the rider is able (and wants) to generate in order to feel safe during the manoeuvre (i.e. not to fall down). The dashed line represents the limit imposed by the constraints. This is not the best optimal braking manoeuvre that really uses at best the rear tyre, however, it is the most realistic one. Additionally, the missing points at the bottom and top of the chart are due to the incoming wheel lift during respectively braking and traction.

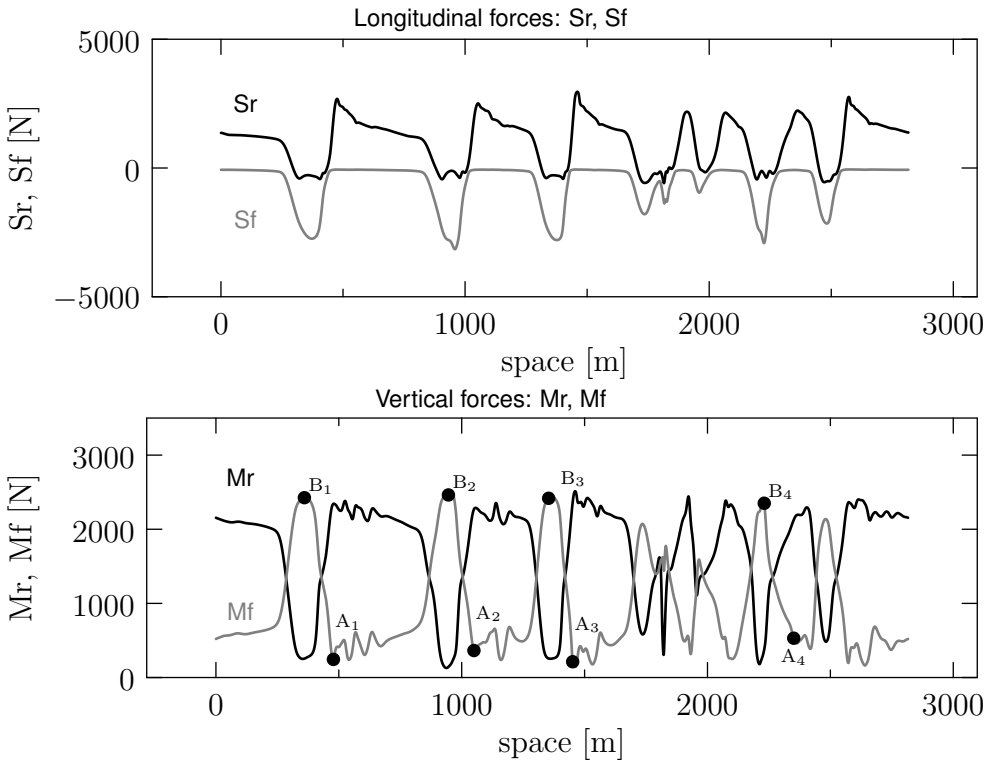


Figure 5.28: Optimal longitudinal forces (top) and vertical forces (bottom) on the circuit of Adria.

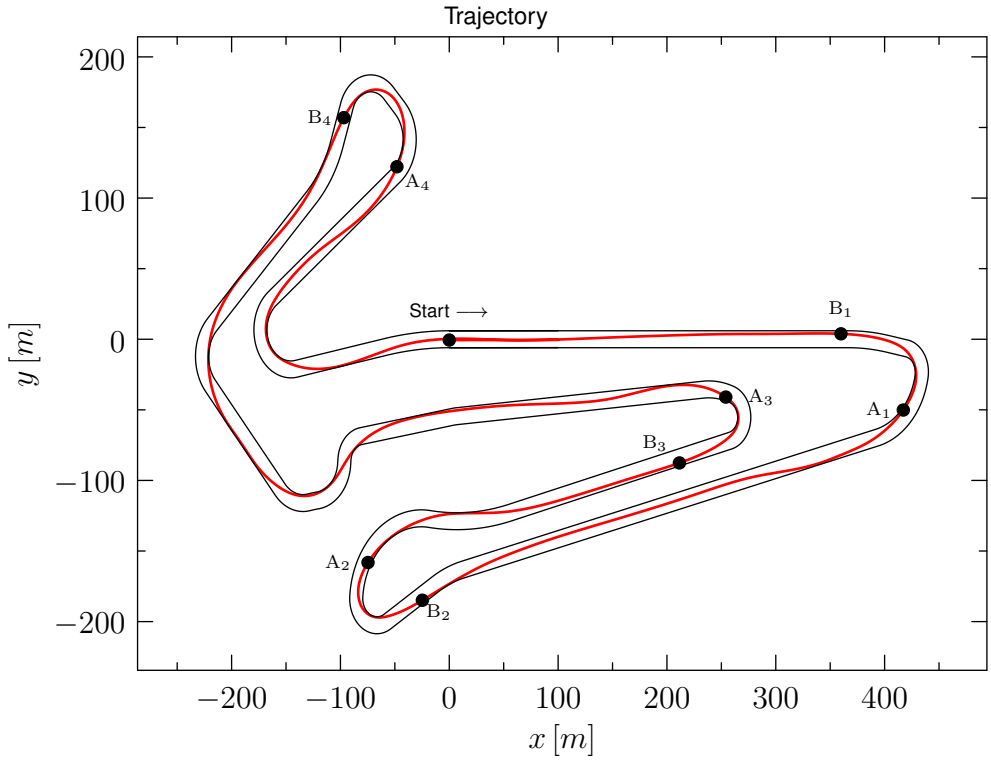


Figure 5.29: The optimal trajectory for the circuit of Adria.

6

CLOTHOIDS FOR ROAD DESIGN

A new algorithm for the solution to the problem of Hermite G^1 interpolation with a clothoid curve is herein proposed, that is a clothoid that interpolates two given points in a plane with assigned unit tangent vectors. The interpolation problem is formulated as a system of three nonlinear equations with multiple solutions which is difficult to solve even numerically. In this work the solution of this system is reduced to the computation of the zeros of only one single nonlinear function in one variable. The location of the relevant zero is tackled analytically: it is provided the interval containing the zero where the solution is proved to exist and to be unique. A simple guess function allows to find that zero with very few iterations in all of the possible instances.

Computing clothoid curves calls for evaluating Fresnel related integrals, asymptotic expansions near critical values are herein conceived to avoid loss of precision. This is particularly important when the solution of the interpolation problem is close to a straight line or an arc of circle. The present algorithm is shown to be simple and compact.

The comparison with literature algorithms proves that the present algorithm converges more quickly and accuracy is conserved in all of the possible instances while other algorithms have a loss of accuracy near the transition zones.

6.1 MOTIVATION

The fitting that allows a curve to interpolate two given points in a plane with assigned tangent directions is called G^1 Hermite interpolation (see Figure 6.1). If the curvatures are also given at the two points, then this is called G^2 Hermite interpolation [MS09]. The G^2 interpolation provides a smoother curve with good properties, at the price of more constraints to be satisfied and this implies heavier computational costs. In several applications (especially in those in real time), the G^1 Hermite interpolation is cost-effective in particular when the discontinuity of the curvature is acceptable. Clothoid curves are used in a variety of applications like the trajectory planning of robots or autonomous vehicles [BBDL06, DC09, DCBM$^+$07, LNRL08, MVHW10, Wil09] or in computer aided design [BLP10, BD12, WMN$^+$01] or in other fields [ALHB08, Dai12]. It is well known that clothoids are extremely useful and this is why they are being studied despite their transcendental form [KDK95, Sto82].
The purpose of this chapter is to describe a new method for the numerical computation of G^1 Hermite interpolation with a single clothoid segment. Nowadays, the best algorithms for solving the G^1 interpolation problem have been proposed by [KFP03] and [WM08, WM09]. An iterative method was proposed by [KFP03]; however, [WM09] remarked that no existence and uniqueness theorem was provided, also because the convergence rate was linear instead of quadratic as in [WM09]. The algorithm proposed by [WM09] performs generally better than [KFP03] in terms of accuracy and number of iterations. It requires to split the procedure in three mutually exclusive cases: straight lines, circles and clothoids, a geometrical fact that helps to understand the problem. For each of the mutually exclusive cases the problem is reduced to find the root of a single nonlinear equation solved using *damped* Newton–Raphson algorithm. However, the root of the nonlinear equations are ill-conditioned near the transition region, e.g. when the clothoid stretches to a straight line or a circle, as shown in the section of numerical tests.

The present algorithm does not need to separate straight lines, circles, clothoids. The G^1 Hermite interpolation is recast in term of computing a *well conditioned* zero of a *unique* nonlinear equation which is proven to exist and to be unique in a prescribed domain. The Newton–Raphson algorithm *without damping* is herein used to solve the nonlinear equation and the additional help of a good initial guess implies that few iterations (less than four) suffice.

The chapter is structured as follows: there are four logical parts, the first is analytic, constructive and leads to the solution of the problem, the second is strictly numeric and implements the algorithm described in the first part, the third discusses a good guess function in order to achieve a reduced number of iteration in all possible cases, the last is a theoretical proof that covers existence and uniqueness of the solution under reasonable hypotheses.
Section 6.2 introduces the mathematical background and the notation used, there is a brief presentation of three possible definitions of the Fresnel integral functions and their momenta with some properties needed later. Section 6.3 defines the interpolation problem from the new analytical point of view. Section 6.4 describes the mathematical passages to reformulate it such that from three equations in three unknowns it reduces to one nonlinear equation in one unknown. A summary of the algorithm and its issues are pointed out, such issues are solved in the following sections. It is given also a pseudo-code of the method. Section 6.5 considers an appropriate guess function to help the Newton–Raphson method to converge in few iterations, allowing the algorithm to be highly performant. Section 6.6 is devoted to answer to the numerical questions that arise when treating the Fresnel integral momenta such as stability and convergence. *Ad hoc* expressions for critical cases are discussed and provided. Section 6.7 covers the theoretical need of a proof of existence and uniqueness of the solution of the nonlinear equation used to solve the interpolation problem. It is explained how to select a valid solution among the infinite

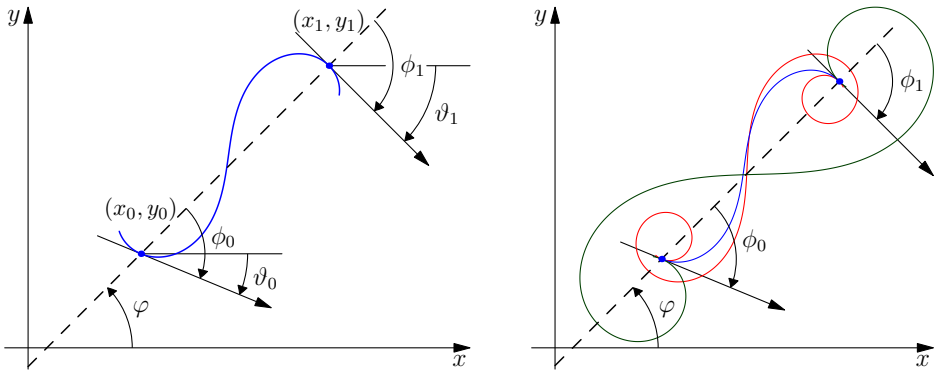


Figure 6.1: Left: G^1 Hermite interpolation schema and notation. Right: some possible solutions.

possibilities and a bounded range where this solution exists and is unique is exhibited. Although this proof is long, technical and not useful from an algorithmic point of view, the authors feel that it is necessary to complete the analysis of the algorithm. Section 6.8 is devoted to numerical tests and comparisons with other methods present in literature. Section 6.9 shows the application of the present algorithm in producing an interpolating clothoid spline that minimizes the jumps of the curvature at the joining points. In the Appendix a pseudo-code completes the presented algorithm for the accurate computation of the Fresnel related integrals.

6.2 SOME PROPERTIES OF FRESNEL INTEGRALS

The fitting problem clearly involves the computation of Fresnel integrals. There are various possible definitions for Fresnel sine $\mathcal{S}(t)$ and cosine $\mathcal{C}(t)$ functions. Here the choice is to follow reference [AS64].

Definition 6.1 (Fresnel integral functions)**.**

$$\mathcal{C}(t) := \int_0^t \cos\left(\frac{\pi}{2}\tau^2\right)\,\mathrm{d}\tau, \qquad \mathcal{S}(t) := \int_0^t \sin\left(\frac{\pi}{2}\tau^2\right)\,\mathrm{d}\tau. \tag{6.1}$$

The literature reports different definitions of Fresnel integrals, such as:

$$\begin{aligned} \widetilde{\mathcal{C}}(t) &:= \int_0^t \cos(\tau^2)\,\mathrm{d}\tau, & \widetilde{\mathcal{S}}(t) &:= \int_0^t \sin(\tau^2)\,\mathrm{d}\tau, \\ \widehat{\mathcal{C}}(\theta) &:= \frac{1}{\sqrt{2\pi}}\int_0^\theta \frac{\cos u}{\sqrt{u}}\,\mathrm{d}u, & \widehat{\mathcal{S}}(\theta) &:= \frac{1}{\sqrt{2\pi}}\int_0^\theta \frac{\sin u}{\sqrt{u}}\,\mathrm{d}u. \end{aligned} \tag{6.2}$$

The following identities allow to switch among these definitions:

$$\begin{aligned} \mathcal{C}(t) &= \int_0^{\frac{\sqrt{2}}{\sqrt{\pi}}t} \cos\left(\tau^2\right)\,\mathrm{d}\tau = \frac{\mathrm{sign}(t)}{\sqrt{2\pi}}\int_0^{\frac{\pi}{2}t^2} \frac{\cos u}{\sqrt{u}}\,\mathrm{d}u, \\ \mathcal{S}(t) &= \int_0^{\frac{\sqrt{2}}{\sqrt{\pi}}t} \sin\left(\tau^2\right)\,\mathrm{d}\tau = \frac{\mathrm{sign}(t)}{\sqrt{2\pi}}\int_0^{\frac{\pi}{2}t^2} \frac{\sin u}{\sqrt{u}}\,\mathrm{d}u. \end{aligned} \tag{6.3}$$

Also momenta of Fresnel integrals are used forward:

$$\mathcal{C}_k(t) := \int_0^t \tau^k \cos\left(\frac{\pi}{2}\tau^2\right)\,\mathrm{d}\tau, \qquad \mathcal{S}_k(t) := \int_0^t \tau^k \sin\left(\frac{\pi}{2}\tau^2\right)\,\mathrm{d}\tau. \tag{6.4}$$

Notice that $\mathcal{C}(t) := \mathcal{C}_0(t)$ and $\mathcal{S}(t) := \mathcal{S}_0(t)$ and that the first momenta are easily obtained:

$$\mathcal{C}_1(t) = \frac{1}{\pi}\sin\left(\frac{\pi}{2}t^2\right), \qquad \mathcal{S}_1(t) = \frac{1}{\pi}\left(1-\cos\left(\frac{\pi}{2}t^2\right)\right). \tag{6.5}$$

It is possible to reduce the integrals (6.4) to a linear combination of standard Fresnel integrals (6.1) with some trigonometric functions. Closed forms via the exponential integral or the Gamma function are also possible, however it is easy to express them as a recurrence. Integrating by parts, the following recurrence is obtained:

$$\begin{aligned}\mathcal{C}_{k+1}(t) &= \frac{1}{\pi}\left(t^k\sin\left(\frac{\pi}{2}t^2\right) - k\,\mathcal{S}_{k-1}(t)\right),\\ \mathcal{S}_{k+1}(t) &= \frac{1}{\pi}\left(k\,\mathcal{C}_{k-1}(t) - t^k\cos\left(\frac{\pi}{2}t^2\right)\right).\end{aligned} \tag{6.6}$$

Recurrence is started by computing standard Fresnel integrals (6.1) and first momenta (6.5). Notice that from recurrence (6.6) it follows that $\mathcal{C}_k(t)$ and $\mathcal{S}_k(t)$ with k odd do not contain Fresnel integrals (6.1) and are combination of elementary functions. It is convenient to introduce now the following functions whose properties are studied in Section 6.6:

$$\begin{aligned}X_k(a,b,c) &:= \int_0^1 \tau^k\cos\left(\frac{a}{2}\tau^2+b\tau+c\right)\,\mathrm{d}\tau,\\ Y_k(a,b,c) &:= \int_0^1 \tau^k\sin\left(\frac{a}{2}\tau^2+b\tau+c\right)\,\mathrm{d}\tau.\end{aligned} \tag{6.7}$$

Notice that, with a simple change of variable, one has the identities

$$\begin{aligned}\int_0^s \tau^k\cos\left(\frac{a}{2}\tau^2+b\tau+c\right)\,\mathrm{d}\tau &= s^{1+k}X_k(as^2,bs,c),\\ \int_0^s \tau^k\sin\left(\frac{a}{2}\tau^2+b\tau+c\right)\,\mathrm{d}\tau &= s^{1+k}Y_k(as^2,bs,c).\end{aligned}$$

which are used in the definition of the fitting problem.

6.3 THE FITTING PROBLEM

Consider the curve which satisfies the following system of ordinary differential equations (ODEs):

$$\begin{aligned}x'(s) &= \cos\vartheta(s), & x(0) &= x_0,\\ y'(s) &= \sin\vartheta(s), & y(0) &= y_0,\\ \vartheta'(s) &= \mathcal{K}(s), & \vartheta(0) &= \vartheta_0,\end{aligned} \tag{6.8}$$

where s is the arc parameter of the curve, $\vartheta(s)$ is the direction of the tangent $(x'(s), y'(s))$ and $\mathcal{K}(s)$ is the curvature at the point $(x(s), y(s))$. When $\mathcal{K}(s) := \kappa' s + \kappa$, i.e. when the curvature changes linearly, the curve is called Clothoid. As a special case, when $\kappa' = 0$ the curve has constant curvature, i.e. is a circle and when both $\kappa = \kappa' = 0$ the curve is a straight line. The solution of ODEs (6.8) is given by:

$$\begin{aligned}x(s) &= x_0 + \int_0^s \cos\left(\frac{1}{2}\kappa'\tau^2+\kappa\tau+\vartheta_0\right)\mathrm{d}\tau = x_0 + sX_0(\kappa' s^2,\kappa s,\vartheta_0),\\ y(s) &= y_0 + \int_0^s \sin\left(\frac{1}{2}\kappa'\tau^2+\kappa\tau+\vartheta_0\right)\mathrm{d}\tau = y_0 + sY_0(\kappa' s^2,\kappa' s,\vartheta_0).\end{aligned} \tag{6.9}$$

Notice that $\frac{1}{2}\kappa' s^2 + \kappa s + \vartheta_0$ and $\kappa' s + \kappa$ are, respectively, the angle and the curvature at the abscissa s. Thus, the problem considered in this chapter is stated next.

Problem 6.2 (G^1 Hermite interpolation)**.** *Given two points* (x_0, y_0) *and* (x_1, y_1) *and two angles* ϑ_0 *and* ϑ_1*, find a curve segment of the form* (6.9) *which satisfies:*

$$\begin{aligned} x(0) &= x_0, \qquad y(0) = y_0, \qquad (x'(0)\,,\, y'(0)) = (\cos\vartheta_0,\, \sin\vartheta_0), \\ x(L) &= x_1, \qquad y(L) = y_1, \qquad (x'(L), y'(L)) = (\cos\vartheta_1,\, \sin\vartheta_1), \end{aligned}$$

where $L > 0$ *is the length of the curve segment.*

The general scheme is shown in Figure 6.1 - left. Notice that Problem 6.2 admits an *infinite* number of solutions. In fact, the angle $\vartheta(s)$ of a clothoid which solves Problem 6.2 satisfies $\vartheta(0) = \vartheta_0 + 2k\pi$ and $\vartheta(L) = \vartheta_1 + 2\ell\pi$ with $k, \ell \in \mathbb{Z}$: different values of k correspond to different interpolant curves that loop around the initial and the final point. Figure 6.1 - right shows possible solutions derived from the same Hermite data.

6.4 RECASTING THE INTERPOLATION PROBLEM

The solution of Problem 6.2 by (6.9) is a zero of the following nonlinear system involving the unknowns L, κ, κ':

$$\begin{cases} x_1 - x_0 - L\, X_0(\kappa' L^2, \kappa L, \vartheta_0) = 0 \\ y_1 - y_0 - L\, Y_0(\kappa' L^2, \kappa L, \vartheta_0) = 0 \\ \vartheta_1 - \left(\frac{1}{2}\kappa' L^2 + \kappa L + \vartheta_0\right) = 0. \end{cases} \tag{6.10}$$

The third equation in (6.10) is linear so that solving it with respect to κ reduces the nonlinear system to

$$\begin{cases} x_1 - x_0 - L\, X_0(\kappa' L^2, \vartheta_1 - \vartheta_0 - \frac{1}{2}\kappa' L^2, \vartheta_0) = 0, \\ y_1 - y_0 - L\, Y_0(\kappa' L^2, \vartheta_1 - \vartheta_0 - \frac{1}{2}\kappa' L^2, \vartheta_0) = 0. \end{cases}$$

An approach based on the solution of a similar nonlinear system is proposed in reference [KFP03], while references [WM08, WM09] point out the criticity of this method by numerical examples. Introducing

$$A = \frac{1}{2}\kappa' L^2, \quad \Delta x = x_1 - x_0, \quad \Delta y = y_1 - y_0, \quad \delta = \vartheta_1 - \vartheta_0, \tag{6.11}$$

the nonlinear system is reduced to the solution of the nonlinear system of two equations in two unknowns, namely L and A:

$$\begin{cases} (*) := \Delta x - L\, X_0(2A, \delta - A, \vartheta_0) = 0, \\ (**) := \Delta y - L\, Y_0(2A, \delta - A, \vartheta_0) = 0. \end{cases} \tag{6.12}$$

Further simplification can be done using polar coordinates for $(\Delta x, \Delta y)$, namely

$$\Delta x = r\cos\varphi, \qquad \Delta y = r\sin\varphi. \tag{6.13}$$

and the well known trigonometric identities

$$\begin{aligned} \sin(\alpha - \beta) &= \sin\alpha\cos\beta - \cos\alpha\sin\beta, \\ \cos(\alpha - \beta) &= \cos\alpha\cos\beta + \sin\alpha\sin\beta. \end{aligned} \tag{6.14}$$

Table 6.1: The fitting algorithm

Function buildClothoid(x_0, y_0, ϑ_0, x_1, y_1, ϑ_1, ϵ)

1 $\Delta x \leftarrow x_1 - x_0$; $\Delta y \leftarrow y_1 - y_0$; compute r, φ from $r\cos\varphi = \Delta x$, $r\sin\varphi = \Delta y$;
2 $\phi_0 \leftarrow$ `normalizeAngle`$(\vartheta_0 - \varphi)$; $\phi_1 \leftarrow$ `normalizeAngle`$(\vartheta_1 - \varphi)$;
3 Set g as $g(A) = Y_0(2A, (\phi_1 - \phi_0) - A, \phi_0)$;
4 Set $A \leftarrow 3(\phi_1 + \phi_0)$; `// In alternative use` (6.17a) `or` (6.17b)
5 **while** $|g(A)| > \epsilon$ **do** $A \leftarrow A - g(A)/g'(A)$ $L \leftarrow r/X_0(2A, \delta - A, \phi_0)$; $\kappa \leftarrow (\delta - A)/L$; $\kappa' \leftarrow (2\,A)/L^2$;
6 **return** κ, κ', L

Function normalizeAngle(ϕ)

1 **while** $\phi > +\pi$ **do**
2 $\quad \phi \leftarrow \phi - 2\pi$
3 **end while**
4 **while** $\phi < -\pi$ **do**
5 $\quad \phi \leftarrow \phi + 2\pi$
6 **end while**
7 **return** ϕ;

From (6.13) and $L > 0$ define two nonlinear functions $f(L, A)$ and $g(A)$, where $g(A)$ does not depend on L, as follows:

$$\begin{aligned} f(L, A) &:= (*)\cdot\cos\varphi + (**)\cdot\sin\varphi = \sqrt{\Delta x^2 + \Delta y^2} - L\,h(A), \\ g(A) &:= \big((*)\cdot\sin\varphi - (**)\cdot\cos\varphi\big)/L = Y_0(2A, \delta - A, \phi_0). \end{aligned} \tag{6.15}$$

where $h(A) := X_0(2A, \delta - A, \phi_0)$, $\phi_0 = \vartheta_0 - \varphi$ and ϕ_1, used later, is defined as $\phi_1 = \vartheta_1 - \varphi$. Supposing to find A such that $g(A) = 0$, then from $f(L, A) = 0$ one computes L, κ and κ' using equations (6.15) and (6.11), respectively. Thus, the solutions of the nonlinear system (6.12) are known if the solutions of the single nonlinear function $g(A)$ of equation (6.15) are determined. The solution of Problem 6.2 is recapitulated in the following steps:

1. Solve $g(A) = 0$;
2. Compute $L = \sqrt{\Delta x^2 + \Delta y^2}/h(A)$;
3. Compute $\kappa = (\delta - A)/L$ and $\kappa' = 2A/L^2$.

This algorithm needs to compute the correct root of $g(A)$ which appropriately solves Problem 6.2 with the length L well defined and positive. These issues are discussed in section 6.7.

The complete algorithm for the clothoid computation is written in the function `buildClothoid` of Table 6.1. This function solves equation (6.15) and builds the coefficients of the interpolating clothoid.

6.5 COMPUTING THE INITIAL GUESS FOR ITERATIVE SOLUTION

The zeros of function $g(A)$ are used to solve the interpolation problem and are approximated by the Newton-Raphson scheme. This algorithm needs "a guess point" to converge to the appropriate solution. Notice that there is an infinite number of solutions of Problem 6.2 and criteria for the selection of a solution are needed. Uniqueness in appropriate range and existence of the root will be discussed in details in section 6.7.

Denote with $\mathcal{A}(\phi_0, \phi_1)$ the selected zero of $g(A)$ as a function of ϕ_0 and ϕ_1. Figure 6.2 shows that $\mathcal{A}(\phi_0, \phi_1)$ is approximated by a plane. A simple approximation of $\mathcal{A}(\phi_0, \phi_1)$ is obtained by $\sin x \approx x$ in $Y_0(2A, \delta - A, \phi_0)$ and thus,

$$g(A) = Y_0(2A, \delta - A, \phi_0) \approx \int_0^1 A\tau^2 + (\delta - A)\tau + \phi_0 \,\mathrm{d}\tau = \frac{\phi_0 + \phi_1}{2} - \frac{A}{6},$$

and solving for A,

$$\mathcal{A}(\phi_0, \phi_1) \approx 3(\phi_0 + \phi_1). \tag{6.16}$$

This approximation is a fine initial point for Newton-Raphson, however better approximation for $\mathcal{A}(\phi_0, \phi_1)$ are obtained by least squares. Invoking reflection and mirroring properties discussed in section 6.7.1 the functional form of the approximation is simplified and results in the two following possible expressions for $\mathcal{A}(\phi_0, \phi_1)$:

$$\mathcal{A}(\phi_0, \phi_1) \approx (\phi_0 + \phi_1)\Big(c_1 + c_2\overline{\phi}_0\overline{\phi}_1 + c_3(\overline{\phi}_0^2 + \overline{\phi}_1^2)\Big), \tag{6.17a}$$

$$\begin{aligned}\mathcal{A}(\phi_0, \phi_1) \approx (\phi_0 + \phi_1)\Big(&d_1 + \overline{\phi}_0\overline{\phi}_1(d_2 + d_3\overline{\phi}_0\overline{\phi}_1) + (\overline{\phi}_0^2 + \overline{\phi}_1^2)(d_4 + d_5\overline{\phi}_0\overline{\phi}_1) \\ &+ d_6(\overline{\phi}_0^4 + \overline{\phi}_1^4)\Big),\end{aligned} \tag{6.17b}$$

where $\overline{\phi}_0 = \phi_0/\pi$, $\overline{\phi}_1 = \phi_1/\pi$. The computed coefficients are reported in Table 6.2 on the left.

Using (6.16), (6.17a) or (6.17b) as the starting point for Newton-Raphson, the solution for Problem 6.2 is found in very few iterations.

The three possible guess functions and their influence in the speed up process of the algorithm were checked in a battery of tests: computing the solution with Newton-Raphson starting with the proposed guesses in a 1024×1024 grid for ϕ_0 and ϕ_1 ranging in $[-0.9999\pi, 0.9999\pi]$ with a tolerance of 10^{-10}, results in the distribution of iterations resumed in Table 6.2 on the right.

Remark 6.3. *For the Newton–Raphson method the iteration is expressed as $A_{k+1} = A_k - g(A_k)/g'(A_k)$, near the root $A^\star$ there is the following well known estimate for the error $e_k = A_k - A^\star$ when $|e_k| \leq r$:*

$$|e_{k+1}| \leq C\,|e_k|^2\,, \qquad C = \frac{\max_{A\in\mathbb{R}} |g''(A)|}{2\min_{A\in[A^\star - r, A^\star + r]} |g'(A)|}.$$

The estimate for the second derivative of $g(A)$ is trivial

$$\begin{aligned}|g''(A)| &= \left| -\int_0^1 (\tau^2 - \tau)^2 \sin(A\tau^2 + (\delta - A)\tau + \phi_0)\,\mathrm{d}\tau \right| \\ &\leq \int_0^1 (\tau^2 - \tau)^2\,\mathrm{d}\tau = \frac{1}{30}.\end{aligned} \tag{6.18}$$

Using Taylor expansion, yields the following estimate of $\min_{A\in[A^\star-r,A^\star+r]}|g'(A)|$*:*

$$\begin{aligned} g'(A) &= g'(A^\star) + (A-A^\star)g''(\zeta),\\ |g'(A)| &\geq |g'(A^\star)| - \frac{|A-A^\star|}{30}, \qquad \textit{for} \quad |A-A^\star|\leq r. \end{aligned}$$

Newton–Raphson is guaranteed to converge when $C\,|e_0|<1$*. This more restrictive condition,*

$$C\,|e_0| \leq \frac{|e_0|\max_{A\in\mathbb{R}}|g''(A)|}{2\min_{A\in[A^\star-r,A^\star+r]}|g'(A)|} \leq \frac{|e_0|}{2\,(30\,|g'(A^\star)|-|e_0|)} < 1,$$

ensures that Newton–Raphson is convergent when $|e_0|<20\,|g'(A^\star)|$*. Let* $g_{\min}$ *be the minimum value of the first derivative of* $|g'(A^\star)|$ *at the root for the angles* ϕ_0,ϕ_1*, then the computation on the previous* 1024×1024 *mesh yields*

$$g_{\min}\approx 0.0505 \tag{6.19}$$

so that the estimate of the convergence radius becomes $r=20g_{\min}\approx1.01$*. On the same mesh, the maximum distance from the computed root with the guess* (6.17b) *results in a maximum distance of about* 0.037*, well below the estimated radius.*

6.6 ACCURATE COMPUTATION OF FRESNEL MOMENTA

The computation of $g(A)$ defined in (6.15) ad $g'(A)$, employed in the Newton iteration, relies on the evaluation of integrals of kind (6.7), in fact, using integration by parts

$$g'(A) = X_1(2A,\delta-A,\phi_0) - X_2(2A,\delta-A,\phi_0). \tag{6.20}$$

Table 6.2: On the left, guess functions interpolation coefficients for guesses (6.17a) and (6.17b). On the right, iteration statistics for different guess functions.

	c	d
1	3.070645	2.989696
2	0.947923	0.716220
3	−0.673029	−0.458969
4		−0.502821
5		0.261060
6		−0.045854

Iter.	Guess (6.16)		Guess (6.17a)		Guess (6.17b)	
1	1025	0.1%	1025	0.1%	1025	0.1%
2	6882	0.7%	10710	1.0%	34124	3.2%
3	238424	22.7%	702534	66.9%	1015074	96.6%
4	662268	63.0%	336356	32.0%	402	0.1%
5	142026	13.5%				

From the trigonometric identities (6.14), integrals (6.7) are rewritten as

$$\begin{aligned} X_k(a,b,c) &= X_k(a,b,0)\cos c - Y_k(a,b,0)\sin c, \\ Y_k(a,b,c) &= X_k(a,b,0)\sin c + Y_k(a,b,0)\cos c. \end{aligned}$$

Defining $X_k(a,b) := X_k(a,b,0)$ and $Y_k(a,b) := Y_k(a,b,0)$ the computation of (6.7) is reduced to the computation of $X_k(a,b)$ and $Y_k(a,b)$. From now on, it is assumed that the standard Fresnel integrals $\mathcal{C}_0$ and $\mathcal{S}_0$ can be computed with high accuracy. For this task one can use algorithms described in reference [Bul67, Tho97] or simply use the available software [PVTF02]. It is convenient to introduce the following quantities

$$\begin{gathered} \sigma := \operatorname{sign}(a), \qquad z := \sigma\frac{\sqrt{|a|}}{\sqrt{\pi}}, \qquad \omega_+ := \frac{b+|a|}{\sqrt{\pi\,|a|}}, \\ \omega_- := \frac{b}{\sqrt{\pi\,|a|}}, \qquad \eta := -\frac{b^2}{2a}, \end{gathered} \tag{6.21}$$

so that it is possible to rewrite the argument of the trigonometric functions in $X_k(a,b)$ and $Y_k(a,b)$ as

$$\frac{a}{2}\tau^2 + b\tau = \frac{\pi}{2}\sigma\left(\tau\frac{\sigma\sqrt{|a|}}{\sqrt{\pi}} + \frac{b}{\sqrt{\pi\,|a|}}\right)^2 - \frac{b^2}{2a} = \frac{\pi}{2}\sigma\big(\tau z + \omega_-\big)^2 + \eta.$$

By using the change of variable $\xi = \tau\, z + \omega_-$ with inverse $\tau = z^{-1}(\xi - \omega_-)$ for $X_k(a,b)$ and the identity (6.14) one has:

$$\begin{aligned} X_k(a,b) &= z^{-1}\int_{\omega_-}^{\omega_+} z^{-k}(\xi-\omega_-)^k \cos\left(\sigma\frac{\pi}{2}\xi^2 + \eta\right)\,\mathrm{d}\xi \\ &= z^{-k-1}\int_{\omega_-}^{\omega_+} \sum_{j=0}^{k}\binom{k}{j}\xi^j(-\omega_-)^{k-j}\cos\left(\frac{\pi}{2}\xi^2 + \sigma\eta\right)\,\mathrm{d}\xi, \\ &= z^{-k-1}\sum_{j=0}^{k}\binom{k}{j}(-\omega_-)^{k-j}\left[\cos\eta\Delta\mathcal{C}_j - \sigma\sin\eta\Delta\mathcal{S}_j\right], \\ &= \frac{\cos\eta}{z^{k+1}}\left[\sum_{j=0}^{k}\binom{k}{j}(-\omega_-)^{k-j}\Delta\mathcal{C}_j\right] - \sigma\frac{\sin\eta}{z^{k+1}}\left[\sum_{j=0}^{k}\binom{k}{j}(-\omega_-)^{k-j}\Delta\mathcal{S}_j\right], \end{aligned} \tag{6.22}$$

where $\Delta\mathcal{C}_j = \mathcal{C}_j(\omega_+) - \mathcal{C}_j(\omega_-)$ and $\Delta\mathcal{S}_j = \mathcal{S}_j(\omega_+) - \mathcal{S}_j(\omega_-)$ are the evaluation of the momenta of the Fresnel integrals as defined in (6.4). Analogously for $Y_k(a,b)$ one has:

$$Y_k(a,b) = \frac{\sin\eta}{z^{k+1}}\left[\sum_{j=0}^{k}\binom{k}{j}(-\omega_-)^{k-j}\Delta\mathcal{C}_j\right] + \sigma\frac{\cos\eta}{z^{k+1}}\left[\sum_{j=0}^{k}\binom{k}{j}(-\omega_-)^{k-j}\Delta\mathcal{S}_j\right]. \tag{6.23}$$

This computation is inaccurate when $|a|$ is small: in fact z appears in the denominator of several fractions. For this reason, for small values of $|a|$, it is better to substitute (6.22) and (6.23) with asymptotic expansions. Notice that the recurrence (6.6) is unstable so that it produces inaccurate results for large k, but only the first two terms are needed, so this is not a problem for the computation of $g(A)$ and $g'(A)$.

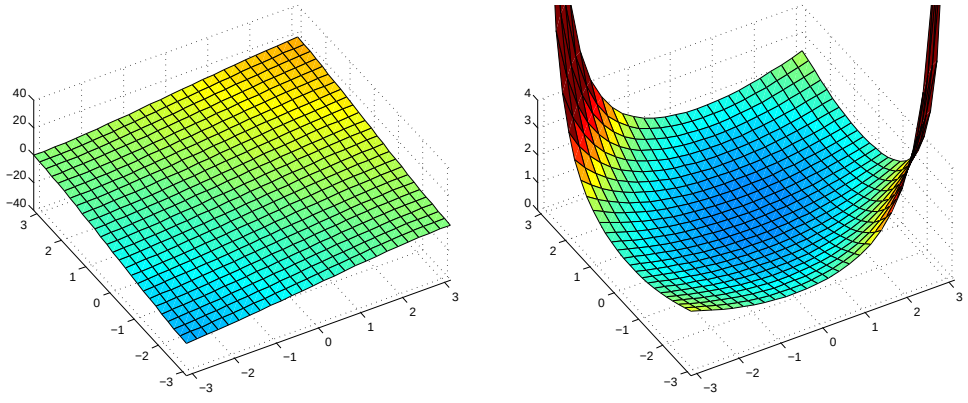


Figure 6.2: Left: the function $\mathcal{A}(\phi_0, \phi_1)$. Notice that $\mathcal{A}(\phi_0, \phi_1)$ is approximately a plane. Right: values of the length L of the computed clothoid as a function of ϕ_0 and ϕ_1. Notice that when angles satisfy $\phi_0 = \pi$, $\phi_1 = -\pi$ or $\phi_0 = -\pi$, $\phi_1 = \pi$ the length goes to infinity. The angles range in $[-\pi, \pi]$.

6.6.1 *Accurate Computation with Small Parameters*

When the parameter $|a|$ is small, identity (6.14) yields the series expansion:

$$\begin{aligned} X_k(a,b) &= \int_0^1 \tau^k \cos\left(\frac{a}{2}\tau^2 + b\tau\right) \mathrm{d}\tau \\ &= \int_0^1 \tau^k \left[\cos\left(\frac{a}{2}\tau^2\right)\cos(b\tau) - \sin\left(\frac{a}{2}\tau^2\right)\sin(b\tau)\right] \mathrm{d}\tau, \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\left(\frac{a}{2}\right)^{2n} X_{4n+k}(0,b) - \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}\left(\frac{a}{2}\right)^{2n+1} Y_{4n+2+k}(0,b), \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\left(\frac{a}{2}\right)^{2n} \left[X_{4n+k}(0,b) - \frac{a\, Y_{4n+2+k}(0,b)}{2(2n+1)}\right], \end{aligned} \tag{6.24}$$

and, analogously, using again identity (6.14):

$$Y_k(a,b) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\left(\frac{a}{2}\right)^{2n} \left[Y_{4n+k}(0,b) + \frac{a\, X_{4n+2+k}(0,b)}{2(2n+1)}\right]. \tag{6.25}$$

From the inequalities:

$$|X_k| \le \int_0^1 |\tau^k|\,\mathrm{d}\tau = \frac{1}{k+1}, \qquad |Y_k| \le \int_0^1 |\tau^k|\,\mathrm{d}\tau = \frac{1}{k+1},$$

the remainder for the series of X_k becomes:

$$\begin{aligned} R_{p,k} &= \left| \sum_{n=p}^{\infty} \frac{(-1)^n}{(2n)!} \left(\frac{a}{2}\right)^{2n} \left[X_{4n+k}(0,b) - \frac{a\, Y_{4n+2+k}(0,b)}{2(2n+1)} \right] \right| \\ &\le \sum_{n=p}^{\infty} \frac{1}{(2n)!} \left(\frac{a}{2}\right)^{2n} \left[\frac{1}{4n+1} + \frac{|a|}{2(2n+1)(4n+3)} \right] \\ &\le \left(\frac{a}{2}\right)^{2p} \sum_{n=p}^{\infty} \frac{(a/2)^{2(n-p)}}{(2(n-p))!} \\ &\le \left(\frac{a}{2}\right)^{2p} \sum_{n=0}^{\infty} \frac{1}{(2n)!} \left(\frac{a}{2}\right)^{2n} = \left(\frac{a}{2}\right)^{2p} \cosh(a). \end{aligned}$$

The same estimate is obtained for the series of Y_k. Both series (6.24) and (6.25) converge fast. For example, if $|a| < 10^{-4}$ and $p = 2$, the error is less than $6.26 \cdot 10^{-18}$ while if $p = 3$ the error is less than $1.6 \cdot 10^{-26}$.

Using a simple recurrence it is possible to compute $X_k(0,b)$ and $Y_k(0,b)$ but it turns out to be unstable. A stable computation is obtained by using an explicit formula based on the Lommel function $s_{\mu,\nu}(z)$ (see reference [SC03]). The explicit formula is:

$$\begin{aligned} X_k(0,b) &= \frac{k\, s_{k+\frac{1}{2},\frac{3}{2}}(b) \sin b + f(b) s_{k+\frac{3}{2},\frac{1}{2}}(b)}{(1+k) b^{k+\frac{1}{2}}} + \frac{\cos b}{1+k}, \\ Y_k(0,b) &= \frac{k\, s_{k+\frac{3}{2},\frac{3}{2}}(b) \sin b + f(b)(2+k) s_{k+\frac{1}{2},\frac{1}{2}}(b)}{(2+k) b^{k+\frac{1}{2}}} + \frac{\sin b}{2+k}, \end{aligned} \tag{6.26}$$

where $k = 1, 2, 3, \ldots$ and $f(b) := b^{-1} \sin b - \cos b$. The Lommel function has the following expansion (see [OLBC10] or reference [Wat44])

$$s_{\mu,\nu}(z) := z^{\mu+1} \sum_{n=0}^{\infty} \frac{(-z^2)^n}{\alpha_{n+1}(\mu,\nu)}, \qquad \alpha_n(\mu,\nu) := \prod_{m=1}^{n} ((\mu+2m-1)^2 - \nu^2), \tag{6.27}$$

and using this expansion in (6.26) results in the next explicit formula for $k = 1, 2, 3, \ldots$:

$$\begin{aligned} X_k(0,b) &= A(b) w_{k+\frac{1}{2},\frac{3}{2}}(b) + B(b) w_{k+\frac{3}{2},\frac{1}{2}}(b) + \frac{\cos b}{1+k}, \\ Y_k(0,b) &= C(b) w_{k+\frac{3}{2},\frac{3}{2}}(b) + D(b) w_{k+\frac{1}{2},\frac{1}{2}}(b) + \frac{\sin b}{2+k}, \end{aligned} \tag{6.28}$$

where

$$\begin{aligned} w_{\mu,\nu}(b) &:= \sum_{n=0}^{\infty} \frac{(-b^2)^n}{\alpha_{n+1}(\mu,\nu)}, \qquad A(b) := \frac{kb \sin b}{1+k}, \\ B(b) &:= \frac{(\sin b - b\cos b) b}{1+k}, \qquad C(b) := -\frac{b^2 \sin b}{2+k}, \\ D(b) &:= \sin b - b \cos b. \end{aligned}$$

Notice that expression (6.28) is continuous in b at $b = 0$.

6.7 THEORETICAL DEVELOPMENT

In this section existence and selection of the appropriate solution are discussed in detail. The computation of L requires only to verify that for $A^\star$ such that $g(A^\star) = 0$ then $h(A^\star) = X_0(2A^\star, \delta - A^\star, \phi_0) \neq 0$. This does not ensure that the computed L is positive; but positivity is obtained by an appropriate choice of $A^\star$.

6.7.1 *Symmetries of the Roots of* $g(A)$

The general analysis of the zeros of $g(A)$ requires the angles ϕ_0 and ϕ_1 to be in the range $(-\pi, \pi)$. It is possible to restrict the domain of search stating the following auxiliary problems:

The reversed problem The clothoid joining (x_1, y_1) to (x_0, y_0) with angles $\vartheta_0^R = -\vartheta_1$ and $\vartheta_1^R = -\vartheta_0$ is a curve with support a clothoid that solves Problem 6.2 but running in the opposite direction (with the same length L). Let $\delta^R = \vartheta_1^R - \vartheta_0^R = -\vartheta_0 + \vartheta_1 = \delta$, it follows that $g^R(A) := Y_0(2A, \delta - A, -\phi_1)$ is the function whose zeros give the solution of the reversed interpolation problem.

The mirrored problem The curve obtained connecting (x_0, y_0) to (x_1, y_1) with angle $\vartheta_0^M = \varphi - \phi_0$ and $\vartheta_1^M = \varphi - \phi_1$ is a curve with support a curve solving the same problem but mirrored along the line connecting the points (x_0, y_0) and (x_1, y_1) (with the same length L). Let $\delta^M = \vartheta_1^M - \vartheta_0^M = -\phi_1 + \phi_0 = -\delta$, it follows that $g^M(A) := Y_0(2A, -\delta - A, -\phi_0)$ is the function whose zeros are the solution of the mirrored interpolation problem.

Lemma (6.4) shows that it is possible to reduce the search of the roots in the domain $|\phi_0| < \phi_1 \leq \pi$. The special cases $\phi_0 \pm \phi_1 = 0$ are considered separately.

Lemma 6.4. *Let* $g(A)$ *and* $h(A)$ *defined in* (6.15) *with*

$$
\begin{aligned}
g^R(A) &:= Y_0(2A, \delta - A, -\phi_1), & g^M(A) &:= Y_0(2A, -\delta - A, -\phi_0),\\
h^R(A) &:= X_0(2A, \delta - A, -\phi_1), & h^M(A) &:= X_0(2A, -\delta - A, -\phi_0),
\end{aligned}
$$

then $g(A) = -g^R(-A)$, $g(A) = -g^M(-A)$, $h(A) = h^R(-A) = h^M(-A)$. *Thus,* $g(A)$ *has the same roots of* $g^R(A)$, $g^M(A)$ *with opposite sign.*

Proof. (omitted) □

Figure 6.3 shows the domain $|\phi_0| < \phi_1 \leq \pi$ with the mirrored and reversed problem. Reflecting and mirroring allows to assume the constraints for the angles described in the following Assumption 6.5.

Assumption 6.5 (Angle domain)**.** *The angles* ϕ_0 *and* ϕ_1 *satisfy the restriction:* $|\phi_0| \leq \phi_1 \leq \pi$ *with ambiguous cases* $|\phi_0| = \phi_1 = \pi$ *excluded (see Figure 6.3).*

This ordering ensures that when $|\phi_0| < \phi_1$ the curvature of the fitting curve is increasing, i.e. $\kappa' > 0$. Notice that if A is the solution of nonlinear system (6.15) then $\kappa' = 2A/L^2$, i.e. the sign of A is the sign of κ' and thus A must be positive. Finally, $\delta = \phi_1 - \phi_0 \geq 0$ with strict inequality when $|\phi_0| < \phi_1$. This assumption is not a limitation because any interpolation problem can be reformulated as a problem satisfying Assumption 6.5. The proof of existence and uniqueness of the fitting problem splits the angle domain in various subregions, while the special case $\phi_0 + \phi_1 = 0$ is performed apart.

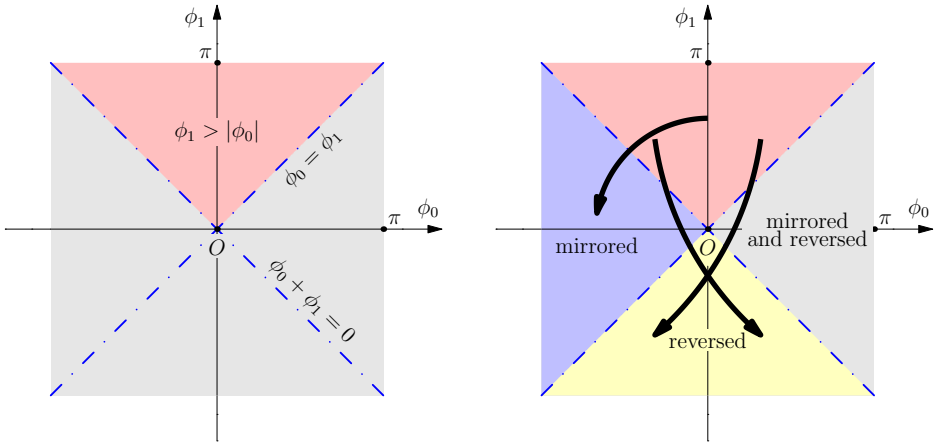


Figure 6.3: Left: the domain $\phi_1 > |\phi_0|$ with the special cases $\phi_1 = \phi_0$ and $\phi_1 + \phi_0 = 0$. Right: the domain mirrored and reversed.

6.7.2 *Localization of the Roots of* $g(A)$

The problem $g(A) = 0$ has in general infinite solutions. The next Theorems show the existence of a *unique* solution in a prescribed range, they are in part new and in part taken from [WM09], here reported without proofs and notation slightly changed to better match our notation. By appropriate transformations these Theorems permit to select the suitable solution and find the interval where the solution is unique. The transformation is contained in the following Lemma:

Lemma 6.6. *The (continuous) functions* $g(A)$ *and* $h(A)$ *defined in* (6.15) *for* $A > 0$*, when* ϕ_0 *and* ϕ_1 *satisfy assumption 6.5, can be written as*

$$g(A) = \frac{\sqrt{2\pi}}{\sqrt{A}} \begin{cases} p\left(\frac{(\delta-A)^2}{4A}\right) & 0 < A \le \delta, \\ q\left(\frac{(\delta-A)^2}{4A}\right) & A \ge \delta; \end{cases} \qquad h(A) = \frac{\sqrt{2\pi}}{\sqrt{A}} \begin{cases} \overline{p}\left(\frac{(\delta-A)^2}{4A}\right) & 0 < A \le \delta, \\ \overline{q}\left(\frac{(\delta-A)^2}{4A}\right) & A \ge \delta, \end{cases}$$

where

$$\begin{aligned} p(\theta) &= \int_\theta^{\theta+\delta} \frac{\sin(u+\phi_0-\theta)}{\sqrt{u}}\,\mathrm{d}u, \quad q(\theta) = p(\theta) + 2\int_0^\theta \frac{\sin(u+\phi_0-\theta)}{\sqrt{u}}\,\mathrm{d}u, \\ \overline{p}(\theta) &= \int_\theta^{\theta+\delta} \frac{\cos(u+\phi_0-\theta)}{\sqrt{u}}\,\mathrm{d}u, \quad \overline{q}(\theta) = \overline{p}(\theta) + 2\int_0^\theta \frac{\cos(u+\phi_0-\theta)}{\sqrt{u}}\,\mathrm{d}u. \end{aligned} \tag{6.29}$$

Proof. Standard trigonometric passages and assumption $A > 0$ yield the following expression for $g(A)$ and $h(A)$:

$$\begin{aligned} \sqrt{A}\,g(A) &= \sqrt{2\pi}\Big[\,(\mathcal{C}(\omega_+) - \mathcal{C}(\omega_-))\sin\eta + (\mathcal{S}(\omega_+) - \mathcal{S}(\omega_-))\cos\eta\Big], \\ \sqrt{A}\,h(A) &= \sqrt{2\pi}\Big[\,(\mathcal{C}(\omega_+) - \mathcal{C}(\omega_-))\cos\eta - (\mathcal{S}(\omega_+) - \mathcal{S}(\omega_-))\sin\eta\Big], \end{aligned}$$

where $\omega_\pm$ and η were previously defined in (6.21) and here take the form

$$\omega_- = \frac{\delta - A}{\sqrt{2\pi A}}, \qquad \omega_+ = \frac{\delta + A}{\sqrt{2\pi A}}, \qquad \eta = \phi_0 - \theta, \qquad \theta = \frac{(\delta - A)^2}{4A}.$$

Combining equivalence (6.3) and the parity properties of $\sin x$ and $\cos x$, $g(A)$ and $h(A)$ take the form:

$$\begin{aligned}\sqrt{A}\, g(A) &= \Delta\widehat{\mathcal{C}} \sin(\phi_0 - \theta) + \Delta\widehat{\mathcal{S}} \cos(\phi_0 - \theta), \\ \sqrt{A}\, h(A) &= \Delta\widehat{\mathcal{C}} \cos(\phi_0 - \theta) - \Delta\widehat{\mathcal{S}} \sin(\phi_0 - \theta),\end{aligned} \tag{6.30}$$

where $\Delta\widehat{\mathcal{C}} := \widehat{\mathcal{C}}(\theta + \delta) - \sigma_- \widehat{\mathcal{C}}(\theta)$, $\Delta\widehat{\mathcal{S}} := \widehat{\mathcal{S}}(\theta + \delta) - \sigma_- \widehat{\mathcal{S}}(\theta)$, $\sigma_- := \operatorname{sign}(\delta - A)$ and $\widehat{\mathcal{C}}(\theta)$ and $\widehat{\mathcal{S}}(\theta)$ are defined in (6.2). By using identities (6.14) equation (6.30) becomes:

$$\begin{aligned}\hat{g}(\theta) = \frac{\sqrt{A}}{\sqrt{2\pi}} g(A) &= \int_0^{\theta+\delta} \frac{\sin(u + \phi_0 - \theta)}{\sqrt{u}}\, \mathrm{d}u - \sigma_- \int_0^{\theta} \frac{\sin(u + \phi_0 - \theta)}{\sqrt{u}}\, \mathrm{d}u, \\ \hat{h}(\theta) = \frac{\sqrt{A}}{\sqrt{2\pi}} h(A) &= \int_0^{\theta+\delta} \frac{\cos(u + \phi_0 - \theta)}{\sqrt{u}}\, \mathrm{d}u - \sigma_- \int_0^{\theta} \frac{\cos(u + \phi_0 - \theta)}{\sqrt{u}}\, \mathrm{d}u.\end{aligned}$$

It is recalled that A must be positive, so that when A ranges into $0 < A < \delta$ then $\sigma_- = 1$, otherwise, when $A > \delta$ then $\sigma_- = -1$. In case $A = \delta$ then $\theta = 0$ and the second integral is 0 and thus $g(\delta) = p(0) = q(0)$ and $h(\delta) = \overline{p}(0) = \overline{q}(0)$. □

The next Theorems characterize the zeros of the functions (6.29) finding intervals where the solution exists and is unique.

Theorem 6.7 (see [WM09] th.2)**.** *Let $0 < -\phi_0 < \phi_1 < \pi$. If $p(0) > 0$ then $p(\theta) = 0$ has no root for $\theta \geq 0$. If $p(0) \leq 0$ then $p(\theta) = 0$ has exactly one root for $\theta \geq 0$. Moreover, the root occurs in the interval $[0, \theta^{\max}]$ where*

$$\theta^{\max} = \frac{\lambda^2}{1 - \lambda^2}(\phi_1 - \phi_0) > 0, \qquad 0 < \lambda = \frac{1 - \cos\phi_0}{1 - \cos\phi_1} < 1. \tag{6.31}$$

Theorem 6.8 (see [WM09] th.3)**.** *Let $-\pi < -\phi_1 < \phi_0 < 0$ and $q(0) > 0$ then $q(\theta) = 0$ has exactly one root in the interval $[0, \pi/2 + \phi_0]$. If $q(0) < 0$ then $q(\theta) = 0$ has no roots in the interval $[0, \pi/2 + \phi_0]$.*

Theorem 6.9 (see [WM09] th.4)**.** *Let $\phi_0 \in [0, \pi)$ and $\phi_1 \in (0, \pi]$, then $q(\theta) = 0$ has exactly one root in $[0, \pi/2 + \phi_0]$, moreover, the root occurs in $[\phi_0, \pi/2 + \phi_0]$.*

The following additional Lemmata are necessary to complete the list of properties of $p(\theta)$ and $q(\theta)$:

Lemma 6.10. *Let $p(\theta)$ and $q(\theta)$ as defined in equation* (6.29)*, then*

- (a) *if $0 \leq \phi_0 \leq \phi_1 \leq \pi$ then*
  - *if $\phi_1 > \phi_0$ then $p(\theta) > 0$ for all $\theta \geq 0$ otherwise $p(\theta) = 0$ for all $\theta \geq 0$;*
  - *$q(\theta) = 0$ for $\theta \in [\phi_0, \pi/2 + \phi_0]$ and the root is unique in $[0, \pi/2 + \phi_0]$;*
- (b) *if $-\pi \leq -\phi_1 < \phi_0 < 0$*
  - *if $p(0) = q(0) \leq 0$ then*
    - *$p(\theta) = 0$ has a unique root θ in $[0, \theta_0]$ with θ_0 defined in* (6.31)*.*
    - *$q(\theta) = 0$ has no roots in the interval $[0, \pi/2 + \phi_0]$;*
  - *if $p(0) = q(0) > 0$ then*
    - *$p(\theta) > 0$ for all $\theta \geq 0$;*
    - *$q(\theta) = 0$ has a unique root in the interval $[0, \pi/2 + \phi_0]$*
- (c) *if $\phi_0 \leq -\pi/2$ then $p(0) = q(0) < 0$.*

Proof. A direct application of Theorems 6.7, 6.8 and 6.9. For point (c), from (6.30) $p(0) = q(0) = \sqrt{\delta}\, g(\delta) = \Delta\widehat{\mathcal{C}} \sin\phi_0 + \Delta\widehat{\mathcal{S}} \cos\phi_0$, in addiction, since $-\pi \leq \phi_0 \leq -\pi/2$, both $\sin\phi_0 \leq 0$ and $\cos\phi_0 \leq 0$ resulting in $p(0) = q(0) < 0$. □

The combination of Lemma 6.4 together with reversed and mirrored problems, proves that any interpolation problem can be reduced to one which satisfies Assumption 6.5. Assumption 6.5 with Lemma 6.6 prove existence and uniqueness of $g(A) = 0$ in a specified range when $\phi_0 + \phi_1 \neq 0$. The case of $\phi_0 - \phi_1 = 0$ follows from the application of Theorem 6.9 for positive angles, because Assumption 6.5 forces $\phi_1 \geq 0$ and excludes the case of equal negative angles. The case $\phi_0 + \phi_1 = 0$ is considered in the following Lemma.

Lemma 6.11. *Let $\phi_0 + \phi_1 = 0$ and $\phi_0 \in (-\pi, \pi)$, then $g(A) = 0$ has the unique solution $A = 0$ in the interval $(-2\pi, 2\pi)$.*

Proof. For $\phi_0 + \phi_1 = 0$ one has $\delta = -2\phi_0$ and

$$\begin{aligned} g(A) &= Y_0(2A, -2\phi_0 - A, \phi_0) \\ &= \int_0^1 \sin(A\tau(\tau-1) + \phi_0(1-2\tau))\,\mathrm{d}\tau \\ &= \int_{-1}^1 \sin(A(z^2-1)/4 - z\phi_0)\frac{\mathrm{d}z}{2}, \quad [\tau = (z+1)/2] \\ &= \int_{-1}^1 \sin(A(z^2-1)/4)\cos(z\phi_0)\frac{\mathrm{d}z}{2} - \int_{-1}^1 \cos(A(z^2-1)/4)\sin(z\phi_0)\frac{\mathrm{d}z}{2}. \end{aligned}$$

Using properties of odd and even functions the rightmost integral of the previous line vanishes yielding

$$g(A) = \int_0^1 \sin(A(z^2-1)/4)\cos(z\phi_0)\,\mathrm{d}z.$$

From this last equality, if $A = 0$ then $g(A) = 0$. If $0 < |A| < 4\pi$, the sign of the quantity $\sin(A(z^2-1)/4)$ is constant; if $|\phi_0| < \pi/2$, then $\cos(z\phi_0) > 0$ and thus $g(A)$ has no roots. For the remaining values of ϕ_0, i.e. $\pi/2 \leq |\phi_0| < \pi$:

$$\int_0^{\pi/(2|\phi_0|)} \cos(z\phi_0)\,\mathrm{d}z = \frac{1}{|\phi_0|}, \qquad \int_{\pi/(2|\phi_0|)}^1 |\cos(z\phi_0)|\,\mathrm{d}z = \frac{1-\sin|\phi_0|}{|\phi_0|} < \frac{1}{|\phi_0|}.$$

If in addition, $0 < |A| < 2\pi$ then $\left|\sin(A(z^2-1)/4)\right|$ is positive and monotone decreasing so that:

$$\begin{aligned} \left| \int_0^{\pi/(2|\phi_0|)} \sin\Big(\frac{A}{4}(z^2-1)\Big)\cos(z\phi_0)\,\mathrm{d}z \right| &\geq \frac{C}{|\phi_0|}, \\ \left| \int_{\pi/(2|\phi_0|)}^1 \sin\Big(\frac{A}{4}(z^2-1)\Big)\cos(z\phi_0)\,\mathrm{d}z \right| &< \frac{C}{|\phi_0|}, \end{aligned}$$

where

$$C = \left| \sin\left(\frac{A}{16\,|\phi_0|^2}\left(\pi^2 - 4\,|\phi_0|^2\right)\right) \right| > 0,$$

and thus $g(A) \neq 0$ for $0 < |A| < 2\pi$ and $|\phi_0| < \pi$. □

It remains to proof that $h(A) > 0$ at the selected root of $g(A) = 0$. This is contained in the main Theorem of this study.

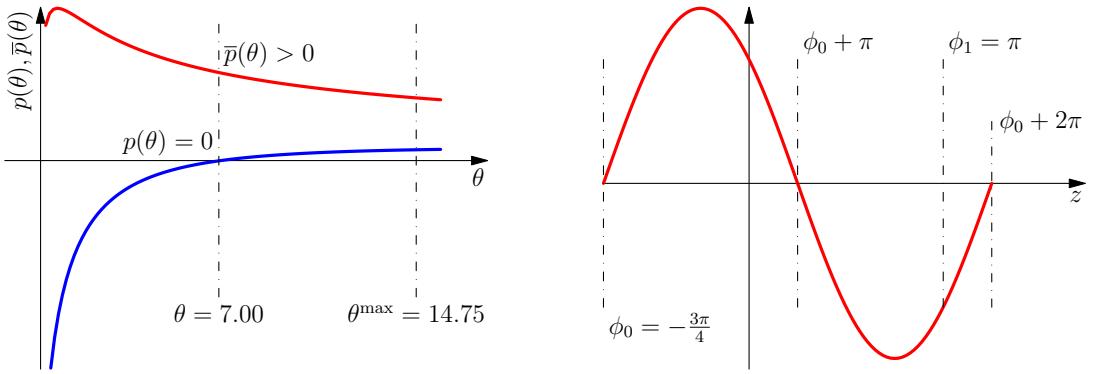


Figure 6.4: Left: functions $p(\theta)$ and $\overline{p}(\theta)$, when p vanishes, $\overline{p}$ is strictly positive. Right: the plot of $\cos z - w \sin z$. In both figures $\phi_0 = -(3/4)\pi$ and $\phi_1 = \pi$.

Theorem 6.12 (Existence and uniqueness of solution for system (6.15))**.** *The function $g(A)$ defined in* (6.15) *when angles ϕ_0 and ϕ_1 satisfy assumption 6.5, admits a unique solution for $A \in (0, A_{\max}]$, where*

$$A_{\max} = \delta + 2\theta_{\max}\left(1 + \sqrt{1 + \delta/\theta_{\max}}\right), \qquad \theta_{\max} = \max\left\{0, \pi/2 + \phi_0\right\}.$$

Moreover, $h(A) > 0$ where $g(A) = 0$.

Proof. The special cases $\phi_0 + \phi_1 = 0$ and $\phi_0 = \phi_1$ were previously considered and in Lemma 6.11. From Lemma 6.10 it follows that the two equations

$$p(\theta) = 0, \quad \text{for } \theta \geq 0, \qquad q(\theta) = 0, \quad \text{for } \theta \in [0, \theta_{\max}],$$

cannot be satisfied by the same θ in the specified range, so that they are mutually exclusive although one of the two is satisfied. Thus $g(A) = 0$ has a unique solution. To find the equivalent range of A, select the correct solution of $(\delta - A)^2 = 4A\theta_{\max}$. The two roots are:

$$\begin{aligned} A_1 &= 2\theta_{\max} + \delta - 2\sqrt{\theta_{\max}^2 + \theta_{\max}\delta} = \frac{\delta^2}{2\theta_{\max} + \delta + 2\sqrt{\theta_{\max}^2 + \theta_{\max}\delta}} \leq \delta \\ A_2 &= 2\theta_{\max} + \delta + 2\sqrt{\theta_{\max}^2 + \theta_{\max}\delta} \geq \delta, \end{aligned}$$

and thus A_2 is used to compute $A_{\max}$.

To check if $h(A) > 0$ when $g(A) = 0$, it suffices to consider the sign of $\overline{p}(\theta)$ and $\overline{q}(\theta)$. Suppose that $p(\theta) = 0$ there is to show that $\overline{p}(\theta) > 0$. For $|\phi_0| < \phi_1 \leq \frac{\pi}{2}$ the cosine in the numerator of $\overline{p}(\theta)$ is always positive, and so is the square root at the denominator, thus the integral $\overline{p}(\theta)$ is strictly positive. Now consider when $-\pi < \phi_0 < -\frac{\pi}{2}$. Using the change of variable $z + \theta - \phi_0$ in (6.29) for all $w \in \mathbb{R}$,

$$\overline{p}(\theta) = \overline{p}(\theta) + wp(\theta) = \int_{\phi_0}^{\phi_1} \frac{\cos z}{\sqrt{z + \theta - \phi_0}}\,\mathrm{d}z = \int_{\phi_0}^{\phi_1} \frac{\cos z - w \sin z}{\sqrt{z + \theta - \phi_0}}\,\mathrm{d}z. \tag{6.32}$$

In particular, it is true for $w = \frac{\cos\phi_0}{\sin\phi_0} > 0$ positive (which, incidentally is always positive because $-\pi < \phi_0 < -\frac{\pi}{2}$), so that the integrand function vanishes for the three values $z = \phi_0, \phi_0 + \pi, \phi_0 + 2\pi$. Moreover, $\cos z - w \sin z$ is strictly positive for $z \in (\phi_0, \phi_0 + \pi)$ and negative for $z \in (\phi_0 + \pi, \phi_0 + 2\pi)$, see Figure 6.4. Thus integral (6.32) can be bound as

$$\overline{p}(\theta) > \int_{\phi_0}^{\phi_0 + 2\pi} \frac{\cos z - w \sin z}{\sqrt{z + \theta - \phi_0}}\,\mathrm{d}z \geq \frac{\int_{\phi_0}^{\phi_0 + 2\pi} \cos z - w \sin z\,\mathrm{d}z}{\sqrt{(\phi_0 + \pi) + \theta - \phi_0}} = 0.$$

If $q(\theta) = 0$ there is to show that $\overline{q}(\theta) > 0$. In this case $A \geq \delta$ and from (6.30) $h(A)/\sqrt{A} = \Delta\widehat{\mathcal{C}}\cos(\phi_0 - \theta) - \Delta\widehat{\mathcal{S}}\sin(\phi_0 - \theta)$, with $\Delta\widehat{\mathcal{C}}, \Delta\widehat{\mathcal{S}} > 0$. If $\theta \in [0, \frac{\pi}{2} + \phi_0]$ then $-\frac{\pi}{2} \leq \phi_0 - \theta \leq 0$, thus the cosine is positive and the sine is negative, hence the whole quantity is strictly positive. □

Corollary 6.13. *All the solutions of the nonlinear system* (6.10) *are given by*

$$L = \frac{\sqrt{\Delta x^2 + \Delta y^2}}{X_0(2A, \delta - A, \phi_0)}, \qquad \kappa = \frac{\delta - A}{L}, \qquad \kappa' = \frac{2\,A}{L^2},$$

where A is any root of $g(A) = Y_0(2A, \delta - A, \phi_0)$ provided that the corresponding $h(A) = X_0(2A, \delta - A, \phi_0) > 0$.

Corollary 6.14. *If the angles ϕ_0 and ϕ_1 are in the range $[-\pi, \pi]$, with the exclusion of the points $\phi_0 = -\phi_1 = \pm\pi$, the solution exists and is unique for $-A_{\max} \leq A \leq A_{\max}$ where*

$$\begin{aligned} A_{\max} &= |\phi_1 - \phi_0| + 2\theta_{\max}\left(1 + \sqrt{1 + |\phi_1 - \phi_0|\,/\theta_{\max}}\right), \\ \theta_{\max} &= \max\Big\{0, \pi/2 + \operatorname{sign}(\phi_1)\phi_0\Big\}. \end{aligned}$$

6.8 NUMERICAL TESTS

The algorithm was implemented and tested in MATLAB and is available at Matlab Central [BF13]. For the Fresnel integrals computation one can use the script of [Tel05]. The first six tests are taken from reference [WM08], where the algorithm is presented; the algorithm is analysed in reference [WM09], moreover, a MATLAB implementation of the algorithm described in [WM08] is used for comparison.

Test 1 $\mathbf{p}_0 = (5, 4)$, $\mathbf{p}_1 = (5, 6)$, $\vartheta_0 = \pi/3$, $\vartheta_1 = 7\pi/6$;

Test 2 $\mathbf{p}_0 = (3, 5)$, $\mathbf{p}_1 = (6, 5)$, $\vartheta_0 = 2.14676$, $\vartheta_1 = 2.86234$;

Test 3 $\mathbf{p}_0 = (3, 6)$, $\mathbf{p}_1 = (6, 6)$, $\vartheta_0 = 3.05433$, $\vartheta_1 = 3.14159$;

Test 4 $\mathbf{p}_0 = (3, 6)$, $\mathbf{p}_1 = (6, 6)$, $\vartheta_0 = 0.08727$, $\vartheta_1 = 3.05433$;

Test 5 $\mathbf{p}_0 = (5, 4)$, $\mathbf{p}_1 = (4, 5)$, $\vartheta_0 = 0.34907$, $\vartheta_1 = 4.48550$;

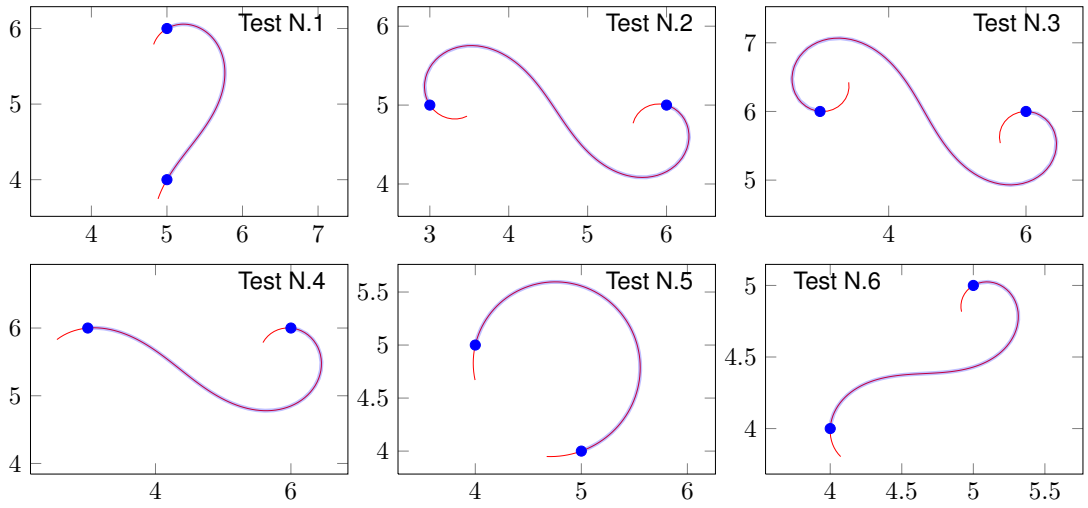
Test 6 $\mathbf{p}_0 = (4, 4)$, $\mathbf{p}_1 = (5, 5)$, $\vartheta_0 = 0.52360$, $\vartheta_1 = 4.66003$.

The accuracy of fit (as in reference [WM08]) is determined by comparing the ending point as computed by both methods, with the given ending points. A tolerance of 10^{-7} and 10^{-14} is used in the stopping criterium for Newton iterations. For all the tests and for both methods, the position error of the solution does not exceed 10^{-14}. Also iterations are comparable, with a small advantage for the present method, and are reported in the Table of Figure 6.5 which also shows the computed curves.

The difference of the present method compared with the algorithm of reference [WM08] are in the transition zone (see e.g. test N.5) where the solution is close to be a circle arc or a segment. In fact, in this situation, the present method performs better without loosing accuracy or increasing in iterations. The following tests, which represent perturbed lines and circular arcs, highlight the differences:

Test 7 $\mathbf{p}_0 = (0, 0)$, $\mathbf{p}_1 = (100, 0)$, $\vartheta_0 = 0.01 \cdot 2^{-k}$, $\vartheta_1 = -0.02 \cdot 2^{-k}$;

Test 8 $\mathbf{p}_0 = (0, -100)$, $\mathbf{p}_1 = (-100, 0)$, $\vartheta_0 = 0.00011 \cdot 2^{-k}$, $\vartheta_1 = \frac{3}{2}\pi - 0.0001 \cdot 2^{-k}$.



	Tolerance 10^{-7}						Tolerance 10^{-14}					
#test	1	2	3	4	5	6	1	2	3	4	5	6
#iter. present method	3	3	2	2	2	3	4	4	3	3	3	4
#iter. method of ref. [WM08]	4	4	4	3	5	4	5	4	5	4	24	4

Figure 6.5: Results of test N.1 up to test N.6. The thick line is the result given by the present method, the thin line is the result obtained by ref. [WM08]. The two results are indistinguishable, so the thin trajectory was prolonged beyond the endpoints to emphasise the overlapping.

Table 6.3 collects the results for $k = 1, 2, \ldots, 10$. The error is computed as the maximum of the norm of the differences at the ending point as computed by the algorithm with the given ending point. The tolerance used for the Newton iterative solver for both the algorithms is 10^{-12}. Notice that the proposed algorithm computes the solution with constant accuracy and few iterations while algorithm in [WM08] loses precision and uses more iterations. The ∞ symbol for iterations in Table 6.3 means that Newton method does not reach the required accuracy and the solution is obtained using the last computed values. The maximum number of allowed iterations was 1000. In Table 6.3, the algorithm of reference [WM08] has a large number of iteration respect to the proposed method. To understand this behaviour, notice that if $f(\theta) = 0$ is the equation to be solved used in reference [WM09] for computing the clothoid curve and $g(A) = 0$ is the equation to be solved in the present method, then the two function f and g with the respective roots $\theta^\star$ and $A^\star$ are connected by the following relations:

$$f(\theta(A))\frac{\sqrt{2\pi}}{\sqrt{A}} = g(A), \qquad \theta(A) = \frac{(\delta - A)^2}{4A}, \qquad \theta^\star = \theta(A^\star), \qquad f(\theta^\star) = g(A^\star) = 0.$$

Both algorithms uses Newton-Raphson method to approximate the roots:

$$\theta_{k+1} = \theta_k - \frac{f(\theta_k)}{f'(\theta_k)}, \qquad A_{k+1} = A_k - \frac{g(A_k)}{g'(A_k)}$$

Denoting by $\epsilon_k = \theta_k - \theta^\star$ and $e_k = A_k - A^\star$ the error near the roots $\theta^\star$ and $A^\star$, at each iteration yields:

$$\epsilon_{k+1} \approx C_f \epsilon_k^2, \qquad e_{k+1} \approx C_g e_k^2, \qquad C_f = -\frac{f''(\theta^\star)}{2f'(\theta^\star)}, \qquad C_g = -\frac{g''(A^\star)}{2g'(A^\star)}.$$

Table 6.3: Test N.7 and N.8 results

	Test N.7				Test N.8			
	Present method		Meek & Walton		Present method		Meek & Walton	
k	iter	Error	iter	Error	iter	Error	iter	Error
1	2	$2{,}6 \cdot 10^{-16}$	30	$1{,}83 \cdot 10^{-6}$	3	$3{,}18 \cdot 10^{-14}$	∞	$3{,}76 \cdot 10^{-9}$
2	2	$1{,}42 \cdot 10^{-14}$	29	$1{,}85 \cdot 10^{-6}$	3	$2{,}01 \cdot 10^{-14}$	∞	$1{,}45 \cdot 10^{-8}$
3	3	0	28	$1{,}38 \cdot 10^{-6}$	2	$2{,}01 \cdot 10^{-14}$	∞	$7{,}47 \cdot 10^{-8}$
4	2	$4{,}33 \cdot 10^{-17}$	27	$9{,}83 \cdot 10^{-7}$	2	$2{,}84 \cdot 10^{-14}$	∞	$3{,}47 \cdot 10^{-8}$
5	2	$5{,}42 \cdot 10^{-18}$	26	$6{,}96 \cdot 10^{-7}$	2	0	∞	$1{,}07 \cdot 10^{-9}$
6	2	0	25	$4{,}92 \cdot 10^{-7}$	2	$1{,}42 \cdot 10^{-14}$	∞	$5{,}53 \cdot 10^{-9}$
7	2	$1{,}35 \cdot 10^{-18}$	24	$3{,}48 \cdot 10^{-7}$	2	$5{,}12 \cdot 10^{-14}$	∞	$2{,}43 \cdot 10^{-7}$
8	2	0	23	$2{,}46 \cdot 10^{-7}$	2	0	∞	$3{,}09 \cdot 10^{-6}$
9	2	0	22	$1{,}74 \cdot 10^{-7}$	2	0	∞	$3{,}25 \cdot 10^{-6}$
10	2	0	21	$1{,}23 \cdot 10^{-7}$	2	$5{,}12 \cdot 10^{-14}$	∞	$4{,}84 \cdot 10^{-7}$

Thus, the speed of convergence of the two methods is related to the constants C_f and C_g, respectively. Large values of the constants reflect a slow convergence. Using estimates (6.19) and (6.18) of remark 6.3 the bound $|C_g| \lesssim 0.66$ is obtained. Joining this with the estimate (6.19) for the minimum of $|g'(A^\star)|$, it follows that the root is always well conditioned and the Newton method converges quickly for the proposed algorithm. Thus the proposed algorithm does not suffer of slow convergence as verified experimentally. To compare the constants C_f and C_g notice that

$$\begin{aligned} \frac{g''(A^\star)}{g'(A^\star)} &= \frac{((A^\star)^2-\delta^2)}{(2A^\star)^2}\frac{f''(\theta^\star)}{f'(\theta^\star)} - \frac{(A^\star)^2-3\delta^2}{A^\star((A^\star)^2-\delta^2)} \\ (2A^\star)^2 C_g &= ((A^\star)^2-\delta^2)C_f + 4A^\star\frac{(A^\star)^2-3\delta^2}{(A^\star)^2-\delta^2} \end{aligned} \tag{6.33}$$

moreover,

- For $A^\star \ll \delta$ equation (6.33) is approximated with $(2A^\star)^2 C_g \approx -\delta^2 C_f - 12A^\star$ and $C_f \approx -4A^\star(C_g A^\star + 3)/\delta^2$. In this case C_f is very low and the algorithm of reference [WM08] converges faster than the proposed one.
- For $A^\star \gg \delta$ equation (6.33) is approximated with $(2A^\star)^2 C_g \approx (A^\star)^2 C_f + 4A^\star$ and $C_f \approx 4(C_g - 1/A^\star)$. Thus, C_f is moderately low and the algorithm of reference [WM08] converges more and less as the proposed one.
- For $A^\star = \delta + \epsilon$ with $\epsilon \approx 0$ equation (6.33) is approximated with $4\delta^2 C_g \approx 2\delta\epsilon C_f - 4\delta^2/\epsilon$ and $C_f \approx 2\delta/\epsilon^2$. Thus, C_f may be huge for small ϵ and the algorithm of reference [WM08] converges slowly or stagnates.

This behaviour is verified in Table 6.3. Notice that when $A^\star \ll \delta$ is true that the algorithm of reference [WM08] is faster but is also true that no more than 4 iterations are necessary for the present algorithm.

6.9 AN APPLICATION

The availability of a fast and reliable routine to compute Hermite G^1 interpolation as a black box, opens the possibility of setting up more structured applications. When computing an interpolating

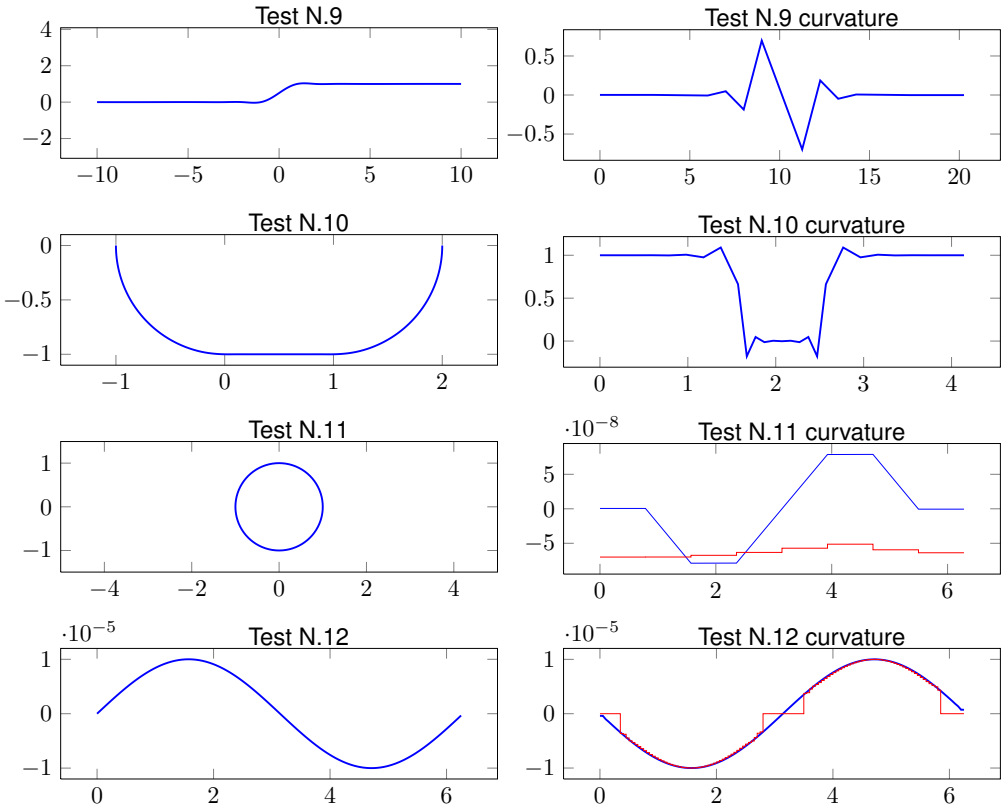


Figure 6.6: Results of the spline test N.9 up to test N.12. Left: interpolating spline. Right: arclength vs curvature. In blue the curvature of the spline obtained with present method, in red the curvature given by algorithm [WM08]. Computed curvatures between the two methods for tests N.9 and N.10 are graphically indistinguishable. In the figure of the curvature of Test N.11, the plot represents the difference from 1 of the curvature, so that eventual jumps are magnified: the red line shows that [WM08] treats those clothoids as circles yielding a piecewise constant curvature. In test N.12, in red the degenerate solution given by algorithm [WM08].

spline, it is possible to take advantage of the linear varying curvature that clothoid curves offer with respect to other splines. In order to achieve the best results in terms of continuity of the curvature, a nonlinear least square problem is set up.

Problem 6.15 (quasi G^2 fitting)**.** *Let $\mathbf{p}_j = (x_j, y_j)$ for $j = 1, \ldots, N$ be assigned points and the free parameters θ_j be the angles associated to point $\mathbf{p}_j$. For each couple of the free parameters θ_j and θ_{j+1} the G^1 Hermite interpolation problem is solved yielding the interpolating clothoid:*

$$\kappa_j = \kappa_j(\theta_j, \theta_{j+1}), \qquad \kappa'_j = \kappa'_j(\theta_j, \theta_{j+1}), \qquad L_j = L_j(\theta_j, \theta_{j+1}),$$

thus, the jump of curvature at point $\mathbf{p}_j$ for $j = 2, 3, \ldots, N-1$ is $\Delta\kappa_j = (\kappa_{j-1} + L_{j-1}\kappa'_{j-1}) - \kappa_j$. The objective function to be minimized is the sum of the squares of the jumps of the curvature at the extrema of each clothoid segments. The curvature at the first and the last point should minimize the the squares of the jump with κ_{bg} and κ_{end},

$$F(\theta_1, \theta_2, \ldots, \theta_N) =$$

$$\frac{1}{\sqrt{N}}\left((\kappa_1 - \kappa_{bg})^2 + (\kappa_{N-1} + L_{N-1}\kappa'_{N-1} - \kappa_{end})^2 + \sum_{j=2}^{N-1}\left((\kappa_{j-1} + L_{j-1}\kappa'_{j-1}) - \kappa_j\right)^2\right)^{1/2}$$

Table 6.4: Results of the interpolating clothoid splines. $\mathrm{it.}$ is the number of iterations used by the MATLAB Levenberg–Marquardt algorithm, F $\mathrm{ev.}$ is the number of evaluations of the objective function, $\mathrm{G^1ev.}$ is the number of evaluations of the routine that gives the G^1 interpolation, $F(\boldsymbol{\theta})$ is the value of the objective at the last computed point, $\mathrm{deg.}$ is the number of times data was considered degenerate. The tolerance for *lsqnonlin* was 10^{-10}.

Test	Present method				Meek & Walton				
	it.	F ev.	G^1 ev.	$F(\boldsymbol{\theta})$	it.	F ev.	G^1 ev.	$F(\boldsymbol{\theta})$	deg.
9	4	70	840	$2{,}8 \cdot 10^{-15}$	29	441	5292	$1{,}3 \cdot 10^{-04}$	290
10	4	113	2938	$1{,}0 \cdot 10^{-12}$	8	257	6682	$7{,}6 \cdot 10^{-13}$	1195
11	4	50	400	$1{,}6 \cdot 10^{-15}$	19	214	1712	$4{,}5 \cdot 10^{-09}$	1497
12	2	381	47625	$6{,}2 \cdot 10^{-20}$	27	3581	447625	$7{,}2 \cdot 10^{-07}$	447506

The quasi G^2 fitting problem *requires to find the angles* $\theta_1, \theta_2, \ldots, \theta_N$ *that minimize* $F(\theta_1, \theta_2, \ldots, \theta_N)$.

Problem 6.15 involves several times the computation of `buildClothoid`. The nonlinear solver adopted in the numerical experiments was Levenberg-Marquardt implemented in `lsqnonlin` of the Optimization Toolbox of MATLAB, no information on the Jacobian was given, hence derivatives were approximated by finite difference. This implies a heavier rely on the evaluation of the objective function itself. Four examples are herein proposed to compare the present algorithm with the algorithm of [WM08].
The tests have the following definition with MATLAB-like syntax, moreover, in all the test $\kappa_{\mathrm{begin}} = \kappa_{\mathrm{end}} = 0$ was chosen.

Test 9 $\boldsymbol{x} = [-10, -7, -4, -3, -2, -1, 0, 1, 2, 3, 4, 7, 10]$ and $\boldsymbol{y} = [0, 0, 0, 0, 0, 0, 0.5, 1, 1, 1, 1, 1, 1]$;

Test 10 $\boldsymbol{x} = [\cos(\boldsymbol{t}), 0.1 : 0.1 : 0.9, 1 - \sin(\boldsymbol{t})]$, $\boldsymbol{y} = [\sin(\boldsymbol{t}), -\mathrm{ones}(1, 9), \cos(\boldsymbol{t})]$, where $\boldsymbol{t} = [\pi : \pi/16 : (3/2)\pi]$;

Test 11 $\boldsymbol{x} = \cos([0 : \pi/4 : 2\pi]) + 10^{-7}\cos([0 : \pi/8 : \pi])$ and $\boldsymbol{y} = \sin([0 : \pi/4 : 2\pi])$;

Test 12 $\boldsymbol{x} = [0 : 0.05 : 2\pi]$ and $\boldsymbol{y} = 10^{-5}\sin([0 : 0.05 : 2\pi])$.

Graphically, they represent two line segments, two arcs of circle joined by a segment, a perturbed circle, a perturbed line (see Figure 6.6). The results are listed in Table 6.4, the residuals are very low for the present method, and the interpolating spline although computed as G^1 gives in practice a G^2 clothoid spline. The performance of the present algorithm yields a residual which is several order lower than the residual of the algorithm present in literature. In computation of tests N.9 up to N.12 the G^1 fitting using algorithm of [WM08] is close to the transition zone where data are considered degenerate and therefore approximated respectively by a circle or a straight line. This results in a low precision fitting which slows down the convergence of Levenberg-Marquardt algorithm. The last column of Table 6.4 counts the number of times the G^1 fitting are considered degenerate. Although there is no direct correlation between the number of iterations and the number of degenerate cases, it is evident that degenerate cases corrupt both the accuracy of the final solution and the convergence speed. These test cases show that the present algorithm performs well also when used as an algorithmic kernel that is repeated several times.

6.10 CONCLUSIONS

An effective solution to the problem of Hermite G^1 interpolation with a clothoid curve is herein described with a full theoretical analysis. The present algorithm does not need the decomposition in mutually exclusive states as in previous geometric works. This introduces numerical instabilities and inaccuracies as it is shown in test N.7 and N.8 of Section 6.6 or test N.11 and N.12 of Section 6.9.

The interpolation problem was reduced to one single function in one variable, making the present algorithm compact, fast and robust. A guess function which allows to find that zero with very few iterations in all possible configurations was provided. Existence and uniqueness of the solution was discussed and proved in Section 6.7. Asymptotic expansions near critical values for Fresnel related integrals are derived to keep the accuracy uniform. Implementation details of the present algorithm are given in appendix using pseudocode and can be easily translated in any programming language.

The algorithm was successfully tested in any possible situation. The accurate computation of the clothoid needs an equally accurate computation of $g(A)$ and $g'(A)$ and thus the accurate computation of Fresnel related functions $X_0(a,b,c)$ and $Y_0(a,b,c)$ with associated derivatives. These functions are a combination of Fresnel and Fresnel momenta integrals which are precise for large $|a|$ and small momenta. For the computation, only the knowledge of the first two momenta are necessary so that the inaccuracy for higher momenta does not pose any problem. A different problem is the computation of these integrals for small values of $|a|$. In this case, demanding (but stable) expansion are used to compute the Fresnel momenta with high accuracy. Finally, a theoretical proof completes the exposition and guarantees the existence of the solution in all possible cases.

The solution of the interpolation problem is uniformly accurate even when close to a straight line or an arc of circle and this was not the case of algorithms found in literature. In fact, even in domains where other algorithms solve the problem, the present method performs better in terms of accuracy and number of iterations. For example, in tests (1-6) proposed by [WM08], the present method requires 3 iterations against 4-5; in critical tests (7-8) the present algorithm converges in all cases in 2-3 iterations (against 20-30 with loss of precision, or no convergence at all after 1000 iterations). It is to point out that critical situations like those, occur in practise every time the Hermite data is acquired with (even a low) corrupting noise and no longer represents straight lines or circles, as was described in the applications of Section 6.9.

6.11 ALGORITHMS FOR THE COMPUTATION OF FRESNEL MOMENTA

In Table 6.11.1 the algorithmic version of the analytical expression derived in the chapter is herein presented. These algorithms are necessary for the computation of the main function `buildClothoid` which takes the input data (x_0, y_0, ϑ_0, x_1, y_1, ϑ_1) and returns the parameters (κ, κ', L) that solve the problem as expressed in equation (6.9). Function `GeneralizedFresnelCS` computes the generalized Fresnel integrals (6.7). It distinguishes the cases of a larger or smaller than a threshold ε. The role and the value of ε are discussed in Section 6.6. Formulas (6.22)-(6.23), used to compute $X_k(a,b)$ and $Y_k(a,b)$ at arbitrary precision when $|a| \geq \varepsilon$, are implemented in function `evalXYaLarge`. Formulas (6.24)-(6.25), used to compute $X_k(a,b)$ and $Y_k(a,b)$ at arbitrary precision when $|a| < \varepsilon$, are implemented in function `evalXYaSmall`. This function requires computation of (6.26) implemented in function `evalXYaZero` which needs (reduced) Lommel function (6.27) implemented in function `rLommel`.

6.11.1 *Pseudocode for the computation of generalized Fresnel integrals*

Pseudocode for the computation of generalized Fresnel integrals (6.7) used for the computation of (6.15) and (6.20).

Function GeneralizedFresnelCS(a, b, c, k)

$\varepsilon \leftarrow 0.01$;
if $|a| < \varepsilon$ **then** $\hat{X}, \hat{Y} \leftarrow$ evalXYaSmall(a,b,k,5) **else** $\hat{X}, \hat{Y} \leftarrow$ evalXYaLarge(a,b,k) **for** $j = 0, 1, \ldots, k-1$ **do**
  $X_j \leftarrow \hat{X}_j \cos c - \hat{Y}_j \sin c$; $Y_j \leftarrow \hat{X}_j \sin c + \hat{Y}_j \cos c$
end for
return X, Y

Function evalFresnelMomenta(t, k)

$C_0 \leftarrow \mathcal{C}(t)$; $S_0 \leftarrow \mathcal{S}(t)$;
$z \leftarrow \pi t^2/2$; $c \leftarrow \cos z$; $s \leftarrow \sin z$;
if $k > 1$ **then** $C_1 \leftarrow s/\pi$; $S_1 \leftarrow (1-c)/\pi$ **if** $k > 2$ **then** $C_2 \leftarrow (t\,s - S_0)/\pi$; $S_2 \leftarrow (C_0 - t\,c)/\pi$ **return** C, S

Function rLommel(μ, ν, b)

$t \leftarrow (\mu+\nu+1)^{-1}(\mu-\nu+1)^{-1}$;
$r \leftarrow t$; $n \leftarrow 1$; $\varepsilon \leftarrow 10^{-50}$;
while $|t| > \varepsilon\,|r|$ **do**
  $t \leftarrow t\dfrac{(-b)}{2n+\mu-\nu+1}\dfrac{b}{2n+\mu+\nu+1}$;
  $r \leftarrow r + t$; $n \leftarrow n+1$
end while
return r

Function evalXYaLarge(a, b, k)

$s \leftarrow a/|a|$; $z \leftarrow \sqrt{|a|/\pi}$; $\ell \leftarrow sb/(z\,\pi)$;
$\gamma \leftarrow -sb^2/(2\,|a|)$; $s_\gamma \leftarrow \sin\gamma$; $c_\gamma \leftarrow \cos\gamma$;
$C^+, S^+ \leftarrow$evalFresnelMomenta($\ell + z, k$);
$C^-, S^- \leftarrow$evalFresnelMomenta(z, k);
$\Delta\mathcal{C} \leftarrow C^+ - C^-$; $\Delta\mathcal{S} \leftarrow S^+ - S^-$;
$X_0 \leftarrow z^{-1}\,(c_\gamma\,\Delta\mathcal{C}_0 - s\,s_\gamma\,\Delta\mathcal{S}_0)$;
$Y_0 \leftarrow z^{-1}\,(s_\gamma\,\Delta\mathcal{C}_0 + s\,c_\gamma\,\Delta\mathcal{S}_0)$;
if $k > 1$ **then**
  $d_c \leftarrow \Delta\mathcal{C}_1 - \ell\Delta\mathcal{C}_0$;
  $d_s \leftarrow \Delta\mathcal{S}_1 - \ell\Delta\mathcal{S}_0$;
  $X_1 \leftarrow (c_\gamma\,d_c - s\,s_\gamma\,d_s)\,/z^2$;
  $Y_1 \leftarrow (s_\gamma\,d_c + s\,c_\gamma\,d_s)\,/z^2$;
end if
if $k > 1$ **then**
  $d_c \leftarrow \Delta\mathcal{C}_2 + \ell(\ell\Delta\mathcal{C}_0 - 2\Delta\mathcal{C}_1)$;
  $d_s \leftarrow \Delta\mathcal{S}_2 + \ell(\ell\Delta\mathcal{S}_0 - 2\Delta\mathcal{S}_1)$;
  $X_2 \leftarrow (c_\gamma\,d_c - s\,s_\gamma\,d_s)\,/z^3$;
  $Y_2 \leftarrow (s_\gamma\,d_c + s\,c_\gamma\,d_s)\,/z^3$;
end if
return X, Y

Function evalXYaZero(b, k)

if $|b| < \varepsilon$ **then**
  $X_0 \leftarrow 1 - \dfrac{b^2}{6}\left(1 - \dfrac{b^2}{20}\right)$;
  $Y_0 \leftarrow \dfrac{b^2}{2}\left(1 - \dfrac{b^2}{6}\left(1 - \dfrac{b^2}{30}\right)\right)$;
else
  $X_0 \leftarrow \dfrac{\sin b}{b}$;
  $Y_0 \leftarrow \dfrac{1 - \cos b}{b}$;
end if
$A \leftarrow b \sin b$;
$D \leftarrow \sin b - b \cos b$;
$B \leftarrow bD$;
$C \leftarrow -b^2 \sin b$;
for $k = 0, 1, \ldots, k-1$ **do**
  $t_1 \leftarrow \texttt{rLommel}\left(k + \frac{1}{2}, \frac{3}{2}, b\right)$;
  $t_2 \leftarrow \texttt{rLommel}\left(k + \frac{3}{2}, \frac{1}{2}, b\right)$;
  $t_3 \leftarrow \texttt{rLommel}\left(k + \frac{3}{2}, \frac{3}{2}, b\right)$;
  $t_4 \leftarrow \texttt{rLommel}\left(k + \frac{1}{2}, \frac{1}{2}, b\right)$;
  $X_{k+1} \leftarrow \dfrac{1}{1+k}\left(kA\,t_1 + B\,t_2 + \cos b\right)$;
  $Y_{k+1} \leftarrow \dfrac{1}{2+k}\left(C\,t_3 + \sin b\right) + D\,t_4$;
end for
return X, Y

Function evalXYaSmall(a, b, k, p)

$\hat{X}, \hat{Y} \leftarrow$ evalXYaZero$(b, k + 4p + 2)$;
$t \leftarrow 1$;
for $j = 0, 1, \ldots, k-1$ **do**
  $X_j \leftarrow X_j^0 - \dfrac{a}{2} Y_{j+2}^0$;
  $Y_j \leftarrow Y_j^0 + \dfrac{a}{2} X_{j+2}^0$;
end for
for $n = 1, 2, \ldots, p$ **do**
  $t \leftarrow (-t\,a^2)/(16n(2n-1))$;
  $s \leftarrow a/(4n+2)$;
  for $j = 0, 1, \ldots, k-1$ **do**
    $X_j \leftarrow X_j + t(\hat{X}_{4n+j} - s\,\hat{Y}_{4n+j+2})$;
    $Y_j \leftarrow Y_j + t(\hat{Y}_{4n+j} + s\,\hat{X}_{4n+j+2})$;
  end for
end for
return X, Y

6.12 APPENDIX: THE FITTING WITH BEZIER CUBICS

6.12.1 *Introduction to the problem*

In this appendix we approximate a given set of points with a class of G^1 curves. The set of points will not be completely random because of the next fact.

Remark 6.16. *The given set $\mathcal{P} = \{\boldsymbol{p}_0, \dots, \boldsymbol{p}_m\}$ of points to be fitted, comes from the sampling of an unknown curve $\bar{\gamma} : [a, b] \to X$ with values in $X = \mathbb{R}^2$ or $\mathbb{R}^3$, $\bar{\gamma}$ can be a closed curve ($\bar{\gamma}(a) = \bar{\gamma}(b)$).*

Since the sampled points are not exactly on the unknown curve, we can not only interpolate them with splines, instead we have to construct a piecewise defined curve that fits them using least squares. Another request we want to satisfy is the continuity of the curve and of its derivative. We have to check this piecewise, with particular attention to the knots that connect every pair of curves.

Among the various families of well known splines, we choose cubic Bezier curves. It is convenient to define a partition $\mathcal{P}$ in $\{\mathcal{P}_k\}$ such that $\cup_{k=1}^{N} \mathcal{P}_k = \mathcal{P}$ and $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset$ if $i \neq j$. We call $n_k = |\mathcal{P}_k|$ the cardinality of each set of the partition. Every $\mathcal{P}_k$ induces a vector of knots (e.g. time intervals) $\mathcal{T}_k = (t_{1,k}, \dots, t_{n_k,k})$. $\mathcal{T}_k$ can be obtained in various ways, the easiest and straight forward is linear interpolation of the points of $\mathcal{P}_k$. In general we have $t_{1,k} = 0$ and $t_{n_k,k} = 1$.

6.12.2 *Minimizing single Bezier curves*

Definition 6.17 (Bezier Curve)**.** *A Bezier curve of degree n in parametric form is defined, starting from $n+1$ points $\boldsymbol{P}_0, \dots, \boldsymbol{P}_n$ in $\mathbb{R}^M$, as*

$$\boldsymbol{B} : [0, 1] \to \mathbb{R}^M \qquad \boldsymbol{B}(t) = \sum_{i=0}^{n} \boldsymbol{P}_i \binom{n}{i} (1-t)^{n-i} t^i$$

where $(1-t)^{n-i}t^i$ is the $i-$th Bernstein's polynomial of degree n. The points $\boldsymbol{P}_i$ are the vertices of the Bernstein's polygon and are called control points*.*

Remark 6.18. *In our problem we will use cubic Bezier curve ($n = 3$), thus splines of the kind*

$$\boldsymbol{B}(t) = (1-t)^3 \boldsymbol{P}_0 + 3t(1-t)^2 \boldsymbol{P}_1 + 3t^2(1-t)\boldsymbol{P}_2 + t^3 \boldsymbol{P}_3 \tag{6.34}$$

The cubic in (6.34) passes through the points $\boldsymbol{P}_0$ and $\boldsymbol{P}_3$ but not through $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, the latter determining the tangents. Defining $\ell \boldsymbol{T}_0 = \boldsymbol{P}_1 - \boldsymbol{P}_0$ and $\ell \boldsymbol{T}_1 = \boldsymbol{P}_3 - \boldsymbol{P}_2$ where ℓ is the length of the Bezier curve and redenoting with $\boldsymbol{P}_1$ the point $\boldsymbol{P}_3$, Bezier curve becomes:

$$\begin{aligned}
\boldsymbol{B}(t) &= b_0(t)\boldsymbol{P}_0 + b_1(t)\boldsymbol{P}_1 + \ell\big(c_0(t)\boldsymbol{T}_0 + c_1(t)\boldsymbol{T}_1\big), \\
b_0(t) &= (1-t)^3, \qquad b_1(t) = t^2(3-2t), \\
c_0(t) &= 3t(1-t)^2, \qquad c_1(t) = 3t^2(t-1),
\end{aligned}$$

where $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ are fixed points. We search for the tangents $\boldsymbol{T}_0$ and $\boldsymbol{T}_1$ that minimize the error $|\boldsymbol{B}(t_i) - \boldsymbol{p}_i|$ for all $\boldsymbol{p}_i \in \mathcal{P}$. We measure the error with the sum of the square of the difference between the approximating spline and the points $\boldsymbol{p}_i \in \mathcal{P}$, i.e. the square of the standard deviation:

$$S(\boldsymbol{T}_0, \boldsymbol{T}_1) = \frac{1}{m+1} \sum_{i=0}^{m} \|\boldsymbol{B}(t_i) - \boldsymbol{p}_i\|^2 . \tag{6.35}$$

Lemma 6.19. *The tangents $\boldsymbol{T}_0$ and $\boldsymbol{T}_1$ that minimize* (6.35) *are functions of $\boldsymbol{p}_{i,k} \in \mathcal{P}$.*

Proof. Differentiating $S(\boldsymbol{T}_0, \boldsymbol{T}_1)$ with respect to the two variables $\boldsymbol{T}_0$ and $\boldsymbol{T}_1$, we obtain a linear system, that can even be solved explicitly. Formally we have

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{T}_0} &= \frac{2}{m+1}\sum_{i=0}^{m} \ell c_0(t_i)(\boldsymbol{B}(t_i) - \boldsymbol{p}_i)\\ \frac{\partial S}{\partial \boldsymbol{T}_1} &= \frac{2}{m+1}\sum_{i=0}^{m} \ell c_1(t_i)(\boldsymbol{B}(t_i) - \boldsymbol{p}_i)\end{aligned}$$

□

6.12.3 *Minimizing piecewise Bezier curve*

Definition 6.20. *A curve $\gamma(t)$ is geometrically continuous (G^n) in a point $t \in [t_0, t_1)$ if exists a parametrization such that the resulting curve is C^n.*

Lemma 6.21 (Continuity in the joints)**.** *The necessary condition for continuity in the joint point of two splines segments $B_{k-1}(t)$ and $B_k(t)$ is*

$$\boldsymbol{P}_{0,k} = \boldsymbol{P}_{3,k-1} \tag{6.36}$$

to have also a continuous derivative must hold

$$\boldsymbol{P}_{1,k} = 2\boldsymbol{P}_{3,k-1} - \boldsymbol{P}_{2,k-1} \tag{6.37}$$

For our scope is enough geometric continuity G^1, thus we will require

$$\boldsymbol{P}_{3,k-1} = \alpha\boldsymbol{P}_{1,k} + (1-\alpha)\boldsymbol{P}_{2,k-1} \tag{6.38}$$

with $\alpha \in (0, 1)$ this means that in the point of connection of two splines the tangents be parallel (proportional) but not necessarily the same.

Proof. It is easy to prove the continuity of $\boldsymbol{B}_{k-1}$ and $\boldsymbol{B}_k(t)$, in facts it should hold $\boldsymbol{B}_{k-1}(1) = \boldsymbol{B}_k(0)$, and this is achieved when $\boldsymbol{P}_{3,k-1} = \boldsymbol{P}_{0,k}$, which is exactly (6.36). To check the continuity of the derivative, we must see if $(1-\alpha)\boldsymbol{B}'_{k-1}(1) = \alpha\boldsymbol{B}'_k(0)$. The derivative of $\boldsymbol{B}_k(t)$ is

$$\begin{aligned}\boldsymbol{B}'_k(t) &= -3(1-t)^2\boldsymbol{P}_{0,k} + (9t^2 - 12t + 3)\boldsymbol{P}_{1,k}\\ &\qquad +(6t - 9t^2)\boldsymbol{P}_{2,k} + 3t^2\boldsymbol{P}_{3,k},\\ \boldsymbol{B}'_k(0) &= 3(\boldsymbol{P}_{1,k} - \boldsymbol{P}_{0,k}),\\ \boldsymbol{B}'_{k-1}(1) &= 3(\boldsymbol{P}_{3,k-1} - \boldsymbol{P}_{2,k-1})\end{aligned}$$

now imposing $\alpha\boldsymbol{B}'_{k-1}(1) = (1-\alpha)\boldsymbol{B}'_k(0)$ and substituting (6.36) yields

$$(1-\alpha)(\boldsymbol{P}_{3,k-1} - \boldsymbol{P}_{2,k-1}) = \alpha(\boldsymbol{P}_{1,k} - \boldsymbol{P}_{0,k})$$

and simplifying terms

$$\boldsymbol{P}_{3,k-1} = \alpha\boldsymbol{P}_{1,k} + (1-\alpha)\boldsymbol{P}_{2,k-1}.$$

Equation (6.38) implies the existence of $\boldsymbol{T}_k$ such that $P_{1,k} = \boldsymbol{P}_k + \alpha\boldsymbol{T}_k$. □

Therefore we can rewrite (6.34) putting the constraints to be G^1 as in (6.36) and (6.37), hence the new Bezier curve becomes

$$\boldsymbol{B}_k(t) = c_0(t)\boldsymbol{P}_k + \ell_k d_0(t)\boldsymbol{T}_k + c_1(t)\boldsymbol{P}_{k+1} + \ell_k d_1(t)\boldsymbol{T}_{k+1}$$

where

$$\begin{aligned} c_0(t) &= b_0(t) + b_1(t) \\ d_0(t) &= b_1(t) \\ c_1(t) &= b_2(t) + b_3(t) \\ d_1(t) &= -b_2(t) \end{aligned}$$

$$\begin{aligned} \boldsymbol{P}_{0,k} &= \boldsymbol{P}_k \\ \boldsymbol{P}_{1,k} &= \boldsymbol{P}_k + \ell_k \boldsymbol{T}_k \\ \boldsymbol{P}_{2,k} &= \boldsymbol{P}_{k+1} - \ell_k \boldsymbol{T}_{k+1} \\ \boldsymbol{P}_{3,k} &= \boldsymbol{P}_{k+1} \end{aligned}$$

Theorem 6.22. *The control points for a piecewise weighted Bezier curve with G^1 continuity conditions can be calculated by minimizing*

$$S = \sum_{k=1}^{N} w_k S_k = \sum_{k=1}^{N} w_k \left(\frac{1}{2} \sum_{i=1}^{n_k} \|\boldsymbol{B}_k(t_{i,k}) - \boldsymbol{p}_{i,k}\|^2 \right).$$

where $w_k > 0$, usually $w_k = 1/n_k$.

6.12.4 *Proof of the theorem*

We start by proving the cyclic case, i.e. when the initial point is equal to the final point (e.g. $\bar{\gamma}$ is a closed curve). The cases with fixed or free extrema are very similar and differs only in the definition of the initial and final spline.

$$S = \sum_{k=1}^{N} w_k S_k = \frac{1}{2} \sum_{k=1}^{N} w_k \sum_{i=1}^{n_k} \|\boldsymbol{B}_k(t_{i,k}) - \boldsymbol{p}_{i,k}\|^2$$

Consider only S_k

$$\begin{aligned} S_k &= \frac{1}{2} \sum_{i=1}^{n_k} \|\boldsymbol{B}_k(t_{i,k}) - \boldsymbol{p}_{i,k}\|^2, \\ &= \frac{1}{2} \sum_{i=1}^{n_k} \left(\boldsymbol{B}_k(t_{i,k})^T \boldsymbol{B}_k(t_{i,k}) + \boldsymbol{p}_{i,k}^T \boldsymbol{p}_{i,k} - 2\boldsymbol{p}_{i,k}^T \boldsymbol{B}_k(t_{i,k}) \right) \end{aligned}$$

we see that

$$\begin{aligned} \boldsymbol{B}_k(t_{i,k}) &= \boldsymbol{P}_{k-1} c_0(t_{i,k}) + \ell_k \boldsymbol{T}_{k-1} d_0(t_{i,k}) + \boldsymbol{P}_k c_1(t_{i,k}) + \ell_k \boldsymbol{T}_k d_1(t_{i,k}), \\ &= \begin{bmatrix} \begin{pmatrix} c_0(t_{i,k}) & \ell_k d_0(t_{i,k}) & c_1(t_{i,k}) & \ell_k d_1(t_{i,k}) \end{pmatrix} \otimes \boldsymbol{I} \end{bmatrix} \begin{pmatrix} \boldsymbol{P}_{k-1} \\ \boldsymbol{T}_{k-1} \\ \boldsymbol{P}_k \\ \boldsymbol{T}_k \end{pmatrix} \end{aligned}$$

therefore

$$S_k = \frac{1}{2} \left(\begin{array}{c} \boldsymbol{P}_{k-1} \\ \boldsymbol{T}_{k-1} \\ \hline \boldsymbol{P}_k \\ \boldsymbol{T}_k \end{array} \right)^T \left(\boldsymbol{A}_k \otimes \boldsymbol{I} \right) \left(\begin{array}{c} \boldsymbol{P}_{k-1} \\ \boldsymbol{T}_{k-1} \\ \hline \boldsymbol{P}_k \\ \boldsymbol{T}_k \end{array} \right) - \boldsymbol{b}_k^T \left(\begin{array}{c} \boldsymbol{P}_{k-1} \\ \boldsymbol{T}_{k-1} \\ \hline \boldsymbol{P}_k \\ \boldsymbol{T}_k \end{array} \right) - c$$

where

$$\boldsymbol{A}_k = \left(\begin{array}{cc|cc} \sum c_0c_0 & \ell_k\sum c_0d_0 & \sum c_0c_1 & \ell_k\sum c_0d_1 \\ \ell_k\sum d_0c_0 & \ell_k^2\sum d_0d_0 & \ell_k\sum d_0c_1 & \ell_k^2\sum d_0d_1 \\ \hline \sum c_1c_0 & \ell_k\sum c_1d_0 & \sum c_1c_1 & \ell_k\sum c_1d_1 \\ \ell_k\sum d_1c_0 & \ell_k^2\sum d_1d_0 & \ell_k\sum d_1c_1 & \ell_k^2\sum d_1d_1 \end{array}\right)$$

$$\boldsymbol{b}_k = \left(\begin{array}{c} \sum c_0(t_{i,k})\boldsymbol{p}_{i,k} \\ \ell_k\sum d_0(t_{i,k})\boldsymbol{p}_{i,k} \\ \hline \sum c_1(t_{i,k})\boldsymbol{p}_{i,k} \\ \ell_k\sum d_1(t_{i,k})\boldsymbol{p}_{i,k} \end{array}\right)^T, \quad c = \sum \boldsymbol{p}_{i,k}^T\boldsymbol{p}_{i,k}$$

the points candidate to be minima are those that $\nabla S = \boldsymbol{0}$. Because of

$$\left(\nabla_{k-1}S_k\ \nabla_k S_k^T\right)^T = \Big(\boldsymbol{A}_k \otimes \boldsymbol{I}\Big)\left(\begin{array}{c} \boldsymbol{P}_{k-1} \\ \boldsymbol{T}_{k-1} \\ \hline \boldsymbol{P}_k \\ \boldsymbol{T}_k \end{array}\right) - \boldsymbol{b}_k$$

where

$$\nabla_k = (\partial_{\boldsymbol{P}_k}^T\ \partial_{\boldsymbol{T}_k}^T), \quad \nabla = (\nabla_0\ \nabla_1\ \cdots\ \nabla_N),$$

hence

$$\nabla S = \sum_{k=1}^{N} w_k \nabla S_k = \begin{pmatrix} w_1\nabla_0 S_1 \\ w_1\nabla_1 S_1 + w_2\nabla_1 S_2 \\ w_2\nabla_2 S_2 + w_3\nabla_2 S_3 \\ \vdots \\ w_{N-1}\nabla_{N-1}S_{N-1} + w_N\nabla_{N-1}S_N \\ w_N\nabla_N S_N \end{pmatrix}$$

we can write $\nabla S = \boldsymbol{M}\boldsymbol{x} - \boldsymbol{Q}$ where $\boldsymbol{M}$ is the matrix of the coefficients, $\boldsymbol{x}$ is the vector of the unknowns, $\boldsymbol{Q}$ is the vector of constants, they will be described better later. Let us expand some terms of the sum over k in order to see what happens at the initial and final points.

$$\begin{aligned} S \;=\; & \frac{w_1}{2}\sum_{i=1}^{n_1}\left\|\boldsymbol{P}_0c_0 + \ell_1\boldsymbol{T}_0d_0 + \boldsymbol{P}_1c_1 + \ell_1\boldsymbol{T}_1d_1 - \boldsymbol{p}_{i,1}\right\|^2 \\ & + \cdots + \\ & + \frac{w_k}{2}\sum_{i=1}^{n_k}\left\|\boldsymbol{P}_{k-1}c_0 + \ell_k\boldsymbol{T}_{k-1}d_0 + \boldsymbol{P}_kc_1 + \ell_k\boldsymbol{T}_kd_1 - \boldsymbol{p}_{i,k-1}\right\|^2 \\ & + \frac{w_{k+1}}{2}\sum_{i=1}^{n_{k+1}}\left\|\boldsymbol{P}_kc_0 + \ell_{k+1}\boldsymbol{T}_kd_0 + \boldsymbol{P}_{k+1}c_1 + \ell_{k+1}\boldsymbol{T}_{k+1}d_1 - \boldsymbol{p}_{i,k}\right\|^2 \\ & + \cdots + \\ & + \frac{w_N}{2}\sum_{i=1}^{n_N}\left\|\boldsymbol{P}_{N-1}c_0 + \ell_N\boldsymbol{T}_{N-1}d_0 + \boldsymbol{P}_Nc_1 + \ell_N\boldsymbol{T}_Nd_1 - \boldsymbol{p}_{i,N}\right\|^2 \end{aligned}$$

in the cyclic case $\boldsymbol{P}_N = \boldsymbol{P}_0$, $\boldsymbol{T}_N = \boldsymbol{T}_0$. The partial derivatives for variables $\boldsymbol{P}_j, \boldsymbol{T}_j$ are:

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{P}_0} &= w_1 \sum_{i=1}^{n_1} (\boldsymbol{P}_0 c_0 + \ell_1 \boldsymbol{T}_0 d_0 + \boldsymbol{P}_1 b_1 + \ell_1 \boldsymbol{T}_2 d_1 - \boldsymbol{p}_{i,1}) c_0 \\ &+ w_N \sum_{i=1}^{n_N} (\boldsymbol{P}_{N-1} c_0 + \ell_N \boldsymbol{T}_{N-1} d_0 + \boldsymbol{P}_0 c_1 + \ell_1 \boldsymbol{T}_0 d_1 - \boldsymbol{p}_{i,N}) c_1 \\ \frac{\partial S}{\partial \boldsymbol{T}_0} &= \ell_1 w_1 \sum_{i=1}^{n_1} (\boldsymbol{P}_0 c_0 + \ell_1 \boldsymbol{T}_0 d_0 + \boldsymbol{P}_1 c_1 + \ell_1 \boldsymbol{T}_1 d_1 - \boldsymbol{p}_{i,1}) d_0 \\ &+ \ell_N w_N \sum_{i=1}^{n_N} (\boldsymbol{P}_{N-1} c_0 + \ell_N \boldsymbol{T}_{N-1} d_0 + \boldsymbol{P}_0 c_1 + \ell_1 \boldsymbol{T}_0 d_1 - \boldsymbol{p}_{i,N}) d_1\end{aligned}$$

In general we have

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{P}_k} &= w_{k-1} \sum_{i=1}^{n_{k-1}} c_1(t_{i,k-1}) \boldsymbol{B}_{k-1}(t_{i,k-1}) \\ &+ w_k \sum_{i=1}^{n_k} c_0(t_{i,k}) \boldsymbol{B}_k(t_{i,k}) \\ \frac{\partial S}{\partial \boldsymbol{T}_k} &= w_{k-1} \ell_{k-1} \sum_{i=1}^{n_{k-1}} d_1(t_{i,k-1}) \boldsymbol{B}_{k-1}(t_{i,k-1}) \\ &+ w_k \ell_k \sum_{i=1}^{n_k} d_0(t_{i,k}) \boldsymbol{B}_k(t_{i,k})\end{aligned}$$

The result is a tridiagonal block system 2×2 with corners, we denote it with $\boldsymbol{M}\boldsymbol{x} = \boldsymbol{Q}$. The vector of the unknowns is $\boldsymbol{x} = (\boldsymbol{P}_0, \boldsymbol{T}_0, \boldsymbol{P}_2, \boldsymbol{T}_2, \ldots, \boldsymbol{P}_{N-1}, \boldsymbol{T}_{N-1})^T$, matrix $\boldsymbol{M}$ and vector $\boldsymbol{Q}$ are the following.

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{L}_1^T & 0 & \ldots & \ldots & \ldots & \boldsymbol{L}_N \\ \boldsymbol{L}_1 & \boldsymbol{D}_2 & \boldsymbol{L}_2^T & 0 & \ldots & \ldots & 0 \\ 0 & \boldsymbol{L}_2 & \boldsymbol{D}_3 & \boldsymbol{L}_3^T & 0 & \ldots & 0 \\ & & & \vdots & & & \\ 0 & & & & \boldsymbol{L}_{N-2} & \boldsymbol{D}_{N-1} & \boldsymbol{L}_{N-1}^T \\ \boldsymbol{L}_N^T & & & & 0 & \boldsymbol{L}_{N-1} & \boldsymbol{D}_N \end{pmatrix}$$

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_1 \\ \boldsymbol{Q}_2 \\ \vdots \\ \boldsymbol{Q}_{N-1} \\ \boldsymbol{Q}_N \end{pmatrix}$$

The single blocks are respectively (omitting the dependance on the knots $t_{i,k}$)

$$\begin{aligned}
\boldsymbol{D}_1 &= w_1 \sum_{i=1}^{n_1} \begin{pmatrix} c_0c_0 & \ell_1 c_0 d_0 \\ \ell_1 c_0 d_0 & \ell_1^2 d_0 d_0 \end{pmatrix} + w_N \sum_{i=1}^{n_N} \begin{pmatrix} c_1 c_1 & \ell_N c_1 d_1 \\ \ell_N c_1 d_1 & \ell_N^2 d_1 d_1 \end{pmatrix} \\
\boldsymbol{D}_k &= w_{k-1} \sum_{i=1}^{n_{k-1}} \begin{pmatrix} c_1 c_1 & \ell_{k-1} c_1 d_1 \\ \ell_{k-1} c_1 d_1 & \ell_{k-1}^2 d_1 d_1 \end{pmatrix} + w_k \sum_{i=1}^{n_k} \begin{pmatrix} c_0 c_0 & \ell_k c_0 d_0 \\ \ell_k c_0 d_0 & \ell_k^2 d_0 d_0 \end{pmatrix} \\
\boldsymbol{L}_N &= w_N \sum_{i=1}^{n_N} \begin{pmatrix} c_0 c_1 & \ell_N c_1 d_0 \\ \ell_N c_0 d_1 & \ell_N^2 d_0 d_1 \end{pmatrix} \\
\boldsymbol{L}_k &= w_k \sum_{i=1}^{n_k} \begin{pmatrix} c_0 c_1 & \ell_k c_1 d_0 \\ \ell_k c_0 d_1 & \ell_k^2 d_0 d_1 \end{pmatrix}
\end{aligned}$$

Finally the vector of the constants $\boldsymbol{Q}$ is

$$\begin{aligned}
\boldsymbol{Q}_1 &= w_1 \sum_{i=1}^{n_1} \begin{pmatrix} c_0 \boldsymbol{p}_{i,1} \\ \ell_1 d_0 \boldsymbol{p}_{i,1} \end{pmatrix} + w_N \sum_{i=1}^{n_N} \begin{pmatrix} c_1 \boldsymbol{p}_{i,N} \\ \ell_N d_1 \boldsymbol{p}_{i,N} \end{pmatrix} \\
\boldsymbol{Q}_k &= w_{k-1} \sum_{i=1}^{n_{k-1}} \begin{pmatrix} c c_1 \boldsymbol{p}_{i,k-1} \\ \ell_{k-1} d_1 \boldsymbol{p}_{i,k-1} \end{pmatrix} + w_k \sum_{i=1}^{n_N} \begin{pmatrix} c_0 \boldsymbol{p}_{i,k} \\ \ell_k d_0 \boldsymbol{p}_{i,k} \end{pmatrix}
\end{aligned}$$

We can notice that $\boldsymbol{M}$ is a symmetric matrix because the blocks $\boldsymbol{D}_k$ are symmetric.

We treat now the non-cyclic case, first with free extrema. This time the matrix $\boldsymbol{M}$ will be $2N + 2 \times 2N + 2$ because we do not connect the first and last point anymore. The central blocks of $\boldsymbol{M}$ are the same. We have only to redefine the first and last row, and the corresponding entries in $\boldsymbol{Q}$.

$$\begin{aligned}
\boldsymbol{D}_1 &= w_1 \sum_{i=1}^{n_1} \begin{pmatrix} c_0 c_0 & \ell_1 c_0 d_0 \\ \ell_1 c_0 d_0 & \ell_1^2 d_0 d_0 \end{pmatrix} \quad \boldsymbol{L}_N = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\
\boldsymbol{L}_1^T &= w_1 \sum_{i=1}^{n_1} \begin{pmatrix} \ell_1 c_0 d_1 & c_0 c_1 \\ \ell_1 c_1 d_0 & \ell_1^2 d_0 d_1 \end{pmatrix} \\
\boldsymbol{Q}_1 &= w_1 \sum_{i=1}^{n_1} \begin{pmatrix} c_0 \boldsymbol{p}_{i,1} \\ \ell_1 d_0 \boldsymbol{p}_{i,1} \end{pmatrix} \quad \boldsymbol{Q}_{N+1} = w_N \sum_{i=1}^{n_N} \begin{pmatrix} c_1 \boldsymbol{p}_{i,N} \\ \ell_N d_1 \boldsymbol{p}_{i,N} \end{pmatrix}
\end{aligned}$$

It remains the case with fixed extrema.

$$\begin{aligned}
\boldsymbol{D}_1 &= \begin{pmatrix} 1 & 0 \\ 0 & w_1 \ell_1^2 \sum_{i=1}^{n_1} d_0 d_0 \end{pmatrix} \quad \boldsymbol{D}_{N+1} = \begin{pmatrix} 1 & 0 \\ 0 & w_N \ell_N^2 \sum_{i=1}^{n_N} d_1 d_1 \end{pmatrix} \\
\boldsymbol{L}_{N+1} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad \boldsymbol{L}_1^T = w_1 \sum_{i=1}^{n_1} \begin{pmatrix} 0 & 0 \\ \ell_1 c_1 d_0 & \ell_1^2 d_0 d_1 \end{pmatrix} \\
\boldsymbol{Q}_1 &= \begin{pmatrix} \boldsymbol{p}_0 \\ w_1 \ell_1 \sum_{i=1}^{n_1} d_0 \boldsymbol{p}_{i,1} \end{pmatrix} \qquad \boldsymbol{Q}_{N+1} = \begin{pmatrix} \boldsymbol{p}_m \\ w_N \ell_N \sum_{i=1}^{n_N} d_1 \boldsymbol{p}_{i,N} \end{pmatrix}
\end{aligned}$$

We are interested in a result of existence and unicity of the solution. We can notice that $\boldsymbol{M}$ is symmetric and positive definite. There is to prove that $\boldsymbol{M}$ is strictly positive definite: S is the sum

of the square of the error in each segment, thus $S = 0$ if and only if all $S_k = 0$. We have to find when S_k is not zero, and this is clearly true when $n_k \geq 4$.

$$\begin{aligned} 2S_k &= \sum_{i=1}^{n_k} \|\boldsymbol{B}_k(t_{i,k}) - \boldsymbol{p}_{i,k}\|^2 , \\ &= \sum_{i=1}^{n_k} \|c_0(t_{i,k})\boldsymbol{P}_k + d_0(t_{i,k})\ell_k\boldsymbol{T}_k + c_1(t_{i,k})\boldsymbol{P}_{k+1} + d_1(t_{i,k})\ell_k\boldsymbol{T}_{k+1} - \boldsymbol{p}_{i,k}\|^2 . \end{aligned}$$

The $k-$th term is

$$\left\| ((c_0, d_0, c_1, d_1) \otimes \boldsymbol{I}) \begin{pmatrix} \boldsymbol{P}_k \\ \ell_k\boldsymbol{T}_k \\ \boldsymbol{P}_{k+1} \\ \ell_k\boldsymbol{T}_{k+1} \end{pmatrix} - \boldsymbol{p}_{i,k}^T \right\|^2 .$$

We put $\boldsymbol{p}_i = 0$ and check when the product vanishes. Writing the previous relations in matrix form yields

$$\left(\begin{pmatrix} c_0(t_{1,k}) & d_0(t_{1,k}) & c_1(t_{1,k}) & d_1(t_{1,k}) \\ c_0(t_{2,k}) & d_0(t_{2,k}) & c_1(t_{2,k}) & d_1(t_{2,k}) \\ \cdots & \cdots & \cdots & \cdots \\ c_0(t_{n_k,k}) & d_0(t_{n_k,k}) & c_1(t_{n_k,k}) & d_1(t_{n_k,k}) \end{pmatrix} \otimes \boldsymbol{I} \right) \begin{pmatrix} \boldsymbol{P}_k \\ \ell_k\boldsymbol{T}_k \\ \boldsymbol{P}_{k+1} \\ \ell_k\boldsymbol{T}_{k+1} \end{pmatrix} = \boldsymbol{0}$$

We want that the unique solution of this linear system be the trivial one. The product is non-zero if the left matrix is full rank, i.e. there exist at least 4 linearly independent rows. This is true if there are at least four distinct knots where we evaluate polynomials c_j, d_j. This completes the proof.

6.12.5 *An Example: reconstruction of the track of Spa-Francorchamps*

We give a final example of the road reconstruction with G^1 Bezier curves, with G^1 clothoids, and with quasi G^2 clothoids. The first picture of Figure 6.7 is the cloud of points obtained by the GPS, the middle picture represents the reconstruction with Bezier curves with cyclic boundary conditions, the right picture with non cyclic conditions.

Figure 6.8 shows the comparison of the reconstruction with non cyclic Bezier curves and with G^1

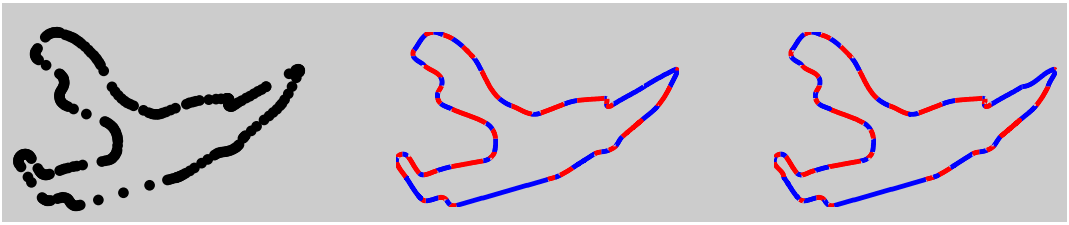


Figure 6.7: From the left: GPS points, fitting with G^1 with cyclic conditions, fitting without cyclic conditions.

clothoids.

Figure 6.9 shows the superposition of the original GPS data with the fitting with clothoids.

Figure 6.10 shows the interpolation of the original points with a quasi G^2 clothoid, the picture below shows the curvature of the fitting. We notice the peaks of the curvature in correspondence of the U curve after Les Combes and the famous La Source curve just before the Eau Rouge Raidillon. The last Figure 6.11 shows the trajectory projected back on the surface of the Earth.

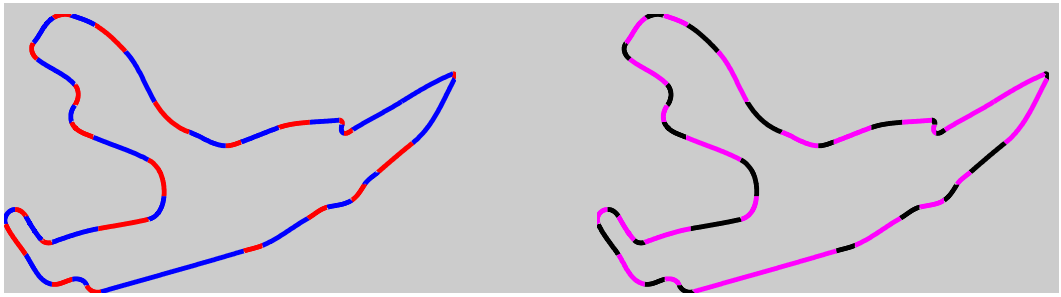


Figure 6.8: Left: G^1 Bezier, right: G^1 clothoids.

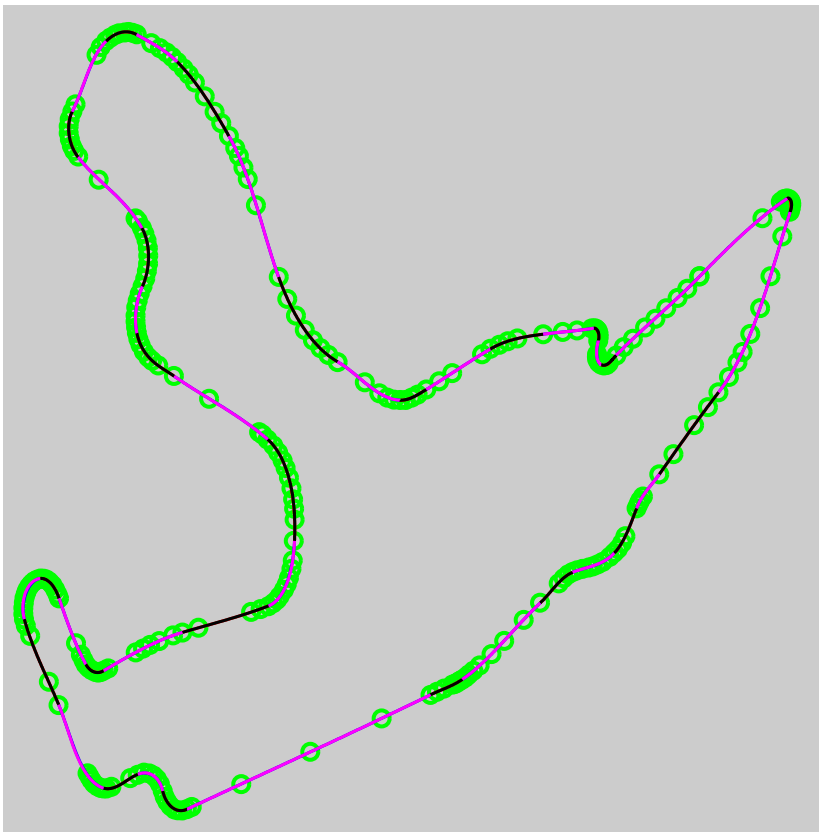


Figure 6.9: The dots are the original GPS data fitted with the clothoids.

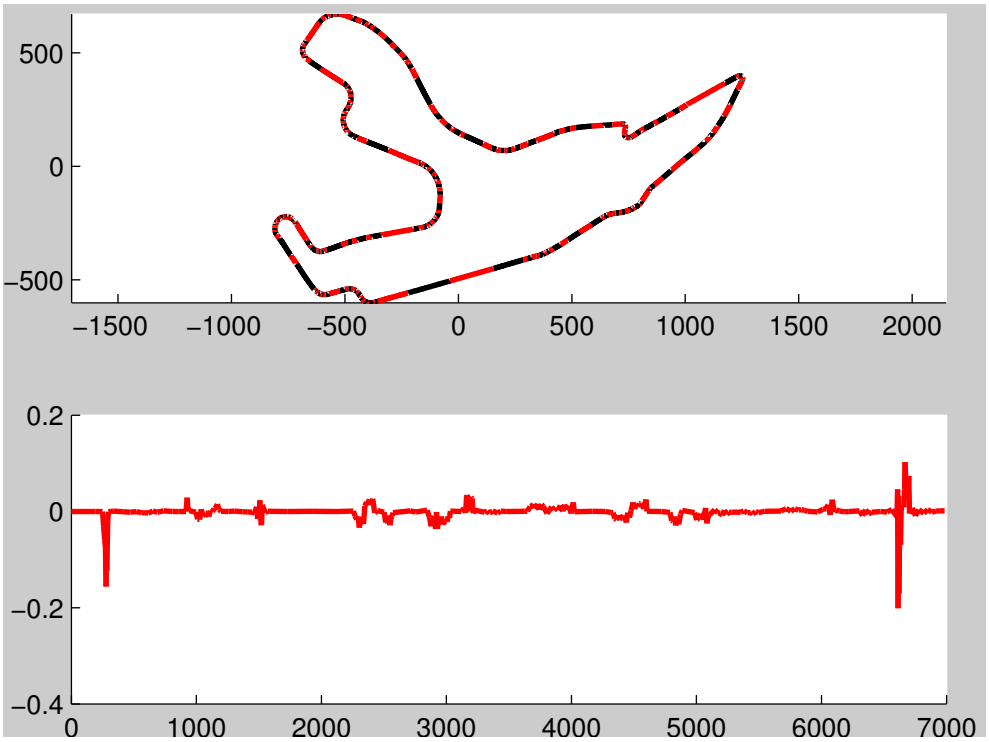


Figure 6.10: Quasi G^2 interpolation and the relative curvature. Units are meters and for the curvature 1/meters.

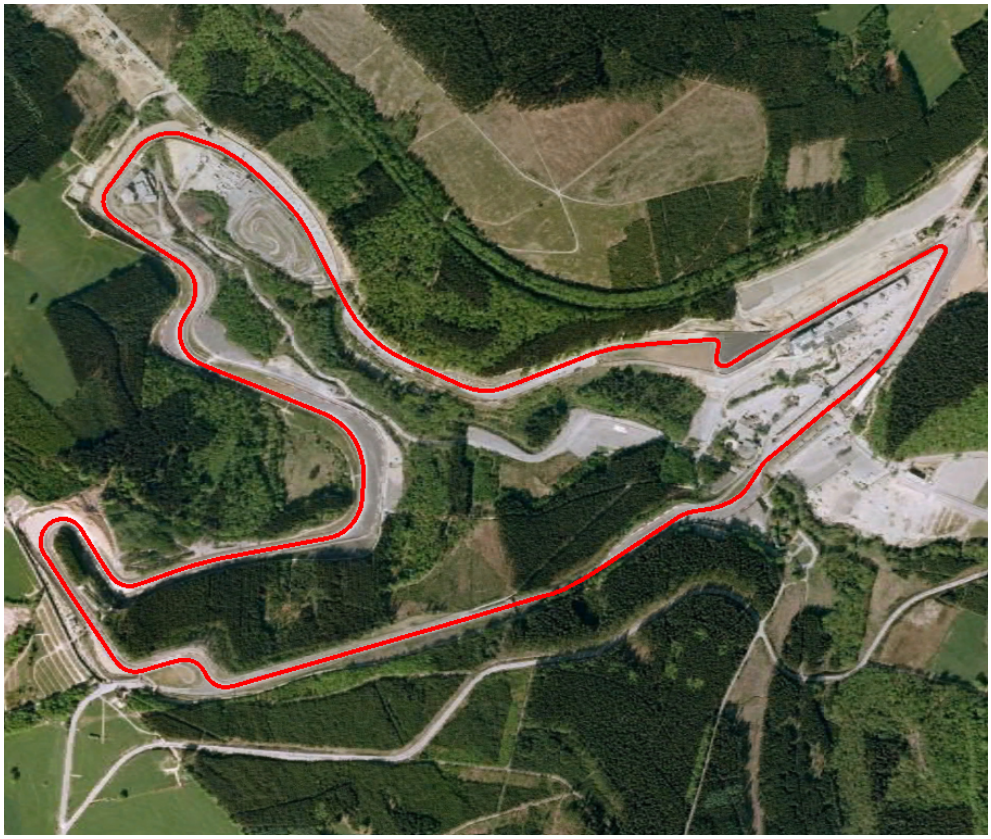


Figure 6.11: Projection of the fitting on the surface.

7

CONCLUSIONS

We have presented a benchmark suite of optimal control problems to validate the solver XOptima proposed by our research group. The suite is supplied with the analytic solution in order to permit a reliable comparison between the numerical and the exact solution. The comparison is enlarged to other three open source software for OCP, Acado, Gpops and Iclocs. Acado uses the multiple shooting algorithm with an SQP, while Gpops uses the pseudo spectral method and Iclocs uses direct collocation. Moreover we collected the numerical results from published articles and books. In particular we choose different approaches to check the difference of the results, thus we have results for each of the direct and indirect methods and for the DPP. We do not compare the performance of the solvers in terms of speed, because it does not make sense to compare different implementations: Acado and XOptima are written in C++ and hence are in general much faster than the Matlab interface for the NLP solver Ipopt used by Gpops and Iclocs. As an *en passant* comment, we just say that the execution time varies from below the second up to few seconds for XOptima, a bunch of seconds up to a minute for Acado, half a minute up to one or two minutes for Gpops and Iclocs. We still remark that those times are not representative because of the different implementations, they are given to just give an idea. We do not compare either the number of iterations employed to converge to the solution, because the various methods are structurally different, e.g. Gpops and Iclocs use subiterations and refinements of the mesh while Acado and XOptima do not have subiterations. Therefore, the only performance criterion used was the precision of the solution in terms of the ratio $(N - E)/E$ where N is the numerical value of the target functional to be minimized and E is the exact value of the target coming from the analytic solution. Here another comment is mandatory: the methods tested have different characteristics, so it is not completely representative to look only at the target value. In facts, for example, Acado uses piecewise constant controls, that are well suited for bang bang problems, but give some inaccuracies when the control is for example a line or a parabola. This is clear looking at Figure 5.4: it is clear that Acado converged (good) to the correct solution, but it can not give a precise solution because of the shape of the control, even using a fine mesh. The opposite occurs for Gpops, that fits the control very precisely and thus gives a very good result (Table 5.1). So it is not enough to consider only the results quantitatively, but we have to check also the quality of the solution in terms of oscillations of the numerical values, e.g. Figure 5.21 shows the ringing of Gpops. The method employed by XOptima makes broad use of penalty functions, yielding a continuous smooth control even in the case of bang bang solutions, to obtain a sharp plot in the points of discontinuity, it is necessary to put severe penalization on the weight of the regularized functions. The best way to obtain a sharp control in those cases and a very quick convergence, is to use homotopy (or continuation). This tool turns out to be fundamental, because it allows to start the numeric solution with very mild penalization of the control obtaining a quick convergence. This first approximate solution is then used as a guess for the states and the control for a new instance of the solver with a more strict requirement on the control or on other states or variables. Practise shows that even if we could solve the problem without applying continuation, the convergence time is dramatically higher than with the homotopy activated. But apart from the speed up of the process, continuation turned out to be the key feature to obtain convergence of XOptima on the hard problems, while the

other solvers were not successful. A limit encountered in the numerical solution of the underwater vehicle, was that Acado, Gpops and Iclocs did not converge, but were also unable to practically handle the required (see [CSMV04]) mesh of 10000 points. With XOptima we could solve the problem up to 20000 points (in a reasonable time of around two minutes), but there was not a significant improvements of the solution, and we decided to report the value of the coarsest mesh yielding a valid solution, that is 2000 points.
We were successful in solving with XOptima the problems it was born for, that is the optimization of the minimum lap time of a high performance vehicle on a race circuit track. The problem is very challenging because of many types of constraints and results in almost 100000 equations on a mesh of 2800 nodes. In this problem it was employed the description of the road obtained with the algorithm presented in chapter 6 and published in [BF13, BF14]. The novelty of the formulation proposed is the proof of existence and uniqueness of the solution, the bound of the number of iterations of the algorithm to produce a satisfactory solution and the analysis of the motivation of the failures of the other state of art algorithms. An important application of the algorithm is the generation of quasi G^2 trajectories, were the jumps of the curvature are in practice negligible from an applicative point of view.

8

BIBLIOGRAPHY

REFERENCES FROM BOOKS

[AF66] Michael Athans and Peter L. Falb. *Optimal control : an introduction to the theory and its applications*. Lincoln Laboratory publications. McGraw-Hill, New York, Saint Louis, San Francisco, 1966.

[AS64] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Number 55 in National Bureau of Standards Applied Mathematics Series. U.S. Government Printing Office, Washington, D.C., 1964.

[Bet01] J.T. Betts. *Practical Methods for Optimal Control Using Nonlinear Programming*. Advances in design and control. Society for Industrial and Applied Mathematics, 2001.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[HNW93] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Solving Ordinary Differential Equations. Springer, 1993.

[HNW96] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Lecture Notes in Economic and Mathematical Systems. Springer, 1996.

[Hul03] D.G. Hull. *Optimal Control Theory for Applications*. Mechanical Engineering Series. Springer, 2003.

[Lib03] D. Liberzon. *Switching in Systems and Control*. Systems & control: foundations & applications. Birkhauser Boston, 2003.

[Luu00] Rein Luus. *Iterative Dynamic Programming*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2000.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[OLBC10] F.W. J. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark, editors. *NIST handbook of mathematical functions*. U.S. Department of Commerce National Institute of Standards and Technology, Washington, DC, 2010. With CD-ROM.

[PVTF02] W.H. Press, W.T. Vetterling, S.A. Teukolsky, and B.P. Flannery. *Numerical Recipes in C++: the art of scientific computing*. Cambridge University Press, New York, NY, USA, 2nd edition, 2002.

[TCS14] Mara Tanelli, Matteo Corno, and Sergio. Savaresi. *Modelling, Simulation and Control of Two-Wheeled Vehicles*. Wiley, Boston, 2014.

[Tho97] W.J. Thompson. *Atlas for Computing Mathematical Functions: An Illustrated Guide for Practitioners with Programs in C and Mathematica with Cdrom*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1997.

[Tro96] J.L. Troutman. *Variational Calculus and Optimal Control With Elementary Convexity*. Undergraduate Texts in Mathematics. Springer Verlag, 1996.

[Wat44] G.N. Watson. *A Treatise on the Theory of Bessel Functions*. Cambridge University Press, Cambridge, England, 1944.

[ZB94] M.I. Zelikin and V.F. Borisov. *Theory of chattering control with applications to astronautics, robotics, economics and engineering*. Birkhauser, Boston, 1994.

[BH69] A. E. Bryson and Y. C. Ho. *Applied Optimal Control*. Blaisdell, 1969.

[Bon03] J.F. Bonnans. *Numerical optimization: theoretical and practical aspects : with 26 figures*. Universitext (1979). Springer, 2003.

[Cha07] Benoit Chachuat. *Nonlinear and Dynamic Optimization: From Theory to Practice*. EPFL, 2007.

[DB78] C. De Boor. *A Practical Guide to Splines*. Number v. 27 in Applied Mathematical Sciences. Springer-Verlag, 1978.

[ET99] Ivar Ekeland and Roger Téman. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[Far02] Gerald Farin. *Curves and surfaces for CAGD: a practical guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 5th edition, 2002.

[FP09] Christodoulos A. Floudas and Panos M. Pardalos, editors. *Encyclopedia of Optimization, Second Edition*. Springer, 2009.

[Gee07] H.P. Geering. *Optimal Control with Engineering Applications*. Springer, 2007.

[Ger12] M. Gerdts. *Optimal Control of ODEs and DAEs*. De Gruyter Textbook. De Gruyter, 2012.

[Hes66] M.R. Hestenes. *Calculus of variations and optimal control theory*. Wiley, New York, 1966.

[Jur96] Velimir. Jurdjevic. Cambridge University Press, 1996.

[Kir70] D.E. Kirk. *Optimal control theory: an introduction*. Prentice-Hall networks series. Prentice-Hall, 1970.

[KS91] Morton I. Kamien and Nancy L. Schwartz. *Dynamic optimization. The calculus of variations and optimal control in economics and management*. Advanced textbooks in economics. North-Holland, 2. ed edition, 1991.

[LC03] L.P. Lebedev and M.J. Cloud. *The Calculus of Variations and Functional Analysis: With Optimal Control and Applications in Mechanics*. Series on stability, vibration and control of systems: Series A. World Scientific, 2003.

[Lib12] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2012.

[Loc01] A. Locatelli. *Optimal Control: An Introduction*. Birkhäuser Basel, 2001.

[MO98] A.A. Milyutin and N.P. Osmolovskii. Calculus of variations and optimal control. *American Mathematical Society*, 1998.

[Nai02] D.S. Naidu. *Optimal Control Systems*. Electrical Engineering Series. Taylor & Francis, 2002.

[Pet68] I.P. Petrov. *Variational methods in optimum control theory*. Mathematics in Science and Engineering. Elsevier Science, 1968.

[Pin93] E.R. Pinch. *Optimal Control and the Calculus of Variations*. Oxford University Press, UK, 1993.

[Pyt99] R. Pytlak. *Numerical Methods for Optimal Control Problems with State Constraints*. Number No. 1707 in Lecture Notes in Mathematics. Springer, 1999.

[Rao96] Singiresu S. Rao. *Engineering Optimization: Theory and Practice*. Wiley-Interscience, 1996.

[SL12] H. Schättler and U. Ledzewicz. *Geometric Optimal Control: Theory, Methods and Examples*. Interdisciplinary applied mathematics. Springer, 2012.

[Tak85] A. Takayama. *Mathematical Economics*. Cambridge University Press, 1985.

[Vin10] R. Vinter. *Optimal Control*. Modern Birkhäuser Classics. Springer, 2010.

[You80] L.C. Young. *Lecture on the Calculus of Variations and Optimal Control Theory*. AMS Chelsea Publishing Company Series. AMS Chelsea Publishing, 1980.

[Zel00] I. Zelikin. *Control Theory and Optimization I: Homogeneous Spaces and the Riccati Equation in the Calculus of Variations*. Control theory and optimization. Springer, 2000.

REFERENCES FROM ARTICLES

[ALHB08] G. Arechavaleta, J-P Laumond, H. Hicheur, and A. Berthoz. An optimality principle governing human walking. *Robotics, IEEE Transactions on*, 24(1):5–14, 2008.

[BBDL03] E. Bertolazzi, F. Biral, and M. Da Lio. Symbolic–numeric efficient solution of optimal control problems for multibody systems. *Journal of Computational Methods in Science and Engineering*, 2(3), 2003.

[BBDL05] E. Bertolazzi, F. Biral, and M. Da Lio. Symbolic–numeric indirect method for solving optimal control problems for large multibody systems. *Multibody System Dynamics*, 13(2):233–252, 2005.

[BBDL06] E. Bertolazzi, F. Biral, and M. Da Lio. Symbolic-numeric efficient solution of optimal control problems for multibody systems. *Journal of Computational and Applied Mathematics*, 185(2):404–421, 2006.

[BBDL07] Enrico Bertolazzi, Francesco Biral, and Mauro Da Lio. real-time motion planning for multibody systems. *Multibody System Dynamics*, 17(2-3):119–139, 2007.

[BBDL+14] Enrico Bertolazzi, Francesco Biral, Mauro Da Lio, Marco Galvani, Paolo Bosetti, Andrea Saroldi, and Fabio Tango. The driver continuous support function in the fp7 interactive project: an implementation based on the co-driver metaphor. 2014.

[BD12] F. Bertails-Descoubes. Super-clothoids. *Computer Graphics Forum*, 31(2pt2):509–518, 2012.

[BF13] E. Bertolazzi and M. Frego. G^1 fitting with clothoids. http://www.mathworks.com/matlabcentral/fileexchange/42113-g1-fitting-with-clothoids, 2013.

[BF14] Enrico Bertolazzi and Marco Frego. G1 fitting with clothoids. *Mathematical Methods in the Applied Sciences*, X(X):18, 2014.

[BL13] Francesco Biral and Roberto Lot. A curvilinear abscissa approach for the lap time optimization of racing vehicles. 2013.

[BLP10] I. Baran, J. Lehtinen, and J. Popović. Sketching clothoid splines using shortest paths. *Computer Graphics Forum*, 29(2):655–664, May 2010.

[BNPS91] R. Bulirsch, E. Nerz, H. J. Pesch, and O. Von Stryk. Combining direct and indirect methods in optimal control: Range maximization of a hang glider. In *Optimal Control, volume 111 of International Series of Numerical Mathematics. Birkhuser*, pages 273–288. Birkhauser Verlag, 1991.

[Bor00] V.F. Borisov. Fuller's phenomenon: Review. *Journal of Mathematical Sciences*, 100(4):2311–2354, 2000.

[Bul67] Roland Bulirsch. Numerical calculation of the sine, cosine and fresnel integrals. *Numerische Mathematik*, 9(5):380–385, 1967.

[Cala] Andrea Calogero. Appunti di calcolo delle variazioni e controllo ottimo. teoria, modelli economici,esercizi e cenni di controllo ottimo stocastico. Technical report, UniversitÃ Milano-Bicocca, Dipartimento di Matematica e Applicazioni.

[Calb] Andrea Calogero. Notes on optimal control theory with economic models and exercises. Technical report, UniversitÃ Milano-Bicocca, Dipartimento di Matematica e Applicazioni.

[CH93] Y. Chen and J. Huang. A new computational approach to solving a class of optimal control problems. *Int. J. Control*, 58:1361–1383, 1993.

[Chy03] M. Chyba. Underwater vehicles: a surprising non time-optimal path. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 3, pages 2750–2755 Vol.3, Dec 2003.

[CSMV04] M. Chyba, H. Sussmann, H. Maurer, and G. Vossen. Underwater vehicles: the minimum time problem. In *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, volume 2, pages 1370–1375 Vol.2, Dec 2004.

[Dai12] J. Daily. Analysis of critical speed yaw scuffs using spiral curves. *SAE Technical Paper*, 2012-01-0606, 2012.

[DC09] R. Dai and J.E. Cochran. Path planning for multiple unmanned aerial vehicles by parameterized cornu-spirals. In *American Control Conference, 2009. ACC '09.*, pages 2391–2396, 2009.

[DCBM+07] M. De Cecco, E. Bertolazzi, G. Miori, R. Oboe, and L. Baglivo. PC-sliding for vehicles path planning and control - design and evaluation of robustness to parameters change and measurement uncertainty. In *ICINCO-RA (2)'2007*, pages 11–18, 2007.

[DM95a] S. A. Dadebo and K.B. Mcauley. On the computation of optimal singular controls. In *Control Applications, 1995., Proceedings of the 4th IEEE Conference on*, pages 150–155, 1995.

[DM95b] S.A. Dadebo and K.B. Mcauley. Dynamic optimization of constrained chemical engineering problems using dynamic programming. *Computers and Chemical Engineering*, 19(5):513–525, 1995.

[FKvW10] P. Falugi, E. Kerrigan, and E. van Wyk. Imperial College London Optimal Control Software user guide - ICLOCS, 2010.

[FO77] J.E. Flaherty and Jr. O'Malley, R. On the computation of singular controls. *Automatic Control, IEEE Transactions on*, 22(4):640–648, 1977.

[GT88] C.J. Goh and L.K. Teo. Control parameterization: a unified approach to optimal control problems with general constraints. *Automatica*, 24:3–18, 1988.

[HFD11] B. Houska, H.J. Ferreau, and M. Diehl. ACADO Toolkit – An Open Source Framework for Automatic Control and Dynamic Optimization. *Optimal Control Applications and Methods*, 32(3):298–312, 2011.

[JGL70] D. Jacobson, Stanley B. Gershwin, and M. Lele. Computation of optimal singular controls. *Automatic Control, IEEE Transactions on*, 15(1):67–73, 1970.

[Joh48] Fritz John. Extremum problems with inequalities as side constraints. *Studies and Essays,*, Courant Anniversary Volume:187–204, 1948.

[KDK95] V.P. Kostov and E.V. Degtiariova-Kostova. Some Properties of Clothoids. Technical Report RR-2752, INRIA, December 1995.

[KFP03] B.B. Kimia, I. Frankel, and A-M. Popescu. Euler spiral for shape completion. *Int. J. Comput. Vision*, 54(1-3):157–180, August 2003.

[KT51] H.W. Kuhn and A.W. Tucker. Nonlinear programming. pages 481–492. Univ. of California Press, Berkeley, 1951.

[LNRL08] L. Labakhua, U. Nunes, R. Rodrigues, and F.S. Leite. Smooth trajectory planning for fully automated passengers vehicles: Spline and clothoid based methods and its simulation. In *Informatics in Control Automation and Robotics*, volume 15 of *Lecture Notes Electrical Engineering*, pages 169–182. Springer Berlin Heidelberg, 2008.

[Luu91] R. Luus. Application of iterative dynamic programming to state constrained optimal control problems. *Hung. J. Ind. Chem.*, 19:245–254, 1991.

[Mar73] C. Marchal. Chattering arcs and chattering controls. *Journal of Optimization Theory and Applications*, 11(5):441–468, 1973.

[Mar75] C. Marchal. Second-order tests in optimization theories. *Journal of Optimization Theory and Applications*, 15(6):633–666, 1975.

[MS09] J. McCrae and K. Singh. Sketching piecewise clothoid curves. *Computers & Graphics*, 33(4):452–461, June 2009.

[MVHW10] M. Manz, F. Von Hundelshausen, and H. J Wuensche. A hybrid estimation approach for autonomous dirt road following using multiple clothoid segments. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 2410–2415, 2010.

[Por07] Frank Porter. Calculus of variations. Technical report, Caltech, 2007. Physics 129a.

[RBD+10] Anil V. Rao, David A. Benson, Christopher Darby, Michael A. Patterson, Camila Francolin, Ilyssa Sanders, and Geoffrey T. Huntington. Algorithm 902: Gpops, a matlab software for solving multiple-phase optimal control problems using the gauss pseudospectral method. *ACM Trans. Math. Softw.*, 37(2):22:1–22:39, April 2010.

[SC03] E.L. Shirley and E.K. Chang. Accurate efficient evaluation of lommel functions for arbitrarily large arguments. *Metrologia*, 40(1):S5, 2003.

[Sto82] J. Stoer. Curve fitting with clothoidal splines. *J. Res. Nat. Bur. Standards*, 87(4):317–346, 1982.

[SW01] Hector J. Sussmann and Jan C. Willems. The brachystrochrone problem and modern control theory. In *Contemporary Trends in Nonlinear Geometric Control Theory and Its, Monroy-Perez (Eds); World Scientific Publishers*. Publishers, 2001.

[Tel05] V. Telasula. Fresnel cosine and sine integral function. `http://www.mathworks.it/matlabcentral`, 2005.

[Wil09] D.K. Wilde. Computing clothoid segments for trajectory generation. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 2440–2445, 2009.

[WM08] D.J. Walton and D.S. Meek. An improved euler spiral algorithm for shape completion. In *Computer and Robot Vision, 2008. CRV '08. Canadian Conference on*, pages 237–244. IEEE, may 2008.

[WM09] D.J. Walton and D.S. Meek. Interpolation with a single cornu spiral segment. *Journal of Computational and Applied Mathematics*, 223(1):86–96, 2009.

[WMN+01] L.Z Wang, K.T Miura, E Nakamae, T Yamamoto, and T.J Wang. An approximation approach of the clothoid curve defined in the interval $[0, \pi/2]$ and its offset by free-form curves. *Computer-Aided Design*, 33(14):1049 – 1058, 2001.

[BC09] Franky Backeljauw and Annie Cuyt. Algorithm 895: A continued fractions package for special functions. *ACM Trans. Math. Softw.*, 36(3):15:1–15:20, July 2009.

[BK97] B. Bonnard and I. Kupka. Generic properties of singular trajectories. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 14(2):167 – 186, 1997.

[CLRS01] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, Cambridge, MA, USA, 2nd edition, 2001.

[Dav99] TG Davis. Total least-squares spiral curve fitting. *Journal of Surveying Engineering-Asce*, 125(4):159–176, NOV 1999.

[FN95] R. T. Farouki and C. A. Neff. Hermite interpolation by pythagorean hodograph quintics. *Math. Comput.*, 64(212):1589–1609, October 1995.

[HS09] Martin Held and Christian Spielberger. A smooth spiral tool path for high speed machining of 2d pockets. *Computer-Aided Design*, 41(7):539 – 550, 2009.

[KH89] Y. Kanayama and B.I. Hartman. Smooth local path planning for autonomous vehicles. In *Robotics and Automation, 1989. Proceedings., 1989 IEEE International Conference on*, pages 1265 –1270 vol.3, may 1989.

[lom] Nist digital library of mathematical functions.

[Mic96] Volker Michel. Singular optimal control - the state of the art. (169), 1996.

[MS78] D.J. Mellefont and R.W.H. Sargent. Calculation of optimal controls of specified accuracy. *Journal of Optimization Theory and Applications*, 25(3):407–414, 1978.

[MW92] D. S. Meek and D. J. Walton. Clothoid spline transition spirals. *Mathematics of Computation*, 59:117–133, July 1992.

[MW04] D. S. Meek and D. J. Walton. A note on finding clothoids. *J. Comput. Appl. Math.*, 170(2):433–453, September 2004.

[MW09] D. S. Meek and D. J. Walton. A two-point $g1$ hermite interpolating family of spirals. *J. Comput. Appl. Math.*, 223(1):97–113, January 2009.

[MZ90] V.F. Borisov M.I. Zelikin. Synthesis in problems of optimal control containing trajectories with participating switchings and singular trajectories of the second order. *jour Mat. Zametki*, 47(1):41–49, 1990.

[MZ93] V.F. Borisov M.I. Zelikin. Regimes with frequented switches in the problems of optimal control. *Selected topics in the theory of oscillations and optimal control theory Trudy Mat. Inst. Steklov.*, 197:95–186, 1993.

[Pav83] Theodosios Pavlidis. Curve fitting with conic splines. *ACM Trans. Graph.*, 2(1):1–31, January 1983.

[Rya84] E.P. Ryan. Optimal feedback control of bilinear systems. *Journal of Optimization Theory and Applications*, 44(2):333–362, 1984.

[SF97] A. Scheuer and Th. Fraichard. Continuous-curvature path planning for car-like vehicles. In *Intelligent Robots and Systems, 1997. IROS '97., Proceedings of the 1997 IEEE/RSJ International Conference on*, volume 2, pages 997–1003. Intelligent Robots and Systems. IROS '97.,, sep 1997.

[Smi11] David M. Smith. Algorithm 911: Multiple-precision exponential integral and related functions. *ACM Trans. Math. Softw.*, 37(4):46:1–46:16, February 2011.

[Sny93] W. Van Snyder. Algorithm 723: Fresnel integrals. *ACM Transactions on Mathematical Software*, 19(4):452–456, December 1993.

[SS90] Dong Hun Shin and Sanjiv Singh. Path generation for robot vehicles using composite clothoid segments. Technical Report CMU-RI-TR-90-31, Robotics Institute, Pittsburgh, PA, December 1990.

[WM96] D.J. Walton and D.S. Meek. A planar cubic bezier spiral. *Journal of Computational and Applied Mathematics*, 72(1):85–100, 1996.

[WM07] D. J. Walton and D. S. Meek. G^2 curve design with a pair of pythagorean hodograph quintic spiral segments. *Comput. Aided Geom. Des.*, 24(5):267–285, July 2007.