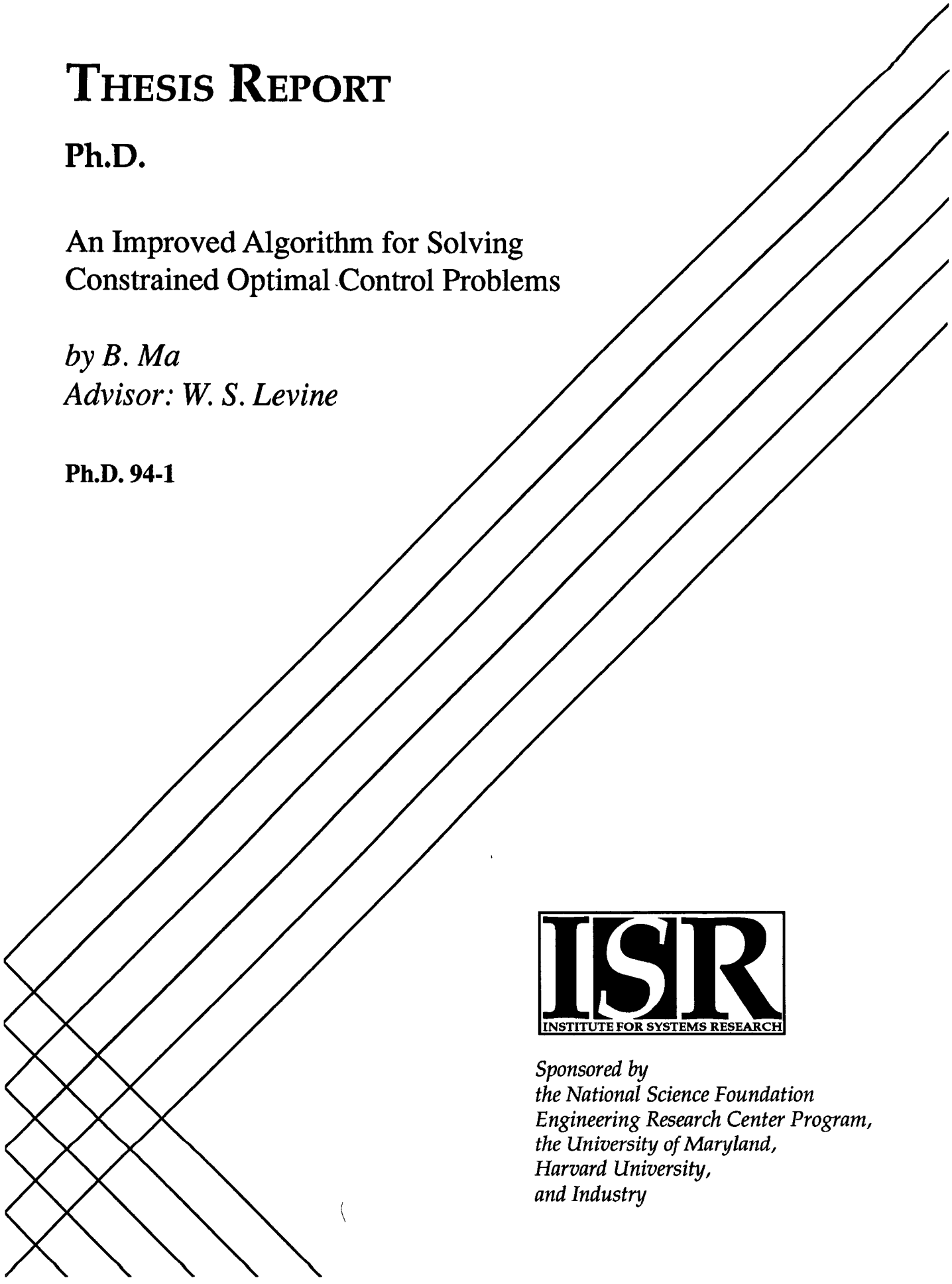
THESIS REPORT

Ph.D.

An Improved Algorithm for Solving
Constrained Optimal Control Problems

by B. Ma
Advisor: W. S. Levine

Ph.D. 94-1



Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Abstract

Title of Dissertation: AN IMPROVED ALGORITHM FOR SOLVING CONSTRAINED OPTIMAL CONTROL PROBLEMS

Baoming Ma, Doctor of Philosophy, 1994

Dissertation directed by: Professor William S. Levine
Department of Electrical Engineering

Motivated by the need to have an algorithm which (1) can solve generally constrained optimal control problems, (2) is globally convergent, (3) has a fast local convergence rate, a new algorithm, which solves fixed end-time optimal control problems with hard control constraints, end-point inequality constraints, and a variable initial state, is developed. This algorithm is based on a second-order approximation to the change of the cost functional due to a change in the control and a change in the initial state. Further approximation produces a simple convex functional. An exact penalty function is employed to penalize any violated end-point inequality constraints. We then show that the solution of the minimization of the convex functional, subject to linearized system dynamics, the original hard control constraints, the original hard initial state constraints, and linearized end-point constraints, generates a descent direction for that exact penalty function.

We then show that the algorithm developed in this dissertation can also solve the following types of optimal control poblems: (1) problems with a free end-time; (2) problems with path constraints; (3) problems with some design parameters that are also to be optimized.

Global convergence properties of a version of the algorithm are analyzed. In particular, it is shown that the algorithm is globally convergent under some conditions.

The local convergence rate of the algorithm can be better than that of the first-order algorithms when some matrices are properly updated.

A version of the algorithm is implemented in a package which is easy to use. A variety of benchmark problems are solved. Finally, the algorithm is employed in solving two challenging biomechanics problems: (1) a human moving his arm from an initial resting position so as to touch and stop at a target with the tip of the index finger while the muscular stress, the joint constraint forces, and the neural excitations are minimized; (2) a human pedaling a stationary bicycle as fast as possible from rest. Those results demonstrate that the algorithm developed in this dissertation is effective in dealing with generally constrained optimal control problems.

AN IMPROVED ALGORITHM FOR SOLVING CONSTRAINED OPTIMAL CONTROL PROBLEMS

by

Baoming Ma

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1994

Advisory Committee:

Professor William S. Levine, Chairman/Advisor
Professor André L. Tits
Professor Robert Newcomb
Professor Wijesuriya P. Dayawansa
Professor R. Bruce Kellogg

Dedication

This dissertation is dedicated to my beloved mother who, sadly, could not share with me this happy moment which she had been awaiting for so long.

Acknowledgements

First, I wish to express my deep appreciation to my advisor Dr. William S. Levine. His constant support and encouragement helped me to overcome many difficult times, both in my work and in my life. His guidance, his deep insight, and unique intuition are the "optimal" driving forces of this dissertation. It has truly been a privilege for me to be his student.

I am grateful to Dr. André L. Tits for his kindness and willingness to help over the years, which always made me feel at home; for his rigor in teaching and research which helped me build up a sound optimization background.

I would like to thank Dr. Wijesuriya P. Dayawansa for being helpful and supportive over the years. I would also like to thank Professor Robert Newcomb and Professor R.B. Kellogg for their generous effort and time spent in reading the manuscript, and for their valuable advice and suggestions concerning my work.

A special thanks goes to Dr. Felix Zajac for his support, patience and understanding, which were the keys to making our collaboration a smooth, productive, and successful one.

To all my dear friends at the CACSE Lab at Maryland: Lei Zhang, Chujen Lin, Craig Lawrence, Gil Yudilevitch, Benjamin Bachrach, John Reilly, Bruce Douglas, Qifeng Wei, Jian Zhou, Peter Yan, Jiqin Pan, Xiaoguang Chen, Jipin He, and Eunsup Sim, I want to thank you all for your caring, help, and the fun we had over the years. You have made my experience here truly enjoyable and memorable.

To all my dear friends at Stanford: Christine Raasch, B.J. Fregly, Lisa Schutte, Pete Loan, Art Kuo, I enjoyed our work together and the fun times. Your vigor and intelligence deeply impressed me. I am looking forward to collaborating with you again in the future.

I am deeply thankful to my parents, who brought me up with great endurance of hardships, with true love and perfect care, and with great expectations. I would also like to thank my sister for her constant support, love and many sacrifices. It is the

deepest sorrow of my life that my beloved mother suddenly passed away six months ago before seeing me finish this dissertation, a happy moment she had been awaiting for so long.

I would like especially to thank my lovely Chong Fu for our wonderful love during the last four years, for her generous support, her mature understanding and her amazing patience.

Thanks also go to the Institute for Systems Research at the University of Maryland for providing an excellent research environment and high quality facilities. This research was financially supported by NIH Grant #NS17662 and NSF Grant CDR-88-03012.

Table of Contents

List of Figures

Part I

Preliminaries

Chapter 1

Introduction

1.1 Role of Optimal Control

Since its birth from a classical subject, the calculus of variations, in the late 1950's, modern optimal control theory has been rapidly developed. In addition to its successful applications in areas which range from designing spacecraft guidance and control systems to applied economics, chemical engineering, nuclear engineering, etc, optimal control theory has also found a role in analyzing animal and human locomotion.

1.2 Locomotion Research That Motivated This Study on Computing Optimal Control

There are two major reasons that could justify the use of optimal control theory in the study of locomotion. The first reason is a logical one. Locomotion is believed to be goal-oriented. Optimal control theory, a study of the control strategy which maximizes or minimizes a certain goal function subject to the constraints imposed by, among others, the dynamics of a system, therefore provides a convenient and natural tool. The second reason is a practical one. In order to understand human or animal neuromotor control strategies, a reasonable and rigorous approach is based on the use of a dynamical model of the musculoskeletal system to predict the muscle excitation signals that produce the movement. Because the number of muscles spanning

each joint usually exceeds the number of degrees of freedom defining joint motion, the human and animal musculoskeletal system is mechanically highly redundant. In addition, many muscles can affect more than one joint, which causes complex dynamical interactions. Therefore, finding the muscle excitation patterns which provide the desired movement is difficult by trial and error for even the simplest case [84,170]. On the other hand, optimal control theory sees no difference between dynamics of muscle activation, dynamics of musculotendon and dynamics of the skeleton. They are simply parts of the system dynamics. It is a tool both unified and systematic.

Human locomotion has always been a subject attracting wide attention, because of its orderly organization and complex coordination. However, studies had been mostly experimental in nature until the late 1960's. A study by Chow and Jacobson [24] was described as "probably the most comprehensive contribution" in applying the optimal control theory in the field of locomotion [47].

Motivated by the need to better understand how the central nervous system coordinates limb movement, Levine, Zajac, and their students have done a significant amount of work in using optimal control theory to study intermuscular control of multijoint movement. Through a variety of increasingly complex models, they have gained insight into the theoretical and computational aspects of optimal control problems involving human and animal musculoskeletal systems.

They began by studying maximum-height jumping by cats and humans. In the case of a simple one-segment, planar baton, driven by an ideal torque generator, a complete analytical solution was derived [78], where the feedback optimal control was expressed in terms of specific controls in different regions of state space. Then, the optimal control problems were solved numerically for a four-segment inverted pendulum with muscles described by models, which are of Hill's type [51,148,167], incorporated in the joint torque generators [20,21,110,148,149]. Then they progressed to the study of pedaling a stationary bicycle as fast as possible, with the belief that it is intermediate in complexity between maximal height jumping and normal locomotion [110,124, 125,126,127,128,148,149].

Qualitative comparisons between the predictions of the model and previously reported experimental findings indicated that the model reproduces major features of both a maximum-height squat jump [110,148] and minimum-time pedaling [124,125, 126,127,128,148,149].

A similar approach was adapted in Giat's work [39], in which the study was to find control patterns which could lift the human arm to hit a target in a plane.

Each of the above problems has a relatively unambiguous performance criterion, so it fits well into the framework of optimal control theory.

1.3 Computational Challenges

The dynamics models of complex systems, especially of those biomechanics systems, always have the following characteristics: high dimension, severe nonlinearity, complex coupling, various constraints and occasionally time-variation. As for the dimension, for example, Sim's jumping model has 24 states and 8 controls; his pedaling model has 48 states, although some of those are dependent, and 16 controls. Both models used a one-state muscle model [148]. Raasch's current pedaling model, in which a new two-state muscle model is used, has 100 states, some of which are dependent, and 18 controls [124,125,126,127,128]. Using the same new muscle model, Giat's arm model has 40 states and 12 controls [39] The nonlinearities are introduced by the generalized gravitational force terms, generalized inertial terms, and by the nonlinear behavior of muscles. The coupling becomes more evident when the mechanical system is closed-loop, and when some muscles can affect more than one joint.

Clearly, except for some special cases [79,80,88], finding closed-form optimal control solutions is almost impossible. A numerical approach then has to be adopted in most cases.

Expectations of a good numerical optimal control algorithm have always been that it should first be "robust", which means (a) it is convergent, (b) it converges to at least a local minimum, (c) it will not crash under any circumstance, and second that it

should be "efficient", which means that it has a fast (local) convergence rate, and third that it should be "versatile", which means that it can handle general optimal control problems: general setting of dynamics, cost function, terminal constraints and path constraints. In reality, those expectations (or standards) are too high to be realized by a single algorithm. There are only a few conceptually implementable optimal control algorithms available which can meet part of the above expectations. Among them, even fewer have been well implemented.

1.4 Objective and Contributions

The goal of this dissertation is to develop a computational algorithm to meet the above challenges. The contributions of this dissertation can be summarized as follows:

- a new algorithm, with an approach similar to the Han-Powell method in finite-dimensional optimization, is devised:
  - it is globally convergent under some conditions;
  - its local convergence rate can be better than that of the first-order algorithms when some matrices are properly updated;
  - it solves optimal control problems with hard control constraints, end-point constraints, a variable initial state, and a variable parameter vector;
  - it can also solve, approximately, path constraints;
- a version of the algorithm is implemented in a package which is easy to use;
- a variety of benchmark problems are solved well and fast by the package;
- the package is also employed in solving two challenging biomechanics problems:
  - a human moving his arm from an initial resting position so as to touch and stop at a target with the tip of the index finger while the muscular stress, the joint constraint forces, and the neural excitations are minimized;
  - a human pedaling a stationary bicycle as fast as possible, from rest.

1.5 Organization

This dissertation is organized into three parts. The first part is the preliminaries, which includes this introductory chapter and Chapter 2 where a comprehensive survey on the techniques and algorithms for solving optimal control problems are given.

In the second part, new algorithms are developed to solve generally constrained optimal control problems.

In Chapter 3, with an approach similar to the Han-Powell method in finite-dimensional optimization, a new algorithm is devised to solve continuous-time optimal control problems where the control variables and the terminal states are constrained. It is first noticed that the summation of the first and second variations is a second-order approximation to the change of the cost functional due to a change in the control. Further approximation produces a simple convex functional. Consequently, solving the original complicated problem can be replaced by solving iteratively a much simplier "direction-finding" subproblem and a line search along the "direction" found. We then show that the solution of the minimization of the convex functional subject to a linearized system dynamics, linearized terminal inequality constraints, and the original control constraint, generates a descent direction of an exact penalty functional. A global convergence analysis is then given.

In Chapter 4, the ideas behind the algorithm developed in Chapter 3 can be further extended to a much more general optimal control problem which has not only hard control constraints and terminal-state inequality constraints, but also a variable initial state vector, some components of which are allowed to vary within a constraint box, while the remaining components are fixed. It will be shown later in the chapter that this problem can include optimal control problems in the most general setting, namely, problems which are subjected to control constraints, path constraints, end-point constraints, a variable initial state, and a variable vector of design parameters, within a fixed/free end-time interval.

A common feature of the algorithms developed in both Chapter 3 and Chap-

ter 4 is that it is required to solve a generally constrained linear quadratic regulator problem (LQR) at each iteration. On the other hand, the constrained LQR problem is important in its own right. The goal of Chapter 5 is to study the following constrained LQR problem: minimize a convex functional, subject to a linear dynamical system, hard control constraints, a constraint box for some initial state variables, and linear end-point constraints. Two special properties which are related to the problem are presented: (1) the existence of an optimal control solution; and (2) the uniqueness of the optimal control solution. In addition, some computational techniques are investigated. In Chapter 6, a variety of benchmark problems are solved by the algorithm.

The third part of this dissertation applies the new algorithm developed in Chapter 3 to two biomechanics problems.

In Chapter 7, the skeletal and muscular dynamics of the human upper extremity is studied by using optimal control theory. The algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed to compute the activity which occurs in each muscle of the upper extremity when the goal is to move the arm from an initial resting position so as to touch and stop at a target with the tip of the index finger while the muscular stress, the joint constraint forces, and the neural excitations are minimized. The results obtained from the simulation describe all the major dynamic events that take place in the upper extremity when the movement is attempted.

In Chapter 8, the skeletal and muscular dynamics of the human lower extremity is studied by using optimal control theory. The algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed to compute the activity which occurs in each muscle of the lower extremity when the goal is to pedal a stationary bicycle as fast as possible.

Finally, in Chapter 9, conclusions are given, and topics for future research are discussed.

Chapter 2

Survey of Computational Methods for Solving Optimal Control Problems

2.1 General Optimal Control Problem

The dynamical system considered is described by the differential equation

$$\dot{x}(t) = f(x(t), u(t), t), \tag{2.1.1}$$

$$x(t_0) = x_0, \tag{2.1.2}$$

which is subject to the control constraints

$$u(t) \in \Omega, \qquad \forall t \in [t_0, t_f], \tag{2.1.3}$$

where Ω is a compact subset of $\mathcal{R}^m$, the path constraints,

$$h(x(t), u(t), t) \leq 0, \tag{2.1.4}$$

for any $t \in [t_0, t_f]$, and the terminal-state constraints,

$$g_{eq}(x(t_f), t_f) = 0, \tag{2.1.5}$$

$$g_{ne}(x(t_f), t_f) \leq 0. \tag{2.1.6}$$

In the above, $x(t) \in \mathcal{R}^n$ is the state of the system at time t, and $u(t) \in \mathcal{R}^m$ the control at time t, with $t \in \mathcal{T} = [t_0, t_f]$. Let the set of admissible controls be,

$$\mathcal{U} = \{ u \mid u : \mathcal{T} \rightarrow \Omega \text{ is continuous a.e. } \}. \tag{2.1.7}$$

Let the set of feasible controls $\mathcal{F}$ consist of all the control functions $u \in \mathcal{U}$ such that the constraints (2.1.4), (2.1.5) and (2.1.6) are all satisfied. We may now formulate a general optimal control problem as follows:

Problem (P). Subject to the dynamical system (2.1.1), the initial condition (2.1.2), the path constraints (2.1.4), the terminal equality constraints (2.1.5), and the terminal inequality constraints (2.1.6), find a control $u \in \mathcal{U}$ such that the cost functional

$$J(u) = K(x(t_f), t_f) + \int_{t_0}^{t_f} L(x(\tau), u(\tau), \tau), d\tau \tag{2.1.8}$$

is minimized over $\mathcal{U}$.

The following conditions are assumed to be satisfied.

Assumption 2.1.1 $f : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T} \to \mathcal{R}^n$, $g_{eq} : \mathcal{R}^n \times \mathcal{T} \to \mathcal{R}^p$, $g_{ne} : \mathcal{R}^n \times \mathcal{T} \to \mathcal{R}^q$, $h : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T} \to \mathcal{R}^r$, $K : \mathcal{R}^n \times \mathcal{T} \to \mathcal{R}$, $L : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T} \to \mathcal{R}$. *$f$, h and L, together with their partial derivatives with respect to each of the components of x and u, are continuous for all $(x, u, t) \in \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T}$. K, g_{eq} and g_{ne} are continuously differentiable with respect to x;*

Assumption 2.1.2 *There exists a positive constant M such that*

$$|f(x, u, t)| \leq M(1 + |x|) \tag{2.1.9}$$

for all $(x, u, t) \in \mathcal{R}^n \times \Omega \times \mathcal{T}$.

Remark: From the theory of differential equations, the system (2.1.1)-(2.1.2) has a unique solution x^u corresponding to each $u \in \mathcal{U}$.

Remark: Because the state x^u is uniquely determined by control u when its initial condition $x(t_0)$ is fixed, the cost functional J dependents only on u. Thus, the above optimal control problem can also be viewed as an abstract optimization problem in function space

$$\min_{u \in \mathcal{F}} J(u)$$

where $\mathcal{F}$ is the feasible control set defined before.

2.2 Analytical Methods

There are three analytical approaches to find an optimal control. The first approach is based on working directly on the variational optimality conditions. Historically, those conditions are the necessary and/or sufficient conditions from the Calculus of Variations — a classical mathematical discipline whose objective was to find, among a family of functions, curves or surfaces, those which possess a certain extremal property [7,37]. Later, the necessary conditions from the Pontryagin minimum principle became dominant [4,120]. A great advantage of the Pontryagin minimum principle is its ability to handle hard constraints on the control, a problem classical variational theory could not readily handle. The second approach is based on the optimality conditions when the optimal control problem is viewed as an abstract optimization in a function space. Among them, the necessary conditions from the Generalized Kuhn-Tucker Theorem [26,85] have often been used. The same family of optimality conditions also includes some second-order necessary and sufficient conditions, which can be used for getting additional information on the optimum. The optimal controls derived by both the first and the second approaches are basically open-loop in nature. The third approach is based on dynamic programming [6,11] which leads to a first-order nonlinear partial differential equation called the Hamilton-Jacobi-Bellman (HJB) Equation and to a description of an optimal feedback control law.

For the first approach, usually, only necessary conditions are studied. Occasionally, some sufficient conditions are examined as well. Application of the necessary conditions which characterize the optimum leaves us with a limited number of candidates for a solution to the problem; it is then sometimes possible, for a problem which is simple and/or special enough, to use a process of elimination to determine which of these candidates is the sought-for solution. Finding candidates satisfying those variational necessary conditions will consequently result in a two-point boundary-value problem of order $2n$, where n is the dimension of the state of the system. An excellent book on this approach was by Athans and Falb [4] where the Pontryagin minimum

principle was used exclusively. In that book, a good variety of problems were solved and the results intuitively interpreted. Another excellent book was by Bryson and Ho [11] where many practical problems have been solved by applying a mixture of the calculus of variations and the Pontryagin minimum principle. Studies of the optimal control problem in the framework of general extremal problems can be found in the books by Ioffe and Tihomirov [58], and Zeidler [171]. Also, a comprehensive presentation of the classical calculus of variations using modern approaches can be found in the book by Cesari [14]. Other good books on this approach include the ones by Leitmann [77], and by Kirk [68].

The second approach has mainly been a theoretical tool rather than a computational one. One reason is that the generalized Kuhn-Tucker theorem is usually applied to a Banach space, which has a dual. People tend to study problems in a Hilbert control space, such as $\mathcal{L}_2^m[t_0, t_f]$, since any real Hilbert space is identified with its dual by a linear isometry. However, a more realistic control space seems to be $\mathcal{L}_\infty^m[t_0, t_f]$, whose dual space is complicated. Another reason is that the generalization of the optimal control problem into a much broader nonlinear functional optimization problem may easily ignore special structures and properties which are unique to the optimal control problem. For a simple case that the magnitude of each control variable is constrained, the corresponding generalized multipliers do not possess clear physical interpretation, and further, they are difficult to compute.

For the third approach, dynamic programming has also mainly been a theoretical tool rather than a computational tool. The reason is that dynamic programming suffers a drawback called the Curse of Dimensionality, which refers to the enormous dimension of the set of all possible final 'paths'.

Unless the system equations, the cost function, and the constraints are quite simple and/or special, finding optimal controls analytically is almost always impossible by any one of the above approaches. Consequently, numerical solutions are necessary.

2.3 Computational Methods — Nonparameterization Approaches

Computational approaches for solving optimal control problems, like those in a closely related field — optimization in finite-dimensional space, are iterative in nature. Their ultimate goals are to find a set of states, costates, controls and multipliers which satisfy certain optimality conditions. Most often, those optimality conditions are the necessary conditions supplied by the Pontryagin minimum principle [4,120] or its variants [59,60,93,105,106,140]. Sometimes, those optimality conditions can be the necessary conditions based on the Generalized Kuhn-Tucker Theorem [26,85]. Different characteristics of the iterative procedures give rise to many different methods.

Although there are a wide variety of methods for the computation of optimal controls, they can be classified into different types. One classification is based on the objectives of the methods at each iteration. Some methods are to find a new set of states, costates, controls and multipliers which satisfy the optimality conditions better than in the previous iteration. One example of such methods is the so-called neighboring extremal algorithm [11]. Most often, however, methods are to find a new set of states, costates, controls and multipliers which make the value of the cost functional smaller than in the previous iterations. Examples of such methods are the gradient method [117], the conditional gradient method [43], the projection method [43], the quasi-Newton's method [30,53,66,99,147,159], differential dynamic programming method [61,95,118], etc.

They can also be classified according to their updating schemes for the controls. Some methods approximate the control trajectories by orthogonal functions or polynomials. That is,

$$u(t) \approx \sum_{i=1}^{N} \alpha_i \phi_i(t)$$

where the $\{\phi_i(t)\}$ is a family of N orthogonal functions or polynomials [40,69,70,71, 81,137,143,150,151,155,156,157,168]. Sometimes, both the control and the state variables are approximated in this way [15,17,54,55,65,69,70,71,83,102,112,131,134,146, 163]. Consequently, the original optimal control problem is converted into an opti-

mization problem, where the updatings are on the parameters $\{\alpha_i\}$. Different selections of the $\{\phi_i(t)\}$ yield many different methods. Typical examples of the orthogonal functions are the Chebyshev [23,82,111,162,163], the Fourier [31,112,113,130,131,132], the Taylor [52,102,114,115,133,134,153], the Walsh [16,17,19,64], the block-pulse [54, 57,129], the Laguerre [22,56,146,165], the Legendre [55], the spline [69,70,71,81], and some other polynomial [40,143,150,151,155,156,157,168]. However, most well known methods update only the control variables, and the update scheme is such that, when at the k-th iteration,

$$u^{k+1} = P(u^k + \lambda^k s^k)$$

where P is a transformation mapping within the admissible control space $\mathcal{U}$, s^k a search direction, and λ^k a suitably chosen stepsize. Different constructions of the mapping P, search direction s^k and stepsize λ^k result in many different methods (algorithms): the gradient method [117], the conditional gradient method [43], the projection method [43], the quasi-Newton's method [30,53,66,99,147,159], etc. Finally, there is a strong variation type of updating scheme on the control

$$\begin{aligned} u^{k+1}(t) &= \overline{u}(t) \qquad \forall t \in I_{\alpha u} \\ u^{k+1}(t) &= u^k(t) \qquad \text{otherwise} \end{aligned}$$

where $I_{\alpha u}$ is a subset of $[t_0, t_f]$ in which the updating on control takes place, and $\overline{u}(t)$ is a control taht has certain properties. One example which uses such an updating scheme is the differential dynamic programming method [61,95,118].

The following is a survey which emphasizes only the major classes of computational methods. Special attention has been given to some selected representative methods.

2.3.1 Gradient Method

Let us recall that an optimal control problem of this chapter can be viewed as an abstract optimization problem

$$\min_{u \in \mathcal{F}} J(u)$$

where the cost functional J is a mapping from a function space $\mathcal{U}$ to the real line $\mathcal{R}$, and $\mathcal{F} \subset \mathcal{U}$ is a feasible control set. If J is defined on a neighborhood of $\bar{u}$ and if there exists a continuous linear mapping $J'(\bar{u}) : \mathcal{U} \to \mathcal{R}$ such that

$$J(\bar{u} + v) = J(\bar{u}) + J'(\bar{u})(v) + o(v), \qquad \forall v \in \mathcal{U} \tag{2.3.1}$$

with

$$\lim_{\|v\| \to 0} \frac{\|o(v)\|}{\|v\|} = 0. \tag{2.3.2}$$

J is called Fréchet differentiable at $\bar{u} \in \mathcal{U}$. If $\mathcal{U}$ is a Hilbert space, then $J'(\bar{u})$ is a linear continuous functional on $\mathcal{U}$, i.e. an element of the dual space $\mathcal{U}^*$. Riesz's representation theorem [85,138] tells us that a real Hilbert space is identified with its dual by a linear isometry. So, there exists an element of $\mathcal{U}$ itself, called the gradient $\nabla J(\bar{u})$, such that $J'(\bar{u})v = \langle \nabla J(\bar{u}), v \rangle$.

The gradient method is an iterative procedure for solving unconstrained optimization problems

$$\min_{u \in \mathcal{U}} J(u) \tag{2.3.3}$$

in a Hilbert space $\mathcal{U}$. The updating formula is $u^{k+1} = u^k - \lambda^k \nabla J(u^k)$, where the search direction, opposite to the direction of the gradient, is a steepest-descent direction.

To apply the gradient method to update the control, a formula for the gradient of the cost functional $J(u)$ must be obtained. It is known that [117]

$$\nabla J(u)(t) = \nabla_u H(x, u, p, t) \tag{2.3.4}$$

where

$$H(x, u, p, t) = L(x, u, t) + \langle p, f(x, u, t) \rangle \tag{2.3.5}$$

the Hamilton function of the system, and

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(x, u, p, t) \tag{2.3.6}$$

$$p(t_f) = \frac{\partial K}{\partial x}(x(t_f), t_f) \tag{2.3.7}$$

defines the costates of the system.

2.3.2 Conditional Gradient Method

Consider now the case where the control is constrained. In order to take advantage of the descent property of the gradient method, one immediate and intuitive extension is to find a direction which is, first, feasible, and second, pointing as close to $-\nabla J(u^k)$ as possible. This method is called the conditional gradient method, or Frank-Wolfe method. That is, for problems

$$\min_{u \in \mathcal{F}} J(u) \tag{2.3.8}$$

where $\mathcal{F}$ is a nonempty, closed, convex subset of a Hilbert space $\mathcal{U}$, the control is updated by

$$u^{k+1} = u^k + \lambda^k(\overline{u}^k - u^k) \tag{2.3.9}$$

where

$$\overline{u}^k = \arg \min_{u \in \mathcal{F}} \langle \nabla J(u^k),\ u - u^k \rangle \tag{2.3.10}$$

and

$$\lambda^k = \arg \min_{\lambda \in [0,1]} J(u^k + \lambda(\overline{u}^k - u^k)). \tag{2.3.11}$$

It has been shown that when an exact stepsize rule is used and J is convex on a convex set $\mathcal{F}$, this algorithm is well-defined and convergent [43].

In optimal control problems, one of the most convenient Hilbert spaces is $\mathcal{L}_2^m[0, t_f]$ equipped with inner product

$$\langle u, v \rangle = \int_{t_0}^{t_f} \sum_{i=0}^{m} u_i(\tau) v_i(\tau) d\tau.$$

So the subproblem of finding $\overline{u}^k$ becomes

$$\begin{aligned}
\overline{u}^k &= \arg \min_{u \in \mathcal{F}} \langle \nabla J(u^k), u - u^k \rangle \\
&= \arg \min_{u \in \mathcal{F}} \int_{t_0}^{t_f} \langle \nabla J(u^k)(\tau), u(\tau) \rangle d\tau \\
&= \arg \min_{u \in \mathcal{F}} \int_{t_0}^{t_f} H_u(x^k(\tau), p^k(\tau), u^k(\tau), \tau)\ u(\tau)\ d\tau. \qquad (2.3.12)
\end{aligned}$$

It is obvious that it can only be solved exactly for special cases.

Example:

Consider a system linear in control,

$$\begin{aligned}
\dot{x}(t) &= f_1(x(t), t) + f_2(x(t), t)u(t), \\
x(t_0) &= x_0.
\end{aligned}$$

The integrand of the cost function is also linear in the control,

$$\min_{u \in \mathcal{F}} J(u) = K(x(t_f), t_f) + \int_{t_0}^{t_f} \Big(L_1(x(\tau), \tau)) + L_2^{\mathsf{T}}(x(\tau), \tau)u(\tau) \Big) d\tau$$

and the feasible control set is

$$\mathcal{F} = \{\ u \mid u_j^{min} \leq u_j(t) \leq u_j^{max}, \quad j = 1, \cdots, m, \quad t \in [t_0, t_f]\ \} \subset \mathcal{L}_2^m[t_0, t_f].$$

The Hamiltonian is

$$H(x, u, p, t) = R(x, t) + S^{\mathsf{T}}(x, p, t)\ u(t)$$

where

$$\begin{aligned}
R(x, p, t) &= L_1(x, t) + p^{\mathsf{T}}(t) f_1(x, t) \\
S(x, p, t) &= L_2(x, t) + f_2^{\mathsf{T}}(x, t)\ p(t).
\end{aligned}$$

Then

$$\begin{aligned}
\overline{u}^k &= \arg \min_{u \in \mathcal{F}} \int_{t_0}^{t_f} H_u(x^k(\tau), p^k(\tau), u^k(\tau), \tau)\ u(\tau)\ d\tau \\
&= \arg \min_{u \in \mathcal{F}} \int_{t_0}^{t_f} S^{\mathsf{T}}(x^k(\tau), p^k(\tau), \tau)\ u^k(\tau)\ d\tau \\
&= \arg \min_{u \in \mathcal{F}} \sum_{j=1}^{m} \int_{t_0}^{t_f} S_j(x^k(\tau), p^k(\tau), \tau)\ u_j^k(\tau)\ d\tau.
\end{aligned}$$

The solution is

$$\overline{u}_j^k(t) = \begin{cases} u_j^{max}, & \text{when } S_j(x^k(t), p^k(t), t) < 0 \\ undefined, & \text{when } S_j(x^k(t), p^k(t), t) = 0 \\ u_j^{min}, & \text{when } S_j(x^k(t), p^k(t), t) > 0 \end{cases}$$

where $t \in [t_0, t_f]$ and $j = 1, \cdots, m$.

Remark: The Pontryagin minimum principle tells us that the optimal control for this example must be bang-bang, as long as $S_j(x^k(t), p^k(t), t)$ could take the value of zero only at isolated points. However, it is interesting to notice that, since u^{k+1} is a convex combination of u^k and $\overline{u}^k$, the conditional gradient method may not always generate bang-bang controls during intermediate iterations.

Remark: Under the assumption that $\mathcal{F}$ is convex, since u^{k+1} is a convex combination of u^k and $\overline{u}^k$, control functions $\{u^k\}$ generated will always be feasible, as long as the initial control function u^0 is feasible.

Remark: For the general linear quadratic problem subject to hard control constraints and linear terminal inequality constraints, which is studied in Chapter 5, the cost functional is convex in the control function, and the feasible control set F is a convex subset of $\mathcal{L}_\infty^m[t_0, t_f]$. In other cases, any sufficiently smooth functional could be approximated locally by a quadratic functional. Then the convergence property mentioned above would still hold for the approximating problem. It reminds us of a common practice in nonlinear programming where convex functions are used to approximate the cost and the constraints leading to a convex programming problem.

Remark: It is interesting to notice that both Sim's bang-bang control algorithm [148] and Mohler's algorithm [101] deal with the problem above. A common advantage of the two algorithms is that all controls generated during intermediate iterations will always be bang-bang, as long as the initial control is chosen to be bang-bang and there is no singularity. Both algorithms are shown to be quite computationally effective [101, 148], even though convergence is not proved in either case.

2.3.3 Projection Method

The basic problem in applying the gradient method to constrained optimization problems is that the gradient may point outside the set of feasible controls. Another way to find a feasible descent direction is the projection method. Consider the following constrained problem

$$\min_{u \in \mathcal{F}} J(u)$$

where $\mathcal{F}$ is a nonempty, closed, convex subset of a Hilbert space $\mathcal{U}$. Let a projection be an operator $P : \mathcal{U} \rightarrow \mathcal{F}$, where for each $v \in \mathcal{U}$ the image Pv is the unique element of $\mathcal{F}$ such that

$$||Pv - v|| \leq ||u - v|| \qquad \forall u \in \mathcal{F},$$

i.e., Pv is the vector in $\mathcal{F}$ which has the smallest distance to v. A popular projection method is the following:

Algorithm:

Step 0 Select a $u_0 \in \mathcal{F}$. Select a parameter $\alpha \in (0, \infty)$. Set $k = 0$.

Step 1 Solve projection $\bar{u}^k = P(u^k - \alpha \nabla J(u^k))$.

Step 2. If its solution is such that $\bar{u}^k = u^k$, stop. Otherwise, go to Step 3.

Step 3. Compute a suitable stepsize $\lambda^k > 0$, according to some stepsize rule.

Step 4. Set $u^{k+1} = u^k + \lambda_k(\bar{u}^k - u^k)$. Set $k = k+1$ and return to Step 1.

Proposition 2.3.1 *Let $J : \mathcal{U} \rightarrow \mathcal{R}$ be Fréchet differentiable.*

(i) If $\hat{u} \in \mathcal{F}$ is optimal, then for each $\alpha \in (0, \infty)$

$$P(\hat{u} - \alpha \nabla J(\hat{u})) = \hat{u}.$$

(ii) If $u^k \in \mathcal{F}$ and

$$v^k = P(u^k - \alpha \nabla J(u^k)) - u^k \neq \theta$$

for some $\alpha > 0$, then

$$\langle \nabla J(u^k),\ v^k \rangle \leq -\frac{1}{2\alpha}||v^k||^2 < 0.$$

Proof: (i) If $\hat{u} \in \mathcal{F}$ is optimal, then

$$\langle \nabla J(\hat{u}),\ u - \hat{u} \rangle \geq 0 \qquad \forall u \in \mathcal{F}.$$

Then, for all $u \in \mathcal{F}$ and each $\alpha \in (0, \infty)$,

$$\begin{aligned} &||u - (\hat{u} - \alpha \nabla J(\hat{u}))||^2 - ||\hat{u} - (\hat{u} - \alpha \nabla J(\hat{u}))||^2 \\ &\quad = \langle u - \hat{u} + \alpha \nabla J(\hat{u}),\ u - \hat{u} + \alpha \nabla J(\hat{u}) \rangle - \langle \alpha \nabla J(\hat{u}),\ \alpha \nabla J(\hat{u}) \rangle \\ &\quad = ||u - \hat{u}||^2 + 2\alpha \langle \nabla J(\hat{u}),\ u - \hat{u} \rangle \geq 0. \end{aligned}$$

That is

$$||\hat{u} - (\hat{u} - \alpha \nabla J(\hat{u}))|| \leq ||u - (\hat{u} - \alpha \nabla J(\hat{u}))||.$$

So

$$\hat{u} = P(\hat{u} - \alpha \nabla J(\hat{u})) \qquad \forall \alpha \in (0, \infty).$$

(ii) As defined,

$$\bar{u}^k = P(u^k - \alpha \nabla J(u^k)).$$

Then from the definition of projection,

$$||\bar{u}^k - (u^k - \alpha \nabla J(u^k))|| \leq ||u^k - (u^k - \alpha \nabla J(u^k))||, \qquad \forall u^k \in \mathcal{F}.$$

Then

$$\begin{aligned} 0 \ &\leq \ \langle \alpha \nabla J(u^k),\ \alpha \nabla J(u^k) \rangle - \langle v^k + \alpha \nabla J(u^k),\ v^k + \alpha \nabla J(u^k) \rangle \\ &= \ -2\alpha \langle \nabla J(u^k),\ v^k \rangle - \ ||v^k||^2. \end{aligned}$$

So

$$\langle \nabla J(u^k),\ v^k \rangle \leq -\frac{1}{2\alpha}||v^k||^2, \qquad \forall u^k \in \mathcal{F}.$$

Remark: From (i) it is clear that when u^k is optimal, $u^{k+1} = u^k$. However the converse is not necessarily true, even if J is convex in the control function u (the second conclusion of Theorem 8.1 and its proof in [43] are not correct).

Remark: The above (ii) tells us that the direction of v^k is really a descent direction.

Remark: It has been shown that if J is convex and Fréchet differentiable on $\mathcal{F}$, $\mathcal{F}$ is bounded, $\nabla J(u)$ Lipschitz continuous on $\mathcal{F}$, the above algorithm is convergent when either the exact step length rule, or Goldstein's rule, or Powell's rule is used [43].

Remark: Finding the projection during each iteration is by no means a trivial matter.

2.3.4 Quasi-Newton's Methods

Newton's method is a powerful computational method for finding the roots of a system of nonlinear equations. The idea is to replace the original equations by their linear approximation at the current estimate of the solution, and then, to solve the linearized version recursively. Its most appealing advantage is the quadratic rate of convergence near the solution. Newton's method is widely used in optimization in both finite-dimensional and infinite-dimensional spaces, because their first-order optimality conditions are in the form of either a system of nonlinear equations or a system of functionals. Consider first the following equality constrained optimization problem in finite-dimensional space

$$\min \quad f(x) \tag{2.3.13}$$

$$\text{s.t.} \quad h(x) = 0. \tag{2.3.14}$$

The first-order necessary condition that x^* be a relative minimum point is that, if x^* is a regular point for the constraints, there exists a λ^* such that

$$\begin{aligned} \nabla f(x^*) + \lambda^{*T} \nabla h(x^*) &= 0 \\ h(x^*) &= 0. \end{aligned}$$

To solve the above nonlinear equations, Newton's method is employed by solving the linearized version recursively. That is, given (x^k, λ^k) the new point (x^{k+1}, λ^{k+1}) is

determined from the equations

$$\begin{aligned} \nabla l(x^k,\lambda^k)^\top + L(x^k,\lambda^k)d^k + \nabla h(x^k)^\top y^k &= 0 \\ h(x^k) + \nabla h(x^k)d^k &= 0 \end{aligned}$$

by setting $x^{k+1} = x^k + d^k$, $\lambda^{k+1} = \lambda^k + y^k$, with Lagrange function $l(x,\lambda) = f(x) + \lambda^\top h(x)$ and $L(x,\lambda)$ its Hessian matrix with respect to x. In matrix form the above equations are

$$\begin{bmatrix} L(x^k,\lambda^k) & \nabla h(x^k)^\top \\ \nabla h(x^k) & 0 \end{bmatrix} \begin{bmatrix} d^k \\ y^k \end{bmatrix} = \begin{bmatrix} -\nabla l(x^k,\lambda^k)^\top \\ -h(x^k) \end{bmatrix}.$$

If $L(x^k,\lambda^k)$ is positive definite, (d^k, y^k) can be solved explicitly (see p430 of [86]):

$$\begin{aligned} d^k &= -L_k^{-1}(I - A_k^\top(A_kL_k^{-1}A_k^\top)^{-1}A_kL_k^{-1})I_k - L_k^{-1}A_k^\top(A_kL_k^{-1}A_k^\top)^{-1}h_k \\ y^k &= (A_kL_k^{-1}A_k^\top)^{-1}(h_k - A_kL_k^{-1}I_k) \end{aligned}$$

where $L_k = L(x^k,\lambda^k)$, $A_k = \nabla h(x^k)$, $I_k = \nabla l(x^k,\lambda^k)^\top$, $h_k = h(x^k)$. It can be shown easily that the d^k constructed above is a descent direction for the simple merit function: $m(x,\lambda) = \frac{1}{2}|\nabla l(x,\lambda)|^2 + \frac{1}{2}|h(x)|^2$. The above suggests a procedure for extending Newton's method to minimization problems with inequality constraints. Consider the problem

$$\min \quad f(x) \tag{2.3.15}$$

$$\text{s.t.} \quad g(x) \le 0. \tag{2.3.16}$$

Given an estimated solution point x^k and estimated Lagrange multipliers μ^k, one solves the quadratic programming problem

$$\min \quad \nabla f(x^k)d + \frac{1}{2}d^\top L_k d \tag{2.3.17}$$

$$\text{s.t.} \quad \nabla g(x^k)d + g(x^k) \le 0. \tag{2.3.18}$$

where L_k is the Hessian matrix for the Lagrange function $l(x^k,\mu^k,\mu^k) = f(x^k) + (\mu^k)^\top g(x^k)$, The new point is determined by $x^{k+1} = x^k + d^k$. The new Lagrange multipliers μ^{k+1} are the Lagrange multipliers of the above quadratic programming problem.

Because the Hessian matrix of the Lagrangian L_k is not always positive definite and L_k^{-1} is generally hard to evaluate, Newton's method is then modified by updating a positive definite matrix to approximate L_k^{-1}, or sometimes, L_k, a method which is called the quasi-Newton's method, or the variable metric method. A number of updating formulae have been suggested. Among them, BFGS is perhaps the most popular one, due to Broyden [10], Fletcher [33], Goldfard [41] and Shanno [145]. It is a modified version of the earlier Davidon's formula [29].

The algorithm described above becomes the famous Han-Powell method [45,46, 86,121]. As a general purpose algorithm, the Han-Powell method has several desirable features. First, it is globally convergent, if the new point is determined by $x^{k+1} = x^k + \lambda^k d^k$, where λ^k is a nonunity stepsize obtained by performing a line search in the direction of d^k, using the merit function: $\theta_r(x) = f(x) + r\sum_{i=1}^{m} \max\{0, g_i(x)\}$. Second, locally near the solution, if the step lengths are taken equal to unity, the iterative process converges superlinearly. Third, like all quasi-Newton's methods, the Han-Powell method employs only first-order information in order to achieve superlinear convergence.

However, extending the quasi-Newton's method to solve optimization problems in infinite-dimensional spaces is by no means a trivial matter. New difficulties are encountered. Let us first consider the unconstrained optimization in Hilbert space

$$\min_{u \in \mathcal{U}} \quad f(u)$$

where f is a functional defined on Hilbert space $\mathcal{U}$. Denote by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ the Banach space of all bounded linear operators mapping Banach space $\mathcal{X}$ into Banach space $\mathcal{Y}$. Let f' and f'' be the first and second Frechet derivatives of f respectively. Then by definition, (1) $f'(u) \in \mathcal{L}(\mathcal{U}, \mathcal{R})$ for each $u \in \mathcal{U}$, (2) $f''(u) \in \mathcal{L}(\mathcal{U}^2, \mathcal{R})$, or equivalently, $f''(u) \in \mathcal{L}(\mathcal{U}, \mathcal{U}^*)$, where $\mathcal{U}^*$ is the dual space of $\mathcal{U}$. Riesz's representation theorem tells us that a real Hilbert space is identified with its dual by a linear isometry. So, for Hilbert space $\mathcal{U}$, there are $\nabla f(u) \in \mathcal{U}$ and $F(u) \in \mathcal{L}(\mathcal{U}, \mathcal{U})$ such that

$$f'(u)(s) = \langle s, \nabla f(u) \rangle \qquad \forall s \in \mathcal{U}$$

and

$$f''(u)(s)(t) = \langle s, F(u)t\rangle \qquad \forall s, t \in \mathcal{U}$$

where $\nabla f(u)$ is called the gradient of f, and $F(u)$ is called the Hessian of f.

The first-order necessary condition that u^* be the minimizing function is that $\nabla f(u^*) = \theta \in \mathcal{U}$. To solve it, the quasi-Newton's method should give the following iteration rule, given α_k a stepsize from exact line search,

$$u^{k+1} = u^k - \alpha_k H^k \nabla f(u^k)$$

where operator H^k is the approximation to the inverse of $F(u^k)$. One of the difficulties in applying the quasi-Newton's method to infinite-dimensional spaces is how to update H^k. Horwitz [53] extended Davidon's formula into a real Hilbert space,

$$H^{k+1} = H^k + \frac{s^k\rangle\langle s^k}{\langle s^k,\ y^k\rangle} - \frac{H^k y^k\rangle\langle H^k y^k}{\langle y^k,\ H^k y^k\rangle} \tag{2.3.19}$$

$$y^k = \nabla f(u^{k+1}) - \nabla f(u^k) \tag{2.3.20}$$

$$s^k = u^{k+1} - u^k \tag{2.3.21}$$

where the initial value of the H operator is chosen to be any strongly positive linear self-adjoint operator in $\mathcal{U}$, $\langle a,\ b\rangle$ and $a\rangle\langle b$ denote the inner and outer dyadic products, respectively, on the given Hilbert space. Note that, if the space is n-dimensional Euclidean space, then

$$\langle a,\ b\rangle = a^\top b \qquad \text{and} \qquad a\rangle\langle b = ab^\top$$

There have been similar attempts to extended the BFGS formula into a real Hilbert space. For example, in [66,99,159]. Horwitz proved that[53], when f is a quadratic cost functional, the sequence $\{u^k\}$ thus obtained converges strongly to the minimizing element u^*. Suppose now the following constrained optimization problem is considered

$$\min_{u\in\Omega} f(u)$$

where f is a functional, and Ω a closed subspace of Hilbert space $\mathcal{U}$. Because $\mathcal{U} = \Omega \oplus \Omega^\perp$, with $\Omega^\perp$ the orthogonal complement of Ω, there is a projection operator P

of $\mathcal{U}$ onto $\Omega^{\perp}$. From a proposition discussed before, the necessary condition for $\hat{u}$ to be an optimal solution is that $P(\hat{u} - \lambda \nabla f(\hat{u})) = \hat{u}$, for each $\lambda \in (0, \infty)$. Suppose now Ω is at least one dimensional. It is known that [85] the projection operator P is also linear and bounded. Then the necessary condition becomes that $P(\nabla f(\hat{u})) = \theta \in \Omega$. For f still being a quadratic cost functional, it is proven in [53] that $\{u^k\}$ converges to u^* when $u^0 \in \Omega$ and $H^0 = P$.

If Ω is just a convex set, not a subspace, the problem becomes more complicated. In [30], an ad hoc approach is adapted to solve optimal control problems with bounded controls, which combines the Horwitz's algorithm for unconstrained problems with a truncation rule and a saturation function. The algorithm is shown to be numerically effective for some examples.

The most general constrained optimization problem in Hilbert space

$$\min_{s \in \mathcal{U}} \quad f(u) \tag{2.3.22}$$

$$s.t. \quad G(u) \leq \theta \tag{2.3.23}$$

where operator $G : \mathcal{U} \rightarrow \mathcal{Y}$, $\mathcal{U}$ and $\mathcal{Y}$ are Hilbert spaces, is equivalent to the problem

$$\min_{s \in \mathcal{U}} \quad f(u) \tag{2.3.24}$$

$$s.t. \quad G(u) \in \mathcal{Y}_{-} \tag{2.3.25}$$

where $\mathcal{Y}_{-}$ is the negative cone in $\mathcal{Y}$. Define the Lagrangian functional as $L(u, \lambda) = f(u) + \lambda^*(G(u))$, where $\lambda^* \in \mathcal{Y}^*$. From the generalized Kuhn-Tucker theorem, under regularity conditions, the necessary condition for u to be an optimal solution is that there exists a $\lambda \in (\mathcal{Y}^*)_{+}$ such that

$$L_u(u, \lambda^*) = f'(u) + \lambda^* G'(u) = \theta \tag{2.3.26}$$

$$\lambda^*(G(u)) = 0. \tag{2.3.27}$$

Notice that the above necessary condition looks exactly like the one in finite-dimensional spaces. The similarity suggests that the quasi-Newton's method for solving the inequality constrained optimization in finite-dimensional space can also be imitated here.

To solve the original problem, one instead solves the following quadratic programming problem

$$\min_{s \in \mathcal{U}} \quad f'(u^k)(s) + \frac{1}{2} B_k(s,s) \tag{2.3.28}$$

$$s.t. \quad G(u^k) + G'(u^k)s^k \leq \theta \tag{2.3.29}$$

where $s^k = u^{k+1} - u^k$, and B_k, an invertible and positive definite bilinear functional, is the approximation of $L_{uu}(u^k, \lambda^k)$. To updated B_k, either Davidon's formula or the BFGS formula discussed above can be used. The above direction-finding problem is generally difficult to solve. In [147], it is solved by solving its dual problem, using the gradient projection method.

2.3.5 Penalty Methods

In nonlinear programming, penalty methods are a family of procedures for approximating constrained problems by unconstrainted problems. They are divided into two classes: exterior and interior. The approximation is accomplished in the case of exterior penalty methods by adding to the objective function a term that assigns a high cost for violation of the constraints and in the case of interior penalty methods by adding a term that favors points interior to the feasible region over those near the boundary. There is a parameter ϵ associated with those methods that determines the severity of the penalty. It is shown in [86] that, as $\epsilon \to 0$, exterior penalty methods approach the solution of the original constrained problem from outside the active constraints, and interior penalty methods from inside the feasible region.

The penalty methods can be easily extended to solve constrained optimal control problems. For the problem setting at the beginning of this chapter, the exterior penalty methods lead to the following family of problems with no path constraint, and no terminal equality and inequality constraint:

$$\min_{u \in \mathcal{U}} \; J(u, \epsilon) = K(x(t_f), t_f) + \int_{t_0}^{t_f} L(x(\tau), u(\tau), \tau) d\tau \; + \; \frac{1}{\epsilon} J_P(u)$$

where

$$J_P(u) = ||g_{eq}(x(t_f), t_f)||^2 + \sum_{j=1}^{q} \max^2\{0,\ g_{ne}^j(x(t_f), t_f)\} + \sum_{j=1}^{r} \int_{t_0}^{t_f} \max^2\{0,\ h^j(x(\tau), u(\tau), \tau)\}\ d\tau \quad (2.3.30)$$

subject to

$$\dot{x}(t) = f(x(t), u(t), t) \quad (2.3.31)$$

$$x(t_0) = x_0 \quad (2.3.32)$$

Russell [139] considered problems where (a) f and L are linear in control u, (b) state trajectories are constrained to a compact set G_x, (c) control values are constrained to a convex and compact set Ω, (d) there is no terminal constraint. Under the assumption that an optimal trajectory is "approximable" from the interior of G_x (see definition 3.1 in [139]), he showed that the following holds. If $u_{\epsilon_k}(\cdot)$ is a solution obtained by the above exterior penalty method and $x_{\epsilon_k}(\cdot)$ the corresponding trajectory, and $\epsilon_k \to 0$, as $k \to \infty$, there is a subsequence of $\{u_{\epsilon_k}(\cdot)\}$ that converges to an optimal control $\hat{u}(\cdot)$ in the weak $\mathcal{L}_2 -$ topology, and the corresponding $\{x_{\epsilon_k}(\cdot)\}$ converges to the corresponding optimal trajectory $\hat{x}(\cdot)$ in the weak $\mathcal{C} -$ topology.

Cullum studied the more general case [27], in which (a) f and L are not necessarily linear in control u, (b) terminal constraints are present. She proved that, under certain conditions, if the above exterior penalty method is used to remove both the state constraints and the terminal constraints of the original problem P, the sequence of unconstrained problems P_{ϵ_k}, approximates problem P^R, the relaxation of P (see definition 4 in [27]). That is, the corresponding sequence of optimal costs of P_{ϵ_k} converges to the optimal cost of P^R, and there is a subsequence of $\{x_{\epsilon_k}(\cdot)\}$ that converges to an optimal trajectory of problem P^R. She also showed that the sequence of P_{ϵ_k} approximates the original problem P if the following additional conditions are satisfied: (a) f is linear in control u; (b) L is a convex function of u for each x; (c) Ω is convex, a case which includes the one Russell studied.

In summary, as in nonlinear programming, penalty methods are used to approximate a constrained optimal control problem by solving a sequence of unconstrained optimal control problems. However, unlike in nonlinear programming, the sequence of solutions generated may not always converge to the solution of the original problem. Finally, penalty methods exhibit slow convergence both in nonlinear programming and in optimal control problems.

2.3.6 Differential Dynamic Programming Methods

Of the computational methods for solving optimal control problem, those that bypass the Pontryagin minimum principle and dynamic programming by using nonlinear programming techniques or their extensions in function space, such as the projection method and some quasi-Newton's methods, do not and cannot fully utilize the special structure of optimal control problems. On the other hand, even though dynamic programming gives a complete solution conceptually, it can hardly be used as a computational method due to the drawback known as the curse of dimensionality.

Jacobson and Mayne [61] have invented differential dynamic programming (DDP) for solving discrete-time and continuous-time optimal control problems, which cleverly combines the principle behind dynamic programming and the conditions supplied by the Pontryagin minimum principle. Differential dynamic programming, which uses an estimate for the change in cost due to a strong variation in control, decreases the cost at each iteration.

Let us recall that, in most algorithms for solving optimal control problems, the new control u^α is normally constructed from the old control u according to a formula of the type

$$u^\alpha = u + \alpha\, s \tag{2.3.33}$$

where s is the search direction and α the step length. The s should be a direction which makes the cost functional descend. Then a stepsize α is to be determined which makes the largest reduction in the cost functional along direction s. However,

a distinguishing feature of DDP is its use of strong variations in control. That is, the new control is calculated according to the formula

$$u^{\alpha}(t) = \overline{u}(t) \qquad \forall t \in I_{\alpha u} \tag{2.3.34}$$

$$u^{\alpha}(t) = u(t) \qquad \text{otherwise} \tag{2.3.35}$$

where $\overline{u}(t)$ minimizes the Hamiltonian function, $I_{\alpha u}$ is a subset of $[0,1]$, the full time duration of the optimal control problems, having a measure of α, and $\alpha \in [0,1]$. Here, $I_{\alpha u}$ may be thought of as a step length. It is easy to see that $u^{\alpha}(t)$ can differ appreciably from $u(t)$ for some t even when the measure of $I_{\alpha u}$ is small. This gives rises to the terminology of "strong variations in control". Strong variations in control provide both a natural and a simple way to handle hard control constraints. It is worth noting that the Pontryagin minimum principle itself was proven by employing two special strong variations in control — temporal variation and spatial variation [4,120].

Gershwin and Jacobson [38], Havira and Lewis [48], and Mayne [94] made further discussions on DDP algorithms for solving discrete-time or continuous-time constrained optimal control problems. Even though all those differential dynamic programming type of algorithms have been found to be computationally effective, convergence results were not available. It was mainly due to the fact that $I_{\alpha u}$ was defined to be the end segment $[1-\alpha, 1]$. With this choice of $I_{\alpha u}$ a decrease in cost cannot be assured for all values of α, thus possibly resulting in a jam up before a local minimum is reached.

Polak and Mayne [95,118] modified the method by choosing $I_{\alpha u}$ to be any subset of $[0,1]$ having the following properties:

- if $\alpha \in [0, \mu(I_u)]$, $I_{\alpha u} \subset I_u$;
- if $\alpha \in (\mu(I_u), 1]$, $I_{\alpha u} \supset I_u$;
- if $\alpha \in [0, \mu(I_u)]$, $\{t \in I_u,\ t' \in I_{\alpha u},\ t < t'\} \Rightarrow \{t \in I_{\alpha u}\}$;
- if $\alpha \in (\mu(I_u), 1]$, $\{t \in I_u,\ t' \in I_{\alpha u} \backslash I_u,\ t < t'\} \Rightarrow \{t \in I_{\alpha u}\}$.

In the case without having terminal state inequality constraints, I_u is defined as [95]:

$$I_u \triangleq \{t \in [0,1] \mid \overline{H}(x^u(t), \bar{u}(t), \lambda^u(t), t) - H(x^u(t), u(t), \lambda^u(t), t) \leq \theta(u)\}$$

where

$$\overline{H}(x^u(t), \bar{u}(t), \lambda^u(t), t) \triangleq \min_{w \in \Omega} H(x^u(t), \omega, \lambda^u(t), t), \tag{2.3.36}$$

and

$$\bar{u}(t) \triangleq \arg\min_{\mu \in \Omega} H(x^u(t), \mu, \lambda^u(t), t), \tag{2.3.37}$$

and

$$\theta(u) \triangleq \int_0^1 (\overline{H}(x^u(t), \bar{u}(t), \lambda^u(t), t) - H(x^u(t), u(t), \lambda^u(t), t))dt. \tag{2.3.38}$$

Now, the subset $I_{\alpha u}$ may appear anywhere in the interval $[0,1]$, rather than just at the end of the interval as in the algorithms in [61]. By constructing $I_{\alpha u}$ this way, the first-order estimate of the change in cost, due to the strong variation in control described above, will always be bounded above by $\alpha\theta(u)$. By letting stepsize α be of *Armijo* type:

$$\alpha \triangleq \beta^{k(u,\overline{u})} \tag{2.3.39}$$

$$k(u,\overline{u}) \triangleq \min\{k \in I \mid J(u^\alpha) - J(u) \leq \beta^k \theta(u)/2, \ \ \beta \in (0,1)\}, \tag{2.3.40}$$

it is then proven that there is a positive $c < \infty$ such that

$$J(u^\alpha) - J(u) \leq -[\theta(u)]^2/4c. \tag{2.3.41}$$

Then, by applying the general convergence theorems of Polak [116], the algorithms were proven to be globally convergent [95,118]. However, it should be noted that those algorithms converge to optimality conditions that are somewhat weaker that those from the Pontryagin minimum principle.

In other developments, Virk [160,161] later extended the method in [95] to systems incorporating delays. Ohno [108] proposed a different type of DDP algorithm for solving discrete-time problems, whose constraints on control and state are

in the forms of equalities and inequalities. In Ohno's algorithm, the Kuhn-Tucker conditions are first applied to the recursive relation of dynamic programming. Then, Newton's method is employed to find the control and the Lagrange multipliers satisfying the condition. However, its convergence, ensured by the locally convergent Newton's method, is obviously not as strong as those convergence properties enjoyed by the algorithms proposed by Polak and Mayne [95,118]. Murray and Yakowitz [103] made a detailed comparison between Newton's method when applied to discrete-time unconstrained optimal control problems, and the method of discrete-time differential dynamic programming described in [61]. Their main conclusions were (1) DDP does not coincide with Newton's method, but (2) they are close enough that they enjoy the same quadratic convergence rate.

2.4 Computational Methods — Parameterization Approachs

The parameterization methods are techniques which convert the original optimal control problem into a finite-dimensional optimization by employing some special functions to approximate control and/or state trajectories. Due to the availability of well-developed nonlinear programming techniques and powerful computers, the parameterization methods are handy in dealing with optimal control problems with general constraints. The parameterization methods are based on the theory of approximation. Different approximation methods give rise to many different parameterization methods.

2.4.1 Methods of Approximating a Function

From the theory of approximation, the partial summation $s_n(t) = \sum_{k=0}^{n} \alpha_k \phi_k(t)$ can be chosen to converge to a well-behaved function $f(t)$ on an interval $[a, b]$, where $\{\alpha_k\}$ is a set of real parameters and $\{\phi_k(t)\}$ is a set of some special functions which are called the basis functions. In other words, when n is sufficiently large, $s_n(t)$ is a good approximation to $f(t)$ on $[a, b]$.

It is evident that a minimum requirement for the approximation to be valid is that $s_n(t)$ converge pointwise to $f(t)$ on $[a,b]$. Uniform convergence, which is stronger than pointwise convergence, is sometimes preferred. All in all, the interval of approximation $[a,b]$ must lie within the convergence domain of $s_n(t)$.

There is a wide class of basis functions possessing certain convergence properties for well behaved $f(t)$. Different choices of basis functions yield different approximation methods.

Approximation by a Polynomial

Consider the basis functions $\{t^k\}$. Then, the corresponding method is to approximate $f(t)$ by an n-th order polynomial. The convergence of the method is assured by the famous Weierstrass theorem (see, for example, [85,122,135]):

Theorem 2.4.1 (Weierstrass) *The space of polynomials is dense in $C[a,b]$ — the space of continuous functions on the interval $[a,b]$. Equivalently speaking, given $f(t)$ a continuous function and ϵ a positive number, there exists a polynomial $P_n(f,t)$ of degree $n(\epsilon)$ such that*

$$|P_n(f,t) - f(t)| < \epsilon \tag{2.4.1}$$

for any $t \in [a,b]$.

A major drawback of the method is that a high order polynomial, which is often needed to achieve a good approximation, can display undesirable oscillatory behavior. Let us suppose a polynomial $P_n(t)$ has order n higher than 1. Connect the two end points $P_n(a)$ and $P_n(b)$ on the curve of $P_n(t)$ by the line:

$$L(t) = P_n(a) + \frac{P_n(b) - P_n(a)}{(b-a)}(t-a).$$

Because $Q_n(t) \triangleq P_n(t) - L(t)$ is still an n-th order polynomial, $Q_n(t)$ can have n zeros according to the Fundamental Theorem of Algebra. Consequently, the curve of $P_n(t)$ can cross line $L(t)$ as many as n times during $[a,b]$, which means that $P_n(t)$ may be very oscillatory.

Approximation by Piecewise Polynomials

Instead of approximating a given function $f(t)$ over an interval $[a, b]$ by a single polynomial, one may subdivide $[a, b]$ by a mesh of points:

$$\Delta: \quad a = t_0 < t_1 < \cdots < t_N = b$$

and approximate $f(t)$ by a different polynomial on each subinterval $[t_{k-1}, t_k]$. The polynomials at different subintervals are usually selected to have low orders to avoid oscillatory behavior.

When the piecewise-polynomials are all of degree 0, the approximation is by piecewise-constants, which is generally discontinuous at each division point t_k. When the piecewise-polynomials are all of degree 1, the approximation is by piecewise linear segments, which can be made continuous at each division point t_k, while the derivative at each t_k is generally discontinuous. This kind of approximation $s_\Delta(t)$ is often called a linear spline with respect to mesh Δ. For some purposes, it is highly desirable that the joints of separate arcs be as "smooth" as possible. Specifically, if it is required that in each subinterval the approximation $s_\Delta(t)$ be a polynomial of maximum degree 3, that $s_\Delta(t)$ agree with $f(t)$ at each of the $N+1$ points $t_0, t_1, \cdots, t_N$, and that the first and second derivatives $s'_\Delta(t)$ and $s''_\Delta(t)$ be continuous on $[a, b]$, then $s_\Delta(t)$ is called a cubic spline with respect to mesh Δ.

In any subinterval $[t_{k-1}, t_k]$, the cubic spline function $s_\Delta(t)$ can be expressed by the following formula, denoting $h_k = t_k - t_{k-1}$,

$$\begin{aligned} s_\Delta(t) = {} & s'_{k-1}\frac{(t-t_k)^2(t-t_{k-1})}{h_k^2} - s'_k\frac{(t-t_{k-1})^2(t-t_k)}{h_k^2} \\ & + f_{k-1}\frac{(t_k-t)^2(t-t_{k-1}+h_k/2)}{2h_k^3} + f_k\frac{(t-t_{k-1})^2(t_k-t+h_k/2)}{2h_k^3}. \end{aligned} \tag{2.4.2}$$

The values $s'_{k-1} \triangleq s_\Delta(t_{k-1})$ and $s'_k \triangleq s_\Delta(t_k)$ are unknowns needing to be identified. It is clear that $s_\Delta(t)$, $s'_\Delta(t)$ and $s''_\Delta(t)$ are all continuous at the interior of each $[t_{k+1}, t_k]$. It is also easy to verify that

$$s_\Delta(t_k-) = s_\Delta(t_k+), \qquad s'_\Delta(t_k-) = s'_\Delta(t_k+)$$

for $k = 1, 2, \cdots, N-1$. That is, both $s_\Delta(t)$ and $s'_\Delta(t)$ are continuous at each interior node t_k. The requirement that $s''_\Delta(t)$ be continuous at each interior node t_k leads to the following linear difference equation [1,8,50]

$$\frac{1}{h_{k+1}} s'_{k+1} + 2\left(\frac{1}{h_k} + \frac{1}{h_{k+1}}\right) s'_k + \frac{1}{h_k} s'_{k-1} = 3\frac{f_k - f_{k-1}}{h_k^2} + 3\frac{f_{k+1} - f_k}{h_{k+1}^2} \qquad (2.4.3)$$

for $k = 1, 2, \cdots, N-1$. Once two appropriate auxiliary conditions on $s'_\Delta(t_k)$ are prescribed, the above difference equation can be solved and its solution serves to determine the cubic spline function $s_\Delta(t_k)$ at each subinterval of $[a, b]$. A common selection of the two auxiliary conditions are the end conditions

$$s'_0 = f'(a), \qquad s'_N = f'(b), \qquad (2.4.4)$$

which means that the derivatives of the cubic function $s_\Delta(t_k)$ agree with the derivatives of $f(t)$ at both end points of $[a, b]$. Another common selection is

$$2s'_0 + s'_1 = 3\frac{(f_1 - f_0)}{h_1}, \qquad s'_{N-1} + 2s'_N = 3\frac{(f_N - f_{N-1})}{h_N}, \qquad (2.4.5)$$

which is equivalent to the end conditions

$$s''_0 = 0, \qquad s''_N = 0. \qquad (2.4.6)$$

A cubic spline which satisfies (2.4.6) is often called a natural cubic spline. Different selections of the two auxiliary conditions give rise to many different forms of cubic functions. A very efficient algorithm to solve (2.4.3) under general end conditions

$$2s'_0 + \mu_0 s'_1 = c_0, \qquad \lambda_N s'_{N-1} + 2s'_N = c_N, \qquad (2.4.7)$$

was given by Ahlberg et al (see p.14 of [1]). The following theorem shows the convergence of the cubic spline with end conditions (2.4.4) (see Theorem 2.3.2 in [1]):

Theorem 2.4.2 *Let $\{\Delta_i\}$ be a sequence of meshes on $[a, b]$ with $\lim_{i\to\infty} \|\Delta_i\| = 0$. Let $f(t)$ be of class $C^1[a, b]$. Let the cubic spline function $s_{\Delta_i}(t)$ satisfy the end conditions in (2.4.4). Then we have,*

$$f^{(p)}(t) - s^{(p)}_{\Delta_i}(t) = o(\|\Delta_i\|^{1-p}) \qquad p = 0, 1, \quad i = \{0, 1, \cdots, \infty\} \qquad (2.4.8)$$

uniformly with respect to t in $[a, b]$.

Notice from the above that the error between the first derivative of the cubic spline $s_\Delta(t)$ and the first derivative of the function being approximated $f(t)$ can always be bounded by a constant for any sequence of meshes, as long as the first derivative of $f(t)$ is continuous. The convergence property of the cubic spline with general end conditions (2.4.7) is discussed in Theorem 2.9.2 in [1]. When function $f(t)$ is smoother, better convergence results can be obtained. See, for example, the following theorem in [44]:

Theorem 2.4.3 *Let $\{\Delta\}$ is a mesh on $[a,b]$. Define*

$$h = \max_k (t_{k+1} - t_k), \tag{2.4.9}$$

and

$$\beta = h / \min_k (t_{k+1} - t_k), \tag{2.4.10}$$

for $k = 0, 1, \cdots, N-1$. Let $f(t)$ be of class $C^4[a,b]$. Let the cubic spline function $s_\Delta(t)$ satisfy the end conditions in (2.4.4) or in (2.4.6). Then we have,

$$\|f^{(p)} - s_\Delta^{(p)}\|_\infty = C_p \|f^{(4)}\|_\infty h^{4-p} \qquad p = 0, 1, 2, 3 \tag{2.4.11}$$

with

$$C_0 = 5/384, \quad C_1 = 1/24, \quad C_2 = 3/8, \quad C_3 = (\beta + \beta^{-1})/2. \tag{2.4.12}$$

There is another way to represent cubic spline functions. Let $\mathcal{P}_3(\Delta)$ denote the space of all cubic spline functions s_Δ with respect to mesh Δ. It can be shown that (see, for example, Theorem 1.17 of [107]) $\mathcal{P}_3(\Delta)$ is a linear space of dimension $N+3$, and it is spanned by the basis

$$\{1, t, t^2, t^3, (t - t_1)_+^3, \cdots, (t - t_{N-1})_+^3\} \tag{2.4.13}$$

where

$$(t - t_k)_+^3 = \max\{0, (t - t_k)^3\} \qquad k = 1, \cdots, N-1. \tag{2.4.14}$$

So, any cubic spline function $s_\Delta \in \mathcal{P}_3(\Delta)$ has a unique representation with respect to the above basis

$$s_\Delta(t) = \sum_{k=0}^{3} a_k t^k + \sum_{k=1}^{N-1} b_k (t - t_k)_+^3, \qquad t \in [a,b]. \tag{2.4.15}$$

Similar to the case in the space of $\mathcal{R}^n$, the basis of the finite-dimensional linear space $\mathcal{P}_3(\Delta)$ is also not unique. A search for restricted support local bases for spline functions leads to the B-splines studied by Schoenberg [144] and by Curry and Schoenberg [28]:

Theorem 2.4.4 *For each $k \in \{-3, -2, -1, 0, 1, \cdots, N-1\}$, there exists a unique cubic spline B_k with respect to mesh Δ such that*

$$B_k(t) \begin{cases} > 0 & \text{if } t \in (t_k, t_{k+4}) \\ = 0 & \text{if } t \notin (t_k, t_{k+4}) \end{cases} \tag{2.4.16}$$

and

$$\int_{t_k}^{t_{k+4}} B_k(t)\, dt = 1. \tag{2.4.17}$$

The spline B_k in above theorem is called the cubic B-spline with support region $[x_k, x_{k+4}]$. The next theorem is due to Curry and Schoenberg [28]:

Theorem 2.4.5 *The set of B-splines*

$$\{B_{-3}, B_{-2}, B_{-1}, B_0, \ldots, B_{N-1}\} \tag{2.4.18}$$

forms a basis of $\mathcal{P}_3(\Delta)$ on $[a, b]$.

It follows from the above theorem that every spline $s_\Delta \in \mathcal{P}_3(\Delta)$ has a unique representation

$$s_\Delta(t) = \sum_{k=-3}^{N-1} \alpha_k B_k(t), \qquad t \in [a, b]. \tag{2.4.19}$$

So, if a function is to be approximated by a cubic spline function corresponding to $N+1$ mesh points, the approximation can then be carried out by a partial summation of $N+3$ B-splines. It is clear that the value of any cubic spline function at a point other than a mesh point can be represented by a linear combination of exactly four B-splines. At the mesh points themselves, the number of basis functions required for this representation reduces to three. Another nice property of the B-splines is that [28] they are of "minimal support", i.e. no basis functions can be found which have smaller regions than those of the B-splines.

When the mesh Δ is on the interval $[0,1]$, and it is equally spaced between adjacent points, the construction of B-splines is greatly simplified. It is easy to check that, the following piecewise cubic polynomial, with $x_0=0$, $x_N=1$, $h=1/N$,

$$B_0(t) = \begin{cases} 0 & t \in (-\infty, t_0] \\ (t-t_0)^3 & t \in (t_0, t_1] \\ h^3 + 3h^2(t-t_1) + 3h(t-t_1)^2 - 3(t-t_1)^3 & t \in (t_1, t_2] \\ 4h^3 - 6h(t-t_2)^2 + 3(t-t_2)^3 & t \in (t_2, t_3] \\ h^3 - 3h^2(t-t_3) + 3h(t-t_3)^2 - (t-t_3)^3 & t \in (t_3, t_4] \\ 0 & t \in (t_4, \infty) \end{cases} \tag{2.4.20}$$

is really a cubic B-spline function, which satisfies the conditions in theorem 2.4.4 [104]. Thus, the complete set of B-splines which forms a basis of $\mathcal{P}_3(\Delta)$ on $[0,1]$ are the following translates of B_0:

$$B_k(t) = B_0(t-kh), \qquad k = -3, -2, -1, 0, \cdots, N-1. \tag{2.4.21}$$

The subspace $\mathcal{P}_3'(\Delta) \subset \mathcal{P}_3(\Delta)$ which is obtained by imposing the additional conditions

$$s_\Delta(0) = \dot{s}_\Delta(0) = \ddot{s}_\Delta(0) = 0, \quad s_\Delta(1) = \dot{s}_\Delta(1) = \ddot{s}_\Delta(1) = 0, \tag{2.4.22}$$

for any $s_\Delta(t) \in \mathcal{P}_3(\Delta)$, has dimension N–3, and is spanned by the basis $\{B_0, \ldots, B_{N-4}\}$. The subspace $\mathcal{P}_3''(\Delta) \subset \mathcal{P}_3(\Delta)$ which is obtained by imposing the additional conditions

$$s_\Delta(0) = \dot{s}_\Delta(0) = \ddot{s}_\Delta(0) = 0 \tag{2.4.23}$$

for any $s_\Delta(t) \in \mathcal{P}_3(\Delta)$, has dimension N, and is spanned by the basis $\{B_0, \ldots, B_{N-1}\}$.

Approximation by Taylor Series

When $f(t)$ behaves better than just continuous, it can be approximated by a special type of polynomial — the Taylor polynomial

$$f(t) \approx f(t_0) + f'(t_0)(t-t_0) + \cdots + f^{(n)}(t_0)\frac{(t-t_0)^n}{n!}. \tag{2.4.24}$$

It is known that, if $f(t)$ is analytic at a point x_0, the Taylor series

$$\sum_{k=0}^{\infty} f^{(k)}(x_0)\frac{(x-x_0)^k}{k!} \tag{2.4.25}$$

converges to $f(t)$ in a neighborhood of t_0 (see, for example, [12]). Major drawbacks of the Taylor series approximation are that, (1) it can only be applied to sufficiently smooth functions, because high-order derivatives are required to form the Taylor polynomial; (2) its convergence rate is often very slow.

Approximation by Fourier Series

Most often, the basis functions $\{\phi_k(t)\}$ are orthogonal functions with respect to some weight function $\{p(t)\}$, which is defined below:

Definition If a set of real functions $\{\phi_k(t)\}$ has the property that over some interval $[a,b]$, finite or infinite,

$$\int_a^b p(t)\phi_m(t)\phi_n(t)\,dt \begin{cases} = 0 & \text{if } m \neq n \\ \neq 0 & \text{if } m = n \end{cases} \tag{2.4.26}$$

then the functions are said to be orthogonal with respect to the weight function $p(t)$ on that interval.

Why are orthogonal functions preferred? Let us assume that the partial summation $\sum_{k=0}^{n}\alpha_k\phi_k(t)$, where $\{\phi_k(t)\}$ is a set of functions orthogonal with respect to a weight function $p(t)$ on an interval $[a,b]$, converges uniformly to a function $f(t)$ on $[a,b]$. So, for any $t\in[a,b]$,

$$f(t) = \sum_{k=0}^{\infty}\alpha_k\phi_k(t). \tag{2.4.27}$$

Multiply both sides of the above equation by $p(t)\phi_k(t)$ and assuming that the order of $\sum_{k=0}^{\infty}$ and $\int_a^b$ can be interchanged, we have

$$\int_a^b f(t)p(t)\phi_k(t)\,dt = \sum_{i=0}^{\infty}\alpha_i\int_a^b p(t)\phi_k(t)\phi_i(t)\,dt. \tag{2.4.28}$$

From the property of orthogonality, we finally have

$$\alpha_k(t) = \frac{\int_a^b f(t)p(t)\phi_k(t)\,dt}{\int_a^b p(t)\phi_k^2(t)\,dt} \tag{2.4.29}$$

for $k = 0, 1, \cdots$. Then, the partial summation $\sum_{k=0}^{n} \alpha_k\phi_k(t)$, with α_k's defined as above, will be a good approximation to $f(t)$ when n is sufficiently large. The simplicity in determining the coefficients, the α_k's, which are called the orthogonal coefficients, is one of many advantages of using orthogonal functions in approximating functions.

However, the orthogonality of the ϕ_k's does not guarantee that the partial summation $\sum_{k=0}^{n} \alpha_k\phi_k(t)$ converges to $f(t)$, or even converges at all. Without an assurance of convergence, the approximation of $f(t)$ by using $\sum_{k=0}^{n} \alpha_k\phi_k(t)$ does not make any mathematical sense, regardless of whatever the ϕ_k's are orthogonal or not.

Consider now trigonomic functions $1, \cos t, \sin t, \cos 2t, \sin 2t, \cdots, \cos kt, \sin kt, \cdots$. These functions are orthogonal with respect to the unit weighting function on the interval $[-\pi, \pi]$. Correspondingly, a function $f(t)$ can be approximated by the following trigonomic polynomial, which is called the Fourier polynomial,

$$f(t) \approx \frac{1}{2}\alpha_0 + \sum_{k=1}^{n}(\alpha_k \cos kt + \beta_k \sin kt) \tag{2.4.30}$$

where the orthogonal coefficients, α_k's and β_k's are obtained by applying (2.4.29)

$$\alpha_k = \frac{\int_{-\pi}^{\pi} f(t)\cos kt\,dt}{\int_{-\pi}^{\pi} \cos^2 kt\,dt} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos kt\,dt, \tag{2.4.31}$$

$$\beta_k = \frac{\int_{-\pi}^{\pi} f(t)\sin kt\,dt}{\int_{-\pi}^{\pi} \sin^2 kt\,dt} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin kt\,dt. \tag{2.4.32}$$

When $n \to \infty$, the Fourier polynomial becomes a Fourier series. The following theorem concerns both the pointwise and the uniform convergence of the Fourier series (see page 106-107 of [142]):

Theorem 2.4.6 *(1) If $f(t)$ is of bounded variation in $[-\pi, \pi]$, then the Fourier series converges to $f(t)$ at every point of continuity and converges to the value $\frac{1}{2}(f(x-0)+f(x+0))$ at every point of discontinuity. At both end points $t=-\pi$ and $t=\pi$, the Fourier series converges to the same value $\frac{1}{2}(f(-\pi)+f(\pi))$; (2) If $f(t)$ is of bounded*

variation in $[-\pi, \pi]$ *and has only a finite number of discontinuities, its Fourier series converges uniformly to* $f(t)$ *in the interior of any interval in which* $f(t)$ *is continuous.*

The Fourier series was originally used to expand periodic functions. For a nonperiodic function of bounded support, the expansion can still be carried out after making a periodic extension of $f(t)$ from $[-\pi, \pi]$ onto the whole t-axis. If $f(-\pi) \neq f(\pi)$, this extension creats 'jump's at every $k\pi$ throughout the t-axis. Because the Fourier series is a linear combination of trigonomic functions which are both periodic and smooth, it is then easy to see that the convergence of the Fourier series should be slow when $f(t)$ has 'jump's. Similarly, because the periodic extension of a nonperiodic function $f(t)$ with $f(-\pi) \neq f(\pi)$ contains 'jump's at every $k\pi$, the convergence of its Fourier series should also be slow no matter how smooth the function is in $[-\pi, \pi]$.

(d) Approximation by Chebyshev Series

A more natural approach to obtain periodicity of a nonperiodic function is through a variable change rather than through a periodic extension. Suppose $f(t)$ is defined on $[-1, 1]$, such a variable change is given by $t = \cos\theta$. Then

$$f(t) = f(\cos\theta) \triangleq g(\theta), \qquad \theta \in [0, \pi]. \tag{2.4.33}$$

Now, $g(\theta)$ is periodic as well as even. Expanding it into a Fourier series, all the sine coefficients, β_k's, become zero and only $\cos kt$ terms remain,

$$g(\theta) = \frac{1}{2}\alpha_0 + \sum_{k=0}^{\infty} \alpha_k \cos k\theta. \tag{2.4.34}$$

By letting

$$T_k(t) \triangleq \cos(k \cos^{-1} t), \tag{2.4.35}$$

the Fourier cosine coefficients, α_k's, become

$$\alpha_k = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\theta)\cos k\theta \, d\theta = \frac{2}{\pi}\int_{-1}^{1} \frac{f(t)T_k(t)}{\sqrt{1-t^2}}\, dt. \tag{2.4.36}$$

It can be easily verified that $T_k(t)$ is a k-th order polynomial which satisfies the following recurrence relation

$$T_0(t) = 1, \qquad T_1(t) = t, \tag{2.4.37}$$

$$T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t), \quad \forall k \geq 1. \tag{2.4.38}$$

It can also be shown that $\{T_k(t)\}$ is a set of orthogonal functions with respect to weight function $(1-x^2)^{-1/2}$ on the interval $[-1,1]$

$$\int_a^b \frac{T_m(t)T_n(t)}{\sqrt{1-t^2}}\,dt = \begin{cases} 0 & \text{if } m \neq n, \\ \pi/2 & \text{if } m = n = 0, \\ \pi & \text{if } m = n \neq 0. \end{cases} \tag{2.4.39}$$

These $T_k(t)$'s are called the Chebyshev polynomials of the first kind [34,122,152,158]. Now, the Fourier expansion (2.4.34) can be rewritten as

$$f(t) = \frac{1}{2}\alpha_0 T_0(t) + \sum_{k=0}^{\infty} \alpha_k T_k(t). \tag{2.4.40}$$

Because

$$\int_{-1}^{1} \frac{T_k(t)}{\sqrt{1-t^2}}\,dt = \int_0^{\pi} \cos^2 k\theta\,d\theta = \begin{cases} \pi & \text{if } k = 0 \\ \pi/2 & \text{if } k > 0 \end{cases} \tag{2.4.41}$$

the α_k's defined in (2.4.36) satisfy (2.4.29). Hence, those α_k's are exactly orthogonal coefficients, and (2.4.40) is exactly an orthogonal expansion of $f(t)$ by the Chebyshev polynomials. Because the Chebyshev series can be viewed as being derived from a corresponding Fourier series by a simple variable change, it should possess similar convergence properties to those of the Fourier series. To study the convergence of the Chebyshev series, let's introduce $s_n(f,t)$ as

$$s_n(f,t) = \frac{1}{2}\alpha_0 T_0(t) + \sum_{k=0}^{n} \alpha_k T_k(t). \tag{2.4.42}$$

The following theorem shows that $s_n(f,t)$ is the least squares approximations of f with respect to the weight function $(1-t^2)^{-1/2}$ (see p.166 of [136]).

Theorem 2.4.7 *If $f(t)$ is continuous on $[-1,1]$, then*

$$\int_{-1}^{1} \Big(f(t) - s_n(f,t)\Big)^2 \frac{dt}{\sqrt{1-t^2}} \leq \int_{-1}^{1} \Big(f(t) - p_n(t)\Big)^2 \frac{dt}{\sqrt{1-t^2}} \tag{2.4.43}$$

for every n-th order polynomial $p_n(t)$, equality holding only for $p_n = s_n(f)$.

Let's now introduce a quantity measuring the smoothness of a function f, called the modulus of continuity of f,

$$\omega_n(\delta, f) \triangleq \sup_{\substack{x_1, x_2 \in [-1,1] \\ |x_1 - x_2| \leq \delta}} |f(x_1) - f(x_2)| \tag{2.4.44}$$

for $\delta > 0$. Also, if $\omega_n(\delta, f) \leq C\delta^\alpha$, where $\alpha > 0$ and C is a constant not depending on δ, f is said to belong to the Lipschitz class of order α. It is clear that f is continuous on $[a, b]$ if, and only if, $\omega_n(\delta) \to 0$, as $\delta \to 0$. The following theorem shows that, under a condition which is stronger than continuity, the unform convergence of the Chebyshev series is guaranteed (see p.168 of [136]).

Theorem 2.4.8 *If the modulus of continuity of a function f satisfies*

$$\lim_{n \to \infty} \omega\left(\frac{1}{n}\right) \ln n = 0, \tag{2.4.45}$$

then the Chebyshev series of f converges uniformly to f on $[-1, 1]$. In particular, if f belongs to the Lipschitz class of order α for some $\alpha \in (0, 1]$, the Chebyshev series of f converges to f uniformly on $[-1, 1]$.

Approximation by Legendre Series

A class of orthogonal functions $\{\phi_k(t)\}$ can be defined by Rodrigue's formula:

$$\phi_k(t) = (1 - t^2)^{-\alpha} \frac{d^k}{dt^k}(1 - t^2)^{k+2}, \tag{2.4.46}$$

which generates orthogonal polynomials on the interval $[-1, 1]$ with respect to the weight function $(1 - t^2)^\alpha$, for $-1 < \alpha < \infty$. These orthogonal polynomials are called the ultra-spherical polynomials. When $\alpha = -1/2$, it generates the Chebyshev polynomials. When $\alpha = 0$, the polynomials being generated are called the Legendre polynomials. It can be shown that a Legendre polynomial $P_k(t)$ is a k-th order polynomial which satisfies the following recurrence relation [152]

$$P_0(t) = 1, \qquad P_1(t) = t, \tag{2.4.47}$$

$$P_{k+1}(t) = \left(\frac{2k+1}{k+1}\right)tP_k(t) - \left(\frac{k}{k+1}\right)P_{k-1}(t). \tag{2.4.48}$$

Also, $\{P_k(t)\}$ is a set of orthogonal functions with respect to the unit weight function on the interval $[-1, 1]$, because

$$\int_a^b P_m(t)P_n(t)\,dt = \begin{cases} 0 & \text{if } m \neq n, \\ 2/(2k+1) & \text{if } m = n. \end{cases} \tag{2.4.49}$$

So, $\{P_k(t)\}$ can be used to approximate function $f(t)$

$$f(t) \approx \sum_{k=0}^{n} \alpha_k P_k(t) \tag{2.4.50}$$

where the orthogonal coefficients, α_k's, are obtained by applying (2.4.29)

$$\alpha_k = \frac{\int_{-1}^{1} f(t)P_k(t)\,dt}{\int_{-1}^{1} P_k^2(t)\,dt} = \frac{2k+1}{2}\int_{-1}^{1} f(t)P_k(t)\,dt. \tag{2.4.51}$$

It has been shown that the resulting Legendre series possess similar convergence properties to those of the Fourier series and the Chebyshev series (see, for example, [142]).

Besides convergence, the rate of the convergence is equally important. If a series converges fast, fewer terms are necessary to approximate a function. Suppose a function $f(t)$ is expanded by the basis functions, $\phi_k(t)$'s on the interval $[a, b]$. That is,

$$f(t) = \sum_{k=0}^{\infty} \alpha_k \phi_k(t). \tag{2.4.52}$$

If the ϕ_k's have been normalized such that

$$\max_{t\in[0,1]} |\phi_k(t)| = 1, \qquad \forall k \geq 0, \tag{2.4.53}$$

the rate of convergence of the above series then depends upon

$$\sum_{k=0}^{\infty} |\alpha_k|. \tag{2.4.54}$$

Small α_k's for large k's means fast convergence. It is noted in [34] that, when the interval is $[-1, 1]$ and k is sufficiently large, the coefficients, α_k, in the Taylor polynomial can be approximated roughly by

$$\alpha_k \approx \frac{1}{k!} f^{(k)}(0); \tag{2.4.55}$$

the coefficients, α_k, in the Legendre polynomial can be approximated by

$$\alpha_k \approx \frac{\sqrt{k\pi}}{2^{k-1}k!} f^{(k)}(0); \tag{2.4.56}$$

the coefficients, α_k, in the Chebyshev polynomial can be approximated by

$$\alpha_k \approx \frac{1}{2^{k-1}k!} f^{(k)}(0). \tag{2.4.57}$$

It is also known that the best convergence rate for any Fourier series is generally in the order of k^{-3}. So, among the above four series, the Taylor has the slowest rate, the Legendre and the Fourier have medium rates, and the Chebyshev has the fastest rate. So, the Chebyshev approximation method generally needs the fewest terms compared to the other three methods.

Approximation by Walsh Series

Another example of orthogonal functions are the Walsh functions on $[0, 1)$. Consider a function defined on the half open unit interval $[0, 1)$ by

$$r_0(t) = \begin{cases} +1 & \text{for } t \in [0, 1/2), \\ -1 & \text{for } t \in [1/2, 1]. \end{cases} \tag{2.4.58}$$

Extend it to the real line by repeating periodically with period 1 and set $r_k(t) \equiv r_0(2^k t)$ for $k = 0, 1, \cdots$. The functions $r_k(t)$ are called the Rademacher functions. The Walsh functions $\{w_k(t)\}$ are obtained by taking all possible products of Rademacher functions. Set $w_0(t) \equiv 1$. To define $w_k(t)$ for $k > 0$, represent the integer k by a dyadic expansion, i.e., find a k' such that

$$k = \sum_{i=0}^{k'} \epsilon_i 2^i \tag{2.4.59}$$

where $\epsilon_{k'} = 1$ and $\epsilon_i = 0$ or 1 for $i = 0, 1, \cdots, k' - 1$. Such a k' obviously satisfies $2^{k'} \leq k \leq 2^{k'+1}$. Set

$$w_k(t) = \prod_{i=0}^{k'} (r_i(t))^{\epsilon_i}. \tag{2.4.60}$$

Each Walsh function $w_k(t)$ takes only the values of 1 or -1, and for $k \geq 1$, is continuous from the right. In particular, the Walsh functions form an orthonormal system in $[0,1)$ with respect to the unit weight functions, because

$$\int_0^1 w_i(t)w_j(t)\,dt \begin{cases} = 0 & \text{if } i \neq j, \\ = 1 & \text{if } i = j. \end{cases} \tag{2.4.61}$$

Correspondingly, function $f(t)$ can be approximated by the following sequence of polynomials, which are called Walsh polynomials,

$$f(t) \approx \sum_{k=0}^{n-1} \alpha_k w_k(t) \triangleq s_n(t) \tag{2.4.62}$$

where the orthogonal coefficients, α_k's, are obtained by applying (2.4.29)

$$\alpha_k = \int_0^1 f(t)w_k(t)\,dt. \tag{2.4.63}$$

Detailed descriptions of the Walsh functions can be found in [25,32,42,109,164]. When $n \to \infty$, the Walsh polynomial becomes a Walsh series. According to Walsh [164], Paley [109], and Fine [32], if $f(t)$ is continuous in $[0,1)$, the series $s_{2^n}(t)$ converges to $f(t)$ uniformly on $[0,1)$. If $f(t)$ is not continuous, there is convergence in mean. The following theorem states the general convergence of $s_n(t)$, instead of $s_{2^n}(t)$ (see p47 of [42]):

Theorem 2.4.9 *If the modulus of continuity of a function f satisfies*

$$\lim_{n\to\infty} \omega\left(\frac{1}{n}\right) \ln n = 0, \tag{2.4.64}$$

then the Walsh series of f converges uniformly to f on $[0,1)$. In particular, if f belongs to the Lipschitz class of order α for some $\alpha \in (0,1]$, the Walsh series for f converges to f uniformly on $[-1,1]$.

Approximation by Block Pulse Functions

Another similar example of orthogonal functions are the block pulse functions on $[0,1)$. Consider functions defined on the half open unit interval $[0,1)$ as

$$\phi_i(t) = \begin{cases} 1 & (i-1)/n \le t < i/n, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4.65}$$

for $i=1,\cdots,n$. The unit interval $[0,1)$ is divided into n equidistant subintervals, and the i-th block pulse function $\phi_i(t)$ has only one rectangular pulse of unit height in the i-th subinterval $[(i-1)/n, i/n)$. It can be easily shown that the block pulse functions form an orthonormal system in $[0,1)$ with respect to the unit weight functions, because

$$\int_0^1 \phi_i(t)\phi_j(t)\,dt \begin{cases} = 0 & \text{if } i \neq j, \\ = 1/n & \text{if } i = j. \end{cases} \tag{2.4.66}$$

Correspondingly, function $f(t)$ can be approximated by the following polynomial, which is called the block pulse polynomial,

$$f(t) \approx \sum_{k=1}^{n} \alpha_k \phi_k(t) \triangleq s_n(t) \tag{2.4.67}$$

where the orthogonal coefficients, α_k's, are obtained by applying (2.4.29)

$$\alpha_k = n \int_0^1 f(t)\phi_k(t)\,dt. \tag{2.4.68}$$

Detailed descriptions of the block pulse functions can be found in [62]. When $n \to \infty$, the block pulse polynomial becomes a block pulse series. It can be shown [62,73] that, for a real piecewise continuous function $f(t)$, $s_n(t)$ converges pointwise to $f(t)$, except possibly at a finite number of discontinuous points.

Approximation by Other Orthogonal Functions

Other typical orthogonal functions include, for example, the Jacobi polynomials, the Chebyshev polynomials of the second kind, the Laguerre polynomials and the Hermite polynomials. Detailed exploration of the analytical properties of these typical

orthogonal functions can be found in [87,154]. These orthogonal functions can in fact be employed in approximating functions.

Notice that, for a function $f(t)$ defined on $[a,b]$, the variable change

$$t = \frac{b+a}{2} + \frac{b-a}{2}t' \tag{2.4.69}$$

transforms $f(t)$ on $[a,b]$ into a new function $g(t')$ on $[-1,1]$

$$g(t') \triangleq f\left(\frac{b+a}{2} + \frac{b-a}{2}t'\right) = f(t). \tag{2.4.70}$$

The variable change

$$t = a + (b-a)t' \tag{2.4.71}$$

transforms $f(t)$ on $[a,b]$ into a new function $g(t')$ on $[0,1]$

$$g(t') \triangleq f\Big(a + (b-a)t'\Big) = f(t). \tag{2.4.72}$$

Therefore, the above results can be applied to functions defined on arbitrary intervals.

2.4.2 Solving Optimal Control Problems by Control Parameterization

Let us consider again the following optimal control problem on the fixed end-time interval $[0,T]$:

$$\min g_0(u) = \omega_0(x(T)) + \int_0^T h_0(x(\tau), u(\tau), \tau)d\tau, \tag{2.4.73}$$

subject to the system constraint

$$\dot{x}(t) = f(x(t), u(t), t), \tag{2.4.74}$$

$$x(0) = x_0, \tag{2.4.75}$$

the control constraints

$$U_i^{min} \leq u_i(t) \leq U_i^{max}, \quad i = 1, \cdots, m, \tag{2.4.76}$$

for any $t \in [0, T]$, and the terminal constraints

$$g_i(u) = \omega_i(x(T)) \leq 0, \qquad i = 1, \cdots, r, \tag{2.4.77}$$

where $x(t) \in \mathcal{R}^n$ is the state vector, $u(t) \in \mathcal{R}^m$ is the control vector, and U^{min}, U^{max} are real constant vectors in $\mathcal{R}^m$.

The most common parameterization method is to approximate the control variables so that the original optimal control problem can be converted into a nonlinear programming problem. In principle, any of the approximation methods introduced before can be employed here. In practice, controls are most often represented by piecewise-polynomials [150,151]. Particularly, they are cubic spline functions [69,70, 71,81,104], and piecewise-constants [40,137,143,155,156,157,168].

To approximate the control variables by piecewise-constants, let's first uniformly divide the time interval $[0, T]$ into N subintervals $[t_{k-1}, t_k]$, with $t_k = kT/N$, and $k = 0, 1, \cdots, N$. During each subinterval, control variables are approximated by constant vectors

$$u(t) \approx \mu^k, \qquad t \in [t_{k-1}, t_k], \tag{2.4.78}$$

for $k = 1, \cdots, N$. By denoting $\mathcal{X}_I$ as an indicator function

$$\mathcal{X}_I(t) = \begin{cases} 1, & \text{when } t \in I \\ 0, & \text{otherwise} \end{cases} \tag{2.4.79}$$

and denoting $I_k = [t_{k-1}, t_k]$, the approximation on control can then be equivalently expressed as

$$u(t) \approx \sum_{k=1}^{N} \mu^k \mathcal{X}_{I_k}(t), \quad t \in [0, T]. \tag{2.4.80}$$

Thus, the control is parameterized by the following vector

$$\xi = [(\mu^1)^\top, \cdots, (\mu^N)^\top]^\top \in \mathcal{R}^\sigma, \tag{2.4.81}$$

where $\sigma = mN$. For the same equally distanced subdivision of $[0, T]$, control variables can be otherwise approximated by $N+3$ B-splines in the following form

$$u(t) \approx \sum_{j=-3}^{N-1} \mu^j B_j(t), \qquad t \in [0, T], \tag{2.4.82}$$

where μ^j's are real constant vectors in $\mathcal{R}^m$. Thus, the control trajectory $u(t) \in \mathcal{R}^m$ during interval $[0,T]$ is parameterized by the following vector

$$\xi = [(\mu^{-3})^\mathsf{T}, \cdots, (\mu^{N-1})^\mathsf{T}]^\mathsf{T} \in \mathcal{R}^\sigma, \tag{2.4.83}$$

where $\sigma = m(N+3)$. Of course, the control variables can also be approximated by the Taylor polynomial, Fourier polynomial, Chebyshev polynomial, Hermite polynomial, Jacobi polynomial, Legendre polynomial, Laguerre polynomial, and Walsh polynomial. The approximations are all in the following form

$$u(t) \approx \sum_{j=1}^{N'} \mu^j \phi_j(t), \qquad t \in [0,T], \tag{2.4.84}$$

where μ^j's are real constant vectors in $\mathcal{R}^m$. Thus, the control trajectory $u(t) \in \mathcal{R}^m$ during interval $[0,T]$ is parameterized by the following vector

$$\xi = [(\mu^1)^\mathsf{T}, \cdots, (\mu^{N'})^\mathsf{T}]^\mathsf{T} \in \mathcal{R}^\sigma, \tag{2.4.85}$$

where $\sigma = mN'$.

After the control is parameterized by any of the above three methods, the original system (2.4.74)-(2.4.75) become

$$\dot{x}(t) = \tilde{f}(x(t), \xi, t), \tag{2.4.86}$$

$$x(0) = x_0, \tag{2.4.87}$$

where $x(t) \in \mathcal{R}^n$ is the state vector, and $\xi \in \mathcal{R}^\sigma$ the parameter vector, and the terminal-state inequality constraints (2.4.77) becomes

$$\tilde{g}_i(\xi) = \tilde{\omega}_i(x(T|\xi)) \leq 0, \qquad i = 1, \cdots, r, \tag{2.4.88}$$

or equivalently, by letting $\tilde{h}_i(x(\tau|\xi), \xi, \tau) \equiv 0$,

$$\tilde{g}_i(\xi) = \tilde{\omega}_i(x(T|\xi)) + \int_0^T \tilde{h}_i(x(\tau|\xi), \xi, \tau) d\tau \leq 0, \qquad i = 1, \cdots, r. \tag{2.4.89}$$

For the piecewise-constants method, the control constraints (2.4.76) becomes

$$U_i^{min} \leq \mu_i^k \leq U_i^{max}, \tag{2.4.90}$$

for $i=1,\cdots,m$, $k=1,\cdots,N$. For the B-splines method, because splines approximation $\sum_{j=-3}^{N-1}\mu^j B_j(t)$ takes the exact value of the approximated function at each node t_k, $k=0,1,\cdots,N$, the control constraints (2.4.76) becomes

$$U_i^{min} \le \sum_{j=-3}^{N-1} \mu_i^j\, B_j(t_k) \le U_i^{max}, \tag{2.4.91}$$

for $i=1,\cdots,m$, $k=0,1,\cdots,N$. For the third type of parameterization method, even though the approximation $\sum_{j=1}^{N'}\mu^j\phi_j(t)$ may not necessarily take the exact value of the approximated function at each node t_k, $k=0,1,\cdots,N$, we still consider that it does. So, the control constraints (2.4.76) becomes

$$U_i^{min} \le \sum_{j=1}^{N'} \mu_i^j\, \phi_j(t_k) \le U_i^{max}, \tag{2.4.92}$$

for $i=1,\cdots,m$, $k=0,1,\cdots,N$. Clearly, among the above three methods, the piecewise-constants approximation results in the simplest expressions of the control constraints.

For each $\xi\in\mathcal{R}^\sigma$, let $x(\cdot|\xi)$ be the corresponding solution of the system (2.4.86)-(2.4.87). So, after the control parameterization, the original problem becomes the following optimal parameter selection problem:

Problem (P_p) Subject to the dynamical system (2.4.86)-(2.4.87), the control constraints (2.4.90) under the piecewise-constants approximation, (2.4.91) under the B-splines approximation, (2.4.92) under the third type of approximation, and the inequality constraints (2.4.89), find a system parameter ξ such that the cost function

$$\tilde{g}_0(\xi) = \tilde{\omega}_0(x(T|\xi)) + \int_0^T \tilde{h}_0(x(\tau|\xi),\xi,\tau)d\tau, \tag{2.4.93}$$

is minimized over $\mathcal{R}^\sigma$.

Clearly, problem (P_p) is just a constrained nonlinear programming problem. In order to solve it, the gradients $\nabla\tilde{g}_i(\xi)$, $i=0,1,\cdots,r$, or, the derivatives $\partial\tilde{g}_i/\partial\xi_j$, $i=0,1,\cdots,r$, $j=1,\cdots,\sigma$, must be supplied. There are two common ways to obtain

those derivatives. One is by the finite difference method. The other is by forming the Hamiltonian functions and the adjoint systems.

Each derivative can be approximated by a forward difference approximation, that is,

$$\frac{\partial \tilde{g}_i}{\partial \xi_j} \approx \frac{\tilde{g}_i(\xi_1 \cdots, \xi_j + \Delta\xi_j, \cdots, \xi_\sigma) - \tilde{g}_i(\xi_1 \cdots, \xi_j, \cdots, \xi_\sigma)}{\Delta\xi_j}, \tag{2.4.94}$$

for $i = 0, 1, \cdots, r$, $j = 1, \cdots, \sigma$. A more accurate approximation could be obtained by using a central difference scheme, but that requires an extra function evaluation: $\tilde{g}_i(\xi_1, \cdots, \xi_j - \Delta\xi_j, \cdots, \xi_\sigma)$. It is clear that the evaluations of all $\tilde{g}_i(\xi)$, $i = 0, \cdots, r$, need only one integration of the system (2.4.74)-(2.4.75), while the evaluations of all $\tilde{g}_i(\xi_1, \cdots, \xi_j + \Delta\xi_j, \cdots, \xi_\sigma)$, $i = 0, \cdots, r$, $j = 1, \cdots, \sigma$ need σ integrations. So, overall $\sigma + 1$ integrations of the system (2.4.74)-(2.4.75) are needed to evaluate all the derivatives. However, all those integrations can be executed simultaneously.

Note that, when the control is parameterized by piecewise-constants, because any change in $u(t)$ during $I_k = [t_{k-1}, t_k]$ will not affect state trajectories before time t_{k-1}, the evaluation of the derivative $\partial \tilde{g}_i / \partial \mu_j^k$, $i = 0, 1, \cdots, r$, $j = 1, \cdots, m$, $k = 1, \cdots, N$, requires the integration of the system equations only from t_{k-1} to t_N, using the value of $x(t_{k-1})$, which corresponds to the nominal value of ξ, as an initial condition [137].

Let us consider another way of obtaining those derivatives. For $i = 0, 1, \cdots, r$, let the corresponding Hamiltonian H_i be defined by

$$H_i(x, \xi, q, t) = \tilde{h}_i(x, \xi, t) + q^\top \tilde{f}(x, \xi, t), \tag{2.4.95}$$

and let $q^i(\cdot|\xi)$ be the solution of the adjoint system

$$-\dot{q}^i(t|\xi) = \frac{\partial H_i}{\partial x}(x(t|\xi), \xi, q^i(t|\xi), t), \tag{2.4.96}$$

$$q^i(T|\xi) = \frac{\partial \tilde{\omega}_i}{\partial x}(x(T|\xi)). \tag{2.4.97}$$

corresponding to an $\xi \in \mathcal{R}^\sigma$. The following theorem shows how to compute the gradients of $\tilde{g}_i(\xi)$'s:

Theorem 2.4.10 *Consider the problem (P_p). The gradient of $\tilde{g}_i(\xi)$ is given as follows:*

$$\frac{\partial \tilde{g}_i(\xi)}{\partial \xi} = \int_0^T \frac{\partial H_i(x(\tau|\xi), \xi, q^i(\tau|\xi), \tau)}{\partial \xi} d\tau \tag{2.4.98}$$

for each $i=0,1,\cdots,r$. Equivalently, the derivative $\partial \tilde{g}_i(\xi)/\partial \xi_j$ is given as follows:

$$\frac{\partial \tilde{g}_i(\xi)}{\partial \xi_j} = \int_0^T \left(\frac{\partial \tilde{h}_i(x(\tau|\xi), \xi, \tau)}{\partial \xi_j} + q^i(\tau|\xi)^\top \frac{\partial \tilde{f}(x(\tau|\xi), \xi, \tau)}{\partial \xi_j} \right) d\tau, \tag{2.4.99}$$

for each $i=0,1,\cdots,r$, $j=1,\cdots,\sigma$.

Especially, when the control is parameterized by piecewise-constants, the expressions for the derivatives can be further simplified. With $\xi=[(\mu^1)^\top,\cdots,(\mu^N)^\top]^\top \in \mathcal{R}^\sigma$, $\sigma=mN$, because

$$\tilde{f}(x(t|\xi), \xi, t) = f(x(t), \sum_{k=1}^N \mu^k \mathcal{X}_{I_k}(t), t), \tag{2.4.100}$$

$$\tilde{h}_0(x(t|\xi), \xi, t) = h_0(x(t), \sum_{k=1}^N \mu^k \mathcal{X}_{I_k}(t), t), \tag{2.4.101}$$

and $\tilde{\omega}_i(\cdot)=\omega_i(\cdot)$, the following corollary is a direct consequence of the above theorem:

Corollary 2.4.1 *Consider the problem (P_p). When the control is parameterized by piecewise-constants, the derivative $\partial \tilde{g}_i(\xi)/\partial \mu^k$ is given as follows:*

$$\frac{\partial \tilde{g}_i(\xi)}{\partial \mu^k} = \int_{t_{k-1}}^{t_k} \left(\frac{\partial h_i(x(\tau|\xi), \mu^k, \tau)}{\partial \mu^k} + q^i(\tau|\xi)^\top \frac{\partial f(x(\tau|\xi), \mu^k, \tau)}{\partial \mu^k} \right) d\tau, \tag{2.4.102}$$

for each $i=0,1,\cdots,r$, $k=1,\cdots,N$.

2.4.3 Solving Optimal Control Problems by Control and State Parameterization

An alternative to the control parameterization is to parameterize both the state and the control variables. Let $\{\phi_k(t)\}$ be a set of some basis functions. For simplicity,

both the state and the control are approximated by partial summations of N terms:

$$x(t) \approx \sum_{k=1}^{N} \mu^k \phi_k(t) \tag{2.4.103}$$

$$u(t) \approx \sum_{k=1}^{N} \beta^k \phi_k(t) \tag{2.4.104}$$

where

$$\mu = [(\mu^1)^\top, \cdots, (\mu^N)^\top]^\top \in \mathcal{R}^{\sigma_x} \tag{2.4.105}$$

$$\beta = [(\beta^1)^\top, \cdots, (\beta^N)^\top]^\top \in \mathcal{R}^{\sigma_u} \tag{2.4.106}$$

with $\sigma_x = nN$, $\sigma_u = mN$. In most cases, the basis functions possess the following integral property

$$\underbrace{\int_a^t \cdots \int_a^t}_{k-times} \phi(t)\,(dt)^k \approx P^k \phi(t) \tag{2.4.107}$$

where P is a square constant matrixs, $\phi^\top(t) = (\phi^1(t), \cdots, \phi^N(t))^\top$. P is called the operational matrix of integration associated with $\phi(t)$. Clearly, the form of P depends on the particular choice of the basis functions.

There are basically two classes of methods utilizing the above property. One class of methods aims at solving the linear quadratic regulator problem. By realizing that the optimal control is a linear feedback of the state, and that the state and the costate are the solution of a linear two-point boundary-value problem, approximating both the state and the control by the forms of (2.4.103) and (2.4.104) will convert the original linear quadratic regulator problem into a pure algebraic problem.

Another class of methods deal with more general problems with nonlinear cost functions, nonlinear dynamical systems, nonlinear path constraints, and nonlinear terminal constraints. By approximating both the state and the control by forms of (2.4.103) and (2.4.104), these methods convert the differential dynamical system into an algebraic equation, so that the original optimal control problem is converted into a constrained nonlinear programming problem in which the algebraic equation serves as an algebraic equality constraint.

Typical examples of the above two classes of methods are, the Taylor series method [52,102,114,115,133,134,153], the Fourier series method [31,112,113,130,131, 132], the Chebyshev series method [23,82,111,162,163], the Legendre series method [22,56,146,165], the Laguerre series method [55], the Hermite series method [65], the Jacobi series method [83], the general orthogonal polynomials method [15], the block pulse functions method [54,57,129], the cubic spline functions method [71], and the Walsh series method [16,17,19,64].

Part II

New Algorithms

Chapter 3

A New Algorithm for Solving Optimal Control Problems with Hard Control Constraints and Terminal Inequality Constraints

3.1 Introduction

Computational techniques for solving optimal control problems, like their close relative — optimization in finite-dimensional space, are iterative in nature. As pointed out by Luenberger [86], referring to the latter case, the theory of iterative algorithms is always dominated by the following three (somewhat overlapping) aspects. The first aspect is the creation of the algorithm itself, which is capable of solving problems as general as possible. The second aspect is the global convergence analysis, which addresses the important question of whether the algorithm, when initiated far from the solution point, will eventually converge to it. The third aspect is the local convergence analysis, which concerns the rate of convergence around a solution. Understandably, computational techniques for solving optimal control problems are also subject to the above three concerns. It is always desirable to devise an algorithm for solving optimal control problems which converges globally as well as fast.

Because a general optimal control problem can be viewed as an optimization of a functional in a general control space subject to system dynamics and some func-

tional equality/inequality constraints, there is then a great deal of similarity in computational techniques between optimization in finite-dimensional space and optimal control problems. For optimal control problems without any constraint, there are gradient methods [11,43,117], the projection method [43], the conjugate gradient methods [74,75,117]. For problems with only nondifferentiable control constraints (for example, the constraints which limits the magnitudes of control variables at any time), there are conditional gradient methods [43,117]. People tend to study problems with only differentiable functional constraints (i.e. the constraints which are differentiable in the function space of control) in a Hilbert control space, such as $L_2^m[t_0, t_f]$. The reason is that any real Hilbert space is identified with its dual by a linear isometry. Under this property, the general Kuhn-Tucker condition can be represented in a much simplified form. This allows the emergence of a family of quasi-Newton (or variable metric) methods [36,53,99,147,159].

However, the most realistic and challenging optimal control problems are the ones with both nondifferentiable control constraints and differentiable functional constraints in the control space $\mathcal{L}_\infty^m[t_0, t_f]$, instead of a Hilbert space $L_2^m[t_0, t_f]$. Polak and Mayne created a number of algorithms to solve the above general problems [95,96,97, 118,119]. Those algorithms are all globally convergent. Some converge to weaker optimality conditions [95,118]. Some converge to stronger conditions [96,97,119]. They are all first-order. Due to the slow convergence locally around the solution it is then desirable to devise an algorithm which not only is globally convergent but also has rate of convergence better than that of the first-order methods.

Despite the emergence of many very successful first-order methods, such as [95, 118], etc, the development of second-order methods has been relatively slow. There are two major reasons for this. One reason is that second-order methods require not only the evaluation of the second-derivatives of the Hamiltonian function at every sampling time, but also their storage. It is quite a burden computationally. Another reason is that the minimization of the summation of the first and second variations is itself not a simple problem. However, because second-order methods generally enjoy rapid,

usually quadratic, convergence around the solution, many attempts have been made.

Early approaches were essentially of the type of neighboring extremal methods (see, for example, [9,67], or the materials in Chapter 6 and Chapter 7 of [11]). Basically, the objectives of those methods were to find the deviations from the nominal control, in the presence of a small disturbance in the initial state and/or in the terminal conditions of the state, so that the revised terminal conditions are met and the second variation is maximized, while the first variation is still kept zero. Consequently, a succession of linear two-point boundary-value problems has to be solved either by finding the transition matrix between the two ends of the boundaries, or by sweeping the two ends by solving a matrix Riccati equation. The major drawbacks of those algorithms are that (1) extremal solutions are often very sensitive to small changes in the unspecified boundary conditions (see Chapter 7 of [11]); (2) the control variables must be unconstrained; (3) $H_{uu}(t)$ is assumed to be nonsingular for all $t \in [t_0, t_f]$; (4) no convergence analysis is provided.

Later, some second-order algorithms began to focus on minimizing the summation of the first and second variations, such as [13,100]. However, all the drawbacks listed above, except (1), still remained.

Motivated by the success of the trust region approach in finite-dimensional optimization, [36] proposed an algorithm to minimize the summation of the first and second variations, where the norm of the control variation is required to be bounded by a positive number, called the trust region. The algorithm is shown to be globally convergent, without assuming the nonsingularity of H_{uu} during intermediate iterations. However, the control variable is still assumed to be unconstrained.

In summary, as far as second-order methods are concerned, there seem to have been relatively few attempts to devise globally convergent algorithms for solving continuous-time optimal control problems, where the control and the terminal state are constrained.

In the late 70's, a promising algorithm, developed by Han [45,46] and Powell

[121], emerged as a general purpose algorithm for solving optimization problems in finite-dimensional space. The method replaces the original problem by a sequence of quadratic programming problems, where the original cost functional is approximated by a quadratic function, with the Hessian being replaced by a positive definite matrix updated by a certain rule, and the inequality constraints approximated by a linear function. One of the most important features of the Han-Powell method is that it is globally convergent. Detailed descriptions of the Han-Powell method can be found, for example, in [86,121].

In this chapter, an algorithm, with an approach similar to the Han-Powell method in finite-dimensional optimization, is devised to solve continuous-time optimal control problems where the control variables and the terminal states are constrained. It is first noticed that the summation of the first and second variations is a second-order approximation to the change of the cost functional due to a change in the control. Further approximation produces a simple convex functional. Consequently, solving the original complicated problem can be replaced by solving iteratively a much simpler "direction-finding" subproblems and a line search along the "direction" found. We then show that the solution of the minimization of the convex functional subject to a linearized system dynamics, linearized terminal inequality constraints, and the original control constraint, generates a descent direction of an exact penalty functional. Global convergence analysis are then given.

Unlike the feasible directions type of algorithms proposed by Mayne and Polak [98], and by Pytlak and Vinter [123], the algorithm in this chapter does not require strict satisfaction of the original feasibility set at any intermediate iterations. However, it will be shown in this chapter that, under certain conditions, the accumulation point of a control sequence will be feasible and satisfy a first-order necessary condition of minimizing the Hamiltonian function. Also, the direction finding subproblem in our algorithm appears to be simpler than the ones in [98] and [123].

A special version of the results in this chapter has appeared in [91], and a complete version would appear in [92].

3.2 Problem Formulation

The dynamical system considered is described by the differential equation, defined on a fixed end-time interval $[t_0, t_f]$,

$$\dot{x}^u(t) = f(x^u(t), u(t), t) \tag{3.2.1}$$

$$x^u(t_0) = x_0 \tag{3.2.2}$$

which is subject to the control constraints

$$u(t) \in \Omega, \qquad \forall\, t \in [t_0, t_f], \tag{3.2.3}$$

where Ω is a compact subset of $\mathcal{R}^m$,

$$\Omega = \{\ \mu \in \mathcal{R}^m \mid U_i^{min} \leq \mu_i \leq U_i^{max},\ i = 1, \cdots, m\ \}, \tag{3.2.4}$$

and the terminal inequality constraints,

$$g_i(u) = h_i(x^u(t_f), t_f) \leq 0, \qquad i = 1, \cdots, r. \tag{3.2.5}$$

In the above, $x^u(t) \in \mathcal{R}^n$ is the state of the system at time $t \in \mathcal{T} = [t_0, t_f]$, which corresponds to the control $u(t) \in \mathcal{R}^m$. Let the set of admissible controls be,

$$\mathcal{U} = \{\ u \mid u : \mathcal{T} \rightarrow \Omega \text{ is continuous a.e. } \} \subset \mathcal{L}_\infty^m[t_0, t_f], \tag{3.2.6}$$

and the set of feasible controls be

$$\mathcal{F} \triangleq \{\ u \mid u \in \mathcal{U};\ g_i(u) \leq 0,\ i = 1, \cdots, r\ \} \subset \mathcal{L}_\infty^m[t_0, t_f]. \tag{3.2.7}$$

Let $\tilde{\mathcal{F}}$ be the set of equivalence classes of functions in $\mathcal{F}$ which are equal almost everywhere on $[t_0, t_f]$. We may now formulate a constrained optimal control problem as follows:

Problem (P). Subject to the dynamical system (3.2.1)-(3.2.2), find a control $u \in \mathcal{U}$ such that the cost functional

$$J(u) = K(x^u(t_f), t_f) + \int_{t_0}^{t_f} L(x^u(\tau), u(\tau), \tau)\, d\tau \tag{3.2.8}$$

is minimized over $\mathcal{F}$.

Throughout, it is understood that the norm of any $w \in \mathcal{R}^p$ for some dimension p is

$$\|w\| = \max_{1 \le i \le p} |w_i|, \tag{3.2.9}$$

the norm of any $w \in \mathcal{L}_\infty^p[t_0, t_f]$ is

$$\|w\| = \operatorname{ess\,sup}_{[t_0,t_f]} \|w(t)\|, \tag{3.2.10}$$

the norm of any $H \in \mathcal{R}^{p \times p}$ is

$$\|H\| = \max_{1 \le i \le p} \sum_{j=1}^{p} |h_{ij}|, \tag{3.2.11}$$

the norm of any $H \in \mathcal{L}_\infty^{p \times p}[t_0, t_f]$ is

$$\|H\| = \operatorname{ess\,sup}_{[t_0,t_f]} \|H(t)\|. \tag{3.2.12}$$

The following conditions are assumed to be satisfied.

Assumption 3.2.1 *$f : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T} \to \mathcal{R}^n$, $h : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T} \to \mathcal{R}^r$, $K : \mathcal{R}^n \times \mathcal{T} \to \mathcal{R}$, $L : \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T} \to \mathcal{R}$. f and L, together with their partial derivatives up to third-order with respect to each of the components of x and u, are continuous for all $(x, u, t) \in \mathcal{R}^n \times \mathcal{R}^m \times \mathcal{T}$. h and K are continuously differentiable with respect to x;*

Assumption 3.2.2 *There exists a positive constant M such that*

$$\|f(x, u, t)\| \le M(1 + \|x\|) \tag{3.2.13}$$

for all $(x, u, t) \in \mathcal{R}^n \times \Omega \times \mathcal{T}$.

Remark: From the theory of differential equations, the system (3.2.1)-(3.2.2) has a unique solution x^u corresponding to each $u \in \mathcal{U}$, and $x^u(t)$ is absolutely continuous (see, e.g. [166]).

3.3 Optimality Conditions

Let the Hamiltonian function of problem (P) be

$$H(x^u, p^u, u, t, p_0) = p_0\, L(x^u, u, t) + (p^u)^\top f(x^u, u, t) \tag{3.3.1}$$

and its minimal function be

$$M(x^u, p^u, t, p_0) = \inf_{u \in \Omega} H(x^u, p^u, u, t, p_0). \tag{3.3.2}$$

Also, let the costate function of problem (P) be

$$\dot{p}^u(t) = -\frac{\partial H}{\partial x}(x^u(t), p^u(t), u(t), t, p_0), \quad \forall [t_0, t_f]. \tag{3.3.3}$$

If problem (P) has no terminal inequality constraint, a necessary condition for optimality is the well-known Pontryagin Maximum Principle [4,120]. A similar maximum principle, which is given below, still holds when problem (P) has terminal constraints in inequality forms [14,58]:

Theorem 3.3.1 *Let $u^* \in \mathcal{F}$ be the optimal solution for the problem (P). Let $x^{u^*}(t)$ be the solution of the system (3.2.1)-(3.2.2) with input $u^*(t)$ during $[t_0, t_f]$. Then, there exists an absolutely continuous costate $p^{u^*}(t) \in \mathcal{R}^n$, $t \in [t_0, t_f]$, which is not identically zero in $[t_0, t_f]$, a nonnegative constant scalar $p_0^* \geq 0$, a constant real vector $\eta^* \in \mathcal{R}^r$, such that, for any $t \in [t_0, t_f]$ (a.e.),*

$$-\dot{p}^{u^*}(t) = \frac{\partial H}{\partial x}(x^{u^*}(t), p^{u^*}(t), u^*(t), t, p_0^*) \tag{3.3.4}$$

$$M(x^{u^*}, p^{u^*}, t, p_0^*) = H(x^{u^*}, p^{u^*}, u^*, t, p_0^*) \tag{3.3.5}$$

$$\frac{dM}{dt}(x^{u^*}(t), p^{u^*}(t), t, p_0^*) = \frac{\partial H}{\partial t}(x^{u^*}(t), p^{u^*}(t), u^*(t), t, p_0^*). \tag{3.3.6}$$

Moreover, the following transversality condition holds,

$$p^{u^*}(t_f) = p_0^* \frac{\partial K}{\partial x}(x^{u^*}(t_f), t_f) + \sum_{i=1}^{r} \eta_i^* \frac{\partial h_i}{\partial x}(x^{u^*}(t_f), t_f) \tag{3.3.7}$$

with

$$\eta_i^* \geq 0 \quad \text{and} \quad \eta_i^*\, h_i(x^{u^*}(t_f), t_f) = 0 \tag{3.3.8}$$

for all $i = 1, \cdots, r$.

Remark If $p_0^* = 0$, then function $L(x^{u^*}, u^*, t)$ will not appear in (3.3.1), nor will $K(x^{u^*}(t_f), t_f)$ in (3.3.7). That means that the above optimality condition of problem (P) is irrelevant to the cost functional $J(u)$, which is obviously "pathological" [4]. Such problems require much more complicated analysis and techniques. We will not treat such problems here. Thus, assume $p_0^* \neq 0$. When $p_0^* \neq 0$ (that is, $p_0^* > 0$), it can be seen easily that the conclusions of the above theorem will not be altered by assuming $p_0^* = 1$. Therefore, throughout this chapter, we only consider the situation when $p_0^* = 1$. For abbreviation, let notation $H(x^u, p^u, u, t)$ be $H(x^u, p^u, u, t, p_0)_{p_0=1}$, and $M(x^u, p^u, t)$ be $M(x^u, p^u, t, p_0)_{p_0=1}$.

3.4 The Algorithm Utilizing Second-Order Information

Let $v(t)$ be the variation of the control, $u^2(t)-u^1(t)$, and let $y^v(t)$ satisfy the following linearized system equations

$$\dot{y}^v(t) = f_x^{(1)}(t)\, y^v(t) + f_u^{(1)}(t)\, v(t) \tag{3.4.1}$$

$$y^v(t_0) = 0 \tag{3.4.2}$$

where $f_x^{(1)}(t)$ and $f_u^{(1)}(t)$ are evaluated at $(x^{u^1}(t),\, u^1(t),\, t)$. Denote

$$\Delta J(u^1)(v) \triangleq \int_{t_0}^{t_f} H_u^{(1)}(t)\, v(t)\, dt \tag{3.4.3}$$

and

$$\begin{aligned}\Delta^2 J(u^1)(v) \triangleq\ & \frac{1}{2}\, y^v(t_f)^\top K_{xx}^{(1)}(t_f)\, y^v(t_f) \\ & + \frac{1}{2}\int_{t_0}^{t_f} (y^v(t)^\top, v(t)^\top) \begin{pmatrix} H_{xx}^{(1)}(t) & H_{xu}^{(1)}(t) \\ H_{ux}^{(1)}(t) & H_{uu}^{(1)}(t) \end{pmatrix} \begin{pmatrix} y^v(t) \\ v(t) \end{pmatrix} dt\end{aligned} \tag{3.4.4}$$

where $H_x^{(1)}(t)$, $H_{xx}^{(1)}(t)$, $H_{xu}^{(1)}(t)$, $H_{ux}^{(1)}(t)$ and $H_{uu}^{(1)}(t)$ are evaluated at $(x^{u^1}(t),\, p^{u^1}(t),\, u^1(t),\, t)$, and $K_{xx}^{(1)}(t_f)$ is evaluated at $(x^{u^1}(t_f),\, t_f)$. In the above, $x^{u^1}(t)$ and $p^{u^1}(t)$ are the state and costate of problem (P) corresponding to control $u^1(t)$. Note that, when there is no terminal inequality constraint (3.2.5), the above $\Delta J(u^1)(v)$ is just the traditional first variation, and $\Delta J(u^1)(v)+\Delta^2 J(u^1)(v)$ the traditional second variation.

If we let $p^{u^1}(t)$ satisfy the following terminal condition,

$$p^{u^1}(t_f) = \frac{\partial K}{\partial x}(x^{u^1}(t_f), t_f), \tag{3.4.5}$$

the following proposition then shows that for any u^1, $u^2 \in \mathcal{U}$, $\Delta J(u^1)(v)$ is a first-order estimate for $J(u^2) - J(u^1)$, and $\Delta J(u^1)(v) + \Delta^2 J(u^1)(v)$ is a second-order estimate:

Proposition 3.4.1 *There exist $c_1, c_2 \in (0, \infty)$ such that, for all u^1, $u^2 \in \mathcal{U}$,*

$$\left| (J(u^2) - J(u^1)) \; - \; \Delta J(u^1)(v) \right| \; \leq \; c_1 \| \, v \, \|^2 \tag{3.4.6}$$

$$\left| (J(u^2) - J(u^1)) \; - \; (\Delta J(u^1)(v) + \Delta^2 J(u^1)(v)) \right| \; \leq \; c_2 \| \, v \, \|^3 \tag{3.4.7}$$

where $v = u^2 - u^1$.

Proof: see Lemma A.8 in Appendix. □

Typically, for most computational techniques seeking the optimal control, an initial control u^0 is selected and a sequence of new controls $u^1, u^2, \cdots, u^k, \cdots$, is generated, each improving upon its predecessor. In viewing the above proposition, a natural and convenient way to find an improving control u^{k+1} at the $k-th$ iteration is to minimize $\Delta J(u^k)(u^{k+1} - u^k)$ or $\Delta J(u^k)(u^{k+1} - u^k) + \Delta^2 J(u^k)(u^{k+1} - u^k)$. The former technique is often called a first-order method, and the latter a second-order method. In what follows, only the second-order method is studied. That is, to solve the original problem (P), the following "direction-finding" subproblem (P_k') is solved repeatedly,

$$(P_k') \quad \min_{u \in \mathcal{U}} \; \Delta J(u^k)(u - u^k) + \Delta^2 J(u^k)(u - u^k)$$

subject to the linearized system,

$$\dot{y}^{v^k}(t) \;\; = \;\; f_x^{(k)}(t)\, y^k(t) + f_u^{(k)}(t)\, v^k \tag{3.4.8}$$

$$y^{v^k}(t_0) \;\; = \;\; 0 \tag{3.4.9}$$

and the linearized terminal inequality constraints,

$$h_i(x^{u^k}(t_f), t_f) + \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f) y^{v^k}(t_f) \leq 0 \qquad i = 1, \cdots, r \tag{3.4.10}$$

with $v^k = u - u^k$.

Algorithm 1.

Step 0. Select a $u_0 \in \mathcal{U}$. Set $k=0$.

Step 1. Compute state x^{u^k} by forward integrating (3.2.1)-(3.2.2).

Step 2. Compute costate p^{u^k} by backward integrating (3.3.3) from terminal condition (3.4.5).

Step 3. Solve the "direction-finding" subproblem (P_k').

Step 4. If its solution is such that $\bar{u}^k = u^k$, stop. Otherwise, go to Step 5.

Step 5. Compute a suitable stepsize $\lambda^k > 0$, according to some stepsize rule.

Step 6. Set $u^{k+1} = u^k + \lambda^k(\bar{u}^k - u^k)$. Set $k=k+1$ and return to Step 2.

However, the above "direction-finding" subproblem (P_k')'s are still difficult to solve. Further simplification is needed, while still preserving the approximation order of two. Because $K_{xx}(t_f)$ is real and symmetric, there exists a constant matrix Λ_1 which is semipositive definite, such that

$$y^v(t_f)^\top K_{xx}(t_f)\, y^v(t_f) \le y^v(t_f)^\top \Lambda_1 y^v(t_f). \tag{3.4.11}$$

Also, because $\begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix}$ is real and symmetric, there exist two time-varying matrices $\Lambda_2(t)$ and $\Lambda_3(t)$ which are positive definite, such that, for any $t \in [t_0, t_f]$,

$$\begin{pmatrix} y^v(t) \\ v(t) \end{pmatrix}^T \begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix} \begin{pmatrix} y^v(t) \\ v(t) \end{pmatrix} \le y^v(t)^\top \Lambda_2(t) y^v(t) + v(t)^\top \Lambda_3(t) v(t). \tag{3.4.12}$$

Let $v = u^2 - u^1$, denote

$$J_2(v) \triangleq K_2^{(1)}(y^v(t_f), t_f) + \int_{t_0}^{t_f} L_2^{(1)}(y^v(\tau), v(\tau), \tau)\, d\tau \tag{3.4.13}$$

with

$$K_2^{(1)}(y^v(t_f), t_f) \;=\; \frac{1}{2} y^v(t_f)^\top \Lambda_1^{(1)} y^v(t_f) \tag{3.4.14}$$

$$L_2^{(1)}(y^v(t),v(t),t) = H_u^{(1)}(t)\,v(t) + \frac{1}{2}y^v(t)^\top \Lambda_2^{(1)}(t)y^v(t) + \frac{1}{2}v(t)^\top \Lambda_3^{(1)}(t)v(t). \tag{3.4.15}$$

Proposition 3.4.2 *There exists a $c \in (0,\infty)$, such that,*

$$J(u^2) - J(u^1) \leq J_2(u^2-u^1) + c\,\|u^2-u^1\|^3 \tag{3.4.16}$$

for all u^1, $u^2 \in \mathcal{U}$.

Proof: The proof is done by applying (3.4.11), (3.4.12) and (3.4.13) to Proposition 3.4.2. □

The above shows that $J_2(u^2-u^1)$ is a "one-way" second order approximation of $J(u^2)-J(u^1)$, when u^2 is close to u^1. Together with the fact that its terminal term and integrand terms are all convex functions, $J_2(u^2-u^1)$ is therefore an easier approximation of $J(u^2)-J(u^1)$ to compute with than $\Delta J(u^k)(u^{k+1}-u^k)+\Delta^2 J(u^k)(u^{k+1}-u^k)$. It reminds us of the similar case in finite-dimensional optimization handled by the quasi-Newton method or the Han-Powell method where the Hessian, which is not necessary positive definite, is iteratively replaced by an updated positive definite matrix to facilitate both the computation and the convergence analysis. Hence , instead of solving the original problem (P), the following new "direction-finding" subproblem (P_k'') is solved repeatedly,

$$(P_k'') \quad \min_{u\in\mathcal{U}} J_2^{(k)}(u-u^k)$$

subject to the linearized system,

$$\dot{y}^{v^k}(t) = f_x^{(k)}(t)\,y^{v^k}(t) + f_u^{(k)}(t)\,v^k \tag{3.4.17}$$

$$y^{v^k}(t_0) = 0 \tag{3.4.18}$$

and the linearized terminal inequality constraints,

$$h_i(x^{u^k}(t_f),t_f) + \frac{\partial h_i}{\partial x}(x^{u^k}(t_f),t_f)y^{v^k}(t_f) \leq 0 \qquad i=1,\cdots,r \tag{3.4.19}$$

where $v^k = u-u^k$.

Algorithm 2.

Step 0. Select a $u_0 \in \mathcal{U}$. Set $k=0$.

Step 1. Compute state x^{u^k} by forward integrating (3.2.1)-(3.2.2).

Step 2. Compute costate p^{u^k} by backward integrating (3.3.3) from terminal condition (3.4.5).

Step 3. Solve the "direction-finding" subproblem (P_k'').

Step 4. If its solution is such that $\bar{u}^k = u^k$, stop. Otherwise, go to Step 5.

Step 5. Compute a suitable stepsize $\lambda^k > 0$, according to some stepsize rule.

Step 6. Set $u^{k+1} = u^k + \lambda^k(\bar{u}^k - u^k)$. Set $k=k+1$ and return to Step 2.

Remark Clearly, set Ω, defined in (3.2.3), is convex. Because $u^k(t), \bar{u}^k(t) \in \Omega$, and

$$\begin{aligned} u^{k+1}(t) &= u^k(t) + \lambda^k(\bar{u}^k(t) - u^k(t)) \\ &= (1-\lambda^k)u^k(t) + \lambda^k \bar{u}^k(t), \end{aligned}$$

a convex combination of $u^k(t)$ and $\bar{u}^k(t)$, we then have $u^{k+1}(t) \in \Omega$ for any $t \in [t_0, t_f]$ and for any $\lambda \in [0,1]$. That is, the algorithm defined above will always generate a sequence of admissible controls, $u^k \in \mathcal{U}$, $k=1,2,\cdots$, as long as the initial control u^0 is admissible. It is also important to note that the control sequence $\{u^k\}_{k=0}^{\infty}$, generated by Algorithm 2, may not always belong to the original feasible set $\mathcal{F}$, defined in (3.2.7). However, as will be seen later on, $\{u^k\}_{k=0}^{\infty}$ would ultimately stop at, or converge to, a feasible control which satisfies some optimality conditions.

An advantage of the above algorithm is that the original constrained nonlinear problem is solved by solving a sequence of constrained linear quadratic optimal control problems, which are much simpler than the original one. Besides, at iteration k, the existence and uniqueness of the solution of the constrained linear quadratic problem is always guaranteed, as long as the feasible set $\mathcal{F}_k$ is not empty (see Chapter 5 for details), which is not the case for problem (P_k'). Also, as will be seen later in this

chapter, the algorithm generates a descent direction of an exact penalty functional, and, under some conditions, the accumulation point of a control sequence satisfies the first-order necessary condition of minimizing the Hamiltonian function. So, in all respects, the above algorithm can be regarded as an analog to the quasi-Newton method or the Han-Powell method in finite-dimensional optimization.

Another advantage of the above algorithm is that, for each subproblem at the k-th iteration, its hamiltonian function, which is quadratic in both state $x(t)$ and control $u(t)$ for any $t \in [t_0, t_f]$, is a strictly convex function in control $u(t)$. So, the optimal control of every subproblem can never be singular.

However, unlike the finite-dimensional case where the "direction-finding" subproblem is a quadratic programming problem which can be solved in a finite number of steps, the exact solution of the above "direction-finding" subproblee (P_k'') at each iteration may require an infinite number of steps. Practically, the iterations can be stopped after some finite number of steps after the optimal solution is approached within a certain accuracy range.

It should also be pointed out that after $K_{xx}(t_f)$ is replaced by its approximation Λ_1, and $\begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix}$ is replaced by the block diagonal matrix $\begin{pmatrix} \Lambda_2(t) & 0 \\ 0 & \Lambda_3((t) \end{pmatrix}$, the second-order local convergence rate may not hold any longer. However, intuition tells us that the new local convergence rate should be at least first-order, and possibly superlinear, depending upon the tightness of the approximations.

3.5 Descent Properties

Let $\bar{u}^k$ be the optimal solution of the "direction-finding" subproblem (P_k''). Let $\bar{v}^k = \bar{u}^k - u^k$, and $y^{\bar{v}^k}$ be the solution of the linearized system (3.4.17)-(3.4.18) with input $\bar{v}^k$. Then, according to Theorem 3.3.1, there exists a constant transversality vector $\eta^k \in \mathcal{R}^r$ satisfying

$$\eta_i^k \geq 0 \quad \text{and} \quad \eta_i^k \left(h_i(x^{u^k}(t_f), t_f) + \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f) y^{\bar{v}^k}(t_f) \right) = 0 \qquad (3.5.1)$$

for all $i = 1, \cdots, r$, and a costate function $q^{\bar{v}^k}(t)$ of the "direction-finding" subproblem (P_k'') satisfying

$$\begin{aligned} -\dot{q}^{\bar{v}^k}(t) &= \frac{\partial H_2}{\partial y}(y^{\bar{v}^k}, q^{\bar{v}^k}, \bar{v}^k, t) \\ q^{\bar{v}^k}(t_f) &= \frac{\partial K_2}{\partial y}(y^{\bar{v}^k}(t_f), t_f) \\ &\quad + \sum_{i=1}^{r} \eta_i^k \frac{\partial}{\partial y}\Big(h_i(x^{u^k}(t_f), t_f) + \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f) y^{\bar{v}^k}(t_f))\Big), \end{aligned}$$

such that, $\bar{u}^k(t)$ minimizes $H_2(y^{\bar{v}^k}(t),\ q^{\bar{v}^k}(t),\ u - u^k(t), t)$ with respect to u at any $t \in [t_0, t_f]$. In the above, H_2 is the Hamiltonian function of the "direction-finding" subproblem (P_k''). So,

$$\begin{aligned} & H_2(y^{\bar{v}^k}(t), q^{\bar{v}^k}(t), \bar{v}^k(t), t) \\ &\triangleq\ L_2(y^{\bar{v}^k}(t), \bar{v}^k(t), t) + q^{\bar{v}^k}(t)^\top (f_x^{(k)}(t)\, y^{\bar{v}^k}(t) + f_u^{(k)}(t)\, \bar{v}^k(t)) \\ &=\ H_u(x^{u^k}(t), u^k(t), p^{u^k}(t), t)\, \bar{v}^k(t) + \frac{1}{2} y^{\bar{v}^k}(t)^\top \Lambda_2^{(k)}(t) y^{\bar{v}^k}(t) \\ &\quad + \frac{1}{2} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t) + q^{\bar{v}^k}(t)^\top (f_x^{(k)}(t)\, y^{\bar{v}^k}(t) + f_u^{(k)}(t)\, \bar{v}^k(t)) \end{aligned} \tag{3.5.2}$$

and

$$-\dot{q}^{\bar{v}^k}(t) = f_x^{(k)}(t)^\top q^{\bar{v}^k}(t) + \Lambda_2^{(k)}(t) y^{\bar{v}^k}(t) \tag{3.5.3}$$

$$q^{\bar{v}^k}(t_f) = \Lambda_1^{(k)} y^{\bar{v}^k}(t_f) + \sum_{i=1}^{r} \eta_i^k \Big(\frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)\Big)^\top \tag{3.5.4}$$

and $\bar{u}^k(t)$ is the solution of

$$\begin{aligned} \min_{\mu \in \Omega} \Big\{ & H_u(x^{u^k}(t), u^k(t), p^{u^k}(t), t)(\mu - u^k(t)) \\ & + \frac{1}{2}(\mu - u^k(t))^\top \Lambda_3^{(k)}(t)(\mu - u^k(t)) \\ & + \frac{1}{2} y^{\bar{v}^k}(t)^\top \Lambda_2^{(k)}(t) y^{\bar{v}^k}(t) + q^{\bar{v}^k}(t)^\top (f_x^{(k)}(t)\, y^{\bar{v}^k}(t) + f_u^{(k)}(t)(\mu - u^k(t)))\Big\}, \end{aligned}$$

for any $t \in [t_0, t_f]$. Equivalently, $\bar{u}^k(t)$ is the solution of the following quadratic programming problem:

$$\min_{U^{min} \le \mu \le U^{max}} \Big\{ \frac{1}{2} \mu^\top \Lambda_3^{(k)}(t) \mu + b^{(k)}(t)^\top \mu \Big\}$$

where

$$b^{(k)}(t) = H_u(x^{u^k}(t), u^k(t), p^{u^k}(t), t)^\top - \Lambda_3^{(k)}(t)u^k(t) + f_u^{(k)}(t)^\top q^{\bar{v}^k}(t),$$

for any $t \in [t_0, t_f]$. By the Kuhn-Tucker Theorem, there exist nonnegative vector functions $\beta^k(t)$ and $\gamma^k(t)$ such that,

$$\begin{aligned} & H_u(x^{u^k}(t), u^k(t), p^{u^k}(t), t)^\top \\ & \quad + \Lambda_3^{(k)}(t)(\bar{u}^k(t) - u^k(t)) + f_u^{(k)}(t)^\top q^{\bar{v}^k}(t) + \beta^k(t) - \gamma^k(t) = 0 \end{aligned} \tag{3.5.5}$$

where

$$\beta_i^k(t)^\top(\bar{u}_i^k(t) - U_i^{max}) = 0 \qquad i = 1, \cdots, m \tag{3.5.6}$$

$$\gamma_i^k(t)^\top(U_i^{min} - \bar{u}_i^k(t)) = 0 \qquad i = 1, \cdots, m \tag{3.5.7}$$

for any $t \in [t_0, t_f]$.

Next, we show that $\bar{v}^k$, the solution of the "direction-finding" subproblem (P_k''), turns out to be a descent direction of the exact penalty functional $\theta_\rho : \mathcal{U} \to \mathcal{R}$,

$$\theta_\rho(u) \triangleq J(u) + \rho \sum_{i=1}^{r} \varphi_i(u) \tag{3.5.8}$$

where

$$\varphi_i(u) \triangleq \max\{0, g_i(u)\} = \max\{0, h_i(x^u(t_f), t_f)\} \tag{3.5.9}$$

and ρ is a positive number.

Theorem 3.5.1 *Let $\bar{u}^k$ be the solution of the "direction-finding" subproblem (P_k''), and $\bar{v}^k(t) = \bar{u}^k(t) - u^k(t)$. If η^k, a constant vector satisfying optimality condition (3.5.1), satisfies*

$$\|\eta^k\| \leq \rho \tag{3.5.10}$$

and $\bar{v}^k \neq 0$, then there exists a $\bar{\lambda}^k \in (0, 1]$, such that, for all $0 < \lambda \leq \bar{\lambda}^k$,

$$\theta_\rho(u^k + \lambda\bar{v}^k) - \theta_\rho(u^k) \leq -\frac{\lambda}{2} \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\, dt < 0. \tag{3.5.11}$$

Proof: The proof here is similar in spirit to the proof of descent property by Han [46] in the finite-dimensional space. Let

$$I^{(k)} \triangleq \{\ i:\ g_i(u^k) > 0\ \} = \{\ i:\ h_i(x^{u^k}(t_f), t_f) > 0\ \},$$

$$\bar{I}^{(k)} \triangleq \{\ i:\ g_i(u^k) = 0\ \} = \{\ i:\ h_i(x^{u^k}(t_f), t_f) = 0\ \},$$

$$\hat{I}^{(k)} \triangleq \{\ i:\ g_i(u^k) < 0\ \} = \{\ i:\ h_i(x^{u^k}(t_f), t_f) < 0\ \}.$$

From Lemma A.1 in the appendix, the mean-value property,

$$\begin{aligned} & g_i(u^k + \lambda\bar{v}^k) - g_i(u^k) \\ &= h_i(x^{u^k+\lambda\bar{v}^k}(t_f), t_f) - h_i(x^{u^k}(t_f), t_f) \\ &= \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k}(t_f)) \\ &\quad + \int_0^1 (1-\tau)\frac{\partial^2 h_i}{\partial x^2}(\bar{x}(\tau, t_f), t_f)(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k}(t_f))^2\ d\tau \end{aligned}$$

where $\bar{x}(\tau, t_f) = x^{u^k}(t_f) + \tau(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k}(t_f))$. According to Lemma A.2 in the appendix, both $x^{u^k}(t_f)$ and $x^{u^k+\lambda\bar{v}^k}(t_f)$ are bounded, because $u^k, u^k + \lambda\bar{v}^k \in \mathcal{U}$, for any $\lambda \in [0,1]$. So, $\bar{x}(\tau, t_f)$ is also bounded, for any $\tau \in [0,1]$ and any $\lambda \in [0,1]$. The continuity of $\frac{\partial h_i}{\partial x}$, $\frac{\partial^2 h_i}{\partial x^2}$, and the boundednesses of $x^{u^k}(t_f)$ and $\bar{x}(\tau, t_f)$, imply that there exist $c', c'' > 0$, such that,

$$\begin{aligned} & g_i(u^k + \lambda\bar{v}^k) - g_i(u^k) \\ &= \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k}(t_f)) + c'\|(u^k + \lambda\bar{v}^k) - u^k\|^2 \\ &\leq \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)y^{\lambda\bar{v}^k}(t_f) + c''\|x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k}(t_f) - y^{\lambda\bar{v}^k}(t_f)\| + c'\|\lambda\bar{v}^k\|^2 \\ &\leq \frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)y^{\lambda\bar{v}^k}(t_f) + c\lambda^2\ \|\bar{v}^k\|^2. \end{aligned}$$

The last inequality comes from Lemma A.3 in the appendix, with some $c > 0$, because $y^{\lambda\bar{v}^k}$ is a solution of the linearized system (3.4.17)-(3.4.18) with input $\lambda\bar{v}^k$, $y^{\lambda\bar{v}^k} = \lambda y^{\bar{v}^k}$. Moreover, from the linearized terminal-state inequality constraints (3.4.10), we have

$$\frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)y^{\bar{v}^k}(t_f) \leq -h_i(x^{u^k}(t_f), t_f) = -g_i(u^k).$$

Therefore, for any $\lambda \in [0,1]$,

$$g_i(u^k + \lambda \bar{v}^k) - g_i(u^k) \leq -\lambda g_i(u^k) + c\lambda^2 \|\bar{v}^k\|^2. \tag{3.5.12}$$

(1) For any $i \in I^{(k)}$. Because $g_i(u)$ is continuous in u, by Lemma A.4 in the appendix, there exists a $\bar{\lambda}_1^k \in (0,1]$, such that, for all $\lambda \in [0, \bar{\lambda}_1^k]$, $g_i(u^k + \lambda \bar{v}^k) > 0$. Then, from (3.5.12),

$$\varphi_i(u^k + \lambda \bar{v}^k) - \varphi_i(u^k) = g_i(u^k + \lambda \bar{v}^k) - g_i(u^k) \leq -\lambda g_i(u^k) + c\lambda^2 \|\bar{v}^k\|^2.$$

(2) For any $i \in \bar{I}^{(k)}$. From (3.5.12), for any $\lambda \in [0,1]$,

$$g_i(u^k + \lambda \bar{v}^k) - g_i(u^k) \leq -\lambda g_i(u^k) + c\lambda^2 \|\bar{v}^k\|^2 = c\lambda^2 \|\bar{v}^k\|^2,$$

we have,

$$\varphi_i(u^k + \lambda \bar{v}^k) - \varphi_i(u^k) \leq c\lambda^2 \|\bar{v}^k\|^2.$$

(3) For any $i \in \hat{I}^{(k)}$. Because $g_i(u)$ is continuous in u, by Lemma A.4 in the appendix, there exists a $\bar{\lambda}_2^k \in (0,1]$, such that, for all $\lambda \in [0, \bar{\lambda}_2^k]$, $g_i(u^k + \lambda \bar{v}^k) < 0$. Then,

$$\varphi_i(u^k + \lambda \bar{v}^k) - \varphi_i(u^k) = 0 - 0 = 0.$$

Therefore, there exist $\bar{\lambda}_1^k, \lambda_2^k \in [0,1)$, $\bar{c}' > 0$, such that, for all $\lambda \in [0, \min\{\bar{\lambda}_1^k, \bar{\lambda}_2^k\}]$,

$$\begin{aligned}
&\theta_\rho(u^k + \lambda \bar{v}^k) - \theta_\rho(u^k) \\
&= J(u^k + \lambda \bar{v}^k) - J(u^k) + \rho \sum_{i=1}^{r} (\varphi_i(u^k + \lambda \bar{v}^k) - \varphi_i(u^k)) \\
&\leq J(u^k + \lambda \bar{v}^k) - J(u^k) - \rho \sum_{i \in I^{(k)}} \lambda g_i(u^k) + \bar{c}' \rho \lambda^2 \|\bar{v}^k\|^2.
\end{aligned} \tag{3.5.13}$$

From Lemma A.5 and (3.5.5),

$$\begin{aligned}
&J(u^k + \lambda \bar{v}^k) - J(u^k) \\
&\quad \leq \lambda \int_{t_0}^{t_f} H_u(x^k(t), u^k(t), p^k(t), t) \bar{v}^k(t)\, dt + \bar{c}'' \lambda^2 \|\bar{v}^k\|^2 \\
&\quad \leq -\lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt - \lambda \int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t) \bar{v}^k(t)\, dt \\
&\qquad -\lambda \int_{t_0}^{t_f} \beta^k(t)^\top \bar{v}^k(t)\, dt + \lambda \int_{t_0}^{t_f} \gamma^k(t)^\top \bar{v}^k(t)\, dt + \bar{c}'' \lambda^2 \|\bar{v}^k\|^2.
\end{aligned}$$

Because

$$\begin{aligned}\beta^k(t)^\top \bar{v}^k(t) &= \beta^k(t)^\top(\bar{u}^k(t) - u^k(t)) \\ &= \underbrace{\beta^k(t)^\top(\bar{u}^k(t) - U^{max})}_{=0} + \underbrace{\beta^k(t)}_{\geq 0}^\top \underbrace{(U^{max} - u^k(t))}_{\geq 0} \geq 0 \quad (3.5.14)\end{aligned}$$

and

$$\begin{aligned}\gamma^k(t)^\top \bar{v}^k(t) &= \gamma^k(t)^\top(\bar{u}^k(t) - u^k(t)) \\ &= \underbrace{\gamma^k(t)^\top(\bar{u}^k(t) - U^{min})}_{=0} + \underbrace{\gamma^k(t)}_{\geq 0}^\top \underbrace{(U^{min} - u^k(t))}_{\leq 0} \leq 0, \quad (3.5.15)\end{aligned}$$

we then have,

$$\begin{aligned}J(u^k + \lambda \bar{v}^k) - J(u^k) \leq & -\lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt \\ & -\lambda \int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t) \bar{v}^k(t)\, dt + \bar{c}'' \lambda^2 \|\bar{v}^k\|^2. \quad (3.5.16)\end{aligned}$$

Because $y^{\bar{v}^k}$, $\bar{u}^k$ is an optimal pair for the "direction-finding" subproblem (P_k''), $y^{\bar{v}^k}(t)$ satisfies (3.4.17) and (3.4.18) with input $\bar{v}^k(t) = \bar{u}^k(t) - u^k(t)$, and the corresponding costate function $q^{\bar{v}^k}$ satisfies (3.5.3) and (3.5.4), with η^k a constant vector satisfying (3.5.1). We then have,

$$\begin{aligned}\frac{d}{dt}\left(q^{\bar{v}^k}(t)^\top y^{\bar{v}^k}(t)\right) &= q^{\bar{v}^k}(t)^\top \underbrace{\dot{y}^{\bar{v}^k}(t)}_{(3.4.17)} + y^{\bar{v}^k}(t)^\top \underbrace{\dot{q}^{\bar{v}^k}(t)}_{(3.5.3))} \\ &= q^{\bar{v}^k}(t)^\top f_u^{(k)}(t) \bar{v}^k(t) - y^{\bar{v}^k}(t)^\top \Lambda_2^{(k)}(t) y^{\bar{v}^k}(t).\end{aligned}$$

So,

$$\begin{aligned}&\int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t) \bar{v}^k(t)\, dt \\ &= \underbrace{q^{\bar{v}^k}(t_f)^\top}_{(3.5.4)} y^{\bar{v}^k}(t_f) - \underbrace{q^{\bar{v}^k}(t_0)^\top y^{\bar{v}^k}(t_0)}_{=0} + \int_{t_0}^{t_f} \underbrace{y^{\bar{v}^k}(t)^\top \Lambda_2^{(k)}(t) y^{\bar{v}^k}(t)}_{\geq 0}\, dt \\ &\geq \underbrace{y^{\bar{v}^k}(t_f)^\top \Lambda_1^{(k)} y^{\bar{v}^k}(t_f)}_{\geq 0} + \sum_{i=1}^{r} \eta_i^k \underbrace{\frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f) y^{\bar{v}^k}(t_f)}_{(3.5.1)} \\ &\geq -\sum_{i \in I^{(k)}} \eta_i^k g_i(u^k) - \sum_{i \in \bar{I}^{(k)}} \eta_i^k \underbrace{g_i(u^k)}_{=0} - \sum_{i \in \hat{I}^{(k)}} \eta_i^k \underbrace{g_i(u^k)}_{<0} \\ &\geq -\sum_{i \in I^{(k)}} \eta_i^k g_i(u^k). \quad (3.5.17)\end{aligned}$$

Combining (3.5.13), (3.5.16) and (3.5.17), we finally have that, there exist $\bar{\lambda}_3^k \in (0,1]$, and $\bar{c}', \bar{c}'' > 0$, such that, for all $\lambda \in [0, \bar{\lambda}_3^k]$,

$$\begin{aligned}
&\theta_\rho(u^k + \lambda \bar{v}^k) - \theta_\rho(u^k) \\
&\leq -\lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt + \lambda \sum_{i \in I^{(k)}} \underbrace{(\eta_i^k - \rho)}_{\leq 0} \underbrace{g_i(u^k)}_{>0} + (\bar{c}'\rho + \bar{c}'')\lambda^2 \|\bar{v}^k\|^2 \\
&\leq -\lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt + (\bar{c}'\rho + \bar{c}'')\lambda^2 \|\bar{v}^k\|^2.
\end{aligned}$$

Hence, there exists a small enough $\bar{\lambda}^k \in (0,1]$, such that, for all $0 < \lambda \leq \bar{\lambda}^k$ and $\bar{v}^k \neq 0$,

$$\theta_\rho(u^k + \lambda \bar{v}^k) - \theta_\rho(u^k) \leq -\frac{\lambda}{2} \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt < 0. \tag{3.5.18}$$

□

Corollary 3.5.1 *Assume that there is no terminal constraint. Let $\bar{u}^k$ be the solution of the "direction-finding" subproblem (P_k''), and $\bar{v}^k(t) = \bar{u}^k(t) - u^k(t)$. If $\bar{v}^k \neq 0$, then there exists a $\bar{\lambda}^k \in (0,1]$, such that, for all $0 < \lambda \leq \bar{\lambda}^k$,*

$$J(u^k + \lambda \bar{v}^k) - J(u^k) \leq -\frac{\lambda}{2} \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt < 0. \tag{3.5.19}$$

Proof: When there is no terminal constraint, $\varphi_i(u)$ is always zero, and the constant vector η^k in Theorem 3.5.1 can be considered as a zero vector. So, $\theta_\rho(u)$ becomes $J(u)$, and condition (3.5.10) is automatically satisfied. The proof is then done by applying Theorem 3.5.1. □

Remark Theorem 3.5.1 shows that, when $\bar{u}^k \neq u^k$, $\bar{v}^k(t) = \bar{u}^k(t) - u^k(t)$ is a descent direction of the exact penalty functional $\theta_\rho(u)$ at the k-th iteration. Corollary 3.5.1 shows that, when there is no terminal constraint and $\bar{u}^k \neq u^k$, $\bar{v}^k = \bar{u}^k - u^k$ is a descent direction of the the cost functional $J(u)$ at the k-th iteration.

Remark The descent properties shown in both Theorem 3.5.1 and Corollary 3.5.1 will always hold, as long as matrix Λ_1 is semi-positive definite, $\Lambda_2(t)$ is semi-positive definite for all $t \in [t_0, t_f]$, and $\Lambda_3(t)$ is positive definite for all $t \in [t_0, t_f]$, regardless

of whether the inequality relationships (3.4.11)-(3.4.12) are satisfied. However, those inequalities are crucial for the rate of convergence of the algorithm. Intuitively, tighter approximations by (3.4.11)-(3.4.12) make the rate of convergence of the algorithm closer to second-order; while looser approximations would destroy the second-order properties and make the algorithm behave more like a first-order algorithm.

3.6 Stepsize Rules

Ideally, the best stepsize at each iteration is the one which minimizes the exact penalty function $\theta_\rho(u^k+\lambda(\bar{u}^k-u^k))$. That is, at the k-th iteration,

$$\lambda^k = \arg\min_{\lambda\geq 0} \theta_\rho(u^k+\lambda(\bar{u}^k-u^k)). \tag{3.6.1}$$

From Theorem 3.5.1, whenever $\bar{u}^k \neq u^k$, the stepsize defined above is a positive number. However, the calculation of the above exact stepsize is very expensive. In practice, inaccurate line search has to be used.

Let us recall that for optimization in finite-dimensional space, a practical and popular line search method is Armijo's rule. Let f be a differentiable function: $\mathcal{R}^n \to \mathcal{R}$. Consider now a search of the smallest integer $l_k \geq 0$, made over the semi-infinite line emanating from x^k in the direction d^k, such that

$$f(x^k + \zeta^{l_k} d^k) - f(x^k) \leq \alpha\zeta^{l_k}\langle \nabla f(x^k), d^k\rangle,$$

where α, ζ are two parameters chosen a priori in $(0,1)$. ζ^{l_k} is then called an Armijo stepsize.

In our optimal control problem, we would like to perform a line search on the exact penalty function $\theta_\rho(u)$, starting from $u = u^k$ in the direction $\bar{v}^k$. However, because $\theta_\rho(u)$ is nondifferentiable with respect to u, the above Armijo's rule cannot be applied directly. Instead, the following Armijo-like line search is adopted: find the smallest integer $l_k \geq 0$ such that

$$\theta_\rho(u^k + \zeta^{l_k}\bar{v}^k) - \theta_\rho(u^k) \leq -\alpha\zeta^{l_k}\int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt. \tag{3.6.2}$$

Clearly, whenever $\bar{v}^k \neq 0$, the above integral is always positive, because $\Lambda_3^{(k)}(t)$ is always a positive definite matrix for any $k > 0$ and any $t \in [t_0, t_f]$. According to Theorem 3.5.1, whenever $\bar{v}^k \neq 0$, there exists a $\bar{\lambda}^k \in (0, 1]$, such that, for all $0 < \lambda \leq \bar{\lambda}^k$,

$$\theta_\rho(u^k + \lambda \bar{v}^k) - \theta_\rho(u^k) \leq -\frac{\lambda}{2} \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t) \bar{v}^k(t)\, dt < 0.$$

So, whenever $\bar{v}^k \neq 0$, the above Armijo-like line search is always well-defined, and the stepsize $\lambda^k = \zeta^{l_k}$ is always a positive number.

3.7 Convergence Properties

In this section, the following three convergence aspects of Algorithm 2 described in this chapter will be discussed;

- If the control function u^k at the k-th iteration satisfies some optimality conditions, would Algorithm 2 terminate automatically? Moreover, does u^k belong to the feasible set $\mathcal{F}$?
- If Algorithm 2 terminates at a finite k-th iteration, does the control function u^k satisfy some optimality conditions? Moreover, does u^k belong to the feasible set $\mathcal{F}$?
- If Algorithm 2 runs for infinite iterations, does the accumulation point u^* of the control function sequence $\{u^k\}_{k=0}^{\infty}$ satisfy some optimality conditions? Moreover, does u^* belong to the feasible set $\mathcal{F}$?

The first two questions are answered by the following theorem. The last question is very involved, and takes the most of this section for the answer.

Theorem 3.7.1 *The necessary and sufficient condition for Algorithm 2 to terminate at a finite k-th iteration is that $u^k \in \mathcal{F}$, and, for any $t \in [t_0, t_f]$, $u^k(t)$ satisfies the Kuhn-Tucker necessary conditions of*

$$\min_{\mu \in \Omega} H(x^{u^k}(t), \mu, \tilde{p}^{u^k}(t), t)$$

where x^{u^k} is the state of the system corresponding to u^k, and $\tilde{p}^{u^k}$ is a costate of the system corresponding to x^{u^k}, u^k, and $\tilde{p}^{u^k}(t_f)$ satisfies transversality condition

$$\tilde{p}^{u^k}(t_f) = \frac{\partial K}{\partial x}(x^{u^k}(t_f), t_f) + \sum_{i=1}^{r} \eta_i^k \left(\frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)\right)^{\mathsf{T}} \tag{3.7.1}$$

with $\eta^k \in \mathcal{R}^r$ a nonnegative constant vector such that

$$\eta_i^k \; h_i(x^{u^k}(t_f), t_f) = 0, \qquad i = 1, \cdots, r. \tag{3.7.2}$$

On the other hand, whenever Algorithm 2 terminates at the k-th iteration, $u^k \in \mathcal{F}$.

Proof: The proof of the "necessary" part is easier than that of the "sufficient" part.

(1) ("necessary" part) Let $\bar{u}^k$ be the optimal solution of the "direction-finding" subproblem (P_k''), corresponding to u^k, x^{u^k}, p^{u^k}, Λ_1^k, Λ_2^k, Λ_3^k. Let $\bar{v}^k = \bar{u}^k - u^k$, and $y^{\bar{v}^k}$ be the solution of the linearized system (3.4.17)-(3.4.18) with input $\bar{v}^k$. Because Algorithm 2 terminates at a finite k-th iteration, we have

$$\bar{u}^k(t) = u^k(t)$$

almost always on $[t_0, t_f]$. By (3.4.17)-(3.4.18), we then have $y^{\bar{v}^k}(t) = 0$ for any $t \in [t_0, t_f]$. Then, according to the optimality conditions of the "direction-finding" subproblem (P_k'') described in section 3.5 and the fact that $y^{\bar{v}^k} = 0$, there exists a function $q^{\bar{v}^k}(t)$, a nonnegative constant vector $\eta^k \in \mathcal{R}^r$ and two nonnegative vector functions $\beta^k(t)$ and $\gamma^k(t)$, such that,

$$\eta_i^k \; h_i(x^{u^k}(t_f), t_f) = 0, \qquad i = 1, \cdots, r, \tag{3.7.3}$$

and

$$-\dot{q}^{\bar{v}^k}(t) = f_x^{(k)}(t)^{\mathsf{T}} q^{\bar{v}^k}(t) \tag{3.7.4}$$

$$q^{\bar{v}^k}(t_f) = \sum_{i=1}^{r} \eta_i^k \left(\frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f)\right)^{\mathsf{T}} \tag{3.7.5}$$

and

$$\nabla_u H(x^{u^k}(t), u^k(t), p^{u^k}(t), t) + f_u^{(k)}(t)^{\mathsf{T}} q^{\bar{v}^k}(t) + \beta^k(t) - \gamma^k(t) = 0 \tag{3.7.6}$$

where

$$\beta_i^k(t)(\bar{u}_i^k(t) - U_i^{max}) = 0, \qquad i = 1, \cdots, m \tag{3.7.7}$$

$$\gamma_i^k(t)(U_i^{min} - \bar{u}_i^k(t)) = 0, \qquad i = 1, \cdots, m \tag{3.7.8}$$

for any $t \in [t_0, t_f]$. Let $\tilde{p}^{u^k}(t) \triangleq p^{u^k}(t) + q^{\bar{v}^k}(t)$. Equation (3.7.6) then becomes

$$\nabla_u H(x^{u^k}(t), u^k(t), \tilde{p}^{u^k}(t), t) + \beta^k(t) - \gamma^k(t) = 0. \tag{3.7.9}$$

Because $p^{u^k}(t)$ satisfies (3.3.3) and (3.4.5), $q^{\bar{v}^k}(t)$ satisfies (3.7.4) and (3.7.5), for any $t \in [t_0, t_f]$, we find that

$$\dot{\tilde{p}}^{u^k}(t) = -\frac{\partial H}{\partial x}(x^{u^k}, \tilde{p}^{u^k}, u^k, t) \tag{3.7.10}$$

$$\tilde{p}^{u^k}(t_f) = \frac{\partial K}{\partial x}(x^{u^k}(t_f), t_f) + \sum_{i=1}^{r} \eta_i^k \left(\frac{\partial h_i}{\partial x}(x^{u^k}(t_f), t_f) \right)^{\mathsf{T}}. \tag{3.7.11}$$

The above means that $\tilde{p}^{u^k}(t)$ is another costate function for the original problem (P) besides $p^{u^k}(t)$, and, above all, $\tilde{p}^{u^k}(t_f)$ satisfies transversality condition (3.7.11) and (3.7.3). Clearly, equations (3.7.9), (3.7.7) and (3.7.8) imply that $(u^k(t), \beta^k(t), \gamma^k(t))$ is a Kuhn-Tucker point of

$$\min_{\mu \in \Omega} H(x^{u^k}(t), \mu, \tilde{p}^{u^k}(t), t)$$

for any $t \in [t_0, t_f]$. The proof of the "necessary" part is now complete.

(2) ("sufficient" part) Because, for any $t \in [t_0, t_f]$, $u^k(t)$ satisfies the Kuhn-Tucker necessary conditions of

$$\min_{\mu \in \Omega} H(x^{u^k}(t), \mu, \tilde{p}^{u^k}(t), t),$$

there exist two nonnegative vector functions $\nu^k(t)$ and $\xi^k(t)$, such that,

$$\nabla_u H(x^{u^k}(t), u^k(t), \tilde{p}^{u^k}(t), t) + \nu^k(t) - \xi^k(t) = 0 \tag{3.7.12}$$

and

$$\nu_i^k(t)(u_i^k(t) - U_i^{max}) = 0, \qquad i = 1, \cdots, m, \tag{3.7.13}$$

$$\xi_i^k(t)(U_i^{min} - u_i^k(t)) = 0, \qquad i = 1, \cdots, m. \tag{3.7.14}$$

Because $\bar{u}^k$ solves the "direction-finding" subproblem (P_k'') and $\bar{v}^k = \bar{u}^k - u^k$, according to the optimality conditions described in section 3.5, there exist nonnegative vector functions $\beta^k(t)$ and $\gamma^k(t)$, such that,

$$\begin{aligned}&\nabla_u H(x^{u^k}(t), u^k(t), p^{u^k}(t), t)\\ &\quad + \Lambda_3^{(k)}(t)(\bar{u}^k(t) - u^k(t)) + f_u^{(k)}(t)^\top q^{\bar{v}^k}(t) + \beta^k(t) - \gamma^k(t) = 0 \qquad (3.7.15)\end{aligned}$$

where

$$\beta_i^k(t)(\bar{u}_i^k(t) - U_i^{max}) = 0, \qquad i = 1, \cdots, m, \qquad (3.7.16)$$

$$\gamma_i^k(t)(U_i^{min} - \bar{u}_i^k(t)) = 0, \qquad i = 1, \cdots, m, \qquad (3.7.17)$$

for any $t \in [t_0, t_f]$. Combining (3.7.12) and (3.7.15), we then have

$$\Lambda_3^{(k)}(t)\bar{v}^k(t) = -f_u^{(k)}(t)^\top q^{\bar{v}^k}(t) - \beta^k(t) + \gamma^k(t) + \nu^k(t) - \xi^k(t).$$

Then

$$\begin{aligned}\int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\, dt &= -\int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\, dt\\ &+ \int_{t_0}^{t_f} \left(-\beta^k(t)^\top \bar{v}^k(t) + \gamma^k(t)^\top \bar{v}^k(t) + \nu^k(t)^\top \bar{v}^k(t) - \xi^k(t)^\top \bar{v}^k(t)\right) dt. \qquad (3.7.18)\end{aligned}$$

Because

$$\begin{aligned}\nu^k(t)^\top \bar{v}^k(t) &= \nu^k(t)^\top(\bar{u}^k(t) - u^k(t))\\ &= \underbrace{\nu^k(t)^\top(U^{max} - u^k(t))}_{=0} + \underbrace{\nu^k(t)}_{\geq 0}^\top (\underbrace{\bar{u}^k(t) - U^{max}}_{\leq 0}) \leq 0 \qquad (3.7.19)\end{aligned}$$

and

$$\begin{aligned}\xi^k(t)^\top \bar{v}^k(t) &= \xi^k(t)^\top(\bar{u}^k(t) - u^k(t))\\ &= \underbrace{\xi^k(t)^\top(U^{min} - u^k(t))}_{=0} + \underbrace{\xi^k(t)}_{\geq 0}^\top (\underbrace{\bar{u}^k(t) - U^{min}}_{\geq 0}) \geq 0, \qquad (3.7.20)\end{aligned}$$

and, from (3.5.14), (3.5.15):

$$\beta^k(t)^\top \bar{v}^k(t) \geq 0, \qquad \gamma^k(t)^\top \bar{v}^k(t) \leq 0, \qquad (3.7.21)$$

we then have,

$$\int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt \leq -\int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\,dt. \tag{3.7.22}$$

Recall, from section 3.5, that

$$\int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\,dt \geq -\sum_{i\in I^{(k)}} \eta_i^k\ h_i(x^{u^k}(t_f), t_f).$$

According to (3.7.2), we know that

$$\sum_{i\in I^{(k)}} \eta_i^k\ h_i(x^{u^k}(t_f), t_f) = 0.$$

So,

$$\int_{t_0}^{t_f} q^{\bar{v}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\,dt \geq 0.$$

Applying the above into (3.7.22), we then have,

$$\int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt \leq 0. \tag{3.7.23}$$

However, since $\Lambda_3^{(k)}(t)$ is a positive definite matrix at any $t\in[t_0, t_f]$, (3.7.23) can hold only when

$$\bar{v}^k(t) = 0,$$

almost always on $[t_0, t_f]$, or

$$\bar{u}^k(t) = u^k(t),$$

almost always on $[t_0, t_f]$, which means Algorithm 2 will terminate automatically at the k-th iteration. The proof of the "sufficient" part is now complete.

(3) ("feasible" part) Clearly, whenever Algorithm 2 terminates at a finite k-th iteration, we have

$$\bar{u}^k(t) = u^k(t)$$

almost always on $[t_0, t_f]$. By (3.4.17)-(3.4.18), we then have $y^{\bar{v}^k}(t) = 0$ for any $t\in[t_0, t_f]$. So, the linear terminal inequality constraints (3.4.19) become

$$h_i(x^{u^k}(t_f), t_f) \leq 0, \qquad i = 1, \cdots, r, \tag{3.7.24}$$

which means that u^k belongs to the feasible control set $\mathcal{F}$ defined in (3.2.7). The proof of the "feasible" part is now complete. □

We now turn to study the case when Algorithm 2 runs for infinite iterations.

Because the set Ω is compact, the control $u(t)$ and the state $x^u(t)$ are uniformly bounded during $[t_0, t_f]$ (see Lemma A.4 in the Appendix), which implies that the exact penalty functional $\theta_\rho(u)$ is bounded below. So, for any sequence $\{u^k\}$ generated by Algorithm 2, the corresponding sequence $\{\theta_\rho^k\}$, which is monotonically decreasing, is bounded below. That means that $\{\theta_\rho^k\}$ has a convergent subsequence. However, the convergence of the sequence $\{\theta_\rho^k\}$ is less important and interesting than the convergence of the control sequence $\{u^k\}$. Next, we would like to find out that, whether the accumulation point of $\{u^k\}$, if there is one, belongs to the feasible set $\mathcal{F}$ and satisfies some optimality conditions.

Let us first denote the optimal control problem

$$\min_{u(t)\in\Omega} \frac{1}{2} y(t_f)^\top \Lambda_1 y(t_f) + \int_{t_0}^{t_f} \left(e(t)^\top v(t) + \frac{1}{2} y(t)^\top \Lambda_2(t) y(t) + \frac{1}{2} v(t)^\top \Lambda_3(t) v(t) \right) dt$$

subject to the linear system

$$\dot{y}(t) = A(t)y(t) + B(t)(u(t) - w(t)) \tag{3.7.25}$$

$$y(t_0) = y_0 \tag{3.7.26}$$

and linear terminal inequality constraints

$$c_i + d_i^\top y(t_f) \leq 0 \qquad i = 1, \cdots, r \tag{3.7.27}$$

by $\Sigma(\Omega, A, B, c, d, e, \Lambda_1, \Lambda_2, \Lambda_3, w, y_0)$. In the above, A, B, c, d, e, Λ_1, Λ_2, Λ_3 and w are all given, and Λ_1, $\Lambda_2(t)$, are convex matrices, and $\Lambda_3(t)$ is a strictly convex matrix. It will be seen in the next chapter that the solution of $\Sigma(\Omega, A, B, c, d, e, \Lambda_1, \Lambda_2, \Lambda_3, w, y_0)$ always exists and is unique.

Definition Let $\bar{u}$ be the unique solution of $\Sigma(\Omega, A, B, c, d, e, \Lambda_1, \Lambda_2, \Lambda_3, w, y_0)$. We say $\bar{u}$ is stable, if, for any $\epsilon > 0$, and any $\Sigma(\Omega, A', B', c', d', e', \Lambda_1', \Lambda_2', \Lambda_3', w', y_0)$ satisfying

$$\max\{\|A-A'\|, \|B-B'\|, \|c-c'\|, \|d-d'\|, \|e-e'\|, \|\Lambda_1-\Lambda_1'\|, \|\Lambda_2-\Lambda_2'\|, \|\Lambda_3-\Lambda_3'\|, \|w-w'\|\} \leq \epsilon,$$

there exists a positive number δ, such that,

$$\|\bar{u}-\bar{u}'\| \leq \delta$$

where $\bar{u}'$ the unique solution of $\Sigma(\Omega, A', B', c', d', e', \Lambda_1', \Lambda_2', \Lambda_3', w', y_0)$.

In order to obtain the global convergence result, we shall make the following additional assumptions throughout the remainder of this section.

(a) For any $\Sigma(\Omega, A, B, c, d, e, \Lambda_1, \Lambda_2, \Lambda_3, w, y_0)$ problem encountered, its solution is always stable.

(b) For any transversality constant $\eta \in \mathcal{R}^r$, its norm is always bounded by ρ, the penalty parameter of the exact penalty functional $\theta_\rho(u)$ defined in (3.5.8).

(c) There exist $\alpha_{11}, \alpha_{12} > 0$, such that, for any k, any $v \in \mathcal{R}^m$,

$$\alpha_{11}\ v^\top v \leq v^\top \Lambda_1^{(k)} v \leq \alpha_{12}\ v^\top v.$$

(d) For positive definite matrices $\Lambda_2^{(k)} \in \mathcal{L}_\infty^{n\times n}[t_0, t_f]$ and $\Lambda_3^{(k)} \in \mathcal{L}_\infty^{m\times m}[t_0, t_f]$, let their updating rules be such that, $\Lambda_2^{(k)}$ and $\Lambda_3^{(k)}$ converge respectively to positive definite matrices Λ_2^* and Λ_3^*, whenever $\begin{pmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{pmatrix}^{(k)}$ converges.

Theorem 3.7.2 *Suppose that the infinite sequence $\{u^k\}$ generated from Algorithm 2 has an accumulation point u^*, that is, there exists a $\{j_k\} \subset \{k\}$, such that,*

$$u^{j_k} \to u^*$$

in the $\mathcal{L}_\infty^m[t_0, t_f]$ topology. Then, $u^* \in \mathcal{F}$, and, for any $t \in [t_0, t_f]$, u^* satisfies the Kuhn-Tucker necessary conditions of

$$\min_{\mu\in\Omega} H(x^{u^*}(t), \mu, p^{u^*}(t), t),$$

where x^{u^*} is the state of the system corresponding to u^*, and $\tilde{p}^{u^*}$ is a costate of the system corresponding to x^{u^*}, u^*, and $\tilde{p}^{u^*}(t_f)$ satisfies transversality condition

$$\tilde{p}^{u^*}(t_f) = \frac{\partial K}{\partial x}(x^{u^*}(t_f), t_f) + \sum_{i=1}^{r} \eta_i^* \left(\frac{\partial h_i}{\partial x}(x^{u^*}(t_f), t_f) \right)^{\mathsf{T}} \tag{3.7.28}$$

with $\eta^* \in \mathcal{R}^r$ a nonnegative constant vector such that

$$\eta_i^* \; h_i(x^{u^*}(t_f), t_f) = 0, \qquad i = 1, \cdots, r. \tag{3.7.29}$$

Proof: Because u^{j_k} converges to u^* in $\mathcal{L}_\infty^m[t_0, t_f]$ topology, both x^{j_k} and p^{j_k} converge in $\mathcal{L}_\infty^n[t_0, t_f]$, according to Lemma A.5. That is

$$u^{j_k} \to u^*, \quad x^{u^{j_k}} \to x^{u^*}, \quad p^{u^{j_k}} \to p^{u^*}.$$

The convergence of u^{j_k}, x^{j_k}, p^{j_k} and the continuity assumptions on H_{xx}, H_{xu}, H_{ux}, H_{uu} then imply that $\begin{pmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{pmatrix}^{(k)}$ converges. So, from condition (d), there exist positive definite matrices $\Lambda_2^* \in \mathcal{L}_\infty^{n \times n}[t_0, t_f]$, $\Lambda_3^* \in \mathcal{L}_\infty^{m \times m}[t_0, t_f]$, such that,

$$\Lambda_i^{j_k} \to \Lambda_i^* \qquad i = 2, 3.$$

Also, from [46], condition (c) implies that there is a $\{j_k'\} \subset \{j_k\}$ and a positive definite matrix $\Lambda_1^* \in \mathcal{R}^{n \times n}$ such that that

$$\Lambda_1^{j_k'} \to \Lambda_1^*.$$

In summary of the above convergence results, we have

$$u^{j_k'} \to u^*, \quad x^{u^{j_k'}} \to x^{u^*}, \quad p^{u^{j_k'}} \to p^{u^*}$$

and

$$\Lambda_1^{j_k'} \to \Lambda_1^*, \quad \Lambda_2^{j_k'} \to \Lambda_2^*, \quad \Lambda_3^{j_k'} \to \Lambda_3^*.$$

Let $\bar{u}^*$ be the solution of the "direction-finding" subproblem (P_k''), corresponding to u^*, x^{u^*}, p^{u^*}, Λ_1^*, Λ_2^*, Λ_3^*. Because the "direction-finding" subproblem (P_k'') is of the type of

$$\Sigma(\Omega, A, B, c, d, e, \Lambda_1, \Lambda_2, \Lambda_3, w, y_0)$$

defined before, its solution is always assumed to be stable. We then have

$$\bar{u}^{j'_k} \to \bar{u}^*.$$

Equivalently, for $v^{j'_k} = u^{j'_k} - \bar{u}^{j'_k}$ and $v^* = u^* - \bar{u}^*$,

$$\bar{v}^{j'_k} \to \bar{v}^*.$$

After reindexing, we may assume

$$u^k \to u^*, \quad \bar{u}^k \to \bar{u}^*, \quad \bar{v}^k \to \bar{v}^*, \quad x^{u^k} \to x^{u^*}, \quad p^{u^k} \to p^{u^*}$$

and

$$\Lambda_1^k \to \Lambda_1^*, \quad \Lambda_2^k \to \Lambda_2^*, \quad \Lambda_3^k \to \Lambda_3^*$$

without losing any generality. Again from the optimality condition of the "direction-finding" subproblem (P''_k), when $y^{\bar{v}^*}$ and $\bar{u}^*$ is its optimal solution pair corresponding to u^*, x^*, p^*, Λ_1^*, Λ_2^*, Λ_3^*, there exists a nonnegative constant vector $\eta^* \in \mathcal{R}^r$ and two nonnegative function vectors $\beta^*(t)$ and $\gamma^*(t)$, such that,

$$\eta_i^* \left(h_i(x^{u^*}(t_f), t_f) + \frac{\partial h_i}{\partial x}(x^{u^*}(t_f), t_f) y^{\bar{v}^*}(t_f) \right) = 0, \quad i = 1, \cdots, r \tag{3.7.30}$$

and

$$-\dot{q}^{\bar{v}^*}(t) = f_x^{(*)}(t)^\top q^{\bar{v}^*}(t) + \Lambda_2^{(*)}(t) y^{\bar{v}^*}(t) \tag{3.7.31}$$

$$q^{\bar{v}^*}(t_f) = \Lambda_1^{(*)} y^{\bar{v}^*}(t_f) + \sum_{i=1}^{r} \eta_i^* \left(\frac{\partial h_i}{\partial x}(x^{u^*}(t_f), t_f) \right)^\top \tag{3.7.32}$$

and

$$\begin{aligned} &\nabla_u H(x^{u^*}(t), u^*(t), p^{u^*}(t), t) \\ &\quad + \Lambda_3^{(*)}(t)(\bar{u}^*(t) - u^*(t)) + f_u^{(*)}(t)^\top q^{\bar{v}^*}(t) + \beta^*(t) - \gamma^*(t) = 0 \end{aligned} \tag{3.7.33}$$

where

$$\beta_i^*(t)(\bar{u}_i^*(t) - U_i^{max}) = 0, \qquad i = 1, \cdots, m \tag{3.7.34}$$

$$\gamma_i^*(t)(U_i^{min} - \bar{u}_i^*(t)) = 0, \qquad i = 1, \cdots, m \tag{3.7.35}$$

for any $t \in [t_0, t_f]$. We now claim that

$$\bar{v}^*(t) = 0$$

almost everywhere on $[t_0, t_f]$. Suppose not. Because

$$\|\eta^*\| \leq \rho,$$

according to condition (b), Theorem 3.5.1 applies. So, there exists a $\bar{\lambda}^* \in (0, \infty)$, such that,

$$\theta(u^* + \lambda \bar{v}^*) - \theta(u^*) \leq -\frac{\lambda}{2} \int_{t_0}^{t_f} \bar{v}^*(t)^\top \Lambda_3^*(t) \bar{v}^*(t)\, dt$$

for any $0 < \lambda \leq \bar{\lambda}^*$. Because the stepsize λ^* is defined by

$$\lambda^* = \arg \min_{0 \leq \lambda \leq 1} \theta(u^* + \lambda(\bar{u}^* - u^*)),$$

it is clear that

$$\lambda^* \geq \min\{\bar{\lambda}^*, 1\} > 0.$$

Let

$$\epsilon^* = \frac{\lambda^*}{2} \int_{t_0}^{t_f} \bar{v}^*(t)^\top \Lambda_3^*(t) \bar{v}^*(t)\, dt.$$

Because $\lambda^* > 0$, $\bar{v}^* \neq 0$, and $\Lambda_3^*(t)$ positive definite for any $t \in [t_o, t_f]$, we have $\epsilon^* > 0$. Then

$$\theta(u^* + \lambda^* \bar{v}^*) - \theta(u^*) \leq -\epsilon^*. \tag{3.7.36}$$

Because $\theta(u)$ is a continuous functional in $u_{[t_0, t_f]}$, and

$$u^k + \lambda^* \bar{v}^k \to u^* + \lambda^* \bar{v}^*,$$

then, for sufficiently large k,

$$\theta(u^k + \lambda^* \bar{v}^k) - \theta(u^* + \lambda^* \bar{v}^*) \leq \frac{\epsilon^*}{2}. \tag{3.7.37}$$

Summing (3.7.36) and (3.7.37), we get

$$\theta(u^k + \lambda^* \bar{v}^k) - \theta(u^*) \leq -\frac{\epsilon^*}{2}. \tag{3.7.38}$$

By the definition of λ^k, we then get

$$\theta(u^k + \lambda^k \bar{v}^k) \leq \theta(u^k + \lambda^* \bar{v}^k) \leq \theta(u^*) - \frac{\epsilon^*}{2}. \tag{3.7.39}$$

So,

$$\theta(u^{k+1}) \leq \theta(u^*), \tag{3.7.40}$$

which contradicts the fact that $\{\theta(u^k)\}$ is a monotonically decreasing sequence. The claim that $\bar{v}^* = 0$ is now proved.

Because now $\bar{u}^* = u^*$, by (3.4.17)-(3.4.18), we then have $y^{\bar{v}^*}(t) = 0$ for any $t \in [t_0, t_f]$. So, the linear terminal inequality constraints (3.4.19) become

$$h_i(x^{u^*}(t_f), t_f) \leq 0, \qquad i = 1, \cdots, r, \tag{3.7.41}$$

which means that u^* belongs to the feasible control set $\mathcal{F}$ defined in (3.2.7). Furthermore, by the fact that $y^{\bar{v}^*} = 0$, the optimality conditions (3.7.30), (3.7.31), (3.7.32) and (3.7.33) become

$$\eta_i^* \; h_i(x^{u^*}(t_f), t_f) = 0, \qquad i = 1, \cdots, r, \tag{3.7.42}$$

and,

$$-\dot{q}^{\bar{v}^*}(t) = f_x^*(t)^\top q^{\bar{v}^*}(t) \tag{3.7.43}$$

$$q^{\bar{v}^*}(t_f) = \sum_{i=1}^{r} \eta_i^* \left(\frac{\partial h_i}{\partial x}(x^{u^*}(t_f), t_f) \right)^\top. \tag{3.7.44}$$

and,

$$\nabla_u H(x^{u^*}(t), u^*(t), p^{u^*}(t), t) + f_u^{(*)}(t)^\top q^{\bar{v}^*}(t) + \beta^*(t) - \gamma^*(t) = 0 \tag{3.7.45}$$

for any $t \in [t_0, t_f]$. Let $\tilde{p}^{u^*}(t) \triangleq p^{u^*}(t) + q^{\bar{v}^*}(t)$. Because $p^{u^*}(t)$ satisfies (3.3.3) and (3.4.5), $q^{\bar{v}^*}(t)$ satisfies (3.7.43) and (3.7.44), we have that

$$\dot{\tilde{p}}^{u^*}(t) = -\frac{\partial H}{\partial x}(x^{u^*}, \tilde{p}^{u^*}, u^*, t) \tag{3.7.46}$$

$$\tilde{p}^{u^*}(t_f) = \frac{\partial K}{\partial x}(x^{u^*}(t_f), t_f) + \sum_{i=1}^{r} \eta_i^* \left(\frac{\partial h_i}{\partial x}(x^{u^*}(t_f), t_f) \right)^\top. \tag{3.7.47}$$

That means that $\tilde{p}^{u^*}(t)$ is another costate function for the original problem (P) besides $p^{u^*}(t)$, and, above all, $\tilde{p}^{u^*}(t_f)$ satisfies transversality condition (3.7.47). Then, (3.7.45) becomes

$$\nabla_u H(x^{u^*}(t), u^*(t), \tilde{p}^{u^*}(t), t) + \beta^*(t) - \gamma^*(t) = 0. \tag{3.7.48}$$

Clearly, (3.7.34), (3.7.35) and (3.7.48) indicate that, $(\bar{u}^*(t), \beta^*(t), \gamma^*(t))$ is a Kuhn-Tucker point of

$$\min_{\mu \in \Omega} H(x^{u^*}(t), \mu, \tilde{p}^{u^*}(t), t),$$

for any $t \in [t_0, t_f]$. The proof is now complete. □

Chapter 4

A New Algorithm for Solving Optimal Control Problems with Hard Control Constraints, End-point Inequality Constraints, and a Variable Initial State

4.1 Introduction

The ideas behind the algorithm developed in Chapter 3 can be further extended to a much more general optimal control problem which has not only hard control constraints and terminal-state inequality constraints, but also a variable initial state vector, some components of which are allowed to vary within a constraint box while the remaining components are fixed. As will be seen later, the problem being considered in this chapter can include the optimal control problems in the most general setting, namely, the problems which are subjected to control constraints, path constraints, end-point constraints, a variable initial state, and a a variable vector of design parameters, within a fixed end-time or free end-time interval.

Similar to the procedure in Chapter 3, the algorithm being described in this chapter is first based on a second-order approximation to the change of the cost functional due to a change in the control and a change in the initial state. Further ap-

proximation produces a simple convex functional. An exact penalty type of function is employed to penalize any violated end-point inequality constraints. We then show that the solution of the minimization of the convex functional, subject to linearized system dynamics, original hard control constraints, original constraint box for some initial state variables, and linearized end-point constraints, generates a descent direction of that exact penalty function.

4.2 Problem Formulation

The dynamical system considered is described by the differential equation, defined on a fixed end-time interval $[t_0, t_f]$,

$$\dot{x}^{u,x_0}(t) = f(x^{u,x_0}(t), u(t), t), \tag{4.2.1}$$

$$x^{u,x_0}(t_0) = x_0. \tag{4.2.2}$$

There are n_{x_0} components of the initial state vector x_0 which are allowed to vary within a constraint box, while the remaining $n - n_{x_0}$ components are fixed. That is, there is an index set $I_{x_0} \subset \{1, \cdots, n\}$ such that

$$x_0 \in \mathcal{S} = \{ \nu \in \mathcal{R}^n \mid X_i^{min} \leq \nu_i \leq X_i^{max},\ i \in I_{x_0};\ \nu_i = x_{0i}(\text{fixed}),\ i \notin I_{x_0} \} \tag{4.2.3}$$

where $\mathcal{S}$ is compact. The dynamical system (4.2.1)-(4.2.2) is also subject to the control constraints

$$u(t) \in \Omega = \{ \mu \in \mathcal{R}^m \mid U_i^{min} \leq \mu_i \leq U_i^{max},\ i = 1, \cdots, m \} \tag{4.2.4}$$

where Ω is compact, and the end-point inequality constraints,

$$g_i(u, x_0) = h_i(x_0, x^{u,x_0}(t_f)) \leq 0, \qquad i = 1, \cdots, r. \tag{4.2.5}$$

In the above, $x^{u,x_0}(t) \in \mathcal{R}^n$ is the state of the system at time $t \in [t_0, t_f]$, which corresponds to both the control $u(t) \in \mathcal{R}^m$ and the initial value of the state x_0. Let the set of admissible controls be,

$$\mathcal{U} = \{ u \mid u : [t_0, t_f] \to \Omega \text{ is continuous a.e. } \} \subset \mathcal{L}_\infty^m[t_0, t_f], \tag{4.2.6}$$

and let $\tilde{\mathcal{U}}$ be the set of equivalence classes of functions in $\mathcal{U}$ which are equal almost everywhere on $[t_0, t_f]$. Let the combined set of feasible controls and feasible initial states be

$$\mathcal{F} \triangleq \{ (u, x_0) \mid u \in \tilde{\mathcal{U}},\ x_0 \in \mathcal{S},\ g_i(u, x_0) \leq 0,\ i = 1, \cdots, r \}. \tag{4.2.7}$$

We may now formulate a constrained optimal control problem as follows:

Problem (P). Subject to the dynamical system (4.2.1)-(4.2.2), find a control $u \in \mathcal{U}$ and an initial state $x_0 \in \mathcal{S}$ such that the cost functional

$$J(u, x_0) = K(x_0,\ x^{u,x_0}(t_f)) + \int_{t_0}^{t_f} L(x^{u,x_0}(\tau), u(\tau), \tau)\, d\tau \tag{4.2.8}$$

is minimized over $\mathcal{F}$.

Throughout, it is understood that the norm of any $w \in \mathcal{R}^p$ for some dimension p is

$$\|w\| = \max_{1 \leq i \leq p} |w_i|, \tag{4.2.9}$$

the norm of any $w \in \mathcal{L}_\infty^p[t_0, t_f]$ is

$$\|w\| = \operatorname{ess\,sup}_{[t_0, t_f]} \|w(t)\|, \tag{4.2.10}$$

the norm of any $H \in \mathcal{R}^{p \times p}$ is

$$\|H\| = \max_{1 \leq i \leq p} \sum_{j=1}^{p} |h_{ij}|, \tag{4.2.11}$$

the norm of any $H \in \mathcal{L}_\infty^{p \times p}[t_0, t_f]$ is

$$\|H\| = \operatorname{ess\,sup}_{[t_0, t_f]} \|H(t)\|. \tag{4.2.12}$$

The following conditions are assumed to be satisfied.

Assumption 4.2.1 $f : \mathcal{R}^n \times \mathcal{R}^m \times [t_0, t_f] \to \mathcal{R}^n$, $h_i : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$, $i = 1, \cdots, r$, $K : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$, $L : \mathcal{R}^n \times \mathcal{R}^m \times [t_0, t_f] \to \mathcal{R}$. *$f$ and L, together with their partial derivatives up to third-order with respect to each of the components of x and u, are continuous for all $(x, u, t) \in \mathcal{R}^n \times \mathcal{R}^m \times [t_0, t_f]$. h_i and K are continuously differentiable with respect to both x_0 and x;*

Assumption 4.2.2 *There exists a positive constant M such that*

$$\|f(x,u,t)\| \leq M(1+\|x\|) \tag{4.2.13}$$

for all $(x,u,t) \in \mathcal{R}^n \times \Omega \times [t_0, t_f]$.

Remark: From the theory of differential equations, the system (4.2.1)-(4.2.2) has a unique solution $x^{u,x_0}(t)$ on interval $[t_0, t_f]$ corresponding to each $u \in \mathcal{U}$ and each $x_0 \in \mathcal{S}$, and $x^{u,x_0}(t)$ is absolutely continuous (see, e.g. [166]).

4.3 Optimality Conditions

Let the Hamiltonian function of problem (P) be

$$H(x^{u,x_0}, p^{u,x_0}, u, t, p_0) = p_0\, L(x^{u,x_0}, u, t) + (p^{u,x_0})^{\top} f(x^{u,x_0}, u, t) \tag{4.3.1}$$

and its minimal function be

$$M(x^{u,x_0}(t), p^{u,x_0}(t), t, p_0) = \inf_{\mu \in \Omega} H(x^{u,x_0}(t), p^{u,x_0}(t), \mu, t, p_0). \tag{4.3.2}$$

Also, let the costate function of problem (P) be

$$\dot{p}^{u,x_0}(t) = -\frac{\partial H}{\partial x}(x^{u,x_0}(t), p^{u,x_0}(t), u(t), t, p_0). \tag{4.3.3}$$

If problem (P) has neither variable initial state nor end-point inequality constraints, a necessary condition for optimality is the well-known Pontryagin Maximum Principle [4, 120]. A similar maximum principle, which is given below, still holds when problem (P) has both variable initial state and end-point inequality constraints [14,58]:

Theorem 4.3.1 *Let* $(u^*, x_0^*) \in \mathcal{F}$ *be the optimal solution for the problem* (P). *Let* $x^{u^*,x_0^*}(t)$ *be the solution of the system (4.2.1)-(4.2.2) with input* $u^*(t)$ *during* $[t_0, t_f]$ *and initial state* x_0^*. *Then, there exists an absolutely continuous costate* $p^{u^*,x_0^*}(t) \in \mathcal{R}^n$, $t \in [t_0, t_f]$, *which is not identically zero in* $[t_0, t_f]$, *a nonnegative constant scalar* $p_0^* \geq 0$,

three constant real vectors $\eta^* \in \mathcal{R}^r$, $\iota^* \in \mathcal{R}^n$, $\kappa^* \in \mathcal{R}^n$, *such that, for any* $t \in [t_0, t_f]$ *(a.e.),*

$$-\dot{p}^{u^*,x_0^*}(t) = \frac{\partial H}{\partial x}(x^{u^*,x_0^*}(t), p^{u^*,x_0^*}(t), u^*(t), t, p_0^*) \quad (4.3.4)$$

$$M(x^{u^*,x_0^*}(t), p^{u^*,x_0^*}(t), t, p_0^*) = H(x^{u^*,x_0^*}(t), p^{u^*,x_0^*}(t), u^*(t), t, p_0^*) \quad (4.3.5)$$

$$\frac{dM}{dt}(x^{u^*,x_0^*}(t), p^{u^*,x_0^*}(t), t, p_0^*) = \frac{\partial H}{\partial t}(x^{u^*,x_0^*}(t), p^{u^*,x_0^*}(t), u^*(t), t, p_0^*). \quad (4.3.6)$$

Moreover, the following transversality conditions hold,

$$p^{u^*,x_0^*}(t_0) = -p_0^* \frac{\partial K}{\partial x_0}(x_0^*, x^{u^*,x_0^*}(t_f)) - \sum_{i=1}^{r} \eta_i^* \frac{\partial h_i}{\partial x_0}(x_0^*, x^{u^*,x_0^*}(t_f)) - (\iota^* - \kappa^*) \quad (4.3.7)$$

and

$$p^{u^*,x_0^*}(t_f) = p_0^* \frac{\partial K}{\partial x_f}(x_0^*, x^{u^*,x_0^*}(t_f)) + \sum_{i=1}^{r} \eta_i^* \frac{\partial h_i}{\partial x_f}(x_0^*, x^{u^*,x_0^*}(t_f)) \quad (4.3.8)$$

with

$$\eta_i^* \geq 0 \quad \text{and} \quad \eta_i^* \, h_i(x_0^*, x^{u^*,x_0^*}(t_f)) = 0 \quad (4.3.9)$$

for all $i = 1, \cdots, r$, *and*

$$\iota_i^* = \kappa_i^* = 0, \qquad \forall i \notin I_{x_0}, \quad (4.3.10)$$

and

$$\iota_i^* \geq 0 \quad \text{and} \quad \iota_i^* (x_{0_i}^* - X_i^{max}) = 0 \quad (4.3.11)$$

$$\kappa_i^* \geq 0 \quad \text{and} \quad \kappa_i^* (X_i^{min} - x_{0_i}^*) = 0 \quad (4.3.12)$$

for all $i \in I_{x_0}$.

Remark If $p_0^* = 0$, then function $L(x^{u^*,x_0^*}(t), u^*(t), t)$ will not appear in (4.3.1), nor will $K(x_0^*, x^{u^*,x_0^*}(t_f))$ in (4.3.7)-(4.3.8). That means that the above optimality conditions of problem (P) are irrelevant to the cost functional $J(u, x_0)$, which is obviously "pathological" [4]. Such problems require much more complicated analysis and techniques. We will not treat such problems here. Thus, assume $p_0^* \neq 0$. When $p_0^* \neq 0$ (that is, $p_0^* > 0$), it can be seen easily that the conclusions of

the above theorem will not be altered by assuming $p_0^* = 1$. Therefore, throughout this chapter, we only consider the situation when $p_0^* = 1$. For abbreviation, let $H(x^{u,x_0}(t), p^{u,x_0}(t), u(t), t)$ be $H(x^{u,x_0}(t), p^{u,x_0}(t), u(t), t, p_0)$ with $p_0 = 1$, and $M(x^{u,x_0}(t), p^{u,x_0}(t), t)$ be $M(x^{u,x_0}(t), p^{u,x_0}(t), t, p_0)$ with $p_0 = 1$.

4.4 The Algorithm Utilizing Second-Order Information

Let $v(t)$ be the variation of the control, $v(t) = u^2(t) - u^1(t)$, w be the variation of the initial state, $w = x_0^2 - x_0^1$, and let $y^{v,w}(t)$ satisfy the following linearized system equations

$$\dot{y}^{v,w}(t) = f_x^{(1)}(t)\, y^{v,w}(t) + f_u^{(1)}(t)\, v(t) \tag{4.4.1}$$

$$y^{v,w}(t_0) = w \tag{4.4.2}$$

where $f_x^{(1)}(t)$ and $f_u^{(1)}(t)$ are evaluated at $(x^{u^1,x_0^1}(t),\, u^1(t),\, t)$. Denote

$$\Delta J(u^1,\, x_0^1)(v,w) \triangleq \left(K_{x_0}(x_0^1, x^{u^1,x_0^1}(t_f)) + p^{u^1,x_0^1}(t_0)^\top\right) w + \int_{t_0}^{t_f} H_u^{(1)}(t)\, v(t)\, dt \tag{4.4.3}$$

and

$$\begin{aligned}\Delta^2 J(u^1,\, x_0^1)(v,w) \triangleq\ & \frac{1}{2}\begin{pmatrix} w \\ y^{v,w}(t_f)\end{pmatrix}^\top \begin{pmatrix} K_{x_0x_0}^{(1)} & K_{x_0x_f}^{(1)} \\ K_{x_fx_0}^{(1)} & K_{x_fx_f}^{(1)}\end{pmatrix}\begin{pmatrix} w \\ y^{v,w}(t_f)\end{pmatrix} \\ & + \frac{1}{2}\int_{t_0}^{t_f} \begin{pmatrix} y^{v,w}(t) \\ v(t)\end{pmatrix}^\top \begin{pmatrix} H_{xx}^{(1)}(t) & H_{xu}^{(1)}(t) \\ H_{ux}^{(1)}(t) & H_{uu}^{(1)}(t)\end{pmatrix}\begin{pmatrix} y^{v,w}(t) \\ v(t)\end{pmatrix} \mathrm{dt}\end{aligned} \tag{4.4.4}$$

where $K_{x_0x_0}^{(1)}(t)$, $K_{x_0x_f}^{(1)}(t)$, $K_{x_0x_f}^{(1)}(t)$ and $K_{x_fx_f}^{(1)}(t)$ are evaluated at $(x_0^1,\, x^{u^1,x_0^1}(t_f))$, and, $H_x^{(1)}(t)$, $H_{xx}^{(1)}(t)$, $H_{xu}^{(1)}(t)$, $H_{ux}^{(1)}(t)$ and $H_{uu}^{(1)}(t)$ are evaluated at $(x^{u^1,x_0^1}(t), p^{u^1,x_0^1}(t), u^1(t), t)$. In the above, $x^{u^1,x_0^1}(t)$ and $p^{u^1,x_0^1}(t)$ are the state and costate of problem (P) corresponding to both the control $u^1(t)$ and the initial state x_0^1. If we let $p^{u^1,x_0^1}(t)$ satisfy the following terminal condition,

$$p^{u^1,x_0^1}(t_f) = \frac{\partial K}{\partial x_f}(x_0^1, x^{u^1,x_0^1}(t_f)), \tag{4.4.5}$$

the following proposition then shows that for any $u^1,\ u^2 \in \mathcal{U}$, $\Delta J(u^1,\, x_0^1)(v,w)$ is a first-order estimate for $J(u^2,\, x_0^2) - J(u^1,\, x_0^1)$, and $\Delta J(u^1,\, x_0^1)(v,w) + \Delta^2 J(u^1,\, x_0^1)(v,w)$ is a second-order estimate:

Proposition 4.4.1 *There exist* $c_1, c_2, c_3, c_4, c_5, c_6, c_7 \in (0, \infty)$, *such that, for any* u^1, $u^2 \in \mathcal{U}$, *any* x_0^1, $x_0^2 \in \mathcal{S}$,

$$\left| J(u^2; x_0^2) - J(u^1; x_0^1) - \Delta J(u^1; x_0^1)(v, w) \right| \leq c_1 \| v \|^2 + c_2 \| v \| \cdot \| w \| + c_3 \| w \|^2 \quad (4.4.6)$$

and

$$\left| J(u^2; x_0^2) - J(u^1; x_0^1) - \Delta J(u^1; x_0^1)(v, w) - \Delta^2 J(u^1; x_0^1)(v, w) \right| \leq c_4 \| v \|^3 + c_5 \| v \|^2 \cdot \| w \| + c_6 \| v \| \cdot \| w \|^2 + c_7 \| w \|^3 \quad (4.4.7)$$

where $v = u^2 - u^1$, $w = x_0^2 - x_0^1$.

Proof: See Lemma B.5 in Appendix. □

In viewing the above proposition, a natural and convenient way to find an improving control u^{k+1} and an improving initial state x_0^{k+1} at the $k-th$ iteration is to minimize $\Delta J(u^k; x_0^k)(u - u^k, x_0 - x_0^k)$ or $\Delta J(u^k; x_0^k)(u - u^k, x_0 - x_0^k) + \Delta^2 J(u^k; x_0^k)(u - u^k, x_0 - x_0^k)$. The former is, of course, a first-order method, and the latter a second-order method. In what follows, only the second-order method is studied. That is, to solve the original problem (P), the following "direction-finding" subproblem (P_k') is solved repeatedly,

$$(P_k') \quad \min_{u \in \mathcal{U},\, x_0 \in \mathcal{S}} \ \Delta J(u^k; x_0^k)(u - u^k, x_0 - x_0^k) + \Delta^2 J(u^k; x_0^k)(u - u^k, x_0 - x_0^k)$$

subject to the linearized system,

$$\dot{y}^{v^k; w^k}(t) = f_x^{(k)}(t)\, y^{v^k; w^k}(t) + f_u^{(k)}(t)\, v^k \quad (4.4.8)$$

$$y^{v^k; w^k}(t_0) = w^k \quad (4.4.9)$$

and the linearized end-point inequality constraints,

$$h_i(x_0^k, x^{u^k; x_0^k}(t_f)) + \frac{\partial h_i}{\partial x_0}(x_0^k, x^{u^k; x_0^k}(t_f)) w^k + \frac{\partial h_i}{\partial x_f}(x_0^k, x^{u^k; x_0^k}(t_f)) y^{v^k; w^k}(t_f) \leq 0 \quad (4.4.10)$$

for $i = 1, \cdots, r$, with $v^k = u - u^k$, $w^k = x_0 - x_0^k$.

Algorithm 1.

Step 0. Select $x_0^0 \in \mathcal{S}$, $u^0 \in \mathcal{U}$. Set $k=0$.

Step 1. Compute state x^{u^k, x_0^k} by forward integrating (4.2.1)-(4.2.2).

Step 2. Compute costate p^{u^k, x_0^k} by backward integrating (4.3.3) from terminal condition (4.2.2).

Step 3. Solve the "direction-finding" subproblem (P_k').

Step 4. If its solution is such that $(\bar{u}^k, \bar{x}_0^k) = (u_0^k, x_0^k)$, stop. Otherwise, go to Step 5.

Step 5. Compute a suitable stepsize $\lambda^k > 0$, according to some stepsize rule.

Step 6. Set $x_0^{k+1} = x_0^k + \lambda^k(\bar{x}_0^k - x_0^k)$, $u^{k+1} = u^k + \lambda^k(\bar{u}^k - u^k)$. Set $k = k+1$ and return to Step 2.

However, the above "direction-finding" subproblem (P_k')'s are still difficult to solve. Further simplification is needed, while still preserving the approximation order of two. Because $\begin{pmatrix} K_{x_0x_0} & K_{x_0x_f} \\ K_{x_fx_0} & K_{x_fx_f} \end{pmatrix}$ is real and symmetric, there exist two constant matrices Λ_1, which is positive definite, and Λ_2, which is semi-positive definite, such that

$$\begin{pmatrix} w \\ y^{v,w}(t_f) \end{pmatrix}^{\top} \begin{pmatrix} K_{x_0x_0} & K_{x_0x_f} \\ K_{x_fx_0} & K_{x_fx_f} \end{pmatrix} \begin{pmatrix} w \\ y^{v,w}(t_f) \end{pmatrix} \leq w^{\top}\Lambda_1 w + y^{v,w}(t_f)^{\top}\Lambda_2 y^{v,w}(t_f). \quad (4.4.11)$$

Similarly, because $\begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix}$ is real and symmetric, there exist two time-varying matrices $\Lambda_3(t)$, which is semi-positive definite, and $\Lambda_4(t)$ which is positive definite, such that, for any $t \in [t_0, t_f]$,

$$\begin{pmatrix} y^{v,w}(t) \\ v(t) \end{pmatrix}^{\top} \begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix} \begin{pmatrix} y^{v,w}(t) \\ v(t) \end{pmatrix} \leq y^{v,w}(t)^{\top}\Lambda_3(t) y^{v,w}(t) + v(t)^{\top}\Lambda_4(t) v(t). \quad (4.4.12)$$

Denote

$$J_2(u^1, x_0^1)(v, w) \triangleq K_2^{(1)}(w, y^{v,w}(t_f)) + \int_{t_0}^{t_f} L_2^{(1)}(y^{v,w}(\tau), v(\tau), \tau)\, d\tau \quad (4.4.13)$$

with

$$K_2^{(1)}(w, y^{v,w}(t_f)) = \left(K_{x_0}(x_0^1, x^{u^1, x_0^1}(t_f)) + p^{u^1, x_0^1}(t_0)^{\top} \right) w$$

$$+\frac{1}{2}w^{\top}\Lambda_1^{(1)}w+\frac{1}{2}y^{v,w}(t_f)^{\top}\Lambda_2^{(1)}y^{v,w}(t_f) \tag{4.4.14}$$

$$\begin{aligned} L_2^{(1)}(y^{v,w}(t),v(t),t) &= H_u^{(1)}(t)\,v(t)+\frac{1}{2}y^{v,w}(t)^{\top}\Lambda_3^{(1)}(t)y^{v,w}(t) \\ &\quad +\frac{1}{2}v(t)^{\top}\Lambda_4^{(1)}(t)v(t). \end{aligned} \tag{4.4.15}$$

Proposition 4.4.2 *There exist $c_1, c_2, c_3, c_4 \in (0,\infty)$, such that, for any u^1, $u^2 \in \mathcal{U}$, any x_0^1, $x_0^2 \in \mathcal{S}$,*

$$\begin{aligned} &J(u^2,x_0^2)-J(u^1,x_0^1)-J_2(u^1,x_0^1)(v,w) \\ &\qquad \leq c_1\|v\|^3+c_2\|v\|^2\cdot\|w\|+c_3\|v\|\cdot\|w\|^2+c_4\|w\|^3 \end{aligned} \tag{4.4.16}$$

where $v = u^2 - u^1$, $w = x_0^2 - x_0^1$.

Proof: The proof is done by applying (4.4.11), (4.4.12), (4.4.13) to Proposition 4.4.1. □

The above shows that $J_2(u^1, x_0^1)(v,w)$ is still a second order approximation of $J(u^2,x_0^2)-J(u^1,x_0^1)$, when (u^2, x_0^2) is close to (u^1, x_0^1). Together with the fact that its terminal term and integrand terms are all convex functions, $J_2(u^1, x_0^1)(v,w)$ is therefore an easier approximation of $J(u^2,x_0^2)-J(u^1,x_0^1)$ to compute with than $\Delta J(u^1, x_0^1)(v,w)+\Delta^2 J(u^1, x_0^1)(v,w)$. It reminds us of the similar case in finite-dimensional optimization handled by the quasi-Newton method or the Han-Powell method where the Hessian, which is not necessary positive definite, is iteratively replaced by an updated positive definite matrix to facilitate both the computation and the convergence analysis. Hence, instead of solving the original problem (P), the following new "direction-finding" subproblem (P_k'') is solved repeatedly,

$$(P_k'') \quad \min_{u\in\mathcal{U},\,x_0\in\mathcal{S}} \; J_2(u^k, x_0^k)(u-u^k)(x_0-x_0^k)$$

subject to the linearized system,

$$\dot{y}^{v^k,w^k}(t) = f_x^{(k)}(t)\,y^{v^k,w^k}(t)+f_u^{(k)}(t)\,v^k \tag{4.4.17}$$

$$y^{v^k,w^k}(t_0) = w^k \tag{4.4.18}$$

and the linearized end-point inequality constraints,

$$h_i(x_0^k,\, x^{u^k;x_0^k}(t_f)) + \frac{\partial h_i}{\partial x_0}(x_0^k,\, x^{u^k;x_0^k}(t_f))w^k + \frac{\partial h_i}{\partial x_f}(x_0^k,\, x^{u^k;x_0^k}(t_f))y^{v^k;w^k}(t_f) \;\leq\; 0 \tag{4.4.19}$$

for $i = 1, \cdots, r$, with $v^k = u - u^k$, $w^k = x_0 - x_0^k$.

Algorithm 2.

Step 0. Select $x_0^0 \in \mathcal{S}$, $u^0 \in \mathcal{U}$. Set $k = 0$.

Step 1. Compute state $x^{u^k;x_0^k}$ by forward integrating (4.2.1)-(4.2.2).

Step 2. Compute costate $p^{u^k;x_0^k}$ by backward integrating (4.3.3) from terminal condition (4.2.2).

Step 3. Solve the "direction-finding" subproblem (P_k'').

Step 4. If its solution is such that $(\bar{u}^k,\, \bar{x}_0^k) = (u_0^k\; x_0^k)$, stop. Otherwise, go to Step 5.

Step 5. Compute a suitable stepsize $\lambda^k > 0$, according to some stepsize rule.

Step 6. Set $x_0^{k+1} = x_0^k + \lambda^k(\bar{x}_0^k - x_0^k)$, $u^{k+1} = u^k + \lambda^k(\bar{u}^k - u^k)$. Set $k = k+1$ and return to Step 2.

Remark Clearly, set Ω defined in (4.2.4) and set $\mathcal{S}$ defined in (4.2.3) are convex. Because $u^k(t), \bar{u}^k(t) \in \Omega$, and, $x_0^k, \bar{x}_0^k \in \mathcal{S}$,

$$u^{k+1}(t) \;=\; u^k(t) + \lambda^k(\bar{u}^k(t) - u^k(t)) \;=\; (1 - \lambda^k)u^k(t) + \lambda^k\bar{u}^k(t)$$

and

$$x_0^{k+1} \;=\; x_0^k + \lambda^k(\bar{x}_0^k - x_0^k) \;=\; (1 - \lambda^k)x_0^k + \lambda^k\bar{x}_0^k.$$

So, we have $u^{k+1}(t) \in \Omega$ for any $t \in [t_0, t_f]$ and for any $\lambda \in [0, 1]$, and, $x_0^{k+1} \in \mathcal{S}$ for any $\lambda \in [0, 1]$. That is, the algorithm defined above will always generate a sequence of admissible controls, $u^k \in \mathcal{U}$, $k = 1, 2, \cdots$, and a sequence of admissible initial states, $x_0^k \in \mathcal{S}$, $k = 1, 2, \cdots$, because both the starting control u^0 and the starting initial state x_0^0

are admissible. It is also important to note that the sequence $\{u^k, x_0^k\}_{k=0}^{\infty}$, generated by Algorithm 2, may not always belong to the original feasible set $\mathcal{F}$, defined in (4.2.7). However, by the arguments similar to the ones in the global convergence analysis on the algorithm in Chapter 3, $\{u^k, x_0^k\}_{k=0}^{\infty}$ should ultimately stop at, or converge to, a feasible control which satisfies some optimality conditions.

Remark It is important to note that there are two kinds of constraints which are dealt with differently by the above algorithm: sets $\mathcal{U}$ and $\mathcal{S}$ are the kinds of constraints which must be satisfied during any intermediate iterations; while the constraints, represented by end-point inequalities (4.2.5), can be violated at intermediate iterations. However, as will be seen in section 4.5, by employing an exact penalty type of function to penalize any violated end-point inequality constraints, we can show that the solution of problem (P_k'') generates a descent direction of that exact penalty function. So, iteration after iteration, the algorithm monotonically decreases the value of that penalty function, until all the terminal-state constraints are satisfied and the cost functional is minimized.

An advantage of the above algorithm is that the original constrained nonlinear problem is solved by solving a sequence of constrained linear quadratic optimal control problems, which are much simpler than the original one. Besides, at iteration k, the existence and uniqueness of the solution of the constrained linear quadratic problem is always guaranteed, as long as the feasible set $\mathcal{F}_k$ is not empty (see Chapter 5 for details), which is not the case for problem (P_k'). Also, as will be seen later in this chapter, the algorithm generates a descent direction of an exact penalty functional. So, in all respects, the above algorithm can be regarded as an analog to the quasi-Newton method or the Han-Powell method in finite-dimensional

Another advantage of the above algorithm is that, for each subproblem at the k-th iteration, its hamiltonian function, which is quadratic in both state $x(t)$ and control $u(t)$ for any $t \in [t_0, t_f]$, is a strictly convex function in control $u(t)$. So, the optimal control of every subproblem can never be singular.

However, unlike the finite-dimensional case where the "direction-finding" subproblem is a quadratic programming problem which can be solved in a finite number of steps, the exact solution of the above "direction-finding" subproblem (P_k'') at each iteration may require an infinite number of steps. Practically, the iterations can be stopped after some finite number of steps after the optimal solution is approached within a certain accuracy range.

It should also be pointed out that after $K_{xx}(t_f)$ is replaced by its approximation Λ_1, and $\begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix}$ is replaced by the block diagonal matrix $\begin{pmatrix} \Lambda_2(t) & 0 \\ 0 & \Lambda_3((t)) \end{pmatrix}$, the second-order local convergence rate may not hold any longer. However, intuition tells us that the new local convergence rate should be at least first-order, and possibly superlinear, depending upon the tightness of the approximations.

4.5 Descent Properties

Let $(\bar{u}^k, \bar{x}_0^k)$ be the optimal solution of the "direction-finding" subproblem (P_k''). Let $\bar{v}^k = \bar{u}^k - u^k$, $\bar{w}^k = \bar{x}_0^k - x_0^k$, and $y^{\bar{v}^k, \bar{w}^k}$ be the solution of the linearized system (4.4.17)-(4.4.18) with input $\bar{v}^k$ and initial state $\bar{w}^k$. Let

$$\pi_i^k(w, y_f) \triangleq h_i(x_0^k, x^{u^k, x_0^k}(t_f)) + \frac{\partial h_i}{\partial x_0}(x_0^k, x^{u^k, x_0^k}(t_f))\, w + \frac{\partial h_i}{\partial x_f}(x_0^k, x^{u^k, x_0^k}(t_f))\, y_f$$

for $i = 1, \cdots, r$. Then, according to Theorem 4.3.1 applied to the "direction-finding" subproblem (P_k''), there exist three constant transversality vectors $\eta^k \in \mathcal{R}^r$, $\iota^k \in \mathcal{R}^n$, $\kappa^k \in \mathcal{R}^n$, satisfying

$$\eta_i^k \geq 0 \quad \text{and} \quad \eta_i^{k\top} \pi_i^k(\bar{w}^k, y^{\bar{v}^k, \bar{w}^k}(t_f)) = 0 \tag{4.5.1}$$

for $i = 1, \cdots, r$, and

$$\iota_i^k = \kappa_i^k = 0, \qquad \forall i \notin I_{x_0}, \tag{4.5.2}$$

and

$$\iota_i^k \geq 0 \quad \text{and} \quad \iota_i^k(\bar{x}_{0_i}^k - X_i^{max}) = 0 \tag{4.5.3}$$

$$\kappa_i^k \geq 0 \quad \text{and} \quad \kappa_i^k(X_i^{min} - \bar{x}_{0_i}^k) = 0 \tag{4.5.4}$$

for all $i \in I_{x_0}$, and a costate function $q^{\bar{v}^k;\bar{w}^k}(t)$ of the "direction-finding" subproblem (P_k'') satisfying

$$
\begin{aligned}
-\dot{q}^{\bar{v}^k;\bar{w}^k}(t) &= \frac{\partial H_2}{\partial y}(y^{\bar{v}^k;\bar{w}^k}(t), q^{\bar{v}^k;\bar{w}^k}(t), \bar{v}^k(t), t) \\
q^{\bar{v}^k;\bar{w}^k}(t_0) &= -\frac{\partial K_2}{\partial w}(\bar{w}^k,\, y^{\bar{v}^k;\bar{w}^k}(t_f)) - (\iota^k - \kappa^k) - \sum_{i=1}^{r} \eta_i^k \frac{\partial}{\partial w}\pi_i^k(\bar{w}^k,\, y^{\bar{v}^k;\bar{w}^k}(t_f)) \\
q^{\bar{v}^k;\bar{w}^k}(t_f) &= \frac{\partial K_2}{\partial y_f}(\bar{w}^k,\, y^{\bar{v}^k;\bar{w}^k}(t_f)) + \sum_{i=1}^{r} \eta_i^k \frac{\partial}{\partial y_f}\pi_i^k(\bar{w}^k,\, y^{\bar{v}^k;\bar{w}^k}(t_f))
\end{aligned}
$$

such that, $\bar{u}^k(t)$ minimizes $H_2(y^{\bar{v}^k;\bar{w}^k}(t),\, q^{\bar{v}^k;\bar{w}^k}(t),\, u-u^k(t), t)$ with respect to u at any $t \in [t_0, t_f]$. In the above, H_2 is the Hamiltonian function of the "direction-finding" subproblem (P_k''). So,

$$
\begin{aligned}
& H_2(y^{\bar{v}^k;\bar{w}^k}(t), q^{\bar{v}^k;\bar{w}^k}(t), \bar{v}^k(t), t) \\
&\triangleq\ L_2(y^{\bar{v}^k;\bar{w}^k}(t), \bar{v}^k(t), t) + q^{\bar{v}^k;\bar{w}^k}(t)^{\mathsf{T}}(f_x^{(k)}(t)\, y^{\bar{v}^k;\bar{w}^k}(t) + f_u^{(k)}(t)\, \bar{v}^k(t)) \\
&=\ H_u(x^{u^k;x_0^k}(t), u^k(t), p^{u^k;x_0^k}(t), t)\, \bar{v}^k(t) + \frac{1}{2} y^{\bar{v}^k;\bar{w}^k}(t)^{\mathsf{T}} \Lambda_3^{(k)}(t) y^{\bar{v}^k;\bar{w}^k}(t) \\
&\quad + \frac{1}{2}\bar{v}^k(t)^{\mathsf{T}} \Lambda_4^{(k)}(t) \bar{v}^k(t) + q^{\bar{v}^k;\bar{w}^k}(t)^{\mathsf{T}}(f_x^{(k)}(t)\, y^{\bar{v}^k;\bar{w}^k}(t) + f_u^{(k)}(t)\, \bar{v}^k(t)) \qquad (4.5.5)
\end{aligned}
$$

and

$$
\begin{aligned}
-\dot{q}^{\bar{v}^k;\bar{w}^k}(t) &= f_x^{(k)}(t)^{\mathsf{T}} q^{\bar{v}^k;\bar{w}^k}(t) + \Lambda_3^{(k)}(t) y^{\bar{v}^k;\bar{w}^k}(t) && (4.5.6) \\
q^{\bar{v}^k;\bar{w}^k}(t_0) &= -K_{x_0}(x_0^k, x^{u^k;x_0^k}(t_f))^{\mathsf{T}} - p^{u^k;x_0^k}(t_0) - \Lambda_1^{(k)} \bar{w}^k && (4.5.7) \\
&\quad -(\iota^k - \kappa^k) - \sum_{i=1}^{r} \eta_i^k \left(\frac{\partial h_i}{\partial x_0}(x_0^k,\, x^{u^k;x_0^k}(t_f)) \right)^{\mathsf{T}} && (4.5.8) \\
q^{\bar{v}^k;\bar{w}^k}(t_f) &= \Lambda_2^{(k)} y^{\bar{v}^k;\bar{w}^k}(t_f) + \sum_{i=1}^{r} \eta_i^k \left(\frac{\partial h_i}{\partial x_f}(x_0^k,\, x^{u^k;x_0^k}(t_f)) \right)^{\mathsf{T}} && (4.5.9)
\end{aligned}
$$

and $\bar{u}^k(t)$ is the solution of

$$
\begin{aligned}
\min_{\mu \in \Omega_u} \Big\{ & H_u(x^{u^k;x_0^k}(t), u^k(t), p^{u^k;x_0^k}(t), t)(\mu - u^k(t)) \\
& + \frac{1}{2}(\mu - u^k(t))^{\mathsf{T}} \Lambda_3^{(k)}(t)(\mu - u^k(t)) \\
& + \frac{1}{2} y^{\bar{v}^k;\bar{w}^k}(t)^{\mathsf{T}} \Lambda_4^{(k)}(t) y^{\bar{v}^k;\bar{w}^k}(t) + q^{\bar{v}^k;\bar{w}^k}(t)^{\mathsf{T}}(f_x^{(k)}(t)\, y^{\bar{v}^k;\bar{w}^k}(t) + f_u^{(k)}(t)(\mu - u^k(t))) \Big\},
\end{aligned}
$$

for any $t \in [t_0, t_f]$. Equivalently, $\bar{u}^k(t)$ is the solution of the following quadratic programming problem:

$$\min_{U^{min} \le \mu \le U^{max}} \left\{ \frac{1}{2} \mu^\top \Lambda_3^{(k)}(t) \mu + b^{(k)}(t)^\top \mu \right\}$$

where

$$b^{(k)}(t) = H_u(x^{u^k; x_0^k}(t), u^k(t), p^{u^k; x_0^k}(t), t)^\top - \Lambda_3^{(k)}(t) u^k(t) + f_u^{(k)}(t)^\top q^{\bar{v}^k, \bar{w}^k}(t),$$

for any $t \in [t_0, t_f]$. By the Kuhn-Tucker Theorem, there exist two nonnegative vector functions $\beta^k(t)$ and $\gamma^k(t)$ such that,

$$\begin{aligned} & H_u(x^{u^k; x_0^k}(t), u^k(t), p^{u^k; x_0^k}(t), t)^\top \\ & \quad + \Lambda_3^{(k)}(t)(\bar{u}^k(t) - u^k(t)) + f_u^{(k)}(t)^\top q^{\bar{v}^k, \bar{w}^k}(t) + \beta^k(t) - \gamma^k(t) = 0 \end{aligned} \tag{4.5.10}$$

where

$$\beta^k(t)^\top (\bar{u}^k(t) - U^{max}) = 0 \tag{4.5.11}$$

$$\gamma^k(t)^\top (U^{min} - \bar{u}^k(t)) = 0 \tag{4.5.12}$$

for any $t \in [t_0, t_f]$.

Next, we show that $(\bar{v}^k, \bar{w}^k)$, the solution of the "direction-finding" subproblem (P_k''), turns out to be a descent direction of the exact penalty functional $\theta_\rho : \mathcal{U} \times \mathcal{S} \to \mathcal{R}$,

$$\theta_\rho(u, x_0) \triangleq J(u, x_0) + \rho \sum_{i=1}^{r} \varphi_i(u, x_0) \tag{4.5.13}$$

where

$$\varphi_i(u, x_0) \triangleq \max\{0, g_i(u, x_0)\} = \max\{0, h_i(x_0, x^{u, x_0}(t_f))\} \tag{4.5.14}$$

and ρ is a positive number.

Theorem 4.5.1 *Let $(\bar{u}^k, \bar{x}_0^k)$ be the solution of the "direction-finding" subproblem (P_k''), and $\bar{v}^k(t) = \bar{u}^k(t) - u^k(t)$, $\bar{w}^k = \bar{x}_0^k - x_0^k$. If η^k, a constant vector satisfying optimality condition (4.3.9), satisfies*

$$\|\eta^k\| \le \rho, \tag{4.5.15}$$

and if $(\bar{v}^k, \bar{w}^k) \neq 0$, *then there exists a* $\bar{\lambda}^k \in (0, \infty)$, *such that, for all* $0 < \lambda \leq \bar{\lambda}^k$,

$$\begin{aligned} &\theta_\rho(u^k + \lambda\bar{v}^k, x_0^k + \lambda\bar{w}^k) - \theta_\rho(u^k, x_0^k) \\ &\quad \leq -\frac{\lambda}{2}(\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k - \frac{\lambda}{2}\int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt \;<\; 0 \end{aligned} \tag{4.5.16}$$

Proof: Let

$$I^{(k)} \triangleq \{\, i :\; g_i(u^k, x_0^k) > 0 \,\} = \{\, i :\; h_i(x_0^k, x^{u^k, x_0^k}(t_f)) > 0 \,\},$$

$$\bar{I}^{(k)} \triangleq \{\, i :\; g_i(u^k, x_0^k) = 0 \,\} = \{\, i :\; h_i(x_0^k, x^{u^k, x_0^k}(t_f)) = 0 \,\},$$

$$\hat{I}^{(k)} \triangleq \{\, i :\; g_i(u^k, x_0^k) < 0 \,\} = \{\, i :\; h_i(x_0^k, x^{u^k, x_0^k}(t_f)) < 0 \,\}.$$

From Lemma A.3, the mean-value property,

$$\begin{aligned} &g_i(u^k + \lambda\bar{v}^k, x_0^k + \lambda\bar{w}^k) - g_i(u^k, x_0^k) \\ &\quad = h_i(x_0^k + \lambda\bar{w}^k, x^{u^k+\lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k}(t_f)) - h_i(x_0^k, x^{u^k, x_0^k}(t_f)) \\ &\quad = \frac{\partial h_i}{\partial x_0}(x_0^k, x^{u^k, x_0^k}(t_f))(\lambda\bar{w}^k) + \frac{\partial h_i}{\partial x_f}(x_0^k, x^{u^k, x_0^k}(t_f))(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k, x_0^k}(t_f)) \\ &\qquad + \int_0^1 (1-\tau)\frac{\partial^2 h_i}{\partial x_0^2}(\bar{x}_0(\tau, t_0), \bar{x}(\tau, t_f))\,d\tau\,(\lambda\bar{w}^k)^2 \\ &\qquad + \int_0^1 (1-\tau)\frac{\partial^2 h_i}{\partial x_f^2}(\bar{x}_0(\tau, t_0), \bar{x}(\tau, t_f))\,d\tau\,(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k, x_0^k}(t_f))^2 \end{aligned}$$

where $\bar{x}(\tau, t_0) = x_0^k + \tau(\lambda\bar{w}^k)$, $\bar{x}(\tau, t_f) = x^{u^k, x_0^k}(t_f) + \tau(x^{u^k+\lambda\bar{v}^k}(t_f) - x^{u^k, x_0^k}(t_f))$. Because $x_0^k, x_0^k + \lambda\bar{w}^k \in \mathcal{S}$, $\bar{w}^k$ is bounded implying $\bar{x}_0(\tau, t_0)$ is bounded for any $\tau \in [0, 1]$ and any $\lambda \in [0, 1]$. Furthermore, from Lemma B.1, because $u^k, u^k + \lambda\bar{v}^k \in \mathcal{U}$, $x_0^k, x_0^k + \lambda\bar{w}^k \in \mathcal{S}$, for any $\lambda \in [0, 1]$, both $x^{u^k, x_0^k}(t_f)$ and $x^{u^k+\lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k}(t_f)$ are bounded. So, $\bar{x}(\tau, t_f)$ is also bounded, for any $\tau \in [0, 1]$ and any $\lambda \in [0, 1]$. Hence, by using in addition the continuity of $\frac{\partial h_i}{\partial x_f}$, $\frac{\partial^2 h_i}{\partial x_0^2}$, and $\frac{\partial^2 h_i}{\partial x_f^2}$, there exist $c_1', c_2', c_3', c_4' > 0$, such that,

$$\begin{aligned} &g_i(u^k + \lambda\bar{v}^k, x_0^k + \lambda\bar{w}^k) - g_i(u^k, x_0^k) \\ &\quad \leq \frac{\partial h_i}{\partial x_0}(x_0^k\; x^{u^k, x_0^k}(t_f))\,(\lambda\bar{w}^k) + \frac{\partial h_i}{\partial x_f}(x_0^k\; x^{u^k, x_0^k}(t_f))\, y^{\lambda\bar{v}^k, \lambda\bar{w}^k}(t_f) \\ &\qquad + c_1' \,\|\, x^{u^k+\lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k}(t_f) - x^{u^k, x_0^k}(t_f) - y^{\lambda\bar{v}^k, \lambda\bar{w}^k}(t_f) \,\| \\ &\qquad + c_2' \,\|\, \lambda\bar{w}^k \,\|^2 + c_3' \,\|\, \lambda\bar{v}^k \,\| \cdot \|\, \lambda\bar{w}^k \,\| + c_4' \,\|\, \lambda\bar{v}^k \,\|^2 \end{aligned}$$

$$\leq \frac{\partial h_i}{\partial x_0}(x_0^k\, x^{u^k,x_0^k}(t_f))\,(\lambda\bar{w}^k) + \frac{\partial h_i}{\partial x_f}(x_0^k\, x^{u^k,x_0^k}(t_f))\, y^{\lambda\bar{v}^k,\lambda\bar{w}^k}(t_f)$$
$$+ c_1\lambda^2 \| \bar{w}^k \|^2 + c_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{w}^k \| + c_3\lambda^2 \| \bar{v}^k \|^2.$$

The last inequality comes from Lemma B.3, with some $c_1, c_2, c_3 > 0$. Because $y^{\lambda\bar{v}^k,\lambda\bar{w}^k}$ is a solution of the linearized system (4.4.17)-(4.4.18) with input $\lambda\bar{v}^k$ and initial state $\lambda\bar{w}^k$, $y^{\lambda\bar{v}^k,\lambda\bar{w}^k} = \lambda y^{\bar{v}^k,\bar{w}^k}$. Moreover, from the linearized end-point inequality constraints (4.4.10), we have

$$\frac{\partial h_i}{\partial x_0}(x_0^k\, x^{u^k,x_0^k}(t_f))\,\bar{w}^k + \frac{\partial h_i}{\partial x_f}(x_0^k\, x^{u^k,x_0^k}(t_f))\, y^{\bar{v}^k,\bar{w}^k}(t_f)$$
$$\leq -h_i(x_0^k\, x^{u^k,x_0^k}(t_f)) = -g_i(u^k, x_0^k).$$

Therefore,

$$g_i(u^k+\lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k) - g_i(u^k, x_0^k)$$
$$\leq -\lambda g_i(u^k, x_0^k) + c_1\lambda^2 \| \bar{w}^k \|^2 + c_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{w}^k \| + c_3\lambda^2 \| \bar{v}^k \|^2. \quad (4.5.17)$$

(1) For any $i \in I^{(k)}$. Because $g_i(u, x_0)$ is continuous at any $u \in \mathcal{U}$, $x_0 \in \mathcal{S}$, according to Lemma B.4, there exists a $\bar{\lambda}_1 > 0$, such that, for all $\lambda \in [0, \bar{\lambda}_1]$, $g_i(u^k+\lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k) > 0$. Then, from (4.5.17),

$$\varphi_i(u^k + \lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k) - \varphi_i(u^k, x_0^k)$$
$$= g_i(u^k + \lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k) - g_i(u^k, x_0^k)$$
$$\leq -\lambda g_i(u^k, x_0^k) + c_1\lambda^2 \| \bar{w}^k \|^2 + c_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{w}^k \| + c_3\lambda^2 \| \bar{v}^k \|^2.$$

(2) For any $i \in \bar{I}^{(k)}$. From (4.5.17),

$$g_i(u^k + \lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k)$$
$$\leq (1-\lambda)\underbrace{g_i(u^k, x_0^k)}_{=0} + c_1\lambda^2 \| \bar{w}^k \|^2 + c_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{w}^k \| + c_3\lambda^2 \| \bar{v}^k \|^2$$
$$= c_1\lambda^2 \| \bar{w}^k \|^2 + c_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{w}^k \| + c_3\lambda^2 \| \bar{v}^k \|^2$$

we then have,

$$\varphi_i(u^k + \lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k) - \varphi_i(u^k, x_0^k)$$
$$= \max\{0, g_i(u^k + \lambda\bar{v}^k, x_0^k+\lambda\bar{w}^k)\} - 0$$
$$\leq c_1\lambda^2 \| \bar{w}^k \|^2 + c_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{w}^k \| + c_3\lambda^2 \| \bar{v}^k \|^2.$$

(3) For any $i \in \hat{I}^{(k)}$. Because $g_i(u, x_0)$ is continuous at any $u \in \mathcal{U}$, $x_0 \in \mathcal{S}$, according to Lemma B.4, there exists a $\bar{\lambda}_2 > 0$, such that, for all $\lambda \in [0, \bar{\lambda}_2]$, $g_i(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) < 0$. Then,

$$\begin{aligned}
&\varphi_i(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - \varphi_i(u^k;\, x_0^k) \\
&\quad = \max\{\,0,\, g_i(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k)\,\} - \max\{\,0,\, g_i(u^k;\, x_0^k)\,\} \;=\; 0 - 0.
\end{aligned}$$

Therefore, there exist $c_1', c_2', c_3' > 0$ such that,

$$\begin{aligned}
&\theta_\rho(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - \theta_\rho(u^k;\, x_0^k) \\
&\quad = J(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - J(u^k;\, x_0^k) + \rho \sum_{i=1}^{r} (\varphi_i(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - \varphi_i(u^k;\, x_0^k)) \\
&\quad \le J(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - J(u^k;\, x_0^k) - \rho\lambda \sum_{i \in I^{(k)}} g_i(u^k;\, x_0^k) \\
&\qquad + c_1' \rho\lambda^2 \,\|\, \bar{w}^k \,\|^2 + c_2' \rho\lambda^2 \,\|\, \bar{v}^k \,\| \cdot \|\, \bar{w}^k \,\| + c_3' \rho\lambda^2 \,\|\, \bar{v}^k \,\|^2. \qquad (4.5.18)
\end{aligned}$$

From the proof of Lemma B.5 and (4.4.6) of Proposition 4.4.1, there exist $c_1'', c_2'', c_3'' > 0$, such that,

$$\begin{aligned}
&J(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - J(u^k;\, x_0^k) \\
&\quad \le \lambda \Big(K_{x_0}(x_0^k, x^{u^k, x_0^k}(t_f)) + p^{u^k, x_0^k}(t_0)^\top \Big) \bar{w}^k + \lambda \int_{t_0}^{t_f} H_u(x^k(t), u^k(t), p^k(t), t)\bar{v}^k(t)\, dt \\
&\qquad + c_1'' \lambda^2 \,\|\, \bar{w}^k \,\|^2 + c_2'' \lambda^2 \,\|\, \bar{v}^k \,\| \cdot \|\, \bar{w}^k \,\| + c_3'' \lambda^2 \,\|\, \bar{v}^k \,\|^2.
\end{aligned}$$

From (4.5.10),

$$\begin{aligned}
&J(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - J(u^k;\, x_0^k) \\
&\quad \le \lambda \Big(K_{x_0}(x_0^k, x^{u^k, x_0^k}(t_f)) + p^{u^k, x_0^k}(t_0)^\top \Big) \bar{w}^k - \lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\, dt \\
&\qquad - \lambda \int_{t_0}^{t_f} q^{\bar{v}^k, \bar{w}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\, dt - \lambda \int_{t_0}^{t_f} \beta^k(t)^\top \bar{v}^k(t)\, dt + \lambda \int_{t_0}^{t_f} \gamma^k(t)^\top \bar{v}^k(t)\, dt \\
&\qquad + c_1'' \lambda^2 \,\|\, \bar{w}^k \,\|^2 + c_2'' \lambda^2 \,\|\, \bar{v}^k \,\| \cdot \|\, \bar{w}^k \,\| + c_3'' \lambda^2 \,\|\, \bar{v}^k \,\|^2.
\end{aligned}$$

Because

$$\begin{aligned}
\beta^k(t)^\top \bar{v}^k(t) \;&=\; \beta^k(t)^\top (\bar{u}^k(t) - u^k(t)) \\
&=\; \underbrace{\beta^k(t)^\top (\bar{u}^k(t) - U^{max})}_{=0} + \underbrace{\beta^k(t)}_{\ge 0}^\top \underbrace{(U^{max} - u^k(t))}_{\ge 0} \;\ge\; 0, \qquad (4.5.19)
\end{aligned}$$

and

$$\begin{aligned}\gamma^k(t)^\top \bar{v}^k(t) &= \gamma^k(t)^\top(\bar{u}^k(t) - u^k(t)) \\ &= \underbrace{\gamma^k(t)^\top(\bar{u}^k(t) - U^{min})}_{=0} + \underbrace{\gamma^k(t)}_{\geq 0}^\top \underbrace{(U^{min} - u^k(t))}_{\leq 0} \leq 0, \qquad (4.5.20)\end{aligned}$$

we then have,

$$\begin{aligned}&J(u^k + \lambda\bar{v}^k;\, x_0^k + \lambda\bar{w}^k) - J(u^k;\, x_0^k) \\ &\leq \lambda\Big(K_{x_0}(x_0^k, x^{u^k;x_0^k}(t_f)) + p^{u^k;x_0^k}(t_0)^\top\Big)\bar{w}^k \\ &\quad -\lambda\int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt - \lambda\int_{t_0}^{t_f} q^{\bar{v}^k;\bar{w}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\,dt \\ &\quad + c_1''\lambda^2 \parallel \bar{w}^k \parallel^2 + c_2''\lambda^2 \parallel \bar{v}^k \parallel \cdot \parallel \bar{w}^k \parallel + c_3''\lambda^2 \parallel \bar{v}^k \parallel^2. \qquad (4.5.21)\end{aligned}$$

Because $(\bar{u}^k;\, \bar{w}^k)$ is an optimal pair for the "direction-finding" subproblem (P_k''), $y^{\bar{v}^k;\bar{w}^k}(t)$ satisfies (4.4.17) and (4.4.18) with input $\bar{v}^k(t) = \bar{u}^k(t) - u^k(t)$ and initial state $\bar{w}^k = \bar{x}_0^k - x_0^k$, and the corresponding costate function $q^{\bar{v}^k;\bar{w}^k}(t)$ satisfies (4.5.6), (4.5.7) and (4.5.9) with three constant vectors η^k, ι^k, κ^k satisfying (4.5.1), (4.5.2), (4.5.3) and (4.5.4). We then have,

$$\begin{aligned}\frac{d}{dt}\Big(q^{\bar{v}^k;\bar{w}^k}(t)^\top y^{\bar{v}^k;\bar{w}^k}(t)\Big) &= q^{\bar{v}^k;\bar{w}^k}(t)^\top \underbrace{\dot{y}^{\bar{v}^k}(t)}_{(4.4.17)} + y^{\bar{v}^k;\bar{w}^k}(t)^\top \underbrace{\dot{q}^{\bar{v}^k}(t)}_{(4.5.6))} \\ &= q^{\bar{v}^k;\bar{w}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t) - y^{\bar{v}^k;\bar{w}^k}(t)^\top \Lambda_2^{(k)}(t) y^{\bar{v}^k;\bar{w}^k}(t).\end{aligned}$$

Therefore,

$$\begin{aligned}&\int_{t_0}^{t_f} q^{\bar{v}^k;\bar{w}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\,dt \\ &= \underbrace{q^{\bar{v}^k;\bar{w}^k}(t_f)}_{(4.5.9)}^\top y^{\bar{v}^k;\bar{w}^k}(t_f) - \underbrace{q^{\bar{v}^k;\bar{w}^k}(t_0)}_{(4.5.7)}^\top \underbrace{y^{\bar{v}^k;\bar{w}^k}(t_0)}_{=\bar{w}^k} + \int_{t_0}^{t_f} \underbrace{y^{\bar{v}^k;\bar{w}^k}(t)^\top \Lambda_2^{(k)}(t) y^{\bar{v}^k;\bar{w}^k}(t)}_{\geq 0}\,dt \\ &\geq \underbrace{y^{\bar{v}^k;\bar{w}^k}(t_f)^\top \Lambda_2^{(k)} y^{\bar{v}^k;\bar{w}^k}(t_f)}_{\geq 0} + (\bar{w}^k)^\top \Lambda_1^{(k)}\bar{w}^k + \Big(K_{x_0}(x_0^k, x^{u^k;x_0^k}(t_f)) + p^{u^k;x_0^k}(t_0)^\top\Big)\bar{w}^k \\ &\quad + \sum_{i=1}^{r} \eta_i^k \underbrace{\Big(\frac{\partial h_i}{\partial x_0}(x_0^k,\, x^{u^k;x_0^k}(t_f))\bar{w}^k + \frac{\partial h_i}{\partial x_f}(x^{u^k;x_0^k}(t_f), t_f) y^{\bar{v}^k;\bar{w}^k}(t_f)\Big)}_{(4.5.1)} + (\iota^k - \kappa^k)^\top \bar{w}^k\end{aligned}$$

$$\geq \left(K_{x_0}(x_0^k, x^{u^k;x_0^k}(t_f)) + p^{u^k;x_0^k}(t_0)^\top\right)\bar{w}^k - \sum_{i=1}^{r} \eta_i^k h_i(x_0^k, x^{u^k;x_0^k}(t_f))$$
$$+ (\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k + (\iota^k - \kappa^k)^\top \bar{w}^k. \tag{4.5.22}$$

From (4.2.5),

$$\int_{t_0}^{t_f} q^{\bar{v}^k;\bar{w}^k}(t)^\top f_u^{(k)}(t)\bar{v}^k(t)\,dt$$
$$\geq \left(K_{x_0}(x_0^k, x^{u^k;x_0^k}(t_f)) + p^{u^k;x_0^k}(t_0)^\top\right)\bar{w}^k - \sum_{i\in I^{(k)}} \eta_i^k g_i(u^k, x_0^k)$$
$$- \sum_{i\in I^{(k)}} \eta_i^k \underbrace{g_i(u^k, x_0^k)}_{=0} - \sum_{i\in \hat{I}^{(k)}} \underbrace{\eta_i^k}_{\geq 0} \underbrace{g_i(u^k, x_0^k)}_{<0} + (\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k + (\iota^k - \kappa^k)^\top \bar{w}^k$$
$$\geq \left(K_{x_0}(x_0^k, x^{u^k;x_0^k}(t_f)) + p^{u^k;x_0^k}(t_0)^\top\right)\bar{w}^k - \sum_{i\in I^{(k)}} \eta_i^k g_i(u^k, x_0^k)$$
$$+ (\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k + (\iota^k - \kappa^k)^\top \bar{w}^k. \tag{4.5.23}$$

Combining (4.5.18), (4.5.21) and (4.5.23), and letting $\bar{c}_1 = c_1'\rho + c_1''$, $\bar{c}_2 = c_2'\rho + c_2''$, we then have

$$\theta_\rho(u^k + \lambda\bar{v}^k, x_0^k + \lambda\bar{w}^k) - \theta_\rho(u^k, x_0^k)$$
$$\leq -\lambda(\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k - \lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt - \lambda(\iota^k - \kappa^k)^\top \bar{w}^k$$
$$+ \lambda \sum_{i\in I^{(k)}} \underbrace{(\eta_i^k - \rho)}_{\leq 0} \underbrace{g_i(u^k, x_0^k)}_{>0} + \bar{c}_1\lambda^2 \| \bar{w}^k \|^2 + \bar{c}_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{v}^k \| + \bar{c}_3\lambda^2 \| \bar{v}^k \|^2$$
$$\leq -\lambda(\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k - \lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt - \lambda(\iota^k - \kappa^k)^\top \bar{w}^k$$
$$+ \bar{c}_1\lambda^2 \| \bar{w}^k \|^2 + \bar{c}_2\lambda^2 \| \bar{v}^k \| \cdot \| \bar{v}^k \| + \bar{c}_3\lambda^2 \| \bar{v}^k \|^2.$$

Because

$$(\iota^k)^\top \bar{w}^k = (\iota^k)^\top (\bar{x}_0^k - x_0^k)$$
$$= \underbrace{(\iota^k)^\top (\bar{x}_0^k - X^{max})}_{=0} + (\underbrace{\iota^k}_{\geq 0})^\top (\underbrace{X^{max} - x_0^k}_{\geq 0}) \geq 0, \tag{4.5.24}$$

and

$$(\kappa^k)^\top \bar{w}^k = (\kappa^k)^\top (\bar{x}_0^k - x_0^k)$$
$$= \underbrace{(\kappa^k)^\top (\bar{x}_0^k - X^{min})}_{=0} + (\underbrace{\kappa^k}_{\geq 0})^\top (\underbrace{X^{min} - x_0^k}_{\leq 0}) \leq 0, \tag{4.5.25}$$

we finally have

$$\begin{aligned}
&\theta_\rho(u^k+\lambda\bar{v}^k,\, x_0^k+\lambda\bar{w}^k) - \theta_\rho(u^k,\, x_0^k) \\
&\quad \le -\lambda\,(\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k - \lambda \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt \\
&\qquad + \bar{c}_1\lambda^2 \,\|\, \bar{w}^k \,\|^2 + \bar{c}_2\lambda^2 \,\|\, \bar{v}^k \,\| \cdot \|\, \bar{v}^k \,\| + \bar{c}_3\lambda^2 \,\|\, \bar{v}^k \,\|^2. \qquad (4.5.26)
\end{aligned}$$

Because the constant matrix $\Lambda_1^{(k)}$ is positive definite, and the time-varying matrix $\Lambda_3^{(k)}(t)$ is positive definite for any $t \in [t_0, t_f]$, there exists a small enough $\bar{\lambda}^k \in (0, \infty)$, such that, whenever $(\bar{v}^k,\, \bar{w}^k) \neq 0$,

$$\begin{aligned}
&\theta_\rho(u^k+\lambda\bar{v}^k,\, x_0^k+\lambda\bar{w}^k) - \theta_\rho(u^k,\, x_0^k) \\
&\quad \le -\frac{\lambda}{2}(\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k - \frac{\lambda}{2} \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt \;<\; 0 \qquad (4.5.27)
\end{aligned}$$

for all $0 < \lambda \le \bar{\lambda}^k$. □

Remark If the initial state x_0 is not allowed to vary, the above Theorem 4.5.1 is reduced to Theorem 3.5.1 shown in Chapter 3.

Remark If the control u is not allowed to vary, the problem considered in this chapter is reduced to the special problem of finding the optimal initial state of a constrained dynamical system, which is itself a useful practical problem. After an easy specialization of the algorithm described in this chapter, it can be used to solve that problem as well. Please see section 4.10

Remark Theorem 4.5.1 shows that, whenever $(\bar{u}^k,\, \bar{w}^k) \neq (u^k,\, w^k)$, $(\bar{v}^k,\, \bar{w}^k) = (\bar{u}^k - u^k,\, \bar{w}^k - w^k)$ is always a descent direction of the exact penalty functional $\theta_\rho(u, x_0)$ at the k-th iteration.

Remark The descent property shown in Theorem 4.5.1 will always hold, as long as matrix Λ_1 is positive definite, Λ_2 is semi-positive definite, $\Lambda_3(t)$ is semi-positive definite for all $t \in [t_0, t_f]$, and $\Lambda_4(t)$ is positive definite for all $t \in [t_0, t_f]$, regardless of whether the inequality relationships (4.4.11)-(4.4.12) are satisfied. However, those

inequalities are crucial for the rate of convergence of the algorithm. Intuitively, tighter approximations by (4.4.11)-(4.4.12) make the rate of convergence of the algorithm closer to second-order; while looser approximations would destroy the second-order properties and make the algorithm behave more like a first-order algorithm.

4.6 Stepsize Rules

Similar to the discussion in Chapter 3, at each iteration k, we would like to perform a line search on the exact penalty function $\theta_\rho(u, x_0)$, starting from $(u, x_0) = (u^k, x_0^k)$ in the direction $(\bar{v}^k, \bar{w}^k)$. The following Armijo-like line search is adopted: find the smallest integer $l_k \geq 0$ such that

$$\theta_\rho(u^k + \zeta^{l_k}\bar{v}^k, x_0^k + \zeta^{l_k}\bar{w}^k) - \theta_\rho(u^k, x_0^k) \leq -\alpha\zeta^{l_k}R^k \tag{4.6.1}$$

where

$$R^k = (\bar{w}^k)^\top \Lambda_1^{(k)} \bar{w}^k + \int_{t_0}^{t_f} \bar{v}^k(t)^\top \Lambda_3^{(k)}(t)\bar{v}^k(t)\,dt \tag{4.6.2}$$

and α, ζ are two parameters chosen a priori in $(0,1)$. ζ^{l_k} is then called an Armijo stepsize. Clearly, whenever $(\bar{v}^k, \bar{w}^k) \neq 0$, the above R^k is always positive, because $\Lambda_1^{(k)}$ is always a positive definite matrix, and $\Lambda_3^{(k)}(t)$ is always a positive definite matrix for any $t \in [t_0, t_f]$. According to Theorem 4.5.1, whenever $(\bar{v}^k, \bar{w}^k) \neq 0$, there exists a $\bar{\lambda}^k \in (0, \infty)$, such that, for all $0 < \lambda \leq \bar{\lambda}^k$,

$$\theta_\rho(u^k + \lambda\bar{v}^k) - \theta_\rho(u^k) \leq -\frac{\lambda}{2}R^k < 0.$$

So, whenever $(\bar{v}^k, \bar{w}^k) \neq 0$, the above Armijo-like line search is always well-defined, and the stepsize $\lambda^k = \zeta^{l_k}$ is always a positive number.

4.7 Special Case: Free End-Time Optimal Control Problem

It will be seen in this section that a generally constrained free end-time optimal control problem can be easily converted to a constrained fixed end-time problem with two more state variables, one of which has a fixed initial value and the other of which has an

initial value which is allowed to vary within an interval. To see this, let us consider the following dynamical system defined on interval $[t_0, t_f]$ where t_f can vary within $[T_1, T_2]$ for some $T_2 > T_1 \geq t_0$:

$$\dot{x}(t) = f(x(t), u(t), t), \tag{4.7.1}$$

$$x(t_0) = x_0. \tag{4.7.2}$$

There are n_{x_0} components of the initial state vector x_0 which are allowed to vary within a constraint box, while the remaining $n - n_{x_0}$ components are fixed. That is, there is an index set $I_{x_0} \subset \{ 1, \cdots, n \}$ such that

$$x_0 \in \mathcal{S} = \{ \nu \in \mathcal{R}^n \mid X_i^{min} \leq \nu_i \leq X_i^{max},\ i \in I_{x_0};\ \nu_i = x_{0i}(\text{fixed}),\ i \notin I_{x_0} \}. \tag{4.7.3}$$

The dynamical system (4.7.1)-(4.7.2) is also subject to the control constraints

$$u(t) \in \Omega = \{ \mu \in \mathcal{R}^m \mid U_i^{min} \leq \mu_i \leq U_i^{max},\ i = 1, \cdots, m \}, \tag{4.7.4}$$

for any $t \in [t_0, t_f]$, and the end-point inequality constraints,

$$g_i(u, x_0) = h_i(x_0, x(t_f)) \leq 0, \qquad i = 1, \cdots, r. \tag{4.7.5}$$

Let the set of admissible controls be,

$$\mathcal{U} = \{ u \mid u : [t_0, t_f] \to \Omega \text{ is continuous a.e. } \} \subset \mathcal{L}_\infty^m[t_0, t_f], \tag{4.7.6}$$

and let $\tilde{\mathcal{U}}$ be the set of equivalence classes of functions in $\mathcal{U}$ which are equal almost everywhere on $[t_0, t_f]$. Let the set of combined feasible controls, feasible initial states and feasible final time be

$$\mathcal{F} \triangleq \{ (u, x_0, t_f) \mid u \in \tilde{\mathcal{U}},\ x_0 \in \mathcal{S},\ t_f \in [T_1, T_2],\ g_i(u, x_0) \leq 0,\ i = 1, \cdots, r \}. \tag{4.7.7}$$

We consider the following free end-time constrained optimal control problem:

Problem ($P_{free-end-time}$). Subject to the dynamical system (4.7.1)-(4.7.2), find a control $u \in \mathcal{U}$, an initial state $x_0 \in \mathcal{S}$, and a final time $t_f > t_0$, such that the cost functional

$$J(u, x_0, t_f) = K(x_0, x(t_f)) + \int_{t_0}^{t_f} L(x(\tau), u(\tau), \tau)\, d\tau \tag{4.7.8}$$

is minimized over $\mathcal{F}$.

In what follows, a transformation due to G. Leitmann [77,118] is used. Let τ be the new time variable defined on $[0,1]$, and the old time variable t be a function of τ, that is, $t=t(\tau)$. Introduce two additional state variables

$$\alpha(\tau) = t(\tau) \quad \text{and} \quad \beta(\tau) = \frac{dt(\tau)}{d\tau}. \tag{4.7.9}$$

Introduce a new state vector $z(\tau)=(x(\tau)^\top, \alpha(\tau), \beta(\tau))^\top \in R^{n+2}$. Then, one has the following new dynamical system defined on the fixed end-time interval $[0,1]$:

$$\dot{x}(\tau) = \beta(\tau) \cdot f(x(\tau), u(\tau), \tau) \tag{4.7.10}$$

$$\dot{\alpha}(\tau) = \beta(\tau) \tag{4.7.11}$$

$$\dot{\beta}(\tau) = 0 \tag{4.7.12}$$

and its initial state is

$$z(0)=(x_0^\top, \alpha(0), \beta(0))^\top. \tag{4.7.13}$$

So, the above dynamical system can be equivalently expressed as the following,

$$\dot{z}(\tau) = f_z(z(\tau), u(\tau), \tau), \tag{4.7.14}$$

$$z(0) = z_0. \tag{4.7.15}$$

The new end-point inequality constraints become

$$g_{z_i}(u, z_0) = g_i(u, x_0) \leq 0, \qquad i = 1, \cdots, r \tag{4.7.16}$$

and

$$T_1 \leq \alpha(1) \leq T_2. \tag{4.7.17}$$

The new initial state constraints become

$$z_0 \in \mathcal{S}_z = \{\ (x_0, \alpha(0), \beta(0)) \in \mathcal{R}^{n+2} \mid x_0 \in \mathcal{S};\ \alpha(0)=t_0\,(\text{fixed});\ \beta(0)\in[0, Z]\ \} \tag{4.7.18}$$

for a big enough real number Z. Also, the new feasible set becomes

$$\mathcal{F}_z \triangleq \{\ (u,\ z_0) \mid u \in \tilde{\mathcal{U}},\ z_0 \in \mathcal{S}_z,\ g_{z_i}(u, z_0) \leq 0,\ i = 1, \cdots, r;\ T_1 \leq \alpha(1) \leq T_2\ \}. \tag{4.7.19}$$

Then, the above free end-time constrained optimal control problem is converted into the following fixed end-time constrained optimal control problem:

Problem ($P_{fixed-end-time}$). Subject to the dynamical system (4.7.14)-(4.7.15), find a control $u \in \mathcal{U}_z$ and an initial state $z_0 \in \mathcal{S}_z$, such that the cost functional

$$\begin{aligned} J(u, z_0) &= K_z(z_0,\ z^{u, x_0}(1)) + \int_0^1 L_z(z^{u, z_0}(\tau), u(\tau), \tau)\, d\tau \\ &= K(x_0,\ x(1)) + \int_0^1 \beta(\tau) \cdot L(x(\tau), u(\tau), \tau)\, d\tau \end{aligned} \tag{4.7.20}$$

is minimized over $\mathcal{F}_z$.

It is clear now that Algorithm 2, described in sections 4.4 and 4.5, can be used to solve constrained free end-time optimal control problem as well.

4.8 Special Case: Optimal Control Problems with Path Constraints

Let us consider the following path constraints,

$$\phi_i(x(t), t) \leq 0, \qquad i = 1, \cdots, l, \quad \forall t \in [t_0, t_f], \tag{4.8.1}$$

where the time interval $[t_0, t_f]$ is fixed. Define:

$$\mathcal{L}_\epsilon(\xi) = \begin{cases} (\xi + \epsilon)^2/4\epsilon & \text{when } \xi \geq -\epsilon \\ 0 & \text{when } \xi < -\epsilon \end{cases} \tag{4.8.2}$$

Introduce l additional state variables: $x_{n+i}(t)$, $i = 1, \cdots, l$, be such that

$$\dot{x}_{n+i}(t) = \mathcal{L}_\epsilon\Big(\phi_l(x(t), t)\Big), \qquad i = 1, \cdots, l, \quad \forall t \in [t_0, t_f].$$

Then, the above state path constraints can be well **approximated** by the following end-point constraints:

$$x_{n+i}(t_f) - \gamma \leq 0, \qquad i = 1, \cdots, l,$$

if the two positive constants ϵ and γ are chosen appropriately.

It is clear now that Algorithm 2, described in sections 4.4 and 4.5, can be used to solve optimal control problems with path constraints.

4.9 Special Case: Optimize a Constrained Optimal Control Problem over Some Design Parameters

In practice, there is often a need to optimize a constrained dynamical system over not only its control and initial state, but also over a number of design parameters. It will be seen in this section that Algorithm 2, described in sections 4.4 and 4.5, can be used to solve that problem as well. To see this, let us consider the following dynamical system defined on a fixed end-time interval $[t_0, t_f]$:

$$\dot{x}^{u,x_0,p}(t) = f(x^{u,x_0,p}(t), p, u(t), t), \tag{4.9.1}$$

$$x^{u,x_0,p}(t_0) = x_0. \tag{4.9.2}$$

In the above, $p \in \mathcal{R}^{n_p}$ is a vector of design parameters which are constrained as follows,

$$p \in \mathcal{S}_p = \{ \rho \in \mathcal{R}^{n_p} \mid P_i^{min} \leq \rho_i \leq P_i^{max},\ i = 1, \cdots, n_p \} \tag{4.9.3}$$

where $\mathcal{S}_p$ is compact. Also, there are n_{x_0} components of the initial state vector x_0 which are allowed to vary within a constraint box, while the remaining $n - n_{x_0}$ components are fixed. That is, there is an index set $I_{x_0} \subset \{ 1, \cdots, n \}$ such that

$$x_0 \in \mathcal{S}_{x_0} = \{ \nu \in \mathcal{R}^n \mid X_i^{min} \leq \nu_i \leq X_i^{max},\ i \in I_{x_0};\ \nu_i = x_{0i}\,(\text{fixed}),\ i \notin I_{x_0} \} \tag{4.9.4}$$

where $\mathcal{S}_{x_0}$ is compact. The dynamical system (4.9.1)-(4.9.2) is also subject to the control constraints

$$u(t) \in \Omega = \{ \mu \in \mathcal{R}^m \mid U_i^{min} \leq \mu_i \leq U_i^{max},\ i = 1, \cdots, m \} \tag{4.9.5}$$

where Ω is compact, and the end-point inequality constraints,

$$g_i(u, x_0, p) = h_i(x_0, p, x^{u,x_0,p}(t_f)) \leq 0, \qquad i = 1, \cdots, r. \tag{4.9.6}$$

In the above, $x^{u,x_0,p}(t) \in \mathcal{R}^n$ is the state of the system at time $t \in [t_0, t_f]$, which corresponds to the control $u(t) \in \mathcal{R}^m$, the initial value of the state $x_0 \in \mathcal{R}^n$, and a vector of design parameters $p \in \mathcal{R}^{n_p}$. Let the set of admissible controls be,

$$\mathcal{U} = \{ u \mid u : [t_0, t_f] \to \Omega \text{ is continuous a.e. } \} \subset \mathcal{L}_\infty^m[t_0, t_f], \tag{4.9.7}$$

and let $\tilde{\mathcal{U}}$ be the set of equivalence classes of functions in $\mathcal{U}$ which are equal almost everywhere on $[t_0, t_f]$. Let the combined set of feasible controls and feasible initial states be

$$\mathcal{F} \triangleq \{ (u, x_0, p) \mid u \in \tilde{\mathcal{U}},\ x_0 \in \mathcal{S}_{x_0},\ p \in \mathcal{S}_p,\ g_i(u, x_0, p) \leq 0,\ i = 1, \cdots, r \}. \tag{4.9.8}$$

We may now formulate a constrained optimal control problem as follows:

Problem ($P_{design-parameters}$). Subject to the dynamical system (4.9.1)-(4.9.2), find a control $u \in \mathcal{U}$, an initial state $x_0 \in \mathcal{S}_{x_0}$, and a vector of design parameters $p \in \mathcal{S}_p$, such that the cost functional

$$J(u, x_0, p) = K(x_0,\ p,\ x^{u,x_0,p}(t_f)) + \int_{t_0}^{t_f} L(x^{u,x_0,p}(\tau), p, u(\tau), \tau)\, d\tau \tag{4.9.9}$$

is minimized over $\mathcal{F}$.

Introduce a new vector $z(t) = (x(t)^\top,\ \hat{x}(t)^\top)^\top \in R^{n+n_p}$, where $\hat{x}(t) \equiv p \in \mathcal{R}^{n_p}$, for any $t \in [t_0, t_f]$. Then, one has the following new dynamical system defined on $[t_0, t_f]$:

$$\dot{x}(t) = f(x(t), \hat{x}(t), u(t), t), \tag{4.9.10}$$

$$\dot{\hat{x}}(t) = 0 \tag{4.9.11}$$

and its initial state is

$$z(0) = (x_0^\top,\ p^\top)^\top. \tag{4.9.12}$$

So, the above dynamical system can be equivalently expressed as the following,

$$\dot{z}^{u,z_0}(t) = f_z(z^{u,z_0}(t), u(t), t), \tag{4.9.13}$$

$$z^{u,z_0}(0) = z_0. \tag{4.9.14}$$

The new end-point inequality constraints become

$$g_{z_i}(u, z_0) = g_i(u, x_0, p) \leq 0, \qquad i = 1, \cdots, r. \tag{4.9.15}$$

The new initial state constraints become

$$z_0 \in \mathcal{S}_z = \{ (x_0, p) \in \mathcal{R}^{n+n_p} \mid x_0 \in \mathcal{S}_{x_0};\ p \in \mathcal{S}_p \}. \tag{4.9.16}$$

Also, the new feasible set becomes

$$\mathcal{F}_z \triangleq \{ (u, z_0) \mid u \in \tilde{\mathcal{U}},\ z_0 \in \mathcal{S}_z,\ g_{z_i}(u, z_0) \leq 0,\ i = 1, \cdots, r \}. \tag{4.9.17}$$

Then, the above constrained optimal control problem with design parameters is converted into the following constrained optimal control problem without design parameters:

Problem ($P_{without-design-parameters}$). Subject to the dynamical system (4.9.13)-(4.9.14), find a control $u \in \mathcal{U}_z$ and an initial state $z_0 \in \mathcal{S}_z$, such that the cost functional

$$J(u, z_0) \;=\; K_z(z_0, z^{u, x_0}(t_f)) + \int_{t_0}^{t_f} L_z(z^{u, z_0}(\tau), u(\tau), \tau)\, d\tau \tag{4.9.18}$$

is minimized over $\mathcal{F}_z$.

It is clear now that Algorithm 2, described in sections 4.4 and 4.5, can be used to solve constrained optimal control problems with design parameters.

4.10 Special Case: Optimize a Constrained Dynamical System over Its Initial State

In practice, there is also a need to optimize a constrained dynamical system over its initial state. It will be seen in this section that Algorithm 2, described in sections 4.4 and 4.5, can be used to solve that problem as well. To see this, let us consider the following dynamical system defined on a fixed end-time interval $[t_0, t_f]$:

$$\dot{x}^{x_0}(t) \;=\; f(x^{x_0}(t), t), \tag{4.10.1}$$

$$x^{x_0}(t_0) \;=\; x_0. \tag{4.10.2}$$

There are n_{x_0} components of the initial state vector x_0 which are allowed to vary within a constraint box, while the remaining $n-n_{x_0}$ components are fixed. That is, there is an index set $I_{x_0} \subset \{1, \cdots, n\}$ such that

$$x_0 \in \mathcal{S} = \{ \nu \in \mathcal{R}^n \mid X_i^{min} \leq \nu_i \leq X_i^{max},\ i \in I_{x_0};\ \nu_i = x_{0i}(\text{fixed}),\ i \notin I_{x_0} \}. \quad (4.10.3)$$

The dynamical system (4.10.1)-(4.10.2) is also subject to the end-point inequality constraints,

$$g_i(x_0) = h_i(x_0, x^{x_0}(t_f)) \leq 0, \qquad i = 1, \cdots, r. \quad (4.10.4)$$

Let the set of feasible initial states be

$$\mathcal{F} \triangleq \{ x_0 \mid x_0 \in \mathcal{S},\ g_i(x_0) \leq 0,\ i = 1, \cdots, r \}. \quad (4.10.5)$$

We consider the following optimization problem:

Problem ($P_{opt-init}$). Subject to the dynamical system (4.10.1)-(4.10.2), find an initial state $x_0 \in \mathcal{S}$ such that the cost functional

$$J(x_0) = K(x_0, x^{x_0}(t_f)) + \int_{t_0}^{t_f} L(x^{x_0}(\tau), \tau)\, d\tau \quad (4.10.6)$$

is minimized over $\mathcal{F}$.

It is obvious that if we do not allow the control u to vary, i.e., by letting $v = 0$, all the previous results shown in sections 4.4 and 4.5 are immediately reduced to results of the above problem. The specialization of those results is omited here because it is a straightforward process.

Chapter 5

Computational Methods of Solving Constrained Linear Quadratic Problems

5.1 Introduction

In Chapter 3 and Chapter 4, the algorithms which can solve generally constrained optimal control problems were developed. A common feature of those algorithms is that it is required to solve a "direction-finding" subproblem, which is a a generally constrained linear quadratic regulator problem (LQR), at each iteration. On the other hand, the constrained LQR problem is important in its own right.

The goal of this chapter is to study the following constrained LQR problem: minimizing a quadratic functional, subject to a linear dynamical system, hard control constraints, hard initial state constraints, and linear end-point constraints. Two special properties which are related to the problem are presented: (1) the existence of an optimal control solution; and (2) the uniqueness of the optimal control solution. In addition, some computational techniques are investigated.

5.2 Problem Formulation

The dynamical system considered is described by the linear differential equation defined on a fixed end-time interval $[t_0, t_f]$:

$$\dot{y}^{u,\zeta}(t) = A(t)y^{u,\zeta}(t) + B(t)(u(t) - \hat{u}(t)), \tag{5.2.1}$$

$$y^{u,\zeta}(0) = \zeta - \hat{\zeta}. \tag{5.2.2}$$

There are n_ζ components of vector ζ which are allowed to vary within a constraint box, while the remaining $n - n_\zeta$ components are fixed. That is, there is an index set $I_\zeta \subset \{1, \cdots, n\}$ such that

$$\zeta \in \mathcal{S} = \{ \nu \in \mathcal{R}^n \mid X_i^{min} \leq \nu_i \leq X_i^{max},\ i \in I_\zeta;\ \nu_i = \zeta_i \,(\text{fixed}),\ i \notin I_\zeta \} \tag{5.2.3}$$

where $\mathcal{S}$ is compact. The dynamical system (5.2.1)-(5.2.2) is also subject to the control constraints

$$u(t) \in \Omega = \{ \mu \in \mathcal{R}^m \mid U_i^{min} \leq \mu_i \leq U_i^{max},\ i = 1, \cdots, m \} \quad \forall t \in [t_0, t_f], \tag{5.2.4}$$

where Ω is compact, and the linear end-point inequality constraints

$$g_i(u, \zeta) = c_i + (d_i)^\top(\zeta - \hat{\zeta}) + (e_i)^\top y^{u,\zeta}(t_f) \leq 0, \qquad i = 1, \cdots, r. \tag{5.2.5}$$

In the above, $\hat{u}(t)$ is a given piecewise continuous function defined on the interval $[t_0, t_f]$, and $\hat{\zeta}$ a given vector in $\mathcal{R}^n$. Also, $y^{u,\zeta}(t) \in \mathcal{R}^n$ is the state of the system (5.2.1)-(5.2.2) on $[t_0, t_f]$, which corresponds to $u(t) \in \mathcal{R}^m$ and $\zeta \in \mathcal{R}^n$; c_i, $i = 1, \cdots, r$, are some given real numbers; d_i and e_i, $i = 1, \cdots, r$, are some given vectors in $\mathcal{R}^n$. Let the set of admissible controls be,

$$\mathcal{U} = \{ u \mid u : [t_0, t_f] \to \Omega \text{ is continuous a.e. } \}, \tag{5.2.6}$$

and let $\tilde{\mathcal{U}}$ be the set of equivalence classes of functions in $\mathcal{U}$ which are equal almost everywhere on $[t_0, t_f]$. Let the set of combined feasible controls and feasible initial states be

$$\mathcal{F} \triangleq \{ (u, \zeta) \mid u \in \tilde{\mathcal{U}},\ \zeta \in \mathcal{S},\ g_i(u, \zeta) \leq 0,\ i = 1, \cdots, r \}. \tag{5.2.7}$$

We may now formulate the optimal control problem as follows:

Problem (P). Subject to the dynamical system (5.2.1)-(5.2.2), find a control $u \in \mathcal{U}$ and an initial state $\zeta \in \mathcal{S}$ such that the cost functional

$$g_0(u, \zeta) = \omega(\zeta, y^{u,\zeta}(t_f)) + \int_{t_0}^{t_f} \Big(L_1(y^{u,\zeta}(\tau), \tau) + L_2(u(\tau), \tau) \Big) d\tau \tag{5.2.8}$$

is minimized over $\mathcal{F}$.

The following conditions are assumed to be satisfied.

Assumption 5.2.1 *$A(t)$ and $B(t)$ are, respectively, real piecewise continuous $n \times n$ and $n \times m$ matrices defined on $[t_0, t_f]$;*

Assumption 5.2.2 *$\omega : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$ is convex and continuously differentiable in both ζ and y_f; $L_1 : \mathcal{R}^n \times [t_0, t_f] \to \mathcal{R}$, is continuous on $\mathcal{R}^n \times [t_0, t_f]$, and convex and continuously differentiable in $\mathcal{R}^n$ for each $t \in [t_0, t_f]$; $L_2 : \mathcal{R}^m \times [t_0, t_f] \to \mathcal{R}$, is continuous on $\mathcal{R}^m \times [t_0, t_f]$, and strictly convex and continuously differentiable in $\mathcal{R}^m$ for each $t \in [t_0, t_f]$.*

Under Assumption 5.2.1, it is well-known from the theory of linear systems that [18,63], for each $u \in \mathcal{U}$, there exists a unique absolutely continuous solution to the system (5.2.1)-(5.2.2) which can be expressed in terms of the fundamental matrix of the differential equation (5.2.1). More precisely, for any $t \in [t_0, t_f]$,

$$y^{u,\zeta}(t) = \Phi(t, t_0)(\zeta - \hat{\zeta}) + \int_{t_0}^{t} \Phi(t, \tau) B(\tau)(u(\tau) - \hat{u}(\tau))\, d\tau \tag{5.2.9}$$

where $\Phi(t, \tau)$, the fundamental matrix of the differential equation (5.2.1), is a real continuous $n \times n$ matrix defined on $t, \tau \in [t_0, t_f]$, and satisfies the following system:

$$\frac{\partial \Phi(t, \tau)}{\partial t} = A(t)\Phi(t, \tau), \quad \forall t, \tau \in [t_0, t_f], \tag{5.2.10}$$

$$\Phi(t_0, t_0) = I. \tag{5.2.11}$$

5.3 Existence of the Optimal Control

Let S_1 be a given subset of $\mathcal{R}^{n+1}$, S_2 be a given subset of $\mathcal{R}^{2n}$. For every $(y,t) \in S_1$, let $U(y,t)$ be a given subset of $\mathcal{R}^m$, and let $M \subset \mathcal{R}^{n+m+1}$ be the set of all (y,u,t) with $(y,t) \in S_1$ and $u \in U(y,t)$. For each fixed $(y,t) \in S_1$, let the extended velocity set $V(y,t) \subset \mathcal{R}^{1+n}$ be the set of all (z^0,z) with $z^0 \geq f_0(y,u,t)$, $z = f(y,u,t)$ for some $u \in U(y,t)$. We consider the problem of the minimization of the cost functional

$$J(u,\zeta) = \omega(\zeta - \hat{\zeta}, y^{u,\zeta}(t_f)) + \int_{t_0}^{t_f} f_0(y^{u,\zeta}(\tau), u(\tau), \tau)\, d\tau \tag{5.3.1}$$

with $y^{u,\zeta}(t) \in \mathcal{R}^n$, $u(t) \in \mathcal{R}^m$, for any $t \in [t_0, t_f]$, satisfying

$$\begin{gathered} \dot{y}^{u,\zeta}(t) = f(y^{u,\zeta}(t), u(t), t), \quad t \in [t_0, t_f] \ \ (a.e.), \\ (y^{u,\zeta}(t), t) \in S_1, \quad u(t) \in U(y^{u,\zeta}(t), t), \quad t \in [t_0, t_f] \ \ (a.e.), \\ (\zeta, y^{u,\zeta}(t_f)) \in S_2, \\ f_0(y^{u,\zeta}(t), u(t), t), \quad \mathcal{L}\text{-integrable in } [t_0, t_f]. \end{gathered} \tag{5.3.2}$$

A pair (u,ζ), which satisfies all the requirements (5.3.2) and $u(\cdot)$ is measurable, is said to be a admissible pair. Let $\mathcal{E}$ be the class of all admissible pairs (u,ζ). The following is the famous Filippov Existence Theorem (see, for example, Section 9.3 of Cesari [14], or, Section 4.2 of Lee and Marcus [76]):

Theorem 5.3.1 (Filippov) *Let S_1 be compact, S_2 closed, M compact, ω lower semicontinuous on S_2, $f_0(y,u,t)$, $f(y,u,t)$ continuous on M. Assume that, for almost all t, the set $V(y,t)$ is convex in $\mathcal{R}^{1+n}$ for each fixed $(y,t) \in S_1$. Then the functional $J(u,\zeta)$ given by (5.3.1) has an absolute minimum in the nonempty class $\mathcal{E}$ of all admissible pairs.*

Next, we are going to apply the above Filippov Existence Theorem to show that, under a mild condition, the existence of a solution of the optimal control problem (P) defined in Section 5.2 is always guaranteed.

Theorem 5.3.2 *Assume that the feasible control set $\mathcal{F}$ is nonvoid. There always exists a solution to the optimal control problem (P) defined in Section 5.2.*

Proof: Since $\Phi(t,\tau)$ is a real continuous $n\times n$ matrix defined on $t,\tau\in[t_0,t_f]$, with fixed t_0 and t_f, there exist a constant $0<k<\infty$ such that

$$\|\Phi(t,t_0)\| \le k$$

for all $t\in[t_0,t_f]$. Then, according to (5.2.9),

$$\|y^{u,\zeta}(t)\| \le k\Big(\|\zeta\| + \|\hat{\zeta}\| + \int_{t_0}^{t_f} \|B(\tau)\|\,\|u(\tau)\|\,d\tau + \int_{t_0}^{t_f} \|B(\tau)\|\,\|\hat{u}(\tau)\|\,d\tau\Big).$$

Because $u(t)\in\Omega$, Ω is compact, $\zeta\in\mathcal{S}$, $\mathcal{S}$ is compact, and, $B(t)$ and $\hat{u}(t)$ are piecewise continuous on $[t_0,t_f]$, there exists another constant $0<k'<\infty$ such that

$$\| y^{u,\zeta}(t)\| \le k'$$

for all $t\in[t_0,t_f]$. So, there exist compact sets S_1, S_2 and M, such that,

$$(y^{u,\zeta}(t),u(t),t)\in M, \quad (y^{u,\zeta}(t),t)\in S_1, \quad t\in[t_0,t_f] \ \ (a.e.),$$
$$(\zeta, y^{u,\zeta}(t_f))\in S_2.$$

Notice that the assumption on the nonvoidness of the feasible control set $\mathcal{F}$ implies the nonvoidness of the class of all admissible pairs $\mathcal{E}$. Based on the above observations and the Assumptions 5.2.1-5.2.2, theorem 5.3.1 will hold if we can show that, for almost all t, the set $V(y,t)$ is convex in $\mathcal{R}^{1+n}$ for each fixed $(y,t)\in S_1$.

For any fixed $(y,t)\in S_1$, select arbitrarily two points q_1, q_2 from the corresponding $V(y,t)$, with $q_1=(z_1^0,z_1)\in V(y,t)$, corresponding to some $u_1(t)\in\Omega$, and $q_2=(z_2^0,z_2)\in V(y,t)$, corresponding to some $u_2(t)\in\Omega$. For an arbitrary $\lambda\in[0,1]$, denote $\bar{u}=\lambda u_1+(1-\lambda)u_2$. Then,

$$\begin{aligned}
&\lambda z_1^0 + (1-\lambda)z_2^0 \\
&\quad\ge\ \lambda(L_1(y,t)+L_2(u_1(t),t)) + (1-\lambda)(L_1(y,t)+L_2(u_2(t),t)) \\
&\quad\ge\ L_1(y,t)+L_2(\lambda u_1(t)+(1-\lambda)u_2(t),t) \\
&\quad=\ L_1(y,t)+L_2(\bar{u}(t),t) \qquad (5.3.3)
\end{aligned}$$

by the convexity of $L_2(u,t)$ in u. Similarly,

$$\lambda z_1 + (1-\lambda)z_2$$

$$
\begin{aligned}
&= \lambda(A(t)y + B(t)(u_1(t) - \hat{u}(t))) + (1-\lambda)(A(t)y + B(t)(u_2(t) - \hat{u}(t))) \\
&= A(t)y + B(t)(\lambda u_1(t) + (1-\lambda)u_2(t)) - B(t)\hat{u}(t) \\
&= A(t)y + B(t)(\bar{u}(t) - \hat{u}(t)). \qquad (5.3.4)
\end{aligned}
$$

Combining relations (5.3.3) and (5.3.4), we then have

$$\lambda q_1 + (1-\lambda)q_2 \in V(y,t)$$

corresponding to control $\bar{u}(t) = \lambda u_1(t) + (1-\lambda)u_2(t) \in \Omega$, since Ω is a convex set. Therefore, the set $V(y,t)$ is convex in $\mathcal{R}^{1+n}$ for each fixed $(y,t) \in S_1$. The proof is then finished when theorem 5.3.1 is applied. □

Remark: Inspecting the above proof of the existence theorem for the optimal control problem (P), it is worth to notice that the convexity assumptions on both $\omega(y^{u,\zeta}(t_f))$ and $L_1(y^{u,\zeta}(t),t)$ have not been used. Therefore, for the following more general problem of the minimization of the cost functional

$$J(u,\zeta) = \omega(y^{u,\zeta}(t_0), y^{u,\zeta}(t_f)) + \int_{t_0}^{t_f} f_0(y^{u,\zeta}(\tau), u(\tau), \tau)\, d\tau \qquad (5.3.5)$$

with $y^{u,\zeta}(t) \in \mathcal{R}^n$, $u(t) \in \mathcal{R}^m$, for any $t \in [t_0, t_f]$, satisfying

$$
\begin{gathered}
\dot{y}^{u,\zeta}(t) = A(y^{u,\zeta}(t),t) + B(t)u(t), \quad t \in [t_0,t_f] \ \ (a.e.), \\
(y^{u,\zeta}(t),t) \in S_1, \quad u(t) \in U(y^{u,\zeta}(t),t), \quad t \in [t_0,t_f] \ \ (a.e.), \\
(\zeta, y^{u,\zeta}(t_f)) \in S_2, \qquad (5.3.6) \\
A(y^{u,\zeta}(t),t),\ B(t),\ \hat{u}(t) \ \text{ piecewise continuous in } [t_0,t_f], \\
f_0(y^{u,\zeta}(t),u(t),t) \ \text{ convex in } u(t), \text{ and } \mathcal{L}\text{–integrable in } [t_0,t_f],
\end{gathered}
$$

it can be easily shown that the extended velocity set $V(y,t)$ is always convex in $\mathcal{R}^{1+n}$ for each fixed $(y,t) \in S_1$. Therefore, as long as S_1, S_2, M being compact, ω lower semi-continuous on S_2, $f_0(y,u,t)$, $f(y,u,t)$ continuous on M, and class $\mathcal{E}$ of all admissible pairs not empty, theorem 5.3.1 holds. Then, the existence of the optimal control of the above problem is guaranteed.

5.4 Uniqueness of the Optimal Control

We know that the general linear quadratic optimal control problem defined in Section 5.2 can be viewed be viewed as an abstract optimization problem in function space

$$\min_{(u,\zeta)\in\mathcal{F}} J(u,\zeta)$$

where $\mathcal{F}$ is the feasible set defined in (5.2.7). The following proposition shows that the problem enjoys some convexity properties.

Proposition 5.4.1 *$J(u,\zeta)$ is a convex functional on $\mathcal{L}_\infty^m[t_0,t_f]\times\mathcal{R}^n$. If we identify all the elements of $\mathcal{L}_\infty^m[t_0,t_f]$ which are equal almost everywhere on $[t_0,t_f]$, $J(u,\zeta)$ becomes a strict convex functional. Moreover, the feasible control set $\mathcal{F}$ is convex.*

Proof: Denote $y^{u,\zeta}(t)$s by the unique solution of the system (5.2.1)-(5.2.2) corresponding to $u(t)$, $\forall t\in[t_0,t_f]$, and ζ. For any $u_1,u_2\in\mathcal{L}_\infty^m[t_0,t_f]$, and any $\zeta_1,\zeta_2\in\mathcal{S}$, and any $\lambda\in[0,1]$

$$\begin{aligned}
&y^{\lambda u_1+(1-\lambda)u_2,\,\lambda\zeta_1+(1-\lambda)\zeta_2}(t)\\
&= \Phi(t,t_0)(\lambda\zeta_1+(1-\lambda)\zeta_2)+\int_{t_0}^{t}\Phi(t,\tau)B(\tau)(\lambda u_1(\tau)+(1-\lambda)u_2(\tau)-\hat{u}(\tau))\,d\tau\\
&= \lambda\Big(\Phi(t,t_0)\zeta_1+\int_{t_0}^{t}\Phi(t,\tau)B(\tau)(u_1(\tau)-\hat{u}(\tau))\,d\tau\Big)\\
&\quad+(1-\lambda)\Big(\Phi(t,t_0)\zeta_2+\int_{t_0}^{t}\Phi(t,\tau)B(\tau)(u_2(\tau)-\hat{u}(\tau))\,d\tau\Big)\\
&= \lambda y^{u_1,\zeta_1}(t)+(1-\lambda)y^{u_2,\zeta_2}(t). \qquad (5.4.1)
\end{aligned}$$

Then,

$$\begin{aligned}
&J(\lambda u_1+(1-\lambda)u_2,\ \lambda\zeta_1+(1-\lambda)\zeta_2)\\
&= \omega(\lambda\zeta_1+(1-\lambda)\zeta_2, y^{\lambda u_1+(1-\lambda)u_2,\,\lambda\zeta_1+(1-\lambda)\zeta_2}(t_f))\\
&\quad+\int_{t_0}^{t_f}L_1(y^{\lambda u_1+(1-\lambda)u_2,\,\lambda\zeta_1+(1-\lambda)\zeta_2}(\tau),\tau)\,d\tau\\
&\quad+\int_{t_0}^{t_f}L_2(\lambda u_1(\tau)+(1-\lambda)u_2(\tau),\tau)\,d\tau\\
&= \omega(\lambda\zeta_1+(1-\lambda)\zeta_2,\ \lambda y^{u_1,\zeta_1}(t_f)+(1-\lambda)y^{u_2,\zeta_2}(t_f))
\end{aligned}$$

$$
\begin{aligned}
&+\int_{t_0}^{t_f} L_2(\lambda u_1(\tau)+(1-\lambda)u_2(\tau),\tau)\,d\tau \\
&+\int_{t_0}^{t_f} L_1(\lambda y^{u_1,\zeta_1}(\tau)+(1-\lambda)y^{u_2,\zeta_2}(\tau),\,\tau)\,d\tau \\
\le\ &\lambda\,\omega(y^{u_1,\zeta_1}(t_f))+(1-\lambda)\,\omega(y^{u_2,\zeta_1}(t_f)) \\
&+\int_{t_0}^{t_f}\Big(\lambda L_1(y^{u_1,\zeta_1}(\tau),\tau)+(1-\lambda)L_1(y^{u_2,\zeta_2}(\tau),\tau)\Big)d\tau \\
&+\int_{t_0}^{t_f}\Big(\lambda L_2(u_1(\tau),\tau)+(1-\lambda)L_2(u_2(\tau),\tau)\Big)d\tau \qquad (5.4.2)
\end{aligned}
$$

by the linearity of system (5.2.1)-(5.2.2), and the convexities of $\omega(y^{u,\zeta}(t_f))$, $L_1(y^{u,\zeta}(t),t)$ and $L_2(u(t),t)$. So,

$$J(\lambda u_1+(1-\lambda)u_2,\lambda\zeta_1+(1-\lambda)\zeta_2)\le\lambda J(u_1,\zeta_1)+(1-\lambda)J(u_2,\zeta_2),$$

which implies that $J(u,\zeta)$ is a convex functional on $\mathcal{L}_\infty^m[t_0,t_f]\times\mathcal{R}^n$.

If, however, we identify all the elements of $\mathcal{L}_\infty^m[t_0,t_f]$ which are equal almost everywhere on $[t_0,t_f]$, let us select arbitrarily $u_1,u_2\in\mathcal{L}_\infty^m[t_0,t_f]$, $u_1\neq u_2$. Then, the measure of the set of t when $u_1(t)$ differs $u_2(t)$ will be a positive number, that is,

$$m(\{\ t\mid u_1(t)\neq u_2(t),\ t\in[t_0,t_f]\ \})>0.$$

Then,

$$
\begin{aligned}
&\int_{t_0}^{t_f} L_2(\lambda u_1(\tau)+(1-\lambda)u_2(\tau),\tau)\,d\tau \\
&\qquad<\int_{t_0}^{t_f}\Big(\lambda L_2(u_1(\tau),\tau)+(1-\lambda)L_2(u_2(\tau),\tau)\Big)d\tau. \qquad (5.4.3)
\end{aligned}
$$

Applying (5.4.3) to (5.4.2), we then have

$$J(\lambda u_1+(1-\lambda)u_2,\lambda\zeta_1+(1-\lambda)\zeta_2)<\lambda J(u_1,\zeta_1)+(1-\lambda)J(u_2,\zeta_2),$$

which implies that $J(u,\zeta)$ is a strict convex function.

To show the convexity of $\mathcal{F}$, let us select arbitrarily (u_1,ζ_1), $(u_2,\zeta_2)\in\mathcal{F}$, and $\lambda\in[0,1]$. Then,

$$u_1(t),u_2(t)\ \in\ \Omega,\qquad\forall t\in[t_0,t_f],$$

and

$$\begin{aligned} c_i + (d_i)^\top \zeta_1 + (e_i)^\top y^{u_1}\zeta_1(t_f) &\leq 0 \\ c_i + (d_i)^\top \zeta_2 + (e_i)^\top y^{u_2}\zeta_2(t_f) &\leq 0 \end{aligned}$$

for $i=1,\cdots,r$. From the convexities of Ω and $\mathcal{S}$, we have

$$\lambda u_1(t) + (1-\lambda)u_2(t) \in \Omega \tag{5.4.4}$$

for all $t\in[t_0,t_f]$, and

$$\lambda\zeta_1 + (1-\lambda)\zeta_2 \in \mathcal{S}. \tag{5.4.5}$$

From the linearity relation (5.4.1), we have

$$\begin{aligned} & c_i + (d_i)^\top(\lambda\zeta_1+(1-\lambda)\zeta_2) + (e_i)^\top y^{\lambda u_1+(1-\lambda)u_2,\,\lambda\zeta_1+(1-\lambda)\zeta_2)}(t_f) \\ &= c_i + (d_i)^\top(\lambda\zeta_1+(1-\lambda)\zeta_2) + (e_i)^\top(\lambda y^{u_1,\zeta_1}(t_f) + (1-\lambda)y^{u_2,\zeta_2}(t_f)) \\ &= \lambda(c_i + (d_i)^\top\zeta_1 + (e_i)^\top y^{u_1}(t_f)) + (1-\lambda)(c_i + (d_i)^\top\zeta_2 + (e_i)^\top y^{u_2}(t_f)) \\ &\leq 0. \end{aligned} \tag{5.4.6}$$

According to (5.4.4), (5.4.5) and (5.4.6), we know that $(\lambda u_1+(1-\lambda)u_2, \lambda\zeta_1+(1-\lambda)\zeta_2)\in\mathcal{F}$, which implies that the feasible control set $\mathcal{F}$ is convex. □

The following proposition concerns with the global property of a local minimizer of a (strictly) convex functional. The "convex" part of the proof can be found in [85], while the "strictly convex" part of the proof is an easy extension to the "convex" part.

Proposition 5.4.2 *Let f be a (strictly) convex functional defined on a convex subset S of a normed space $\mathcal{X}$. If $\bar{x}$ is a local minimizer of f, then $\bar{x}$ is a (unique) global minimizer of f.*

Proof: Since $\bar{x}$ is a local minimizer of f, there exists an open set N containing $\bar{x}$ such that

$$f(\bar{x}) \leq f(x), \qquad \forall x \in S\cap N.$$

Since set N is open, for any $x \in S$, $x \neq \bar{x}$, we can always find an α, $0 < \alpha < 1$, such that $\bar{x} + \alpha(x - \bar{x}) \in S \cap N$. We then have

$$f(\bar{x}) \leq f(\bar{x} + \alpha(x - \bar{x})) = f((1 - \alpha)\bar{x} + \alpha x)$$

(1) If f is convex, then

$$f(\bar{x}) \leq (1 - \alpha)f(\bar{x}) + \alpha f(x).$$

Since $0 < \alpha < 1$, we then have

$$f(\bar{x}) \leq f(x),$$

which implies that $\bar{x}$ is a global minimizer of f.

(2) If f is strictly convex, then

$$f(\bar{x}) < (1 - \alpha)f(\bar{x}) + \alpha f(x).$$

Since $0 < \alpha < 1$, we then have

$$f(\bar{x}) < f(x),$$

which implies that $\bar{x}$ is the unique global minimizer of f. □

Theorem 5.4.1 *If we identify all the elements of $\mathcal{L}_\infty^m[t_0, t_f]$ which are equal almost everywhere on $[t_0, t_f]$, then any local minimum (u^*, ζ^*) of the optimal control problem (P) will be the unique global minimum.*

Proof: It is a direct consequence of proposition 5.4.1 and proposition 5.4.2. □

5.5 Computational Methods

In this section, we only concentrate on a special case of the problem formulated in Section 5.2:

$$\omega(\zeta, y^{u,\zeta}(t_f)) = \frac{1}{2}\zeta^\top K_1 \zeta + \frac{1}{2} y^{u,\zeta}(t_f)^\top K_2\, y^{u,\zeta}(t_f) \tag{5.5.1}$$

$$L_1(y^{u,\zeta}(t), t) = \frac{1}{2} y^{u,\zeta}(t)^\top Q(t) y^{u,\zeta}(t) \tag{5.5.2}$$

$$L_2(u(t), t) = T(t)^\top u(t) + \frac{1}{2} u(t)^\top R(t) u(t) \tag{5.5.3}$$

where K_1 is symmetric and positive definite, K_2 symmetric and semi-positive definite, $Q(t)$ symmetric and semi-positive definite for any $t \in [t_0, t_f]$, $R(t)$ symmetric and positive definite for any $t \in [t_0, t_f]$.

If there is neither control constraint nor end-point inequality constraint nor variable initial state, the optimal control problem we study is just the classical linear quadratic regulator problem, whose optimal control is in state feedback form and can be computed by solving matrix Riccati equations [2,4,72]. Equivalently, for a constrained problem whose (u, ζ) is interior to $\mathcal{F}$, the problem can be treated as though it is an unconstrained one. If, however, some constraints could be satisfied on their boundaries, other methods should be used.

Let us first consider the problem where there is only the control constraint (5.2.6), i.e. there is neither the end-point inequality constraint (5.2.7) nor the variable initial state (5.2.3). In [5], Barnes developed an algorithm for computing the optimal control of such problems. In his algorithm, the original problem (P) is replaced by a sequence of subproblems of minimizing the first order Taylor approximation of the cost functional (5.2.8). Consequently, the subproblems are the ones whose Hamiltonian functions are linear in control. He has shown that, if the subproblems are solved exactly and iteratively, and if the controls are updated in the right way, the sequence of control functions will converge pointwise to the unique optimal control. However, a very important prospect has been overlooked. That is, if there exists singularity at some k-th iteration, which is likely to happen, the subproblem could not be solved easily. So, the effectiveness of the algorithm is seriously affected.

Due to the availability of well-developed nonlinear programming techniques and powerful computers, a practical and handy way to handle the generally constrained optimal control problem seems to be the one which converts the original problem into a finite-dimensional optimization. In Chapter 2, many commonly used parameterization techniques have been surveyed. In this chapter, only control parameterization techniques are used to solve the problem formulated in Section 5.2.

5.5.1 A Parameterization Method

In this section, a sequence of approximating problems (P_p), $p = 1, \cdots, \infty$, are constructed from problem (P) by discretizing each control variable by piecewise constants. For each p, the interval $[t_0, t_f]$ is subdivided by a mesh of N_p+1 points:

$$\Delta^p : \quad t_0 = t_0^p < t_1^p < \cdots < t_k^p < \cdots < t_{N_p}^p = t_f.$$

The partition $\mathcal{I}^p$, corresponding to each Δ^p, is defined by

$$\mathcal{I}^p = \{ I_k^p : \ k = 1, \cdots, N_p \},$$

where $I_k^p = [t_{k-1}^p, t_k^p]$. The Δ^p's are chosen such that the following two properties are satisfied: (i) Δ^{p+1} is a refinement of Δ^p; (ii) by denoting $|I_k^p| = |t_k^p - t_{k-1}^p|$,

$$\lim_{p \to \infty} \max_{k=1,\cdots,N_p} |I_k^p| = 0.$$

Corresponding to each partition $\mathcal{I}^p$, let $\mathcal{U}^p$ be a set which consists of all the piecewise constant controls expressed by

$$u^p(t) = \sum_{k=1}^{N_p} \mu^{p,k} \mathcal{X}_{I_k^p}(t) \tag{5.5.4}$$

for $t \in [t_0, t_f]$, where $\mathcal{X}_I$ is the indicator function defined in Chapter 2. When $u(t)$ is approximated by $u^p(t)$ during $[t_0, t_f]$, the control function is then parameterized by the vector $[(\mu^{p,1})^\top, \cdots, (\mu^{p,N_p})^\top]^\top$, $\mu^{p,k} \in \mathcal{R}^m, k = 1, \cdots, N_p$. Let

$$\xi^p = [(\mu^{p,1})^\top, \cdots, (\mu^{p,N_p})^\top, \zeta^\top]^\top \in R^{\sigma_p}$$

where $\sigma_p = mN_p + n$. After the control is parameterized by piecewise constants, the original system (5.2.1)-(5.2.2) becomes

$$\dot{y}(t|\xi) = A(t)y(t|\xi) + B(t) \sum_{k=1}^{N_p} \mu^{p,k} \mathcal{X}_{I_k^p}(t) + \hat{B}(t) \tag{5.5.5}$$

$$y(t_0|\xi) = \zeta, \tag{5.5.6}$$

where $\hat{B}(t) = B(t)\hat{u}(t)$, and the control constraints defined in (5.2.4) becomes

$$U_i^{min} \le \mu_i^{p,k} \le U_i^{max}, \quad i = 1, \cdots, m, \ k = 1, \cdots, N_p. \tag{5.5.7}$$

The constraints on initial state are still the same as before

$$\zeta \in \mathcal{S} = \{\, \nu \in \mathcal{R}^n \mid X_i^{min} \le \nu_i \le X_i^{max},\ i \in I_\zeta;\ \ \nu_i = \zeta_i(\text{fixed}),\ i \notin I_\zeta \,\}. \qquad (5.5.8)$$

Let $y^{u,\zeta}(\cdot|\xi^p)$ be the solution of the above system (5.5.5)-(5.5.6). Then, the linear end-point inequality constraints (5.2.5) becomes

$$\tilde{g}_i(\xi^p) = c_i + (d_i)^\top \zeta + (e_i)^\top y(t_f|\xi^p) \le 0, \qquad i = 1, \cdots, r. \qquad (5.5.9)$$

Let Ξ^p be the set of all those ξ^p vectors which satisfy the constraints (5.5.7) and (5.5.9). We may now formulate the approximating problem as follows:

Problem (P_p). Subject to the dynamical system (5.5.5)-(5.5.6), find a vector $\xi^p = [(\mu^{p,1})^\top, \cdots, (\mu^{p,N_p})^\top, \zeta^\top]^\top$ such that the cost functional

$$\begin{aligned}
\tilde{g}_0(\xi^p) \ &= \ \frac{1}{2}\zeta^\top K_1 \zeta + \frac{1}{2} y(t_f|\xi)^\top K_2\, y(t_f|\xi) \\
&\quad + \int_{t_0}^{t_f} \Big(\frac{1}{2} y(\tau|\xi)^\top Q(\tau) y(\tau|\xi) + T(\tau)^\top \sum_{k=1}^{N_p} \mu^{p,k} \mathcal{X}_{I_k^p}(\tau) \Big) d\tau \\
&\quad + \int_{t_0}^{t_f} \frac{1}{2} \Big(\sum_{k=1}^{N_p} \mu^{p,k} \mathcal{X}_{I_k^p}(\tau) \Big)^\top R(\tau) \Big(\sum_{k=1}^{N_p} \mu^{p,k} \mathcal{X}_{I_k^p}(\tau) \Big) d\tau \qquad (5.5.10)
\end{aligned}$$

is minimized over Ξ^p.

In the following, we assume that the interval $[t_0, t_f]$ is subdivided uniformly by N_p+1 points. According to (5.2.9), the solution of the system (5.5.5)-(5.5.6) becomes

$$y(t_k|\xi) = G(k)\xi - \hat{G}(k)$$

where

$$\begin{aligned}
G(k) \ &= \ \Big[\int_{t_0}^{t_1} \Phi(t_k, \tau) B(\tau)\, d\tau, \cdots, \int_{t_{k-1}}^{t_k} \Phi(t_k, \tau) B(\tau)\, d\tau,\ 0, \ldots,\ 0,\ \Phi(t_k, t_0) \Big] \\
\hat{G}(k) \ &= \ \int_{t_0}^{t_k} \Phi(t_k, \tau) B(\tau) \hat{u}(\tau)\, d\tau
\end{aligned}$$

Thus, by using the rectangular integration rule, the cost functional in (5.5.10) becomes the following quadratic function

$$\tilde{g}_0(\xi^p) \ = \ \frac{1}{2}\zeta^\top K_1 \zeta + \frac{1}{2} \Big(G(N_p)\xi - \hat{G}(N_p) \Big)^\top K_2 \Big(G(N_p)\xi - \hat{G}(N_p) \Big)$$

$$+\frac{h}{2}\sum_{k=0}^{N_p-1}\Big(G(k)\xi-\hat{G}(k)\Big)^{\mathsf{T}}Q(t_k)\Big(G(k)\xi-\hat{G}(k)\Big)$$
$$+h\sum_{k=0}^{N_p-1}T(t_k)^{\mathsf{T}}\mu^{p,k}+\frac{h}{2}\sum_{k=0}^{N_p-1}(\mu^{p,k})^{\mathsf{T}}R(t_k)\,\mu^{p,k}.$$

The above problem (P_p) can then be approximated by the following quadratic programming problem:

Problem (QP_p). Find a vector $\xi^p=[(\mu^{p,1})^{\mathsf{T}},\cdots,(\mu^{p,N_p})^{\mathsf{T}},\zeta^{\mathsf{T}}]^{\mathsf{T}}\in\mathcal{R}^{mN_p+n}$, such that

$$\min_{\xi}\quad \tilde{g}_0(\xi)=\frac{1}{2}\xi^{\mathsf{T}}M\xi+\gamma^{\mathsf{T}}\xi \tag{5.5.11}$$
$$\text{s.t.}\quad \tilde{g}_i(\xi)=\alpha_i+\beta_i^{\mathsf{T}}\xi\ \le\ 0\qquad i=1,\cdots,r \tag{5.5.12}$$

where

$$M=h\sum_{k=0}^{N_p-1}G(k)^{\mathsf{T}}Q(t_k)G(k)+G(N_p)^{\mathsf{T}}K_2\,G(N_p)+h\ \text{diag}[\,R(t_0),\cdots,R(t_{N_p-1}),K_1\,],$$
$$\gamma=h\,[\,T(t_0)^{\mathsf{T}},\cdots,T(t_{N_p-1})^{\mathsf{T}},0\,]^{\mathsf{T}}-h\sum_{k=0}^{N_p-1}G(k)^{\mathsf{T}}Q(t_k)\,\hat{G}(k)-G(N_p)^{\mathsf{T}}K_2\,\hat{G}(N_p),$$

and

$$\alpha_i\ =\ c_i-e_i^{\mathsf{T}}\hat{G}(N_p),\qquad \beta_i\ =\ G(N_p)^{\mathsf{T}}e_i+[\,0,\cdots,0,\,d_i^{\mathsf{T}}\,]^{\mathsf{T}}$$

for $i=1,\cdots,r$.

Notice that, because $R(t_k)$ is symmetric and positive definite for any $k=0,\cdots,N_p-1$, K_1 is symmetric and positive definite, $Q(t_k)$ is symmetric and semi-positive definite for any $k=0,\cdots,N_p-1$, and K_1 is symmetric and semi-positive definite, matrix M is therefore symmetric and positive definite. The problem (QP_k) above is therefore a standard quadratic programming problem, which can be solved efficiently by the active set method within finite steps.

Chapter 6

Numerical Examples

6.1 Introduction

To apply the algorithm described in Chapter 3 or in Chapter 4, it is essential to solve the "direction-finding" subproblem (P_k''). For the special case where there is neither control constraint, nor end-point constraint, nor variable initial state, the "direction-finding" subproblem (P_k'') is just a classical time-varying LQR problem which can be solved by integrating two Riccati equations. For the case where there is only the control constraint (5.2.6), but there is neither the end-point inequality constraint (5.2.7) nor the variable initial state (5.2.3), it can be solved by a convergent algorithm proposed by Barnes [5], or by an effective first-order strong-variation algorithm proposed by Mayne and Polak [95].

For optimal control problems with terminal constraints, a convenient way to solve the "direction-finding" subproblem (P_k'') is to parameterize the control variables by piecewise constants so that (P_k'') is converted to a quadratic programming problem (see Chapter 5 for details). The advantages of this convertion are that (1) any solution of a quadratic programming problem with N variables can be obtained in no more than N steps; (2) quadratic programming problem is usually solved by an active set method which involves only matrix manipulation, so it can be solved efficiently, (3) there are many good codes available. Some of them are freely available, for example,

a Fortran code *qld.f* by Schittkowski and Powell.

For optimal control problems with path constraints, they can always be converted, approximatedly, into optimal control problems with only terminal constraints by introducing some extra state variables (see Section 4.8 for details).

In this chapter, five examples, which have fixed initial state, are solved by applying the algorithm described in Chapter 3. The examples in section 6.2 have neither terminal nor path constraint; while ones in section 6.3 have either terminal or path constraints.

6.2 Without Terminal and Path Constraints

In this section, three examples without terminal and path constraint are solved. As a comparison, four types of simulations are tried:

- type-A: using only the first-order strong-variation algorithm proposed by Mayne and Polak [95];
- type-B: using the algorithm described in Chapter 3 whose "direction-finding" subproblem (P_k'') is solved by Mayne and Polak's first-order strong-variation algorithm;
- type-C: using the algorithm described in Chapter 3 whose "direction-finding" subproblem (P_k'') is solved by quadratic programming;
- type-D: using the algorithm described in Chapter 3 whose "direction-finding" subproblem (P_k'') is solved by integrating two Riccati equations (for unconstrained LQR).

In all cases, the stopping criteria are set to be $-\delta H^{min} \leq 10^{-3}$ and $-\theta(u) \leq 10^{-2}$, where [95]

$$\delta H^{min}(u) = \sup_{t\in[t_0,t_f]} \Big(\overline{H}(t) - H(x(t), p(t), u(t), t) \Big)$$

$$\theta(u) = \frac{1}{\Delta}\int_{t_0}^{t_f}\Big(\overline{H}(t) - H(x(t), p(t), u(t), t)\Big)\,dt$$

where $\Delta = t_f - t_0$ and

$$\overline{H}(t) = \min_{\mu\in\Omega} H(x(t), p(t), \mu, t), \qquad \forall t \in [t_0, t_f].$$

Clearly, $\delta H^{min}(u)$ and $\theta(u) \leq 0$ for any $u \in \mathcal{U}$. The approximation schemes for matrices Λ_1, $\Lambda_2(t)$ and $\Lambda_3(t)$ are quite primitive:

$$\Lambda_1 = \lambda_k I_{n\times n}$$
$$\Lambda_2(t) = \max\{1,\, 0.01\lambda_h(t)\} I_{n\times n}$$
$$\Lambda_3(t) = \max\{1,\, 0.01\lambda_h(t)\} I_{m\times m}$$

for any $t \in [t_0, t_f]$. In the above, λ_k is the largest eigenvalue of matrix K_{xx}, $\lambda_h(t)$ is the largest eigenvalue of matrix $\begin{pmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{pmatrix}$, for $t \in [t_0, t_f]$. In the following, N^* is denoted by the number of iterations before termination, and u^* by the final control.

Example 1. Consider the optimal control problem

$$\min\ J(u) = \frac{1}{2}\int_0^5 \Big(x_1^2(\tau) + x_2^2(\tau) + u^2(\tau)\Big)\,d\tau$$

subject to dynamics described by Van der Pol's equation,

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = x_2(t)(1 - x_1^2(t)) - x_1(t) + u(t)$$

with the initial condition

$$x(0) = (1.5,\ 1.5)^{\mathsf{T}}.$$

The initial control has been set to be

$$u^0(t) \equiv 0, \qquad \forall t \in [0, 5].$$

The sampling number is set to be 200. Notice that, the first-order algorithm alone terminates before reaching the stopping criteria with no further improvement, and,

type	N^*	$J(u^*)$	$\theta(u^*)$	$\delta H^{min}(u^*)$
A	45	4.423059	-0.001134	-0.003152
B	11	4.421777	-0.000438	-0.002431
C	4	4.420835	-0.000002	-0.000022
D	5	4.423215	-0.000941	-0.003902

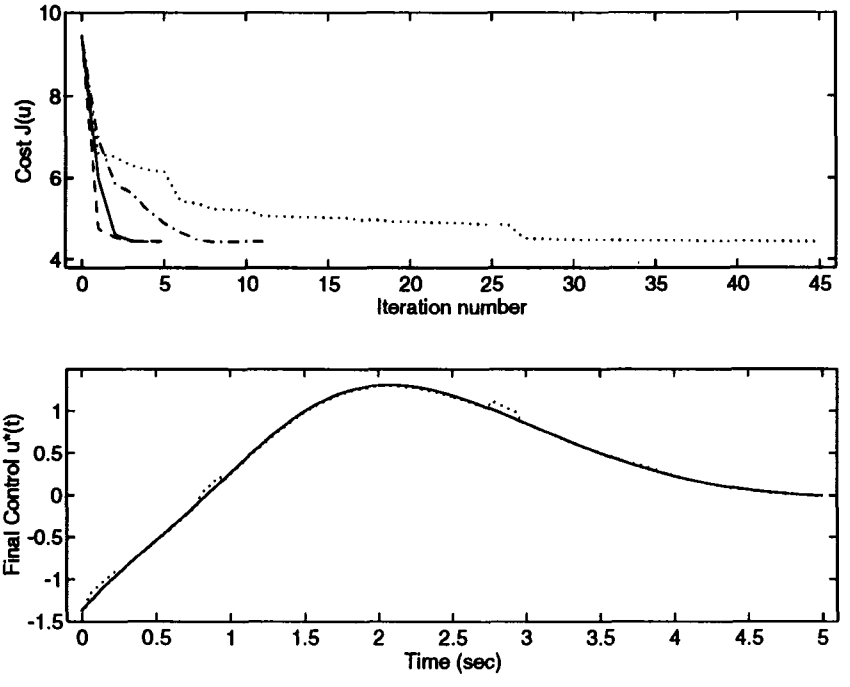


Figure 6.1: Results of Example 1, dotted line for type-A, dashdot line for type-B, solid line for type-C, dashed line for type-D.

it converges much slower than the other three types of simulations. However, please be aware that only the number of iterations are compared here, not the CPU time. The new algorithm described in this dissertation is, of course, more computational intensive than first-order algorithms.

Example 2. Consider the same dynamics and cost functional as in Example 1, except that the control is now constrained by

$$-1 \leq u(t) \leq 1, \qquad \forall t \in [0, 5].$$

Notice that, the first-order algorithm alone converges much slower than the other two types of simulations. Once again, please be aware that only the number of iterations are compared here, not the CPU time. The new algorithm described in this dissertation is, of course, more computational intensive than first-order algorithms.

Example 3. Consider the same dynamics and control constraint as in Example 2, except that the cost functional is a nonlinear one

$$\min_{-1 \leq u \leq 1} J(u) = \frac{1}{2} \int_0^5 \Big(\cos^2(x_1(\tau)) + \sin^2(x_2(\tau)) \Big) \, d\tau.$$

Obviously, because the Hamiltonian is linear in the control, the optimal control must be bang-bang (assuming there is no singularity). Notice that, final values of $-\theta(u^*)$ from both type-A and type-B are smaller than 10^{-3}, while the final values of $-\delta H^{min}(u^*)$ are larger than 10^{-3}. Notice that, type-A and type-B terminate before reaching the stopping criteria with no further improvement, and, the first-order algorithm alone converges slower than the other two types of simulations. Once again, please be aware that only the number of iterations are compared here, not the CPU time. The new algorithm described in this dissertation is, of course, more computational intensive than first-order algorithms.

type	N^*	$J(u^*)$	$\theta(u^*)$	$\delta H^{min}(u^*)$
A	40	4.553104	-0.000312	-0.001221
B	9	4.553078	-0.000037	-0.000804
C	10	4.553684	-0.000156	-0.009200

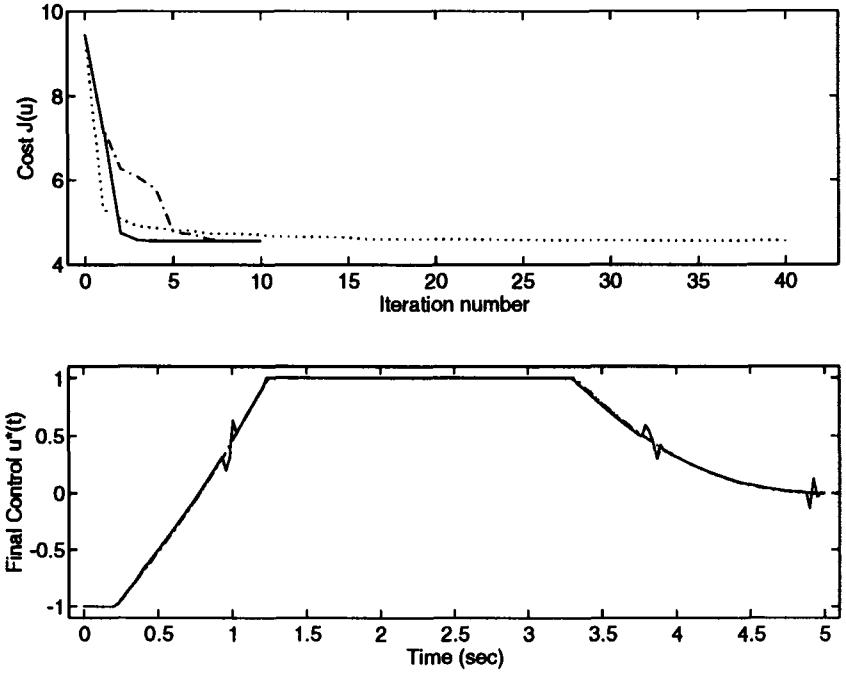


Figure 6.2: Results of Example 2, dotted line for type-A, dashdot line for type-B, solid line for type-C.

type	N^*	$J(u^*)$	$\theta(u^*)$	$\delta H^{min}(u^*)$
A	17	2.166657	-0.003459	-0.138943
B	6	2.149522	-0.000155	-0.016705
C	11	2.149594	-0.000085	-0.005073

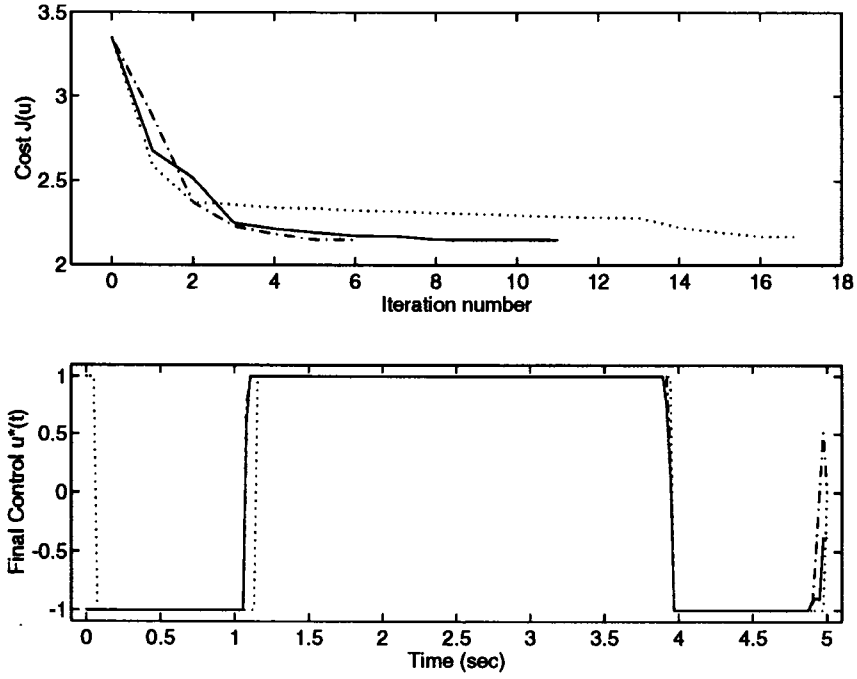


Figure 6.3: Results of Example 3, dotted line for type-A, dashdot line for type-B, solid line for type-C.

N^*	ϵ	γ	$J(u^*)$	$x_3^*(1) - \gamma$
14	0.02	0.01	0.1763	-1.074×10^{-6}

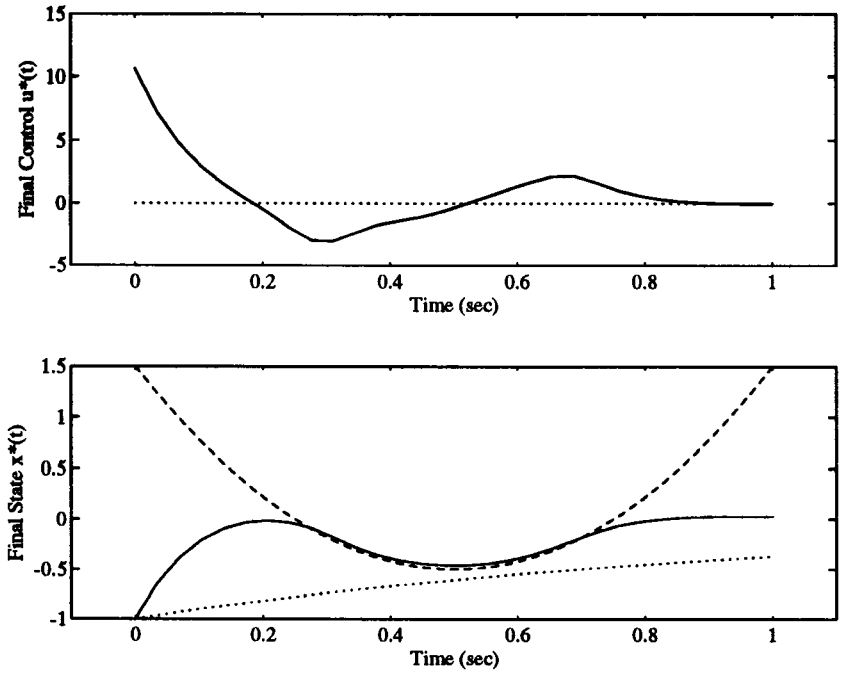


Figure 6.4: Results of Example 4, dotted line for initial results, solid line for final results, dashed line for path constraint.

6.3 With Terminal And/Or Path Constraints

The following two examples have either terminal or path constraints. In both cases, the parameter ρ in the exact penalty function $\theta_\rho(u)$, defined in Chapter 3, is set to be 5. In the following, N^* is denoted by the number of iterations before termination, and u^* by the final control.

Example 4. Consider the optimal control problem

$$\min\ J(u) = \int_0^1 \left(x_1^2(\tau) + x_2^2(\tau) + 0.005 * u^2(\tau) \right) d\tau$$

subject to the dynamical system

$$\dot{x}_1(t) = x_2(t)$$
$$\dot{x}_2(t) = -x_1(t) + u(t)$$

with the initial condition

$$x(0) = (0,\ -1)^\top,$$

the control constraint

$$-20 \leq u(t) \leq 20, \qquad \forall t \in [0,1],$$

and the continuous state path constraint

$$\phi(x(t),t) = x_2(t) - 8(t-0.5)^2 + 0.5 \leq 0, \qquad \forall t \in [0,1].$$

According to Section 4.8, the above state path constraint can be converted into a state terminal constraint by the following transcription: Define

$$\mathcal{L}_\epsilon(\xi) = \begin{cases} (\xi+\epsilon)^2/4\epsilon & \text{when } \xi \geq -\epsilon \\ 0 & \text{when } \xi < -\epsilon \end{cases} \tag{6.3.1}$$

Let a new state variable $x_3(t)$ be such that

$$\dot{x}_3(t) = \mathcal{L}_\epsilon(\phi(x(t),t)), \qquad \forall t \in [0,1].$$

Then, the above state path constraint can be well approximated by the following terminal state constraint:

$$x_3(1) - \gamma \leq 0,$$

if the two positive constants ϵ and γ are chosen appropriately. The initial control has been set to be

$$u^0(t) \equiv 0, \qquad \forall t \in [0,1].$$

The sampling number is set to be 30.

Example 5. In [141], a realistic and complex problem problem of transfering containers from a ship to a cargo truck at the port of Kobe was considered. The container

crane is driven by a hoist motor and a trolley drive motor. For safety reasons, the objective is to minimize the swing during and at the end of the transfer. The problem can be modeled as the following optimal control problem [155]:

$$\min\ J(u) = 4.5 \int_0^1 \Big(x_3^2(\tau) + x_6^2(\tau) \Big)\, d\tau$$

subject to the dynamical system

$$\begin{aligned}
\dot{x}_1(t) &= 9x_4(t) \\
\dot{x}_2(t) &= 9x_5(t) \\
\dot{x}_3(t) &= 9x_6(t) \\
\dot{x}_4(t) &= 9(17.27x_3(t) + u_1(t)) \\
\dot{x}_5(t) &= 9u_2(t) \\
\dot{x}_6(t) &= -\frac{9}{x_2(t)}\Big(27.08x_3(t) + 2x_5(t)x_6(t) + u_1(t)\Big)
\end{aligned}$$

with the initial state condition

$$x(0) = (0,\ 22,\ 0,\ 0,\ -1,\ 0)^\top,$$

the terminal state condition

$$x(1) = (10,\ 14,\ 0,\ 2.5,\ 0,\ 0)^\top,$$

the control constraints

$$\begin{aligned}
-2.83 \le u_1(t) \le 2.83, \qquad & \forall t \in [0,1], \\
-0.81 \le u_2(t) \le 0.71, \qquad & \forall t \in [0,1],
\end{aligned}$$

and the continuous state path constraints

$$\begin{aligned}
-2.5 \le x_4(t) \le 2.5, \qquad & \forall t \in [0,1], \\
-1.0 \le x_5(t) \le 1.0, \qquad & \forall t \in [0,1].
\end{aligned}$$

For two appropriately chosen positive constants ϵ and γ, the above terminal state condition can be well approximated by the following terminal state inequality constraint:

$$(x_1(1) - 10)^2 + (x_2(1) - 14)^2 + x_3^2(1) + x_4^2(1) + (x_5(1) - 2.5)^2 + x_6^2(1) \le \gamma_1.$$

Let a new state variable $x_7(t)$ be such that

$$\dot{x}_7(t) = \overline{\mathcal{L}}_\epsilon(x(t), t)$$

for any $t \in [0, 1]$, where

$$\begin{aligned}\overline{\mathcal{L}}_\epsilon(x(t), t) = \mathcal{L}_\epsilon(x_4(t) - 2.5)) + \mathcal{L}_\epsilon(-x_4(t) - 2.5)) \\ + \mathcal{L}_\epsilon(x_5(t) - 1.0)) + \mathcal{L}_\epsilon(-x_4(t) - 1.0))\end{aligned}$$

for any $t \in [0, 1]$, where $\mathcal{L}_\epsilon(\cdot)$ is as defined in the last example. Then, the above state path constraint can be well approximated by the following terminal state constraint:

$$x_7(1) - \gamma_2 \leq 0.$$

The initial control has been set to be

$$u^0(t) \equiv (2.0,\ 0.5)^\top, \qquad \forall t \in [0, 1].$$

The sampling number is set to be 70.

N^*	$\epsilon=\gamma_1=\gamma_2$	$J(u^*)$	$(x_1^*(1), x_2^*(1), x_3^*(1), x_4^*(1), x_5^*(1), x_6^*(1))$	$x_7^*(1)-\gamma_2$
60	0.005	0.00472	(9.942,14.003,-0.030,2.489,-0.002,-0.025)	1.631×10^{-7}

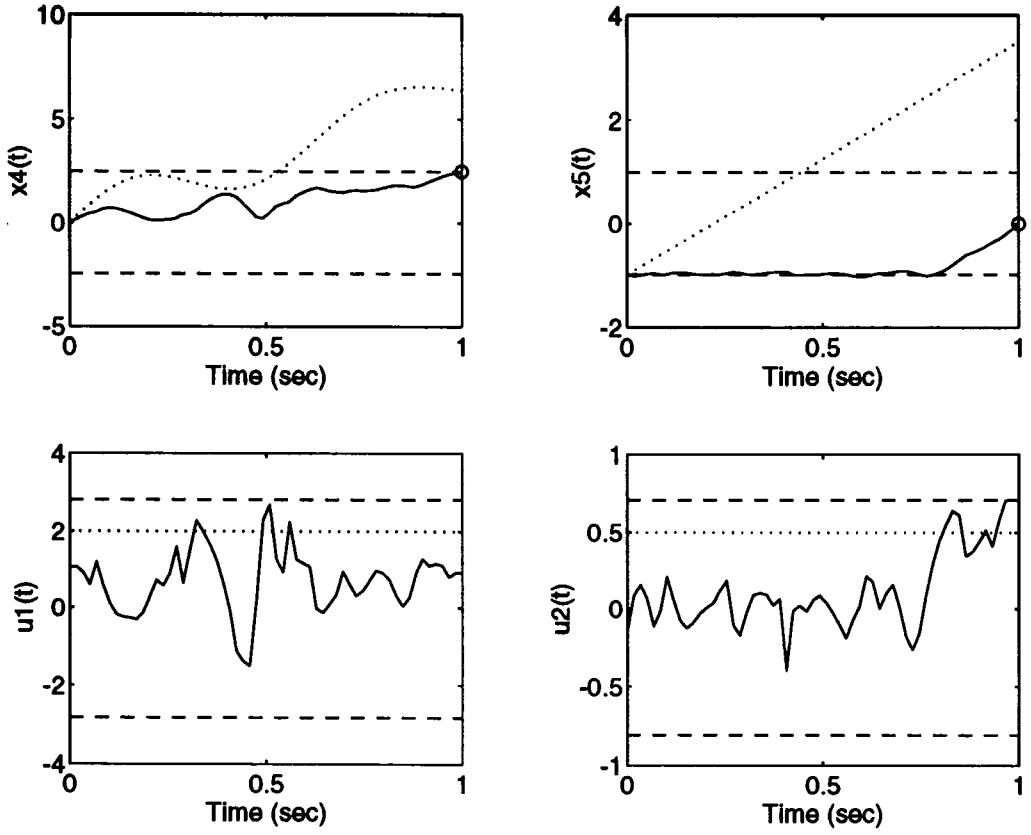


Figure 6.5: Results of Example 5, dotted line for initial results, solid line for final results, dashed line for path constraint.

Part III

Applications to Biomechanics

Chapter 7

The Optimal Control of a Movement of the Human Upper Extremity

7.1 Introduction

Because the number of muscles spanning each joint usually exceeds the number of degrees of freedom defining joint motion, the human and animal musculoskeletal system is mechanically highly redundant. In addition, many muscles can affect more than one joint, which causes complex dynamical interactions. Therefore, finding the muscle excitation patterns which provide the desired movement is difficult by trial and error for even the simplest case [84,170]. On the other hand, optimal control theory provides a unified and systematic tool for solving such problems, provided the performance measure is known.

Due to the complexity of human locomotion, dynamics models always have the following characteristics: high dimension, severe nonlinearity, complex coupling, and constraints. The nonlinearities are introduced by the generalized gravitational force terms, generalized inertial terms, and by the nonlinear behavior of muscles. The coupling becomes more evident when the mechanical system is closed-loop, and when some muscles can affect more than one joint. The constraints include the limits on controls, the joint limits and the terminal constraints. Clearly, except for some special cases [79,80,88], finding closed-form optimal control solutions is almost impossible. A

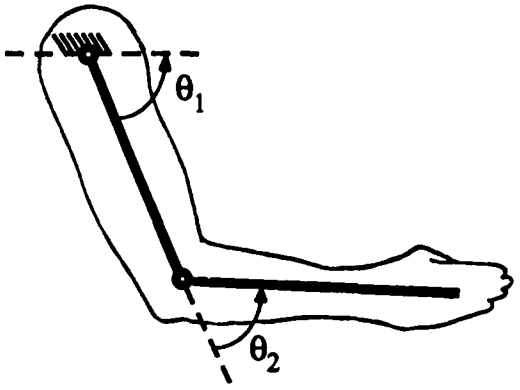


Figure 7.1: The human upper right extremity modeled as two rigid segments – the arm and the forearm, including the hand, moving in the vertical plane of the scapula.

numerical approach has to be adopted in most cases.

In this chapter, the skeletal and muscular dynamics of the human upper extremity are studied by using optimal control theory. The algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed to compute the activity which occurs in each muscle of the upper extremity when the goal is to move the arm from an initial resting position so as to touch and stop at a target with the tip of the index finger while the muscular stress, the joint constraint forces, and the neural excitations are minimized. The results obtained from the simulation describe all the major dynamic events that take place in the upper extremity when the movement is attempted.

Most results in this chapter have appeared in [89,90].

7.2 Neuro-Musculo-Skeletal Model

The human upper extremity is modeled as a two-segment, planar, articulated linkage, with the arm as one segment and the forearm including the hand as the other. The proximal end of the arm is also assumed to be a pin joint (see Figure 7.1). All movement of the system is restricted to the vertical plane of the scapula.

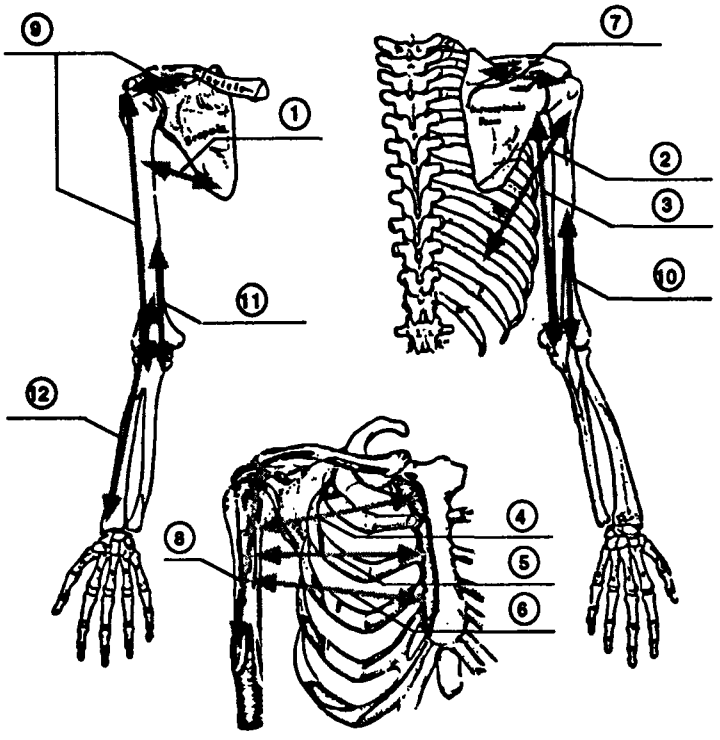


Figure 7.2: The schematic description of the actuation system that represents the human upper extremity musculature, which is from Anderson's "Grant's Atlas of Anatomy".

A total of twelve upper extremity musculotendon units provide the actuation. They are: (1) teres major, (2) latissimus dorsi, (3) triceps brachii (long and lateral heads), (4) pectoralis major (clavicular head), (5) pectoralis major (upper sternal head), (6) pectoralis major (lower sternal head), (7) supra spinatus, (8) middle deltoid, (9) biceps brachii, (10) triceps brachii (medial head, (11) brachialis, and (12) brachioradialis (see Figure 7.2 which is from [3]). The biceps brachii, brachialis and brachioradialis, which are on the anterior side of the elbow joint, are the three main elbow flexors. The triceps, which is on the posterior side of the elbow joint, is a main elbow extensor.

Each musculotendon actuator is modeled as a two-state, lumped-parameter entity, in series with tendon. Driving the musculotendon model is a first-order represen-

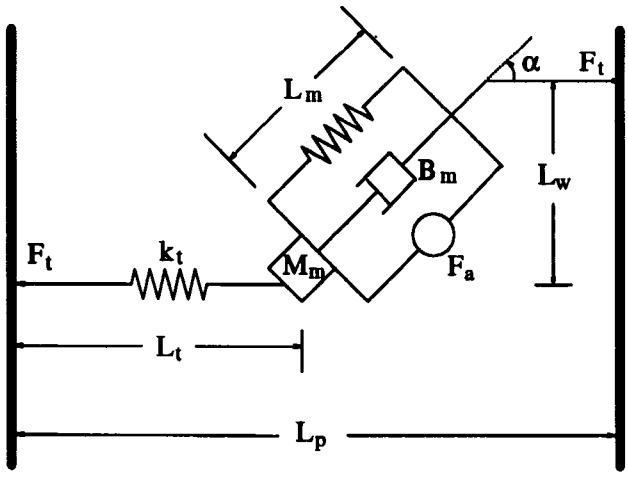


Figure 7.3: Mechanical representation of a muscle.

tation of excitation-contraction (activation) dynamics.

7.2.1 Skeletal Dynamics

The dynamical equations governing the motion of the upper extremity were derived using Newton's laws. The analysis yields the following differential equation:

$$A(\theta)\,\ddot{\theta} = B(\theta)\,\dot{\theta}^2 + C(\theta)g + D(\theta)F_t \tag{7.2.1}$$

where θ, $\dot{\theta}$, $\ddot{\theta}$ are 2×1 vectors of segment angular displacements, velocities, and accelerations; F_t is a 12×1 vector of tendon forces; $D(\theta)$ is a 2×12 moment-arm matrix which transforms muscle forces into joint torques; $A(\theta)$ is a 2×2 inertia matrix; $B(\theta)\,\dot{\theta}^2$ is a 2×1 vector describing both the centrifugal and Coriolis effects; $C(\theta)$ is a 2×1 vector describing the gravitational effect, and g is the gravity constant. The details of equation (7.2.1) can be found in [39].

7.2.2 Musculotendon Dynamics

Figure 7.3 depicts the lumped parameter model for the muscle and tendon. It contains a spring-like tendon through which the muscle force is exerted on the bones to generate movement. The muscle is in series (off axis by the pinnation angle α) with the tendon,

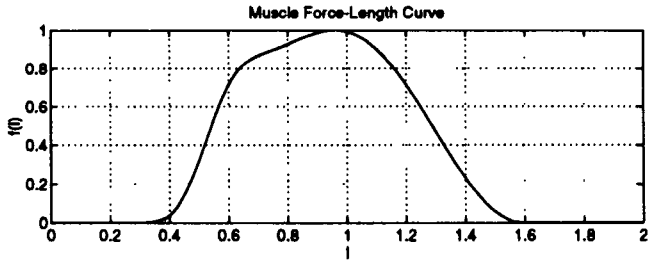


Figure 7.4: Force–length relation.

and is assumed to consist of three components: an active force generator representing the muscle contracile mechanism, a parallel elastic component representing the passive muscle elasticity, and a parallel damping component representing the viscousity of the muscle fiber fluids. The pathlength L_p is the total length between two end-points where the muscle is attached. This model is a modification, by including muscle mass, by He and Levine [49] of the dimensionless model developed by Zajac et al [169]. It was assumed that all muscle fibers are equal in length, parallel to each other, and oriented at the same pinnation angle α. The muscle volume was assumed to remain constant, or equivalently, muscle thickness L_w was assumed to remain constant during stretch or contraction.

The tendon is modeled as a nonlinear spring which exhibits an exponential force–length relation at small tendon lengths, and a linear relation for longer tendon lengths:

$$F_t(L_t) = \begin{cases} A_t(e^{k_{t_e}(L_t - L_{t_0})} - 1) & L_{t_0} \leq L_t \leq L_{t_c} \\ k_t(L_t - L_{t_c}) + F_t(L_{t_c}) & L_t > L_{t_c} \end{cases} \tag{7.2.2}$$

where F_t is the tendon force, L_t the tendon length, L_{t_0} the tendon length at rest, L_{t_c} the tendon length at which the tendon force shifts from a nonlinear relation to a linear one, and A_t, k_{t_e}, k_t are stiffness coefficients.

The active muscle force is the product of three independent factors: (1) the force–length relation $f(\bar{L}_m)$, (2) the force–velocity relation $g(\bar{L}_m)$, and (3) the muscle

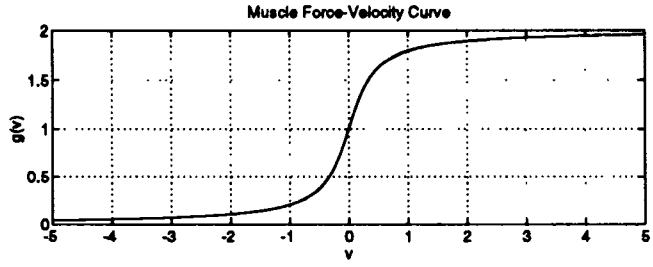


Figure 7.5: Force–velocity relation.

activation level a:

$$F_a = F_z f(\bar{L}_m) g(v) a \tag{7.2.3}$$

where $\bar{L}_m = L_m / L_z$, $v = \bar{L}_m / V_{max}$, F_a is the active muscle force, F_z the maximum isometric force, L_z the muscle length at which F_z is achieved, V_{max} the maximum shortening velocity, a the muscle activation level. The force–length relation is shown in Figure 7.4. There can be approximated by many different curve fitting techniques. Currently, the following curve-fitting formula is used:

$$f(\bar{L}_m) = \sin(b_1 \bar{L}_m^2 + b_2 \bar{L}_m + b_3) \tag{7.2.4}$$

where $b_1 = -0.9062$, $b_2 = 4.5009$, and $b_3 = -2.0239$. The force–velocity relation is shown in Figure 7.5. There can also be approximated by many different curve fitting techniques. Currently, the following curve-fitting formula is used:

$$g(v) = 1 + \arctan(c_1 v^3 + c_2 v^2 + c_3 v) \tag{7.2.5}$$

where $c_1 = -1.3166$, $c_2 = -0.4027$, and $c_3 = 2.4541$. The passive muscle force is assumed to take effect at length $L_m > L_z$, and is generated by a stretch of muscle fiber without electrical stimulation, increases exponentially with respect to muscle fiber length L_m [110]:

$$F_p(\bar{L}_m) = A_p(e^{k_{p_e}(L_m - L_z)} - 1) \tag{7.2.6}$$

where F_p is the passive muscle force, L_m the length of muscle, and A_p, k_{p_e} the stiffness coefficients. The damping force satisfies that

$$F_d = k_d \dot{\bar{L}}_m \tag{7.2.7}$$

where k_d is the damping coefficient. Although F_d has negligible effect on the muscle dynamics, it was included in this musculotendon model for the sake of completeness. The total force of a muscle is the sum of the passive force F_p, the active force F_a and the damping force F_d. According to Figure 7.3, the musculotendon dynamics is the following:

$$M_m \ddot{L}_m = F_t \cos\alpha - (F_p + F_a + B_m \dot{L}_m)\cos^2\alpha + \frac{M_m \dot{L}_m^2}{L_m}\tan^2\alpha \tag{7.2.8}$$

and

$$\alpha = \arcsin\frac{L_w}{L_m} \tag{7.2.9}$$

where L_w is the muscle thickness which is assumed to be a constant. Detailed derivation of the above equation can be found in [49].

The above musculotendon model includes two sets of parameters. One set uses non-specific muscle parameters. These parameters are assumed to be identical across all the muscles modeled. The other set includes following four muscle-specific parameters: F_z the maximum isometric force, L_z the muscle length at which F_z is achieved, M_m the muscle mass, and L_{t_0} the tendon length at rest. The details of the parameters of the twelve muscles studied can be found in [39].

7.2.3 Activation Dynamics

The activation dynamics describes the relation between the neural excitation to the muscle and its mechanical activation. The corresponding physiological process within the muscle is the release, diffusion, and uptake of the Calcium ions that control the production of sliding forces between the intermuscular filaments. The most important characteristics of activation dynamics are the different time constants for activation and deactivation, the low pass filter property, and the saturation of activation. It is assumed that the activation dynamics is independent of muscle contraction dynamics. The following first-order, nonlinear differential equation is used to describe the

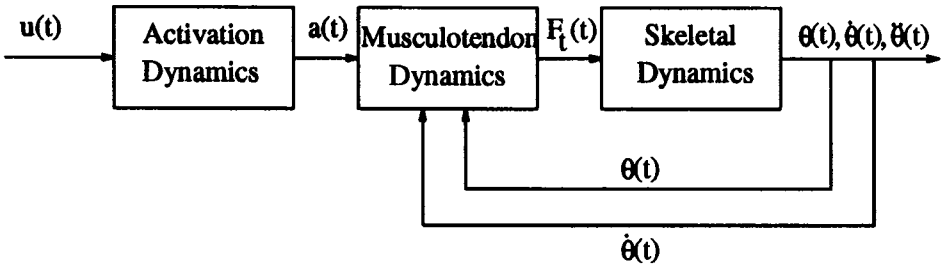


Figure 7.6: System diagram of the neuro-musculo-skeletal control system

activation dynamics:

$$\dot{a}(t) = \frac{1}{\tau_{\text{rise}}}(a_{max} - a(t))u(t) + \frac{1}{\tau_{\text{fall}}}(a_{min} - a(t))(1 - u(t)) \tag{7.2.10}$$

with

$$0 \leq u(t) \leq 1 \tag{7.2.11}$$

where $u(t)$, $\forall t \in [t_0, t_f]$, taken as the rectified electromyogram (EMG), is the neural excitation to the muscle, $a(t)$ the excitation level of the muscle, τ_{rise} the rising time constant of the excitation, τ_{fall} the falling time constant of the excitation, $a_{max}(=1)$ the higher limit of the activation level, $a_{min}(=0)$ the lower limit of the activation level. Obviously, $0 \leq a(t) \leq 1$. The above activation dynamics has the following features: (1) the mechanical activation follows EMG asymptotically, and is bounded between 0 and 1; (2) the rising of activation is faster than its decaying ($\tau_{\text{rise}} < \tau_{\text{fall}}$); (3) different muscles can have different time constants τ_{rise}, τ_{fall}, i.e. faster muscles have faster responses.

7.2.4 Complete Dynamics

It is clear that tendon force is the interface between the musculotendon dynamics and the skeletal dynamics, muscle activation couples the activation dynamics and the musculotendon dynamics. Figure 7.6 illustrates such relations. By combining the equations of the skeletal, musculotendon and activation dynamics, we obtain the complete dynamics for the neuro-musculo-skeletal control system (NMSCS) of the

human upper extremity. It can be described by the following vector form equation:

$$\dot{x}(t) = f(x(t), u(t), t), \tag{7.2.12}$$

$$x(t_0) = x_0. \tag{7.2.13}$$

In the above, $x = [x_1, \cdots, x_{40}]$ is the state vector of the system which consists of angles and angular velocities of both the arm and the forearm, the lengths, velocities and activations of the twelve muscles described above. $u = [u_1, \cdots, u_{12}]$ is the control vector of the system, where u_i is the neural excitation signal of the i-th muscle.

7.3 Optimal Control Formulation

The purpose of this chapter is to find the activity which occurs in each muscle of the upper extremity when the goal is to move the arm from an initial resting position so as to touch and stop at a target with the tip of the index finger while the muscular stress, the joint constraint forces, and the neural excitations are minimized. This can be precisely formulated as the following constrained optimal control problem:

Problem. Subject to the dynamical system (7.2.12)-(7.2.13), the hard control constraints $0 \leq u_i \leq 1, i = 1, \cdots 12$, and four terminal-state equality constraints $x_1(t_f) = \theta_1^f$, $x_2(t_f) = \theta_2^f$, $\dot{x}_1(t_f) = \dot{x}_2(t_f) = 0$, find a control vector $u = [u_1, \cdots, u_{12}]^\top$ such that the cost functional

$$J(u) = \int_{t_0}^{t_f} h(t)\, dt \tag{7.3.1}$$

with

$$\begin{aligned} h(t) = {} & s(t)^\top W_1 s(t) + F^{A/B}(t)^\top W_2\, F^{A/B}(t) + F^{N/A}(t)^\top W_3\, F^{N/A}(t) \\ & + T^{A/B}(t)^\top W_4\, T^{A/B}(t) + T^{N/A}(t)^\top W_5\, T^{N/A}(t) + u(t)^\top W_6 u(t) \end{aligned} \tag{7.3.2}$$

is minimized over control space.

In the expression of the cost integrand (7.3.2) above, the vector $s = [s_1, \cdots, s_{12}]^\top$ represents the muscular stress: $s_i = F_t^i / F_z^i$, where F_t^i is the i-th tendon force, and F_z^i

the maximum isometric force of the i-th muscle, $i = 1, \cdots, 12$. The vectors $F^{A/B} = [F_1^{A/B},\ F_2^{A/B},\ F_3^{A/B}]$ and $F^{N/A} = [F_1^{N/A},\ F_2^{N/A},\ F_3^{N/A}]$, consist of three directional components of the bearing forces at the elbow joint and the shoulder joint respectively; the vectors $T^{A/B} = [T_1^{A/B},\ T_2^{A/B},\ T_3^{A/B}]$ and $T^{N/A} = [T_1^{N/A},\ T_2^{N/A},\ T_3^{N/A}]$ consist of three directional components of the bending moments at the elbow joint and the shoulder joint respectively. Bending moment is the muscle's tendency to rotate a body segment about an axis that is perpendicular to the joint axis, and bearing force is a bone-to-bone interactive force at a joint, which includes shear, stress and tensile forces (see pages 72-74 of [39] for details). The vector $u = [u_1, \cdots, u_{12}]^\top$ represents the neural excitation. W_1 and W_6 are 12×12 diagonal constant weighting matrices, and W_i, $i = 2, 3, 4, 5$ are 3×3 diagonal constant weighting matrices.

The following discussion elaborates the rational for adopting the above cost criterion. When the upper extremity moves in a motion plane due to muscular activity, each active muscle produces forces parallel to its longitudinal axis. In general, the forces produced by different muscles are not parallel nor do they lie in the same plane. Thus, they can have components that produce moments about an axis which is perpendicular to the joint axis. Such moments do not make any contribution to the joint's motion, and at least some of them may be harmful to the integrity of the joint because they exert pressure which attempts to bend the assumed joint axis. Also, for the most part, the bearing forces at the joint center are probably compressive, which would also have a destructive impact on the integrity of the joint. When a "comfortable" movement is performed, it is reasonable to expect a self-convenient muscular harmony with minimal self-destructive effects. It is for this reason that the minimization of joint bearing forces, bending moments and muscle activations is sought in our attempt at predicting muscular activity.

Observing the dynamical system (7.2.12)-(7.2.13) of the musculo-skeletal model of the human upper extremity, it has the following characteristics: high dimension ($=40$), nonlinearity (which is introduced by the generalized gravitational force terms, generalized inertial terms, and the nonlinear behavior of muscles), and various con-

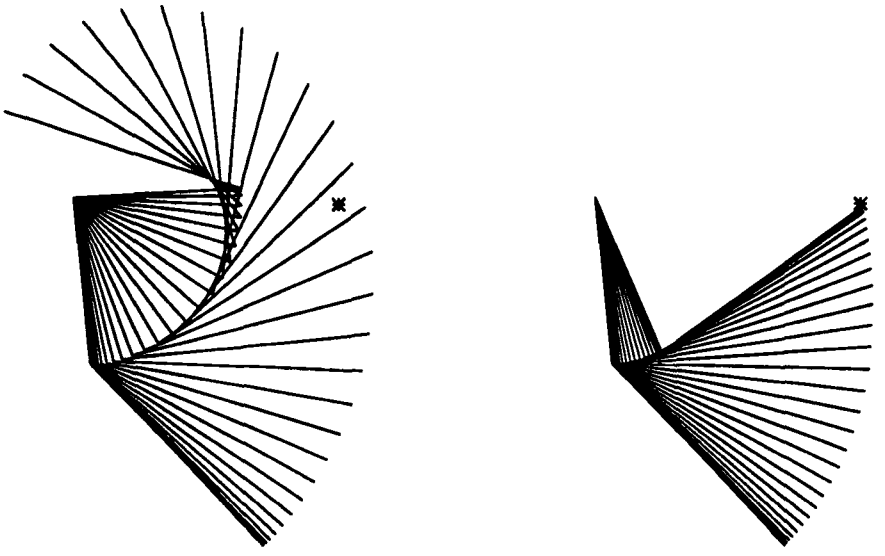


Figure 7.7: Stick figures — the left is under the initial control, the right is under the final control.

straints (which include the limits on controls, the joint limits and the terminal constraints). Clearly, finding closed-form optimal control solutions is impossible. A numerical approach has to be adopted here. In the following, the algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed in the computation.

7.4 Results

The initial state in (7.2.13) is selected to represent a rest condition of the musculoskeletal system of the human upper extremity: the initial angular velocities of the two segments are zero, the initial muscle fiber lengths of all twelve muscles are equilibria of their corresponding musculotendon dynamics (7.2.8), the initial muscle velocities and initial levels of activations of all twelve muscles are zero.

To start the optimal control algorithm, the following initial control pattern is chosen: the 7th, 8th, 9th, 11th, and 12th muscles, i.e. supra spinatus, middle deltoid,

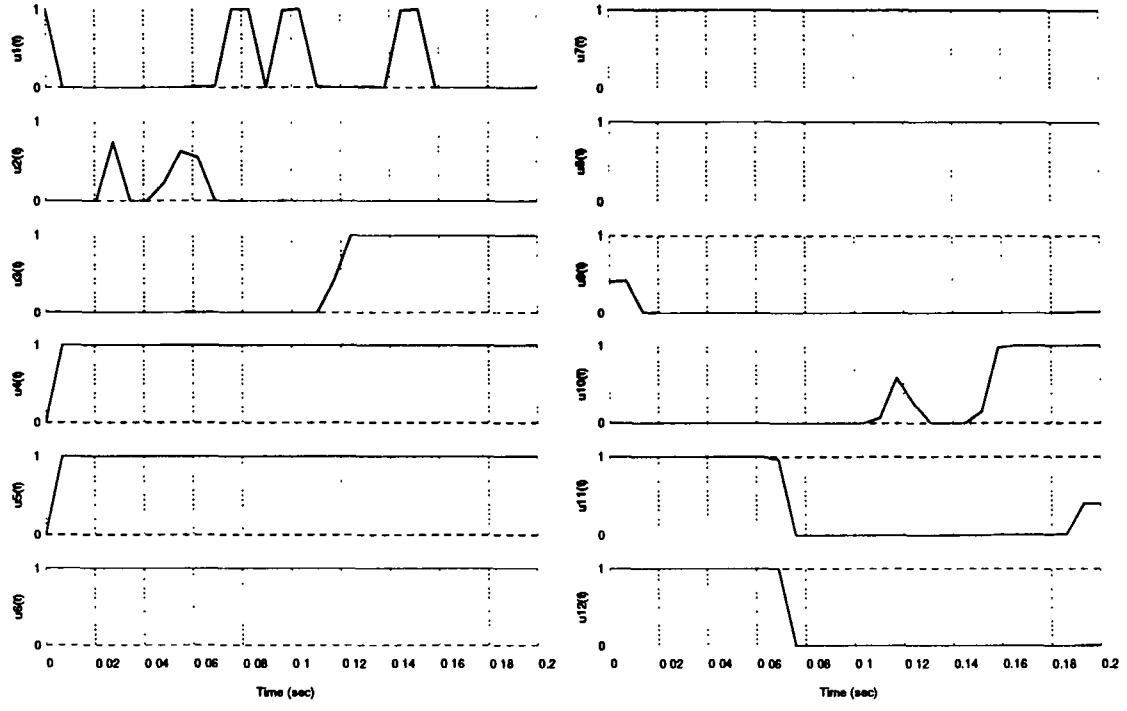


Figure 7.8: Plots for the final (in solid lines) and the initial (in dotted line) control trajectories.

biceps brachii, brachialis, and brachioradialis, are all on, the rest of the seven muscles, i.e. teres major, latissimus dorsi, triceps brachii (long and lateral heads), pectoralis major (clavicular head), pectoralis major (upper sternal head), pectoralis major (lower sternal head), and triceps brachii (medial head), are all off. It stops after 37 iterations.

Although most of the dynamic events that occur during the execution of the investigated motion cannot be confirmed empirically – for example the muscle tension histories and the joint constraint forces, it is possible to obtain data for angular trajectories. Film records were made in order to document the movement at the joints. The purpose was to compare the recorded signals with those generated by the optimal control algorithm.

In Figure 7.7, the left part shows the stick figures of the movements of the upper extremity under the initial control described above, the right part shows the stick figures of under the final control. Figure 7.9 plots the angular trajectories and

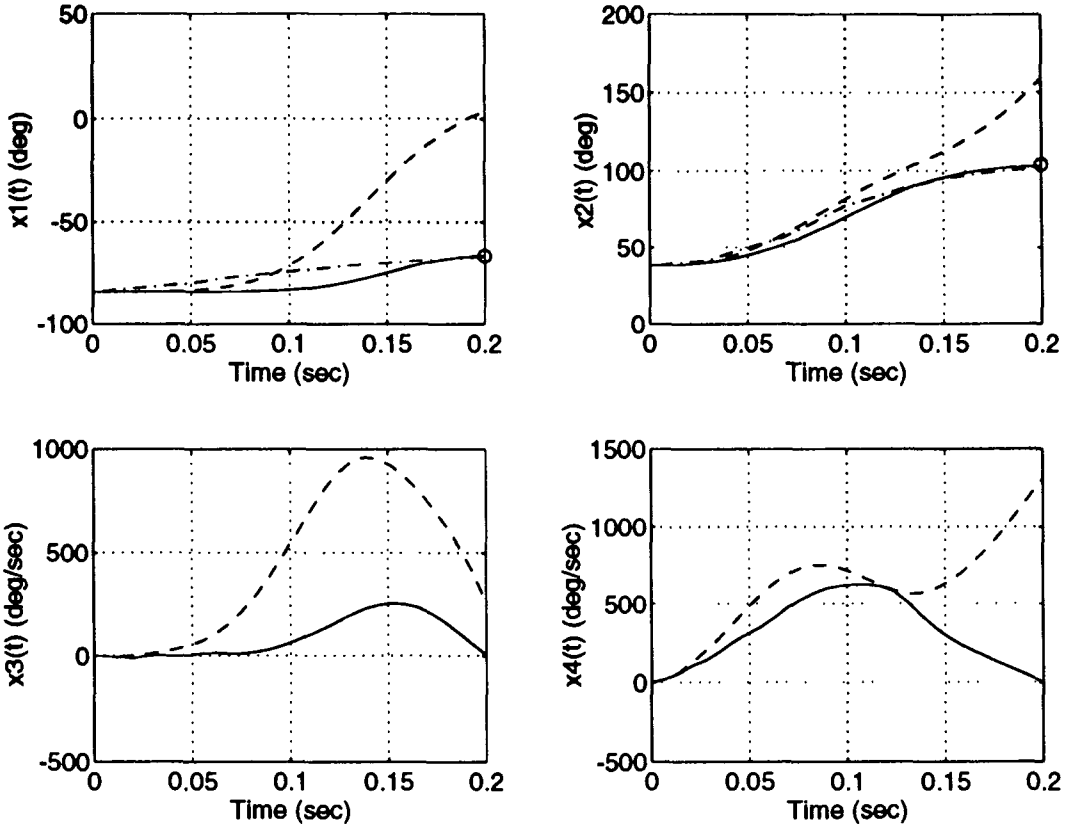


Figure 7.9: Pots for angles $\theta_1(=x_1)$, $\theta_2(=x_2)$, and angular velocities $\dot{\theta}_1(=x_3)$, $\dot{\theta}_2(=x_4)$, where the solid lines correspond to the final control, the dashed lines correspond to the initial control, the dash-dot lines correspond to the experimental data obtained from film records.

angular velocity trajectories computed by the optimal control algorithm, as well as the angular trajectories corresponding to the experiment data obtained from film records. Notice that the angular trajectories from the simulation resemble very closely those from the experiment. It is clear from both Figure 7.7 and Figure 7.8 that the final control achieves the goal very well: to excite each muscle in the upper extremity so that it moves to touch a specific target with the tip of the index finger with zero velocity.

There are two typical muscular activities: synergism and antagonism. Synergism refers to an action combined by several muscles to share a required effort, while antagonism refers to opposition between muscles whose muscular forces could result in a segment move. Figure 7.8 plots both the initial and final control patterns, showing a great deal of difference between them. The computed final control demonstrates the quantitative patterns of the synergistic activities that take place in the musculature of the upper extremity when the investigated movement is performed.

It is interesting to see that the optimal patterns of the 4-th muscle, pectoralis major (clavicular head), the 5-th muscle, pectoralis major (upper sternal head), the 6-th muscle, pectoralis major (lower sternal head), the 7-th muscle, supra spinatus, the 8-th muscle, middle deltoid, are all entirely on during the mission. It is, however, a little surprising that the optimal pattern of the 9-th muscle, biceps brachii, is entirely off during the mission. The most likely explanation is that the biceps brachii is a two joint muscle that is normally very strong. In our simulations, almost any actuation of this muscle cause the arm to overshoot the target. The frequent on-off activity of the 1-th muscle, teres major, indicates that it is very sensitive to the "soft targeting" requirement of the mission.

Also, because our complete goal is to achieve the above soft targeting while minimize the muscular stress, the joint constraint forces, and the neural excitations, one can immediately expect that all the muscle velocities must be zero at the end of the mission, which exactly happened in our simulation results.

7.5 Conclusions

A musculo-skeletal model of the human upper extremity is studied by using optimal control theory. The algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed to compute the muscular activity which occurs in each muscle of the upper extremity when the goal is to move the arm from an initial resting position so as to touch a specific target with the tip of the index finger with zero velocity while the muscular stress, the joint constraint forces, and the neural excitations are minimized. The simulation results demonstrate the quantitative patterns of the synergistic and antagonistic activities that take place in the musculature of the upper extremity when the investigated movement is performed. The angular trajectories from the simulation match closely with those recorded experimentally.

The joint angles are not especially sensitive measures of movement. It would be much better to compare measured joint angular rates and accelerations, tendon forces, and EMG with those predicted by the analysis. Heretofore there was little incentive to make such measurements because models were too primitive to predict even joint angles. Now it is worthwhile to collect more detailed experimental data to test the analytical model.

Chapter 8

The Optimal Control of Human Pedaling a Stationary Bicycle

8.1 Introduction

One of many important goals of biomechanics is to improve understanding of how the human central nervous system (CNS) coordinates muscles during multi-joint lower limb movements. The study of humans pedaling a stationary bicycle as fast as possible was motivated by the belief that this task is intermediate in complexity between human maximal height jumping and human normal locomotion. Pedaling, like maximal height jumping, requires only minimal attention to stability. Pedaling, like walking, is a periodic task involving out of phase coordination of legs.

This ongoing research has been under a collaboration between Dr. F.E. Zajac's group at the Stanford University and Dr. W.S. Levine's group here at the University of Maryland for many years. The strength of the group at Stanford has been on modeling and experiments, and the strength of our group has been on optimization and optimal control.

The skeletal and muscular dynamics of the human lower extremity has been studied by using optimal control theory. Previously, an algorithm, which was designed only to solve optimal control problems of bang-bang type, had been employed to solve

the pedaling problem [148,149]. Recently, a first-order strong-variation algorithm developed by Mayne and Polak [95], has been employed extensively to solve the pedaling problem [124,125,126,127,128].

In this chapter, the algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed to compute the activity which occurs in each muscle of the lower extremity when the goal is to pedal a stationary bicycle at maximum speed, starting from rest. The results demonstrate the effectiveness of this new algorithm.

8.2 Neuro-Musculo-Skeletal Model

The linkage model consists of two legs, each having thigh, shank, and foot-pedal segments, connected by a stationary pelvis and a crank (see Figure 8.1, which was made by C.C. Raasch from Dr. F.E. Zajac's group). The ergometer flywheel is modeled by an inertial/frictional load [35]. For simulations which allow backpedaling of the crank, the flywheel is modeled as a separate rotating body connected to the crank with a switchable constraint.

A total of nine lower limb muscles provide the forces required to move each leg. They are: SOL(SOL,OPF), TA, GAS, VAS, RF, HAM(SM,BFlh), GMAX(GM,AM), IL(ILIACUS,PSOAS), and BFsh (see Figure 8.2, which was made by C.C. Raasch from Dr. F.E. Zajac's group). Each musculotendon actuator is the same two-state model as described in Chapter 7. Also, driving the musculotendon model is the same first-order excitation-contraction dynamics as described in Chapter 7. As a result, the complete dynamics of the neuro-musculo-skeletal system of human pedaling a stationary bicycle is described by the following differential equation in vector form:

$$\dot{x}(t) = f(x(t), u(t), t), \tag{8.2.1}$$

$$x(t_0) = x_0. \tag{8.2.2}$$

In the above, $x = [x_1, \cdots, x_{100}]$ is the state vector of the system which consists of angles

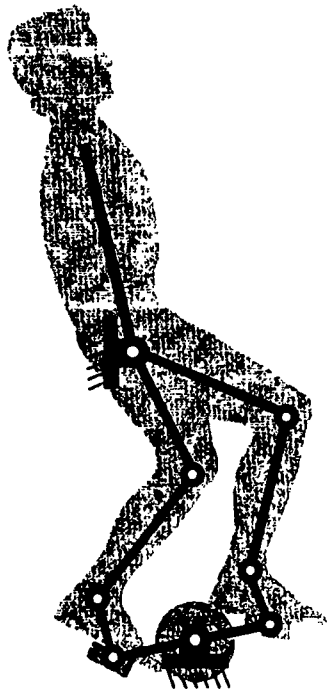


Figure 8.1: The linkage model of human pedaling a stationary bicycle.

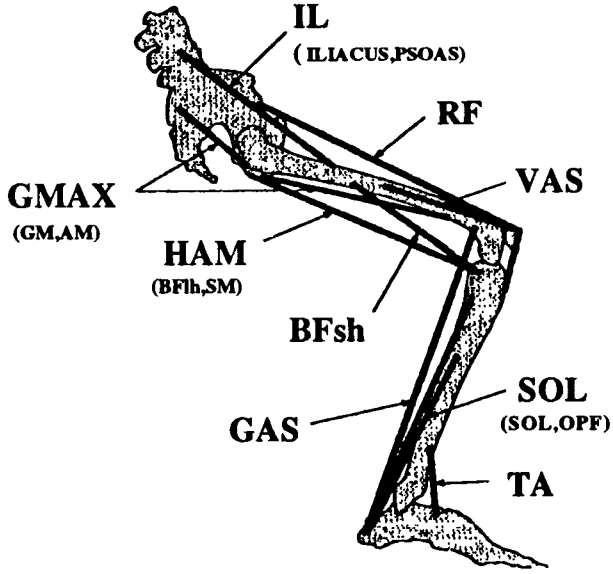


Figure 8.2: The schematic description of the actuation system that represents the human lower extremity musculature.

and angular velocities of each segment, the lengths, velocities and activations of the 18 muscles described above, $u=[u_1,\cdots,u_{18}]$ is the control vector of the system, where u_i is the neural excitation signal of the i-th muscle. Of course, because the mechanical system is a closed-loop linkage, many of the 100 state variables are dependent. For the details of the neuro-musculo-skeletal model, please see Raasch's dissertation [124].

8.3 Optimal Control Formulation

The purpose of this chapter is to find the activity which occurs in each muscle of the lower extremity when the goal is to pedal a stationary bicycle at maximum speed, starting from rest. This can be precisely formulated as the following constrained optimal control problem:

Problem. Subject to the dynamical system (8.2.1)-(8.2.2), the hard control constraints $0 \le u_i \le 1, i=1,\cdots 18$, find a control vector $u=[u_1,\cdots,u_{18}]^\top$ such that the cost functional

$$J(u) = -\Big(\theta_{\text{crank angle}}(t_f) - \theta_{\text{crank angle}}(t_0)\Big)^2 \tag{8.3.1}$$

is minimized over control space.

Similar to the biomechanics problem in Chapter 7, the musculo-skeletal model of the human pedaling a stationary bicycle also has the following characteristics: high dimension ($=100$), nonlinearity (which is introduced by the generalized gravitational force terms, generalized inertial terms, and the nonlinear behavior of muscles), and various constraints (which include the limits on controls, the joint limits and the terminal constraints). Clearly, finding closed-form optimal control solutions is impossible. A numerical approach has to be adopted here.

In the following, the algorithm, which was developed in Chapter 3 and is capable of handling generally constrained optimal control problems, is employed in the computation.

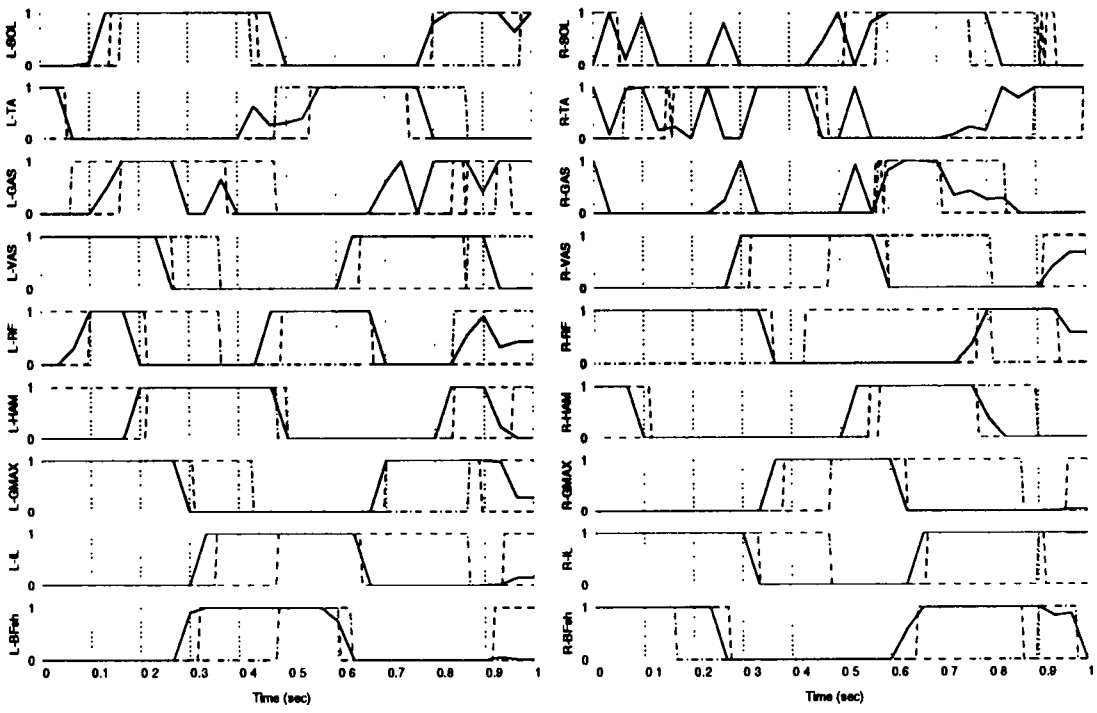


Figure 8.3: Plots for control patterns: solid line for the final control obtained by using the algorithm developed in Chapter 3, dash-dotted line for the initial control, dotted line for the optimal control obtained by C.C Raasch.

8.4 Results

The algorithm started from a set of initial control patterns which make the crank rotate 364.01 degrees and which are plotted in Figure 8.3 in dash-dotted lines. The number of discretization points of time was 30. The final control patterns, which make the crank rotate 477.44 degrees and which are plotted in Figure 8.3 in solid lines, were obtained in 6 iterations. That is, a final improvement of 113.43 degrees of the crank progress was obtained in 6 iterations.

For comparison, the first-order strong-variation algorithm developed by Mayne and Polak [95] was used to solve the same problem. The final control patterns, which make the crank rotate only 408.27 degrees, were obtained in 8 iterations. That is, a final improvement of 44.26 degrees of the crank progress was obtained in 8 iterations.

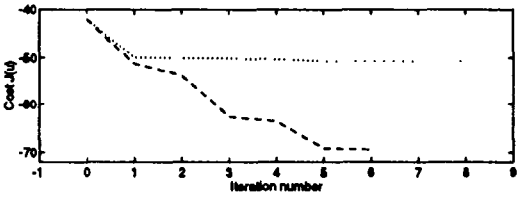


Figure 8.4: Plots of the cost versus iterations by the algorithm developed in Chapter 3 (in solid line) and by the first-order strong-variation algorithm developed by Mayne and Polak (in dotted line).

Figure 8.4 shows the simulation performance of the two algorithms.

This problem had also been solved extensively by C.C. Raasch. After tedious work moving along with the evolution of the system model, the optimal control was obtained by using the first-order strong-variation algorithm developed by Mayne and Polak [95] and by trial and error [124,125,126,127,128]. The optimal control patterns make the crank rotate 491.90 degrees, an improvement of 127.89 degrees of the crank progress. The optimal control patterns are plotted in Figure 8.3 in dotted lines.

Notice that, even though the number of discretization points of time was selected as few as 30, the final crank angle obtained by the algorithm developed in Chapter 3 was not far away from the optimal one. Better results can be expected when the discretization of time is made finer.

Judging from the above facts and from the performance curves shown in Figure 8.4, the algorithm developed in Chapter 3 can be considered as a success.

Chapter 9

Conclusions and Future Research

In this dissertation, computational methods and techniques of optimal control were studied. Motivated by the need to have an algorithm which not only is convergent but also has a fast local convergence rate, a new algorithm was developed. It was shown in the previous chapters that, first, the algorithm is globally convergent under some conditions, second, its a local convergence rate can be better than that of the first-order algorithms when some matrices are properly updated, and third, it is able to handle optimal control problems in the most general setting, namely, problems which are subjected to control constraints, path constraints, end-point constraints, a variable initial state, and a variable vector of design parameters, within a fixed/free end-time interval.

A version of the algorithm is implemented into a package which is easy to use. A variety of benchmark problems have been solved. Finally, the algorithm was employed in solving challenging biomechanics problems: (1) human moving his arm from an initial resting position so as to touch and stop at a target with the tip of the index finger while the muscular stress, the joint constraint forces, and the neural excitations are minimized; (2) human pedaling a stationary bicycle as fast as possible.

There is still much work remaining to be done. Future research is needed in the following four aspects:

First, global convergence analysis is only done on the algorithm in Chapter 3

which can handle optimal control problems with control constraints and terminal-state constraints. For the algorithm in Chapter 4 which can handle the most general optimal control problems subjected to control constraints, path constraints, end-point constraints, a variable initial state, and a variable parameter vector in a fixed end-time or free end-time interval, global convergence analysis remains to be done.

Second, the package, which implements the algorithm in Chapter 3, needs to be extended to include the algorithm in Chapter 4 as well. By then, both the constrained free end-time problems and the constrained problems with variable initial state and parameters can all be solved.

Third, the current approximation schemes for choosing the matrices Λ's appearing in both Chapter 3 and Chapter 4 are too primitive. As discussed in Chapter 3 and Chapter 4, the local convergence rate of the algorithms there depends crucially on the tightness of the approximations of those matrices. The tighter those approximations are, the closer the convergence rates of the algorithms are to second order. Let us recall that, in the Han-Powell algorithm, which is a general purpose algorithm for solving optimization problems in finite-dimensional space, the Hessian is replaced by a positive definite matrix updated by a certain rule. When the BFGS rule is adapted, the algorithm will have a superlinear convergence rate locally near the solution, if the step lengths are taken equal to unity. Similarly, to help the algorithms in Chapter 3 and Chapter 4 to achieve a local convergence rate better than that of the first-order algorithms, more sophisticated schemes than the ones shown in Chapter 6 are needed to be developed for those matrices Λ's.

Fourth, a common feature of the algorithms in Chapter 3 and Chapter 4 is that it is required to solve a generally constrained linear quadratic regulator problem (LQR) at each iteration. At the present time, those "direction-finding" subproblems are converted to quadratic programming problems after the controls are approximated by piecewise constants. The drawback of this approach is that the size of the quadratic programming problems can become very big when the dimension of the state of the optimal control problem and the number of discretization points of time are large.

Because the constrained LQR problem, as formulated in Chapter 6, is important in its own right, better techniques to solve the constrained LQR problem which can exploit as much of the intrinsic properties of the problem as possible, rather than simply converting it to a quadratic programming problem, are needed.

Appendix A

Some Basic Results Used in Chapter 3

Lemma A.1 *Suppose $\phi(t)$ has continuous derivatives up to the third-order, during $t \in [0,1]$, then*

$$\begin{aligned} \phi(1) - \phi(0) - \phi'(0) &= \int_0^1 (1-t)\phi''(t)dt \\ \phi(1) - \phi(0) - \phi'(0) - \frac{1}{2}\phi''(0) &= \int_0^1 \frac{1}{2}(1-t)^2\phi'''(t)dt. \end{aligned}$$

Proof: It can be easily checked by integrating by parts. □

Lemma A.2 *Suppose $f: \mathcal{R}^n \to \mathcal{R}$ has continuous derivatives up to the third-order, then for any $x \in \mathcal{R}^n$, $y \in \mathcal{R}^n$,*

$$\begin{aligned} & f(y) - f(x) - f_x(x)(y-x) \\ & \qquad = \int_0^1 (1-t) f_{xx}(x + t(y-x))(y-x)^2 dt \qquad \text{(A.1)} \\ & f(y) - f(x) - f_x(x)(y-x) - \frac{1}{2} f_{xx}(x)(y-x)^2 \\ & \qquad = \int_0^1 \frac{1}{2}(1-t)^2 f_{xxx}(x + t(y-x))(y-x)^3 dt. \qquad \text{(A.2)} \end{aligned}$$

Proof: Let $\phi(t) = f(x + t(y-x))$ for some $t \in [0,1]$. Then

$$\begin{aligned} \phi'(t) &= f_x(x + t(y-x))(y-x) \\ \phi''(t) &= f_{xx}(x + t(y-x))(y-x)^2 \\ \phi'''(t) &= f_{xxx}(x + t(y-x))(y-x)^3. \end{aligned}$$

Noting that $\phi(1)=f(y)$, $\phi(0)=f(x)$, $\phi'(0)=f_x(x)(y-x)$ and $\phi''(0)=f_{xx}(x)(y-x)^2$, the proof is then completed when Lemma A.1 is applied. □

Lemma A.3 *Suppose F: $\mathcal{R}^n \times \mathcal{R}^m \to \mathcal{R}$ has continuous derivatives up to the third-order, then for any $x^{u^2}, x^{u^1} \in \mathcal{R}^n$, any $u^2, u^1 \in \mathcal{R}^m$, with $\Delta x = x^{u^2} - x^{u^1}$, $\Delta u = u^2 - u^1$, $\bar{x}(t) = x^{u^1} + t(x^{u^2} - x^{u^1})$, $\bar{u}(t) = u^1 + t(u^2 - u^1)$,*

$$
\begin{aligned}
&F(x^{u^2}, u^2) - F(x^{u^1}, u^1) = \\
&\quad F_x(x^{u^1}, u^1)\Delta x + F_u(x^{u^1}, u^1)\Delta u + \int_0^1 (1-t)F_2(t)\,dt \qquad \text{(A.3)} \\
&F(x^{u^2}, u^2) - F(x^{u^1}, u^1) = \\
&\quad F_x(x^{u^1}, u^1)\Delta x + F_u(x^{u^1}, u^1)\Delta u + \frac{1}{2}F_{xx}(x^{u^1}, u^1)\Delta x^2 \\
&\quad + \frac{1}{2}F_{xu}(x^{u^1}, u^1)\Delta x \Delta u + \frac{1}{2}F_{ux}(x^{u^1}, u^1)\Delta u \Delta x \\
&\quad + \frac{1}{2}F_{xx}(x^{u^1}, u^1)\Delta u^2 + \int_0^1 \frac{1}{2}(1-t)^2 F_3(t)\,dt \qquad \text{(A.4)}
\end{aligned}
$$

where,

$$
\begin{aligned}
F_2(t) &= F_{xx}(\bar{x}(t), \bar{u}(t))\Delta x^2 + F_{ux}(\bar{x}(t), \bar{u}(t))\Delta u \Delta x \\
&\quad + F_{xu}(\bar{x}(t), \bar{u}(t))\Delta x \Delta u + F_{uu}(\bar{x}(t), \bar{u}(t))\Delta u^2 \\
F_3(t) &= F_{xxx}(\bar{x}(t), \bar{u}(t))\Delta x^3 + F_{uxx}(\bar{x}(t), \bar{u}(t))\Delta u \Delta x^2 \\
&\quad + F_{xxu}(\bar{x}(t), \bar{u}(t))\Delta x^2 \Delta u + F_{xux}(\bar{x}(t), \bar{u}(t))\Delta x \Delta u \Delta x \\
&\quad + F_{uux}(\bar{x}(t), \bar{u}(t))\Delta u^2 \Delta x + F_{uxu}(\bar{x}(t), \bar{u}(t))\Delta u \Delta x \Delta u \\
&\quad + F_{xuu}(\bar{x}(t), \bar{u}(t))\Delta x \Delta u^2 + F_{uuu}(\bar{x}(t), \bar{u}(t))\Delta u^3.
\end{aligned}
$$

Proof: Let $\phi(t) = F(x^{u^1} + t(x^{u^2} - x^{u^1}), u^1 + t(u^2 - u^1))$ for some $t \in [0,1]$. Then

$$
\begin{aligned}
\phi'(t) &= F_x(\bar{x}(t), \bar{u}(t))\Delta x + F_u(\bar{x}(t), \bar{u}(t))\Delta u \\
\phi''(t) &= F_{xx}(\bar{x}(t), \bar{u}(t))\Delta x^2 + F_{ux}(\bar{x}(t), \bar{u}(t))\Delta u \Delta x \\
&\quad + F_{xu}(\bar{x}(t), \bar{u}(t))\Delta x \Delta u + F_{uu}(\bar{x}(t), \bar{u}(t))\Delta u^2 \\
\phi'''(t) &= F_{xxx}(\bar{x}(t), \bar{u}(t))\Delta x^3 + F_{uxx}(\bar{x}(t), \bar{u}(t))\Delta u \Delta x^2 \\
&\quad + F_{xxu}(\bar{x}(t), \bar{u}(t))\Delta x^2 \Delta u + F_{xux}(\bar{x}(t), \bar{u}(t))\Delta x \Delta u \Delta x
\end{aligned}
$$

$$+F_{uux}(\bar{x}(t),\bar{u}(t))\Delta u^2\Delta x + F_{uxu}(\bar{x}(t),\bar{u}(t))\Delta u\Delta x\Delta u$$
$$+F_{xuu}(\bar{x}(t),\bar{u}(t))\Delta x\Delta u^2 + F_{uuu}(\bar{x}(t),\bar{u}(t))\Delta u^3.$$

The proof then follows by recognizing $\phi(1)$, $\phi(0)$, $\phi'(0)$ and $\phi''(0)$ from the above expressions and by applying Lemma A.1. □

Lemma A.4 *There exists an $N \in (0,\infty)$ such that, for all $u \in \mathcal{U}$*

$$\| x^u(t) \| \leq N \tag{A.5}$$
$$\| p^u(t) \| \leq N \tag{A.6}$$

Proof: See Proposition 6.2 in [95]. □

Lemma A.5 *There exists a $c \in (0,\infty)$ such that, for all $u^2, u^1 \in \mathcal{U}$*

$$\sup_{t\in[t_0,t_f]} \| x^{u^2}(t) - x^{u^1}(t) \| \leq c \| u^2 - u^1 \| \tag{A.7}$$
$$\sup_{t\in[t_0,t_f]} \| p^{u^2}(t) - p^{u^1}(t) \| \leq c \| u^2 - u^1 \| \tag{A.8}$$

Proof: See Proposition 6.3 in [95]. □

Lemma A.6 *There exists a $c \in (0,\infty)$ such that, for all $u^2, u^1 \in \mathcal{U}$*

$$\sup_{t\in[t_0,t_f]} \| x^{u^2}(t) - x^{u^1}(t) - y^v(t) \| \leq c \| u^2 - u^1 \|^2 \tag{A.9}$$

where $y^v(t)$ is defined in (3.4.1).

Proof: Let $\epsilon(t) = x^{u^2}(t) - x^{u^1}(t) - y^v(t)$. Applying (A.1), together with $v = u^2 - u^1$ and $\epsilon(t_0) = 0$,

$$\begin{aligned}\epsilon(t) &= \int_{t_0}^{t} (\dot{x}^{u^2}(s) - \dot{x}^{u^1}(s) - \dot{y}^v(s))\, ds \\ &= \int_{t_0}^{t} (f(x^{u^2}(s), u^2(s), s) - f(x^{u^1}(s), u^2(s), s) - f_x(x^{u^1}(s), u^1(s), s)\, y^v(s))\, ds \\ &\quad + \int_{t_0}^{t} (f(x^{u^1}(s), u^2(s), s) - f(x^{u^1}(s), u^1(s), s) - f_u(x^{u^1}(s), u^1(s), s)\, v(s))\, ds\end{aligned}$$

$$
\begin{aligned}
&= \int_{t_0}^{t} (f_x(x^{u^1}(s), u^2(s), s)\,(x^{u^2}(s) - x^{u^1}(s)) - f_x(x^{u^1}(s), u^1(s), s)\,y^v(s))\,ds \\
&+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{xx}(\bar{x}(\tau,s), u^2(s), s)(x^{u^2}(s) - x^{u^1}(s))^2\,d\tau\,ds \\
&+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{uu}(x^{u^1}(s), \bar{u}(\tau,s), s)(u^2(s) - u^1(s))^2\,d\tau\,ds \\
&= \int_{t_0}^{t} f_x(x^{u^1}(s), u^1(s), s)\,\epsilon(s)\,ds \\
&+ \int_{t_0}^{t} (f_x(x^{u^1}(s), u^2(s), s) - f_x(x^{u^1}(s), u^1(s), s))(x^{u^2}(s) - x^{u^1}(s))\,ds \\
&+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{xx}(\bar{x}(\tau,s), u^2(s), s)(x^{u^2}(s) - x^{u^1}(s))^2\,d\tau\,ds \\
&+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{uu}(x^{u^1}(s), \bar{u}(\tau,s), s)(u^2(s) - u^1(s))^2\,d\tau\,ds \\
&= \int_{t_0}^{t} f_x(x^{u^1}(s), u^1(s), s)\,\epsilon(s)\,ds \\
&+ \int_{t_0}^{t} f_{xu}(x^{u^1}(s), \bar{u}(s), s)(u^2(s) - u^1(s))(x^{u^2}(s) - x^{u^1}(s))\,ds \\
&+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{xx}(\bar{x}(\tau,s), u^2(s), s)(x^{u^2}(s) - x^{u^1}(s))^2\,d\tau\,ds \\
&+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{uu}(x^{u^1}(s), \bar{u}(\tau,s), s)(u^2(s) - u^1(s))^2\,d\tau\,ds
\end{aligned}
$$

where $\bar{x}(\tau,s) = x^{u^1}(s) + \tau(x^{u^2}(s) - x^{u^1}(s))$, $\bar{u}(\tau,s) = u^1(s) + \tau(u^2(s) - u^1(s))$, and, $\bar{u}(s) = u^1(s) + \tau(s)(u^2(s) - u^1(s))$, for some $\tau(s) \in [0,1]$. Hence, by making use of the continuity of f_x, f_{xx} and f_{uu} on the compact set $\{\,\|x\| \le N\,\} \times \Omega \times [t_0, t_f]$, together with the fact that $x^{u^2}(s) - x^{u^1}(s)$ is uniformly bounded by $\|u^2 - u^1\|$ on $[t_0, t_f]$ by Lemma A.5, then there exists $c', c'' \in (0, \infty)$, such that

$$
\| \epsilon(t) \| \;=\; c' \int_{t_0}^{t} \| \epsilon(\tau) \|\, ds \;+\; c'' \| u^2 - u^1 \|^2.
$$

The result then follows from the Bellman-Gronwall inequality. □

Lemma A.7 *Cost functional $J(u)$, and terminal constraints functionals $g_i(u)$, $i = 1, \cdots, r$ are all continuous in $u \in \mathcal{U}$.*

Proof: According to Assumption 3.2.1, K_x is continuous in $x(t_f)$, L_x and L_u are continuous in $x(t)$ and $u(t)$. So, for bounded control $u \in \mathcal{U}$, its corresponding state

x^u is also bounded, by Lemma A.4. Then for any $u^1, u^2 \in \mathcal{U}$, there exist constants $c_1, c_2, c_3, c_4 > 0$, such that,

$$\begin{aligned}
&|J(u^2) - J(u^1)| \\
&\leq |K(x^{u^2}(t_f), t_f) - K(x^{u^1}(t_f), t_f)| + \int_{t_0}^{t_f} |L(x^{u^2}, u^2, t) - L(x^{u^1}, u^1, t)|\, dt \\
&\leq c_1 \|x^{u^2}(t_f) - x^{u^1}(t_f)\| + \int_{t_0}^{t_f} (c_2 \|x^{u^2}(t) - x^{u^1}(t)\| + c_3 \|u^2(t) - u^1(t)\|)\, dt \\
&\leq c_4 \|u^2 - u^1\|,
\end{aligned}$$

which implies the continuity of $J(u)$ in $u \in \mathcal{U}$. Similarly, the continuity of $g_i(u)$, $i = 1, \cdots, r$ in $u \in \mathcal{U}$ can also be proved. □

Lemma A.8 *There exists $c_1, c_2 \in (0, \infty)$ such that, for all u^1, $u^2 \in \mathcal{U}$,*

$$\left| (J(u^2) - J(u^1)) \; - \; \Delta J(u^1)(v) \right| \; \leq \; c_1 \| \, v \, \|^2 \tag{A.10}$$

$$\left| (J(u^2) - J(u^1)) \; - \; (\Delta J(u^1)(v) + \Delta^2 J(u^1)(v)) \right| \; \leq \; c_2 \| \, v \, \|^3 \tag{A.11}$$

where $v = u^2 - u^1$.

Proof: We first prove (A.11). From cost functional (3.2.8) and costate equation (3.3.3),

$$\begin{aligned}
&J(u^2) - J(u^1) \\
&= K(x^{u^2}(t_f), t_f) - K(x^{u^1}(t_f), t_f) - \int_{t_0}^{t_f} p^{u^1}(t)^\top (\dot{x}^{u^2}(t) - \dot{x}^{u^1}(t))\, dt \\
&\quad + \int_{t_0}^{t_f} H(x^{u^2}(t), u^2(t), p^{u^1}(t), t) - H(x^{u^1}(t), u^1(t), p^{u^1}(t), t)\, dt \\
&= K(x^{u^2}(t_f), t_f) - K(x^{u^1}(t_f), t_f) - p^{u^1}(t)^\top (x^{u^2}(t) - x^{u^1}(t)) \Big|_{t_0}^{t_f} \\
&\quad + \int_{t_0}^{t_f} H(x^{u^2}(t), u^2(t), p^{u^1}(t), t) - H(x^{u^1}(t), u^1(t), p^{u^1}(t), t)\, dt \\
&\quad + \int_{t_0}^{t_f} \dot{p}^{u^1}(t)^\top (x^{u^2}(t) - x^{u^1}(t))\, dt.
\end{aligned}$$

Because $x^{u^2}(t_0) = x^{u^1}(t_0) = x_0$, together with the terminal condition (3.4.5) of the costate,

$$J(u^2) - J(u^1)$$

$$
\begin{aligned}
&= K(x^{u^2}(t_f), t_f) - K(x^{u^1}(t_f), t_f) - K_x(x^{u^1}(t_f), t_f)(x^{u^2}(t_f) - x^{u^1}(t_f)) \\
&\quad + \int_{t_0}^{t_f} (H(x^{u^2}(t), u^2(t), p^{u^1}(t), t) - H(x^{u^1}(t), u^1(t), p^{u^1}(t), t))\, dt \\
&\quad - \int_{t_0}^{t_f} H_x(x^{u^1}(t), u^1(t), p^{u^1}(t), t)(x^{u^2}(t) - x^{u^1}(t))\, dt.
\end{aligned}
$$

Applying equation (A.2) to function K, and equation (A.4) to function H,

$$
\begin{aligned}
&J(u^2) - J(u^1) \\
&= \int_{t_0}^{t_f} H_u^{(1)}(t)\,(u^2(t) - u^1(t))\, dt \\
&\quad + \frac{1}{2}\,(x^{u^2}(t_f) - x^{u^1}(t_f))^{\mathsf{T}}\, K_{xx}^{(1)}(t_f)\,(x^{u^2}(t_f) - x^{u^1}(t_f)) \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f} \begin{pmatrix} x^{u^2}(t) - x^{u^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} H_{xx}^{(1)}(t) & H_{xu}^{(1)}(t) \\ H_{ux}^{(1)}(t) & H_{uu}^{(1)}(t) \end{pmatrix} \begin{pmatrix} x^{u^2}(t) - x^{u^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix} dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xxx}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta x^3(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uxx}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta u(t)\Delta x^2(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xxu}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta x^2(t)\Delta u(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xux}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta x(t)\Delta u(t)\Delta x(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uux}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta u^2(t)\Delta x(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uxu}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta u(t)\Delta x(t)\Delta u(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xuu}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta x(t)\Delta u^2(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uuu}(\bar{x}(\tau,t), \bar{u}(\tau,t), p^{u^1}(t), t)\Delta u^3(t)\, d\tau\, dt \\
&\quad + \frac{1}{2}\int_0^1 (1-\tau)^2 K_{xxx}(\bar{x}(\tau,t_f), t_f)\Delta x^3(t_f)\, d\tau
\end{aligned}
$$

where $\Delta x(t) = x^{u^2}(t) - x^{u^1}(t)$, $\Delta u(t) = u^2(t) - u^1(t)$, $\bar{x}(\tau,t) = x^{u^1}(t) + \tau(x^{u^2}(t) - x^{u^1}(t))$, $\bar{u}(\tau,t) = u^1(t) + \tau(u^2(t) - u^1(t))$, and $H_x^{(1)}(t)$, $H_{xx}^{(1)}(t)$, $H_{xu}^{(1)}(t)$, $H_{ux}^{(1)}(t)$ and $H_{uu}^{(1)}(t)$ are evaluated at $(x^{u^1}(t),\, p^{u^1}(t),\, u^1(t),\, t)$, and $K_{xx}^{(1)}(t_f)$ is evaluated at $(x^{u^1}(t_f),\, t_f)$. Hence, by making use of the continuity of K_{xxx} on the compact set $\{\, \|x\| \le N \,\} \times [t_0, t_f]$, and of H_{xxx}, H_{uxx}, H_{xxu}, H_{xux}, H_{uux}, H_{uxu}, H_{xuu} and H_{uuu} on the compact set

$\{\,\|x\| \le N\,\} \times \Omega \times \{\,\|p\| \le N\,\} \times [t_0, t_f]$, together with the fact that $x^u(t)$ is uniformly bounded on $[t_0, t_f]$ by Lemma A.4, and that $x^{u^2}(t) - x^{u^1}(t)$ is uniformly bounded by $\|u^2 - u^1\|$ on $[t_0, t_f]$ by Lemma A.5, then there exists a $c' \in (0, \infty)$, such that

$$\left| \, (J(u^2) - J(u^1)) - \Delta J(u^1)(u^2 - u^1) - \tilde{\Delta}^2 J(u^1)(u^2 - u^1) \, \right| \;\le\; c' \, \|u^2 - u^1\|^3$$

where

$$\begin{aligned}
&\tilde{\Delta}^2 J(u^1)(u^2 - u^1) \\
&\quad = \frac{1}{2} \, (x^{u^2}(t_f) - x^{u^1}(t_f))^\top K_{xx}^{(1)}(t_f) \, (x^{u^2}(t_f) - x^{u^1}(t_f)) \\
&\qquad + \frac{1}{2} \int_{t_0}^{t_f} \begin{pmatrix} x^{u^2}(t) - x^{u^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix}^\top \begin{pmatrix} H_{xx}^{(1)}(t) & H_{xu}^{(1)}(t) \\ H_{ux}^{(1)}(t) & H_{uu}^{(1)}(t) \end{pmatrix} \begin{pmatrix} x^{u^2}(t) - x^{u^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix} dt. \qquad \text{(A.12)}
\end{aligned}$$

To complete the proof of (A.11), it remains to prove that there exists a $c'' \in (0, \infty)$, such that

$$\left| \, \tilde{\Delta}^2 J(u^1)(u^2 - u^1) \;-\; \Delta^2 J(u^1)(u^2 - u^1) \, \right| \;\le\; c'' \, \|u^2 - u^1\|^3. \qquad \text{(A.13)}$$

Let $\epsilon(t) = x^{u^2}(t) - x^{u^1}(t) - y^v(t)$, from (A.12) above and (3.4.4),

$$\begin{aligned}
&\tilde{\Delta}^2 J(u^1)(u^2 - u^1) \;-\; \Delta^2 J(u^1)(u^2 - u^1) \\
&\quad = -\epsilon(t_f)^\top K_{xx}^{(1)}(t_f) \, (x^{u^2}(t_f) - x^{u^1}(t_f)) \;+\; \frac{1}{2} \, \epsilon(t_f)^\top K_{xx}^{(1)}(t_f) \, \epsilon(t_f) \\
&\qquad - \int_{t_0}^{t_f} \epsilon(t)^\top H_{xx}^{(1)}(t) \, (x^{u^2}(t) - x^{u^1}(t)) \, dt \;+\; \frac{1}{2} \int_{t_0}^{t_f} \epsilon(t)^\top H_{xx}^{(1)}(t) \, \epsilon(t) \, dt \\
&\qquad - \int_{t_0}^{t_f} \epsilon(t)^\top H_{xu}^{(1)}(t) \, (u^2(t) - u^1(t)) \, dt.
\end{aligned}$$

By making use of the continuity of K_{xxx} on the compact set $\{\,\|x\| \le N\,\} \times [t_0, t_f]$, and of H_{xxx} and H_{uuu} on the compact set $\{\,\|x\| \le N\,\} \times \Omega \times \{\,\|p\| \le N\,\} \times [t_0, t_f]$, together with the fact that $x^u(t)$ is uniformly bounded on $[t_0, t_f]$ by Lemma A.4, and that $x^{u^2}(t) - x^{u^1}(t)$ is uniformly bounded by $\|u^2 - u^1\|$ on $[t_0, t_f]$ by Lemma A.5, and that $\epsilon(t)$ is uniformly bounded by $\|u^2 - u^1\|^2$ on $[t_0, t_f]$ by Lemma A.6, (A.13) then follows. The proof of (A.11) is now complete.

Finally, by making use of the continuity of K_{xx} on the compact set $\{\,\|x\| \le N\,\} \times [t_0, t_f]$, and of H_{xx}, H_{xu}, H_{ux} and H_{uu} on the compact set $\{\,\|x\| \le N\,\} \times \Omega \times \{\,\|p\| \le$

$N\} \times [t_0, t_f]$, together with the fact that $y^v(t)$ is uniformly bounded by $\|u^2 - u^1\|$ on $[t_0, t_f]$ by Lemma A.6, the proof of (A.10) then follows. □

Lemma A.9 *Let $x(t)$ be the solution of*

$$\begin{aligned} \dot{x}(t) &= A(t)\,x(t) + B(t)\,u(t) \\ x(t_d) &= d, \end{aligned}$$

and $x'(t)$ be the solution of

$$\begin{aligned} \dot{x}'(t) &= A'(t)\,x'(t) + B'(t)\,u'(t) \\ x'(t_d) &= d'. \end{aligned}$$

In the above, t_d could be any value in $[t_0, t_f]$. Assume that $A(t)$, $A'(t)$, $B(t)$, $B'(t)$ are continuous, and that there exist positive numbers M_a, M_b, M_u such that

$$\|A(t)\|, \|A'(t)\| \le M_a, \quad \|B(t)\|, \|B'(t)\| \le M_b, \quad \|u(t)\|, \|u'(t)\| \le M_u$$

for any $t \in [t_0, t_f]$. Then, there exists a positive number c such that

$$\|x - x'\| \le c\,\epsilon$$

where

$$\epsilon = \max\{\|A - A'\|, \|B - B'\|, \|u - u'\|, \|d - d'\|\}.$$

Proof: From the continuity of $A(t)$, $A'(t)$, $B(t)$ and $B'(t)$, and the uniform boundedness of $A(t)$, $A'(t)$, $B(t)$, $B'(t)$, $u(t)$ and $u'(t)$, Lemma A.4 holds. That is, there exists a positive number M_x such that

$$\|x(t)\|, \|x'(t)\| \le M_x$$

for any $t \in [t_0, t_f]$. Consider the case when $t_d \le t \le t_f$. From the differential equations,

$$\begin{aligned} &\|x(t) - x'(t)\| \\ &\le \|d - d'\| + \int_{t_d}^{t} (\|A(\tau)x(\tau) - A'(\tau)x'(\tau)\| + \|B(\tau)u(\tau) - B'(\tau)u'(\tau)\|)\,d\tau \end{aligned}$$

$$
\begin{aligned}
&\leq \|d-d'\| + \int_{t_d}^{t} \|A(\tau)-A'(\tau)\|\|x(\tau)\|\,d\tau + \int_{t_d}^{t} \|A'(\tau)\|\|x(\tau)-x'(\tau)\|\,d\tau \\
&\quad + \int_{t_d}^{t} \|B(\tau)-B'(\tau)\|\|u(\tau)\|\,d\tau + \int_{t_d}^{t} \|B'(\tau)\|\|u(\tau)-u'(\tau)\|\,d\tau \\
&\leq c'\epsilon + \int_{t_d}^{t} M_a\|x(\tau)-x'(\tau)\|\,d\tau
\end{aligned}
$$

where

$$c' = 1 + (t_f - t_0)(M_x + M_b + M_u).$$

Applying Gronwall Inequality, we then have

$$\|x(t)-x'(t)\| \leq (c'\epsilon)\exp\left[\int_{t_d}^{t} M_a\,dt\right] \leq c\,\epsilon$$

where

$$c = c'\exp[M_a(t_f - t_d)].$$

The case when $t_0 \leq t \leq t_d$ follows from the same approach as above. □

Appendix B

Some Basic Results Used in Chapter 4

Lemma B.1 *There exists an $N \in (0, \infty)$ such that, for any $u \in \mathcal{U}$, any $x_0 \in \mathcal{S}$,*

$$\| x^{u,x_0}(t) \| \leq N \tag{B.1}$$

$$\| p^{u,x_0}(t) \| \leq N \tag{B.2}$$

Proof: For any $u \in \mathcal{U}$, any $x_0 \in \mathcal{S}$, because

$$\begin{aligned} \| x^{u,x_0}(t) \| &\leq \| x_0 \| + \int_{t_0}^{t} \| f(x^{u,x_0}(\tau), u(\tau), \tau) \| \, d\tau \\ &\leq \| x_0 \| + \int_{t_0}^{t} M \left(1 + \| x^{u,x_0}(\tau) \|\right) d\tau \\ &\leq \max_{i \in I_{x_0}} \{ \| X_i^{min} \|, \| X_i^{max} \| \} + M(t_f - t_0) + M \int_{t_0}^{t} \| x^{u,x_0}(\tau) \| \, d\tau, \end{aligned}$$

inequality (B.1) follows from the Bellman-Gronwall inequality. Similarly, because

$$\begin{aligned} p^{u,x_0}(t) &= K_{x_f}(x_0, x^{u,x_0}(t_f)) \\ &\quad - \int_{t_f}^{t} \Big(L_x(x^{u,x_0}(\tau), u(\tau), \tau) + p^{u,x_0}(\tau)^{\mathsf{T}} f_x(x^{u,x_0}(\tau), u(\tau), \tau) \Big) \, d\tau. \end{aligned}$$

Hence, by making use of inequality (B.1), and the continuity of K_{x_f} on the compact set $\{ \| x \| \leq N \} \times \{ \| x \| \leq N \}$, and the continuity of L_x and f_x on the compact set $\{ \| x \| \leq N \} \times \Omega \times [t_0, t_f]$, there exists $c_1, c_2 \in (0, \infty)$ such that,

$$\| p^{u,x_0}(t) \| \leq c_1 + c_2 \int_{t_f}^{t} \| p^{u,x_0}(\tau) \| \, d\tau.$$

Inequality (B.2) then follows from the Bellman-Gronwall inequality. □

Lemma B.2 *There exists $c_1, c_2 \in (0,\infty)$ such that, for any $u^1, u^2 \in \mathcal{U}$, any $x_0^1, x_0^2 \in \mathcal{S}$,*

$$\sup_{t\in[t_0,t_f]} \| x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \| \leq c_1 \| u^2 - u^1 \| + c_2 \| x_0^2 - x_0^1 \| \qquad \text{(B.3)}$$

$$\sup_{t\in[t_0,t_f]} \| p^{u^2,x_0^2}(t) - p^{u^1,x_0^1}(t) \| \leq c_1 \| u^2 - u^1 \| + c_2 \| x_0^2 - x_0^1 \| \qquad \text{(B.4)}$$

Proof: For any $u^1, u^2 \in \mathcal{U}$, any $x_0^1, x_0^2 \in \mathcal{S}$,

$$\begin{aligned}
&x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \\
&= x_0^2 - x_0^1 + \int_{t_0}^t \Big(f(x^{u^2,x_0^2}(s), u^2(s), s) - f(x^{u^1,x_0^1}(s), u^1(s), s) \Big)\, ds \\
&= x_0^2 - x_0^1 + \int_{t_0}^t \Big(f(x^{u^2,x_0^2}(s), u^2(s), s) - f(x^{u^1,x_0^1}(s), u^2(s), s) \Big)\, ds \\
&\quad + \int_{t_0}^t \Big(f(x^{u^1,x_0^1}(s), u^2(s), s) - f(x^{u^1,x_0^1}(s), u^1(s), s) \Big)\, ds \\
&= x_0^2 - x_0^1 + \int_{t_0}^t f_x(\bar{x}(s), u^2(s), s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))\, ds \\
&\quad + \int_{t_0}^t f_u(x^{u^1,x_0^1}(s), \bar{u}(s), s)(u^2(s) - u^1(s))\, ds
\end{aligned}$$

where $\bar{x}(s) = x^{u^1,x_0^1}(s) + \tau_1(s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))$, $\bar{u}(s) = u^1(s) + \tau_2(s)(u^2(s) - u^1(s))$. for some $\tau_1(s), \tau_2(s) \in [0,1]$. Hence, by making use of the continuity of f_x and f_u on the compact set $\{ \| x \| \leq N \} \times \Omega \times [t_0, t_f]$, there exists $c_1, c_2 \in (0,\infty)$ such that,

$$\begin{aligned}
&\| x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \| \\
&\leq \| x_0^2 - x_0^1 \| + \int_{t_0}^t c_1 \| x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s) \|\, ds + \int_{t_0}^t c_2 \| u^2(s) - u^1(s) \|\, ds.
\end{aligned}$$

Inequality (B.3) follows from the Bellman-Gronwall inequality. Similarly, for any $u^1, u^2 \in \mathcal{U}$, any $x_0^1, x_0^2 \in \mathcal{S}$,

$$\begin{aligned}
&p^{u^2,x_0^2}(t) - p^{u^1,x_0^1}(t) \\
&= K_{x_f}(x_0^2, x^{u^2,x_0^2}(t_f))^\top - K_{x_f}(x_0^1, x^{u^1,x_0^1}(t_f))^\top \\
&\quad - \int_{t_f}^t \Big(H_x(x^{u^2,x_0^2}(s), u^2(s), p^{u^2,x_0^2}(s), s)^\top - H_x(x^{u^1,x_0^1}(s), u^1(s), p^{u^1,x_0^1}(s), s)^\top \Big)\, ds \\
&= K_{x_f}(x_0^2, x^{u^2,x_0^2}(t_f))^\top - K_{x_f}(x_0^1, x^{u^2,x_0^2}(t_f))^\top \\
&\quad + K_{x_f}(x_0^1, x^{u^2,x_0^2}(t_f))^\top - K_{x_f}(x_0^1, x^{u^1,x_0^1}(t_f))^\top
\end{aligned}$$

$$
\begin{aligned}
&-\int_{t_f}^{t}\Big(H_x(x^{u^2,x_0^2}(s),u^2(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}-H_x(x^{u^1,x_0^1}(s),u^2(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}\Big)\,ds\\
&-\int_{t_f}^{t}\Big(H_x(x^{u^1,x_0^1}(s),u^2(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}-H_x(x^{u^1,x_0^1}(s),u^1(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}\Big)\,ds\\
&-\int_{t_f}^{t}\Big(H_x(x^{u^1,x_0^1}(s),u^1(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}-H_x(x^{u^1,x_0^1}(s),u^1(s),p^{u^1,x_0^1}(s),s)^{\mathsf{T}}\Big)\,ds\\
=\;&K_{x_fx_0}(\bar{x}_0,x^{u^2,x_0^2}(t_f))^{\mathsf{T}}(x_0^2-x_0^1)+K_{x_fx_f}(x_0^1,\bar{x}_{t_f})^{\mathsf{T}}(x^{u^2,x_0^2}(t_f)-x^{u^1,x_0^1}(t_f))\\
&-\int_{t_f}^{t}H_{xx}(\bar{x}(s),u^2(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}(x^{u^2,x_0^2}(s)-x^{u^1,x_0^1}(s))\,ds\\
&-\int_{t_f}^{t}H_{xu}(x^{u^1,x_0^1}(s),\bar{u}(s),p^{u^2,x_0^2}(s),s)^{\mathsf{T}}(u^2(s)-u^1(s))\,ds\\
&-\int_{t_f}^{t}H_{xp}(x^{u^1,x_0^1}(s),u^1(s),\bar{p}(s),s)^{\mathsf{T}}(p^{u^2,x_0^2}(s)-p^{u^1,x_0^1}(s))\,ds
\end{aligned}
$$

where $\bar{x}_0=x_0^1+\tau_1(x_0^2-x_0^1)$, $\bar{x}_f=x^{u^1,x_0^1}(t_f)+\tau_2(x^{u^2,x_0^2}(t_f)-x^{u^1,x_0^1}(t_f))$, for some $\tau_1,\tau_2\in[0,1]$, and, $\bar{x}(s)=x^{u^1,x_0^1}(s)+\tau_3(s)(x^{u^2,x_0^2}(s)-x^{u^1,x_0^1}(s))$, $\bar{u}(s)=u^1(s)+\tau_4(s)(u^2(s)-u^1(s))$, $\bar{p}(s)=p^{u^1,x_0^1}(s)+\tau_5(s)(p^{u^2,x_0^2}(s)-p^{u^1,x_0^1}(s))$, for some $\tau_3(s),\tau_4(s),\tau_5(s)\in[0,1]$. Hence, by making use of the continuity of H_{xx}, H_{xu} and H_{xp} on the compact set $\{\,\|x\|\le N\,\}\times\Omega\times\{\,\|x\|\le N\,\}\times[t_0,t_f]$, and of $K_{x_fx_0}$ and $K_{x_fx_f}$ on $\{\,\|x\|\le N\,\}\times\{\,\|x\|\le N\,\}$, there exists $c_1,c_2,c_3,c_4,c_5\in(0,\infty)$ such that,

$$
\begin{aligned}
&\|\,p^{u^2,x_0^2}(t)-p^{u^1,x_0^1}(t)\,\|\\
&\quad\le c_1\,\|\,x_0^2-x_0^1\,\|+c_2\,\|\,x^{u^2,x_0^2}(t_f)-x^{u^1,x_0^1}(t_f)\,\|+\int_{t_0}^{t}c_3\,\|\,x^{u^2,x_0^2}(s)-x^{u^1,x_0^1}(s)\,\|\,ds\\
&\qquad+\int_{t_0}^{t}c_4\,\|\,p^{u^2,x_0^2}(s)-p^{u^1,x_0^1}(s)\,\|\,ds+\int_{t_0}^{t}c_5\,\|\,u^2(s)-u^1(s)\,\|\,ds.
\end{aligned}
$$

Inequality (B.4) follows from using (B.3) and the Bellman-Gronwall inequality. □

Lemma B.3 *There exists $c_1,c_2,c_3\in(0,\infty)$ such that, for any $u^1,u^2\in\mathcal{U}$ and any $x_0^1,x_0^2\in\mathcal{S}$,*

$$
\sup_{t\in[t_0,t_f]}\|\,x^{u^2,x_0^2}(t)-x^{u^1,x_0^1}(t)-y^{v,w}(t)\,\|\;\le\;c_1\,\|\,v\,\|^2+c_2\,\|\,v\,\|\cdot\|\,w\,\|+c_3\,\|\,w\,\|^2 \qquad \text{(B.5)}
$$

where $v=u^2-u^1$ and $w=x_0^2-x_0^1$.

Proof: Let $\epsilon(t)=x^{u^2,x_0^2}(t)-x^{u^1,x_0^1}(t)-y^{v,w}(t)$. Because $\epsilon(t_0)=x_0^2-x_0^1-w=0$,

$$
\epsilon(t)=\int_{t_0}^{t}(\dot{x}^{u^2,x_0^2}(s)-\dot{x}^{u^1,x_0^1}(s)-\dot{y}^{v,w}(s))\,ds
$$

$$= \int_{t_0}^{t} (f(x^{u^2,x_0^2}(s), u^2(s), s) - f(x^{u^1,x_0^1}(s), u^2(s), s) - f_x(x^{u^1,x_0^1}(s), u^1(s), s)\, y^{v,w}(s))\, ds$$
$$+ \int_{t_0}^{t} (f(x^{u^1,x_0^1}(s), u^2(s), s) - f(x^{u^1,x_0^1}(s), u^1(s), s) - f_u(x^{u^1,x_0^1}(s), u^1(s), s)\, v(s))\, ds$$

Applying Lemmas A.3 and B.2,

$$\epsilon(t) = \int_{t_0}^{t} (f_x(x^{u^1,x_0^1}(s), u^2(s), s)\,(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s)) - f_x(x^{u^1,x_0^1}(s), u^1(s), s)\, y^{v,w}(s))\, ds$$
$$+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{xx}(\bar{x}(\tau,s), u^2(s), s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))^2\, d\tau\, ds$$
$$+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{uu}(x^{u^1,x_0^1}(s), \bar{u}(\tau,s), s)(u^2(s) - u^1(s))^2\, d\tau\, ds$$
$$= \int_{t_0}^{t} f_x(x^{u^1,x_0^1}(s), u^1(s), s)\, \epsilon(s)\, ds$$
$$+ \int_{t_0}^{t} (f_x(x^{u^1,x_0^1}(s), u^2(s), s) - f_x(x^{u^1,x_0^1}(s), u^1(s), s))(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))\, ds$$
$$+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{xx}(\bar{x}(\tau,s), u^2(s), s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))^2\, d\tau\, ds$$
$$+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{uu}(x^{u^1,x_0^1}(s), \bar{u}(\tau,s), s)(u^2(s) - u^1(s))^2\, d\tau\, ds$$
$$= \int_{t_0}^{t} f_x(x^{u^1,x_0^1}(s), u^1(s), s)\, \epsilon(s)\, ds$$
$$+ \int_{t_0}^{t} f_{xu}(x^{u^1,x_0^1}(s), \bar{u}(s), s)(u^2(s) - u^1(s))(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))\, ds$$
$$+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{xx}(\bar{x}(\tau,s), u^2(s), s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))^2\, d\tau\, ds$$
$$+ \int_{t_0}^{t}\int_0^1 (1-\tau) f_{uu}(x^{u^1,x_0^1}(s), \bar{u}(\tau,s), s)(u^2(s) - u^1(s))^2\, d\tau\, ds$$

where $\bar{x}(\tau,s) = x^{u^1,x_0^1}(s) + \tau(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))$, $\bar{u}(\tau,s) = u^1(s) + \tau(u^2(s) - u^1(s))$, and, $\bar{u}(s) = u^1(s) + \tau(s)(u^2(s) - u^1(s))$ for some $\tau(s) \in [0,1]$. Hence, by making use of the continuity of f_x, f_{xx}, f_{xu} and f_{uu} on the compact set $\{ \| x \| \le N \} \times \Omega \times [t_0, t_f]$, and by Lemma B.2, then there exists c', c'', c''', $c'''' \in (0, \infty)$, such that

$$\| \epsilon(t) \| = c' \int_{t_0}^{t} \| \epsilon(\tau) \|\, ds + c'' \| u^2 - u^1 \|^2 + c''' \| u^2 - u^1 \| \cdot \| x_0^2 - x_0^1 \| + c'''' \| x_0^2 - x_0^1 \|^2.$$

The result then follows from the Bellman-Gronwall inequality. □

Lemma B.4 *Cost functional $J(u, x_0)$, and terminal constraint functionals $g_i(u, x_0)$, $i = 1, \cdots, r$ are all continuous at any $u \in \mathcal{U}$, $x_0 \in \mathcal{S}$.*

Proof: For any $u^2, u^1 \in \mathcal{U}$, and any $x_0^2, x_0^1 \in \mathcal{S}$,

$$
\begin{aligned}
&J(u^2, x_0^2) - J(u^1, x_0^1) \\
&= K(x_0^2, x^{u^2,x_0^2}(t_f)) - K(x_0^1, x^{u^2,x_0^2}(t_f)) + K(x_0^1, x^{u^2,x_0^2}(t_f)) - K(x_0^1, x^{u^1,x_0^1}(t_f)) \\
&\quad + \int_{t_0}^{t_f} \Big(L(x^{u^2,x_0^2}(s), u^2(s), s) - L(x^{u^1,x_0^1}(s), u^2(s), s) \Big)\, ds \\
&\quad + \int_{t_0}^{t_f} \Big(L(x^{u^1,x_0^1}(s), u^2(s), s) - L(x^{u^1,x_0^1}(s), u^1(s), s) \Big)\, ds \\
&= K_{x_0}(\bar{x}_0, x^{u^2,x_0^2}(t_f))(x_0^2 - x_0^1) + \int_{t_0}^{t_f} L_x(\bar{x}(s), u^2(s), s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))\, ds \\
&\quad + K_{x_f}(x_0^1, \bar{x}_f)(x^{u^2,x_0^2}(t_f) - x^{u^2,x_0^2}(t_f)) + \int_{t_0}^{t_f} L_u(x^{u^1,x_0^1}(s), \bar{u}(s), s)(u^2(s) - u^1(s))\, ds
\end{aligned}
$$

where $\bar{x}_0 = x_0^1 + \tau_1(x_0^2 - x_0^1)$, $\bar{x}_f = x^{u^1,x_0^1}(t_f) + \tau_2(x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f))$, for some $\tau_1, \tau_2 \in [0,1]$, and, $\bar{x}(s) = x^{u^1,x_0^1}(s) + \tau_3(s)(x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s))$, $\bar{u}(s) = u^1(s) + \tau_4(s)(u^2(s) - u^1(s))$, for some $\tau_3(s), \tau_4(s) \in [0,1]$. Hence, by making use of the continuity of K_{x_0} and K_{x_f} on the compact set $\{ \| x \| \le N \} \times \{ \| x \| \le N \}$, and of L_x and L_u on the compact set $\{ \| x \| \le N \} \times \Omega \times [t_0, t_f]$, there exist constants $c_1, c_2, c_3, c_4 > 0$, such that,

$$
\begin{aligned}
| J(u^2, x_0^2) - J(u^1, x_0^1) | \;\le\; & c_1 \| x_0^2 - x_0^1 \| + c_2 \| x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f) \| \\
& + \int_{t_0}^{t_f} c_3 \| x^{u^2,x_0^2}(s) - x^{u^1,x_0^1}(s) \|\, ds + \int_{t_0}^{t_f} c_4 \| u^2(s) - u^1(s) \|\, ds.
\end{aligned}
$$

By making use of Lemma B.2, there exists $c_1', c_2' > 0$, such that,

$$
| J(u^2, x_0^2) - J(u^1, x_0^1) | \le c_1' \| x_0^2 - x_0^1 \| + c_2' \| u^2 - u^1 \|,
$$

which implies the continuity of $J(u, x_0)$ in both u and x_0. The continuity of $g_i(u, x_0)$, $i = 1, \cdots, r$ in both $u \in \mathcal{U}$ and $x_0 \in \mathcal{S}$ can be proved similarly. □

Lemma B.5 *There exists $c_1, c_2, c_3, c_4, c_5, c_6, c_7 \in (0, \infty)$, such that, for any u^1, $u^2 \in \mathcal{U}$ and any x_0^1, $x_0^2 \in \mathcal{S}$,*

$$
\Big| J(u^2, x_0^2) - J(u^1, x_0^1) - \Delta J(u^1, x_0^1)(v, w) \Big| \;\le\; c_1 \| v \|^2 + c_2 \| v \| \cdot \| w \| + c_3 \| w \|^2 \quad \text{(B.6)}
$$

and

$$
\begin{aligned}
&\Big| J(u^2, x_0^2) - J(u^1, x_0^1) - \Delta J(u^1, x_0^1)(v, w) - \Delta^2 J(u^1, x_0^1)(v, w) \Big| \\
&\qquad \le\; c_4 \| v \|^3 + c_5 \| v \|^2 \cdot \| w \| + c_6 \| v \| \cdot \| w \|^2 + c_7 \| w \|^3 \quad \text{(B.7)}
\end{aligned}
$$

where $v = u^2 - u^1$, $w = x_0^2 - x_0^1$.

Proof: We first prove (B.7). From cost functional (4.2.8), Hamiltanian (4.3.1) with $p_0 = 1$, and costate equation (4.3.3),

$$\begin{aligned}
&J(u^2, x_0^2) - J(u^1, x_0^1) \\
&= K(x_0^2, x^{u^2,x_0^2}(t_f)) - K(x_0^1, x^{u^1,x_0^1}(t_f)) - \int_{t_0}^{t_f} p^{u^1,x_0^1}(t)^\top (\dot{x}^{u^2,x_0^2}(t) - \dot{x}^{u^1,x_0^1}(t))\, dt \\
&\quad + \int_{t_0}^{t_f} \Big(H(x^{u^2,x_0^2}(t), u^2(t), p^{u^1,x_0^1}(t), t) - H(x^{u^1,x_0^1}(t), u^1(t), p^{u^1,x_0^1}(t), t) \Big)\, dt \\
&= K(x_0^2, x^{u^2,x_0^2}(t_f)) - K(x_0^1, x^{u^1,x_0^1}(t_f)) - p^{u^1,x_0^1}(t)^\top (x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t))\Big|_{t_0}^{t_f} \\
&\quad + \int_{t_0}^{t_f} \dot{p}^{u^1,x_0^1}(t)^\top (x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t))\, dt \\
&\quad + \int_{t_0}^{t_f} \Big(H(x^{u^2,x_0^2}(t), u^2(t), p^{u^1,x_0^1}(t), t) - H(x^{u^1,x_0^1}(t), u^1(t), p^{u^1,x_0^1}(t), t) \Big)\, dt.
\end{aligned}$$

Because $x^{u^2,x_0^2}(t_0) = x_0^2$, $x^{u^1,x_0^1}(t_0) = x_0^1$, together with the terminal condition (4.4.5) of the costate,

$$\begin{aligned}
&J(u^2, x_0^2) - J(u^1, x_0^1) \\
&= K(x_0^2, x^{u^2,x_0^2}(t_f)) - K(x_0^1, x^{u^1,x_0^1}(t_f)) \\
&\quad - K_{x_f}(x_0^1, x^{u^1,x_0^1}(t_f))(x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f)) + p^{u^1,x_0^1}(t_0)^\top (x_0^2 - x_0^1) \\
&\quad + \int_{t_0}^{t_f} \Big(H(x^{u^2,x_0^2}(t), u^2(t), p^{u^1,x_0^1}(t), t) - H(x^{u^1,x_0^1}(t), u^1(t), p^{u^1,x_0^1}(t), t) \Big)\, dt \\
&\quad - \int_{t_0}^{t_f} H_x(x^{u^1,x_0^1}(t), u^1(t), p^{u^1,x_0^1}(t), t)(x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t))\, dt.
\end{aligned}$$

Applying Lemma B.2 and Lemma B.3

$$\begin{aligned}
&J(u^2, x_0^2) - J(u^1, x_0^1) \\
&= \Big(K_{x_0}(x_0^1, x^{u^1,x_0^1}(t_f)) + p^{u^1,x_0^1}(t_0)^\top \Big)(x_0^2 - x_0^1) + \int_{t_0}^{t_f} H_u^{(1)}(t)\,(u^2(t) - u^1(t))\, dt \\
&\quad + \frac{1}{2} \begin{pmatrix} x_0^2 - x_0^1 \\ x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f) \end{pmatrix}^\top \begin{pmatrix} K_{x_0x_0}^{(1)} & K_{x_0x_f}^{(1)} \\ K_{x_fx_0}^{(1)} & K_{x_fx_f}^{(1)} \end{pmatrix} \begin{pmatrix} x_0^2 - x_0^1 \\ x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f) \end{pmatrix} \\
&\quad + \frac{1}{2} \int_{t_0}^{t_f} \begin{pmatrix} x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix}^\top \begin{pmatrix} H_{xx}^{(1)}(t) & H_{xu}^{(1)}(t) \\ H_{ux}^{(1)}(t) & H_{uu}^{(1)}(t) \end{pmatrix} \begin{pmatrix} x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix} dt
\end{aligned}$$

$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xxx}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta x^3(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uxx}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta u(t)\Delta x^2(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xxu}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta x^2(t)\Delta u(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xux}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta x(t)\Delta u(t)\Delta x(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uux}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta u^2(t)\Delta x(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uxu}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta u(t)\Delta x(t)\Delta u(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{xuu}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta x(t)\Delta u^2(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_{t_0}^{t_f}\int_0^1 (1-\tau)^2 H_{uuu}(\bar{x}(\tau,t),\bar{u}(\tau,t),p^{u^1,x_0^1}(t),t)\Delta u^3(t)\,d\tau\,dt$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_0x_0x_0}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x^3(t_0)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_0x_0x_f}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x^2(t_0)\Delta x(t_f)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_0x_fx_0}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x(t_0)\Delta x(t_f)\Delta x(t_0)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_fx_0x_0}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x(t_f)\Delta x^2(t_0)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_fx_fx_0}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x^2(t_f)\Delta x(t_0)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_fx_0x_f}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x(t_f)\Delta x(t_0)\Delta x(t_f)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_0x_fx_f}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x(t_0)\Delta x^2(t_f)\,d\tau$$
$$+\frac{1}{2}\int_0^1 (1-\tau)^2 K_{x_fx_fx_f}(\bar{x}(\tau,t_0),\bar{x}(\tau,t_f))\Delta x^3(t_f)\,d\tau$$

where $\Delta x(t)=x^{u^2,x_0^2}(t)-x^{u^1,x_0^1}(t)$, $\Delta u(t)=u^2(t)-u^1(t)$, $\bar{x}(\tau,t)=x^{u^1,x_0^1}(t)+\tau(x^{u^2,x_0^2}(t)-x^{u^1,x_0^1}(t))$, $\bar{u}(\tau,t)=u^1(t)+\tau(u^2(t)-u^1(t))$, and, $H_u^{(1)}(t)$, $H_{xx}^{(1)}(t)$, $H_{xu}^{(1)}(t)$, $H_{ux}^{(1)}(t)$ and $H_{uu}^{(1)}(t)$ are evaluated at $(x^{u^1,x_0^1}(t),\ p^{u^1,x_0^1}(t),\ u^1(t),\ t)$, and $K_{x_0x_0}^{(1)}$, $K_{x_0x_f}^{(1)}$, $K_{x_fx_0}^{(1)}$, and $K_{x_fx_f}^{(1)}$ are evaluated at $(x_0^1, x^{u^1,x_0^1}(t_f))$. Hence, by making use of the continuity of $K_{x_0x_0x_0}$, $K_{x_0x_0x_f}$, $K_{x_0x_fx_0}$, $K_{x_fx_0x_0}$, $K_{x_fx_fx_0}$, $K_{x_fx_0x_f}$, $K_{x_0x_fx_f}$ and $K_{x_fx_fx_f}$ on the compact set $\{\|x\|\le N\}\times\{\|x\|\le N\}$, and of H_{xxx}, H_{uxx}, H_{xxu}, H_{xux}, H_{uux}, H_{uxu}, H_{xuu} and H_{uuu} on the compact set $\{\|x\|\le N\}\times\Omega\times\{\|p\|\le N\}\times[t_0,t_f]$,

together with the fact that $x^{u,x_0}(t)$ is uniformly bounded on $[t_0,t_f]$ by Lemma B.1, and that $x^{u^2,x_0^2}(t)-x^{u^1,x_0^1}(t)$ is uniformly bounded by $\|u^2-u^1\|$ and $\|x_0^2-x_0^1\|$ on $[t_0,t_f]$ by Lemma B.2, then there exists $c_1',c_2',c_3',c_4' \in (0,\infty)$, such that

$$\left| J(u^2,x_0^2)-J(u^1,x_0^1)-\Delta J(u^1,x_0^1)(v,w)-\tilde{\Delta}^2 J(u^1,x_0^1)(v,w) \right| \\ \leq c_1' \| v \|^3 + c_2' \| v \|^2 \cdot \| w \| + c_3' \| v \| \cdot \| w \|^2 + c_4' \| w \|^3$$

where $v = u^2 - u^1$, $w = x_0^2 - x_0^1$, and

$$\tilde{\Delta}^2 J(u^1)(v,w) \\ = \frac{1}{2} \begin{pmatrix} x_0^2 - x_0^1 \\ x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f) \end{pmatrix}^{\top} \begin{pmatrix} K_{x_0x_0}^{(1)} & K_{x_0x_f}^{(1)} \\ K_{x_fx_0}^{(1)} & K_{x_fx_f}^{(1)} \end{pmatrix} \begin{pmatrix} x_0^2 - x_0^1 \\ x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f) \end{pmatrix} \\ + \frac{1}{2} \int_{t_0}^{t_f} \begin{pmatrix} x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix}^{\top} \begin{pmatrix} H_{xx}^{(1)}(t) & H_{xu}^{(1)}(t) \\ H_{ux}^{(1)}(t) & H_{uu}^{(1)}(t) \end{pmatrix} \begin{pmatrix} x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) \\ u^2(t) - u^1(t) \end{pmatrix} dt. \quad \text{(B.8)}$$

To complete the proof of (B.7), it remains to prove that there exist $c_1'', c_2'', c_3'', c_4'' \in (0,\infty)$, such that

$$\left| \tilde{\Delta}^2 J(u^1,x_0^1)(v,w) - \Delta^2 J(u^1,x_0^1)(v,w) \right| \\ \leq c_1'' \| v \|^3 + c_2'' \| v \|^2 \cdot \| w \| + c_3'' \| v \| \cdot \| w \|^2 + c_4'' \| w \|^3. \quad \text{(B.9)}$$

Let $\epsilon(t) = x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t) - y^{v,w}(t)$, from (B.8) above and the definition of $\Delta^2 J(u^1,x_0^1)(v,w)$ in (4.4.4),

$$\tilde{\Delta}^2 J(u^1,x_0^1)(v,w) - \Delta^2 J(u^1,x_0^1)(v,w) \\ = \epsilon(t_f)^{\top} K_{x_fx_0}^{(1)} w + \epsilon(t_f)^{\top} K_{x_fx_f}^{(1)} (x^{u^2,x_0^2}(t_f) - x^{u^1,x_0^1}(t_f)) - \frac{1}{2} \epsilon(t_f)^{\top} K_{x_fx_f}^{(1)} \epsilon(t_f) \\ + \int_{t_0}^{t_f} \epsilon(t)^{\top} H_{xu}^{(1)}(t) (u^2(t) - u^1(t)) \, dt + \int_{t_0}^{t_f} \epsilon(t)^{\top} H_{xx}^{(1)}(t) (x^{u^2,x_0^2}(t) - x^{u^1,x_0^1}(t)) \, dt \\ - \frac{1}{2} \int_{t_0}^{t_f} \epsilon(t)^{\top} H_{xx}^{(1)}(t) \, \epsilon(t) \, dt.$$

By making use of the continuity of $K_{x_fx_0}^{(1)}$ and $K_{x_fx_f}^{(1)}$ on the compact set $\{ \| x \| \leq N \} \times \{ \| x \| \leq N \}$, and of H_{xx} and H_{xu} on the compact set $\{ \| x \| \leq N \} \times \Omega \times \{ \| p \| \leq N \} \times [t_0,t_f]$, together with Lemma B.1, Lemma B.2, and Lemma B.3, (B.9) then follows. The proof of (B.7) is now complete.

Finally, by making use of the continuity of $K_{x_0x_0}$, $K_{x_0x_f}$, $K_{x_fx_0}$, $K_{x_fx_f}$ on the compact set $\{ \| x \| \leq N \} \times \{ \| x \| \leq N \}$, and of H_{xx}, H_{xu}, H_{ux} and H_{uu} on the compact set $\{ \| x \| \leq N \} \times \Omega \times \{ \| p \| \leq N \} \times [t_0, t_f]$, together with the uniform boundedness of $y^{v,w}(t)$ by Lemma B.3, inequality (B.6) then follows immediately. □

References

1. AHLBERG, J. H., NILSON, E. N. AND WALSH, J. L. *The Theory of Splines and Their Applications*. Academic Press, Inc., New York, 1967.

2. ANDERSON, B. D. O. AND MOORE, J. B. *Optimal Control — Linear Quadratic Methods*. Prentice-Hall, Englewood Cliffs, N.J., 1990.

3. ANDERSON, J. E. In *Grant's Atlas of Anatomy*. The Williams and Wilkins Co., Baltimore, 1978.

4. ATHANS, M. AND FALB, P. L. *Optimal Control*. McGraw-Hill, New York, 1966.

5. BARNES, E. R. An extension of Gilbert's algorithm for computing optimal controls. *Journal of Optimization Theory and Applications* 7(1971), 420–443.

6. BELLMAN, R. E. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

7. BLISS, G. *Calculus of Variations*. Mathematical Association of America, The Open Court Publishing Campany, LaSalle, ILL.

8. BOOR, C. D. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978.

9. BREAKWELL, J. V., SPEYER, J. L. AND BRYSON, A. E. Optimization and control of nonlinear systems using the second variation. *SIAM Journal on Control 12* (1963), 193–223.

10. BROYDEN, C. G. The convergence of a class of double-rank minimization algorithm, 1, General considerations, and 2, The new algorithm. *Journal of the Institute of Mathematics and Its Applications* 61, 3 (1970).

11. BRYSON, A. E. AND HO, Y. C. *Applied Optimal Control*. Hemisphere Publishing Co., 1975.

12. BUCK, R. C. *Advanced Calculus, third edition*. McGraw-Hill, New York, 1978.

13. BULLOCK, T. E. AND FRANKLIN, G. F. A Second-order feedback method for optimal control computations. *IEEE Trans. Automatic Control AC-12* (1967), 666–673.

14. CESARI, L. *Optimization — Theory and Applications*. Springer-Verlag, New York, 1983.

15. CHANG, Y. F. AND LEE, T. T. General orthogonal polynomials approximation of the linear-quadratic-gaussian control design. *Int. J. Control 43* (1986), 1879–1895.

16. CHEN, C. F. AND HSIAO, C. H. Design of piecewise constant gains for optimal control via Walsh functions. *IEEE Trans. Automatic Control AC-20* (1975), 596–603.

17. CHEN, C. F. AND HSIAO, C. H. Walsh series analysis in optimal control. *Int. J. Control 216* (1975), 881–897.

18. CHEN, C. T. *Linear System Theory and Design*. Holt, Rinehart and Winston, New York, 1984.

19. CHEN, W. L. AND SHIH, Y. P. Analysis and optimal control of time-varying linear systems via Walsh functions. *Int. J. Control 27* (1978), 917–932.

20. CHIU, H. Y. The optimal control of a three-segment jump model of a maximal height jump. University of Maryland, M.S. Thesis, 1984.

21. CHO, Y. M. The optimal control of multi-segment inverted pendulum. University of Maryland, Ph.D. Dissertation, 1983.

22. CHOU, J. H. Application of Legendre series to the optimal control of integro-differential equations. *Int. J. Control 45* (1987), 269–277.

23. CHOU, J. H. AND HORNG, I. R. Application of Chebyshev polynomials to the optimal control of time-varying linear systems. *Int. J. Control 41* (1985), 135–144.

24. CHOW, D. AND JACOBSON, D. Studies of human locomotion via optimal programming. *Math. Biosci. 28* (1971), 239–306.

25. CORRINGTON, M. S. Solution of differential and integral equations with Walsh functions. *IEEE Trans. Circuit Theory CT-205* (1980), 470–476.

26. CRAVEN, B. D. *Mathematical Programming and Control Theory*. London: Chapman and Hall, 1978.

27. CULLUM, J. *Penalty functions and nonconvex continuous optimal control problems.* In *Computing Methods in Optimization Problems — 2*, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, Eds. Academic Press, New York, 1969, 55–66.

28. CURRY, H. B. AND SCHOENBERG, I. J. On polya frequency functions. IV. The fundamental spline functions and their limits.. *J. Analyse Math. 17* (1966), 71–107.

29. DAVIDON, W. C. Variable metric methods of minimization. *Research and Development Report ANL-5990.*

30. EDGE, E. R. AND POWERS, W. F. Function-space quasi-Newton algorithms for optimal control problems with bounded control and singular arcs. *Journal of Optimization Theory and Applications 20* (1976), 455–479.

31. ENDOW, T. Optimal control via fourier series of operational matrix of integration. *IEEE Trans. Automatic Control AC-347* (1989), 770–773.

32. FINE, N. J. On the Walsh functions. *Trans. Amer. Math. Soc. 65* (1949), 372–414.

33. FLETCHER, R. A new approach to variable metric algorithms. *Computer Journal 133* (1970).

34. FOX, L. AND PARKER, I. B. *Chebyshev Polynomials in Numerical Analysis.* Oxford University Press, 1968.

35. FREGLY, B. J. AND ZAJAC, F. E. Issues in the modeling and control of biomechanical systems. *Winter Annual Meeting of the ASME 17* (1989), 29–33.

36. FUKUSHIMA, M. AND YAMAMOTO, Y. A second-order algorithm for continuous time nonlinear optimal control problems. *IEEE Trans. Automatic Control AC-31* (1986), 673–676.

37. GELFAND, I. M. AND FOMIN, S. V. *Calculus of Variations*. Prentice-Hall, Englewood Cliffs, N.J., 1963.

38. GERSHWIN, S. B. AND JACOBSON, D. H. A discrete-time differential dynamic programming algorithm with application to optimal orbit transfer. *AIAA Journal 8* (1970), 1616–1626.

39. GIAT, Y. Prediction of muscular synergism and antagonism of human upper extremity: A dynamic optimization approach. University of Maryland, Ph.D. Dissertation, 1990.

40. GOH, C. J. AND TEO, K. L. Control parametrization: a unified approach to optimal control problems with general constraints. *Automatica 24* (1988), 3–18.

41. GOLDFARD, D. A family of variable metric methods derived by variational means. *Mathematics of Computation 24* (1970), 23–26.

42. GOLUBOV, B., EFIMOV, A. AND SKVORTSOV, V. *Walsh Series and Transforms — Theory and Applications*. Kluwer Academic Publishers, Netherlands, 1987.

43. GRUVER, W. A. AND SACHS, E. *Algorithmic Methods in Optimal Control*. Boston: Pitman Advanced Publishing Program, 1980.

44. HALL, C. A. AND MEYER, W. W. Optimal error bounds for cubic spline interpolation. *Journal of Approximation 16* (1976), 105–122.

45. HAN, S. P. Superlinearly convergent variable metric algorithm for general nonlinear programming problems. *Mathematical Programming 11* (1976), 263–282.

46. HAN, S. P. A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications 22* (1977), 297–309.

47. HATZE, H. The complete optimization of human motion. *Math. Biosci. 28*(1976), 99–135.

48. HAVIRA, R. M. AND LEWIS, J. B. Computation of quantized controls using differential dynamic programming. *IEEE Trans. Automatic Control AC-17*(1972), 191–196.

49. HE, J. A feedback control analysis of the neuro-musculo-skeletal control system of a cat hindlimb. University of Maryland, Ph.D. Dissertation, 1988.

50. HILDEBRAND, F. B. *Introduction to Numerical Analysis, second edition.* McGraw-Hill, New York, 1974.

51. HILL, A. V. *First and last experiments in muscle mechanics.* Cambridge Univ. Press, Cambridge, 1970.

52. HORNG, I. R., CHOU, J. H. AND TSAI, R. Y. Taylor series analysis of linear optimal control systems incorporating observers. *Int. J. Control 44*(1986), 1265–1272.

53. HORWITZ, L. B. AND SARACHIK, P. E. Davidon's method in Hilbert space. *SIAM Journal on Applied Mathematics 164*(1968), 676–695.

54. HSU, N. S. AND CHENG, B. Analysis and optimal control of time-varying linear systems via block-pulse functions. *Int. J. Control 33*(1981), 1107–1122.

55. HWANG, C. AND CHEN, M. Y. Laguerre series direct method for variational problems. *Journal of Optimization Theory and Applications 39*(1983), 143–149.

56. HWANG, C. AND CHEN, M. Y. Analysis and optimal control of time-varying linear systems via shifted Legendre polynomials. *Int. J. Control 41*(1985), 1317–1330.

57. HWANG, C. AND SHIH, Y. P. Optimal control of delay systems via block-pulse functions. *Journal of Optimization Theory and Applications 45*(1985), 101–112.

58. IOFFE, A. D. AND TIHOMIROV, V. M. *Theory of Extremal Problems*. North-Holland Publishing Co., Netherlands, 1979.

59. ITO, S. AND SHIMIZU, K. Necessary conditions for constrained optimal control problems via mathematical programming. *Numer. Funct. Anal. and Optimiz. 11* (1990), 267–281.

60. JACOBSON, D. H., LELE, M. M. AND SPEYER, J. L. New necessary conditions of optimality for control problems with state-variable inequality constraints. *J. Math. Anal. Appl. 35* (1971), 255–284.

61. JACOBSON, D. H. AND MAYNE, D. Q. *Differential Dynamic Programming*. American Elsevier Pub. Comp., New York, 1970.

62. JIANG, Z. H. AND SCHAUFELBERGER, W. *Block Pulse Functions and Their Applications in Control Systems*. Springer-Verlag, New York, 1992.

63. KAILATH, T. *Linear Systems*. Prentice-Hall, Englewood Cliffs, N.J., 1980.

64. KAWAJI, S. AND TADA, R. I. Walsh series analysis in optimal control systems incorporating observers. *Int. J. Control 373* (1983), 455–462.

65. KEKKERIS, G. T. AND PARASKEVOPOULOS, P. N. Hermite series approach to optimal control. *Int. J. Control 47* (1988), 557–567.

66. KELLEY, C. T. AND SACHS, E. W. Quasi-Newton methods and unconstrained optimal control problems. *SIAM Journal of Control and Optimization 256* (1987), 1503–1516.

67. KELLY, H. J., KOPP, R. E. AND MOYER, H. G. *A trajectory optimization technique based upon the theory of the second variation*. In *Celestial Mechanics and Astrodynamics*, G. Leitmann, Ed. Academic Press, New York, 1962.

68. KIRK, D. E. *Optimal Control Theory — An Introduction*. Prentice-Hall, Englewood Cliffs, N.J., 1970.

69. KRAFT, D. *Comparing mathematical programming algorithms based on Lagrangian functions for solving optimal control problems.* In *Control Applications of Nonlinear Programming*, H. E. Rauch, Ed. Pergamon Press,, New York, 1980, 71–84.

70. KRAFT, D. Finite-difference gradients versus error-quadrature gradients in the solution of parameterized optimal control problems. *Optimal Control Applications and Methods 2* (1981), 191–199.

71. KRAFT, D. *On converting optimal control problems into nonlinear programming problems.* In *Computational Mathematical Programming — NATO ASI Series, Vol. F15*, K. Schittkowski, Ed. Springer-Verlag, New York, 1985, 261–280.

72. KWAKERNAAK, H. AND SIVAN, R. *Linear Optimal Control Systems.* Wiley, New York, 1972.

73. KWONG, C. P. AND CHEN, C. F. The convergence properties of block-pulse series. *Int. J. Systems. Sci. 12* (1981), 745–751.

74. LASDON, L. S. Conjugate direction methods for optimal control. *IEEE Trans. Automatic Control AC-15* (1970), 267–268.

75. LASDON, L. S., MITTER, S. K. AND WAREN, A. V. The conjugate gradient method for optimal control problems. *IEEE Trans. Automatic Control AC-12* (1967), 132–138.

76. LEE, E. B. AND MARCUS, L. *Foundations of Optimal Control Theory.* John Wiley and Sons, Inc., 1967.

77. LEITMANN, G. *The Calculus of Variations and Optimal Control — An Introduction.* Plenum Press, New York, 1981.

78. LEVINE, W. S., CHRISTODOULOU, M. AND ZAJAC, F. E. On propelling a rod to a maximum vertical or horizontal distance. *Automatica 19* (1983).

79. LEVINE, W. S., MA, B. AND ZAJAC, F. E. The dynamics and optimal controls for some simplified models of pedaling a stationary bicycle. Presented at *Proc. of 1989 Conference on Information Science and Systems* (1989).

80. LEVINE, W. S., ZAJAC, F. E., BELZER, M. R. AND ZOMLEFER,, M. R. Ankle controls that produce a maximal vertical jump when other joints are locked. *IEEE Trans. Automatic Control AC-28* (1983), 1008–1016.

81. LITT, F. X. AND DELCOMMUNE, J. Implementation of spline approximations algorithms in numerical optimal control. Presented at *IFAC Applications of Nonlinear Programming to Optimization and Control*, Palo, Alto, CA, USA (1983).

82. LIU, C. C. AND SHIH, Y. P. Analysis and optimal control of time-varying systems via Chebyshev polynomials. *Int. J. Control 38* (1983), 1003–1012.

83. LIU, C. C. AND SHIH, Y. P. System analysis, parameter estimation, and optimal regulator design of linear system via Jacobi series. *Int. J. Control 42* (1985), 211–224.

84. LOEB, G. E. AND LEVINE, W. S. *Linking musculoskeletal mechanics to sensorimotor neurophysiology.* In *Multiple Muscle System - Biomachanics and Movement Organization*, J. M. Winters and S. L. Woo, Eds. Springer-Verlag, New York, 1990, 165–181.

85. LUENBERGER, D. G. *Optimization by Vector Space Methods.* Wiley, New York, 1969.

86. LUENBERGER, D. G. *Linear and Nonlinear Programming.* Addison-Wesley Publishing Company, 1984.

87. LUKE, Y. *The Special Functions And Their Approximations.* Academic Press, Inc., New York, 1969.

88. MA, B. The dynamics and time optimal control for the skeletal system of humans pedaling a stationary bicycle. University of Maryland, M.S. Thesis, 1989.

89. MA, B., GIAT, Y. AND LEVINE, W. S. An optimally controlled movement of the human arm. Presented at *Proc. of 13th Southern Biomedical Engineering Conference* (1994).

90. MA, B., GIAT, Y. AND LEVINE, W. S. The optimal control of a movement of the human upper extremity. Presented at *Proc. of 1994 IFAC Symposium on Modeling and Control in Biomedical Systems* (March, 1994).

91. MA, B. AND LEVINE, W. S. An algorithm for solving control constrained optimal control problems. Presented at *Proc. of 32nd IEEE Conference on Decision and Control (CDC)*, San Antonio (1993).

92. MA, B. AND LEVINE, W. S. An algorithm for solving optimal control problems with control and terminal-state constraints",. Presented at *(Submitted to) the 33nd IEEE Conference on Decision and Control (CDC)*, Orlando (1994).

93. MAKOWSKI, K. AND NEUSTADT, L. W. Optimal control problems with mixed control-phase variable equality and inequality constraints. *SIAM J. Control 12* (1974), 184–228.

94. MAYNE, D. Q. Differential dynamic programming — A unified approach to the optimization of dynamic systems. *Control and Dynamic Systems 10* (1973).

95. MAYNE, D. Q. AND POLAK, E. First order strong variation algorithms for optimal control. *Journal of Optimization Theory and Applications 16* (1975), 277–301.

96. MAYNE, D. Q. AND POLAK, E. An exact penalty function algorithm for optimal control problems with control and terminal constraints, Part 1. *Journal of Optimization Theory and Applications 32* (1980), 211–247.

97. MAYNE, D. Q. AND POLAK, E. An exact penalty function algorithm for optimal control problems with control and terminal constraints, Part 2. *Journal of Optimization Theory and Applications 32* (1980), 345–365.

98. MAYNE, D. Q. AND POLAK, E. An exact penalty function algorithm for control problems with state and control constraints. *IEEE Trans. Automatic Control AC-32* (1987).

99. MAYORGA, R. V. AND QUINTANA, V. H. A family of variable metric methods in function space, Without exact line searches. *Journal of Optimization Theory and Applications 31* (1980), 303–329.

100. MITTER, S. K. Successive approximation methods for the solution of optimal control problems. *Automatica 3* (1966), 135–149.

101. MOHLER, R. R. *Bilinear Control Processes*. Academic Press, Inc., 1973.

102. MOUROUTSOS, S. G. AND SPARIS, P. D. Taylor series approach to system identification, analysis, and optimal control. *Journal of Franklin Institute 319* (1985), 359–371.

103. MURRAY, D. M. AND YAKOWITZ, S. J. Differential dynamic programming and Newton's method for discrete optimal control problems. *Journal of Optimization Theory and Applications 43* (1984), 395–414.

104. NEUMAN, C. P. AND SEN, A. A suboptimal control algorithm for constrained problems using cubic splines. *Automatic 19* (1973), 601–613.

105. NEUSTADT, L. W. *Optimization*. Princeton University Press, Princeton, 1976.

106. NORRIS, D. O. Nonlinear programming applied to state-constrained optimization problems. *J. Math. Anal. Appl. 43* (1973), 261–272.

107. NURNBERGER, G. *Approximation by Spline Functions*. Springer-Verlag, New York, 1989.

108. OHNO, K. A new approach to differential dynamic programming for discrete time systems. *IEEE Trans. Automatic Control AC-231* (1978).

109. PALEY, R. E. A. C. A remarkable series of orthogonal functions. *Proc. of London Math. Soc., 2nd series, 34* (1932), 241–279.

110. PANDY, M. G., ZAJAC, F. E., SIM, E. AND LEVINE, W. S. An optimal control model for maximum-height human jumping. *Journal of Biomechanics 23* (1990), 1185–1198.

111. PARASKEVOPOULOS, P. N. Chebyshev series approach to system identification, analysis and optimal control. *Journal of Franklin Institute 3162* (1983), 135–157.

112. PARASKEVOPOULOS, P. N. The operational matrices of integration and differentiation for the Fourier sine-cosine and exponential series. *IEEE Trans. Automatic Control AC-32* (1987), 648–651.

113. PARASKEVOPOULOS, P. N., SPARIS, P. D. AND MOUROUTSOS, S. G. The fourier series operational matrix of integration. *International Journal of Systems Science 16* (1985), 171–176.

114. PARASKEVOPOULOS, P. N., TSIRIKOS, A. S. AND ARVANITIS, K. G. New Taylor series approach to state-space analysis and optimal control of linear systems. *Journal of Optimization Theory and Applications 71* (1991), 315–340.

115. PERNG, M. H. An effective approach to the optimal control problem for time-varying linear systems via Taylor series. *Int. J. Control 44* (1986), 1225–1231.

116. POLAK, E. *Computational Methods in Optimalization: A Unified Approach.* Academic Press, New York, 1971.

117. POLAK, E. An historical survey of computational methods in optimal control. *SIAM Review 15* (1973), 553–584.

118. POLAK, E. AND MAYNE, D. Q. First order strong variation algorithms for optimal control with terminal inequality constraints. *Journal of Optimization Theory and Applications 16* (1975), 303–325.

119. POLAK, E. AND MAYNE, D. Q. A feasible directions algorithm for optimal control problems with control and terminal inequality constraints. *IEEE Trans. Automatic Control AC-22* (1977), 741–751.

120. PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMKRELIDZE, R. V. AND MISCHENKO, E. F. *The Mathematical Theory of Optimal Processes*. Intersciences Publishers, Inc., New York, 1962.

121. POWELL, M. J. D. Algorithms for nonlinear constraints that use lagrangian functions. *Mathematical Programming 14* (1978), 224–248.

122. POWELL, M. J. D. *Approximation Theory and Methods*. Cambridge University Press, 1981.

123. PYTLAK, R. AND VINTER, R. B. A feasible directions type algorithm for optimal control problems with hard state and control constraints. Presented at *Proc. of 32nd IEEE Conference on Decision and Control (CDC)*, San Antonio (1993).

124. RAASCH, C. C. The use of musculoskeletal models and optimization to study coordination strategies and synergies in cycling. Stanford University, Ph.D. Dissertation, 1994.

125. RAASCH, C. C., MA, B., ZAJAC, F. E. AND LEVINE, W. S. Muscle coordination of maximum-speed pedaling based on modeling and kinesiological data. Presented at *Society for Neuroscience Abstracts* (1993).

126. RAASCH, C. C., MA, B., ZAJAC, F. E. AND LEVINE, W. S. Importance of biarticular muscle control to smooth pedaling. Presented at *Second World Congress of Biomechanics*, Amsterdam (1994).

127. RAASCH, C. C., MA, B., ZAJAC, F. E. AND LEVINE, W. S. Use of an optimal control model to study normal and constrained control of maximum-speed pedaling. Presented at *Proc. of 1994 IFAC Symposium on Modeling and Control in Biomedical Systems* (March, 1994).

128. RAASCH, C. C., ZAJAC, F. E., MA, B., LEVINE, W. S., DAIRAGHI, C. A. AND STEVENSON, P. J. The use of optimal control and kinesiological data to study muscle coordination of pedaling. *Proc. of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society 15* (1993), 1147–1148.

129. RAO, V. P. AND RAO, K. R. Optimal feedback control via block-pulse functions . *IEEE Trans. Automatic Control AC-26* (1979), 372–374.

130. RAZZAGHI, M. Solution of linear two-point boundary value problems via fourier series and application to optimal control of linear systems. *Journal of Franklin Institute 3264* (1989), 523–533.

131. RAZZAGHI, M. Optimal control of linear time-varying systems via fourier series. *Journal of Optimization Theory and Applications 65* (1990), 375–384.

132. RAZZAGHI, M. AND ARABSHAHI, A. Analysis of linear time-varying systems and bilinear systems via fourier series. *Int. J. Control 503* (1989), 889–898.

133. RAZZAGHI, M. AND RAZZAGHI, M. Taylor series direct method for variational problems. *Journal of Franklin Institute 325* (1988), 125–131.

134. RAZZAGHI, M. AND RAZZAGHI, M. Solution of linear two-point boundary value problems via Taylor series. *Journal of Franklin Institute 326* (1989), 511–521.

135. RICE, J. R. *The Approximation of Functions*. Addison-Wesley Publishing Company, 1984.

136. RIVLIN, T. J. *Chebyshev Polynomials — From Approximation Theory to Algebra and Number Theory, second edition*. John Wiley and Sons, Inc., 1990.

137. ROSEN, O. AND LUUS, R. Evaluation of gradients for piecewise constant optimal control. *Computers Chem. Engng. 154* (1991), 273–281.

138. ROYDEN, H. L. *Real Analysis, third edition*. Macmillan Publishing Company, New York, 1988.

139. RUSSELL, D. L. Penalty functions and bounded phase coordinate control. *SIAM Journal of Control 2* (1965), 409–422.

140. RUSSELL, D. L. The Kuhn-Tucker conditions in Banach space with an application to control theory. *J. Math. Anal. Appl. 15* (1966), 200–212.

141. SAKAWA, Y. AND SHINDO, Y. Optimal Control of Container Cranes. *Automatica 18* (1982), 257–266.

142. SANSONE, G. *Orthogonal Functions.* Intersciences Publishers, Inc., New York, 1959.

143. SARGENT, R. W. H. AND SULLIVAN, G. R. *The development of an efficient optimal control package.* In *Optimization Techniques — Part 2*, J. Stoer, Ed. Springer-Verlag, New York, 1978, 158–168.

144. SCHOENBERG, I. J. Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math. 4* (1946), Part A, 45–99; Part B, 112-141.

145. SHANNO, D. F. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation 24* (1970), 647–656.

146. SHIH, D. H. AND KUNG, F. C. Optimal control of deterministic systems via shifted Legendre polynomials. *IEEE Trans. Automatic Control AC-31* (1986), 451–454.

147. SHIMIZU, K. AND ITO, S. Constrained optimization in Banach space and generalized dual quasi-Newton algorithms for state-constrained optimal control problems. Presented at *Proc. of 1991 ACC* (1991).

148. SIM, E. The application of optimal control theory for analysis of human jumping and pedaling. University of Maryland, Ph.D. Dissertation, 1988.

149. SIM, E., MA, B., LEVINE, W. S. AND ZAJAC, F. E. Some results on the neuromuscular controls involved in pedaling a bicycle at maximum speed. Presented at *Proc. of 1989 ACC* (1989).

150. SIRISENA, H. R. Computation of optimal controls using a piecewise polynomial parameterization. *IEEE Trans. Automatic Control AC-18* (1973), 409–411.

151. SIRISENA, H. R. AND TAN, K. S. Computation of constrained optimal controls using parameterization techniques. *IEEE Trans. Automatic Control AC-19* (1974), 431–433.

152. SNYDER, M. A. *Chebyshev Methods in Numerical Approximation.* Prentice-Hall, Englewood Cliffs, N.J., 1966.

153. SPARIS, P. D. AND MOUROUTSOS, S. G. Analysis and optimal control of time-varying linear systems via Taylor series. *Int. J. Control 41* (1985), 831–842.

154. SZEGO, G. *Orthogonal Polynomials, fourth edition.* American Mathematical Society, Providence, Rhode Island, 1975.

155. TEO, K. L., GOH, C. J. AND WONG, K. H. *A Unified Computational Approach to Optimal Control Problems.* Longman Scientific and Techanical, 1991.

156. TEO, K. L. AND JENNINGS, L. S. Nonlinear optimal control problems with continuous state inequality constraints. *Journal of Optimization Theory and Applications 63* (1989), 1–22.

157. TEO, K. L. AND WOMERSLEY, R. S. A control parameterization algorithm for optimal control problems involves linear systems and linear terminal inequality constraints. *Numer. Funct. Anal. And Optimiz. 63* (1983), 291–313.

158. TODD, J. *Survey of Numerical Analysis.* McGraw-Hill, New York, 1962.

159. TURNER, P. R. AND HUNTLEY, E. Variable metric methods in Hilbert space with applications to control problems. *Journal of Optimization Theory and Applications 19* (1976), 381–400.

160. VIRK, G. S. A strong variation algorithm for delay systems. *Journal of Optimization Theory and Applications 45* (1985), 295–312.

161. VIRK, G. S. Digital implementation of strong variational algorithms. *Optimal Control Applications and Methods 6* (1985), 211–233.

162. VLASSENBROECK, J. A chebyshev polynomial method for optimal control with state constraints. *Automatica 244* (1988), 499–506.

163. VLASSENBROECK, J. AND DOOREN, R. V. A chebyshev technique for solving nonlinear optimal control problems. *IEEE Trans. Automatic Control AC-334* (1988), 333–340.

164. WALSH, J. L. A closed set of normal orthogonal functions. *Amer. J. Math. 45* (1923), 5–24.

165. WANG, M. L. AND CHANG, R. Y. Optimal control of lumped-parameter systems via shifted Legendre polynomials approximation. *Journal of Optimization Theory and Applications 45* (1985), 313–324.

166. WARGA, J. *Optimal Control of Differential and Functional Equations.* Academic Press, Inc., 1972.

167. WINTERS, J. M. *Hill-based muscle model: A system engineering perspective.* In *Multiple Muscle System - Biomachanics and Movement Organization* , J. M. Winters and S. L. Woo, Eds. Springer-Verlag , New York , 1990, 121–148 .

168. WONG, K. H. Nonlinearly constrained optimal control problems. *J. Austral. Math. Soc. Ser. B 33* (1992), 507–530.

169. ZAJAC, F. *Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control.* In *CRC Critical Reviews in Biomedical Engineering*, J. R. Bourne, Ed. CRC Press, Inc., Boca Raton, FL, 1988.

170. ZAJAC, F. E. AND WINTERS, J. M. *Modeling muscluloskeletal movement systems: Joint and body-segment dynamics, Musculotendinous actuation, and Neuromuscular control.* In *Multiple Muscle System - Biomachanics and Movement Organization*, J. M. Winters and S. L. Woo, Eds. Springer-Verlag, New York, 1990, 121–148.

171. ZEIDLER, E. *Nonlinear Functional Analysis and Its Applications, Part III — Variational Methods and Optimization.* Springer-Verlag, 1985.