The Dissertation Committee for Matthew Wade Harris certifies that this is the approved version of the following dissertation:

Lossless Convexification of Optimal Control Problems

Committee:

Behçet Açıkmeşe, Supervisor

Maruthi R. Akella

Ari Arapostathis

Efstathios Bakolas

David G. Hull

LOSSLESS CONVEXIFICATION OF OPTIMAL CONTROL PROBLEMS

by

Matthew Wade Harris, B.S., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2014

In memory of

Margie, JB, Sue, and JT

ACKNOWLEDGMENTS

It is not often that a young man has the opportunity to associate himself with a recognized leader in his chosen field of endeavor. For this privilege, I am deeply indebted to Professor David G. Hull. I thank Dr. Hull for sharing his enthusiasm for optimal control, demand for precision and clarity, and meaningful life lessons.

I thank Dr. Behçet Açıkmeşe for introducing me to convex optimization and its practical applications. His talent and mathematical agility constantly challenged me to expand my skill set.

I thank my brother Brad for leading the way. He paved my path through youth sports, high school basketball, undergraduate and graduate school, and marriage. Without his friendship and guidance, I would not have known the way to go. I will likely continue to follow his lead.

I thank my mother and father, Nancy and Tom, for their perpetual love and support. They provided endless opportunities for me as a child and proved to be top notch role models for me as an adult. Watching them has taught me who I want to be.

I thank my wife Whitney for her love and friendship. For her, no achievement is too small to celebrate and no failure is too large to cause worry. She is a daily source of inspiration.

LOSSLESS CONVEXIFICATION OF OPTIMAL CONTROL PROBLEMS

Publication No. _____

Matthew Wade Harris, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Behçet Açıkmeşe

This dissertation begins with an introduction to finite-dimensional optimization and optimal control theory. It then proves lossless convexification for three problems: 1) a minimum time rendezvous using differential drag, 2) a maximum divert and landing, and 3) a general optimal control problem with linear state constraints and mixed convex and non-convex control constraints. Each is a unique contribution to the theory of lossless convexification. The first proves lossless convexification in the presence of singular controls and specifies a procedure for converting singular controls to the bang-bang type. The second is the first example of lossless convexification with state constraints. The third is the most general result to date. It says that lossless convexification holds when the state space is a strongly controllable subspace. This extends the controllability concepts used previously, and it recovers earlier results as a special case. Lastly, a few of the remaining research challenges are discussed.

TABLE OF CONTENTS

LIST OF FIGURES ix

CHAPTER I: INTRODUCTION 1

CHAPTER II: FINITE-DIMENSIONAL OPTIMIZATION 9

A. Problem Description 10

B. Duality Theory 12

C. Linear Programming 17

D. Convex Programming 25

E. Normality in Convex Programming 30

CHAPTER III: OPTIMAL CONTROL THEORY 34

A. Problem Description 36

B. Generating the Terminal Cone 42

1. Temporal Perturbations 42

2. Spatial Perturbations 44

C. Properties of the Terminal Cone 52

D. Properties of the Adjoint System 58

E. Properties of the Hamiltonian 61

F. The Transversality Condition 64

G. Optimal Control with State Constraints 68

CHAPTER IV: RENDEZVOUS USING DIFFERENTIAL DRAG . . . 72
A. Problem Description . . . 75
B. Lossless Convexification . . . 82
C. Solution Method . . . 89
D. Results . . . 91
E. Summary and Conclusions . . . 96

CHAPTER V: MAXIMUM DIVERT AND LANDING . . . 97
A. Problem Description . . . 99
B. Lossless Convexification . . . 104
C. Transformation of Variables . . . 112
D. Results . . . 115
E. Summary and Conclusions . . . 119

CHAPTER VI: A GENERAL RESULT FOR LINEAR SYSTEMS . . 120
A. Problem Description . . . 122
B. Linear System Theory . . . 127
C. Lossless Convexification . . . 131
D. Example 1: Minimum Fuel Planetary Landing . . . 137
E. Example 2: Minimum Time Rendezvous . . . 140
F. Summary and Conclusions . . . 143

CHAPTER VII: FINAL REMARKS . . . 144

REFERENCES . . . 146

LIST OF FIGURES

Figure 1: Functions with unattainable costs. 11

Figure 2: Geometry of a bounded linear programming problem. . . . 18

Figure 3: Geometry of an unbounded linear programming problem. . 18

Figure 4: Geometry of sets $\mathcal{A}$ and $\mathcal{B}$. 26

Figure 5: Optimal solutions in the original and lifted space. 40

Figure 6: Suboptimal, optimal, and impossible solutions. 40

Figure 7: Temporal perturbations. 42

Figure 8: Linear approximation of temporal variations. 44

Figure 9: Simplest spatial perturbation. 45

Figure 10: Effect of a spatial perturbation. 47

Figure 11: Linear approximation of a spatial perturbation. 48

Figure 12: Linear approximation of multiple spatial perturbations. . . 48

Figure 13: Two spatial perturbations. 49

Figure 14: Convex hull of spatial perturbations. 51

Figure 15: Terminal cone. 52

Figure 16: Terminal wedge and vector. 53

Figure 17: Actual terminal points in a warped ball. 55

Figure 18: Separating hyperplane and normal vector. 57

Figure 19: Backward evolution of hyperplanes. 59

Figure 20: Illustration of sets. 66

Figure 21: x_1-x_2 phase plane. 78

Figure 22: ωx_3-ωx_4 phase plane. 79
Figure 23: Double integrator switch curve. 81
Figure 24: Harmonic oscillator switch curve. 82
Figure 25: x_1-x_2 phase plane with $M = 1$. 92
Figure 26: ωx_3-ωx_4 phase plane with $M = 1$. 93
Figure 27: Target spacecraft control with $M = 1$. 93
Figure 28: Chaser spacecraft control with $M = 1$. 93
Figure 29: Target spacecraft control $u_0(t)$ with $M = 4$. 95
Figure 30: Chaser spacecraft control $u_1(t)$ and $u_2(t)$ with $M = 4$. . . . 95
Figure 31: Chaser spacecraft control $u_3(t)$ and $u_4(t)$ with $M = 4$. . . . 95
Figure 32: Maximum divert landing scenario. 101
Figure 33: Two-dimensional non-convex thrust constraint. 102
Figure 34: Relaxed thrust constraints. 104
Figure 35: Comparison of positions using different weights. 116
Figure 36: Comparison of velocities using different weights. 117
Figure 37: Thrust magnitude for maximum divert. 118
Figure 38: Mass profile for maximum divert. 118
Figure 39: Two-dimensional non-convex control constraint. 123
Figure 40: Two-dimensional constraint with linear inequalities. 123
Figure 41: Three-dimensional constraint. 124
Figure 42: State trajectory for constrained landing. 139
Figure 43: Control trajectory for constrained landing. 140
Figure 44: State trajectory for constant altitude rendezvous. 142
Figure 45: Control trajectory for constant altitude rendezvous. 143

CHAPTER I:
INTRODUCTION

The topic of this dissertation is lossless convexification, which is the study of convex optimization problems and their *equivalence* with non-convex problems. Given a non-convex optimization problem of interest, the simplest case of lossless convexification consists of only two steps: 1) proposing a convex problem and 2) proving that the convex problem has the same solution as the non-convex problem. This two step process can be complicated in any number of ways. We will explore these complications in Chapters IV through VI.

Simply put, the motivation for lossless convexification is that non-convex problems are more difficult to solve than convex problems. Typical numerical methods for non-convex problems require a good initial guess, do not guarantee convergence, and do not certify global optimality [1, p. 9]. Numerical methods for convex problems correct these deficiencies [2]. Additionally, current research with customized methods indicates orders of magnitude improvement in computation time [3–5].

Thus, lossless convexification has very practical implications in engineering. The most notable successes are the 2012 and 2013 flight tests with Masten Space Systems, Inc. [6, 7]. By means of a lossless convexification, optimal trajectories were computed onboard and successfully flown by the Xombie rocket. A more encompassing theory, including the results in this dissertation, facilitates broader opportunities and greater successes in engineering practice.

Lossless convexification was introduced by Açıkmeşe and Ploen in 2007 [8]. They proved lossless convexification for a fuel optimal planetary landing problem. The problem contained a number of state constraints. However, the proof was completed with simplifying assumptions stating that the state constraints could not be active over intervals, i.e., they could only be active at a finite number of points. This is a strong assumption since it cannot be verified before solving the problem.

This work was extended by Blackmore et al. in 2010 [9]. They developed a prioritized scheme in which landing occurred at the specified final point if possible and at the nearest possible location otherwise. Lossless convexification was used in both cases under the same state constraint assumptions as before.

A more general result was obtained in 2011 by Açıkmeşe and Blackmore [10]. Interest shifted away from a specific landing problem, and lossless convexification was proved for an optimal control problem with convex cost, linear dynamics, and a non-convex control constraint. It was here where it was first seen why convexification works and how it is tied to system properties. It was shown, under a few other assumptions, that lossless convexification holds if the linear system is controllable. This is a very powerful result since most systems are engineered to be controllable.

This result was extended to nonlinear systems in 2012 by Blackmore et al. [11]. Although a general condition was stated regarding lossless convexification, it was very difficult to verify since it depends on gradient matrices maintaining full rank. These gradients in turn depend on the optimal state and control trajectories. Even so, some special cases were treated rigorously.

Attention returned to planetary landing in the work by Açıkmeşe et al. in 2013 [12]. They focused on an additional thrust pointing constraint. Proof of lossless convexification could not be completed in the typical way. This was because one could only prove that the optimal control belonged on the boundary of the control set not the extremal points of the control set. As a workaround, they introduced a small perturbation to the problem and then completed the proof. In this sense, the result is non-rigorous.

This brief literature survey encapsulates the state of lossless convexification. Many open questions remain since lossless convexification has only been proven for a relatively small class of problems. A few of the more important questions are the following.

1. How does one proceed when optimal solutions are non-unique and convexification can only be proven for some solutions?

2. How does one address the planetary landing problem with state constraints?

3. How does one generalize the previous result [10] for problems with state constraints; and under what conditions can pointing constraints be handled rigorously?

Chapters IV, V, and VI answer these three questions, respectively. These three chapters have been published in journal form [13–15], where additional details and results are provided. We now give a brief introduction to the upcoming chapters. More details and references are given at the start of each chapter.

Chapter II: Finite-Dimensional Optimization

This chapter introduces the mathematical theory of finite-dimensional optimization. It starts with a general, nonlinear optimization problem and then specializes the results for linear and convex programming problems. Attention is paid to the concepts of attainability, strong duality, and normality. The chapter concludes with a result connecting the three for convex problems.

Chapter III: Optimal Control Theory

This chapter introduces the mathematical theory of optimal control. It starts with a basic optimal control problem and statement of necessary conditions. The conditions are proved again paying attention to the concept of normality. The proof is not too different than the original of Pontryagin, although the measure-theoretic concepts are avoided by considering only piecewise continuous controls. The chapter concludes by stating an optimal control problem with explicit time dependence and state constraints and the associated necessary conditions.

Chapter IV: Rendezvous Using Differential Drag

This chapter presents lossless convexification for the rendezvous problem of multiple spacecraft using only relative aerodynamic drag. The work is unique because it is only proven that the convexification can work – not that it must. The reason is that singular controls exist when there are more than two spacecraft. A result due to LaSalle on bang-bang controls is invoked, and a constructive procedure for converting singular controls to the bang-bang type is specified. This guarantees lossless convexification.

Chapter V: Maximum Divert and Landing

This chapter presents lossless convexification for the maximum divert and landing of a single engine rocket. The work is unique because it is the first to prove lossless convexification with state constraints. Numerical results show cases where two of the three state constraints are active simultaneously. The proof is complicated by the state constraints since they make the adjoint differential equations inhomogeneous and state boundary arcs are similar to singular arcs, which are typically undesirable in lossless convexification.

Chapter VI: A General Result for Linear Systems

This chapter presents lossless convexification for a class of optimal control problems governed by linear differential equations, linear state constraints, and mixed convex and non-convex control constraints. The work is unique because it is the most general result to date. It says that convexification holds whenever the state space is a strongly controllable subspace. This extends the controllability concept used by Açıkmeşe and Blackmore [10], and it recovers their result as a special case. The work naturally handles pointing constraints and answers the question of when lossless convexification can be done without having to perturb the problem.

Chapter VII: Final Remarks

This chapter briefly explores open questions in lossless convexification and anticipates future challenges. These challenges include theory and practice, and in particular, the challenge of reshaping the engineering community's concept of real-time path planning and control.

The notation used throughout the remaining chapters is mostly standard. In the context of optimal control, functions are denoted by the parenthetical dot notation or just the symbol. For example, $x(\cdot)$, or just x, is an element of a function space such as the space of piecewise continuous functions. The notation $x(t)$ means the function $x(\cdot)$ evaluated at t, and it is a finite-dimensional vector.

If a function has many arguments, the bracket notation is used. For example, a function with three arguments $f(\cdot,\cdot,\cdot)$ is sometimes written as $f[\cdot]$. If all three arguments depend on time, then the function evaluated at time t can be written as $f[t]$. This is only done when it will not lead to confusion.

If a scalar valued function $f(\cdot)$ is differentiable at the point $x \in \mathbb{R}^n$, then its partial derivative is given by the column vector

$$\partial_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \tag{1}$$

If a vector valued function $g(\cdot)$ that maps to $\mathbb{R}^m$ is differentiable at the point $x \in \mathbb{R}^n$, then its partial derivative is given by the matrix

$$\partial_x g(x) = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1(x)}{\partial x_n} & \cdots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} \tag{2}$$

The time derivative of a function $x(\cdot)$ is denoted with an over dot as $\dot{x}(\cdot)$. The time derivate evaluated at time t is $\dot{x}(t)$.

If the function $x(\cdot)$ is differentiable at t, then it is true that

$$x(t+h) = x(t) + \dot{x}(t)h + o(h) \tag{3}$$

for any h so long as $t+h$ is the domain of $x(\cdot)$. The term $o(h)$ is a remainder term, and the little "o" notation indicates the property

$$\lim_{h \to 0} \frac{o(h)}{h} = 0 \tag{4}$$

Similar statements hold if the function is only differentiable from one side.

Each chapter contains optimization/optimal control problems that are of interest. These problems are denoted P0, P1, and so on. Each problem statement contains the performance index to be minimized or maximized along with all of the constraints. Unless stated otherwise, we use the words minimum and optimal to mean global minimum. The feasible sets for each of these problems are denoted $\mathcal{F}_0$, $\mathcal{F}_1$, and so on. The optimal solution sets are $\mathcal{F}_0^*$, $\mathcal{F}_1^*$, and so on. We repeatedly use the fact that $\mathcal{F}^* \subset \mathcal{F}$ since all optimal solutions must be feasible.

Optimal control problems are infinite-dimensional optimization problems. They must be discretized, or converted to finite-dimensional optimization problems, in order to be solved numerically. An example of discretization for the planetary landing problem is given by Açıkmeşe and Ploen [8]. The excellent paper by Hull covers the topic at a general level [16]. For the example problems herein, the simplest discretization is used. The time interval is divided into equally spaced subintervals, the control is assumed piecewise constant,

and the constraints are enforced at each node. The finite-dimensional problem is then solved using one of the following software packages: SDPT3 [17], SeDuMi [18], Gurobi [19], or CVX [20]. The topic of discretization is not discussed further in this dissertation.

Each of the above software packages implements a numerical method specifically suited for linear or convex programming problems (or more generally semidefinite programming problems). The most powerful methods are the primal-dual interior point methods. There are a number of excellent papers and books on the methods [1, 2, 21–23]. The most important properties of interior point methods are the polynomial complexity and certification of global optimality. Polynomial complexity means that the number of arithmetic operations required to solve the problem is bounded above by a polynomial function of the problem size (number of constraints and number of variables). Certification of global optimality can be made because the primal and dual problems are solved simultaneously. Equally important, the methods can certify infeasibility of the primal and dual problems. This means that the method can recognize in polynomial time whether or not the problem is feasible and if the optimal solution is bounded. This is critically important in real-time applications. The topic of numerical methods is not discussed further in this dissertation. However, Chapter II introduces finite-dimensional optimization using duality theory. This is a good place to start since it is the foundation of all primal-dual interior point methods.

CHAPTER II:
FINITE-DIMENSIONAL OPTIMIZATION

In this chapter, we introduce finite-dimensional optimization – also called parameter optimization. The subject has this name because it is concerned with finding the best points, or parameters, to minimize a function. These points are elements of a finite-dimensional set.

Finite-dimensional optimization is a mature subject with a long history. Our goal here is to prove some of the standard results and to emphasize a few subtle, interesting aspects of the theory. These aspects include attainability, strong duality, and normality. Even in a finite-dimensional setting, these issues are at times difficult to address. They are even more so in the infinite-dimensional setting.

Although there are many ways of proving the results herein, we do so from a duality theory perspective. The chapter begins with the problem statement followed by the formulation of the dual problem. This leads to a generic set of optimality conditions called the Karush-Kuhn-Tucker (KKT) conditions. We then specialize this result to linear and convex programming problems paying close attention to attainability and strong duality. Finally, dual attainability, strong duality, and normality are connected in Lemma 5.

Popular references for much of this material are the books by Berkovitz [24] and Boyd [1]. Many of the results here can be found in one or the other.

A. Problem Description

Consider the finite-dimensional optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{subj. to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned} \tag{1}$$

The optimization variable is $x \in \mathbb{R}^n$, and the cost function is $f : \mathbb{R}^n \supset \Omega_f \to \mathbb{R}$. The inequality constraint function is $g : \mathbb{R}^n \supset \Omega_g \to \mathbb{R}^p$. The equality constraint function is $h : \mathbb{R}^n \supset \Omega_h \to \mathbb{R}^q$. The problem domain

$$\Omega = \Omega_f \ \cap \ \Omega_g \ \cap \ \Omega_h \tag{2}$$

is assumed to be open and nonempty. For the time being, we make no constraint qualifications regarding differentiability, existence of feasible solutions, or otherwise.

For reasons to be made explicit shortly, this problem is called the primal problem. The feasible set $\mathcal{P}$ is the set of all points x that satisfy the constraints.

$$\mathcal{P} = \{x \in \Omega : g(x) \leq 0, \ h(x) = 0\} \tag{3}$$

This set may be empty, finite, or infinite corresponding to no feasible solutions, finitely many feasible solutions, or infinitely many feasible solutions.

We can now define the optimal cost to account for each of these scenarios. The optimal cost is given by p^* where

$$p^* = \begin{cases} +\infty, & \#\mathcal{P} = 0 \\ \inf\{f(x) : x \in \mathcal{P}\}, & \#\mathcal{P} > 0 \end{cases} \tag{4}$$

In words, this means that the optimal cost for an infeasible problem is infinite. The optimal cost for a feasible problem is the greatest lower bound on all feasible solutions. The goal is to find an optimal point x^* such that $p^* = f(x^*)$. Thus, it is clear that by optimal we mean global minimum.

Note that p^* always exists but is not always attainable meaning that there is not always a feasible point x^* such that $p^* = f(x^*)$. In such cases, an optimal solution does not exist. For example, the function $f(x) = x^3$ is unbounded below so that $p^* = -\infty$. The function e^{-x} has a greatest lower bound of zero, but no optimal solution exists since zero is not attainable. These two situations are shown graphically in Figure 1.

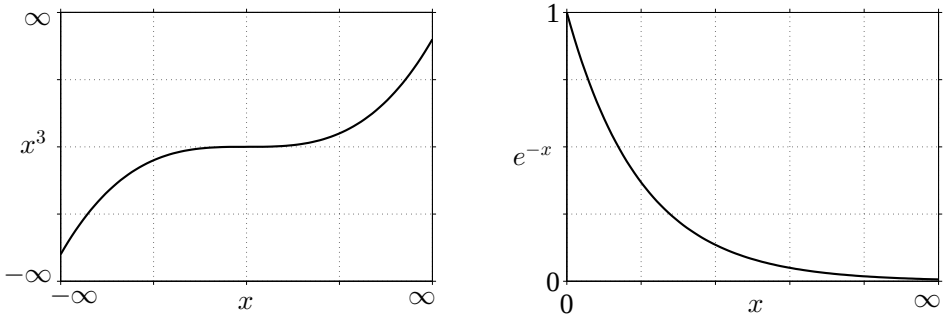

Figure 1: Functions with unattainable costs.

In line with these observations, the attainable optimal solution set is

$$\mathcal{P}^* = \{x^* \in \mathcal{P} : f(x^*) \leq f(x) \; \forall x \in \mathcal{P}\} \tag{5}$$

When the optimal cost is attainable, all of the optimal solutions belong to the set $\mathcal{P}^*$. If the optimal cost is not attainable, then the set $\mathcal{P}^*$ is empty and no optimal solutions exist.

B. Duality Theory

We now introduce duality theory to set the stage for deriving optimality conditions. The duality approach begins with less restrictive assumptions than a classical variational approach, and its results can be specialized to obtain the classical KKT conditions as will be shown. We define the Lagrangian $\mathcal{L} : \Omega \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ as

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^T g(x) + \nu^T h(x) \tag{6}$$

The Lagrange multipliers are $\lambda \in \mathbb{R}^p$ and $\nu \in \mathbb{R}^q$. They are also called the dual variables. The Lagrange dual function, or just dual function, $\ell : \mathbb{R}^p \times \mathbb{R}^q \to \bar{\mathbb{R}}$ is defined to be

$$\ell(\lambda, \nu) = \inf_{x \in \Omega} \mathcal{L}(x, \lambda, \nu) = \inf_{x \in \Omega} f(x) + \lambda^T g(x) + \nu^T h(x) \tag{7}$$

Note that x is not required to belong to the feasible set $\mathcal{P}$, and that the dual function takes the value of $-\infty$ if the Lagrangian is unbounded below in x.

We now prove that the dual function lower bounds the primal cost. This is called weak duality.

Lemma 1. *For any $\lambda \geq 0$ and any ν, $\ell(\lambda, \nu) \leq p^*$.*

Proof. First, consider the case when the primal problem is infeasible. Then $p^* = +\infty$ and the inequality is satisfied trivially. Next, consider the case when the primal problem is feasible. Let $\hat{x}$ be a feasible point and let $\lambda \geq 0$. It follows that

$$\lambda^T g(\hat{x}) + \nu^T h(\hat{x}) \leq 0 \tag{8}$$

since each term in the first product is non-positive and each term in the second product is zero. Therefore, the Lagrangian is bounded above as

$$\mathcal{L}(\hat{x}, \lambda, \nu) = f(\hat{x}) + \lambda^T g(\hat{x}) + \nu^T h(\hat{x}) \leq f(\hat{x}) \tag{9}$$

By taking the infimum, it is clear that the value of the Lagrangian can only be decreased. Thus,

$$\ell(\lambda, \nu) = \inf_{x \in \Omega} \mathcal{L}(x, \lambda, \nu) \leq \mathcal{L}(\hat{x}, \lambda, \nu) \leq f(\hat{x}) \tag{10}$$

Since $\ell(\lambda, \nu) \leq f(\hat{x})$ for all feasible $\hat{x}$, it follows from the definition of p^* that the dual function satisfies $\ell(\lambda, \nu) \leq p^*$. □

Weak duality holds but is trivial when $\ell(\lambda, \nu) = -\infty$. Upon defining the set

$$\Gamma = \{(\lambda, \nu) : \ell(\lambda, \nu) > -\infty\} \tag{11}$$

it is clear that weak duality provides a nontrivial lower bound only when $\lambda \geq 0$ and $(\lambda, \nu) \in \Gamma$. A natural question is, "What is the best lower bound that can be obtained from the dual function?" This question leads to another optimization problem called the Lagrange dual problem, or simply the dual problem.

$$\begin{aligned} \max \quad & \ell(\lambda, \nu) \\ \text{subj. to} \quad & \lambda \geq 0 \\ & (\lambda, \nu) \in \Gamma \end{aligned} \tag{12}$$

All of the statements regarding the primal have analogs for the dual. For example, the dual feasible set is

$$\mathcal{D} = \{(\lambda, \nu) \in \Gamma : \lambda \geq 0\} \tag{13}$$

The attainable optimal solution set is

$$\mathcal{D}^* = \{(\lambda^*, \nu^*) \in \mathcal{D} : \ell(\lambda^*, \nu^*) \geq \ell(\lambda, \nu) \; \forall (\lambda, \nu) \in \mathcal{D}\} \tag{14}$$

and the optimal dual cost, denoted d^*, is given by

$$d^* = \begin{cases} -\infty, & \#\mathcal{D} = 0 \\ \sup\{\ell(\lambda, \nu) : (\lambda, \nu) \in \mathcal{D}\}, & \#\mathcal{D} > 0 \end{cases} \tag{15}$$

By invoking weak duality, we can make two "dual" statements: 1) a feasible, unbounded primal implies an infeasible dual and 2) a feasible, unbounded dual implies an infeasible primal.

The optimal dual cost d^* is the best lower bound on the optimal primal cost p^* that can be obtained from the dual function. In terms of the dual cost, weak duality can be stated as $d^* \leq p^*$. Weak duality is very important since it has been derived under very general terms. The difference $p^* - d^*$ is called the optimal duality gap since it gives the smallest gap between the optimal primal cost and dual cost.

Strong duality occurs when the optimal duality gap is zero such that $d^* = p^*$. Strong duality does not always hold, but it does hold under certain constraint qualifications (CQs). Some of the more popular CQs are the linear CQ, Slater's CQ for convex problems, the linear independence CQ, and the Mangasarian-Fromowitz CQ. When strong duality does hold, we can state some generic optimality conditions for differentiable problems.

Theorem 1 (KKT Conditions)**.** *Assume that f, g, and h are differentiable. If 1) the primal attains a minimum at x^*, 2) the dual attains a maximum at (λ^*, ν^*), and 3) strong duality holds, then the following system is solvable:*

$$\begin{aligned} g(x^*) &\leq 0 \\ h(x^*) &= 0 \\ \lambda^* &\geq 0 \\ \lambda^{*T} g(x^*) &= 0 \\ \partial_x f(x^*) + \partial_x g(x^*)\lambda^* + \partial_x h(x^*)\nu^* &= 0 \end{aligned} \tag{16}$$

Proof. The three hypotheses imply that x^* and (λ^*, ν^*) are feasible so that $g(x^*) \leq 0$, $h(x^*) = 0$, and $\lambda^* \geq 0$. Because the optimal costs are attainable and strong duality holds, $f(x^*) = \ell(\lambda^*, \nu^*)$. Consequently,

$$\begin{aligned} f(x^*) &= \ell(\lambda^*, \nu^*) \\ &= \inf_{x \in \Omega} \left(f(x) + \lambda^{*T} g(x) + \nu^{*T} h(x) \right) \\ &\leq f(x^*) + \lambda^{*T} g(x^*) + \nu^{*T} h(x^*) \\ &\leq f(x^*) \end{aligned} \tag{17}$$

The second line is the definition of the dual function at the optimal dual pair. The third line follows because of the infimum. The fourth line follows from the fact that $\lambda^* \geq 0$, $g(x^*) \leq 0$, and $h(x^*) = 0$. It is obvious that the first and fourth lines hold with equality. Consequently, the third line also holds with equality, and we can make two very important conclusions:

1. The point x^* also minimizes $\mathcal{L}(x, \lambda^*, \nu^*)$ over $x \in \Omega$.

2. The product $\lambda^{*T} g(x^*) = 0$.

When the functions f, g, and h are differentiable, the first conclusion indicates that

$$\partial_x \mathcal{L}(x^*, \lambda^*, \nu^*) = \partial_x f(x^*) + \partial_x g(x^*) \lambda^* + \partial_x h(x^*) \nu^* = 0 \tag{18}$$

since it is an unconstrained minimization problem. □

The conditions (16) are frequently stated as "necessary conditions for x^* to be optimal." However, it is clear that the conditions are only necessary under three assumptions: 1) the optimal primal cost is attainable, 2) the optimal dual cost is attainable, and 3) strong duality holds.

The question of how one can verify these assumptions is important. In the next two sections, we will look at linear and convex programming problems for which there exist the linear CQ and Slater's CQ, respectively. These CQs allow us to remove the dual attainability and strong duality assumptions and strengthen the KKT conditions to be necessary and sufficient.

C. Linear Programming

In this section, we explore the linear programming problem. The problem carries this name because the cost function and constraints are linear. The linear structure can be exploited to arrive at a much stronger statement than the generic KKT conditions in Theorem 1. In particular, the dual attainability and strong duality assumptions can be removed. Also, the conditions can be strengthened to be necessary and sufficient.

One of the standard forms for linear programming problems is stated below alongside its dual problem.

$$\begin{array}{lll} \min & c^T x & \\ \text{subj. to} & Ax \leq b & \end{array} \qquad\qquad \begin{array}{lll} \max & -b^T\lambda & \\ \text{subj. to} & \lambda \geq 0 & \\ & A^T\lambda + c = 0 & \end{array} \tag{19}$$

A sketch of such a problem for two variables is shown in Figure 2.

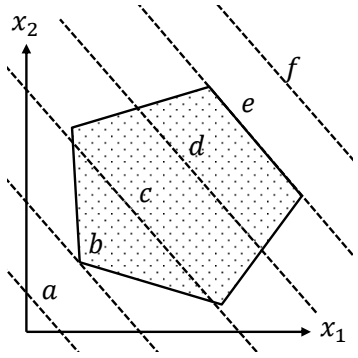

Figure 2: Geometry of a bounded linear programming problem.

Each edge of the polygon represents one of the inequality constraints in $Ax \leq b$. The six dashed lines represent contours of constant cost. If, for example, the f contour has the greatest cost and cost decreases toward a, then the optimal cost is b and the solution is uniquely attained at the apex. On the other hand, if the a contour has the greatest cost and cost decreases toward f, then the optimal cost is e and any point on the edge aligned with e is an optimal solution. The constraints do not have to form a closed polygon. Such an example is shown in Figure 3.

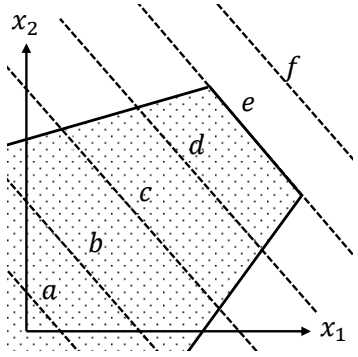

Figure 3: Geometry of an unbounded linear programming problem.

Now, if the f contour has the greatest cost and cost decreases toward a, then the optimal cost is unbounded and no solution is attained. If the direction of decreasing cost reverses, we again have any point on the edge aligned with e attaining the optimal cost.

In an effort to prove KKT conditions for the linear programming problem, we now prove two results known as Farkas' Lemma and Farkas' Corollary. Such results are frequently referred to as theorems of the alternative since they state that exactly one of two systems must be solvable.

Lemma 2. *Exactly one of the following is solvable:*

1. $Ax \leq 0, \qquad c^T x < 0, \quad x \in \mathbb{R}^n$

2. $A^T y + c = 0, \quad y \geq 0, \qquad y \in \mathbb{R}^p$

Proof. Proving that exactly one of the statements is solvable is logically equivalent to proving that both cannot have a solution and that one of them not having a solution implies the other does.

We first show that both cannot have a solution by contradiction. Suppose that both are solvable. Multiplying statement 1 with y implies that $y^T Ax \leq 0$ since $y \geq 0$. Multiplying statement 2 with x implies that $y^T Ax = -c^T x > 0$ since $c^T x < 0$. The quantity $y^T Ax$ cannot be greater than zero and less than or equal to zero. Thus, both statements cannot have a solution.

We now prove that statement 2 not having a solution implies that statement 1 does. Define the set $\mathcal{Q}$ as

$$\mathcal{Q} = \{s : s = A^T y = \sum a_i y_i, y \geq 0\} \tag{20}$$

where a_i are the columns of A^T. Thus, statement 2 is solvable if and only if $-c \in \mathcal{Q}$. Suppose statement 2 does not have a solution. Because $\mathcal{Q}$ is closed and convex [24, p. 62], there is a hyperplane strictly separating $\{-c\}$ and $\mathcal{Q}$ [24, p. 49], i.e., there exist an $\alpha \neq 0$ and β such that

$$\alpha^T(-c) < \beta \quad \text{and} \quad \alpha^T s < \beta \; \forall s \in \mathcal{Q} \tag{21}$$

Since $\mathcal{Q}$ contains the zero element, we know that $\beta > 0$. It is also true that $a_i\lambda \in \mathcal{Q}$ for all $\lambda > 0$. Consequently,

$$\alpha^T(\lambda a_i) < \beta \quad \forall \lambda > 0 \quad \text{such that} \quad \alpha^T a_i < \beta/\lambda \quad \forall \lambda > 0 \tag{22}$$

Since $\beta > 0$, as $\lambda \to \infty$, we get $\alpha^T a_i \leq 0$. By setting $x = \alpha$, it follows that $x^T(-c) > \beta$ such that $c^T x < 0$. Finally, because $\alpha^T a_i \leq 0$, we get that $a_i^T x \leq 0$ for all i. This implies $Ax \leq 0$. Thus, statement 1 is solvable. □

Corollary 1. *Exactly one of the following is solvable:*

1. $Ax \leq b, \qquad x \in \mathbb{R}^n$
2. $A^T y = 0, \quad b^T y < 0, \quad y \geq 0, \quad y \in \mathbb{R}^p$

Proof. We first show that both cannot have a solution. Suppose that both are solvable. Multiplying statement 1 with y implies that $y^T Ax \leq y^T b < 0$ since $y^T b \leq 0$. Multiplying statement 2 with x implies that $y^T Ax = 0$. The quantity $y^T Ax$ cannot be less than zero and equal to zero. Thus, both statements cannot have a solution.

We now prove that statement 2 not having a solution implies that statement 1 does. Note that statement 2 is equivalent to

$$A^T y = 0, \quad b^T y = -\gamma, \quad y \geq 0 \quad \text{for some } \gamma > 0 \tag{23}$$

The two equalities can be combined in matrix form such that

$$\begin{bmatrix} A^T \\ b^T \end{bmatrix} y = \begin{bmatrix} 0 \\ -\gamma \end{bmatrix}, \quad y \geq 0 \quad \text{for some } \gamma > 0 \tag{24}$$

Suppose this is false. Then, from Lemma 2, there exist x and λ such that $Ax + b\lambda \leq 0$ and $\gamma\lambda < 0$. Since $\gamma > 0$, we know that $\lambda < 0$ and

$$A \begin{pmatrix} x \\ \hline -\lambda \end{pmatrix} \leq b \tag{25}$$

Thus, statement 1 is solvable. □

Using Farkas' Lemma and Corollary, the KKT theorem for linear programming problems can be proved. The only assumption required is primal attainability. The assumptions on dual attainability and strong duality are removed. A discussion of attainability and strong duality follows the proof.

Theorem 2 (KKT Conditions for Linear Programming)**.** *The primal attains a minimum at x^* if and only if the following system is solvable:*

$$Ax^* \leq b, \quad \lambda^* \geq 0, \quad \lambda^{*T}(Ax^* - b) = 0, \quad A^T\lambda^* + c = 0 \tag{26}$$

Note: The system in (26) is the same as (16) in Theorem 1.

Proof. The system in (26) is equivalent to

$$Ax^* \leq b, \quad -\lambda^* \leq 0, \quad c^T x^* + b^T \lambda^* \leq 0, \quad A^T \lambda^* \leq -c, \quad -A^T \lambda^* \leq c \quad (27)$$

Case 1: ($\Longrightarrow$) Suppose that the primal attains a minimum at x^* and that (27) does not have a solution. From Corollary 1, and after some algebra, the system (27) does not have a solution provided the following system does.

$$A^T u + cw = 0, \quad Av + bw \geq 0, \quad b^T u - c^T v < 0, \quad u, v, w \geq 0 \qquad (28)$$

Suppose that $w = 0$. Then, $A^T u = 0$, $Av \geq 0$, and either $b^T u < 0$ or $c^T v > 0$. If $b^T u < 0$, then Corollary 1 implies that primal is infeasible, which contradicts the hypothesis that the primal attains a minimum. If $c^T v > 0$, then for all $\gamma > 0$, $A(x - \gamma v) \leq b$. Further, since $c^T v > 0$, we have $c^T(x - \gamma v) \to -\infty$ as $\gamma \to \infty$. This again contradicts the hypothesis that the primal attains a minimum. Thus $w \neq 0$. Suppose that $w > 0$. Dividing through by w gives

$$A^T \left(\frac{u}{w}\right) + c = 0, \quad A\left(-\frac{v}{w}\right) \leq b, \quad c^T \left(-\frac{v}{w}\right) < -b^T \left(\frac{u}{w}\right) \qquad (29)$$

This violates weak duality and implies that (28) does not have a solution. Thus, (27) must be solvable contradicting the original hypothesis.

Case 2: ($\Longleftarrow$) Suppose that (26) is solvable. Then, for any feasible x,

$$c^T x - c^T x^* = \lambda^{*T} A x^* - \lambda^{*T} A x \geq \lambda^{*T}(Ax - b) = 0 \qquad (30)$$

Thus, the primal attains a minimum at x^*. □

We are now in a position to discuss attainability and strong duality. It is shown that the optimal primal cost is attainable if it is finite and that strong duality holds unless the primal and dual are both infeasible.

Lemma 3. *If the optimal primal cost is finite, then it is attainable.*

Proof. We will prove the contrapositive. Suppose that the optimal primal cost is not attainable. Theorem 2 implies that (26) does not have a solution. *Case 1* in the proof of that theorem indicates that (26) not having a solution implies 1) the primal is infeasible, 2) the primal is unbounded below, or 3) weak duality does not hold. Since weak duality must hold, it must be that the primal is infeasible or unbounded below. In either case, the cost is infinite. □

Lemma 4. *If the primal or dual is feasible, then strong duality holds.*

Proof. There are three cases: 1) the primal and dual are feasible, 2) the primal is feasible and the dual is infeasible, and 3) the primal is infeasible and the dual is feasible.

Case 1: Suppose that the primal and dual are feasible and that strong duality does not hold, i.e., $p^* > d^*$. Then, there is no x such that

$$Ax \leq b \quad \text{and} \quad c^T x \leq d^* \tag{31}$$

Corollary 1 implies there is a $\lambda \geq 0$ and $\gamma \geq 0$ for which

$$A^T\lambda + \gamma c = 0 \quad \text{and} \quad b^T\lambda < -\gamma d^* \tag{32}$$

If $\gamma = 0$, then $A^T\lambda = 0$ and $b^T\lambda < 0$. Consequently, Corollary 1 says that

$Ax \leq b$ does not have a solution, which contradicts the hypothesis. Thus, $\gamma \neq 0$. If $\gamma > 0$, then we can divide through by γ to get

$$A^T\left(\frac{\lambda}{\gamma}\right) + c = 0 \quad \text{and} \quad -b^T\left(\frac{\lambda}{\gamma}\right) > d^* \tag{33}$$

This cannot happen by definition of dual optimality. Thus, strong duality must hold.

Case 2: Suppose that the primal is feasible and the dual is infeasible. From Lemma 2, dual infeasibility implies there is an x such that $Ax \leq 0$ and $c^T x < 0$. Let $\hat{x}$ be a primal feasible point. Then, for all $\gamma > 0$,

$$A(\hat{x} + \gamma x) \leq b \tag{34}$$

Further, since $c^T x < 0$, we have that $c^T(\hat{x} + \gamma x) \to -\infty$ as $\gamma \to \infty$. Thus, $p^* = -\infty$ and strong duality holds.

Case 3: Suppose that the primal is infeasible and the dual is feasible. From Corollary 1, primal infeasibility implies there is a λ such that $A^T\lambda = 0$, $b^T\lambda < 0$, and $\lambda \geq 0$. Let $\hat{\lambda}$ be a dual feasible point. Then, for all $\gamma > 0$,

$$A^T(\hat{\lambda} - \gamma\lambda) + c = 0 \tag{35}$$

Further, since $b^T\lambda < 0$, we have that $b^T(\hat{\lambda} - \gamma\lambda) \to \infty$ as $\gamma \to \infty$. Thus, $d^* = \infty$ and strong duality holds. □

D. Convex Programming

We now turn our attention to the convex programming problem. The convex problem is one in which the cost and constraint functions are convex. This immediately implies that the equality constraints must be affine, i.e., $h(x) = Ax - b = 0$. With regard to this constraint, we can assume without loss of generality that matrix A is onto. If not, then there are redundant constraints that can be removed or inconsistent constraints that preemptively make the problem infeasible.

Convex problems can be more complicated than linear problems because of nonlinearities in the cost or inequality constraint functions. Loosely speaking, convexity ensures that the cost function is curved upward on a domain without holes or indentations. Like linear programming problems, convex programming problems have enough structure to significantly strengthen the generic KKT conditions in Theorem 1. In particular, the three hypotheses of Theorem 1 reduce to primal attainability and a constraint qualification, and the conditions become necessary and sufficient. These conditions are at the core of numerical methods for convex problems.

We begin the analysis by considering the two sets

$$\begin{aligned} \mathcal{A} &= \{(t,u,v) : \exists\ x \in \Omega \ \text{ s.t. } f(x) \leq t, g(x) \leq u, Ax - b = v\} \\ \mathcal{B} &= \{(s,0,0) : s < p^*\} \end{aligned} \tag{36}$$

Set $\mathcal{A}$ is nonempty and captures the cost, t, and amount of constraint violation, (u,v), for a given point x. If the point x is a feasible point, then u and v are

zero. It can be shown that $\mathcal{A}$ is convex when f and g are convex. Further, the optimal primal cost is

$$p^* = \inf\{t : (t, 0, 0) \in \mathcal{A}\} \tag{37}$$

which is consistent with our earlier definition. Set $\mathcal{B}$ is nonempty provided p^* is greater than $-\infty$. In this case, it is easy to show that $\mathcal{B}$ is convex and does not intersect $\mathcal{A}$. The geometry is illustrated in Figure 4.

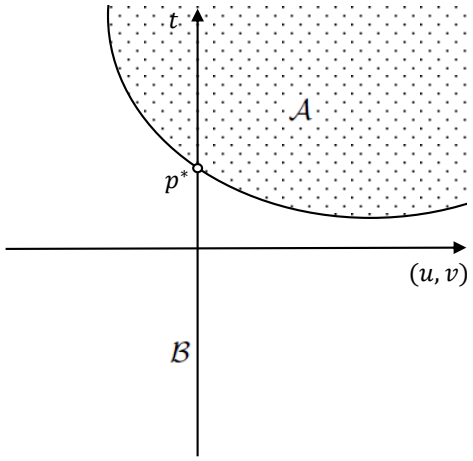

Figure 4: Geometry of sets $\mathcal{A}$ and $\mathcal{B}$.

The set $\mathcal{A}$ is the shaded region in the upper right, and set $\mathcal{B}$ is the line segment along the t-axis below p^*. If the point $(p^*, 0, 0)$ is attainable, then it belongs to $\mathcal{A}$ but not $\mathcal{B}$. If it is not attainable, then it does not belong to either. This geometry and separation of convex sets motivates the following constraint qualification, which guarantees dual attainability and strong duality when the primal problem is strictly feasible. Strictly feasible points are those satisfying $g(x) < 0$ and $h(x) = 0$.

Theorem 3 (Slater's Constraint Qualification)**.** *If there exists an $\hat{x} \in \Omega$ that is strictly feasible and the optimal primal cost is finite, then the dual attains a maximum and strong duality holds.*

Proof. By hypothesis, the primal problem is feasible and p^* is finite such that the sets $\mathcal{A}$ and $\mathcal{B}$ are nonempty disjoint convex sets. Hence, they can be separated by a hyperplane [24, p. 53], i.e., there is a vector $(\lambda_0, \lambda, \nu) \neq 0$ and a scalar α such that

$$\begin{aligned} (t,u,v) \in \mathcal{A} &\implies \lambda_0 t + \lambda^T u + \nu^T v \geq \alpha \\ (t,u,v) \in \mathcal{B} &\implies \lambda_0 t + \lambda^T u + \nu^T v \leq \alpha \end{aligned} \tag{38}$$

From the first condition, we deduce that $(\lambda_0, \lambda) \geq 0$. Otherwise, the terms $\lambda_0 t + \lambda^T u$ would be unbounded below. From the second condition, we deduce that $\lambda_0 t \leq \alpha$ for all $t < p^*$, which implies $\lambda_0 p^* \leq \alpha$. Consequently,

$$(t,u,v) \in \mathcal{A} \implies \lambda_0 t + \lambda^T u + \nu^T v \geq \lambda_0 p^*, \ (\lambda_0, \lambda) \geq 0 \tag{39}$$

This statement can be rewritten in terms of x: For any $x \in \Omega$,

$$\lambda_0 f(x) + \lambda^T g(x) + \nu^T (Ax - b) \geq \lambda_0 p^*, \quad (\lambda_0, \lambda) \geq 0 \tag{40}$$

Suppose that $\lambda_0 = 0$. Then, for any $x \in \Omega$,

$$\lambda^T g(x) + \nu^T (Ax - b) \geq 0 \tag{41}$$

At the strictly feasible point $\hat{x} \in \Omega$, $\lambda^T g(\hat{x}) \geq 0$. Because $g(\hat{x}) < 0$ and $\lambda \geq 0$,

we conclude that $\lambda = 0$. Because λ_0, λ, and ν cannot all be zero, the vector ν must be nonzero such that

$$\begin{aligned} \nu^T(Ax - b) &\geq 0 \quad \forall x \in \Omega \\ \nu^T(A\hat{x} - b) &= 0 \end{aligned} \tag{42}$$

Because Ω is open, there is a neighborhood of $\hat{x}$ in Ω. Thus, the inequality can be made negative unless $A^T\nu = 0$. This is impossible since A is an onto matrix. We conclude that $\lambda_0 \neq 0$.

Suppose that $\lambda_0 > 0$. Dividing through by λ_0 gives

$$\mathcal{L}(x, \lambda/\lambda_0, \nu/\lambda_0) \geq p^* \quad \forall x \in \Omega \tag{43}$$

After taking the infimum, we get $\ell(\lambda/\lambda_0, \nu/\lambda_0) \geq p^*$. Combining this with weak duality gives $\ell(\lambda^*, \nu^*) = p^*$ where $\lambda^* = \lambda/\lambda_0$ and $\nu^* = \nu/\lambda_0$. □

Theorem 4 (KKT Conditions for Convex Programming)**.** *Assume that f and g are differentiable. If Slater's CQ holds, then the primal attains a minimum at x^* if and only if the following system is solvable:*

$$\begin{aligned} g(x^*) &\leq 0 \\ Ax^* &= b \\ \lambda^* &\geq 0 \\ \lambda^{*T} g(x^*) &= 0 \\ \partial_x f(x^*) + \partial_x g(x^*)\lambda^* + A^T\nu^* &= 0 \end{aligned} \tag{44}$$

Note: The system in (44) is the same as (16) in Theorem 1.

Proof. Case 1: ($\implies$) Suppose that the primal attains a minimum at x^*. Then the optimal primal cost p^* is finite. Slater's CQ implies that the dual attains a maximum at (λ^*, ν^*) and strong duality holds. Thus, the hypotheses of Theorem 1 are satisfied and the system (16), which is the same as (44), is solvable.

Case 2: ($\impliedby$) Suppose that (44) holds. The first two conditions imply that x^* is feasible. Because the Lagrangian $\mathcal{L}(x, \lambda^*, \nu^*)$ is convex in x and its derivative is zero at x^*, it attains a minimum there. Thus,

$$\begin{aligned} \ell(\lambda^*, \nu^*) &= \mathcal{L}(x^*, \lambda^*, \nu^*) \\ &= f(x^*) + \lambda^{*T} g(x^*) + \nu^{*T}(Ax^* - b) \\ &= f(x^*) \end{aligned} \tag{45}$$

This indicates that the primal attains a minimum at x^*, the dual attains a maximum at (λ^*, ν^*), and strong duality holds. □

In the linear programming section, we followed the KKT theorem with a statement that attainability occurs if the optimal primal cost is finite. Unfortunately, no such statement can be made for convex problems. This is easily seen by trying to minimize the convex function $f(x) = e^{-x}$ without constraints. As discussed earlier and shown in Figure 1, the optimal primal cost is finite but cannot be attained.

E. Normality in Convex Programming

In proving Theorem 1, we used three assumptions: 1) primal attainability, 2) dual attainability, and 3) strong duality. The first assumption is expected. The reasons for the second and third assumptions are less clear since they involve the dual variables, which are not directly part of the original problem. For convex problems, Slater's CQ was invoked, and this served as a sufficient condition for dual attainability and strong duality. We are now interested in removing all assumptions except primal attainability and proving new optimality conditions for convex problems. Doing so leads to what are generally called the Fritz John (FJ) conditions.

Theorem 5 (FJ Conditions for Convex Programming)**.** *Assume that f and g are differentiable. If the primal attains a minimum at x^*, then the following system is solvable:*

$$\begin{aligned} (\lambda_0^*, \lambda^*, \nu^*) &\neq 0 \\ g(x^*) &\leq 0 \\ Ax^* &= b \\ (\lambda_0^*, \lambda^*) &\geq 0 \\ \lambda^{*T} g(x^*) &= 0 \\ \lambda_0^* \partial_x f(x^*) + \partial_x g(x^*)\lambda^* + A^T \nu^* &= 0 \end{aligned} \tag{46}$$

Proof. The optimal point x^* is a feasible point. Thus, $g(x^*) \leq 0$ and $Ax^* = b$. Everything from the proof of Theorem 3 remains true up to and including (40). That is, there is a vector $(\lambda_0^*, \lambda^*, \nu^*) \neq 0$ satisfying $(\lambda_0^*, \lambda^*) \geq 0$ and

$$\lambda_0^* f(x) + \lambda^{*T} g(x) + \nu^{*T}(Ax - b) \geq \lambda_0^* p^*, \quad (\lambda_0^*, \lambda^*) \geq 0, \quad \forall x \in \Omega \tag{47}$$

Case 1: Suppose that $\lambda_0^* = 0$. Then,

$$\lambda^{*T} g(x) + \nu^{*T}(Ax - b) \geq 0 \quad \forall x \in \Omega \tag{48}$$

At the optimal point x^*, the second terms goes to zero. Since $\lambda^{*T} \geq 0$ and $g(x^*) \leq 0$, we conclude that the product must be zero: $\lambda^{*T} g(x^*) = 0$. Since the above inequality is zero at x^*, it is minimized at that point. Thus,

$$\partial_x g(x^*)\lambda^* + A^T \nu^* = 0 \tag{49}$$

which implies that the system (46) is solvable with $\lambda_0^* = 0$.

Case 2: Suppose that $\lambda_0^* > 0$. Dividing through by λ_0^* gives

$$\mathcal{L}(x, \lambda^*/\lambda_0^*, \nu^*/\lambda_0^*) \geq p^* \quad \forall x \in \Omega \tag{50}$$

Consequently, $\ell(\lambda^*/\lambda_0^*, \nu^*/\lambda_0^*) \geq p^*$. Combining this with weak duality gives $\ell(\lambda^*/\lambda_0^*, \nu^*/\lambda_0^*) = p^*$. Therefore, the dual attains a maximum at $(\lambda^*/\lambda_0^*, \nu^*/\lambda_0^*)$ and strong duality holds. The hypotheses of Theorem 1 are satisfied implying that the system (46) is solvable with $\lambda_0^* = 1$. □

Remark 1. *If a solution satisfies Equation 46 with $\lambda_0^* = 1$, then the solution is called a normal solution. If a solution satisfies Equation 46 and λ_0^* must be zero, then the solution is called an abnormal solution. Note that some authors denote any solution with $\lambda_0^* = 0$ as an abnormal solution, and solutions where λ_0^* must be zero as a strictly abnormal solution.*

Theorem 1 indicates that dual attainability and strong duality imply normality – even for non-convex problems. The proof of Theorem 5 suggests that the reverse implication is also true for convex problems. This result is formalized in Lemma 5.

Lemma 5. *Suppose the primal attains a minimum at x^*. Dual attainability and strong duality hold if and only if normality holds.*

Proof. If dual attainability and strong duality hold, Theorem 1 implies that normality holds. If normality holds, then the last of (46) indicates that $\partial_x \mathcal{L}(x^*, \lambda^*, \nu^*) = 0$. Because $\mathcal{L}$ is convex in x, it attains a minimum at x^*. Thus,

$$\mathcal{L}(x^*, \lambda^*, \nu^*) = \inf_{x \in \Omega} \mathcal{L}(x, \lambda^*, \nu^*) \tag{51}$$

Weak duality implies that $f(x^*) \geq d^*$. Additionally, by definition of the dual cost, $d^* \geq \ell(\lambda^*, \nu^*)$. Thus, the following inequalities hold.

$$\begin{aligned} f(x^*) &\geq d^* \\ &\geq \ell(\lambda^*, \nu^*) \\ &= \mathcal{L}(x^*, \lambda^*, \nu^*) \\ &= f(x^*) \end{aligned} \tag{52}$$

It is obvious that the first and fourth lines hold with equality such that the intermediate lines do as well. Thus, $f(x^*) = \ell(\lambda^*, \nu^*)$, i.e., the dual attains a maximum at (λ^*, ν^*) and strong duality holds. $\square$

The above result cannot be generalized for non-convex problems since (51) does not necessarily hold. Additionally, there are known examples where normality holds but strong duality does not.

CHAPTER III:
OPTIMAL CONTROL THEORY

The purpose of this chapter is to state and prove necessary conditions for global optimality of optimal control problems. These necessary conditions fall within the scope of the now famous work by Pontryagin, and they are a result of what is commonly called Pontryagin's principle, the maximum principle, the minimum principle, or some combination.

Optimal control problems are infinite-dimensional optimization problems. This is because the objective is to find the best control functions to minimize an integral. These control functions are elements of an infinite-dimensional space – a function space – such as the space of continuous functions, piecewise continuous functions, or measurable functions. For this reason, optimal control theory is more involved than finite-dimensional optimization theory.

Optimal control theory has a rich history and is an outgrowth of the classical calculus of variations. Centuries of work culminated in the 1950s and 1960s with the results of Pontryagin and his colleagues [25]. We will prove their result for control functions belonging to the space of piecewise continuous functions. This is weaker than their original result, but it is sufficient for our purposes.

In the author's view, one of the more important aspects of the maximum principle is its removal of normality assumptions (assumptions on the existence of control variations). Any conditions that operate under normality assump-

tions without qualification are not actually necessary, and so this aspect is an important one. The literature indicates that McShane first resolved the normality issue in the calculus of variations in 1939 [26]. Even after the work of McShane and Pontryagin, normality assumptions persisted in singular optimal control theory [27–30] until 1977 when Krener removed them [31].

As in finite-dimensional optimization, the issues of attainability and normality are interesting. There are many examples where the cost is lower bounded but not attainable. The simplest example is to minimize the control effort required to drive a stable, linear time-invariant system to the origin in free final time. Any feasible solution can be improved by extending the final time, but the lower bound of zero control effort is not attainable. The issue of attainability has been addressed by Berkovitz under certain convexity and compactness assumptions [32, Ch. 3].

There are also many problems that have abnormal solutions. The simplest example is to minimize the time it takes to drive a harmonic oscillator to the origin with a bounded control. For certain initial conditions, the global optimal solution is unique and abnormal. If one were to study this physically motivated problem under a normality assumption, he would incorrectly conclude that no optimal solution exists. The issue of normality is typically addressed on a case by case basis. The details of attainability and normality in optimal control are beyond the scope of this chapter.

The primary references for this chapter are the original work by Pontryagin [25] and the more recent exposition by Liberzon [33]. We first work with a basic optimal control problem followed by a problem with state constraints.

A. Problem Description

In optimal control, one controls a dynamic system over a finite interval $[t_0, t_f]$ so that a given performance index is minimized. The control function $u(\cdot)$ is assumed to belong to the space of piecewise continuous functions, i.e.,

$$u(\cdot) \in \mathcal{PC} : [t_0, t_f] \to \Omega \subset \mathbb{R}^m \tag{1}$$

Thus, it is continuous except at a finite number of points where it is discontinuous. The points of continuity are called regular points; the points of discontinuity are called irregular points. Without loss of generality, we assume that $u(\cdot)$ is continuous from the left, i.e.,

$$u(t_p) = \lim_{t \uparrow t_p} u(t) \tag{2}$$

for all irregular points t_p. The set Ω is the control constraint set, and it is assumed to be fixed. Dynamic systems of interest take the form

$$\dot{x}(t) = f(x(t), u(t)), \quad x(t_0) = a, \quad x(t_f) = b \tag{3}$$

so that the state evolves according to ordinary differential equations.

Remark 1. *When writing differential equations, the equality sign is used to mean "equal almost everywhere." The reason is that the integral of a piecewise continuous function is only differentiable at the regular points of the function [34, p. 133].*

Consequently, the state function $x(\cdot)$ is absolutely continuous, i.e.,

$$x(\cdot) \in \mathcal{AC} : [t_0, t_f] \to \mathbb{R}^n \tag{4}$$

It is assumed that $f(\cdot, \cdot) \in \mathcal{C}$ and $f(\cdot, \omega) \in \mathcal{C}^1$ for each fixed ω. In words, it is continuous in both arguments and continuously differentiable with respect to the first.

Remark 2. *These assumptions in conjunction with piecewise continuity of* $u(\cdot)$ *guarantee local existence and uniqueness for solutions of Equation 3. These assumptions can be relaxed using local Lipschitz arguments [33, p. 83-86].*

The cost is given by the scalar quantity $J = x_0(t_f)$ where $x_0(\cdot)$ satisfies the differential equation

$$\dot{x}_0(t) = \ell(x(t), u(t)), \quad x_0(t_0) = 0 \tag{5}$$

The function $\ell(\cdot, \cdot)$ shares the same properties as $f(\cdot, \cdot)$. Given the existence and uniqueness assumptions, the cost is a function of the control function, i.e., $J = J[u(\cdot)]$.

The feasible set $\mathcal{F}$ is the set of all control functions that satisfy the constraints.

$$\mathcal{F} = \{u(\cdot) : u(\cdot) \text{ satisfies Equation 1 and } x(\cdot) \text{ satisfies Equation 3}\} \tag{6}$$

This set may be empty, finite, or infinite corresponding to no feasible solutions, finitely many feasible solutions, or infinitely many feasible solutions. As in

finite-dimensional optimization, the optimal cost is given by J^* where

$$J^* = \begin{cases} +\infty, & \#\mathcal{F} = 0 \\ \inf\{J[u(\cdot)] : u(\cdot) \in \mathcal{F}\}, & \#\mathcal{F} > 0 \end{cases} \tag{7}$$

In words, this means that the optimal cost for an infeasible problem is infinite. The optimal cost for a feasible problem is the greatest lower bound on all feasible solutions. The goal is to find an optimal control $u^*(\cdot)$ such that $J^* = J[u^*(\cdot)]$. Thus, it is clear that by optimal we mean global minimum. As mentioned in the introduction to this chapter, attainability is an important question, but it is beyond the scope of this chapter. Nonetheless, the attainable optimal solution set is

$$\mathcal{F}^* = \{u^*(\cdot) : J[u^*(\cdot)] \leq J[u(\cdot)] \; \forall u(\cdot) \in \mathcal{F}\} \tag{8}$$

When the optimal cost is attainable, all of the optimal solutions belong to the set $\mathcal{F}^*$. If the optimal cost is not attainable, then the set $\mathcal{F}^*$ is empty and no optimal solutions exist.

For the sake of brevity, it is common to drop the function space and differentiability requirements and state the problem more concisely. With those requirements still in force, a concise statement of the basic optimal control

problem (BOCP) is below.

$$\begin{aligned} \min \quad & J = x_0(t_f) \\ \text{subj. to} \quad & \dot{x}_0(t) = \ell(x(t), u(t)), \quad x_0(t_0) = 0 \\ & \dot{x}(t) \ = f(x(t), u(t)), \quad x(t_0) = a, \quad x(t_f) = b \\ & u(t) \in \Omega, \quad t_0 \text{ fixed}, \quad t_f \text{ free} \end{aligned} \tag{BOCP}$$

Before stating the necessary conditions that we will prove in the following sections, it is convenient to introduce the following functions

$$y(t) = \begin{bmatrix} x_0(t) \\ x(t) \end{bmatrix}, \quad q(t) = \begin{bmatrix} p_0(t) \\ p(t) \end{bmatrix}, \quad g(y(t), u(t)) = \begin{bmatrix} \ell(x(t), u(t)) \\ f(x(t), u(t)) \end{bmatrix} \tag{9}$$

such that $\dot{y}(t) = g(y(t), u(t))$. Then, the Hamiltonian is defined to be

$$\begin{aligned} \mathcal{H}(x(t), u(t), p(t), p_0(t)) &= \langle q(t), g(y(t), u(t)) \rangle \\ &= \langle p_0(t), \ell(x(t), u(t)) \rangle + \langle p(t), f(x(t), u(t)) \rangle \end{aligned} \tag{10}$$

Lastly, we define the terminal sets

$$\begin{aligned} S_b &= \{x : x = b\} \\ S_b' &= \{(x_0, x) : x_0 \in \mathbb{R} \text{ and } x \in S_b\} \\ S_b'' &= \{(x_0, x) : x_0 < J^* \text{ and } (x_0, x) \in S_b'\} \end{aligned} \tag{11}$$

The final point must belong to S_b since $x(t_f) = b$, and S_b' simply lifts S_b into the x_0 direction. The set S_b'' consists only of those points in S_b' that have lower cost than the optimal. Figures 5 and 6 illustrate the geometric significance.

Note the different axes used in these figures. They are used repeatedly throughout the rest of the chapter.

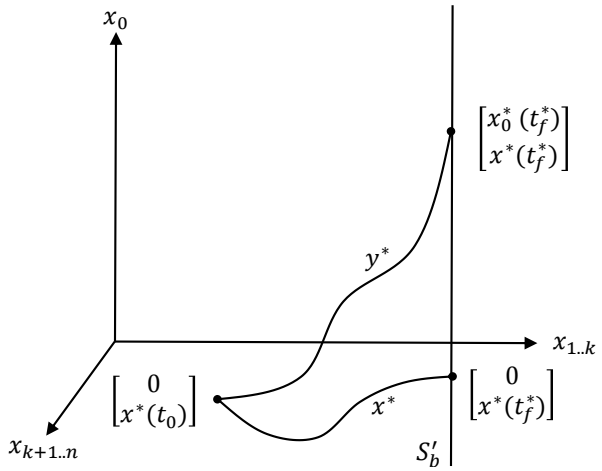

Figure 5: Optimal solutions in the original and lifted space.

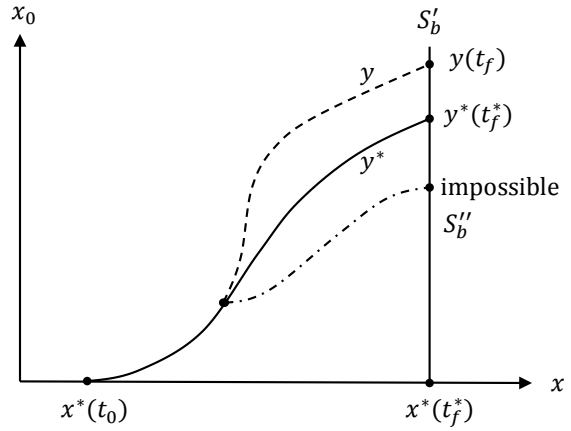

Figure 6: Suboptimal, optimal, and impossible solutions.

We can now state necessary conditions for BOCP. The original result is due to Pontryagin. The next few sections prove this theorem by investigating properties of the terminal cone, adjoint system, and Hamiltonian.

Theorem 1 (Necessary Conditions for BOCP)**.** *If BOCP attains a minimum at $u^*(\cdot)$, then the following system is solvable:*

i) the normality condition

$$p_0^* \leq 0 \tag{12}$$

ii) the non-triviality condition

$$(p_0^*, p^*(t)) \neq 0 \quad \forall t \in [t_0, t_f^*] \tag{13}$$

iii) the differential equations

$$\begin{aligned} \dot{x}_0^*(t) &= \ell(x^*(t), u^*(t)) \\ \dot{x}^*(t) &= \partial_p \mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*) \\ -\dot{p}^*(t) &= \partial_x \mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*,) \end{aligned} \tag{14}$$

iv) the pointwise maximum condition

$$u^*(t) = \arg\max_{\omega \in \Omega} \mathcal{H}(x^*(t), \omega, p^*(t), p_0^*,) \quad \text{a.e. } t \in [t_0, t_f^*] \tag{15}$$

v) the Hamiltonian condition

$$\mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*) = 0 \quad \forall t \in [t_0, t_f^*] \tag{16}$$

vi) the boundary conditions

$$x_0^*(t_0) = 0, \quad x^*(t_0) = a, \quad x^*(t_f^*) = b \tag{17}$$

B. Generating the Terminal Cone

1. Temporal Perturbations

In this section, we explore how the final point of the optimal trajectory $y^*(\cdot)$ varies with respect to changes in the final time. In all that follows, ϵ is a constant, positive scalar. The perturbed final time is

$$t_f = t_f^* + \epsilon\tau \tag{18}$$

where τ is a scalar that can be positive or negative. For every τ, we define a perturbed control function $u(\cdot)$, which generates $y(\cdot)$. The perturbed control is

$$u(t) = \begin{cases} u^*(t), & t_0 \le t \le t_f^* \\ u^*(t_f^*), & t_f^* < t \le t_f \end{cases} \tag{19}$$

Thus, it is well-defined for all perturbation times. This is illustrated graphically in Figure 7.

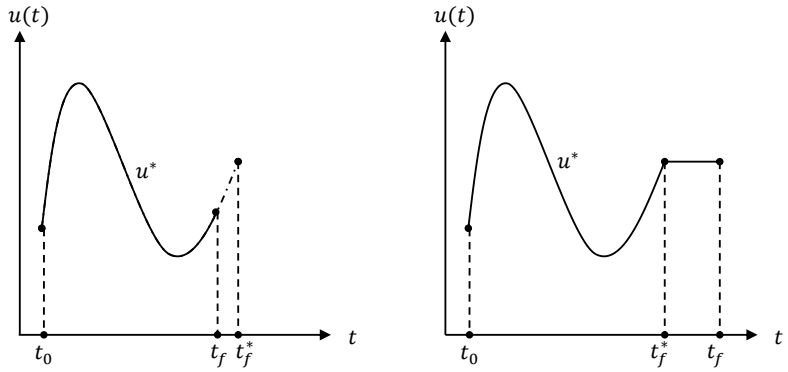

Figure 7: Temporal perturbations.

In exploring the affect of perturbing the final time, we first recognize that the point t_f^* is a regular point of $u^*(\cdot)$ and $u(\cdot)$. Therefore, the time derivatives of $y^*(\cdot)$ and $y(\cdot)$ are well-defined at t_f^*. When $\tau < 0$, it is known that $y(t_f) = y^*(t_f)$. Thus,

$$\begin{aligned} y(t_f) &= y^*(t_f^* + \epsilon\tau) \\ &= y^*(t_f^*) + g(y^*(t_f^*), u^*(t_f^*))\epsilon\tau + o(\epsilon) \end{aligned} \tag{20}$$

When $\tau > 0$, we have

$$\begin{aligned} y(t_f) &= y(t_f^* + \epsilon\tau) \\ &= y(t_f^*) + g(y(t_f^*), u(t_f^*))\epsilon\tau + o(\epsilon) \end{aligned} \tag{21}$$

At the optimal final time, it is true that $y(t_f^*) = y^*(t_f^*)$ and $u(t_f^*) = u^*(t_f^*)$. Hence, when $\tau > 0$,

$$y(t_f) = y^*(t_f^*) + g(y^*(t_f^*), u^*(t_f^*))\epsilon\tau + o(\epsilon) \tag{22}$$

which is the same as for the case when $\tau < 0$. By defining the linear function $\delta(\tau) = g(y^*(t_f^*), u^*(t_f^*))\tau$, the perturbed final point can be simply written as

$$y(t_f) = y^*(t_f^*) + \epsilon\delta(\tau) + o(\epsilon) \tag{23}$$

The set

$$\Delta_t = \{\xi : \xi = y^*(t_f^*) + \epsilon\delta(\tau),\ \epsilon \text{ fixed}\} \tag{24}$$

represents a linear approximation to the set of all reachable states achievable by perturbing the optimal final time. Figure 8 shows the set Δ_t in context of

the optimal trajectory.

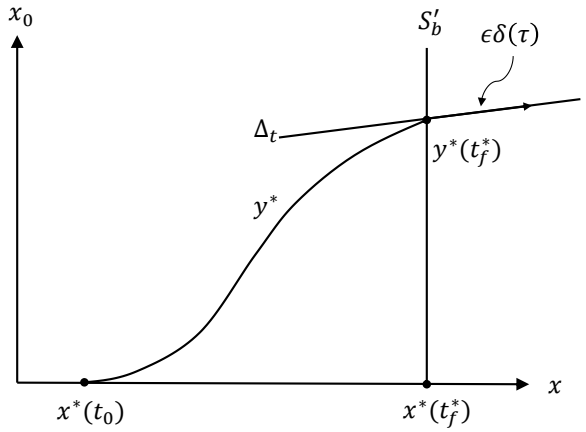

Figure 8: Linear approximation of temporal variations.

2. Spatial Perturbations

In this section, we explore how the trajectory varies with respect to changes in the control. These changes will be constructed by introducing pulses away from the optimal control. Let I be the interval

$$I = (t_p - \epsilon\sigma, t_p] \subset (t_0, t_f^*) \tag{25}$$

where t_p is a regular point of $u^*(\cdot)$ and σ is a positive scalar. Let ω be an arbitrary element of the control set Ω. Then the simplest spatial control perturbation is

$$u(t) = \begin{cases} u^*(t), & t \notin I \\ \omega, & t \in I \end{cases} \tag{26}$$

This scenario is shown in Figure 9.

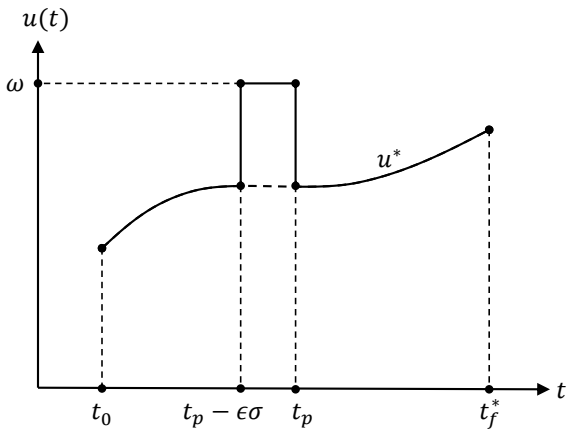

Figure 9: Simplest spatial perturbation.

We now want to study how the perturbed trajectory deviates from the optimal on the interval I. Because t_p is a regular point, we can write

$$\begin{aligned} y^*(t_p - \epsilon\sigma) &= y^*(t_p) - \dot{y}^*(t_p)\epsilon\sigma + o(\epsilon) \\ &= y^*(t_p) - g(y^*(t_p), u^*(t_p))\epsilon\sigma + o(\epsilon) \end{aligned} \tag{27}$$

Note that the point $t_p - \epsilon\sigma$ is not a regular point of $u(\cdot)$ since, by construction, it is discontinuous there. Nonetheless, we can use the right-sided derivative, $\dot{y}^+(t_p - \epsilon\sigma)$, to write

$$\begin{aligned} y(t_p) &= y(t_p - \epsilon\sigma) + \dot{y}^+(t_p - \epsilon\sigma)\epsilon\sigma + o(\epsilon) \\ &= y(t_p - \epsilon\sigma) + g(y(t_p - \epsilon\sigma), \omega)\epsilon\sigma + o(\epsilon) \end{aligned} \tag{28}$$

Because $y(t_p - \epsilon\sigma) = y^*(t_p - \epsilon\sigma)$, it follows that

$$y(t_p) = y^*(t_p) - g(y^*(t_p), u^*(t_p))\epsilon\sigma + g(y^*(t_p - \epsilon\sigma), \omega)\epsilon\sigma + o(\epsilon) \tag{29}$$

The third term on the right hand side can be expanded using the series notation about t_p. After doing so, the above equation becomes

$$y(t_p) = y^*(t_p) + \gamma_p(\omega)\epsilon\sigma + o(\epsilon) \tag{30}$$

where

$$\gamma_p(\omega) = g(y^*(t_p), \omega) - g(y^*(t_p), u^*(t_p)) \tag{31}$$

This perturbed value at time t_p propagates forward to the optimal final time. To characterize this effect, we introduce the function $\eta(\cdot) \in \mathcal{AC} : [t_p, t_f^*] \to \mathbb{R}^{n+1}$ such that

$$y(t) = y^*(t) + \epsilon\eta(t) + o(\epsilon) \tag{32}$$

for which it is already known that $\eta(t_p) = \gamma_p(\omega)\sigma$. On the interval $[t_p, t_f^*]$, $u(\cdot)$ and $u^*(\cdot)$ share the same regular and irregular points. Differentiating at the regular points time and solving for the first-order part gives

$$\dot{\eta}(t) = \partial_y^T g(y^*(t), u^*(t))\eta(t), \quad \eta(t_p) = \gamma_p(\omega)\sigma \tag{33}$$

which, as with other differential equations, only holds almost everywhere. The solution of such a linear system is frequently written in terms of the state transition matrix, i.e., $\eta(t) = \Phi(t, t_p)\gamma_p(\omega)\sigma$, where $\Phi(t, t_p)$ satisfies

$$\dot{\Phi}(t, t_p) = \partial_y^T g(y^*(t), u^*(t))\Phi(t, t_p), \quad \Phi(t_p, t_p) = I \tag{34}$$

Thus, at the optimal final time, the perturbed state is given by

$$y(t_f^*) = y^*(t_f^*) + \Phi(t_f^*, t_p)\gamma_p(\omega)\epsilon\sigma + o(\epsilon) \tag{35}$$

Defining $\delta(\omega, I) = \Phi(t_f^*, t_p)\gamma_p(\omega)\sigma$, we have

$$y(t_f^*) = y^*(t_f^*) + \epsilon\delta(\omega, I) + o(\epsilon) \tag{36}$$

A sketch of the first-order behavior is shown in Figure 10.

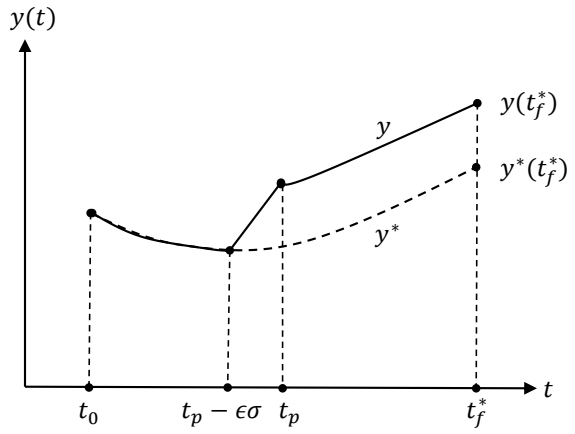

Figure 10: Effect of a spatial perturbation.

The set

$$\Delta_s(\omega, t_p) = \{\xi : \xi = y^*(t_f^*) + \epsilon\delta(\omega, I),\ \epsilon \text{ fixed}\} \tag{37}$$

represents a linear approximation of all reachable states achievable by the simplest spatial perturbation. The elements of the set are a linear function of σ. This set is illustrated in Figure 11.

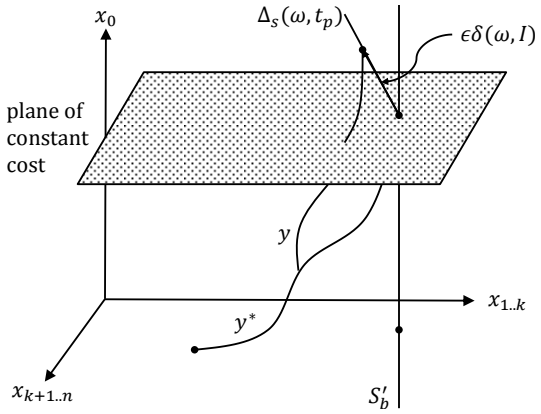

Figure 11: Linear approximation of a spatial perturbation.

As indicated by the notation, the set $\Delta_s(\omega, t_p)$ depends on the particular control ω and the time t_p. There are an infinite number of times that could be used as well as different controls (finite or infinite depending on the control set). We now define the set Δ_s to be the union of all possible sets of the form $\Delta_s(\omega, t_p)$. That is, Δ_s is a linear approximation of all reachable states achievable by all the possible simplest perturbations. Figure 12 shows a simple illustration of this set.

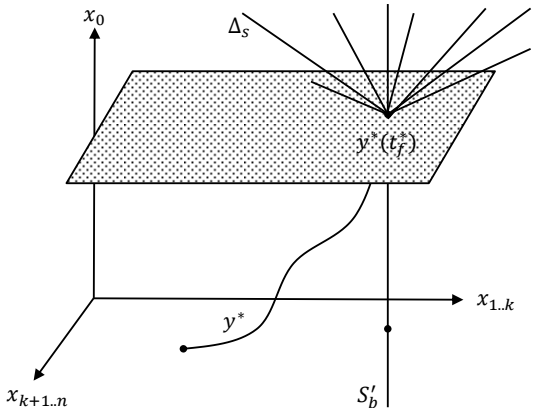

Figure 12: Linear approximation of multiple spatial perturbations.

Note that Δ_s is a cone emanating from the optimal final point $y^*(t_f^*)$, and it may be convex or not depending on the problem.

We now construct a perturbed control that contains two pulses over the distinct intervals I_1 and I_2.

$$I_1 = (t_1 - \epsilon\sigma_1, t_1] \quad \text{and} \quad I_2 = (t_2 - \epsilon\sigma_2, t_2] \tag{38}$$

Again, it must be assumed that t_1 and t_2 are regular points of the optimal control $u^*(\cdot)$. It is also assumed that the intervals do not overlap and that I_1 precedes I_2. This scenario is shown in Figure 13.

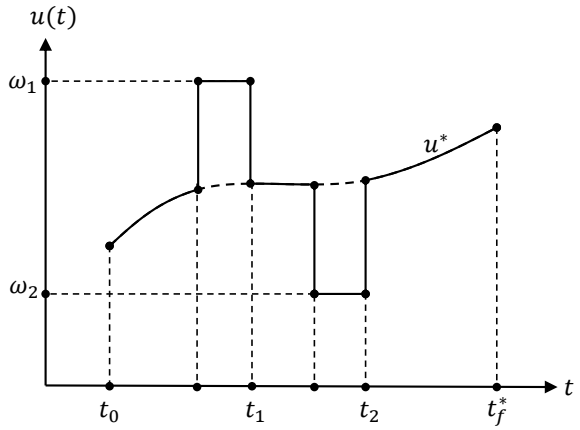

Figure 13: Two spatial perturbations.

Using the formulas just derived and properties of the state transition matrix, it follows that

$$y(t_f^*) = y^*(t_f^*) + \epsilon\delta(\omega_1, I_1) + \epsilon\delta(\omega_2, I_2) + o(\epsilon) \tag{39}$$

That is, two spatial control perturbations "add" together to affect the final point. A simple inductive argument shows that this is true for any finite number of perturbations. This requires some notational complexity but is given by Pontryagin [25, p. 86-92].

Conversely, it is easy to see that a final perturbation of the form

$$y(t_f^*) = y^*(t_f^*) + \epsilon r_1 \delta(\omega_1, I_1) + \epsilon r_2 \delta(\omega_2, I_2) + o(\epsilon) \tag{40}$$

with $r_1, r_2 \geq 0$ is possible by simply changing the intervals to

$$I_1' = (t_1 - \epsilon r_1 \sigma_1, t_1] \quad \text{and} \quad I_2' = (t_2 - \epsilon r_2 \sigma_2, t_2] \tag{41}$$

By choosing all values of r_1 and r_2 such that $r_1 + r_2 = 1$, we construct all possible convex combinations, i.e., the convex hull of the sets $\Delta_s(\omega_1, t_1)$ and $\Delta_s(\omega_2, t_2)$. This hull is a linear approximation to the set of all reachable states generated by the two pulses at t_1 and t_2.

We then extend this idea to include all convex combinations of points in the set Δ_s. What results is the convex hull of Δ_s, denoted $\text{co}(\Delta_s)$. The set $\text{co}(\Delta_s)$ is a linear approximation to the set of all reachable states generated by all possible convex combinations of simple perturbations. The set is also a convex cone emanating from the point $y^*(t_f^*)$. Figure 14 provides an illustration.

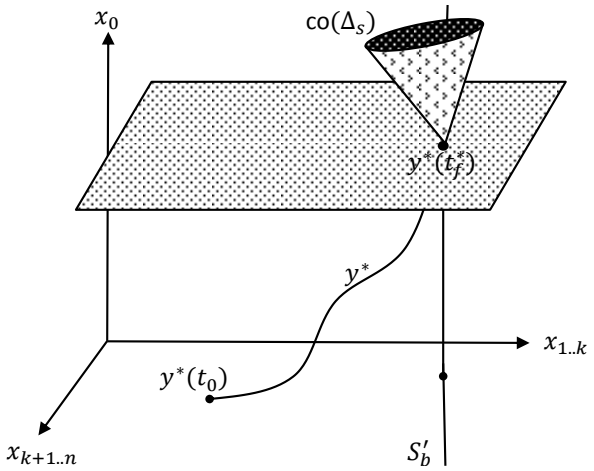

Figure 14: Convex hull of spatial perturbations.

We now construct a linear approximation to the set of all reachable states generated by affine combinations of elements in Δ_t and co(Δ_s). This set is given by

$$\Delta = \{\xi : \xi = y^*(t_f^*) + \epsilon r_0 \delta(\tau) + \epsilon \sum_{i=1}^{k} r_i \delta(\omega_i, I_i)\} \tag{42}$$

where $r_i \geq 0$. The fact that any element in co(Δ_s) can be constructed using a finite number of pulses has been proved by Carathéodory [24, p. 41]. The set Δ is convex, and because of its special significance, it is called the terminal cone. It is shown in Figure 15. It is the infinite wedge between the two half-planes.

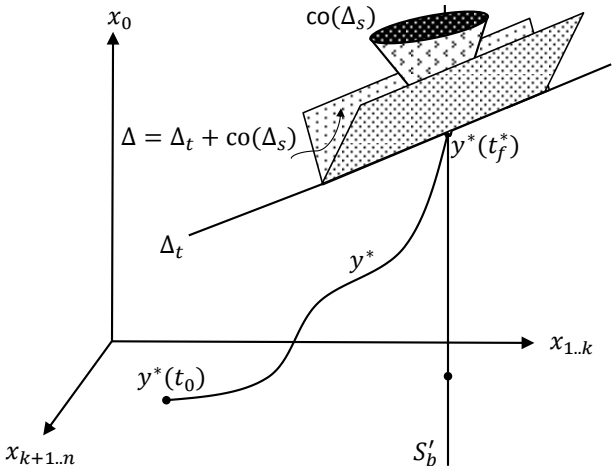

Figure 15: Terminal cone.

C. Properties of the Terminal Cone

The significance of the terminal cone is the following: for every point $\xi \in \Delta$, there is a perturbation of the optimal control such that the terminal point of the perturbed trajectory is within an order of epsilon, i.e., there is a perturbed control generating $y(t_f)$ such that

$$y(t_f) = \xi + o(\epsilon) \tag{43}$$

where $y(t_f)$ may or may not be in the terminal cone.

We now use the fact that $u^*(\cdot)$ is the optimal control, and hence, no other control gives a lower cost. Geometrically, no other trajectory intersects the terminal set S_b' at a lower point. Since the terminal cone Δ is a linear approximation of the reachable set, it is expected that the terminal cone should face "upward." This must be proved.

Consider the vector $\pi = (-1, 0, \ldots, 0) \in \mathbb{R}^{n+1}$, which points downward. Let all of the points on the half-ray emanating from $y^*(t_f^*)$ in the direction of π be denoted by the set Π.

$$\Pi = \{\xi : \xi = y^*(t_f^*) + r\pi, r \geq 0\} \tag{44}$$

This set is illustrated below in Figure 16.

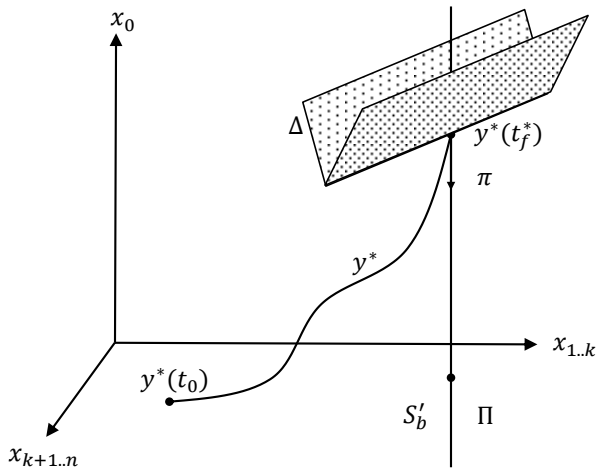

Figure 16: Terminal wedge and vector.

We must prove the following lemma.

Lemma 1. *The set* Π *does not intersect the interior of* Δ.

Before proving this lemma, we will outline a popular line of reasoning, which is incorrect. Suppose that the lemma were false and that Π did intersect the interior of Δ. Then there exists a control $u(\cdot)$ that generates

$$y(t_f) = y^*(t_f^*) + \epsilon r\pi + o(\epsilon) \tag{45}$$

for some $r > 0$. Writing $y(t_f)$ in terms of its components yields

$$\begin{aligned} J[u(\cdot)] &= J[u^*(\cdot)] - \epsilon r + o(\epsilon) \\ x(t_f) &= x^*(t_f^*) + o(\epsilon) \end{aligned} \tag{46}$$

The perturbed control clearly gives a lower cost since $\epsilon, r > 0$. Therefore, some may conclude, this contradicts the definition of an optimal control. However, it is important to note that the perturbed control is not feasible since $x(t_f) \neq x^*(t_f^*)$. At this point, another popular approach is to make a normality assumption. This was done in the calculus of variations until 1939 when McShane resolved the issue [26]. It was also done in singular optimal control theory until Krener resolved the issue there [31]. Lemma 1 is proved below.

Proof. Suppose that the lemma is false and that Π does intersect the interior of Δ. Then there exists a point ξ such that $\xi \in \Pi$ and $\xi \in \text{int}(\Delta)$. This implies that the point must be of the form

$$\xi = y^*(t_f^*) + \epsilon r \pi \tag{47}$$

for some $r > 0$. By definition of interior, the second inclusion implies that there is a ball centered at ξ with radius ϵ, $B(\xi, \epsilon)$, such that $B(\xi, \epsilon) \subset \text{int}(\Delta)$. Further, it implies that all elements of $B(\xi, \epsilon)$ have the form

$$y^*(t_f^*) + \epsilon\gamma \tag{48}$$

where γ is a function of the perturbation parameters such as τ, ω, and I.

Corresponding to each element in $B(\xi, \epsilon)$, there is a point of the form

$$y^*(t_f^*) + \epsilon\gamma + o(\epsilon) \tag{49}$$

which is the actual terminal point, as opposed to the linear approximation. The set of actual terminal points is denoted $\tilde{B}(\xi, \epsilon)$. Since it is only $o(\epsilon)$ away from $B(\xi, \epsilon)$, it can be thought of as a warping as shown in Figure 17.

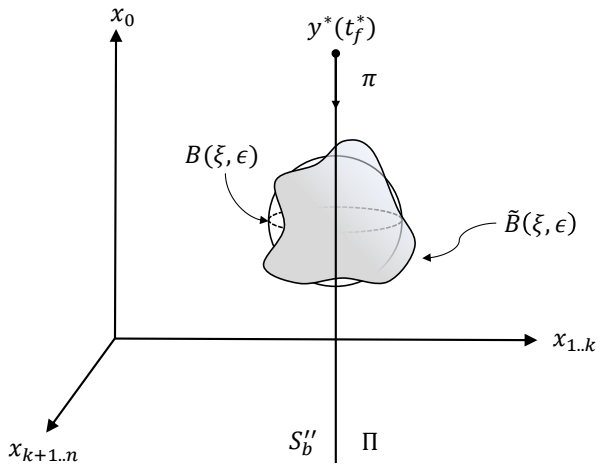

Figure 17: Actual terminal points in a warped ball.

As shown in Figure 17, the set of actual terminal points $\tilde{B}(\xi, \epsilon)$ intersects S_b'' (note that $\xi \in S_b''$ since $r > 0$). If this were true, a contradiction would be reached since those points are feasible and generate a lower cost.

However, we must prove $\tilde{B}(\xi, \epsilon)$ and S_b'' have a non-empty intersection, i.e., that $\tilde{B}(\xi, \epsilon)$ does not have a hole in it. This is done by showing that, for all sufficiently small ϵ, $\tilde{B}(\xi, \epsilon)$ contains the ball $B(\xi, (1-\alpha)\epsilon)$ for any $\alpha \in (0, 1)$.

By definition of $o(\epsilon)$,

$$||o(\epsilon)|| < \alpha\epsilon, \quad \alpha \in (0,1) \tag{50}$$

for all sufficiently small ϵ. We define the function $F : B(\xi,\epsilon) \to \tilde{B}(\xi,\epsilon)$, which is the mapping from an approximated terminal point to the actual terminal point. That is, given a point $c \in B(\xi,\epsilon)$, the point $F(c) \in \tilde{B}(\xi,\epsilon)$ and

$$F(c) = c + o(\epsilon) \tag{51}$$

Additionally, F is continuous since the terminal points depend continuously on the perturbation parameters. Let q be an arbitrary point in $B(\xi,(1-\alpha)\epsilon)$, and consider the map G with domain $B(\xi,\epsilon)$.

$$G(p) = p - F(p) + q \tag{52}$$

For all sufficiently small ϵ, $||G(p) - q|| < \alpha\epsilon$, and hence, $G(p) \in B(\xi,\epsilon)$. Thus, G is a continuous map from $B(\xi,\epsilon)$ to itself. Brouwer's fixed point theorem indicates that G has a fixed point [34, p. 203], i.e., there exists a point $c \in B(\xi,\epsilon)$ such that $G(c) = c$.

Consequently, $F(c) = q$, i.e., q is an actual terminal point. Because q was arbitrary, we conclude that for all sufficiently small ϵ, $B(\xi,(1-\alpha)\epsilon) \subset \tilde{B}(\xi,\epsilon)$. Therefore, $\tilde{B}(\xi,\epsilon)$ and S_b'' have a non-empty intersection. This contradicts optimality and completes the proof of Lemma 1. □

Corollary 1. *There is a hyperplane that separates the sets* Π *and* Δ.

Proof. Both sets are convex and Π does not intersect int(Δ). The convexity of Δ ensures that 1) int(Δ) is convex or 2) int(Δ) is empty [24, p. 39]. In the first case, there is a hyperplane that properly separates Π and Δ [24, p. 53]. In the second case, there is a hyperplane that contains Δ and trivially separates Π and Δ [35, p. 239]. □

The existence of a separating hyperplane implies that there is a non-zero vector, which we define to be $q^*(t_f^*)$, such that

$$\langle q^*(t_f^*), \beta \rangle \leq 0 \quad \forall \beta \text{ s.t. } y^*(t_f^*) + \beta \in \Delta \tag{53}$$

$$\langle q^*(t_f^*), \pi \rangle \geq 0 \tag{54}$$

The hyperplane and normal vector are sketched in Figure 18.

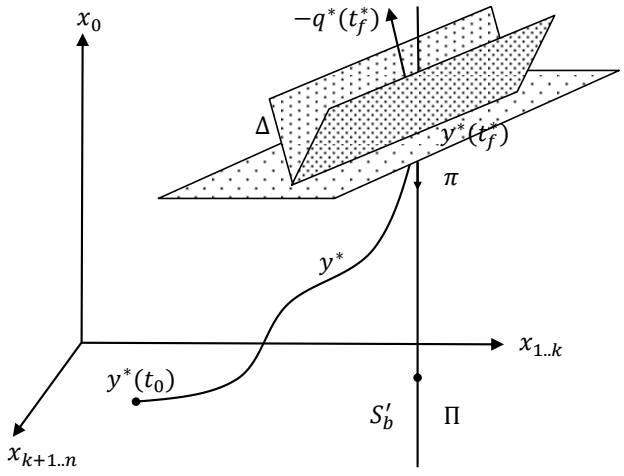

Figure 18: Separating hyperplane and normal vector.

D. Properties of the Adjoint System

The inequalities in Equation 53 and 54 will now be used in conjunction with the adjoint system. Two linear systems of the form $\dot{x}_1(t) = Ax_1(t)$ and $\dot{x}_2(t) = -A^T x_2(t)$ are called adjoint to each other. Adjoint systems satisfy the property that their inner product is constant. This is seen by direct calculation.

$$\begin{aligned} \frac{d}{dt}\langle x_2(t), x_1(t)\rangle &= \langle \dot{x}_2(t), x_1(t)\rangle + \langle x_2(t), \dot{x}_1(t)\rangle \\ &= (-A^T x_2(t))^T x_1(t) + x_2(t)Ax_1(t) = 0 \end{aligned} \tag{55}$$

Recalling the linear system in Equation 33, we define the adjoint system as

$$\dot{q}^*(t) = -\partial_y g(y^*(t), u^*(t))q^*(t) \tag{56}$$

We can now prove the following lemma regarding this system.

Lemma 2. *The function $p_0^*(\cdot)$ is constant and non-positive, the adjoint vector $(p_0^*, p^*(t)) \neq 0\ \forall t \in [t_0, t_f^*]$, and $p^*(\cdot)$ satisfies the differential equation*

$$-\dot{p}^*(t) = \partial_x \mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*(t)) \tag{57}$$

Proof. Expanding Equation 56 into component form gives the following differential equations.

$$\begin{aligned} \dot{p}_0^*(t) &= 0 \\ -\dot{p}^*(t) &= \partial_x \mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*(t)) \end{aligned} \tag{58}$$

Hence $p_0^*(\cdot)$ is constant and Equation 57 is satisfied. The inequality in Equation 54 yields $p_0^* \leq 0$, i.e., p_0^* is non-positive. Finally, since $q^*(t)$ is the solution of a linear homogeneous differential equation and its boundary condition is non-zero, it is non-zero for all time. That is,

$$(p_0^*, p^*(t)) \neq 0 \quad \forall t \in [t_0, t_f^*] \tag{59}$$

This completes the proof. □

Geometrically, $q^*(t)$ represents the normal to the separating hyperplane evolving backward from the final point. The coordinate p_0^* is the vertical component. The backward evolution is illustrated in Figure 19. Note that $-q^*(t)$ is plotted for illustration purposes.

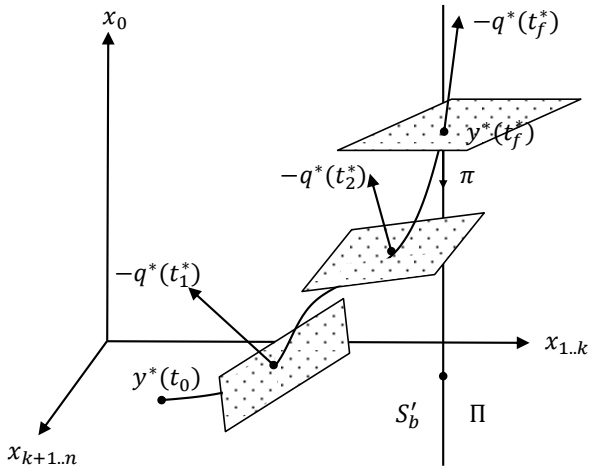

Figure 19: Backward evolution of hyperplanes.

We now invoke Corollary 1 and properties of the adjoint system to prove the famous pointwise maximization condition.

Lemma 3. *The optimal control $u^*(\cdot)$ maximizes the Hamiltonian almost everywhere. That is,*

$$u^*(t) = \arg\max_{\omega \in \Omega} \mathcal{H}(x^*(t), \omega, p^*(t), p_0^*) \tag{60}$$

for almost every $t \in [t_0, t_f^*]$.

Proof. Recall that any element $\xi \in \Delta$ that is generated by the simplest spatial perturbation, a single pulse, has the form

$$\xi = y^*(t_f^*) + \epsilon \Phi(t_f^*, t_p)\gamma_p(\omega) \tag{61}$$

The second term on the right is the linear perturbation, and the separation property in Equation 53 indicates that

$$\langle q^*(t_f^*), \epsilon \Phi(t_f^*, t_p)\gamma_p(\omega)\rangle \geq 0 \tag{62}$$

Since $\Phi(\cdot, t_p)$ is the state transition matrix for the linear system in Equation 33, we can invoke the property of adjoints to get

$$\langle q^*(t_p), \gamma_p(\omega)\rangle \geq 0 \tag{63}$$

Expanding both elements of the inner product into component form and rearranging gives the inequality

$$\mathcal{H}(x^*(t_p), u^*(t_p), p^*(t_p), p_0^*) \geq \mathcal{H}(x^*(t_p), \omega, p^*(t_p), p_0^*) \tag{64}$$

However, both the control element ω and pulse time t_p were arbitrary except for t_p being a regular point of $u^*(\cdot)$. Thus, the optimal control must maximize the Hamiltonian at all regular points. This leads to the famous pointwise maximization condition

$$u^*(t) = \arg\max_{\omega \in \Omega} \mathcal{H}(x^*(t), \omega, p^*(t), p_0^*) \tag{65}$$

which holds for almost all t. □

Remark 3. *Lemma 3 can be strengthened to hold everywhere [25, p. 102]. This strengthened result is not needed here and is not pursued any further.*

E. Properties of the Hamiltonian

We will prove that the Hamiltonian is zero everywhere. This is done by showing that the Hamiltonian at the final point is zero, it is a continuous function of time, and it has zero derivative almost everywhere.

Lemma 4. *The Hamiltonian at the final time is zero.*

Proof. The separation property in Equation 53 applies, in particular, to the case when $\beta = \delta(\tau)$ and $y^*(t_f^*) + \delta(\tau) \in \Delta$. Recalling the linear function $\delta(\tau) = g(y^*(t_f^*), u^*(t_f^*))\tau$ and that τ can be positive or negative, the separation inequality leads to $\langle q^*(t_f^*), \delta(\tau) \rangle = 0$. In terms of the Hamiltonian, we have

$$\mathcal{H}(x^*(t_f^*), u^*(t_f^*), p^*(t_f^*), p_0^*) = 0 \tag{66}$$

This completes the proof. □

Lemma 5. *The Hamiltonian is a continuous function of time.*

Proof. The Hamiltonian $\mathcal{H}(x^*(\cdot), u^*(\cdot), p^*(\cdot), p_0^*)$ is continuous at the regular points of $u^*(\cdot)$ since it is a continuous function of each of its arguments and they are all continuous functions of time. Let t_p be an irregular point of $u^*(\cdot)$. Define

$$\omega_1 = \lim_{t \uparrow t_p} u^*(t) \quad \text{and} \quad \omega_2 = \lim_{t \downarrow t_p} u^*(t) \tag{67}$$

Let $\tau > 0$ such that $(t_p + \tau)$ is a regular point. Application of the pointwise maximization condition in Lemma 3 gives

$$\mathcal{H}(x^*(t_p+\tau), u^*(t_p+\tau), p^*(t_p+\tau), p_0^*) \geq \mathcal{H}(x^*(t_p+\tau), u^*(t_p-\tau), p^*(t_p+\tau), p_0^*) \tag{68}$$

Taking the limit of both sides as τ goes to zero gives

$$\mathcal{H}(x^*(t_p), \omega_2, p^*(t_p), p_0^*) \geq \mathcal{H}(x^*(t_p), \omega_1, p^*(t_p), p_0^*) \tag{69}$$

since limits preserve weak inequalities [36, p. 43] and $x^*(\cdot)$ and $p^*(\cdot)$ are continuous. Switching the time arguments above and repeating the limit process leads to the inequality

$$\mathcal{H}(x^*(t_p), \omega_1, p^*(t_p), p_0^*) \geq \mathcal{H}(x^*(t_p), \omega_2, p^*(t_p), p_0^*) \tag{70}$$

Thus, we conclude that equality holds. Provided $u^*(t_p)$ is defined to equal its left limit, ω_1, or its right limit, ω_2, continuity holds since the left and right limits equal $\mathcal{H}(x^*(t_p), u^*(t_p), p^*(t_p), p_0^*)$. □

Lemma 6. *The Hamiltonian has zero derivative almost everywhere.*

Proof. A consequence of Equation 64 is the following inequality. Let t_1 and t_2 be regular points of $u^*(\cdot)$ in the open interval (t_0, t_f^*). Then

$$\begin{aligned} & \mathcal{H}(x^*(t_2), u^*(t_2), p^*(t_2), p_0^*) - \mathcal{H}(x^*(t_1), u^*(t_2), p^*(t_1), p_0^*) \\ \geq\ & \mathcal{H}(x^*(t_2), u^*(t_2), p^*(t_2), p_0^*) - \mathcal{H}(x^*(t_1), u^*(t_1), p^*(t_1), p_0^*) \\ \geq\ & \mathcal{H}(x^*(t_2), u^*(t_1), p^*(t_2), p_0^*) - \mathcal{H}(x^*(t_1), u^*(t_1), p^*(t_1), p_0^*) \end{aligned} \tag{71}$$

Consider the case when $t_2 > t_1$. Then, using Equation 71,

$$\lim_{t_2 \to t_1} \frac{\mathcal{H}(x^*(t_2), u^*(t_2), p^*(t_2), p_0^*) - \mathcal{H}(x^*(t_1), u^*(t_1), p^*(t_1), p_0^*)}{t_2 - t_1}$$

$$\leq \lim_{t_2 \to t_1} \frac{\mathcal{H}(x^*(t_2), u^*(t_2), p^*(t_2), p_0^*) - \mathcal{H}(x^*(t_1), u^*(t_2), p^*(t_1), p_0^*)}{t_2 - t_1} \tag{72}$$

$$\begin{aligned} = \ & \langle \partial_x \mathcal{H}(x^*(t_1), u^*(t_1), p^*(t_1), p_0^*), \dot{x}^*(t_1) \rangle \\ & + \langle \partial_p \mathcal{H}(x^*(t_1), u^*(t_1), p^*(t_1), p_0^*), \dot{p}^*(t_1) \rangle \quad = \quad 0 \end{aligned}$$

That is, the first limit of Equation 72 must be less than or equal zero. Repeating this limit argument using the bottom two lines of Equation 71 says that the first line must be greater than or equal zero. Thus, the limit from the right is zero.

This process can be repeated with $t_2 < t_1$. Because the first line is the definition of the time derivative of the Hamiltonian, and the limit is zero from both directions, we conclude that the time derivative is zero at all regular

points of $u^*(\cdot)$, i.e.,

$$\frac{d}{dt}\mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*) = 0 \tag{73}$$

for almost every $t \in [t_0, t_f^*]$. □

The above properties of the Hamiltonian lead to the following corollary.

Corollary 2. *The Hamiltonian is zero everywhere.*

Proof. This follows directly from the fact that the Hamiltonian at the final point is zero, the time derivative is zero almost everywhere, and it is a continuous function of time. Thus,

$$\mathcal{H}(x^*(t), u^*(t), p^*(t), p_0^*) = 0 \tag{74}$$

for all $t \in [t_0, t_f^*]$. □

Remark 4. *Note that Equation 17 and the first two equations of Equation 14 are just restatements of the constraints in BOCP. Thus, Lemma 2, Lemma 3, and Corollary 2 complete the proof Theorem 1.*

F. The Transversality Condition

We now relax BOCP so that the final point must belong not to a point set, but to a smooth $(n-p)$-fold in $\mathbb{R}^n$ [37, p. 92-95]. The manifold is characterized

by the function $b(\cdot) \in \mathcal{C}^1 : \mathbb{R}^n \to \mathbb{R}^p$. Thus, the new problem, BOCP', is

$$
\begin{aligned}
\min \quad & J = x_0(t_f) && \text{(BOCP')} \\
\text{subj. to} \quad & \dot{x}_0(t) = \ell(x(t), u(t)), \quad x_0(t_0) = 0 \\
& \dot{x}(t) \; = f(x(t), u(t)), \quad x(t_0) = a, \quad b(x(t_f)) = 0 \\
& u(t) \in \Omega, \quad t_0 \text{ fixed}, \quad t_f \text{ free}
\end{aligned}
$$

Note that the only change from BOCP is this final boundary condition on $x(\cdot)$. Only one element in the preceding arguments is affected by this change. It is the proof of Lemma 1. We will complete the proof for BOCP' and obtain one more separation inequality in addition to those in Equations 53 and 54.

It is convenient to define the following sets.

$$
\begin{aligned}
S_b &= \{x : b(x) = 0\} \\
S_b' &= \{(x_0, x) : x_0 \in \mathbb{R} \text{ and } x \in S_b\} \qquad (75) \\
S_b'' &= \{(x_0, x) : x_0 < J^* \text{ and } (x_0, x) \in S_b'\}
\end{aligned}
$$

The set S_b is the terminal set, and S_b' simply lifts S_b in the x_0 direction. The set S_b'' is a subset of S_b'. It consists only of those points in S_b' that have lower cost than the optimal. The set Π is redefined to be

$$
\Pi = \{\xi : \xi = y^*(t_f^*) + r\pi + \begin{bmatrix} 0 \\ s \end{bmatrix}, r \geq 0, s \in \mathcal{T}(S_b \,|\, x^*(t_f^*))\} \qquad (76)
$$

where $\mathcal{T}(S_b \,|\, x^*(t_f^*))$ is the tangent space of S_b at the point $x^*(t_f^*)$. Some aspects of these sets are captured in Figure 20.

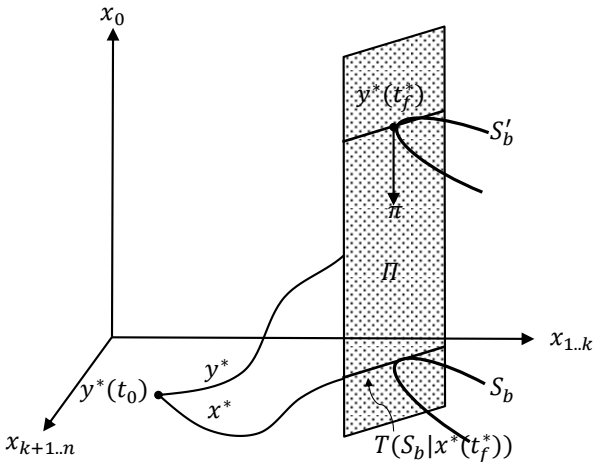

Figure 20: Illustration of sets.

With this new definition of Π, we prove Lemma 1.

Proof (Lemma 1 for BOCP'). Suppose that the lemma is false and that Π does intersect the interior of Δ. Then there exists a point ξ such that $\xi \in \Pi$ and $\xi \in \text{int}(\Delta)$. This implies that ξ must be of the form

$$\xi = y^*(t_f^*) + \epsilon r\pi + \begin{bmatrix} 0 \\ s \end{bmatrix} \tag{77}$$

for some $r > 0$ and some $s \in \mathcal{T}(S_b \,|\, x^*(t_f^*))$. And as before, there is the ball $B(\xi, \epsilon) \subset \text{int}(\Delta)$ and the associated warped ball $\tilde{B}(\xi, \epsilon)$. To reach a contradiction, we must show that $\tilde{B}(\xi, \epsilon)$ intersects S_b''.

The same fixed point arguments hold such that $B(\xi, (1-\alpha)\epsilon) \subset \tilde{B}(\xi, \epsilon)$. Because Π and S_b'' are tangent to each other along the direction of π, the distance from ξ to S_b'' is of order $o(\epsilon)$. Thus, for all sufficiently small ϵ, $B(\xi, (1-\alpha)\epsilon)$ intersects S_b''. This completes the proof since $B(\xi, (1-\alpha)\epsilon) \subset \tilde{B}(\xi, \epsilon)$. □

Since Lemma 1 remains true, Corollary 1 does as well. Recognizing the fact that $0 \in \mathcal{T}(S_b \,|\, x^*(t_f^*))$, we recover the inequalities in Equations 53 and 54. By setting $r = 0$, we obtain a third inequality

$$\langle q^*(t_f^*), \begin{bmatrix} 0 \\ s \end{bmatrix} \rangle \geq 0 \quad \forall s \in \mathcal{T}(S_b \,|\, x^*(t_f^*)) \tag{78}$$

For every $s \in \mathcal{T}(S_b \,|\, x^*(t_f^*))$ we also have $-s \in \mathcal{T}(S_b \,|\, x^*(t_f^*))$. Thus, the inequality must be satisfied with strict equality. Expanding into component form gives the transversality condition

$$\langle p^*(t_f^*), s \rangle = 0 \quad \forall s \in \mathcal{T}(S_b \,|\, x^*(t_f^*)) \tag{79}$$

The only change that must be made to Theorem 1 to adapt it to BOCP' is statement *vi*. It now reads

vi) the boundary conditions

$$x_0^*(t_0) = 0,\ x^*(t_0) = a,\ b(x^*(t_f^*)) = 0,\ \langle p^*(t_f^*), s \rangle = 0\ \forall s \in \mathcal{T}(S_b \,|\, x^*(t_f^*)) \tag{80}$$

Remark 5. *The transversality condition says that $p^*(t_f^*)$ is orthogonal to the null space of $\partial_x^T b(x^*(t_f^*))$. Hence, there is a non-zero vector ν such that $p^*(t_f^*) = \partial_x b(x^*(t_f^*))\nu$.*

A number of generalizations can be reached from this point. For example, problems with time dependence, terminal and integral costs, and final time constraints can easily be transformed to the above form so that necessary conditions can be stated [33, p. 130-134].

G. Optimal Control with State Constraints

Attention is now turned to optimal control problems with state constraints. In these problems, the state cannot evolve arbitrarily in $\mathbb{R}^n$. It must evolve in some subset. Necessary conditions for such problems are not simple generalizations of the previous results, and their history and proof are beyond the scope of this chapter. The exposition here is an adaptation of a recent survey paper [38].

Consider the following optimal control problem with explicit time dependence and state constraints (OCPSC).

$$\begin{aligned} \min \quad & J = m(t_f, x(t_f)) + \textstyle\int_{t_0}^{t_f} \ell(t, x(t), u(t))\, dt \\ \text{subj. to} \quad & \dot{x}(t) = f(t, x(t), u(t)), \quad x(t_0) = a, \quad t_0 \text{ fixed} \\ & g(t, x(t), u(t)) \leq 0, \qquad h(t, x(t)) \leq 0 \\ & b(t_f, x(t_f)) = 0, \qquad c(t_f, x(t_f)) \leq 0 \end{aligned} \tag{OCPSC}$$

It is assumed that each of the above functions is continuously differentiable with respect to all of its arguments – even the control. This assumption is quite strong and can be relaxed, but it is sufficient for our purposes. Additionally, it is assumed that the gradient of all active constraints are linearly independent. See Hartl et al. for more details [38].

The control constraints are now given in function form. The control set is

$$\Omega(t, x(t)) = \{\omega : g(t, x(t), \omega) \leq 0\} \tag{81}$$

The intervals where the state constraints $h(t, x(t))$ are satisfied with strict equality are called boundary arcs. The intervals where they are satisfied with strict inequality are called interior arcs. The time instances joining boundary and interior arcs are called junction times and denoted by τ. Lastly, inequality constraints on the final point are also present with $c(t_f, x(t_f)) \leq 0$.

Before stating the necessary conditions, it is convenient to define three functions.

$$\mathcal{H}(t, x(t), u(t), p(t), p_0) = \langle p_0 \ell(t, x(t), u(t)) \rangle + \langle p(t), f(t, x(t), u(t)) \rangle \tag{82}$$

$$\begin{aligned} \mathcal{L}(t, x(t), u(t), p(t), p_0, \lambda(t), \nu(t)) = \mathcal{H}(t, x(t), u(t), p(t), p_0) \\ + \langle \lambda(t), g(t, x(t), u(t)) \rangle + \langle \nu(t), h(t, x(t)) \rangle \end{aligned} \tag{83}$$

$$\begin{aligned} \mathcal{G}(t_f, x(t_f), p_0, \xi, \mu) = \langle p_0, m(t_f, x(t_f)) \rangle + \langle \xi, b(t_f, x(t_f)) \rangle \\ + \langle \zeta, c(t_f, x(t_f)) \rangle + \langle \mu, h(t_f, x(t_f)) \rangle \end{aligned} \tag{84}$$

Special cases of the result below have been studied by Jacobson, Lele, and Speyer [39] and Maurer [40]. A more modern exposition can be found in the survey paper [38]. The necessary conditions for OCPSC are below.

Theorem 2 (Necessary Conditions for OCPSC)**.** *If OCPSC attains a minimum at $u^*(\cdot)$ such that $x^*(\cdot)$ has a finite number of junction times, then the following system is solvable:*

i) the normality condition

$$p_0^* \leq 0 \tag{85}$$

ii) the non-triviality condition

$$(p_0^*, p^*(t)) \neq 0 \quad \forall t \in [t_0, t_f^*] \tag{86}$$

iii) the differential equations

$$\begin{aligned} \dot{x}^*(t) &= \partial_p \mathcal{L}(t, x^*(t), u^*(t), p^*(t), p_0^*, \lambda^*(t), \nu^*(t)) \\ -\dot{p}^*(t) &= \partial_x \mathcal{L}(t, x^*(t), u^*(t), p^*(t), p_0^*, \lambda^*(t), \nu^*(t)) \\ \dot{\mathcal{H}}(t, x^*(t), u^*(t), p^*(t), p_0^*) &= \partial_t \mathcal{L}(t, x^*(t), u^*(t), p^*(t), p_0^*, \lambda^*(t), \nu^*(t)) \end{aligned} \tag{87}$$

iv) the pointwise maximum condition

$$u^*(t) = \arg \max_{\omega \in \Omega(t, x^*(t))} \mathcal{H}(t, x^*(t), \omega, p^*(t), p_0^*) \quad \text{a.e. } t \in [t_0, t_f^*] \tag{88}$$

v) the stationary condition

$$\partial_u \mathcal{L}(t, x^*(t), u^*(t), p^*(t), p_0^*, \lambda^*(t), \nu^*(t)) = 0 \quad \text{a.e. } t \in [t_0, t_f^*] \tag{89}$$

vi) the complementary slackness conditions

$$\begin{aligned}
&g(t,x^*(t),u^*(t)) \leq 0, \quad \lambda^*(t) \leq 0 \qquad \langle \lambda^*(t), g(t,x^*(t),u^*(t)) \rangle = 0 \\
&h(t,x^*(t)) \leq 0, \qquad\qquad \nu^*(t) \leq 0, \qquad \langle \nu^*(t), h(t,x^*(t)) \rangle = 0 \\
&c(t_f^*,x^*(t_f^*)) \leq 0, \qquad \zeta^* \leq 0, \qquad\quad \langle \zeta^*, c(t_f^*,x^*(t_f^*)) \rangle = 0 \\
&h(\tau^*,x^*(\tau)) \leq 0, \qquad \eta^*(\tau^*) \leq 0, \quad \langle \eta^*(\tau^*), h(\tau^*,x^*(\tau)) \rangle = 0 \\
&h(t_f^*,x^*(t_f^*)) \leq 0, \qquad \mu^* \leq 0, \qquad\quad \langle \mu, h(t_f^*,x^*(t_f^*)) \rangle = 0
\end{aligned} \tag{90}$$

vii) the jump conditions

$$\begin{aligned}
&p(\tau_-^*) = p(\tau_+^*) + \partial_x h(\tau^*,x^*(\tau^*))\eta(\tau^*) \quad \forall \tau^* \\
&\mathcal{H}(\tau_-^*,x(\tau^*),u(\tau_-^*),p(\tau_-^*),p_0) = \mathcal{H}(\tau_+^*,x(\tau^*),u(\tau_+^*),p(\tau_+^*),p_0) \\
&\qquad\qquad - \partial_t h(\tau^*,x^*(\tau^*))\eta(\tau^*) \quad \forall \tau^*
\end{aligned} \tag{91}$$

viii) the boundary conditions

$$\begin{aligned}
&x(t_0) = a, \quad b(t_f^*,x^*(t_f^*)) = 0, \quad c(t_f^*,x^*(t_f^*)) \leq 0 \\
&p^*(t_f^*) = \partial_x \mathcal{G}(t_f^*,x^*(t_f^*),p_0^*,\xi^*,\mu^*) \\
-&\mathcal{H}(t_f^*,x^*(t_f^*),u^*(t_f^*),p^*(t_f^*),p_0^*) = \partial_t \mathcal{G}(t_f^*,x^*(t_f^*),p_0^*,\xi^*,\mu^*)
\end{aligned} \tag{92}$$

The statement of these conditions concludes this chapter on optimal control. We now have the tools to study lossless convexification.

CHAPTER IV:
RENDEZVOUS USING DIFFERENTIAL DRAG

This chapter presents a method to solve the minimum time rendezvous problem of multiple spacecraft using differential drag. The method is based on lossless convexification and reduces the problem to solving a fixed number of linear programming (LP) problems. The problem is to rendezvous any number of chaser spacecraft with a single target spacecraft in low earth orbit using only the relative aerodynamic drag between the chaser and target vehicles. Attached to each spacecraft are drag plates that can be deployed or not. It is the actuation of these plates that generates the relative aerodynamic drag. The objective is to minimize the total time to achieve rendezvous.

There are a number of reasons to use differential drag as a control in formation flight [41]. First, fuel savings are possible since thrust is not used or used only for corrections. Second, plumes and jet firings may be harmful to other spacecraft in the vicinity. Third, and generally speaking, the use of thrust requires more complicated mechanical systems and so the system hardware can be simplified. Interest in rendezvous using differential drag is evident in the work of the Canadian and Japanese Space Agencies and Israel Aerospace Industries [42,43]. Convex optimization for rendezvous and proximity operations has also been introduced recently by Lu and Liu [44].

Leonard et al. [41] studied the problem of station keeping using differential drag and used the Clohessy-Wiltshire-Hill equations [45] to describe the relative motion. They introduced a transformation that separated the equations of motion into that of a double integrator and a harmonic oscillator coupled only through the control. It is important to note that optimal strategies for the decoupled systems are well understood [37, Ch. 7], but a general analysis for the coupled system remains an open problem. In the end, Leonard et al. developed an analytical, sub-optimal control law.

Bevilacqua and Romano [46] studied the problem of rendezvous using differential drag and used the Schweighart-Sedwick equations [47] to describe the relative motion. These equations are of the same form as the Clohessy-Wiltshire-Hill equations, but they include an averaged J2 effect more suitable for maneuvers that occur over several periods. They developed an analytical, sub-optimal control law for two vehicles. The approach was then extended to the rendezvous of multiple vehicles by sequentially applying the two-vehicle control law. Bevilacqua et al. [48] also solved the problem by using a combination of differential drag and thrust for the final approach.

In this chapter, it is assumed that the relative motion of each chaser with respect to the target is accurately described by the Schweighart-Sedwick equations. The equations are then simplified through two transformations. At this point, the optimization problem is most naturally solved as a mixed integer nonlinear programming (MINLP) problem because the feasible control set is $\{-1, 0\}$ and the final time is free. Such problems are generally difficult to solve, and for this reason, the control set is relaxed from $\{-1, 0\}$ to $[-1, 0]$,

i.e., from a non-convex point set to a convex interval. Feasible controls for the relaxed problem are not necessarily feasible for the original problem. The proof of lossless convexification shows that a minimum time control of the relaxed problem exists that is also a minimum time control of the original problem.

This relaxation and proof are different than others in lossless convexification. Typical proofs hinge on the non-existence of singular controls. Singular controls do exist in the current problem. However, 1) it is shown that singular controls that are bang-bang also exist and 2) a constructive procedure for converting non-bang-bang singular controls to bang-bang controls is given. These points make the work a unique contribution.

Because the final time is free, lossless convexification of the control set only reduces the problem to a nonlinear programming (NLP) problem. However, for a given final time, the problem of finding the optimal control reduces to a LP problem. When a feasible solution exists, LPs can be solved with guaranteed convergence to the global minimum in polynomial time. When feasible solutions do not exist, the program returns infeasible in polynomial time. Thus, the problem reduces to a one-dimensional search for the least final time so that the program returns a feasible answer.

Upon completion, some controls will be singular, and so the procedure mentioned above is implemented to make all singular controls be bang-bang controls. The same line search could have been used to convert the MINLP problem to a sequence of integer linear programming (ILP) problems, thus avoiding the need for convexification. However, ILPs do not share the polynomial time convergence rate and are generally NP-hard [49].

The LP method developed here is programmed using a custom solver. Mattingley and Boyd [3] have shown that LP problems can be solved two to three orders of magnitude faster using custom code than with standard algorithms. Such significant speed ups are not evident here, but the customized solver does out perform all others tested.

A. Problem Description

The Schweighart-Sedwick equations [47] describe the linearized relative motion of two generic spacecraft in a local vertical local horizontal frame. In deriving the equations, it is assumed that the reference orbit is circular, the spacecraft are close to each other compared to their orbital radius, and the only forces acting on the spacecraft are gravity and aerodynamic drag. The gravity force is composed of a two-body effect and an averaged effect of the J2 perturbation over one orbit. The aerodynamic drag appears as part of the control term since it is the primary mechanism for maneuvering the spacecraft [48]. The equations are

$$\begin{bmatrix} \dot{z}_1(t) \\ \dot{z}_2(t) \\ \dot{z}_3(t) \\ \dot{z}_4(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ b & 0 & 0 & a \\ 0 & 0 & 0 & 1 \\ 0 & -a & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ z_4(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ a_D \end{bmatrix} [u(t) - u_0(t)] \tag{1}$$

The origin of the frame is located at the target spacecraft. The coordinates z_1 and z_2 are the relative position and velocity in the vertical or radial direction. The coordinates z_3 and z_4 are the relative position and velocity in the hori-

zontal or tangential direction. The positive constants a and b are associated with a particular reference path about which the motion was linearized.

The constant a_D is the magnitude of acceleration associated with opening the drag panels.

$$a_D = \tfrac{1}{2}\beta\rho_r V_r^2 \tag{2}$$

The air density associated with the reference orbit is ρ_r. The speed associated with a spacecraft in the reference orbit is V_r, and β is the differential ballistic coefficient between the spacecraft when one vehicle has the drag plates deployed and the other does not. In the calculation of drag, it is assumed that the air density is constant throughout the maneuver and that the speeds of the spacecraft are reasonably approximated as their reference orbital speed. The assumptions are reasonable since motion about the reference is expected to be small. The control of the chaser spacecraft is u, and the control of the target spacecraft is u_0. At a given time, each control can take values in the set $\{-1, 0\}$ corresponding to the cases where the drag plates are deployed or not. It follows that the effective control, $u_e(t) = u(t) - u_0(t)$, can only take one of three values at a given time.

1. If the drag plates on the chaser are deployed and the drag plates on the target are not, then $u_e(t) = -1$.

2. If the drag plates on the target are deployed and the drag plates on the chaser are not, then $u_e(t) = +1$.

3. If the drag plates on both spacecraft are in the same position, then $u_e(t) = 0$.

The system matrix in Equation 1 is defective so that it cannot be diagonalized. Nonetheless, the system can be reduced to a double integrator and harmonic oscillator by a similarity transform. The columns of the transformation matrix P consist of the generalized eigenvectors of the system matrix. After defining $\omega^2 = a^2 - b$,

$$P = \begin{bmatrix} 0 & -a & 0 & -1/a \\ 0 & 0 & \omega^2/a & 0 \\ b & 0 & 1 & 0 \\ 0 & b & 0 & 1 \end{bmatrix} \tag{3}$$

The new states are given by $y(t) = P^{-1}z(t)$. Differentiating and substituting for $\dot{z}(t)$ gives

$$\begin{bmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \\ \dot{y}_3(t) \\ \dot{y}_4(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\omega^2 & 0 \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{bmatrix} + \begin{bmatrix} 0 \\ -a_D/\omega^2 \\ 0 \\ a^2 a_D/\omega^2 \end{bmatrix} [u(t) - u_0(t)] \tag{4}$$

It is now obvious that the first two states form a double integrator, the last two states form a harmonic oscillator, and the two are coupled by the control variables. For convenience, one more transformation is introduced.

$$Q = \frac{\omega^2}{a_D} \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \omega/a^2 & 0 \\ 0 & 0 & 0 & 1/a^2 \end{bmatrix} \tag{5}$$

The final states are given by $x(t) = Qy(t)$. Differentiating and substituting for $\dot{y}(t)$ gives

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \dot{x}_4(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \omega \\ 0 & 0 & -\omega & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} [u(t) - u_0(t)] \tag{6}$$

This system of differential equations describes the relative motion of one chaser with respect to the target spacecraft. Henceforth, the system matrix in Equation 6 is labeled as A and the control influence matrix is labeled as B.

In these transformed coordinates, the motion of each state can be simply understood by studying the phase plane. Figure 21 shows the x_1-x_2 phase plane. The solid curves represent paths of constant effective control $u_e(t) = -1$. The dashed curves represent paths of constant effective control $u_e(t) = +1$. Paths for zero effective control are not shown, but are horizontal lines.

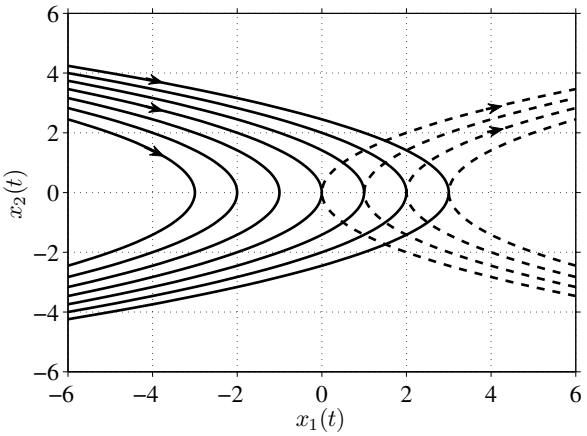

Figure 21: x_1-x_2 phase plane.

Figure 22 shows the ωx_3-ωx_4 phase plane. The solid circles centered at $(-1, 0)$ represent paths of constant effective control $u_e(t) = -1$. The dashed circles centered at $(+1, 0)$ represent paths of constant effective control $u_e(t) = +1$. Paths for zero effective control are not shown, but are circles centered at the origin. The arrows indicate that all motion occurs in a clockwise direction.

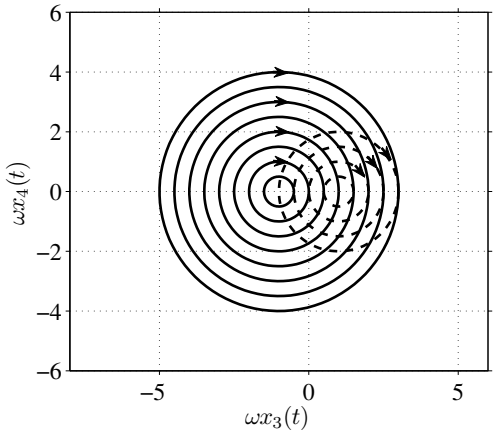

Figure 22: ωx_3-ωx_4 phase plane.

Now considering the case where there are M chasers, the motion of each chaser relative to the target is again given by Equation 6, which is written below for i^{th} chaser.

$$\dot{x}_i(t) = Ax_i(t) + Bu_i(t) - Bu_0(t), \quad \forall i = 1, \ldots M \tag{7}$$

The subscript i indicates that the equation describes the motion of the i^{th} chaser relative to the target. It is evident that each chaser is coupled to the others by the target control u_0. Thus the motion of the entire system, including

all $M+1$ vehicles, must be considered as a whole. Upon defining the matrices

$$\mathcal{A} = \begin{bmatrix} A & & & \\ & A & & \\ & & \ddots & \\ & & & A \end{bmatrix} \quad \text{and} \quad \mathcal{B} = \begin{bmatrix} -B & B & & & \\ -B & 0 & B & & \\ \vdots & & & \ddots & \\ -B & 0 & 0 & & B \end{bmatrix} \tag{8}$$

the motion of the entire system can be described by using the new, augmented state and control vectors $x := [x_1^T, \ldots, x_M^T]^T$ and $u := [u_0, u_1, \ldots, u_M]^T$, which satisfy the dynamics $\dot{x}(t) = \mathcal{A}x(t) + \mathcal{B}u(t)$. This system will be of primary concern in the upcoming theoretical developments. In developing results for the special case of only one chaser, it will be made clear that the results hold only when $M = 1$.

Attention is turned to the optimal control problem. Considering the system equations in the previous paragraph, a formal statement of the minimum time rendezvous problem is below.

$$\begin{aligned} \min \quad & J = \textstyle\int_{t_0}^{t_f} 1 \, dt \\ \text{subj. to} \quad & \dot{x}(t) \;=\; \mathcal{A}x(t) + \mathcal{B}u(t) \\ & x(t_0) = x_0, \;\; x(t_f) = 0 \\ & u_i(t) \in \{-1, 0\} \quad \forall i = 0, \ldots, M \end{aligned} \tag{P0}$$

A full analysis for the minimum time transfer to the origin is currently unavailable, but a number of important facts are still known [37, Ch. 7]. The discussion begins by setting $M = 1$ and considering the uncoupled double inte-

grator and harmonic oscillator – for which a complete analysis is possible. For the double integrator system, the minimum time control consists of at most one switch. Figure 23 shows the switch curve in the x_1-x_2 plane. If the state initially lies below the switch curve, the optimal strategy is to apply $u_e(t) = +1$ until the state intersects the switch curve, and then switch to $u_e(t) = -1$ until the state reaches the origin. If the state initially lies above the switch curve, the optimal strategy is to apply $u_e(t) = -1$ and then $u_e(t) = +1$.

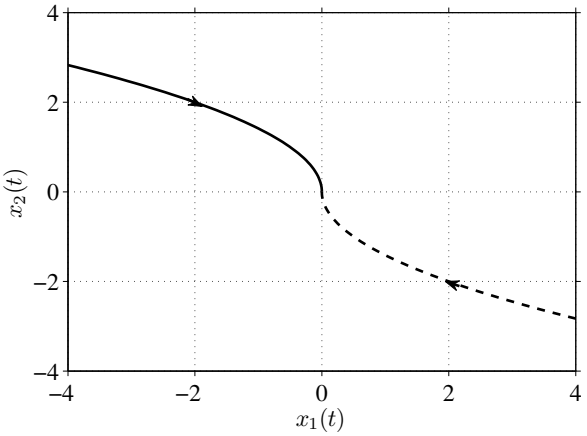

Figure 23: Double integrator switch curve.

Analyzing the harmonic oscillator is more difficult since the number of switches cannot be bounded a priori. Figure 24 shows the switch curve in the ωx_3-ωx_4 plane. If the state lies below the switch curve, the optimal strategy is to apply $u_e(t) = +1$ until the state intersects the switch curve, and then switch to $u_e(t) = -1$. If the state lies above the switch curve, the optimal strategy is to apply $u_e(t) = -1$ and then $u_e(t) = +1$. This process repeats until the state reaches the origin.

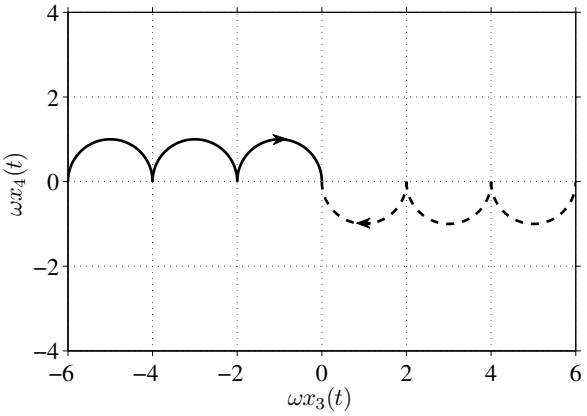

Figure 24: Harmonic oscillator switch curve.

Solving the double integrator and harmonic oscillator systems simultaneously can be done for some sets of initial conditions, but a general analysis is unavailable. It follows that a general analysis for $M > 1$ is also unavailable. This motivates the lossless convexification and solution method introduced in the next two sections.

B. Lossless Convexification

Note that the control set in problem P0 is not convex, which introduces numerical challenges. At the risk of introducing infeasible solutions, the control

set is relaxed from $\{-1, 0\}$ to $[-1, 0]$, and a new problem is formulated.

$$\begin{aligned} \min \quad & J = \textstyle\int_{t_0}^{t_f} 1 \, dt \\ \text{subj. to} \quad & \dot{x}(t) \;=\; \mathcal{A}x(t) + \mathcal{B}u(t) \\ & x(t_0) = x_0, \;\; x(t_f) = 0 \\ & u_i(t) \in [-1, 0] \quad \forall i = 0, \ldots, M \end{aligned} \tag{P1}$$

With Theorem 1 of Chapter III in mind, the Hamiltonian and differential equations for the relaxed problem P1 are

$$\begin{aligned} \mathcal{H}(x(t), u(t), p(t), p_0) &= p_0 + \langle p(t), \mathcal{A}x(t) \rangle + \langle p(t), \mathcal{B}u(t) \rangle \\ \dot{x}(t) &= \mathcal{A}x(t) + \mathcal{B}u(t) \\ -\dot{p}(t) &= \mathcal{A}^T p(t) \end{aligned} \tag{9}$$

Upon defining the switching functions $\varphi_i(t) = \langle p(t), \mathcal{B}_i \rangle$, the pointwise maximum condition indicates that the optimal controls are

$$u_i(t) = \begin{cases} -1, & \varphi_i(t) < 0 \\ 0, & \varphi_i(t) > 0 \\ \text{singular}, & \varphi_i(t) = 0 \end{cases} \tag{10}$$

The aim of the remainder of this section is to prove that an optimal solution of the relaxed problem P1 exists that is also an optimal solution of the original problem P0.

The following definitions regarding the optimal control fix some nomenclature used throughout the rest of the chapter.

Definition 1. *A control $u_i(t)$ is a bang-bang control if it only takes values of -1 and 0.*

Definition 2. *A control $u_i(t)$ is a non-singular control if the switching function $\varphi_i(t)$ is non-zero for almost every t.*

Remark 1. *It follows from these definitions and Eq. 10 that a non-singular control is a bang-bang control.*

Definition 3. *A control $u_i(t)$ is a singular control if the switching function $\varphi_i(t)$ is zero for almost every t.*

Remark 2. *For relaxed problem P1, if a control is non-singular (singular) on any interval, then it is non-singular (singular) on $[t_0, t_f]$. This follows from the fact that the switching functions are independent of the state and control trajectories.*

It is obvious that non-singular solutions for the relaxed problem P1 are feasible solutions for the original problem P0. However, it is shown that when $M > 1$, singular solutions must exist for the relaxed problem P1. Thus, solutions for the relaxed problem P1 are not necessarily feasible for the original problem P0. Below, several lemmas and proofs are given to help synthesize the optimal control problem, and the bang-bang principle due to LaSalle [50] is invoked to prove that bang-bang singular controls exist.

In light of this fact, the section culminates with Theorem 1 stating that an optimal solution of the relaxed problem P1 exists that is an optimal solution

of the original problem P0. The next two lemmas assert the existence of a minimum time control and the existence of a non-trivial adjoint.

Lemma 1. *If a feasible control exists, then a minimum time control exists.*

Proof. See Pontryagin [25, p. 127]. □

Lemma 2. *The adjoint $p(\cdot)$ is never zero.*

Proof. If $p(\cdot)$ is zero somewhere, it is zero everywhere since it is the solution of a homogeneous differential equation. Because the Hamiltonian is zero everywhere, p_0 must be zero. This violates the non-triviality condition. Thus, $p(\cdot)$ is never zero. □

The next four lemmas assess several important controllability criterion of the linear system. It is first shown in Lemma 3 that the (A, B) and $(\mathcal{A}, \mathcal{B})$ pairs are controllable. It is then shown in Lemma 4 that upon removing any control from the system, the linear system is still controllable. Finally, in Lemma 5, it is shown that the linear system is normal when $M = 1$ and not normal when $M > 1$, which is used in Lemma 6 to prove existence of singular and non-singular controls.

Lemma 3. *The (A, B) and $(\mathcal{A}, \mathcal{B})$ pairs are controllable.*

Proof. The determinant of the controllability matrix is

$$\det(C) = \det([B\ AB\ A^2B\ A^3B]) = \omega^5 \neq 0 \tag{11}$$

Thus, the controllability matrix is full row rank, $\text{rank}(C) = 4$, and the (A, B) pair is controllable. Define the matrix $\tilde{C} = [B\ AB\ \ldots A^{4M-1}B]$, for which

$\text{rank}(\tilde{C}) = 4$ since the (A, B) pair is controllable. The controllability matrix for the $(\mathcal{A}, \mathcal{B})$ pair is then

$$\mathcal{C} = \begin{bmatrix} -\tilde{C} & \tilde{C} & & \\ -\tilde{C} & 0 & \tilde{C} & \\ \vdots & & & \ddots \\ -\tilde{C} & 0 & 0 & \tilde{C} \end{bmatrix} \tag{12}$$

It is obvious that the M blocks of rows are linearly independent. Thus, the controllability matrix is full row rank, $\text{rank}(\mathcal{C}) = 4M$, and the $(\mathcal{A}, \mathcal{B})$ pair is controllable. □

Definition 4. *The control influence matrix with the j^{th} column removed is denoted $\mathcal{B}_{\sim j}$. The corresponding controllability matrix is denoted $\mathcal{C}_{\sim j}$.*

Lemma 4. *The $(\mathcal{A}, \mathcal{B}_{\sim j})$ pair is controllable for any j.*

Proof. Denote the j^{th} block of columns as $\mathcal{C}_j$. Any block of columns can be written as a linear combination of the others, i.e.,

$$\mathcal{C}_j = -\sum_{\substack{i=1 \\ i \neq j}}^{M+1} \mathcal{C}_i \tag{13}$$

Thus, the controllability matrix with the j^{th} block of columns removed, $\mathcal{C}_{\sim j}$, is full rank,

$$\text{rank}(\mathcal{C}_{\sim j}) = \text{rank}(\mathcal{C}) = 4M \tag{14}$$

It follows that upon removing any control variable, the $(\mathcal{A}, \mathcal{B}_{\sim j})$ pair remains controllable. □

Definition 5. *The linear system is normal if the* $(\mathcal{A}, \mathcal{B}_j)$ *pair is controllable for all* j*. That is,* $\mathcal{A}$ *is controllable with every column of* $\mathcal{B}$*.*

Lemma 5. *The linear system is normal when* $M = 1$*. The linear system is not normal when* $M > 1$*.*

Proof. It is easy to see that $\text{rank}(\mathcal{C}_j) = 4$. Thus, only when $M = 1$ is $\text{rank}(\mathcal{C}_j) = 4M \ \forall j$. It follows that the linear system is normal when $M = 1$, and the linear system is not normal when $M > 1$. □

Lemma 6. *i) Singular controls cannot exist when* $M = 1$ *and at least one singular control must exist when* $M > 1$*. ii) At least two controls are bang-bang.*

Proof. i) A necessary and sufficient condition for the non-existence of singular controls is that the linear system be normal [37, p. 399]. In light of Lemma 5, singular controls cannot exist when $M = 1$, and at least one singular control must exist when $M > 1$.

ii) Suppose that all controls are singular. Then $\varphi_j(t) = 0$ a.e. $t \in [t_0, t_f]$ for all j. Differentiating each switching function $4M - 1$ times implies that $\mathcal{C}_j^T p(t) = 0$ a.e. $t \in [t_0, t_f]$ for all j. However, Lemmas 2 and 3 imply $\mathcal{C}^T p(t) \neq 0 \ \forall t \in [t_0, t_f]$. Thus, in light of Remark 2, there exists at least one j such that $\varphi_j(t) \neq 0$ a.e. $t \in [t_0, t_f]$, i.e., at least one control is bang-bang. Upon removing the j^{th} column, the matrix $\mathcal{C}_{\sim j}$ remains full rank (see Lemma 4). Applying the same logic as before, at least one more control is bang-bang. It follows that at least two controls are bang-bang. □

Existing proofs of lossless convexification hinge on proving the non-existence of singular controls. The proof here is complicated by the fact that singular controls do exist when $M > 1$, and hence, not all extremal controls for relaxed problem P1 are feasible controls for the original problem P0. The bang-bang principle due to LaSalle [50] helps resolve this issue.

Lemma 7. *If a minimum time control exists, then a minimum time control that is bang-bang exists.*

Proof. See LaSalle [50]. □

A consequence of Lemma 7 is that a minimum time control for relaxed problem P1 exists that is also a feasible solution for original problem P0. This leads to the following theorem.

Theorem 1. *A minimum time control of relaxed problem P1 exists that is also a minimum time control for the original problem P0.*

Proof. Note that the cost functions for P0 and P1 are the same. Since $\mathcal{F}_1 \subset \mathcal{F}_2$, $J_2^* \leq J_1^*$. Lemma 7 implies that $\mathcal{K} = \mathcal{F}_1 \cap \mathcal{F}_2^*$ is non-empty. Since the cost of every trajectory in $\mathcal{K}$ is J_2^* and $\mathcal{K} \subset \mathcal{F}_1$, it must be that $J_1^* \leq J_2^*$. The two inequalities imply $J_1^* = J_2^*$. Thus, every trajectory in $\mathcal{K}$ is in $\mathcal{F}_1^*$, i.e., $\mathcal{K} \subset \mathcal{F}_1^*$. That is, there is an optimal control for P1 that is optimal for P0. □

Thus, it has been proved that the relaxation *can be* lossless but is not necessarily so. The crux of the proof is Lemma 7, which is only a statement existence – not uniqueness.

C. Solution Method

This section describes the solution method to solve the problem P1 to ensure that it is a solution of P0. It is the combination of the results above and this solution method that make the relaxation a lossless convexification. Essentially, the method answers the question of how to replace non-bang-bang controls with bang-bang controls. For this reason, it is not required if $M = 1$ since singular controls do not exist according to Lemma 6.

The first step is to discretize the problem so that it can be solved numerically. For any fixed final time, this numerical problem is a linear programming problem. Thus, the minimum time can be found by solving a sequence of linear programs. At the minimum time, the linear program returns *an* optimal control. According to Lemma 6, some controls may not be bang-bang. The final step is to replace the non-bang-bang control with bang-bang controls.

This procedure is first described in words. Then, the specific steps of the procedure are enumerated for precision. The first step is to solve the optimization problem as described above. According to Lemma 6, at least one control will be singular and at least two controls will be bang-bang. Removing one of the system equations associated with a bang-bang control eliminates one of the control variables from the problem. Upon solving this reduced system with the other bang-bang control as a known function of time, Lemma 4 guarantees that at least one more control will be made bang-bang. This process of repeatedly reducing the system equations and obtaining a bang-bang control is continued until all controls are bang-bang. Precise statement of the procedure is below.

1. Define $I = \{1, \ldots, M\}$.

2. Solve the optimization problem considering all systems $\dot{x}_i \ \forall i \in I$ (see Eq. 7).

3. Obtain at least two bang-bang controls, u^*_α and u^*_β, and the minimum time, t^*_f. NOTE: The two bang-bang controls do *not* need to be specified a priori. They are a product of optimization in Step 2.

4. Update I. $I = I - \{\beta\}$. Define $\gamma = \alpha$.

5. Solve the optimization problem considering systems $\dot{x}_i, \forall i \in I$ with u^*_γ a known function of time.

6. Obtain at least one bang-bang control, u'_α, and a final time, $t'_\alpha \leq t^*_f$.

7. Define
$$u^*_\alpha = \begin{cases} u'_\alpha, & t \in [t_0, t'_\alpha] \\ u^*_\gamma, & t \in (t'_\alpha, t^*_f] \end{cases}. \tag{15}$$

8. Update I. $I = I - \{\gamma\}$. Define $\gamma = \alpha$.

9. If $I = \{\}$, end. Else, go to Step 5.

Upon completing the procedure, all controls are bang-bang and the problem is solved.

A brief comment is in order on why an equation can be removed each time. As an example, consider the case where there are two spacecraft, and let u^*_1 and u^*_2 be the two bang-bang controls associated with the minimum time t^*_f.

Upon defining $\xi = x_2 - x_1$, the system equations become

$$\dot{x}_1(t) = Ax_1(t) + B[u_1^*(t) - u_0(t)] \tag{16}$$

$$\dot{\xi}(t) = A\xi(t) + B[u_2^*(t) - u_1^*(t)] \tag{17}$$

Thus, the ξ dynamics are completely determined by the two bang-bang controls. Eliminating this system leaves only the x_1 system. Solving the optimization problem with u_1^* as a known function of time gives u_0', which is bang-bang, and a new final time $t_0' \leq t_f^*$. The optimal control u_0' can be extended to t_f^* as in Step 7 above. But this has no effect on the ξ system. Thus, it is possible to remove the x_2 system as described above. This generalizes to the case for arbitrary M.

D. Results

The method is first tested with $M = 1$ (one target and one chaser) using the following constants and initial conditions taken from the literature [46].

$$\begin{gathered} a = 8.24 \text{ 1/hr} \\ b = 50.90 \text{ 1/hr}^2 \\ a_D = 0.59 \text{ km/hr}^2 \\ z(0) = [-0.53 \text{ km}, 0.25 \text{ km/hr}, -0.48 \text{ km}, 3.31 \text{ km/hr}] \end{gathered} \tag{18}$$

The integration step size was set at three minutes. The linear programs were solved using the custom LP code CVXGEN [3], and the one-dimensional search

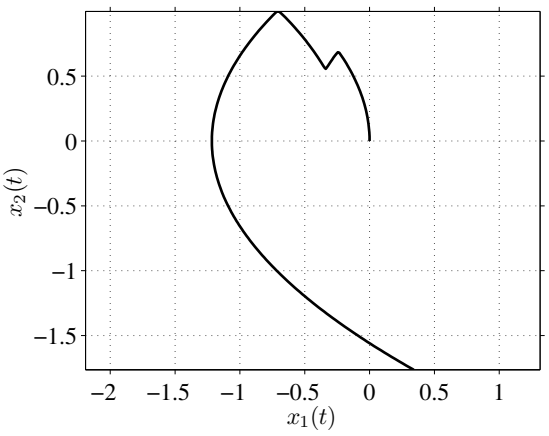

Figure 25: x_1-x_2 phase plane with $M = 1$.

for the minimum time converged in 4 iterations. The minimum time is 4.09 hours. Figure 25 shows the x_1-x_2 phase plane. The three switches are obvious, and the segments associated with positive and negative controls resemble that described earlier in Figure 21.

Figure 26 shows the ωx_3-ωx_4 phase plane. The three switches are not so obvious here. The trajectory first completes the circle on the right, moves to the bottom left, redirects upward, and then moves to the origin along a circular arc. The fact that the trajectory circles the points $(\pm 1, 0)$ resembles the motion described earlier in Figure 22.

Figure 27 shows the control history for the target, $u_0(t)$, and Figure 28 shows the control history for the chaser, $u_1(t)$. Each control switches three times between 0 and -1.

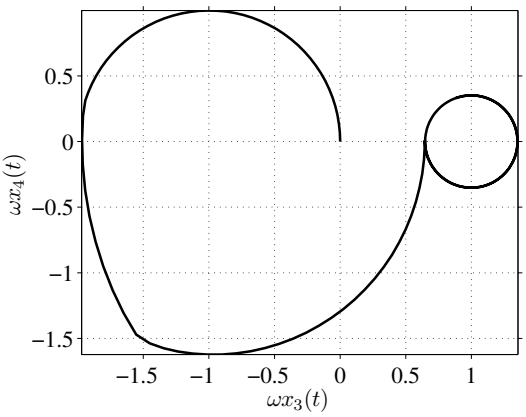

Figure 26: ωx_3-ωx_4 phase plane with $M = 1$.

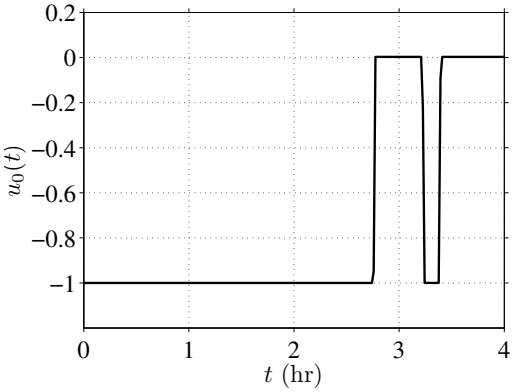

Figure 27: Target spacecraft control with $M = 1$.

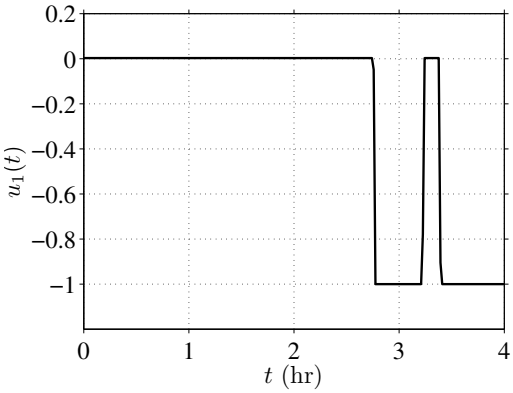

Figure 28: Chaser spacecraft control with $M = 1$.

To help place this in context with existing results, we compare with an analytical sub-optimal approach [46]. The linear programming approach described here took 0.006 seconds to run. The analytical approach, for simplicity's sake, took 0 seconds. However, the linear programming solution is simpler in that the resulting flight is about 1.5 hours shorter and requires two fewer control switches. Further comparisons with other numerical methods have been made elsewhere [13].

Next, the method is tested with $M = 4$ (one target and four chasers). Again, all of the data is from the literature [46]. The initial conditions are

$$\begin{aligned} z_1(0) &= [-0.53 \text{ km}, 0.25 \text{ km/hr}, -0.48 \text{ km}, 3.31 \text{ km/hr}] \\ z_2(0) &= [0.53 \text{ km}, 0.25 \text{ km/hr}, -0.48 \text{ km}, -3.31 \text{ km/hr}] \\ z_3(0) &= [0.38 \text{ km}, 0.25 \text{ km/hr}, -0.38 \text{ km}, -2.30 \text{ km/hr}] \\ z_4(0) &= [0.28 \text{ km}, 0.25 \text{ km/hr}, 0.44 \text{ km}, -1.69 \text{ km/hr}] \end{aligned} \tag{19}$$

The one-dimensional search plus elimination of non-bang-bang controls required 33 iterations in 0.029 seconds for the custom LP solver. The flight time is 8.55 hours. Compared to the analytical approach, the final time reduces by about ten hours from 18.16 hours to 8.55 hours. The control histories for the target and four chasers appear in Figures 29-31 to show that the non-bang-bang controls have been eliminated and all controls belong to the original set $\{-1, 0\}$. In both examples, it was observed that no collisions occurred.

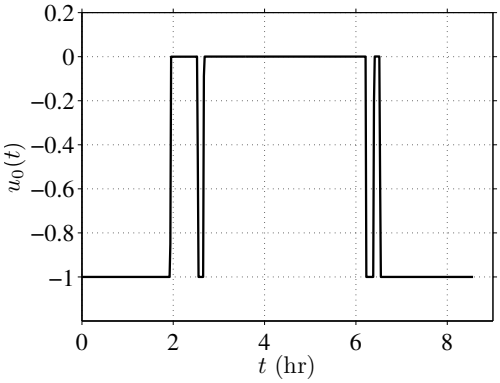

Figure 29: Target spacecraft control $u_0(t)$ with $M = 4$.

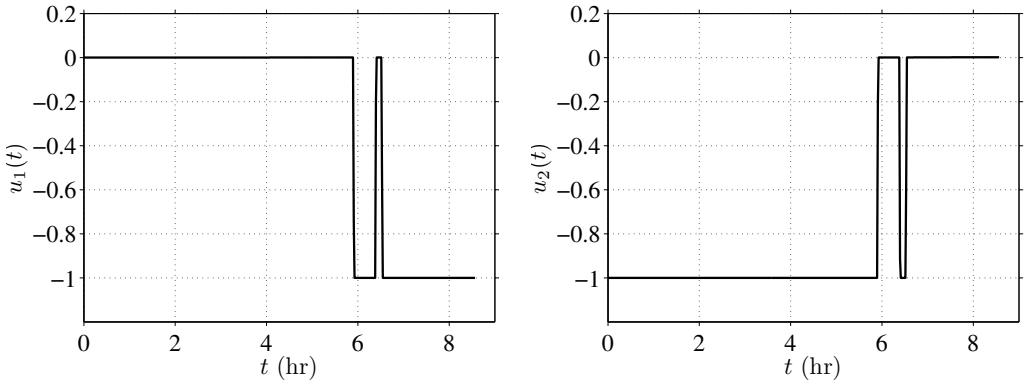

Figure 30: Chaser spacecraft control $u_1(t)$ and $u_2(t)$ with $M = 4$.

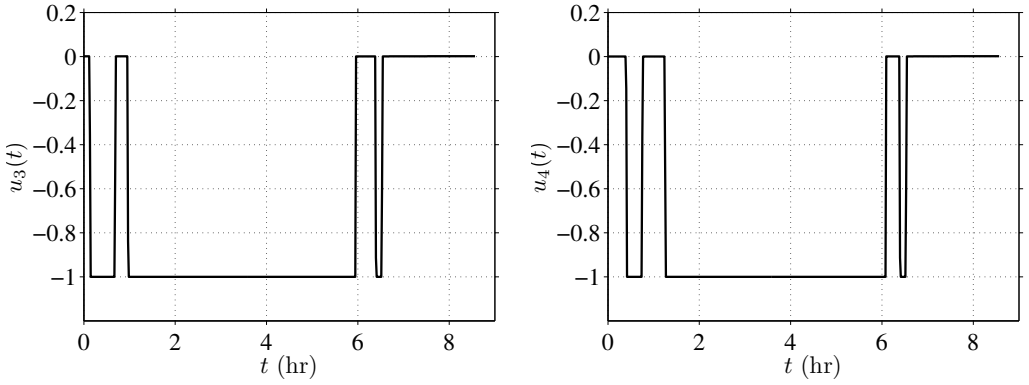

Figure 31: Chaser spacecraft control $u_3(t)$ and $u_4(t)$ with $M = 4$.

It is interesting that the line search converged in four iterations in the two vehicle example and 33 iterations in the five vehicle example. The primary reason for the increase is that steps five through eight of the solution method described are unnecessary in the two vehicle example. In the five vehicle example, these particular steps are not only necessary, but they are repeated three times in the elimination of the non-bang-bang controls. In fact, in the five vehicle example, the first five iterations solve the optimization problem and the next 28 replace non-bang-bang controls with bang-bang controls. Nonetheless, the general convergence properties of the algorithm are not affected by the increase in vehicles, i.e., the guaranteed, polynomial time convergence is derived from the fact that a finite number of linear programs must be solved – not from the special case of there being two vehicles.

E. Summary and Conclusions

In this chapter, we proved lossless convexification for the minimum time rendezvous of any number of spacecraft using differential drag. The work is unique because it did so by specifying a procedure to ensure that singular controls are bang-bang in nature. This is in contrast to typical proofs of lossless convexification that rely on the non-existence of singular controls. Compared to existing methods, our linear programming based method yields significantly shorter mission times at the expense of hundredths or thousandths of seconds in computation time. It is concluded that the theory of lossless convexification offers practical advantages in the differential drag rendezvous problem.

CHAPTER V:
MAXIMUM DIVERT AND LANDING

This chapter presents a method to solve the problem of performing a maximum divert maneuver and landing. The method is based on lossless convexification and reduces the problem to solving a fixed number of convex programming problems. This is the first time lossless convexification has been established for problems with active linear and quadratic state constraints. The problem is to land safely on the surface of a planet. In mid-course, the vehicle is to abort and retarget to a landing site as far from the nominal as physically possible. The divert trajectory must satisfy a number of state and control constraints on the velocity and thrust magnitude.

Large divert type maneuvers have been considered for Mars pinpoint landing [8, 9]. Trajectories generated by similar convex optimization algorithms have recently been flight tested to demonstrate their effectiveness. As an example, flight tests of a vehicle from Masten Space Systems, Inc. showed significant improvements in divert capabilities by using similar algorithms [6]. Their vehicle "Xombie" flew up to 750 meters and landed safely in three consecutive flight tests. Prior flights of Xombie flew diverts less than 100 meters. In these flight tests, velocity constraints were imposed to keep aerodynamic forces below a threshold to ensure structural integrity. Thrust constraints were imposed to point the sensors to the ground. Existing results in convexification do not apply to this problem [8, 9, 12].

Planetary landing problems have long provided motivation for research. Hull showed that minimum fuel vertical take-off/landing trajectories consist only of coast and maximum thrust arcs regardless of the mathematical form of the gravitational force [51, 52]. He used a technique based on Green's Theorem developed by Miele [53]. Meditch analyzed the vertical landing problem with constant gravitational acceleration using the maximum principle [54]. He derived a simple switching function so that the maximum thrust arc begins at the correct time for a safe landing.

More recently, the research has focused on the landing problem that includes the descent phase so that the problem is two or three-dimensional. For example, Hull solves the problem by assuming a flat planet and constant thrust [55, 56]. After a small angle approximation, he finds an analytical solution using throttling to satisfy the boundary conditions. The same ideas can be used in the ascent problem [57]. Analytical solutions are well-suited for guidance applications because of their simplicity, and they provide a certain physical insight into the problem and solution. Numerous other simple solutions exist in the literature, for example, Apollo guidance [58], IGM [59, 60], and PEG [61].

Other recent work has focused more on numerical methods allowing more complicated models and constraints to be incorporated. For example, the problem has been converted to a nonlinear parameter optimization and solved using direct collocation and direct multiple shooting [62]. These direct numerical methods are attractive because explicit use of necessary conditions is not required. In this setting, the infinite-dimensional optimal control problem is

converted to a finite-dimensional parameter optimization problem and solved by a nonlinear programming method [16]. In general, such methods do not offer convergence guarantees.

This work focuses on lossless convexification of the control set in the presence of state constraints. Because of nonlinear dynamics, this does not lead to a convex problem directly. A transformation of variables is introduced to rigorously linearize the dynamics. Unfortunately, this introduces non-convexity back into the problem elsewhere and an approximation must be made. This approximation is conservative in the sense that the new feasible set is contained in the original feasible set. This is desirable for practical reasons, and the numerical example indicates that the approximation is good.

A. Problem Description

The problem is to perform a maximum distance divert from the specified target location and land safely on the planet's surface. It is assumed that 1) the only forces acting on the vehicle are the thrust and gravity forces and 2) the vehicle is sufficiently close to the planet to warrant a flat planet model where the acceleration due to gravity is constant, and 3) the maneuver time is sufficiently short that rotation of the planet can be ignored. The equations of motion under these assumptions are

$$\begin{aligned} \dot{r}(t) &= v(t) \\ \dot{v}(t) &= T(t)/m(t) - g \\ \dot{m}(t) &= -\alpha||T(t)|| \end{aligned} \tag{1}$$

The components of r are the range, cross range, and altitude, and the components of v are their respective rates. For example, r_1 is the range, and v_1 is the range rate. The symbol m denotes the mass. The thrust is T, and the gravity is g, which is constant and points only in the altitude direction. The positive constant α is the engine constant describing the mass flow rate.

All of the initial conditions for the system are specified and denoted with a subscript zero. The vehicle is required to land softly on the surface; thus, the final altitude and final velocities are specified and denoted with a subscript f.

$$\begin{aligned} &r(t_0) = r_0, \qquad v(t_0) = v_0, \qquad m(t_0) = m_0 \\ &r_3(t_f) = 0, \qquad v(t_f) = 0 \end{aligned} \tag{2}$$

The final range, cross range, and mass are free; however, the final mass cannot be less than the dry mass of the vehicle, i.e., $m(t_f) \geq m_c$.

The problem is to perform a maximum divert from the specified target location; that is, given a nominal landing site, divert to a landing site as far from the nominal as physically possible. For convenience, the nominal landing site is specified to be the origin. Mathematically, the performance index for the problem is

$$\max \quad w_1 r_1(t_f) + w_2 r_2(t_f) \tag{3}$$

where the constants w_1 and w_2 are non-zero weights. By picking the signs and magnitudes of the weights in the performance index, the analyst can design the trajectory. As an example, consider the scenario described in Figure 32.

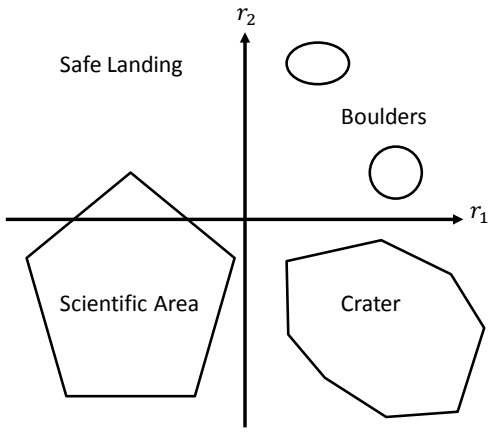

Figure 32: Maximum divert landing scenario.

The axes are the range and cross range, and the nominal landing site is at the origin. The terrain associated with the first quadrant contains a number of boulders and irregular ground unsuitable for a safe landing. The third quadrant and lower portion of the second quadrant contains the location of scientific interest, and it is best to leave this area undisturbed for research. The fourth quadrant contains a large crater, which is also unsuitable for a safe landing. Thus, the best location for divert is the upper portions of the second quadrant. In this case, the weights are chosen as $w_1 < 0$, $w_2 > 0$, and $|w_2| > |w_1|$.

Another scenario where maximum divert trajectories are useful is when a lander is asked to land near a desired location, $(r_{1_d}, r_{2_d}, 0)$, but it is physically impossible to get there. In that case, the weights can be chosen as $w_1 = r_{1d} - r_{1_0}$ and $w_2 = r_{2d} - r_{2_0}$ to ensure maximum divert towards the target location.

There are a number of safety requirements for landing problems. A common one is that the velocity of the vehicle not exceed a critical value. For

example, it is not desirable for the speed to approach the speed of sound, which leads to instabilities in the thrusters. There are three velocity constraints added to the problem: 1) a range velocity constraint, 2) a cross range velocity constraint, and 3) a total speed constraint.

$$|v_1(t)| \leq V_a, \quad |v_2(t)| \leq V_b, \quad ||v(t)|| \leq V_c \tag{4}$$

The quantities V_a, V_b, and V_c are constant values describing the upper bounds. Finally, the thrust magnitude cannot exceed a lower and upper bound. The lower bound exists because the engine cannot operate reliably below the bound. The upper bound exists because arbitrarily large thrusts are impossible.

$$\rho_1 \leq ||T(t)|| \leq \rho_2 \tag{5}$$

This thrust magnitude constraint is not convex. This type of constraint looks like the annulus in Figure 33 where the shaded region is the admissible region.

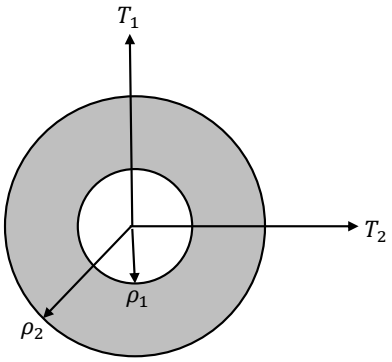

Figure 33: Two-dimensional non-convex thrust constraint.

The following assumptions are also made.

Assumption 1. *The velocity bounds satisfy* $V_a^2 + V_b^2 > V_c^2$ *such that the range and cross range velocity constraints cannot be active simultaneously.*

Assumption 2. *The thrust bounds satisfy* $\rho_1 \leq ||m_c g||$ *and* $||m_0 g|| \leq \rho_2$ *such that the thrust force can always cancel the gravitational force.*

For convenience, the performance index and all constraints are collected here in the statement of the original problem P0. Without loss of generality, the maximization problem has been converted to a minimization problem.

$$
\begin{aligned}
\min \quad & J = -w_1 r_1(t_f) - w_2 r_2(t_f) && && \text{(P0)} \\
\text{subj. to} \quad & \dot{r}(t) = v(t), && r(t_0) = r_0,\ r_3(t_f) = 0 \\
& \dot{v}(t) = T(t)/m(t) - g, && v(t_0) = v_0,\ v(t_f) = 0 \\
& \dot{m}(t) = -\alpha ||T(t)||, && m(t_0) = m_0,\ m_c \leq m(t_f) \\
& |v_1(t)| \leq V_a, \quad |v_2(t)| \leq V_b, \quad ||v(t)|| \leq V_c \\
& \rho_1 \leq ||T(t)|| \leq \rho_2
\end{aligned}
$$

As stated, problem P0 is highly constrained and non-convex. The non-convexity arises because of 1) the lower bound on the thrust magnitude and 2) the non-linear dynamics. The rest of this chapter addresses lossless convexification of the thrust constraint and a transformation of variables for the nonlinear dynamics.

B. Lossless Convexification

The purpose of this section is to prove a lossless convexification of the non-convex thrust constraint in Equation 5. The proposed relaxation is to replace Equation 5 with the two constraints

$$||T(t)|| \leq \Gamma(t) \quad \text{and} \quad \rho_1 \leq \Gamma(t) \leq \rho_2 \tag{6}$$

where $\Gamma(t)$ is a scalar slack variable. The variable $\Gamma(t)$ is also inserted into the mass dynamics, but all other constraints are kept the same such that the only source of non-convexity is the nonlinear dynamics. Geometrically, this relaxation is obtained by pulling the annulus of Figure 33 out along the Γ direction. This is shown in Figure 34.

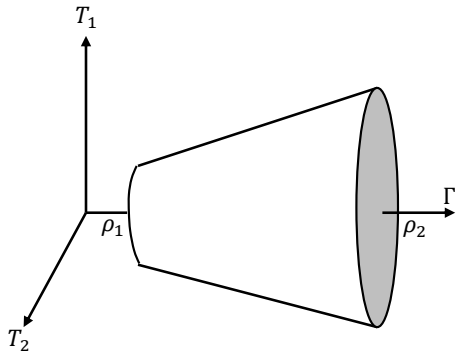

Figure 34: Relaxed thrust constraints.

The thrust constraints are now convex, however, optimal solutions of the relaxed problem are not necessarily optimal solutions of the original problem. For example, the control $(T(t), \Gamma(t)) = (0, \rho_1)$ is feasible in the relaxed prob-

lem, but $T(t) = 0$ is not feasible in the original problem. The goal is to show that any optimal solution satisfies $||T(t)|| = \Gamma(t)$, in which case, the relaxation is lossless, i.e., optimal solutions of the relaxed problem are indeed optimal solutions of the original problem. In Figure 34, lossless convexification holds when the thrust is always on the boundary of the cone shape.

The relaxed problem P1 is now stated. All of the constraints which are non-differentiable in P0 are written in a form so that they are differentiable as required by Theorem 2 of Chapter III.

$$
\begin{aligned}
\min \quad & J = -w_1 r_1(t_f) - w_2 r_2(t_f) && && \text{(P1)}\\
\text{subj. to} \quad & \dot{r}(t) = v(t), && r(t_0) = r_0,\ r_3(t_f) = 0 \\
& \dot{v}(t) = T(t)/m(t) - g, && v(t_0) = v_0,\ v(t_f) = 0 \\
& \dot{m}(t) = -\alpha\Gamma(t), && m(t_0) = m_0,\ m_c \leq m(t_f) \\
& v_1(t) \leq V_a, \quad -v_1(t) \leq V_a, \quad v_2(t) \leq V_b, \quad -v_2(t) \leq V_b \\
& ||v(t)||^2 \leq V_c^2, \quad ||T(t)||^2 \leq \Gamma^2(t), \quad \rho_1 \leq \Gamma(t) \leq \rho_2
\end{aligned}
$$

The goal is to now prove lossless convexification of the thrust constraint. To synthesize the relaxed problem, the useful conditions from Theorem 2 of Chapter III are listed below. The Hamiltonian is

$$\mathcal{H}[t] = p_r(t)^T v(t) + p_v(t)^T (T(t)/m(t) - g) - \alpha p_m(t)\Gamma(t) \tag{7}$$

The Lagrangian is

$$\begin{aligned}\mathcal{L}[t] = \mathcal{H}[t] &+ \lambda_1(t)(||T(t)||^2 - \Gamma^2(t)) + \lambda_2(t)(\rho_1 - \Gamma(t)) + \lambda_3(t)(\Gamma - \rho_2) \\ &+ \nu_1^+(t)(v_1(t) - V_a) + \nu_1^-(t)(-v_1(t) - V_a) \\ &+ \nu_2^+(t)(v_2(t) - V_b) + \nu_2^-(t)(-v_2(t) - V_b) + \nu_3(t)(||v(t)||^2 - V_c^2)\end{aligned} \tag{8}$$

The endpoint function is

$$\begin{aligned}\mathcal{G}[t] = &-p_0 w_1 r_1(t_f) - p_0 w_2 r_2(t_f) + \xi_{r_3} r_3(t_f) + \xi_v^T v(t_f) \\ &+ \zeta_m(m_c - m(t_f)) + \mu_1^+(v_1(t_f) - V_a) + \mu_1^-(-v_1(t_f) - V_a) \\ &+ \mu_2^+(v_2(t_f) - V_b) + \mu_2^-(-v_2(t_f) - V_b) + \mu_3(||v(t_f)||^2 - V_c^2)\end{aligned} \tag{9}$$

The adjoint differential equations are

$$\begin{aligned}\dot{p}_r(t) &= 0 \\ \dot{p}_v(t) &= -p_r(t) - e_1(\nu_1^+(t) - \nu_1^-(t)) - e_2(\nu_2^+(t) - \nu_2^-(t)) - 2\nu_3(t)v(t) \\ \dot{p}_m(t) &= p_v(t)^T T(t)/m^2(t)\end{aligned} \tag{10}$$

The stationary conditions are

$$\begin{aligned}\partial_T \mathcal{L}[t] &= p_v(t)/m(t) + 2\lambda_1(t)T(t) = 0 \\ \partial_\Gamma \mathcal{L}[t] &= -\alpha p_m(t) - 2\lambda_1(t)\Gamma(t) - \lambda_2(t) + \lambda_3(t) = 0\end{aligned} \tag{11}$$

The complementary slackness conditions for the control constraints are

$$\begin{array}{lll} ||T(t)||^2 - \Gamma^2(t) \leq 0, & \lambda_1(t) \leq 0, & \lambda_1(t)(||T(t)||^2 - \Gamma^2(t)) = 0 \\ \rho_1 - \Gamma(t) \leq 0, & \lambda_2(t) \leq 0, & \lambda_2(t)(\rho_1 - \Gamma(t)) = 0 \\ \Gamma(t) - \rho_2 \leq 0, & \lambda_3(t) \leq 0, & \lambda_3(t)(\Gamma(t) - \rho_2) = 0 \end{array} \tag{12}$$

The complementary slackness conditions for the state constraints are

$$\begin{array}{lll} v_1(t) - V_a \leq 0, & \nu_1^+(t) \leq 0, & \nu_1^+(t)(v_1(t) - V_a) = 0 \\ -v_1(t) - V_a \leq 0, & \nu_1^-(t) \leq 0, & \nu_1^-(t)(-v_1(t) - V_a) = 0 \\ v_2(t) - V_b \leq 0, & \nu_2^+(t) \leq 0, & \nu_2^+(t)(v_2(t) - V_b) = 0 \\ -v_2(t) - V_b \leq 0, & \nu_2^-(t) \leq 0, & \nu_2^-(t)(-v_2(t) - V_b) = 0 \\ ||v(t)||^2 - V_c^2 \leq 0, & \nu_3(t) \leq 0, & \nu_3(t)(||v(t)||^2 - V_c^2) = 0 \end{array} \tag{13}$$

The complementary slackness conditions at the final time are

$$\begin{array}{lll} m_c - m(t_f) \leq 0, & \zeta_m \leq 0, & \zeta_m(m_c - m(t_f)) = 0 \\ v_1(t_f) - V_a \leq 0, & \mu_1^+ \leq 0, & \mu_1^+(v_1(t_f) - V_a) = 0 \\ -v_1(t_f) - V_a \leq 0, & \mu_1^- \leq 0, & \mu_1^-(-v_1(t_f) - V_a) = 0 \\ v_2(t_f) - V_b \leq 0, & \mu_2^+ \leq 0, & \mu_2^+(v_2(t_f) - V_b) = 0 \\ -v_2(t_f) - V_b \leq 0, & \mu_2^- \leq 0, & \mu_2^-(-v_2(t_f) - V_b) = 0 \\ ||v(t_f)||^2 - V_c^2 \leq 0, & \mu_3 \leq 0, & \mu_3(||v(t_f)||^2 - V_c^2) = 0 \end{array} \tag{14}$$

The transversality conditions are

$$\begin{aligned} p_r(t_f) &= -e_1 p_0 w_1 - e_2 p_0 w_2 + e_3 \xi_{r_3} \\ p_v(t_f) &= \xi_v + e_1(\mu_1^+ - \mu_1^-) + e_2(\mu_2^+ - \mu_2^-) + 2\mu_3 v(t_f) \\ p_m(t_f) &= -\zeta_m \end{aligned} \tag{15}$$

Finally, because the final time is free and the problem does not depend explicitly on time, the Hamiltonian is identically zero, i.e.,

$$\mathcal{H}[t] = 0 \; \forall t \tag{16}$$

With the optimality conditions clearly stated, the goal is to now show that any optimal solution of P1 is also an optimal solution of P0.

Lemma 1. *i) If $T(\cdot)$ is feasible for P0, then there exists a $\Gamma(\cdot)$ such that $(T(\cdot), \Gamma(\cdot))$ is feasible for P1. ii) If $(T(\cdot), \Gamma(\cdot))$ is feasible for P1 and $||T(t)|| = \Gamma(t)$ $\forall t$, then $T(\cdot)$ is feasible for P0.*

Proof. i) Suppose that $T(\cdot)$ is feasible for P0. Define $\Gamma(t) := ||T(t)||$ $\forall t$. It follows that $\rho_1 \leq \Gamma(t) \leq \rho_2$ $\forall t$, which implies that $(T(\cdot), \Gamma(\cdot))$ is feasible for P1. ii) Suppose that $(T(\cdot), \Gamma(\cdot))$ is feasible for P1 and $||T(t)|| = \Gamma(t)$ $\forall t$. Since $\rho_1 \leq ||T(t)|| \leq \rho_2$ $\forall t$, it follows that $T(\cdot)$ is feasible for P0. □

Lemma 2. *If $(T(\cdot), \Gamma(\cdot))$ is optimal for P1, then $||T(t)|| = \Gamma(t)$ $\forall t$.*

Proof. Suppose that $(T(\cdot), \Gamma(\cdot))$ is optimal for P1 and there exists a t where $||T(t)|| < \Gamma(t)$. Because $T(\cdot)$ and $\Gamma(\cdot)$ are piecewise continuous, there exists

an interval, $[\tau_1, \tau_2] \subset [t_0, t_f]$, where $||T(t)|| < \Gamma(t)\ \forall t \in [\tau_1, \tau_2]$. The first of the slackness conditions in Equation 12 implies that $\lambda_1(t) = 0$, and it follows from Equation 11 that $p_v(\cdot)$ and its derivatives are zero on the interval. Solving for p_r using Equation 10 gives

$$p_r = -e_1(\nu_1^+(t) - \nu_1^-(t)) - e_2(\nu_2^+(t) - \nu_2^-(t)) - 2\nu_3(t)v(t) \tag{17}$$

where the time dependence of p_r has been dropped since it is constant. Computing $\ddot{p}_v(t)$ and pre-multiplying with $\dot{v}(t)^T$ gives

$$\begin{aligned} \dot{v}(t)^T\ddot{p}_v(t) = -\dot{v}(t)^T e_1(\dot{\nu}_1^+(t) - \dot{\nu}_1^-(t)) - \dot{v}(t)^T e_2(\dot{\nu}_2^+(t) - \dot{\nu}_2^-(t)) \\ - 2\dot{\nu}_3(t)\dot{v}(t)^T v(t) - 2\nu_3(t)\dot{v}(t)^T\dot{v}(t) = 0 \end{aligned} \tag{18}$$

Note that the first three terms on the right hand side are each zero. For example, looking at the first term, if the the velocity constraint on v_1 is inactive, then $\nu_1^+(t) = \nu_1^-(t) = \dot{\nu}_1^+(t) = \dot{\nu}_1^-(t) = 0$. If the constraint is active, then $v_1(t) = \pm v_a$ and $\dot{v}_1(t) = 0$. In either case, the term is zero. Similar arguments apply to the second and third terms. Thus, the fourth term must also be zero, which implies that

$$\nu_3(t) = 0 \quad \text{or} \quad \dot{v}(t) = 0 \tag{19}$$

The remainder of the proof addresses these two cases.

Case 1. Suppose that $\nu_3(t) = 0$. Then $p_r(t)$ becomes

$$p_r = -e_1(\nu_1^+(t) - \nu_1^-(t)) - e_2(\nu_2^+(t) - \nu_2^-(t)) \tag{20}$$

Since p_r is constant, the values above can be equated with the transversality conditions. Writing in component form gives

$$\begin{aligned} p_{r_1} &= -(\nu_1^+ - \nu_1^-) = -w_1 p_0 \\ p_{r_2} &= -(\nu_2^+ - \nu_2^-) = -w_2 p_0 \\ p_{r_3} &= \xi_{r_3} = 0 \end{aligned} \tag{21}$$

Because $V_a^2 + V_b^2 > V_c^2$ by Assumption 1, the range and cross range velocity constraints cannot be active simultaneously. For sake of argument, and without loss of generality, suppose the range constraint is the inactive one such that $\nu_1^+ = \nu_1^- = 0$. It follows immediately, since w_1 and w_2 are nonzero, that $p_0 = 0$ and that $\nu_2^+ = \nu_2^- = 0$. Thus, it has been shown that $p_v(t) = p_r(t) = 0$. Finally, because of the Hamiltonian condition in Equation 16, it is true that $p_m(t) = 0$. This violates the non-triviality condition since

$$(p_0, p_r(t), p_v(t), p_m(t)) = 0 \tag{22}$$

The case when $\nu_3(t) = 0$ has been ruled out.

Case 2. Suppose that $\dot{v}(t) = 0$ and $\nu_3(t) \neq 0$. For this to be true, the thrust acceleration must equal the gravitational acceleration, i.e., $T(t) = m(t)g$, which by Assumption 2, satisfies $\rho_1 \leq ||T(t)|| \leq \rho_2$. The pointwise maximum condition reduces to

$$\max \quad -\alpha p_m(t)\Gamma(t) \tag{23}$$

subject to $||T(t)|| \leq \Gamma(t)$ and $\rho_1 \leq \Gamma \leq \rho_2$. If $p_m(t) > 0$, it is clear that $\Gamma(t)$

should be made as small as possible, i.e., $||T(t)|| = \Gamma(t)$. It is now shown that $p_m(t)$ is indeed positive. Equation 16 along with $p_v(t) = 0$ gives

$$p_r^T v(t) = \alpha p_m(t)\Gamma(t) \tag{24}$$

Expanding the left hand side gives

$$p_r^T v(t) = -(\nu_1^+(t) - \nu_1^-(t))v_1(t) - (\nu_2^+(t) - \nu_2^-(t))v_2(t) - 2\nu_3(t)||v(t)||^2 \tag{25}$$

Note that $||v(t)|| = V_c$, since otherwise $\nu_3(t) = 0$. Expanding the above equation gives

$$p_r^T v(t) = -\nu_1^+(t)v_1(t) + \nu_1^-(t)v_1(t) - \nu_2^+(t)v_2(t) + \nu_2^-(t)v_2(t) - 2\nu_3(t)V_c^2 \tag{26}$$

Each term on the right hand side is non-negative. For example, if $v_1(t) < V_a$, then $\nu_1^+(t) = 0$. If $v_1(t) = V_a$, then $\nu_1^+(t) \leq 0$ such that $-\nu_1^+(t)v_1(t) \geq 0$. Similar arguments hold for the other terms. Finally, because $\nu_3(t)$ is strictly negative, the right hand side is strictly positive. Thus, using Equation 23, $p_m(t) > 0$ and $||T(t)|| = \Gamma(t)$, which contradicts the original hypothesis.

Thus, the lemma has been established and $(T(\cdot), \Gamma(\cdot))$ being optimal for P1 implies $||T(t)|| = \Gamma(t)\ \forall t$. □

The following theorem is the main result of this section. It states that optimal solutions of P1 are optimal solutions of P0.

Theorem 1. *If* $(T(\cdot), \Gamma(\cdot))$ *is optimal for P1, then* $T(\cdot)$ *is optimal for P0.*

Proof. Suppose that $(T(\cdot), \Gamma(\cdot))$ is optimal for P1. Lemma 2 implies that $||T(t)|| = \Gamma(t)\ \forall t$. Lemma 1 implies that $T(\cdot)$ is feasible for P0. Because the optimal solution of P1 is a feasible solution for P1, it must be that $J_0^* \leq J_1^*$. Similarly, from Lemma 1, feasible solutions for P0 define feasible solutions for P1. Thus, $J_1^* \leq J_0^*$. The two inequalities imply equality: $J_1^* = J_0^*$. Therefore, $T(\cdot)$ is optimal for P0. □

The theorem is a statement of lossless convexification of the thrust constraint, and it is true because all optimal solutions of P1 are on the boundary of the relaxed control set, i.e., $||T(t)|| = \Gamma(t)$. The boundary of the relaxed set generates feasible controls for the original problem, and hence, optimal controls for the original problem.

The fact that the controls are on the boundary is not surprising – in fact, the maximum principle says that an optimal control maximizes the Hamiltonian pointwise. This is what motivated the relaxation moving from Figure 33 to Figure 34. A projection of the optimal control set of P1 onto the $T-$plane (in Figure 34) coincides with the feasible set of P0 (in Figure 33).

C. Transformation of Variables

After the control relaxation, the only source of non-convexity is the nonlinear equations of motion. A transformation of variables is made to rigorously linearize the equations, but, unfortunately, this causes the control bounds to become time-varying and non-convex. These bounds are then approximated

in a conservative sense so that feasible solutions are still feasible in the original problem. The transformation is

$$\sigma(t) = \frac{\Gamma(t)}{m(t)}, \quad u(t) = \frac{T(t)}{m(t)}, \quad z(t) = \ln(m(t)) \tag{27}$$

which is well-defined since the mass is strictly positive. This leads to the linearized equations of motion

$$\begin{aligned} \dot{r}(t) &= v(t) \\ \dot{v}(t) &= u(t) + g \\ \dot{z}(t) &= -\alpha\sigma(t) \end{aligned} \tag{28}$$

The relaxed thrust constraints in Equation 6 become

$$||u(t)|| \leq \sigma(t) \tag{29a}$$

$$\rho_1 e^{-z(t)} \leq \sigma(t) \leq \rho_2 e^{-z(t)} \tag{29b}$$

The inequality in Equation 29a is convex. The left inequality in Equation 29b defines a convex region, but the right inequality does not. Further, even though the lower constraint is convex, it does not fit within the structure of a second-order cone problem. For this reason, it is approximated as a quadratic constraint using a Taylor series. The upper constraint is approximated as a linear constraint so that it is convex. The Taylor series approximation is based

on the reference trajectory

$$z_r(t) = \begin{cases} \ln(m_0 - \alpha\rho_2 t) & 0 \le t \le (m_0 - m_c)/(\alpha\rho_2) \\ \ln(m_c) & \text{otherwise} \end{cases} \tag{30}$$

which is a lower bound on $z(t)$ at time t. This guarantees that the approximation is conservative [8]. Then the approximation of Equation 29b is

$$\begin{aligned} \rho_1 e^{-z_r(t)} \left[1 - (z(t) - z_r(t)) + \frac{1}{2}(z(t) - z_r(t))^2\right] &\le \sigma(t) \\ &\le \rho_2 e^{-z_r(t)} \left[1 - (z(t) - z_r(t))\right] \end{aligned} \tag{31}$$

This leads to the statement of the convex problem P2. It is emphasized that this problem is not equivalent to problems P0 and P1 due to the conservative approximations just made. It is true however that this problem is convex and its optimal solutions are feasible solutions to P0 and P1.

$$\begin{aligned} \min \quad & J = -w_1 r_1(t_f) - w_2 r_2(t_f) && \text{(P2)} \\ \text{subj. to} \quad & \dot{r}(t) = v(t), && r(t_0) = r_0,\ r_3(t_f) = 0 \\ & \dot{v}(t) = u(t) - g, && v(t_0) = v_0,\ v(t_f) = 0 \\ & \dot{z}(t) = -\alpha\sigma(t), && z(t_0) = \ln(m_0),\ \ln(m_c) \le z(t_f) \\ & v_1(t) \le V_a, \quad -v_1(t) \le V_a, \quad v_2(t) \le V_b, \quad -v_2(t) \le V_b \\ & ||v(t)||^2 \le V_c^2, \quad ||u(t)||^2 \le \sigma^2(t) \\ & \rho_1 e^{-z_r(t)} \left[1 - (z(t) - z_r(t)) + \tfrac{1}{2}(z(t) - z_r(t))^2\right] \le \sigma(t) \\ & \sigma(t) \le \rho_2 e^{-z_r(t)} \left[1 - (z(t) - z_r(t))\right] \end{aligned}$$

D. Results

The convexification and solution method are now tested in a specific landing scenario using CVX [20,63]. The initial conditions are set for a nominal landing (c.f. [9]), at which point it is decided to perform a maximum divert maneuver. The boundary conditions are

$$\begin{aligned} r(0) &= [2000,\ 500,\ 1500]\ \text{m}, & r_3(t_f) &= 0\ \text{m} \\ v(0) &= [-25,\ -10,\ -20]\ \text{m/s}, & v(t_f) &= [0,\ 0,\ 0]\ \text{m/s} \\ m(0) &= 1905\ \text{kg}, & m(t_f) &\geq 1505\ \text{kg} \end{aligned} \tag{32}$$

The initial conditions are fixed. The final range and cross range are free; however, the final altitude, velocities, and mass must satisfy constraints. The planet of interest is Mars, and the constant acceleration vector due to gravity is $g = [0,\ 0,\ -3.71]\ \text{m/s}^2$. The state constraints and thrust constants are

$$\begin{aligned} |v_1(t)| &\leq 45\ \text{m/s}, & \rho_1 &= 4972\ \text{N} \\ |v_2(t)| &\leq 45\ \text{m/s}, & \rho_2 &= 13260\ \text{N} \\ ||v(t)|| &\leq 50\ \text{m/s}, & \alpha &= 4.53 \times 10^{-4}\ \text{s/m} \end{aligned} \tag{33}$$

Note that the range rate and cross range rate constraint magnitudes do not have to be equal, and Assumption 1 is satisfied.

To illustrate different landing scenarios, three sets of weights have been chosen. In the first scenario, the goal is to land near the middle of the second quadrant using the weights $w_1 = -1$ and $w_2 = 1$. In the second scenario,

the goal is to push the trajectory 'upward' using the weights $w_1 = -1$ and $w_2 = 10$. In the third scenario, the goal is to push the trajectory 'downward' using the weights $w_1 = -10$ and $w_2 = 1$. Results are illustrated in Figure 35.

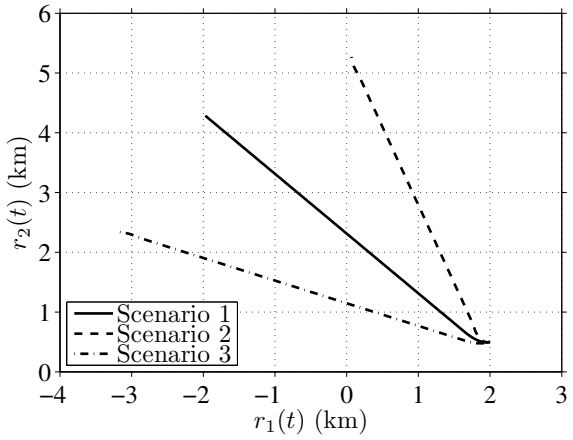

Figure 35: Comparison of positions using different weights.

To show that the different state constraints become active in the different scenarios, plots of the total speed, range rate, and cross range rate are shown in Figure 36 on page 117.

Finally, to illustrate that the thrust constraints are also satisfied, the thrust magnitude is plotted in Figure 37 on page 118 for the first scenario where $w_1 = -1$ and $w_2 = 1$. The horizontal dashed lines that bound the thrust correspond to ρ_1 and ρ_2. It is evident that the constraint is in fact satisfied. The interval where the thrust is not at the upper boundary corresponds to a state boundary arc where the velocity constraints are active. The fact that this intermediate thrust is always admissible follows from Assumption 2.

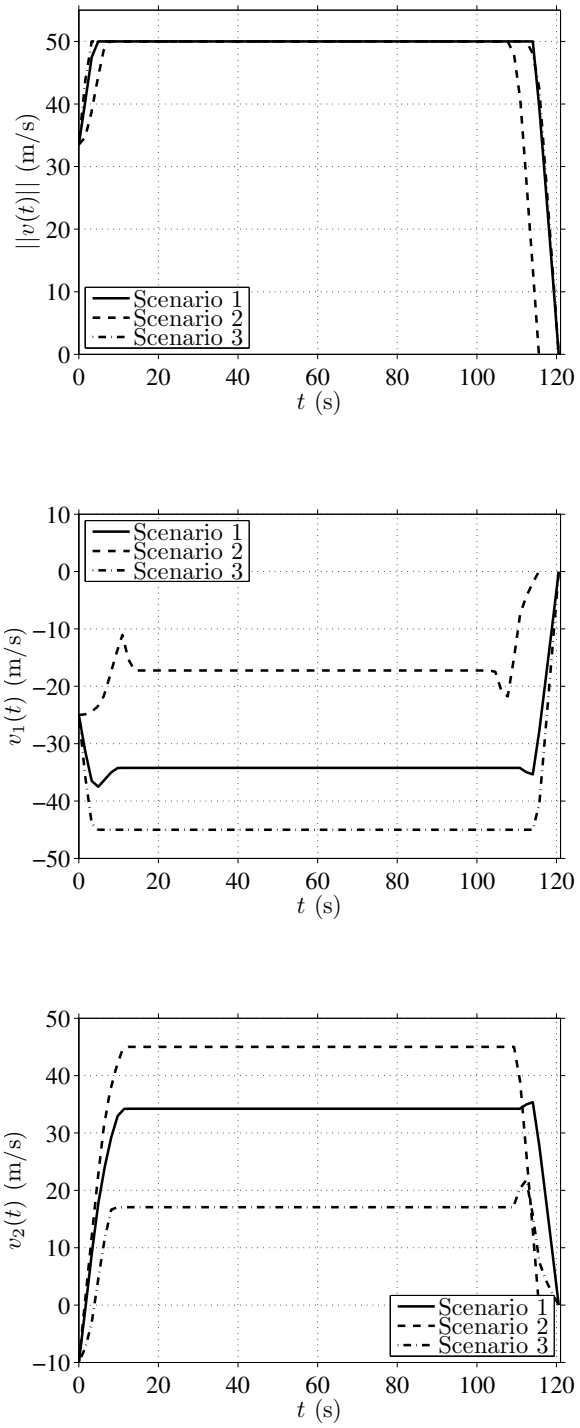

Figure 36: Comparison of velocities using different weights.

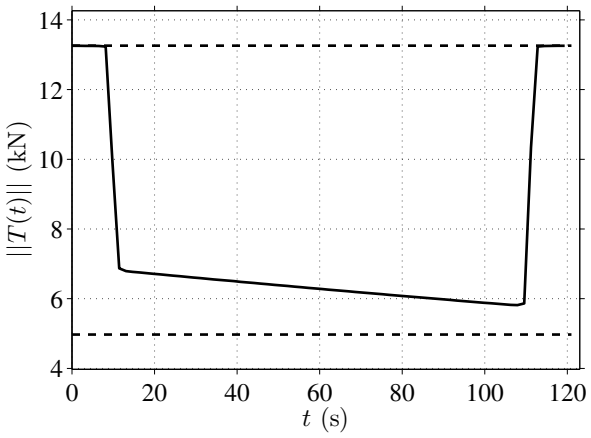

Figure 37: Thrust magnitude for maximum divert.

Figure 38 shows the mass profile. The mass decreases to the dry mass of the vehicle. This is expected since the goal is to maximize the distance traveled. If mass were left over, the vehicle could have traveled further.

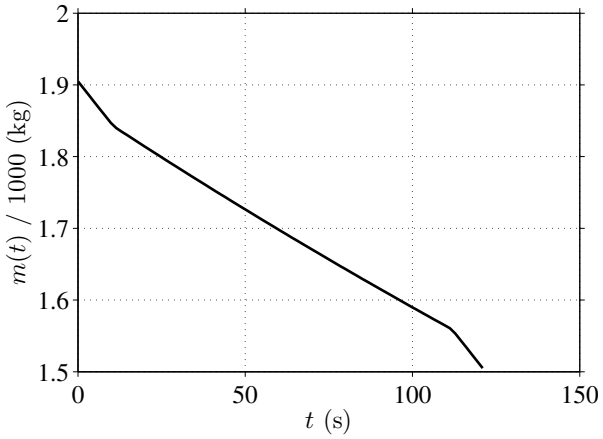

Figure 38: Mass profile for maximum divert.

E. Summary and Conclusions

In this chapter, we proved lossless convexification of a thrust constraint in the maximum divert and landing problem. The work is unique because it did so in the presence of active linear and quadratic state constraints. Additionally, it serves as a mathematical foundation for current flight tests that use lossless convexification and convex optimization in the guidance routine. The evolution of real-time convex optimization depends on strong theoretical and numerical guarantees. This chapter enlarges the class of problems that can be solved with these guarantees. It is concluded that lossless convexification offers practical advantages in the planetary landing problem.

CHAPTER VI:
A GENERAL RESULT FOR LINEAR SYSTEMS

This chapter presents lossless convexification for a class of optimal control problems with linear state constraints and mixed convex and non-convex control constraints. It is the most general result to date since 1) convexification is tied not to any one problem, but to a class of problems that share some system properties, and 2) the results simplify to the known special cases.

Previous convexification results are all tied to controllability of the dynamic system [10, 11]. The generality of our current result is best seen in this light. By bringing the notions of strongly controllable/observable subspaces into the picture [64, p. 182-188], we naturally incorporate the dynamics, control constraints, and restricted state space into the convexification result. The new result states that convexification holds when the state space is a strongly controllable subspace for the system (dynamics and control constraints). When there are no additional control and state constraints, the state space becomes $\mathbb{R}^n$ and strong controllability reduces to the standard notion of controllability. This indicates a deep and useful connection with multivariable systems theory.

There are numerous problems that belong to the class studied here and provide strong motivation for this work. In the planetary landing problem [65], a spacecraft lands using thrusters, which produce a force vector whose magnitude is bounded above and below. The lower bound introduces a non-convex control constraint. The landing problem also incorporates a number of state

constraints, e.g., altitude limits and landing cones, which can be written as linear state constraints. The rendezvous problem in low earth orbit is based on the Clohessy-Wiltshire-Hill equations and fits squarely within the class of problems studied here [45]. That problem has pointing constraints, thrust magnitude constraints, and more. Additionally, the Clohessy-Wiltshire-Hill equations decouple into a double integrator and harmonic oscillator. Integrators and oscillators are prevalent in many mechanical systems, and so our results here apply to many practical problems.

The rest of the chapter formally states the problem of interest and assumptions on which convexification can be proved. A few results from linear systems theory are needed such as controllability and observability subspaces. These are documented and expanded upon as necessary. Finally, lossless convexification is proved by analyzing convex relaxations of the original problem. Unfortunately, the most intuitive relaxation cannot be analyzed simply, and so variable transformations are introduced and the dimension of the system is reduced. It is in this reduced state that convexification is proved. While this process is laborious and less than ideal, it does not weaken the result. That is, one can still claim that convexification holds for the most intuitive relaxation.

A. Problem Description

This section introduces the optimal control problem that is of primary interest. It is labeled as problem P0.

$$\begin{aligned} \min \quad & J = m(t_f, x(t_f)) + \int_{t_0}^{t_f} \ell(\kappa(u(t))) \, dt \\ \text{subj. to} \quad & \dot{x}(t) = Ax(t) + Bu(t) + Ew(t), \quad x(t_0) = x_0 \\ & 0 < \rho_1 \leq \kappa(u(t)) \leq \rho_2, \quad Cu(t) \leq d \\ & x(t) \in \mathcal{X}, \quad b(t_f, x(t_f)) = 0 \end{aligned} \tag{P0}$$

The typical convexification result does not include the linear control constraint $Cu(t) \leq d$ nor the state constraint $x(t) \in \mathcal{X}$. The linear control constraint is important since it is one way to enforce pointing type constraints in a landing problem. The state constraint is important since most practical problems have constraints such as altitude limits, velocity limits, or something similar. The following constraint qualifications are made:

- The function $m(\cdot, \cdot)$ is affine in both arguments.
- The functions $\ell(\cdot)$ and $\kappa(\cdot)$ are convex and strictly positive except possibly at the origin.
- The set $\mathcal{X}$ is a linear subspace of $\mathbb{R}^n$ with dimension k_x.
- The function $b(\cdot, \cdot)$ is affine in both arguments.

Additionally, for consistency with Theorem 2 of Chapter III, it is assumed that each function is differentiable with respect to all of its arguments.

Problem P0 is a non-convex optimal control problem since the control inequality constraint $0 < \rho_1 \leq \kappa(u(t)) \leq \rho_2$ is non-convex. In two dimensions, this non-convex constraint looks like an annulus as in Figure 39.

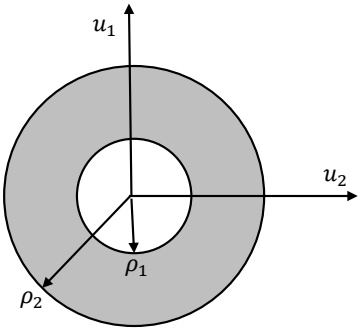

Figure 39: Two-dimensional non-convex control constraint.

The shaded region is the set of admissible controls, and it is clearly non-convex. Problem P0 also includes linear inequality constraints on the control of the form $Cu(t) \leq d$. In the two-dimensional example, each row of C represents a line cutting through the annulus. For example, when C has two rows C_1 and C_2, the constraints are illustrated in Figure 40.

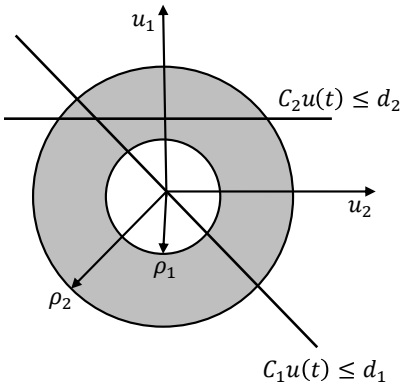

Figure 40: Two-dimensional constraint with linear inequalities.

The set of admissible controls becomes the shaded region below the C_2 constraint and left of the C_1 constraint. The control set remains non-convex in this case.

The purpose of the rest of this chapter is to show that a convex relaxation of problem P0 can be solved to obtain a solution for P0. The relaxation is motivated by geometric insight. In two dimensions, the annulus of Figure 40 can be lifted to a convex region by introducing a third dimension and extending the annulus in this direction. See Figure 41.

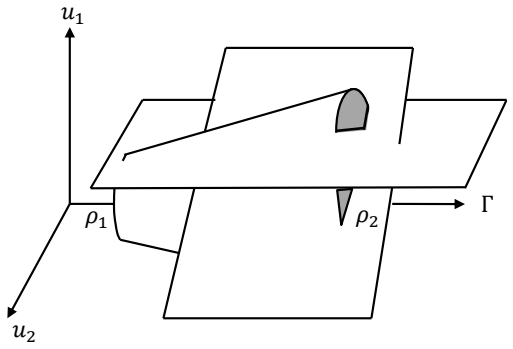

Figure 41: Three-dimensional constraint.

This solid cone shape extends from ρ_1 to ρ_2 along the Γ axis. The two planes that intersect the cone are the two linear inequality constraints that have also been extended in the Γ direction. The set of admissible controls is now all points in the cone that are also below the planes. This particular relaxation has introduced controls that may be inadmissible in P0 since points inside the cone are not necessarily admissible. For example, the point $(u_1(t), u_2(t), \Gamma(t)) = (0, 0, \rho_2)$ is admissible in the relaxation although $(u_1(t), u_2(t)) = (0, 0)$ is not admissible in P0. In lossless convexification, it is shown that this cannot occur.

With this motivation, the relaxed convex problem is stated below as P1.

$$\begin{aligned} \min \quad & J = m(t_f, x(t_f)) + \textstyle\int_{t_0}^{t_f} \ell(\Gamma(t))\, dt \\ \text{subj. to} \quad & \dot{x}(t) = Ax(t) + Bu(t) + Ew(t), \quad x(t_0) = x_0 \\ & 0 < \rho_1 \leq \Gamma(t) \leq \rho_2, \qquad \kappa(u(t)) \leq \Gamma(t) \\ & Cu(t) \leq d, \quad x(t) \in \mathcal{X}, \quad b(t_f, x(t_f)) = 0 \end{aligned} \tag{P1}$$

Under certain conditions, solutions of P1 are also solutions of P0. Thus, the convex problem can be solved instead of the non-convex problem. These sufficient conditions are stated below in Assumption 1.

Assumption 1. *i) P0 is time-invariant and there exist friends F and G such that $\mathcal{X}$ is the strongly controllable subspace for the linear system $(A + BF, BG, CF, CG)$; or, ii) P0 is time-varying, $\mathcal{X}$ is A-invariant, the matrix*

$$\begin{bmatrix} \partial_x m(t_f, x(t_f)) & \partial_x b(t_f, x(t_f)) \\ \partial_t m(t_f, x(t_f)) + \ell(\Gamma(t_f)) & \partial_t b(t_f, x(t_f)) \end{bmatrix} \tag{1}$$

is full column rank, and there exists a friend G such that $\mathcal{X}$ is the strongly controllable subspace for the linear system $(A, BG, 0, CG)$.

The problem is considered time-invariant if time does not explicitly appear in any of the functions defining the problem. It is time-variant otherwise. All of the conditions in Assumption 1 can be checked a priori, and they are natural extensions of the standard controllability conditions required for convexification of problems without state constraints [10].

Physically, the conditions require that the dynamic system be controllable on the restricted subspace – not the entire state space – meaning the system can transfer between any two points in the subspace without leaving the subspace. The conditions are satisfied for many applications since controllability is designed into the system for practical reasons. The notions of a "friend" of a linear system and a strongly controllable subspace are introduced in the next section.

The full rank condition in Assumption 1 is a strengthening of the non-triviality condition. To see this, we expand the transversality conditions in Theorem 2 of Chapter III. Rearranging and recognizing that there are no inequality constraints at the final point gives

$$\begin{bmatrix} p(t_f) \\ -p(t_f)^T f[t_f] \end{bmatrix} = \begin{bmatrix} \partial_x m(t_f, x(t_f)) & \partial_x b(t_f, x(t_f)) \\ \partial_t m(t_f, x(t_f)) + \ell(\Gamma(t_f)) & \partial_t b(t_f, x(t_f)) \end{bmatrix} \begin{bmatrix} p_0 \\ \xi \end{bmatrix} \tag{2}$$

If $p(t_f) = 0$, then $p_0 = 0$. This violates the non-triviality condition. Thus, under the full rank assumption, $p(t_f) \neq 0$, which indicates that the separating hyperplane cannot be horizontal. It must be tilted. Refer to Chapter III for geometric illustrations of the adjoint vector and the separating hyperplane.

Given Assumption 1, the formal statement for lossless convexification of P0 to P1 is Theorem 1. The proof is the goal of the remainder of this chapter.

Theorem 1. *If Assumption 1 is satisfied, then optimal solutions of P1 are optimal solutions of P0.*

B. Linear System Theory

This section presents important concepts from linear systems theory including those in Assumption 1. Only those concepts that are critical to the proof of lossless convexification are covered in this section.

The first task is to establish the meaning of a "friend". This is done by considering the following linear system, which is the same as that in problems P0 and P1.

$$\dot{x}(t) = Ax(t) + Bu(t) + Ew(t), \quad x(t_0) = x_0 \tag{3}$$

As in those problems, the state $x(t)$ is restricted to evolve in a subspace $\mathcal{X}$ of dimension $k_x \leq n$. By simple extensions of [64, p. 82-85], it can be shown that the state evolves in $\mathcal{X}$ if and only if the control has the form

$$u(t) = Fx(t) + Gv(t) + Hw(t) \tag{4}$$

where $v(t)$ is a new control variable. The matrices F, G, and H are the so-called friends, and they must belong to the following sets.

$$\begin{aligned} \mathcal{F}(\mathcal{X}) &:= \{F : (A + BF)\mathcal{X} \subset \mathcal{X}\} \\ \mathcal{G}(\mathcal{X}) &:= \{G : \text{im } BG \subset \mathcal{X}\} \\ \mathcal{H}(\mathcal{X}) &:= \{H : (E + BH)\mathcal{W} \subset \mathcal{X}\} \end{aligned} \tag{5}$$

The second task is to establish the meaning of strongly controllable and strongly observable subspaces. To do so, we must introduce the standard linear

system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{6}$$

Note that we are reusing some of these variables. No connection with P0 or P1 is implied at this point. Together with Equation 3, this linear system is denoted $\Sigma = (A, B, C, D)$. By definition, an initial point is a strongly controllable point if the origin is instantaneously reachable by an impulsive input. The set of all such points forms the strongly controllable subspace, denoted $\mathcal{C}(\Sigma)$. If the strongly controllable subspace is the entire space, then the system is said to be strongly controllable. Strong observability is explicitly used later so it is formally defined [64, p. 164].

Definition 1. *The system Σ is strongly observable if for every initial condition $x(t_0)$ and every control input $u(\cdot)$, $y(t) = 0\ \forall t \geq t_0$ implies that $x(t_0) = 0$.*

It is important to know that Σ being strongly controllable is equivalent to $\Sigma^T = (A^T, C^T, B^T, D^T)$ being strongly observable [64, p. 192].

We also have the following theorem, which sometimes serves as a definition of the strongly controllable subspace.

Theorem 2. *$\mathcal{C}(\Sigma)$ is the smallest subspace $\mathcal{V}$ for which there exists a linear map K such that*

$$(A + KC)\mathcal{V} \subset \mathcal{V} \quad \text{and} \quad \text{im}(B + KD) \subset \mathcal{V} \tag{7}$$

Proof. See Trentelman et al. [64, p. 185]. □

This theorem leads us to a very important fact used later to prove convexification. Suppose that $A\mathcal{C}(\Sigma) \subset \mathcal{C}(\Sigma)$, i.e., $\mathcal{C}(\Sigma)$ is A-invariant, $\mathrm{im}B \subset \mathcal{C}(\Sigma)$, and let T be a basis adapted to $\mathcal{C}(\Sigma)$. Then the differential equations for Σ can be written as

$$\begin{bmatrix} \dot{\zeta}(t) \\ \dot{\sigma}(t) \end{bmatrix} = \begin{bmatrix} \tilde{A} & * \\ 0 & * \end{bmatrix} \begin{bmatrix} \zeta(t) \\ \sigma(t) \end{bmatrix} + \begin{bmatrix} \tilde{B} \\ 0 \end{bmatrix} u(t) \tag{8}$$

where the star quantities are possibly non-zero and $\sigma(t) = 0 \; \forall t$ if and only if $x(t) \in \mathcal{C}(\Sigma) \; \forall t$. Likewise, the output equation can be written as

$$y(t) = [\tilde{C} \; * \;] \begin{bmatrix} \zeta(t) \\ \sigma(t) \end{bmatrix} + \tilde{D}u(t) \tag{9}$$

The system of matrices with tilde and starred quantities is denoted $\bar{\Sigma} = (\bar{A}, \bar{B}, \bar{C}, \bar{D})$, and the system with only tilde quantities is similarly denoted $\tilde{\Sigma} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$. With this notation, we can state the following theorem.

Theorem 3. *If* $A\mathcal{C}(\Sigma) \subset \mathcal{C}(\Sigma)$ *and* $\mathrm{im}B \subset \mathcal{C}(\Sigma)$*, then* $\tilde{\Sigma}$ *is strongly controllable.*

Proof. Suppose that $A\mathcal{C}(\Sigma) \subset \mathcal{C}(\Sigma)$ and $\mathrm{im}B \subset \mathcal{C}(\Sigma)$. By Theorem 2, there exists a matrix K such that

$$(A + KC)\mathcal{C}(\Sigma) \subset \mathcal{C}(\Sigma) \text{ and } \mathrm{im}(B + KD) \subset \mathcal{C}(\Sigma) \tag{10}$$

Thus, for every $x \in \mathcal{C}(\Sigma)$ and every $u \in \mathbb{R}^m$ there exist $p, q \in \mathcal{C}(\Sigma)$ such that

$$Ax + KCx = p \quad \text{and} \quad Bu + KDu = q \tag{11}$$

By using basis T, it follows that for every $\zeta \in \mathbb{R}^{k_c}$ and every $u \in \mathbb{R}^m$ there exist $\tilde{p}, \tilde{q} \in \mathbb{R}^{k_c}$, where k_c is the dimension of $\mathcal{C}(\Sigma)$, such that

$$\tilde{A}\zeta + \tilde{K}\tilde{C}\zeta = \tilde{p} \quad \text{and} \quad \tilde{B}u + \tilde{K}\tilde{D}u = \tilde{q} \tag{12}$$

Therefore, when $\mathcal{V} = \mathbb{R}^{k_c}$, there exists a linear map $\tilde{K}$ satisfying Equation 7. It remains to be shown that $\mathbb{R}^{k_c}$ is the smallest such $\mathcal{V}$. Suppose it is not such that there exists a subspace $\mathcal{L}$ with dimension less than k_c and a $\hat{K}$ satisfying

$$(\tilde{A} + \hat{K}\tilde{C})\mathcal{L} \subset \mathcal{L} \quad \text{and} \quad \text{im}(\tilde{B} + \hat{K}\tilde{D}) \subset \mathcal{L} \tag{13}$$

Let $\mathcal{M} \subset \mathcal{C}(\Sigma)$ be a subspace of $\mathbb{R}^n$ with the same dimension as $\mathcal{L}$, and let M be a basis for $\mathbb{R}^n$ adapted to the chain $\mathcal{M}, \mathcal{C}(\Sigma)$ such that $M = T$. By similar arguments as before, it can be shown that for every $x \in \mathcal{M}$ and every $u \in \mathbb{R}^m$ there exist $p, q \in \mathcal{M}$ such that

$$Ax + KCx = p \quad \text{and} \quad Bu + KDu = q. \tag{14}$$

Thus, there exists a linear map K for $\mathcal{M} \subset \mathcal{C}(\Sigma)$ satisfying Equation 7, which contradicts the definition of $\mathcal{C}(\Sigma)$. Thus, $\tilde{\Sigma}$ is strongly controllable. $\square$

C. Lossless Convexification

The purpose of this section is to prove Theorem 1, which says that optimal solutions of P1 are optimal solutions of P0. This is the main result of the chapter. To complete the proof, two more problems, P2 and P3, will be considered. The strategy here is to prove a property about optimal solutions of P3, link the solutions of P0-P3, and then show that this implies convexification between P0 and P1. The problems P2 and P3 are used only in the proofs. They have no role in the numerical solution process.

The problem P2 is obtained from P1 by replacing the state constraint $x(t) \in \mathcal{X}$ with the constraint specified in Equation 4 and repeated here.

$$u[t] = Fx(t) + Gv(t) + Hw(t) \tag{15}$$

The bracket notation is used to emphasize that the control is no longer independent. It is a function of other variables. The closed loop system is

$$\dot{x}(t) = A_F x(t) + B_G v(t) + E_H w(t) \tag{16}$$

where $A_F = A + BF$, $B_G = BG$, and $E_H = E + BH$. Problem P2 is

$$\begin{aligned}
\min \quad & J = m(t_f, x(t_f)) + \textstyle\int_{t_0}^{t_f} \ell(\Gamma(t))\, dt && \text{(P2)} \\
\text{subj. to} \quad & \dot{x}(t) = A_F x(t) + B_G v(t) + E_H w(t), \quad x(t_0) = x_0 \\
& \kappa(u[t]) \leq \Gamma(t), \quad 0 < \rho_1 \leq \Gamma(t) \leq \rho_2 \\
& Cu[t] \leq d, \qquad\quad b(t_f, x(t_f)) = 0
\end{aligned}$$

The final problem, labeled P3, is obtained by reducing the state space to $\mathcal{X}$ by using the basis as in Equation 8. In that case, the evolution of the system on $\mathcal{X}$ can be expressed as follows

$$\dot{\zeta}(t) = \tilde{A}_F \zeta(t) + \tilde{B}_G v(t) + \tilde{E}_H w(t) \tag{17}$$

Additionally, the control constraint can be expressed as

$$\tilde{u}[t] = \tilde{F}\zeta(t) + Gv(t) + Hw(t) \tag{18}$$

Again, the bracket notation indicates that $\tilde{u}$ is dependent on other variables. Similar adaptations hold for the terminal cost and terminal constraint since they are affine by assumption. By using the new state variable ζ, these are written as $\tilde{m}(t_f, \zeta(t_f))$ and $\tilde{b}(t_f, \zeta(t_f))$, respectively. The final problem is

$$\begin{aligned}
\min \quad & J = \tilde{m}(t_f, \zeta(t_f)) + \textstyle\int_{t_0}^{t_f} \ell(\Gamma(t))\, dt \\
\text{subj. to} \quad & \dot{\zeta}(t) = \tilde{A}_F \zeta(t) + \tilde{B}_G v(t) + \tilde{E}_H w(t), \quad \zeta(t_0) = \zeta_0 \\
& \kappa(\tilde{u}[t]) \leq \Gamma(t), \quad 0 < \rho_1 \leq \Gamma(t) \leq \rho_2 \\
& C\tilde{u}[t] \leq d, \qquad \tilde{b}(t_f, \zeta(t_f)) = 0
\end{aligned} \tag{P3}$$

This problem fits the appropriate form so that Theorem 2 of Chapter III applies. Results connecting problems P0 through P3 are now given in a sequence of lemmas leading to the main result. Then Theorem 1, the main theorem of the paper, is proved. Note that the control variables in P2 and P3 are $v(\cdot)$ and $\Gamma(\cdot)$. The original control $u(\cdot)$ no longer appears.

Lemma 1 connects the feasible sets for the original problem P0 and its convex relaxation P1.

Lemma 1. *i) If* $u(\cdot) \in \mathcal{F}_0$*, then there exists a* $\Gamma(\cdot)$ *such that* $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1$*. ii) If* $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1$ *and* $\kappa(u(t)) = \Gamma(t)\ \forall t$*, then* $u(\cdot) \in \mathcal{F}_0$*.*

Proof. i) Suppose $u(\cdot) \in \mathcal{F}_0$ and define $\Gamma(t) = \kappa(u(t))\ \forall t$. Then $\rho_1 \leq \Gamma(t) \leq \rho_2\ \forall t$ such that all constraints of P1 are satisfied and $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1$. ii) Since $\rho_1 \leq \kappa(u(t)) \leq \rho_2\ \forall t$, all constraints of P0 are satisfied and $u(\cdot) \in \mathcal{F}_0$. □

In a similar fashion, Lemmas 2 and 3 connect the optimal solutions of problems P1, P2, and P3.

Lemma 2. $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1^*$ *if and only if there exists a* $v(\cdot)$ *such that the pair* $(v(\cdot), \Gamma(\cdot)) \in \mathcal{F}_2^*$*.*

Proof. The result follows from the facts that 1) the performance indices for P1 and P2 are the same and 2) $x(t) \in \mathcal{X}$ if and only if there exists a $v(t)$ such that $u(t) = Fx(t) + Gv(t) + Hw(t)$ [64, p. 82-85]. □

Lemma 3. $(v(\cdot), \Gamma(\cdot)) \in \mathcal{F}_2^*$ *if and only if* $(v(\cdot), \Gamma(\cdot)) \in \mathcal{F}_3^*$*.*

Proof. The result follows from the fact that the only difference between the problems is a coordinate adaptation (see Equation 8 and the related discussion). Thus, the cost is the same and the constraints are equivalent in the new space. □

We are now finally to the point of synthesizing problem P3 using the optimality conditions given in Theorem 2 of Chapter III. The goal is to characterize the optimal solutions with a link back to problems P0 and P1 so that

lossless convexification can be proved. To do so, we define the Hamiltonian, Lagrangian, and endpoint functions specifically for P3.

$$\mathcal{H}[t] = p_0\ell(\Gamma(t)) + p(t)^T(\tilde{A}_F\zeta(t) + \tilde{B}_G v(t) + \tilde{E}_H w(t)) \tag{19}$$

$$\begin{aligned}\mathcal{L}[t] = \mathcal{H}[t] &+ \lambda_1(t)(\kappa(\tilde{u}[t]) - \Gamma(t)) + \lambda_2(t)(\rho_1 - \Gamma(t)) \\ &+ \lambda_3(t)(\Gamma(t) - \rho_2) + \lambda_4(t)^T(C\tilde{u}[t] - d)\end{aligned} \tag{20}$$

$$\mathcal{G}[t_f] = p_0\tilde{m}(t_f, \zeta(t_f)) + \xi^T\tilde{b}(t_f, \zeta(t_f)) \tag{21}$$

The adjoint differential equation is

$$-\dot{p}(t) = \tilde{A}_F^T p(t) + \tilde{F}^T C^T \lambda_4(t) \tag{22}$$

The stationary conditions are

$$\begin{aligned}\partial_\Gamma\mathcal{L}[t] &= p_0\partial_\Gamma\ell(\Gamma(t)) - \lambda_1(t) - \lambda_2(t) + \lambda_3(t) = 0 \\ \partial_v\mathcal{L}[t] &= \tilde{B}_G^T p(t) + \lambda_1(t)\partial_v\kappa(\tilde{u}[t]) + G^T C^T\lambda_4(t) = 0\end{aligned} \tag{23}$$

The complementary slackness conditions are

$$\begin{array}{lll}\kappa(\tilde{u}[t]) - \Gamma(t) \le 0, & \lambda_1(t) \le 0, & \lambda_1(t)(\kappa(\tilde{u}[t]) - \Gamma(t)) = 0 \\ \rho_1 - \Gamma(t) \le 0, & \lambda_2(t) \le 0, & \lambda_2(t)(\rho_1 - \Gamma(t)) = 0 \\ \Gamma(t) - \rho_2 \le 0, & \lambda_3(t) \le 0, & \lambda_3(t)(\Gamma(t) - \rho_2) = 0 \\ C\tilde{u}[t] - d \le 0, & \lambda_4(t) \le 0, & \lambda_4(t)^T(C\tilde{u}[t] - d) = 0\end{array} \tag{24}$$

Lemma 4. *If Assumption 1 is satisfied, then* $(v(\cdot), \Gamma(\cdot)) \in \mathcal{F}_3^*$ *implies* $\kappa(\tilde{u}[t]) = \Gamma(t)\ \forall t$.

Proof. The proof is handled in two cases. *Case 1* is for the time-invariant problem, and *Case 2* is for the time-varying problem.

Case 1. Suppose that $(v(\cdot), \Gamma(\cdot)) \in \mathcal{F}_3^*$ and there exists a t such that $\kappa(\tilde{u}[t]) < \Gamma(t)$. Because $\tilde{u}[\cdot]$ and $\Gamma(\cdot)$ are piecewise continuous in time, and $\kappa(\cdot)$ is differentiable, there exists an interval $[\tau_1, \tau_2] \subset [t_0, t_f]$ where $\kappa(\tilde{u}[t]) < \Gamma(t)$ for all $t \in [\tau_1, \tau_2]$. Equation 24 implies that $\lambda_1(t) = 0$ on this interval. Equations 22 and 23 become

$$\begin{aligned} -\dot{p} &= \tilde{A}_F^T p(t) + \tilde{F}^T C^T \lambda_4(t) \\ 0 &= \tilde{B}_G^T p(t) + G^T C^T \lambda_4(t) \end{aligned} \tag{25}$$

By part i) of Assumption 1, $\mathcal{X}$ is the strongly controllable subspace for $(A + BF, BG, CF, CG)$. Theorem 3 implies that the system $(\tilde{A}_F, \tilde{B}_G, C\tilde{F}, CG)$ is strongly controllable, i.e., $(\tilde{A}_F^T, \tilde{F}^T C^T, \tilde{B}_G^T, G^T C^T)$ is strongly observable. Strong observability implies $p(\tau_1) = 0$. Because the problem is time-invariant, the Hamiltonian is identically zero. This means that $p_0 = 0$ since $\ell(\Gamma(\tau_1))$ cannot be zero. This violates the non-triviality condition. Thus, $\kappa(\tilde{u}[t]) = \Gamma(t)\ \forall t$.

Case 2. By part ii) of Assumption 1, $\mathcal{X}$ is A-invariant such that $F = 0$ is a friend. Thus, Equation 25 becomes

$$\begin{aligned} -\dot{p} &= \tilde{A}^T p(t) \\ 0 &= \tilde{B}_G^T p(t) + G^T C^T \lambda_4(t) \end{aligned} \tag{26}$$

Again, Theorem 3 implies that the system $(\tilde{A}, \tilde{B}_G, 0, CG)$ is strongly con-

trollable, i.e., $(\tilde{A}^T, 0, \tilde{B}_G^T, G^T C^T)$ is strongly observable. Strong observability implies $p(\tau_1) = 0$. Since $p(\cdot)$ is the solution of a homogeneous equation, $p(t_f) = 0$. It follows from the full rank condition in Assumption 1 that $(p_0, p(t_f)) = 0$, which violates the non-triviality condition (see Equation 2). Thus, $\kappa(\tilde{u}(t)) = \Gamma(t)\ \forall t$. □

Lemma 5. *If Assumption 1 is satisfied, then the pair* $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1^*$ *implies* $\kappa(u(t)) = \Gamma(t)\ \forall t$.

Proof. Suppose $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1^*$ and there exists a t such that $\kappa(u(t)) < \Gamma(t)$. Then, from Lemmas 2 and 3, there exists a $v(\cdot)$ such that $(v(\cdot), \Gamma(\cdot)) \in \mathcal{F}_3^*$ with $\kappa(\tilde{u}[t]) < \Gamma(t)$. This contradicts Lemma 4. □

The following theorem establishes lossless convexification between P0 and P1. It is a main result of the paper and says that optimal solutions of P1 are also optimal solutions of P0. It is the same as Theorem 1.

Theorem 4. *If Assumption 1 is satisfied, then the pair* $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1^*$ *implies* $u(\cdot) \in \mathcal{F}_0^*$.

Proof. Suppose that $(u(\cdot), \Gamma(\cdot)) \in \mathcal{F}_1^*$. Lemma 5 implies that $\kappa(u(t)) = \Gamma(t)\ \forall t$. Consequently, problems P0 and P1 have the same cost function. Lemma 1 implies that $u(\cdot) \in \mathcal{F}_0$. Thus, $J_0^* \leq J_1^*$. Similarly, Lemma 1 also implies that $J_1^* \leq J_0^*$. Thus, $J_0^* = J_1^*$. Because the cost functions are the same, the cost of $u(\cdot)$ in P0 is $J_1^* = J_0^*$. Thus, $u(\cdot) \in \mathcal{F}_0^*$. □

We now use this result to solve two example problems from the aerospace engineering field.

D. Example 1: Minimum Fuel Planetary Landing

The first example is the planetary landing problem. In the final descent phase, it is assumed that 1) the vehicle is close enough to the surface that gravity is constant, 2) the thrust forces dominate the aerodynamic forces, and 3) a known time-varying disturbance acts on the system. In this case, the equations of motion are

$$\ddot{x}(t) = -g + u(t) + w(t) \tag{27}$$

The first component of x, denoted x_1, is the range. The altitude and cross range are x_2 and x_3, respectively. The components x_4, x_5, and x_6 are the range rate, altitude rate, and cross range rate. Near the surface of Mars, the gravity vector is approximately $g = [0 \ -3.71 \ \ 0]$ m/s^2. It is assumed that the disturbance is a sinusoidal function of time of the form $w(t) = [\sin(t) \ 0 \ \cos(t)]$ m/s^2. The problem is to transfer the vehicle from its initial condition to the landing site with zero final velocity, e.g.,

$$\begin{aligned} x(0) &= [400 \ 400 \ 300] \text{ m}, \quad x(t_f) = [0 \ 0 \ 0] \text{ m} \\ \dot{x}(0) &= -[10 \ 10 \ 75] \text{ m/s}, \quad \dot{x}(t_f) = [0 \ 0 \ 0] \text{ m/s} \end{aligned} \tag{28}$$

For safety reasons, it is also required that the vehicle not approach the landing site with too steep or too shallow an approach angle. A 45 degree approach in the altitude/range plane is specified and can be written as

$$x_1(t) - x_2(t) = 0 \tag{29}$$

The control magnitude is bounded above and below since the thrusters cannot operate reliably below this bound.

$$2 \leq ||u(t)|| \leq 10 \text{ m/s}^2 \tag{30}$$

The goal is to achieve the landing and minimize the fuel consumption, i.e.,

$$\min J = \int_0^{t_f} ||u(t)|| \, dt \tag{31}$$

The optimal control problem is defined by Equations 27 through 31 and fits within the structure of P0 given in Equation P0. Note that no linear control constraints are present, i.e., no C. The convex relaxation is easily obtained by relaxing Equation 30 to the two constraints

$$||u(t)|| \leq \Gamma(t) \quad \text{and} \quad 2 \leq \Gamma(t) \leq 10 \tag{32}$$

and minimizing $J = \int_0^{t_f} \Gamma(t) \, dt$. For the state to evolve on the plane $x_1(t) - x_2(t) = 0$, the time derivative of the constraint must also be zero, i.e., $x_4(t) - x_5(t) = 0$. Thus, the subspace $\mathcal{X}$ is given as

$$\mathcal{X} = \text{null} \left(\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} \right) \tag{33}$$

The friends F, G, and H are

$$F = 0, \quad G = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{34}$$

The fact that these are in fact friends and that $\mathcal{X}$ is the strongly controllable subspace for the system $(A+BF, BG, CF, CG)$ can be checked using the tests in Trentelman et al. [64, p. 182-188].

The numerical simulations are carried out using SDPT3 [17]. Figure 42 shows the state trajectory. The trajectory begins at the top center, ends at the origin in the bottom right, and evolves on the 45 degree plane.

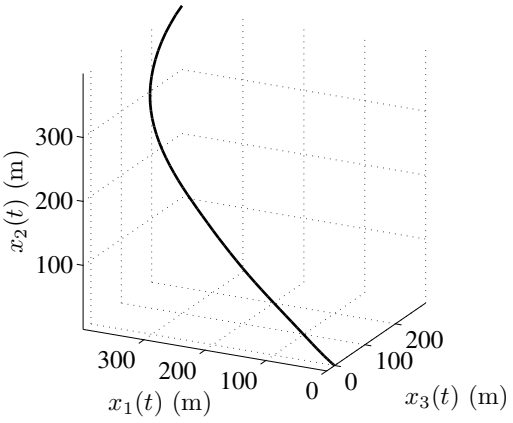

Figure 42: State trajectory for constrained landing.

The thrust magnitude is shown in Figure 43. The upper constraint is active along the initial and final arcs, and the control is in the interior during the middle arc. The oscillations are caused by the time-varying disturbance.

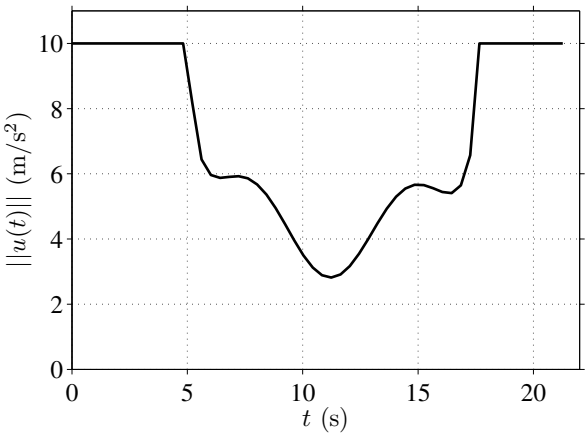

Figure 43: Control trajectory for constrained landing.

E. Example 2: Minimum Time Rendezvous

A second example is the rendezvous of two spacecraft at constant altitude using low thrust. It is assumed that the motion of the chaser spacecraft relative to the target spacecraft is accurately described by the Clohessy-Wiltshire-Hill equations [45] so the equations of motion are

$$\begin{aligned} \ddot{x}(t) &= 3\omega^2 x(t) + 2\omega\dot{y}(t) + u_1(t) \\ \ddot{y}(t) &= -2\omega\dot{x}(t) + u_2(t) \\ \ddot{z}(t) &= -\omega^2 z(t) + u_3(t) \end{aligned} \tag{35}$$

The states x, y, and z are the altitude, range, and cross range, respectively, and $\dot{x}$, $\dot{y}$, $\dot{z}$ are the rates. The orbital mean motion is $\omega = 4$ hr^{-1}, and corresponds to a near circular, low earth orbit. The problem is to rendezvous the two vehicles, i.e., bring them together with zero relative velocity. The boundary

conditions are

$$\begin{aligned} [x(0)\ y(0)\ z(0)] &= [0\ 2\ 1] \text{ km}, & [x(t_f)\ y(t_f)\ z(t_f)] &= [0\ 0\ 0] \text{ km} \\ [\dot{x}(0)\ \dot{y}(0)\ \dot{z}(0)] &= [0\ -0.5\ -0.25] \text{ km/s}, & [\dot{x}(t_f)\ \dot{y}(t_f)\ \dot{z}(t_f)] &= [0\ 0\ 0] \text{ km/s} \end{aligned} \tag{36}$$

For safety reasons, viewing angles, etc., it is required that the rendezvous maneuver take place at constant altitude. This constraint is written simply as $x_1(t) = 0$. As in the previous example, the thrust magnitude is bounded above and below

$$3 \leq ||u(t)|| \leq 5 \text{ km/hr}^2 \tag{37}$$

It is also required that the chaser spacecraft not point in the cross range direction more than $\theta = 45$ degrees.

$$u_3(t) - u_2(t) \tan\theta \leq 0 \tag{38}$$

The goal is to achieve the rendezvous and minimize the time of flight so that the cost function is simply $J = \int_{t_0}^{t_f} 1\ dt$. This optimal control problem also fits within the structure of P0, and the convex relaxation of the control constraints is obtained by relaxing the control constraint to

$$||u(t)|| \leq \Gamma(t) \quad \text{and} \quad 3 \leq \Gamma(t) \leq 5 \tag{39}$$

and minimizing the same cost $J = \int_{t_0}^{t_f} 1\ dt$.

For the state to evolve on the plane $x(t) = 0$, the time derivative of the constraint must also be zero, i.e., $\dot{x}(t) = 0$. Thus, the subspace $\mathcal{X}$ is given as

$$\mathcal{X} = \text{null}\left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}\right) \tag{40}$$

The friends F, G, and H are

$$F = \begin{bmatrix} 1 & 0 & 0 & 1 & -8 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 9 & 1 & 1 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{41}$$

Figure 44 shows the range and cross range. The trajectory begins in the upper right, oscillates, and terminates at the origin on the left.

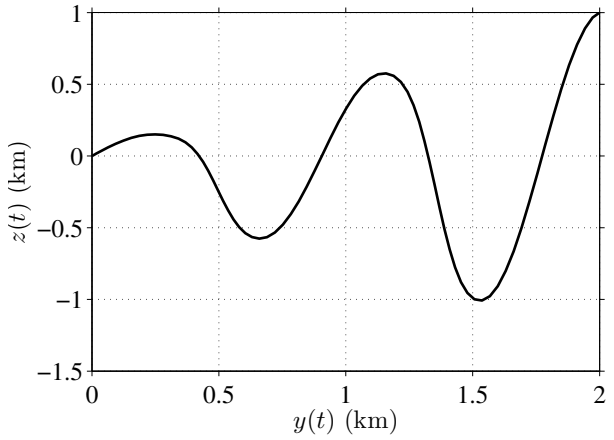

Figure 44: State trajectory for constant altitude rendezvous.

The thrust angle is shown in Figure 45, and it is evident that the angle satisfies the point constraint of 45 degrees.

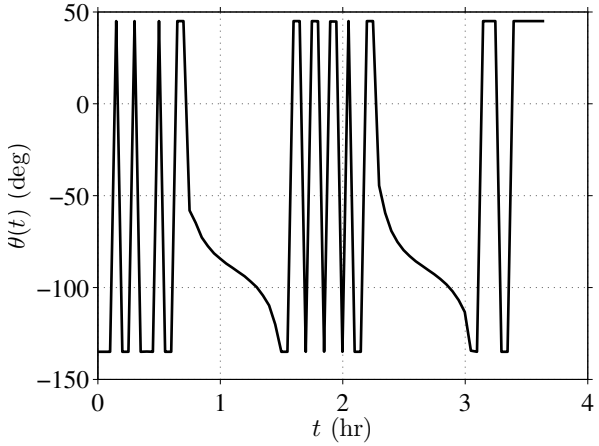

Figure 45: Control trajectory for constant altitude rendezvous.

The altitude plot is not shown since it is identically zero. The thrust magnitude is not shown, but it is constant along the upper boundary.

F. Summary and Conclusions

In this chapter, the theory of lossless convexification was generalized to optimal control problems with mixed non-convex and convex control constraints and linear state constraints. This was done by introducing the strongly controllable subspaces and studying a sequence of convex relaxations of the original problem. As a consequence, a larger class of non-convex problems can be solved as a convex problem. The work is significant because this class of problems includes several important practical applications.

CHAPTER VII:
FINAL REMARKS

Significant work on lossless convexification in the last seven years has culminated in numerous publications [8–15] and two successful flight tests [6, 7]. Results have progressed from specific applications to general theoretical results with state constraints and mixed control constraints. However, there are many avenues for continued research.

The first is to continue exploring state constraints. The analysis of Chapter VI is laborious because it depends on certain coordinate representations. Recent work by Sussmann indicates a very natural description of optimal control on manifolds in a coordinate free way [66]. It is possible that the proof of lossless convexification will be much more direct in this new setting. It may also open the door to stronger results since one can more easily identify system properties.

Another avenue is to forgo optimal control problems and prove lossless convexification directly with the finite-dimensional optimization problem. This is desirable since the finite-dimensional problem is the one solved in the end. However, recent attempts to do so have not been fruitful. The reason is that the finite-dimensional problem is essentially a fixed final time problem. Most convexification results are with free final time wherein the Hamiltonian is zero everywhere. Future researchers should look for a constraint qualification or a critical final time for which convexification works with any lesser final time.

Lossless convexification is also important in the design of Markov decision processes [67,68]. Through convexification, it is possible to design probabilistic and decentralized control strategies for large, multi-agent systems using convex optimization. Applications include meteorology, oceanography, and aerospace engineering. Work in this area has already begun, but considerable progress can still be made especially with applications.

Finally, and most importantly, lossless convexification and the resulting algorithms must be pushed into industry. Management is not always interested in mathematical eloquence. Flight tests and numerical experiments are the best way to convince skeptics. Any opportunity for either should be accepted. Because this process has already started [6, 7], the future is promising.

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[2] L. Vandenberghe and S. Boyd, "Semidefinite Programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[3] J. Mattingley and S. Boyd, "CVXGEN: A Code Generator for Embedded Convex Optimization," *Optimization and Engineering*, vol. 13, no. 1, pp. 1–27, 2012.

[4] E. Chu, N. Parikh, A. Domahidi, and S. Boyd, "Code Generation for Embedded Second-Order Cone Programming," in *European Control Conference*, (Zurich, Switzerland), July 2013.

[5] A. Domahidi, E. Chu, and S. Boyd, "ECOS: An SOCP Solver for Embedded Systems," in *European Control Conference*, (Zurich, Switzerland), July 2013.

[6] B. Açıkmeşe, M. Aung, J. Casoliva, S. Mohan, A. Johnson, D. Scharf, D. Masten, J. Scotkin, A. Wolf, and M. W. Regehr, "Flight Testing of Trajectories Computed by G-FOLD: Fuel Optimal Large Divert Guidance Algorithm for Planetary Landing," in *AAS/AIAA Spaceflight Mechanics Meeting*, (Kauai, Hawaii), February 2013.

[7] D. P. Scharf, M. W. Regehr, D. Dueri, B. Açıkmeşe, G. M. Vaughan, J. Benito, H. Ansari, M. Aung, A. Johnson, D. Masten, S. Nietfeld, J. Casoliva, and S. Mohan, "ADAPT: Demonstrations of Onboard Large-Divert Guidance with a Reusable Launch Vehicle," in *IEEE Aerospace Conference*, (Big Sky, Montana), March 2014.

[8] B. Açıkmeşe and S. R. Ploen, "Convex Programming Approach to Powered Descent Guidance for Mars Landing," *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 5, pp. 1353–1366, 2007.

[9] L. Blackmore, B. Açıkmeşe, and D. P. Scharf, "Minimum Landing Error Powered Descent Guidance for Mars Landing using Convex Optimization," *Journal of Guidance, Control, and Dynamics*, vol. 33, July-August 2010.

[10] B. Açıkmeşe and L. Blackmore, "Lossless Convexification for a Class of Optimal Control Problems with Nonconvex Control Constraints," *Automatica*, vol. 47, no. 2, pp. 341–347, 2011.

[11] L. Blackmore, B. Açıkmeşe, and J. M. Carson, "Lossless Convexification of Control Constraints for a Class of Nonlinear Optimal Control Problems," *Systems and Control Letters*, vol. 61, pp. 863–371, 2012.

[12] B. Açıkmeşe, J. M. Carson, and L. Blackmore, "Lossless Convexification of Nonconvex Control Bound and Pointing Constraints of the Soft Landing Optimal Control Problem," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 6, pp. 2104–2113, 2013.

[13] M. W. Harris and B. Açıkmeşe, "Minimum Time Rendezvous of Multiple Spacecraft Using Differential Drag," *Journal of Guidance, Control, and Dynamics*, 2014. Accepted.

[14] M. W. Harris and B. Açıkmeşe, "Maximum Divert for Planetary Landing Using Convex Optimization," *Journal of Optimization Theory and Applications*, 2014. Accepted.

[15] M. W. Harris and B. Açıkmeşe, "Lossless Convexification of Non-Convex Optimal Control Problems for State Constrained Linear Systems," *Automatica*, 2014. Accepted.

[16] D. G. Hull, "Conversion of Optimal Control Problems into Parameter Optimization Problems," *Journal of Guidance, Control, and Dynamics*, vol. 20, no. 1, pp. 57–60, 1997.

[17] K. C. Toh, M. J. Todd, and R. H. Tutuncu, "SDPT3 – a Matlab Software Package for Semidefinite Programming," *Optimization Methods and Software*, vol. 11, no. 1, pp. 545–581, 1999.

[18] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB Toolbox for Optimization over Symmetric Cones," *Optimization Methods and Software*, vol. 17, no. 6, pp. 1105–1154, 2002.

[19] Gurobi Optimization, Inc., *Gurobi Optimization Reference Manual*, 2012.

[20] M. Grant and S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming, Version 2.0*, 2012.

[21] J. Peng, C. Roos, and T. Terlaky, *Self-Regularity: A New Paradigm for Primal-Dual Interior-Point Algorithms*. Princeton Series in Applied Mathematics, 2001.

[22] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Methods in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[23] G. B. Dantzig, *Linear Programming and Extensions*. Princeton University Press, 1963.

[24] L. D. Berkovitz, *Convexity and Optimization in* $\mathbb{R}^n$. Wiley-Interscience, 2001.

[25] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*. Gordon and Breach Science Publishers, 1986. Originally published by Interscience Publishers in 1962.

[26] E. J. McShane, "On Multipliers for Lagrange Problems," *American Journal of Mathematics*, vol. 61, pp. 809–819, October 1939.

[27] H. J. Kelley, "A Second Variation Test for Singular Extremals," *AIAA Journal*, vol. 2, pp. 1380–1382, August 1964.

[28] R. E. Kopp and H. G. Moyer, "Necessary Conditions for Singular Extremals," *AIAA Journal*, vol. 3, pp. 1439–1444, August 1965.

[29] B. S. Goh, "Necessary Conditions for Singular Extremals involving Multiple Control Variables," *SIAM Journal on Control*, vol. 4, no. 4, pp. 716–731, 1966.

[30] H. M. Robbins, "A Generalized Legendre-Clebsch Condition for the Singular Cases of Optimal Control," *IBM Journal*, pp. 361–372, July 1967.

[31] A. J. Krener, "The High Order Maximal Principle and its Application to Singular Extremals," *SIAM Journal on Control and Optimization*, vol. 15, pp. 256–293, February 1977.

[32] L. D. Berkovitz, *Optimal Control Theory*. Springer-Verlag, 1975.

[33] D. Liberzon, *Calculus of Variations and Optimal Control Theory*. Princeton University Press, 2012.

[34] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[35] A. de la Fuente, *Mathematical Methods and Models for Economists*. Cambridge University Press, 2000.

[36] W. R. Wade, *An Introduction to Analysis*. Prentice Hall, 2000.

[37] M. Athans and P. L. Falb, *Optimal Control: An Introduction to the Theory and Its Applications*. McGraw-Hill, Inc., 1966.

[38] R. F. Hartl, S. P. Sethi, and R. G. Vickson, "A Survey of the Maximum Principles for Optimal Control Problems with State Constraints," *SIAM Review*, vol. 37, pp. 181–218, June 1995.

[39] D. H. Jacobson, M. M. Lele, and J. L. Speyer, "New Necessary Conditions of Optimality for Control Problems with State Variable Inequality Constraints," *Journal of Mathematical Analysis and Applications*, vol. 35, pp. 255–284, August 1971.

[40] H. Maurer, "On Optimal Control Problems with Bounded State Variables and Control Appearing Linearly," *SIAM Journal on Control and Optimization*, vol. 15, pp. 345–362, May 1977.

[41] C. L. Leonard, W. M. Hollister, and E. V. Bergmann, "Orbital Formationkeeping with Differential Drag," *Journal of Guidance, Control, and Dynamics*, vol. 12, pp. 108–113, January-February 1989.

[42] A. DeRuiter, J. Lee, and A. Ng, "A Fault-Tolerant Magnetic Spin Stabilizing Controller for the JC2Sat-FF Mission," in *AIAA Guidance, Navigation, and Control Conference*, (Honolulu, Hawaii), August 2008.

[43] B. S. Kumar and A. Ng, "A Bang-Bang Control Approach to Maneuver Spacecraft in a Formation with Differential Drag," in *AIAA Guidance, Navigation, and Control Conference*, (Honolulu, Hawaii), August 2008.

[44] P. Lu and X. Liu, "Autonomous Trajectory Planning for Rendezvous and Proximity Operations by Conic Optimization," *Journal of Guidance, Control, and Dynamics*, vol. 36, pp. 375–389, March-April 2012.

[45] W. H. Clohessy and R. S. Wiltshire, "Terminal Guidance System for Satellite Rendezvous," *Journal of Aerospace Systems*, vol. 27, no. 9, pp. 653–658, 1960.

[46] R. Bevilacqua and M. Romano, "Rendezvous Maneuvers of Multiple Spacecraft Using Differential Drag Under J2 Perturbation," *Journal of Guidance, Control, and Dynamics*, vol. 31, pp. 1595–1607, November-December 2008.

[47] S. A. Schweighart and R. J. Sedwick, "High-Fidelity Linearized J2 Model for Satellite Formation Flight," *Journal of Guidance, Control, and Dynamics*, vol. 25, pp. 1073–1080, November-December 2002.

[48] R. Bevilacqua, J. S. Hall, and M. Romano, "Multiple Spacecraft Rendezvous Maneuvers by Differential Drag and Low Thrust Engines," *Celestial Mechanics and Dynamical Astronomy*, vol. 106, pp. 69–88, January 2010.

[49] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Dover, 1998.

[50] J. P. LaSalle, "Study of the Basic Principles Underlying the Bang-Bang Servo," Tech. Rep. GER-5518, Goodyear Aircraft Corporation, July 1953.

[51] D. G. Hull, "Thrust Programs for Minimum Propellant Consumption During Take-off and Landing Maneuvers of a Rocket Vehicle in a Vacuum," Tech. Rep. 59, Boeing Scientific Research Laboratories, Flight Sciences Laboratory, July 1962.

[52] D. G. Hull, "Thrust Programs for Minimum Propellant Consumption During the Vertical Take-off and Landing of a Rocket," *Journal of Optimization Theory and Applications*, vol. 1, pp. 53–69, July 1967.

[53] A. Miele, "Application of Green's Theorem to the Extremization of Linear Integrals," Tech. Rep. 40, Boeing Scientific Research Laboratories, Flight Sciences Laboratory, 1961.

[54] J. S. Meditch, "On the Problem of Optimal Thrust Programming For a Lunar Soft Landing," *IEEE Transactions on Automatic Control*, vol. AC-9, no. 4, pp. 477–484, 1964.

[55] D. G. Hull, "Optimal Guidance for Quasi-Planar Lunar Descent With Throttling." Presented at the 21st AAS/AIAA Space Flight Mechanics Meeting, AAS 11-169, February 2011.

[56] D. G. Hull, "Optimal Guidance for Quasi-Planar Lunar Descent With Throttling," in *Advances in the Astronautical Sciences Series*, vol. 140, (San Diego, CA), Proceedings of the 21st AAS/AIAA Spaceflight Mechanics Meeting, Univelt, Inc., 2011.

[57] D. G. Hull and M. W. Harris, "Optimal Solutions for Quasi-Planar Ascent Over a Spherical Moon," *Journal of Guidance, Control, and Dynamics*, vol. 35, pp. 1218–1223, July-August 2012.

[58] A. R. Klumpp, "Apollo Lunar Descent Guidance," *Automatica*, vol. 10, pp. 133–146, 1974.

[59] D. T. Martin, R. F. Sievers, R. M. O'Brien, and A. L. Rice, "Saturn V Guidance, Navigation, and Targeting," *Journal of Spacecraft and Rockets*, vol. 4, no. 7, pp. 891–898, 1967.

[60] D. C. Chandler and I. E. Smith, "Development of the Iterative Guidance Mode with its Applications to Various Vehicles and Missions," *Journal of Spacecraft and Rockets*, vol. 4, no. 7, pp. 898–903, 1967.

[61] R. L. McHenry, T. J. Brand, A. D. Long, B. F. Cockrell, and J. R. Thibodeau, "Space Shuttle Ascent Guidance, Navigation, and Control," *Journal of the Astronautical Sciences*, vol. 27, no. 1, pp. 1–38, 1979.

[62] U. Topcu, J. Casoliva, and K. D. Mease, "Fuel Efficient Powered Descent Guidance for Mars Landing," *Journal of Spacecraft and Rockets,*, vol. 44, no. 2, 2007.

[63] M. Grant and S. Boyd, "Graph Implementations for Nonsmooth Convex Programs," in *Recent Advances in Learning and Control* (V. Blondel, S. Boyd, and H. Kimura, eds.), Lecture Notes in Control and Information Sciences, pp. 95–110, Springer-Verlag Limited, 2008.

[64] H. L. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control Theory for Linear Systems*. Springer, 2001.

[65] B. Açıkmeşe and S. R. Ploen, "A Powered Descent Guidance Algorithm for Mars Pinpoint Landing," in *AIAA Guidance, Navigation, and Control Conference*, (San Francisco, California), August 2005.

[66] H. J. Sussmann, "Set Separation, Approximating Multicones, and the Lipschitz Maximum Principle," *Journal of Differential Equations*, vol. 243, no. 2, pp. 448–488, 2007.

[67] A. Arapostathis, R. Kumar, and S. Tangirala, "Controlled Markov Chains with Safety Upper Bound," *IEEE Transactions on Automatic Control*, vol. 48, no. 7, pp. 1230–1234, 2003.

[68] B. Açıkmeşe, N. Demir, and M. W. Harris, "Convex Necessary and Sufficient Conditions for Density Safety Constraints in Markov Chain Synthesis," *IEEE Transactions on Automatic Control*, 2014. In Review.