Title
Techniques for solving nonlinear optimal control problems

Permalink
https://escholarship.org/uc/item/4dn7d13p

Author
Han, Seung Hak

Publication Date
2015

Peer reviewed|Thesis/dissertation
Techniques for solving nonlinear optimal control problems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Engineering Sciences (Mechanical Engineering)

by

Seung Hak Han

Committee in charge:

Professor William M. McEneaney, Chair
Professor Robert R. Bitmead
Professor Maurício de Oliveira
Professor Patrick J. Fitzsimmons
Professor Daniel M. Tartakovsky

2015
Copyright
Seung Hak Han, 2015
All rights reserved.
The dissertation of Seung Hak Han is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015
DEDICATION

To my beloved parents
TABLE OF CONTENTS

Signature Page ................................................................. iii
Dedication ........................................................................ iv
Table of Contents ............................................................. v
List of Figures ................................................................. vii
List of Tables ................................................................. viii
Acknowledgements ........................................................... ix
Vita ................................................................................... x
Abstract of the Dissertation ............................................... xi

Chapter 1  Introduction ...................................................... 1

Chapter 2  Fundamental solutions for two-point boundary value problems in orbital mechanics .................................................. 5
  2.1 Introduction ................................................................ 5
  2.2 Problem statement and Fundamental solution .................... 6
  2.3 Optimal control problem ............................................. 9
  2.4 Differential game formulation .................................... 13
    2.4.1 Revisiting the payoff ........................................... 14
    2.4.2 Existence and uniqueness of optimal controls: $c < \infty$ .. 16
    2.4.3 Existence and uniqueness of optimal controls: $c = \infty$ .. 22
    2.4.4 Hamilton-Jacobi-Bellman PDE ................................ 24
  2.5 The fundamental solution in terms of Riccati equation solutions 26
  2.6 The maximization over $\alpha$ ....................................... 29
    2.6.1 Derivative of $W^{\alpha,c}$ with respect to $\alpha$ ............... 30
    2.6.2 An approximate solution ..................................... 35
    2.6.3 Error analysis .................................................... 38
    2.6.4 First-order necessary condition for maximization .......... 41
  2.7 Example ..................................................................... 45

Chapter 3  Solution of an Optimal Sensing and Interception Problem Using Idempotent Methods ........................................... 47
  3.1 Introduction ........................................................... 47
  3.2 Problem definition .................................................... 50
    3.2.1 System dynamics ................................................. 51
    3.2.2 Payoff and value ................................................ 53
  3.3 Idempotent based dynamic programming ......................... 54
  3.4 Efficient projection and pruning ................................... 58
    3.4.1 $L_{\infty}$ error bounds ........................................... 64
LIST OF FIGURES

Figure 2.1: All trajectories with time axis ............................ 46
Figure 2.2: Truth and solution of TPBVP ............................. 46

Figure 3.1: UAV/UGS/Interceptor problem .......................... 48
Figure 3.2: An example of $\nabla_{T-2}(q, x^i)$ over a simplex slice in coordinates $(q^2, q^3, q^5)$ with $x^i = 5$ ........................................... 59

Figure 4.1: Example solutions: $c = 1, d = 0.1, k_0 = 1$ ........... 89
Figure 4.2: Example solutions: $c = 1, d = 0.1, k_0 = 0.25$ ....... 89
Figure 4.3: Example solutions: $c = 1, d = 0.02, k_0 = 0.01$ ....... 89
**LIST OF TABLES**

Table 3.1: Probability transition matrix ............................................. 66
Table 3.2: Confusion matrix ................................................................. 67
Table 3.3: Monte Carlo testing results .................................................... 67
ACKNOWLEDGEMENTS

First and foremost I would like to express my deepest gratitude to my advisor, Professor William M. McEneaney, the greatest teacher I’ve ever had in all my years of education. He has provided me with thoughtful guidance, patience, and warm encouragement during my graduate studies, and his passion for scientific researches has been an excellent role model for me.

I gratefully acknowledge the members of my committee, Professor Robert R. Bitmead, Maurício de Oliveira, Patrick J. Fitzsimmons, and Daniel M. Tartakovsky, for taking the time of their busy schedules and for their invaluable consideration and insightful comments, which improved this dissertation.

Of course, nothing can compare with the gratitude I feel to my parents, Kwang Soo Han and Myung Jin Kim, who instilled in me the joy of learning, inspired me to be the best I could be, supported me in reaching my goals, and loved me unconditionally always. To my brother, Dr. Seung Ryong Han, whose path I followed closely, thank you for leading the way. I would also like to thank my Macho in heaven.

I also wish to extend my thanks to my parents-in-law, Sung Sick Kim and Soon Ye Kang, for their kind support and encouragement.

My most profound thanks go to my wife, Jung Yeon Kim. I have been deeply indebted to my wife for her support, encouragement, quiet patience and unwavering love. Last but not least, I thank my son, Aiden Jeewoo Han. His adorable smiles can always cheer me up and make all my efforts worthy.

This dissertation includes reprints of the following papers:

Seung Hak Han, William M. McEneaney, “Fundamental solutions for two-point boundary value problems in orbital mechanics”, submitted to Applied Mathematics & Optimization (Chapter 2)

Seung Hak Han, William M. McEneaney, “Solution of an Optimal Sensing and Interception Problem Using Idempotent Methods”, in Recent Advances in Research on Unmanned Aerial Vehicles, Springer-Verlag, 2013 (Chapter 3)

William M. McEneaney, Seung Hak Han, “Optimization Formulation and Monotonic Solution Method for the Witsenhausen Problem”, Automatica, 2015 (Chapter 4)

The dissertation author was the primary author or co-author in these publications and Professor McEneaney directed and supervised the research.
VITA

2006 B.S. in Mechanical Engineering, Sogang University, Seoul, Korea
2006 B.S. in Electronic Engineering, Sogang University, Seoul, Korea
2006-2008 Engineer, Samsung Electronics, Suwon, Korea
2009-2015 Graduate Student Researcher, University of California, San Diego.
2010 M.S. in Engineering Sciences (Mechanical Engineering), University of California, San Diego.
2015 Ph.D. in Engineering Sciences (Mechanical Engineering), University of California, San Diego.

PUBLICATIONS

Seung Hak Han, William M. McEneaney, “Fundamental solutions for two-point boundary value problems in orbital mechanics”, submitted to Applied Mathematics & Optimization.


ABSTRACT OF THE DISSERTATION

Techniques for solving nonlinear optimal control problems

by

Seung Hak Han

Doctor of Philosophy in Engineering Sciences (Mechanical Engineering)

University of California, San Diego, 2015

Professor William M. McEneaney, Chair

This dissertation is a study of problem-solving techniques for optimization problems, and considers both theoretical and practical issues. Theoretical approaches based on the dynamical systems and corresponding cost criteria are developed, and efficient computational methods to implement these theoretical developments are explored.

We begin with a two-point boundary value problem (TPBVP) in orbital mechanics in the context of the principle of least action. In particular, the motion of a single body under the influence of the gravitational potential generated by $N$ other large celestial bodies is examined. We derive the fundamental solution from the value function for an optimal control formulation, which may be used to solve the TPBVPs for various boundary conditions. A mathematical theory allowing one to interpret such optimization problems as differential games is obtained, and in consequence, the problem is converted
into a convex-concave minimax problem over infinite dimensional spaces. Further, an approximate subproblem is introduced, and corresponding first-order necessary conditions are obtained, where these are used in the numerical method.

Secondly, we address a motivational application in the realm of military command and control involving an unmanned aerial vehicle and an interceptor ground vehicle defending the military base against an intruder, where the model includes imperfect information obtained from unattended ground sensors. Due to the presence of observation errors, stochastic models are employed and a dynamic programming approach is taken where the value is backward propagated using a min-plus curse-of-dimensionality-free algorithm. A greedy pruning algorithm to attenuate the associated complexity growth, computationally enhanced by use of the double description method, is developed.

Lastly, the Witsenhausen counterexample is examined as an optimization problem over the space of quantile functions that is a convex cone in $L_2(0,1)$. Calculus of variations methods are applied, which aids in the development of necessary conditions for minima, where the conditions are beyond first-order. A numerical method is obtained, which generates a sequence of approximations with monotonically decreasing cost, based on the necessary conditions.
Chapter 1

Introduction

We are confronted with various decisions to make on a daily basis, and try to make the best decisions possible. There have been many efforts towards such problems of decision making for many years by engineers, mathematicians, economists, operations researchers and many others in the field of optimization in a variety of contexts. Techniques and strategies for analyzing problems vary according to the characteristics of the systems of interest and the corresponding objectives. The main goal of this work is to present problem-solving techniques for various classes of control problems arising from theoretical and practical issues in the realm of optimization. Dynamical systems in three different classes of optimal control problems are addressed in the next three chapters. In each chapter, we provide a model-based approach to develop an optimal control to minimize the performance measure (e.g., cost, payoff, action, etc.), a mathematical expression which when minimized subject to satisfaction of system constraints, indicates that the corresponding system achieves the desired performance. Various functional analytic principles are utilized to derive an approach suitable for each optimization problem, including dynamic programming, duality principle, the calculus of variations, Pontryagin’s minimum principle and so on. Also, efficient computational methods based on the theoretical development are explored, and every chapter concludes with one or more examples, many of which are numerical.

We first address a deterministic, continuous-time and continuous-space optimal control problem of minimizing a non-quadratic action functional of a linear dynamical system governed by conservative dynamics. The problem is explored via the principle of least action, wherein the system evolves so as to minimize the action functional where
the action functional is given by the kinetic energy minus the potential energy, integrated over time, satisfying system constraints (cf., [12, 13]). The key aspect of the theoretical development in this class of problems is that one may take a dynamic game approach, which allows the problem of minimizing non-quadratic action functional to be converted to a convex-concave minimax problem seeking a stationary point. In the case where the cost is described by a quadratic functional, rather than a non-quadratic one, the minimizing control problem is common in classical control theory (cf., [5, 22]). There, one is required to solve associated Riccati equations to obtain the desired optimal control problem solution, which gives rise to fruitful analytical techniques and numerical analyses (cf., [31, 33, 39]).

The second problem is the one in coordinated command and control seeking optimal strategies for an unmanned aerial vehicle and an interceptor ground vehicle that maximize the survivability of a military base against an intruder under imperfect observation. The problem belongs to the class of stochastic, discrete-time and discrete-space problems minimizing a linear form of payoff where the state of the system consists not only of the physical state, but also a component describing the information state. In particular, discrete-time dynamical systems are modelled via linear state-space equations, and the information state is modelled as an observation-conditioned stochastic process. The payoff has a linear form over an $N$-dimensional probability simplex domain. We develop an idempotent-based numerical approach in the class of curse-of-dimensionality-free methods where the form of the payoff is preserved under backward dynamic programming (cf., [32, 35, 36, 39]). However, the difficulty associated with the approach lies in an extreme curse-of-complexity, wherein the number of terms in the min-plus expansion grows very rapidly as one propagates. Therefore, efficient projection and pruning methods are also developed for attenuating the complexity growth resulting from this class of methods (cf. [34, 36, 39, 45]).

The last problem is a two-stage linear quadratic Gaussian team problem with nonclassical information structure. In particular, the Witsenhausen counterexample is considered, which is a two-stage decentralized stochastic control problem. There is incomplete communication between two decision makers, the controller and the estimator, and as a result, the observation of the controller is not available to the estimator which characterizes the difficulty of the problem. The cost functional consists of a measure of the controller effort and the expected squared estimate error, and the controller is
attempting not only to minimize its own effort, but also to aid the estimator through its control action. Although it has a quadratic cost structure, Witsenhausen demonstrated that the optimal policy admits a nonlinear optimal solution that outperforms all linear control policies and also proved the existence of such a solution (cf., [55]). The nonlinear optimal policy and corresponding optimal costs are still unknown so that numerous approaches have been made to solve the problem since Witsenhausen posed it. In search of an optimal law in this chapter, the analysis is performed using an associated quantile function (cf., [14]), rather than the controller itself or its distribution, so that calculus of variations methods may be applied, which aids in the development of the necessary conditions for minima.

The contents and structure of the thesis are organized as follows.

In Chapter 2, we consider a two-point boundary value problem (TPBVP) in orbital mechanics involving a small body (e.g., a spacecraft or asteroid) and N larger bodies. The least action principle TPBVP formulation is converted into an initial value problem via the addition of an appropriate terminal cost to the action functional. The latter formulation is used to obtain a fundamental solution, which may be used to solve the TPBVP for a variety of boundary conditions within a certain class. In particular, the method of convex duality allows one to interpret the least action principle as a differential game, where an opposing player maximizes over an indexed set of quadratics to yield the gravitational potential. In the case where the time duration is less than a specific bound, there exists a unique critical point for the resulting differential game, which yields the fundamental solution given in terms of the solutions of associated Riccati equations. Further, the differentiability of the solutions of Riccati equations with respect to the maximizing player is examined, which aids in the development of the first-order necessary condition for a maximizing subproblem, where these are used in the numerical method.

In Chapter 3, we consider a problem where one desires to intercept a vehicle before it reaches a specific (target) location. The vehicle travels on a road network. There are unattended ground sensors (UGSs) on the road network. The data obtained by a UGS is uploaded to an unmanned aerial vehicle (UAV) when the UAV overflies the UGS. There is an interceptor vehicle which may travel the road network. The interceptor uses intruder state estimates based on the data uploaded from the UGSs to the UAV. There is a negative payoff if the intruder is intercepted prior to reaching the target location.
location, and a positive payoff if the intruder reaches the target location without having been previously intercepted. The intruder position is modeled as a Markov chain. The observations are corrupted by random noise. A dynamic programming approach is taken where the value is propagated using a min-plus curse-of-dimensionality-free algorithm. The double-description method for convex polytopes is used to reduce computational loads.

In Chapter 4, the Witsenhausen counterexample is examined. The problem is reduced to an optimization problem over the space of quantile functions. The optimization criterion is the sum of two functionals. The first, representing the control cost, is a simple quadratic. The second, representing the expected squared estimation error, has a more complex structure over this space. Calculus of variations methods are applied, and necessary conditions are generated. Aspects of the structure of the problem and the solution are discussed. A numerical method generating a sequence of solution approximations with monotonically decreasing cost is constructed, based on the necessary conditions. The limit of the method satisfies the necessary conditions.
Chapter 2

Fundamental solutions for two-point boundary value problems in orbital mechanics

2.1 Introduction

We examine the motion of a single body under the influence of the gravitational potential generated by \( N \) other celestial bodies, where the mass of the first body is negligible relative to the masses of the other bodies, and we suppose that the \( N \) large bodies are on known trajectories. The single, small body follows a trajectory satisfying the principle of stationary action (cf., [12, 13]), where under certain conditions, the stationary-action trajectory coincides with the least-action trajectory. This allows such problems in dynamics to be posed, instead, in terms of optimal control problems with vastly simplified dynamics. Here, we are specifically interested in two-point boundary value problems (TPBVPs). From the solution of certain optimal control problems, we will obtain fundamental solutions for classes of TPBVPs.

In the case of a quadratic potential function, the control problem takes a linear-quadratic form. Although the gravitational potential is not quadratic, one may take a dynamic game approach, where an inner optimization problem is posed in a linear-quadratic form. In particular, the non-quadratic control problem is converted into a differential game where the minimizing, outer player controls the velocity, and the maximizing, inner player controls the potential energy term (cf., [4, 11]). It will be demon-
strated that for the case where the time duration is less than a specified bound, the action functional is strictly convex in the velocity control. The action functional is naturally concave in the potential-energy control. We will find that because of the very special form of this problem, one can invert the order of minimum and maximum so that the maximizing player is the outer player. For any potential-energy (outer player) control, the minimizing trajectory is the unique stationary point, and the least action is obtained by solution of associated Riccati equations. This leaves only a control problem for the outer player. We use a numerical method to maximize this concave function. As an aid in this maximization, we also obtain the derivative of the Riccati-equation solution with respect to the potential-energy control.

In Section 2.2, we define the orbital mechanics problem of interest, and develop its relevant least action principle. In Sections 2.3 and 2.4, the problem is reformulated into the aforementioned equivalent differential game in a linear-quadratic form. We provide a bound on the duration that guarantees the existence and uniqueness of the solution. Then, in next two sections, we examine two subproblems of the differential game, separately. Specifically, in Section 2.5, we demonstrate that the minimizing subproblem in the differential game is solved via associated Riccati equations, and in Section 2.6, the derivative of the solution of Riccati equations with respect to the maximizing player is examined. Further, an approximate subproblem is introduced, and corresponding first-order necessary conditions are obtained, where these are used in the numerical method. An error analysis is also provided. Lastly, in Section 2.7, an example is given.

2.2 Problem statement and Fundamental solution

We consider a small body, moving among a set of $N$ other bodies in $\mathbb{R}^3$. The only forces to be considered are gravitational. The single body has negligible mass in relation to the masses of the other bodies, and consequently has no effect on their motion. In particular, we suppose that the $N$ bodies are moving along already-known trajectories. We will obtain fundamental solutions of TPBVPs for the motion of the small body. Note that, for a problem involving dynamical systems, we use the term fundamental solution to indicate an object, which once obtained for a specific time-horizon, allows solution of the problem for varying input data by an operation on the object and given specific data that does not require re-propagation over time. (See [38] for further discussion.) The concept will become more clear further below.
The set of \( N \) bodies may be indexed as \( N = \{1, 2, \ldots, N\} \). Throughout, for integers \( a \leq b \), we will use \([a, b]\) to denote \( \{a, a+1, \ldots, b-1, b\} \). We assume that the larger bodies are spherical with spherically symmetric densities. As for a given total mass, the specific radial density profiles of the bodies do not affect the resulting trajectories (for small-body paths not intersecting the larger bodies), we may, without loss of generality assume that the larger bodies each have uniform density. For \( i \in N \), let \( \rho_i \) and \( R_i \) denote the (uniform) density and radius of larger body \( i \). Obviously, the mass of each body is given by \( m_i = \frac{4}{3} \pi \rho_i R_i^3 \).

Let \( \zeta^i_r(\tau) \) denote the position of the center of body \( i \) at time \( \tau \geq 0 \). We suppose that \( \zeta \triangleq \{\zeta^i\}_{i \in N} \in \tilde{\mathcal{Z}} \triangleq \{\{\zeta^i\}_{i \in N} \mid \zeta^i \in C([0, \infty); \mathbb{R}^3) \forall i \in N\} \), where \( \tilde{\mathcal{Z}} \) will be equipped with the usual (supremum) norm. Assuming that collision between bodies does not occur, we define the subset of \( \tilde{\mathcal{Z}} \) given by \( \mathcal{Z} = \{\zeta \in \tilde{\mathcal{Z}} \mid |\zeta^i_\tau - \zeta^j_\tau| > R_i + R_j \ \forall \tau \geq 0, \forall i \neq j \in N\} \).

For simplicity, the small body is considered as a point particle with mass \( \bar{m} \). Suppose that the position of the small body at time \( \tau \) is denoted by \( \xi_\tau \), where also, we will use \( x \in \mathbb{R}^3 \) to denote generic position values. We model the dynamics of the small body position as

\[
\dot{\xi}_\tau = u_\tau, \quad \xi_0 = x,
\]

where \( u = u_\tau \in U^\infty \triangleq \{u : [0, \infty) \rightarrow \mathbb{R}^3 \mid u_{[0, \tau]} \in L_2([0, \tau); \mathbb{R}^3) \ \forall \tau \in [0, \infty)\} \) where \( u_{[0, \tau]} \) denotes the restriction of the function to domain \([0, \tau)\).

The kinetic energy, \( \hat{T} \), for generic velocity, \( v \), is given by

\[
\hat{T}(v) = \frac{1}{2} \bar{m} |v|^2 \ \forall v \in \mathbb{R}^3.
\]

Let \( \mathcal{Y} \triangleq \{\{y^i\}_{i \in N} \mid y^i \in \mathbb{R}^3 \ \forall i \in N\} \). Given \( i \in N \) and \( Y \in \mathcal{Y} \), the potential energy between the small body at \( x \) and body \( i \) at \( y^i \), \( \hat{V}_i(x, y^i) \), is given by

\[
-\hat{V}_i(x, y^i) = \begin{cases} \frac{G \bar{m} \bar{m_i} 3R_i^2 - |x - y^i|^2}{2R_i^3} & \text{if } x \in B_{R_i}(y^i), \\ \frac{G \bar{m} \bar{m_i}}{|x - y^i|} & \text{if } x \not\in B_{R_i}(y^i), \end{cases}
\]

where \( G \) is the universal gravitational constant. We define the total potential energy \( \hat{V} : \mathbb{R}^3 \times \mathcal{Y} \rightarrow \mathbb{R} \) as \( \hat{V}(x, Y) = \sum_{i \in N} \hat{V}_i(x, y^i) \). We remark that we include the gravitational potential here within the extended bodies as the finiteness and smoothness of the potential are relevant at technical points in the theory, in spite of the infeasibility of small body trajectories that pass through the other bodies.
We remind the reader that we will obtain fundamental solutions for the TPBVPs through a game-theoretic formulation. The game will appear through application of a generalization of convex duality to a control-problem formulation. With that in mind, we define the action functional $J^0 : [0, \infty) \times \mathbb{R}^3 \times U^\infty \times Z \to \mathbb{R}$ as

$$J^0(t, x, u, \zeta) = \int_t^0 T(u_r) - \nabla(\xi_r, \zeta_r) \, dr,$$

(2.3)

where

$$\nabla = \hat{\nabla}/\bar{m}, \quad \nabla_i = \hat{\nabla}_i/\bar{m} \quad \text{and} \quad T = \hat{T}/\bar{m},$$

(2.4)

and $\xi$ satisfies (2.1).

Adding a terminal cost to $J^0$ will yield a control problem equivalent to a TPBVP, where we can manipulate the terminal condition in the TPBVP by adjusting the terminal cost, and we will have initial condition $\xi_0 = x$. For background on this approach to TPBVPs for conservative systems, see [16, 37, 38]. Given generic terminal cost $\bar{\psi} : \mathbb{R}^3 \to \mathbb{R}$, let

$$\bar{W}(t, x, \zeta) \triangleq \inf_{u \in U^\infty} \bar{J}(t, x, u, \zeta).$$

(2.5)

Suppose that the desired terminal position of the small body is denoted by $z \in \mathbb{R}^3$, i.e., $\xi_t = z$. Then, we define an associated reachability problem via the value function $\hat{W} : [0, \infty) \times \mathbb{R}^3 \times Z \times \mathbb{R}^3 \to \mathbb{R}$ by

$$\hat{W}(t, x, \zeta, z) \triangleq \inf_{u \in U^\infty} \left\{ J^0(t, x, u, \zeta) \mid \xi_t = z \right\},$$

(2.6)

where $\xi$ satisfies (2.1), and the function $\hat{W} : [0, \infty) \times \mathbb{R}^3 \times Z \to \mathbb{R}$ by

$$\hat{W}(t, x, \zeta) \triangleq \inf_{z \in \mathbb{R}^3} \left\{ \hat{W}(t, x, \zeta, z) + \bar{\psi}(z) \right\}.$$  

(2.7)

**Theorem 2.2.1.** $\bar{W}(t, x, \zeta) = \hat{W}(t, x, \zeta)$ for all $t \geq 0$, $x \in \mathbb{R}^3$, and $\zeta \in Z$.

**Proof.** Let $t > 0$, $x \in \mathbb{R}^3$, and $\zeta \in Z$. By (2.6) and (2.7),

$$\hat{W}(t, x, \zeta) = \inf_{z \in \mathbb{R}^3} \inf_{u \in U^\infty} \left\{ J^0(t, x, u, \zeta) + \bar{\psi}(z) \mid \xi_t = z \right\} = \inf_{z \in \mathbb{R}^3} \inf_{u \in U^\infty} \bar{J}(t, x, u, \zeta) = \inf_{u \in U^\infty} \bar{J}(t, x, u, \zeta) = \bar{W}(t, x, \zeta).$$

$\Box$
For the development of the fundamental solution, it is useful to introduce a terminal cost that takes the form of a min-plus delta-function. Let \( \psi^\infty : \mathbb{R}^3 \times \mathbb{R}^3 \to [0, \infty] \) (where throughout we let \([0, \infty] = [0, \infty) \cup \{\infty\}\)) be given by

\[
\psi^\infty(y, z) = \delta^-(y - z) = \begin{cases} 
0 & \text{if } y = z, \\
\infty & \text{otherwise},
\end{cases}
\]

where \( \delta^- \) denotes the min-plus “delta-function” (cf., [30]). We define the finite time-horizon payoff \( \bar{J}^\infty : [0, \infty) \times \mathbb{R}^3 \times U^\infty \times Z \times \mathbb{R}^3 \to \mathbb{R} \cup \{\infty\} \) by

\[
\bar{J}^\infty(t, x, u, \zeta, z) = J^0(t, x, u, \zeta) + \psi^\infty(\xi_t, z),
\]

and the corresponding value function as

\[
\bar{W}^\infty(t, x, \zeta, z) = \inf_{u \in U^\infty} \bar{J}^\infty(t, x, u, \zeta, z),
\]

where \( \xi \) satisfies (2.1). The proof of the following is nearly identical to that of Proposition 2.11 in [38], and so is not included.

**Theorem 2.2.2.** For all \( t \geq 0, x \in \mathbb{R}^3 \) and \( \zeta \in Z \),

\[
\bar{W}^\infty(t, x, \zeta, z) = \widetilde{W}(t, x, \zeta, z) < \infty,
\]

and

\[
W(t, x, \zeta) = \inf_{z \in \mathbb{R}^3} \left\{ \bar{W}^\infty(t, x, \zeta, z) + \bar{\psi}(z) \right\}.
\]

It is seen that given \( \bar{W}^\infty \), the value function \( W \) of (2.5) for terminal cost \( \bar{\psi} \) can be evaluated via (2.7). Therefore, \( \bar{W}^\infty \) may be regarded as a fundamental solution.

### 2.3 Optimal control problem

We will find it helpful to define a value function \( \bar{W}^c \) with quadratic terminal cost \( \psi^c \), and demonstrate that the limit property, \( \lim_{c \to \infty} \bar{W}^c = \bar{W}^\infty \) holds. For \( c \in [0, \infty) \), let \( \psi^c : \mathbb{R}^3 \times \mathbb{R}^3 \to [0, \infty) \) be given by

\[
\psi^c(x, z) = \frac{c}{2} |x - z|^2.
\]

We define the finite time-horizon payoff, \( \bar{J}^c : [0, \infty) \times \mathbb{R}^3 \times U^\infty \times Z \times \mathbb{R}^3 \to \mathbb{R} \), by

\[
\bar{J}^c(t, x, u, \zeta, z) = J^0(t, x, u, \zeta) + \psi^c(\xi_t, z),
\]

where \( J^0 \) is given by (2.3), and corresponding value function,
\begin{align}
\overline{W}(t, x, \zeta, z) &= \inf_{u \in U_{t}} \overline{J}(t, x, u, \zeta, z). \tag{2.11}
\end{align}

**Lemma 2.3.1.** The potential energy $\overline{V}(x, Y)$ is globally Lipschitz continuous in $x$, i.e., for any $Y \doteq \{y^{i}\}_{i \in \mathcal{N}} \in \mathcal{Y}$, there exists $K_{L} = K_{L}(\{m_{i}, R_{i}\}_{i \in \mathcal{N}}) < \infty$ such that

\begin{equation}
|\overline{V}(x, Y) - \overline{V}(\hat{x}, Y)| \leq K_{L}|x - \hat{x}| \quad \forall x, \hat{x} \in \mathbb{R}^{3}. \tag{2.12}
\end{equation}

Further, there exists $K^{3}_{L} = K^{3}_{L}(\{m_{i}, R_{i}\}_{i \in \mathcal{N}}) < \infty$ such that

\begin{equation}
|\overline{V}(x, Y)| \leq K^{3}_{L}(1 + |x|) \quad \forall x \in \mathbb{R}^{3}. \tag{2.13}
\end{equation}

**Proof.** By the definition of $\overline{V}_{i}$, we see that there exists $D_{V} = D_{V}(\{m_{i}, R_{i}\}_{i \in \mathcal{N}}) < \infty$ such that

\begin{equation}
0 < -\overline{V}(x, Y) \leq D_{V} \quad \forall x \in \mathbb{R}^{3}, \forall Y \in \mathcal{Y}. \tag{2.14}
\end{equation}

Given $y^{i} \in \mathbb{R}^{3}$, in the cases where $x, \hat{x} \in B_{R_{i}}(y^{i})$ and $x, \hat{x} \notin B_{R_{i}}(y^{i})$, by (2.2), (2.4) and the mean value theorem, we have

\begin{equation}
|\overline{V}_{i}(x, y^{i}) - \overline{V}_{i}(\hat{x}, y^{i})| \leq \frac{Gm_{i}}{R_{i}^{2}}|x - \hat{x}|. \tag{2.15}
\end{equation}

Lastly, suppose without loss of generality that $x \notin B_{R_{i}}(y^{i})$ and $\hat{x} \in B_{R_{i}}(y^{i})$. Let $x^{\lambda} = x(\lambda) = \lambda x + (1 - \lambda)\hat{x}$ for $\lambda \in [0, 1]$. Then, there exists $\lambda \in (0, 1]$ such that $|x^{\lambda} - y^{i}| = R_{i}$. Note that for such $x^{\lambda} \in \mathbb{R}^{3}$, the potential is given by

\begin{equation}
-\overline{V}_{i}(x^{\lambda}, y^{i}) = Gm_{i} \frac{3R_{i}^{2} - |x^{\lambda} - y^{i}|^{2}}{2R_{i}^{3}} = \frac{Gm_{i}}{|x^{\lambda} - y^{i}|}. \tag{2.16}
\end{equation}

Therefore, by the triangle inequality,

\begin{equation}
|\overline{V}_{i}(x, y^{i}) - \overline{V}_{i}(\hat{x}, y^{i})| \leq |\overline{V}_{i}(x, y^{i}) - \overline{V}_{i}(x^{\lambda}, y^{i}) + |\overline{V}_{i}(x^{\lambda}, y^{i}) - \overline{V}_{i}(\hat{x}, y^{i})|,
\end{equation}

which by using the results of two previous cases with (2.16),

\begin{equation}
\leq \frac{Gm_{i}}{R_{i}^{2}} \left[|x - x^{\lambda}| + |x^{\lambda} - \hat{x}| \right] = \frac{Gm_{i}}{R_{i}^{2}} |x - \hat{x}| \tag{2.17}
\end{equation}

where the last equality is obtained since $x^{\lambda}$ is a point on the straight-line between $x$ and $\hat{x}$. Consequently, combining (2.15) and (2.17) yields the first assertion, and letting $\hat{x} = 0$, the second assertion is obtained by (2.12) and (2.14). \hfill \Box

**Theorem 2.3.2.**

\begin{equation}
\overline{W}^\infty(t, x, \zeta, z) = \lim_{c \to \infty} \overline{W}^{c}(t, x, \zeta, z) = \sup_{c \in [0, \infty)} \overline{W}^{c}(t, x, \zeta, z).
\end{equation}

where the convergence is uniform on $\overline{B}_{i} \times \overline{B} \times \mathcal{Z} \times \overline{B}$ for any compact $\overline{B}_{i} \subset [0, \infty)$ and compact $\overline{B} \subset \mathbb{R}^{3}$. 
Proof. Let \( t > 0 \). Suppose that given \( x, z \in \mathbb{R}^3 \), the straight-line control from \( x \) to \( z \) is given by \( u_r^s = (1/t)[z - x] \) for all \( r \in [0, t] \), and we let the corresponding trajectory be denoted by \( \xi^s \). Then,

\[
|\xi^s_r| = |x + \int_0^r u^s_\rho \, d\rho| = |x + \frac{r}{t}[z - x]| \leq |x| + |z|
\]

for all \( r \in [0, t] \). By (2.13) and (2.18),

\[
\int_0^t -V(\xi^s_r, \zeta_r) \, dr \leq K^1_L \int_0^t (1 + |\xi^s_r|) \, dr \leq K^1_L (1 + |x| + |z|) t.
\]

Noting that \( \xi^s_t = z \), for \( c \in [0, \infty) \),

\[
\tilde{J}^c(t, x, u^s, \zeta, z) = J^0(t, x, u^s, \zeta),
\]

which by the definition of \( u^s \) and (2.19),

\[
\leq \frac{1}{2t} |z - x|^2 + K^1_L (1 + |x| + |z|) t \leq D_1 (1 + |x|^2 + |z|^2),
\]

for an appropriate choice of \( D_1 = D_1(t) < \infty \).

On the other hand, by definition, given \( c \in (0, \infty) \) and \( \varepsilon \in (0, 1] \), there exists \( u^{c, \varepsilon} \in \mathcal{U}_\infty \) such that

\[
\tilde{J}^c(t, x, u^{c, \varepsilon}, \zeta, z) \leq \tilde{W}^c(t, x, \zeta, z) + \varepsilon.
\]

Let \( \xi^{c, \varepsilon} \) be the trajectory corresponding to \( u^{c, \varepsilon} \). By the non-negativity of \( T \) and \( -V \) and (2.21),

\[
\frac{1}{2} |\xi^{c, \varepsilon}_t - z|^2 \leq \tilde{W}^c(t, x, \zeta, z) + \varepsilon,
\]

which by the suboptimality of \( u^c \) with respect to \( \tilde{W}^c \),

\[
\leq \tilde{J}^c(t, x, u^c, \zeta, z) + \varepsilon,
\]

which by (2.20),

\[
\leq D_1 (1 + |x|^2 + |z|^2) + 1 \leq \frac{1}{2} [\tilde{D}(1 + |x| + |z|)]^2,
\]

for an appropriate choice of \( \tilde{D} = \tilde{D}(t) < \infty \). This implies that

\[
|\xi^{c, \varepsilon}_t - z| \leq \frac{\tilde{D}(1 + |x| + |z|)}{\sqrt{c}}.
\]

Let

\[
\hat{u}^{c, \varepsilon}_r = u^{c, \varepsilon}_r + \frac{1}{t}[z - \xi^{c, \varepsilon}_t], \quad \forall r \in [0, t],
\]

(2.24)
which yields \( \tilde{\xi}_t^{c,\varepsilon} = z \) where \( \tilde{\xi}^{c,\varepsilon} \) denotes the trajectory corresponding to \( \hat{u}^{c,\varepsilon} \). Then, by (2.23) and (2.24),

\[
|\xi_r^{c,\varepsilon} - \tilde{\xi}_t^{c,\varepsilon}| \leq \frac{1}{t} \int_0^r |z - \xi_t^{c,\varepsilon}| \, dr \leq \frac{r \tilde{D}(1 + |x| + |z|)}{t \sqrt{c}}
\]  

(2.25)

for all \( r \in [0, t] \). By (2.12) and (2.25),

\[
\left| \int_0^r -\nabla (\xi_r^{c,\varepsilon}, \zeta_r) + \nabla (\tilde{\xi}_t^{c,\varepsilon}, \zeta_r) \, dr \right| \leq K_L \int_0^t |\xi_r^{c,\varepsilon} - \tilde{\xi}_t^{c,\varepsilon}| \, dr \leq \frac{K_L \tilde{D}(1 + |x| + |z|) t}{2 \sqrt{c}}.
\]  

(2.26)

Now, by (2.3), (2.10) and (2.21),

\[
\int_0^t T(u_r^{c,\varepsilon}) - \nabla (\xi_r^{c,\varepsilon}, \zeta_r) + \psi^c(\xi_r^{c,\varepsilon}, z) \, dr \leq \hat{W}^c(t, x, \zeta, z) + \varepsilon.
\]

By the definition of \( T \) and the non-negativity of \( -\nabla \psi^c \), this implies

\[
\frac{1}{2} \|u^{c,\varepsilon}\|_{L_2(0,t)}^2 \leq \hat{W}^c(t, x, \zeta, z) + \varepsilon \leq \frac{1}{2} (1 + |x| + |z|)^2,
\]

where the last bound follows by (2.22). This implies that

\[
\|u^{c,\varepsilon}\|_{L_2(0,t)} \leq \tilde{D}(1 + |x| + |z|).
\]  

(2.27)

Noting that \( |a|^2 - |b|^2 < |a - b| (|a| + |b|) \) for \( a, b \in \mathbb{R}^3 \),

\[
\left| \int_0^t T(u_r^{c,\varepsilon}) - \hat{T}(\hat{u}_r^{c,\varepsilon}) \, dr \right| \leq \frac{1}{2} \int_0^t \|u_r^{c,\varepsilon} - \hat{u}_r^{c,\varepsilon}\| \, \left[ \|u_r^{c,\varepsilon}\| + \|\hat{u}_r^{c,\varepsilon}\| \right] \, dr,
\]

which by (2.24) and the triangle inequality,

\[
\leq \frac{1}{2t} |z - \xi_t^{c,\varepsilon}| \int_0^t 2|u_r^{c,\varepsilon}| + \frac{1}{t} |z - \xi_t^{c,\varepsilon}| \, dr,
\]

which by applying Hölder’s inequality,

\[
\leq \frac{1}{2t} |z - \xi_t^{c,\varepsilon}| \left[ 2 \sqrt{t} \|u^{c,\varepsilon}\|_{L_2(0,t)} + |z - \xi_t^{c,\varepsilon}| \right],
\]

which by (2.23) and (2.27),

\[
\leq \frac{\tilde{D}(t)(1 + |x| + |z|)^2}{\sqrt{c}},
\]  

(2.28)

for all \( x, z \in \mathbb{R}^3 \) and all \( c \in [1, \infty) \), for an appropriate choice of \( \hat{D} = \tilde{D}(t) < \infty \). Therefore, by (2.26), (2.28) and the non-negativity of \( \psi^c \), we have

\[
\hat{J}^c(t, x, u^{c,\varepsilon}, \zeta, z) - \hat{J}^c(t, x, \hat{u}^{c,\varepsilon}, \zeta, z) \geq -\frac{D_2(t)(1 + |x| + |z|)^2}{\sqrt{c}}
\]  

(2.29)

for proper choice of \( D_2(t) < \infty \). The suboptimality of \( \hat{u}^{c,\varepsilon} \) with respect to \( \hat{W}^\infty \) combined with (2.29) yields

\[
\hat{W}^\infty(t, x, \zeta, z) - \frac{D_2(t)(1 + |x| + |z|)^2}{\sqrt{c}} \leq \hat{J}^c(t, x, u^{c,\varepsilon}, \zeta, z) \leq \hat{W}^c(t, x, \zeta, z) + \varepsilon
\]
where the last inequality follows by (2.21). Since this is true for all \( \varepsilon \in (0, 1] \),
\[
W^c(t, x, \zeta, z) \geq W^\infty(t, x, \zeta, z) - \frac{D_2(t)(1 + |x| + |z|)^2}{\sqrt{c}}. \tag{2.30}
\]
Next, we examine the monotonicity of \( W^c \) with respect to \( c \). Given \( t > 0; x, z \in \mathbb{R}^3; u \in \mathcal{U}^\infty \) and \( \zeta \in \mathbb{Z} \), note that for \( c_1 \leq c_2 \leq \infty \), by the definitions of \( \bar{J}^\infty \) of (2.8) and \( \bar{J}^c \) of (2.10),
\[
\bar{J}^{c_1}(t, x, u, \zeta, z) \leq \bar{J}^{c_2}(t, x, u, \zeta, z),
\]
which easily yields
\[
W^{c_1}(t, x, \zeta, z) \leq W^{c_2}(t, x, \zeta, z) \quad \forall c_1 \leq c_2 \leq \infty.
\]
Combining this with (2.30) implies
\[
W^\infty(t, x, \zeta, z) - \frac{D_2(t)(1 + |x| + |z|)^2}{\sqrt{c}} \leq W^c(t, x, \zeta, z) \leq W^\infty(t, x, \zeta, z)
\]
for all \( x, z \in \mathbb{R}^3, \zeta \in \mathbb{Z}, t > 0 \) and \( c \in [1, \infty) \).

\[ \Box \]

### 2.4 Differential game formulation

Recall that the gravitational potential does not take a quadratic form in the position variable. In the case where the potential energy does take a quadratic form, the fundamental solution may be obtained through the solution of associated differential Riccati equations (DREs) [38, 37]. In order to exploit that Riccati-solution form, we will take a duality-based approach to gravitation. That is, we will express the additive inverse of the gravitational potential as the pointwise maximum over an indexed set of quadratics. Extending this to time-dependent trajectories, the action functional will take the form of a max-plus integral, over potential-energy controls, of quadratic action functionals. We will find that the control problem is converted to a zero-sum differential game where the velocity controller is the minimizing player, and the potential-energy controller is an opposing, maximizing player. Although at first, this may appear to lead to additional complications, the ability to exploit the DRE solution form yields significant benefits.

By [38], Lemma 4.1, (see also [37]),
\[
-\nabla_i(x, y^i) = \sup_{\hat{a} \in [0, \sqrt{2/3}R^-_i]} \mu_i \left[ \hat{a} - \frac{\hat{a}^3}{2} |x - y^i|^2 \right] \tag{2.31}
\]
for all $x, y^i \in \mathbb{R}^3$ such that $|x - y^i| \geq R_i$, where $\mu_i \equiv G m_i (\frac{3}{2})^{3/2}$. Now, let

$$\hat{\alpha} \equiv \sqrt{\frac{2}{3}R_i^{-1}}$$

(2.32)

and $|x - y^i| \leq R_i$. Recalling (2.2),

$$-\nabla_i(x, y^i) = G m_i \frac{3R_i^2 - |x - y^i|^2}{2R_i^3},$$

which upon a little calculation, yields

$$= \mu_i \left[ \hat{\alpha}^3 - \frac{3}{2} \hat{\alpha}^3 |x - y^i|^2 \right].$$

(2.33)

Further, let

$$\hat{\alpha} = \{ \hat{\alpha} \}_{i \in \mathbb{N}} \supseteq \{ \tilde{\alpha} \}_{i \in \mathbb{N}}$$

where $\mu_i \equiv G m_i (\frac{3}{2})^{3/2}$.

### 2.4.1 Revisiting the payoff

Let

$$A \equiv \{ \hat{\alpha} \}_{i \in \mathbb{N}} \supseteq \{ \tilde{\alpha} \}_{i \in \mathbb{N}} \ni \tilde{\alpha} \in (0, \sqrt{2/3}R_i^{-1}) \quad \forall i \in \mathbb{N}.$$

Then, using Theorem 2.4.1, the potential energy $-\nabla$ may be represented by

$$-\nabla(x, Y) = \sum_{i \in \mathbb{N}} -\nabla_i(x, y^i) \equiv \max_{\hat{\alpha} \in A} \left\{ -\hat{\nabla}(x, Y, \hat{\alpha}) \right\}$$

(2.34)

where

$$-\hat{\nabla}(x, Y, \hat{\alpha}) \equiv \sum_{i \in \mathbb{N}} \mu_i \left[ \hat{\alpha}^3 - \frac{3}{2} \hat{\alpha}^3 |x - y^i|^2 \right].$$

(2.35)

Further, the payoff (2.10) may be written as

$$\tilde{J}^c(t, x, u, \zeta, z) = \int_0^t T(u_r) - \nabla(\xi_r, \zeta_r) \, dr + \psi^c(\xi_t, z)$$

(2.36)

$$= \int_0^t T(u_r) + \max_{\hat{\alpha} \in A} \{ -\hat{\nabla}(\xi_r, \zeta_r, \hat{\alpha}) \} \, dr + \psi^c(\xi_t, z).$$

(2.37)
Given $t > 0$, let

$$\hat{A}^t \doteq C([0, t]; \mathcal{A}) \quad \text{and} \quad A^t \doteq L_\infty([0, t]; \mathcal{A}).$$

(2.38)

Also, for $\alpha \in A^t$, $r \in [0, t]$, $x \in \mathbb{R}^3$ and $Y \in \mathcal{Y}$, let

$$-V^\alpha(r, x, Y) \doteq -\hat{V}(x, Y, \alpha_r) = \sum_{i \in \hat{N}} \mu_i \left[ \alpha_r^i - \frac{(\alpha_r^i)^3}{2} |x - y^i|^2 \right].$$

(2.39)

Given $c \in [0, \infty]$, let $J^c : [0, \infty) \times \mathbb{R}^3 \times U^\infty \times A^t \times \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ be given by

$$J^c(t, x, u, \alpha, \zeta, z) \doteq \int_0^t T(u_r) - V^\alpha(r, \xi_r, \zeta_r) \, dr + \psi^c(\xi_t, z)$$

$$\doteq J^0(t, x, u, \alpha, \zeta) + \psi^c(\xi_t, z).$$

(2.40)

(2.41)

Let $\hat{\alpha}^* : \mathbb{R}^3 \times \mathbb{R}^n \to A$ be given by $\hat{\alpha}^*(x, Y) \doteq \{[\hat{\alpha}^*]^i(x, y^i)\}_{i \in \hat{N}}$ where

$$[\hat{\alpha}^*]^i(x, y^i) = \arg\max_{\hat{\alpha} \in [0, \sqrt{2/3} R^{-1}]} \mu_i \left[ \hat{\alpha} - \frac{\hat{\alpha}^3}{2} |x - y^i|^2 \right] = \sqrt{2/3} \min\{R^{-1}, |x - y^i|^{-1}\}$$

(2.42)

for all $x, y^i \in \mathbb{R}^3$ and all $i \in \hat{N}$. Let

$$\alpha^*_r = \alpha^*(r; u_r, \zeta_r) = \{[\alpha^*_r]^i| i \in \hat{N}\},$$

(2.43)

where the $i^{th}$ element of $\alpha^*$ is given by

$$[\alpha^*_r]^i = [\hat{\alpha}^*]^i(\xi_r, \zeta_r^i) \quad \forall r \in [0, t],$$

(2.44)

where $\xi_r = x + \int_0^r u_\rho \, d\rho$.

**Theorem 2.4.2.** Let $t > 0$; $c \in [0, \infty)$; $x, z \in \mathbb{R}^3$; and $\zeta \in \mathcal{Z}$. For any $u \in U^\infty$,

$$J^c(t, x, u, \zeta, z) = \max_{\alpha \in \hat{A}^t} J^c(t, x, u, \alpha, \zeta, z)$$

$$= \max_{\alpha \in \hat{A}^t} J^c(t, x, u, \alpha, \zeta, z) = J^c(t, x, u, \alpha^*, \zeta, z),$$

(2.45)

(2.46)

where $\alpha^*$, depending on $u$, is given by (2.43). Further,

$$\overline{J}^c(t, x, \zeta, z) = \inf_{u \in U^\infty} \max_{\alpha \in \hat{A}^t} J^c(t, x, u, \alpha, \zeta, z)$$

$$= \inf_{u \in U^\infty} \max_{\alpha \in \hat{A}^t} J^c(t, x, u, \alpha, \zeta, z).$$

(2.47)

(2.48)

**Proof.** Fix $t > 0$, $x, z \in \mathbb{R}^3$ and $\zeta \in \mathcal{Z}$. Given $u \in U^\infty$, let $\xi$ denote the state trajectory corresponding to $u$ with $\xi_0 = x$. By (2.38) and (2.39), given any $\alpha \in A^t$, $\alpha_r$ is suboptimal in the maximization in (2.37) at each $r \in [0, t]$, and in particular,

$$J^c(t, x, u, \zeta, z) \geq \int_0^t T(u_r) - \hat{V}(\xi_r, \zeta_r, \alpha_r) \, dr + \psi^c(\xi_t, z) = J^c(t, x, u, \alpha, \zeta, z).$$
As this is true for all $\alpha \in A^t$,
\[
\bar{J}^c(t, x, u, \zeta, z) \geq \max_{\alpha \in A^t} J^c(t, x, u, \alpha, \zeta, z).
\tag{2.49}
\]
Then, by (2.34), (2.39), (2.42) and (2.44),
\[
-\nabla(\xi_r, \zeta_r) = -V^{\alpha^*}(r, \xi_r, \zeta_r) \quad \forall r \in [0, t]
\tag{2.50}
\]
and by (2.36), (2.37), (2.40) and (2.50),
\[
\bar{J}^c(t, x, u, \zeta, z) = J^c(t, x, u, \alpha^*, \zeta, z) \leq \sup_{\alpha \in A^t} J^c(t, x, u, \alpha, \zeta, z)
\tag{2.51}
\]
Consequently, combining (2.49) and (2.51) yields (2.45), which immediately implies (2.47).

For $r \in [0, t]$ and $i \in \mathcal{N}$, let $d^i_r = |\xi_r - \zeta^i_r|$. Fix $s \in [0, t]$ and $\varepsilon_d > 0$. By the continuity of $\zeta$ and $\xi$ (the latter being guaranteed by $u \in U^\infty$), there exists $\delta_d > 0$ such that for all $i \in \mathcal{N}$ and all $r \in (s - \delta_d, s + \delta_d) \cap [0, t]$, $|d^i_r - d^i_s| < \varepsilon_d$. Using (2.42) and (2.44), one easily sees that this implies $||[\alpha^*]^i - [\alpha^*_s]^i|| < \varepsilon_d/R_i^2$ for all $r \in (s - \delta_d, s + \delta_d) \cap [0, t]$. Consequently, $\alpha^* \in \tilde{A}^t$, which implies (2.46) and (2.48).

Note that the non-quadratic control problem has been converted into a differential game that has a linear-quadratic form of the potential energy.

2.4.2 Existence and uniqueness of optimal controls: $c < \infty$

We will see that $J^c(t, x, \cdot, \cdot, \zeta, z)$ is strictly convex-concave over the velocity and potential energy control sets within a certain time-horizon bound, and that there exists a unique point over $U^\infty \times A^t$.

We first study the question of existence of optimal velocity controls by examining the smoothness, convexity and coercivity of the payoff. We will obtain a bound on the time duration which will be sufficient to guarantee the convexity and coercivity of the payoff $J^c$ and the uniqueness of the stationary-action trajectory.

Let $t > 0$. We define an operator $B : L^2(0, t) \to L^2(0, t)$ as
\[
[Bv](r) = \int_0^r v_\rho \, d\rho \quad \forall r \in [0, t].
\tag{2.52}
\]
$B$ is obviously linear. Moreover,
\[
\|Bv\|_{L^2(0, t)}^2 = \int_0^t r \|v_\rho\|_{L^2(0, t)}^2 \, dr \leq \int_0^t \left[\int_0^r |v_\rho| \, d\rho\right]^2 \, dr,
\]
which by applying Hölder's inequality to the inner integral,
\[ \leq \int_0^t r \, dr \|v\|^2_{L_2(0,t)} = \frac{t^2}{2} \|v\|^2_{L_2(0,t)}. \] (2.53)

Let \( x, z \in \mathbb{R}^3; c \in [0, \infty); \alpha \in \mathcal{A}^i \), and \( \zeta \in \mathcal{Z} \). Let \( u \in \mathcal{U}^\infty \) and \( \xi \) be the corresponding trajectory. Then, using (2.52), we may rewrite (2.39) as

\[ \int_0^t -V^\alpha(r, \xi_r, \zeta_r) \, dr = \sum_{i \in \mathcal{N}} \mu_i \int_0^t \alpha^i_r \, dr - \sum_{i \in \mathcal{N}} \frac{\mu_i}{2} \int_0^t (\alpha^i_r)^3 |\dot{x}_r + [Bu](r) - \zeta^i_r|^2 \, dr \]

where \( \dot{x}_r = x \) for all \( r \in [0, t] \). Letting

\[ [B^i_\alpha u](r) = \mu_i^{1/2} (\alpha^i_r)^{3/2} \int_0^r u_\rho \, d\rho = \mu_i^{1/2} (\alpha^i_r)^{3/2} [Bu](r), \]

we also note that

\[ \langle B^i_\alpha u, B^j_\alpha u \rangle_{L_2(0,t)} = \int_0^t \mu_i (\alpha^i_r)^3 \|Bu\|^2_2 \, dr, \]

which since \( \alpha^i_r \in (0, \sqrt{2/3} R_i^{-1}) \) for all \( r \in [0, t] \),

\[ \leq \frac{Gm_i}{R_i^3} \int_0^t \|Bu\|^2_2 \, dr \leq \frac{Gm_i}{2R_i^3} \|u\|^2_{L_2(0,t)}, \] (2.54)

where the last inequality follows by (2.53). Then, for \( i \in \mathcal{N} \), we have

\[ U^i(u, \alpha, \dot{x}, \zeta) = \frac{1}{2} \int_0^t \mu_i (\alpha^i_r)^3 |\dot{x}_r - \zeta^i_r|^2 \, dr \]

\[ + \int_0^t \mu_i (\alpha^i_r)^3 (\dot{x}_r - \zeta^i_r) \cdot [Bu](r) \, dr + \frac{1}{2} \int_0^t \mu_i (\alpha^i_r)^3 \|Bu\|^2_2 \, dr \]

\[ \simeq \frac{1}{2} \langle w^i, \dot{x}\rangle_{L_2(0,t)} + \langle w^i, B^i_\alpha u \rangle_{L_2(0,t)} + \frac{1}{2} \langle B^i_\alpha u, B^i_\alpha u \rangle_{L_2(0,t)} \]

where \( w^i_r = w^i_r(\alpha, \dot{x}, \zeta) = \mu_i^{1/2} (\alpha^i_r)^{3/2} (\dot{x}_r - \zeta^i_r) \) for all \( i \in \mathcal{N} \) and \( r \in [0, t] \). For \( \nu = \{\nu^i\}_{i \in \mathcal{N}}, \nu \simeq \{\nu^i\}_{i \in \mathcal{N}} \subset L_2(0,t) \), define the inner product (with associated norm)

\[ \langle \nu, \nu \rangle_{L_2(0,t)} \equiv \sum_{i \in \mathcal{N}} \langle \nu^i, \nu^i \rangle_{L_2(0,t)}. \] (2.56)

Then, letting \( w = \{w^i\}_{i \in \mathcal{N}} \) and \( B_\alpha u = \{B^i_\alpha u\}_{i \in \mathcal{N}} \), we may rewrite (2.54) as

\[ \int_0^t -V^\alpha(r, \xi_r, \zeta_r) \, dr = S(\alpha) - \frac{1}{2} \langle w, w \rangle_{L_2(0,t)} - \langle w, B_\alpha u \rangle_{L_2(0,t)} - \frac{1}{2} \langle B_\alpha u, B_\alpha u \rangle_{L_2(0,t)}, \]

so that \( \bar{J}^0 \) given in (2.41) may be rewritten as

\[ J^0(t, x, u, \alpha, \zeta) = \frac{1}{2} \langle u, u \rangle_{L_2(0,t)} + \tilde{S}(\alpha) - \langle w, B_\alpha u \rangle_{L_2(0,t)} - \frac{1}{2} \langle B_\alpha u, B_\alpha u \rangle_{L_2(0,t)}, \] (2.57)
where \( \hat{S}(\alpha) = \hat{S}(\alpha; \hat{x}, \zeta) \equiv S(\alpha) - \frac{1}{2} \langle w, w \rangle_{L^2(0,t)} \). Further, by (2.52),
\[
\psi^c(\xi_t, z) = \frac{\xi}{2} |x - z + [Bu](t)|^2;
\]
which by letting \( \hat{z}_r = z \) for all \( r \in [0,t] \),
\[
\psi^c(\xi_t, z) = \frac{\xi}{2} |x - z|^2 + \langle c(\hat{x} - \hat{z}), u \rangle_{L^2(0,t)} + \frac{\xi}{2} |[Bu](t)|^2.
\]  

\[ \tag{2.58} \]

Theorem 2.4.3. Let \( t > 0; \ c \in [0,\infty); \ x, z \in \mathbb{R}^3; \ \alpha \in \mathcal{A}^t, \) and \( \zeta \in \mathcal{Z}. \) Then, \( J^0(t, x, u, \alpha, \zeta) \) and \( J^c(t, x, u, \alpha, \zeta, z) \) are Fréchet differentiable with respect to \( u. \)

Proof. Let \( u, v \in U^\infty. \) Then, by (2.41), (2.57) and (2.58),
\[
J^c(t, x, u + v, \alpha, \zeta, z) - J^c(t, x, u, \alpha, \zeta, z) = \langle v, [Bu](t) \rangle_{L^2(0,t)} + \frac{1}{2} \langle [Bu](t), v \rangle_{L^2(0,t)} - \langle Bu, v \rangle_{L^2(0,t)} + \frac{\xi}{2} |[Bu](t)|^2,
\]
which by letting \( B^*_\alpha \) be the adjoint of \( B_\alpha \) (cf., [25]),
\[
= \langle c(\xi_t - z) - B^*_\alpha w + (I - B^*_\alpha B_\alpha)u, v \rangle_{L^2(0,t)}
+ \frac{1}{2} \left[ \langle v, [Bu](t) \rangle_{L^2(0,t)} - \langle B_\alpha v, B_\alpha v \rangle_{L^2(0,t)} + c |[Bu](t)|^2 \right];
\]  

(2.59)

where \( I \) denotes the identity operator. This implies that letting \( DJ^c_\alpha(u) = c(\xi_t - z) - B^*_\alpha w + (I - B^*_\alpha B_\alpha)u, \)
\[
|J^c(t, x, u + v, \alpha, \zeta, z) - J^c(t, x, u, \alpha, \zeta, z) - \langle DJ^c_\alpha(u), v \rangle_{L^2(0,t)}| \leq \frac{1}{2} \left| \langle v, [Bu](t) \rangle_{L^2(0,t)} - \langle B_\alpha v, B_\alpha v \rangle_{L^2(0,t)} + c |[Bu](t)|^2 \right| \leq C_u \|v\|^2_{L^2(0,t)}
\]

for proper choice of \( C_u = C_u(t, \tilde{t}) < \infty. \) Since this is true for all \( u, v \in U^\infty, \) \( J^c \) is Fréchet differentiable with Fréchet derivative \( DJ^c_\alpha. \) Similarly, using (2.57), one easily sees that \( J^0 \) is Fréchet differentiable with Fréchet derivative \( DJ^0_\alpha(u) = (I - B^*_\alpha B_\alpha)u - B^*_\alpha w. \)

We also have the following, and do not include the obvious proof.

Lemma 2.4.4. Let \( t > 0; \ x, z \in \mathbb{R}^3; \ \zeta \in \mathcal{Z}; \ \alpha \in \mathcal{A}^t, \) and \( c \in [0,\infty). \) \( J^0(t, x, \cdot, \zeta), \)
\( \tilde{J}(t, x, \cdot, \alpha, \zeta), \tilde{J}^c(t, x, \cdot, \alpha, \zeta, z), \) and \( J^c(t, x, \cdot, \alpha, \zeta, z) \) are continuous on \( U^\infty. \)

Theorem 2.4.5. Let
\[
\tilde{t} = \left[ \sum_{i \in N} \frac{G_{m_i}}{2R_i^2} \right]^{-1/2}.
\]  

(2.60)

Let \( x, z \in \mathbb{R}^3; \ c \in [0,\infty), \) and \( \zeta \in \mathcal{Z}. \) If \( t \in (0, \tilde{t}), \) \( \tilde{J}^0(t, x, u, \alpha, \zeta) \) and \( J^c(t, x, u, \alpha, \zeta, z) \)
are strictly convex quadratic and coercive in \( u \) for any \( \alpha \in \mathcal{A}^t. \)
Proof. Considering the quadratic terms in $u$ in the definition (2.57) of $\tilde{J}^0$, by (2.55) and (2.56),
\[
\frac{1}{2} \langle u, u \rangle_{L^2(0,t)} - \frac{1}{2} \langle B_\alpha u, B_\alpha u \rangle_{L^2(0,t)} \geq \frac{1}{2} \left[ 1 - \sum_{i \in N} \frac{Gm_i}{2R_i^3} \right] \| u \|^2_{L^2(0,t)} > 0 \tag{2.61}
\]
if
\[
t < \bar{t} \doteq \left[ \sum_{i \in N} \frac{Gm_i}{2R_i^3} \right]^{-1/2}.
\]
That is, $\tilde{J}^0(t, x, \cdot, \alpha, \zeta)$ is coercive and strictly convex if $t \in (0, \bar{t})$. Further, from (2.58), we note that $\psi^c(\xi_t, z)$ is convex quadratic in $u$. Consequently, the strict convexity and coercivity of $J^c(t, x, \cdot, \alpha, \zeta, z)$ are guaranteed for $t \in (0, \bar{t})$.

Remark 2.4.6. The condition (2.60) in Theorem 2.4.5 may be overly restrictive. If one can assume a greater minimum distance from the bodies than their respective $R_i$, say $\delta_i > R_i$ for $i \in [1, N]$, then this could be relaxed, replacing the $R_i$ with the $\delta_i$. Also, we mention that the condition allows one to seek minima rather than stationary points; consideration of the stationary-point case could be an area for future research.

Henceforth, throughout the paper, $\bar{t}$ is used to denote the time upper bound given by (2.60).

By Theorems 2.4.2 and 2.4.5, we have:

Corollary 2.4.7. Let $x, z \in \mathbb{R}^3; c \in [0, \infty)$, and $\zeta \in \mathcal{Z}$. For $t \in (0, \bar{t})$, $\tilde{J}^c(t, x, u, \zeta, z)$ is strictly convex in $u$.

Lemma 2.4.8. Let $x, z \in \mathbb{R}^3; c \in [0, \infty)$, and $\zeta \in \mathcal{Z}$. Then, for $t > 0$, $\tilde{J}^c(t, x, \cdot, \zeta, z)$ is coercive in $U^\infty$.

Proof. For any $u \in U^\infty$, by the non-negativity of $-\nabla$ and $\psi^c$,
\[
\tilde{J}^c(t, x, u, \zeta, z) \geq \frac{1}{2} \| u \|^2_{L^2(0,t)},
\]
which implies the coercivity of $\tilde{J}^c(t, x, \cdot, \zeta, z)$.

Combining Theorem 2.4.5, Corollary 2.4.7, Lemmas 2.4.4 and 2.4.8 immediately yields the following uniqueness property (cf., [24]).

Theorem 2.4.9. Let $t < \bar{t}; x, z \in \mathbb{R}^3; c \in [0, \infty)$. Then, there exists a unique optimal velocity control in the definition (2.11) of $\tilde{W}^c(t, x, \zeta, z)$. For any $\alpha \in \mathcal{A}^t$, there exists a unique optimal velocity control of $J^c(t, x, \cdot, \alpha, \zeta, z)$.
Next we will examine the concavity of $J^c(t, x, u, \cdot, \zeta, z)$, which guarantees the existence of a unique potential energy control in the maximization in (2.45).

**Lemma 2.4.10.** For $t > 0; c \in [0, \infty); x, z \in \mathbb{R}^3; \zeta \in \mathcal{Z}$ and $u \in \mathcal{U}^\infty$, $J^c(t, x, u, \alpha, \zeta, z)$ and $\bar{J}^0(t, x, u, \alpha, \zeta)$ are strictly concave in $\alpha$.

**Proof.** Let $t > 0; c \in [0, \infty); x, z \in \mathbb{R}^3$ and $\zeta \in \mathcal{Z}$. Given $\alpha \in \mathcal{A}$, let $\dot{\alpha} \in L^\infty([0, t]; \mathbb{R}^3)$ be such that $\alpha \pm \dot{\alpha} \in \mathcal{A}$ and $\delta \in [-1, 1]$. Then, for any $u \in \mathcal{U}^\infty$,

$$J^c(t, x, u, \alpha + \delta \dot{\alpha}, \zeta, z) + J^c(t, x, u, \alpha - \delta \dot{\alpha}, \zeta, z) - 2J^c(t, x, u, \alpha, \zeta, z) = J^0(t, x, u, \alpha + \delta \dot{\alpha}, \zeta) + J^0(t, x, u, \alpha - \delta \dot{\alpha}, \zeta) - 2J^0(t, x, u, \alpha, \zeta)$$

$$= \int_0^t -\tilde{V}(\xi_r, \zeta_r, \alpha_r + \delta \dot{\alpha}_r) - \tilde{V}(\xi_r, \zeta_r, \alpha_r - \delta \dot{\alpha}_r) + 2\tilde{V}(\xi_r, \zeta_r, \alpha_r) dr,$$

which, using (2.35),

$$= \int_0^t \sum_{i \in \mathcal{N}} \mu_i \left[ - (\alpha^i_r + \delta \dot{\alpha}^i_r)^3 - (\alpha^i_r - \delta \dot{\alpha}^i_r)^3 + 2(\alpha^i_r)^3 \right] |\xi_r - \zeta^i_r|^2 / 2 dr$$

$$= -3\delta^2 \int_0^t \sum_{i \in \mathcal{N}} \mu_i \alpha^i_r(\dot{\alpha}^i_r)^2 |\xi_r - \zeta^i_r|^2 dr,$$

which completes the proof.

**Theorem 2.4.11.** Let $t \in (0, \infty); x, z \in \mathbb{R}^3; c \in [0, \infty)$ and $\zeta \in \mathcal{Z}$. Let $u^\dagger \in \mathcal{U}^\infty$, and let the corresponding trajectory be denoted by $\xi^\dagger$. Let $\alpha^*_r = \alpha^*_r(r; x, u^\dagger_r, \zeta_r) \equiv \tilde{\alpha}^*(\xi^\dagger_r, \zeta_r)$ for all $r \in [0, t]$ where $\alpha^*$ is given in (2.42). Then, $u^\dagger$ is a stationary point of $\bar{J}^c(t, x, \cdot, \alpha^*, \zeta, z)$ if and only if $u^\dagger$ is a stationary point of $J^c(t, x, \cdot, \alpha^*, \zeta, z)$.

**Proof.** Let $\nu \in \mathcal{U}^\infty$ and $\delta > 0$. Letting $\xi^{\dagger, \nu}$ denote the trajectory corresponding to $u^\dagger + \delta \nu$. We examine differences in the direction $\nu$ from $u^\dagger$. Recall from (2.34) that

$$-\nabla_x \tilde{V}(x, Y) = \max_{\tilde{\alpha} \in \mathcal{A}} \{-\tilde{V}(x, Y, \tilde{\alpha})\}$$

where the maximum is uniquely attained at $\tilde{\alpha}^*(x, Y)$. Consequently,

$$-\nabla_x \tilde{V}(x, Y) = -\nabla_x \tilde{V}(x, Y, \tilde{\alpha}^*(x, Y)),$$

and with this, the first-order difference in the potential-energy term is

$$-\tilde{V}(\xi^{\dagger, \nu}_r, \zeta_r, \alpha^*_r) + \tilde{V}(\xi^{\dagger}_r, \zeta_r, \alpha^*_r) = -\delta \nabla_x \tilde{V}(\xi^{\dagger}_r, \zeta_r, \alpha^*_r) \cdot (\xi^{\dagger, \nu}_r - \xi^{\dagger}_r) + \mathcal{O}(\delta^2)$$

$$= -\delta \nabla_x \tilde{V}(\xi^{\dagger}_r, \zeta_r, \alpha^*_r(\xi^{\dagger}_r, \zeta_r)) \cdot (\xi^{\dagger, \nu}_r - \xi^{\dagger}_r) + \mathcal{O}(\delta^2)$$

$$= -\nabla \tilde{V}(\xi^{\dagger}_r, \zeta_r) + \tilde{V}(\xi^{\dagger}_r, \zeta_r) + \mathcal{O}(\delta^2). \quad (2.62)$$
Now,

\[ J^c(t, x, u^\dagger + \delta \nu, \alpha^*, \zeta, z) - J^c(t, x, u^\dagger, \alpha^*, \zeta, z) \]

\[ = \int_0^t \left[ T(u^\dagger_r + \delta \nu_r) - T(u^\dagger_r) - \tilde{V}(\xi^\dagger_r, \zeta_r, \alpha^*) + \tilde{V}(\xi^\dagger_r, \zeta_r, \alpha^*) \right] \, dr + \psi^c(\xi^\dagger_r, \zeta_r, \alpha^*) \, dr + \psi^c(\xi^\dagger_t, \zeta_t, \alpha^*) , \]

which by (2.62),

\[ = \bar{J}^c(t, x, u^\dagger + \delta \nu, \zeta, z) - \bar{J}^c(t, x, u^\dagger, \zeta, z) + O(\delta^2). \quad (2.63) \]

By (2.63), we have the desired result.

\[ \square \]

**Theorem 2.4.12.** Given \( t < \bar{t}; \, x, z \in \mathbb{R}^3; \, \zeta \in \mathcal{Z} \) and \( c \in [0, \infty) \), let \( u^c* \) be the unique minimizer of \( \bar{J}^c(t, x, \cdot, \zeta, z) \) over \( U^\infty \), and \( \xi^c* \) be the corresponding trajectory. Let \( \alpha^*_r = \tilde{\alpha}^*(\xi^c*, \zeta_r) \) for all \( r \in [0, t] \) where \( \tilde{\alpha}^* \) is given in (2.42). Then,

\[ \bar{W}^c(t, x, \zeta, z) = J^c(t, x, u^c*, \alpha^*, \zeta, z) = \min_{u \in U^\infty} \max_{\alpha \in \mathcal{A}^t} J^c(t, x, u, \alpha, \zeta, z) \]

\[ = \max_{\alpha \in \mathcal{A}^t} \min_{u \in U^\infty} J^c(t, x, u, \alpha, \zeta, z) \quad (2.64) \]

Further,

\[ u^c* = \arg\min_{u \in U^\infty} \max_{\alpha \in \mathcal{A}^t} J^c(t, x, u, \alpha, \zeta, z) \quad (2.66) \]

\[ \alpha^* = \arg\max_{\alpha \in \mathcal{A}^t} \min_{u \in U^\infty} J^c(t, x, u, \alpha, \zeta, z). \quad (2.67) \]

**Proof.** Let \( t, x, z, \zeta, c \) be as indicated in the theorem statement. Then, we note that \( \min_{u \in U^\infty} J^c(t, x, u, \alpha, \zeta, z) \) and \( \max_{\alpha \in \mathcal{A}^t} J^c(t, x, u, \alpha, \zeta, z) \) exist for all \( u \in U^\infty \) and \( \alpha \in \mathcal{A}^t \) by Theorems 2.4.2 and 2.4.9 and Lemma 2.4.10. By Lemma 2.4.4 and Theorem 2.4.11, \( u^c* \) is a stationary point of \( J^c(t, x, \cdot, \alpha^*, \zeta, z) \). Further, by the uniqueness given in Theorem 2.4.9, \( u^c* = \arg\min_{u \in U^\infty} J^c(t, x, u, \alpha^*, \zeta, z) \). Also, note that by Theorem 2.4.5, \( J^c(t, x, \cdot, \zeta, c) \) is strictly convex in \( u \) for all \( \alpha \in \mathcal{A}^t \), which implies that \( \max_{\alpha \in \mathcal{A}^t} J^c(t, x, u, \alpha, \zeta, z) \) is strictly convex in \( u \), which implies that the argmin of (2.66) is single-valued, and one has (2.66). Similarly, using Lemma 2.4.10, one obtains (2.67).

By Lemma 2.4.10 and Theorem 2.4.2,

\[ \alpha^* = \arg\max_{\alpha \in \mathcal{A}^t} J^c(t, x, u^c*, \alpha, \zeta, z), \]

which implies that

\[ \max_{\alpha \in \mathcal{A}^t} J^c(t, x, u^c*, \alpha, \zeta, z) = J^c(t, x, u^c*, \alpha^*, \zeta, z) = \min_{u \in U^\infty} J^c(t, x, u, \alpha^*, \zeta, z). \]
This implies
\[
\min_{u \in U} \max_{\alpha \in A} \mathcal{J}^c(t, x, u, \alpha, \zeta, z) \leq \mathcal{J}^c(t, x, u^c, \alpha^*, \zeta, z) \leq \max_{\alpha \in A} \min_{u \in U} \mathcal{J}^c(t, x, u, \alpha, \zeta, z),
\] (2.68)

Also, by the usual reordering inequality, one has
\[
\max_{\alpha \in A} \min_{u \in U} \mathcal{J}^c(t, x, u, \alpha, \zeta, z) \leq \min_{u \in U} \max_{\alpha \in A} \mathcal{J}^c(t, x, u, \alpha, \zeta, z). \quad (2.69)
\]

Combining (2.68) and (2.69), one has
\[
\min_{u \in U} \max_{\alpha \in A} \mathcal{J}^c(t, x, u, \alpha, \zeta, z) = \mathcal{J}^c(t, x, u^c, \alpha^*, \zeta, z) = \max_{\alpha \in A} \min_{u \in U} \mathcal{J}^c(t, x, u, \alpha, \zeta, z).
\]

Combining this with (2.45) completes the proof. \( \Box \)

### 2.4.3 Existence and uniqueness of optimal controls: \( c = \infty \)

**Remark 2.4.13.** Recall that the TPBVP corresponds to the case \( c = \infty \). Given \( x, z \in \mathbb{R}^3 \) and \( t > 0 \), let \( \tilde{U}^\infty_{t,x,z} = \{ u \in U^\infty | \int_0^t u_r \, dr = z - x \} \). Recall from Theorem 2.2.2 that
\[
W^\infty(t, x, \zeta, z) = \inf_{u \in \tilde{U}^\infty_{t,x,z}} J^0(t, x, u, \zeta).
\]

Also, by Lemma 2.4.4 and Theorem 2.4.5, \( J^0(t, x, u, \zeta) \) is continuous, coercive and strictly convex in \( u \in \tilde{U}^\infty_{t,x,z} \) if \( t < \bar{t} \), which implies that there exists unique optimal velocity control \( u^* \in \tilde{U}^\infty_{t,x,z} \subset U^\infty \) in the definition (2.9) of \( W^\infty(t, x, \zeta, z) \) (cf., [24]), i.e.,
\[
u^* = \arg\min_{u \in U^\infty} \tilde{J}^\infty(t, x, u, \zeta, z) = \arg\min_{u \in \tilde{U}^\infty_{t,x,z}} \tilde{J}^\infty(t, x, u, \zeta, z),
\] (2.70)

where the corresponding trajectory, \( \xi^* \), is the solution of the TPBVP. In an analogous fashion to the proof of Theorem 2.4.11, \( u^* \) is a stationary point of \( \tilde{J}^0(t, x, \cdot, \cdot, \tilde{\alpha}^*, \zeta) \) over \( \tilde{U}^\infty_{t,x,z} \) where \( \tilde{\alpha}^*_r = \tilde{\alpha}^*(\xi^*_r, \zeta_\infty) \) for all \( r \in [0, t] \).

Further, we may represent the fundamental solution in terms of \( W^\alpha,c \) with a limit property.

**Theorem 2.4.14.** Let \( t < \bar{t} \). For \( u \in U^\infty; \alpha \in A^t; x, z \in \mathbb{R}^3 \), and \( \zeta \in \mathcal{Z} \), let
\[
\mathcal{J}^\infty(t, x, u, \alpha, \zeta, z) \doteq J^0(t, x, u, \alpha, \zeta) + \psi^\infty(\xi_t, z).
\] (2.71)

Then,
\[
W^\infty(t, x, \zeta, z) = \max_{\alpha \in A^t} \min_{u \in U^\infty} \mathcal{J}^\infty(t, x, u, \alpha, \zeta, z) = \max_{\alpha \in A^t} W^\alpha,c(t, x, \zeta, z).
\] (2.72)
Further, given \( \alpha \in \mathcal{A}^t \),

\[
W^{\alpha, \infty}(t, x, \zeta, z) = \lim_{{c \to \infty}} W^{\alpha, c}(t, x, \zeta, z) = \sup_{{c > 0}} W^{\alpha, c}(t, x, \zeta, z)
\]

where the convergence is uniform on \( \bar{\mathcal{B}} \times \mathcal{Z} \times \bar{\mathcal{B}} \) for any compact \( \bar{\mathcal{B}} \subset \mathbb{R}^3 \).

**Proof.** Let \( u^* \in \tilde{U}^\infty_{t,x,z} \) be as per (2.70), and \( \xi^* \) be the corresponding trajectory. Let

\[
\tilde{\alpha}^*_r = \tilde{\alpha}^*_r(\xi^*_r, \zeta_r) \quad \text{for all } r \in [0, t].
\]

Then, by the definitions of \( \tilde{U}^\infty_{t,x,z} \) and \( J^\infty \),

\[
\inf_{{u \in U^\infty}} J^\infty(t, x, u, \tilde{\alpha}^*, \zeta) = \inf_{{u \in \tilde{U}^\infty_{t,x,z}}} \tilde{J}^0(t, x, u, \tilde{\alpha}^*, \zeta),
\]

which by Theorem 2.4.5 and Lemma 2.4.4,

\[
= \min_{{u \in \tilde{U}^\infty_{t,x,z}}} \tilde{J}^0(t, x, u, \tilde{\alpha}^*, \zeta) = \tilde{J}^0(t, x, u^*, \tilde{\alpha}^*, \zeta). \tag{2.73}
\]

where the last equality follows by Remark 2.4.13. Noting that the terminal state \( \xi_t = z \) corresponding to all \( u \in \tilde{U}^\infty_{t,x,z} \), by Theorem 2.4.2, we may have

\[
J^0(t, x, u, \zeta) = \max_{\alpha \in \mathcal{A}^t} \tilde{J}^0(t, x, u, \alpha, \zeta) = \tilde{J}^0(t, x, u^*, \tilde{\alpha}^*, \zeta) \quad \forall u \in \tilde{U}^\infty_{t,x,z} \tag{2.74}
\]

where \( \tilde{\alpha}^*_r = \tilde{\alpha}^*_r(\xi^*_r, \zeta_r) \) for \( r \in [0, t] \) and \( \xi^*_r \) denotes the trajectory corresponding to \( u \).

Noting \( \xi^*_t = z \), by (2.71), (2.73) and (2.74),

\[
\max_{\alpha \in \mathcal{A}^t} J^\infty(t, x, u^*, \alpha, \zeta, z) = J^\infty(t, x, u^*, \tilde{\alpha}^*, \zeta, z) = \min_{{u \in \tilde{U}^\infty_{t,x,z}}} J^\infty(t, x, u, \alpha, \zeta, z), \tag{2.75}
\]

which yields in an analogous fashion to the proof of Theorem 2.4.12 that

\[
\min_{{u \in \tilde{U}^\infty_{t,x,z}}} \max_{\alpha \in \mathcal{A}^t} J^\infty(t, x, u, \alpha, \zeta, z) = \max_{\alpha \in \mathcal{A}^t} \min_{{u \in \tilde{U}^\infty_{t,x,z}}} J^\infty(t, x, u, \alpha, \zeta, z), \tag{2.75}
\]

and \((u^*, \tilde{\alpha}^*)\) is the unique solution of (2.75) over \( \tilde{U}^\infty_{t,x,z} \times \mathcal{A}^t \). Consequently, by Remark 2.4.13 and (2.74),

\[
\mathcal{W}^\infty(t, x, \zeta, z) = \min_{{u \in \tilde{U}^\infty_{t,x,z}}} J^0(t, x, u, \zeta) = \min_{{u \in \tilde{U}^\infty_{t,x,z}}} \max_{\alpha \in \mathcal{A}^t} \tilde{J}^0(t, x, u, \alpha, \zeta),
\]

which by the definitions of \( \tilde{U}^\infty_{t,x,z} \) and \( J^\infty \) and (2.75),

\[
= \min_{{u \in \tilde{U}^\infty_{t,x,z}}} \max_{\alpha \in \mathcal{A}^t} J^\infty(t, x, u, \alpha, \zeta, z) = \max_{\alpha \in \mathcal{A}^t} \min_{{u \in \tilde{U}^\infty_{t,x,z}}} J^\infty(t, x, u, \alpha, \zeta, z),
\]

which since \( J^\infty(t, x, u, \alpha, \zeta, z) = \infty \) for all \( u \in U^\infty \setminus \tilde{U}^\infty_{t,x,z} \) and any \( \alpha \in \mathcal{A}^t \),

\[
= \max_{\alpha \in \mathcal{A}^t} \min_{{u \in U^\infty}} J^\infty(t, x, u, \alpha, \zeta, z),
\]

which completes the first assertion.

Regarding the last assertion, similar to Theorem 2.3.2, we have monotonicity of \( \mathcal{W}^{\alpha, c} \) in \( c \), and we do not include the analogous proof. \( \Box \)
Recall from Theorem 2.3.2 that $W^c(t, x, \zeta, z)$ approaches $W^\infty(t, x, \zeta, z)$ as $c \to \infty$. The following lemma shows that the optimal velocity controls for $c < \infty$ also approach those for $c = \infty$ as $c \to \infty$.

**Theorem 2.4.15.** Let $t < \bar{t}$ and $c \in [0, \infty)$. Let $x, z \in \mathbb{R}^3$ and $\zeta \in \mathbb{Z}$. Let $u^*$ and $u^{c,*}$ be the least action points in the definitions (2.9) of $W^\infty(t, x, \zeta, z)$ and (2.11) of $W^c(t, x, \zeta, z)$, respectively. Then, there exists $\tilde{D} = \tilde{D}(t, \bar{t}) < \infty$ such that

$$\|u^* - u^{c,*}\|_{L^2(0,t)}^2 \leq \frac{\tilde{D}(1 + |x| + |z|)^2}{\sqrt{c}}.$$ 

**Proof.** Let $u^*, u^{c,*}$ be as per the statement. Let $\xi_{c,*}$ be the trajectory corresponding to $u^{c,*}$, and $\alpha_r^* \doteq \bar{\alpha}^*(\xi_{c,*}, \zeta_r)$ for all $r \in [0, t]$ where $\bar{\alpha}^*$ is given in (2.42). Then, by (2.29),

$$\frac{D_2(t)(1 + |x| + |z|)^2}{\sqrt{c}} \geq J^\infty(t, x, u^*, \zeta, z) - J^c(t, x, u^{c,*}, \zeta, z),$$

which since $\xi_{c,*}^* = z$,

$$J^c(t, x, u^*, \zeta, z) - J^c(t, x, u^{c,*}, \zeta, z),$$

which by the suboptimality and optimality of $\alpha^*$ with respect to $J^c(t, x, u^*, \alpha, \zeta, z)$ and $J^c(t, x, u^{c,*}, \alpha, \zeta, z)$, respectively,

$$\geq J^c(t, x, u^*, \zeta, z) - J^c(t, x, u^{c,*}, \alpha^*, \zeta, z),$$

which by (2.59) and the optimality of $u^{c,*}$ with respect to $J^c(t, x, u, \alpha^*, \zeta, z)$,

$$\geq \frac{1}{2}\left[\langle u^* - u^{c,*}, u^* - u^{c,*}\rangle_{L^2(0,t)} - \langle B_{\alpha^*} u^* - u^{c,*}, B_{\alpha^*} u^* - u^{c,*}\rangle_{L^2(0,t)}\right],$$

which by (2.61),

$$\geq \frac{1}{2}(1 - (t/\bar{t})^2)\|u^* - u^{c,*}\|_{L^2(0,t)}^2,$$

which completes the proof. \qed

### 2.4.4 Hamilton-Jacobi-Bellman PDE

The Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE) problem associated with our $\alpha$-indexed control problem is

$$0 = -\frac{\partial}{\partial r} W(r, x, \zeta, z) + \inf_{v \in \mathbb{R}^3} \left\{\frac{1}{2}|v|^2 - V^\alpha(t - r, x, \zeta_{t-r}) + v^T \nabla_x W(r, x, \zeta, z)\right\}$$

$$= -\frac{\partial}{\partial r} W(r, x, \zeta, z) - \inf_{v \in \mathbb{R}^3} H^\alpha(t - r, x, v, \zeta, \nabla_x W(r, x, \zeta, z))$$

(2.76)

where the initial conditions, indexed by $z \in \mathbb{R}^3$, corresponding to $\mathcal{W}^\alpha_c$ are

$$W(0, x, \zeta, z) = \psi^\alpha(x, z) \quad \forall x \in \mathbb{R}^3,$$

(2.77)
and $\nabla_x W$ represents the gradient with respect to the space variable.

For $t > 0$, let

$$D_t = C([0, t] \times \mathbb{R}^3) \cap C^1((0, t) \times \mathbb{R}^3).$$

Suppose that $W \in D_t$ satisfies (2.76) and (2.77). Since

$$\frac{1}{2}|v|^2 + v^T \nabla_x W(r, x, \zeta, z) \geq \frac{1}{2}|v|^2 - |v||\nabla_x W(r, x, \zeta, z)|,$$

the coercivity and convexity of the Hamiltonian imply that

$$- \inf_{v \in \mathbb{R}^3} H^\alpha(t - r, x, v, \zeta, \nabla_x W(r, x, \zeta, z)) = - \min_{v \in \mathbb{R}^3} H^\alpha(t - r, x, v, \zeta, \nabla_x W(r, x, \zeta, z)),$$

which since $H^\alpha$ is quadratic in $v$,

$$= -H^\alpha(t - r, x, v^*, \zeta, \nabla_x W(r, x, \zeta, z)) \quad (2.78)$$

where $v^* = -\nabla_x W(r, x, \zeta, z)$.

**Theorem 2.4.16.** Let $t > 0$; $c \in [0, \infty)$; $x, z \in \mathbb{R}^3$; $\zeta \in \mathcal{Z}$ and $\alpha \in \mathcal{A}^t$. Suppose that $W(\cdot, \cdot, \zeta, z) \in D_t$ satisfies (2.76) and (2.77), and $\nabla_x W(t, \cdot, \zeta, z)$ is globally Lipschitz in $x$. Then, $W(t, x, \zeta, z) = J^c(t, x, \tilde{u}^{c, \tau}, \alpha, \zeta, z)$ for the input $\tilde{u}^{c, \tau} = \tilde{u}(r, \tilde{\tau})$ with $\tilde{\tau}$ given by (2.1) with $\tilde{u}(r, x) = -\nabla_x W(t - r, x, \zeta, z)$ and $\zeta_0 = x$. Consequently, $W(t, x, \zeta, z) = \mathcal{W}^{\alpha, c}(t, x, \zeta, z)$.

**Proof.** Let $W$ and $\tilde{u}$ be as asserted. Let $\tilde{u}^{c, \tau} = \tilde{u}(r, \tilde{\tau})$ for all $r \in [0, t]$, and $\tilde{\tau}$ be the corresponding trajectory. Then, by (2.78), we may rewrite (2.76) as

$$0 = -\frac{\partial}{\partial t} W(r, \tilde{\zeta}^{c, \tau}, \zeta, z) - H^\alpha(t - r, \tilde{\zeta}^{c, \tau}, \tilde{u}^{c, \tau}, \zeta, \nabla_x W(r, \tilde{\zeta}^{c, \tau}, \zeta, z))$$

$$= -\frac{\partial}{\partial t} W(r, \tilde{\zeta}^{c, \tau}, \zeta) + \nabla_x W(r, \tilde{\zeta}^{c, \tau}, \zeta, z) \cdot \tilde{u}^{c, \tau} + \frac{1}{2} |\tilde{u}^{c, \tau}|^2 - V^\alpha(t - r, \tilde{\zeta}^{c, \tau}, \zeta_{-r})$$

$$= -\frac{\partial}{\partial t} W(r, \tilde{\zeta}^{c, \tau}, \zeta) + T(\tilde{u}_{-r}) - V^\alpha(t - r, \tilde{\zeta}^{c, \tau}, \zeta_{-r}).$$

Integrating with respect to $r$ over $[0, t]$ yields

$$0 = W(0, \tilde{\zeta}_{t}^{c, \tau}, \zeta, z) - W(t, x, \zeta, z) + \int_{0}^{t} T(\tilde{u}_{r}^{c, \tau}) - V^\alpha(t - r, \tilde{\zeta}^{c, \tau}, \zeta_{-r}) \, dr,$$

or equivalently, by (2.77) and letting $s = t - r$,

$$W(t, x, \zeta, z) = \int_{0}^{t} T(\tilde{u}_{s}^{c, \tau}) - V^\alpha(s, \tilde{\zeta}^{c, \tau}, \zeta_{s}) \, ds + \psi^{c}(\tilde{\zeta}_{t}^{c, \tau}, z)$$

$$= J^c(t, x, \tilde{u}^{c, \tau}, \alpha, \zeta, z),$$

which by Theorem 2.4.9,

$$= \mathcal{W}^{\alpha, c}(t, x, \zeta, z),$$

which completes the proof. \qed
2.5 The fundamental solution in terms of Riccati equation solutions

Given \( c \in [0, \infty), r \leq t, \alpha \in \mathcal{A}^t, \) and \( \zeta \in \mathcal{Z}, \) we look for a solution, \( \mathcal{W}^{\alpha,c} \), of the form

\[
\mathcal{W}^{\alpha,c}(r, x, \zeta, z) \doteq \frac{1}{2} \left[ p^c_r x \cdot x + 2 q^c_r x \cdot z + r^c_r z \cdot z + 2 h^c_r x \cdot x + 2 l^c_r z + \gamma^c_r \right]
\]

(2.79)

where \( p^c, q^c, r^c, h^c, l^c, \) and \( \gamma^c \) depend implicitly on given \( \alpha \) and \( \zeta, \) and satisfy the respective initial value problems:

\[
\begin{align*}
p^c_r &= -[p^c_r]^2 - \sum_{i \in \mathcal{X}} \mu_i [\alpha^i_{t-r}]^3, & p^c_0 &= c, \\
q^c_r &= -p^c_r q^c_r, & q^c_0 &= -c, \\
r^c_r &= -[h^c_r]^2, & r^c_0 &= c, \\
h^c_r &= -p^c_r h^c_r + \sum_{i \in \mathcal{X}} \mu_i [\alpha^i_{t-r}]^3 \zeta^i_{t-s}, & h^c_0 &= 0_{n \times 1}, \\
l^c_r &= -q^c_r h^c_r, & l^c_0 &= 0_{n \times 1}, \\
\gamma^c_r &= -[p^c_r]^2 + \sum_{i \in \mathcal{X}} \mu_i \left\{ 2 \alpha^i_{t-s} - (\alpha^i_{t-s})^3 |\zeta^i_{t-s}|^2 \right\}, & \gamma^c_0 &= 0,
\end{align*}
\]

(2.80)

where \( 0_{m \times k} \) denotes the zero matrix of size \( m \times k. \)

Lemma 2.5.1. Let \( t < \bar{t} \). Then, for any \( \alpha \in \mathcal{A}^t \) and any \( c \in [0, \infty), \) the solution of (2.80) exists on \([0, t).\)

Proof. Let \( \alpha \in \mathcal{A}^t \) and \( c \in [0, \infty). \) Note that since \( \dot{p}^c_r < 0 \) for all \( r \in [0, t], \)

\[
p^c_r \leq p^c_0 = c \quad \forall r \in [0, t].
\]

(2.81)

From (2.60), we have \( \sum_{i \in \mathcal{X}} \mu_i [\alpha^i_r]^3 \leq 2\bar{t}^{-2}, \) and then \( \dot{p}^c_r \geq -[p^c_r]^2 - 2\bar{t}^{-2} \) for all \( r \in [0, t]. \)

Consider

\[
\dot{p}^c_r = -[\dot{p}^c_r]^2 - 2\bar{t}^{-2}, \quad \dot{p}^c_0 = c.
\]

(2.82)

Then, \( \dot{p}^c_r \geq \dot{p}^c_0, \) which implies that

\[
p^c_r \geq \dot{p}^c_r \quad \forall r \in [0, t].
\]

(2.83)

The analytical solution of (2.82) is given by \( \dot{p}^c_r = -\bar{t}^{-1} \tan \left( \bar{t}^{-1} (\hat{c}_1 + r) \right) \) where \( \bar{t} = \bar{t}/\sqrt{2} \) and \( \hat{c}_1 \equiv \bar{t} \tan^{-1}(-c\bar{t}). \) Since \( \tan^{-1}(-c\bar{t}) \in (-\pi/2, 0) \) and \( t < \bar{t}, \) we see that \( \bar{t}^{-1}(\hat{c}_1 + r) = \tan^{-1}(-c\bar{t}) + r\bar{t}^{-1} < \sqrt{2} < \frac{\pi}{4} \) for all \( r \in [0, t], \) which implies that there exists \( \bar{D}_p < \infty \) such that

\[
\dot{p}^c_r \geq -\bar{D}_p \quad \forall r \in [0, t].
\]

(2.84)
Combining (2.81), (2.83) and (2.84), there exist \( \hat{D}_p^0, \hat{D}_p^1 < \infty \) such that \( |p_r^c| < \hat{D}_p^0, |\hat{p}_r^c| < \hat{D}_p^1 \) for all \( r \in [0, t] \), which implies by Picard-Lindelöf theorem (cf. [18]), that there exists a unique solution. The existence and uniqueness of the remaining initial value problems follows easily. \( \square \)

**Theorem 2.5.2.** Let \( t \in [0, \bar{t}) \) and \( c \in [0, \infty) \). Then, \( W^\alpha,c(r, x, \zeta, z) = \tilde{W}^\alpha,c(r, x, \zeta, z) \) for all \( x, z \in \mathbb{R}^3, \zeta \in \mathcal{Z} \), and \( r \in [0, t] \).

**Proof.** It will be sufficient to show that \( \tilde{W}^\alpha,c \) satisfies the conditions of Theorem 2.4.16. Note first that Lemma 2.5.1 implies \( \tilde{W}^\alpha,c(\cdot, \cdot, \zeta, z) \in \mathcal{D}_t \). Also by Lemma 2.5.1, there exists \( D_p < \infty \) such that \( |p_r^c| < D_p \) for all \( r \in [0, t] \). For \( x, \hat{x} \in \mathbb{R}^3 \), note that

\[
|\nabla_x \tilde{W}^\alpha,c(r, x, \zeta, z) - \nabla_x \tilde{W}^\alpha,c(r, \hat{x}, \zeta, z)| \leq |p_r^c||x - \hat{x}| \leq D_p|x - \hat{x}|
\]

which implies that \( \nabla_x W^\alpha,c(t, \cdot, \cdot, z) \) is globally Lipschitz continuous in \( x \).

Let \( \mathcal{P} \doteq \mathbb{R}^{10} \). We define \( \hat{P}_r^c \in \mathcal{P} \) as

\[
\hat{P}_r^c = (p_r^c, q_r^c, r_r^c, (h_r^c)', (l_r^c)', \gamma_r^c)', (2.85)
\]

and accordingly, \( \hat{C} \doteq \hat{P}_0^c \). For \( x, z \in \mathbb{R}^3 \), we define \( X : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathcal{P} \) as

\[
X(x, z) = (x \cdot x, 2x \cdot z, z \cdot z, 2x', 2z', 1)'.
\]

Suppose that \( \hat{\rho}, \hat{\eta} \in \mathcal{P} \) are given by

\[
\hat{\rho} \doteq (\rho_1, \rho_2, \rho_3, \hat{\rho}_4, \hat{\rho}_5, \rho_6)', \quad \hat{\eta} \doteq (\eta_1, \eta_2, \eta_3, \hat{\eta}_4, \hat{\eta}_5, \eta_6)'
\]

(2.87)

where \( \rho_j, \eta_j \in \mathbb{R} \) for \( j \in \{1, 2, 3, 6\} \) and \( \hat{\rho}_j, \hat{\eta}_j \in \mathbb{R}^3 \) for \( j \in \{4, 5\} \). We define \( f : \mathcal{P} \times \mathcal{P} \rightarrow \mathcal{P} \) as

\[
f(\hat{\rho}, \hat{\eta}) \doteq -(\rho_1 \eta_1, \rho_1 \eta_2, \rho_2 \eta_2, \rho_1 \hat{\eta}_4, \rho_2 \hat{\eta}_4, \hat{\rho}_1 \cdot \hat{\eta}_4)',
\]

(2.88)

and for \( \hat{\alpha} = \{\hat{\alpha}^i\}_{i \in \mathcal{N}} \in \mathcal{A} \) and \( Y = \{y^i\}_{i \in \mathcal{N}} \in \mathcal{Y} \), define \( \Gamma : \mathcal{A} \times \mathcal{Y} \rightarrow \mathcal{P} \) as

\[
\Gamma(\hat{\alpha}, Y) \doteq \sum_{i \in \mathcal{N}} \mu_i (-\hat{\alpha}^i)^3, 0, 0, (\hat{\alpha}^i)^3(y^i)', 0_{1 \times n}, Y_i)',
\]

(2.89)

where \( Y_i \doteq \{2\hat{\alpha}^i - [\hat{\alpha}^i] \hat{\alpha}^i | y^i |^2 \} \). Then, we may rewrite (2.79) as

\[
\hat{W}^\alpha,c(r, x, \zeta, z) = \frac{1}{2}X(x, z) \cdot \hat{P}_r^c,
\]

(2.90)

and note that (2.80) is equivalent to
\[ \hat{P}^r_c = f(\hat{P}^r_c, \hat{P}^c) + \Gamma(\alpha_{t-r}, \zeta_{t-r}) \quad \text{with} \quad \hat{P}^0_c = \bar{C}. \quad (2.91) \]

Now, from (2.90), we note that
\[ \frac{\partial}{\partial r} \hat{W}^{\alpha,c}(r, x, \zeta, z) = \frac{1}{2} X(x, z) \cdot \hat{P}^c, \quad (2.92) \]
and with a bit of work, one may verify that
\[ -|\nabla_x \hat{W}^{\alpha,c}(r, x, \zeta, z)|^2 = X(x, z) \cdot f(\hat{P}^r_c, \hat{P}^c). \quad (2.94) \]

Also, collecting like terms, we have
\[ -V^\alpha(t - r, x, \zeta_{t-r}) \]
\[ = \frac{1}{2} \sum_{i \in N} \mu_i \left[ -(\alpha_i^{t-r})^3 x \cdot x + x \cdot (\alpha_i^{t-r})^3 \zeta_i^{t-r} + \{2\alpha_i^{t-r} - (\alpha_i^{t-r})^3 \zeta_i^{t-r}|^2 \} \right], \]
\[ = \frac{1}{2} X(x, z) \cdot \Gamma(\alpha_{t-r}, \zeta_{t-r}), \quad (2.95) \]
where the last equality follows by (2.86) and (2.89). Consequently, substituting (2.92) – (2.95) in the right-hand side of the PDE (2.76) yields
\[ 0 = -\frac{\partial}{\partial r} \hat{W}^{\alpha,c}(r, x, \zeta, z) - H^\alpha(t - r, x, -\nabla_x \hat{W}^{\alpha,c}(r, x, \zeta, z), \zeta, \nabla_x \hat{W}^{\alpha,c}(r, x, \zeta, z)) \]
\[ = \frac{1}{2} X(x, z) \cdot \left[ -\hat{P}^c + f(\hat{P}^r_c, \hat{P}^c) + \Gamma(\alpha_{t-r}, \zeta_{t-r}) \right], \]
which implies (2.79) is a solution of HJB PDE (2.76), and by Theorems 2.4.5 and 2.4.16, \( \hat{W}^{\alpha,c}(r, x, \zeta, z) = \hat{W}^{\alpha,c}(r, x, \zeta, z) \) for all \( r \in [0, t] \), with \( t \in [0, \bar{t}) \).

Recall from Theorems 2.3.2 and 2.4.14 that the fundamental solution of interest is obtained through the \( c \to \infty \) limit of \( \hat{W}^{\alpha,c} \). Consequently, we have that for \( t < \bar{t} \), by Theorems 2.4.14 and 2.5.2,
\[ \bar{W}^\infty(t, x, \zeta, z) = \sup_{\alpha \in \mathcal{A}^t} \lim_{c \to \infty} \frac{1}{2} X(x, z) \cdot \hat{P}^c(\alpha, \zeta) = \sup_{\alpha \in \mathcal{A}^t} \frac{1}{2} X(x, z) \cdot \hat{P}^\infty(t, \alpha, \zeta). \quad (2.96) \]
Further, letting \( \mathcal{G}_t = \{ \hat{P}^\infty(t, \alpha, \zeta) | \alpha \in \mathcal{A}^t \} \), the fundamental solution (2.96) can be represented by
\[ \bar{W}^\infty(t, x, \zeta, z) = \sup_{P \in \mathcal{G}_t} \frac{1}{2} X(x, z) \cdot P. \]
Also note that by the linearity in \( P \) of the expression inside the supremum,
\[ \bar{W}^\infty(t, x, \zeta, z) = \sup_{P \in \mathcal{G}_t} \frac{1}{2} X(x, z) \cdot P. \]
where \( \langle G_t \rangle \) denotes the convex hull of \( G_t \). Further, by Theorem 2.4.14, there exists \( P^* \in \langle G_t \rangle \) such that

\[
P^* = \arg\max_{P \in \langle G_t \rangle} \frac{1}{2} X(x, z) \cdot P = \arg\max_{P \in \partial \langle G_t \rangle} \frac{1}{2} X(x, z) \cdot P
\]

where \( \partial \langle G_t \rangle \) denotes the boundary of \( \langle G_t \rangle \).

### 2.6 The maximization over \( \alpha \)

Recall that by Lemma 2.4.10, \( J^c(t, x, u, \cdot, \zeta, z) \) is strictly concave over \( \alpha \in \mathcal{A}^t \), and consequently, by definition (2.65), so is \( \mathcal{W}^{c}(t, x, \zeta, z) \). Combining this with Theorem 2.4.12, we see that there exists a unique maximizer. Further, by Theorem 2.5.2, \( \mathcal{W}^{\alpha, c} = \tilde{\mathcal{W}}^{\alpha, c} \). Using this last representation, in Corollary 2.6.4 below, we will see that \( \mathcal{W}^{c}(t, x, \zeta, z) \) is Fréchet differentiable. Consequently, in searching for the unique maximum, we may utilize algorithms which require knowledge of the derivative, possibly through first-order necessary conditions. We next obtain this differentiability and an expression for the derivative.

Recall from (2.65) and Theorem 2.5.2 that

\[
\mathcal{W}^{c}(t, x, \zeta, z) = \max_{\alpha \in \mathcal{A}^t} \tilde{\mathcal{W}}^{\alpha, c}(r, x, \zeta, z) = \max_{\alpha \in \mathcal{A}^t} \frac{1}{2} X(x, z) \cdot \hat{P}^c_t(\alpha, \zeta).
\]

Then, letting

\[
\hat{\alpha}^{c,*} = \arg\max_{\alpha \in \mathcal{A}^t} \frac{1}{2} X(x, z) \cdot \hat{P}^c_t(\alpha, \zeta),
\]

by Theorems 2.4.12 and 2.5.2, \( \hat{\alpha}^{c,*} \equiv \tilde{\alpha}^*(\xi^{c,*}, \zeta) \) where \( \xi^{c,*} \) and \( \tilde{\alpha}^* \) are indicated in the statement of Theorem 2.4.12. Further, recall from Theorem 2.4.14 that the fundamental solution of (2.72) has a unique solution. Further, by Theorem 2.4.15, we may see that the solution of \( \mathcal{W}^{c} \) uniformly converges to that of the fundamental solution as \( c \to \infty \).

That is, letting \( \tilde{\alpha}^* \) be the maximizing solution of \( J^{\infty} \) of (2.72), we know from (2.42) that for proper choice of \( \tilde{D}_\alpha = \tilde{D}_\alpha(t, t, \{R_i\}_{i \in \mathcal{N}}) < \infty \),

\[
\|\hat{\alpha}^{c,*} - \alpha^*\|_{L_2(0,t)}^2 \leq \tilde{D}_\alpha \|u^{c,*} - u^*\|_{L_2(0,t)}^2 \leq \frac{\tilde{D}_\alpha \tilde{D}(1 + |x| + |z|)^2}{\sqrt{c}},
\]

where the last inequality follows by Theorem 2.4.15.

We will demonstrate the existence of derivatives of \( \mathcal{W}^{\alpha, c} \) with respect to \( \alpha \). In order to develop a numerical scheme, we will consider the maximizing problem over a finite-dimensional subspace of \( \mathcal{A}^t \). Then, we may equivalently search for the point where
the derivative is zero. If one can compute the derivative efficiently, this may become an efficient means of location the maximum, and therefore solving our problem.

Let $L < \infty$, and suppose $x, z, \zeta_i \in \overline{B}_L(0)$ for all $i \in \mathcal{N}$ and all $r \in [0, t]$. Henceforth, we will work on this compact domain. Accordingly, we define the subset of $\mathcal{Z}$ given by $\mathcal{Z}_L = \{ \zeta \in \mathcal{Z} | |\zeta_i^r| \leq L \forall r \geq 0, \forall i \in \mathcal{N} \}$.

Henceforth, we will work on this compact domain. Accordingly, we define the subset of $\mathcal{Z}$ given by $\mathcal{Z}_L = \{ \zeta \in \mathcal{Z} | |\zeta_i^r| \leq L \forall r \geq 0, \forall i \in \mathcal{N} \}$.

We assume that for $t < \bar{t}$, there exists $x, z \in \overline{B}_L(0); \zeta \in \mathcal{Z}_L; \bar{c} = \bar{c}(t, x, z) < \infty$, and $\bar{c} = \bar{c}(t, x, z) < \infty$ such that for any $\varepsilon \in (0, \bar{\varepsilon}]$, $c > \bar{c}$ and any $\varepsilon$--optimal in the definition (2.11) of $W^c$, we have

$$|\xi_{c}^{r} - \zeta_{c}^r| > R_i, \quad \forall r \in [0, t], \forall i \in \mathcal{N} \quad (A.N1)$$

where $\xi^{c}$ denotes the corresponding trajectory.

### 2.6.1 Derivative of $W^{\alpha,c}$ with respect to $\alpha$

We first note some simple miscellaneous bounds that will be used below. For $\hat{\rho}, \hat{\eta} \in \mathcal{P}$ given as (2.87), using the bilinearity of $f$ in (2.88), we define

$$\ell_1(\hat{\eta}) = f_{\hat{\rho}}(\hat{\rho}, \hat{\eta}), \quad \ell_2(\hat{\rho}) = f_{\hat{\eta}}(\hat{\rho}, \hat{\eta}), \quad (2.97)$$

where the subscripts on $f$ denote differentiation, and $\ell_1, \ell_2$ are introduced to emphasize the dependence on only one variable each. Then, from the definition of $f$, we have

$$\|\ell_1(\hat{\eta})\|_F = [|\eta_1|^2 + 2|\eta_2|^2 + 3|\eta_4|^2]^{1/2} \leq \sqrt{3}|\hat{\eta}|, \quad (2.98)$$

$$\|\ell_2(\hat{\rho})\|_F = [5|\rho_1|^2 + 4|\rho_2|^2 + |\bar{\rho}_4|^2]^{1/2} \leq \sqrt{5}|\hat{\rho}|. \quad (2.99)$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Further, note that for $\sigma, \omega \in \mathcal{P}$,

$$f(\sigma, \sigma) - f(\omega, \omega) = [\ell_1(\sigma) + \ell_2(\omega)](\sigma - \omega), \quad (2.100)$$

$$\|\ell_1(\sigma) - \ell_1(\omega)\|_F \leq \sqrt{3}|\sigma - \omega|. \quad (2.101)$$

By Lemma 2.5.1, given $t < \bar{t}$, $c \in [0, \infty)$ and $\zeta \in \mathcal{Z}_L$, we may choose $K_p = K_p(c, t, \zeta) < \infty$ such that

$$|\hat{P}_c^{r}(\alpha, \zeta)| \leq K_p \quad \forall r \in [0, t], \alpha \in \mathcal{A}^t \quad (2.102)$$

where $\hat{P}_c^{r}(\cdot, \zeta)$ are the solutions of (2.91). Further, combining (2.98), (2.99) and (2.102) yields

$$\|\ell_1(\hat{P}_c^{r}(\alpha, \zeta))\|_F, \|\ell_2(\hat{P}_c^{r}(\alpha, \zeta))\|_F \leq \sqrt{5}K_p = K_1 \quad \forall r \in [0, t], \alpha \in \mathcal{A}^t. \quad (2.103)$$
Also, by examining (2.89), we note that for \( \tilde{\alpha} = \{\tilde{\alpha}_i\}_{i \in \mathcal{N}} \in \mathcal{A} \) and \( Y \in \mathcal{Y}_L = \{(y^j)_{i \in \mathcal{N}} \in \mathcal{Y} \mid |y^j| \leq L, \forall i \in \mathcal{N}\}, \)
\[
\|\Gamma_\alpha(\tilde{\alpha}, Y)\|_F^2 = 9 \sum_{i \in \mathcal{N}} [((\tilde{\alpha}_i)^2]^2 \{1 + |y^i|^2 + |y^i|^4\},
\]
where the subscripts on \( \Gamma \) indicate differentiation with respect to the first variable \( \alpha \). Then, by the definitions of \( \mathcal{A} \) and \( \mathcal{Y}_L \), we see that there exists \( \hat{K}_\gamma^1 = K_\gamma^1(\{R_i\}_{i \in \mathcal{N}}, L) \leq \infty \) such that
\[
\|\Gamma_\alpha(\tilde{\alpha}, Y)\|_F \leq K_\gamma^1 \quad \forall \tilde{\alpha} \in \mathcal{A}, \ Y \in \mathcal{Y}_L. \tag{2.104}
\]
Further, for \( \tilde{\alpha}_1 = \{\tilde{\alpha}_1^i\}_{i \in \mathcal{N}}, \tilde{\alpha}_2 = \{\tilde{\alpha}_2^i\}_{i \in \mathcal{N}} \in \mathcal{A} \) and \( Y \in \mathcal{Y}_L, \)
\[
\|\Gamma_\alpha(\tilde{\alpha}_1, Y) - \Gamma_\alpha(\tilde{\alpha}_2, Y)\|_F^2 = 9 \sum_{i \in \mathcal{N}} [((\tilde{\alpha}_1^i)^2 - (\tilde{\alpha}_2^i)^2]^2 \{1 + |y^i|^2 + |y^i|^4\},
\]
which implies, by the definitions of \( \mathcal{A} \) and \( \mathcal{Y}_L \), that there exists \( \hat{K}_\gamma = \hat{K}_\gamma(\{R_i\}_{i \in \mathcal{N}}, L) \leq \infty \) such that this is
\[
\leq [\hat{K}_\gamma]^2 \sum_{i \in \mathcal{N}} |\tilde{\alpha}_1^i - \tilde{\alpha}_2^i|^2 = \left\{\hat{K}_\gamma |\tilde{\alpha}_1 - \tilde{\alpha}_2|\right\}^2. \tag{2.105}
\]

Lemma 2.6.1. For \( t < \tilde{t}, c \in [0, \infty), \tilde{\alpha}, \hat{\alpha} \in \mathcal{A}^t \) and \( \zeta \in \mathcal{L}_L \), there exists \( \hat{C}_1 < \infty \) such that
\[
|\hat{P}_r^c(\tilde{\alpha}, \zeta) - \hat{P}_r^c(\hat{\alpha}, \zeta)| \leq \hat{C}_1 \|	ilde{\alpha} - \hat{\alpha}\|_{L_2(0,t)}
\]
for all \( r \in [0,t] \) where \( \hat{P}_r^c(\tilde{\alpha}, \zeta) \) and \( \hat{P}_r^c(\hat{\alpha}, \zeta) \) are the solutions of (2.91) with \( \alpha = \tilde{\alpha} \) and \( \alpha = \hat{\alpha} \), respectively.

Proof. By (2.91) and the triangle inequality,
\[
|\hat{P}_r^c(\tilde{\alpha}, \zeta) - \hat{P}_r^c(\hat{\alpha}, \zeta)| \leq \int_0^t |f(\hat{P}_r^c(\tilde{\alpha}, \zeta), \hat{P}_r^c(\hat{\alpha}, \zeta)) - f(\hat{P}_r^c(\tilde{\alpha}, \zeta), \hat{P}_r^c(\hat{\alpha}, \zeta))| \nonumber \\
+ |\Gamma(\tilde{\alpha}_{t-\nu}, \zeta_{t-\nu}) - \Gamma(\hat{\alpha}_{t-\nu}, \zeta_{t-\nu})| d\nu,
\]
which by (2.100) and (2.104),
\[
\leq \int_0^t \left[\|\ell_1(\hat{P}_r^c(\tilde{\alpha}, \zeta))\|_F + \|\ell_2(\hat{P}_r^c(\hat{\alpha}, \zeta))\|_F \right] \nonumber \\
\cdot |\hat{P}_r^c(\tilde{\alpha}, \zeta) - \hat{P}_r^c(\hat{\alpha}, \zeta)| + K_\gamma |\tilde{\alpha}_{t-\nu} - \hat{\alpha}_{t-\nu}| d\nu,
\]
which by (2.103) and Hölder’s inequality,
\[
\leq 2K_1^1 \int_0^t |\hat{P}_r^c(\tilde{\alpha}, \zeta) - \hat{P}_r^c(\hat{\alpha}, \zeta)| d\nu + K_\gamma \sqrt{t} \|	ilde{\alpha} - \hat{\alpha}\|_{L_2(0,t)}.
\]
Using Gronwall’s inequality, this implies
\[
|\hat{P}_r^c(\hat{\alpha}, \zeta) - \hat{P}_r^c(\hat{\alpha}, \zeta)| \leq K_1 \sqrt{t} \exp(2K_1^t) \|\hat{\alpha} - \hat{\alpha}\|_{L_2(0,t)}
\]
for all \( r \in [0,t] \).

\[\square\]

**Lemma 2.6.2.** Let \( c \in [0, \infty) \). Then \( W^{\alpha,c}(t, x, \zeta, z) \) is Lipschitz continuous in \( \alpha \) on bounded subsets of \([0, \hat{t}] \times \mathbb{R}^3 \times \mathbb{Z}_L \times \mathbb{R}^3\).

**Proof.** Let \( t \in (0, \hat{t}); x, z \in \mathcal{B}_L(0) \) and \( \zeta \in \mathbb{Z}_L \). Then, by Theorem 2.5.2,
\[
|W^{\alpha,c}(t, x, \zeta, z) - W^{\hat{\alpha},c}(t, x, \zeta, z)| \leq \frac{1}{2} |X(x, z)| |\hat{P}_r^c(\hat{\alpha}, \zeta) - \hat{P}_r^c(\hat{\alpha}, \zeta)|,
\]
which by Lemma 2.6.1,
\[
\leq \hat{K}_\alpha \|\hat{\alpha} - \hat{\alpha}\|_{L_2(0,t)}
\]
for proper choice of \( \hat{K}_\alpha = \hat{K}_\alpha(\hat{C}_1, L) < \infty \).

\[\square\]

Letting \( \mathcal{A}_o \) denote the interior of \( \mathcal{A} \), we define \( \mathcal{A}_o^t \equiv L_\infty([0, \hat{t}] ; \mathcal{A}_o) \). Given \( \zeta \in \mathbb{Z}_L \) and \( c \in [0, \infty) \), we will obtain a representation for the derivative of \( \hat{P}_r^c(\alpha, \zeta) \) with respect to \( \alpha \in \mathcal{A}_o^t \). For \( s \in [0, t] \) and \( i \in \mathcal{N} \), consider
\[
\pi_{s,i}^r \equiv \frac{d\pi_{s,i}^r}{dr} = f(\pi_{s,i}^r, \hat{P}_r^c) + f(\hat{P}_r^c, \pi_{s,i}^r) + \Gamma_{\alpha,i}(\alpha_{t-r}, \zeta_{t-r}) \tag{2.106}
\]
for all \( r \in (s, t) \) with \( \pi_{s,i}^0 = 0_{M \times 1} \) where
\[
\Gamma_{\alpha,i}(\alpha_{t-r}, \zeta_{t-r}) \equiv \frac{\partial \Gamma(\alpha_{t-r}, \zeta_{t-r})}{\partial \alpha^i}. \tag{2.107}
\]
The following lemma demonstrates the desired representation for the derivative.

**Lemma 2.6.3.** Given \( \alpha \in \mathcal{A}_o^t \), let \( h \in \mathcal{A}_o^t \) such that \( \alpha + h \in \mathcal{A}_o^t \). Let \( c \in [0, \infty) \) and \( \zeta \in \mathbb{Z}_L \). Then, there exists \( \hat{C}_2 < \infty \) such that letting \( \hat{P}_{h,r}^c = \hat{P}_r^c(\alpha + h, \zeta) \),
\[
\left| \hat{P}_{h,r}^c - \hat{P}_r^c - \int_0^r \left( -\frac{d\pi_{s,i}^r}{ds} \right) h_{t-s} \, ds \right| \leq \hat{C}_2 \|h\|_{L_2(0,t)}^2 \tag{2.108}
\]
for all \( r \in [0,t] \) where \( \hat{P}_r^c(\cdot, \zeta) \) is the solution of (2.91).

**Proof.** Using the integral form of (2.106) and its initial condition, for \( r \in [s, t) \),
\[
\pi_{s,i}^r = \int_s^r f(\pi_{\nu,i}^r, \hat{P}_r^c) + f(\hat{P}_r^c, \pi_{\nu,i}^r) + \Gamma_{\alpha,i}(\alpha_{t-\nu}, \zeta_{t-\nu}) \, d\nu.
\]
Differentiating, and using (2.97), we have
\[
\frac{d\pi^s}{ds} = -\Gamma_\alpha(\alpha_{t-s}, \zeta_{t-s}) + \int_s^r \left[ \ell_1(\hat{P}^c_\nu) + \ell_2(\hat{P}^c_\nu) \right] \frac{d\pi^s}{ds} d\nu. \tag{2.109}
\]

Letting $\Delta \hat{P}^c_\nu = \hat{P}^c_{h,r} - \hat{P}^c_{r}$, we define
\[
\phi_r = \Delta \hat{P}^c_r - \int_0^r \frac{d\pi^s}{ds} h_{t-s} ds \quad \forall r \in [0, t]. \tag{2.110}
\]

Letting $\Delta \Gamma_{t-r} = \Gamma(\alpha_{t-r} + h_{t-r}, \zeta_{t-r}) - \Gamma(\alpha_{t-r}, \zeta_{t-r})$, we note that by (2.91), differentiation $\Delta \hat{P}^c_r$ with respect to $r$ is given by
\[
\Delta \hat{P}^c_r = \hat{P}^c_{h,r} - \hat{P}^c_r = f(\hat{P}^c_{h,r}, \hat{P}^c_{r}) - f(\hat{P}^c_{r}, \hat{P}^c_{r}) + \Delta \Gamma_{t-r},
\]
which by (2.100),
\[
= [\ell_1(\hat{P}^c_r) + \ell_2(\hat{P}^c_r)] \Delta \hat{P}^c_r + \Delta \Gamma_{t-r}. \tag{2.111}
\]

Also, note that by (2.109),
\[
\frac{d}{dr} \int_0^r -\frac{d\pi^s}{ds} h_{t-s} ds = \int_0^r \Gamma_\alpha(\alpha_{t-s}, \zeta_{t-s}) h_{t-s} ds - \frac{d}{dr} \int_0^r \left[ \ell_1(\hat{P}^c_\nu) + \ell_2(\hat{P}^c_\nu) \right] \frac{d\pi^s}{ds} d\nu h_{t-s} ds,
\]
which, using the Leibniz integral rule,
\[
= \Gamma_\alpha(\alpha_{t-r}, \zeta_{t-r}) h_{t-r} - \int_0^r \frac{d}{dr} \int_s^r \left[ \ell_1(\hat{P}^c_\nu) + \ell_2(\hat{P}^c_\nu) \right] \frac{d\pi^s}{ds} d\nu h_{t-s} ds
\]
\[
= \Gamma_\alpha(\alpha_{t-r}, \zeta_{t-r}) h_{t-r} + \int_0^r \left[ \ell_1(\hat{P}^c_\nu) + \ell_2(\hat{P}^c_\nu) \right] \frac{d\pi^s}{ds} h_{t-s} ds. \tag{2.112}
\]

Next, differentiating (2.110) with respect to $r$ yields
\[
\dot{\phi}_r = \Delta \dot{\hat{P}}^c_r - \frac{d}{dr} \int_0^r -\frac{d\pi^s}{ds} h_{t-s} ds, \tag{2.113}
\]
which by (2.111) and (2.112),
\[
= [\ell_1(\hat{P}^c_{h,r}) + \ell_2(\hat{P}^c_{r})] \Delta \hat{P}^c_r
\]
\[
- [\ell_1(\hat{P}^c_r) + \ell_2(\hat{P}^c_r)] \int_0^r -\frac{d\pi^s}{ds} h_{t-s} ds + \Delta \Gamma_{t-r} - \Gamma_\alpha(\alpha_{t-r}, \zeta_{t-r}) h_{t-r}
\]
\[
= [\ell_1(\hat{P}^c_r) + \ell_2(\hat{P}^c_r)] \phi_r + [\ell_1(\hat{P}^c_{r,h}) - \ell_1(\hat{P}^c_{r})] \Delta \hat{P}^c_r + \Delta \Gamma_{t-r} - \Gamma_\alpha(\alpha_{t-r}, \zeta_{t-r}) h_{t-r}
\]
where the last equality follows by (2.110). Note that by (2.101),
\[
\int_0^r \left[ \ell_1(\hat{P}^c_{r,h}) - \ell_1(\hat{P}^c_{r}) \right] \Delta \hat{P}^c_r \nu \leq \sqrt{3} \int_0^r |\Delta \hat{P}^c_r|^2 d\nu,
\]
which by Lemma 2.6.1,
for proper choice of $C_2 < \infty$.

Note that by the integral mean value theorem (cf., Ch. 9 in [29]),

$$\Delta \Gamma_{t-\nu} = \Gamma(\alpha_{t-\nu} + h_{t-\nu}, \zeta_{t-\nu}) - \Gamma(\alpha_{t-\nu}, \zeta_{t-\nu}) = \left[ \int_0^1 \Gamma_\alpha(\alpha_{t-\nu} + s h_{t-\nu}) \, ds \right] h_{t-\nu}$$

for all $\nu \in [0, t]$. Then,

$$|\Delta \Gamma_{t-\nu} - \Gamma_\alpha(\alpha_{t-\nu}, \zeta_{t-\nu})h_{t-\nu}| \leq \int_0^1 \|\Gamma_\alpha(\alpha_{t-\nu} + s h_{t-\nu}) - \Gamma_\alpha(\alpha_{t-\nu}, \zeta_{t-\nu})\|_F \, ds \, |h_{t-\nu}|$$

$\leq \frac{1}{2} \tilde{K}_\gamma \|h\|_2 |h_{t-\nu}|^2,$

where the last bound follows by (2.105). This implies that

$$\int_0^t |\Delta \Gamma_{t-\nu} - \Gamma_\alpha(\alpha_{t-\nu}, \zeta_{t-\nu})h_{t-\nu}| \, d\nu \leq \frac{1}{2} \tilde{K}_\gamma \int_0^t |h_{t-\nu}|^2 \, d\nu$$

$$= \frac{1}{2} \tilde{K}_\gamma \|h\|_2^2 |h_{t-\nu}|^2,$$

Substituting (2.114) and (2.115) into the integration of (2.113) yields

$$\phi_r \leq \int_0^r 2K_1^2 |\phi_\nu| \, d\nu + (C_2 + C_3) \|h\|_2^2 |h_{t-\nu}|$$

By Gronwall’s inequality, this implies

$$\phi_r \leq (C_2 + C_3)t \exp(2K_1 t) \|h\|_2^2 |h_{t-\nu}|$$

for all $r \in [0, t]$.

**Corollary 2.6.4.** Let $c \in [0, \infty)$ and $t \in (0, \hat{t})$. Then, $W^{\alpha,c}(t, x, \zeta, z)$ is Fréchet differentiable with respect to $\alpha \in A^t_0$ on bounded subsets of $\mathbb{R}^3 \times Z_L \times \mathbb{R}^3$.

**Proof.** Let $x, z \in \overline{B}_L(0)$ and $\zeta \in Z_L$. Given $\alpha \in A^t_0$, let $h \in A^t_0$ such that $\alpha + h \in A^t_0$. Let

$$\phi^W_r \equiv W^{\alpha+h,c}(r, x, \zeta, z) - W^{\alpha,c}(r, x, \zeta, z) - \frac{1}{2} X(x, z) \cdot \int_0^t \left( -\frac{d\pi_r^s}{ds} \right) h_{t-s} \, ds,$$

where $X$ is given in (2.86), and which by Theorem 2.5.2 and (2.90),

$$= \frac{1}{2} X(x, z) \cdot \hat{P}_r^c(\alpha + h, \zeta) - \frac{1}{2} X(x, z) \cdot \hat{P}_r^c(\alpha, \zeta) - \frac{1}{2} X(x, z) \cdot \int_0^t \left( -\frac{d\pi_r^s}{ds} \right) h_{t-s} \, ds.$$

Then, by Lemma 2.6.3,

$$|\phi^W_r| \leq \frac{1}{2} |X(x, z)| C_2 \|h\|_2^2 |h_{t-\nu}|,$$

which completes the proof. □
2.6.2 An approximate solution

We will consider piecewise constant potential energy controls rather than the $L_\infty$ elements of $A^t$, as this will allow us to compute numerically the derivatives of interest.

Let $t \in (0, \bar{t})$ where $\bar{t}$ is as per (2.60). Let $K \in \mathbb{N}$ denote the number of time intervals contained in the given time duration $[0, t]$, and $A^K$ denote the $K^{th}$ Cartesian power of $A$. We may choose a norm on $\hat{A}^K$ as

$$\|\hat{\alpha}\|_2 = \sqrt{\frac{t}{K} \sum_{k \in \mathcal{K}} |\hat{\alpha}_k|^2}^{1/2} \text{ for } \hat{\alpha} = \{\hat{\alpha}_k\}_{k \in \mathcal{K}} \in \hat{A}^K. \quad (2.116)$$

Let $\tau = t/K$ denote the length of each time interval. Letting $\tau_0 = 0$, we define

$$\tau_k = k\tau \quad \text{and} \quad I_k = (\tau_{k-1}, \tau_k) \quad \forall k \in \mathcal{K} = [1, K].$$

We define the set of piecewise constant functions defined on $[0, t)$ relative to grid $\{\tau_k | k \in [0, K]\}$ as

$$A^t_K = \{\alpha \in A^t | \forall k \in \mathcal{K}, \exists \hat{\alpha}_k \in A \text{ s.t. } \alpha_r = \hat{\alpha}_k \forall r \in I_k\}.$$ 

Further, we define the one-to-one and onto linear mapping $L^K : \hat{A}^K \to A^t_K$ such that for $\hat{\alpha} = \{\hat{\alpha}_k\}_{k \in \mathcal{K}} \in \hat{A}^K$, letting $\hat{\alpha} = L^K(\hat{\alpha})$,

$$\hat{\alpha}_r = \hat{\alpha}_k \quad \forall r \in I_k, \quad \forall k \in \mathcal{K}. \quad (2.117)$$

From Lemma 2.4.10 and the definition of $L^K(\cdot)$, one immediately obtains the following:

**Lemma 2.6.5.** Let $K \in \mathbb{N}$. For all $t > 0$, $c \in [0, \infty); x, z \in \mathbb{R}^3$, $\zeta \in Z_L$, and $u \in U^\infty$, $\mathcal{W}^{L^K(\hat{\alpha}), c}(t, x, \zeta, z)$ and $J^c(t, x, u, L^K(\hat{\alpha}), \zeta, z)$ are strictly concave in $\hat{\alpha} \in \hat{A}^K$.

**Lemma 2.6.6.** Let $t < \bar{t}$ and $K \in \mathbb{N}$. Let $c \in [0, \infty); x, z \in \overline{B}_L; \zeta \in Z_L$. Then, for any $\tilde{u} \in U^\infty$,

$$\sup_{\hat{\alpha} \in \hat{A}^K} J^c(t, x, \tilde{u}, L^K(\hat{\alpha}), \zeta, z) = \max_{\hat{\alpha} \in \hat{A}^K} J^c(t, x, \tilde{u}, L^K(\hat{\alpha}), \zeta, z).$$

Further, letting $\hat{\xi}$ be the trajectory corresponding to $\tilde{u}$,

$$\hat{\alpha}^* = \{\hat{\alpha}_k^*\}_{k \in \mathcal{K}} \doteq \argmax_{\hat{\alpha} \in \hat{A}^K} J^c(t, x, \tilde{u}, L^K(\hat{\alpha}), \zeta, z)$$

where

$$\hat{\alpha}_k^* = \sqrt{2/3} \min \left\{R_i^{-1}, \sqrt{\tau} \|\zeta^i - \zeta^j\|_{L^2(\tau_{k-1}, \tau_k)}^{-1} \right\}$$

for all $i \in \mathcal{N}$ and $k \in \mathcal{K}$.
Proof. Let \( \tilde{u} \in U^\infty \) and \( \tilde{\zeta} \) be the corresponding trajectory. By the independent sum of integrals over the segments,

\[
\sup_{\alpha \in A^K} J^c(t, x, \tilde{u}, \mathcal{L}^K(\tilde{\alpha}), \zeta, z) = \frac{1}{2} \|	ilde{u}\|^2_{L^2(0, t)} + \psi^c(\tilde{\zeta}, z) + \sup_{\alpha \in A^K} \sum_{k \in \mathcal{K}} \int_{I_k} -V L^K(\tilde{\alpha})(r, \tilde{\zeta}_r, \zeta_r) \, dr
\]

\[
= \frac{1}{2} \|	ilde{u}\|^2_{L^2(0, t)} + \psi^c(\tilde{\zeta}, z) + \sum_{(i, k) \in \mathcal{N} \times \mathcal{K}} \sup_{\alpha_i \in \{0, \sqrt{2/3} R_i^{-1}\}} \sup_{\tilde{\zeta}} V_i^k(\tilde{\alpha}, \tilde{\zeta}, \zeta)
\]

where

\[
V_i^k(\tilde{\alpha}, \tilde{\zeta}, \zeta) \triangleq \int_{I_k} \mu_i \left[ \tilde{\alpha}_i^k - \frac{3}{2} |\tilde{\alpha}_i^k|^2 |\tilde{\zeta}_r - \zeta_i^r|^2 \right] \, dr.
\]

Note that

\[
\frac{d}{d\tilde{\alpha}_i^k} V_i^k(\tilde{\alpha}, \tilde{\zeta}, \zeta) = \mu_i \left[ \tau - \frac{3}{2} |\tilde{\alpha}_i^k|^2 \int_{I_k} |\tilde{\zeta}_r - \zeta_i^r|^2 \, dr \right].
\]

From this, it is not difficult to show that

\[
\tilde{\alpha}_{i, k}^* = \tilde{a}_{i, k}^*(\tilde{\zeta}, \zeta) \triangleq \arg\max_{\alpha_i^k \in \{0, \sqrt{2/3} R_i^{-1}\}} V_i^k(\tilde{\alpha}, \tilde{\zeta}, \zeta)
\]

\[
= \sqrt{2/3} \min \left\{ R_i^{-1}, \sqrt{\tau} \|\tilde{\zeta}_r - \zeta_i^r\|_{L^2(t_{\tau_k}, \tau_k)}^{-1} \right\}.
\]  \( \Box \)

Let \( \ddot{\alpha} : \mathbb{R}^3 \times \mathbb{R}^{3\mathcal{N}} \to \ddot{A}^\mathcal{K} \) be given by

\[
\ddot{\alpha}(\cdot, \cdot) \triangleq \{\ddot{\alpha}_{i, k}(\cdot, \cdot)\}_{k \in \mathcal{K}}
\]  \( \text{(2.119)} \)

where \( \ddot{\alpha}_{i, k}(\cdot, \cdot) \triangleq \{\ddot{a}_{i, k}(\cdot, \cdot)\}_{i \in \mathcal{N}} \) and \( \ddot{a}_{i, k} \) is given by (2.118) for all \( i \in \mathcal{N} \) and \( k \in \mathcal{K} \).

Lemma 2.6.7. Let \( t < \bar{t} \) and \( K \in \mathcal{N} \). Let \( x, z \in \mathbb{R}^3; \zeta \in \mathbb{Z}_L \) and \( c \in [0, \infty) \). Then, there exists a unique stationary point of \( J^c(t, x, \cdot, \mathcal{L}^K(\cdot), \zeta, z) \) over \( U^\infty \times \ddot{A}^\mathcal{K} \).

Proof. Recall that by Theorem 2.4.5 and Lemma 2.4.4, \( J^c(t, x, u, \mathcal{L}^K(\cdot), \zeta, z) \) is continuous, coercive and strictly convex in \( u \). Then, by (2.117), it is easy to see that \( \max_{\tilde{\alpha} \in A^K} J^c(t, x, u, \mathcal{L}^K(\tilde{\alpha}), \zeta, z) \) is also continuous, coercive and strictly convex in \( u \), where the existence of the maximum follows from Lemma 2.6.6. This guarantees the existence of the unique optimal velocity control,

\[
\hat{\tilde{u}}^c = \arg\min_{u \in U^\infty} \max_{\tilde{\alpha} \in \ddot{A}^\mathcal{K}} J^c(t, x, u, \mathcal{L}^K(\tilde{\alpha}), \zeta, z).
\]
This implies
\[
\min_{u \in U} \max_{\tilde{\alpha} \in \tilde{A}^K} J^c(t, x, u, L^K(\tilde{\alpha}), \zeta, z) = \max_{\tilde{\alpha} \in \tilde{A}^K} \min_{u \in U} J^c(t, x, u, L^K(\tilde{\alpha}), \zeta, z) = J^c(t, x, \tilde{\alpha}^{c, *}, L^K(\tilde{\alpha}^{c, *}), \zeta, z) \tag{2.120}
\]
where \(\tilde{\alpha}\) is given by (2.119) and \(\tilde{\alpha}^{c, *}\) denotes the trajectory corresponding to \(\tilde{\alpha}^{c, *}\). Then, by an argument similar to that of the proof of Theorem 2.4.9 and we do not include the repetitive details), letting \(\tilde{\alpha}^{c, *} \doteq \tilde{\alpha}(\tilde{\alpha}^{c, *}, \zeta)\),
\[
J^c(t, x, \tilde{\alpha}^{c, *}, L^K(\tilde{\alpha}^{c, *}), \zeta, z) = \min_{u \in U} J^c(t, x, u, L^K(\tilde{\alpha}^{c, *}), \zeta, z). \tag{2.121}
\]
Combining (2.120) and (2.121) yields in an analogous fashion to the proof of Theorem 2.4.12 that
\[
\min_{u \in U} \max_{\tilde{\alpha} \in \tilde{A}^K} J^c(t, x, u, L^K(\tilde{\alpha}), \zeta, z) = \max_{\tilde{\alpha} \in \tilde{A}^K} \min_{u \in U} J^c(t, x, u, L^K(\tilde{\alpha}), \zeta, z), \tag{2.122}
\]
and \((\tilde{\alpha}^{c, *}, \tilde{\alpha}^{c, *})\) is the unique solution of (2.122) over \(U^\infty \times \tilde{A}^K\).

\[\blacksquare\]

**Remark 2.6.8.** Let \(t > 0\) and \(K \in \mathbb{N}\). For \(\tilde{\alpha} = \{\tilde{\alpha}_k\}_{k \in K} \in \tilde{A}^K\), letting \(\hat{\alpha} = L^K(\tilde{\alpha})\),
\[
\|\hat{\alpha}\|^2_{L^2(0, t)} = \int_0^t |\hat{\alpha}_r|^2 dr = \sum_{k \in K} \int_{I_k} |\tilde{\alpha}_r|^2 dr = t \sum_{k \in K} |\tilde{\alpha}_k|^2 = \|\tilde{\alpha}\|^2_2,
\]
where the last equality follows by (2.116). Therefore, \(L^K\) is an isomorphism between two normed spaces, \((\tilde{A}^K, \|\cdot\|_2)\) and \((\mathcal{A}^K_t, \|\cdot\|_{L^2(0, t)})\) (cf., [23]). Further, by the linearity of \(L^K\), for \(\tilde{\alpha}, \hat{\alpha} \in \tilde{A}^K\),
\[
\|L^K(\tilde{\alpha}) - L^K(\hat{\alpha})\|_{L^2(0, t)} = \|L^K(\tilde{\alpha} - \hat{\alpha})\|_{L^2(0, t)} = \|\tilde{\alpha} - \hat{\alpha}\|_2.
\]

Since \(\tilde{A}^K\) and \(\mathcal{A}^K_t\) are isomorphic, the next corollaries follow immediately from Lemmas 2.6.1 and 2.6.2, respectively.

**Corollary 2.6.9.** For \(t \leq \bar{t}\), \(c \in [0, \infty)\), and \(\zeta \in \mathcal{Z}_L\), there exists \(\tilde{C}_1 < \infty\) such that
\[
|\tilde{P}^c_r(L^K(\tilde{\alpha}), \zeta) - \tilde{P}^c_r(L^K(\hat{\alpha}), \zeta)| \leq \tilde{C}_1 \|\tilde{\alpha} - \hat{\alpha}\|_2
\]
for all \(r \in [0, \bar{t}]\) where \(\tilde{P}^c(\cdot, \zeta)\) is the solution of (2.91).

**Corollary 2.6.10.** Let \(c \in [0, \infty)\) and \(K \in \mathbb{N}\). \(W^{L^K(\tilde{\alpha}), c}(t, x, \zeta, z)\) is Lipschitz continuous in \(\tilde{\alpha} \in \tilde{A}^K\) on bounded subsets of \([0, \bar{t}] \times \mathbb{R}^3 \times \mathcal{Z}_L \times \mathbb{R}^3\).
2.6.3 Error analysis

Computationally, we can approximate the maximum over \( \mathcal{A}^t \) (equivalently, \( \hat{\mathcal{A}}^{t} \)), which yields \( \hat{W}^e \), by taking a maximum over \( \mathcal{A}_K^t \), which due to the above-noted isomorphism, is equivalent to a maximum over \( \mathcal{A}^K \). Letting

\[
\hat{W}_K^e(t, x, \zeta, z) \triangleq \max_{\alpha \in \mathcal{A}^K} \mathcal{W}^{\mathcal{L}(\hat{\alpha})} \cdot c(t, x, z, \zeta) = \min_{u \in \mathcal{U}} \max_{\alpha \in \mathcal{A}^K} \mathcal{J}^c(t, x, u, \mathcal{L}(\hat{\alpha}), \zeta), \tag{2.123}
\]

for all \( t \in [0, T] \); \( x, \zeta \in \mathbb{R}^d \); \( \zeta \in \mathcal{Z}_L \), we will demonstrate that \( \hat{W}_K^e \to \hat{W}^e \) as \( K \to \infty \).

Importantly, we will also demonstrate that the optimal velocity controls converge as \( K \to \infty \), which implies that the optimal trajectories converge.

**Lemma 2.6.11.** Let \( t < T \); \( L < \infty \); \( x, \zeta \in \overline{B}_L(0) \), \( c \in [0, \infty) \) and \( \zeta \in \mathcal{Z}_L \). Then,

\[
\hat{W}_K^e(t, x, \zeta, z) = \lim_{K \to \infty} \hat{W}_K^e(t, x, \zeta, z) = \lim_{K \to \infty} \max_{\alpha \in \mathcal{A}^K} \mathcal{W}^{\mathcal{L}(\hat{\alpha})} \cdot c(t, x, \zeta, z). \tag{2.124}
\]

Letting \( \alpha^* = \arg\max_{\alpha \in \mathcal{A}^t} \mathcal{W}^{\alpha} \cdot c(t, x, \zeta, z) \) and \( \hat{\alpha}_K^* = \arg\max_{\alpha \in \mathcal{A}^K} \mathcal{W}^{\mathcal{L}(\hat{\alpha})} \cdot c(t, x, \zeta, z) \) for \( K \in \mathbb{N} \),

\[
\hat{W}_K^e(t, x, \zeta, z) - \hat{W}_K^e(t, x, \zeta, z) \leq K_\alpha \| \alpha^* - \mathcal{L}(\hat{\alpha}_K^*) \|_{L_2(0,t)},
\]

and

\[
\mathcal{L}(\hat{\alpha}_K^*) \to \alpha^* \quad \text{as} \quad K \to \infty.
\]

**Proof.** Given \( K \in \mathbb{N} \) and \( \alpha \in \hat{\mathcal{A}}^t \), let

\[
\beta_i^b(K, \alpha) \triangleq \frac{1}{t/K} \int_{I_k} \alpha^i \, dp \quad \forall i \in N \quad \forall k \in K. \tag{2.125}
\]

For \( r \in [0, t] \), let \( \hat{\alpha}_K = \{ \hat{\alpha}_K \} \in \mathcal{A}_K^t \) such that

\[
\hat{\alpha}_k^i(r) = \sum_{k \in K} \beta_i^b(K, \alpha^*) \mathbf{1}_{I_k}(r) \quad \forall i \in N \quad \tag{2.126}
\]

where \( \mathbf{1} \) denotes an indicator function. Note that for any \( r \in [0, t] \) and \( K \in \mathbb{N} \), there exist \( k = k(r, K) \in K \) and \( \delta^+_K = \delta^+_K(r), \delta^-_K = \delta^-_K(r) \geq 0 \) such that \( r \in I_k = [t_{k-1}, t_k] \equiv [r - \delta^-_K, r + \delta^+_K] \) where \( \delta^+_K + \delta^-_K = t/K \). Also, recalling from Theorem 2.4.2 that \( \alpha^* \) is uniformly continuous in \([0, t]\), given \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
|\alpha^*_\rho - \alpha^*_r| < \varepsilon \quad \text{if} \quad \rho \in B_\delta(r) \quad \forall i \in N.
\]

This implies that for any \( r \in [0, t] \), there exists \( K_\varepsilon < \infty \) such that

\[
|\alpha^*_\rho - \alpha^*_r| < \varepsilon \quad \forall \rho \in [r - \delta^-_K, r + \delta^+_K] \subset B_\delta(r). \tag{2.127}
\]
for all $K > K_\varepsilon$. Further, by (2.126),
\[
|\hat{\alpha}_K^i(r) - [\alpha_r^*]^i| = |\beta_k^i(K, \alpha^*) - [\alpha_r^*]^i|
\]
which by (2.125) and (2.127),
\[
\leq \frac{1}{\delta_+^K + \delta_-^K} \int_{r-\delta_-^K}^{r+\delta_+^K} [\alpha_r^*]^i - [\alpha_r^*]^i \, d\rho = \varepsilon
\]
for all $K > K_\varepsilon$, which implies that $\hat{\alpha}_K$ converges pointwise to $\alpha^*$ as $K \to \infty$. Further, since $W^\alpha,c$ is Lipschitz continuous in $\alpha$, given $\varepsilon > 0$, there exists $\hat{K}_\varepsilon < \infty$ such that for all $K > \hat{K}_\varepsilon$,
\[
\varepsilon \geq \overline{W}^c(t, x, \zeta, z) - W^{\hat{\alpha}_K,c}(t, x, \zeta, z),
\]
which by the suboptimality of $\hat{\alpha}_K$ with respect to $\overline{W}^c_K$,
\[
\geq \overline{W}^c(t, x, \zeta, z) - \overline{W}^c_K(t, x, \zeta, z),
\]
proving the first assertion. The second assertion follows directly from Lemma 2.6.2.

For the final assertion, let $\check{\alpha}_K \in \tilde{A}^K$ such that $\check{\alpha}_K = \mathcal{L}^K(\check{\alpha}_K)$, and let $\hat{\alpha}_K^*$ be as per the Lemma statement. By the optimality of $\check{\alpha}_K$ with respect to $\overline{W}^c_K$,
\[
0 \leq W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z) - W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z),
\]
which by the suboptimality of $\mathcal{L}^K(\check{\alpha}_K^*)$ with respect to the definition (2.65) of $\overline{W}^c$,
\[
\leq \overline{W}^c(t, x, \zeta, z) - W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z),
\]
which by (2.129), for $K > \hat{K}_\varepsilon$,
\[
\leq \varepsilon.
\]

This implies that
\[
\lim_{K \to \infty} \{ W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z) - W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z) \} = 0.
\]

Therefore, given $\tilde{\varepsilon} \in (0, 1)$, there exists $\tilde{K}_\varepsilon < \infty$ such that
\[
\tilde{\varepsilon} \geq W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z) - W^{\mathcal{L}^K(\check{\alpha}_K),c}(t, x, \zeta, z)
\]
for all $K > \tilde{K}_\varepsilon$, which by the strong concavity asserted in Lemma 2.6.5, there exists $C_\alpha > 0$ such that
\[
\geq C_\alpha \| \check{\alpha}_K^* - \check{\alpha}_K \|^2_2,
\]
which by Remark 2.6.8,
\[ \| \mathcal{L}^K(\hat{\alpha}_K^*) - \hat{\alpha}_K \|^2_{L_2(0,t)} = C_\alpha \| \mathcal{L}^K(\hat{\alpha}_K^*) - \hat{\alpha}_K \|^2_{L_2(0,t)}, \]

which implies that \( \| \mathcal{L}^K(\hat{\alpha}_K^*) - \hat{\alpha}_K \|_{L_2(0,t)} \to 0 \) as \( K \to \infty \). Further, noting that

\[ \| \alpha^* - \mathcal{L}^K(\hat{\alpha}_K^*) \|_{L_2(0,t)} \leq \| \alpha^* - \hat{\alpha}_K \|_{L_2(0,t)} + \| \hat{\alpha}_K - \mathcal{L}^K(\hat{\alpha}_K^*) \|_{L_2(0,t)}, \]

applying (2.128) and (2.130) to the above completes the last assertion. \( \square \)

**Theorem 2.6.12.** Let \( t < T; c \in [0, \infty); L < \infty; x, z \in \overline{B}_L(0) \) and \( \zeta \in Z_L \). Given \( K \in \mathbb{N} \), suppose that \( (u^{c,*}, \alpha^*) \in U^\infty \times A^t \) and \( (\bar{u}^{c,*}, \bar{\alpha}_K^*) \in \mathcal{U}_K \times \mathcal{A}^K \) are the solutions of \( W^c(t, x, \zeta, z) \) of (2.64) and \( \mathcal{W}^{\mathcal{K}}(t, x, \zeta, z) \) of (2.123), respectively. Then, there exists \( \hat{D}^u = \hat{D}^u(t, L) < \infty \) such that

\[ \| u^{c,*} - \bar{u}^{c,*} \|_{L_2(0,t)} \leq \hat{D}^u \| \alpha^* - \mathcal{L}^K(\bar{\alpha}_K^*) \|_{L_2(0,t)}. \]

**Proof.** Let \( \alpha^*, \bar{\alpha}_K^*, u^{c,*} \) and \( \bar{u}^{c,*} \) be as asserted, and let \( \xi^{c,*} \) and \( \bar{\xi}^{c,*} \) denote the trajectories corresponding to \( u^{c,*} \) and \( \bar{u}^{c,*} \), respectively. By Theorems 2.4.16 and 2.5.2 and (2.93), for all \( r \in [0, t] \),

\[ u^{c,*}_r = -\nabla_x W^{\alpha^*,c}(t - r, \xi^{c,*}_r, \zeta) = -\nabla_x X(\xi^{c,*}_r, z) \cdot \hat{P}^{c,c}_{t-r}(\alpha^*, \zeta), \quad (2.131) \]

\[ \bar{u}^{c,*}_r = -\nabla_x \mathcal{L}^K(\bar{\alpha}_K^*,c)(t - r, \bar{\xi}^{c,*}_r, \zeta) = -\nabla_x X(\bar{\xi}^{c,*}_r, z) \cdot \hat{P}^{c,c}_{t-r}(\mathcal{L}^K(\bar{\alpha}_K^*), \zeta). \quad (2.132) \]

Note that for \( r \in [0, t] \),

\[ |\xi^{c,*}_r| \leq |x| + \int_0^r |\bar{u}^{c,*}_\rho| \, d\rho, \]

which by Hölder’s inequality,

\[ \leq |x| + \sqrt{t} \| \bar{u}^{c,*} \|_{L_2(0,t)}, \]

and by the definition of \( \mathcal{W}^c \), (2.11), and the nonnegativity of \( \mathcal{V} \), this is

\[ \leq |x| + [2t \mathcal{W}^c(t, x, \zeta, z)]^{1/2}, \]

which by (2.11) and (2.20),

\[ \leq |x| + [2t D_1(1 + |x|^2 + |z|^2)]^{1/2} \leq \hat{D}^x \quad (2.133) \]

for proper choice of \( \hat{D}^x = \hat{D}^x(t, L) < \infty \). Also, note that by the definition of \( X(\cdot, \cdot) \), for all \( r \in [0, t] \),

\[ \| \nabla_x X(\xi^{c,*}_r, \zeta) \|_F \leq 2(1 + |\xi^{c,*}_r| + |z|) \leq 2(1 + \hat{D}^x + L) = \hat{D}^x(t, L) \quad (2.134) \]

where the last bound follows by (2.133). Similarly,

\[ \| \nabla_x X(\xi^{c,*}_r, z) - \nabla_x X(\bar{\xi}^{c,*}_r, z) \|_F = 2|\xi^{c,*}_r - \bar{\xi}^{c,*}_r| \leq 2 \int_0^r |u^{c,*}_\rho - \bar{u}^{c,*}_\rho| \, d\rho. \quad (2.135) \]
Then, by (2.131) and (2.132),

$$|u^c_r - \bar{u}^c_r| = |\nabla_x X(\xi^c, z) \cdot \hat{P}^c_{t-r}(\alpha^*, \zeta) - \nabla_x X(\tilde{\xi}^c, z) \cdot \hat{P}^c_{t-r}(\mathcal{L}^K(\tilde{\alpha}_K), \zeta)|,$$

which by the triangle inequality,

$$\leq \|\nabla_x X(\xi^c, z) - \nabla_x X(\tilde{\xi}^c, z)\|_F |\hat{P}^c_{t-r}(\alpha^*, \zeta) - \hat{P}^c_{t-r}(\mathcal{L}^K(\tilde{\alpha}_K), \zeta)|$$

which by (2.102), (2.134), (2.135) and Lemma 2.6.1,

$$\leq 2K_p \int_0^r |u^c_{\rho^*} - \bar{u}^c_{\rho^*}| \, d\rho + \tilde{D}^z\hat{C}_1 \|\alpha^* - \mathcal{L}^K(\tilde{\alpha}_K)\|_{L_2(0, t)}.$$

By Gronwall’s inequality, this implies

$$|u^c_r - \bar{u}^c_r| \leq \tilde{D}^z\hat{C}_1 \exp(2K_p t) \|\alpha^* - \mathcal{L}^K(\tilde{\alpha}_K)\|_{L_2(0, t)}$$

for all $r \in [0, t]$, which completes the proof.

Recall that $u^*$ defined by (2.70) yields the fundamental solution, $W^\infty(t, x, z, \zeta).$ Combining Theorems 2.4.15 and 2.6.12, we can obtain a bound on the error in the resulting path, induced by using the approximations $c < \infty$ and piecewise constant $\alpha$. That is, we have:

**Corollary 2.6.13.** Let $t < \bar{t}; c \in [0, \infty); L < \infty; x, z \in B_L(0);$ and $\zeta \in Z_L.$ Given $K \in \mathbb{N}$, suppose that $u^*$ is the least action point in definition (2.70) of $W^\infty(t, x, z, \zeta)$, and that $(\tilde{u}^c, \tilde{\alpha}_K) \in \mathcal{U}^\infty \times \hat{\mathcal{A}}^K$ is the solution of (2.123). Then, there exist $\tilde{D} = \tilde{D}(t, \bar{t}) < \infty$ and $\tilde{D}^u = \tilde{D}^u(t, L) < \infty$ such that

$$\|u^* - \tilde{u}^c\|_{L_2(0, t)} \leq \frac{\tilde{D}(1 + |x| + |z|)^2}{\sqrt{c}} + \tilde{D}^u \|\alpha^* - \mathcal{L}^K(\tilde{\alpha}_K)\|_{L_2(0, t)}.$$

### 2.6.4 First-order necessary condition for maximization

For $K \in \mathbb{N}$, let

$$A^K_0 = \{\tilde{\alpha} = \{\tilde{\alpha}_i\}_{i \in \mathbb{N}} \mid \tilde{\alpha}_i \in (0, \sqrt{2/3}(R_i + 1/K)^{-1}) \quad \forall i \in \mathbb{N} \}.$$

Then, given $t < \bar{t}$, there exists $\hat{K} = \hat{K}(t) \in \mathbb{N}$ such that

$$t < \left[ \sum_{i \in \mathbb{N}} \frac{Gm_i}{2(R_i + 1/K)^3} \right]^{-1/2} < \bar{t} \quad \forall K > \hat{K}. \quad (2.136)$$
Let \( \hat{A}_0^K \) be the \( K \)th Cartesian product of \( A_0^K \), we see that for \( K \geq \hat{K} \), the coercivity and strict convexity of \( J^c(t,x,\cdot,\mathcal{L}^K(\tilde{\alpha}^o),\zeta,z) \) holds for any \( \tilde{\alpha}^o = \{\tilde{\alpha}^o_k\}_{k \in K} \in \hat{A}_0^K \). Further, by the arguments similar to that of the proofs of Lemma 2.6.7 and Lemma 2.6.11, there exists a unique stationary point of \( J^c(t,x,\cdot,\mathcal{L}^K(\cdot),\zeta,z) \) over \( U^\infty \times \hat{A}_0^K \), and

\[
W^c(t,x,\zeta,z) = \lim_{K \to \infty} \max_{\hat{\alpha}^o \in \hat{A}_0^K} W^{\mathcal{L}^K}(\hat{\alpha}^o),c(t,x,\zeta,z),
\]

and letting \( \hat{\alpha}^{o,*} = \argmax_{\hat{\alpha}^o \in \hat{A}_0^K} W^{\mathcal{L}^K}(\hat{\alpha}^o),c(t,x,\zeta,z), \)

\[
\mathcal{L}^K(\hat{\alpha}^{o,*}) \to \alpha^* \quad \text{as} \quad K \to \infty.
\]

Given \( t < \hat{t} \), we fix \( K > \hat{K}(t) \) throughout where \( \hat{K}(\cdot) \) is given in (2.136), and with a slight abuse of notation, let \( \hat{\alpha} = \hat{\alpha}^o \) and \( \hat{\alpha}^* = \hat{\alpha}^{o,*} \). We will demonstrate the existence of the derivative of \( W^{\mathcal{L}^K(\cdot),c} \) with respect to \( \hat{\alpha} \in \hat{A}_0^K \). Then, the maximum is achieved at the point where this derivative is zero (cf., [28]).

Lemma 2.6.14. Given \( \hat{\alpha} \in \hat{A}_0^K \), let \( \hat{\delta} = \{\hat{\delta}_k\}_{k \in K} \in \hat{A}_0^K \) be such that \( \hat{\alpha} + \hat{\delta} \in \hat{A}_0^K \). Let \( c \in [0,\infty) \) and \( \zeta \in \mathbb{Z}_L \). Then, there exists \( \hat{C}_2 < \infty \) such that

\[
\left| P^c_r(\mathcal{L}^K(\hat{\alpha} + \hat{\delta}),\zeta) - P^c_r(\mathcal{L}^K(\hat{\alpha}),\zeta) - \sum_{k \in K} \int_{(0,r) \cap \mathcal{I}_k} \left( -\frac{d\pi^s}{ds} \right) ds \hat{\delta}_{K+1-k} \right| \leq \hat{C}_2\|\hat{\delta}\|^2
\]

for all \( r \in (0,t) \) where \( P^c_r(\cdot,\zeta) \) and \( \pi^s = [\pi^s_i]_{i \in \mathcal{N}} \) are given by (2.91) and (2.106), respectively, driven by \( \mathcal{L}^K(\hat{\alpha}) \).

Proof. Letting \( \hat{\alpha} = \mathcal{L}^K(\hat{\alpha}) \) and \( h = \mathcal{L}^K(\hat{\delta}) \), by the linearity of \( \mathcal{L}^K \),

\[
\mathcal{L}^K(\hat{\alpha} + \hat{\delta}) = \mathcal{L}^K(\hat{\alpha}) + \mathcal{L}^K(\hat{\delta}) = \hat{\alpha} + h.
\]

(2.137)

Note that for \( r \in (0,t) \),

\[
\int_0^r \left( -\frac{d\pi^s}{ds} \right) h_{t-s} ds = \sum_{k \in K} \int_{(0,r) \cap \mathcal{I}_k} \left( -\frac{d\pi^s}{ds} \right) ds \hat{\delta}_{K+1-k}.
\]

(2.138)

Also, by Remark 2.6.8,

\[
\|h\|^2_{L^2(0,t)} = \|\mathcal{L}^K(\hat{\delta})\|_{L^2(0,t)} = \|\hat{\delta}\|^2.
\]

(2.139)

Substituting (2.137) – (2.139) into (2.108) completes the proof.

By Corollaries 2.6.9 and 2.6.14, \( P^c(\mathcal{L}^K(\cdot),\zeta) \in C([0,t] \times \hat{A}^K ; \mathcal{P}) \cap C^1((0,t) \times \hat{A}_0^K ; \mathcal{P}) \) (where we recall \( \mathcal{P} = \mathcal{H}^{10} \)).
For \( \alpha \in \mathcal{A}^K_0 \), \( i \in \mathcal{N} \) and \( k \in \mathcal{K} \), consider

\[
\dot{\Pi}_{r^k-1,i} = f(\Pi_{r}^{\tau_{k-1,i}}, \dot{P}_{r}^{c,k}) + f(\dot{P}_{r}^{c,k}, \Pi_{r}^{\tau_{k-1,i}}) + \Gamma(\alpha_{K+1-k}, \zeta_{t-r}) \tag{2.140}
\]

for all \( r \in \mathcal{I}_k \) where \( \Pi_{\tau_{k-1}} = 0_{M \times 1} \) (where \( f \) is given in (2.88) and \( \Gamma \) in (2.107)), and

\[
\dot{P}_{r}^{c,k} = f(\dot{P}_{r}^{c,k}, \Pi_{r}^{\tau_{k-1}}) + \Gamma(\alpha_{K+1-k}, \zeta_{t-r}) \tag{2.141}
\]

for all \( r \in \mathcal{I}_k \) and \( k \in \mathcal{K} \) with \( \dot{P}_{r}^{c,k} = \dot{P}_{r}^{c,k-1} = \dot{P}_{r}^{c} \) where \( \dot{P}_{r}^{c} \) is the solution of (2.91) with \( \alpha = \mathcal{L}^{K}(\hat{\alpha}) \).

**Proposition 2.6.15.** For \( \alpha \in \mathcal{A}^K_0 \), let \( \hat{\alpha} = \mathcal{L}^{K}(\hat{\alpha}) \).

\[
\frac{\partial \dot{P}_{r}^{c,k}}{\partial \hat{\alpha}} = \frac{\partial \dot{P}_{r}^{c,k}}{\partial \alpha_{K+1-k}} = \Pi_{r}^{\tau_{k-1}} \tag{2.142}
\]

for all \( r \in \mathcal{I}_k \) and \( k \in \mathcal{K} \) where \( \Pi_{\tau_{k-1}} = [\Pi_{\tau_{k-1,i}}]_{i \in \mathcal{N}} \in \mathbb{R}^{M \times N} \), and \( \Pi_{\tau_{k-1,i}} \) and \( \dot{P}_{r}^{c,k} \) are the solutions of (2.140) and (2.141), respectively.

**Proof.** Let \( i \in \mathcal{N} \) and \( k \in \mathcal{K} \). Then, we note that \( \dot{P}_{r}^{c}(\hat{\alpha}, \zeta) = \dot{P}_{r}^{c,k}(\mathcal{L}^{K}(\hat{\alpha}), \zeta) \) and \( \hat{\alpha}_{t-r} = \hat{\alpha}_{K+1-k} \) for all \( r \in \mathcal{I}_k \). Further, for \( r \in \mathcal{I}_k \), by (2.141),

\[
\frac{\partial \dot{P}_{r}^{c,k}(\mathcal{L}^{K}(\hat{\alpha}), \zeta)}{\partial \hat{\alpha}^{K+1-k}} = \frac{\partial \dot{P}_{r}^{c,k}(\hat{\alpha}, \zeta)}{\partial \hat{\alpha}^{i}} = \frac{\partial \dot{P}_{r}^{c}(\hat{\alpha}, \zeta)}{\partial \hat{\alpha}^{i}},
\]

which by Corollary 2.6.14,

\[
\hat{\alpha}_{t-r}^{\tau_{k-1,i}}
\]

where

\[
\hat{\alpha}_{t-r}^{\tau_{k-1,i}} = f(\hat{\alpha}_{t-r}^{\tau_{k-1,i}}, \dot{P}_{r}^{c}) + f(\dot{P}_{r}^{c}, \hat{\alpha}_{t-r}^{\tau_{k-1,i}}) + \Gamma(\hat{\alpha}_{t-r}, \zeta_{t-r})
\]

\[
= f(\hat{\alpha}_{t-r}^{\tau_{k-1,i}}, \dot{P}_{r}^{c}) + f(\dot{P}_{r}^{c}, \hat{\alpha}_{t-r}^{\tau_{k-1,i}}) + \Gamma(\hat{\alpha}_{t-r}, \zeta_{t-r}).
\]

Replacing \( \hat{\alpha}_{t-r}^{\tau_{k-1,i}} \) with \( \dot{\Pi}_{r}^{\tau_{k-1,i}} \) completes the proof. \( \square \)

We are now in a position to obtain the necessary condition.

**Theorem 2.6.16.** Let \( x, z \in \mathcal{B}_{L}(0) \) and \( \zeta \in \mathcal{Z}_{L} \). For \( c \in [0, \infty) \),

\[
\hat{\alpha}^{*} = \arg\max_{\hat{\alpha} \in \mathcal{A}^{K}_{0}} \mathcal{W}^{\mathcal{L}^{K}(\hat{\alpha}), c}(t, x, \zeta, z) \quad \text{if and only if}
\]

\[
\dot{F}(\hat{\alpha}^{*}) = 0_{N \times K}
\]
where the \((i, j)\)th elements of \(\hat{F}(\hat{\alpha}^*)\) are given by

\[
\hat{F}_{ij}(\hat{\alpha}^*) = \frac{1}{2} X(x, z) \cdot \frac{\partial \hat{P}_c^c(L^K(\hat{\alpha}^*), \zeta)}{\partial \hat{\alpha}_j^i}
\tag{2.143}
\]

\[
= \begin{cases} 
\frac{1}{2} X(x, z) \cdot \Pi_{\tau_{K-1}}^{K-1} & \text{if } j = 1, \\
\frac{1}{2} X(x, z) \cdot \mathcal{D}_{K+2-j} \Pi_{\tau_{K-1+j}}^{K-1} & \text{if } j \in [2, K[. 
\end{cases}
\tag{2.144}
\]

**Proof.** By the differentiability and the concavity in \(\hat{\alpha}\) given in Corollary 2.6.14 and Lemma 2.6.5, we have the assertions with the exception of the last representation for \(\hat{F}_{ij}(\hat{\alpha}^*)\).

For \(k \in K\), note that

\[
\frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} = \frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} = \frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} = \cdots = \frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} \Pi_{\tau_{k}}^{K-1},
\tag{2.145}
\]

where the last term follows from (2.142) of Proposition 2.6.15. Letting

\[
\phi_r^{k+1} = \frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} \quad \forall \tau \in I_{k+1}, \quad \forall k \in [1, K-1[, 
\tag{2.146}
\]

\[
\phi_r^{k+1} = \frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} = \frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} f(\hat{\alpha}_{K+1-k}, \hat{P}_c^c)
\]

with \(\phi_r^{k+1} = I_M\) where \(I_M\) denotes the identity matrix of size \(M\) and the \(m\)th column of \(\phi_r^{k+1}\) is given by

\[
\phi_{m,r}^{k+1} = f(\phi_{m,r}^{k+1}) + f(\hat{P}_c^c, \phi_{m,r}^{k+1})
\]

for all \(m \in [1, M]\) where \(\phi_r^{k+1}\) is given by (2.141). Substituting (2.146) into (2.145), we have

\[
\frac{\partial \hat{P}_c^c}{\partial \hat{\alpha}_{K+1-k}} = \begin{cases} 
\phi_r^K \phi_{r,K-1} \phi_{r,K-2} \cdots \phi_{r,k+1} \Pi_{\tau_{k}}^{K-1} \mathcal{D}_{k+1} \Pi_{\tau_{k}}^{K-1} & \text{if } k \in [1, K-1[, \\
\Pi_{\tau_{K}}^{K-1} & \text{if } k = K. 
\end{cases}
\]

Substituting these expressions into (2.143), we obtain the last representation. \(\square\)

**Remark 2.6.17.** In the above algorithm development, we solve the problem with \(c < \infty\), whereas the TPBVP requires \(c = \infty\). The error induced by using \(c < \infty\) is indicated in Theorem 2.4.15. However, when applying standard solution methods (e.g., Runge-Kutta methods) in solution of (2.80), (2.85), or equivalently, (2.91), taking very large \(c\)
approaching $\infty$, leads to difficulties. Here, we note a small practical point. An approximation for solution of (2.91) for $c = \infty$, $\hat{\alpha} \in \hat{A}_0^K$ and $\zeta \in \mathbb{Z}_L$, on an arbitrarily short initial time segment, say $r \in (0, \bar{\tau})$ with $\bar{\tau} << 1$, is given by

$$\hat{\nu}_r^\infty(L^K(\hat{\alpha}), \zeta) = (r^{-1}, -r^{-1}, r\hat{\sigma}^T, -r\hat{\sigma}^T, \frac{1}{3}\|\hat{\sigma}\|^2 r^3 + \hat{\eta} r)'$$

where

$$\hat{\sigma} = \frac{1}{2} \sum_{i \in \mathcal{N}} \hat{\alpha}_i K^i \zeta_i, \quad \hat{\eta} = \sum_{i \in \mathcal{N}} \mu_i \{2\hat{\alpha}_i K^i - (\hat{\alpha}_i K^i)^3 |\zeta_i|^2\}$$

which the reader may easily verify, and we do not attempt a bound on the error. Then, for $r > \bar{\tau}$, one may continue with a standard method.

### 2.7 Example

As an example, we solve a TPBVP where initial and terminal positions are specified. Suppose that five large bodies are moving clockwise on circular orbits with radii of 1.5, 3, 4.5, 6 and 7 AU. The masses of bodies are given by $[2, 4, 9, 10, 7] \times 10^{31}$ kg. In order to ensure a sensible problem, and for reasons of error estimation, we generate a trajectory, $\hat{\xi}$, of the small body by forward propagation of Newton’s second law formulation from an initial position/velocity pair given by $x = [0, 1, 0.02]^T$ AU, $u_0 = [35, 104, 0.3]^T$ km/s from initial time $r = 0$ to terminal time, $r = t = 11$ days. This yields a terminal position, $z = [0.179, 5.265, 0.0078]^T$. We reconstruct the trajectory from solution of the TPBVP given by $x$, $z$ and $t$. We remark that the body masses and duration form an exaggerated example, constructed such that the dynamics lead to an interesting trajectory. Further, we see that for all $i \in \mathcal{N}$, there exists $\delta_i > 0$ such that $|\hat{\xi}_r - \zeta^i| > \delta_i$ for all $r \geq 0$. Assuming the radius of body $R_i < \delta_i$ for all $i \in \mathcal{N}$, by Remark 2.4.6, the time duration that guarantees the convexity in the velocity control is given by $\bar{t} = 13.5$ [Day]. We use piecewise constant potential energy controls over $[0, \bar{t}]$ with a uniform grid $\{\tau_k|k \in [0, K]\}$, with $K = 25$. The small body is required to move through the intermediate points $\hat{\xi}_{t/4}, \hat{\xi}_{t/2}$ and $\hat{\xi}_{3t/4}$, where

$$\hat{\xi}_{t/4} = [0, 0.132, 1.708, -0.017], \quad \hat{\xi}_{t/2} = [0.121, 2.681, -0.011], \quad \hat{\xi}_{3t/4} = [0.134, 3.958, 0.0024]$$

The maximizing $\hat{\alpha}^* \in \hat{A}_0^K$ is obtained numerically by a gradient ascent method using Theorem 2.6.16 and Remark 2.6.17. Figure 2.1 depicts the true trajectory of the small body as well as the trajectories of five large bodies. Figure 2.2 depicts the true trajectory.
\( \dot{\xi} \), and the approximate solution of the TPBVP. As the bulk of the motion is in the first two coordinates, the trajectories are projected onto the plane generated by these coordinates.

Chapter 2, in full, has been submitted to Applied Mathematics & Optimization for publication of the material. The dissertation author was the primary author of this paper.

**Figure 2.1:** All trajectories with time axis

**Figure 2.2:** Truth and solution of TPBVP
Chapter 3

Solution of an Optimal Sensing and Interception Problem Using Idempotent Methods

3.1 Introduction

We consider a problem in coordinated command and control. The problem belongs to a class of problems where the state of the system consists not only of the physical state, but also a component describing the information state [40]. Although the payoff will depend entirely on the physical state, achievement of the desired physical-state goal depends on the controls used to advance the information state process. There are controls which affect both the physical- and information-state processes, and controls which only affect the latter. One difficult aspect of these problems is that the information state is typically a high-dimensional object. In the problem studied herein, even though the physical state will take values in a finite set, the information state will be modeled in terms of a probability distribution over that finite set, and lies in a Euclidean space where the dimension is the cardinality of the finite set. Consequently, one needs computational tools which can handle control problems over high-dimensional spaces. Natural approaches are the algorithms in the class of idempotent curse-of-dimensionality-free methods [35, 36, 32]. Although such methods have been applied to similar information state problems before [40, 47], here we combine those techniques with the “double-description method” [45], which will prove to be a fruitful combination in terms of
Figure 3.1: UAV/UGS/Interceptor problem

computational-effort reduction.

The specific application under consideration is as follows, where we note that it may be helpful to refer to Figure 3.1. There is a single intruder vehicle. We would like to intercept that vehicle before it reaches a specific target location, labeled as the base in the figure. The intruder may only move along the given road network. There are unattended ground sensors (UGSs) placed along the road network. The intruder is unaware of the locations of these sensors. When the intruder passes through a UGS location, the data is acquired and stored by the UGS. We have an unmanned aerial vehicle (UAV) which is monitoring the situation. When the UAV overflies a USG, the data from the UGS is uploaded to the UAV. We also have an interceptor ground vehicle which moves along the road network. We will use a discrete-time, discrete-state model of the physical system state. The problem has an exit-time aspect. The exit time is the first time when either the interceptor and the intruder are co-located or the intruder reaches the base. We will be attempting to minimize the payoff. There will be a positive payoff if the intruder reaches the base, and a negative payoff if the interceptor is co-located with the intruder prior to the intruder reaching the base.

We will employ stochastic models. Purely game-theoretic and more general stochastic game models could certainly be considered. The reason for applying a stochastic approach over a game-theoretic approach is that we expect significant observation
errors, and that these would be best modeled as random processes. The more general
stochastic game approach might be optimal, but the computational burden is already
significant. One possible alternative which might be computationally tractable is a risk-
averse stochastic model, as these can prove effective in the context of stochastic games
[41].

With a stochastic control problem formulation, we will have an information state
which takes the form of an observation-conditioned probability distribution over the
space of possible intruder locations. As we assume full communication between the UAV,
the interceptor and possibly a central processing location, we do not distinguish where the
computations may be occurring. Updates to the conditional distribution will be through
Bayes’ rule when observations are processed. As the state-space is finite, we model the
intruder movement as a Markov chain. The state of the system at any given moment has
both a physical and an informational component. Of course the informational component
is the conditional distribution over a set of locations. The physical component will consist
of the locations of the intruder, the UAV and the interceptor. Our controls at each time
step will be the next UAV location and the next interceptor location. We will assume
that the UAV can move from any location to any other location in one step, whereas in
one step, both the interceptor and the intruder are constrained to move only between
adjacent nodes of the road-network graph.

As our main focus will be on the application of idempotent curse-of-dimensionality
-free methods to this class of control problems, we will make some other simplifications
so as to reduce the technical complexity. Most notable is the following. Each UGS takes
data each time the intruder passes by. However, our control system only becomes aware
of the observations taken by a UGS when the UAV overflies that UGS. For example,
at time $t_5$, we may obtain data from UGS 7, where it observed the intruder at time $t_4$.
Then, at time $t_6$, we might obtain data from UGS 3, where it observed the intruder at time $t_4$.
If we have already processed the time $t_4$ data, updating the system with the
time $t_2$ data becomes a bit tedious. One could maintain the full history of the conditional probability distributions so that one could go back and fold in the $t_2$ data for
time $t_2$ and then propagate forward back up to time $t_4$ to then process that data, and
further propagate forward to the current time. Although the requisite algorithm would
be complex, it would not substantially affect the total computational cost, which is the
central point of the discussion here. Consequently, we restrict ourselves to models where
we only process newly uploaded observations which are more recent than the most recent previously uploaded observation. A comprehensive discussion of the issues related to the sensor data timing can be found in [21].

3.2 Problem definition

Assume that there are \( N_s \) UGS locations in the region under consideration. (As above, it may be useful to refer to Figure 3.1.) The set of these locations may be indexed as \( \mathcal{N}_s := \{1, 2, \ldots, N_s\} \). Throughout, for integers \( a \leq b \), we will use \([a, b]\) to denote \( \{a, a+1, a+2, \ldots b\} \). We denote the target, or base, location as \( N_b = N_s + 1 \) and let \( \mathcal{N}_b := \{1, N_b[=]1, N_s + 1\} \). We will let the states of the intruder and the interceptor take values in the finite set \( \mathcal{N}_b \), with the addition of one more state value to be indicated just below. The road network will be interpreted as a graph with the set of nodes being \( \mathcal{N}_b \), and an edge, \((i, j)\) (equivalently \((j, i)\) ), existing between nodes \( i \) and \( j \) if there exist roads connecting those nodes without passing through another UGS location. We let the neighborhood, or set of immediately accessible states, from node \( i \in \mathcal{N}_b \), be the set of nodes, \( j \in \mathcal{N}_b \) such that there exists an edge \((i, j)\) connecting them. and denote this set as \( \mathcal{I}_i \). Recall that we use a discrete-time model. For simplicity, we let the time-step be one, and let the time be denoted by \( t \in \mathcal{T} := \{0, 1, 2, \ldots\} \). We can model the intruder motion process as a Markov chain on \( \mathcal{N}_b \), with transition probability matrix, \( \mathbb{P} \in \mathbb{R}^{\mathcal{N}_b \times \mathcal{N}_b} \). In particular, we will take \( \mathcal{N}_b \) to be an absorbing state.

The interceptor also moves on \( \mathcal{N}_b \). However, in this case, we will have an interceptor control which will allow the interceptor to move from \( i \in \mathcal{N}_b \) to any \( j \in \mathcal{T}_i \) in one time step. Recall that the intruder is “captured” if the interceptor and intruder are co-located. Because of this, we find it helpful to add another location to the model. We suppose that if the interceptor and intruder are co-located at time, \( t \), then the intruder moves instantaneously with probability one to an exit-state location denoted by \( \mathcal{N}_e \). Therefore, we add this location to \( \mathcal{N}_b \), and let \( \mathcal{N}_e := \{1, \mathcal{N}_e[=]1, N_s + 2\} \).

The intruder and interceptor states at time \( t \in \mathcal{T} \) are denoted by \( x^i_t \in \mathcal{N}_e \) and \( x^i_t \in \mathcal{N}_b \), respectively. The interceptor control at time \( t \), is denoted by \( u^i_t \in \mathcal{T}^{x^i_t} \subseteq \mathcal{N}_b \).

The interceptor dynamics are simply \( x^i_{t+1} = u^i_t \). The UAV state at time \( t \) is denoted by \( x^o_t \). As the UAV may travel from any node to any other in one time step, we let the UAV control at time \( t \) be \( u^o_t \in \mathcal{N}_b \), and the UAV dynamics are simply \( x^o_{t+1} = u^o_t \).

Our information at time \( t \) on the intruder states in \( \mathcal{N}_e \) is described by probability
distribution \( q_t \). In particular, the \( j^{th} \) component of \( q_t \), \( q^j_t \), is the probability that the intruder is at \( j \in \mathcal{N}_e \) at time \( t \). Distribution \( q_t \) lies in probability simplex \( S^{\mathcal{N}_e} \) where

\[
S^{\mathcal{N}_e} \doteq \left\{ q \in \mathbb{R}^{\mathcal{N}_e} \mid q^j \in [0, 1] \ \forall j \in \mathcal{N}_e \ \text{and} \ \sum_{j \in \mathcal{N}_e} q^j = 1 \right\}.
\]

### 3.2.1 System dynamics

At this point, we have indicated some components of the problem statement. First, note that the system state at time \( t \) will be \((x^o_t, x^i_t, x^R_t, q_t)\), and the controls will be \((u^o_t, u^i_t)\). We have the dynamics for \( x^o \) and \( x^i \), which are

\[
x^{o}_{t+1} = u^o_t \quad \text{and} \quad x^{i}_{t+1} = u^i_t,
\]

where

\[
u^o_t \in \mathcal{N}_s \quad \text{and} \quad u^i_t \in \mathcal{I}^{x^i_t} \subseteq \mathcal{N}_b.
\]

We have indicated that \( x^R \) propagates as a Markov chain, with nominal transition matrix \( \mathcal{P} \in \mathbb{R}^{\mathcal{N}_b \times \mathcal{N}_b} \). However, because of the potential for interception, we now extend and modify that transition matrix. We suppose that some interceptor control, \( u^i_t \) is given. For any \( u^i \in \mathcal{N}_b \), we let \( \widehat{\mathcal{P}}^{u^i} \in \mathbb{R}^{\mathcal{N}_e \times \mathcal{N}_e} \) be as follows.

\[
\widehat{\mathcal{P}}^{u^i}_{j,k} = \begin{cases} 
\mathcal{P}^{j,k} & \text{if } j < \mathcal{N}_b, k \neq u^i \\
0 & \text{if } j < \mathcal{N}_b, k = u^i \\
\mathcal{P}^{j,k} & \text{if } j = \mathcal{N}_b, k = \mathcal{N}_b \\
\mathcal{P}^{j,u^i} & \text{if } j < \mathcal{N}_b, k = \mathcal{N}_e \\
0 & \text{if } j = \mathcal{N}_e, k \neq \mathcal{N}_e \text{ or } j = \mathcal{N}_b, k \neq \mathcal{N}_b \\
1 & \text{if } j = \mathcal{N}_e, k = \mathcal{N}_e \text{ or } j = \mathcal{N}_b, k = \mathcal{N}_b.
\end{cases}
\]

Lastly, we need to describe the dynamics of the information state, \( q \). We suppose that at each time step, the observation update occurs prior to the dynamics update, where the latter is described by the above transition matrix, \( \widehat{\mathcal{P}}^{u^i} \). We now describe the observation update, where this will be performed via Bayes’ rule.

Each UGS may record information regarding the intruder when the intruder is in its range of detection, such as the time, position, speed, direction, and type of intruder. As our focus is on the development of an efficient numerical algorithm, we will use a simple UGS observation model. We suppose that if the UAV control at time \( t \) is
\( u_t^o \in \mathcal{N}_s \), then the UAV immediately uploads the (random variable) observation at time \( t \), \( y_t \), made by the UGS. We let \( y_t \) take values in \( \mathcal{Y} = \{0, 1\} \), where \( y_t = 0 \) is the observation that no intruder passed the node \( j = u_t^o \) at time \( t \) and \( y_t = 1 \) is the observation that the intruder passed there at that time. Again, for a more complete analysis of observation processing where multiple time-step information may be uploaded to the UAV, see [21].

We suppose that the UGSs have known false alarm and misdetection rates, and that there exists noise in the communication system between the UAV and the UGSs, also with known parameters, although communication dropout is not considered. With this, we obtain an overall confusion matrix. In particular, we let \( R_j^{u^o,y} \) denote the probability of observation \( y \in \mathcal{Y} \) given the UAV is at \( u^o \in \mathcal{N}_s \) and the intruder is at \( j \in \mathcal{N}_e \). That is,
\[
R_j^{u^o,y} = P_{u^o}(y_t = y | x_t^R = j),
\]
where the subscript \( u^o \) indicates the dependence of the particular UAV sensing control. Let \( R^{u^o,y} \) be the vector of length \( N_e \) with components \( R_j^{u^o,y} \). Throughout, for generic vector, \( \mathbf{R} \), we let \( D(\mathbf{R}) \) denote the diagonal matrix with diagonal entries given by the components of \( \mathbf{R} \). More specifically, \( D(R^{u^o,y}) \) denotes the \( N_e \times N_e \) matrix with components \( [D(R^{u^o,y})]_{jj} = R_j^{u^o,y} \), and such that \( [D(R^{u^o,y})]_{ij} = 0 \) for \( i \neq j \).

Suppose that under control \( u_t^o \), the observation \( y_t = y \in \mathcal{Y} \) is made. Then, by Bayes' rule, the posteriori probability is given by
\[
\hat{q}_t = P_{u_t^o}(x_t^R = j | y_t = y) = \frac{P_{u_t^o}(y_t = y | x_t^R = j)P(x_t^R = j)}{\sum_{j \in \mathcal{N}_e} P_{u_t^o}(y_t = y, x_t^R = j)}
\]
which by (3.2) and the definition of \( R^{u_t^o,y} \),
\[
\hat{q}_t = \frac{R_j^{u_t^o,y} q_j^t}{R^{u_t^o,y} q_t}.
\]
Hence, by the definition of \( D(\cdot) \),
\[
\hat{q}_t = \frac{1}{R^{u_t^o,y} q_t} D(R^{u_t^o,y}) q_t = \hat{\beta}^{u_t^o,y}(q_t)
\]
where \( \hat{q}_t \in S^{N_e} \).

Consequently, given the current interceptor state \( x_t^i \) and the interceptor control \( u_t^i \in \mathcal{I}^{x_t^i} \), the combined dynamics and observation update becomes
\[
q_{t+1} = \left( \tilde{\mathbf{H}}^{u_t^i} \right)^T \hat{\beta}^{u_t^o,y}(q_t).
\]
We remark that there is an implicit observation by the interceptor as well as the UAV observation discussed above. For example, if there is no intruder at \( x_{i+1} = u_i \), then there is an implicit interceptor-observation that the intruder is not at this location. One should assume that this observation is noise-free as we also assume the intruder is captured with probability one if they are co-located. In this case, letting \( y = 0 \) denote the observation that the intruder is not at the interceptor location, one is led to \( R^{u_i,y} = R^{u_i,0} \) (analogous to \( R^{u_i,y} \) above) which is one if \( j \neq u_i \) and zero otherwise.

Because of the form of \( \hat{\Pi}^{u_i} \) above, including these additional observation dynamics into the backward dynamic program below yields absolutely no difference in the dynamic programming update. (Of course, in forward simulation, one does need to include this implicit observation.)

### 3.2.2 Payoff and value

We impose a finite time-horizon, \( T \), on the problem. We will not start from basic principles in definition of the payoff and value function. Proceeding from that point to the problem definition we use below would involve a substantial argument using the Separation Principle, which is outside the scope of our interests here. Let \( T^- = ]0, T] \) and \( T^− = ]0, T − 1[ \). For \( t \in T^- \), let

\[
\mathcal{U}_i = \{ \{u_r\}_{r=t}^{T-1} : [N_b \times S^{N_e}]^{T−t} \rightarrow [N_b]^{T−t} \mid \text{if } (x^i, q)_{t,r} \in [N_b \times S^{N_e}]^{T−t} \\
\text{and } r \in [t, T−1[ \text{ are such that } (x^i, q)_{t,r} = (\tilde{x}^i, \tilde{q})_{t,r}, \\
\text{then } u_i(x^i, q_r) = u_i(\tilde{x}^i, \tilde{q}_r) \}.
\]

We define \( \mathcal{U}_o \) similarly, but with \( N_b \) in the range replaced by \( N_e \). Let \( \beta \in \mathbb{R}^{N_e} \) where \( \beta_{N_b} = c_b \), \( \beta_{N_e} = c_e \), and \( \beta_i = 0 \) otherwise. Here, \( c_b \) denotes the payoff obtained in the case that the intruder reaches the base, and \( c_e \) denotes the payoff in the case that the intruder is intercepted prior to reaching the base. Let \( \psi : S^{N_e} \rightarrow \mathbb{R} \) be given by

\[
\psi(q) = \beta \cdot q.
\]

For \( t \in T^- \), let \( \mathcal{Y}_t \) denote the observation process \( y_{t,T−1[} \), taking values in \( \mathcal{Y} \), where we suppose the \( y_t | x_t^R \) form a sequence of independent random variables. For \( t \in T^- \), the payoff \( \mathcal{J}_t : S^{N_e} \times N_e \times \mathcal{U}_i \times \mathcal{U}_o \rightarrow \mathbb{R} \) is given by

\[
\mathcal{J}_t(q, x^i, u^i, u^o) = \mathbb{E}_{y_{t,T−1[}} \psi(q_t),
\]
where \( q \) satisfies (3.4) with \( q_t = q \), and \( x^0, x^i \) satisfy (3.1) with \( x^i_t = x^i \). For \( t \in T \), the value \( V_t : S^{N_e} \times N_e \rightarrow \mathbb{R} \) is given by
\[
V_t(q, x^i) = \inf_{u^i \in U_t} \inf_{u^o \in U_t^o} J_t(q, x^i, u^i, u^o).
\]

### 3.3 Idempotent based dynamic programming

To solve the state feedback problem, backward dynamic programming will be enabled via an idempotent-based algorithm, not subject to the curse-of-dimensionality. As it is standard given the value function definition (3.5), we present the dynamic programming principle without proof. For \( t \in T \) we have
\[
V_t(q, x^i) = \min_{u^o \in N_s} \min_{u^i \in I_{x^i}^t} \mathbb{E}_{y_t} \left\{ \sum_{y \in Y} \min_{z \in Z_{x^i_t+1}} \left[ (\tilde{D}^{u^o} \tilde{R} u^i, y_t) \tilde{u}^{u^o, y_t}(q), u^o \right] \right\},
\]
where the expectation is over the set of possible observations at time \( t \).

For any set \( Z \) and positive integer \( M \), let \( |Z| \) be the cardinality of \( Z \) and \( \mathcal{P}^M(Z) \) be the set of all sequence of length \( M \) with elements from \( Z \). Let the cardinality of the observation set \( Y \) be denoted by \( N_y \), and given \( x^i, N_u^{x^i} = |N_s \times I_{x^i}| \). Given \( x^i \) and time \( t \), we denote the index set of coefficient vectors of functionals by \( Z_{x^i}^t = \{1, |Z_{x^i}^t|\} \). It will be helpful to define \( \mathcal{G} : \mathbb{R}^{N_s} \rightarrow \mathbb{R}^{N_e} \) given by
\[
\mathcal{G}(b) = (b^T, c_b, c_e)^T,
\]
and \( \mathcal{G}^{-1} : \text{Range}(\mathcal{G}) \rightarrow \mathbb{R}^{N_s} \) given by
\[
\mathcal{G}^{-1}(\bar{b}) = \bar{b}_{|1, N_s|},
\]
where \( \bar{b}_{|1, N_s|} \) denotes the vector consisting of the first \( N_s \) components of \( \bar{b} \).

**Lemma 3.3.1.** Suppose there exist index sets \( Z_{x^i}^t \) (of cardinality \( Z_{x^i}^t \)) for all \( x \in N_b \), and \( b_{x^i}^{t+1} \in \mathbb{R}^{N_s} \) for all \( x \in N_b \) and all \( z \in \mathcal{Z}_{x^i}^t \) such that
\[
V_{t+1}(q, x^i) = \min_{z \in \mathcal{Z}_{x^i}^t} \bar{b}_{t+1}^{x^i, z} \cdot q, \quad \forall q \in S^{N_e}, \ \forall x^i \in N_b,
\]
where \( \bar{b}_{t+1}^{x^i, z} = \mathcal{G}(b_{x^i}^{t+1}) \). Then,
\[
V_t(q, x^i) = \left\{ \min_{u^o \in N_s} \min_{u^i \in I_{x^i}} \sum_{y \in Y} \min_{z \in \mathcal{Z}_{x^i}^{t+1}} \left[ (D(\tilde{R}^{u^o, y}) \tilde{R}^{u^i, z}, y_t) \tilde{u}^{u^o, y_t}(q), u^o \right] \right\}
+ c_b q^{N_b} + c_e q^{N_e},
\]
where \( \mathcal{P}^{u^i} \) denotes the first \( N_s \) rows of \( \mathcal{P}^{u^i} \), \( \mathcal{R}^{u^o,y} \) denotes the first \( N_s \) elements of \( \mathcal{R}^{u^o,y} \), and \( \mathcal{H} : S^{N_e} \rightarrow \mathbb{R}^{N_s} \) be given by \( \mathcal{H}(q) = q^{1:N_s} \) (i.e., the first \( N_s \) elements of \( q \) for any \( q \in S^{N_e} \)).

**Proof.** First, fix any \( u^o_t = u^o \in \mathcal{N}_b \), and let \( y_t \) be the resulting observation. Suppose \( x_t^R \) is distributed according to \( q \). We have

\[
P_{u^o}(y_t = y) = \sum_{i \in \mathcal{N}_e} P_{u^o}(y_t = y, x_t^R = i) = \sum_{i \in \mathcal{N}_e} P_{u^o}(y_t = y | x_t^R = i) P(x_t^R = i) = \sum_{i \in \mathcal{N}_e} R_{i}^{u^o,y} q^i = \mathcal{R}^{u^o,y} \cdot q.
\]  

(3.7)

Now recall from (3.6) that

\[
\mathcal{V}_t(q, x^i) = \min_{u^o \in \mathcal{N}_b} \min_{u^i \in \mathcal{Z}^i} E_{y_t} \left\{ \mathcal{V}_{t+1} + \left( \mathcal{P}^{u^i} \right)^T \mathcal{R}^{u^o,y}(q), u^i \right\},
\]

which with the assumed form,

\[
= \min_{u^o \in \mathcal{N}_b} \min_{u^i \in \mathcal{Z}^i} E_{y_t} \left\{ \min_{z \in \mathcal{Z}^{u^i}_{t+1}} \left( \mathcal{P}^{u^i} \right)^T \mathcal{R}^{u^o,y}(q) \right\}
= \min_{u^o \in \mathcal{N}_b} \min_{u^i \in \mathcal{Z}^i} \sum_{y \in \mathcal{Y}} \left\{ \min_{z \in \mathcal{Z}^{u^i}_{t+1}} \left( \mathcal{P}^{u^i} \right)^T \mathcal{R}^{u^o,y}(q) \right\} P_{u^o}(y_t = y).
\]

Using (3.3) and (3.7), this becomes

\[
\mathcal{V}_t(q, x^i) = \min_{u^o \in \mathcal{N}_b} \min_{u^i \in \mathcal{Z}^i} \sum_{y \in \mathcal{Y}} \left\{ \min_{z \in \mathcal{Z}^{u^i}_{t+1}} \left( \tilde{b}_{t+1}^{u^i,z} \right)^T \left( \mathcal{P}^{u^i} \right)^T \mathcal{R}^{u^o,y} \cdot q \right\} R_{u^o,y} \cdot q.
\]

(3.8)

Now, for \( u^i \in \mathcal{N}_b \), let \( \mathcal{P}^{u^i} \) denote the first \( N_s \) rows of \( \mathcal{P}^{u^i} \). Note that \( \mathcal{P}^{u^i}_{N_b,j} = \delta_{N_b,j} \) and \( \mathcal{P}^{u^i}_{N_e,j} = \delta_{N_e,j} \), where \( \delta_{i,j} \) denotes the Dirac delta function. Then, for any \( q \in S^{N_e} \),

\[
\left( \mathcal{P}^{u^i} \right)^T D(\mathcal{R}^{u^o,y}) q = \left( \mathcal{P}^{u^i} \right)^T \mathcal{H}(D(\mathcal{R}^{u^o,y}) q) + q',
\]

(3.9)

where \( q' \) is the \( N_e \)-dimensional vector given by

\[
q' = q'(u^o, y) = (0^T, R_{N_b}^{u^o,y} q_{N_b}, R_{N_e}^{u^o,y} q_{N_e})^T.
\]
Substituting (3.9) into (3.8) yields
\[
\nabla_x(q, x) = \min_{u^i \in \mathcal{N}_u} \min_{u^i \in \mathcal{I}^x} \sum_{y \in \mathcal{Y}} \min_{z \in \mathcal{Z}_{t+1}^{x_i}} \left[ \left( \overline{b}_{t+1}^{x_i, z} \right)^T \mathcal{H}(D(R^{u^i, y})q) + q' \right],
\]
which upon noting that \( \overline{b}_{t+1}^{x_i, z} \mathcal{N}_b = c_b \) and \( \overline{b}_{t+1}^{x_i, z} \mathcal{N}_c = c_e \),
\[
= \min_{u^i \in \mathcal{N}_u} \min_{u^i \in \mathcal{I}^x} \sum_{y \in \mathcal{Y}} \min_{z \in \mathcal{Z}_{t+1}^{x_i}} \left[ \left( \overline{b}_{t+1}^{x_i, z} \right)^T \mathcal{H}(D(R^{u^i, y})q) \right.
+ c_b R^{u^i, y}_{N_b} q_{N_b} + c_e R^{u^i, y}_{N_c} q_{N_c} \left. \right],
\]
and noting that \( \sum_{y \in \mathcal{Y}} R_{j}^{u^i, y} = 1 \) for all \( j \in \mathcal{N}_c \), this is
\[
= \left\{ \min_{u^i \in \mathcal{N}_u} \min_{u^i \in \mathcal{I}^x} \sum_{y \in \mathcal{Y}} \min_{z \in \mathcal{Z}_{t+1}^{x_i}} \left[ \left( \overline{b}_{t+1}^{x_i, z} \right)^T \mathcal{H}(D(R^{u^i, y})q) \right] \right\}
+ c_b q_{N_b} + c_e q_{N_c},
\]
and letting \( \overline{R}_{u^i, y} \) denote the first \( N_s \) elements of \( R^{u^i, y} \),
\[
= \left\{ \min_{u^i \in \mathcal{N}_u} \min_{u^i \in \mathcal{I}^x} \sum_{y \in \mathcal{Y}} \min_{z \in \mathcal{Z}_{t+1}^{x_i}} \left[ (D(\overline{R}_{u^i, y}) \overline{b}_{t+1}^{x_i, z})^T \mathcal{H}(q) \right] \right\}
+ c_b q_{N_b} + c_e q_{N_c}.
\]

\[\Box\]

**Theorem 3.3.2.** Suppose there exist index sets \( \mathcal{Z}_{t+1}^{x_i} \) (of cardinality \( \mathcal{Z}_{t+1}^{x_i} \)) for all \( x \in \mathcal{N}_b \), and \( \overline{b}_{t+1}^{x_i, z} \in \mathcal{I}^{N_s} \) for all \( x \in \mathcal{N}_b \) and all \( z \in \mathcal{Z}_{t+1}^{x_i} \) such that
\[
\nabla_{t+1}(q, x) = \min_{z \in \mathcal{Z}_{t+1}^{x_i}} G(b_{t+1}^{x_i, z}) \cdot q, \quad \forall q \in S^{N_c} \forall x \in \mathcal{N}_b.
\]
Let \( \mathcal{Z}_{t+1}^{x_i} = \sum_{u^i \in \mathcal{I}^x} | \mathcal{Z}_{t+1}^{x_i} | N_u \mathcal{N}_a \) and \( \mathcal{Z}_{t+1}^{x_i} = \mathcal{Z}_{t+1}^{x_i} \). Let \( \mathcal{M} \) be a one-to-one, onto mapping from \( \mathcal{N}_x \times \mathcal{I}^x \times \mathcal{P}^{N_u} \left( \mathcal{Z}_{t+1}^{x_i} \right) \rightarrow \mathcal{Z}_{t+1}^{x_i} \), where the notation \( \mathcal{P}^{N_u}(Z) \) denotes the set of sequences of length \( N_u \) of elements of \( Z \). Lastly, let \( \overline{R}_{u^i} \) and \( D(\overline{R}_{u^i, y}) \) be as defined.
in Lemma 3.3.1. Then,
\[
\mathcal{V}_t(q, x^i) = \min_{z \in \mathcal{Z}_t^x} \left[ b_t^{x^i, z} \cdot \mathcal{H}(q) \right] + c_b q^N_b + c_e q^N_e
\tag{3.10}
\]
\[
= \min_{z \in \mathcal{Z}_t^x} \mathcal{G}(b_t^{x^i, z}) \cdot q, \quad \forall q \in S^{N_e} \forall x^i \in \mathcal{N}_b,
\tag{3.11}
\]
where, for each \( z \in \mathcal{Z}_t^x \),
\[
b_t^{x^i, z} = \sum_{y \in \mathcal{Y}} D(\tilde{R}^{u^o, y}) \tilde{T}^{u^i} \mathcal{G}(b_{t+1}^{u^i z^y}),
\tag{3.12}
\]
and \((u^o, u^i, \{z_y\}_{y \in \mathcal{Y}}) = \mathcal{M}^{-1}(z)\).

**Remark 3.3.3.** It is important to note that this implies that the value function may be propagated purely by the algebraic operations given in (3.12). The set of vectors \( \{b_t^{x^i, z} | z \in \mathcal{Z}_t^x, x \in \mathcal{N}_b\} \) completely define the function \( \mathcal{V}_t \).

**Proof.** From Lemma 3.3.1, we have
\[
\mathcal{V}_t(q, x^i) = \left\{ \min_{u^o \in \mathcal{N}_s} \min_{u^i \in \mathcal{I}^{x^i}} \left( \min_{y \in \mathcal{Y}} \min_{z \in \mathcal{Z}_{t+1}^{u^i}} \left[ (D(\tilde{R}^{u^o, y}) \tilde{T}^{u^i} b_{t+1}^{u^i z^y}) \right]^T \mathcal{H}(q) \right) \right\}
\]
\[
+ c_b q^N_b + c_e q^N_e,
\tag{3.13}
\]
where we recall \( b_{t+1}^{x^i, z} = \mathcal{G}(b_{t+1}^{x^i, z}) \). It is handiest to employ the min-plus algebra form here. We remind the reader that the min-plus algebra (semifield) is defined on \( \mathbb{R} \cup \{+\infty\} \), with addition and multiplication operations given by \( a \oplus b = \min\{a, b\} \) and \( a \otimes b = a + b \) (cf., [1, 7, 31]). Using this formulation, (3.13) becomes
\[
\mathcal{V}_t(q, x^i) = \left\{ \bigoplus_{u^o \in \mathcal{N}_s} \bigoplus_{u^i \in \mathcal{I}^{x^i}} \bigotimes_{y \in \mathcal{Y}} \bigoplus_{z \in \mathcal{Z}_{t+1}^{u^i}} \left[ (D(\tilde{R}^{u^o, y}) \tilde{T}^{u^i} b_{t+1}^{u^i z^y}) \right]^T \mathcal{H}(q) \right\}
\]
\[
+ c_b q^N_b + c_e q^N_e.
\]
Using the min-plus distributive property (cf., [39, 35]), this becomes

\[ V_t(q, x^i) = \left\{ \bigoplus_{u^o \in N_s} \bigoplus_{u^i \in \mathcal{I}^i} \bigoplus_{\{z\} \in \mathcal{P}^N_y(Z_{t+1}^i)} \left[ D(\tilde{R}^{u^o,y}) \tilde{P}^{u^i} b_{t+1}^{u^i} z_y \right]^T \mathcal{H}(q) \right\} + c_b q^N_b + c_e q^N_e \]

\[ = \left\{ \min_{u^o \in N_s} \min_{u^i \in \mathcal{I}^i} \min_{\{z\} \in \mathcal{P}^N_y(Z_{t+1}^i)} \left[ \sum_{y \in \mathcal{Y}} D(\tilde{R}^{u^o,y}) \tilde{P}^{u^i} b_{t+1}^{u^i} z_y \right]^T \mathcal{H}(q) \right\} + c_b q^N_b + c_e q^N_e \]

\[ = \left\{ \min_{(u^o,u^i,\{z\}) \in N_s \times \mathcal{I}^i \times \mathcal{P}^N_y(Z_{t+1}^i)} \left[ \sum_{y \in \mathcal{Y}} D(\tilde{R}^{u^o,y}) \tilde{P}^{u^i} b_{t+1}^{u^i} z_y \right]^T \mathcal{H}(q) \right\} + c_b q^N_b + c_e q^N_e , \]

and using (3.12), this is

\[ = \min_{z \in Z_{t+1}^i} \left[ b_{t+1}^{x^i,z} \cdot \mathcal{H}(q) \right] + c_b q^N_b + c_e q^N_e = \min_{z \in Z_{t+1}^i} \mathcal{G}(b_{t+1}^{x^i,z}) \cdot q, \]

where \( Z_{t+1}^i \) is as asserted. \( \square \)

It is worth noting that \( V_t \) is a concave, piecewise linear function of its first argument, the controller’s probability distribution regarding the intruder position. For a problem with 8 UGSs in the region, a slice of \( V_t \) is depicted in Fig. 3.2, where \( c_e = -100 \) and \( c_b = 100 \).

### 3.4 Efficient projection and pruning

Although the curse-of-dimensionality is avoided with the idempotent-based dynamic program, another computational complexity issue arises, which may be referred to as the curse-of-complexity. Note that a naive gridding of the state-space imposes the curse-of-dimensionality where this might not actually be present in the solution. As trivial examples, note that a linear functional over \( \mathbb{R}^N \) can be represented exactly with only \( N \) numbers, while using a grid-based representation (over only a rectangular region) would require a set of \( L^N \) numbers where \( L \) would be the number of grid-points per dimension. At the other extreme, a one-dimensional Brownian path would require
Figure 3.2: An example of $V_{T-2}(q, x^i)$ over a simplex slice in coordinates $(q^2, q^3, q^5)$ with $x^i = 5$

a very large number of data points for reasonable approximate representation, even though $N = 1$. We see that complexity and dimension are not necessarily related, and the curse-of-dimensionality is only an artifact of the classical representation form. With certain idempotent-based algorithms, a low-complexity solution approximation might be obtained regardless of state-space dimension if the solution is indeed low-complexity. We do not make this more rigorous here.

The curse-of-complexity associated with the idempotent-based approach is evident in the extremely rapid growth of $Z^x_t$ as one propagates backwards. In practice, it is often found that most of the linear functionals provide either zero additional value or very little. Also, [54], suggests that one should expect at worst polynomial growth rather than exponential.

In order to reduce the solution complexity, one should project the solution approximation, at each step, down onto a lower-dimensional min-plus subspace. It is known that the optimal projection is actually obtained through pruning of the set of linear functionals (cf., [34]). Consequently, our optimal complexity attenuation problem reduces to an optimal pruning problem. In particular, one seeks the optimal projection
onto a subspace of dimension, say $N$. Given the above, we see that this is obtained by optimally pruning $Z^x_t$ to a subset of size $N$. As the computations for selection of the optimal subset are costly, we apply a greedy (suboptimal) approach. One begins with the functional which obtains the minimum value over $S^{N_e}$. Given a subset of size $k < N$, one chooses the next functional from the remaining set, according to which additional functional would maximally reduce the pointwise minimum in an $L_\infty$ sense. This process is repeated until one has a subset of size $N$.

There are two significant issues in this process. First, one must be able to evaluate the $L_\infty$ change in the pointwise minimum which would be obtained by the inclusion of a new linear functional. This is more easily computed if one has the vertices of the piecewise linear, concave functional formed by the previously selected set of functionals. The second issue is the sequential computation of these vertices. We discuss each of these issues separately.

We begin with the first issue: the maximally-greedy, selection of the next linear functional to add to the current set. Here, the optimality is taken in terms of the $L_\infty$ difference between the pointwise minimum of the set of $k$ linear functionals and the pointwise minimum of the set of $k+1$ linear functionals obtained by the addition of one more functional. Motivation for selection of the $L_\infty$ norm can be found in [34].

We now describe the greedy algorithm, and the computational efficiencies that may be obtained in implementation. Suppose we have $\bar{V}_t$ in the form (3.10), Let $e(q) = c_b q^{N_b} + c_e q^{N_e}$, and note that we may then write $\bar{V}_t$ as

$$\bar{V}_t(q, x^i) = \min_{z \in Z^x_t} \left[ b^{x^i,z} \cdot \mathcal{H}(q) \right] + e(q) = \tilde{V}_t(\mathcal{H}(q), x^i) + e(q).$$ (3.14)

Fix any $x^i$. Suppose we wish to find an approximately optimal projection/pruning of $\tilde{V}_t(\cdot, x^i)$ with complexity $N < Z^x_t$, that is an approximation

$$\tilde{V}^N_t(\mathcal{H}(q), x^i) = \min_{z \in \tilde{Z}} \left[ b^{x^i,z} \cdot \mathcal{H}(q) \right] \quad \forall q \in S^{N_e},$$ (3.15)

where $|\tilde{Z}| = N$. Again, by [34] the optimal projection is obtained by pruning. The maximally-greedy (suboptimal) algorithm we employ is as follows. First, we select $z_1 \in \text{argmin} \left\{ \min_{q \in S^{N_e}} \left[ b^{x^i,z_1} \cdot q \right] \mid z \in Z^x_t \right\}$, and let $Z'_1 = \{z_1\}$. Given $Z'_k$, we select $z_{k+1}$, and let $Z'_{k+1} = Z'_k \cup \{z'_{k+1}\}$, as follows.
Let
\[ z_{k+1} \in \text{argmax} \left\{ \max_{q \in S^{N_e}} \left[ \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot z} \cdot \mathcal{H}(q)) - \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot z} \cdot \mathcal{H}(q)) \right] \mid z \in Z^{x_{t+1}}_t \right\} \]
\[ = \text{argmax} \left\{ \max_{q \in S^{N_e}} \left[ \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot q} - \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot z} \cdot q) \right] \mid z \in Z^{x_{t+1}}_t \right\} \]
\[ = \text{argmin} \left\{ \min_{q \in S^{N_e}} \left[ \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot q} - \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot z} \cdot q) \right] \mid z \in Z^{x_{t+1}}_t \right\} \]
\[ = \tilde{Z}'_{k+1}. \]

Suppose
\[ \min_{q \in S^{N_e}} \left[ \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot q} - \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot z} \cdot q) \right] < 0, \quad (3.16) \]
i.e., that there is a \( z \in Z^{x_{t+1}}_k \setminus \tilde{Z}'_k \) which improves the pointwise minimum. (Otherwise, nothing will be gained by the addition of more functionals.) Then
\[ \tilde{Z}'_{k+1} = \text{argmin} \left\{ \min_{q \in S^{N_e}} \left[ b^{x_{t+1} \cdot q} - \min_{z \in Z^{x_{t+1}}_k} (b^{x_{t+1} \cdot z} \cdot q) \right] \mid z \in Z^{x_{t+1}}_t \right\} \]
\[ = \text{argmin} \left\{ \min_{q \in S^{N_e}} \left[ b^{x_{t+1} \cdot q} - \gamma(q, \tilde{Z}'_k) \right] \mid z \in Z^{x_{t+1}}_t \right\}. \quad (3.17) \]

Now we examine algorithms for efficient computation of the right-hand side of (3.17). For \( Z'_k \subseteq Z^{x_{t+1}}_t \), let \( \mathcal{V}_k = V_k(Z'_k) \) be the set of vertices generated by hyperplanes associated to the linear functionals \( b^{x_{t+1} \cdot z} \cdot q \) for \( z \in Z'_k \) as well as the boundary planes of \( S^{N_e} \). Let \( Q(\mathcal{V}_k) \) be the set of \( q \in S^{N_e} \) associated to each vertex in \( \mathcal{V}_k \), i.e.,
\[ Q(\mathcal{V}_k) \doteq \{ q \in S^{N_e} \mid (q, \gamma(q, Z'_k)) \in \mathcal{V}_k \}. \]

It is not difficult to see that
\[ \tilde{Z}'_{k+1} = \text{argmin} \left\{ \min_{q \in Q(\mathcal{V}_k)} \left[ b^{x_{t+1} \cdot q} - \gamma(q, \tilde{Z}'_k) \right] \mid z \in Z^{x_{t+1}}_t \right\}. \quad (3.18) \]

Note that
\[ \min_{z \in Z^{x_{t+1}}_t} \min_{q \in Q(\mathcal{V}_k)} \left[ b^{x_{t+1} \cdot q} - \gamma(q, \tilde{Z}'_k) \right] \]
\[ = \min_{q \in Q(\mathcal{V}_k)} \left\{ \min_{z \in Z^{x_{t+1}}_t} \left[ b^{x_{t+1} \cdot q} - \gamma(q, \tilde{Z}'_k) \right] \right\}, \]
which by Theorem 3.3.2,
\[ = \min_{q \in Q(\mathcal{V}_k)} \left\{ \min_{(u^o, u^i) \in \mathcal{N}_x \times I^{x_i}} \min_{(z_{l+1})} \left[ \sum_{y \in \mathcal{Y}} b^{z_{l+1} \cdot u^o, u^i, \zeta, y} \cdot q - \gamma(q, \tilde{Z}'_k) \right] \right\}, \quad (3.19) \]
where \( b^{z_{l+1} \cdot u^o, u^i, \zeta, y} = D(\tilde{\mathcal{R}}^{u^o, y} \tilde{P}^{u^i} \mathcal{G}(b^{z_{l+1} \cdot \zeta, y}) \) for all \( x^i \in \mathcal{N}_e \), all \( (u^o, u^i) \in \mathcal{N}_x \times I^{x_i} \) and all \( (\zeta, y) \in P^{N_y}(Z^{x_{l+1}}_t) \times \mathcal{Y} \).
Theorem 3.4.1. Given finite index set $Z$, $x^i \in \mathcal{N}_b$, $(u^o, u^i) \in \mathcal{N}_s \times \mathcal{I}^{x^i}$, $q \in \mathcal{S}^{N_s}$ and $y \in \mathcal{Y}$, let
\[
\zeta_y = \zeta_y^{u^o, u^i}(q, Z) \in \arg\min_{\zeta \in Z} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q.
\] (3.20)

Also let
\[
\tilde{Z}^{u^o, u^i}(q, Z) = \{ \zeta_y \} = \{ \zeta_y \}_{y \in Y} \in \mathcal{P}^{N_y}(Z), \quad \zeta_y \text{ satisfies (3.20) } \forall y \in \mathcal{Y}.
\] (3.21)

Then, for any finite index set $Z$, $x^i \in \mathcal{N}_b$, $(u^o, u^i) \in \mathcal{N}_s \times \mathcal{I}^{x^i}$ and $q \in \mathcal{S}^{N_s}$,
\[
\min_{\{ \zeta \} \in \mathcal{P}^{N_y}(Z)} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right] = \min_{\{ \zeta \} \in \tilde{Z}^{u^o, u^i}(q, Z)} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right].
\]

Proof. Fix index set $Z$, $(u^o, u^i) \in \mathcal{N}_s \times \mathcal{I}^{x^i}$ and $q \in \mathcal{S}^{N_s}$. By set inclusion, it is immediate that
\[
\min_{\{ \zeta \} \in \mathcal{P}^{N_y}(Z)} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right] \leq \min_{\{ \zeta \} \in \tilde{Z}^{u^o, u^i}(q, Z)} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right].
\]

We prove the reverse inequality. Let $\{ \tilde{\zeta} \} = \{ \tilde{\zeta}_y \}_{y \in \mathcal{Y}} \in \mathcal{P}^{N_y}(Z)$ and $\{ \zeta \} = \{ \zeta_y \}_{y \in \mathcal{Y}} \in \tilde{Z}^{u^o, u^i}(q, Z)$. By (3.21), for each $y \in \mathcal{Y}$
\[
\tilde{b}_t^{x^i, u^o, u^i, \tilde{\zeta}, y} \cdot q \geq \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q,
\]
which implies
\[
\sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \tilde{\zeta}, y} \cdot q \geq \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q.
\]

Since this is true for all $\{ \tilde{\zeta} \} \in \mathcal{P}^{N_y}(Z)$,
\[
\min_{\{ \zeta \} \in \mathcal{P}^{N_y}(Z)} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right] \geq \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q
\]
\[
= \min_{\{ \zeta \} \in \tilde{Z}^{u^o, u^i}(q, Z)} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right].
\]

Employing Theorem 3.4.1 in (3.19) yields
\[
\min_{x \in \mathcal{Z}^{x^i}} \min_{q \in \mathcal{Q}(\mathcal{V}_k)} \left[ b_t^{x^i, z} \cdot q - \gamma(q, Z_k) \right]
\]
\[
= \min_{q \in \mathcal{Q}(\mathcal{V}_k)} \left\{ \min_{(u^o, u^i) \in \mathcal{N}_s \times \mathcal{I}^{x^i}} \left[ \min_{\{ \zeta \} \in \tilde{Z}^{u^o, u^i}(q, Z_{k+1}^{x^i})} \left[ \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y} \cdot q \right] - \gamma(q, Z_k) \right] \right\},
\]
and letting $\tilde{b}_t^{x^i, u^o, u^i, \{ \zeta \}} = \sum_{y \in \mathcal{Y}} \tilde{b}_t^{x^i, u^o, u^i, \zeta, y}$, this is
\begin{align}
= \min_{q \in Q} \left\{ \min_{(w^u, u^d) \in \mathcal{N} \times \mathcal{Q}} \left( \min_{(q^*, u^a, u^d, \{\zeta\}) \in \tilde{Z}^a} \min_{x^i} \left( b^a_{t,x^i} - q^* - \gamma(q^*, \mathcal{Z}^a) \right) \right) \right\}, \tag{3.22}
\end{align}

Suppose the minimum on (3.22) is achieved at \( q^*, u^a, u^d, \{\zeta^*\} \). Then, one lets
\[ z_{k+1} = M^{-1}(u^a, u^d, \{\zeta^*\}) \] and \( \mathcal{Z}^{a}_{k+1} = \mathcal{Z}^a_k \cup \{z_{k+1}\} \). We continue the procedure until \( k + 1 = N \) (or until the left-hand side of (3.16) is zero). The resulting set of indices is \( \tilde{Z} \). We let \( \tilde{V}^N_t \) be given by (3.15). As in (3.14),(3.15), for each \( x^i \in \mathcal{N} \) we obtain the approximation
\begin{align}
\tilde{V}_t(q, x^i) & \simeq \tilde{V}^N_t(q, x^i) \\
& = \min_{z \in \tilde{Z}} \left[ b^{x^i,z} \cdot \mathcal{H}(q) \right] + c_b q^N + c_e q^N. \tag{3.23}
\end{align}

For further backward propagation, we could replace \( \mathcal{Z}^{a}_{k} \) with \( \tilde{Z} \), and \( \tilde{V}_t \) with \( \tilde{V}^N_t(q, x^i) \). Each step in the resulting approximate, idempotent form of the dynamic programming consists of two parts: the backward propagation according to Theorem 3.3.2, followed by the projection down to a set of cardinality \( N \). Let the backward propagation via idempotent representation operator (Theorem 3.3.2) be denoted by \( S \). That is, \( \tilde{V}_t = S[\tilde{V}_{t+1}] \). Let the above min-plus projection operator onto \( N \)-dimensional min-plus subspace be denoted by \( \Pi^N \), where above, this is \( \tilde{V}^N_t = \Pi^N[\tilde{V}_t] \). Together, the approximate idempotent backward dynamic program is
\begin{align}
\tilde{V}_t = \Pi^N[S[\tilde{V}_{t+1}]], \tag{3.24}
\tilde{V}_T(q, x^i) = \tilde{\beta} \cdot q \quad \forall q \in \mathcal{N}, \forall x^i \in \mathcal{N} \tag{3.25}
\end{align}

The efficient implementation of the projection substep of the above algorithm requires an ability to sequentially produce the vertices of the convex polytope generated by the \( \mathcal{Z}^{a}_{k} \) as one iterates over \( k \in ]1, N[ \). More specifically, the polytope is the hypograph of \( \gamma(\cdot, \mathcal{Z}^{a}_{k}) \) over \( \mathcal{N} \). The sequential computation of this set of vertices is very efficiently performed by the double description method \cite{45}. Let us describe the double description approach to convex polytope representation a bit. Note that the polytope may be uniquely represented either by the set of hyperplanes (more exactly, half-spaces) defining it, or by its set of vertices. The key to the algorithm is that one keeps track of both the set of half-spaces and the set of vertices at each step. The half-spaces defining the hypograph of \( \gamma(\cdot, \mathcal{Z}^{a}_{k}) \) are those indexed by the functionals, i.e., the \( z \in \mathcal{Z}^{a}_{k} \), plus the half-space boundaries of \( \mathcal{N} \). If we let \( w \in \mathcal{R} \) denote the range dimension and \( q^l \in \mathcal{N} \subset \mathcal{R} \) denote the domain dimensions, each \( z_j \in \mathcal{Z}^{a}_{k} \) corresponds to a half-space \( w - b^{x^i,z_j} \cdot q^l \leq 0 \). (For our purposes, we do not need to include a “bottom” for
At each additional \( z_{k+1} \), we are adding a half-space constraint, \( w - b_{k}^{x^i,z_{k+1}} \cdot q' \leq 0 \), with corresponding face \( w - b_{k}^{x^i,z_{k+1}} \cdot q' = 0 \). Recall the set of vertices, prior to the additional \( z_{k+1} \), is denoted by \( \mathcal{V}_{k} \). One needs to determine the resulting new set of vertices, \( \mathcal{V}_{k+1} \), efficiently. The double description method does this.

The hyperplane \( w - b_{k}^{x^i,z_{k+1}} \cdot q' = 0 \) partitions \( \mathcal{V}_{k} \) into two subsets, \( \mathcal{V}_{k}^{+} \) and \( \mathcal{V}_{k}^{-} \), where
\[
\mathcal{V}_{k}^{-} = \{ (q'_\alpha, w_\alpha) \in \mathcal{V}_{k} \mid w_\alpha - b_{k}^{x^i,z_{k+1}} \cdot q'_\alpha < 0 \} \quad \text{and} \quad \mathcal{V}_{k}^{+} = \{ (q'_\alpha, w_\alpha) \in \mathcal{V}_{k} \mid w_\alpha - b_{k}^{x^i,z_{k+1}} \cdot q'_\alpha > 0 \}.
\]
We ignore the special case of equality in our brief discussion here. Those vertices in \( \mathcal{V}_{k}^{+} \) are dropped; they are outside the new polytope. The vertices in \( \mathcal{V}_{k}^{-} \) are retained. Where there is \((q'_\alpha, w_\alpha) \in \mathcal{V}_{k}^{-}\) and \((q'_\beta, w_\beta) \in \mathcal{V}_{k}^{+}\) which are joined by an edge of the polytope, one must compute the resulting new vertex formed by the intersection of that edge and the new hyperplane \( w - b_{k}^{x^i,z_{k+1}} \cdot q' = 0 \). The union of this set of new vertices and \( \mathcal{V}_{k}^{-} \) is \( \mathcal{V}_{k+1} \).

### 3.4.1 \( \mathcal{L}_\infty \) error bounds

Recall from Theorem 3.3.2 and (3.24), that the ideal idempotent backward propagation is
\[
\overline{V}_t = \mathcal{S}[\overline{V}_{t+1}],
\]
and the approximate idempotent backward propagation is
\[
\hat{V}_t = \Pi^N [\mathcal{S}[\hat{V}_{t+1}]],
\]
both with the same terminal condition, \( \hat{V}_T(q, x^i) = \overline{V}_T(q, x^i) = \bar{\beta} \cdot q \). Let the one-step projection error be
\[
e^N_t = \left\| \hat{V}_t - \mathcal{S}[\hat{V}_{t+1}] \right\|_\infty = \left\| \Pi^N [\mathcal{S}[\hat{V}_{t+1}]] - \mathcal{S}[\hat{V}_{t+1}] \right\|_\infty,
\]
where \( \left\| \hat{V}_t \right\|_\infty = \max_{x \in \mathcal{N}_e} \sup_{q \in \mathcal{S}[\hat{V}_t]} |\hat{V}_t(q, x)| \). Let the total error be denoted as \( \bar{e}_t = \left\| \hat{V}_t - \overline{V}_t \right\|_\infty \). Then,
\[
\bar{e}_t = \left\| \hat{V}_t - \overline{V}_t \right\|_\infty \\
\leq \left\| \hat{V}_t - \mathcal{S}[\hat{V}_{t+1}] \right\|_\infty + \left\| \mathcal{S}[\hat{V}_{t+1}] - \overline{V}_t \right\|_\infty \\
= \left\| \hat{V}_t - \mathcal{S}[\hat{V}_{t+1}] \right\|_\infty + \left\| \mathcal{S}[\hat{V}_{t+1}] - \mathcal{S}[\overline{V}_{t+1}] \right\|_\infty \\
= e^N_t + \left\| \mathcal{S}[\hat{V}_{t+1}] - \mathcal{S}[\overline{V}_{t+1}] \right\|_\infty
\]
which by Theorem 4.2 in [47],
\[ \leq e_t^N + \| \hat{V}_{t+1} - V_{t+1} \|_\infty = e_t^N + \bar{e}_{t+1}. \]

Applying an induction argument, one finds

\[ e_t^N \leq \sum_{r=t}^{T-1} e_r^N. \]

This is, of course, a conservative bound on the error growth rate. If the worst-case projection errors occurred at different \((q, x^i)\) locations at different steps, then the errors would not grow additively. Secondly, the probability mass moves toward nodes \(N_b\) and \(N_e\) (where there is no error in the value) as one propagates forward, and so we can expect reduction in error effects from this.

### 3.5 Examples

In order to give a sense of the value of the above tools as well as their computational tractability, we exercise the tools on an example. The results appear in Table 3.3. In particular, we apply the above techniques to the problem depicted in Figure 3.1, where there are 8 UGSs, one UAV, one interceptor and one intruder. The road network and UGS locations are exactly as depicted in the figure. The payoffs are \(-100\) points if the interceptor is co-located with the intruder prior to the intruder reaching the base \((c_e = -100)\), \(100\) points if the intruder gets to the base \((c_b = 100)\), and zero otherwise. The interceptor initial position is indicated in column 1. The initial probability distribution regarding the intruder is uniform over the locations other than \(N_b\) and \(N_e\), where we note that \(x_i^i = N_b\) in the last row. The final column is the value function at time \(T - 7\), computed by the above algorithm, with \(N = 20\) functional hyperplanes retained at each step. The second column is a check on the value function computation, obtained by applying the optimal control obtained during the value function (idempotent backward dynamic program) computation. The number of runs in the Monte Carlo tests per data point is 300,000. The initial intruder position was generated according to the uniform distribution indicated above. Regarding observation errors, the probability of a false negative ran between 0.004 and 0.044, depending on location, and the probability of a false positive similarly ran from 0.004 to 0.044. The probability transition matrix is large, and we do not include all the data. Comparing the second and last columns, we see that there is good correspondence between the Monte Carlo results and the computed value.
### Table 3.1: Probability transition matrix

<table>
<thead>
<tr>
<th>( x_t^R )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.20</td>
<td>0.45</td>
<td>0.35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>0.60</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.45</td>
<td>0.25</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0</td>
<td>0.15</td>
<td>0.10</td>
<td>0</td>
<td>0.60</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.10</td>
<td>0.25</td>
<td>0</td>
<td>0.05</td>
<td>0</td>
<td>0.35</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td>0.10</td>
<td>0</td>
<td>0.02</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
<td>0.20</td>
<td>0</td>
<td>0.40</td>
<td>0.30</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
<td>0</td>
<td>0.15</td>
<td>0</td>
<td>0.75</td>
</tr>
<tr>
<td>Base</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Columns 3 and 4 indicate expected payoffs for heuristic controllers, also obtained by the Monte Carlo method with the same number of runs. The first heuristic controller is as follows. At each step, the UAV moves to the location with the a priori highest probability of containing the intruder. The interceptor moves to the most likely next position of the intruder based on the current observation-conditioned distribution and the probability transition matrix. The results with this heuristic appear in column 3, and are substantially inferior to those obtained with the computed optimal controls. Note from Figure 3.1, that the intruder must pass through either node 7 or node 8 just prior to reaching the base. The second heuristic controller exploits this. In the second heuristic controller, the interceptor oscillates between nodes 7 and 8, while the UAV acts exactly as in the first heuristic. In the case where the initial interceptor position is 6, we let the interceptor move first to node 7, and then begin oscillating. In the case where the initial interceptor position is 5 or 9 = \( N_b \), we let the interceptor move first to node 8, and then begin oscillating. We did not apply the second heuristic in cases where the initial interceptor position was 1, 2, 3 or 4. Clearly, the second heuristic controller also substantially underperforms relative to the computed optimal controls.
Table 3.2: Confusion matrix

<table>
<thead>
<tr>
<th>UAV control, $v^o$</th>
<th>Probability</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False negative</td>
<td>False positive</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.044</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.029</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.015</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.018</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.038</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.060</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.008</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.004</td>
<td>0.018</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Monte Carlo testing results

<table>
<thead>
<tr>
<th>Initial interceptor position, $x^i$</th>
<th>Monte Carlo</th>
<th>Payoff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monte Carlo</td>
<td>Heuristic 1</td>
</tr>
<tr>
<td>1</td>
<td>-7.7</td>
<td>23.9</td>
</tr>
<tr>
<td>2</td>
<td>-25.8</td>
<td>8.4</td>
</tr>
<tr>
<td>3</td>
<td>-26.7</td>
<td>17.7</td>
</tr>
<tr>
<td>4</td>
<td>-14.5</td>
<td>20.8</td>
</tr>
<tr>
<td>5</td>
<td>-39.7</td>
<td>-4.9</td>
</tr>
<tr>
<td>6</td>
<td>-33.6</td>
<td>5.8</td>
</tr>
<tr>
<td>7</td>
<td>-50.4</td>
<td>4.1</td>
</tr>
<tr>
<td>8</td>
<td>-54.0</td>
<td>-13.5</td>
</tr>
<tr>
<td>Base</td>
<td>-40.6</td>
<td>-12.5</td>
</tr>
</tbody>
</table>
3.6 Closing comments

A high-dimensional partially-observed stochastic control problem with observation-process control has been defined and motivated. In addition to the UAV observation process, the interceptor also provides both physical-domain control and observational data, and so has a dual-control nature. The Separation Principle is applied to create a control problem formulation with state space which is essentially $S^{N_s} \times N_s$. This is a high-dimensional problem for moderately-sized examples, and so classical methods may not be computationally tractable.

A min-plus curse-of-dimensionality-free algorithm is developed for the problem. The associated curse-of-complexity is attenuated through repeated projection to a low-dimensional min-plus subspace. The optimal projection is obtained by pruning, and a greedy pruning algorithm, computationally enhanced by use of the double description method, is developed.

This has been, so far, demonstrated to perform well on problems with up to ten UGSs ($N_s = 10$). Here, an example with $N_s = 8$ is presented. We see that the resulting controls greatly outperform a reasonable heuristic. It is shown that the errors grow at most linearly with time-horizon length. However, that bound is clearly excessively conservative for long time horizons.

The actual per-step errors expected as a function of representation complexity are unknown, and likely problem-dependent. Development of an approach to estimation of these errors is one open problem. The possibility of additional computational efficiencies is another open problem. On a more theoretical note, demonstration of the applicability of the Separation Principle (or counterexample to such) for this problem class would be a useful advance.

Chapter 3, in full, is a reprint of the material as it appears in Recent Advances in Research on Unmanned Aerial Vehicles, Springer-Verlag, 2013. The dissertation author was the primary author of this chapter.
Chapter 4

Optimization Formulation and Monotonic Solution Method for the Witsenhausen Problem

4.1 Introduction

We examine the structure of the Witsenhausen counterexample/problem [55] and its solution. The problem can be viewed as a benchmark in the area of networked control problems, which is of course an area of current interest. The problem formulation is straight-forward. Although it is typically viewed as a two time-step problem, we effectively transform it into a static optimization problem. However, it is the original interpretation of the problem that makes it a problem of substantial interest to the controls community. That is, a controller receives a random variable with Gaussian density, and acts to modify this input. An observation of the resulting output random variable is corrupted by additive Gaussian noise. Finally, an estimator attempts to determine the output from this observation. The goal is to minimize the sum of a measure of the controller effort and the expected squared estimate error. Consequently, the controller is attempting not only to minimize its own effort, but also to aid the estimator through its control action, and it is the tension between these competing costs that yields the interesting problem structure.

We find it useful to work with the associated quantile function, rather than the controller itself or its distribution. With this transformation, the problem can be reduced to
minimization of a criterion over a cone in $L_2(0,1)$. The optimization criterion is the sum of two functionals. The first, representing the control cost, is a simple quadratic, the form being well-known from Monge-Kantorovich theory. The second, representing the expected squared estimation error, has a more complex structure over this space.

The contributions of the paper are as follows.

- The problem is reformulated as minimization over a cone in $L_2(0,1)$, which aids in the development of necessary conditions.
- It is demonstrated that under a condition regarding the zeros of a certain functional, neither of the cost terms, separately, have isolated local minima.
- Necessary conditions which are beyond first-order are rigorously obtained. The optimization is over a quantile function, $G$, and an estimator, $e$. The conditions will hold if variation in either $G$ or $e$ separately implies that the total cost cannot decrease.
- A numerical method is obtained, which generates a sequence of approximations with monotonically decreasing cost, and which terminates only at a solution of the above conditions.
- A structure in the solution, where jumps in $G$ correspond to switching among solutions of a scalar nonlinear equation will be seen in the resulting plots.

In Section 4.2, we review the problem, and very briefly indicate references to earlier results. In Section 4.3, the problem is reformulated in terms of optimization over quantile functions. In Section 4.4, we examine the two components of the cost separately. In Section 4.5, we obtain the aforementioned necessary conditions (where we allow variation in both the quantile function and the estimator). Then, in Section 4.6, the method is obtained, and its properties are indicated. Lastly, in Section 4.7, we present and examine some numerical results.

### 4.2 Background and Definition

The problem formulation is quite simple, and as noted above, one might place it in the arena of optimization rather than control, as one could argue that the problem does not have the time-structure which distinguishes control from optimization. The problem is as follows. The first input is a scalar normal random variable, $W \sim \mathcal{N}(0,c)$, and we let its range be denoted as $\mathcal{W} \subseteq \mathbb{R}$. A “controller”, $\zeta : \mathcal{W} \to \mathbb{R}$, acts additively on the first input, generating output $X = W + \zeta(W)$. We assume that $\zeta(\cdot)$ is measurable, where
we recall that the existence of an optimal, measurable control is already well-known (cf., [56, 55]). An observation, $Y = X + W^o$ is made, where $W^o \sim \mathcal{N}(0, d)$, and we let the range of $Y$ be denoted by $\mathcal{Y} \triangleq \mathbb{R}$. The estimator generates estimate $e(Y)$, knowing $Y$, but not $\zeta(W)$. Note that we assume the estimator does know the control strategy to be followed, $\zeta(\cdot)$, but not the actual control applied. The payoff to be minimized is

$$J(\zeta(\cdot), e(\cdot)) \triangleq \mathbb{E} \left\{ k_0 |\zeta(W)|^2 + |X - e(Y)|^2 \right\},$$

where $k_0 \in [0, \infty)$. Due to the squared-error form of the second term on the right, the optimal estimate is the conditional expectation, $\hat{e}_Y \triangleq \mathbb{E}[X \mid Y]$, (cf., [55]), and given that, we let

$$J(\zeta) \triangleq \mathbb{E} \left\{ k_0 |\zeta(W)|^2 + |X - \hat{e}_Y|^2 \right\}. \quad (4.2)$$

Clearly, the solution depends only on the three parameters, $c$, $d$, and $k_0$. Upon examining (4.2), we see that an optimal control must also have finite variance. Consequently, we take the control space to be

$$\mathcal{Z} \triangleq \{ \zeta : \mathcal{W} \to \mathbb{R} \mid \text{measurable, } \mathbb{E}[\zeta^2(W)] < \infty \}. \quad (4.3)$$

We let

$$V = V(c, d, k_0) = \inf_{\zeta \in \mathcal{Z}} J(\zeta) = \inf_{\zeta \in \mathcal{Z}} J(\zeta; c, d, k_0). \quad (4.4)$$

In this form, we see that the problem reduces to an (infinite-dimensional) optimization problem.

A great deal of quite interesting work has used this problem as a basis for development (cf., [3, 2, 10, 17, 20, 19, 26, 27, 44, 48, 49, 51, 56] and the references therein). Of particular relevance to the analysis here is [26]. In [26], the authors assume a signaling structure (originally suggested in [55]) for the controller, where the controller acts to make $X$ take on one of a small finite set of possible values, the selection of which is based on input $W$. This allows the estimator to correctly identify $X$ with high probability, particularly if the gap between possible $X$ values is relatively large compared with $\sqrt{d}$. Using a more general approach here (see also [42, 56]), we find that optimal solutions with structure similar to this, but not exactly such, emerge naturally in an interesting region of parameter space, while solutions similar to normal random variables, corresponding to nearly linear controllers, occur in another region.
4.3 Transformation to Quantile Representation

We will find it helpful to optimize not over the controller, but instead over the resulting distribution of $X$, $F_X$. Further, in many places we will find it helpful to perform the bulk of the analysis not with the distribution function, but with the corresponding quantile function, which we denote as $G$. The reason is that the minimization is then taken over a cone in $L_2(0,1)$, rather than over the space of probability distributions. Consequently, it is helpful to review the transformation between the two representations.

Let $\mathcal{F}$ denote the space of probability distribution functions on $\mathbb{R}$ with finite second moments. Let $\mathcal{G}$ denote the space of square-integrable quantile functions given by

$$\mathcal{G} = \left\{ G : (0,1) \to \mathbb{R} \mid \int_{(0,1)} G^2(u) \, du < \infty, \text{ monotonically increasing, continuous on the left, limits on the right} \right\}.$$ 

Given $F \in \mathcal{F}$, for all $u \in (0,1)$ let

$$\mathcal{I}[F](u) = \inf\{x \mid F(x) \geq u\}.$$ 

(4.5)

Theorem 4.3.1. $\mathcal{I}$ is a bijection from $\mathcal{F}$ to $\mathcal{G}$.

This result is classical, and hence we do not include a proof. For the smooth case, one may see [14], Sec. 1.4. For the more general case, one can refer to, for instance, [6], Sec. 1.5. We note that [6] uses the right-continuous representations of the elements of $\mathcal{G}$.

Given $G \in \mathcal{G}$, for all $x \in \mathbb{R}$ let

$$\mathcal{J}[G](x) \triangleq \begin{cases} \sup\{u \mid G(u) \leq x\} & \text{if } \{u \mid G(u) \leq x\} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 4.3.2. $\mathcal{J} = \mathcal{I}^{-1}$.

Again, we refer the reader to standard texts (cf., [6, 14] as above) for more details and proof.

Remark 4.3.3. For purposes of intuition, it is helpful to consider the simple smooth, strictly increasing case. In this case, with $G \equiv \mathcal{I}[F]$, we may write $G = F^{-1}$ where the inverse function is interpreted in the classical sense, and one has, formally,

$$du = \frac{du}{dx} \, dx = \frac{dF}{dx} \, dx = f(x) \, dx.$$
with \( f \) denoting the corresponding density. For further intuition, we remark that \( G(u) = \bar{x} \in \mathbb{R} \) for all \( u \in (0, 1) \) corresponds to a probability mass of one at \( \bar{x} \). Also, discontinuities in \( G \) correspond to flat sections in \( F \), i.e., intervals where the control places no mass.

**Remark 4.3.4.** Further, for square-integrable \( H \),

\[
E[H(X)] = \int_{\mathbb{R}} H(x) \, dF(x) = \int_{(0,1)} H(G(u)) \, du,
\]

where this equivalence holds in the general case, and we do not include the proof (cf., [14], Sec. 3.4).

We now examine the structure of the cost criterion, \( J \), using a quantile functional representation. Let

\[
\hat{A}(\zeta) \doteq E \{ k_0 |\zeta(Y)|^2 \} \quad \text{and} \quad \hat{B}(\zeta) \doteq E \{ |X - \hat{e}_y|^2 \}.
\]

(4.7)

First, we look at \( \hat{B} \). Note that the conditional expectation of \( X \) given \( Y = y \) is

\[
\hat{e}_y \doteq E \{ X \mid Y = y \} = K_1/K_0 = \frac{1}{\int_{\mathbb{R}} h_d(x, y) \, dF_X(x)} \int_{\mathbb{R}} xh_d(x, y) \, dF_X(x)
\]

(4.8)

where

\[
h_d(x, y) \doteq \frac{1}{\sqrt{2\pi d}} \exp \left[ -\frac{(x - y)^2}{2d} \right].
\]

Employing change of variables (4.6), this becomes

\[
\hat{e}_y = \hat{e}_y(G) = \frac{\int_{(0,1)} G(u)h_d(G(u), y) \, du}{\int_{(0,1)} h_d(G(u), y) \, du},
\]

(4.9)

where \( G = \mathcal{I}[F_X] \) and \( du \) indicates integration with respect to Lebesgue measure, and we will occasionally suppress the argument of \( \hat{e}_y \). Further, noting that for Borel measurable \( \mathcal{C} \subseteq \mathbb{R} \), \( P(Y \in \mathcal{C}) = \int_{\mathbb{R}} \int_{\mathbb{R}} h_d(x, y) \, dF_X(x) \, dy \), we see that

\[
\hat{B}(\zeta) = \int_{\mathbb{R}} \int_{\mathbb{R}} |x - \hat{e}_y|^2 h_d(x, y) \, dF_X(x) \, dy
\]

\[
= \int_{\mathbb{R}} \int_{(0,1)} |G(u) - \hat{e}_y|^2 h_d(G(u), y) \, du \, dy \doteq B(G).
\]

(4.10)

where the last equality is due to a change of variables. Next, we look to represent \( \hat{A} \) in terms of the quantile function corresponding to \( X \). Due to the explicit presence of \( \zeta \), one should examine the transformation carefully. Note that \( \zeta \) wishes to transform input \( W \) into some form (presumably more useful to the estimator), \( X \). Considering (4.10),
we see that the expected estimator error, $B$, depends on $\zeta$ only through the resulting distribution, $F_X$, or equivalently, the quantile function $G$. Consequently, $\zeta$ would like to generate any given $F_X$, with the minimum squared effort given by $\hat{A}$. It seems intuitively clear that in order to minimize cost, for any given $F_X$, one would choose a monotonically increasing $\zeta(\cdot)$. However, to be completely rigorous, we do not presume this form, but find this form, along with the representation of the problem, in terms of the quantile function.

Let $F_W$ denote the distribution corresponding to (normally distributed) input $W$, and let $G_W = \mathcal{I}[F_W]$, which is of course, in $C^\infty(\mathbb{R})$. The following result is an application of Monge-Kantorovich theory. For a proof, see for instance, [43, 8, 9, 46, 50, 53]. (Also, for a self-contained sketch of a proof, one may see [42].)

**Theorem 4.3.5.** For any $G \in \mathcal{G}$,

$$\min_{\zeta \in \hat{Z}(G)} \mathbb{E}[\zeta^2] = \int_{(0,1)} [G(u) - G_W(u)]^2 du,$$

where $\hat{Z}(G) \doteq \{ \zeta \in \mathcal{Z} | \mathcal{I}[F_X] = G \}$.

It is worth noting that an optimal control in Theorem 4.3.5 is obtained as $\zeta(w) = \hat{X}(w) - w$ where

$$\hat{X}(w) = G(F_W(w)), \quad \forall w \in \mathbb{R}, \quad (4.11)$$

and in fact

$$\int_{(0,1)} [G(u) - G_W(u)]^2 du = \int_{\mathbb{R}} [\hat{X}(w) - w]^2 P(dw). \quad (4.12)$$

Given Theorem 4.3.5, it is natural to define

$$A(G) \doteq \int_{(0,1)} [G(u) - G_W(u)]^2 du. \quad (4.13)$$

Recalling (4.4), we have

$$V = V(c, d, k_0) = \inf_{\zeta \in \mathcal{Z}} J(\zeta)$$

$$= \inf_{G \in \mathcal{G}} \min_{\zeta \in \hat{Z}(G)} \left\{ k_0 \mathbb{E}[|\zeta(W)|^2] + \mathbb{E}[|X - \hat{e}_Y|^2] \right\}$$

$$= \inf_{G \in \mathcal{G}} \left\{ k_0 \int_{(0,1)} [G(u) - G_W(u)]^2 du + \int_{\mathbb{R}} \int_{(0,1)} [G(u) - \hat{e}_y(G)]^2 h_d(G(u), y) du dy \right\}$$

$$\doteq \inf_{G \in \mathcal{G}} \bar{J}(G) = \inf_{G \in \mathcal{G}} J(G; c, d, k_0),$$
where the third equality follows from Theorem 4.3.1, and the fourth equality follows by Theorem 4.3.5, (4.7) and (4.10). Noting the above form, we may write

\[ \bar{J}(G) = k_0 A(G) + B(G), \]  

which defines \( A \) and \( B \). Further, given an optimal \( G \), one can construct the corresponding controller \( \zeta \) as given just above (4.11).

### 4.4 Comments on the Components

In this section, we make some remarks on the problem structure. First, in the quantile representation, the \( A \) functional is simply a quadratic, with minimum at \( G = G_W \). We will also see that although \( B \) is more complex, under an assumption regarding the zeros of a certain function, \( B \) possesses no isolated local minima. Consequently, it is only the combination of the two functionals which produces difficulties for optimization. In this section, we demonstrate the above result regarding \( B \). This result, Theorem 4.4.4, is not required for the necessary conditions and monotonic algorithm of later sections, but is included because it helps in elucidation of the structure of the functionals over quantile space.

We say \( G \in \mathcal{G} \) is antisymmetric (around 1/2) if \( G(u) = -G(1-u) \) for almost every \( u \in (0,1) \). (Alternatively, \( G(1/2-\delta) = -G(1/2+\delta) \) for almost every \( \delta \in (0,1/2) \).) Similarly, \( F \in \mathcal{F} \) is antisymmetric (around range value 1/2) if \( F(-x) = 1 - F(x) \) for all \( x \in \mathbb{R} \). Both of these correspond to a symmetric density function when such exists. Let

\[ \mathcal{G}^a \doteq \{ G \in \mathcal{G} \mid G \text{ is antisymmetric} \}. \]

The following is obvious.

**Lemma 4.4.1.** The minimum of \( B \) (as well as the minimum of \( A \)) over \( \mathcal{G} \) is attained on \( \mathcal{G}^a \).

Fix some \( G \in \mathcal{G} \). We will consider certain \( L_2 \) variations around \( G \). Let \( \gamma \in L_2(0,1) \) (with specific form to follow), and \( \delta \in \mathbb{R} \). Recall from (4.8) that \( \hat{e}_y = K_1/K_0 \)
where $K_1, K_0$ are given there. By standard computations,

$$
K_1(G + \delta \gamma) - K_1(G) = \delta \int_{(0,1)} h_d(y, G(v)) \left[ 1 + G(v) \left( \frac{y - G(v)}{d} \right) \right] \gamma(v) dv + \mathcal{O}(\delta^2) ,
$$

(4.15)

$$
K_0(G + \delta \gamma) - K_0(G) = \delta \int_{(0,1)} h_d(y, G(v)) \left( \frac{y - G(v)}{d} \right) \gamma(v) dv + \mathcal{O}(\delta^2) .
$$

(4.16)

Taking a similar differential in (4.8), and employing (4.15) and (4.16), one obtains

$$
\hat{e}_y(G + \delta \gamma) - \hat{e}_y(G) = \delta \int_{(0,1)} h_d(y, G(v)) \left\{ 1 + \left[ G(v) - \frac{K_1(y, G)}{K_0(y, G)} \right] \left( \frac{y - G(v)}{d} \right) \right\} \gamma(v) dv + \mathcal{O}(\delta^2)
$$

(4.17)

where, for clarity, we remark that when $G(v)$ appears as an argument, it indicates dependence on $G$ evaluated at $v$, and when $G$ appears as an argument with no argument of its own, this indicates dependence on the entire function.

Continuing with this process, we find

$$
B(G + \delta \gamma) - B(G) = \delta \int_{(0,1)} \int_{\mathbb{R}} h_d(y, G(u)) \left[ 2(G(u) - \hat{e}_y(G)) + |G(u) - \hat{e}_y(G)|^2 \left( \frac{y - G(u)}{d} \right) \right] dy \gamma(u) du + \mathcal{O}(\delta^2),
$$

(4.18)

where

$$
\Delta^1_y(y, G) \triangleq \int_{(0,1)} 2(\hat{e}_y(G) - G(v)) h_d(y, G(v)) dv,
$$

(4.19)

which by integration and (4.9),

$$
= 2[\hat{e}_y(G)K_0(G) - K_1(G)] = 0 \quad \forall y \in \mathbb{R}, \quad G \in \mathcal{G}.
$$

(4.20)

Employing (4.20) in (4.18), we obtain

$$
B(G + \delta \gamma) - B(G) = \delta \int_{(0,1)} \int_{\mathbb{R}} h_d(y, G(u)) \cdot \left[ 2(G(u) - \hat{e}_y(G)) + \frac{y - G(u)}{d} |G(u) - \hat{e}_y(G)|^2 \right] dy \cdot \gamma(u) du + \mathcal{O}(\delta^2)
$$

$$
= \delta \int_{(0,1)} \int_{\mathbb{R}} b(G, G(u)) \gamma(u) du + \mathcal{O}(\delta^2).
$$

(4.21)
It is worthwhile explicitly noting that for \( \alpha \in \mathbb{R} \),
\[
b(G, \alpha) = \int_{\mathbb{R}} \left\{ h_d(y, \alpha) \left[ 2(\alpha - \hat{e}_y(G)) \frac{y - \alpha}{d} |\alpha - \hat{e}_y(G)|^2 \right] \right\} dy,
\]
which is clearly \( C^\infty \) in \( \alpha \).

Now, recalling that \( G \) is monotonically increasing, there exists at most a countably infinite number of discontinuities. Consequently, there exists a finite or countably infinite set of open intervals \( \{(\beta_k, \beta_{k+1})\}_{k \in \mathbb{K}} \) such that \( G \) is continuous on each open interval and such that \( (0, 1) \setminus \bigcup_{k \in \mathbb{K}} (\beta_k, \beta_{k+1}) \) consists of at most a countably infinite number of points. Suppose \( G \) is not constant on \( (\beta_k, \beta_{k+1}) \). Without loss of generality, \( \beta_k \geq 1/2 \) and \( G(\beta_k) > 0 \). (More specifically, by the antisymmetry, any change we make in \( G \) at \( u > 1/2 \) is presumed to be mirrored at \( \bar{u} = 1 - u \), and for compactness of presentation, we do not include the details.) Then, there exist \( \mu_0, \nu_0 \in (\beta_k, \beta_{k+1}) \) such that \( \mu_0 < \nu_0 \) and \( G(\mu_0) < G(\nu_0) \). Suppose there exists \( \bar{\alpha} \in (G(\mu_0), G(\nu_0)) \) such that \( \varepsilon = |b(G, \bar{\alpha})| > 0 \).

By the continuity of \( G \), there exists \( \bar{c} \in (\mu_0, \nu_0) \) such that \( G(\bar{c}) = \bar{\alpha} \). Let
\[
\mu_1 = \inf \{ \mu \in (\mu_0, \bar{c}) | |b(G, G(\mu))| > \varepsilon/2 \}, \\
\nu_1 = \sup \{ \nu \in (\bar{c}, \nu_0) | |b(G, G(\nu))| > \varepsilon/2 \}.
\]

By the continuity of \( b(G, \cdot) \) and \( G \), \( \mu_1 < \bar{c} < \nu_1 \) and \( G(\mu_1) < G(\nu_1) \). Now, \( G \in \mathcal{G} \) and continuous on \( (\beta_k, \beta_{k+1}) \) implies that there exists a Borel measure, \( \lambda^G \), such that \( G(u) - G(\mu_1) = \int_{(u, \mu_1]} d\lambda^G(v), \quad \forall u \in [\mu_1, \nu_1] \). Let
\[
\gamma(u) = \begin{cases} 
0 & \text{if } u \in (0, \mu_1], \\
G(u) - G(\mu_1) & \text{if } u \in (\mu_1, \nu_1), \\
G(\nu_1) - G(\mu_1) & \text{if } u \in [\nu_1, 1).
\end{cases}
\]

Now, with this choice of \( \bar{\alpha} \) and \( \gamma \), and supposing \( b(G, \bar{\alpha}) < 0 \), from (4.21) with \( \delta > 0 \) one has
\[
B(G + \delta \gamma) - B(G) \leq \frac{-\delta \varepsilon}{2} \int_{(\mu_1, \nu_1]} \gamma(u) du + O(\delta^2)
\]
\[
= \frac{-\delta \varepsilon}{2} \int_{(\mu_1, \nu_1]} \left[ G(u) - G(\mu_1) \right] du + O(\delta^2) < 0,
\]
for \( \delta \) sufficiently small. Similarly, if \( b(G, \bar{\alpha}) > 0 \), with \( \delta \in (-1, 0) \),
\[
B(G + \delta \gamma) - B(G) \leq \frac{-\delta \varepsilon}{2} \int_{(\mu_1, \nu_1]} \left[ G(u) - G(\mu_1) \right] du + O(\delta^2) < 0
\]
for $\delta$ sufficiently close to zero. Consequently, in either case, $G$ cannot be optimal. The case where $\beta_k < 1/2$ and $G(\beta_k) < 0$ is similar. We see that if $G$ is optimal and not constant on some $(\beta_k, \beta_{k+1})$, then $b(G, G(u)) = 0$ for all $u \in (\beta_k, \beta_{k+1})$. As it appears technically demanding to prove, for the present, we assume:

$b(G, \alpha)$ has only isolated zeros as a function of $\alpha$ for any $G \in \mathcal{G}$. \hspace{1cm} (A.1)

The reader may choose to examine (4.22) for an understanding of the motivation behind this assumption. We also note that Assumption (A.1) is used only in the remainder of this section. If $G$ is not constant on $(\beta_k, \beta_{k+1})$, then by (A.1) and the continuity of $G$ over this interval, there exists $u \in (\beta_k, \beta_{k+1})$ such that $b(G, G(u)) \neq 0$, and so $G$ cannot be optimal. We have:

**Lemma 4.4.2.** Assume (A.1). Suppose $B$ has a local minimum at $\bar{G} \in \mathcal{G}$. Then, $\bar{G}$ is piecewise constant.

Of the piecewise constant quantile functions, the entirely constant function is important. Suppose there exists $\bar{g} \in \mathbb{R}$ such that $G(u) = \bar{g}$ for all $u \in (0, 1)$. Then, one easily obtains $\hat{e}_y(G) = \bar{g}$ for all $y \in \mathbb{R}$. Consequently, $|\hat{e}_y(G) - G(u)|^2 = 0$ for all $y, u$, and we see $B(G) = 0$.

Now, suppose $G$ is not constant. Then, noting the monotonicity, there exist $\varepsilon > 0$ and $0 < \bar{u} < \bar{v} < 1$ such that $G(\bar{v}) - G(\bar{u}) = \varepsilon$. This implies

\[
G(u) \leq G(\bar{u}) \quad \forall u \in (0, \bar{u}],
\]

\[
G(u) \geq G(\bar{v}) \quad \forall u \in [\bar{v}, 1).
\]

Therefore, since $\hat{e}_y(G)$ is independent of $u$, for any $y \in \mathbb{R}$,

either \quad $|G(u) - \hat{e}_y(G)| \geq \frac{\varepsilon}{2}$ \quad $\forall u \in [\bar{v}, 1)$,

or \quad $|G(u) - \hat{e}_y(G)| \geq \frac{\varepsilon}{2}$ \quad $\forall u \in (0, \bar{u}]$.

Employing this in (4.10), one finds

\[
B(G) \geq \int_{\mathbb{R}} \min \left\{ \int_{[\bar{v},1]} \frac{\varepsilon^2}{4} h_d(y, G(u)) \, du, \int_{[0,\bar{u}]} \frac{\varepsilon^2}{4} h_d(y, G(u)) \, du \right\} \, dy > 0.
\]

Consequently, we have

**Lemma 4.4.3.** If $G$ is constant, then $B(G) = 0$; otherwise, $B(G) > 0$. 
At this point, one knows that any \( G \) that minimizes \( B \) is piecewise constant, and that one may restrict the search for minima to \( G^a \). One also knows that constant functions yield the minimum, with the constant function, \( G_0(u) \equiv 0 \) being the minimizer within \( G^a \). We have not yet shown that there do not exist other local minima. We briefly indicate this result; a sketch of the proof appears in the appendix.

**Theorem 4.4.4.** Assume (A.1). Neglecting the absolute minimizer, \( \bar{G}_0(u) \equiv 0 \), there are no other local minimizers of \( B \) over \( G^a \).

The above simple results indicate something of the structure of the optimization problem. One desires to minimize the sum of a quadratic, \( A \), with minimum at \( G = G_W \), and a functional, \( B \), with the somewhat odd structure indicated here. This interplay is what leads to the variety of solutions one finds over the parameter space, where the “signaling” \( \xi \) (cf., [26]) are those where the \( B \) component plays a more significant role than the cases where the solution looks closer to a normal random variable. Note that although neither \( A \) nor \( B \) have local minima other than their global minima, this does not imply that \( \bar{J} = k_0 A + B \) does not possess extraneous local minima.

### 4.5 Necessary Conditions and Solution Form

We remark that existence of a minimizer is already known (cf., [56, 55]). Here, we examine what can be learned from some necessary conditions which are more than first-order necessary conditions. Again, we remark that the use of the space of square-integrable quantile functions as the space over which to do optimization appears to be quite helpful in generation of necessary conditions, as \( G \) is a convex cone in \( L_2(0,1) \). In particular, the calculus of variations arguments are more easily formulated in a Hilbert space such as \( L_2(0,1) \), and convex cones are well-understood in this context (cf., [15]).

The section will culminate with the necessary conditions, as indicated in Theorem 4.5.7 and Proposition 4.5.8. However, there are some interesting results regarding the structure of the minimizer that are obtained along the way (Lemma 4.5.3 and Theorem 4.5.6).

Let \( \mathcal{L} \doteq L_2(\mathbb{R}; \mathbb{R}) \). We define \( \bar{B} : \mathcal{G} \times \mathcal{L} \to \mathbb{R} \cup \{ +\infty \} \) by

\[
\bar{B}(G,e) \doteq \int_{(0,1)} \int_{\mathbb{R}} |G(u) - e(y)|^2 h_d(G(u),y) \, dy \, du,
\]

where we note that here we are not restricting \( e(y) \) to \( \hat{e}_y \) (in contrast to (4.10)). Also let

\[
\bar{J}(G,e) \doteq k_0 A(G) + \bar{B}(G,e).
\]
Remark 4.5.1. Note that $\tilde{B}$ and $\tilde{J}$ are convex quadratic in $e$. Of course, $A$ is convex quadratic in $G$. Lastly, letting $\tilde{B}': \mathcal{F} \times \mathcal{L} \to \mathbb{R}$ be given by

$$\tilde{B}'(F,e) = \tilde{B}(I[F],e) = \int_{\mathbb{R}} \int_{\mathbb{R}} |x - e(y)|^2 h_d(x,y) dF(x) dy,$$

where the last equality follows by (4.10), and we see that $\tilde{B}'$ is linear in $F$.

For clarity, we remind the reader that

$$V(c,d,k_0) = \inf_{G \in \mathcal{G}} \tilde{J}(G) = \inf_{G \in \mathcal{G}} [k_0 A(G) + B(G)],$$

and recalling that $\hat{e}_y = \mathbb{E}\{X|Y = y\}$ is the optimal mean square estimator, this becomes

$$= \inf_{G \in \mathcal{G}}\min_{e \in \mathcal{L}} [k_0 A(G) + \tilde{B}(G,e)]. \tag{4.24}$$

We seek necessary conditions for the minimum in (4.24). As the minimum over $e \in \mathcal{L}$ is well-understood, we immediately have necessary condition (with change of variables given by (4.9))

$$e(y) = \hat{e}_y = \frac{\int_{(0,1)} G(u) h_d(G(u),y) du}{\int_{(0,1)} h_d(G(u),y) du}, \tag{4.25}$$

for a.e. $y \in \mathbb{R}$.

Next, we begin determination of necessary conditions on $G$. Let $G \in \mathcal{G}$ and $g \in L_2(0,1)$, and consider $G + \delta g \in L_2(0,1)$ for $\delta \in \mathbb{R}$ such that $G + \delta g$ remains in $\mathcal{G}$. We have

$$\tilde{J}(G + \delta g,e) - \tilde{J}(G,e)$$

$$= k_0 \int_{(0,1)} [G(u) + \delta g(u) - G_W(u)]^2 - [G(u) - G_W(u)]^2 du$$

$$+ \int_{(0,1)} \int_{\mathbb{R}} \left\{ [G(u) + \delta g(u) - e(y)]^2 - [G(u) - e(y)]^2 \right\} \cdot h_d(y,G(u) + \delta g(u)) dy du$$

$$+ \int_{(0,1)} \int_{\mathbb{R}} [G(u) - e(y)]^2 [h_d(y,G(u) + \delta g(u)) - h_d(y,G(u))] dy du$$

$$= \delta \int_{(0,1)} \left\{ 2k_0(G(u) - G_W(u)) + \int_{\mathbb{R}} \left[ 2(G(u) - e(y)) + (G(u) - e(y))^2 \left( \frac{y - G(u)}{d} \right) \right] \cdot h_d(y,G(u)) dy \right\} g(u) du + O(\delta^2)$$

$$= \delta \int_{(0,1)} \left\{ 2k_0(G(u) - G_W(u)) + \int_{\mathbb{R}} \left[ 2(G(u) - e(y)) + (G(u) - e(y))^2 \left( \frac{y - G(u)}{d} \right) \right] \cdot h_d(y,G(u)) dy \right\} \int_{(0,u)} dg(v) du + O(\delta^2).$$
Consequently, a necessary condition is
\[
0 \leq \delta \int_{(0,1)} \left\{ 2k_0(G(u) - G_W(u)) + \int_{\mathbb{R}} \left[ 2(G(u) - e(y)) + (G(u) - e(y))^2 \left( \frac{y - G(u)}{d} \right) \right] \cdot h_d(y, G(u)) \, dy \right\} \int_{(0,u)} d(g(v)) \, du,
\]
for all \( \delta \in \mathbb{R} \) and \( g \in \mathcal{G} \) such that \( G + \delta g \in \mathcal{G} \).

The right-hand side of (4.26) is a linear functional on the set of Borel measures on \((0,1)\), and we seek a more standard form. Consider a general functional with form (4.26) given by
\[
L(g) = \int_{(0,1)} \left[ H(u) \int_{(0,u)} d(g(v)) \right] du,
\]
for appropriate \( H \). Applying integration by parts yields
\[
L(g) = \left[ \int_{(0,1)} H(v) \, dv \right] \left[ \int_{(0,1)} d(g(u)) \right] - \int_{(0,1)} \int_{(0,u)} H(v) \, dv \, d(g(u))
= \int_{(0,1)} \left[ \int_{(u,1)} H(v) \, dv \right] d(g(u)).
\]
Employing (4.27) in (4.26), one obtains
\[
0 \leq \delta \int_{(0,1)} \left\{ \int_{(u,1)} \left[ 2k_0(G(v) - G_W(v)) + \int_{\mathbb{R}} \left[ 2(G(v) - e(y)) + (G(v) - e(y))^2 \left( \frac{y - G(v)}{d} \right) \right] \cdot h_d(y, G(v)) \, dy \right\] \, dv \right\} \, du,
\]
for all \( \delta \in \mathbb{R} \) and \( g \in \mathcal{G} \) such that \( G + \delta g \in \mathcal{G} \). Now, for all \( \delta > 0 \) and all \( g \in \mathcal{G} \), one has \( G + \delta g \in \mathcal{G} \). Consequently, (4.28) implies one must have
\[
0 \leq \Lambda(u) \doteq \int_{(u,1)} \left[ 2k_0(G(v) - G_W(v)) + \int_{\mathbb{R}} \left[ 2(G(v) - e(y)) + (G(v) - e(y))^2 \left( \frac{y - G(v)}{d} \right) \right] \cdot h_d(y, G(v)) \, dy \right] \, dv,
\]
for all \( u \in (0,1) \). We indicate more about the optimal solution.

**Lemma 4.5.2.** If \( G \in \mathcal{G} \) is optimal, then \( \lim_{u \uparrow 1} G(u) = +\infty \) and \( \lim_{u \downarrow 0} G(u) = -\infty \).
Proof. Suppose there exists $D < \infty$ such that $|G(u)| \leq D$ for all $u \in (0, 1)$. Then, one easily sees that there exists $D_1 < \infty$ such that $\Lambda(u) \leq D_1(1 - u) - \int_{(u, 1)} G_W(v) \, dv < 0$ for $u$ sufficiently close to one, which violates necessary condition (4.29). The lower limit condition then follows from Lemma 4.4.1.

We remark that the above lemma implies that an optimal $G$ cannot correspond to a density, $f$, with compact support. Also, as the distribution is monotonically increasing, by a standard result, it can have at most a countable number of discontinuities. Consequently, one trivially has:

**Lemma 4.5.3.** $G$ can have at most a countable number of intervals on which it is constant.

Lemma 4.5.3 will be subsumed by the stronger result, Theorem 4.5.6, below. However, it is worth noting that Lemma 4.5.3 is obtained without the substantial additional machinery that will be needed for Theorem 4.5.6.

Now, for compactness of notation, let

$$T^1(\gamma, \gamma_W) = T^1(\gamma, \gamma_W; e) = 2k_0(\gamma - \gamma_W) + \int_{\mathbb{R}} \left[ 2(\gamma - e(y)) + (\gamma - e(y))^2 \left( \frac{y - \gamma}{d} \right) \right] h_d(y, \gamma) \, dy.$$  

**Theorem 4.5.4.** Fix any $e \in \mathcal{L}$, and suppose $G \in \mathcal{G}$ minimizes $\tilde{J}(\cdot, e)$. Then $G$ satisfies

$$0 = \lambda(u) = \lambda(u; e) = T^1(G(u), G_W(u); e)$$

$$\triangleq 2k_0(G(u) - G_W(u)) + \int_{\mathbb{R}} \left[ 2(G(u) - e(y)) + (G(u) - e(y))^2 \left( \frac{y - G(u)}{d} \right) \right] h_d(y, G(u)) \, dy,$$

for all $u \in (0, 1)$.

Proof. Fix $e \in \mathcal{L}$, and suppose $G \in \mathcal{G}$ minimizes $\tilde{J}(\cdot, e)$. The proof proceeds in multiple steps.

Step 1.0 Obtaining an integral necessary condition: Note that (4.28) holds for all $\delta > 0$ and all $g$ such that $dg$ is a finite, positive measure on $(0, 1)$. Taking $dg$ concentrated at some $u \in (0, 1)$ in (4.28), we find that a necessary condition is

$$0 \leq \int_{(u, 1)} T^1(G(v), G_W(v)) \, dv \quad \forall u \in (0, 1).$$  

(4.31)
Alternatively, let $0 < u_1 < u_2 < 1$, and for $u \in (0, 1)$, let
\[
 g^{G}_{u_1, u_2}(u) = \begin{cases} 
 0 & \text{if } u \in (0, u_1) \\
 G(u) - G(u_1) & \text{if } u \in [u_1, u_2) \\
 G(u_2) - G(u_1) & \text{if } u \in [u_2, 1),
\end{cases}
\] (4.32)

where we note
\[
 G + \delta g^{G}_{u_1, u_2} \in G \quad \forall \delta \in [-1, \infty).
\]

Consequently, by (4.28) we have
\[
 0 \leq \delta \int_{(0, 1)} \int_{(u, 1)} T^1(G(v), G_W(v)) \, dv \, dg^{G}_{u_1, u_2}(u)
\]
for all $\delta \in [-1, \infty)$, which implies that a necessary condition is
\[
 0 = \int_{(0, 1)} \int_{(u, 1)} T^1(G(v), G_W(v)) \, dv \, dg^{G}_{u_1, u_2}(u) 
\] (4.33)
for all $0 < u_1 < u_2 < 1$.

**Step 2.0 Obtaining a point-wise in $u$ necessary condition**: Now, by Lemma 4.5.3, there exists a countable sequence of points of discontinuity of $G$, say $\{u^d_{k, k} \}_{k \in \mathbb{N}}$. From this and the right-continuity of $G$, one sees that $(0, 1) = \bigcup_{k \in \mathbb{N}} I_k$, where the $I_k$ have forms $I_k = (a_k, b_k]$ and $I_k = [a_k, b_k]$, and in the special case of $b_k = 1$, either $I_k = (a_k, b_k) = (a_k, 1)$ or $I_k = [a_k, b_k) = [a_k, 1)$. For simplicity of discussion, we henceforth mainly ignore the special case, as the analysis is essentially unchanged. Here (excepting of course for the case $b_k = 1$), we have $b_k = u^d_{k, k}$ for all $k$. (It may be worth remarking that the case $I_k = \{b_k\} = \{u^d_{k, k}\}$ is not possible due to the right-continuity.)

**Step 2.1 Proving $\lambda(u) \leq 0$**: Let $d \in (0, 1)$. Then $d \in I_k$ for some $k \in \mathbb{N}$. Suppose
\[
 \lambda(d) = T^1(G(d), G_W(d)) = \varepsilon > 0.
\] (4.34)

Let
\[
 \varepsilon = \inf\{ u \in (a_k, d] | T^1((G(v), G_W(v)) > \varepsilon / 2 
\] (4.35)
for all $v \in (u, d]$, and
\[
 d = \inf\{ u \in (a_k, d] | G(v) = G(d) \forall v \in (u, d]\}.
\] (4.36)

Suppose $d \leq \varepsilon$, and let $\alpha = G(d)$. Then $G(v) = \alpha$ for all $v \in (d, d]$, and note that by the fact that $G_W$ is strictly increasing, $T^1(G(v), G_W(v)) = T^1(\alpha, G_W(v))$ is strictly
decreasing on $(d, d]$. By (4.34) and the assumption that $d < c$, $T^1(G(c), G_W(c)) > \varepsilon$, which, combined with the continuity of $G$ on $I_k$, contradicts (4.34). Therefore,

$$c < d \leq d.$$  \hfill (4.37)

Then, by (4.33)

$$0 = \int_{(0,1)} \int_{(u,1)} T^1(G(v), G_W(v)) dv \, dg_{E,d}^G(u).$$  \hfill (4.38)

By (4.31), $0 \leq \int_{(d,1)} T^1(G(v), G_W(v)) dv$. Combining this with (4.35), we see that

$$\int_{(u,1)} T^1(G(v), G_W(v)) dv > \frac{\varepsilon}{2}(d - u), \quad \forall u \in [c, d],$$

which implies

$$\int_{(0,1)} \int_{(u,1)} T^1(G(v), G_W(v)) dv \, dg_{E,d}^G(u) > \frac{\varepsilon}{2} \int_{(0,1)} (d - u) \, dg_{E,d}^G(u)$$

$$= \frac{\varepsilon}{2} \int_{(c,d)} (d - u) \, dg_{E,d}^G(u).$$  \hfill (4.39)

Now, note that by (4.35)–(4.37), $G(c) < G(d) = G(d')$. Therefore, by continuity, there exists $d'' \in (c, d)$ such that $G(c) < G(d'') = [G(c) + G(d')]/2$. Using this in (4.39) yields

$$\int_{(0,1)} \int_{(u,1)} T^1(G(v), G_W(v)) dv \, dg_{E,d}^G(u) > \frac{\varepsilon}{2} \int_{(c,d')} (d - d'') \, dg_{E,d}^G(u)$$

$$\geq \frac{\varepsilon}{2} \int_{(c,d')} (d - d'') \, dg_{E,d}^G(u) = \frac{\varepsilon}{2} (d - d')(G(d') - G(c)) > 0.$$  \hfill (4.40)

However, (4.38) and (4.40) form a contradiction, and therefore we must have

$$\lambda(u) = T^1(G(u), G_W(u)) \leq 0 \quad \forall u \in (0,1).$$  \hfill (4.41)

Step 2.2 Proving $\lambda(u) \geq 0$: Suppose $T^1(G(d), G_W(d)) = -\varepsilon < 0$ for some $d \in (0,1)$. Again $d_k \in I_k$ for some $k$. By continuity, there exists $\xi < d$ such that

$$T^1(G(u), G_W(u)) < \frac{-\varepsilon}{2} \quad \forall u \in (\xi, d).$$  \hfill (4.42)

Combining (4.41) and (4.42), we have

$$\int_{(\xi,1)} T^1(G(v), G_W(v)) dv = \int_{(\xi,d]} T^1(G(v), G_W(v)) dv + \int_{(d,1)} T^1(G(v), G_W(v)) dv$$

$$< \frac{-\varepsilon}{2} (d - \xi) < 0,$$

which contradicts (4.31). Consequently, using (4.41), $\lambda(u) = 0$ for all $u \in (0,1)$. \hfill □
We now have the first-order necessary condition for the minimization of $\tilde{J}(G,e)$ with respect to the direction $G$. As for the direction, $e$, we simply recall condition (4.25). So far, we have the following.

**Corollary 4.5.5.** Suppose $G,e$ minimize $\tilde{J}$. Then they satisfy (4.25) and (4.30) (where (4.30) holds for all $u \in (0,1)$).

The following is immediate from Theorem 2 of [56]. Here, we also include a direct proof which follows immediately from the necessary condition of Theorem 4.5.4.

**Theorem 4.5.6.** If $G \in \mathcal{G}$ is optimal, it does not have any segments where it is constant.

**Proof.** We consider only the interval $[1/2,1)$, and note that one may use the antisymmetry for $u \in (0,1/2)$. Suppose there exists $(d, \bar{d}] \subset (1/2,1)$ and $\alpha \in \mathbb{R}$ such that $G(u) = \alpha$ for all $u \in (d, \bar{d}]$. Then, $\lambda(u) = T^1(\alpha, G_W(u))$ is strictly decreasing on $(d, \bar{d}]$, which contradicts the necessary condition of Theorem 4.5.4.

It is worth remarking that the necessary conditions in Corollary 4.5.5 are stronger than first-order necessary conditions, as we know that $\hat{e}_y$ is not only a critical point, but also the unique minimizer. Still we can go further rather easily. Now, for any $u \in (0,1)$ and $e \in \mathcal{L}$, let

$$\Gamma(u,e) = \{ \gamma \in \mathbb{R} \mid T^1(\gamma, G_W(u); e) = 0 \}. \quad (4.43)$$

Comparison with (4.30) implies that our necessary conditions are now $G(u) \in \Gamma(u,e)$ for all $u \in (0,1)$ and (4.25).

In order to further guarantee that the points correspond to minima rather than only critical points, we must have

$$0 \leq \frac{d}{d\gamma} T^1(\gamma, G_W(u); e)$$

$$= 2k_0 + \int_{\mathbb{R}} \left\{ \left[ 2 + 2(\gamma - e(y)) \left( \frac{y - \gamma}{d} \right) - \frac{(\gamma - e(y))^2}{d} \right] + \left[ 2(\gamma - e(y)) + \frac{y - \gamma}{d} (\gamma - e(y))^2 \right] \frac{y - \gamma}{d} \right\} h_d(y, \gamma) \, dy. \quad (4.44)$$

Obviously, if (4.44) did not hold at $u$, then a small adjustment of $G(u)$ would lower $\tilde{J}$. Consequently, we take the subset of $\Gamma(u,e)$ given by

$$\hat{\Gamma}(u,e) = \{ \gamma \in \Gamma(u,e) \mid (4.44) \text{ is satisfied} \}. \quad (4.45)$$
One expects that at each \((u, e)\), \(\hat{\Gamma}(u, e)\) will be a finite set of points in \((u, 1)\), although we do not attempt a proof. Lastly, we select the minimizing element of \(\hat{\Gamma}(u, e)\) for each \(u \in (0, 1)\). That is, we have condition

\[
G(u) \in \arg\min_{\gamma \in \hat{\Gamma}(u, e)} \left\{ k_0(\gamma - G_W(u))^2 + \int_R (\gamma - e(y))^2 h_d(y, \gamma) \, dy \right\} \quad \forall u \in (0, 1), \tag{4.46}
\]

where we note that existence of the minimizer follows from continuity and compactness.

We have arrived at our final set of necessary conditions:

**Theorem 4.5.7.** Suppose \(G, e\) minimize \(\tilde{J}\). Then they satisfy (4.25) and (4.46).

These conditions are beyond first-order necessary conditions (although still not sufficient), as is indicated in the following result.

**Proposition 4.5.8.** Suppose \((\bar{G}, \bar{e}) \in G \times L\) are such that \(\bar{G}\) satisfies (4.46) with \(e = \bar{e}\). Suppose there exists \(A \subseteq (0, 1), \mu_L(A) > 0\) where \(\mu_L\) denotes Lebesgue measure, such that \(G(\bar{u}) \notin \hat{\Gamma}(\bar{u}, \bar{e}(\cdot))\) for all \(\bar{u} \in A\). Then \(\tilde{J}(G, \bar{e}) > \tilde{J}(\bar{G}, \bar{e})\). Alternatively, suppose \(\bar{e}\) satisfies (4.25) with \(G = \bar{G}\). Then, for any other \(e \in L\), \(\tilde{J}(\bar{G}, e) \geq \tilde{J}(\bar{G}, \bar{e})\).

**Proof.** By the definition of \(\hat{\Gamma}(u, \bar{e})\), one finds that

\[
T^0(G(u), G_W(u); \bar{e}) > T^0(\bar{G}(u), G_W(u); \bar{e}) \quad \forall u \in A,
\]

where

\[
T^0(\gamma, \gamma_W; e) = k_0(\gamma - \gamma_W)^2 + \int_R (\gamma - e(y))^2 h_d(y, \gamma) \, dy.
\]

(It may be helpful to note that \(T^1(\gamma, \gamma_W; e) = \frac{d}{d\gamma}T^0(\gamma, \gamma_W; e)\).) Let

\[
D_n = \{u \in A \mid T^0(G(u), G_W(u); \bar{e}) > T^0(\bar{G}(u), G_W(u); \bar{e}) + 1/n\},
\]

for all \(n \in \mathbb{N}\). Then, \(D_n \subseteq D_{n+1}\) for all \(n\), and \(\bigcup_{n \in \mathbb{N}} D_n = A\). Consequently (cf., [52]), there exists \(\hat{n} \in \mathbb{N}\) such that

\[
\mu_L(D_n) \geq \mu_L(A)/2. \tag{4.47}
\]

Now,

\[
\tilde{J}(G, \bar{e}) - \tilde{J}(\bar{G}, \bar{e}) = \int_{(0, 1)} T^0(G(u), G_W(u); \bar{e}) \, du - T^0(\bar{G}(u), G_W(u); \bar{e}) \, du
\]

\[
= \int_A T^0(G(u), G_W(u); \bar{e}) - T^0(\bar{G}(u), G_W(u); \bar{e}) \, du
\]

\[
\geq \int_{D_{\hat{n}}} (1/\hat{n}) \, du \geq \frac{\mu_L(A)}{2\hat{n}},
\]
where the final inequality follows from (4.47). That completes the proof of the first assertion. The second is immediate from (4.8)–(4.9), and standard results regarding the conditional expectation (cf., [55]).

4.6 Monotonic Method

The conditions developed in Section 4.5 allow one to develop a corresponding numerical method. This method bears some resemblance to a policy iteration algorithm. Each step consists of two parts. Recall that a functional, $G$, corresponds to a set of controls, $ζ$. The minimization part of the step consists of computing an improved control policy, in the form of a $G$ determined from (4.46). Given such a $G$, the second part of the step consists of solving for a new observation-conditioned estimate, $\hat{e} = \hat{e}(G)$, from (4.9) (or equivalent).

We now describe the algorithm more completely. Suppose we have iterate $(G^n, e^n) \in G \times L$, with associated cost $V^n = \tilde{J}(G^n, e^n)$. For the first half of the next step, let $G^{n+1}$ be given by (4.46) with $e = e^n$, and $V_{1}^{n+1} = \tilde{J}(G^{n+1}, e^n)$. Then, for the second half, let $e^{n+1}$ be given by (4.25) with $G = G^{n+1}$, and $V^{n+1} = \tilde{J}(G^{n+1}, e^{n+1})$. By Proposition 4.5.8, $V^{n+1} \leq V_{1}^{n+1} \leq V^n$.

Consequently, the algorithm results in a sequence of iterates with monotonically decreasing cost.

Also, if $G^{n+1} = G^n$, then we will have $e^{n+1} = e^n$. Then, $G^{n+1}$ satisfies (4.46) with $e = e^{n+1}$, and $e^{n+1}$ satisfies (4.25) with $G = G^{n+1}$. Therefore, $(G^{n+1}, e^{n+1})$ satisfies the necessary conditions (4.25),(4.46). We summarize this as:

**Theorem 4.6.1.** The cost, $\tilde{J}(G^n, e^n)$ for a sequence of iterates generated by the above algorithm is monotonically decreasing. Further, if $G^{n+1} = G^n$, then the pair $(G^{n+1}, e^{n+1})$ satisfies the necessary conditions (4.25),(4.46), and $(G^{n+m}, e^{n+m}) = (G^n, e^n)$ for all $m \geq 0$.

**Remark 4.6.2.** More generally, the authors expect that if there exist $(\bar{G}, \bar{e}) \in G \times L$ such that $(G^n, e^n) \rightarrow (\bar{G}, \bar{e})$ in say an $L_2$ sense, for some sequence (or subsequence) of iterates, then the pair $(\bar{G}, \bar{e})$ will satisfy the necessary conditions (4.25),(4.46). However, it appears that the proof may be rather technical, requiring estimates of certain higher-order moments, and we do not obtain this.
The above algorithm is necessarily robust. It is, however, not extraordinarily fast (at least as currently encoded by the authors), with speeds roughly on par with that of a partially heuristic gradient descent algorithm (also encoded by the authors). Efficiency and accelerations were not explored. The authors believe that development of a faster, robust approximation algorithm may be a main remaining task for the simple one-step version of this problem class. Of course, theory for time-dependent generalizations of the problem remains of interest.

4.7 Numerical Results and Comments

We include some examples which roughly indicate the classes of forms of the solutions. In all cases, \( G \) and \( e \) were each approximated at 200 grid points. (Although one could speed the algorithm by making use of the anti-symmetry, this was not done, and these grid points were uniformly distributed across the entire relevant domains.) With abuse of notation, we denote the final iterate as \((\bar{G}, \bar{e})\). The initial iterate, \( G^0 \), was taken to be \( G_W \). The top plot of Figure 4.1 depicts the approximate solution, \( \hat{G}(u) \) as a solid curve, corresponding to \( c = 1.0, d = 0.1 \) and \( k_0 = 1.0 \). The “x” markers (directly on top of the solid curve) indicate the set \( \hat{\Gamma}(u, \bar{e}) \). The bottom plot in the figure depicts the corresponding \( \bar{e}(y) \). Figures 4.2 and 4.3 depict the same information in the cases of \( c = 1.0, d = 0.1 \) and \( k_0 = 0.25 \), and \( c = 1.0, d = 0.02 \) and \( k_0 = 0.01 \), respectively. These three cases give some sense of the types of behavior that one encounters in the optimal solutions. One can see how discontinuities in the quantile-space solution (i.e., flat sections in the distribution of the optimal \( X \)) correspond to switching among the solutions in the set \( \hat{\Gamma}(u, \bar{e}) \). Also and interestingly, the algorithm converges most slowly in the third example, where the solution is essentially of simple signaling form.

We remark that one can easily see the form of the optimal distribution, \( F(x) \), from the \( \hat{G}(u) \) plots by visually flipping them around the diagonal axis. Also, note that the nearly horizontal segments of \( \hat{G} \) in Figure 4.3 are not, in fact, constant, and as discussed further above, the corresponding optimal \( F \) is not piecewise constant.

Appendix: Sketch of proof of Theorem 4.4.4

Here we find it convenient to work with the density “function” represented in terms of Dirac delta functions. By Lemma 4.4.2, it is sufficient to consider only piecewise
Figure 4.1: Example solutions: $c = 1, d = 0.1, k_0 = 1$  

Figure 4.2: Example solutions: $c = 1, d = 0.1, k_0 = 0.25$  

Figure 4.3: Example solutions: $c = 1, d = 0.02, k_0 = 0.01$
constant \(G\), and by the monotonicity of \(G\), there can be at most a countably infinite number of such segments. We break the proof into several cases.

1.0 A finite number of constant segments: First we consider the case where there are only a finite number of masses, i.e., where \(G\) is piecewise constant with only a finite number of segments. Let \(S^N = \{\lambda \in [0,1]^N \mid \sum_{k=1}^N \lambda_k = 1\}\). Let the density be

\[
f^1(x) = \sum_{k=1}^N \lambda_k \delta_{x_k}(x),
\]

where \(\lambda \in S^N\), \(x_k < x_{k+1}\) for all \(k\), and \(\delta_z(x)\) denotes the Dirac delta function with mass one at \(x = z\). Let the corresponding distribution and quantile functions be denoted by \(F^1\) and \(G^1\), respectively. We will show that one can construct a path from \(G^1\) to the minimizer, such that \(B\) is monotonically decreasing along the path. Recall from (4.10) that

\[
B(G^1) = \sum_{k=1}^N \int_{\mathbb{R}} [\hat{e}_y(G^1) - x_k]^2 h_d(x_k, y) \, dy \, \lambda_k \triangleq \tau_k \lambda_k.
\]

Let \(\tilde{k} \in \arg\min_{k \in 1, N} \{\tau_k\}\) and \(m_{\tilde{k}} = \sum_{k \neq \tilde{k}} \lambda_k\), and let \(\hat{\lambda} : [0, m_{\tilde{k}}] \to S^N\) be given by

\[
\hat{\lambda}_k(\omega) = \begin{cases} 
\lambda_k + \omega & \text{if } k = \tilde{k}, \\
\lambda_k - (\lambda_k/m_{\tilde{k}})\omega & \text{otherwise}.
\end{cases}
\]

Let \([\hat{f}(\omega)](x) = \sum_{k=1}^N \hat{\lambda}_k(\omega) \delta_{x_k}(x)\), and let \([\hat{G}(\omega)](\cdot) : (0, 1) \to \mathbb{R}\) for all \(\omega \in [0, m_{\tilde{k}}]\) denote the corresponding quantile function. That is, \(\hat{G} : [0, m_{\tilde{k}}] \to G\), and we further remark that \(\hat{G}(\omega) = \mathcal{I}[\hat{F}(\hat{\lambda}(\omega))]\), where for \(\lambda \in S^N\), \([\hat{F}(\lambda)](x) = \sum_{k=1}^N \lambda_k H_{x_k}(x)\) for all \(x \in \mathbb{R}\), and the \(H_{x_k}\) denote the appropriate right-continuous heaviside functions. Let \(\overline{B}(\omega) \triangleq B(\hat{G}(\omega))\). Then,

\[
\overline{B}(\omega) = \sum_{k=1}^N \int_{\mathbb{R}} [\hat{e}_y(\hat{\lambda}(\omega)) - x_k]^2 h_d(x_k, y) \, dy \, \hat{\lambda}_k(\omega)
\]

\[
\triangleq \sum_{k=1}^N \hat{e}_k(\hat{\lambda}(\omega)) \hat{\lambda}_k(\omega),
\]

where for \(\lambda \in S^N\), \(\hat{e}_y(\lambda) \triangleq \hat{e}_y(\mathcal{I}[\hat{F}(\lambda)])\), and hence \(\hat{e}_y(\hat{\lambda}(\omega)) \triangleq \hat{e}_y(\hat{G}(\omega))\). Differentiating, and using (4.19) and (4.20), we have

\[
\frac{dB}{d\omega} = \sum_{k=1}^N \frac{\partial B}{\partial \lambda_k} \frac{d\lambda_k}{d\omega} + \int_{\mathbb{R}} \Delta^1(y, \hat{G}(\omega)) \sum_{j=1}^N \frac{d\lambda_j}{d\omega} \frac{d\hat{\lambda}_j}{d\omega} \, dy
\]

\[
= \sum_{k=1}^N \frac{\partial B}{\partial \hat{\lambda}_k} \frac{d\hat{\lambda}_k}{d\omega} = \sum_{k=1}^N \hat{c}_k(\hat{\lambda}(\omega)) \frac{d\hat{\lambda}_k}{d\omega} = \hat{c}(\hat{\lambda}(\omega)) \sum_{j \neq k} \frac{\lambda_j}{m_{\tilde{k}}} \hat{c}_j(\hat{\lambda}(\omega)).
\]

(4.48)
Now, note that letting $\tau_k : S^N \times \mathbb{R} \to \mathbb{R}$ be given by $\tau_k(\lambda, y) = (\tilde{c}_y(\lambda) - x_k)h_d(x_k, y)$ and $K_0(\lambda, y) = \sum_{i=1}^{N} h_d(x_i, y)\lambda_i$, we have

$$\frac{d\tilde{c}_k}{d\lambda_j} = \int_{\mathbb{R}} \frac{-2}{K_0(\lambda, y)} \tau_k(\lambda, y) \tau_j(\lambda, y) dy.$$ 

Consequently, one finds

$$\frac{d}{d\omega} [\tilde{c}_k(\tilde{\lambda}(\omega)) - \sum_{j \neq k} \frac{\lambda_j}{m_k} \tilde{c}_j(\tilde{\lambda}(\omega))] = \sum_{k=1}^{N} \frac{d}{d\lambda_k} [\tilde{c}_k - \sum_{j \neq k} \frac{\lambda_j}{m_k} \tilde{c}_j] d\tilde{\lambda}_k$$

$$= \int_{\mathbb{R}} \frac{-2}{K_0(\lambda, y)} \left[ [\tau_k(\tilde{\lambda}, y)]^2 - 2 \sum_{j \neq k} \tau_k(\tilde{\lambda}, y) \tau_j(\tilde{\lambda}, y) \frac{\lambda_j}{m_k} \right] dy$$

$$+ \sum_{j \neq k \neq k} \tau_j(\tilde{\lambda}, y) \tau_k(\tilde{\lambda}, y) \frac{\lambda_k}{m_k} \frac{\lambda_j}{m_k} dy \leq 0.$$ 

Of course, this implies that for all $\omega \in [0, m_k]$

$$\tilde{c}_k(\tilde{\lambda}(\omega)) - \sum_{j \neq k} \frac{\lambda_j}{m_k} \tilde{c}_j(\tilde{\lambda}(\omega)) \leq \tilde{c}_k - \sum_{j \neq k} \frac{\lambda_j}{m_k} c_j \leq c_k \left[ 1 - \sum_{j \neq k} \frac{\lambda_j}{m_k} \right] \leq 0, \quad (4.49)$$

where the last inequality follows by the choice of $\tilde{k}$. Combining (4.48) and (4.49), we have

$$\frac{d\mathcal{B}}{d\omega} \leq 0 \quad \forall \omega \in [0, m_k].$$

Noting that $\omega = m_k$ corresponds to a density which is a single delta-function at $x_k$, or equivalently a constant quantile function, we see that $\mathcal{B}(m_k) = 0$, and by Lemma 4.4.3, this is the minimizer. This completes the proof in the case of a piecewise constant $G$ with a finite number of segments.

2.0 A countably infinite number of constant segments: Now we turn to the case where $G$ consists of a countably infinite number of constant segments. Here, there may not be a minimizing $\tilde{k}$. We will use the metric on

$$\mathcal{G}^0 \equiv \{G \in \mathcal{G} \mid G = \mathcal{I}[F], \text{ where } F \in \mathcal{F} \text{ and } F \text{ is piecewise constant with a countable number of segments}\},$$

given by $d(G^1, G^2) \equiv \|\mathcal{I}^{-1}[G^1] - \mathcal{I}^{-1}[G^2]\|_{L_{\infty}(\mathbb{R})}$, and we note that if $\tilde{f}^1, \tilde{f}^2$ denote the corresponding densities, $d(G^1, G^2) \leq \|\tilde{f}^1 - \tilde{f}^2\|_{L_1(\mathbb{R})}$. 
2.1 The subcase where not all $c_k$ are identical: First, we consider the case where there does not exist $\bar{c}$ such that $c_k = \bar{c}$ for all $k \in \mathbb{N}$. Without loss of generality,

$$f^1(x) = \sum_{k=1}^{\infty} \lambda_k \delta_{x_k}(x),$$

where $\lambda_k > 0$ for all $k$, and let $G^1$ denote the corresponding quantile function. Let $N_0 \geq \min\{k \geq 1 \mid \exists i, j \leq k \text{ s.t. } c_i \neq c_j\}$. Let $j_0 \in \arg\min_{k \leq N_0} c_k$ and $i_0 \in \arg\max_{k \leq N_0} c_k$ Let $\varepsilon = \lambda_{i_0} (c_{i_0} - c_{j_0})$. Choose $N_1 \geq N_0$ such that

$$m_{N_1} = \sum_{k=N_1+1}^{\infty} \lambda_k \leq \max \left\{ \frac{\varepsilon}{4}, \frac{1}{2}, \frac{\varepsilon}{4B(G^1)} \right\}. \quad (4.50)$$

Let $f^2(x) = \sum_{k=1}^{N_1} \lambda_k^2 \delta_{x_k}(x)$ where $\lambda_k^2 = \lambda_k/(1 - m_{N_1})$ for all $k \leq N_1$, and let $G^2$ denote the corresponding quantile function. Note that by (4.50), $\|f^1 - f^2\|_{L_1(\mathbb{R})} \leq \frac{\varepsilon}{2}$, and that consequently,

$$d(G^1, G^2) \leq \frac{\varepsilon}{2}. \quad (4.51)$$

Also,

$$B(G^2) = \sum_{k=1}^{N_1} \int_{\mathbb{R}} [\tilde{e}_y(G^2) - x_k]^2 h_d(x_k, y) dy \frac{\lambda_k}{1 - \bar{m}_{N_1}}, \quad (4.52)$$

which by the definition of $\tilde{e}_y$ and the nonnegativity of $c_k, \lambda_k$,

$$\leq \sum_{k=1}^{N_1} \int_{\mathbb{R}} [\tilde{e}_y(G^1) - x_k]^2 h_d(x_k, y) dy \frac{\lambda_k}{1 - \bar{m}_{N_1}} + \sum_{k=N_1+1}^{\infty} c_k \lambda_k$$

$$= \frac{1}{1 - \bar{m}_{N_1}} \sum_{k=1}^{N_1} c_k \lambda_k + \sum_{k=N_1+1}^{\infty} c_k \lambda_k \leq B(G^1) + \frac{\bar{m}_{N_1}}{1 - \bar{m}_{N_1}} B(G^1) < B(G^1) + \frac{\varepsilon}{2},$$

where the last inequality follows by (4.50). For this case, we redefine $\tilde{\lambda}$ as

$$\tilde{\lambda}_k(\omega) = \begin{cases} \lambda_k^2 + \omega & \text{if } k = j_0, \\ \lambda_k^2 - \omega & \text{if } k = i_0, \\ \lambda_k^2 & \text{otherwise,} \end{cases}$$

for $\omega \in [0, \lambda_{i_0}]$. Similar to above, let $[\tilde{f}(\omega)](x) = \sum_{k=1}^{N_1} \tilde{\lambda}_k(\omega) \delta_{x_k}(x)$, and let $\tilde{G}(\omega) \in G^0$ denote the corresponding quantile functions. Again as above, let $\tilde{B}(\omega) \doteq B(\tilde{G}(\omega))$.

Then,

$$\tilde{B}(\omega) = \sum_{k=1}^{N_1} \int_{\mathbb{R}} [\tilde{e}_y(\tilde{\lambda}(\omega)) - x_k]^2 h_d(x_k, y) dy \tilde{\lambda}_k(\omega) \doteq \sum_{k=1}^{N_1} \tilde{e}_k(\tilde{\lambda}(\omega)) \tilde{\lambda}_k(\omega).$$
Proceeding similarly to the case of a finite number of segments, one finds
\[
\frac{dB}{d\omega} = \tilde{c}_{j_0}(\tilde{\lambda}(\omega)) - \tilde{c}_{i_0}(\tilde{\lambda}(\omega)) \leq \tilde{c}_{j_0}(\tilde{\lambda}(0)) - \tilde{c}_{i_0}(\tilde{\lambda}(0)),
\]
for all \(\omega \in [0, \lambda_{i_0}]\). Consequently,
\[
B(\lambda_{i_0}) - B(0) \leq \tilde{c}_{j_0} - \tilde{c}_{i_0} = -\varepsilon. \tag{4.53}
\]

Also, by the definition of \(\tilde{f}\),
\[
\|\tilde{f}(\lambda_{i_0}) - \tilde{f}(0)\|_{L_1(\mathbb{R})} = 2\lambda_{i_0} = \frac{2\varepsilon}{\tilde{c}_{j_0} - \tilde{c}_{i_0}},
\]
which implies,
\[
d(\tilde{G}(\lambda_{i_0}), G_2) \leq \frac{2\varepsilon}{\tilde{c}_{j_0} - \tilde{c}_{i_0}}. \tag{4.54}
\]

Combining (4.51) and (4.54)
\[
d(\tilde{G}(\lambda_{i_0}), G_1) \leq \left[\frac{1}{2} + \frac{2}{\tilde{c}_{j_0} - \tilde{c}_{i_0}}\right]\varepsilon.
\]

On the other hand, by (4.52) and (4.53),
\[
B(\tilde{G}(\lambda_{i_0})) = B(\lambda_{i_0}) \leq B(G_1) - \frac{\varepsilon}{2}.
\]
Since this is true for all \(\varepsilon > 0\), we are finished with this case.

2.2 The subcase of identical \(c_k\): Lastly, we consider the case where \(G_1\) consists of a countably infinite number of constant segments and there exists \(\bar{c}\) such that \(c_k = \bar{c}\) for all \(k \in \mathbb{N}\). In this case, all the \(c_k\) achieve the minimum, and the proof is nearly identical to the case of a finite number of segments. Let \(S^\infty = \{\lambda \in \ell_1 \mid \lambda_k \in [0, 1] \forall k \in \mathbb{N}, \sum_{k=1}^{\infty} \lambda_k = 1\}\). Let \(m_k = \sum_{k=2}^{\infty} \lambda_k\), and let \(\tilde{\lambda} : [0, m_k] \to S^\infty\) be given by
\[
\tilde{\lambda}_k(\omega) = \begin{cases} 
\lambda_k + \omega & \text{if } k = 1, \\
\lambda_k - (\lambda_k/m_k)\omega & \text{otherwise.}
\end{cases}
\]

Let \(\tilde{f}, \tilde{F}\) and \(\tilde{G}\) be defined similarly to the finite case. Let
\[
\overline{B}(\omega) = \sum_{k=1}^{\infty} \int_{\mathbb{R}} \left[\tilde{e}_y(\tilde{\lambda}(\omega)) - x_k\right]^2 h_d(x_k, y) dy \tilde{\lambda}_k(\omega) \equiv \sum_{k=1}^{\infty} \tilde{c}_k(\tilde{\lambda}(\omega)) \tilde{\lambda}_k(\omega),
\]
where for \(\lambda \in S^\infty\), \(\tilde{e}_y(\lambda) = \tilde{e}_y(\mathbb{I}[(\tilde{F}(\lambda))])\), and hence \(\tilde{e}_y(\tilde{\lambda}(\omega)) = \tilde{e}_y(\tilde{G}(\omega))\). As in the finite case, we find
\[
\frac{d\overline{B}}{d\omega} = \tilde{c}_1(\tilde{\lambda}(\omega)) - \sum_{k=2}^{\infty} \frac{\lambda_j}{m_k} \tilde{c}_k(\tilde{\lambda}(\omega)), \tag{4.55}
\]
and
\[
\frac{d}{d\omega} \left[ \tilde{c}_1(\tilde{\lambda}(\omega)) - \sum_{j=2}^{\infty} \frac{\lambda_j}{m_k} \tilde{c}_j(\tilde{\lambda}(\omega)) \right] \leq 0 \quad \forall \omega \in [0, m_k].
\]

As above, this implies that for all \( \omega \in [0, m_k] \),
\[
\tilde{c}_1(\tilde{\lambda}(\omega)) - \sum_{j=2}^{\infty} \frac{\lambda_j}{m_k} \tilde{c}_j(\tilde{\lambda}(\omega)) \leq c_1 \left[ 1 - \sum_{j=2}^{\infty} \frac{\lambda_j}{m_k} \right] \leq 0. \tag{4.56}
\]

Combining (4.55) and (4.56), we have
\[
\frac{dB}{d\omega} \leq 0 \quad \forall \omega \in [0, m_k].
\]

As in the finite case, \( \omega = m_k \) corresponds to a constant quantile function, and consequently, \( B(m_k) \) is the minimizer.

Chapter 4, in full, is a reprint of the material as it appears in Automatica, 2015. The dissertation author was the co-author of this paper.
Bibliography


